

STEPTOE & JOHNSON ^{LLP}

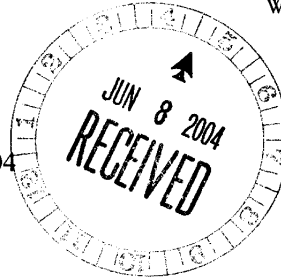
ATTORNEYS AT LAW

Samuel M. Sipe Jr.
202.429.6486
ssipe@steptoe.com

1330 Connecticut Avenue, NW
Washington, DC 20036-1795
Tel 202.429.3000
Fax 202.429.3902
steptoe.com

211318

June 8, 2004



Via HAND DELIVERY

The Honorable Vernon A. Williams
Secretary
Surface Transportation Board
1925 K Street, NW
Washington, DC 20423

ENTERED
Office of Proceedings

JUN 08 2004

Part of
Public Record

PUBLIC VERSION

Re: **STB Docket No. 42071, *Otter Tail Power Company v. The Burlington Northern and Santa Fe Railway Company***

Dear Secretary Williams:

Enclosed for filing on behalf of defendant The Burlington Northern and Santa Fe Railway Company ("BNSF") in the above-captioned matter are the original and ten copies of the Public version of Defendant's Motion to Strike. BNSF is also filing three copies of a CD containing a PDF version of this Motion. BNSF is filing a Highly Confidential version of this Motion under separate cover today.

Please date stamp the extra copy of this cover letter and Motion and return it to the messenger who delivered this filing.

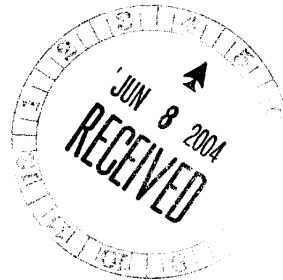
Sincerely,

Samuel M. Sipe, Jr.

Enclosures

cc: Counsel for Complainant

**BEFORE THE SURFACE
TRANSPORTATION BOARD**



OTTER TAIL POWER COMPANY,

Complainant,

v.

**THE BURLINGTON NORTHERN AND
SANTA FE RAILWAY COMPANY,**

Defendant.

Docket No. 42071

DEFENDANT'S MOTION TO STRIKE

Pursuant to 49 CFR §1117.1, Defendant The Burlington Northern and Santa Fe Railway Company ("BNSF") hereby moves to strike new evidence that was submitted by the Complainant Otter Tail Power Company ("Otter Tail") in its April 29, 2004 Rebuttal Evidence. The new evidence that is the subject of this Motion falls into two categories: new evidence that exceeds the scope of proper rebuttal in a SAC case and new evidence that does not meet the standards for the admissibility of expert testimony. The first category of new evidence -- evidence that exceeds the proper scope of rebuttal -- includes the following: (a) evidence relating to Otter Tail's purchases of coal from certain southern PRB mines; (b) evidence relating to SARR capacity and operating statistics based on Otter Tail's new rebuttal string program -- the fourth string program it has used in this proceeding; (c) Otter Tail's changes to the SARR's yard configuration; (d) new evidence on the inflation index to be used in the DCF analysis; and

(e) Otter Tail's modified calculation of the SARR's cost of capital. The second category -- evidence that does not meet the admissibility standards for expert testimony -- includes Otter Tail's statistical analysis of rerouted traffic, contained in Otter Tail's Rebuttal Narrative at III-A-49 to III-A-52. These issues are discussed further below.

I. LEGAL STANDARDS GOVERNING NEW EVIDENCE ON REBUTTAL

The Board has established strict limits on the submission of new evidence on rebuttal. The complainant in a SAC case has the burden of proof and the right to open and close the evidentiary record. However, to streamline SAC proceedings and to ensure that a reliable record is created through the adversarial testing of evidentiary submissions, the Board limits the scope of evidence that can be submitted by the complainant on rebuttal. Several recent decisions by the Board have addressed the scope of rebuttal in SAC cases and have clarified the rules governing the submission of new evidence on rebuttal.

The first and most basic rule governing rebuttal is that a complainant's rebuttal filings cannot be used as an opportunity to submit evidence that a complainant could have and should have presented on opening.¹ This rule follows directly from the fact that the complainant in a rate case has the burden of proof on SAC issues. As the party with the burden of proof, the complainant must present its entire case-in-chief on opening.² Moreover, the complainant cannot rely on a bare bones recitation of the assumptions underlying its SAC calculations.

¹ *General Procedures for Presenting Evidence in Stand-Alone Cost Rate Cases*, Ex Parte No. 347 (Sub-No. 3), slip op. at 5 (served Mar. 12, 2001) ("*General Procedures*"); *Duke Energy Corp. v. CSX Transp., Inc.*, STB Docket No. 42070, slip op. at 3 (served March 25, 2003) ("*Duke/CSX P*").

² *General Procedures* at 5; *Duke Energy Corp. v. Norfolk Southern Ry. Co.*, STB Docket No. 42069, slip op. at 13-4 (served Nov. 5, 2003) ("*Duke/NS*"); *Wisconsin Power and Light v. Union Pacific R.R. Co.*, STB Docket No. 42051, slip op. at 72 (served Sept. 12, 2001).

Complainants must submit in their case-in-chief the support for their SAC assumptions.³

Logically, this is particularly true where complainants seek to test the boundaries of the SAC standard or to present unprecedented SAC assumptions. Allowing a complainant to hold back evidence supporting its SAC assumptions until rebuttal would encourage gamesmanship by the complainant and would deprive the defendant of a meaningful opportunity to reply to the complainant's case.⁴

The second rule governing rebuttal evidence is that a complainant cannot significantly alter the configuration of the SARR or alter the core assumptions on which its case-in-chief is based.⁵ Thus, in *Duke/CSX I* the Board struck from the record new evidence submitted by the complainant on rebuttal relating to a redesigned tunnel and a new yard location on grounds that the evidence improperly changed basic SAC assumptions in the complainant's opening evidence.⁶ The Board stated that if the complainant has a valid reason to seek to change the configuration of its SARR or the core assumptions underlying its SAC calculations, it must seek permission from the Board to introduce new evidence. Moreover, if permission to make such

³ *Duke/NS* at 15 ("[T]he shipper must plan to submit its best, least-cost, fully supported case on opening. It may not hold back to see the railroad's reply evidence before finalizing or supporting its own case, as an opportunity to correct deficiencies in its opening evidence is not assured.").

⁴ As the Board stated in a recent decision: "We are increasingly troubled by the submission of incomplete or erroneous evidence on opening in a SAC case and a complainant's reliance upon an opportunity to address deficiencies through later evidentiary submissions, to which a defendant has no opportunity to respond. The interests of fairness and orderly handling of a case dictate that parties submit their best evidence on opening, so that each party has a fair opportunity to reply to the other's evidence." *Public Service Co. of Colorado d/b/a Xcel Energy v. BNSF*, STB Docket No. 42057, slip op. at 2 (served April 4, 2003) ("*Xcel*").

⁵ *Duke/NS* at 15; *Duke/CSX I* at 3, *Duke Energy Corp. v. CSX Transp., Inc.*, STB Docket No. 42070, slip op. at 14 (served Feb. 3, 2004) ("*Duke/CSX II*").

⁶ *Duke/CSX I* at 4.

changes is granted, “[f]airness dictates that [the defendant] would need an opportunity to address the feasibility and cost of a new SARR design and how that change might alter other evidence in the record.”⁷

In addition to these two fundamental rules, the Board has identified three permissible responses by a complainant to a defendant’s reply evidence.⁸ First, the complainant may present further argument in support of the assumptions it made on opening in response to a defendant’s criticism of the complainant’s opening evidence. However, to be consistent with the first rule governing rebuttal evidence described above and to prevent a sandbagging of defendant, the complainant cannot submit new evidence in support of its original assumptions unless it first shows that it could not have anticipated the need for such supporting evidence when it filed its opening evidence. Second, the complainant can accept the defendant’s alternative evidence and incorporate it into the complainant’s rebuttal SAC calculations. Third, where the defendant has shown that the complainant’s opening evidence was deficient, the complainant can modify its opening assumptions or methodologies to cure the deficiency but only if it first demonstrates that any alternative proposed by the defendant is infeasible, unrealistic or unsupported.⁹ In other words, once a defendant shows that the complainant’s opening evidence is deficient on a

⁷ *Id.* at 4.

⁸ *Duke/NS* at 14-5.

⁹ *Id.* See also *Carolina Power and Light v. Norfolk Southern Ry. Co.*, STB Docket No. 42072, slip op. at 69-70 (served Dec. 23, 2003) (“*CP&L*”) (holding that the alternative rebuttal evidence offered by the complainant that raised its MOW staffing levels by more than 60% may not be considered by the Board where the complainant did not demonstrate that the railroad’s alternative MOW staffing proposal was unrealistic or infeasible); *Duke/NS* at 49-50 (holding that the railroad’s yard configuration would be used and that the changes made by the complainant on rebuttal would be disregarded because it had not demonstrated that the railroad’s yard configuration would be “inappropriate” for meeting SARR needs).

particular issue, the defendant is entitled to a presumption that its alternative SAC evidence on that issue is reasonable, and the complainant must overcome that presumption before it can submit new evidence on the issue.

In several areas described below, Otter Tail's rebuttal evidence oversteps the boundaries of permissible rebuttal evidence.

II. THE BOARD SHOULD STRIKE OTTER TAIL'S EVIDENCE ON HISTORICAL MOVEMENTS OF COAL FROM MINES SOUTH OF CORDERO

As BNSF explained in its October 8, 2003 Reply Evidence, the issue traffic that Otter Tail included in the stand-alone network did not include any coal movements originating from PRB mines south of Cordero. *See* BNSF Reply Nar. at III.A-20. Otter Tail routed all issue traffic north out of the PRB; therefore that traffic does not share any facilities with the large volume of coal traffic included in the OTRR traffic group that originates at southern PRB mines (from Cordero to Antelope) and is routed south from these mines to Converse Yard. Since the issue traffic does not share any facilities with this southern PRB mine traffic, the southern PRB mine traffic should be excluded from the traffic group under the traffic grouping principles established in the Board's *Coal Rate Guidelines* and subsequent case law. *Id.* at III.A-20 to III.A-23.

In its rebuttal, Otter Tail takes issue with BNSF's statement that this southern PRB mine traffic does not share facilities with the issue traffic and it presents new evidence, not included on opening, regarding historical movements of coal to the Big Stone plant. Specifically, Otter Tail asserts that historically it shipped coal to Big Stone from two southern PRB mines *not included as mine origins for the issue traffic* in its SAC presentation -- the Black Thunder and Rochelle mines. Otter Tail claims that small volumes of Black Thunder coal were shipped to Big Stone in 2000, 2001, and 2003 and a small volume of coal (500 tons, according to Otter

Tail's electronic workpaper) was shipped from Rochelle to Big Stone in 1988. *See* Otter Tail Reb. Nar. at III-A-24.

Otter Tail appears to believe that this new evidence of historical coal movements from southern PRB mines redeems its inclusion in the OTRR shipper group of the southbound movements of coal originating at the southern PRB mines. This premise is wrong. Any evidence that Otter Tail used or will use southern PRB mine coal is irrelevant to Board's consideration of Otter Tail's SAC evidence. Otter Tail made the assumption in its SAC presentation (both on opening and on rebuttal) that no issue traffic would originate at those southern PRB mines. Regardless of the reason for basing its SAC presentation on such an assumption (which could include {

}, lack of any plans to use such coal in the future, or a desire to simplify the SARR's operating plans), Otter Tail based its evidence on an assumption that no issue traffic originated at the southern PRB mines and it cannot change that assumption now.

Even if the new evidence was potentially relevant to Otter Tail's SAC presentation, Otter Tail's evidence of historical shipments of coal from Black Thunder and Rochelle to Big Stone constitutes improper rebuttal. Otter Tail's SARR involved a significant departure from established SAC principles and precedent by having a large volume of cross-over traffic in the stand-alone shipper group that shared no facilities with the issue traffic. As BNSF pointed out in its Supplemental Reply Narrative, the *Coal Rate Guidelines* explained that the purpose of the SAC test is "to compute the rate a competitor in the market-place would need to charge in serving a captive shipper or group of shippers who benefit from sharing joint and common

costs.”¹⁰ Sharing of costs entails the sharing of facilities. Thus, the sharing of facilities is a fundamental requirement governing the grouping of traffic in a SAC test and that requirement has been in place since the inception of the SAC test. If Otter Tail had evidence when it filed its Opening Evidence that it believed supported the inclusion of traffic in the stand-alone shipper group notwithstanding that the traffic shared *no* facilities with the issue traffic, it should have presented that evidence on opening. By holding back evidence on this issue until rebuttal, Otter Tail deprives BNSF of an opportunity to respond fully to the evidence and thereby precludes the development of an adequate record on the critical issue of the SARR’s configuration and traffic group.

One of the most compelling reasons for limiting the complainant’s submission of new evidence on rebuttal is that the complainant’s knowledge that the defendant does not have an opportunity to reply to the evidence may encourage the submission of misleading evidence on rebuttal. In fact, the new evidence that Otter Tail has submitted here is highly misleading. Otter Tail states that it “has shipped PRB coal originated at Rochelle mine to Big Stone” in its narrative but fails to mention that it only shipped 500 tons of coal from that mine in 1988, 14 years before the hypothetical OTRR begins operations.¹¹ As to coal shipped from the Black Thunder mine to Big Stone, the volumes shipped in 2000 were test volumes and, according to documents produced by Otter Tail in discovery, that test burn was {

¹⁰ *Coal Rate Guidelines*, 1 I.C.C.2d at 528. See BNSF Supp. Reply Nar. at III.A-7 to III.A-13.

¹¹ See Otter Tail Rebuttal electronic workpaper “FERC PRB Big Stone.xls.”

}¹² This test burn result suggests that {
} And while Otter Tail claims that 136,000 tons of
Black Thunder coal was shipped to Big Stone in 2001, its supporting FERC data (confirmed by
the traffic data produced by BNSF in this case) shows only one train of Black Thunder coal
moving to the plant in 2001, consistent with the { }

The most troubling aspect of Otter Tail's new evidence is the coincidence that after
moving no Black Thunder coal since 2001 (and after Otter Tail chose not to include any Black
Thunder coal originations in the SAC presentation) Otter Tail suddenly shipped four trainloads
of Black Thunder coal between October 18, 2003 and October 22, 2003,¹³ 10 to 14 days *after*
BNSF filed its October 8, 2003 Reply evidence. In that Reply, BNSF challenged the inclusion of
southern PRB coal traffic that exits the OTRR at Converse on the basis that this southern PRB
traffic -- including large volumes of traffic originating at Black Thunder -- shares no facilities
with the issue traffic moving on the OTRR. The timing of these four Black Thunder shipments
is particularly suspect {

¹² See July 12, 2000 Memo from Jeff Endrizzi to Tyler Albertson at page 1. This document was produced by Otter Tail at OTP 002345 to 2353 and is included as Exhibit 1.

¹³ See Verified Statement of Stephen Behn (hereafter "Behn V.S.") at 2. This verified statement is included as Exhibit 2.

¹⁴ See Behn V.S. at 1.

} shipped four train loads of coal from Black Thunder. No additional shipments from Black Thunder to Big Stone have occurred since October 22, 2003.¹⁶

Given the possibility that the October 2003 shipments were intended to game the SAC results by creating the appearance that Otter Tail intended to source future coal needs from the Black Thunder mine, the Board cannot give any weight to these movements of Black Thunder coal.

III. THE BOARD SHOULD STRIKE OTTER TAIL'S REBUTTAL EVIDENCE BASED ON ITS NEW STRING PROGRAM.

On rebuttal, Otter Tail uses the *fourth* string program that has been presented in this case to assess capacity requirements for the OTRR and to calculate operating statistics. Otter Tail's June 13, 2003 Opening Evidence was based on a string program that has now been superceded three times. Even before BNSF had a chance to respond to Otter Tail's June 13, 2003 string program analysis, Otter Tail identified flaws in its string program and submitted a new string program in its July 14, 2003 Errata filing. These changes, along with other changes made to Otter Tail's Opening Evidence, necessitated a significant delay in the filing of BNSF's Reply Evidence. Otter Tail then submitted a new string program in support of its January 9, 2004 Supplemental Opening Evidence. Otter Tail called this third string program its "rebuttal" string program, suggesting that no more changes would be made to the string program. However, in

¹⁵ *Id.* at 2.

¹⁶ *Id.*

response to BNSF's March 22, 2004 Supplemental Reply Evidence, Otter Tail made further changes to the string program which it incorporated in its April 29, 2004 Rebuttal Evidence as its fourth string program analysis in this case.

The Board cannot accept any evidence that is based on a new string program presented for the first time on rebuttal. As noted in the attached verified statement of BNSF's experts Karl Leopold and Terence Monks, the new string program contains over one hundred changes to the lines of code as compared to the so-called "rebuttal" string program that Otter Tail claimed would be the final iteration of the string program in this case.¹⁷ That prior string program already included substantial changes from the string programs used in Otter Tail's Opening Evidence and Errata filing. A program that must be changed every time it is used is inherently unreliable. In other cases where complainants have used the string program developed by L.E. Peabody & Associates, the same pattern has emerged -- the complainant uses a string program on opening but replaces it with another string program on rebuttal when flaws in the program are identified by the defendant.¹⁸ The persistent flaws in the string program render it unusable in a SAC analysis.

Even if the string program were not inherently unreliable, the Board could not accept a new string program on rebuttal because of the difficulty in evaluating the output and statistics generated by the string program. As BNSF has explained previously, a fundamental problem

¹⁷ See Verified Statement of Karl Leopold and Terrance Monks ("Leopold/Monks V.S."), included as Exhibit 3.

¹⁸ See, e.g., Complainant's Supplemental Rebuttal Evidence, *Arizona Electric Power Coop., Inc. v. The Burlington No. and Santa Fe Ry. Co. and Union Pac. R.R. Co.*, STB Docket No. 42058, at III-C-11 to 12 (filed April 2, 2004); Complainant's Rebuttal Evidence, *Public Service Co. of Colorado d/b/a Xcel Energy v. The Burlington No. and Santa Fe Ry. Co.*, STB Docket No. 42057, at III-B-35 to 36 (filed May 19, 2003); Defendant's Brief, *Duke Energy Corp v. CSX Transp., Inc.*, STB Docket No. 42070, at 18 (filed April 14, 2003).

with the string program is its lack of transparency. The operation of trains in the string program cannot be observed and the operating statistics for the modeled trains are produced and aggregated through an obscure and unreliable process that has produced erroneous results in the past. As a result, the validity of a string program analysis and the statistics generated by the string program can be evaluated only with a detailed and extensive review of the program code and the output files. While it requires a massive effort by the defendant to carry out this evaluation, it is possible for a defendant to carry out such an analysis on reply when the string program is submitted in the complainant's opening evidence. The submission of a new string program on rebuttal precludes this evaluation.

The difficulty in evaluating the string program output is compounded by the fact that the string program does not automatically identify all train conflicts. As BNSF confirmed at the string program conference in January 2004, a manual review must be made each time the string program is run to determine whether the model failed to identify train collisions. BNSF has found numerous instances in the past where mistakes were made in this manual review and train conflicts were overlooked. (Apparently, other defendants have found the same mistakes, thus producing the constantly changing string programs in other cases.) In fact, even without doing a thorough review here, BNSF has identified a collision that Otter Tail's consultants failed to identify or resolve. *See Leopold/Monks V.S.* at 2.

Moreover, when train conflicts are identified, they must be resolved through manual intervention. The string program operator must make decisions after the program has been run about how to modify train movements to avoid conflicts such as overloaded sidings and yards. But the string program is never run again after these manual adjustments are made to determine whether the manual changes produced unintended ripple effects elsewhere on the SARR lines.

As a result, a massive audit must be made of each new string program run to determine whether problems were overlooked or whether manual changes produced new problems. The Board relies on the parties to carry out these complex evaluations of another party's evidence. The Board cannot be expected to carry out such a review when a new string program is submitted for the first time on rebuttal. But without such a review, the string program evidence cannot be relied upon for any purpose. Again, based on only a preliminary audit of Otter Tail's new string program analysis, BNSF identified instances in which Otter Tail's manual adjustments produced conflicts elsewhere on the system. *See Leopold/Monks V.S.* at 2-3.

Another major problem with the string program is the complex and unverifiable manner in which it aggregates operating statistics from the modeled operations of individual trains. As described in BNSF's Supplemental Reply Evidence, at the January 2004 string program conference, BNSF identified certain problems related to the string program's accumulation of loading time and its aggregation of operating statistics.¹⁹ One of the changes that Otter Tail made in its string program on rebuttal was to address this issue. Indeed, Otter Tail admits that the so-called rebuttal string program in its January 9, 2004 filing had this defect and it claims that it fixed the problem in its new rebuttal string program. Otter Tail submitted a reconciliation file that purportedly shows that the string program now accurately accounts for various time elements and accurately aggregates operating statistics. *Otter Tail Reb. Nar.* at III-C-29. However, as discussed in the attached *Leopold/Monks V.S.*, the Board cannot confirm the

¹⁹ Specifically, BNSF identified inconsistencies between the string program output files showing specific train movements and the output file summarizing the OTRR's operating statistics. Those inconsistencies confirmed that the string program failed to record loading time for all trains and incorrectly aggregated the OTRR's operating statistics. *See BNSF Supp. Reply. Narr.* At III.C-7 to III.C-9.

accuracy of Otter Tail's supposed reconciliation because Otter Tail's output files cannot be replicated using the string program instructions. *See Leopold/Monks V.S. at 3.*

In addition, Otter Tail admits on rebuttal that its first rebuttal string program failed to accumulate loading times on all trains and it purports to fix this problem in the new rebuttal string program by adding six hours of loading time to each empty train returning to the mine on the peak day. But Otter Tail's peak day is determined by the number of *loaded* trains departing the mines on that day, not by the number of *empty* trains returning to the mines. Otter Tail's use of empty trains to accumulate loading time for the peak day seriously distorts the operating statistics since the number of loaded trains on the peak day substantially exceeds the number of empty trains on that day. Otter Tail's string program analysis shows that 51 loaded trains were dispatched from the mines on the peak day versus 33 empty trains returning to the mines that day. *See Leopold/Monks V.S. at 3-4.* By adding loading time to 33 empty trains instead of the 51 loaded trains, Otter Tail substantially understated its locomotive and car requirements on its peak day and consequently understated its operating costs in the SAC analysis.

Due to the complexity of the string program, the Board cannot permit the complainant to offer a new string program analysis on rebuttal and thereby avoid the scrutiny that would be required before any weight could be given to the string program analysis. Since Otter Tail's capacity evidence and operating statistics are based on a string program analysis presented for the first time on rebuttal, all of Otter Tail's evidence on those issues should be struck.

IV. THE BOARD SHOULD STRIKE OTTER TAIL'S NEW EVIDENCE ON THE FARGO YARD

In its Opening Evidence, Otter Tail included in OTRR's traffic group a large volume of rerouted freight traffic that moves today on BNSF's northern line through Minot, North Dakota. As BNSF explained in its Reply Evidence, BNSF currently performs significant intra- and inter-

train switching operations on this traffic at Minot. BNSF Reply Nar. at III.B-43 to 45, III.C-7 to 9. Otter Tail rerouted this traffic onto the SARR's Glendive/Mandan line so that it no longer passes through the Minot Yard, but Otter Tail failed to make any provision for the intra-/inter-train switching that needs to be performed on the traffic. In BNSF's Reply Evidence, BNSF added the trackage required to perform this switching at the Glendive Yard, the principal yard on the northern lines of the OTRR. *Id.* at III.B-44 to 46; *see* Otter Tail Opening Nar. at III-B-8.

In BNSF's Supplemental Reply Evidence, BNSF excluded all rerouted non-coal traffic for reasons explained at III.A-20 to III.A-27 of BNSF's Supplemental Reply Narrative. As a result of the exclusion of this traffic, the OTRR would no longer need to perform intra- and inter-train switching on non-coal traffic at Glendive. Therefore, for purposes of BNSF's Supplemental Reply, BNSF eliminated the yard tracks at Glendive that were to be dedicated to this switching activity. *See* BNSF Supp. Reply Nar. at III.B-12.

On rebuttal, Otter Tail objected to the exclusion of the rerouted non-coal traffic and it included that traffic in its Rebuttal Evidence. However, it acknowledged the need to perform intra- and inter-train switching on this traffic as well as the need for facilities to carry out these functions. Otter Tail Reb. Nar. at III-B-15. It did not accept BNSF's evidence on the facilities required to handle this switching activity but relocated the additional track provided for this intra-/inter-train switching from the Glendive Yard to the Fargo Yard, and it reduced the total amount of additional track provided for this switching from the 10.1 miles provided by BNSF to 6.7 miles of track. *Id.* at III-B-15 to 16.

Otter Tail is not entitled to modify the configuration of its SARR on rebuttal. As noted above, the Board in *Duke/CSX I* found that configuration changes like that proposed by Otter

Tail here are not permissible on rebuttal.²⁰ The Board made it clear in that decision that a complainant must first seek permission from the Board to make changes in the configuration of its SARR. Otter Tail did not follow the prescribed procedures for modifying the configuration of the SARR on rebuttal, and the new evidence should therefore be struck.²¹ In any event, there would be no need to provide for any additional yard facilities to handle intra-and inter-train switching activities if the Board accepts BNSF's position that no rerouted non-coal traffic should be included in the SAC analysis, since those switching activities are only performed on the rerouted traffic.

V. THE BOARD SHOULD STRIKE OTTER TAIL'S NEW EVIDENCE ON A PRODUCTIVITY ADJUSTMENT TO THE DCF INFLATION INDEX

Otter Tail's evidence on operating cost inflation is contained in a single sentence in its Opening Evidence, where Otter Tail states without any explanation or supporting evidence that the annual inflation forecast it uses in the DCF analysis "is based on a forecast of the RCAF-A from DRI-WEFA's RCAF forecast dated June 2001." Otter Tail Opening Nar. at III-G-3. On rebuttal, Otter Tail seeks to make up for its failure to provide evidentiary support for this assertion by submitting new evidence in two areas.

First, it submits 75 pages of workpapers containing articles that discuss various ongoing efforts by the railroad industry to improve productivity. This material is offered in support of

²⁰ *Duke/CSX I* at 4.

²¹ By submitting new evidence on yard location, Otter Tail avoids scrutiny of many cost and feasibility issues associated with this new SARR configuration. For example, since the Fargo Yard is located in downtown Fargo, it would be necessary to determine, among other things, whether buildings would need to be demolished and whether overpasses and bridges would need to be expanded. To illustrate the extent of the problems generated by Otter Tail's new yard proposal, BNSF's construction cost expert Mr. Phillips prepared an aerial photograph of downtown Fargo showing where the new Fargo Yard would be constructed. *See* Exhibit 4.

Otter Tail's use of the RCAF-A since, according to Otter Tail, it indicates that the SARR can be expected to achieve productivity improvements over the DCF period. *See* Otter Tail Reb. Nar. at III-G-4 to III-G-11. Second, Otter Tail presents for the first time on rebuttal an alternative inflation forecast that is based on what Otter Tail characterized as a productivity factor embedded in the Energy Information Administration's ("EIA") forecast of coal transportation rates. *Id.* at III-G-11 to III-G-13. While Otter Tail purports to present this alternative index as support for the "appropriateness of using the RCAF-A in this case," *id.* at III-G-11, Otter Tail actually uses the supposed EIA productivity factor to develop a modified inflation index, which it then uses in its DCF analysis. *Id.* at III-G-13. Otter Tail's entire discussion of inflation on rebuttal should be struck on grounds that it is new evidence that exceeds the scope of permissible rebuttal.

When Otter Tail filed its Opening Evidence on June 13, 2003, it knew that the Board had rejected the use of the RCAF-A as the basis for forecasting a SARR's future inflation. In the *TMPA* decision, served on March 24, 2003, the Board stated clearly that it would not apply a productivity-adjusted inflation index in a SAC case in the absence of evidence identifying the specific productivity improvements a SARR could expect to achieve: "In the absence of any evidence showing specific productivity improvements for unit coal train operations that would affect the [SARR], we use the RCAF-U to index the operating expenses for the [SARR]." ²² Otter Tail was clearly on notice when it filed its Opening Evidence that its use of the RCAF-A had to be supported by specific evidence of productivity improvements that a SARR could expect. If any evidence supported its use of a productivity-adjusted inflation index, Otter Tail

²² *TMPA* at 161.

should have submitted that evidence on opening. It cannot evade scrutiny of that evidence by submitting it for the first time on rebuttal.²³

Otter Tail's discussion of an alternative inflation index based on the EIA's supposed productivity factor must also be struck. As discussed previously, the Board's rules governing rebuttal evidence permit a complainant to modify its opening assumptions only in response to a defendant's reply evidence and only if the complainant first demonstrates that the defendant's reply evidence is infeasible, unrealistic or unsupported.²⁴ Otter Tail's proposed EIA adjustment is not a reply to BNSF's evidence, which simply used the Board's approach in prior cases to apply the RCAF-U, but rather was an afterthought. Moreover, Otter Tail could not possibly demonstrate that BNSF's Reply Evidence based on the use of the RCAF-U is infeasible or unsupported, since the Board has adopted the RCAF-U in several recent SAC cases. There is no basis in the Board's rules governing rebuttal to accept a new approach to the calculation of inflation in the DCF model that is tendered for the first time on rebuttal.

Moreover, it would be inappropriate to accept new evidence based on the EIA's productivity assumptions without giving the defendant an opportunity to address the relevance of

²³ There are obvious defects with Otter Tail's new evidence that BNSF would have addressed in detail if the evidence had been presented in a timely manner. For example, the articles included in Otter Tail's workpapers discuss prospective locomotive technologies, but Otter Tail makes no effort to show the extent to which those new technologies might reduce SAC costs, the additional costs that the SARR would have to incur to acquire the new technologies, when the new technologies might become available, or whether the SARR, which uses long-term locomotive leases, would even be able to take advantage of the new technologies. The mere possibility of technological advances in rail and locomotive equipment, which is undeniable, cannot support the use of the RCAF-A index in a SAC test. Even if new technologies can be expected to come into existence at some point in the future, that would not justify use of the RCAF-A index, and it would not justify the use of any productivity adjustment starting in the early years of the DCF analysis.

²⁴ *Duke/NS* at 14-15.

EIA's productivity assumptions in the context of a SAC case. BNSF has addressed those issues in another pending SAC case and has shown that there are several aspects of the EIA productivity factor that make it inappropriate for use in a SAC case.²⁵ The Board could not accept new evidence on a complex issue like the forecasting of inflation without a full development of the record. But a record cannot be developed when the complainant holds back evidence and alternative methodologies until rebuttal, when the defendant does not have an opportunity to respond.

VI. THE BOARD SHOULD REJECT OTTER TAIL'S PROPOSED CHANGES TO THE SARR'S COST OF CAPITAL

Otter Tail also proposed on rebuttal a modification to its original cost of capital assumptions. On both opening and rebuttal, Otter Tail calculated the SARR's cost of capital using separate factors for the cost of equity and the cost of debt. As to the equity portion of its cost of capital calculations, Otter Tail's Rebuttal Evidence simply updated its Opening Evidence using the same approach it presented on opening, which BNSF had accepted. BNSF does not oppose this update. As to the cost of debt, however, Otter Tail modified the approach it used on opening, which used the industry average cost of debt for the period 1999-2001. On rebuttal, Otter Tail used the 1999-2001 industry average only through 2002 and then "assumed that the

²⁵ As BNSF explained in its AEP Texas Reply Evidence, use of the EIA productivity factor in a SAC analysis suffers from the same basic flaw that makes the RCAF-A inappropriate for use in a SAC case, namely that it assesses productivity improvements that have been achieved by existing railroads. In addition, the EIA productivity analysis is not a true measure of rail productivity, but rather is a comparison of rail rate escalation to a measure of input price inflation, thus reflecting numerous developments in rail markets that are completely unrelated to productivity that a SARR could expect. In addition, even if the SARR could be expected to achieve some productivity over the DCF period, those productivity improvements could not be expected until much later in the DCF period, particularly given the extremely low operating costs of the SARR at the beginning of the DCF period compared to the operating costs of real world railroads on which the EIA analysis is based. See BNSF's May 24, 2004 Reply Evidence in *AEP Texas North Company v. BNSF*, Docket No. 41191 (Sub-No.1), at III.G-3 to III.G-13.

OTRR will refinance the debt portion of its capital structure in 1Q03” at an interest rate of 4.99%. Otter Tail Reb. Nar. at III-G-2.

Otter Tail offers no justification for changing its methodology on rebuttal, nor could it. Otter Tail’s new approach is clearly not a response to BNSF’s Reply Evidence, since BNSF accepted Otter Tail’s opening methodology. Moreover, Otter Tail provides no evidence at all in support of its assumed refinancing. It provides no evidence that the SARR would be able to refinance a debt burden of the size of the SARR’s debt; no evidence about the interest rates that might apply to such a refinancing;²⁶ and no evidence about the financing fees that would be required if a refinancing were feasible. The Board would have to reject Otter Tail’s unsupported refinancing assumption even if it had been sponsored on opening.

VII. THE BOARD SHOULD STRIKE OTTER TAIL’S STATISTICAL ANALYSIS OF TRANSIT TIMES ON REROUTED TRAFFIC UNDER *DAUBERT* PRINCIPLES

Otter Tail submitted a new statistical analysis at pages III-A-49 through III-A-52 of its Rebuttal Narrative which purports to demonstrate that the OTRR’s admittedly longer average transit times for rerouted non-coal traffic only *appear* to be longer than the transit times for the same traffic on the real-world BNSF. Indeed, Otter Tail asserts that its analysis demonstrates that from a statistical perspective – at “a 99 percent confidence level” – “there is no difference between the OTRR’s transit times and the BNSF’s transit times.” Otter Tail Reb. Nar. at III-A-52. In fact, an appropriate statistical analysis employing standard statistical techniques confirms

²⁶ Otter Tail relies on a March 29, 2004 filing by the Association of American Railroads in Ex Parte 558 (Sub-No. 7), for its updating of the cost of equity. But that same filing shows that interest rates increase with the length of the repayment period. *See* Appendix B to the Verified Statement of AAR Director - Accounting and Finance David F. Miller, included with the AAR filing. It is obvious that any refinancing of the SARR’s debt would present complications that Otter Tail has not even begun to address.

that the OTRR's transit times are longer, on average, than those for the real-world BNSF and that the difference in transit times is statistically significant at the 99 percent level.

As explained in the Verified Statement of Dr. Matthew G. Mercurio, attached hereto as Exhibit 5, Otter Tail did not employ reliable statistical methods for its analysis, nor would any competent statistician accept the results it proffers. Otter Tail ignores standard statistical methods for determining whether the means of two distributions are the same and instead invents a test that produces absurd results and fails to distinguish in a statistically meaningful way between the two populations being compared. Otter Tail's analysis exhibits confusion about the meaning of statistical terms and the use of statistical tools to test hypotheses. In the end, Otter Tail reaches conclusions that cannot be supported when standard statistical tools are employed. Otter Tail's statistical "evidence" does not meet the admissibility standards for expert evidence and should be struck by the Board.

Well-established law governing the testimony of expert witnesses supports the principle that "junk science" like Otter Tail's statistical analysis should not be admissible in STB proceedings. The Board has not itself adopted threshold standards for the admissibility of expert testimony, but it should be guided in evaluating the probative value of proffered expert evidence by the same standards used to determine the admissibility of such evidence in federal court proceedings.²⁷ Under Rule 702 of the Federal Rules of Evidence, there are threshold standards governing the admissibility of expert testimony. Such testimony may be presented only if

- (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3)

²⁷ The Board's rule governing the admissibility of evidence specifically references admissibility "under the rules of evidence governing proceedings in matters not involving trial by jury in the courts of the United States." 49 C.F.R. § 1114.1. The Federal Rules of Evidence apply to all "proceedings in the courts of the United States," *i.e.*, to both jury and non-jury trials.

the witness has applied the principles and methods reliably to the facts of the case.

Expert testimony like Otter Tail's that does not meet these standards of reliability should not be considered by the trier of fact (here the Board).

Rule 702 reflects a pronounced trend in federal court litigation to preclude expert testimony that is not reliable. *See Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993)(establishing guidelines for determining admissibility of expert scientific evidence); *General Electric Co. v. Joiner*, 522 U.S. 136 (1997) (establishing standard of review for admissibility determinations); *Kumho Tire Co. v. Carmichael*, 526 U.S. 137 (1999) (applying *Daubert* principles to expert evidence based on technical or other specialized knowledge). Administrative agencies, like the courts, must avoid using unreliable evidence proffered by experts as the basis of their determinations. Accordingly, the Board should strike Otter Tail's statistical analysis on the grounds that it is unreliable and does not meet the standards of admissibility for expert evidence.

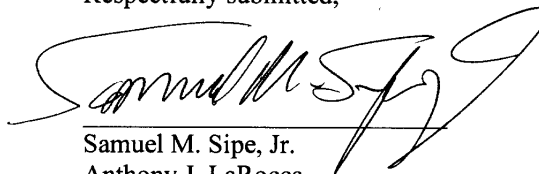
VIII. CONCLUSION

For the reasons set forth above, the Board should strike evidence contained in Otter Tail's April 29, 2004 Rebuttal Evidence in the following areas: (a) new evidence relating to Otter Tail's purchases of coal from certain southern PRB mines; (b) evidence relating to SARR capacity and operating statistics based on Otter Tail's new rebuttal string program; (c) new evidence relating to the SARR's yard configuration; (d) new evidence on the inflation index to be used in the DCF analysis; (e) Otter Tail's modified calculation of the SARR's cost of capital; and (f) Otter Tail's statistical analysis of rerouted traffic.

OF COUNSEL

Richard E. Weicher
Michael E. Roper
THE BURLINGTON NORTHERN
AND SANTA FE RAILWAY COMPANY
2500 Lou Menk Drive
P. O. Box 961039
Ft. Worth, TX 76131-0039
Phone (817) 352-2368
Facsimile (817) 352-2397

Respectfully submitted,

A handwritten signature in black ink, appearing to read "Samuel M. Sipe, Jr.", written over a horizontal line.

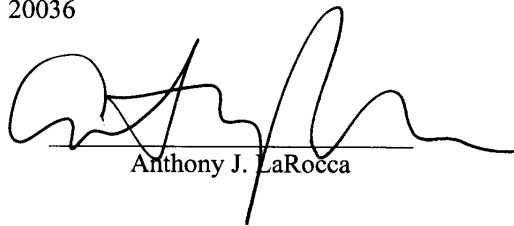
Samuel M. Sipe, Jr.
Anthony J. LaRocca
Linda S. Stein
STEPTOE & JOHNSON, LLP.
1330 Connecticut Ave., N.W.
Washington, D.C. 20036
Phone (202) 429-6486
Facsimile (202) 429-3902

June 8, 2004

CERTIFICATE OF SERVICE

I hereby certify that a true and correct copy of the foregoing was served by messenger on
this 8th day of June, 2004 upon the following counsel for Otter Tail Power Company:

Nicholas J. DiMichael
Jeffrey O. Moreno
Michael H. Higgins
David E. Benz
Thompson Hine LLP
1920 N Street, N.W.
Washington, D.C. 20036



Anthony J. LaRocca

Exhibit 1 Has Been Redacted

Exhibit 2 Has Been Redacted

**BEFORE THE
SURFACE TRANSPORTATION BOARD**

OTTER TAIL POWER COMPANY,

Complainant,

v.

**THE BURLINGTON NORTHERN AND
SANTA FE RAILWAY COMPANY,**

Defendant.

Docket No. 42071

**VERIFIED STATEMENT OF
KARL M. LEOPOLD AND TERENCE J. MONKS**

We are Karl M. Leopold and Terence J. Monks. We have been retained by The Burlington Northern and Santa Fe Railway Company (BNSF) in this proceeding to assist in the evaluation of Otter Tail's evidence. We have previously submitted testimony in support of BNSF's evidence in this proceeding and our qualifications are presented therein.¹ In particular, we have submitted testimony that evaluates Otter Tail's string programs and the conclusions that Otter Tail reached based on its string program analyses. In our prior submissions, we pointed out several fundamental flaws in the string program analyses that Otter Tail sponsored in its Opening and Supplemental Opening Evidence.

After Otter Tail filed its Rebuttal Evidence on April 29, 2004, BNSF asked us to compare the string program submitted in support of Otter Tail's Supplemental Opening Evidence with the latest program submitted with Otter Tail's Rebuttal Evidence. We identified over 100 changes to

¹ See BNSF Rep. Nar. at IV-24 and IV-32 to 33; BNSF Supp. Rep. Nar. at IV-17 and IV-23 to 24.

the code that runs the string program.² We determined that all but one of the many input files, which contain data used by the string program, had been modified.³ As a result of the changes to the input files and program code, all of the string program output files contain new data.

We understand that the Board's rules do not provide BNSF with an opportunity to submit detailed evidence in response to a complainant's rebuttal evidence. Nevertheless, BNSF asked us to carry out a preliminary review of Otter Tail's new string program to determine if Otter Tail had fixed some of the more fundamental problems that we had previously identified. Therefore, we performed a limited evaluation of the new string program output to determine whether Otter Tail successfully resolved train conflicts when it manually inspected the program's output files. The need for extensive manual review and manipulation of the string program output was one of the principal flaws that we have described in our prior testimony. *See* BNSF Supp. Reply Nar. at III.B-3 to 4. Without carrying out an extensive analysis, we identified a head-on collision occurring on a single-track section of the OTRR between W. Echeta and E. Kendrick. This collision was identified by reviewing the string diagram for this section of the OTRR and confirming that the segment between those two locations is single track with no passing sidings.⁴ The existence of this collision was enough to confirm that Otter Tail had not addressed one of the critical flaws in the string program -- its inability to detect and resolve collisions.

We also reviewed the manual changes that Otter Tail made to the movement of trains to determine whether those changes caused any problems elsewhere on the system. Without carrying out a full review, we determined that Otter Tail's adjustments allowed an empty train to

² *See* BNSF electronic workpaper "source CSDiff.htm."

³ *See* BNSF electronic workpaper "Folder diffs.htm."

⁴ *See* BNSF electronic workpaper "STRING_47.xls" at worksheets "Model Induced Collision" and "Chart A."

pass another empty train on a single-track segment between E. Lodge Grass and MP 740. On that same segment, Otter Tail's adjustments also caused a head-on collision between one of the empty trains and an oncoming loaded train.⁵

BNSF also asked us to evaluate Otter Tail's claim that the new string program accurately aggregates the OTRR's peak day service units. In BNSF's Supplemental Reply Evidence, we showed that the string program failed to aggregate operating statistics properly, and that the errors tended to substantially understate statistics such as loading time. BNSF Supp. Rep. Nar. at III.C-7 to 9. To evaluate Otter Tail's claim that it had fixed this problem, we first attempted to replicate Otter Tail's string program results. We ran the string program using the instructions and run-time parameters provided in Otter Tail's "ReadMe.xls" file and compared the output files from that run with the output files provided in Otter Tail's Rebuttal Evidence. The output files were substantially different as shown in the "TIMESL_diffs" and "TIMESE_diffs" files in BNSF's electronic workpaper "Benchmark.zip." The fact that we cannot replicate Otter Tail's string program results suggests that Otter Tail either failed to provide correct instructions or failed to provide the input files and/or string program code that it actually used to generate its rebuttal evidence. In either case, Otter Tail's results and therefore its claim that the string program generates accurate operating statistics cannot be verified.

Finally, we showed in BNSF's Supplemental Reply Evidence that the string program's operating statistics ignored loading time for the majority of the OTRR trains. BNSF Supp. Reply Nar. at III.C-8 to 9. Otter Tail claims to have fixed this problem by accumulating six hours of loading time for each empty train returning to the mines on the OTRR's peak day. Otter Tail Reb. Nar. at III-C-29. To check this assertion, we reviewed the string program files to determine

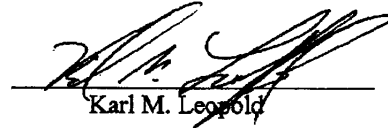
⁵ See BNSF electronic workpaper "STRING_47_BASE.xls" at worksheet "Manually Induced Collision."

the number of loaded trains departing the mines on the OTRR's peak day and the number of empty trains arriving at the mines. We determined that the flow of trains is imbalanced with 51 loaded trains departing the mines and only 33 empty trains arriving at the mines.⁶ By adding mine loading time to the empty trains on its peak day, Otter Tail understates the actual amount of loading time its trains would incur on that day.

⁶ See Otter Tail Reb. electronic workpaper "Service Hours Reconciliation (Base Case).xls" at worksheet "PeakDay Dispatch" (sorted by the field "type").

I declare under penalty of perjury that I have read the foregoing Verified Statement and that the contents thereof are true and correct. Further, I certify that I am qualified and authorized to sponsor this testimony.

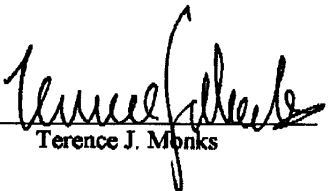
Executed on June 8, 2004

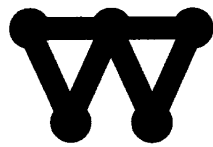


Karl M. Leopold

I declare under penalty of perjury that I have read the foregoing Verified Statement and that the contents thereof are true and correct. Further, I certify that I am qualified and authorized to sponsor this testimony.

Executed on June 8, 2004

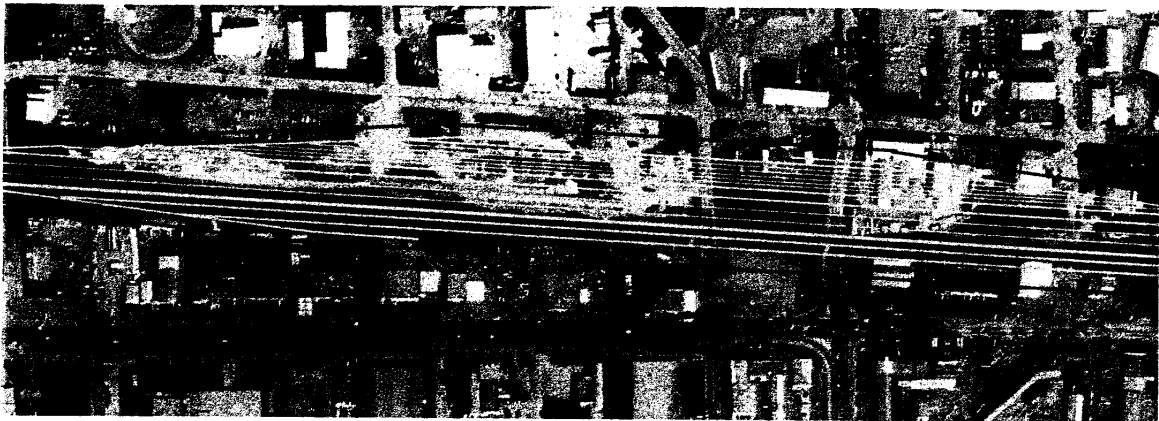

Terence J. Monks



**Ralph Whitehead
Associates, Inc.**

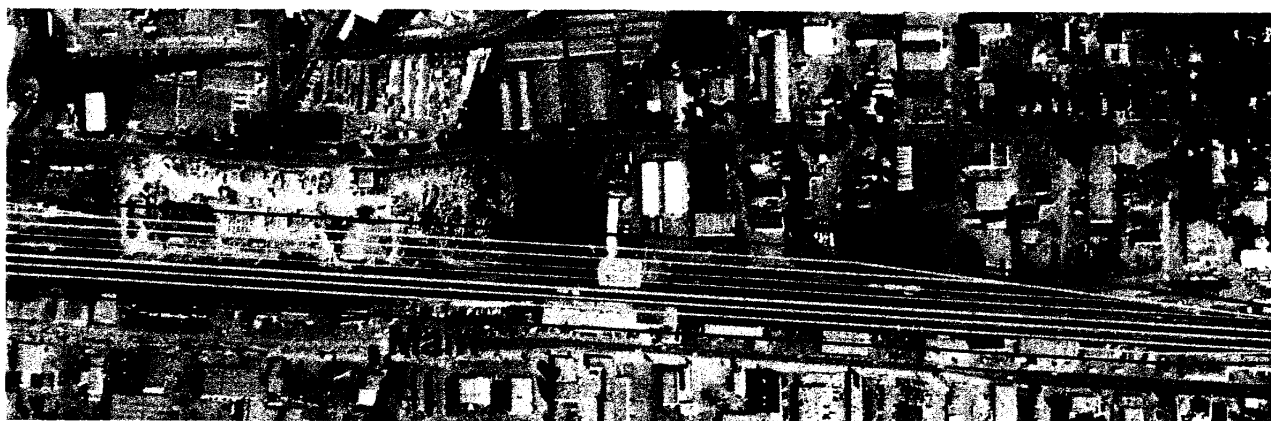


FARGO YARD



Railroad Tracks
Right-of-Way Limits





***Otter Tail Power vs. BNSF
STB Docket No. 42071***



**BEFORE THE
SURFACE TRANSPORTATION BOARD**

OTTER TAIL POWER COMPANY,	)	
	)	
Complainant,	)	
	)	
v.	)	
	)	
THE BURLINGTON NORTHERN AND	)	Docket No. 42071
SANTA FE RAILWAY COMPANY,	)	
	)	
Defendant.	)	
	)	

Verified Statement of Matthew G. Mercurio

I am Matthew G. Mercurio. I am a Manager of FTI Consulting, Inc., with offices at 1201 Eye Street, N.W., Suite 400, Washington, DC 20005. I was awarded a Ph.D. in economics by Princeton University in 1996. In addition, I hold a Masters Degree in economics, also from Princeton University, and a Bachelor of Arts degree in mathematics and economics from Boston University.

From 1996 through 2002, I held the positions of Senior Economist and then Vice President at Economists Incorporated, a consulting firm in Washington, DC, that specializes in antitrust and regulatory economics. At Economists Incorporated I specialized in empirical econometric and statistical analyses of microeconomic problems. I developed econometric models for competitive analysis, damage estimates, cost modeling, and survey analysis. In 2003, I joined FTI Consulting as a Manager in the Network Industries Strategies group where I have continued with substantially the same types of work.

As my background information demonstrates, I am familiar with the use of statistical methods and mathematical tools in the analysis of data. I have been retained by The Burlington

Northern and Santa Fe Railway Company (BNSF) in this proceeding to assist in the evaluation of statistical analyses offered by Otter Tail Power Company (Otter Tail).

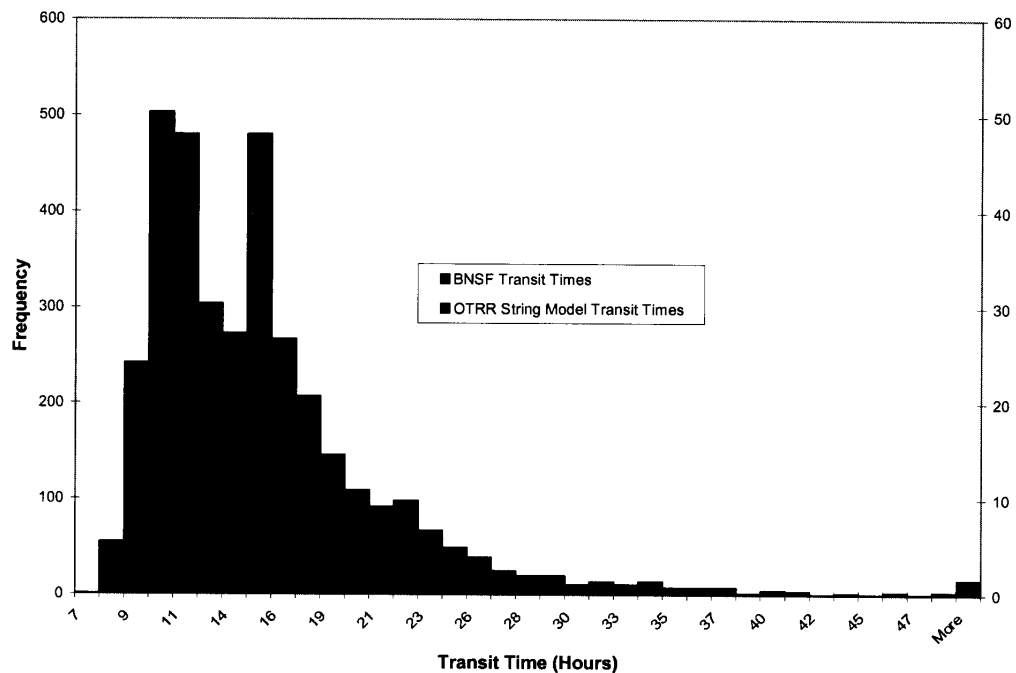
In particular, I have been asked to review the statistical analysis contained in Otter Tail's Rebuttal Narrative at III-A-49 through III-A-52. Otter Tail offers this analysis to demonstrate that the transit times for rerouted non-coal traffic produced by Otter Tail's string program are not "statistically different" from real-world BNSF transit times for the same trains.¹ Otter Tail asserts that "there is a strong statistical inference (a 99 percent confidence level) that there is no difference between the OTRR's transit times and the BNSF's transit times." Otter Tail Rebuttal Nar. at III-A-52.

I conclude that Otter Tail did not employ valid statistical methods in comparing OTRR transit times to real-world BNSF times. Otter Tail's analysis reveals a fundamental misunderstanding of statistical methods of hypothesis testing. Otter Tail ignored standard, accepted methods for evaluating the similarity between two sets of data. Instead, Otter Tail based its conclusions on a "test" that is incapable of detecting differences between data sets in a meaningful or useful manner. As a result, Otter Tail's statistical analysis is unsound and unreliable. No competent statistician would employ Otter Tail's methodology or accept Otter Tail's conclusion that the OTRR transit times are statistically indistinguishable from the real-world BNSF transit times. Indeed, standard statistical tests demonstrate that the OTRR transit times are significantly longer, on average, than real-world BNSF times.

Otter Tail's analysis supposedly demonstrates that the two distributions shown in Figure 1 below are statistically indistinguishable:

¹ The string program models the operation of these trains over a route that is different from the real-world route of movement.

Figure 1
Distribution of Actual BNSF and OTRR String Program Transit Times²



First, I will explain the errors in Otter Tail’s arguments that these two distributions are no different from one another, statistically. I then show that the proper statistical test confirms that the means of these two distributions are different at 99.9 percent confidence.

Otter Tail attempted to demonstrate that no statistical significance should be attached to the fact that the BNSF transit times have a lower mean – indicating shorter transit times – than the OTRR transit times. As a preliminary matter, it is worth noting that comparing means rather

² Source: BNSF electronic workbook “Confidence_Intervals_Corrected.xls.”

than medians³ is more likely to produce a favorable result for Otter Tail because the average BNSF transit time is distorted (increased) by a relatively small number of outliers, *i.e.*, movements with extremely long transit times, in the BNSF distribution. Outliers are to be expected in the BNSF data because it is a real-world distribution which covers a full year of movements. BNSF trains experience track outages, breakdowns, scheduled maintenance, and other events that are reflected in the real world times. The OTRR distribution, on the other hand, does not have such outliers because it covers a single week and is derived from a model that intentionally excludes real-world variation both in terms of events such as maintenance and other track outages and in terms of the time spent on routine tasks such as fueling, switching, and changing crews in the yards. The limitations programmed into the Otter Tail string model thus produce an artificially “tight” distribution. This point alone seriously undermines Otter Tail’s arguments regarding the variability of transit times.

Despite the obvious difference between the two distributions – visible in Figure 1 to even a casual observer – Otter Tail seeks to convince the Board that average BNSF transit times are not, in fact, shorter than average Otter Tail transit times. There are reliable, well-known statistical methods for comparing the means of two distributions. Otter Tail chose not to adopt such an approach, however. Instead, Otter Tail proposes a test of similarity that does not work and that appears to exhibit substantial confusion on Otter Tail’s part concerning the meaning of basic statistical concepts such as “confidence interval” and “confidence level.”

Otter Tail’s test is as follows: Otter Tail recognizes that, for population distributions such as the transit time distributions here, over 99 percent of the values will fall within plus or

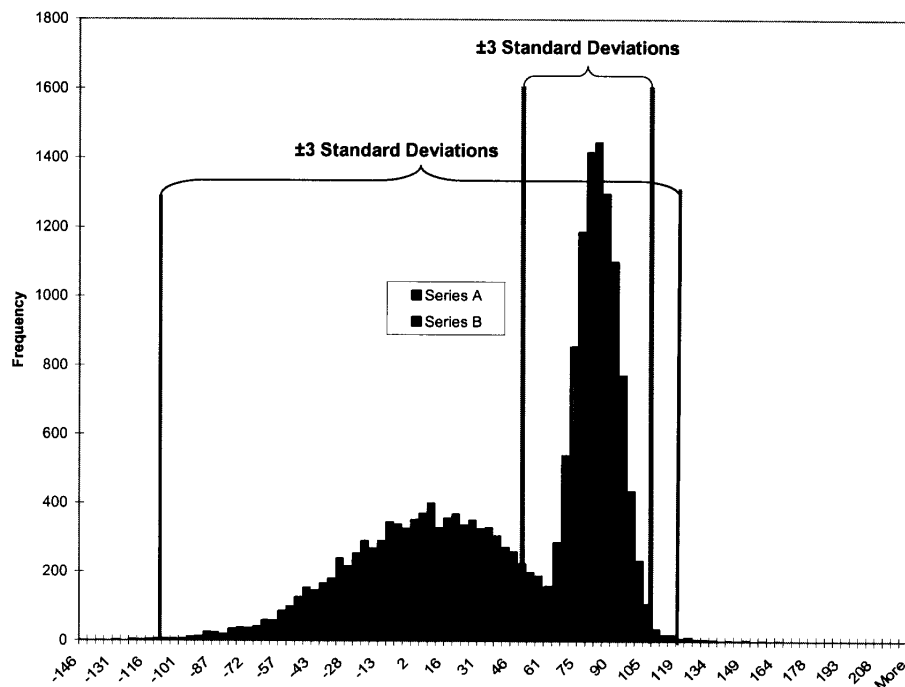
³ The median defines the point at which half of the points are lower and half the points are higher and, unlike the mean, is not overly sensitive to outliers. BNSF trains have a median transit time of 12.05 hours while the median transit time for OTRR trains is 14.27 hours.

minus three standard deviations of the mean.⁴ The outcome of Otter Tail's test is determined by whether the ± 3 standard deviation range of one distribution completely encompasses the ± 3 standard deviation range for the second distribution. If the distributions overlap in this manner, Otter Tail asserts that it is appropriate to conclude at a 99 percent confidence level that there is "no difference" between the transit times.

Otter Tail's test is facially invalid because it leads to absurd results and fails to distinguish in a statistically meaningful way between the populations being compared. The invalidity of Otter Tail's test can be illustrated with a simple example. Under Otter Tail's test, one would be forced to conclude that the two hypothetical distributions shown in Figure 2 below are identical, which is obviously not the case.

⁴ The standard deviation is a measure of the how tightly a distribution is clustered about its mean.

Figure 2
Hypothetical Distributions with Overlap of ± 3 Standard Deviations



In Figure 2, the ± 3 standard deviation range for each distribution has been marked in the appropriate color. Because the upper bound of distribution B is to the left of the upper bound for distribution A, Otter Tail would say that distribution B “lie[s] within the range” of distribution A. *See Otter Tail Rebuttal Nar. at III-A-51.* This would lead Otter Tail to conclude, despite the obvious difference, that there is a 99 percent probability that there is no difference between the distributions.

Otter Tail’s test also produces internally inconsistent results that render it invalid. Under Otter Tail’s overlap test, one would be forced to conclude, based on Figure 2, that while the average of distribution B is identical to the average of distribution A because of the overlap, the average of distribution A is not identical to the average of distribution B. This is because

distribution A is not encompassed within distribution B. In other words, Otter Tail's test indicates $B=A$ but $A \neq B$. A valid statistical test could not involve such internally inconsistent conclusions.

Although Otter Tail claims that it performed its analysis at a "99 percent confidence level," it did not actually do so. Rather, it is merely attempting to cloak its analysis in the language of statistics. What Otter Tail has actually done is construct 99 percent bounds for two distributions, the BNSF transit times and the OTRR transit times.⁵ Otter Tail knows that, for Normal distributions, 99 percent of the values will fall within the "confidence interval" identified. Because Otter Tail has identified the range within which 99 percent of the values for each distribution lie, it apparently believes that any comparison it makes between the distributions will involve a 99 percent level of certainty. Otter Tail is incorrectly equating the standard error bounds it has identified for two separate populations to a confidence level for a statistical test, with the attendant implication that any comparison it makes *between* the two populations will have only a 1 percent chance of error.

In simplified terms, Otter Tail's apparent reasoning can be expressed as three logical statements:

- (1) $X < 99 \text{ percent of Distribution A} < Y$
- (2) $X < 99 \text{ percent of Distribution B} < Y$

Therefore:

- (3) There is a 99 percent chance that $A=B$.

⁵ Otter Tail's use of the label "confidence interval" for these bounds is totally inappropriate since true confidence intervals are defined with regard to estimated parameters based on underlying distributions, not with regard to the distributions themselves.

The leap of faith that must be made to reach the conclusion in statement (3) is obvious. As was shown above in Figure 2, it is entirely possible for two very different distributions to fall between specified endpoints.

There is simply no basis, mathematical or otherwise, for concluding that the 99 percent error bounds that Otter Tail has identified permit any conclusion to be drawn about the two distributions with a 1% chance of error (a 99% confidence *level*). Otter Tail's overlap test reveals a serious misconception concerning the meaning of statistical confidence levels and how such confidence levels relate to hypothesis testing. In statistics, confidence level has a precise definition: it is $1-\alpha$, where α is the probability that a hypothesis being tested will be rejected even if it is true.⁶ Hence, a statistician might say that a hypothesis that the means of two distributions are identical has been rejected at a 95 percent confidence level. In such a case, there is a 5% chance that the hypothesis has been rejected in error. Similarly, at a 99 percent confidence level, there is a 1% chance that the hypothesis has been rejected in error. Otter Tail is not conducting a test at a 99 percent confidence level as that term is used by statisticians and it has established no basis upon which to draw any conclusion about the probability that its proposed test produces correct results.

Otter Tail may have been attempting to calculate confidence intervals for the means of the OTRR and BNSF distributions rather than identifying the range for each distribution within which 99 percent of the values fall. If so, Otter Tail's calculations represent a crude conceptual error. The confidence interval for the mean of a distribution, which is essentially a statistical measure of how precisely the "true" mean of a sampled population has been estimated, is not the same as the spread around the mean for the underlying distribution that Otter Tail identified for

⁶ Greene, William H., *Econometric Analysis*, at 125-26 (2nd ed. 1993)(included as Attachment A to this verified statement).

the entire population. This is because the standard error of the mean is related to but not the same as the standard deviation of the distribution as a whole. Otter Tail employs the following formula, where μ is the mean and σ is the standard deviation of the population:

$$(1) \mu_{[\text{Transit Times}]} \pm 3 \cdot \sigma_{[\text{Transit Times}]}$$

In other words, OTP's calculations imply that the confidence interval around the mean transit time estimate is defined by three standard deviations of the transit times. This formula does not take into account the number of sample points used to generate the mean observation. The correct formula is:

$$(2) \mu_{[\text{Transit Times}]} \pm 3 \cdot \sigma_{[\text{Mean Transit Times}]}$$

The variance of the mean transit times is equal to the variance of transit times divided by N, the total number of observations. Thus the standard deviation of the mean transit times is equal to the standard deviation of transit times divided by $\sqrt{N}$:

$$(3) \sigma_{[\text{Mean Transit Times}]} = \sigma_{[\text{Transit Times}]} / \sqrt{N}$$

Because BNSF's mean transit times are based on a large number of observations, these mean estimates are extremely precise. This precision is reflected in the small confidence intervals (when correctly computed), as shown in Table 1 below.⁷ Table 1 is based on Otter Tail's Table III-A-3, Otter Tail Rebuttal Nar. at III-A-51.

⁷ Source: BNSF electronic workbook "Confidence_Intervals_Corrected.xls."

Table 1
Comparison of BNSF and OTRR Effective Transit Times Between Fargo, ND and Snowden, MT

	BNSF Full-Year Fargo-Snowden Transit Times Via Minot 2/				OTRR 10/11 to 10/18 Fargo-Snowden Transit Times Via Glendive			
	<u>Train Type 3/</u> (1)	<u>Average</u> (5)	<u>Correct Three Std. Dev. 4/</u> (6)	<u>OTP 99 Percent Confidence Interval 5/</u> (7)	<u>Correct 99 Percent Confidence Interval 6/</u> (7)	<u>Average</u> (8)	<u>Correct Three Std. Dev. 4/</u> (9)	<u>OTP 99 Percent Confidence Interval 6/</u> (10)
1. Bare Table Intermodal ("B")		10.74	1.22	25.34	11.96	7/	7/	7/
2. Loaded Unit Grain ("G")		15.54	0.81	33.55	16.36	16.71	1.02	21.04
3. High Priority Manifest ("H")		11.07	0.22	17.18	11.29	13.95	1.00	17.55
4. Normal Priority Manifest ("M")		19.71	0.94	43.04	20.65	13.01	1.51	18.22
5. Premium Service Intermodal ("P")		18.43	1.24	40.32	19.67	14.80	3.77	21.33
6. Stack Intermodal ("S")		12.92	0.31	27.78	13.23	14.23	0.69	19.23
7. Other Unit Train ("U")		14.33	6.12	31.64	20.45	7/	7/	7/
8. Vehicle Train ("V")		12.07	0.44	22.83	12.50	13.41	1.54	17.62
9. Empty Unit Grain Train ("X")		11.28	0.81	23.13	12.08	14.37	0.84	18.99
10. Average		13.77	0.24	31.38	14.01	14.37	0.50	19.95

1/ Total transit time from BNSF origin to BNSF destination was developed from BNSF 2002 revenue files.

2/ Total BNSF transit time between Fargo and Snowden via Minot was developed from BNSF 2002 train movement records.

3/ Includes trains moving between Fargo and Snowden and Snowden and Fargo.

4/ The standard deviation indicates how the numbers in a data set vary from the average. Approximately 99.73 percent of the values in a distribution fall within three standard deviations of the mean.

5/ Average transit time plus three (3) standard deviations of the transit time.

6/ Average transit time plus three (3) standard deviations of the mean transit time.

7/ This train type did not move during Otter Tail's String Model study period.

Table 1 demonstrates that OTP's confidence interval calculations are grossly understated compared to the true values. Table 1 clearly shows that when the standard deviations of the means are calculated properly, the overlap of "confidence intervals" on which Otter Tail based its analysis disappears. For example, the upper limit of the ± 3 standard deviation range for H trains on the BNSF is 11.29 hours while the upper limit for H trains on the OTRR is 14.95 hours. The OTRR range is no longer encompassed by the range for the BNSF distribution.

The above discussion demonstrates that Otter Tail's statistical analysis is invalid and would not be accepted by a competent statistician. A qualified statistician would have employed the two-sample t-Test of mean equality to determine whether the OTRR's transit times are, on average, longer than those of the real-world BNSF.⁸ That test permits a statistician to mathematically test the hypothesis that the means of two distributions are the same. A "t-statistic" is calculated for the two populations, and the probability that the calculated value is equal to or greater than the standard value (*i.e.*, assuming the means are the same) is determined. If the probability that the t-statistic would be greater is large, there is a high confidence level that the means of the distributions are not the same.⁹ Table 2 below presents the results of these t-Tests by train type and for the total average transit times.¹⁰ As the table demonstrates, had Otter Tail performed the correct statistical test, it would have concluded that the two data sets are significantly different. The table shows that Otter Tail's transit times are longer than those for

⁸ It has been conclusively demonstrated that this test is superior to comparing confidence intervals. Schenker, Nathaniel and Gentleman, Jane F., "On Judging the Differences by Examining the Overlap Between Confidence Intervals," *The American Statistician*, Vol. 55 No 3, pp. 182-186 (2001) (included as Attachment B to this verified statement).

⁹ Maddala, G.S., *Introduction to Econometrics*, at 28-30 (2nd ed. 1992)(included as Attachment C to this verified statement).

¹⁰ Source: BNSF electronic workpaper "Confidence_Intervals_Corrected.xls."

the real-world BNSF at a 99.87 percent confidence level. Otter Tail concluded, of course, that average transit times were *the same* at a purported 99 percent confidence level.

The same results hold true generally if the t-Test is performed individually by train type. For train types H, S, V, and X, OTRR transit times are longer at better than a 99 percent confidence level. For train type G, OTRR transit times are longer at a 98.93 percent confidence level. The only apparent exceptions to the rule that OTRR transit times are longer are M and P trains. For M trains, BNSF pointed out in its Supplemental Reply Evidence that there is no basis for comparison between the OTRR transit times and BNSF real-world times since the real-world times include a substantial amount of switching time while the OTRR times do not. *See* BNSF Supplemental Reply Nar. at III.A-24. As a result, the result of the t-Test for this train type is not meaningful and must be disregarded. For P trains, the t-Test does not permit a clear conclusion at the standard 95 percent confidence level that the OTRR would move P trains faster than the real-world BNSF. Given that P trains experience switching which Otter Tail may not have accounted for correctly, *see* BNSF Supplemental Reply Nar. at III.A-24, and that the sample size for P trains is extremely small (3 trains in Otter Tail's study period), it would not be appropriate to draw strong conclusions regarding the apparent difference between P train transit times.

Table 2
Comparison of BNSF and OTRR Transit Times Between Fargo, ND and Snowden, MT

<u>Train Type 2/</u> (1)	<u>BNSF Full Year Fargo- Snowden Transit Times Via Minot 1/ Average</u> (2)	<u>OTRR 10/11 to 10/18 Fargo- Snowden Transit Times Via Glendive Average</u> (3)	<u>t-Test</u> (4)	<u>Implied Confidence Level</u> (4)
1. Bare Table Intermodal ("B")	10.74	<u>3/</u>	<u>3/</u>	<u>3/</u>
2. Loaded Unit Grain ("G")	15.54	16.71	0.0107	98.930%
3. High Priority Manifest ("H")	11.07	13.95	<0.0001	>99.999%
4. Normal Priority Manifest ("M")	19.71	13.01	<0.0001	>99.999%
5. Premium Service Intermodal ("P")	18.43	14.80	0.0888	91.121%
6. Stack Intermodal ("S")	12.92	14.23	<0.0001	>99.999%
7. Other Unit Train ("U")	14.33	<u>3/</u>	<u>3/</u>	<u>3/</u>
8. Vehicle Train ("V")	12.07	13.41	0.0007	99.928%
9. Empty Unit Grain Train ("X")	11.28	14.37	<0.0001	>99.999%
10. Average	13.77	14.37	<u>0.0013</u>	<u>99.87%</u>

1/ Total BNSF transit time between Fargo and Snowden via Minot was developed from BNSF 2002 train movement records.

2/ Includes trains moving between Fargo and Snowden and Snowden and Fargo.

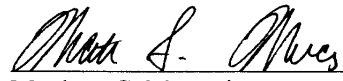
3/ This train type did not move during Otter Tail's String Model study period.

In sum, I conclude that Otter Tail's statistical analysis of transit times is flawed and is inconsistent with established principles and standards applied by qualified statisticians. Otter Tail's methods are incapable of testing whether the means of two distributions are different. An analysis of BNSF and OTRR transit times that employs accepted statistical procedures, however, demonstrates that OTRR transit times are significantly longer than those for the real-world BNSF.

Verification of Matthew G. Mercurio

I, Matthew G. Mercurio, verify under penalty of perjury that the foregoing is true and correct. Further, I certify that I am qualified and authorized to file this Verified Statement.

Executed on June 7, 2004


Matthew G. Mercurio



itself is a rule, stated in terms of the data, that dictates whether the null hypothesis should be rejected or not. For example, the hypothesis might state that a parameter is equal to a specified value. The decision rule might state that the hypothesis should be rejected if a sample estimate of that parameter is too far away from that value (where "far" remains to be defined). More generally, the classical, or Neyman–Pearson, methodology involves partitioning the sample space into two regions. If the observed data (i.e., the test statistic) fall in the **rejection region** (sometimes called the **critical region**), the null hypothesis is rejected; if they fall in the **acceptance region**, it is not.

4.8.1. Testing Procedures

Since the sample is random, the test statistic, however defined, is also random. The same test procedure can lead to different conclusions in different samples. As such, there are two ways such a procedure can be in error:

1. **Type I error.** The procedure may lead to rejection of the null hypothesis when it is true.
2. **Type II error.** The procedure may fail to reject the null hypothesis when it is false.

To continue the previous example, there is some probability that the estimate of the parameter will be quite far from the hypothesized value, even if the hypothesis is true. This might cause a type I error.

Definition. The probability of a type I error is the **size** of the test. This is conventionally denoted α and is also called the **significance level**.

The size of the test is under the control of the analyst. It can be changed just by changing the decision rule. Indeed, the type I error could be eliminated altogether just by making the rejection region very small. But this would come at a cost. By eliminating the probability of a type I error, that is, by making it unlikely that the hypothesis is rejected, we must increase the probability of a type II error. Ideally, we would like both probabilities to be as small as possible. However, it is clear that there is a tradeoff between the two. The best we can hope for is that for a given probability of type I error, the procedure we choose will have as small a probability of type II error as possible.

Definition. The **power** of a test is the probability that it will correctly lead to rejection of a false null hypothesis:

$$\text{power} = 1 - \beta = 1 - \text{Prob}(\text{type II error}). \quad (4-59)$$

For a given significance level, α , we would like β to be as small as possible. Since β is defined in terms of the alternative hypothesis, it depends on the true value of the parameter.

EXAMPLE 4.30

For testing $H_0: \mu = \mu^0$ in a normal distribution with known variance, σ^2 , the decision rule is to reject the hypothesis if the absolute value of the z statistic, $\sqrt{n}(\bar{x} - \mu^0)/\sigma$, exceeds the predetermined critical value. For a test at the 5 percent significance level, we set the critical value at 1.96. The power of the test is, therefore, the probability that the absolute value of the test statistic will exceed 1.96 given that the true value of μ is, in fact, not μ^0 . This depends on the alternative value of μ , as shown in Figure 4.7. Notice that for this test the power is equal to the size at the point where μ equals μ^0 . As might be expected, the test becomes more powerful the farther the true mean is from the hypothesized value.

Construct a 95 percent confidence interval for μ . Assuming that the sample of 25 is from a normal distribution,

$$\text{Prob}\left(-2.064 < \frac{5(\bar{x} - \mu)}{s} < 2.064\right) = 0.95,$$

where 2.064 is the critical value from a t distribution with 24 degrees of freedom. Thus, the confidence interval is

$$1.63 \pm \frac{2.064(0.51)}{5} \quad \text{or} \quad [1.4195, 1.8405].$$

Remark. Had the parent distribution not been specified, it would have been natural to use the standard normal distribution instead, perhaps relying on the central limit theorem. But a sample size of 25 is small enough that the more conservative t distribution might still be preferable.

The theory does not prescribe exactly how to choose the endpoints for the confidence interval. An obvious criterion is to minimize the width of the interval. If the sampling distribution is symmetric, the symmetric interval is the best one. If the sampling distribution is not symmetric, however, this procedure will not be optimal.

EXAMPLE 4.29

The chi-squared distribution is used to construct a confidence interval for the variance of a normal distribution. Using the data from Example 4.28, we find that the usual procedure would use

$$\text{Prob}\left(12.4 < \frac{24s^2}{\sigma^2} < 39.4\right) = 0.95,$$

where 12.4 and 39.4 are the 0.025 and 0.975 cutoff points from the chi-squared(24) distribution. This leads to the 95 percent confidence interval [0.1581, 0.5032]. By making use of the asymmetry of the distribution, a narrower interval can be constructed. Allocating 4 percent to the left-hand tail and 1 percent to the right instead of 2.5 percent to each, the two cutoff points are 13.4 and 42.9, and the resulting 95 percent confidence interval is [0.1455, 0.4659].²¹

Finally, the confidence interval can be manipulated to obtain a confidence interval for a function of a parameter. For example, based on the preceding, a 95 percent confidence interval for σ would be $[\sqrt{0.1581}, \sqrt{0.5032}] = [0.3976, 0.7094]$.

4.8. Hypothesis Testing

The second major group of statistical inference procedures is hypothesis tests. The classical testing procedures are based on constructing a statistic from a random sample that will enable the analyst to decide, with reasonable confidence, whether or not the data in the sample would have been generated by a hypothesized population. The formal procedure involves a statement of the hypothesis, usually in terms of a "null" or maintained hypothesis and an alternative, conventionally denoted H_0 and H_1 , respectively. The procedure

²¹ Formal conditions for finding the narrowest interval are considered in Exercise 19.



On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals

Nathaniel SCHENKER and Jane F. GENTLEMAN

To judge whether the difference between two point estimates is statistically significant, data analysts often examine the overlap between the two associated confidence intervals. We compare this technique to the standard method of testing significance under the common assumptions of consistency, asymptotic normality, and asymptotic independence of the estimates. Rejection of the null hypothesis by the method of examining overlap implies rejection by the standard method, whereas failure to reject by the method of examining overlap does not imply failure to reject by the standard method. As a consequence, the method of examining overlap is more conservative (i.e., rejects the null hypothesis less often) than the standard method when the null hypothesis is true, and it mistakenly fails to reject the null hypothesis more frequently than does the standard method when the null hypothesis is false. Although the method of examining overlap is simple and especially convenient when lists or graphs of confidence intervals have been presented, we conclude that it should not be used for formal significance testing unless the data analyst is aware of its deficiencies and unless the information needed to carry out a more appropriate procedure is unavailable.

KEY WORDS: Efficiency; Inference; Power; Test of significance; Two-sample problem; Type I error.

1. INTRODUCTION

To judge whether the difference between two point estimates is statistically significant, data analysts sometimes examine the overlap between the two associated confidence intervals. If there is no overlap, the difference is judged significant, and if there is overlap, the difference is not judged significant. This method is simple to use because it is easy to compare boundary values from two confidence intervals to see whether the intervals overlap. Moreover, it seems natural to inspect graphed confidence intervals for overlap. Over a number of years, however, we have tried to discourage the practice of examining overlap to assess significance among authors of articles that we have reviewed, because the procedure can lead to mistaken conclusions.

A nonexhaustive search of abstracts and articles in the health sciences found more than 60 articles in which the method of examining overlap was used or recommended to demonstrate

significance and/or lack of significance, either formally or informally. Recent examples were found in 22 different sources: *American Heart Journal*; *American Journal of Epidemiology*; *Annals of the Rheumatic Diseases*; *British Medical Journal*; *Cancer Epidemiology Biomarkers & Prevention*; *Cancer Research*; *CDC Surveillance Summaries*; *Clinical and Experimental Allergy*; *Clinical Cancer Research*; *Clinical Neuropharmacology*; *Emerging Infectious Diseases*; *Health Education Research*; *International Journal of Epidemiology*; *Journal of Clinical Epidemiology*; *Journal of Medical Entomology*; *Journal of Medical Genetics*; *Journal of the National Cancer Institute*; *Journal of Nuclear Medicine*; *The Clinical Journal of Pain*; *The Cochrane Library*; *The Journal of Nutrition*; and *The Lancet*. Respective examples of articles are: Abergel et al. (1998); Sont et al. (2001); Gotzsche (2000); Appleby and Bell (2000); Millikan et al. (2000); Martinez et al. (1999); Kann et al. (2000); Ezeamuzie et al. (1999); Tersmette et al. (2001); Inzelberg et al. (2000); Chalmers and Salmon (2000); McBride (2000); Wong et al. (1999); Mancuso, Peterson, and Charlson (2001); Debboun et al. (2000); Slavotinek et al. (1999); Djordjevic, Stellman, and Zang (2000); Eising and von der Ohe (1998); Herr et al. (1998); Collins et al. (2001); Vatassery, Smith, and Quach (1998); and Mehta, Eikelboom, and Yusuf (2000).

Discussion of properties of the method of examining overlap in the literature is limited and contradictory. For example, an exchange in the *Journal of the American Academy of Dermatology* (Bigby and Gadenne 1996; Rahlfs 1997; Bigby and Gadenne 1997) offered contradictory advice concerning the method. A follow-up letter in that journal by Cole and Blair (1999) provided a more accurate assessment of the method along with some Monte Carlo results for a few situations.

This article considers the common task of comparing estimates that are assumed to be consistent, asymptotically normal, and asymptotically independent. It is shown that rejection of the null hypothesis by the method of examining overlap implies rejection by the standard method, whereas failure to reject by the method of examining overlap does not imply failure to reject by the standard method. Thus, the method of examining overlap is more conservative (i.e., rejects the null hypothesis less often) than the standard method when the null hypothesis is true, and it mistakenly fails to reject the null hypothesis more frequently than does the standard method when the null hypothesis is false. We expand on the work of Cole and Blair (1999) by providing more theoretical insight into the properties of the methods and by developing asymptotic results on efficiency, probability of Type I error, and power for a broad range of situations. Our goal is to provide an accessible discussion for statistical practitioners in subject-matter areas and for statistical consultants. Throughout the article, we provide simple examples and figures that can be used in consulting situations to illustrate the results.

Nathaniel Schenker is Senior Scientist for Research and Methodology, National Center for Health Statistics, Hyattsville, MD 20782 (E-mail: nschenker@cdc.gov). Jane F. Gentleman is Director, Division of Health Interview Statistics, National Center for Health Statistics, Hyattsville, MD 20782 (E-mail: jgentleman@cdc.gov). The authors thank Jennifer Madans of the National Center for Health Statistics for her encouragement.

Section 2 describes our general setup and then derives the main result underlying the conservatism and low power of the method of examining overlap relative to the standard method. Section 3 presents asymptotic numerical comparisons of the methods. Section 4 briefly discusses another simple but sometimes misleading method of examining overlap that we have seen used. Section 5 discusses extensions and other aspects of our results.

2. SETUP AND MAIN RESULT

2.1 Definitions and Assumptions

Suppose that we wish to make inference about the difference between two population quantities, Q_1 and Q_2 , which are estimated by $\hat{Q}_1$ and $\hat{Q}_2$, and suppose that the associated standard errors, SE_1 and SE_2 , are estimated by $\widehat{SE}_1$ and $\widehat{SE}_2$.

Our development assumes that we have consistent estimates and that $\hat{Q}_1$ and $\hat{Q}_2$ are asymptotically normal and independent. More formally, we assume that $\hat{Q}_1 \xrightarrow{P} Q_1$, $\hat{Q}_2 \xrightarrow{P} Q_2$, $\widehat{SE}_1/SE_1 \xrightarrow{P} 1$, $\widehat{SE}_2/SE_2 \xrightarrow{P} 1$, $(\hat{Q}_1 - Q_1)/SE_1 \xrightarrow{D} Z_1$, and $(\hat{Q}_2 - Q_2)/SE_2 \xrightarrow{D} Z_2$ as the sample sizes increase to ∞ , where Z_1 and Z_2 are independent standard normal random variables.

It follows from standard theorems on convergence (see, e.g., Cramér 1946, pp. 254–255) that $(\hat{Q}_1 - Q_1)/\widehat{SE}_1$, $(\hat{Q}_2 - Q_2)/\widehat{SE}_2$, and $[(\hat{Q}_1 - \hat{Q}_2) - (Q_1 - Q_2)]/\sqrt{\widehat{SE}_1^2 + \widehat{SE}_2^2}$ have limiting standard normal distributions. Based on these three approximate pivotal quantities (Brownlee 1965, sec. 2.16), nominal 95% confidence intervals for Q_1 , Q_2 , and $Q_1 - Q_2$ are given by

$$\hat{Q}_1 \pm 1.96\widehat{SE}_1, \quad (1)$$

$$\hat{Q}_2 \pm 1.96\widehat{SE}_2, \quad (2)$$

and

$$(\hat{Q}_1 - \hat{Q}_2) \pm 1.96\sqrt{\widehat{SE}_1^2 + \widehat{SE}_2^2}, \quad (3)$$

respectively.

By the duality between $100(1 - \alpha)\%$ confidence intervals and significance tests at level α (Bickel and Doksum 1977, p. 178),

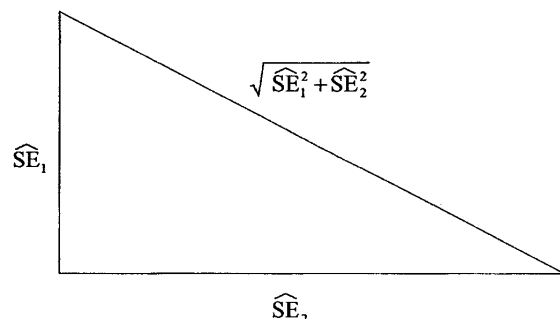


Figure 1. Right triangle depicting the relationship between the estimated standard errors of $\hat{Q}_1$, $\hat{Q}_2$, and $\hat{Q}_1 - \hat{Q}_2$.

a test at nominal level .05 of the null hypothesis that $Q_1 - Q_2 = 0$ can be carried out by examining whether the nominal 95% interval (3) contains 0. The null hypothesis is rejected if and only if the interval does not contain 0. We will refer to this as the “standard” method. In contrast, the “overlap” method conducts a test at nominal level .05 by rejecting the null hypothesis if and only if the nominal 95% intervals (1) and (2) do not overlap.

2.2 Example: Comparing Proportions

Suppose we wish to compare two large populations with respect to the proportions of people with a given attribute. Independent simple random samples, each of size 200, have been drawn, and 112 people in the first sample have the attribute, whereas 88 people in the second sample have the attribute.

The estimated proportions and their estimated standard errors are $\hat{Q}_1 = 112/200 = .56$, $\hat{Q}_2 = 88/200 = .44$, $\widehat{SE}_1 = \sqrt{(.56)(.44)/200} = .0351$, and $\widehat{SE}_2 = \sqrt{(.44)(.56)/200} = .0351$, respectively. (Finite population corrections have been ignored since the populations are large.) Thus, intervals (1) and (2) for the population proportions Q_1 and Q_2 are $.56 \pm (1.96)(.0351) = [.49, .63]$ and $.44 \pm (1.96)(.0351) = [.37, .51]$, respectively. The overlap method leads us to declare that the proportions are not significantly different.

In contrast, since interval (3) for the difference $Q_1 - Q_2$, $(.56 - .44) \pm 1.96\sqrt{(.0351)^2 + (.0351)^2} = [.02, .22]$, does not contain 0, the standard method would lead us to declare that the proportions are significantly different.

2.3 Main Result

The preceding example illustrates that the overlap method can fail to reject the null hypothesis when the standard method would reject it. In general, rejection of the null hypothesis by the overlap method implies rejection by the standard method, but not vice versa. Thus, the overlap method is more conservative and less powerful than the standard method.

To derive this result, we begin by noting that intervals (1) and (2) overlap if and only if the interval

$$(\hat{Q}_1 - \hat{Q}_2) \pm 1.96(\widehat{SE}_1 + \widehat{SE}_2) \quad (4)$$

contains 0. Thus, the difference between the standard method and the overlap method is only in the width of the underlying intervals (3) and (4). The ratio of the width of interval (4) to the width of interval (3) is

$$\frac{\widehat{SE}_1 + \widehat{SE}_2}{\sqrt{\widehat{SE}_1^2 + \widehat{SE}_2^2}}. \quad (5)$$

Figure 1 displays a right triangle depicting the relationship between the estimated standard errors of $\hat{Q}_1$, $\hat{Q}_2$, and $\hat{Q}_1 - \hat{Q}_2$. Expression (5) is equal to the ratio of the sum of the lengths of the two sides adjacent to the right angle to the length of the hypotenuse. It is clear from the figure that the ratio is greater than 1, which demonstrates our result.

It is also straightforward to show that ratio (5) achieves its maximum value, $\sqrt{2}$, when the two adjacent sides of the triangle in Figure 1 are of equal length, and that it approaches its minimum value, 1, as one adjacent side increases in length rel-

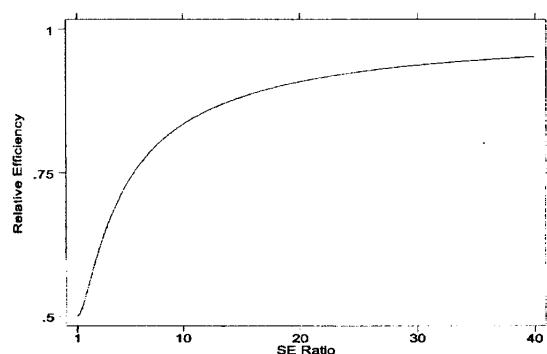


Figure 2. Asymptotic relative efficiency of the overlap method to the standard method.

ative to the other. Thus, the overlap method is expected to be more deficient relative to the standard method when SE_1 and SE_2 are nearly equal, and less deficient when one standard error is large relative to the other.

3. ASYMPTOTIC COMPARISONS

To illustrate the deficiencies of the overlap method due to the length of its underlying interval relative to the standard method, we provide asymptotic numerical comparisons (as the sample sizes increase to ∞) concerning efficiency, probability of Type I error, and power. Our results all follow from standard theorems on convergence (see, e.g., Cramér 1946, pp. 254–255). Derivations are available from the authors upon request. We investigate performance under various limiting values of the ratio of one standard error to the other. Our results when the limiting value of SE_2/SE_1 is k are the same as those when the limiting value of SE_1/SE_2 is k , so we simply refer to the limiting value as the “SE ratio” and consider only ratios that are greater than or equal to 1.

3.1 Efficiency

The asymptotic relative efficiency (ARE) of the overlap method to the standard method may be defined as the limit in probability of the ratio of the squared width of interval (3) to the squared width of interval (4), that is, the limit in probability of the squared inverse of expression (5). For an SE ratio of k , the ARE is given by $(1 + k^2)/(1 + k)^2$. The ARE is displayed in Figure 2 for a range of values of the SE ratio. As anticipated at the end of Section 2.3, the minimum ARE, $1/2$, occurs when the SE ratio is 1. As the SE ratio increases, the ARE approaches 1.

Suppose, as in the example of Section 2.2, that the standard errors of $\hat{Q}_1$ and $\hat{Q}_2$ are of the form $\sigma_1/\sqrt{n_1}$ and $\sigma_2/\sqrt{n_2}$, that is, inversely proportional to the square roots of the sample sizes underlying $\hat{Q}_1$ and $\hat{Q}_2$. Then the ARE can be interpreted as the asymptotic ratio of the sample sizes that the standard method and the overlap method would need to produce underlying intervals of the same width and to achieve the same power. For example, when the SE ratio is 1, the standard method needs asymptotically only half of the sample sizes that the overlap method needs.

3.2 Probability of Type I Error and Power

Since the limiting power of either the standard method or the overlap method is 1 for any fixed, nonzero hypothesized alternative value of $Q_1 - Q_2$, we consider a sequence of alternative values such that $(Q_1 - Q_2)/\sqrt{SE_1^2 + SE_2^2}$ approaches a finite value, which we call the “standardized difference,” as the sample sizes increase. Thus, we effectively specify alternative values of $Q_1 - Q_2$ that tend to 0, and we force the powers of the two methods to converge to values that are less than 1 and not necessarily equal. This allows a more meaningful comparison of the two methods. [See Cox and Hinkley (1974, pp. 317–318), for similar uses of sequences of alternative hypotheses.]

For an SE ratio of k and a standardized difference of d , when tests of significance are conducted at nominal level .05, the asymptotic powers of the standard method and the overlap method are given by

$$\Phi(-1.96 + d) + \Phi(-1.96 - d), \quad (6)$$

and

$$\Phi\left(\frac{-1.96(1+k)}{\sqrt{1+k^2}} + d\right) + \Phi\left(\frac{-1.96(1+k)}{\sqrt{1+k^2}} - d\right), \quad (7)$$

respectively, where Φ denotes the standard normal cumulative distribution function.

The asymptotic probability of Type I error is just the asymptotic power when the standardized difference is 0 (i.e., when the null hypothesis is true), so expressions (6) and (7) can be used to calculate the asymptotic probability of Type I error by setting $d = 0$. Figure 3 displays asymptotic probabilities of Type I error for the standard method and overlap method for a range of values of the SE ratio. The probability for the standard method is equal to the nominal level of .05 regardless of the value of the SE ratio, whereas the probability for the overlap method is smaller than .05 and does not even reach .03 until the SE ratio nears the value 9.

Figure 4 displays the asymptotic powers of the standard method and the overlap method as functions of the standardized difference, for SE ratios of 1 and 8. For each method and SE ratio, the asymptotic power for a standardized difference of

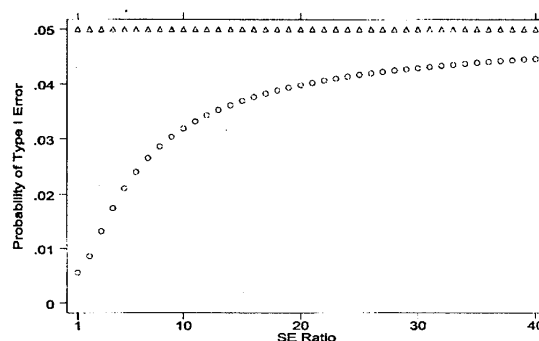


Figure 3. Asymptotic probabilities of Type I error for the overlap method and the standard method when the nominal level of tests is .05. $\circ$ = overlap method; $\triangle$ = standard method.

5. DISCUSSION

5.1 Relationship to Study Design

If a study has been designed specifically for the comparison of two population quantities, then it is likely that the two associated standard errors will not be very different and thus that the overlap method will have substantial deficiencies. For example, consider a stratified random sample with a major goal being the comparison of two stratum means for a variable. If we ignore finite population corrections, the variance of the difference between the stratum estimates is of the form $\sigma_1^2/n_1 + \sigma_2^2/n_2$, where σ_1^2 and σ_2^2 are the stratum variances and n_1 and n_2 are the sample sizes. The variance of the difference is minimized for a given total sample size by setting $n_1/n_2 = \sigma_1/\sigma_2$. Thus, provided that σ_1 and σ_2 do not differ drastically, the individual standard errors, $SE_1 = \sigma_1/\sqrt{n_1}$ and $SE_2 = \sigma_2/\sqrt{n_2}$ will not have very different relative sizes. Moreover, the effects of any differences between σ_1 and σ_2 on the individual standard errors will be mitigated somewhat by the fact that more of the sample is allocated to the stratum with the higher variance. Conversely, when a study has not been designed for a specific comparison, as might be the case with a general-purpose survey, the relative sizes of the standard errors have more of a chance of being very different, in which case the overlap method could be less deficient.

5.2 Effects of Correlation

Another consideration is the effect of correlation between the point estimates $\hat{Q}_1$ and $\hat{Q}_2$, which is not an issue when the estimates are independent, as we have assumed to be true asymptotically up to now. In the presence of a non-zero correlation, say ρ , the standard error of $\hat{Q}_1 - \hat{Q}_2$ is $\sqrt{SE_1^2 + SE_2^2 - 2\rho SE_1 SE_2}$. Thus, if $\rho > 0$, then the overlap method tends to be even more conservative and less powerful, relative to a method that uses an appropriate estimate of standard error, than has been indicated by our results. If $-1 < \rho < 0$, then the overlap method still tends to be deficient, but less so than has been indicated by our results. The degree of deficiency decreases as the correlation becomes more negative. In the unlikely case of a correlation equal to -1 , the standard error of $\hat{Q}_1 - \hat{Q}_2$ reduces to $SE_1 + SE_2$, and the interval (4) underlying the overlap method corresponds to an interval formed using an appropriate estimate of standard error. It follows that, in the case of perfect negative correlation, the overlap method will tend to be appropriate.

Examples of situations in which positive correlations can occur are blocked experiments, longitudinal studies, and analyses of data in which there is overlap in the sets of elements used to compute the two estimates being compared. Complex survey designs can result in either positive or negative correlations, even when there is no overlap in the sets of elements used for the two estimates. For example, in a comparison of males with females, a correlation could result from the two groups having households in common.

5.3 Conclusions

The overlap method is simple, and it is convenient when lists or graphs of confidence intervals are presented. It can be use-

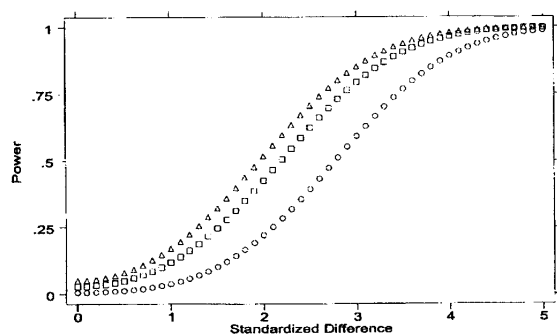


Figure 4. Asymptotic powers of the overlap method and the standard method when the nominal level of tests is .05. $\circ$ = overlap method (SE ratio = 1); $\square$ = overlap method (SE ratio = 8); $\triangle$ = standard method.

d is equal to that for a standardized difference of $-d$, so we only show values for standardized differences that are greater than or equal to 0. Moreover, for a given value of the standardized difference, the asymptotic power of the standard method does not depend on the SE ratio, which is why there is only one graph for the standard method.

For the SE ratio of 1, large differences in asymptotic power occur for intermediate values of the standardized difference. For example, for a standardized difference of 2.8, which corresponds to the difference between Q_1 and Q_2 being about the same size as the width of interval (1) for Q_1 or interval (2) for Q_2 , the power of the standard method is .80, whereas the power of the overlap method is .51. The power of the overlap method is much closer to that of the standard method when the SE ratio is 8. For example, for a standardized difference of 2.8, the power of the overlap method is .73.

4. EXAMINING WHETHER EITHER INTERVAL CONTAINS THE OTHER POINT ESTIMATE

Although it is not the focus of this article, we briefly mention another method that we have seen used, which declares a difference to be significant if and only if interval (1) does not contain the value $\hat{Q}_2$ and interval (2) does not contain the value $\hat{Q}_1$. {Another version declares significance if one interval does not contain the other point estimate; see, e.g., Cromwell et al. (1996).} Under the assumptions of Section 2.1, it can be shown that this method tends to be anti-conservative; that is, it rejects the null hypothesis more often than it should when the null hypothesis is true.

As a simple example, consider the setup of Section 2.2, but suppose that the first and second samples have, respectively, 108 and 92 people with the attribute. Then the point estimates are $\hat{Q}_1 = .54$ and $\hat{Q}_2 = .46$, and the confidence intervals for Q_1 and Q_2 are $[.47, .61]$ and $[.39, .53]$, respectively. The method of examining whether either interval contains the other point estimate leads one to declare that the two proportions are significantly different. The standard analysis, however, declares the difference to be nonsignificant since the interval for $Q_1 - Q_2$, $[-.02, .18]$, contains 0.

ful as a quick and relatively rough method for exploratory data analysis. It should not be regarded as an optimal method for significance testing, however, given its conservatism and low power relative to the standard method in the common situation that we have considered. Thus, the overlap method should not be used for formal significance testing unless the data analyst is aware of its deficiencies and unless the information needed to carry out a more appropriate procedure is unavailable.

[Received June 2000. Revised April 2001.]

REFERENCES

- Abergel, E., Linhart, A., Chatellier, G., Garipey, J., Ducardonnet, A., Diebold, B., and Menard, J. (1998), "Vascular and Cardiac Remodeling in World Class Professional Cyclists," *American Heart Journal*, 136, 818–823.
- Appleby, J., and Bell, A. (2000), "Reporting NHS Performance: How Did the Media Perform?" *British Medical Journal*, 321, 248.
- Bickel, P. J., and Doksum, K. A. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics*. San Francisco: Holden-Day.
- Bigby, M., and Gadenne, A. (1996), "Understanding and Evaluating Clinical Trials," *Journal of the American Academy of Dermatology*, 34, 555–590.
- (1997), Reply to comments on "Understanding and Evaluating Clinical Trials," *Journal of the American Academy of Dermatology*, 37, 804–805.
- Brownlee, K. A. (1965), *Statistical Theory and Methodology in Science and Engineering*. New York: Wiley.
- Chalmers, R. M., and Salmon, R. L. (2000), "Primary Care Surveillance for Acute Bloody Diarrhea, Wales," *Emerging Infectious Diseases*, 6, 412–414.
- Cole, S. R., and Blair, R. C. (1999), "Overlapping Confidence Intervals," *Journal of the American Academy of Dermatology*, 41, 1051–1052.
- Collins, S. L., Edwards, J. E., Moore, R. A., and McQuay, H. J. (2001), "Single Dose Dextropropoxyphene, Alone and with Paracetamol (Acetaminophen), for Postoperative Pain" (Cochrane Review), in *The Cochrane Library*, 1, 2001, Oxford: Update Software.
- Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*. London: Chapman and Hall.
- Cramér, H. (1946), *Mathematical Methods of Statistics*. Princeton: Princeton University Press.
- Cromwell, G. L., Davis, G. W., Morrow, W. E., Primo, R. A., Rozeboom, D. W., Sims, M. D., Stanisiewski, E. P., and Ho, C. H. (1996), "Efficacy of the Antimicrobial Compound U-82,127 as a Growth Promoter for Growing-Finishing Pigs," *Journal of Animal Science*, 74, 1284–1287.
- Debboun, M., Strickman, D., Solberg, V. B., Wilkerson, R. C., McPherson, K. R., Golenda, C., Keep, L., Wirtz, R. A., Burge, R., and Klein, T. A. (2000), "Field Evaluation of Deet and a Piperidine Repellent Against *Aedes Communis* (Diptera: Culicidae) and *Simulium Venustum* (Diptera: Simuliidae) in the Adirondack Mountains of New York," *Journal of Medical Entomology*, 37, 919–923.
- Djordjevic, M. V., Stellman, S. D., and Zang, E. (2000), "Doses of Nicotine and Lung Carcinogens Delivered to Cigarette Smokers," *Journal of the National Cancer Institute*, 92, 106–111.
- Eising, E. G., and von der Ohe, M. R. (1998), "Differentiation of Prolonged Colonic Transit Using Scintigraphy with Indium-111-Labeled Polystyrene Pellets," *Journal of Nuclear Medicine*, 39, 1062–1066.
- Ezeamuzie, C. I., Al-Ali, S. F., Al-Dowaisan, A., Khan, M., Hijazi, Z., and Thomson, M. S. (1999), "Reference Values of Total Serum IgE and Their Significance in the Diagnosis of Allergy Among the Young Adult Kuwaiti Population," *Clinical and Experimental Allergy*, 29, 375–381.
- Gotzsche, P. C. (2000), "Do Patients with Osteoarthritis Get the Clinical Research They Need?" *Annals of the Rheumatic Diseases*, 59, 407–408.
- Herr, K. A., Mobily, P. R., Kohout, F. J., and Wagenaar, D. (1998), "Evaluation of the Faces Pain Scale for Use with the Elderly," *The Clinical Journal of Pain*, 14, 29–38.
- Inzelberg, R., Carasso, R. L., Schechtman, E., and Nispeanu, P. (2000), "A Comparison of Dopamine Agonists and Catechol-O-Methyltransferase Inhibitors in Parkinson's Disease," *Clinical Neuropharmacology*, 23, 262–266.
- Kann, L., Kinchen, S. A., Williams, B. I., Ross, J. G., Lowry, R., Grunbaum, J. A., and Kolbe, L. J. (2000), "Youth Risk Behavior Surveillance—United States, 1999," *CDC Surveillance Summaries*, MMWR 2000, 49(SS-5).
- Mancuso, C. A., Peterson, M. G. E., and Charlson, M. E. (2001), "Comparing Discriminative Validity Between a Disease-Specific and a General Health Scale in Patients with Moderate Asthma," *Journal of Clinical Epidemiology*, 54, 263–274.
- Martinez, M. E., Maltzman, T., Marshall, J. R., Einspahr, J., Reid, M. E., Sampliner, R., Ahnen, D. J., Hamilton, S. R., and Alberts, D. S. (1999), "Risk Factors for Ki-Ras Protooncogene Mutation in Sporadic Colorectal Adenomas," *Cancer Research*, 59, 5181–5185.
- McBride, N. (2000), "The Western Australian School Health Project: Comparing the Effects of Intervention Intensity on Organizational Support for School Health Promotion," *Health Education Research*, 15, 59–72.
- Mehta, S. R., Eikelboom, J. W., and Yusuf, S. (2000), Reply to comments on "Intracranial Haemorrhage with Bolus Thrombolytic Agents," *The Lancet*, 356, 1850.
- Millikan, R., Pittman, G., Tse, C.-K., Savits, D. A., Newman, B., and Bell, D. (2000), "Glutathione S-Transferases M1, T1, and P1 and Breast Cancer," *Cancer Epidemiology Biomarkers & Prevention*, 9, 567–573.
- Rahlf, V. W. (1997), Comment on "Understanding and Evaluating Clinical Trials," *Journal of the American Academy of Dermatology*, 37, 803–804.
- Slavotinek, A., Rosenberg, M., Knight, S., Gaunt, L., Fergusson, W., Killoran, C., Clayton-Smith, J., Kingston, H., Campbell, R. H., Flint, J., Donnai, D., and Biesecker, L. (1999), "Screening for Submicroscopic Chromosome Rearrangements in Children with Idiopathic Mental Retardation Using Microsatellite Markers for the Chromosome Telomeres," *Journal of Medical Genetics*, 36, 405–411.
- Sont, W. N., Zielinski, J. M., Ashmore, J. P., Jiang, H., Krewski, D., Fair, M. E., Band, P. R., and Létoimeau, E. G. (2001), "First Analysis of Cancer Incidence and Occupational Radiation Exposure Based on the National Dose Registry of Canada," *American Journal of Epidemiology*, 153, 309–318.
- Tersmette, A. C., Petersen, G. M., Offerhaus, G. J. A., Falatko, F. C., Brune, K. A., Goggins, M., Rozenblum, E., Wilentz, R. E., Yeo, C. J., Cameron, J. L., Kern, S. E., and Hruban, R. H. (2001), "Increased Risk of Incident Pancreatic Cancer Among First-Degree Relatives of Patients with Familial Pancreatic Cancer," *Clinical Cancer Research*, 7, 738–744.
- Vatassery, G. T., Smith, W. E., and Quach, H. T. (1998), "α-Tocopherol in Rat Brain Subcellular Fractions is Oxidized Rapidly During Incubations with Low Concentrations of Peroxynitrite," *The Journal of Nutrition*, 128, 152–157.
- Wong, C. M., Hu, Z. G., Lam, T. H., Hedley, A. J., and Peters, J. (1999), "Effects of Ambient Air Pollution and Environmental Tobacco Smoke on Respiratory Health of Non-Smoking Women in Hong Kong," *International Journal of Epidemiology*, 28, 859–864.



2.9 Testing of Hypotheses

We list here some essential points regarding hypothesis testing.

1. A statistical hypothesis is a statement about the values of some parameters in the hypothetical population from which the sample is drawn.

2. A hypothesis which says that a parameter has a specified value is called a *point hypothesis*. A hypothesis which says that a parameter lies in a specified interval is called an *interval hypothesis*. For instance, if μ is the population mean, then $H: \mu = 4$ is a point hypothesis. $H: 4 \leq \mu \leq 7$ is an interval hypothesis.

3. A hypothesis test is a procedure that answers the question of whether the observed difference between the *sample value* and the *population value* hypothesized is real or due to chance variation. For instance, if the hypothesis says that the population mean $\mu = 6$ and the sample mean $\bar{y} = 8$, then we want to know whether this difference is real or due to chance variation.

4. The hypothesis we are testing is called the *null hypothesis* and is often denoted by H_0 .² The alternative hypothesis is denoted by H_1 . The probability of rejecting H_0 when, in fact, it is true, is called the *significance level*. To test whether the observed difference between the data and what is expected under the null hypothesis H_0 is real or due to chance variation, we use a *test statistic*. A desirable criterion for the test statistic is that its sampling distribution be tractable, preferably with tabulated probabilities. Tables are already available for the normal, t , χ^2 , and F distributions, and hence the test statistics chosen are often those that have these distributions.

5. The *observed significance level* or *P-value* is the probability of getting a value of the test statistic that is as extreme or more extreme than the observed value of the test statistic. This probability is computed on the basis that the null hypothesis is correct. For instance, consider a sample of n independent observations from a normal population with mean μ and variance σ^2 . We want to test

$$H_0: \mu = 7 \quad \text{against} \quad H_1: \mu \neq 7$$

The test statistic we use is

$$t = \frac{\sqrt{n}(\bar{y} - \mu)}{S}$$

which has a t -distribution with $(n - 1)$ degrees of freedom. Suppose that $n = 25$, $\bar{y} = 10$, $S = 5$. Then under the assumption that H_0 is true, the observed value of t is $t_0 = 3$. Since high positive values of t are evidence against the null hypothesis H_0 , the P -value is [since degrees of freedom $(n - 1) = 24$]

$$P = \text{Prob}(t_{24} > 3)$$

This is the observed significance level.

6. It is common practice to say simply that the result of the test is (statistically) significant or not significant and not report the actual P -values. The meaning of the two terms is as follows:

²This terminology is unfortunate because "null" means "zero, void, insignificant, amounts to nothing, etc." A hypothesis $\mu = 0$ can be called a null hypothesis, but a hypothesis $\mu = 100$ should not be called a "null" hypothesis. However, this is the standard terminology that was introduced in the 1930s by the statisticians Jerzy Neyman and E. S. Pearson.

Statistically significant. Sampling variation is an unlikely explanation of the discrepancy between the null hypothesis and sample values.

Statistically insignificant. Sampling variation is a likely explanation of the discrepancy between the null hypothesis and the sample value.

Also, the terms *significant* and *highly significant* are customarily used to denote "significant at the 0.05 level" and "significant at the 0.01 level" respectively. However, consider two cases where the P -values are 0.055 and 0.045, respectively. Then in the former case one would say that the results are "not significant" and in the latter case one would say that the results are "significant," although the sample evidence is marginally different in both cases. Similarly, two tests with P -values of 0.80 and 0.055 will both be considered "not significant," although there is a tremendous difference in the compatibility of the sample evidence with the null hypothesis in the two cases. That is why many computer programs print out P -values.³

7. *Statistical significance* and *practical significance* are not the same thing. A result that is highly significant statistically may be of no practical significance at all. For instance, suppose that we consider a shipment of cans of cashews with expected mean weight of 450 g. If the actual sample mean of weights is 449.5 g, the difference may be practically insignificant but could be highly statistically significant if we have a large enough sample or a small enough sampling variance (note that the test statistic has $\sqrt{n}$ in the numerator and S in the denominator). On the other hand, in the case of precision instruments, a part is expected to be of length 10 cm and a sample had a mean length of 9.9 cm. If n is small and S is large, the difference could not be statistically significant but could be practically very significant. The shipment could simply be useless.

8. It is customary to reject the null hypothesis H_0 when the test statistic is statistically significant at a chosen significance level and not to reject H_0 when the test statistic is not statistically significant at the chosen significance level. There is, however, some controversy on this issue which we discuss later in item 9. In reality, H_0 may be either true or false. Corresponding to the two cases of reality and the two conclusions drawn, we have the following four possibilities:

Result of the Test	Reality	
	H_0 Is True	H_0 Is False
Significant (reject H_0)	Type I error or α error	Correct conclusion
Not significant (do not reject H_0)	Correct conclusion	Type II error or β error