BIOMETRICS 38, 597–611
September 1982

Nonparametric Analysis of Covariance by Matching

Dana Quade

Department of Biostatistics, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27514, U.S.A.

SUMMARY

The basic problem under consideration is the comparison of treatments with respect to a response Y when a covariable X is taken into account. Various methods involving matching may be regarded as compromises between the standard analysis of covariance and the standard analysis of independent matched pairs. First, there is no need to restrict attention to independent matched pairs, but rather all matched pairs may be incorporated. Then, if X is concomitant, that is, if its distribution is the same regardless of treatment, methods may be used which are based ultimately on randomization although in practice they are based on analysis of variance. When X is not concomitant, methods related to partial correlation (between Y and 'treatment', given X) are applicable. All methods considered may use either the actual magnitudes of Y or analogues of their ranks.

1. Introduction

Given data $(Y_{ij}, \mathbf{X}_{ij})$ for $i = 0, 1, \ldots, k$ and $j = 1, \ldots, n_i$, the task at hand is to compare the $k+1$ samples with respect to the response Y while taking account of the covariable $\mathbf{X} = (X_1, \ldots, X_m)'$, where $m \geq 1$. In particular, do some samples more than others show a tendency toward higher values of Y for fixed $\mathbf{X}$? If so, what are the magnitudes of such tendencies? Note that only univariate responses are considered, although the covariable may be multivariate. Also, attention is restricted to one-way layouts. For convenience, the sample with $i = 0$ is referred to as the 'control' and those with $i = 1, \ldots, k$ as 'treatments'.

The two standard approaches to analyzing such data are first, through classical analysis of covariance, in which a least squares or maximum likelihood approach is applied, on the assumption of a complicated model; and second, by formation of independent matched groups, which allows a simple analysis under weak assumptions but may discard valuable information. The purposes of this paper are to discuss the philosophical basis for these standard methods, and the difference between them; and to suggest new compromise methods (analysis of covariance by matching, and analysis of matched differences).

2. Assumptions

It will be assumed throughout either that the observations form *independent random samples* (Assumption IA) or that they come from a *completely randomized design* (Assumption IB). In this section we discuss some other assumptions which are commonly made in performing analysis of covariance.

Let the joint distribution of Y and $\mathbf{X}$, under Assumption IA, be

$$J_i(Y, \mathbf{X}) = F_i(\mathbf{X}) C_i(Y \mid \mathbf{X}),$$

Key words: Concomitant variables; Adjusted values; Matched pairs; Matched groups; Caliper matching; Category matching; Ranks; Dental caries; Randomization; Confidence interval.

where F_i is the marginal distribution of X and C_i is the conditional distribution of Y given $\mathbf{X}$. When attention is directed to the testing of a hypothesis, it will be that the conditional distribution is the same in all populations: that is, simply,

$$H_0: C_i(Y \mid \mathbf{X}) = C(Y \mid \mathbf{X}), \qquad i = 0, \ldots, k.$$

This hypothesis is stated nonparametrically, in that nothing whatever is assumed about the form of the common conditional distribution function C.

However, the more interesting question of estimating the magnitudes of the differences requires a definition of what is meant by treatment effects, and this may involve *parallelism* (Assumption II). The populations are said to be parallel if

$$C_i(Y \mid \mathbf{X}) = C(Y - \delta_i \mid \mathbf{X}), \qquad i = 0, 1, \ldots, k,$$

where $\delta_0 = 0$ by definition; then $\delta_1, \ldots, \delta_k$ are called the 'effects' of the k treatments. (It should be obvious that the meaning of the effects can be usefully altered by suitable transformation of Y; for example, for the treatments to produce constant percentage changes in response, one may transform Y to log Y before making Assumption II.) On the assumption that the populations are parallel, the null hypothesis can be expressed as

$$H_0: \delta_1 = \cdots = \delta_k = 0.$$

One is now also in a position to test nonnull hypotheses such as

$$H(\boldsymbol{\phi}): \boldsymbol{\delta} = \boldsymbol{\phi},$$

where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_k)'$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)'$. This is done by defining new variables

$$Q_0 = Y_0, \qquad Q_1 = Y_1 - \phi_1, \ldots, Q_k = Y_k - \phi_k,$$

whose treatment effects vanish if and only if $H(\boldsymbol{\phi})$ is true: testing $H(\boldsymbol{\phi})$ is then equivalent to testing the null hypothesis with respect to the new response Q. Furthermore, it is well established that the set of all values in the parameter space which a test accepts is a confidence set for the parameter: that is, given that $H(\boldsymbol{\phi})$ can be tested, the values $\boldsymbol{\phi}$ which the test accepts form a confidence set for $\boldsymbol{\delta}$.

The standard 'parametric' analysis of covariance also requires *homogeneity* (Assumption III): usually this means that the conditional distribution depends on $\mathbf{X}$ only through $\psi(\mathbf{X}; \boldsymbol{\beta})$, the regression function of Y on $\mathbf{X}$, as follows:

$$C_i(Y \mid \mathbf{X}) = C_i\{Y - \psi(\mathbf{X}; \boldsymbol{\beta})\},$$

where the distribution function $C_i(z)$ has median 0 for $i = 0, \ldots, k$. A possible weakening of this assumption is to allow Y to depend on $\mathbf{X}$ also through a scale function σ, so that

$$C_i(Y \mid \mathbf{X}) = C_i[\{Y - \psi(\mathbf{X}; \boldsymbol{\beta})\}/\sigma(\mathbf{X}; \boldsymbol{\gamma})].$$

The standard analysis, on the contrary, strengthens Assumption III to include *linearity* (Assumption IV), which means that

$$\psi(\mathbf{X}; \boldsymbol{\beta}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m,$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)'$. The final requirement for the classical analysis is that the conditional distribution function of Y given $\mathbf{X}$ be *normal* (Assumption V). Then, putting together the Assumptions I–V, of independent random samples (or a completely randomized design), parallelism, homogeneity, linearity and normality, one arrives at the usual model

$$Y_{ij} = \delta_i + \beta_0 + \beta_1 X_{ij1} + \cdots + \beta_m X_{ijm} + E_{ij}, \qquad i = 0, \ldots, k, \quad j = 1, \ldots, n_i,$$

where $\delta_0 = 0$ and the E_{ij} are independent normal variables with mean 0 and common standard deviation σ.

The methods to be proposed in this paper do not require homogeneity, linearity and normality, but for exact results in small samples they depend upon the *concomitance* of $\mathbf{X}$ (Assumption VI), which is not required mathematically for the classical analysis. This means that the marginal distribution of $\mathbf{X}$ is the same in all populations: that is,

$$F_i(\mathbf{x}) = F(\mathbf{x}), \qquad i = 0, 1, \ldots, k,$$

where, however, nothing is stated about the form of F. Under this assumption, the joint distribution function of Y and $\mathbf{X}$ is

$$J_i(Y, \mathbf{X}) = F(\mathbf{X}) C_i(Y \mid \mathbf{X}), \qquad i = 0, 1, \ldots, k,$$

and the null hypothesis that the C_i are equal can be re-expressed as

$$H_0: J_i(Y, \mathbf{X}) = J(Y, \mathbf{X}), \qquad i = 0, 1, \ldots, k.$$

Thus, the concomitance of $\mathbf{X}$ turns the test of H_0 into a special case of the comparison of multivariate samples, the alternative to identify of the joint distributions being a difference in the distribution of the single component Y.

Note that if Assumption IB (rather than IA) holds, then the value of $\mathbf{X}$ is first established for each experimental unit, after which the units are randomly assigned to be controls or to have any particular treatment applied. In that case concomitance is a fact which can logically deduced, and need not be assumed separately. Whether deduced or assumed, however, concomitance justifies using a randomization analysis in comparing the samples.

3. Concepts of Control

What does it mean to 'control for' or 'take account of' a covariable $\mathbf{X}$? In this paper consideration is limited to situations where the design has already been determined and the data are in hand, so that the only question is how to proceed with the analysis. Then there seem to be two quite different concepts in the literature: *adjustment for* $\mathbf{X}$, which underlies the classical analysis of covariance, and *holding* $\mathbf{X}$ *constant*, which underlies procedures based on matching or blocking.

To adjust for $\mathbf{X}$, one first establishes a relationship—a regression function $\psi(\mathbf{X})$—which is suitable for predicting Y from $\mathbf{X}$ in the sense that the residual $R = Y - \psi(\mathbf{X})$ is concentrated about zero as closely as possible according to some reasonable criterion. The 'adjusted value' Y', adjusted to the common X-value $\mathbf{x}^*$, is

$$Y' = Y + \psi(\mathbf{x}^*) - \psi(\mathbf{X}) = \psi(\mathbf{x}^*) + R.$$

The hypothesis of no treatment effects is then tested by comparing the adjusted values from the different samples (or, what is equivalent mathematically, the residuals).

The intuitive idea behind the method of adjustment is that one may make equitable comparisons among responses at different covariate values by substituting the responses which presumably would have been observed at the common covariate value $\mathbf{x}^*$. The method stands or falls on the quality of the adjustment—in particular, on whether the function ψ, or at least its form, is well-chosen, assuming its parameters can be suitably estimated. The analysis thus obviously involves the assumption of homogeneity. However, the assumption of concomitance may be needed also. That is because, when $\mathbf{X}$ is not concomitant, although the standard analysis of covariance may still be justifiable from a

purely mathematical standpoint, there is often a nagging doubt concerning its application. For the classical method the assumptions of homogeneity and linearity establish a structure under which one adjusts the responses observed at different values of $\mathbf{X}$ to a common value which may not be at all typical for some samples; one may wonder how to interpret hypothetical responses at an atypical common covariate value $\mathbf{x}^*$. This amounts to extrapolating data and involves all the usual pitfalls of such a procedure.

A standard approach in the spirit of holding $\mathbf{X}$ constant is to form independent matched groups such that each contains m_i observations from the ith sample, $i = 0, 1, \ldots, k$, with all $M = \sum m_i$ observations within the group having the same value of $\mathbf{X}$, or at least sufficiently close to the same that comparisons of responses within the group are considered equitable. (The most common special case is that of matched pairs, where $k = 1$, $M = 2$, and $m_0 = m_1 = 1$.) Various simple methods are available for dealing with data of this form. Matching may be extremely wasteful, however, particularly if the sample sizes are small, since then there will be many observations which cannot be matched. Furthermore, matching generally involves an element of arbitrariness: suppose, for example, that a treated observation can be matched with either of two controls, but m_0 has been taken equal to 1; then one control must be chosen for comparison and the other ignored.

One can avoid these disadvantages of matching, at the expense of some complication in the analysis, by categorizing the covariable and forming blocks, each of which contains all the observations with covariable values in a given category. This often involves another kind of arbitrariness, however. Suppose the covariable is age of human subjects, categorized in ten-year intervals such as 20–29, 30–39, etc., and subjects A, B and C are, respectively, aged 24, 28 and 30 years. Then the analysis involves the comparison between A and B, yet surely for almost any purpose it would be better to compare B and C. Thus, it is intuitively preferable to use 'caliper matching' (based on some tolerance, say ε, the maximum amount by which two covariable values are allowed to differ) rather than 'category matching'; standard analyses, however, require discrete matched groups.

The methods of analysis proposed in this paper are based on holding $\mathbf{X}$ constant by caliper matching, but utilizing *all* matches: in the example of the preceding paragraph, with the tolerance set at 5 years, say, they would compare B's response with those of both A and C, while yet not comparing A with C.

4. Analysis of Covariance by Matching

For cases where concomitance holds, the following general principle was enunciated by Quade (1967):

> If the hypothesis is true then the populations are all identical and the samples can be pooled. So use the pooled sample to determine a relationship through which Y can be predicted from X. Then compare each observed response Y_{ij} with the value which would be predicted for it from the corresponding X_{ij}, and assign it a score Z_{ij}, positive if Y_{ij} is greater than predicted, and negative if smaller. Finally, compare the populations by performing an ordinary one-way analysis of variance of the scores.

An intuitively obvious candidate for predicting Y given $\mathbf{X} = \mathbf{x}$ is the average of the values of Y observed for cases where $\mathbf{X}$ is at least approximately equal to $\mathbf{x}$. To make this more specific, let there be defined for all pairs $(\mathbf{x}_1, \mathbf{x}_2)$ of possible values of $\mathbf{X}$ some distance function $D(\mathbf{x}_1, \mathbf{x}_2)$, and let the tolerance ε be chosen such that the values $\mathbf{x}_1$ and $\mathbf{x}_2$ are regarded as approximately equal, or 'matched', if and only if $D(\mathbf{x}_1, \mathbf{x}_2) \leqslant \varepsilon$. Then, writing

$I\{\cdot\}$ for the indicator function of an event, the predicted value of Y_{ij} is

$$\hat{Y}_{ij} = \sum_{i'=0}^{k} \sum_{j'=1}^{n_{i'}} Y_{i'j'} I\{D(\mathbf{X}_{ij}, \mathbf{X}_{i'j'}) \leqslant \varepsilon\} / M_{ij},$$

where

$$M_{ij} = \sum_{i'=0}^{k} \sum_{j'=1}^{n_{i'}} I\{D(\mathbf{X}_{ij}, \mathbf{X}_{i'j'}) \leqslant \varepsilon\}$$

is the number of observations in all samples which match observation (i, j), and $\hat{Y}_{ij}$ is their mean response. The corresponding score is the difference $Y_{ij} - \hat{Y}_{ij}$. Under the null hypothesis of identical conditional (and hence, using the concomitance assumption, joint) distributions, these scores are interchangeable, and hence it is correct asymptotically (for large sample sizes) to use them in an analysis of variance (see Quade, 1965). This may be called 'analysis of covariance by matching'. Note that if the tolerance ε is so increased that all pairs are matched, then $M_{ij} \equiv N$ and $\hat{Y}_{ij} = \bar{Y}$ (the overall mean) for all (i, j), the scores are just the deviations from the mean, and the test reduces to ordinary analysis of variance. Thus the procedure proposed here is a generalization of the analysis of variance.

For an illustration of the analysis of covariance by matching, consider the artificial data of Table 1. In this conveniently small numerical example (more realistic examples appear in §7) there is one treatment ($k = 1$), with $n_1 = 3$ observations, and there are $n_0 = 4$ controls. The covariable X is univariate, and the tolerance for matching is $\varepsilon = 10$. The first observation $(Y_{01}, X_{01}) = (258, 17)$ is matched with itself (by definition) and four others $(Y_{02}, X_{02}) = (240, 20)$, $(Y_{03}, X_{03}) = (242, 21)$, $(Y_{11}, X_{11}) = (644, 11)$, and $(Y_{12}, X_{12}) = (526, 25)$], so $M_{01} = 5$. The mean response for these five observations is 382, which is then $\hat{Y}_{01}$, and the corresponding score is $258 - 382$ or -124. The other scores in the Table may be checked similarly. The mean scores are -86 for controls and 152 for treated. Analysis of variance of these scores is of course equivalent to a t test, which produces $t = 2.74$ with 5 df, and an approximate significance level $P = .020$ for testing against a one-sided alternative. For such small sample sizes the analysis of variance approximation cannot be relied upon, but an exact randomization analysis is feasible. In this example the observed result is the second most significant possible among the $7!/4!\,3! = 35$ reallocations of the observations between treatment and control, so the exact P value is $2/35 = .057$ (one sided).

Table 1

Small example ($n_0 = 4$, $n_1 = 3$) to illustrate calculations

Data			Analysis of variance				Matched differences		
j	X_{ij}	Y_{ij}	M_{ij}	$\hat{Y}_{ij}$	$Y_{ij} - \hat{Y}_{ij}$	S_{ij}	M_{ij}	D_{ij}	T_{ij}
Control observations ($i = 0$)									
1	17	258	5	382	−124	.00	2	654	2
2	20	240	5	382	−142	−.80	2	690	2
3	21	242	5	382	−140	−.40	2	686	2
4	38	888	2	826	62	.50	1	−124	−1
Treated observations ($i = 1$)									
1	11	644	4	346	298	.75	3	1192	3
2	25	526	5	406	120	.40	3	838	3
3	32	764	3	726	38	.00	1	−124	−1

Assuming that the two populations are parallel, the point estimate of the treatment effect is the difference between the mean scores, $152-(-86)=238$. In this situation ($k=1$), it is easy to construct the randomization confidence distribution for the treatment effect. After relabeling the observations to have a single subscript $i=1,\ldots,N$, define $u_i=0$ if the ith observation is a control and $u_i=1$ if it is treated, and let t_i be the proportion treated among those observations which match the ith. For any nontrivial permutation $\mathbf{v}=(v_1,\ldots,v_N)'$ of $\mathbf{u}=(u_1,\ldots,u_N)'$ let

$$\delta(\mathbf{v})=\frac{\sum(u_i-v_i)(Y_i-\hat{Y}_i)}{\sum(u_i-v_i)(u_i-t_i)}.$$

Then the values $\delta(\mathbf{v})$ corresponding to the $(N!/n_0!\,n_1!)-1$ possible values of $\mathbf{v}$ are the boundary points of the confidence distribution, and each gap in the array of values of δ has equal confidence $n_0!\,n_1!/N!$. (For a proof see the Appendix.) In the example, if the observations are ordered as in Table 1, then the values of u_i are $(0,0,0,0,1,1,1)$, the values of t_i are $(\frac{2}{5},\frac{2}{5},\frac{2}{5},\frac{1}{2},\frac{1}{4},\frac{2}{5},\frac{2}{3})$, and for $\mathbf{v}=(0,1,0,1,0,1,0)$, say, the corresponding δ is

$$\delta(\mathbf{v})=\frac{0(-124)-(-142)+0(-140)-(62)+(298)+0(120)+(38)}{0(-\frac{2}{5})-(-\frac{2}{5})+0(-\frac{2}{5})-(-\frac{1}{2})+(\frac{3}{4})+0(\frac{3}{5})+(\frac{1}{3})}$$

$$=\frac{416}{(119/60)}=209.75.$$

The full array of 34 values of δ is

−28.80 52.73 120.00 128.73 129.82 188.80 200.67 208.74 209.75
213.33 220.44 220.56 220.91 221.23 221.33 226.59 242.73
243.46 244.00 244.62 245.46 253.85 260.00 262.00 298.96 317.21
318.14 318.58 319.65 325.58 328.14 366.96 380.87 382.61

From this it follows that a one-sided interval with confidence coefficient 28/35 (say), or 80%, extends from 200.67 to $+\infty$.

An analogous procedure can be obtained which generalizes the rank analysis of variance, or Kruskal–Wallis test. For $i=0,1,\ldots,k$ and $j=1,\ldots,n_i$ define scores

$$S_{ij}=\sum_{i'=0}^{k}\sum_{j'=1}^{n_{i'}}\operatorname{sgn}(Y_{ij}-Y_{i'j'})I\{D(\mathbf{X}_{ij},\mathbf{X}_{i'j'})\leqslant\varepsilon\}/M_{ij};$$

in words, S_{ij} is the proportion of observations $(Y_{i'j'},X_{i'j'})$ which have $Y_{i'j'}<Y_{ij}$, minus the proportion with $Y_{i'j'}>Y_{ij}$, among those in all the samples combined which are matched with (Y_{ij},X_{ij}). Then an analysis of variance can be performed using these scores, or a randomization test if the samples are sufficiently small. That this test does indeed generalize the Kruskal–Wallis test can be seen by choosing ε sufficiently large that all pairs of observations are matched: then each $M_{ij}=N$ and $S_{ij}=\{2R_{ij}-(N+1)\}/N$, where R_{ij} is the rank of Y_{ij} among the N values of Y. Note that S_{ij} is related to the score $Y_{ij}-\hat{Y}_{ij}$ of the previous test by a change of $Y_{ij}-Y_{i'j'}$ to $\operatorname{sgn}(Y_{ij}-Y_{i'j'})$ in the definition. This test may be called 'rank analysis of covariance by matching'.

The scores S_{ij} have also been entered into Table 1. For example, $S_{01}=0$ because, considering the five observations matched with (Y_{01},X_{01}), the two controls other than (Y_{01},X_{01}) itself have smaller values of Y and the two treated observations have larger

values. Given the values of S_{ij}, a little computation will verify that the exact one-sided P value is 5/35 or .143 for these data.

There does not appear to be any direct method for constructing a confidence distribution corresponding to rank analysis of covariance by matching. It is not difficult to see, however, that if $k = 1$, then the possible boundary points for confidence intervals are the differences between responses in matched between-sample pairs. In the example, these treated − control differences are $\{-124, 268, 284, 286, 386, 402, 404\}$, and the interval $(-124, \infty)$ has confidence coefficient 34/35 or 97% for the treatment effect δ.

5. Analysis of Matched Differences

In this section the assumption of concomitance is not required mathematically. Except for a few remarks at the end, however, attention is restricted to the situation where there is only one treatment ($k = 1$).

A reasonable overall measure of the effect of the treatment is the sample 'matched difference in mean',

$$\bar{D} = \frac{\sum_j \sum_{j'} (Y_{1j} - Y_{0j'}) I\{D(\mathbf{X}_{1j}, \mathbf{X}_{0j'}) \leq \varepsilon\}}{\sum_j \sum_{j'} I\{D(\mathbf{X}_{1j}, \mathbf{X}_{0j'}) \leq \varepsilon\}}.$$

The population matched difference in mean, which this estimates, is the conditional expected difference between the response of a treated observation and the response of a control observation, given that they are matched on $\mathbf{X}$ (within tolerance ε). An alternative measure is the 'matched difference in probability',

$$\bar{T} = \frac{\sum_j \sum_{j'} \operatorname{sgn}(Y_{1j} - Y_{0j'}) I\{D(\mathbf{X}_{1j}, \mathbf{X}_{0j'}) \leq \varepsilon\}}{\sum_j \sum_{j'} I\{D(\mathbf{X}_{1j}, \mathbf{X}_{0j'}) \leq \varepsilon\}}.$$

This estimates the population matched difference in probability, which is the conditional probability that a treated observation will show a greater response than a control observation, minus the conditional probability that it will show a smaller response, given that they are matched on $\mathbf{X}$ (within tolerance ε).

Either matched difference is essentially an average of conditional differences at fixed values of $\mathbf{X}$. This is most clearly seen in the special case where $\mathbf{X}$ is discrete, with possible values $\{v_k\}$, say, and the tolerance allowed is $\varepsilon = 0$. Then

$$\bar{D}_K = \sum_j \sum_{j'} (Y_{1j} - Y_{0j'}) I\{\mathbf{X}_{1j} = \mathbf{X}_{0j'} = v_k\}/M_k$$

and

$$\bar{T}_k = \sum_j \sum_{j'} \operatorname{sgn}(Y_{1j} - Y_{0j'}) I\{\mathbf{X}_{1j} = \mathbf{X}_{0j'} = v_k\}/M_k$$

are the sample conditional differences at v_k, where

$$M_k = \sum_j \sum_{j'} I\{\mathbf{X}_{1j} = \mathbf{X}_{0j'} = v_k\}$$

is the number of treatment − control pairs with $\mathbf{X} = v_k$; and the corresponding matched differences are the simple weighted averages $\bar{D} = \sum w_k \bar{D}_k$ and $\bar{T} = \sum w_k \bar{T}_k$ where $w_k = M_k / \sum M_k$. The definitions of $\bar{D}$ and $\bar{T}$ given in the preceding paragraph are easily seen to generalize these.

Even if $\mathbf{X}$ is not discrete, one may define

$$\delta(\mathbf{x}) = \int\int (Y_1 - Y_0)\, dC_0(Y \mid \mathbf{X} = \mathbf{x})\, dC_1(Y \mid \mathbf{X} = \mathbf{x})$$

and

$$\theta(\mathbf{x}) = \int\int \operatorname{sgn}(Y_1 - Y_0)\, dC_0(Y \mid \mathbf{X} = \mathbf{x})\, dC_1(Y \mid \mathbf{X} = \mathbf{x})$$

as the conditional differences in mean and in probability, respectively, given $\mathbf{X} = \mathbf{x}$ (with tolerance zero). The corresponding averages

$$\bar{\delta} = \tfrac{1}{2}\int \delta(\mathbf{x})\, d\{F_0(\mathbf{x})F_1(\mathbf{x})\} \quad \text{and} \quad \bar{\theta} = \tfrac{1}{2}\int \theta(\mathbf{x})\, d\{F_0(\mathbf{x})F_1(\mathbf{x})\}$$

may be called the 'partial differences'. These may be regarded as ideal measures of differences between the populations when the covariable is completely controlled for (or 'partialled out'). In the discrete case with tolerance zero, the population matched differences are identical with them; otherwise, matched differences are only imperfect approximations.

Even in the most general case, $\bar{D}$ and $\bar{T}$ may be useful summary statistics; however, if the populations are parallel, then other approaches seem reasonable. In this case $\delta(\mathbf{x})$ and $\theta(\mathbf{x})$ do not depend on $\mathbf{x}$, and the partial differences are equal to their common values; furthermore, $\bar{\delta}$ is then the same as the treatment effect δ defined previously. Thus, assuming parallelism, one might estimate δ by that constant which, when subtracted from the response of each treated observation, makes $\bar{D}$ or $\bar{T}$ equal to zero. It may be seen that these statistics are respectively the 'mean matched difference' and the 'median matched difference' between treatment and control in matched pairs. The first of these, however, is just $\bar{D}$ itself: that is, the matched difference in mean is the same as the mean matched difference.

A matched or partial difference defined as an average conditional difference is analogous to a partial correlation defined as an average conditional correlation. To make this analogy more explicit, let Z be a dummy variable taking on the value 1 for treated observations and 0 for controls, and let a pair of observations be defined as relevant if it consists of one treated and one control observation, matched on $\mathbf{X}$. Then the matched difference in probability is formally identical to the matched partial correlation coefficient between Y and Z given X, as defined by Quade (1974). This formal equivalence between $\bar{T}$ and the matched partial correlation coefficient makes it possible to use computer programs designed for the latter; for example, see Johnson, Quade and Langston (1981). Furthermore, the classical analysis of covariance is formally equivalent to testing the significance of the product-moment partial correlation between Y and Z given $\mathbf{X}$. Note also that in general if $(Y_0, Z_0, \mathbf{X}_0)$ and $(Y_1, Z_1, \mathbf{X}_1)$ are two observations then

$$\mathrm{E}\{(Z_1 - Z_0)(Y_1 - Y_0) \mid \mathbf{X}_0 = \mathbf{X}_1 = \mathbf{x}\} = 2\,\mathrm{cov}(Y, Z \mid \mathbf{X} = \mathbf{x}),$$

but putting $Z_1 = 1$ and $Z_0 = 0$ yields

$$\mathrm{E}\{(Y_1 - Y_0) \mid \mathbf{X}_0 = \mathbf{X}_1 = \mathbf{x}\} = \delta(\mathbf{x});$$

thus the conditional difference in means is formally proportional to a conditional correlation.

Both matched differences may be recognized as ratios of two-sample U-statistics: for

example, their common denominator is the U-statistic

$$\frac{1}{n_0 n_1}\sum_i\sum_{i'} I\{D(\mathbf{X}_{1i}, \mathbf{X}_{0i'}) \leqslant \varepsilon\},$$

which is a consistent estimator of the probability that a treated and a control observation chosen at random will be matched within tolerance ε. From standard theory of U-statistics, then, the two matched partial difference statistics are asymptotically normally distributed as the sample sizes increase, with standard deviations estimable by extending the method of components of Sen (1960) as in Quade (1974). For $i = 0, 1$ and $j = 1, \ldots, n_i$ let

$$M_{ij} = \sum_{j'=1}^{n_{1-i}} I\{D(\mathbf{X}_{ij}, \mathbf{X}_{1-i,j'}) \leqslant \varepsilon\}$$

be the number of observations matching $\mathbf{X}_{ij}$ in the other sample. (Note that this is *not* the same as the M_{ij} of the preceding section.) Then

$$M = \sum_{j=1}^{n_0} M_{0j} = \sum_{j=1}^{n_1} M_{1j}$$

is the total number of matched pairs.

Let

$$D_{ij} = (-1)^{1-i}\sum_{j'=1}^{n_{1-i}} (Y_{ij} - Y_{1-i,j'}) I\{D(\mathbf{X}_{ij}, \mathbf{X}_{1-i,j'}) \leqslant \varepsilon\};$$

then $\bar{D} = \sum D_{0j}/M = \sum D_{1j}/M$. Its standard error may be written

$$\mathrm{SE}(\bar{D}) = \frac{1}{M}\left\{\sum_{i=0}^{1}\sum_{j=1}^{n_i} M_{ij}^2\left(\frac{D_{ij}}{M_{ij}} - \bar{D}\right)^2\right\}^{\frac{1}{2}},$$

from which one sees that $\mathrm{var}(\bar{D})$ is essentially proportional to the dispersion of the 'microestimates' D_{ij}/M_{ij}. A less intuitive formula, but more convenient for computation, is

$$\mathrm{var}(\bar{D}) = \frac{1}{M^4}\sum_i\left\{M^2\sum_j D_{ij}^2 - 2M\sum_j D_{ij}\sum_j D_{ij}M_{ij} + \left(\sum_j D_{ij}\right)^2\sum_j M_{ij}^2\right\}.$$

Note that if ε is increased to such an extent that all pairs are considered matched, then $M = n_0 n_1$, $\bar{D} = \bar{y}_1 - \bar{y}_0$, and with some algebra it can be established that

$$\mathrm{var}(\bar{D}) = \frac{\sum (Y_{0j} - \bar{Y}_0)^2}{n_0^2} + \frac{\sum (Y_{1j} - \bar{Y}_1)^2}{n_1^2}.$$

This is asymptotically equivalent to the usual formula for the standard error of $\bar{y}_1 - \bar{y}_0$; to achieve exact equivalence in finite samples one could adjust the two preceding formulas by inserting the factor $n_i/(n_i - 1)$ after the first summation. The standard error may be used in the obvious manner to construct a confidence interval for the population matched difference in mean.

Similar remarks hold for the matched difference in probability. With

$$T_{ij} = (-1)^{1-i}\sum_{j'=1}^{n_{1-i}} \mathrm{sgn}(Y_{ij} - Y_{1-i,j'}) I\{D(\mathbf{X}_{ij}, \mathbf{X}_{1-i,j'}) \leqslant \varepsilon\},$$

$\bar{T} = \sum T_{0j}/M = \sum T_{1j}/M$, and its standard error is obtained by substituting T_{ij} for D_{ij} in the

formulas given above. In addition, with infinite tolerance,

$$\bar{T} = \frac{1}{n_0 n_1} \sum_i \sum_{i'} \operatorname{sgn}(Y_{1i} - Y_{0i'}) = \frac{U_1 - U_0}{n_0 n_1},$$

where U_0 and U_1 are the ordinary Mann–Whitney statistics for comparing the treated observations with the controls; then $\mathrm{SE}(\bar{T})$ is equivalent to the usual standard error for $\mathrm{pr}_{\mathrm{est}}(Y_1 > Y_0)$ calculated without the assumption that the treatment has no effect.

To illustrate the computations involved, consider the example of the preceding section. The values of M_{ij}, D_{ij} and T_{ij} are shown in the rightmost section of Table 1. For this analysis $M = 7$ is the total number of *between-sample* matched pairs, and the matched differences are $\bar{D} = 1906/7 = 272.29$ (in mean) and $\bar{T} = \frac{5}{7} = .71$ (in probability); the median matched difference is 286. The corresponding standard errors, from the formulas presented above, are $\mathrm{SE}(\bar{D}) = 101.86$ and $\mathrm{SE}(\bar{T}) = .41$. (The sample sizes are of course too small for these to be meaningful.)

Consider finally the situation where $k > 1$. It is always possible, and may be useful, to calculate matched differences for each pair of samples. Suppose first that the treatments have a natural ordering, such that a monotonic trend in the treatment effects might be expected. Then the (matched) partial correlation between Y and Z (= sample number) given $\mathbf{X}$ may be a suitable descriptive statistic; it is in fact a weighted average of the matched differences. The corresponding test of the null hypothesis that the treatments have no effect is formally the same as the test for no population matched partial correlation, as presented by Quade (1974). Supposing instead that the treatments have no ordering, it would be possible to perform a test by constructing a quadratic form in the matched differences, asymptotically distributed as χ^2 on k df under the null hypothesis, but the details will not be given here.

6. Matched-Pair Charts

A useful diagram for illustrating the comparison of treated and control observations is the 'pair chart' (Quade, 1973). To construct this chart, first draw a rectangle of width n_0 units and height n_1 units. Starting from its lower left corner, draw a line one unit to the right (upwards) if the smallest response in the combined samples is a control (treated) observation. Then, starting from the end of this line, draw another line to the right (upwards) if the second smallest response is a control (treated) observation. Continue through all the observations, thus producing a path to the upper right corner of the rectangle. Each unit square of the pair chart compares one control with one treated observation; the square is below (above) the path according as the control (treated) observation has the greater response.

By shading or blanking out the squares corresponding to unmatched pairs, an ordinary pair chart may be converted into a 'matched-pair chart' which takes account of the covariable. The matched difference in probability is then the difference between the proportions above and below the path among the remaining squares. In addition, if each unshaded square is labeled with the numerical value of the response difference for the pair of observations which it represents, the me(di)an matched difference is the me(di)an of the labels.

Figure 1 thus illustrates the matched differences for the example considered previously. It shows graphically how the treatment–control comparisons favor the treatment in a 9 to 3 (or 3 to 1) ratio when the covariable is ignored, and how this ratio becomes 6 to 1 when

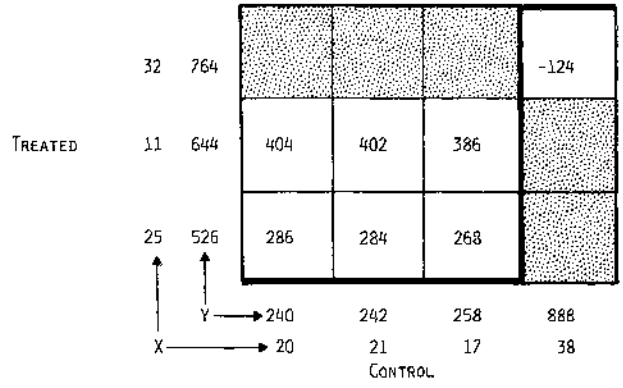

Figure 1. Matched-pair chart ($N_0 = 4$, $N_1 = 3$).

the covariable is taken into account; this description corresponds to the matched difference in probability. Also, a rough 'eyeball' estimate of the average of the displayed labels can be made: this approximates the me(di)an matched difference, which estimates the treatment effect if the populations are parallel.

7. Examples

As a first example consider the data on p. 386 of Snedecor and Cochran (1980). There are two groups of women, 11 from Iowa (arbitrarily called the 'controls') and 19 from Nebraska; the response Y is cholesterol concentration (mg/ml), and the covariable X is age (years). Snedecor and Cochran checked the classical assumptions of homoscedasticity and parallelism of regression lines for these data, and found no evidence of violation. The classical analysis of covariance then yields the point estimate 28.7 for the treatment effect (i.e. difference in cholesterol levels between the two States); the corresponding test statistic is $F = 3.00$ with (1, 27) df, and hence $P = .095$. For the analysis of covariance by matching, allowing a tolerance on age of $\varepsilon = 10$ years, the estimated treatment difference is 20.4; this corresponds to $F = 1.89$ with (1, 28) df, and hence $P = .180$. For the rank analysis of covariance by matching, $F = 1.58$ and $P = .220$. The corresponding matched-pair chart is Fig. 2. It shows 73 between-sample matched pairs, out of 209 total, among which 49 indicate higher cholesterol in Nebraska and only 24 in Iowa, so the matched difference in probability is $(49 - 24)/73 = .342$, with standard error .236; the median matched difference is 17. The matched difference in mean, or mean matched difference, is 28.1 ± 17.5. By any analysis, there is no significant difference between the States.

A second example involves data obtained by Cartwright, Lindahl and Bawden (1968) in their experiment to evaluate the effectiveness of topically-applied stannous fluoride (SF) and acid-phosphate fluoride (APF) in reducing the incidence of dental caries, as compared with a placebo treatment of distilled water (W). The data for the 69 female children who completed the two-year study are given in Table 2, in which B and A represent the number of decayed, missing or filled teeth (DMFT) before and after the study, respectively; the response to be analyzed is the increment $Y = A - B$. (It is noted that one increment is negative, which must indicate some observational or recording error, but no correction will be made.) Covariables to be taken into account are first, B, and second, X, the age in years at the beginning of the study. A third possible covariable, with values 1, 2

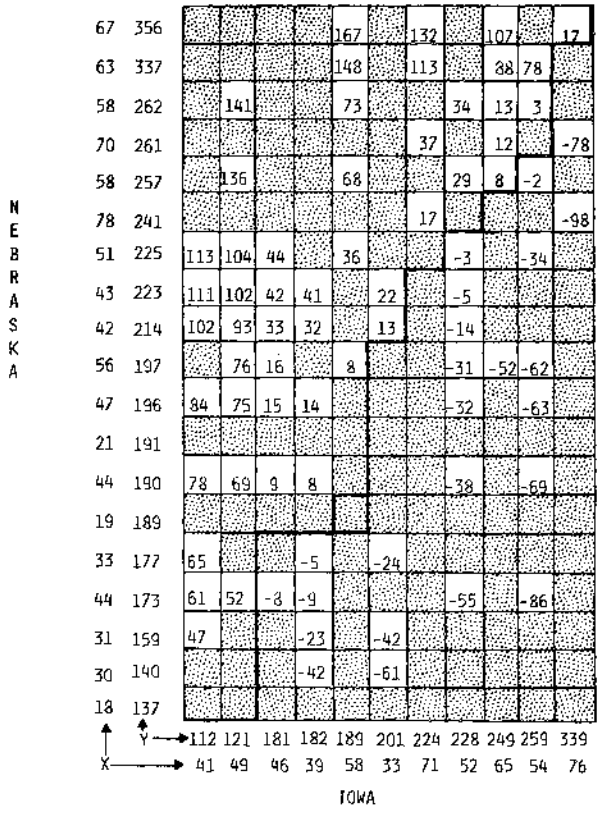

Figure 2. Matched-pair chart for comparing cholesterol levels in women from Iowa and Nebraska, taking account of age.

and 3 representing three institutions where the children resided, is also presented, but this variable will not be used, although it is as easy to match on as any other (whereas for the standard analysis it would have to be replaced by two dummy variables).

The results for this example are as follows. By the standard analysis of covariance, the adjusted treatment effects are -1.01 for SF and -1.73 for APF; the F statistic is 10.12 on (2, 64) df, giving $P < .001$. Allowing tolerances of one year on age and one unit on initial DMFT, analysis of covariance by matching yields point estimates of $-.61$ and -1.07 for the treatment effects, and $F = 6.04$ on (2, 66) df, with $P = .004$; the corresponding rank analysis gives $F = 4.04$, with $P = .022$. Since the treatments are unordered, no overall test is performed, but the pairwise matched differences are presented:

Pair	In mean	In probability
SF − W	$-1.12 \pm .44$	$-.58 \pm .23$
APF − W	$-1.65 \pm .52$	$-.65 \pm .20$
APF − SF	$-.56 \pm .43$	$-.18 \pm .26$

This example illustrates, by the way, that matched differences are not generally transitive. Note that the overall conclusions are essentially the same by any of the analyses: a significant reduction in caries associated with fluoride treatment, and a suggestion of greater reduction with APF than with SF.

Table 2
An experiment to evaluate the effectiveness of fluoride in preventing caries

Serial number	Institution	Age X (yrs)	DMFT B	DMFT A	Group	Serial number	Institution	Age X (yrs)	DMFT B	DMFT A	Group
1	1	13	7	11	W	36	2	16	8	8	APF
2	1	17	20	24	W	37	3	8	2	4	W
3	1	16	21	25	W	38	3	16	13	18	W
4	1	13	1	2	W	39	3	14	9	12	W
5	1	10	3	7	W	40	3	16	15	18	W
6	1	17	20	23	W	41	3	12	13	17	W
7	1	13	9	13	W	42	3	8	2	5	W
8	1	9	2	4	W	43	3	14	9	12	W
9	1	14	11	13	SF	44	3	9	4	6	SF
10	1	14	15	18	SF	45	3	15	10	14	SF
11	1	11	7	10	APF	46	3	14	7	11	SF
12	1	15	17	17	APF	47	3	13	14	15	SF
13	1	11	9	11	APF	48	3	12	7	10	SF
14	1	7	1	5	APF	49	3	12	3	6	SF
15	1	11	3	7	APF	50	3	14	9	12	SF
16	2	16	10	14	W	51	3	13	8	10	SF
17	2	16	13	17	W	52	3	14	19	19	SF
18	2	7	3	4	W	53	3	14	10	13	SF
19	2	11	4	7	W	54	3	14	10	12	APF
20	2	15	4	9	W	55	3	11	7	11	APF
21	2	14	15	18	SF	56	3	14	13	12	APF
22	2	11	6	8	SF	57	3	9	5	8	APF
23	2	9	4	6	SF	58	3	11	1	3	APF
24	2	17	18	19	SF	59	3	12	8	9	APF
25	2	14	11	12	SF	60	3	14	4	5	APF
26	2	13	9	9	SF	61	3	10	4	7	APF
27	2	9	4	7	SF	62	3	12	14	14	APF
28	2	9	5	7	SF	63	3	11	8	10	APF
29	2	15	11	14	SF	64	3	11	3	5	APF
30	2	10	4	6	SF	65	3	16	11	12	APF
31	2	10	4	4	APF	66	3	15	16	18	APF
32	2	15	7	7	APF	67	3	10	8	8	APF
33	2	11	0	4	APF	68	3	6	0	1	APF
34	2	9	3	3	APF	69	3	7	3	4	APF
35	2	9	0	1	APF						

8. Discussion

It should be pointed out that the word 'analysis' in the title of this paper is not just part of the stock phrase 'analysis of covariance'. By Assumption I, attention has been restricted to the analysis of data already in hand rather than to the method of sampling. The question arises, however, as to what could be done if the samples were matched by design. The methods of this paper can be applied in such a situation, but with subtle differences in interpretation, particularly concerning the population about which inference can be made. Whether one *should* choose matched samples is another question, the answer to which depends in part on the costs involved; for some recent work on this, using the matched difference in probability, see the Ph.D. thesis of J. R. Schoenfelder (at the University of North Carolina at Chapel Hill, 1981). Schoenfelder also investigates the asymptotic

relative efficiency of the corresponding test when the covariable is discrete, and suggests a modification to improve it.

Another question which has not been considered in this paper concerns the decision as to whether a pair is matched or not. This might be regarded as a substantive question, rather than a statistical one: the statistician may call two observations matched if the investigator familiar with the subject matter area does so. If ties on the covariable are plentiful, one may as well let the tolerance be zero. Otherwise, the proper choice is a compromise: too small a tolerance leaves an insufficient number of matched pairs for analysis, while too large a tolerance makes only an incomplete adjustment for the covariable. Some researchers might prefer a generalized procedure in which pairs are not dichotomously classified as 'matched' (and hence used) or 'not matched' (and hence discarded), but are weighted according to their closeness with respect to the covariable. Thus, one might want to give a pair of observations $(Y_1, \mathbf{X}_1)$ and $(Y_2, \mathbf{X}_2)$, with $D(\mathbf{X}_1, \mathbf{X}_2) = d$, the weight $w(d) = \exp(-d)$, say. This would cause no mathematical difficulty whatever: simply substitute $w(d)$ for $I\{d \leqslant \varepsilon\}$, which is just a special case of $w(d)$, in all the defining formulas of this paper.

Acknowledgements

This research was supported in part by Grant HS04175 from the National Center for Health Services Research. Thanks are due to Vance Cartwright for permission to use the dental data presented in §7.

Résumé

On considère le problème de la comparaison de traitements pour une réponse Y quand une covariable X est prise en compte. Plusieurs méthodes utilisant l'appariement sont en fait des compromis entre l'analyse de covariance classique et l'analyse de paires indépendantes classique. D'abord, aucune nécessité de se limiter aux paires indépendantes, mais toutes peuvent être introduites. Alors, si X est une variable concomitante, c'est-à-dire si sa distribution est la même quel que soit le traitement, on peut employer des méthodes fondées en fin de compte sur la randomisation, bien qu'elles soient en pratique fondées sur l'analyse de variance. Quand X n'est pas une variable concomitante, on peut appliquer des méthodes proches de la corrélation partielle (entre Y et le traitement, à X donnée). Pour toutes ces méthodes, on peut utiliser les valeurs Y ou leurs rangs.

References

Cartwright, H. V., Lindahl, R. L. and Bawden, J. W. (1968). Clinical findings on the effectiveness of stannous fluoride and acid phosphate fluoride as caries reducing agents in children. *Journal of Dentistry for Children* **35,** 36–40.

Johnson, R. E., Quade, D. and Langston, R. D. (1981). The matched-pair partial correlation procedure: a user's guide to PROC MATPAR. SAS Technical Report S-120. Cary, North Carolina: SAS Institute.

Quade, D. (1965). On analysis of variance for the k-sample problem. *Annals of Mathematical Statistics* **37,** 1747–1758.

Quade, D. (1967). Rank analysis of covariance. *Journal of the American Statistical Association* **62,** 1187–1200.

Quade, D. (1973). The pair chart. *Statistica Neerlandica* **27,** 29–45.

Quade, D. (1974). Nonparametric partial correlation. In *Measurement in the Social Sciences: Theories and Strategies*, H. M. Blalock, Jr (ed.), 369–398. Chicago: Aldine.

Sen, P. K. (1960). On some convergence properties of U-statistics. *Calcutta Statistical Association Bulletin* **10,** 1–18.

Snedecor, G. W. and Cochran, W. G. (1980). *Statistical Methods*, 7th ed. Ames: Iowa State University Press.

Received May 1981; *revised October* 1981

APPENDIX

Finding the Randomization Confidence Distribution for the Treatment Effect ($k = 1$)

Given n_0 control and n_1 treated observations, let Q be any test criterion for the null hypothesis of no difference attributable to the treatment. The randomization test involves calculation of Q for each hypothetical data set obtainable by reallocating the observations to the samples. Its P-value is then equal to $(1+A)/C$, where A is the number of reallocations producing a hypothetical Q as extreme as that observed, or more so. There are $C-1$ reallocations, where $C = N!/n_0!\, n_1!$ and $N = n_0 + n_1$. These can be put in one–one correspondence with the nontrivial permutations, say $\mathbf{v} = (v_1, \ldots, v_N)$, of $\mathbf{u} = (u_1, \ldots, u_N)$, where the observations are indexed by the single subscript $i = 1, \ldots, N$, and $u_i = 0$ or 1 according as (Y_i, X_i) is a control or is treated. To construct a confidence interval requires testing the hypothesis $H(\delta)$ that the treatment effect has the value δ; but this is equivalent to testing the null hypothesis after first substracting δ from each treated observation. So let $Q(\mathbf{v}, \delta)$ be the value of Q for testing $H(\delta)$ calculated from the hypothetical data set obtained by that reallocation of observations which corresponds to $\mathbf{v}$. Then A is the number of times that $Q(\mathbf{v}, \delta) \geqslant Q(\mathbf{u}, \delta)$, the value observed in the actual dataset.

In analysis of covariance by matching, after subtracting δ from the observations which were treated, the ith score is

$$(Y_i - u_i\delta) - \frac{\sum_j (Y_j - u_j\delta) I\{D(X_i, X_j) \leqslant \varepsilon\}}{\sum_j I\{D(X_i, X_j) \leqslant \varepsilon\}}.$$

Recall that

$$\hat{Y}_i = \sum_j Y_j I\{D(X_i, X_j) \leqslant \varepsilon\}/M_i,$$

where $M_i = \sum_j I\{D(X_i, X_j) \leqslant \varepsilon\}$ is the number of observations (including itself) which match (Y_i, X_i), and define

$$t_i = \sum_j u_j I\{D(X_i, X_j) \leqslant \varepsilon\}/M_i,$$

the proportion treated among the observations which match (Y_i, X_i). Then the ith score can be re-expressed as

$$(Y_i - u_i\delta) - (\hat{Y}_i - t_i\delta).$$

Now, a convenient test criterion, equivalent to Student's t statistic calculated from the scores, is the sum of the scores for the treated observations. In the hypothetical dataset obtained from the reallocation corresponding to $\mathbf{v}$, this criterion is

$$\begin{aligned} Q(\mathbf{v}, \delta) &= \sum_i v_i\{(Y_i - u_i\delta) - (\hat{Y}_i - t_i\delta)\} \\ &= \sum_i v_i(Y_i - \hat{Y}_i) - \delta \sum_i v_i(u_i - t_i), \end{aligned}$$

and $Q(\mathbf{v}, \delta) \geqslant Q(\mathbf{u}, \delta)$ if and only if

$$\sum_i (v_i - u_i)(Y_i - \hat{Y}_i) - \delta \sum_i (v_i - u_i)(u_i - t_i) \geqslant 0,$$

or

$$\delta \leqslant \frac{\sum_i (u_i - v_i)(Y_i - \hat{Y}_i)}{\sum_i (u_i - v_i)(u_i - t_i)} = \delta(\mathbf{v}), \quad \text{say.}$$

Thus, A is the number of boundary points $\delta(\mathbf{v})$ which are greater than or equal to the hypothesized effect; the P-value is $1/C$ for δ greater than the maximum $\delta(\mathbf{v})$, $2/C$ in the gap between the maximum and the next largest value of $\delta(\mathbf{v})$, and so on; this establishes the confidence distribution.

References

On Analysis of Variance for the K-Sample Problem

Dana Quade

The Annals of Mathematical Statistics, Vol. 37, No. 6. (Dec., 1966), pp. 1747-1758.

Stable URL:

http://links.jstor.org/sici?sici=0003-4851%28196612%2937%3A6%3C1747%3AOAOVFT%3E2.0.CO%3B2-A

Rank Analysis of Covariance

Dana Quade

Journal of the American Statistical Association, Vol. 62, No. 320. (Dec., 1967), pp. 1187-1200.

Stable URL:

http://links.jstor.org/sici?sici=0162-1459%28196712%2962%3A320%3C1187%3ARAOC%3E2.0.CO%3B2-3