Research Note

Bonferroni's Bound — A Control of Significance Level Errors in Speech Pathology and Audiology Research

C V Kass, Ph D (Witwatersrand)
Department of Statistics
University of the Witwatersrand, Johannesburg
M Marks Wahlhaus, M A (Log.) (Witwatersrand)
Department of Speech Pathology and Audiology
University of the Witwatersrand, Johannesburg

ABSTRACT

Many studies in behavioural sciences, such as speech pathology and audiology, involve statistical hypothesis testing. Repeated tests are made, for example, of judge reliability in assessing the disorder, or within subject variability, or between subject comparisons over several measures of the disorder or types of treatment. If the error rate of the statistical test is only controlled for each individual test, the overall error rate is magnified and the chance of reporting a significant result where none exists, arises. This paper addresses this potential problem, by noting some common procedures that inherently guard against this pitfall, and suggesting a simple, albeit conservative, solution for other cases.

OPSOMMING

Talle studies in die gedragwetenskappe, soos spraakheelkunde en oudiologie, betrek statistiese hipotesetoetsing. Herhaaldelike toetse word uitgevoer, byvoorbeeld, van die betroubaarheid van beoordelaars by die evaluering van 'n afwyking, of intervergelykings van proefpersone ten opsigte van metings van die afwyking of van die tipe behandeling wat toegepas is. Indien die foutvoorkoms van die statistiese toets slegs vir elke individuele toets gekontroleer word, word die totale foutvoorkoms vergroot en ontstaan die moontlikheid dat 'n betekenisvolle resultaat opgeteken word waar daar in werklikheid geen resultaat bestaan nie. Hierdie artikel spreek hierdie potensiële probleem aan deur sommige statistiese prosedures wat inherent teen hierdie valstrik waak, te vermeld en deur 'n eenvoudige, hoewel konserwatiewe oplossing vir ander gevalle aan die hand te doen.

OVERVIEW

The usual research in speech pathology, social, psychological and medical sciences typically advocates a significance level of 5% for reporting results or theories as being established. That is, for example, if a new therapy is to be deemed better than an established regime, then the statistical analysis of an observed "improvement" (eg. a **decrease** in the frequency of stuttering) must show that such a result could not be ascribed to natural variation in the subject's stuttering frequency except with a 5% chance. In other words, if the stutterer was tested repeatedly (over many weeks) **without** the new therapy, only one time in twenty (equivalent to 5% of the time) would he show such a marked decrease in stuttering frequency as evidenced on the single test performed after the new therapy.

From the viewpoint of an individual researcher, a significant improvement of the 5% level is satisfactory as it protects her from advocating a new therapy that is no better than the existing one. An unhappy alternate interpretation is that if a hundred researchers all over the world decided to test this new therapy, then five of them could be expected to detect significant improvement **even if the new therapy is ineffectual.** This is a consequence of statistical testing in modern times. Similarly it is understood that for every hundred journal articles that use statistics and claim an improvement of some technique or difference between two approaches that are significant at the 5% level, it is expected that five such articles will be erroneous — although statistically there is no way of knowing which five they are, nor even if there are precisely five in error. For this reason researchers try to claim significant results at more extreme levels. That is, instead of using the 5% level (implying a one in twenty chance of claiming a false positive), the 1% (one in a hundred), 0.1% (one in a thousand) or more extreme level is used to indicate how small the chance is of an erroneous conclusion by the researcher. There is a problem in being too stringent, namely, if very small significance levels are used, false negatives increase; that is, the smaller the significance level the larger the probability of finding no improvement in a new therapy when in fact it really is efficacious.

The above considerations are usually well known to the researcher. Less understood are the implications when an **individual researcher applies many statistical tests** (in contrast to many researchers applying a single test as dis-

© SASHA 1988

cussed above). Some variations of this theme will now be discussed.

ONE RESEARCHER APPLYING MANY TESTS

The previous example of testing a new therapy to decrease stuttering can be extended as follows: Should the researcher consider the therapy to be conducive to reducing some stuttering behaviours but not others, she could subdivide the types of stutter into a number of auditory categories: gasp, glottal stop, laryngealization ... etc. If twenty categories were for instance decided upon the naive approach would be to apply twenty statistical tests, each at the 5% significance level. A simple extension of the original experiment in which the stuttered words are themselves categorised by the five initial phonemes [f], [s], [t], [m] and [h] and the twenty categories of stutter analysed within each of the five word types would result in one hundred statistical tests. In the latter case the analogy between a hundred researchers each employing one test, and the individual researcher applying one hundred tests, is complete. Even if the new therapy is valueless, five of the behaviour/word-type combinations can be expected to show a significant decrease in frequency (i.e. clinical improvement); and further, one of these will even be significant at the 1% level. While publication of false positives due to many researchers working in isolation is accepted (that is, it is not expected of a researcher in South Africa to anticipate other researchers in the country or throughout the world when performing her statistical tests, any more than they would take into account her research when performing theirs), it is believed that each researcher should include in her reckoning the other statistical tests that she herself performs, at least within a single research topic. Failing "protection" in this way spurious significant results would likely be found.

There are a number of ways of keeping the overall significance level (i.e. probability of Type I error over several statistical tests) down to a pre-specified level, depending on the situation. One of the simplest and most versatile is to use Boole's Inequality (also known as the first Bonferroni Inequality), (Feller 1968). In essence in its simplest form it states that if the number of tests to be performed is n, and the overall significance level to be contained is p, then each individual test should be performed at the p/n level.

Thus, in the first example above where the researcher was about to perform twenty (n = 20) tests each at the 5% (p = 5%) level, each test should have been performed at the 0.25% (p/n = 5%/20 = 0.25% = 1/4%) level, only then could any significant result be claimed to be truly meaningful at the 5% level. Equivalently she would ensure that her overall error rate was at most 5% (i.e. overall probability of a Type I error is at most 5%), by performing each individual test at the 1/4% level.

Similarly, if she wished to validly test all one hundred behaviour/word-type combinations at the overall 5% level of significance, each individual test could only be claimed to be significant if it attened the 0.05% (5%/100 = 0.05%) = (1 in 2000) level.

Such tests (at the 1/4% or 0.05% as appropriate) may be called modified 5% level tests that safeguard against the inflation of the probability of a Type I error. Indeed, without this modification it is almost sure (99.4%) that the Type I error (i.e. claim a false positive) will be committed when performing a hundred tests each at the 5% significance level. One possible difficulty in using Boole's Inequality as described, is that the necessary statistical tables may not be easily accessible. Thus, in the above example where a 0.25% significance level was postulated as providing the required protection, the critical values of the chosen test statistic at the 0.25% level may not be published in commonly used tables or appendices. A statistician may be able to provide a reference to superior tables, or a (possible complicated) procedure either to interpolate in tables or to access a computer approximation. Some tables have been generated specifically for use with Boole's Inequality, eg. Bailey (1977) gives tables so designed for use with the various forms of the t-test. Another approach is to **lower** the necessary significance level to a value for which the required critical values are tabulated. This will induce a similar proportional reduction in the overall significance level. Thus, for example, if the critical values of the 0.25% level are not available but those for the 0.1% level are (i.e. reduced by a factor of 2.5 from 0.25% to 0.1%), then the use of the latter tables will cause a concomitant reduction in the overall significance level from 5% to 2% (since 5%/2.5 = 2%), a more stringent level.

AN EXAMPLE INVOLVING COMPARISONS OF VARIABLES

Suppose the researcher wishes to examine the association between ten visual behaviours (i.e. behaviours that can be observed by eye, eg. a jaw jerk, eye flutter, furrowed brow, ... etc) that are manifested during or just prior to a stuttered word.

In this case it may be deemed appropriate to examine a matrix of pairwise comparisons, eg. a matrix of correlations or other measures of association or even 'distances' between pairs (as is used in cluster analysis).

In the case of product-moment correlations a statistical package like SAS (SAS Institute Inc. (1985)) offers an overall test of whether or not all the pairwise comparisons can be considered insignificantly different from zero. Unhappily, rejection of the hypothesis that all the comparisons do not differ significantly from zero, does not indicate which of the comparisons show a significant difference, and so the test, while valuable for certain problems, is incomplete as far as the hypothetical researcher into stuttering behaviours is concerned.

The global test can thus only indicate whether further analysis of the correlation severally may be profitable. If the global test is rejected at the 5% significance level, then individual tests may be performed. As before, use of Boole's Inequality is recommended.

In the example of ten visual behaviours (or ten judges) there are 45 pairwise comparisons. Thus each of the 45 tests should be executed at the 0.1111% level (5%/45 = 0.1111%), or perhaps more conveniently at the slightly lower 0.1% level.

Note that the procedure of considering these 45 tests each at the 0.1% level is a valid method for containing the overall significance level at most 5% for **any** such matrix of paired comparisons (eg. rank correlations) and not just product-moment correlations. In the field of speech pathology and audiology judge agreement is often a relevant research issue. The assessment inter-judge reliability poses problems with a similar structure to the example which has been examined, i.e. the association at stuttering behaviour.

AN EXAMPLE INVOLVING ANALYSIS OF VARIANCE

As a final example consider a researcher who measures stuttering frequency on a number of subjects before and after five different therapies (eg. a control group 'time heals' therapy, a fluency-based approach, a stuttering-modification approach, psycho-therapy treatment using psychopharmocological drugs). The null hypothesis that all these methods are equally effective (or equally ineffective) based on appropriate measures, is a standard Analysis of Variance problem. Built into this technique are tests of all the alternative sub-hypotheses including those of one or two therapies being different to one, two, three or even all the other therapies. Like the previous case of correlations examined above, rejection of the null hypothesis does not indicate which of the many alternative sub-hypotheses may be significant. Again if the sub-hypotheses of interest (eg. the control group is worse off than any of the others, the drug therapy is better than the others, the last two are better than the first three, ... etc.) can be listed and numbered (let there be n of them) then Boole's Inequality can be invoked as before to give an overall significance level of p, by conducting each individual appropriate t-test and the p/n level. This approach is, however, only recommended if the number of sub-hypotheses of interest is fairly small (eg. up to n = 4 say).

The reason for eschewing Boole's Inequality for larger n in this case is that more powerful tests have been developed although some require specialised statistical tables. These tests come in various forms and are under such headings as 'multiple range tests', 'multiple comparisons' and 'simultaneous testing procedures'. A discussion of the more popular tests may be found in Winer (1971), while a comprehensive review including more modern procedures can be found in Miller (1981). Standard Statistical packages such as SAS (SAS Institute Inc. (1985)) offer methods such as Duncan's multiple range test, Gabriel's multiple comparison procedure, Tukey's studentized range test, among others, on request.

CONCLUSION

It is clear from this study that each researcher should ensure that her overall significance level is controlled within an accepted bound, when performing multiple tests whether implicitly or explicitly. Some statistical procedures and their associated computer programs contain such built-in protection, eg. Analysis of Variance. Other techniques unfortunately do not normally provide such a safeguard, nor do the associated computer programs supply a caveat; the one- and two-sample t-test and their non-parametric counterparts such as the Wilcoxon test fall into this class. In these cases the use of Boole's Inequality as described in this paper is recommended.

Each researcher should be aware that performing many tests each at an accepted significance level, could lead to an unacceptable increase in the Type I error; that is "discoveries" may be made which are, indeed, due merely to chance fluctuations. It is unfortunate that within the purview of behavioural research, the simple protection against this type of error outlined above, is not more widely used.

REFERENCES

Bailey, B.J.R. Tables of the Bonferroni t Statistics *Journal of the American Statistical Association*. 469—478, 1977.

Feller, W. *An Introduction to Probability Theory and its Applications*. Wiley, New York, 1968.

Miller, R.G. *Simultaneous Statistical Inference*. 2nd ed. Springer, New York, 1981.

SAS Institute Inc. *SAS User's Guide: Statistics, Version 5 Edition*. Cary, NC, 1985.

Winer, *Statistical Principles in Experimental Design*. 2nd ed. McGraw Hill, New York, 1971.

DO IT YOURSELF .....

INSERTION GAIN INSTRUMENT

If you have a suitable IBM compatible computer and if you are familiar with it, Acoustimed can now offer you a "do it yourself" insertion gain instrument. We sell you a kit which you assemble yourself without any special tools.

The equipment has all the features of the HA-2000 II system but is in a less expensive housing and we save on installation costs. Any support which you may need is given over the telephone or in our offices — saving you thousands of Rands on the world's most versatile hearing aid analyzer.

FEATURES:

Complex test signals
Fast pure tone sweep
Speech weighted signals
Transients, bursts, continuous signals
Built in signal synthesizer
Real time analysis
Time delay spectrometry
"Prescription" calculations are programmable
Auto-correlation for noise reduction
Signal averaging and spectrum averaging
RMS, peak and crest factor displayed
Linear response probe microphone
Data management with sophisticated data base program
Easy to use Acodat programming language
Word processor with graphics facility
Mailing list programs
Invoicing programs
Calendar/scheduling program

No other system offers all these features.

Write or call for a descriptive booklet.

NB.
This is a marketing experiment for which we have prepared two instruments. We reserve the right to request that you bring your computer to us for assembly, demonstration and instruction.

ACOUSTIMED (PTY) LTD.
327 Bosman Building
Cor. Eloff and Bree Streets
Johannesburg Tel: (011) 337-2977

INFORMATION FOR CONTRIBUTORS

The South African Journal of Communication Disorders publishes reports and papers concerned with research, or critically evaluative theoretical, or therapeutic issues dealing with disorders of speech, voice, hearing or language, or on aspects of the processes underlying these.

The South African Journal of Communication Disorders will not accept material which has been published elsewhere or that is currently under review by other publications.

All contributions are reviewed by at least two consultants who are not provided with author identification.

Form of Manuscript. Authors should submit four neatly typewritten manuscripts in triple spacing with wide margins which should not exceed much more than 25 pages. Each page should be numbered. The *first page* of *two* copies should contain the title of the article, name of author/s, highest degree and address or institutional affiliation. The first page of the remaining two copies should contain only the title of the article. The *second* page of all copies should contain only an *abstract* (100 words) which should be provided in both English and Afrikaans. Afrikaans abstracts will be provided for overseas contributors. All paragraphs should start at the left margin and not be indented.

Major headings, where applicable, should be in the order of METHOD, RESULTS, DISCUSSION, CONCLUSION, ACKNOWLEDGEMENTS and REFERENCES.

Tables and Figures should be prepared on separate sheets (one per table/figure). Figures, graphs and line drawings must be originals, in black ink on good quality white paper. Lettering appearing on these should be uniform and professionally done, bearing in mind that such lettering should be legible after a 50% reduction in printing. On no account should lettering be typewritten on the illustration. Any explanation or legend should not be included in the illustration but should appear below it. The titles of tables and figures should be concise but explanatory. The title of tables appears above, and of figures below. Tables and figures should be numbered in order of appearance (with Arabic numerals). The amount of tabular and illustrative material allowed will be at the discretion of the Editor (usually not more than 6).

References. References should be cited in the text by surname of the author and date, e.g. Van Riper (1971). Where there are more than two authors, *et al.* after the first author will suffice. The names of all authors should appear in the Reference List. References should be listed alphabetically in triple-spacing at the end of the article. For acceptable abbreviations of names of journals, consult the fourth issue (October) of *DSH ABSTRACTS* or *The WorldList of Scientific Periodicals.* The number of references used should not exceed much more than 20.

Note the following examples:

Locke, J.L. Clinical Psychology: The Explanation and Treatment of Speech Sound Disorders. *J. Speech Hear. Disord.*, 48, 339-341, 1983.

Penrod, J.P. Speech Discrimination Testing. In J. Katz (Ed.) *Handbook of Clinical Audiology*, 3rd ed., Baltimore: Williams & Wilkins, 1985.

Van Riper, C. *The Nature of Stuttering.* Englewood Cliffs, New Jersey: Prentice-Hall, 1971.

Proofs. Galley proofs will be sent to the author wherever possible. Corrections other than typographical errors will be charged to the author.

Reprints. 10 reprints without covers will be provided free of charge.

All manuscripts and correspondence should be addressed to:

The Editor,
South African Journal of Communication Disorders,
South African Speech and Hearing Association,
P.O. Box 31782, Braamfontein 2017, South Africa.

INLIGTING VIR BYDRAERS

Die Suid-Afrikaanse Tydskrif vir Kommunikasieafwykings publiseer verslae en artikels oor navorsing, of krities evaluerende artikels oor die teoretiese of terapeutiese aspekte van spraak-, stem-, gehoor- of taalafwykings, of oor aspekte van die prosesse onderliggend aan hierdie afwykings.

Die Suid-Afrikaanse Tydskrif vir Kommunikasieafwykings sal nie materiaal aanvaar wat reeds elders gepubliseer is, of wat tans deur ander publikasies oorweeg word nie.

Alle bydraes word deur minstens twee konsultante nagegaan wat nie ingelig is oor die identiteit van die skrywer nie.

Formaat van die Manuskrip. Skrywers moet vier netjies getikte manuskripte in 3-spasiëring en met breë kantlyn indien, en dit moet nie veel langer as 25 bladsye wees nie. Elke bladsy moet genommer wees.

Op die *eerste bladsy* van 2 afskrifte moet die titel van die artikel, die naam van die skrywer/s, die hoogste graad behaal en die adres of naam van hulle betrokke instansie verskyn. Op die eerste bladsy van die oorblywende twee afskrifte moet slegs die titel van die artikel verskyn. Die *tweede* bladsy van alle afskrifte moet slegs 'n *opsomming* (100 woorde) in beide Engels en Afrikaans bevat. Afrikaanse opsommings sal vir buitelandse bydraers voorsien word. Alle paragrawe moet teenaan die linkerkantlyn begin word en moet nie ingekeep word nie.

Hoofopskrifte moet, waar dit van toepassing is, in die volgende volgorde wees: METODE, RESULTATE, BESPREKING, GEVOLGTREKKING, ERKENNINGS en VERWYSINGS.

Tabelle en Figure moet op afsonderlike bladsye verskyn (een bladsy per tabel/illustrasie). Figure, grafieke en lyntekeninge moet oorspronklike weergawes wees en moet in swart ink op wit papier van 'n hoë gehalte gedoen word.

Letterwerk wat hierop verskyn moet eenvormig wees, professioneel gedoen word en daar moet in gedagte gehou word dat dit leesbaar moet wees na 'n 50%-verkleining in drukwerk. Letterwerk by die illustrasie moet onder geen omstandighede getik word nie. Verklarings of omskrywings moet nie in die illustrasie nie, maar daaronder verskyn. Die byskrifte van tabelle moet bo-aan verskyn en dié van figure onderaan. Tabelle en figure moet in die volgorde waarin hulle verskyn, genommer word (met Arabiese syfers). Die hoeveelheid materiaal in die vorm van tabelle en illustrasies wat toegelaat word, word deur die redakteur bepaal (gewoonlik nie meer as 6 nie).

Verwysings. Verwysings in die teks moet voorsien word van die skrywer se van en die datum, bv. Van Riper (1971). Waar daar meer as twee skrywers is, sal *et al.* na die eerste skrywer voldoende wees. Die name van alle skrywers moet in die Verwysingslys verskyn. Verwysings moet alfabeties in 3-spasiëring aan die einde van die artikel gerangskik word. Vir die aanvaarde afkortings van tydskrifte se titels, raadpleeg die vierde uitgawe (Oktober) van *DSH ABSTRACTS* of *The World List of Scientific Periodicals.* Die getal verwysings wat gebruik is, moet nie veel meer as 20 wees nie.

Let op die volgende voorbeelde:

Locke, J.L. Clinical Phonology: The Explanation and Treatment of Speech Sound Disorders. *J. Speech Hear. Disord.*, 48, 339-341, 1983.

Penrod, J.P. Speech Discrimination Testing. In J. Katz (Ed.) *Handbook of Clinical Audiology*, 3de ed., Baltimore: Williams & Wilkins, 1985.

Van Riper, C. *The Nature of Stuttering.* Englewood Cliffs, New Jersey: Prentice Hall, 1971.

Proewe. Galeiproewe sal waar moontlik aan die skrywer gestuur word. Die onkoste van veranderings, behalwe tipografiese foute, sal deur die skrywer self gedra moet word.

Herdrukke. 10 herdrukke sonder omslae sal gratis verskaf word.

Alle manuskripte en korrespondensie moet gerig word aan:

Die Redakteur,
Die Suid-Afrikaanse Tydskrif vir Kommunikasieafwykings.
Die Suid-Afrikaanse Vereniging vir Spraak- en Gehoorheelkunde,
Posbus 31782,
Braamfontein 2017, Suid-Afrika.