Statistical Science
2002, Vol. 17, No. 3, 256–285

Parity: Implementing the Telecommunications Act of 1996

Colin Mallows

Abstract. We discuss various technical problems that have arisen in attempting to implement the Telecommunications Act of 1996, the purpose of which was to ensure fair competition in the local telecommunications market. We treat the interpretation of the "parity" requirement, testing the parity hypothesis, the effect of correlation, disaggregation and reaggregation, "balancing," benchmarks, payment schedules and some computational problems. Also we discuss the difficulty of working in an adversarial (rather than scientific) environment.

Key words and phrases: Adversarial environment, modified t, permutation tests, correlation, disaggregation and reaggregation, balancing, benchmarks, optimal payments, Neyman–Pearson.

1. INTRODUCTION

The concept of "parity" arises when one company (company B) enters into a contract to provide services to the customers of a competitive company (company A). The concept was incorporated into the Telecommunications Act of 1996, which mandated that an incumbent local exchange carrier (ILEC), that is, an established local telephone company, company B (think Bell), must provide, on request, for a fair price, certain services (such as responding to customer requests for installation or repair) to the customers of a competitive local exchange carrier (CLEC), company A (think AT&T), where these services are "... at least equal in quality to that provided by the local exchange carrier to itself ...."

If company B is found to be in violation, various penalties can be imposed by the Federal Communications Commission (FCC) up to withholding permission to enter or remain in the long-distance market.

Clearly the situation imposes a conflict of interest on company B, which is strongly motivated to give poor service to company A's customers in the hope that they will become dissatisfied and will migrate to company B. To prevent this from happening, it is necessary to establish formal procedures to test whether parity of service is being provided and to prescribe financial penalties if noncompliance is detected. Establishing these procedures has been the topic of lengthy discussions and negotiations, supervised by the various state Public Service Commissions (PSCs) and the Federal Communications Commission. Several intriguing statistical issues have arisen, which this paper will discuss. We will not discuss the approach known as mechanism design, which economists have developed to deal with similar externality problems (see, e.g., Mas-Colell, Whinston and Green, 1995, Chapter 23).

For each instance of customer service, one or more service quality measurements (SQMs) can be recorded. These measurements can be classified into several categories. For one version of such a classification, proposed in Louisiana, see Appendix 2. In each of these categories many measures can be defined, each of which may be further subdivided according to geographic location, type of service, type of customer and so on. For example, it is recorded whether a customer's line is out of service for more than four hours. These SQMs are summarized in monthly reports. Definition of these measurements has been the topic of extensive negotiations between the ILECs (B companies) and CLECs (A companies), supervised by the various commissions. The need for extreme care was made clear

Colin Mallows is a member of Avaya Labs Research, Basking Ridge, New Jersey 07920 (e-mail: colinm@research.avayalabs.com).

by an experience in New York, where the ILEC procedure for handling a delayed installation request had the effect of the request being ignored completely once the delay exceeded a certain threshold. The result was that the CLECs and the PSC were receiving many complaints, while the ILEC reports were still showing acceptable performance. A referee has remarked that the use of thresholds to guide policy has introduced distortions in other contexts, for example, to reduce the poverty count, it is optimal for the government to give money to the people who are just below the threshold, rather than those who need it most. See the discussion under the heading "Benchmarks."

Since SQMs are not precisely predictable and their values in successive months exhibit irregular variation, it is natural (for a statistician) to regard these SQMs as realizations of random variables, X and Y, with distributions F_A and F_B for the customers of company A and company B, respectively. We are thus led to formulate the problem as that of testing the hypothesis that these two distributions are the same. A complication is that there are many different SQMs, and many of them must be disaggregated into small cells, which may number in the hundreds or even thousands. Also, the testing problem recurs in successive reporting periods (months). Measures will exhibit nonstationary behavior, in part because the ILEC management is continually modifying its procedures. There may be many incoming competitive companies: we need tests for each company separately and for the group as a whole. Finally, we must consider what schedule of penalties (called *incentives*) should apply when discrimination is detected.

There are cases where no comparable measurements on customers of company B can be identified. In this case, it is necessary to set up "benchmarks" to quantify satisfactory performance. These raise further problems.

While this formulation is natural for a statistician, it is not without its deficiencies. In reality, observations are not independent: it makes little sense to assume that the specifications remain constant throughout a calendar month, changing to a new value at midnight on the last day of the month. However, the data do exhibit irregular unpredictable variability, and the simplest way to deal with this is by way of a probabilistic model.

Further, this statistical formulation may not seem natural to a lawyer. The legal language does not mention variability, but says simply that company B's performance for company A's customers must be "at least equal in quality to" its service to its own customers. The first problem we must discuss is "What does the legal language mean?" In subsequent sections we will discuss the problem of working in an adversarial (rather than scientific) environment, testing the parity hypothesis, correlation, disaggregation and reaggregation, balancing, benchmarks, payment schedules and some computational problems. Along the way we will draw attention to several open issues. Since the purpose of this paper is solely to explain the technical issues, I have chosen not to assign authorship to the various positions that have been taken by the parties. Appendix 1 describes an elegant (but unrealistic) solution to a simplified version of the problem of defining an optimal payment function.

2. WHAT IS PARITY?

One reading of the law suggests that if comparable measurements of some service to customers of companies A and B are $X_1 \le X_2 \le \cdots \le X_n$ and $Y_1 \le Y_2 \le \cdots \le Y_n$, respectively (where large values are "bad"), then company B is in violation unless $X_i \le Y_i$ for all i. However, what if the samples are of different sizes? How do we decide whether the samples are of "at least equal quality"? One could compare the sample means, but from company A's point of view, a simple average may not be relevant, since bad service for one customer (who may decide to leave company A) cannot be balanced by good service for another (who expects good service anyway). The parties have disputed what the law means. One way this could be resolved is for the parties to agree on the financial utility of each service measurement. For example, the companies might be able to agree that the (dis)utility of a SQM of magnitude a is some function $u(a)$. Perhaps this could be calculated from (estimates of) the discounted total value of that customer, multiplied by the probability that this SQM value will cause the customer to migrate to the other company. Presumably $u(a)$ will be a rapidly growing function of a. Then we could compare $\text{ave}(u(X))$ with $\text{ave}(u(Y))$. This approach is difficult to implement even in principle, because if service is sufficiently bad, so many customers of company A may decide to leave that company A cannot continue in business; the disutility of the "last straw" customer is effectively infinite.

Some parties have suggested that parity exists whenever the application of some statistical test results in a finding of "not significant." Thus they do not distinguish between parity and a test of parity. A counter to this argument can be made by pointing out that to

choose a statistical test, we must (1) choose a statistic (Z), (2) choose a type-1 error (α) and (3) be able to find the critical value (ζ) such that when parity holds, Z will exceed ζ with probability α. If we do not distinguish between the test and the parity concept itself, it is not clear that there is any rational basis for choosing the statistic. Of course the usual statistical approach is to consider a class of alternative hypotheses and to choose the statistic to give good power for these alternatives. We will need to consider what alternatives are relevant; see the discussion under the heading "Balancing."

The most satisfying definition of parity for a statistician is that the CLEC observations and the ILEC observations should be exchangeable. (This definition could make sense even if the processes that generate the SQMs were not stationary.) However, so far this concept has not been accepted by any commissions. It is possible to explain this concept in nontechnical terms: one says that the law implies that it should not be possible to distinguish, by looking at the data, the CLEC observations from the ILEC observations. Then the concept of a permutation test can be explained by considering the special case where there is just one CLEC observation, and many (say m) ILEC observations. If exchangeability holds, the CLEC observation will be larger than all m ILEC observations with probability exactly $1/(m+1)$, so this gives us a permutation test with size $1/(m+1)$.

3. THE ADVERSARIAL ENVIRONMENT

The environment that faces the statistician involved in these problems seems not to have been dealt with. The most relevant discussion I have seen is in a paper by Mann (2000). Mann pointed out the distinction between a "testifying expert" and a "consulting expert." The former prepares a report and appears before a court (or commission) to defend it. He may be cross-examined by the opposing attorney. He may not be aware of the existence of a consulting expert, whom the attorneys may rely on for technical advice, but are reluctant to put forward as a witness. Mann (and other authors in the same volume) discussed the ethical issues that arise. One quotation must suffice:

> One question that should arise in the mind of the statistician ... is what happens if their analysis does not produce results that retaining counsel and client would like. ... the answer is that if timing permits, you will in all likelihood be replaced. To the extent that your analysis is proper and thorough, this should not be a professional concern. Regrettably, many attorneys act as if they believe they will eventually find a statistician to defend whatever position they have taken. More regrettably, they may often be correct (Mann, 2000, page 253).

Most companies do not maintain an in-house team of statistical experts, and so need to hire consultants to advise them and to testify before the commissions. My own experience has been untypical in that I worked in an established research laboratory (funded by AT&T) and had no direct responsibility for arguing the CLEC case. It has been my view (and this view was met with approval by the lawyers at AT&T) that my most productive role is to give advice to my employer regarding the statistical issues, to propose effective and ethical methods for dealing with the issues, and to provide criticism of proposals submitted by our adversaries. I tried (not with complete success) to avoid testifying as an AT&T advocate. In my view it is important that the role of statistical experts should be kept separate from that of advocates, since their effectiveness depends in large measure on whether their comments are seen as being based only on technical issues and not on the desires of their employers. On the few occasions when I had to testify, I found the experience to be much more stressful and less rewarding technically. Much more enjoyable and, I think, effective, have been appearances before staff employed by the commissions, whose role is to advise the commissioners on technical issues. In these sessions there is opportunity for interaction between the opposing experts.

Commissions are reluctant to accept arguments that cannot be found in standard references (e.g., Finkelstein and Levin, 1990; a referee remarks that almost all judges will have access to the Federal Judicial Center's Manual on Scientific Evidence, which cites several statistics texts). This causes difficulty since many of the problems that arise are not treated in such works. The use of likelihood and of Bayes' theorem [which is discussed in detail in several papers in Gastwirth (2000) in the context of DNA evidence] will not be appropriate because the ILEC will not agree to any nonzero prior probability of violation. The confrontational procedures that are employed make it very difficult for the opposing technical experts to discuss their views freely, although on at least one occasion the parties were instructed by a state commission to produce a joint statement summarizing the statistical issues and proposing resolutions. This exercise was very productive techni-

cally and resulted in resolution of almost all the outstanding points of disagreement.

4. TESTING

For the rest of this paper we accept the straightforward statistical view of the problem, so that the observations of any SQM in any month are random variables $X_1 \le X_2 \le \cdots \le X_m$ and $Y_1 \le Y_2 \le \cdots \le Y_n$ drawn from distributions F_A and F_B for the customers of company A and company B, respectively. The observations may have range $\{0, 1\}$, for example, when we are counting missed installation appointments, in which case F_A and F_B are simply binomial distributions. First we consider the problem of testing the hypothesis that the two distributions F_A and F_B are the same.

The problem has several nonstandard aspects. First, at least for some measures, company B can control (at some expense to itself) both F_A and F_B. Each company can propose a test procedure and can attempt to convince the commission, which is concerned with protecting the public interest, that its proposal is fair and reasonable. Company A's proposal will be based on incomplete knowledge of F_B and of the strategies company B might be able to employ. Company A wants the test procedure to be such that company B cannot manipulate the two distributions to company A's disadvantage without being detected. Company B's interests are exactly the opposite. Once a procedure has been accepted, it will be execute in good faith, but within the limits of the law each company will attempt to do whatever is necessary to serve its shareholders. For example, company B will always find it uneconomic to attempt to give very high quality service to all customers: there is no reason why company B cannot arrange to give poor service to some fraction of A's customers and equally poor service to the same fraction of its own customers, if by so doing it can cause migration of "good" A customers and "bad" B customers. This strategy requires company B to be able to identify good and bad customers. This may be easy; for example, business accounts are typically more profitable than residential accounts. To combat this strategy, it is necessary to disaggregate the data, comparing business A only with business B. There are other reasons for disaggregating; see below.

Another peculiarity is the nonstandard form of the alternatives that are of most concern. What company A wants to avoid is having some large fraction of its customers getting poor service, even if some other fraction receives unusually good service. Suppose that under standard conditions some SQM has (at least approximately) a Gaussian distribution. Then alternatives of high concern would include Gaussians with larger means, larger variances or both. A "slippage" alternative $[F_A(x) = F_B(x - a)]$ or a "stochastic dominance" alternative $[F_A(x) < F_B(x)$ for all $x]$ may not be appropriate, since neither considers the "increased variance" possibility.

5. SOME HYPOTHESES AND TESTS

Let us consider just one type of service and measurements taken in just one month, namely X_i, $i = 1, \ldots, m$, and Y_j, $j = 1, \ldots, n$. Suppose that the standard ILEC procedures make F_B approximately Gaussian, with mean μ_B and variance σ_B^2. A standard one-sided two-sample t-test would reject the null hypothesis $F_A = F_B$ when $Z > c$, where c is some critical value and

$$Z = (\bar{X} - \bar{Y})/S,$$

where $\bar{X}$ and $\bar{Y}$ are the sample means for observations on company A and company B, respectively, and S^2 is one of

$$S_1^2 = S_{\text{pooled}}^2\left(\frac{1}{m} + \frac{1}{n}\right),$$

$$S_2^2 = \frac{s_X^2}{m} + \frac{s_Y^2}{n},$$

where s_X^2 and s_Y^2 are the two sample variances.

According to most textbooks, S_1 is appropriate if we can assume that (under the alternative hypothesis) $\sigma_A^2 = \sigma_B^2$; otherwise, S_2 should be used. However, S_2 is used in a test of whether the means are equal, allowing the variances to be different under the null hypothesis. Our null hypothesis requires both the means and the variances to be equal, while allowing both to differ under the alternative. So in our situation, neither of these classical tests is appropriate.

The statistic Z depends on five quantities: the two sample numbers, the two sample means and the estimate of the variance of the difference between these means. Company B is highly motivated to make some fraction of the company A measurements large if it can do this without making Z large or giving poor service to its own best customers. It can achieve this by any or all of the following measures:

1. Keeping $\bar{X}$ small by making some of the company A measurements small.
2. Increasing $\bar{Y}$ (by giving poor service to some of its own customers, perhaps those it would least mind losing or whose loyalty is sure).
3. Increasing either or both of s_X and s_Y.

Company A cannot do anything about options 1 and 2, but it can remove option 3 by proposing to use, instead of S_1 or S_2, the nonstandard

$$S_3^2 = s_Y^2\left(\frac{1}{m} + \frac{1}{n}\right).$$

With this choice for S, the statistic Z still has a t-distribution under the null hypothesis (with n degrees of freedom), but now Company B can get no advantage by increasing the variance of F_A.

This asymmetric form of Student's t was suggested by Brownie, Boos and Hughes-Oliver (1990) for use in designed experiments where there is reason to expect the application of a treatment to increase both the mean and the variance of the response. They showed that in this situation the modified form is more powerful than the classical "pooled" t. This increase in power is balanced by a decrease in the chance of detecting alternatives where the variance has decreased. In the present context, this is acceptable because such alternatives do not imply that many customers of company A will receive bad service. Use of this "modified t" has been accepted by several state commissions and the FCC.

Another way to detect an increase in the variance is to employ a separate test, using the two sample variances. However, since we need to make a single decision, we have to combine this test with the test of means in some way to arrive at a composite test. While such a composite test would be effective in detecting an increase in the variance unaccompanied by a change in the mean, in the present context we judged that this would reduce the power against shift alternatives too much. Also, use of a nonparametric test (based on ranks) was judged to throw away too much power (this may have been a mistake) and also to be hard to sell to commissions and to the companies concerned. Note that the modified test described above does have some useful power for detecting variance increases even when the mean has not changed.

Podgor and Gastwirth (1994) have reviewed 14 alternative tests for this situation, including the Brownie, Boos and Hughes-Oliver test, a composite test owing to Lepage (1971) and several rank-based tests. They found that some rank tests have reasonable power against a variety of alternatives. Also of course t-tests are not robust. Is there a useful rank-based analog of our modified t?

6. PERMUTATION TESTS

Permutation tests are attractive, since (one can argue) they test exactly the right hypothesis, namely that the ILEC and CLEC observations are exchangeable, without relying on any shape assumptions. However, if thousands of tests need to be applied each month, the computations are daunting. When there are many observations, the modified t-test (see above) gives adequate accuracy. For smaller numbers, an adjustment has been developed (Balkin and Mallows, 2001, following the method of Johnson, 1978) that allows for skewness in the populations. In one application, this adjustment worked adequately down to samples of size 10. For samples smaller than this, exact calculation of the permutation test may be feasible. Of course, for counted measures there is no need for permutation calculations, since the exact hypergeometric probability can be calculated readily.

We have found that for small samples, modified t sacrifices too many degrees of freedom to be useful. In permutation calculations with small samples, it is preferable to use classical two-sample t, which cuts down the computing considerably, since for each permutation it is necessary to compute only the sum of the selected CLEC sample values.

The following questions relating to the use of permutation tests have been studied to some extent, but elude satisfactory answers.

1. If we compute both a t-test and a permutation test, and are allowed to use the result we like best, what will be the true type-1 error?
2. How far apart can the two tests be? What is the mean square difference between the p values?
3. Can useful permutation p values be approximated using some cleverly designed set of permutations?
4. The permutation test based on the standard t-statistic is equivalent to that based on $\bar{X}$. Can a useful approximation to the permutation p value be obtained from the moments of the permutation distribution of $\bar{X}$? [These moments can be obtained directly from the moments of the combined samples; formulas through degree 4 can be found in Kendall and Stuart (1963).]

7. CORRELATION

The statistical approach runs into difficulty because of the possibility of correlation. We can identify at least four kinds of correlation.

1. Correlation between different measures. This is sometimes built into the definitions of the measures themselves; for example, we may record both the average number of days it takes for some operation to be executed and the proportion of cases that take longer than 10 days. These measures are clearly not independent.
2. Correlation over time. A slow drift over several months can be interpreted either as a realization of a highly correlated but stationary process or as a deterministic drift in an independent process.
3. We can have within-measure correlation, sometimes called the *backhoe effect*. A single mechanical fault (perhaps a cable cut caused by a careless backhoe operator) can cause several customer complaints, all of which will be resolved at once. If these complaints are counted as if they were independent, their effect is inflated.
4. Records are kept in distinct geographic areas and there may be correlation between these areas, for example, due to effects of climate.

8. DISAGGREGATION AND REAGGREGATION

To avoid Simpson's paradox, it is necessary to disaggregate each SQM into homogeneous cells. Otherwise, a spurious signal of lack of parity for one class of customers (one market) might be generated or a genuine lack-of-parity effect may be canceled out. For example, data for business and residential customers may be systematically different, and if the mix of activity is different for the two companies, naive pooling could lead to an indication of apparent discrimination even if we have parity everywhere. Also pooling may lead to an important effect being swamped.

It may be necessary to disaggregate into very many cells; in one region, for the SQM order completion interval, which counts the number of days it takes to service a customer's order, it was found appropriate to define as many as 72 cells for each of 221 wire centers (roughly, a wire center corresponds to a three-digit exchange) so that there are $221 \times 72 = 15{,}912$ possible cells. Many of these cells will contain no data, but there can easily be as many as 1000 occupied cells. If we want a single statistic to summarize the performance on this SQM, we need to consider how to test at the cell level and how to aggregate into an overall statistic. The individual cells may have very small numbers of observations, so we cannot rely on normality assumptions. A referee has remarked that the problem of incorporating two tests of significance into a single overall test has arisen in other legal cases. An elementary discussion of the Environmental Protection Agency's use of both t and sign tests appears in Gastwirth (1988), and this case is also mentioned by Finkelstein and Levin (1990).

The need for care in this arena is evidenced by the fact that a technique called the *replicate variance* method, that can be found in a standard sample survey text (Wolter, 1985) and that seems at first sight to offer a straightforward approach to the aggregation problem, turns out to be completely inappropriate. In the sample survey context, one selects a large number (perhaps 200) of primary units and assembles them into 30 groups. Within each primary unit, one draws a random sample of secondary units, measures some attribute (in our case the difference between ILEC and CLEC performance) and computes weighted averages within each group. A weighted average of these attributes is an overall estimate. An estimate of the variance of this overall estimate is obtained from the dispersion of the within-group averages. Hence an overall t-statistic can be formed. Some parties have suggested that this technique should be used here, treating the primary units (here the complete set of wire centers in the geographical region) as if they had been drawn randomly from some superpopulation of wire centers and the secondary units (here the various cells within wire centers) as if they had also been drawn randomly from superpopulations. However, in our context, the wire centers were not drawn randomly; they are simply all the wire centers that exist in a geographical region. Similarly, the cells were not drawn randomly; they are the subcategories of service into which the wire centers have been disaggregated.

The replicate variance methodology assumes that all differences between wire centers and between cells within wire centers are random, and all this randomness is allowed to contribute to the final variance estimate. The result is that systematic effects, if they exist, are allowed to cancel out and are allowed to inflate the variance estimate. Numerical calculations show that if large systematic differences occur in only a few cells, the power of the final t-test can be smaller than its size. This is because both the numerator and denominator of

the t-statistic are increased, so that, in the limit, t approaches unity and the power approaches zero (provided the critical value is greater than 1). We need a different approach.

In work with data from one ILEC, it was found feasible to proceed as follows. First, we obtain for each cell a variable that should be (approximately) standard normal under the parity hypothesis. Here is a procedure for doing this.

PROCEDURE. When the number of observations is sufficiently small, perform an exact permutation test using complete enumeration of the permutation distribution. (For counted variables this reduces to the hypergeometric distribution, which can be applied exactly for all sample sizes.) For intermediate sample sizes on measured variables, perform an approximate permutation test by sampling the permutation distribution. Transform the p values obtained from these permutation tests into normal quantiles. For large sample sizes, calculate the modified t-statistic described above, find the p value from the t-distribution and convert to a normal quantile.

These calculations give us a set of statistics, one for each cell, that under the null (parity) hypothesis have approximate standard normal distributions. We call these the cell Z's. How should we aggregate them into a single statistic to measure the overall performance of the SQM? We want the method to have the following properties:

(a) The method should provide a single index, which is on a standard scale.

(b) If the entries in the cells are exactly proportional, the aggregate index should be very nearly the same as if we had not disaggregated.

(c) The contribution of each cell should depend on the numbers of ILEC and CLEC observations in that cell.

(d) As far as possible, cancellation should not be allowed to occur.

(e) The index should be a continuous function of the observations.

The motivation for requirement (d) is that, for example, the ILEC should not be able to discriminate against CLEC business customers while avoiding detection by discriminating in favor of CLEC residential customers. The motivation for requirement (e) is that we do not want the final result to depend on minor details in the data. A small change in the data should induce only a small change in the result.

One approach would be simply to count how many of the cell Z's exceed some chosen critical value, perhaps 1.96 or 3.09. This method satisfies requirements (a) and (d), but not (b), (c) or (e). Another possibility is to form a weighted average of the Z's. The weights should depend on the sample sizes. Requirement (c) will be satisfied if we take

$$\text{ave}\, Z = \sum wZ \Big/ \sqrt{\sum w^2},$$

where

$$w = 1/\sqrt{1/n_{\text{ILEC}} + 1/n_{\text{CLEC}}}.$$

This average Z statistic is on a standard normal scale (approximately) and satisfies all the requirements except (d). To meet this requirement, we perform a truncation operation in which Z is replaced by

$$Z^* = \max(0, Z)$$

and the final aggregate is adjusted appropriately:

$$Z^* = \Big(\sum wZ^* - M\Big)/V,$$

where

$$M = \sum w/\sqrt{2\pi}, \quad V = (1/2 - 1/(2\pi)) \sum w^2.$$

The final statistic Z^* is only approximately Gaussian, but if there are many cells, the approximation will be good. The exact distribution could be derived if required. Mulrow (2001a) considered adjusting this procedure to allow for the skewness of the individual Z^*'s.

An alternative method for avoiding cancellation is simply to discard cells for which $Z < 0$ and to form an adjusted weighted average of the remaining cells. This gives much greater power for detecting violations in a small number of cells at the cost of violating requirement (e). This method also has the defect that the final statistic cannot be assigned the full weight associated with all the measurements.

9. BALANCING

A subject that has been discussed at great length in several jurisdictions is that of choosing the appropriate size of the test. Appeals to the authority of textbooks have not resolved the issue. The ILECs want the type-1 error probability, α, to be small, to avoid incurring unfair penalties when they are providing parity service, while the CLECs want the power ($= 1 - \beta$, where β is the type-2 error probability) to be large to ensure that violations, if they occur, are detected with high probability. These desires are in

direct conflict. Progress toward agreement has been made by introducing a balancing concept, to which both parties have agreed, at least in some jurisdictions. The idea is that the parties should agree to define some particular nonnull hypothesis, say H_1, that represents a "substantial" degree of departure from parity, and then choose the test to equate α to β for this alternative. The choice of H_1 is clearly of critical importance and ideally would be made only after careful study of the response of customers to various degrees of violation. Such studies have not been made. It may be possible to accumulate such data when the market has become truly competitive. It is to be hoped that eventually data will be collected and analyzed, so that the choice of H_1 can be based on real experience. At present we must rely on the judgement of telephony experts.

To facilitate the choice of H_1, we define a standardized shift, δ, which measures the difference between the two distributions F_A and F_B. For measured variables, where we propose to use modified t as defined above, this is taken to be simply

$$\delta = \big(\text{mean}(F_A) - \text{mean}(F_B)\big)/\text{s.d.}(F_B).$$

Then agreement will be attainable if the parties can agree on what value to take for δ, which may be an easier task than choosing H_1 directly.

What should δ be? This is not a question that a statistician can answer, but statisticians can perform computations to aid in such judgements. Here is one such aid. Assume F_B is standard Gaussian, $F_B(\cdot) = \Phi(\cdot)$, and suppose $F_A(\cdot)$ is simply shifted by an amount δ [i.e., $F_A(\cdot) = \Phi(\cdot - \delta)$]. Consider the level of service that is enjoyed by all but a small fraction of company B's customers, say 1% of them; these customers all have SQM < 2.33. Then with a shift of δ, the proportion of company A's customers who receive service beyond 2.33 is $1 - \Phi(2.33 - \delta)$ and we can tabulate this as in Table 1.

TABLE 1
Effect of various values of δ

δ	P (> 2.33) (%)
0	1.000
0.125	1.286
0.25	1.893
0.5	3.390
1.0	9.076

For counted variables, such as missed appointments, we can use the arcsin–square-root transform to get to an approximate "normal shift" scenario. This suggests that for such SQMs we should take

$$\delta = 2\big(\arcsin(\sqrt{p_{\text{CLEC}}}) - \arcsin(\sqrt{p_{\text{ILEC}}})\big).$$

Now we can calculate what various values of delta imply. Calculations such as these can help to calibrate the level of violation described by δ and so help to judge what value of delta should be chosen.

Difficulties arise when we try to apply the balancing idea to an aggregated statistic. Suppose we have N measures, with sample sizes m_i and n_i for the ILEC and CLEC, respectively, with everything Normal, with known variances, so that we may ignore small-sample-size issues. For the ith measure we have a test statistic

$$Z_i = (\text{ILECmean}_i - \text{CLECmean}_i)/\sigma_i\sqrt{(1/m_i + 1/n_i)}.$$

We aggregate these by forming

$$Z_{\text{agg}} = \sum_i w_i Z_i \Big/ \sqrt{\sum w_i^2},$$

where $w_i = 1/\sqrt{1/m_i + 1/n_i}$ (we ignore the cancellation effect).

Let us consider only shift alternatives. Then a typical hypothesis is described by a set of shifts $\delta_i\sigma_i$, $i = 1, \ldots, N$. The statistic Z_i is then Normal with mean $-w_i\delta_i$ and variance 1, and the aggregate statistic Z_{agg} is Normal with mean

$$E(Z_{\text{agg}}) = -\sum w_i^2\delta_i \Big/ \sqrt{\sum w_i^2}$$

and variance 1.

How should we define the meaningful alternative? The difficulty is that now the alternative must be specified not by a single δ, but by a whole vector of $N\delta$'s. The power of the aggregate test depends only on $E(Z_{\text{agg}})$. One possibility is to take

$$\delta_i = \delta, \quad i = 1, \ldots, N.$$

For this alternative, the mean of Z_{agg} is $-\delta\sqrt{\sum w_i^2}$. If m_i is large for each i, then w_i is approximately $\sqrt{n_i}$ and the mean of Z_{agg} is about $-\delta\sqrt{\sum n_i}$. This is clearly the right answer for this case, since we have effectively a single set of data with sample sizes $\sum m_i$ and $\sum n_i$.

For a numerical example, suppose $N = 100$, and take $\delta = 0.1$. If each m_i is very large and each n_i is 100, we have a total sample size of 10,000: $E(Z_{\text{agg}})$ is $100\delta = 10$, ζ is about -5 and α is about 3×10^{-7}. This is not a sensible result.

Other sets of δ's are relevant. We could argue that the CLEC is hurt when any delta is positive, so a relevant alternative might be

$$\delta_1 = \delta, \quad \delta_i = 0, \quad i = 2, \ldots, N,$$

or any permutation of this. For the numerical example above, $E(Z_{\text{agg}}) = \delta_1 = 0.1$ and we get α near 0.25. This also seems unreasonable. We could get a more reasonable answer by arguing that to be meaningful, the single nonzero δ needs to be rather larger than when this is the only test that is being made. If we say it has to be N times larger, we get back to the previous result (because the average delta is again 0.1), but we could argue (arbitrarily) that δ_1 needs only to be as large as $0.1\sqrt{N}$ to be meaningful, which leads to $E(Z_{\text{agg}}) = 1$ and $\zeta = -0.72, \alpha = 0.24$.

Alternatively, we could look at cases where, say, M of the δ's are zero, while the remaining $N - M$ are all equal to some common value. All such arguments seem arbitrary and unconvincing. As yet there is no agreement on how to handle this aggregated case.

10. BENCHMARKS

When no ILEC analog exists for some CLEC measure, it is necessary to set up a benchmark level of performance. We should remark that some parties have argued (and this argument has been accepted by some commissions) that setting a benchmark does not require any probabilistic argument. Here we will continue to work within the standard statistical framework in which observed irregularities are modeled as a random process. We need to consider separately the case where an SQM is an actual measurement or merely a proportion.

10.1 Measured Variables

An example of a measured variable for which it has been determined (in Louisiana) that no sufficiently close ILEC analog exists is "firm order confirmation timeliness (mechanized only)," for which the parties have agreed that a benchmark of 95% within 4 hours is appropriate. This criterion has two components, a value (here 4 hours), which relates to a standard of performance, and a percentage (here 95%) that specifies how frequently the standard must be met.

Usually benchmarks are not based on careful study of data, but are arrived at by compromise and intuition on the part of the experts doing the formulation. It may be that any quantitative analysis will give undue weight to what are essentially arbitrary numbers.

There is a divergence of opinion as to what the benchmark value and benchmark percentage represent. One view is that the value is a limit such that any performance that fails to reach this level is unacceptable. In this view, the ILEC should be aiming at perfect performance (all cases dealt with in less than 4 hours); the benchmark percentage is set somewhat below 100% as a concession, to allow (informally) for random variation. An alternative view is that the benchmark value and percentage together set (implicitly, see below) a level of performance that is a "target" for the ILEC, and that probabilistic arguments are appropriate to evaluate where the percentage should be set. In my view, however the benchmark rule is interpreted, probabilistic arguments are relevant, because the observed data exhibit irregular variability which can only be discussed using a probability model.

10.2 Counted Variables

Some variables are not based on measurements, but simply on counts; for example, "percent of due dates missed," for which the agreed benchmark (in Louisiana) is less than 10%. For such a measure the benchmark has only one component: there is no value, only a percentage. Of course, one could formally agree to record a nonmiss as 0 and a miss as 1, so that the value would be 1, but it is not appropriate to regard this value as a limit, since an attempt to attain perfect performance would incur unreasonable expense. (The ILEC would have to assign unreasonably many workers to handle the CLEC requests.) Nevertheless, we can discuss both the measured and the counted cases at the same time by concentrating on the percentage component of the benchmark rule, assuming the value (for a measured variable) has been set using engineering judgement. This type of benchmark is reminiscent of the four-fifths rule of the U.S. EEOC (1987) which tests for evident discrimination (in employment, e.g.) against any group by checking whether the observed rate for that group is less than 4/5 of the rate for the group with the highest rate. A referee has remarked that a recent case illustrating the use of a test of significance rather than a benchmark rule is Bew v. City of Chicago (252 F. 3d 89; 7th Circuit, 2001). The effect of sample size has been discussed in the legal literature.

For both measured and counted variables, the benchmark rule claims violation if the observed proportion of failures exceeds the specified percentage. (As in our first example above, the rule may be stated in terms of nonfailures rather than failures, in which case violation

is claimed if this proportion falls short of the stated percentage.) While this rule is easy to state, it introduces challenging difficulties. The benchmark rule does not specify a null hypothesis or a type-1 error rate. Further, in many cases benchmark rules will be applied many times to different subcells of the data, and it will be necessary to aggregate the results over many cells to form an overall criterion. If all we have are the rules themselves, it is not clear how this can be done to yield an overall test with known properties.

A major weakness of the benchmark rule is that its performance depends strongly on the number of cases that enter into the observed proportion. For example, if the benchmark proportion is 95% and $n < 20$, then to avoid the violation result, there must be no failures. If we assume that these n cases are independent, so that the observed number of misses is Binomial (n, p) for some p, then to achieve $P(\text{violation}) = 0.01$ when $n = 20$, the ILEC must aim at $p = 0.99241$; when $n = 100$, the ILEC can relax to $p = 0.98185$; and when $n = 1000$, the ILEC need only achieve $p = 0.96387$. It is not clear that these requirements correspond to the intended properties of the rule, which says nothing about relaxing the standard when more cases are considered. ILECs have complained that applying the benchmark rule as stated imposes an unreasonable burden when the number of cases is small.

Here we will discuss the proportion kind of benchmark rule; the discussion applies also to the measured case once the value is decided. To discuss the properties of such a rule, we need to distinguish several quantities. First, the proportion of bad observations in the data. Let us call this proportion SF (for sample fraction) so that

$$SF = K/n = \frac{\text{number of bad sample values}}{\text{total number in the sample}}.$$

The benchmark criterion says that SF should not exceed some specified value BP, the benchmark proportion, which is usually taken to be some conventional value such as 5 or 10%. Thus if $SF \le BP$, the benchmark rule says "no violation." In the contrary case, if $SF > BP$, it says "violation."

We assume that the sample observations are random, independent and have some probability p of being bad. The benchmark rule, which uses the sample fraction SF, can be thought of being a replacement for a procedure that we would prefer to use, which would decide between violation and nonviolation according to whether $p > BP$ or $p \le BP$. Since p is not observable directly, in using the benchmark rule we are replacing p by SF and hoping for the best.

TABLE 2
Type-1 errors (α's) for the benchmark rule with $BP = 0.10$, with the null-hypothesis value of p_0 determined so that $\alpha = 0.05$ for $N = 20$ and 1000

n	$N = 20$ $p_0 = 0.042169$	$N = 1000$ $p_0 = 0.085714$
10	0.06389	0.20921
20	0.05	0.24288
100	0.00335	0.23719
200	0.00013	0.19527
1000	10^{-15}	0.05
2000	10^{-28}	0.01138

The final quantity to be considered is the probability P_v that the benchmark rule says "violation." P_v depends on n, p and BP. Once these are given, P_v is simply a binomial probability, which can take only $n + 1$ different values, and so (in general) cannot be made to be close to some target such as 0.05.

The benchmark rule specifies only BP and does not specify a null-hypothesis value of p. So we cannot state the size and power of the procedure. The following argument has been proposed for determining these quantities. Pick a reference sample size, say N, and a value for the type-1 error, α. Then determine p_0 so that

$$P(SF > BP \mid N, p_0) = \alpha.$$

The idea is that the benchmark rule implicitly defines the null-hypothesis value p_0. With p_0 in hand, we can now evaluate the type-1 error for other values of n and the power for other values of p. Taking $BP = 0.10$ and $\alpha = 0.05$, with $N = 20$ and 1000, we get the values in Table 2. We see that the value of α depends strongly on n when $n > N$ and is very small if n is much larger than N. It is much less sensitive for $n < N$ and is not monotone in n. (This effect is in addition to those that occur when $n \times BP$ is not an integer.) In using this approach, it would seem to be important to choose N near the typical values of n that occur.

11. UPDATING A BENCHMARK RULE

Once data have been collected for several months, we may want to reconsider whether the benchmark rule has been defined appropriately. If the rule has resulted in many failures, the ILEC may complain that it is too stringent and should be relaxed. (The CLEC may disagree, claiming that the ILEC's performance

needs to be improved.) If the rule has been "passed" easily every month, the CLEC may want to tighten the rule so that the ILEC is not tempted to relax its procedures. (The ILEC may disagree.) It seems clear that there can be no automatic criterion for updating the benchmark rule. Any argument for updating must depend on engineering judgement, just as when the rule was originally set up, but once data are available, those judgements can be informed by analysis of the data, which may lead to an understanding of why there are so many failures or easy passes.

12. PAYMENT SCHEDULES

One can question whether any of these testing problems are relevant to the problem facing the commissions. Once an ILEC has been found to be in violation, financial penalties must be assessed. These are of two types: first, penalties paid to an individual CLEC in recognition of damage to its operations. These are usually called Tier I payments. Second, punitive fines paid to the state treasury when a pervasive pattern of violation, affecting the whole CLEC industry, is detected. A direct attack on the problem of choosing the payment schedule seems more relevant than consideration of tests without specifying how they are to be used, although one can have sympathy for an attempt to split the problem into manageable pieces. Given a specification of an ILEC distribution, the relevant quantity for judging a proposed payment schedule is the distribution of the payment as a function of each possible CLEC distribution. This is analogous to considering the power function of a test, although here it may not be sufficient to consider only the expected payment. For an oversimplified approach to this problem, see Appendix 1.

Many payment schedules have been proposed. Their intent is to induce the ILEC management to alter its practices, and this requires that the payments be large enough to affect the ILEC's bottom line. The problem of designing a payment schedule has many complications. One of these is the choice between a per-transaction rule, where the payment is proportional to the number of CLEC customers who are hurt, and a per-measure rule, where penalties are assessed as soon as a statistically significant violation occurs, regardless of the number of customers involved. Another factor is the desirability of making payments escalate when violations are chronic, occurring month after month.

Some of the payment plans that have been proposed are extremely complicated. Here is an example, which at one time was endorsed by both sides in California proceedings. Consider a single SQM, for which we will obtain a Z-score each month. The plan involves three critical values, which we call $c1$, $c2$ and $c3$. A month in which the value of the Z-score is between $c1$ and $c2$ is called a level-1 miss, between $c2$ and $c3$ is an intermediate or level-2 miss, and greater than $c3$ is a severe or level-3 miss. Three successive misses (at any level) constitute a chronic miss situation.

The plan involves four payment amounts, which we call $a1$, $a2$, $a3$ and $a4$ for (respectively) unforgiven level-1 and -2 misses, level-3 and chronic misses (which are not forgivable). The plan has a schedule for issuing forgivenesses, namely one forgiveness is issued every k months ($k = 6$ is suggested). Forgivenesses can only be used for level-1 and -2 misses, and cannot be used in any month if the previous month was a miss. A forgiveness cancels the penalty for the month in which it is used; it does not cancel the event that a miss occurred. There is a cap (taken as two) on the number of available forgivenesses.

The consequences of each plan can be studied by making assumptions about the process P that is generating the Z-scores. For example, we may assume that the process is in parity or that we have deviation from parity by any desired amount. Here we consider only models in which successive months are independent. A more relevant model would allow for the possibility of correlation, specifically negative correlation, which we expect would be introduced by management's actions following a missed month.

The consequences of the payment plan depend on the scenario P and the eight parameters $(c1, c2, c3, a1, a2, a3, a4, k)$. For any given scenario, we can calculate the probabilities

$$
\begin{aligned}
p0 &= P(Z < c1) && \text{(probability of no miss)},\\
p1 &= P(c1 < Z < c2) && \text{(probability of a level-1 miss)},\\
p2 &= P(c2 < Z < c3) && \text{(probability of a level-2 miss)},\\
p3 &= P(c3 < Z) && \text{(probability of a level-3 miss)}.
\end{aligned}
$$

Then the consequences depend on $(P, c1, c2, c3)$ only through the values of $(p0, p1, p2, p3)$.

For each set of values of $(p0, p1, p2, p3, a1, a2, a3, a4, k)$ we could simulate the payment scheme, obtaining an estimate of the expected payout under one scenario. This would allow us to estimate both the long-run stationary behavior and the initial transient. However, obtaining results that are accurate enough to be useful requires a large amount of computation. Since

TABLE 3

State	Current month	Payment
$x\ 0\ 0\ i$	No miss	No payment
$x\ 0\ 1f\ i$	Forgiven level-1 miss	No payment
$x\ 0\ 1n\ i$	Not-forgiven level-1 miss	Pay a_1
$x\ 0\ 2f\ i$	Forgiven level-2 miss	No payment
$x\ 0\ 2n\ i$	Not-forgiven level-2 miss	Pay a_2
$x\ 0\ 3\ i$	Unforgivable level-3 miss	Pay a_3
$0\ m\ 0\ i$	No miss	No payment
$0\ m\ 1\ i$	Unforgivable level-1 miss	Pay a_1
$0\ m\ 2\ i$	Unforgivable level-2 miss	Pay a_2
$0\ m\ 3\ i$	Unforgivable level-3 miss	Pay a_3
$m\ m\ 0\ i$	No miss	No payment
$m\ m\ m\ i$	Chronic miss	Pay a_4

the process is a Markov chain (if we assume independence between months), we can compute the exact transition matrix and hence get exact results.

How many states can the system be in? We need to keep track of the outcomes for three consecutive months, including whether a forgiveness has just been issued, the number of forgivenesses currently available and the number of months since the last forgiveness was issued. This makes a total of $4 \cdot 4 \cdot 4 \cdot 2 \cdot 3 \cdot k$ states, which is 2304 if $k = 6$. Fortunately we can collapse these and work with no more than 36 states as follows. In Table 3, the first three columns under the "state" heading refer to the most recent three months: 0 means no miss, 1, 2 and 3 mean misses at the appropriate levels, x means anything, f and n mean that the latest miss has been forgiven or not forgiven, respectively, and m means a miss at any level. In the fourth column, i stands for 0, 1 or 2, the number of forgivenesses currently available. If we ignore temporarily the schedule for issuing forgivenesses, this gives 36 cases, namely those shown in Table 3.

Let A be the 36×36 transition matrix that in the (i, j) cell gives the probability of being in state i at month t, given that we were in state j in month $t-1$ and that a new forgiveness is not due to be issued. These probabilities depend on the chosen values of $p0$, $p1$, $p2$, $p3$. It turns out that two states, namely $x\ 0\ 1n\ 2$ and $x\ 0\ 2n\ 2$ are inaccessible, since if the most recent level-1 or level-2 miss was not forgiven, it must have been that no forgiveness was available; since only a single forgiveness can be issued, there is no way two forgivenesses could now be available. Deleting these states, A is a 34×34 matrix of transition probabilities.

Assume that the plan is started January 1. We start with a clean slate—no history of misses and no forgivenesses available. A forgiveness is issued January 15, so we are now in state $x\ 0\ 0\ 1$. The initial state vector $v(0)$ has a 1 in the first position and zeros elsewhere. At the end of January, a test result will become available and the new state vector is given by $v(1) = Av(0)$. At the end of February, the state vector will be $v(2) = A^2v(0)$. This goes on through the end of June. On July 15 a new forgiveness is issued. We can handle this by defining a new transition matrix B which simply maps a state with, respectively, 0, 1 or 2 available forgivenesses into the corresponding state with 1, 2 or 2 forgivenesses. Thus just before the July result is processed, the state vector is $Cv(0)$, where $C = BA^6$. After $6N$ months, the state vector will be $C^Nv(0)$. As N increases, this will converge to the dominant right eigenvector of C, which we call v.

The system we are studying is not stationary, even asymptotically, since the state vector depends on how many months it is since the last forgiveness was issued. Nevertheless, the average payment (asymptotically) can be calculated as

$$\bar{v} = \left(v + Av + A^2v + \cdots + A^{(k-1)}v\right)/k.$$

Once we have calculated $\bar{v}$, we can apply any desired schedule of payments and find the long-run average payment. We could also calculate the initial transient behavior.

Computations using this approach were carried out for four versions of the parameters $(c1, c2, c3)$, eight scenarios and two payment schedules. The results were very helpful in understanding the effect of the various components of the plan. For example, the effect of the "forgiveness" rule is to greatly reduce the average payment when parity holds, while not affecting it very much when consistent violations occur. The cap on the number of forgivenesses prevents the ILEC from accumulating credit to be used later. The definition and the amount of the intermediate payments are not very critical. Assessing small payments for small violations will reduce the average payment for large violations.

A similar approach was applied to some Tier II proposals. In one version, there were 344 states and straightforward simulation would have been impractical, but an exact calculation like that above went through without difficulty.

13. COMPUTATION

ILECs have complained that producing the necessary monthly reports and analyses is an unfair computational burden. Their procedures have developed

over many decades to provide management with techniques for efficient monitoring and control of their enterprises. The methods of statistical process control (which were invented by Walter Shewhart at AT&T in 1925) are used very widely in the industry. However, it is not clear that these techniques address the problem of detecting violations of parity month by month. We could have stationarity with parity being violated consistently or nonstationarity with parity holding every month.

The ILECs have little spare capacity for calculating the thousands of tests that may be needed each month, with reports being generated for each CLEC (there may be dozens of these) and for the aggregate of the CLECs. Clearly an automated system is needed; a system that requires manual intervention at any stage is not practical. In some cases major revision of the recording processes is needed; for example, for some SQMs the practice has been to record only averages, and new programming is needed to obtain the sample variances.

14. DISCUSSION

This paper has shown how intensive study of a practical and important problem can lead to novel formulations and techniques. Several research questions have been raised, including those listed at the end of the "Permutation Tests" section and the problem of applying the balancing approach to an aggregated statistic.

APPENDIX 1. AN IDEALIZED PROBLEM

We formulate an idealized version of the problem of choosing an optimum payment function. Our solution will be obtained using a new variation on the Neyman–Pearson lemma. We use the notation $[A]$ for the indicator function of the event A, so that $[A] = 1$ when A is true, and $[A] = 0$ when A is false. Also the notation $(x)^+$ stands for the function that equals x when x is positive, and is zero when x is negative (or zero).

Suppose we have an observation X, taking values in an arbitrary space, a simple null hypothesis H_0 [with density $f_0(x)$ relative to some measure] and a simple alternative H_1 [with density $f_1(x)$]. A protagonist can control which hypothesis is true. We want a nonnegative payment function $g(x)$, representing a penalty that must be paid when the observation is x, such that when H_0 is true, the expected payment is small and when H_1 is true, the expected payment is large. Imposition of this penalty will provide an incentive for the protagonist to make H_0 true.

Formally, we require that for some given constant m_1,

$$E\big(g(X)|H_0\big) = m_1 \tag{1}$$

while

$$E\big(g(X)|H_1\big) = \text{maximum}. \tag{2}$$

Unfortunately, for most interesting specifications this formulation makes no sense, because we can make $E(g(X)|H_1)$ arbitrarily large by making $g(x)$ huge on a small set where the likelihood ratio $L(x) = f_1(x)/f_0(x)$ is large. For example, when f_0 denotes the standard Gaussian density and $f_1(x) = f_0(x-1)$ denotes a unit shift alternative, if we take $g(x) = [x > \xi]m_1/(1 - F_0(\xi))$, we satisfy (1) and we can make (2) arbitrarily large by taking ξ large. This payment function is not reasonable, because the probability that the payment is incurred is very small under both H_0 and H_1.

Suppose we impose a side condition, making g bounded, by g_1, say. Clearly g_1 must not be smaller than m_1. Writing $\phi(x) = g(x)/g_1$ we have exactly the setup of the classical Neyman–Pearson test, since now we can interpret $\phi(x)$ as the probability that H_0 is rejected. The Neyman–Pearson lemma shows that the optimal g is $g(x) = [L(x) > c]g_1$ for some c. This is not a satisfactory solution to our problem because g is not continuous; a very small change in x can lead to a large change in $g(x)$.

We can get a different result by imposing a different side condition, namely

$$E\big(g(X)^2|H_0\big) = m_2, \tag{3}$$

where m_2 is specified. This condition controls the variance of $g(X)$ under H_0. A variational argument just like that of Neyman and Pearson shows that there is an optimal g, having the form

$$g(x) = k\big(L(x) - c\big)^+ \tag{4}$$

for some k and c. This payment function is continuous, is zero when $x < c$, satisfies (1) and (3) (if k and c are chosen correctly), and maximizes (2). It is clear that c depends only on the coefficient of variation of $g(X)$ under H_0, $\text{CV} = m_2/m_1^2 - 1$. We describe some examples.

Example 1: Monotone Likelihood Ratio

If x is one-dimensional and the likelihood ratio is monotone (increasing), we can write

$$g(x) = k\big(L(x) - L(\xi)\big)^+, \tag{5}$$

where k and ξ are determined by the two conditions

$$k\big(\alpha_1 - L(\xi)\alpha_0\big) = m_1, \tag{6}$$

$$k^2\big(\alpha_2 - 2L(\xi)\alpha_1 + L(\xi)^2\alpha_0\big) = m_2, \tag{7}$$

where

$$\begin{aligned}\alpha_0 &= \int_\xi^\infty f_0(x)\,d\mu(x),\\ \alpha_1 &= \int_\xi^\infty f_1(x)\,d\mu(x),\\ \alpha_2 &= \int_\xi^\infty \frac{f_1(x)^2}{f_0(x)}\,d\mu(x).\end{aligned} \tag{8}$$

The optimal expected payment under H_1 is

$$\begin{aligned}k\big(\alpha_2 - L(\xi)\alpha_1\big) &= \frac{m_2}{k + L(\xi)m_1}\\ &= m_1\left(\frac{\alpha_2 - L(\xi)\alpha_1}{\alpha_1 - L(\xi)\alpha_0}\right).\end{aligned} \tag{9}$$

The critical value ξ is determined by

$$\frac{m_2}{m_1^2} = \frac{\alpha_2 - 2L(\xi)\alpha_1 + L(\xi)^2\alpha_0}{(\alpha_1 - L(\xi)\alpha_0)^2}. \tag{10}$$

Thus we can view $g(x)$ as being determined by the moments m_1, m_2, by the pair m_1, ξ or by the pair m_1, α_0.

Example 2: Exponential Scale

Some simplification occurs if we assume H_0 is exponential, $f_0(x) = e^{-x}$ for $x > 0$, with a scale alternative:

$$f_\theta = \theta e^{-\theta x}. \tag{11}$$

We suppose $0 < \theta < 1$, so that the mean is $1/\theta > 1$. We have

$$\alpha_0 = e^{-\xi}, \quad \alpha_1 = e^{-\theta\xi} \tag{12}$$

while α_2 is infinite unless $1/2 < \theta$, in which case we have

$$\alpha_2 = \frac{\theta^2}{2\theta - 1}e^{(1-2\theta)\xi}. \tag{13}$$

The critical value ξ is determined by

$$\frac{m_2}{m_1^2} = \frac{2\theta}{2\theta - 1}e^{\xi} \tag{14}$$

so that ξ is not independent of θ. The optimal expected payment under H_θ is

$$m_1\frac{\theta}{2\theta - 1}e^{(1-\theta)\xi}. \tag{15}$$

No comparable simplification occurs for the normal shift model.

Example 3: The Binomial Case

Assume that K is binomial (p, n) with $p = p_0$ under H_0 and $p = p_1$ under H_1, with $p_1 > p_0$. This is a model of the benchmark situation. The likelihood ratio is monotone, so the optimal payment function is

$$g(k) = a(\theta^k - c)^+ \tag{16}$$

for some a and c [determined by the moment conditions (2) and (4)], where $\theta = p_1(1 - p_0)/p_0(1 - p_1)$.

Example 4: Uniform, Linear Alternative

Assume that H_0 specifies that X is uniform in $(0, 1)$, while for some a in $(0, 2)$, H_a makes the density of X

$$f_a(x) = 1 + a(x - 1/2). \tag{17}$$

In this case the optimal payment function is piecewise linear,

$$g(x) = ka(x - \xi)^+, \tag{18}$$

where k and ξ are determined from the moment conditions

$$\begin{aligned}m_1 &= ka(1 - \xi)^2/2,\\ m_2 &= k^2a^2(1 - \xi)^3/3,\end{aligned} \tag{19}$$

that is,

$$k = \frac{9m_2^2}{8am_1^3}, \quad \xi = 1 - \frac{4m_1^2}{3m_2}. \tag{20}$$

Notice that in this case ξ does not depend on a. The payment function in (13) maximizes the expected payment under all alternatives H_a, $0 < a \le 2$. The expected payment is

$$\begin{aligned}E\big(g(X)|H_a\big) &= m_1\left(1 + a\left(\frac{1}{2} - \frac{4m_1^2}{9m_2}\right)\right)\\ &= m_1\left(1 + \frac{a}{6} + \frac{a\xi}{3}\right).\end{aligned} \tag{21}$$

Discussion

While this formulation is much too simple to be of real interest, it does suggest some general principles. For example, it turns out that when the observation is real-valued, the optimal payment function will be a continuous function of the observation. This is clearly desirable. Second, the size of the payment can and should increase (smoothly) with the size of the violation. The solutions may suggest acceptable approximations, for example, in the binomial case we might choose to use a piecewise linear function

$$g(k) = a(k - c)^+. \tag{22}$$

Many questions remain, including extending the theory to handle composite alternatives and the simultaneous consideration of several measures. We may need to estimate an optimal payment function when the densities are not known a priori. The relationship of this approach to that used by economists needs to be worked out. We leave all these considerations for future work.

APPENDIX 2. SOME SERVICE QUALITY MEASUREMENTS

Various sets of measures have been adopted in different jurisdictions. Some of the measures proposed before the Louisiana PSC are given in the following list. We will not attempt to explain the jargon. The measures were classified into several sections, including pre-ordering, ordering, provisioning, maintenance, billing, trunk blockage, coordinated customer conversions and collocation. Each measure may have several components and is to be reported for each of several hundred geographical regions.

- Examples of benchmarked measures include:
  - Ordering: Firm order confirmation timeliness, 95% within 4 hours
  - Collocation: Percent of due dates missed, less than 10%
- Examples of measures where an ILEC analog could be defined include:
  - Provisioning: Order completion interval
    Percent missed installation appointments
    Percent provisioning troubles within 4 days
  - Maintenance: Customer trouble report rate
    Percent repeat troubles within 30 days
  - Billing: Invoice accuracy
    Mean time to deliver invoices

ACKNOWLEDGMENTS

We thank R. Bell, J. L. Gastwirth, E. Mulrow and M. Kalb for their helpful comments. The comments of two referees and an editor were very helpful. The views expressed in this paper are my own and do not represent the official policy of any company.

REFERENCES

BALKIN, S. D. and MALLOWS, C. L. (2001). An adjusted, asymmetric two-sample t-test. *Amer. Statist.* **55** 203–206.

BROWNIE, C., BOOS, D. and HUGHES-OLIVER, J. (1990). Modifying the t and ANOVA F tests when treatment is expected to increase variability relative to controls. *Biometrics* **46** 259–266.

FINKELSTEIN, M. O. and LEVIN, B. (1990). *Statistics for Lawyers*. Springer, New York.

GASTWIRTH, J. L. (1988). *Statistical Reasoning in Law and Public Policy* **2** 611–620. Academic Press, New York. (See especially pages 614–616.)

GASTWIRTH, J. L., ed. (2000). *Statistical Science in the Courtroom*. Springer, New York.

JOHNSON, N. J. (1978). Modified t tests and confidence intervals for asymmetrical populations. *J. Amer. Statist. Assoc.* **73** 536–544.

JOHNSON, N. L. and KOTZ, S. (1969). *Discrete Distributions*. Houghton Mifflin, Boston.

JOHNSON, N. L. and KOTZ, S. (1994). Further comments on Matveychuk and Petunin's generalized Bernoulli model, and nonparametric tests of homogeneity. *J. Statist. Plann. Inference* **41** 61–72.

KENDALL, M. G. and STUART, A. (1963). *The Advanced Theory of Statistics* **2**. Griffin, London.

LEPAGE, Y. (1971). A combination of Wilcoxon's and Ansari–Bradley's statistics. *Biometrika* **58** 213–217.

MANN, C. R. (2000). Statistical consulting in the legal environment. In *Statistical Science in the Courtroom* (J. L. Gastwirth, ed.). Springer, New York.

MAS-COLELL, A., WHINSTON, M. and GREEN, J. (1995). *Microeconomic Theory*. Oxford Univ. Press.

MATVEYCHUK, S. A. and PETUNIN, Y. I. (1990/1991). A generalization of the Bernoulli model arising in order statistics, I and II. *Ukrainian Math. J.* **42** 459–466; **43** 728–734.

MULROW, E. (2001a). Personal communication.

PODGOR, M. J. and GASTWIRTH, J. L. (1994). On non-parametric and generalized tests for the two-sample problem with location and scale change alternatives. *Statistics in Medicine* **13** 747–758.

U.S. EEOC (1987). Report 29 C.F.R. 607.4(D).

WOLTER, K. M. (1985). *Introduction to Variance Estimation*. Springer, New York.

Comment

Joseph L. Gastwirth and Weiwen Miao

Abstract. This valuable article by Dr. Mallows not only illustrates how challenging new problems arise in real-world applications, it also shows how constraints imposed by the adversarial process adopted by the regulatory and legal system create substantial barriers to objective resolution of scientific issues. The first part of our comment will briefly review the context of the problem. Then we will comment on some of the interesting statistical problems that are discussed in Dr. Mallows' paper. The third section concerns the difficulties statisticians and scientists face when they become involved in the legal environment with its focus on "winning" rather than finding the truth. The final section relates the relevance of the methods discussed by Dr. Mallows to the March 4, 2002 decision of the U.S. Supreme Court concerning competition in the power industry along with a few additional remarks.

1. LEGAL BACKGROUND

The purpose of the 1996 Telecommunications Act was to reduce the barriers to entry in various markets of the industry. In a sense, it can be regarded as a continuation of the government's antimonopoly policy that led to the consent decree in which AT&T was split into various companies, including the "local" Bell phone companies. Perritt (2000) quoted the purpose of the Act summarized by the Federal Communications Commission (FCC):

> (1) opening the local exchange and exchange access markets to competitive entry; (2) promoting increased competition in telecommunications markets that are already open to competition, including the long distance services market; and (3) reforming our system of universal service so that universal service is preserved and advanced as the local exchange and exchange access markets move from monopoly to competition.

The Act covers many aspects of the industry, including billing practices and privacy issues, beyond the scope of the article. Looking up the statute and related background material, we found that its legislative history required 21 volumes and already there have been several important legal cases, which are cited at the end of this comment. We will focus on the sections implementing the deregulation of the phone industry that led to the statistical problems discussed by Dr. Mallows.

In return for opening up local markets to competition in a fair and nondiscriminatory manner [Section 251(c)], the Bell companies [incumbent local exchange carriers (ILECs)] were given the opportunity to offer long-distance service to calls originating in their service area, subject to approval from the FCC (Section 271). This approval depends in part on the ILEC being in compliance with Section 251 and related parts of the statute. Section 251(c) (2) of the Act, which defines the parity concept is augmented by Section 251(c) (3), which states that the ILEC must provide to any requesting telecommunications carrier access to network elements on an unbundled basis with rates, terms and conditions that are just, reasonable and nondiscriminatory (Perritt, 2001, page 95). An ILEC shall provide such unbundled network elements in a manner that allows requesting carriers to combine such elements in order to provide such telecommunications service.

Joseph L. Gastwirth is Professor of Statistics and Economics, George Washington University, Washington, DC 20052 and Visiting Scientist, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute (e-mail: jlgast@gwu.edu). Weiwen Miao is Assistant Professor, Department of Mathematics and Computer Science, Macalester College, St. Paul, Minnesota 55105 (e-mail: miao@macalester.edu).

The fact that the word nondiscriminatory occurs twice in this section suggests, albeit indirectly, that methodology useful for examining "parity" may have parallels in methodology that has been found useful in examining data arising in equal pay and related cases concerning discrimination. Several issues arising in both applications will be discussed further in the next section as the similarity between "benchmark" rules and the Equal Employment Opportunity Commission's four-fifth's rule is noted in the article.

The article notes that the local companies have little capacity to calculate the many statistical tests that may be needed to monitor compliance. It would appear that provisions for sharing the cost of monitoring compliance could be built into the system. Since these costs were created by the Act itself, they might well be included as an appropriate cost that may be used by the ILEC in pricing its services to a competitive local exchange carrier (CLEC). The Supreme Court discussed these costs in AT&T et al. v. Iowa Utilities Board (1999). The subsequent implementation of the decision by the Eighth Circuit in GTE v. FCC (2000) noted that state commissions have the authority to determine pricing methods and they may include "recovery mechanisms for legitimate costs."

2. STATISTICAL ISSUES

The core of the problem is the translation of the nondiscriminatory or parity requirement. Let X and Y be a service quality measurement (SQM) to the CLEC's customers and the ILEC's *own* customers, respectively. Let $X_1 \le X_2 \le \cdots \le X_n$ and $Y_1 \le Y_2 \le \cdots \le Y_n$ be *ordered* samples from X and Y (where large values are "bad"). Suppose that X and Y are continuous and independent. Then when parity holds, that is, when X and Y have the same distribution, by applying a theorem from Gnedenko and Mihalevic (1952) (see also Drion, 1952; Takacs, 1964), we have

$$P(X_1 < Y_1, X_2 < Y_2, \ldots, X_n < Y_n) = \frac{1}{n+1}.$$

Since X and Y are continuous, $P(X_i = Y_j) = 0$ for any i and j. So, if we adopt the rule of "company B (ILEC) is in violation unless $X_i \le Y_i$, for all i" described in the article, then even when parity holds, the probability of finding the ILEC in violation of parity is $1 - 1/(n+1) = n/(n+1)$. Thus, as the sample size n increases, the probability of finding a parity violation goes to 1. Clearly this would be unfair to the ILEC.

Dr. Mallows properly observes that the usual pooled t-test and the Welch t-test are not appropriate for testing parity because the null hypothesis requires the first two moments of X and Y to be equal, but the alternatives allow both to change. We conducted a small simulation to examine the comparative performance of four parametric t-tests and four nonparametric tests. The four parametric t-tests are the pooled t-test, the Welch t-test, the modified t-test (Brownie, Boos and Hughes-Oliver, 1990) and the adjusted t-test (Balkin and Mallows, 2001). The four nonparametric tests are the Wilcoxon test, the Savage scores test, the normal scores test and a test that has high relative efficiency for the gamma family, ranging from the exponential to the normal (Gastwirth and Mahmoud, 1986). We call it the GM test in our study. The results in our Table 1 are based on 10,000 simulations. In the table, n_1 refers to the sample size of X and n_2 refers to the sample size of Y. As pointed out in Balkin and Mallows (2001), the distributions of X and Y are usually skewed and the sample sizes are very different with $n_1 \ll n_2$. Our simulation focused on the exponential distribution and on the influence of the unequal sample sizes with $n_1 + n_2 = 100$.

Table 1 clearly shows that unbalanced sample sizes diminish the power of all tests considered. The greater the imbalance in the sample sizes, the lower the power of all the tests. Uneven sample sizes also influence the type-1 error of the tests. For the pooled t-test and the four nonparametric tests, the greater the difference in sample sizes, the larger the actual level. When the sample sizes are equal, the levels of these five tests are very close to the nominal 0.05. Even in unbalanced settings, the excess in their levels over 0.05 is not great. The other three tests do not perform well: the actual type-1 error of the modified t-test and the adjusted t-test is higher than 0.05 for all sample size combinations. The Welch t-test has a size much smaller than 0.05 for uneven sample sizes; consequently, its power is also much lower than the other tests. The pooled t-test works well when sample sizes are equal, but its type-1 error is larger than 0.05 for unbalanced sample sizes. Among the four nonparametric tests, the Savage test and the GM test have the highest power, but their sizes slightly exceed 0.05. When the sample sizes are not extremely uneven, the size of the GM test is close to the nominal level of 0.05. The GM test procedure, which is designed to have high power relative to the optimum tests for members of the Gamma family, achieves this for exponential distributions. It loses a small amount

TABLE 1
Type-1 error and power for various two-sample tests—exponential data

		Parametric t-tests				Nonparametric tests			
(n_1, n_2)	β	Pool	Welch	Modified	Adjusted	Wilcoxon	Savage	Normal score	GM test
(5, 95)	1	0.072	0.012	0.077	0.096	0.055	0.069	0.055	0.062
	1.5	0.301	0.062	0.313	0.355	0.205	0.289	0.231	0.269
	2	0.542	0.118	0.554	0.599	0.370	0.520	0.421	0.482
	2.5	0.710	0.193	0.719	0.750	0.506	0.685	0.580	0.650
	3	0.820	0.252	0.829	0.852	0.616	0.801	0.698	0.765
(10, 90)	1	0.066	0.016	0.075	0.094	0.052	0.064	0.053	0.060
	1.5	0.405	0.156	0.430	0.477	0.286	0.388	0.313	0.357
	2	0.721	0.382	0.739	0.776	0.542	0.698	0.598	0.669
	2.5	0.883	0.581	0.895	0.910	0.727	0.865	0.783	0.839
	3	0.955	0.720	0.960	0.967	0.839	0.943	0.883	0.925
(15, 85)	1	0.062	0.021	0.072	0.090	0.051	0.060	0.051	0.056
	1.5	0.475	0.259	0.507	0.549	0.346	0.453	0.378	0.422
	2	0.825	0.601	0.844	0.867	0.673	0.807	0.718	0.779
	2.5	0.948	0.815	0.955	0.964	0.845	0.937	0.887	0.921
(25, 75)	1	0.059	0.033	0.074	0.094	0.050	0.057	0.052	0.056
	1.5	0.582	0.437	0.629	0.673	0.444	0.562	0.477	0.533
	2	0.911	0.834	0.929	0.944	0.800	0.901	0.839	0.885
(50, 50)	1	0.050	0.049	0.080	0.104	0.050	0.050	0.049	0.048
	1.5	0.642	0.641	0.736	0.786	0.545	0.635	0.566	0.616
	2	0.961	0.960	0.977	0.985	0.906	0.957	0.918	0.946

of power compared to the Savage test and the pooled t-test, but has greater level robustness.

We also ran simulations for normal data and obtained similar results. In summary, our studies show that both the Savage test and the GM test work reasonably well in uneven sample size situations. The Savage test, however, has less level robustness. In the problems discussed by Dr. Mallows, the precise form of the skewed distribution is usually not known and likely varies by SQMs. Our simulations suggest that using a robust nonparametric test such as the GM test or the robust LePage-type tests described in Podgor and Gastwirth (1994) should have type-1 error close to the nominal level and still possess high power under different situations.

Our simulation studies indicate that highly imbalanced sample sizes noticeably decrease the power of statistical tests. Dr. Mallows kindly informed us that the available data are really observational in nature. Such data are often analyzed as though they arise from a random process and possible effects of the deviation from the assumed model are assessed by sensitivity analysis (Rosenbaum, 2002). If the highly unequal samples sizes are due to the fact that the vast majority of customers remain with the ILEC, that is, the CLEC has a very small market share, then this imbalance would be inherent in the data. On the other hand, if the ILEC makes it difficult for its customers to switch providers, then the small sample of the CLEC customers in the available data actually reflects the policy of the ILEC under scrutiny. Indeed, Young, Dreazen and Blumenstein (2002) indicated that many local phone companies (the ILECs) did that, for example, one imposed a \$111.86 fee (subsequently removed after pressure from a regulatory commission). Solomon (2002) reported that while the local companies (Bells) are anxious to crack the long-distance market, they have been fined by both the states and the federal government for not first providing an equal playing field to rivals for local service, as federal regulations require. As we have seen, such practices that are inconsistent with the requirements of Section 251(c) (3) also make it difficult to detect parity violations. The same problem arises in equal employment cases, where the hiring process determines the samples of minority and majority employees available to study fairness of promotion. Examining these employment practices separately can lead to anomalous results (Gastwirth, 1997). Judge I. Goldberg's dissent in Watson v. Fort Worth Bank (1986), where he notes that an employer who discriminates in hiring is unlikely to become a saint when

making promotions, shows that courts are aware of this problem, so regulatory authorities should also examine the reasons for the highly imbalanced sample sizes.

Problems concerning differences in the "upper tail" also arise in medical applications where a drug may be effective only on a portion of the population. This type of alternative is similar to the one here, that is, an ILEC may provide poor service only to the desirable customers of the CLEC. The methods reviewed by Freidlin and Korn (2002) might also be explored.

The problem of balancing the type-1 and type-2 errors also has arisen in equal employment cases. In that context, Dawson (1980) questioned the imbalance in the choice of values for the two types of error and recommended that they be equal. Dr. Mallows has provided a more detailed analysis and discussed the magnitude of the change (δ) that should be used to calculate the type-2 error probability (β).

The author's discussion of the inappropriateness of the replicate variance (RV) method reminds us that just because a method is readily available or a computer program exists to implement it does not mean that one should use it uncritically. The RV method is used to obtain variances and covariances for parameters estimated from complex survey data. In that application, a jackknife-type estimator is used, that is, one could sample from all primary sampling units (PSU's) and then reestimate them by deleting one PSU at a time. The covariance matrix of these estimates is calculated and used as the final estimate of the covariance matrix.

A superpopulation model is appropriate when there is an underlying variability in the process generating the finite population sampled from. Then there is a need to incorporate both the variability in the generation process and the sampling error from the sampled finite population. The article by Graubard and Korn (2002) describes procedures that have been developed for superpopulation models. There is no superpopulation in the problem Dr. Mallows discusses as the focus is on deciding whether or not the CLEC is receiving the same level of service as the ILEC in a particular geographical area.

The question of which summary statistic to use is very important and the author indicates why it has been so difficult to come to an agreement. We only wish to mention an additional complication, namely that some of the various measures are dependent. This may allow the ILEC more room to manipulate the data: the ILEC can provide equal service to those measures that are related and occur several times as components of the summary test hoping that they will hide another measure of poor service to customers of the CLEC. Dr. Mallows also observed that correlation over time as well as common causes of poor service need to be accounted for in deriving the appropriate distribution of the test statistics.

3. THE ADVERSARIAL ENVIRONMENT

The negative experience that Dr. Mallows had with the legal and regulatory system unfortunately is all too common. As statistical scientists, we are primarily concerned with uncovering the actual facts and interpreting them properly. This includes providing measures of uncertainty and discussing various strengths and weaknesses of a study or analysis. The legal system restricts the nature of the information it allows to be submitted as evidence to ensure that it is reliable and relevant to the case at hand. According to Burns (1999), the legal system has developed its own concept of "legal truth." Trial procedures rely heavily on the ability of each side to marshal the facts most favorable to its view of the case and trusts that the truth will be brought out in the process of cross-examination. Thus, the partisanship noted by Dr. Mallows is an inherent part of the legal system.

The ethical system that lawyers are taught also differs greatly from the scientific ideal. Burns (1999, page 76) described a sample case often used in the professional responsibility class in law schools. Briefly, the lawyer believes her client is innocent of the murder of his ex-wife. The man's landlady saw him return home shortly before the shooting and he was in his room when the police came shortly afterward. The client does inform his lawyer that he did leave his room for a few minutes to get some fresh air. If the client does not take the witness stand, his lawyer is free to argue that the evidence will show that the man never left his room and therefore could not have committed the crime. Burns then says that his lawyer may conduct a destructive cross-examination on any witness, however truthful, who testifies that they saw her client outside his room during the relevant time. While Burns is bothered by this, many nonlawyers might conclude that teaching lawyers that such practices are "ethical" and "responsible" contradicts the claim made by the legal community that the adversarial process is another approach to finding the truth.

The nation's most cited legal scholar, Judge Posner (2001), observed that the significance and social value of cross-examination is often misunderstood. He observes that although "cross-examination *can* destroy a

witness's credibility, it rarely does so." The reason is that the individuals whose credibility would be seriously tarnished are not likely to be called to testify. The implication of Judge Posner's view is that cross-examination is a deterrent to people who would lie or try to cover up evidence on the stand. The example from legal ethics classes, however, is likely to *deter* honest citizens who are eyewitnesses to a crime or destruction of evidence from testifying.

One of the difficulties scientists face in this system is that the lawyers are under *no obligation* to inform us of all the relevant data, who else they may have consulted or whatever other information they have. This asymmetry in disclosure of information is also noted in a recent National Research Council report (2002). During the discovery process, each side has the opportunity to depose or question each other's witnesses in preparation for the trial. This process is supposed to reduce the possibility of unfair surprise and in cases involving statistical evidence may encourage both parties to agree on a common database. Unfortunately, the discovery process allows the lawyers a great deal of flexibility without judicial oversight. Depositions are typically carried out in a law office, with only the witness and lawyers for both parties present. While the lawyer who plans to use your testimony will try to be of assistance, as Mann (2000) pointed out, their first duty is to their client. Worse yet, there appears to be an increase in problems with the discovery system. These can readily occur in situations where one party has better access to or possesses the information. Over the last few years, the *National Law Journal* has published articles concerning the possible destruction or hiding of relevant information or documents by DuPont, Wal-Mart and the auto manufacturers. As the local companies will collect the data in the application discussed by Dr. Mallows, we describe similar problems occurring when the needed information is under the control of one party.

A recent toxic tort case involving polychlorinated biphenyl (PCB) exposure illustrates the importance of a full and fair discovery process and the incentive a party that controls the information has to hide or destroy it. At issue is whether a producer's employees or the surrounding community were exposed to toxic chemicals. The *time* a producer knew or should have known of a potential harm from exposure is of critical importance, because the law expects one to act prudently in light of the state of knowledge at the time. Monsanto and Solutia, the two companies that owned the plant, were found liable for damages. The firms argued that the plant was shut down in 1971, immediately after it was concluded that PCBs were potentially carcinogenic. Grunwald (2002) reported that in 1966 and 1969 Monsanto had observed that fish exposed to the plant's output died. The documents and testimony about these pre-1971 events, clarifying when the firms knew there was a risk from PCBs, was a factor in the deliberations of the jury. Clearly, the defendants and/or their lawyers had an incentive to destroy or hide this early evidence.

Gastwirth (1991) described another issue we may face as experts. After being deposed in an equal employment case, "more" employment records were found by the defendant and provided to the plaintiff. It may be reasonable to assume that records and documents that are missed during the first search by a defendant are missing at random. After the defendant knows the evidence that needs to be countered at trial, for example, after deposing the plaintiff's expert, it is less plausible that unfavorable documents have the same probability of being located than favorable ones. The problems experts face in criminal cases can even be more bothersome (Geisser, 2000).

We agree with Dr. Mallows that the adversarial and confrontational process used in the legal system discourages a scientific dialogue aimed at resolving the issues. Indeed, it may well encourage unethical behavior, especially as strong sanctions are rarely imposed (Nesson, 1991). We would be remiss, however, if we failed to mention that our own profession might also benefit from stronger ethical guidelines. Horton (2001) cited a study of medical statisticians. While the response rate was low (37%), half of the respondents knew of at least one fraudulent project done in the previous 10 years, 26% described deceptive reporting of data and 19% knew of data suppression. This should motivate our profession to take action that would enable individuals who observe such practices to come forward without jeopardizing their livelihood.

4. FINAL REMARKS

Dr. Mallows has illustrated how a variety of critical statistical issues arose in an important application, for example, multiple comparisons, dependent observations, the potential effect of change in the process after the ILEC has been found in violation of parity and the essential role that the power of a test has in deciding which procedure should be used. Not surprisingly, similar issues have arisen in the context of equal employment litigation. For example, the loss of power

in highly unbalanced samples for comparing SQMs arose there. As the author's discussion of the "replicate variance" method demonstrates, one must think carefully about the assumptions underlying a method developed for one purpose before using it on data from another one. Appropriate modifications may be necessary.

At first glance, it might seem that the application discussed is unique. On March 4, 2002, the U.S. Supreme Court decided another case that may well generate similar problems. Two cases, New York v. Federal Energy Regulatory Commission and Enron Power Marketing Inc. v. Federal Energy Regulatory Commission, were consolidated. The issue concerned regulations established to enhance "open access" in energy transmission that the Federal Energy Regulatory Commission (FERC) established. The purpose of the rule was to encourage lower electricity rates by structuring an orderly transition to competitive bulk power markets.

FERC ordered unbundling of wholesale and retail generation and transmission, and required local power companies to apply a single tariff for the transmission of its own wholesale sales and purchases and those of competitors. It also said that if a public utility voluntarily offers or a state requires unbundled retail access, the retail customer *must* obtain its unbundled transmission service under a nondiscriminatory transmission tariff on file with the Commission. FERC did not require local utilities to separate generation from transmission costs to retail customers if neither of the above situations applied. The court sided with the Commission on both issues.

To implement the nondiscriminatory open access provisions, it is reasonable that not only will the tariffs need to be identical, but other measures of service should also be equal. Thus, similar problems, albeit with different SQMs, may require the same kind of rigorous statistical investigation that Dr. Mallows and his colleagues on both sides of the regulatory hearing gave to assessing parity in telephone service.

Another important legal development that statisticians thinking of participating in legal proceedings should be aware of is the heightened scrutiny that courts are giving to expert testimony. Three major cases have dealt with the screening the trial judge should give to proposed testimony to ensure its reliability before admitting it into evidence. Good sources for a description of these cases and their impact on expert testimony are Kaye (2001), who discussed econometric testimony in an antitrust case, and Rosenblum (2000), who focused on their effect on equal employment cases.

ACKNOWLEDGMENTS

The authors thank Dr. Barry Graubard for a most helpful discussion. The study in Section 2 was supported in part by a grant from the National Science Foundation.

CASES CITED

AT&T Corp. et al. v. Iowa Utilities Board et al., 119 S. Ct. 721 (1999).

Enron Power Marketing, Inc. v. Federal Energy Regulatory Commission (2002).

GTE Service Corp. v. FCC, 295 F. 3d 413, 427 (D.C. Cir. 2000).

New York et al. v. Federal Energy Regulatory Commission (2002).

Watson v. Ft. Worth Bank, 798 F. 2d 791 (5th Cir. 1986).

Comment

Edward J. Mulrow

One wonders if the framers of the Telecommunications Act of 1996 realized the growth in statistical thinking that would take place across the country as state public service and public utility commissions struggled with the task of implementation. Dr. Mallows has certainly done an excellent job of describing this process and of personally making major contributions to it. It has been a great pleasure working alongside him.

Dr. Mallows in his exposition ably tackles both the application of statistical thinking and its context. In dis-

Edward J. Mulrow, formerly with Ernst & Young LLP, is Senior Manager in the Statistical and Probability Sampling Services practice of PricewaterhouseCoopers LLP, Washington, DC 20005 (e-mail: edward.j.mulrow@us.pwcglobal.com).

cussing this challenge, he well describes the mathematical issues in framing the inference, the problems of estimation and a discussion of the distribution of the statistics, but he does much more too. He also discusses what he described in his 1997 Fisher Memorial Lecture as "the zeroth problem: considering the relevance of the observed data, and other data that might be observed, to the substantive problem" (Mallows, 1998). I hope to maintain the same flavor with my comments, and I will try to supplement Dr. Mallows' discussion with some of my own experiences and the reflections on them.

1. BACKGROUND

I, along with colleagues at Ernst & Young (E&Y), have been involved in several state proceedings on the local exchange carrier service parity issue. We were retained by an incumbent local exchange carrier (ILEC) to act as an independent third party in deriving a statistical approach that would meet with a state commission staff's approval. When the project began, we requested two things from the ILEC:

1. That the team have access to the performance measure data so that they could be explored before a final methodology was chosen.
2. That the client allow the team to report its findings to the regulatory bodies involved, even if they were not flattering to the ILEC.

Our client agreed to these conditions, asking only that we secure a level playing field for them. A natural consequence of this client arrangement is that we were able to proceed collaboratively—something that we feel has led to a better understanding of both the data and the issues than would otherwise have been the case.

Naturally, we did not always agree with Dr. Mallows, especially at the beginning since we were coming at the problem using somewhat different tools. Dr. Mallows (1998) pointed out, "In a complex problem, it is possible for ethical analysts to take opposing positions. But this style of thinking is what statisticians should be trained to do." With the data as teacher, we narrowed our differences by laying out the arguments on each side of the issue and concentrating on what the data implied. In the end, we believed that we had a statistical solution that fairly addresses the concerns of all the stakeholders: the ILEC, the competitive local exchange carriers (CLECs), and the state commissions.

Philosophical differences of opinion still exist over the appropriateness of some statistical techniques, and I will briefly address these below. The approach now agreed upon makes these differences largely moot, however. A bigger issue has been differences between the parties over subject matter input that is needed to carry out the statistical methodology. Dr. Mallows points out in the "Balancing" section of his paper that statisticians cannot provide this input, but they can perform analysis that describes the consequences of different input choices. I will also address these "subject matter" issues and suggest some alternatives in ways to think about the problem to help commissioners decide the issues.

2. STATISTICAL ISSUES

The most common way to deal with the analysis of the data is to use all transactions that have been completed in the month of interest. There is no sample that is taken and there is no randomized assignment of treatments to units. As Rubin (2000) stated:

> The key problem for inference is that, for any individual unit, we get to observe the value of the potential outcome under only one of the possible treatments, namely the treatment actually assigned, and the potential outcome under the other treatment is missing. Thus, under this straightforward perspective, inference for causal effects is a missing-data problem. When this definition of causal effects is applied to a set of individuals, a complication is that the potential outcomes for a particular unit may depend on the treatments assigned the other units.

Thus, we have an observational study and we look at the data as a sample of the service process—a process that is very complex. When we (E&Y) initially examined what other jurisdictions were considering for data analysis, we did not see anyone advocating a methodology that tried to take the complexities of the process into account. Two things concerned us:

1. Identifying important covariates so as to reduce bias.
2. Allowing for dependencies that may be present due to the way in which the telecommunications industry is structured.

The team proposed a random group replicate variance method at a Louisiana Public Service Commission workshop as a way to handle these concerns. As Dr. Mallows points out, it is a technique that is often used for variance estimation with sample survey

data. Since some members of the E&Y team come from sampling backgrounds, this was a natural way to think about the problem. After listening to objections from Dr. Mallows and others at the workshop, the methodology was revised to jackknife the random group estimates. Both approaches are described in Wolter (1985). Jackknifing differs from the random group variance method mainly in that jackknifing protects better against biases that can be important in small samples.

An additional adjustment to the method is also needed to satisfy Dr. Mallows' property (d) in the "Disaggregation and Reaggregation" section of his paper. Once both these (minor) adjustments are made, we argue that we have two models—both of which might be considered adequate to test for parity. Now in an observational study based on operating data, such as we have in this context, there is no entirely satisfactory constructive way to choose, as there would be in a real experiment or sampling setting, between these two competing models (or other plausible models that might have been found by other researchers). Clearly another less model-sensitive approach was needed.

The team felt that the balancing approach that Dr. Mallows and others were advocating provided such an alternative since it does not require that pre-specified alpha levels be used. Without a doubt, balancing makes the decision process less sensitive to the choice of a model. By this I mean that the hypothesis testing process will end up with the same conclusion of parity or nonparity with a balancing approach (under either model). The evaluation of the type-1 and -2 errors does depend upon the model, but the decision does not. The size of the test and its power can be very different depending on the model that underlies the inference, but not necessarily the decision.

Consider for a moment how one would determine the "balancing critical value" for a test statistic based on a resampling method such as the jackknife. Our choice was to employ a concept similar to that of a design effect in sampling (Kish, 1965). In this case, we would calculate the standard error for the difference in the ILEC and CLEC performance measures using the jackknife approach as well as the approach Dr. Mallows describes. We then calculate the balancing critical value for the jackknifed test statistic by rescaling the balancing critical value for the truncated Z model using the ratio of the jackknife standard error to the standard error of the truncated Z model. Now, the jackknifed test statistic will be greater than the critical value (and we would reject the null hypothesis) if and only if the truncated Z-test statistic is greater than its balancing critical value.

Once we noticed this about the technique we were advocating, we concluded that Dr. Mallows' methodology was just as appropriate for the data as our own. Furthermore, since our team strongly stressed from the beginning the need for disaggregation to homogeneous subgroups in making comparisons, we had real concerns about the stability of a jackknife estimator employed directly for individual subdomains—something the Public Service Commission wanted to do. It is also the case that while the truncated Z methodology is complicated, it is simpler to explain to commissioners and easier to implement than the jackknife methodology. Therefore, a compromise was reached with Dr. Mallows, and we could move on to putting our energies into creating a process that would be able to perform all the necessary calculations in a timely manner. For more details on this process see Balkin (2001).

To sum up the process, it was exploratory at the beginning, choosing safe tools that are known to work in complex data settings. Then there was a period of disagreement and rethinking with the use of confirmatory tools (some new). We did not fully explore alternatives that were being suggested in regulatory workshops that we were not involved with. Instead we concentrated on adjusting the techniques that we had offered up. This was followed by convergence, perhaps too quickly, to an approach that we felt was robust against what we know we do not know.

3. SUBJECT MATTER ISSUES

In the state hearings that I have been involved in, the ILEC, the CLECs and the state commissions have accepted the truncated Z statistical methodology that Dr. Mallows describes in the "Disaggregation and Reaggregation" section along with the balancing methodology. However, there are two important subject matter decisions that need to be made to implement the methodology: (1) the reaggregation level for the truncated Z statistic and (2) the alternative hypothesis used to determine the balancing critical value.

With respect to the issue of the reaggregation level, one needs to understand how the stakeholders are thinking about the issues. It is quite natural for a statistician to consider using a "global" decision process in the presence of multiple test results. Since we know that testing error exists even when the null hypothesis is true, we understand that it is very likely that a

test failure will be observed when hundreds (or possibly thousands) of tests are performed at the same time. However, for the regulators the issue is one of determining discrimination that is harmful to local competition and they do not want to forgive a testing failure if it truly represents poor service to a CLEC's customers.

This would not be a big issue if the aggregate statistic could not mask discrimination at lower testing levels, but we have not found one yet that cannot be "gamed." Dr. Mallows has been very creative in deriving the truncated Z statistic, yet it is still possible to provide good service to a CLEC in enough cells that would result in truncated Z scores of 0 so that they outweigh one or two bad cells. For this reason, the CLECs have voiced concern about letting the aggregation level of the truncated Z get too high.

What this really boils down to is determining the disaggregation level where it is important to detect discrimination. In Georgia, the state Public Service Commission has decided to determine discrimination based on the type of business entry that a CLEC uses in the marketplace. For instance, if a CLEC is reselling plain old telephone service (POTS), then all cell results involving these types of services, be they business customers, residential customers or other types of special services under resale POTS, are aggregated together using Dr. Mallows' truncated Z. In contrast, Florida has decided that it wants discrimination determined at much lower levels, for example, residential resale POTS where a dispatch call is made to a customer with less than 10 circuits.

In looking at the problem in this light, we are really dealing with the issue of bias reduction in a set of observational data. Many techniques exist to reduce such bias; see Hinkins (2001) for a further discussion of this issue. If a commission (or other subject matter expert) determines that, in terms of opening up the market to competition, the appropriate place to monitor discrimination is at a much higher level than where like-to-like comparisons need to be made, then we may want to question whether the computationally intensive truncated Z aggregate test statistic that Dr. Mallows describes is necessary. Would a simpler process suffice? In one jurisdiction, a commission staff member rejected the idea that any further disaggregation of the data is necessary even though the ILEC was willing to perform the extensive calculations. The main reason that the truncated Z was rejected is that it was felt that the process would be too hard for a commission to monitor. The potential benefits in bias reduction that might be obtained do not, in this particular staffer's mind, outweigh the cost of monitoring the system.

The most contentious issue, however, is the choice of the alternative hypothesis for the balancing methodology. Dr. Mallows describes the situation quite well. Mulrow (2001b) provided extra insight into the way a balanced test works. In the simple situation where we are using the modified t statistic described by Dr. Mallows, a balanced parity test is equivalent to "benchmarking" the effect size of the disparity. That is, the null hypothesis of parity is rejected whenever

$$\frac{\bar{X} - \bar{Y}}{s_Y} > \frac{\delta_a}{2}.$$

The left-hand side of this relationship is known as an effect size called Glass' d in the metaanalysis literature (Hunter and Schmidt, 1995). On the right-hand side, δ_a denotes the size of distribution shift that is assumed for the balancing alternative hypothesis.

Note that sample size plays no role here (unless we evaluate the size of the error probabilities). If the estimated size of the disparity is less than $\delta_a/2$, then the difference between the average performance in servicing CLEC and ILEC customers is considered immaterial. In other words, the difference has no practical significance.

This concept should be one that subject matter experts can address. One of the problems though is that these experts are not used to thinking about standardized differences of this nature. As Dr. Mallows suggests, statisticians can provide analyses of the situation to see if they can bridge the gap between the input the methodology calls for and the input with which the subject matter expert is familiar. Another possibility is to change the concept upon which the methodology is based.

Many subject matter experts whom I have talked with are used to thinking in terms of percent increase (or decrease). In this case the effect size size would be measured as

$$\frac{\bar{X} - \bar{Y}}{\bar{Y}}.$$

The concept of an immaterial difference in performance can be restated as the difference in average performance being less than a $100\rho\%$ increase. Now if a subject matter expert determines an appropriate value of ρ, then we can determine the parameter of the alternative hypothesis by

$$\delta_a = \frac{2\rho}{\mathrm{CV}_Y},$$

where CV_Y is the coefficient of variation of the ILEC distribution. One problem that this approach has is determining the coefficient of variation for the ILEC distribution. Should we use a value based on previous knowledge of the process or should we estimate it from the current data? If we choose the latter, should we use a global value for the performance measurement of interest or should we estimate the CV for each cell? Another problem that arises is how we would translate this type of approach to the case of counted variables.

4. OTHER ISSUES

Dr. Mallows has summarized a number of issues that have arisen in the implementation of the Telecommunications Act of 1996. I have focused on some of the more contentious issues from my experiences. I do appreciate the fact that Dr. Mallows has addressed issues such as benchmarks and penalty payments. These are topics that have been viewed by some as nonstatistical. Dr. Mallows' treatment of benchmarks is very creative. It is unfortunate that, in all the proceedings that I have knowledge of, parties have chosen to ignore the statistical aspects of the benchmark comparison problem. Similarly, statisticians have not been asked to play a role in the determination of penalty amounts. In my view it is unwise to ignore the statistical methodology being used to determine parity/discrimination when deriving the penalty payment mechanism.

Better solutions to such problems as determining the alternative hypothesis for balancing might be found if the penalty mechanism is also considered in the solution. It is perhaps too simple an idea, but the choice of δ may not be as contentious if payments for statistical test failures are simultaneously considered. If an ILEC wants δ_a large so that the probability of a type-1 error is not very high for small samples, then a high penalty when there is a failure may be reasonable. Likewise, CLECs that want smaller values of δ_a should be willing to accept smaller penalty payments. This gets at the idea of balancing the expected loss to each side when a testing error occurs.

Another issue is the way many of us have gone about solving this problem. Of the hundreds of measures that are being dealt with, only two or three are measured variables; most are counted variables. However, almost all of the analytic work and discussion has dealt with measured variables. The assumption is that once we solve the problem of comparing means, the solution to the proportion or rate of a counted variable will be immediate. However, it is not always the case that the solution that is derived for a mean is readily applicable to a proportion or rate. Dr. Mallows has given one example in discussing the balancing alternative hypothesis for a counted variable. We must take time to consider the meaning of the alternative for a proportion or rate. If we had first developed a methodology for a proportion measure, would we immediately think of using the arcsine–square-root transformation? Maybe we should have concentrated on developing a counted variable concept that is easier for a regulatory commissioner to understand.

Similarly, we must be careful in applying the modified t statistic developed for mean measures to proportion measures. The modified t uses only the ILEC variance to determine the standard error of the difference in the ILEC–CLEC mean. Intuitively this is because there are two potential ways for an ILEC to discriminate when looking at the length of time it takes to complete an order. It is thought that the ILEC might be able to independently control the location and spread of the CLEC service time distribution, so the modified t provides a way to make a test of mean differences more sensitive to situations where the CLEC variance is larger than the ILEC variance. From a theoretical perspective, the modified t is more powerful than the pooled t against compound (mean and variance) discriminatory alternatives.

Do we want to use the same idea with counted variables? There is only one parameter that the ILEC can control here, and there are discriminatory alternatives for which the modified t will have lower power than a test statistic based on a pooled variance (a CLEC proportion larger than the ILEC proportion which is greater than $1/2$). When sample sizes are small for a proportion measure, the recommended "exact" test, Fisher's exact test, is a pooled test (Lehmann, 1986). Why should we switch to a nonpooled test when sample sizes are large? I ask these questions because in some jurisdictions, for example, New York, the modified version of the test statistic is used for proportion measures. This does not seem right to me.

5. CONCLUSION

Dr. Mallows has made many significant contributions to solving the problem of implementing the Telecommunications Act of 1996. He has been exemplary in the way he has handled himself in the adversarial environment of regulation and he has been an equally good collaborator. I thank Dr. Mallows for a well thought out discussion of the statistical methodologies that have been considered and in some cases

employed by jurisdictions across the country. On the surface this problem seems simple, but due to the nature of the data, the requirements imposed by many regulators and the number of different stakeholders, this problem has many complexities. For those who are looking for examples of real-world problems to bring to their students, this is one that illustrates many of the challenges faced by our profession.

Even with all the work that has been done, there is no sense in which we would claim to have fully captured the operating complexity of the telecommunication industry, even for one ILEC. Still the methods described by Dr. Mallows fit the principles agreed to by the parties and hence can be considered acceptable, if not optimal. In our view they meet the dictum of Tukey of being "roughly right."

ACKNOWLEDGMENTS

I thank Sandy Balkin, Susan Hinkins and Fritz Scheuren for their help in preparing my comments.

Comment

Daniel R. Shiman

Colin Mallows has played an important role in the restructuring of the local telephone market. As one who has followed developments in the industry, I will provide some background information on the regulatory environment in which this work has been done, and then discuss some of the issues he raises and some of the challenges facing statisticians in this area.

For most of the twentieth century the telephone service market was treated as a natural monopoly: prices and service quality were carefully regulated by government agencies and competition was discouraged from entering the market. AT&T was the dominant player in this market, owning most of the local phone lines (about 80% of the nation's lines), the local exchanges serving those lines, and virtually all of the long-distance market for connecting phone calls between cities. AT&T's monopoly position began to crumble in the 1970s, due to MCI's entry into the long-distance market and the Department of Justice's antitrust lawsuit against AT&T. Because of this pressure AT&T agreed in 1982 to a settlement of the suit in the modified final judgement (MFJ). Under the MFJ, AT&T agreed to divest itself of its local exchange telephone companies, which were considered the source of its monopoly power, while it kept its long-distance and manufacturing businesses. These local exchange companies were organized into seven regional Bell operating companies (known as BOCs, RBOCs or baby Bells). The BOCs kept their monopoly in the local exchange market, but were not permitted to enter the long-distance market and had to provide equal access to all long-distance carriers.

As a result of the MFJ, by the mid-1990s the long-distance market, defined as calls made between local calling areas called local access and transport areas (LATAs), had become essentially a competitive market. Entry was fairly easy and local phone companies were required to provide equal access to their networks to all long-distance carriers. Meanwhile local phone companies, which controlled the local lines (called loops) that allowed customers access to the phone network from their telephones and the switches that routed local phone calls, retained monopoly control of the local exchange market. This split between local and long-distance markets, each with its own set of companies and regulations, is unique to the United States. Elsewhere the local exchange provider also provides domestic long-distance service on an integrated basis under a single regulatory regime.

The Telecommunications Act of 1996 (the Act) called for a broad restructuring of the local telephone exchange industry to open it to competition. Companies that had in the past provided local phone service [known as incumbent local exchange carriers (ILECs)] were required to open up their networks for use by competitors attempting to enter the local exchange market [called competitive local exchange carriers (CLECs)]. Specifically, under Section 251 of the

Daniel R. Shiman is an Economist in the Competition Policy Division of the Wireline Competition Bureau at the Federal Communications Commission, Washington, DC 20554 (e-mail: DSHIMAN@fcc.gov). Opinions expressed are those of the author alone and do not represent the views or policies of the FCC or its commissioners.

Act, ILECs are required, among other things, to interconnect with CLECs (so that CLEC customers could call ILEC customers and vice versa), allow CLECs to resell ILEC retail telecommunications services at a discount and provide access to piece parts of their networks as separate unbundled network elements (called UNEs) at cost-based rates. The Federal Communications Commission (FCC) and state public utility commissions have had to develop detailed regulations to implement the Act, due to the extraordinary complexity of the industry, the significant degree of restructuring of the industry called for by the Act and the natural reluctance of ILECs to negotiate interconnection agreements with CLECs and invest substantial resources in changing their systems and processes to enable competitors to enter their market.

Thus Section 251 effectively provides for three modes of entry for CLECs into the local exchange market, depending on how much of their own facilities they provide: full facilities, partial facilities and resale of the ILECs' services. Full facilities-based carriers provide all of their own facilities and thus require only interconnection with the ILEC. Partial facilities-based carriers purchase use of parts of the ILECs network as UNEs, usually the most expensive parts of the network to duplicate such as the loop, and combine those with elements of their own network that they have built. Resale carriers resell the ILEC's services at a modest discount from the ILEC's retail prices. CLECs are also allowed to purchase from the ILEC all of the elements of the network necessary to provide customer service, without facilities of their own, which is called the UNE platform (UNE-P). In a technical sense UNE-P is identical to resale. UNE-P is especially popular with CLECs, both because it is a faster and easier method of entering the market than constructing their own facilities, and because the price they pay the ILEC for the package of service, which is based on the cost of building the UNEs, is usually significantly lower than purchasing the services through resale.

The three key facilities needed to provide local phone service are the loop, the switch and transport. The loop is the copper wire or optic fiber connection running from the customer's premises to the local phone company's central office. The switch, located in the central office, is used to route traffic to the proper destination. Interoffice transport, which is carried over optic fiber running from the central office to other local central offices or to long-distance carriers' points-of-presence (POPs), carries traffic destined to customers served by other central offices or located in other LATAs.

The Act places a particular focus on opening up the networks of the largest local phone companies, the BOCs, which were the local phone companies spun off from the breakup of AT&T in 1984. After several mergers there are now four BOCs, which control 93% of the nation's phone lines. Under Section 271 of the Act, the BOCs are permitted to enter the long-distance market in their region if they successfully demonstrate to the FCC that they have met a series of legal requirements, which effectively measure whether they have opened up their local wireline markets to competition. Because they perceived that entering the long-distance market would be profitable, the BOCs have invested significant resources in making the necessary changes to their systems, negotiating interconnection agreements with CLECs and attempting to gain state and FCC approval of their applications to provide long-distance service. To demonstrate compliance with the market-opening provisions of the Act to state and federal regulators, the BOCs have agreed to, among other things, make UNEs and interconnection available at state-approved cost-based rates; provide performance metric data according to definitions and standards set by the state commission, to demonstrate they are providing nondiscriminatory service to CLECs; abide by a performance plan, set up by the state commission, which provides for automatic payment of penalties if the BOCs provide poor performance to CLECs, as measured by the performance metrics; and submit to a third party test, in which an independent evaluator assesses whether the BOCs systems have been opened to CLEC use. Much of this was set up by state commissions in open proceedings and collaborative workshops with BOC and CLEC participation.

Thus a new regulatory regime is being developed to ensure ILEC compliance with their obligations under the Act. A cornerstone of this new regime is the development of an extensive scheme of performance metrics [called service quality measurements (SQMs) by some]. The metrics measure the performance of ILECs in handling a variety of tasks, such as how quickly and efficiently they process and provision orders from CLECs for service. These metrics have been developed at the state level and can vary substantially in their definitions and standards from state to state (although they are usually fairly similar for states in a region). Most state plans have about 40–70 metrics, and most of these metrics are disaggregated by the different

kinds of services provided, yielding 500–2500 submetrics (750 seems typical).

Much work has gone into creating a set of statistical tests to help evaluate this large set of performance metrics. Statistical tests have been developed for three purposes: (1) to assist state regulators and the FCC in evaluating the metric data measuring the ILECs' commercial performance in providing services to the CLECs; (2) for use in performance plans; and (3) to help the third party tester evaluate the ability of ILECs to process CLEC orders. The more intensive negotiations have typically concerned the statistical tests used in the evaluation of commercial data and in the performance plan. Because the third party test is similar to a controlled experimental study, the design of statistical tests there tends to be less controversial and often is left to the tester.

Dr. Mallows points out the importance of defining "parity" before choosing a statistical test to determine if parity has been violated. The Act requires the ILEC to provide service "that is at least equal in quality to that provided by the local exchange carrier to itself ... or any other party ... on rates, terms, and conditions that are just, reasonable, and nondiscriminatory" [47 U.S.C. Section 251(c)(2)]. For services which the ILEC provides to both CLECs and itself, such as provisioning new phone service to CLEC and retail customers, ILEC performance to CLECs can be compared with the performance it provides to itself, to determine that service to CLECs is at least equal in quality, that is, there is parity. For services where there is no retail analog, a benchmark is used to determine if the ILEC is providing an acceptable level of performance [the FCC adopted the standard that it should provide CLECs "a meaningful opportunity to compete" in Ameritech Michigan 271 Order, 12 FCC Rcd (1997) at 20618-19].

When designing statistical tests, it is also important to understand how regulators intend to use the performance metrics. In their evaluation of commercial data to make a legal determination whether there is discrimination, regulators have not generally adopted the approach advocated by some to use an automatic rule, relying heavily on statistical tests, to test for parity. Instead statistical tests have been used mostly to rule out that random chance caused an observed difference in performance for a parity metric. For example, the FCC in its evaluation of BOC performance for Section 271 applications has not relied solely on statistical significance in its determination whether there is discrimination, but has also considered whether an observed difference in performance would have been "competitively significant," that is, have a competitive impact [Bell Atlantic New York 271 Order, 15 FCC Rcd (1999) at 3976, paragraph 59].

The legal, informational and process constraints regulators labor under must also be taken into consideration. Because analysis and decision-making will often be vested in nonstatisticians, for the evaluation of commercial data a regulator will often prefer a statistical test that is easy to understand and apply. Tests involving a simple rule, with a straightforward and unambiguous result that laypeople can apply, while allowing a clear interpretation of the original metric data, are preferred. The widely used "modified Z test" has these properties.

In the design of a performance plan, technical issues must be dealt with up front, since there is ordinarily no judgement involved in determining payments while the plan is in operation. This requires more careful thought on how to handle all the technical issues, but it also gives more freedom to rely on complicated technical techniques that are difficult to explain to laypeople, such as the "balanced averaged disaggregated truncated adjusted modified Z" plan (my name for it) that Dr. Mallows describes in the disaggregation and balancing sections of his paper, which he helped develop and is now in use.

Statisticians have to recognize that solutions must be workable and easy to implement. I have seen proposed solutions that were technically infeasible or required too much information to be practical. Not all problems can be solved by statisticians alone. For some problems it is important to get input from people familiar with the technical, business and legal sides of the issues, such as in the choice of the alternative hypothesis in a balancing approach.

Dr. Mallows points out the inherent conservatism of commissions. Yet by necessity this regulatory arena is fairly open to innovative statistical techniques, because this is a new regulatory regime with its own set of standards, the problems encountered are unusual and often lack textbook solutions (partly because the data are nonnormal, the sample sizes are often small and results can be significantly affected by correlation and confounding factors) and the range of solutions to be considered must fit legal and process constraints. A variety of state commissions are working on these problems, suggesting there is room for different viewpoints and experimentation. For many of the statistical techniques he describes, Dr. Mallows played an important role in their development and adoption. The nonstandard

modified t-test has been used by most state commissions and the FCC for parity metrics, under the name "modified Z-test" since it is usually used for large sample sizes when the mean is approximately normally distributed [Bell Atlantic New York 271 Order, 15 FCC Rcd (1999) at 4182, Appendix B].

There are many interesting challenges that remain for statisticians. For example, there is the problem of finding the right level of disaggregation. Too much can lead to small sample sizes, inconclusive results and an excess of information to evaluate, while too little can lead to Simpson's paradox (in which aggregation can reverse the direction of the relationship between two variables). Since many CLECs are located in particular niches of the market, such that the geographic and product distribution of their customers is quite different from that of the ILEC, Simpson's paradox is a serious concern here. Dr. Mallows' solution of taking a weighted average of truncated disaggregated Z scores may work for a performance plan, but regulators may not find it as useful for evaluating commercial data, since it involves a significant transformation of the data and is difficult to explain to laypeople.

In the design of performance plans one of the most difficult problems has been the control of the overall type-1 error rate. Many plans attempt to minimize spurious payments by the ILEC for metrics for which it is providing parity, either through reducing the type-1 error rate for individual metrics or by using a plan of forgiveness of a certain number of metric failures. The calculations usually assume that under the null hypothesis $\mu_A = \mu_B$ for each metric, yet it has been pointed out (Initial report on OSS performance results replication and assessment, in proceeding R.97-10-016/I.97-10-017, California PUC Telecommunications Division, June 15, 2001) that p values for many metrics for a given period for an ILEC were not distributed as predicted, even for metrics with good performance to CLECs. It appears then that the correct null is $\mu_A \geq \mu_B$. If one of the goals of a plan is to avoid "cancellation," in which good performance on one metric cancels out bad performance on another, then calculating the proper number of forgivenesses becomes difficult if the ILEC may be providing superior performance to CLECs in a large number of less important areas of service.

The choice of performance metrics and statistical tests used could have a significant impact on the growth of competition in this industry, and on hundreds of millions of dollars in potential revenue (from CLEC entry into the local market and BOC entry into long distance) and possible penalty payments (from the penalty plans). There is much room for innovative work to find new methods of testing for parity under the constraints described here and to better present the data to decision-makers with new kinds of summary statistics and graphs.

Rejoinder

Colin Mallows

I thank the discussants for their kind words, and for the additional background and references.

Dan Shiman explains what the problem looks like from the perspective of the FCC. While to the ILECs and CLECs the commissions may seem to be all-powerful, they also are subject to hobbling constraints. I strongly endorse his final comment that there are still many opportunities for innovative work in this area.

Gastwirth and Miao have performed an interesting Monte Carlo study. Presumably their beta index gives the scale of the alternative being considered. Unfortunately legal proceedings have tremendous inertia, so that once a procedure such as modified t has been accepted, it is difficult to make changes. (But not impossible; the idea of using a permutation test was introduced at a late stage. As Dan Shiman remarks, some commissions do recognize that we are in uncharted territory.)

Gastwirth and Miao's remarks on the adversarial environment are very much to the point. Regarding the disclosure issue, ILECs routinely label their data as "company private," so that as a statistician working for a CLEC, I had no access to it; this made it difficult to identify the crucial technical problems and to make reasoned arguments. An ILEC may have no intention to deceive; it has sound business reasons for being reluctant to share its data. At one stage the Louisiana Commission recognized this difficulty and I was allowed access to some BellSouth data, but only after I had signed a nondisclosure agreement. No one

else at AT&T was allowed to see the data or even my report until it had been released by BellSouth, and I had to work on a PC that was not connected to any network, and was "scrubbed" when I was finished. It was the experience of working with these data that led to many of the ideas that I have described in the paper and eventually to resolution of almost all the technical issues in the Louisiana proceedings.

I welcome Ed Mulrow's description of the stages he and his colleagues went through in arriving at their present positions. He and I started out as adversaries in formal proceedings, but once we were able to work together in scientific mode we found it possible to agree on almost all the technical issues. Our collaboration has been enjoyable and productive.

ADDITIONAL REFERENCES

BALKIN, S. D. (2001). Testing the compliance hypothesis in a self-effectuating system. *ASA Proceedings of the Quality and Productivity Section*. Amer. Statist. Assoc., Alexandria, VA.

BURNS, R. P. (1999). *A Theory of the Trial*. Princeton Univ. Press.

DAWSON, D. (1980). Are statisticians being fair to employment discrimination plaintiffs? *Jurimetrics Journal* **20** 1–20.

DRION, E. F. (1952). Some distribution-free tests for the difference between two empirical cumulative distribution functions. *Ann. Math. Statist.* **23** 563–574.

FREIDLIN, B. and KORN, E. L. (2002). A testing procedure for survival data with few responders. *Statistics in Medicine* **21** 65–78.

GASTWIRTH, J. L. (1991). Comment on Nesson. *Cardozo Law Review* **13** 817–829.

GASTWIRTH, J. L. (1997). Statistical evidence in discrimination cases. *J. Roy. Statist. Soc. Ser. A* **160** 289–303.

GASTWIRTH, J. L. and MAHMOUD, H. (1986). An efficiency robust nonparametric test for scale change for data from a gamma distribution. *Technometrics* **28** 81–84.

GEISSER, S. (2000). Statistics, litigation, and conduct unbecoming. In *Statistical Science in the Courtroom* (J. L. Gastwirth, ed.) 71–85. Springer, New York.

GNEDENKO, B. V. and MIHALEVIC, V. S. (1952). On the distribution of the number of excesses of one empirical distribution function over another. English translation (1961). *Selected Translations in Mathematics, Statistics and Probability* **1** 83–85. Amer. Math. Soc., Providence, RI.

GRAUBARD, B. I. and KORN, E. L. (2002). Inference for superpopulation parameters using sample surveys. *Statist. Sci.* **17** 73–96.

GRUNWALD, M. (2002). Monsanto held liable for PCB dumping. *Washington Post*, Feb. 23, A1.

HINKINS, S. (2001). Statistical issues and responses to the mandate of the Telecommunications Act of 1996. *ASA Proceedings of the Quality and Productivity Section*. American Statistical Association, Alexandria, VA.

HORTON, R. (2001). The clinical trial: Deceitful, disputable, unbelievable, unhelpful, and shameful—What next? *Controlled Clinical Trials* **22** 593–604.

HUNTER, J. E. and SCHMIDT, F. L. (1995). *Methods of Meta-Analysis*: *Correcting Error and Bias in Research Findings*. Sage, Thousand Oaks, CA.

KAYE, D. H. (2001). The dynamics of *Daubert*: Methodology, conclusions, and fit in statistical and econometric studies. *Virginia Law Review* **87** 1933–2018.

KISH, L. (1965). *Survey Sampling*. Wiley, New York.

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York.

MALLOWS, C. (1998). The zeroth problem. *Amer. Statist.* **52** 1–9.

MULROW, E. J. (2001b). Balancing type I and II error probabilities: Making materiality an integral part of hypothesis testing. *ASA Proceedings of the Quality and Productivity Section*. American Statistical Association, Alexandria, VA.

NATIONAL RESEARCH COUNCIL (2002). *The Age of Expert Testimony*: *Science in the Courtroom*. National Research Council, Washington, DC.

NESSON, C. R. (1991). Incentives to spoliate evidence in civil litigation: The need for vigorous judicial action. *Cardozo Law Review* **13** 791–816.

PERRITT, Jr., H. H. (2001). *Law and the Information Superhighway*. Aspen Law and Business, Gaithersburg, MD.

POSNER, R. A. (2001). *Frontiers of Legal Theory*. Harvard Univ. Press.

ROSENBAUM, P. (2002). *Observational Studies*, 2nd ed. Springer, New York.

ROSENBLUM, M. (2000). On the evolution of analytical proof, statistics, and the use of experts in EEO litigation. In *Statistical Science in the Courtroom* (J. L. Gastwirth, ed.) 161–194. Springer, New York.

RUBIN, D. B. (2000). Statistical inference for causal effects in epidemiological studies via potential outcomes. XL Scientific Meeting of the Italian Statistical Society, Specialized Session 9. Available at http://www.ds.unifi.it/sis2000/estese/estese_e.htm

SOLOMON, D. (2002). States probe Qwest's secret deals to expand long-distance service. *Wall Street Journal*, April 29, A1.

TAKACS, L. (1964). An application of a ballot theorem in order statistics. *Ann. Math. Statist.* **35** 1356–1358.

YOUNG, S., DREAZEN, Y. and BLUMENSTEIN, R. (2002). How effort to open local phone market helped the baby Bells. *Wall Street Journal*, Feb. 11, A1.