AD-P003 853

NUMERICAL VALIDATION OF TUKEY'S CRITERION FOR CLINICAL TRIALS AND SEQUENTIAL TESTING

Charles R. Leake
USA Concepts Analysis Agency
ATTN: CSCA-RQR
8120 Woodmont Avenue
Bethesda, Maryland 20814

Abstract. A basic problem in conducting either clinical or sequential trials is to determine which or when statistical significance for a predetermined level of α has occurred. The criterion of

Tukey's $$\alpha_T = \alpha/k$$ for k nonoverlapping comparisons

for k nonoverlapping comparisons is mentioned in a paper by Tukey (1). The consequences of not using ~~this~~ criterion are developed. The use of this criterion might be too stringent, however, and an alternative statistic is given.

Introduction. Tukey (1) presented a paper at the Birnbaum Memorial Symposium in May 1977. This paper was later published in Science. In this paper, Tukey mentions a criterion to determine whether or not one can say that he has observed statistical significance other than some random noise when making a number of comparisons on a set of data. This criterion, with all apologies to Professor Tukey, has been bestowed with the name Tukey's Criterion through common usage in a number of circles in the military analytical community.

The criterion is, for a given level of significance say α where k is the total number of plausible comparisons, $\alpha_T = \alpha/k$. Thus, if one observes a difference which has a probability of occurring of α_T or less when one is comparing k nonoverlapping classes (or subsets) of a sample space, then one can say that this difference is statistically significant at the α level.

The converse shows why this is necessary. Table 1 gives a sample of the probability of not reaching statistical significance at $\alpha = .05$ (5%) and $\alpha_T = .05/k$ for a selected number of comparisons, as well as the probability of observing at least one statistical significance for $\alpha = .05$. Clearly for a fixed α level, the greater the number of comparisons which one makes, the more obvious it becomes that one will observe at least one statistical significance. Thus, the practice of conducting a test, making pair-wise comparisons, and reporting the significances for a fixed level of α shows a certain statistical naivety. When done deliberately, it raises an obvious ethical question. To quote Tukey on this subject,

> "The moral seems to me to be abundantly clear: Knowing that, for one class of patient, a clinical inquiry has reached some specific level of significance, such as 4%, is not evidence of the same strength as knowing that a focused clinical trial, involving a prechosen question, has reached exactly that level of significance, even if both the inquiry and the trial involved the same number of patients exposed to risk, and the same total number of end points, distributed in the same way." (1, p. 681)

Table 1. Sample of Probabilities of Not Reaching Significance

Sample number of comparisons	Probability of not reaching significance		Probability of at least one significance at 5%
	At 5%	At 5%/k	
1	95.0	95.0	5.0
2	90.2	95.1	9.8
3	85.7	95.1	14.3
4	81.5	95.1	18.5
5	77.4	95.1	22.6
10	60.0	95.1	40.0
20	35.8	95.1	64.2
50	7.7	95.1	92.3
100	0.6	95.1	99.4

What then can one do, when one is conducting an inquiry on a set of data that might not even have been created by the inquirer? There is one obvious answer to this question, use Tukey's Criterion to determine which comparisons are statistically significant.

In order to use Tukey's Criterion, one must first divide α by the number of comparisons to be made. Let's assume for illustrative purposes that $\alpha = .05$ and k, the number of comparisons is 20. It follows, then, that the α-level, adjusted for Tukey's Criterion becomes $\alpha_T = \alpha/k = .05/20 = .0025$.

Thus, the probability of rejecting H_0 is not .05 but .0025 when α is adjusted in accordance with Tukey's Criterion. The effect of this change in α-level is reflected by a corresponding change in the rejection region of the statistic being used. For example, if a Z-score is being used, for α =

.05, the critical Z is 1.64. On the other hand, if $\alpha_T = .0025$, as Tukey's Criterion specifies, then the critical Z is 2.81. Thus, the observed difference must be over 1.1σ greater than would be required if the α were not adjusted for Tukey's Criterion. As a result, the data may not be compatible with such a requirement for statistical significance. Another would be to use another statistic such as Scheffe comparisons in conjunction with an analysis of variance. However, in order to use analysis of variance, there are certain data requirements such as equal or proportional cell size in a two or more way analysis of variance. That available data Joes not always lend itself to such an analysis goes without saying.

It appears more likely that choosing either of these alternatives is unsatisfactory to the inquirer. Either Tukey's Criterion is too stringent, or one does not have the required prerequisities for an analysis of variance or a similar nonparametric substitute. What then?

Alternative Statistic. An examination of the problem raised by Tukey leads to an alternative approach to attempt to attach meaning to making comparisons on a set of data. Consider the following problem:

How many observed statistical significances made on k, nonoverlapping, and statistically independent comparisons must be made in order to say that the number observed has less than a 5% probability of occurring?

The answer to this question can be found by using the binomial distribution and solving for x, where

$$b(x:N,1-\alpha) < .05.$$

As shown in this inequality, x is the desired number of statistical significances, N the number of comparisons, and α, the significance level.

If this number of statistical significances is achieved, one could imply that factors other than chance were involved in obtaining that number of statistical significances. Moreover, this statistic could be used for parametric and nonparametric comparisons as well as a substitute method for an analysis of variance where such an analysis was unfeasible due to sample considerations.

Table 2, which was obtained by using the binomial theorem for $n \leq 100$, is shown below for $\alpha = .05$. For $n > 100$, a normal approximation of the binomial theorem can be used. The number of observed significances were obtained from a binomial table (2). This table, or the one below, can be used for $N \leq 100$ to determine whether or not the number of observed statistical significances occurred by chance alone.

Table 2. Number of Observed Statistical Significances for $\alpha = .05$ for N Comparisons to Occur with Less than 5% Probability

N, number of comparisons	Observed significances
1	1
2-7	2
8-17	3
18-28	4
29-40	5
41-53	6
54-66	7
67-79	8
80-96	9
97-100	10
n 100	Use normal approximation

REFERENCES

1. Tukey, J.W., "Some Thoughts on Clinical Trials, Especially Problems of Mutiplicity", Science, 18 Nov 77, pp. 679-684.

2. USAAMSAA, Cumulative Binomial Probability Distribution (AD 502418), 1979.