Research Note:
Statistical Testing of Single-Subject Word- and Phoneme- Recognition Scores

George G. Woodworth

Department of Statistics and Actuarial Science
and
Department of Preventive Medicine, Division of Biostatistics

University of Iowa, Iowa City, Iowa

April 1994

ABSTRACT

Thornton and Raffin's critical intervals for comparing binomial single-subject open-set speech recognition scores are equivalent to the arcsin-transformed z-test for independent proportions. However, it is necessary to adjust for context effects when comparing test scores in which subjects' responses to stimulus words or phrases can receive multinomial scores (e.g., phonemes in monosyllables or words in sentences). Boothroyd and Nittouer suggested a simple, approximate sample size adjustment for context effects, which works well in practice, but sometimes produces incorrect inferences. Neither method is valid for comparing two open-set speech recognition scores, whether binomial or multinomial, when the same stimulus list was used on both occasions. This paper places Thornton and Raffin's and Boothroyd and Nittouer's methods in a general class of randomization tests first proposed by R.A. Fisher in 1935. Such tests compare the difference between two scores to the expected variation induced by random selection of the stimulus list.

Randomization tests provide exact significance levels (p-values) for comparing open-set different-list speech recognition scores; however special software is needed to compute exact p-values. Approximations and bounds for the exact p-values are discussed, including Thornton and Raffin's arcsine approximation for binomial data, the z approximation to Fisher's randomization test, and Boothroyd and Nittouer's (BN) binomial approximation with sample size adjustment. The accuracy the approximations is compared for 160 pairwise comparisons of Iowa Sentence scores in 23 Cochlear implant wearers. For statistical tests conducted at the 5% level, BN produced about 10 percent false positives and 6 percent false negatives vis-a-vis the exact test while the z-approximation produced no false positives and less than 2 percent false negatives. The results were similar for inferences at the 1% significance level.

1 This study was supported by research grant DC00242 from the National Institutes of Health/NIDCD, grant RR59 from the General Clinical Research Centers Program, Division of Research Resources, NIH, and the Iowa Lions Sight and Hearing Foundation.
Comparison of same-list, open-set speech recognition scores is a paired-sample problem. The binomial case reduces to McNemar's test; i.e., the binomial test applied to the reduced set of stimuli which received different responses on the two occasions. Exact significance levels for the multinomial case are provided by a test of bivariate symmetry described by Kurtz, Landis, and Koch. Convenient approximations and bounds for exact p-values are discussed.

Exact and approximate methods are illustrated and approximations are assessed with Sentence Test data from the Iowa Cochlear Implant Project.

**INTRODUCTION**

Many speech-recognition tests are administered as a list of independent stimuli and scored as percent of speech units correct (phonemes, words, etc). Examples of word tests are the NU6 Monosyllabic Word Test and the CID Auditory Test W-22. Examples of sentence tests are the CID Everyday Sentence Test and the Iowa Laser Videodisc Sentence Test Without Context (Tyler, et. al., 1983 and 1986). The NU6 test involves 50 monosyllabic consonant-vowel-consonant (CVC) words presented in random order; four alternate word lists are available. Responses are scored either as percent words correct out of 50 or as percent phonemes correct out of 150. The W-22 test involves 50 words presented in random order; four alternate word lists are available. W-22 is generally scored as percent words correct but could in principle also be scored as percent phonemes correct. The CID sentence test consists of 10 alternate lists of 10 sentences. Typically, two lists are presented in a test session, responses are scored as percent of keywords correct. The Iowa sentence test consists of five lists of 20 sentences of 4 to 7 words each (excluding articles). Each list has a total of about 87 words (range 83 - 91). Typically a single list is presented in a test session. Sentences are presented in random order and scored as percent words correct.

The importance of the binomial model in the statistical analysis of speech recognition tests has long been recognized. Thornton and Raffin (1978) clarified the basis for its application in audiology, carefully restricting the types of speech recognition tests to which the binomial model can be validly applied to those in which stimuli are presented independently and are scored correct or incorrect, and different stimulus lists were used to obtain the scores to be compared. Boothroyd and Nittrouer (1988) demonstrated experimentally that the binomial model is invalid for tests in which stimuli are scored by parts, for example sentences scored as percent words correct or words scored as percent phonemes correct and proposed a straightforward adjustment to the sample size, replacing the number of scored speech units per stimulus with an effective number of independent units, j. For example, Iowa Sentences have an average j-value of 1.82 for experienced cochlear implant wearers (range 1.2 - 4.7). Consequently for the typical subject, percent correct
scores for a 20-sentence (87-word) test would behave approximately like scores on a test with only 36 independent items (1.82 x 20).

Thornton and Raffin's method is based on treating a stimulus list as a random sample from a larger pool of equivalent stimuli (e.g., common monosyllables or everyday sentences). Statistically significant difference between two test scores must therefore be attributed to effects or sources of variation other than random selection of stimuli. These include a real change in speech recognition ability between the two occasions, distractions during the testing session, factors which affect the subject's motivation or concentration, etc. It has been reported that for stimuli presented in noise to naive subjects (Sever, 1985), the binomial variance component due to random selection of stimuli appears to miss a significant part of the variability between scores. One possible explanation is that the novelty of the task produces a steep learning curve or quickly fatigues the subject.

Thornton and Raffin's procedure for binomially scored stimuli is equivalent to the well-known arc-sin-transformed z test for independent proportions and is a refined version of the Test for Significant Differences between Proportions (Shearer, 1982), which some investigators (e.g. Spivak and Waltzman, 1990) have used to compare a subject's speech recognition scores. For stimuli scored by counting correct parts of the stimulus (e.g. phonemes in words) Shearer's and Thornton and Raffin's binomial tests should not be used2, Fisher's randomization test (Conover, 1983, 328-330) provides exact significance levels (p-values).

Data from a cochlear implant wearer will be used to illustrate the methods. The subject identified as IC02 was fitted with a Clarion multichannel cochlear implant, which can be programmed to stimulate in either analog or pulsatile modes. Table 1, shows the information needed to compute exact or approximate statistical significance levels between any two testing occasions. For example, after three day's experience with the device, the Iowa Sentence test was administered under analog and pulsatile modes using lists D and B, respectively. Scores were 17.44% (analog) and 44.32% (pulsatile), expressed as percent words correct.

2 Shearer appears not to understand that the binomial test requires independent binary responses; his "Computational Example" of the binomial procedure is invalid (Shearer, 1982, page 93). In that hypothetical example, 10 subjects each respond to 12 diphthong presentations and 20 vowel presentations. Shearer treats the 100 diphthongs and 200 vowel presentations as independent groups of responses. This is clearly incorrect, since responses by the same individual are likely to be more consistent than responses by different individuals. The correct way to analyze those observations would have been to reduce each subject's data to two numbers -- the proportion of diphthongs correct and the proportion of vowels correct. These ten pairs of numbers would then be compared using either the t-Test For Related Measures or the Matched Pairs Wilcoxon Test (Shearer, 1982, pages 96 and 98, respectively).
The exact statistical significance of the difference between the two scores is $p = 0.0217$ (two-sided), computed using the exact permutation (PE/EX) test in the StatXact package (Mehta and Patel, 1992). This method of computing $p$-values is the gold standard against which simpler approximations are to be compared.

Boothroyd and Nittrouer (BN) proposed a simple, convenient approximate test based on treating percent words correct as a binomial random variable with $n$ replaced by $n'$, the effective number of independent units per sentence ($j$) times the number of sentences (20). The approximate $p$-value can then be computed using binomial methods (e.g. Fisher's exact test, Thornton and Raffin's critical intervals, or the $z$-test for independent proportions with or without arcsine normalization, ). With this adjustment, the approximate significance of the difference between the day 3 analog and pulsatile scores is $p = 0.0120$, using the $z$-test with arcsine normalization. In this case the BN approximation would produce the same conclusion as the exact test and has the advantage of not requiring special software or detailed knowledge of the number of correct words for each stimulus sentence. The generally more accurate $z$-approximation recommended in this paper yields $p = 0.0224$. The naive binomial test, based on $n$'s of 86 and 88 yields a gross overstatement of the significance, $p < 0.0001$. 
Table 1. Post-Implant Sound-Only Sentence Test History\(^1\) for Subject IC02.

<table>
<thead>
<tr>
<th>Date</th>
<th>8/26/</th>
<th>8/26/</th>
<th>8/27/</th>
<th>8/27/</th>
<th>11/16</th>
<th>2/15/</th>
<th>5/4/9</th>
<th>8/16/</th>
<th>8/18/</th>
<th>1/20/</th>
<th>1/20/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience(^2)</td>
<td>2D</td>
<td>2D</td>
<td>3D</td>
<td>3D</td>
<td>3M</td>
<td>6M</td>
<td>9M</td>
<td>12M</td>
<td>12M</td>
<td>18M</td>
<td>18M</td>
</tr>
<tr>
<td>Mode(^3)</td>
<td>P</td>
<td>A</td>
<td>A</td>
<td>D</td>
<td>B</td>
<td>F</td>
<td>A</td>
<td>B</td>
<td>F</td>
<td>D</td>
<td>A</td>
</tr>
<tr>
<td>List</td>
<td>E</td>
<td>A</td>
<td>D</td>
<td>B</td>
<td>F</td>
<td>A</td>
<td>B</td>
<td>F</td>
<td>D</td>
<td>A</td>
<td>E</td>
</tr>
<tr>
<td>Sentence(^4)</td>
<td>Correct Words(^5) / Total Words(^6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0 / 6</td>
<td>0 / 5</td>
<td>0 / 3</td>
<td>0 / 3</td>
<td>4 / 5</td>
<td>5 / 5</td>
<td>1 / 3</td>
<td>5 / 5</td>
<td>3 / 3</td>
<td>5 / 5</td>
<td>6 / 6</td>
</tr>
<tr>
<td>2</td>
<td>0 / 4</td>
<td>1 / 4</td>
<td>0 / 5</td>
<td>4 / 4</td>
<td>5 / 6</td>
<td>4 / 4</td>
<td>4 / 4</td>
<td>5 / 6</td>
<td>5 / 5</td>
<td>4 / 4</td>
<td>3 / 4</td>
</tr>
<tr>
<td>3</td>
<td>0 / 4</td>
<td>1 / 4</td>
<td>0 / 4</td>
<td>2 / 4</td>
<td>4 / 4</td>
<td>1 / 4</td>
<td>2 / 4</td>
<td>2 / 4</td>
<td>4 / 4</td>
<td>4 / 4</td>
<td>4 / 4</td>
</tr>
<tr>
<td>4</td>
<td>1 / 3</td>
<td>0 / 4</td>
<td>0 / 5</td>
<td>4 / 4</td>
<td>3 / 3</td>
<td>4 / 4</td>
<td>4 / 4</td>
<td>3 / 3</td>
<td>4 / 5</td>
<td>4 / 4</td>
<td>3 / 3</td>
</tr>
<tr>
<td>5</td>
<td>0 / 4</td>
<td>0 / 4</td>
<td>0 / 4</td>
<td>0 / 5</td>
<td>5 / 5</td>
<td>4 / 4</td>
<td>4 / 5</td>
<td>5 / 5</td>
<td>4 / 4</td>
<td>4 / 4</td>
<td>4 / 4</td>
</tr>
<tr>
<td>6</td>
<td>0 / 4</td>
<td>0 / 4</td>
<td>1 / 4</td>
<td>1 / 5</td>
<td>5 / 5</td>
<td>4 / 4</td>
<td>5 / 5</td>
<td>5 / 5</td>
<td>3 / 4</td>
<td>4 / 4</td>
<td>4 / 4</td>
</tr>
<tr>
<td>7</td>
<td>2 / 4</td>
<td>0 / 6</td>
<td>1 / 4</td>
<td>4 / 6</td>
<td>6 / 6</td>
<td>6 / 6</td>
<td>4 / 4</td>
<td>4 / 4</td>
<td>6 / 6</td>
<td>6 / 6</td>
<td>4 / 4</td>
</tr>
<tr>
<td>8</td>
<td>0 / 3</td>
<td>0 / 6</td>
<td>0 / 5</td>
<td>2 / 4</td>
<td>5 / 5</td>
<td>6 / 6</td>
<td>3 / 4</td>
<td>5 / 5</td>
<td>5 / 5</td>
<td>6 / 6</td>
<td>1 / 3</td>
</tr>
<tr>
<td>9</td>
<td>5 / 5</td>
<td>0 / 4</td>
<td>0 / 5</td>
<td>4 / 4</td>
<td>3 / 4</td>
<td>4 / 4</td>
<td>1 / 5</td>
<td>4 / 4</td>
<td>5 / 5</td>
<td>4 / 5</td>
<td>5 / 5</td>
</tr>
<tr>
<td>10</td>
<td>3 / 4</td>
<td>0 / 4</td>
<td>0 / 3</td>
<td>2 / 4</td>
<td>0 / 5</td>
<td>4 / 4</td>
<td>4 / 4</td>
<td>1 / 5</td>
<td>3 / 3</td>
<td>4 / 4</td>
<td>4 / 4</td>
</tr>
<tr>
<td>11</td>
<td>3 / 3</td>
<td>0 / 4</td>
<td>3 / 4</td>
<td>0 / 3</td>
<td>5 / 5</td>
<td>4 / 4</td>
<td>3 / 3</td>
<td>2 / 5</td>
<td>4 / 4</td>
<td>4 / 4</td>
<td>3 / 3</td>
</tr>
<tr>
<td>12</td>
<td>0 / 5</td>
<td>0 / 4</td>
<td>0 / 5</td>
<td>4 / 4</td>
<td>3 / 5</td>
<td>4 / 4</td>
<td>4 / 4</td>
<td>5 / 5</td>
<td>4 / 5</td>
<td>4 / 5</td>
<td>5 / 5</td>
</tr>
<tr>
<td>13</td>
<td>0 / 4</td>
<td>0 / 4</td>
<td>1 / 3</td>
<td>0 / 5</td>
<td>5 / 5</td>
<td>4 / 4</td>
<td>3 / 5</td>
<td>5 / 5</td>
<td>3 / 3</td>
<td>4 / 4</td>
<td>4 / 4</td>
</tr>
<tr>
<td>14</td>
<td>4 / 4</td>
<td>0 / 4</td>
<td>0 / 5</td>
<td>1 / 5</td>
<td>3 / 3</td>
<td>4 / 4</td>
<td>1 / 5</td>
<td>3 / 3</td>
<td>4 / 5</td>
<td>4 / 4</td>
<td>4 / 4</td>
</tr>
<tr>
<td>15</td>
<td>1 / 6</td>
<td>0 / 6</td>
<td>0 / 5</td>
<td>2 / 5</td>
<td>0 / 3</td>
<td>1 / 6</td>
<td>3 / 5</td>
<td>1 / 3</td>
<td>4 / 5</td>
<td>2 / 6</td>
<td>6 / 6</td>
</tr>
<tr>
<td>16</td>
<td>3 / 3</td>
<td>0 / 5</td>
<td>0 / 4</td>
<td>1 / 5</td>
<td>1 / 4</td>
<td>1 / 5</td>
<td>3 / 5</td>
<td>2 / 4</td>
<td>2 / 4</td>
<td>1 / 5</td>
<td>3 / 3</td>
</tr>
<tr>
<td>17</td>
<td>0 / 4</td>
<td>0 / 4</td>
<td>0 / 5</td>
<td>0 / 4</td>
<td>1 / 4</td>
<td>4 / 4</td>
<td>2 / 4</td>
<td>4 / 4</td>
<td>5 / 5</td>
<td>4 / 4</td>
<td>4 / 4</td>
</tr>
<tr>
<td>18</td>
<td>0 / 5</td>
<td>0 / 7</td>
<td>4 / 4</td>
<td>3 / 4</td>
<td>1 / 3</td>
<td>7 / 7</td>
<td>1 / 4</td>
<td>0 / 3</td>
<td>4 / 4</td>
<td>7 / 7</td>
<td>5 / 5</td>
</tr>
<tr>
<td>19</td>
<td>4 / 4</td>
<td>1 / 3</td>
<td>1 / 4</td>
<td>4 / 5</td>
<td>4 / 4</td>
<td>1 / 3</td>
<td>5 / 5</td>
<td>4 / 4</td>
<td>0 / 4</td>
<td>3 / 3</td>
<td>4 / 4</td>
</tr>
<tr>
<td>20</td>
<td>0 / 4</td>
<td>0 / 5</td>
<td>4 / 5</td>
<td>3 / 4</td>
<td>2 / 4</td>
<td>4 / 5</td>
<td>3 / 4</td>
<td>4 / 4</td>
<td>5 / 5</td>
<td>5 / 5</td>
<td>4 / 5</td>
</tr>
</tbody>
</table>

% Words\(^7\) | 31.33 | 3.30 | 17.44 | 44.32 | 70.93 | 86.81 | 67.05 | 80.23 | 84.88 | 90.11 | 96.39 |
% Sentences | 25.00 | 0.00 | 5.00 | 20.00 | 45.00 | 80.00 | 35.00 | 65.00 | 60.00 | 85.00 | 90.00 |

1. The table does not include pre-implant testing and does not include audio-visual or vision-only testing.
2. Time in days (D) or months (M) since connection of the processor.
3. Major processing strategy: analog (A) or pulsatile (P). Other processor parameters may also vary between tests (waveform, sequential or parallel electrode addressing, pulse width, pulse frequency).
4. Each list consists of 20 different sentences. No sentence appears on two lists.
5. Number of words correctly recognized.
6. Number of words in sentence (not counting articles)
7. Percent of words correct, percent of whole sentences correct.

A different approach is needed to compare scores obtained from administration of the same list. In this situation, the exact significance level (p-value) is obtained from McNemar's test or its multinomial extension, the Exact Test for Bivariate Symmetry (Kuritz, Landis and Koch, 1988: Mehta and Patel, 1992). Table 2 shows the data layout for comparing subject IC02’s 3 month score (70.93) with the 12 month score (80.23), both obtained with list F. For example 2 sentences had 0 words correct at 3 months and 1 word correct at 12 months.
The exact (extended) McNemar (MC/EX) test of the StatXact package (Mehta and Patel 1992) was used to compute the exact significance level (p=.1573, two sided) of the difference between the 3- and 12 month scores. Binomial methods, even with the Boothroyd-Nittrouer adjustment, are not valid for comparing open-set, same-list scores since they ignore the correlation between responses induced by the common stimulus list (r=.736). However, under the reasonable assumption that the correlation is positive, analyzing the data as if the lists were different produces a p-value larger than the exact value. Thus McNemar's test might find a significant difference when the invalid (separate-list) method did not, but the reverse could not happen. Consequently the separate-list method is conservative provided responses to the same list are positively correlated. In the present case, the separate list method (i.e. Fisher's randomization test) produced p= .5407.

Table 3: Scores for Sentence List F at Three and Twelve Months Post Connection

<table>
<thead>
<tr>
<th>Correct Words per Sentence</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 Months:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
</tbody>
</table>

An implication of this work is that more detailed data are required to compute exact significance levels for multinomially scored speech tests. For this reason, clinicians and researchers who plan to compare single-subject scores will need to maintain more detailed records of subjects' responses. When such detail is lacking and only the percent correct scores are recorded, approximate methods are available but may produce incorrect inferences. Since McNemar's test requires a level of data detail that is unlikely to be practical, this paper assumes that same-list data will be analyzed by more conservative separate-list methods.

Theory

Comparison of Separate-List Scores

For simplicity, whole stimuli will be called sentences and scored subunits will be called words. There are L sentences per list, and the pair (Ci, Wi) denotes the number of correct words and number of possible words in the ith sentence. For example, for 3 month data (Table 1) the pairs are (4,5), (5,6), (2,4), etc.
The null hypotheses is that (C,W) pairs obtained at the earlier and later test occasions are samples from the same bivariate population. Put another way, the null hypothesis is that the subject responded to the two lists as if they had been administered in a single session. If the null hypothesis is rejected and the second score is significantly larger than the first, the implication is that the second set of responses was sampled from a population with a greater proportion of correct words. One interpretation of the significant change is that the subject's speech recognition ability improved, but other explanations need to be considered such as learning the test materials, fatigue or inattention, temporary health problems, transient problems with the implant or speech processor, etc.

Fisher's randomization test compares univariate means and a data manipulation is needed to convert bivariate (C,W) pairs to univariate date. Let Pw be the proportion of correct words in the combined sample (all 40 sentences from the two test administrations to be compared). For each sentence compute Di=Ci-PwxWi; i.e., the observed minus the expected number of correct words in the ith sentence. For the combined three and six month data, Pw=140/177=.7910, so for the first sentence in the 3-month data D equals 4 - 5x.7910 = .0452. The exact test determines if the mean of the D-values changed significantly between the two test administrations (see Conover, 1980, section 5.11 for a general explanation and additional references, Mehta, 1982 for a description of the computer algorithm, and Mehta and Patel, 1992 for its microcomputer implementation).

Preservation of words correct for each sentence, as illustrated in Table 1, provides sufficient data for computing at a later date any exact test or any approximate test for same-list or separate-list comparisons. While this does not seem unduly burdensome given the wide availability of personal computers, there are several reasons for pursuing approximate methods. First, software for exact computation is not yet widely available, second, data beyond the total percent correct may not have been retained.

**Analysis of Separate-List Scores**

A1 thru A20 and B1 through B20 denote the number of correct words for administrations of lists A and B to the same subject. Taking 6 and 9 months scores from Table 2 as an example, the A's are 5 4 4 4 4 5 5 4 4 4 4 4 1 1 4 7 1 4 and the B's are 1 4 1 4 4 5 4 3 1 4 3 4 3 1 3 3 2 1 5 3. The combined sample of both lists put together is denoted Z1 through Z40. In this case the Z's (from smallest to largest) are 1 1 1 1 1 1 1 1 2 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5 5 6 6 7. The null hypothesis implies that A1 thru A20 and B1 thru B20 are statistically equivalent to random samples from the combined sample, Z1 through Z40. Approximations to p-values are derived from this concept using finite sampling theory.

The order of presentation of sentences is randomized, consequently aggregated data is sufficient for all statistical computations. Figure 1 is a symbolic representation of the aggregated data such as table 1 and
contains sufficient information for comparing separate-list scores (but not same-list scores, since correlations cannot be deduced).

<table>
<thead>
<tr>
<th>n</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>List A</td>
<td>n1</td>
<td>f10</td>
<td>f11</td>
<td>f12</td>
<td>f13</td>
<td>f14</td>
<td>f15</td>
<td>f16</td>
</tr>
<tr>
<td>List B</td>
<td>n2</td>
<td>f20</td>
<td>f21</td>
<td>f22</td>
<td>f23</td>
<td>f24</td>
<td>f25</td>
<td>f26</td>
</tr>
<tr>
<td>Combined</td>
<td>N</td>
<td>f0</td>
<td>f1</td>
<td>f2</td>
<td>f3</td>
<td>f4</td>
<td>f5</td>
<td>F6</td>
</tr>
</tbody>
</table>

The exact test is computed from the two by six distribution table, f10 through f26, (see Conover, 1980, section 5.11 for a general explanation and additional references, Mehta, 1982 for a description of the algorithm, and Mehta and Patel, 1992 for implementation).

Approximate tests are derived from finite sampling theory. All involve $\bar{Z}$ and $S\bar{Z}$, the mean and standard deviation of the combined sample.

$$\bar{Z} = \frac{\sum_{i=1}^{N} Z_i}{N}$$

and

$$S_{\bar{Z}} = \sqrt{\frac{\sum_{i=1}^{N} (Z_i - \bar{Z})^2}{N}}$$

Or using aggregated data, where C is the length of the longest sentence,

$$Z = \sum_{c=1}^{C} f_c \times c / N$$

and

$$S_{\bar{Z}} = \sqrt{\left(\sum_{c=0}^{C} f_c \times c^2 / N\right) - \bar{Z}^2} .$$

If sentence level data are not available, then

$$\bar{Z} = (n_1 \times P_{w1} + n_2 \times P_{w2}) / (2L),$$

where L is the number of sentences per list, and the standard deviation can be estimated by Boothroyd's method,

$$S_{\bar{Z}} = \frac{1}{L} \sqrt{\frac{j_1^2 \times V_1}{2j_1} + \frac{j_2^2 \times V_2}{2j_2} + D^2} ,$$

For each list, upper case J is the average number of words per sentence (J=\(n/L\)), lower case j is Boothroyd-Nittrouer effective number of units per average sentence, V is the binomial variance (V=\(P_w \times (1-P_w)\)), and D is the deviance,

$$D^2 = (n_1 \times P_{w1} - n_2 \times P_{w2})^2 / 2$$
Let $\overline{A}$ and $\overline{B}$ denote the mean of $A_1$ thru $A_{20}$ and $B_1$ thru $B_{20}$, respectively; i.e., the average numbers of correct words per sentence. Finite sampling theory (see e.g. ref) states that under the null hypothesis, $\overline{A}$ has mean $\overline{Z}$ and standard deviation $\sqrt{\frac{2L - 1}{n}}$. Since $\overline{B} = 2\overline{Z} - \overline{A}$ and $\overline{A} = n_1 \times P_{w_1} / L$ and $\overline{B} = n_1 \times P_{w_2} / L$ it follows that $P_{w_1} - P_{w_2} = \frac{L \times N}{n_1 \times n_2} \left( \overline{A} - \overline{Z} \right) + \overline{Z} \times \frac{n_1 - n_2}{n_1 + n_2}$. For constant length stimuli (e.g. CVC monosyllables) the second term is exactly zero; however, for variable length stimuli it is nonzero and, since $n_1$ and $n_2$ are random, will introduce a small additional source of variation. For the purpose of deriving an approximation that component of variation will be ignored.

### Inappropriate use of the binomial model leads to incorrect inferences

The previous examples have shown that the binomial model substantially understates the true standard deviation for sentence tests. This is not simply a minor technicality, but can actually lead to incorrect conclusions. To illustrate the seriousness of the problem, Iowa sentence test scores for 23 subjects in the Iowa Cochlear Implant Project at 1, 9 and 18 months post-connection were statistically contrasted using both the exact standard deviation (Formula 4) and the invalid binomial standard deviation (Formula 5). Three comparisons were made for each subject: 1 vs. 9 months, 9 vs. 18 months and 1 vs. 18 months. Differences were converted into z-scores and significance levels (p-values) were computed using exact and invalid standard deviations. (Exact standard deviations could not be computed for 12 of the 69 comparisons due to lack of data on responses to individual sentences). Table 3 summarizes the results, with p-values grouped into four commonly used categories: .001 or less, .001 to .01, .01 to .05, and .05 or greater. Because of the large number of tests, most investigators would select type I error (alpha) levels of .001 or smaller. An investigator adopting the .001 criterion and employing the invalid binomial model would have reported 51 significant gains out of 57. However, four of those inferences would have been incorrect (binomial p less than .001, exact p greater than .001).

It is important to understand that these incorrect inferences were not the random false positives which an investigator controls by selecting a small Type I error level (alpha). They are, instead, easily avoidable computational errors produced by the use of an incorrect model. To put these errors in perspective, recall that the type I error rate (alpha) was set at .001. Since three correlated tests were performed per subject the actual type I error rate is .003 or less (by Bonferroni's inequality). In other words, there should be, on average, three or fewer spuriously significant differences per 1000 subjects, or about 1 in 300. However, according to Table 3, inappropriate application of the binomial formula produced four erroneous declarations of significance at the .001 level, distributed, as it happens, over four of the of the 23 subjects. Thus spurious
significance occurred for about one subject in six, not the expected one in three hundred, an error rate 50 times higher than expected. Needless to say, use of the correct standard deviation, Formula (4), would reduce the rate of spurious significance to the expected 1 in 300.

The importance of using correct formulas is now clear. Unfortunately, they require data which may not have been routinely recorded. Once the need to record more detailed data is recognized, it is comparatively easy to implement: if the same stimulus list is to be used on all occasions, then scores for individual stimuli must be preserved, as illustrated in Table 1; if different stimulus lists are to be used on different testing occasions, then it is not necessary to save individual responses, but Formula (2) must be used to compute the standard deviation before responses to individual stimuli are discarded. Implementation of these procedures require a one-time programming effort if responses are already being collected by computer. However, there is no way to retroactively apply formula (2) or (4) to previous test sessions for which responses to individual stimuli have already been discarded. The next section supplies a conservative procedure for analyzing such data.

**ANALYZING INCOMPLETE DATA**

Computation of Formulas (2), (3) and (4) requires that subjects' responses to individual stimuli be recorded, as was done in Figure 1. Unfortunately the need to retain such microdata has not been recognized, with the result that responses to individual stimuli may not have been permanently recorded. Without such data it is not possible to compute exact standard deviations; however, even without data on responses to individual stimuli, it is possible to compute conservative estimates of standard deviations, which can validly be used for significance testing and confidence intervals. But these standard deviation estimates can be quite conservative, resulting in inefficient use of data, particularly when comparing scores from two test sessions. Therefore, an objective of this paper is to alert audiological investigators to the need to retain complete response data, particularly if single-subject, longitudinal comparisons are to be made.

Although there is no way to recover lost information, the data which are routinely recorded (percent of correct phonemes or words) do permit the computation of conservative estimates of the standard deviation of single-subject percent-correct scores and of single-subject change scores. For example, the knowledge that subject IN8 achieved 64.9% correct on the Iowa sentence test at one month is sufficient to guarantee that the standard deviation of his one month score could not exceed 5.1 percentage points, regardless of his the pattern of responses to individual sentences. Since the exact standard deviation was 4.15 (see Figure 4), the
penalty for not recording IN8's responses to individual sentences would be a 23\% increase in the estimated standard deviation.

Conservative standard error estimates can safely be used in statistical inference. For example, a conservative 95\% confidence interval for subject IN8's true score at one month is 64.9 ± 2 × 5.1. It is conservative in the sense that the correct confidence interval, 64.9 ± 2 × 4.15, is about 23\% narrower. The use of conservative estimates of the standard deviation produces confidence intervals which are wider than they need to be. Similarly, when a conservative standard deviation is used to compute a test statistic (a z-score, for example) the p-value is larger than it would be if the exact standard deviation were used. Thus a significant result with a test constructed from a conservative standard deviation would have been even more significant if the exact standard deviation had been used. On the other hand, a borderline insignificant result could have been significant if the exact standard deviation had been used. In other words, the specificity of the statistical test is preserved but its sensitivity is reduced.

A conservative estimate of the standard deviation of the difference of two percent-correct scores can also be computed. For example, subject IN8's one- and nine-month scores were 64.9\% and 84.5\%. This information by itself is sufficient to guarantee that the standard deviation of the difference of these two scores is no greater than 7.6. On the other hand, the exact standard deviation was 4.15 (see Figure 5), so in this case the data analyst would have to pay a rather stiff penalty for loss of data regarding IN8's responses to individual sentences: the conservative estimate would be about 80\% larger than the exact estimate. The practical consequence is that confidence intervals would be 80\% wider than necessary and that the subject's one- to nine- month gain would appear less significant than it actually is.

The exact standard deviation is substantially lower in this case because IN8's one- and nine-month responses are highly consistent: for fifty-three out of one hundred sentences, the number of correct words was identical on both occasions. This consistency lowers the standard deviation of the difference (indeed, if responses had been identical on both occasions, then the standard deviation of the difference would have been zero). As it is, the correlation between the percent of words which IN8 recognized in each of the 100 sentences at one and nine months is .43, which substantially reduces the standard deviation of the difference. Since the conservative standard deviation estimate lacks any information about the correlation between responses on successive occasions it must be based on the worst case assumption of a strong negative correlation between scores on the two occasions.
It may seem puzzling that the author and other statisticians criticize the use of underestimated standard deviations (as provided by the binomial model) but regard the use of overestimated (conservative) standard deviations as valid. The explanation is simple: standard deviation indicates the precision of a speech discrimination score (or of the difference between two scores), the smaller the standard deviation, the smaller the probable error of the score. Consequently, a conservative estimate, a number which the standard deviation cannot exceed, sets a limit to the probable error of a score. On the other hand a lower bound for the standard deviation places no limit on the size of the error.

Tables of conservative standard deviations are presented below. These were computed by finding the theoretical maximum of Formula (2) or (4) using the mathematical technique linear programming. Maxima were computed assuming that percent correct scores were known, but responses to individual stimuli were not. Readers wishing details of the algorithm used to compute these tables may contact the author.

Table 4 is intended for use with NU6 phoneme scores. This test is administered as 50 CVC monosyllables presented in random order and scored as the percent of phonemes correct out of 150. The table can be used for single scores as well as for differences of two scores. Conservative standard deviations for single test scores are found in the left margin. For example, referring to Table 4, a NU6 phoneme score of 80% correct has a standard deviation no greater than 5.71 percentage points. For scores under 50%, deduct the score from 100 before entering the table. For example, a score of 36% has the same standard deviation as a score of 64%, that is 6.86 percentage points. For odd or fractional scores, use interpolation. For example to obtain the standard deviation of a score of 87.4, interpolate between rows 86 (SD = 4.96) and 88 (SD=4.64); i.e., the interpolated standard deviation is .3×4.96+7×4.64 = 4.74.

Conservative standard deviations for the difference of two scores obtained with the same stimulus list are recorded in the body of the table. To find the standard deviation of the difference, locate the larger of the two scores in the left margin and read the standard deviation from the column corresponding to the difference between the two scores. For example a subject had NU6 phoneme scores of 72% at nine months and 86% at eighteen months, a difference of 14 percentage points. The conservative standard deviation of that difference is 9.02, found at the intersection of row 86 and column 14. If the score combination is not found in the table, work with percent incorrect rather than percent correct. I.e., subtract each score from 100 before entering the table (this will not change the difference score). For example, a subject scored 24% at one month and 48% at nine months. The scores are not found in the table and must therefore be converted to percent incorrect scores: 76% and 52%. The larger score is 76% and the difference is 24, so the conservative
standard deviation, 11.63, is found at the intersection of row 76 and column 24. For odd or fractional scores, two-way interpolation can be used.

Table 4 should be used only for NU6 phoneme scores on 50-word lists and only if the subject's responses to individual words are unavailable. When these data are available Formula (2) should be used to compute the standard deviation of a single NU6 phoneme score, and Formula (4) should be used to compute the standard deviation of the difference of two such scores if the same word list is used for both administrations.

If different word lists were presented, then Formula (3) should be used, if possible using exact standard deviations computed from Formula (2), otherwise using conservative standard deviations read from the left margin of Table 4 (if Formula (2) cannot be used due to incomplete data). For example, a subject scored 48% phonemes correct at month one on NU6 list I and 74% phonemes correct at month nine on NU6 list II. Conservative standard deviations from the right margin of Table 4 are SD1=7.14 and SD2=6.27. Inserting these numbers in Formula (3) yields a conservative standard deviation for the difference,

\[ SD_{\text{difference}} = \sqrt{7.14^2 + 6.27^2} = 9.50 \]

Recall that if both test sessions had used the same word list, then the conservative standard deviation, 12.06, would be obtained from the body of table 4 at the intersection of row 74 and column 26 (the difference of the two scores).

Table 5 is provided for use with the Iowa sentence test. It can be used as printed for scores obtained from administration of the full, 100-sentence list. With a simple adjustment this table can also be used for shorter tests built from randomly selected Iowa sentences. The table is based explicitly upon the distribution of numbers of words in Iowa sentences and should not be used for other sentence tests. Readers wishing tables for other tests should contact the author.

For tests containing fewer than the full set of 100 Iowa sentences, the standard deviations in Table 4 must be multiplied the square root of 100 and divided by the square root of the number of sentences administered. For a 20-sentence test score of 62% the conservative standard deviation is \( 10 \times \frac{5.13}{\sqrt{20}} = 11.5 \).

Needless to say, if responses to individual sentences are available, then Formula (2) should be used in preference to Table 5.
For differences of two scores, Tables 5 assumes that the same stimuli (words or sentences) were administered to the subject on both testing occasions. If a different subset of sentences was administered on the second occasion, the standard deviation is obtained using Formula (3) as explained above. For example, a subject achieved 62% on a 20-sentence subtest of the Iowa sentence test administered at 1 month and 84% at 9 months on a different 20-sentence subtest. The conservative standard deviations (adjusted for test length) are $10 \times 5.13 / \sqrt{20} = 11.5$ at one month and $10 \times 4.14 / \sqrt{20} = 9.3$ at nine months. According to Formula (3) the standard deviation of the difference of independent scores is the square root of the sum of their squared standard deviations; that is, $SD = \sqrt{[11.5^2 + 9.3^2]} = 15.3$. In this case, the difference between one and nine month scores, 22 percentage points, is less than two standard deviations and would not be statistically significant ($z=22/15.3=1.4$, $p=.16$). As before, if responses to individual sentences are available, then Formula (2) should be used in preference to the right margin of Table 5.

**CONCLUSIONS**

The binomial distribution is not valid for all speech recognition tests scored as percent correct.

The standard deviation of a single-subject score, or of a single-subject difference score is the basis for statistical inference about individual performance. The binomial model yields valid single-subject, single-occasion standard deviations only when randomization units coincide with scoring units. It yields valid standard deviations for differences between scores on two test sessions only if different stimulus lists were used.

When either of these conditions fails, the computation of valid standard deviations requires information about the subject's responses to individual stimuli, data which are often not routinely preserved.

Incomplete data in the form of whole-test percent-correct scores do provide sufficient information to place upper limits on the standard deviation of a single score or the difference of two scores. These overestimated standard deviations can be used as the basis for conservative statistical inferences (confidence intervals which are wider than necessary and significance tests which are less sensitive than they could be). However, particularly in the case of difference scores, exact standard deviations can be substantially smaller than their conservative approximations, resulting in sharper statistical inferences.
More efficient use of data leads to better clinical decisions, to sharper research findings and to reduced response burdens for patients. Thus although preserving responses to individual stimuli is an added burden for the data manager it is important to do so to make efficient use of subjects' time. Preserving the required information is particularly simple to implement when stimuli are presented under computer control and responses are either machine-scored or scores are keyed-in by the audiologist after each response. In this case software can be added to compute the standard deviation at the end of the test session in addition to the percent correct score. Additionally, if the same stimulus list is to be reused in later test sessions, stimuli and responses can be written to an archival file and later linked (as in Figure 1) with data from other sessions using the same stimulus list.

REFERENCES


Mathematical Appendix

Formulas

\[ Z = \sum_{i=1}^{N} Z_i / N \]

and

\[ S_z = \sqrt{\sum_{i=1}^{N} (Z_i - \bar{Z})^2 / N} \]

Or using aggregated data, where \( C \) is the length of the longest sentence,

\[ \bar{Z} = \sum_{c=1}^{C} f_c \times c / N \]

and

\[ S_z = \sqrt{\left( \sum_{c=0}^{C} f_c \times c^2 / N \right) - \bar{Z}^2} \]

If sentence level data are not available, then

\[ \bar{Z} = (n_1 \times Pw_1 + n_2 \times Pw_2) / (2L), \]

where \( L \) is the number of sentences per list,

and the standard deviation can be estimated by Boothroyd's method,

\[ S_z = \frac{1}{L} \sqrt{\frac{J_1^2 \times V_1}{2J_1} + \frac{J_2^2 \times V_2}{2J_2} + D^2}, \]

For each list, upper case \( J \) is the average number of words per sentence (\( J=n/L \)), lower case \( j \) is Boothroyd-Nittouer effective number of units per average sentence, \( V \) is the binomial variance (\( V=Pw \times (1-Pw) \)), and \( D \) is the deviance,

\[ D^2 = (n_1 \times Pw_1 - n_2 \times Pw_2)^2 / 2 \]