A Note on a Method for the Analysis of Significances en masse

PAUL SEEGER
University of Uppsala, Sweden

This note concerns the derivation of the p-mean significance levels, in the case of independent tests, for a mass-significance method developed by Eklund [1]. The solution is reached by formulating and solving an urn problem. Some comparisons are made with the p-mean significance levels of Duncan's multiple range test.

In three seminar papers from 1961–1963 [1] Eklund suggested the following solution to what he called the mass-significance problem: In large exploratory investigations it is desirable to keep the proportion of false significances low, at most equal to a small value k. Consider therefore the variable

$$y = \frac{\text{number of false significances}}{\text{number of significances}}$$

where the denominator is observed but the numerator has to be predicted. Both numerator and denominator are functions of the level of significance α', which is supposed to be used for each of N tests. Eklund's method consists in determining α' so that $y \leq k$, where k is predetermined. The observed number of significances at the level of α' may be denoted by $n(\alpha')$. Eklund considered three alternatives for the numerator. If the null hypothesis is true for each of the N tests we can predict the number of false significances to be $N\alpha'$. This is the most conservative of Eklund's alternatives. The method consists in finding a significance level α' for the individual test so that

$$\frac{N\alpha'}{n(\alpha')} \leq k$$

or

$$n(\alpha') \geq \frac{N\alpha'}{k} \tag{1}$$

Like the technique for making multiple comparisons based on Bonferroni's inequality [5], Eklund's method is only used to determine the level of significance for the individual test; it can be applied to N tests of any kind. One starts making the tests at the level $\alpha' = k$. If the criterion (1) is not satisfied, a lower value of

Received June 1967.

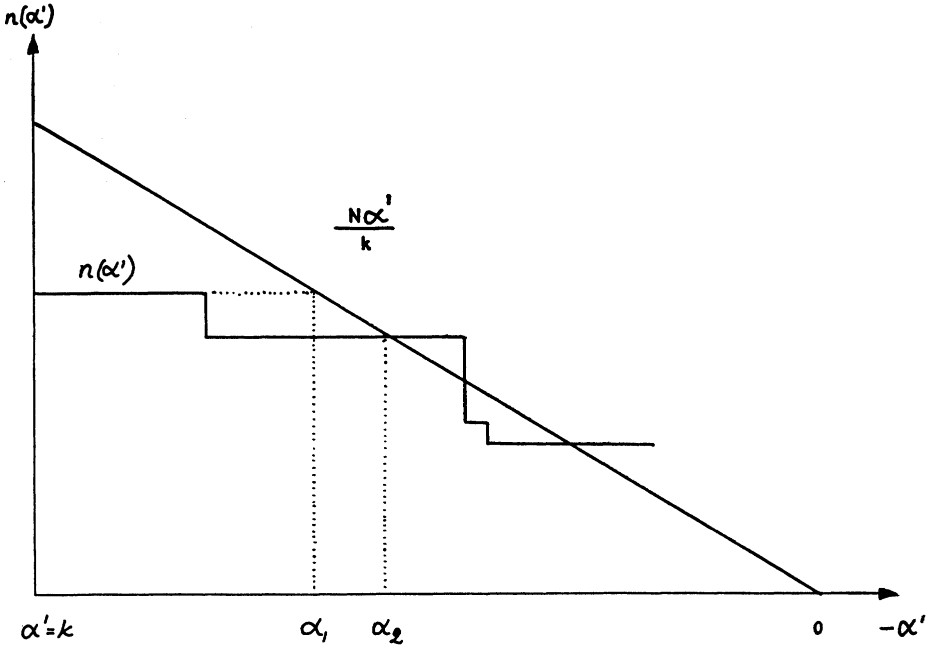

FIGURE 1
The number of significances $n(\alpha')$ plotted against the individual level of significance α'.

α' is tried until it is satisfied; that is until the curve in Fig. 1 touches the straight line. If this happens for the first time for $\alpha' = \alpha_0$, we say that we have obtained $n(\alpha_0)$ "mass-significances" ("approved" significances). In fact only a few α'-values need be investigated, as $n(\alpha')$ can not increase with decreasing α'. After a certain $\alpha' = \alpha_1$, where $n(\alpha_1) < N\alpha_1/k$ (and $n(\alpha') < N\alpha'/k$ for all $\alpha' > \alpha_1$), it is easy to compute the largest $\alpha' < \alpha_1$ which could give $n(\alpha_1) \geq n(\alpha') \geq N\alpha'/k$. Thus we investigate only $\alpha' = k$, α_1 and α_2 for the case of Fig. 1.

The individual significance levels of Eklund's method are determined not only by the number of tests, as, for example, in the method based on Bonferroni's inequality, but also by the observations themselves. This latter characteristic means that including some tests whose null hypotheses have very low prior probability of being true may increase the number of significances that exists among the original tests. However, requirement (1) seems to be a reasonable one, if measures are taken after every significant test (but not after non-significant ones). Only a small fraction of these measures are then unjustified.

As a referee has pointed out the properties of Eklund's method would certainly be greatly illuminated by a decision-theoretic discussion. In one of his alternatives Eklund, in a way, gave different prior probabilites to different null hypotheses when he assumed a certain (known) number $N_0 \leq N$ of them to be true. These aspects will not be considered here.

Eklund's method has also been discussed in a paper by Eklund and Seeger in 1965 [2] and in a monograph by Seeger 1966 [3]. In these publications the method is compared to some methods for making multiple comparisons. It is found to give about the same results as Duncan's multiple range test in two very large investigations with $N = 703$ and 1540 respectively. In [3] it is shown that when the N tests are independent, the probability of at least one significant result when all N null hypotheses are true is equal to k. That is, the experiment-wise error rate is k when the null hypothesis is true. It lies in the nature of the method that the probability of rejecting some true null hypotheses will be greater when some of the other hypotheses are not true. In order to characterize the behaviour of the method when some of N independent tests have untrue null hypotheses, we shall now compute the type of p-mean significance levels introduced by Duncan [4] and described by Miller [5].

Let us denote the "over-all" null hypothesis for the N tests by

$$H_0 : \mu_1 = 0, \qquad \mu_2 = 0, \cdots, \mu_N = 0.$$

The p-mean significance level is now defined as the maximum probability of falsely rejecting:

$$H^{(p)}_0 : \mu_1 = 0, \qquad \mu_2 = 0, \cdots, \mu_p = 0$$

for any values of $\mu_{p+1} \cdots \mu_N$. More precisely this can be written

$$\alpha_p = \sup_{\mu_{p+1} \ldots \mu_N} P\{D(\mu_1 \neq 0 \cup \mu_2 \neq 0 \cup \cdots \cup \mu_p \neq 0 \mid H_0^{(P)}$$

where $D(\mu_1 \neq 0 \cup \mu_2 \neq 0 \cup \cdots \cup \mu_p \neq 0)$ stands for the decision to reject at least one of the p null hypotheses.

The α_p of Eklund's method are now computed in the same way as for $\alpha_N = \alpha$ in [3]. We assume that the supremum values are taken for such values of $\mu_{p+1}, \mu_{p+2} \ldots \mu_N$ that the corresponding $N - p$ tests are certain to be significant at any level, for example, the level $\alpha' = k \cdot (N - p)/N$. This is obviously true for the case of testing means where the supremum values are taken as μ_{p+1}, $\mu_{p+2} \cdots \mu_N$ approaches $\pm \infty$

At least one significance among the other p tests is obtained if, either the p tests are significant at the level k,

or if $p - 1$ tests are significant at the level $k \cdot (N - 1)/N$,

or if $p - 2$ tests are significant at the level $k \cdot (N - 2)/N$,

.
.
.

or if $p - \nu$ tests are significant at the level $k \cdot (N - \nu)/N$,

.
.
.

or if 1 test is significant at the level $k \cdot (N - p + 1)/N$.

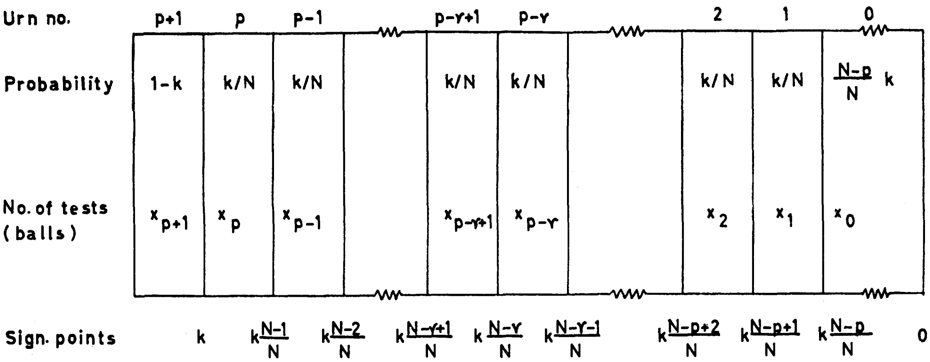

FIGURE 2

The probability that none of these events occur can be obtained if we consider the following urn problem:

We have $p + 2$ urns separated by the significance points $k \cdot (N - \nu)/N$, $\nu = 0, 1, \cdots, p$ (see Fig. 2.).

Now p balls are thrown at random into these $p + 2$ urns. For each urn the probability of receiving a thrown balls is given in Fig. 2. We want to find the probability of the following being true:

$$x_0 = 0, \qquad x_0 + x_1 = 0, \qquad x_0 + x_1 + x_2 \leq 1,$$
$$x_0 + x_1 + x_2 + x_3 \leq 2, \cdots, x_0 + x_1 + x_2 + x_3 + \cdots + x_p \leq p - 1,$$
$$x_0 + x_1 + x_2 + x_3 + \cdots + x_p + x_{p+1} = p.$$

It can be seen that if these requirements are fullfilled none of the p events above occur. Thus

$$1 - \alpha_p = P\left(\sum_{i=1}^{p+1} x_i = p\right) \cdot P\left(\sum_{i=1}^{m} x_i \leq m - 1; m = 1, 2, \cdots, p \;\middle|\; \sum_{i=1}^{p+1} x_i = p\right)$$

The first factor in this product is

$$P\left(\sum_{i=1}^{p+1} x_i = p\right) = \left(1 - k + \frac{p}{N} \cdot k\right)^p$$

The second factor can be computed as in [3], where the method was given by Docent Bengt Rosén, Uppsala. Let us introduce

$$\psi(s, p) = P\left(\sum_{i=1}^{m} x_i \leq m - 1; m = 1, 2 \cdots, p \;\middle|\; \sum_{i=1}^{p} x_i = s\right)$$

It is easily shown by induction in p that when s balls are thrown independently in p equally probable urns:

$$\psi(s; p) = 1 - \frac{s}{p}; \qquad s = 0, 1, 2, \cdots p.$$

From this follows that :

$$P\left(\sum_{i=1}^{m} x_i \le m - 1; m = 1, 2, \cdots, p \;\middle|\; \sum_{i=1}^{p+1} x_i = p\right)$$

$$= \underset{s}{E}\left\{\psi(s; p) \;\middle|\; \sum_{i=1}^{p+1} x_i = p\right\} = 1 - \frac{E\left(s \;\middle|\; \sum_{i=1}^{p+1} x_i = p\right)}{p}$$

$$= 1 - \frac{p \cdot \frac{pk}{N} \Big/ \left(1 - k + \frac{pk}{N}\right)}{p} = 1 - \frac{pk}{N\left(1 - k + \frac{pk}{n}\right)}$$

and that

$$1 - \alpha_p = \left(1 - k + \frac{pk}{N}\right)^p \cdot \frac{N - Nk + pk - pk}{N\left(1 - k + \frac{pk}{N}\right)} = (1 - k)\left\{1 - k\left(1 - \frac{p}{N}\right)\right\}^{p-1}$$

$$\alpha_p = 1 - (1 - k)\left\{1 - k\left(1 - \frac{p}{N}\right)\right\}^{p-1} \tag{2}$$

In the special case where $p = N$ we obtain the experimentwise error rate under the null hypothesis:

$$\alpha_N = \alpha = k.$$

It can also be seen that

$$\alpha_1 = k$$

$$\alpha_p > k; \qquad 1 < p < N$$

Thus, if we introduce some new tests, for which the null hypotheses are not true, we tend to increase the number of significances among the original tests (compare Fig. 3).

As a comparison, it might be mentioned that Newman–Keuls' test [5] has the same value α on all p-mean significance levels. In his range test Duncan uses the p-mean significance levels

$$\alpha_p = 1 - (1 - \alpha)^{p-1}.$$

However, as we have considered independent tests in deriving the expression for Eklund's method, we must, for the sake of comparison, mention that Duncan's motivation for his p-mean significance levels leads to:

$$\alpha_p = 1 - (1 - \alpha)^p, \tag{3}$$

if the tests are independent; the levels are based on degrees of freedom. Comparing the limits of (2) and (3) we find that when $N \to \infty$, and (if $k = \alpha$)

(i) p is finite:

$$\alpha_p^E \to 1 - (1 - k)^p = \alpha_p^D$$

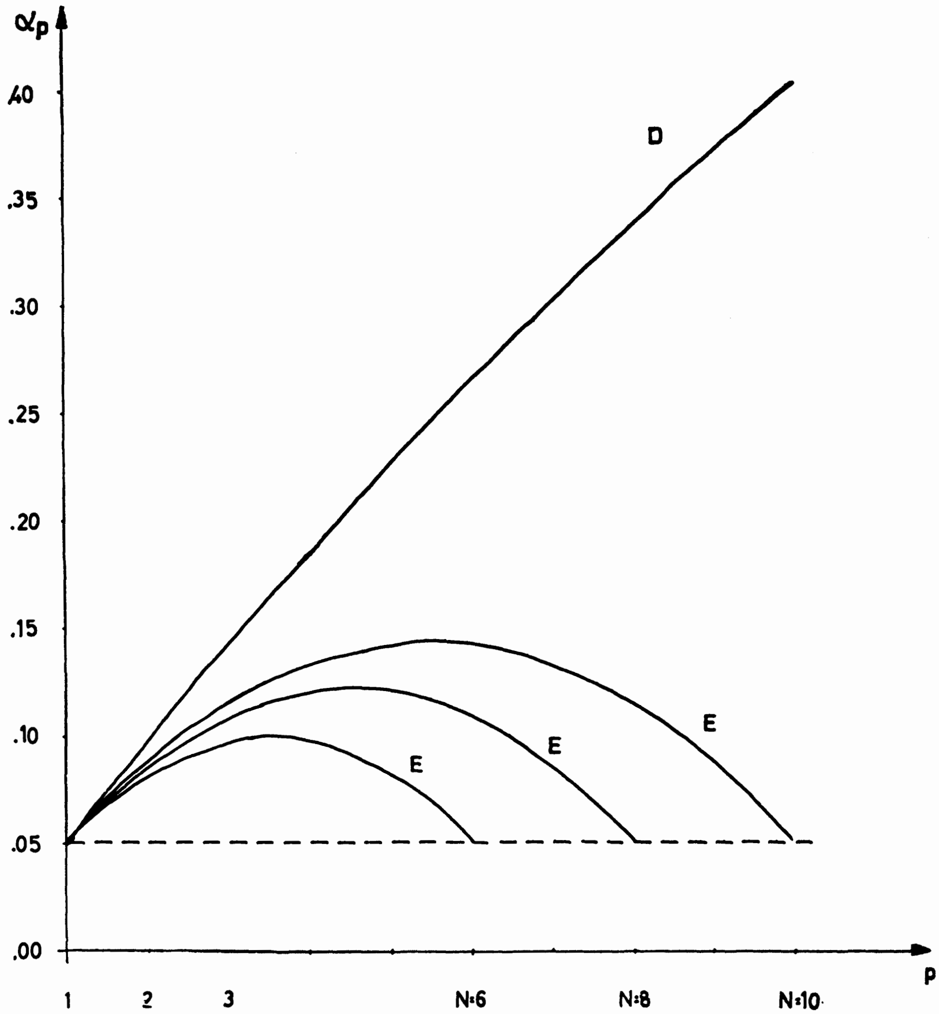

FIGURE 3

(ii) $p \to \infty$, so that $N - p \to \infty$:

$$\alpha_p^E \to 1 \ \alpha_p^D \to 1$$

(iii) $p \to \infty$, so that $N - p$ is finite:

$$\alpha_p^E \to 1 - (1 - k)e^{-k(N-p)} \alpha_p^D \to 1$$

The upper bound 1 is not very satisfactory, but it must be remembered that this bound has been attained for the most unfavourable cases, when all means, except the p in question, are infinite. This will, of course, never happen in practice.

It is easily seen from formulas (2) and (3) that

$$\alpha_p^E \leq \alpha_p^D$$

TABLE 1

p	No. of (dependent) tests with true null hypotheses for Eklund's method	α_p^E	No. of independent tests with true null hypotheses for Duncan's method	α_p^D
2	1		1	$1 - (1 - \alpha)^1$
3	3		2	$1 - (1 - \alpha)^2$
.	.		.	.
.	.		.	.
.	.		.	.
.	.		.	.
p	$\binom{p}{2}$		$p - 1$	$1 - (1 - \alpha)^{p-1}$
.	.		.	.
.	.		.	.
.	.		.	.
a	$\binom{a}{2}$		$a - 1$	$1 - (1 - \alpha)^{a-1}$

for any values of p and N. Their curves are compared for $\alpha = k = 0.05$ and $N = 6, 8, 10$ in Fig. 3. The difference seems to be very large. However, the comparison does not do justice to Duncan's method, which is intended for all possible pairwise comparisons among a means. The number of tests in Eklund's method would then be $\binom{a}{2}$, and they are not independent. Duncan utilizes the fact that the a means can be compared with the help of $a - 1$ independent tests. Fig. 3 can be said to illustrate the behaviour of Eklund's method as compared to making repeated t-tests without heightening the level of significance (the multiple-t-method).

It would be better to compare Eklund's and Duncan's methods by considering the p-mean significance levels in Table 1.

At least an appreciation of these levels can be obtained if we use formula (2) changing p to $\binom{p}{2}$ and N to $\binom{a}{2}$; although it is derived for independent tests. (Intuitively, it seems to give upper limits when comparing means by double-sided tests.) For $\alpha = k = 0.05$ and $a = 5$ the probabilities are given in Table 2.

TABLE 2

p	No. of dependent tests with true null hypotheses	α_p^E	No. of independent tests with true null hypotheses	α_p^D
2	1	0.0500	1	0.0500
3	3	0.1153	2	0.0975
4	6	0.1413	3	0.1426
5	10	0.0500	4	0.1855

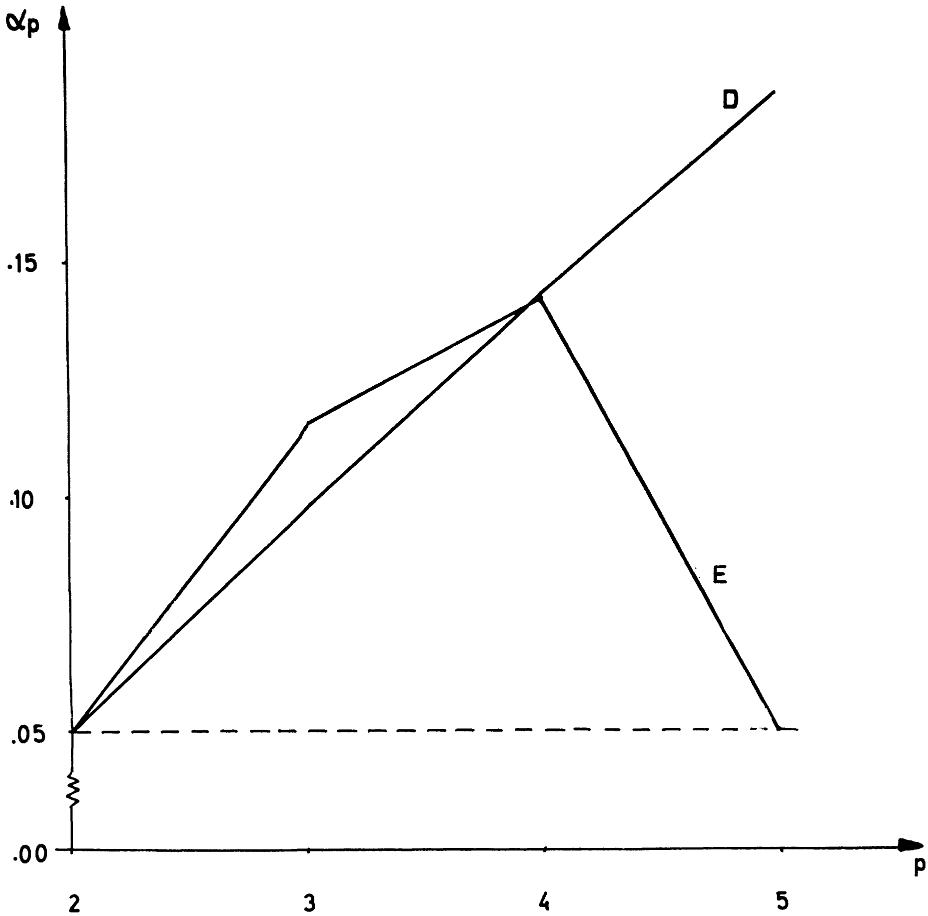

FIGURE 4

A graphical illustration is given in Fig. 4. In this figure Newman–Keuls' method would be represented by a horizontal line through $\alpha_p = 0.05$ and Scheffé's and Tukey's well-known methods [5] by curves rising from a value near zero (depending on the degrees of freedom for the estimate of error) when $p = 2$ to 0.05 when $p = 5$. The technique based on Bonferroni's inequality would give a curve rising from 0.005 to 0.05.

REFERENCES

1. EKLUND, G., 1961–1963. *Massignifikansproblemet.* Unpublished seminar papers, Uppsala University Institute of Statistics.
2. EKLUND, G. and SEEGER, P., 1965. *Massignifikansanalys.* Statistisk Tidskrift Stockholm, third series vol. 4, 355–365.
3. SEEGER, P., 1966. *Variance Analysis of Complete Designs. Some practical aspects.* Almqvist & Wiksell, Stockholm.
4. DUNCAN, D. B., 1955. Multiple range and multiple *F*-tests. *Biometrics 11,* 1–42.
5. MILLER, R. G., 1966. *Simultaneous Statistical Inference.* McGraw-Hill, New York.

LINKED CITATIONS

References

[4] **Multiple Range and Multiple F Tests**
David B. Duncan
Biometrics, Vol. 11, No. 1. (Mar., 1955), pp. 1-42.
Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28195503%2911%3A1%3C1%3AMRAMFT%3E2.0.CO%3B2-P