*Oceanogr. Mar. Biol. Annu. Rev.*, 1993, **31**, 177–221
A. D. Ansell, R. N. Gibson and Margaret Barnes, *Editors*
UCL Press

# USING FIXED-EFFECTS MODEL MULTIVARIATE ANALYSIS OF VARIANCE IN MARINE BIOLOGY AND ECOLOGY

CRAIG R. JOHNSON[1,2] and CHRISTOPHER A. FIELD[3]
[1]*Marine Ecology Laboratory, Department of Fisheries and Oceans, Bedford Institute of Oceanography, Dartmouth, Nova Scotia, Canada B2Y 4AZ*
[2]*Department of Zoology, University of Queensland, St. Lucia, Queensland 4072, Australia (address for reprints)*
[3]*Department of Mathematics, Statistics and Computing Science, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5*

ABSTRACT The robustness and power of four commonly used MANOVA statistics (the Pillai-Bartlett trace ($V$), Wilks' Lambda ($W$), Hotelling's trace ($T$), Roy's greatest root ($R$)) are reviewed and their behaviours demonstrated by Monte Carlo simulations using a one-way fixed effects design in which assumptions of the model are violated in a systematic way under different conditions of sample size ($n$), number of dependent variables ($p$), number of groups ($k$), and balance in the data. The behaviour of Box's $M$ statistic, which tests for covariance heterogeneity, is also examined. The behaviours suggest several recommendations for multivariate design and for application of MANOVA in marine biology and ecology, viz. (1) Sample sizes should be equal. (2) $p$, and to a lesser extent $k$, should be kept to a minimum insofar as the hypothesis permits. (3) Box's $M$ statistic is rejected as a test of homogeneity of covariance matrices. A suitable alternative is Hawkins' (1981) statistic that tests for heteroscedasticity and non-normality simultaneously. (4) To improve agreement with assumptions, and thus reliability of tests, reduction of $p$ (e.g. by PCA or MDS methods) and/or transforming data to stabilise variances should be attempted. (5) The $V$ statistic is recommended for general use but the others are more appropriate in particular circumstances. For Type I errors, the violation of the assumption of homoscedasticity is more serious than is non-normality and the $V$ statistic is clearly the most robust to variance heterogeneity in terms of controlling level. Kurtosis reduces the power of all statistics considerably. Loss of power is dramatic if assumptions of normality and homoscedasticity are violated simultaneously. (6) The preferred approach to multiple comparison procedures after MANOVA is to use Bonferroni-type methods in which the total number of comparisons is limited to the fewest possible. If all possible comparisons are required an alternative is to use the $V$ statistic in the overall test and the $R$ statistic in a follow-up simultaneous test procedure. We recommend following a significant MANOVA result with a canonical discriminant analysis. (7) Classical parametric MANOVA should not be used with data in which high levels of variance heterogeneity cannot be rectified or in which sample sizes are unequal and assumptions are not satisfied. We discuss briefly alternatives to parametric MANOVA.

# INTRODUCTION

## GENERAL

Marine biologists and ecologists frequently need to compare among groups in which there are several response variables. Examples could range from examining the response of a multispecies community to particular treatments, which includes examining multispecies responses to environmental impacts by testing for a significant interaction effect in a Before-After/Control-Impact (BACI) design (see Stewart-Oaten *et al.*, 1986); to investigating multiple physiological responses to treatments; to comparing several physical parameters of vocalisation among designated behaviour contexts; to analysing univariate repeated-measures data in which the repeated measures (or successive differences between them) are treated as separate response variables (see Barker & Barker, 1984). Inferential statistics are now fundamental to ecological methodology and the analysis of variance (ANOVA) is arguably the most widely applied parametric technique. When hypotheses focus on multivariate responses the appropriate parametric test is often the multivariate analysis of variance (MANOVA). Independent univariate ANOVAs could be conducted on each of $p$ response variables after adjusting the nominated significance level ($\alpha$) to control for compounding of Type I error ($p$ independent univariate ANOVAs will inflate $\alpha$ to $\alpha_{\text{new}}$ where $\alpha_{\text{new}} = 1\text{-}(1\text{-}\alpha)^p$). However, although $\alpha$ can be controlled by applying the Bonferroni adjustment ($\alpha_{\text{adj}} = \alpha/p$; see Harris, 1985), this reduces power considerably when $p$ is large. Further problems associated with conducting several independent ANOVAs are that information is lost if there are interactions among variates, which is usually the case in ecology, and that groups which separate clearly in multidimensional space may overlap and not be distinct when single dimensions are considered in isolation.

This paper is intended as a practical guide to using fixed effects model (Model I) MANOVA. For those without a statistical background, a glossary is given in Appendix 1. Our recommendations are based on the robustness and power of the commonly used MANOVA statistics. This approach reflects the viewpoint that although it is inappropriate to ignore assumptions about the distributions of statistics, since many test statistics are robust to some violations of their theoretical requirements, it is equally improper to invariably reject a test as invalid if its underlying assumptions are not met exactly. Because biological data rarely conform to the theoretical requirements of the tests used on them, it becomes essential to know whether tests are robust, and how to describe raw data to make them suitable to use in statistical tests.

In contrast to univariate ANOVA, in which the robustness and power of the $F$-statistic are well understood (see comprehensive review by Glass *et al.*, 1972; also Scheffe, 1959; Srivastava, 1959; Tiku, 1971; Ito, 1980; Underwood, 1981), these properties of the several MANOVA statistics that are multivariate generalisations of the $F$-statistic are less well known. Some standard multivariate texts do not, or barely, broach the subject at all (e.g. Morrison, 1976; Srivastava & Carter, 1983; Hair *et al.*, 1987; Krzanowski, 1988), some extrapolate inductively from the univariate case (e.g. Cooley & Lohnes, 1971; Press, 1972), while others discuss broadly the kinds of violations likely to be most serious but do not compare the different statistics (Marriot, 1974). More recent texts (e.g. Green, 1979; Barker & Barker 1984; Harris, 1985; Stevens, 1986;

Tatsuoka, 1988; Tabachnick & Fidell, 1989) offer recommendations, which vary in depth and opinion, based on results of studies that have addressed the question specifically. This is due in part to the paucity and relative recentness of such studies; the problem has been addressed comprehensively only by Olson (1974), although there have been limited investigations by Ito & Schull (1964), Schatzoff (1966), Mardia (1971), Korin (1972), Bird & Hadzi-Pavlovic (1983), and others. Much of this work was reviewed by Olson (1976) and, in a more technical account, by Ito (1980), but both focused on the question of appropriate choice of test statistic and did not attempt to provide overall guidelines for all steps from initial design through to completed MANOVA, nor did they present an ecological or biological perspective. Barker & Barker's (1984) text is broader in scope (but not biological) and is a useful companion to this article.

## THE APPROACH

Here we review the robustness and power of the four most widely used MANOVA statistics, viz. the Pillai-Bartlett trace ($V$), Hotelling's trace ($T$), Wilks' Lambda ($W$), and Roy's largest root ($R$) (see Appendix 2). Because of the technical and often piecemeal nature of published accounts, we present results of a Monte Carlo study of a one-way fixed-effects (Model I) multivariate model to demonstrate the behaviour of the statistics in terms of their power and rates of Type I error when underlying assumptions are violated in a systematic way under certain conditions of sample size ($n$), number of groups ($k$), number of variables ($p$), and balance in the data. We also investigate the behaviour of Box's $M$ statistic (Box, 1949; see also Cooley & Lohnes, 1971; Morrison, 1976), which is designed to test the assumption of homoscedasticity among dispersion matrices, i.e. $M$ is a multivariate analogue of Bartlett's test for equality of variances in the univariate case. Our own and published results lead us to construct guidelines for applying MANOVA to ecological and biological data in which we present (1) conclusions about the most reliable test statistics, (2) recommendations for multivariate experimental design, (3) a statistic that tests for non-normality and heteroscedasticity simultaneously and which is better suited to this purpose than is the $M$ statistic, (4) suggestions for reducing dimensionality and transforming data so that tests may be conducted reliably, and (5) methods for multiple range tests (comparing among multivariate means) after MANOVA. Although discussion is focused on the one-way model, most of our comments and conclusions will hold for multiway problems and for general regresssion models of which analysis of variance is a special case.

## ASSUMPTIONS OF THE MANOVA MODEL

Assuming correct experimental design with replication and independent sampling (e.g. see Sokal & Rohlf, 1981; Steel & Torrie, 1981; Hurlbert, 1984; Stewart-Oaten *et al.*, 1986; Underwood, 1990), the critical assumptions of univariate ANOVA are a normal distribution of the error terms about a mean of zero, and identical group variances. The corresponding assumptions in the multivariate case are multinormality of error terms and homogeneity among group covariance (or dispersion) matrices (if there are $p$ variates then the dispersion matrix is a symmetrical $p \times p$ matrix in which the variances are the diagonal elements, and the covariances the off-diagonal elements). To state this more formally, consider a one-way fixed effects model in which there are $k$ groups with $n_i$ observations on the $i^{th}$ group. The $j^{th}$ observation in group $i$, $\mathbf{y}_{ij}$, is a vector of length $p$, where $p$ is the number of response variates. The model can be written as:

$$\mathbf{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \mathbf{e}_{ij}$$

where $i = 1, \ldots k$ and $j = 1, \ldots n_i$, $\boldsymbol{\mu}$ is the grand mean vector which defines the centroid, $\boldsymbol{\alpha}_i$ is the group effect vector or fixed deviation of group $i$ from the grand mean, and $\mathbf{e}_{ij}$ is the vector of error terms or random deviations from the expected value of $\boldsymbol{\mu} + \boldsymbol{\alpha}_i$.

In the fixed effects model the test is $H_o$: $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \ldots = \boldsymbol{\alpha}_k$ versus $H_1$:$\boldsymbol{\alpha}_i$s not all equal. In carrying out the test the assumption is made that the $\mathbf{e}_{ij}$s are independent and identically distributed and have a multivariate normal distribution with mean 0 and covariance $\Sigma$ (note that there is no requirement that the $p$ variates of any single multivariate observation $\mathbf{y}_{ij}$ are independent). In a more general design the assumption on the errors remains the same, but $\boldsymbol{\alpha}_i$ will be replaced by a more complicated form. For example, in a two-way model with interaction we have:

$$\mathbf{y}_{ijm} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_m + (\boldsymbol{\alpha\beta})_{im} + \mathbf{e}_{ijm}$$

where the $\mathbf{e}_{ijm}$s satisfy the above assumption. Note that normality is required for the distribution of the test statistic, but it is not needed to justify the use of least squares fitting. In contrast, homogeneity of variances is required for the justification of least squares which is based on equal precision of observations.

## SIMULATION METHODS

### GENERAL

For the simulations, random data were drawn from $N(0,1)$ populations (i.e. normal about a mean of zero and with unit variance) generated using the IMSL package (routine GGNSM; start seed set by function of the computer clock), and the test statistics were calculated using the SPSS 9 software package. Each simulation consisted of 200 runs which provided sufficient resolution to give unambiguous results (see Appendix 3). In all cases the testing was carried out with a nominal level (probability of Type 1 error) of $\alpha = 0.10$.

Since the MANOVA test statistics are affine-invariant, i.e. they behave identically whether covariance matrices are expressed in raw or canonical form (proof in Appendix 4), group covariance structure was set in canonical form. Thus, the covariance matrix of any group with $p$ variables was set initially as the $p \times p$ identity matrix, $\mathbf{I}$, where

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & . & . & 0 \\ 0 & 1 & 0 & . & . & 0 \\ 0 & 0 & 1 & . & . & 0 \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & 1 \end{bmatrix}$$

### HETEROSCEDASTICITY

Variance heterogeneity ($z$) was introduced into only one of the $k$ groups in any one design, and into all $p$ variables of that group, so that the covariance structure of the 'contaminated' group was given by $z\mathbf{I}$, where $z$ is a scalar $\geq 1$. The consequences of heteroscedasticity of this pattern (i.e. concentrated with respect to group, but diffuse with respect to dimension) are worse than when heterogeneity is concentrated in less than $p$ dimensions, but less than when heterogeneity occurs in a greater number of groups (Korin, 1972; Olson, 1974). The magnitude of heterogeneity was set sequentially at $z = 1$ (i.e. data homoscedastic), and then 5, 10, 15, 20, 30, 40, and 50. In examining power, only homoscedastic ($z = 1$) and mildly heteroscedastic data ($z = 20$) were used.

### NORMALITY

Since the marginal densities are normal and the covariance matrix is the identity, the joint density is multivariate normal (Anderson, 1984). Non-normality was generated using a normal mixture of $N(0,1)$ probability $= 0.8$ and $N(0,9)$ probability $= 0.2$, and therefore our consideration of non-normality included the component of kurtosis, but not skewness. When examining the effect of non-normality all variables in all groups were generated from this normal mixture.

It is noted that our simulations of non-normality may be conservative in that effects of skewness may be worse than those of kurtosis. However, since skewness usually manifests as strong heterogeneity and effects of heterogeneity are much stronger than those of non-normality, the hiatus in not considering skewness is unlikely to affect conclusions.

### NON-CENTRALITY

Except in producing the power curves, the population means of all $p$ variables in all $k$ groups were zero (since data were from $N(0,1)$ or $N(0,9)$ populations). In generating the power curves, non-centrality was introduced by increasing the population means of all variables in only one of the $k$ groups by increments of 0.2 to a maximum of 2.0, i.e. the distribution of non-centrality was concentrated in one group but affected all dimensions (since the centroids, or multivariate means, of the groups would align along a single dimension in multivariate space, this is a type of concentrated non-centrality). This procedure is simpler than the usual way of determining power functions in which a standardised measure of the distance between group means is described by a non-centrality parameter (e.g. see Schatzoff, 1966; Lee, 1971; Olson, 1974; Morrison, 1976). We include the results on power from these studies in our discussion of the effect of violation of assumptions. Unless stated otherwise, variance heterogeneity (when present) and non-centrality were coincident in the same group.

### SAMPLE SIZE

Sample size was usually $n = 10$ observations per variable unless specified differently. In unbalanced designs, the sample size of one group was half that of the other $k$-1 groups which were equal, e.g. 5, 10, 10, 10,... or 10, 20, 20,...

### A NOTE ON SIMULATION STRUCTURE

There are two principal reasons why multivariate simulations are not straightforward. First, many of the parameters (e.g. variance heterogeneity and noncentrality) can be defined in different ways and their definition depends on subjective judgements by the experimenter. Second, in systematically varying a particular parameter it is extremely difficult to avoid confounding the effects of the parameter of interest with other factors, e.g. changing the number of groups, or number of variables, or number of replicates, also changes the error degrees of freedom ($df_{\text{error}}$) which will in itself affect the power of a test. Thus, if the number of groups is changed, to keep $df_{\text{error}}$ constant some other parameter must be changed to counter the effect of changing the number of groups, so the problem of confounding is not avoided. Clearly, the experimenter must make important decisions about how to structure simulations, and results need to be interpreted cautiously.

We used the ratio of maximum variance to minimum variance as a measure of heteroscedasticity. This quantity is computed easily and is more intuitive than alternatives based on the sum of eigenvalues of covariance matrices. Thus, for a particular amount of heterogeneity the amount of variance contamination in the contaminated group was fixed regardless of the number of dimensions or groups, e.g. we consider that the amount of heterogeneity in (**I**,**I**,10**I**) and in (**I**,**I**,**I**,**I**,**I**,**I**,**I**,**I**,10**I**) to be equivalent. While it may be argued that in this example the amount of heterogeneity is different, it was our judgement that the experimenter most often considers the ratio of maximum to minimum variance as a 'natural' measure, and hence our graphs use this ratio on the $x$-axis.

Our simulations with respect to power were not intended to be exhaustive but were kept simple deliberately to give a comparison of the four statistics under simple deviations from the null hypothesis. The primary concern was to find a procedure which gave the most robust behaviour under deviations from the model and then to undertake some calculations on power to cover common cases to ensure that power is not compromised badly. It is clear that power is affected by distances between centroids, degrees of freedom, and the relative magnitude of the error variation. In our simulations these effects are confounded in that we have not kept $df_{\text{error}}$ constant, but we emphasise that an ecologist would not strive for constant $df_{\text{error}}$. For example, having selected $p$ appropriate response variables and (usually) the maximum number of replicates ($n$) that is practicable, if an experimenter was to change the number of treatment groups, it is unlikely that $p$ or $n$ would be changed simply to achieve constant $df_{\text{error}}$. However, there is a need to undertake a more exhaustive study of power with the most important factors controlled in an orthogonal design.

## ROBUSTNESS OF MANOVA STATISTICS

The effects of violations of the assumptions of covariance homogeneity and multinormality on the rates of Type I error and power of the four MANOVA statistics are summarised in Table I. Overall, heterogeneous covariance structure is the most serious violation. Moderate levels of heterogeneity reduces significantly the power of all four statistics, and inflates dangerously the rates of Type I error of Roy's $R$, Wilk's $W$, and Hotelling's $T$ statistics. The deleterious

TABLE I

Summary of properties of MANOVA statistics $V, T, W$ and $R$. Abbreviations are ER = exceedance rates, df = degrees of freedom, $n$ = sample size, $p$ = number of dimensions, $k$ = number of groups. Numbers in parentheses refer to references: (1) Simulations in this paper; (2) Bird & Hadzi-Pavlovic 1983; (3) Ito & Schull, 1964; (4) Korin 1972; (5) Lee 1971; (6) Mardia 1971; (7) Olson 1974; (8) Olson 1976; (9) Pillai & Jayachandran 1967; (10) Schatzoff 1966. Note that some statements are composed from several sources, and thus each work cited may not support every aspect of the statement with which it is associated.

|  | Property of statistic |  |
|---|---|---|
| Violation | Rate of Type I error | Power |
| No violation | Occurs at nominated level, e.g. $\alpha = 0.05$ | Differences in power among statistics are small, especially for large samples (1,5,7,9,10). $V$, $W$ and $T$ are equivalent for very large sample sizes (1,7,10); empirically when $(df_{\text{error}}/10p) \geq df_{\text{hypothesis}}$ (8)<br>$V \geq W \geq T \geq R$ for diffuse noncentrality, i.e. when most groups differ from each other in most dimensions (2,5,7,8,9; noncentrality is a measure of the difference among group mean vectors)<br>$R \geq T \geq W \geq V$ for concentrated noncentrality, i.e. when noncentrality occurs in only one group (1,5,7,8,9)<br>For constancy of standardised noncentrality parameter, power tends to decrease (slightly) with increasing $p$ and $k$ (5,7)<br>For constancy of magnitude of difference in means in one group from all others, power tends to increase (slightly) with increasing $p$ and $k$ (1) |
| Heteroscedasticity<br>(= variance heterogeneity) | $V$ robust to moderate levels of heterogeneity in balanced data (1,2,7) | Moderate levels of heterogeneity reduces greatly power of all statistics, but for balanced data $V$ is least affected (1,7) |

TABLE I—*Continued*

| Violation | Property of statistic | |
|---|---|---|
| | Rate of Type I error | Power |
| | None of the statistics is robust if design is unbalanced (1,3). If heteroscedasticity is in group with smallest $n$, rates of Type I error are liberal (1,3,7) but if heteroscedasticity is in group with largest $n$ then rates of Type I error become conservative (3) | Differences in lower end of power curves reflect effect of heterogeneity on Type I error rates, thus, for $V$, lower end of curve is least affected, and for $R$, $T$ and $W$ upper end is least affected (1,7) |
| | $W$, $T$ and $R$, in order of increasing liberality, are poorly robust for small $n$ even in balanced designs (1,2,4,7) | ER at bottom end of curves of all statistics increase greatly with heterogeneity if data are unbalanced, reflecting effect of heterogeneity on Type I error rates (1) |
| | $W$, $T$ and $V$ are equivalent for large $n$ and thus robust for large $n$ if data are balanced (1,3,7,8,10) | Magnitude of deleterious effects are smaller (but still serious) when heterogeneity and noncentrality are not coincident with respect to group (1,7) |
| | Liberal ER of $W$, $T$ and $R$ with heterogeneity are exacerbated by increasing $p$ and $k$ (1,2,3,4,7). For $V$, ER increase slightly with $p$ when $p > k$ (1), and increase with $k$ approximately when $k > p$ but decrease with increasing $k$ approximately when $k < p$ (7) | |
| | Deleterious effects on Type I error are greatest when distribution of heterogeneity is diffuse with respect to dimension and concentrated with respect to group (4,7) | |
| Non-normality | All statistics are robust; ER are conservative, but the effect is not serious (1,6,7) | Power of all statistics is reduced considerably and by similar magnitude (1,7) |
| | | Conservative ER when there are no differences among group means reflect effect on Type I error (1) |
| Heteroscedasticity and non-normality (balanced data) | Effects are similar to heteroscedasticity in normal data (1,7) | Perturbation of power curves is extreme for all statistics (1,7) |
| | | ER are relatively invariant with increasing noncentrality, and are set by the effect of heterogeneity on Type I error rates (1) |

effect on Type I error rates of $R$, $W$ and $T$ is made much worse by unequal sample sizes and increasing numbers of dependent (= response) variables. In contrast, Pillai's $V$ statistic is much more robust to heterogeneity, and Type I error rates increase only slightly (but consistently) with dimensionality only when (approximately) $p > k$ for equal sample sizes. However, in line with the others, the $V$ statistic also performs badly with respect to Type I error rates if covariance matrices are heteroscedastic and sample sizes are unequal. In unbalanced designs, if heterogeneity occurs in the group with smallest $n$, then level becomes excessively liberal, but if heterogeneity occurs in the largest group, rates of Type I error are made conservative. Non-normality (kurtosis) has a much less serious impact on rates of Type I error, which are made more conservative. However, kurtosis reduces severely the power of all statistics, and if the assumptions of multinormality and homogeneous covariance matrices are both violated, power is reduced to negligible levels. A detailed account and illustration of the behaviours of the statistics follows.

## TYPE I ERROR

### *Summary*

For most of the statistics, with the exception of Pillai's $V$, rates of Type I error increase to undesirable levels when data are heteroscedastic and the number of dependent variables is large. The harmful effect of variance heterogeneity is made worse if sample size is small and especially if sample sizes are unequal.

### *Effect of heteroscedasticity*

Exceedance rates of the $T$, $W$ and $R$ statistics are affected adversely by variance heterogeneity unless the following conditions are met: (i) if sample size is small, the number of groups ($k$) and variables ($p$) must be small, and the data balanced (cf. bottom left graphs in Figs 1, 2 and 3), or (ii) if $p$ and $k$ are large, the data must be balanced, and sample size large ($n = 50$ in Fig 4). For these statistics, as dimensionality ($p$) increases, rates of Type I error rise rapidly with increasing heteroscedasticity, even with balanced data (Figs 1 and 2). Roy's greatest root ($R$) is the least robust.

The undesirable effects of heteroscedasticity are countered in part by increasing (but maintaining equal) sample sizes (Fig 4). The effect of increasing $n$ in reducing the likelihood of producing too many significant results is greatest for $T$ and $W$, and for $n = 50$ their exceedance rates are almost identical to those of the more robust $V$ statistic. This is because $V$, $W$ and $T$ are asymptotically equivalent for very large samples (Schatzoff, 1966; Olson, 1974). From his simulations, Olson (1976) suggested that the three may be considered equivalent for $n$ sufficiently large that $(df_{\text{error}}/10p) \geq df_{\text{hypothesis}}$, where $df$ = degrees of freedom.

For $T$, $W$ and $R$, increasing the numbers of groups is much less serious than increasing $p$, but when dimensionality is low ($p = 2$), increases in $k$ result in unacceptably high rates of Type I error at moderate to high levels of variance heterogeneity (approximately $z \geq 20$; Figs 1 and 2). Increasing the number of groups when $p \geq 5$ has virtually no effect on the rates of Type I error of the $W$ and $T$ statistics, but at this level of $p$ exceedance rates are much greater than the nominated significance level anyway. Exceedance rates for Roy's $R$ statistic

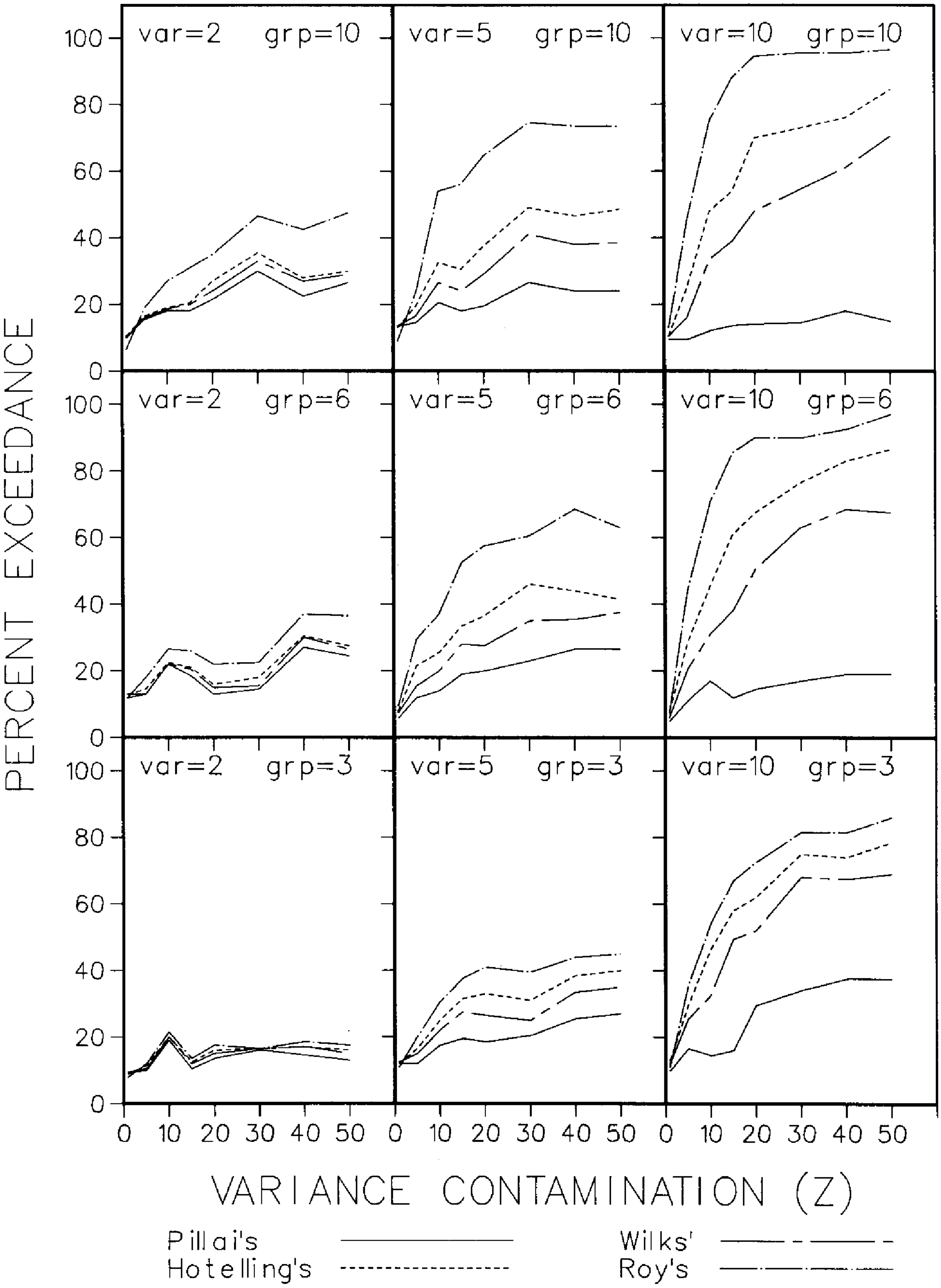


Fig 1.—Effect of heteroscedasticity on rates of Type I error in MANOVA statistics under conditions of different numbers of groups (grp) and variables within groups (var), when data are balanced and normal. In all cases $n = 10$.

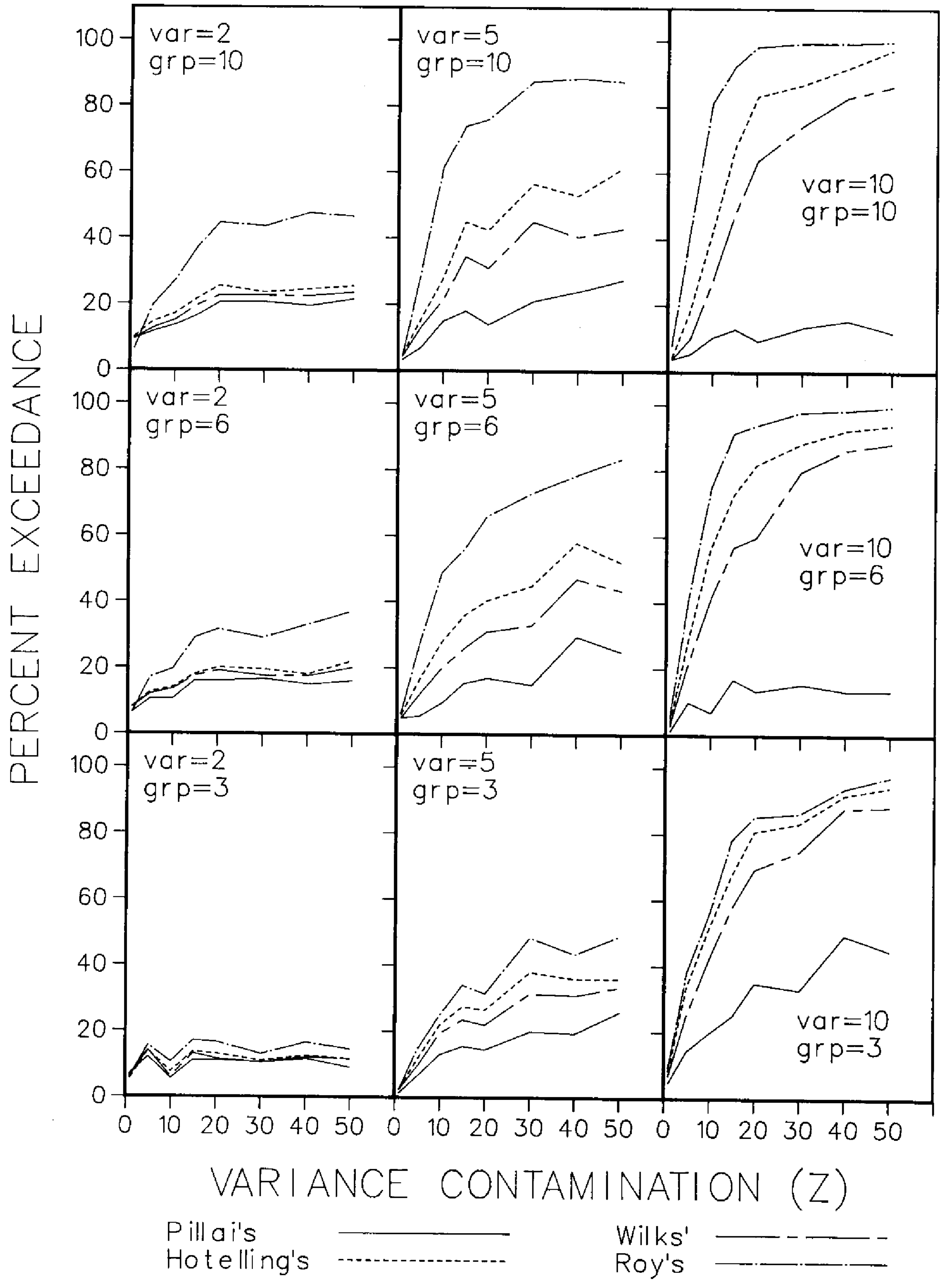


Fig 2.—Effect of heteroscedasticity on rates of Type I error in MANOVA statistics under conditions of different numbers of groups (grp) and variables within groups (var), when data are balanced but non-normal. In all cases $n = 10$.

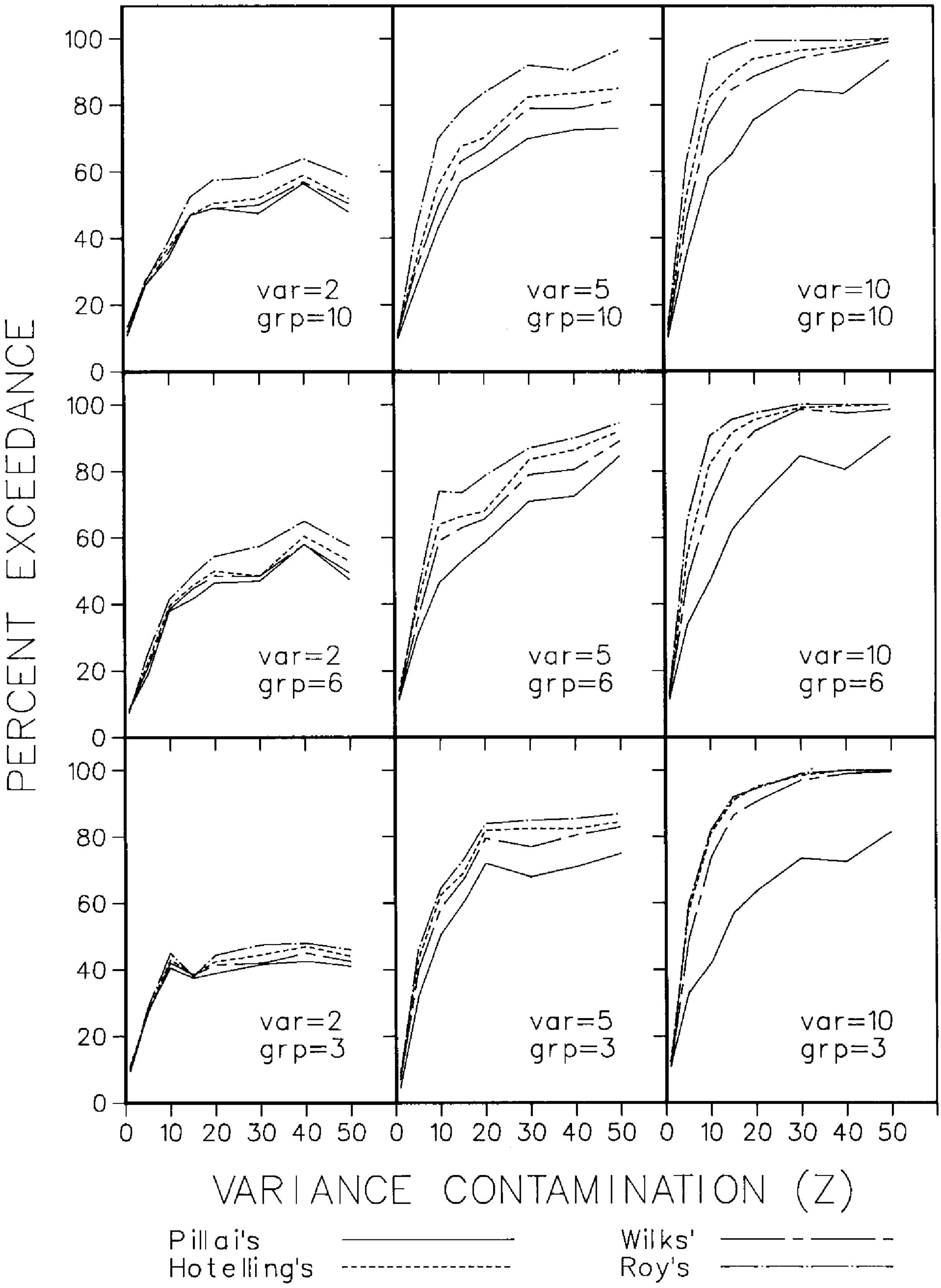


Fig 3.—Effect of heteroscedasticity on rates of Type I error in MANOVA statistics under conditions of different numbers of groups (grp) and variables within groups (var), when data are normal but unbalanced. For groups 1,2,3,... $n = 5,10,10,\ldots$ respectively.

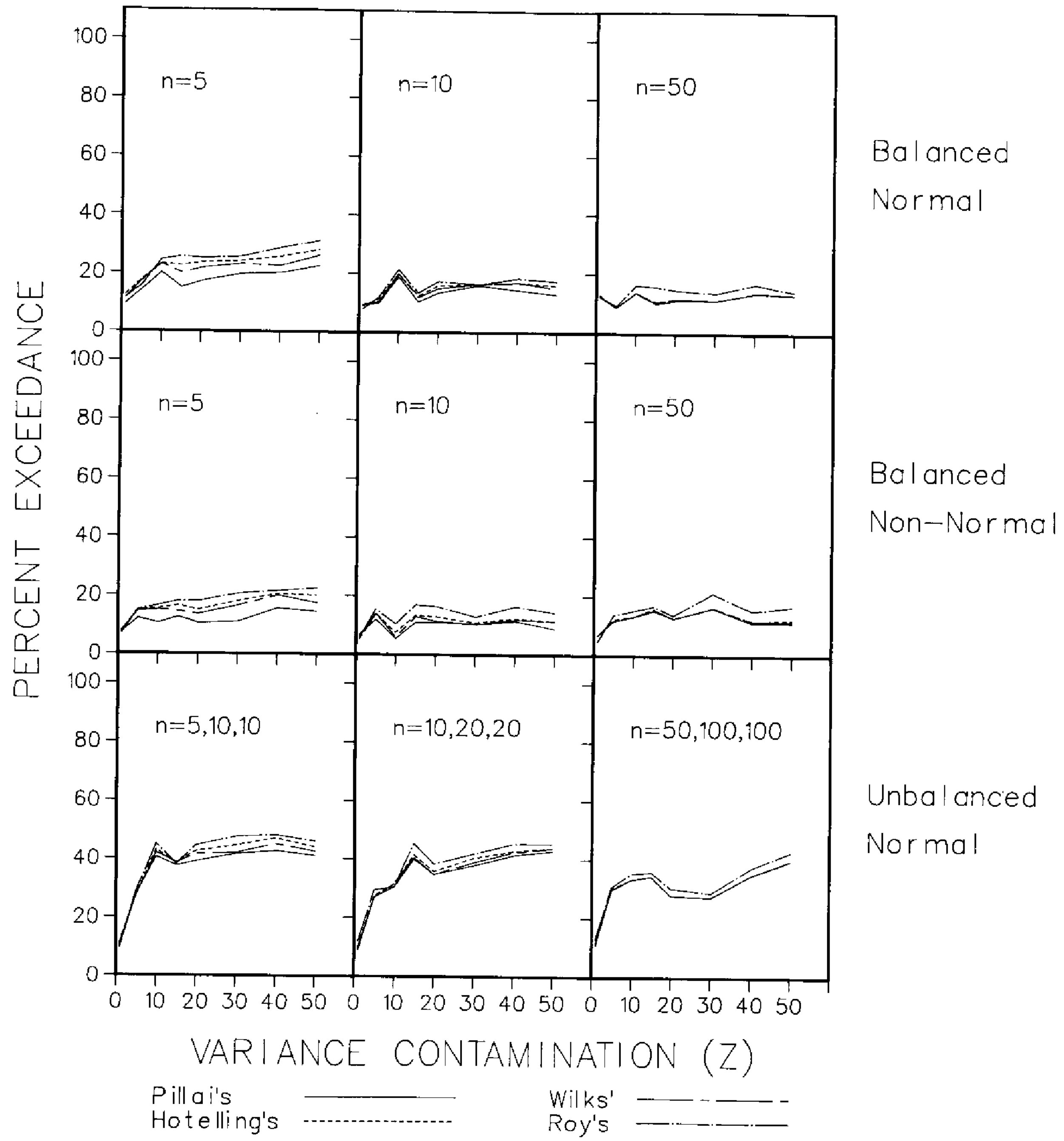


Fig 4.—Effect of sample size ($n$) and heteroscedasticity on rates of Type I error in MANOVA statistics under certain conditions of normality and balance in the data. In all cases the number of groups ($k$) = 3, and number of variables ($p$) = 2.

increase with numbers of groups regardless of dimensionality. Note that in our demonstration of the effect of heteroscedasticity, we have presented a 'moderate case' scenario in which scale variability occurs in all dimensions of only one group. The magnitude of the deleterious effects is less when heterogeneity occurs in fewer than the total of $p$ dimensions in a single group, and greater if it occurs in all $p$ dimensions in more than one group (Korin, 1972; Olson, 1974).

The exception to this general pattern is Pillai's criterion ($V$) which is the most robust to departures from covariance homogeneity, providing that sample sizes are equal. When the number of groups is small ($k$ = 3), this statistic behaves similarly to the others in that increasing dimensionality increases the likelihood of Type I error, although at a much lower rate. In contrast, when the number of groups is large ($k = 10$), adding more variables reduces Type I error rates (Figs 1 and 2). Olson (1974) examined the effect of $p$ and $k$ in detail and found that exceedance rates for Pillai's $V$ increase with $k$ when $k > p$, but decrease with increasing $k$ for values of $k < p$.

Note that in focusing on the number of groups ($k$), number of dimensions ($p$) and sample size ($n$) as the parameters likely to be most useful to practitioners, $df_{\mathrm{error}}$ has not been kept constant. However, given the dramatic effects on Type I error, it is unlikely that results of simulations in which $df_{\mathrm{error}}$ were kept constant would differ qualitatively from those presented here.

### *Effect of non-normality*

Non-normality (kurtosis) on its own has relatively little effect on Type I error rates. For balanced non-normal data with homogeneous covariance structure, rates of Type I error of all statistics are more conservative than the nominated significance level (cf. Figs 1 and 2 at $z = 1$). However, any effect of kurtosis in compensating for heterogeneity is of small importance because the tendency of kurtosis to yield too few significant results is overshadowed greatly by the large exceedance rates caused by variance heterogeneity.

Introducing heteroscedasticity into non-normal data with balanced sample sizes produces behaviour similar to that for heteroscedasticity in normal data. For all statistics except Pillai's $V$, which is little affected, Type I error rates with heteroscedastic data are slightly greater for non-normal than for multinormal data, but only when dimensionality is large ($p = 10$; cf. Figs 1 and 2).

### *Effect of unbalanced data*

If data are multinormal and homoscedastic, unequality in sample size on its own does not affect the nominal significance level. However, if data are both unbalanced and heteroscedastic, rates of Type I error are affected severely and may be excessively liberal or conservative depending on whether the divergent variance structure occurs in the group with the smallest or largest sample size (Ito & Schull, 1964).

When sample sizes are unequal (but multinormality satisfied) and variance heterogeneity occurs in the group with smallest $n$, as in our simulations, the Type I error rates of all statistics increases significantly for all levels of heteroscedasticity ($z > 1$), dimensionality and number of groups. Even when the departure from covariance homogeneity is minimal, and sample sizes are large, and there are few groups ($k = 3$) and variables ($p = 2$), the rate of Type I error of all the statistics exceeds the nominal level by an unacceptable amount (Figs 3 and 4). This dramatic and undesirable effect is exacerbated by increasing the number of variables, and to a lesser extent by adding more groups (Fig 3). Conversely, if heteroscedasticity occurs in the group with the largest sample size, rates of Type I error are also affected adversely, but in this situation are overly conservative (Ito & Schull, 1964).

## POWER

### *Summary*

The ability of all statistics to detect real differences among multivariate group means is influenced dramatically by both variance heterogeneity and non-normality. Perturbation of power curves is extreme if both assumptions are violated (Fig 5). The effects of variance heterogeneity on power are slightly more harmful when the divergent variance structure and non-centrality are coincident in the same group.

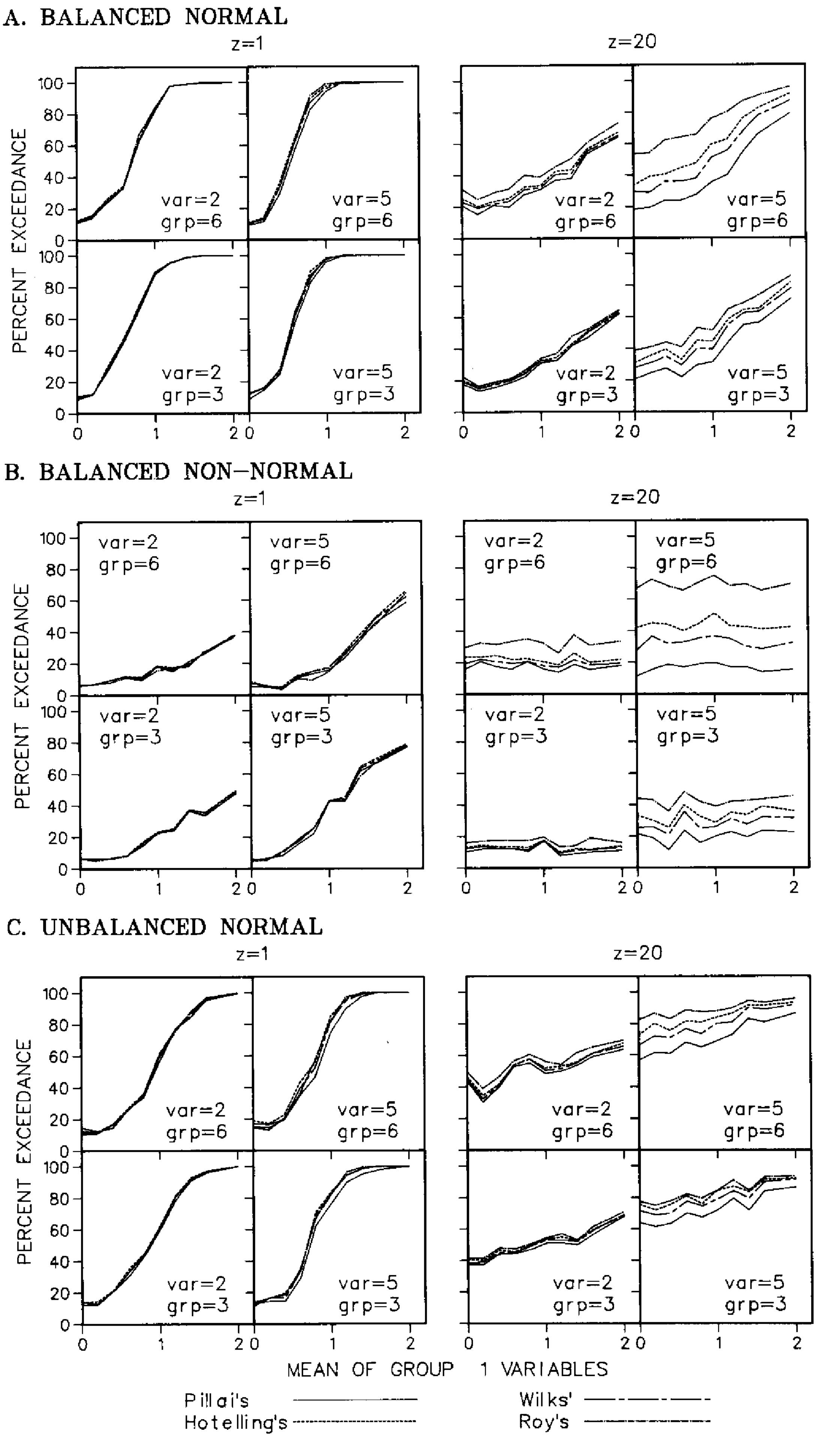
Violations usually cause serious loss of power at moderate to high levels of non-centrality. However, exceedance rates are excessive at the lower end of power curves for statistics whose rates of Type I error are elevated by violations and for this reason it is important to interpret power curves cautiously. If the power curve shows an elevated value at the null hypothesis, then the fact that power is high for non-null values is of little comfort. It is our view that the experimenter must first ensure that the level of the test is appropriate before undertaking power comparisons. In Figs 5–7 it is important to realise that if a procedure has a higher power for some level of non-centrality, this is only useful when the test has the appropriate level of Type I error. The rate of Type I error is given by the value of the power curve when non-centrality is 0, i.e. at the left hand end of the curves in Figs 5–7.

When data satisfy the requirements of multinormality and equal covariance structure, all test statistics demonstrate similar but not identical power (Figs 5 and 6). The small differences in power are sufficiently consistent to rank the statistics $V \geq W \geq T \geq R$ for small $n$ when non-centrality is diffuse, i.e. when group centroids are spread in all dimensions (Schatzoff, 1966; Pillai & Jayachandran, 1967; Lee, 1971; Olson, 1974). For concentrated non-centrality (i.e. when group centroids largely align along a single axis, as in our simulations) the order is reversed, and also the power of all statistics is slightly greater. However, when assumptions are met, for most practical purposes differences in power are small.

### *Effect of heteroscedasticity*

Moderate levels of variance heterogeneity ($z = 20$) affects significantly and adversely the power of all statistics irrespective of normality and balance, although Pillai's criterion is least affected (Fig 5, right hand side). The deleterious effects on power worsen with increasing heteroscedasticity (Olson, 1974). When data are balanced and multinormal, larger sample sizes ostensibly compensate for the serious loss of power with heteroscedasticity, e.g. power is similar for $n = 5$ with homoscedastic data, and for $n = 50$ with heteroscedastic data (Fig 6A). However, the increase in power with group size may not be solely an effect of sample size, since in our simulations the 'amount' of non-centrality is also influenced by sample size.

When there are no differences among multivariate group means (Fig 5 when the mean of Group 1 variables = 0), exceedance rates are set by the rate of Type I error. Thus, variance heterogeneity affects greatly the lower (i.e. left hand) end of the power curves for Hotelling's ($T$), Wilks' ($W$) and Roy's ($R$) statistics (Fig 5A, right hand side) because their rates of Type I error increase rapidly with heteroscedasticity (Figs 1–3). In contrast, because Pillai's criterion



A. BALANCED NORMAL

B. BALANCED NON-NORMAL

C. UNBALANCED NORMAL

($V$) is more robust (Figs 1–3), the lower end of the power curve for $V$ is less affected and the reduction in power is most apparent at the top end of the curve (Fig 5A).

### *Effect of non-normality*

The same qualitative pattern holds for non-normal as for heteroscedastic data but the harmful effects on power are less severe. Whereas non-normality has minimal effect on rates of Type I error (Fig 2), it reduces dramatically the power of all statistics (cf. left hand side of Figs 5A and 5B). Poor ability to discern differences among group means in non-normal (but homoscedastic) data might be ameliorated to some degree by increasing sample size (Fig 6B, but note that increasing sample size for the same value of group means also increases the amount of non-centrality). Conservative exceedance rates when there are no differences among group means (Fig 5B, left hand side) illustrate the effect of non-normality suppressing Type I error rates when data are homoscedastic (cf. Figs 1 and 2 at $z = 1$).

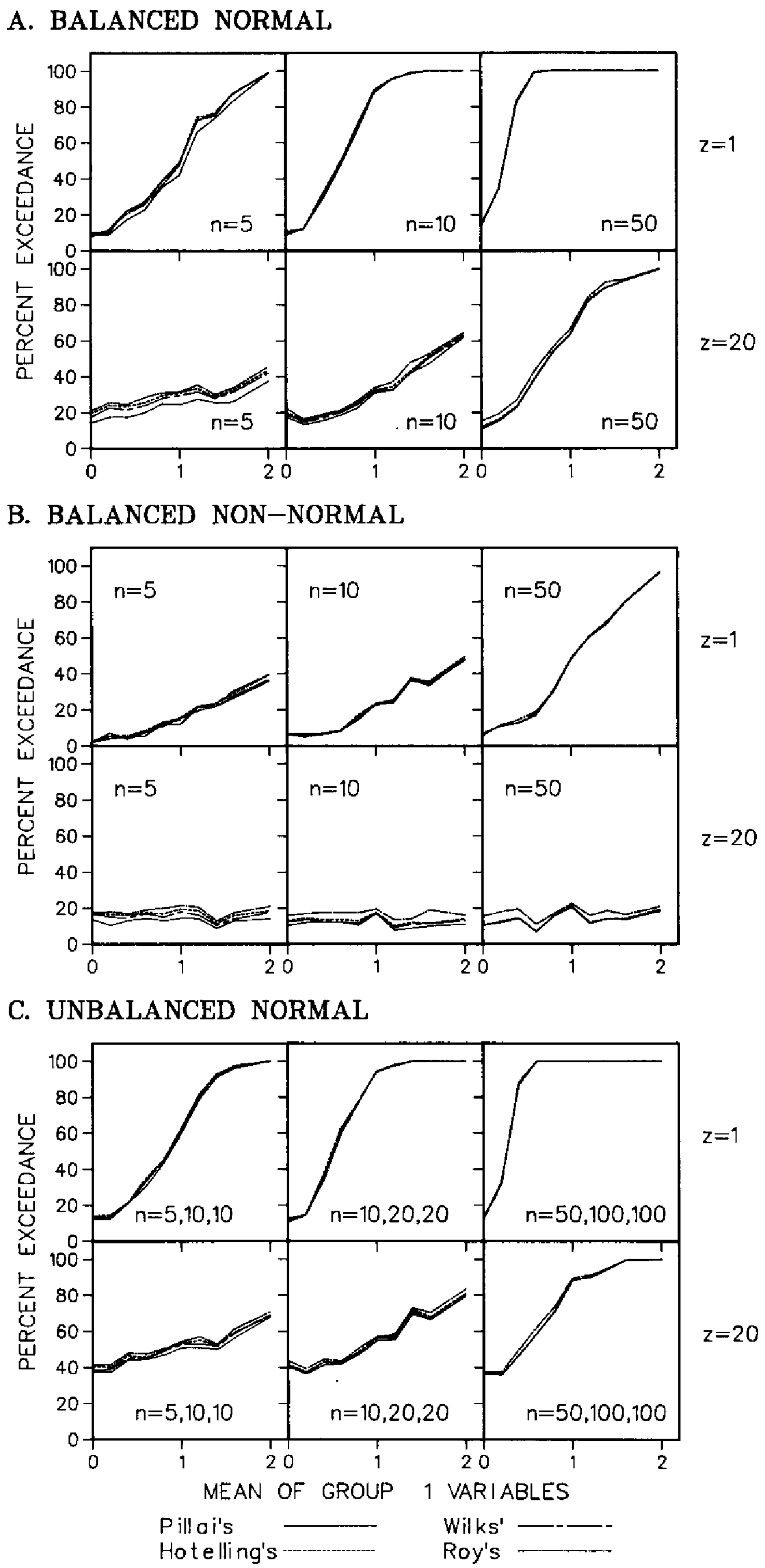
When non-normality and heteroscedasticity are introduced simultaneously, the effect on power is extreme and the curves are virtually flat (right hand side of Fig 5B). Moreover, in these circumstances increasing the sample size (and thus the level of non-centrality) does not improve power (Fig 6B). With heteroscedastic non-normal data, exceedance rates of statistics $R$, $T$ and $W$ increase to unacceptable levels with increasing dimensionality ($p > 2$) at very low levels of non-centrality, reflecting the sensitivity of their Type I error rates to violations of this kind. Only Pillai's $V$ tends not to yield too many significant results at low levels of non-centrality, but at higher levels of non-centrality the power of $V$ is reduced greatly (Fig 5B, right hand side).

### *Effect of unbalanced data*

The most serious effect of unbalance in data is when covariance matrices are unequal. Because unbalance in data leads to excessive rates of Type I error when variance structure is heterogeneous (Fig 3), exceedance rates at the lower end of power curves under these conditions are excessive, and the slope of power curves for all statistics is slight (Fig 5C, right hand side).

For multinormal homoscedastic data, interpretation of our results is not straightforward since unbalance in data is confounded with the effect of sample size, and therefore with the 'amount' of non-centrality. For a structure of two variables in each of three groups, with non-centrality occurring only in the first group, power at different sample sizes is ranked $n = 10{,}20{,}20 > n = 10{,}10{,}10 >> N = 5{,}10{,}10$ (cf. left hand side of Figs 5A and 5C, and Fig 6C). The

Fig 5.—Power curves of MANOVA statistics under certain conditions of covariance heterogeneity, normality and balance in the data, numbers of groups (grp), and numbers of variables within groups (var). Variance heterogeneity, if present at all, is coincident with non-centrality; $z = 1$ indicates homoscedastic covariance structure, $z = 20$ is moderately heteroscedastic (refer to Figs 1–4). In balanced designs $n = 10$, and in unbalanced designs the sample sizes of groups 1,2,3,... are $n = 5{,}10{,}10{,}\ldots$ respectively. Exceedance rates $> 0.10$ when data are heteroscedastic but when there are no differences among means (i.e. mean of Group 1 variables = 0) reflect rates of Type I error.



A. BALANCED NORMAL

B. BALANCED NON–NORMAL

C. UNBALANCED NORMAL

Fig 6.—Effect of sample size on power of MANOVA statistics under certain conditions of covariance heterogeneity, and normality and balance in the data. Variance heterogeneity, if present at all, is coincident with non-centrality; $z = 1$ indicates homoscedastic covariance structure, $z = 20$ is moderately heteroscedastic (refer to Figs 1–4). In all cases the number of groups ($k$) = 3, and number of variables ($p$) = 2.

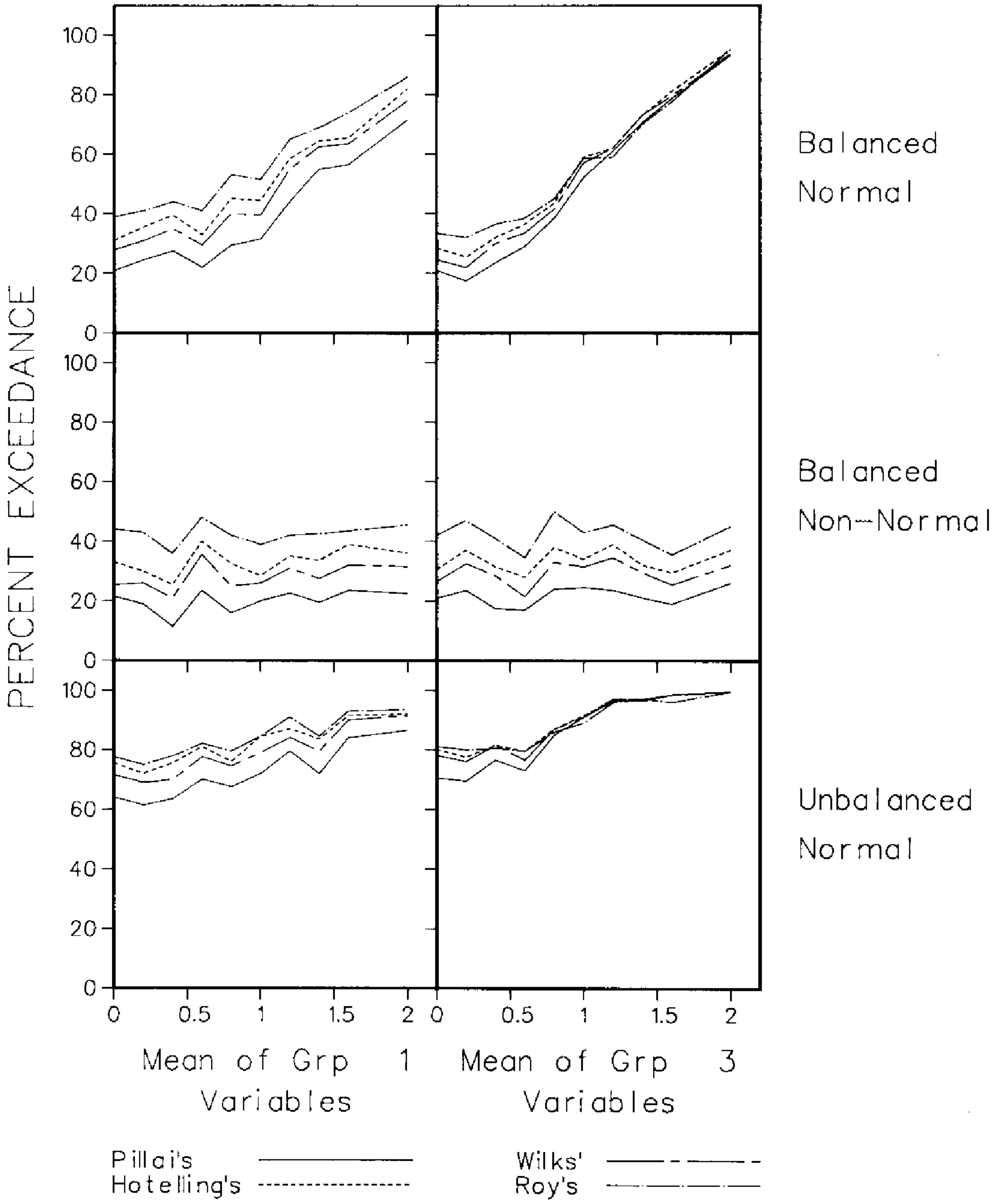


Fig 7.—Effect of coincidence, with respect to group, of variance heterogeneity and non-centrality on power of MANOVA statistics under certain conditions of normality and balance in the data. In all cases data were moderately heteroscedastic ($z = 20$; refer to Figs 1–4), the number of groups ($k$) = 3, number of variables ($p$) = 5, sample size $n = 10$ except in unbalanced designs where $n = 5,10,10$ for groups 1,2,3, respectively. Variance heterogeneity was always introduced into Group 1, and therefore was coincident with non-centrality when the population means of Group 1 were > 0, but was not coincident when the population means of variables in Group 3 exceeded zero. Note that, if present at all, heteroscedasticity and non-centrality occurred in all dimensions of one group (but not necessarily the same group), thus the two were always coincident with respect to dimension.

essential result is that differences among these curves are relatively small. The (slightly) greater power of the structure $n = 10,20,20$ over $n = 10,10,10$ implies that the difference is due to different sample sizes, since the 'amount' of non-centrality is less in the first structure than in the latter. Thus, even if loss of power can be attributed to unbalance in the data, it is likely that this could be ameliorated to some extent by increasing sample size. Predictably, our results suggest that power will be poorest when non-centrality is coincident in the group with the smallest sample size.

### *Effect of coincidence of heteroscedasticity and non-centrality*

Reduction in power with heteroscedasticity is less pronounced when non-centrality occurs in a group other than that containing the variance heterogeneity (Fig 7). However, if data are both non-normal and heterogeneous the small reprieve for power from non-coincidence with respect to groups is lost and power curves are equally poor irrespective of coincidence (in our simulations non-centrality always occurred in all dimensions of one group, and thus was always coincident with variance heterogeneity with respect to dimension). Olson (1974) provides a detailed discussion of the effect of coincidence of non-centrality and heterogeneous variances. He concluded similarly that the effect of heterogeneity in reducing power is greatest when heterogeneity and non-centrality are coincident with respect to both group and dimension. Olson's results show that the structure used in some of our simulations, in which non-centrality and heterogeneity occur in all dimensions of the same group, is the worst case for loss of power.

### *Effect of number of groups (k) and dimensionality (p)*

It is difficult to assess the effect of dimensionality ($p$) and number of groups ($k$) on power from our simulations because changing $p$ or $k$ also changes the amount of non-centrality (particularly when non-centrality is distributed diffusely among all dimensions of only one group) and $df_{\text{error}}$. Thus, the slight increase in power with dimensionality (Fig 5A, left hand side) cannot unequivocally be attributed to the effect of dimensionality alone. We found power little affected by increasing the total number of groups (Fig 5A, left hand side). Conversely, Lee (1971) and Olson (1974) found that increases in $p$ or $k$ tended to reduce power when non-centrality structure and values of their non-centrality parameters were held constant. The divergence with our simulations does not indicate conflicting results but reflects a different choice of arbitrary standard. In our study we kept constant the magnitude of the difference (from zero) of the means of all dimensions of one group, whereas both Lee and Olson maintained a constant value of a standardised non-centrality measure. The important result of all three studies is that the effect on power of dimensionality and number of groups when assumptions are satisfied is small.

## ROBUSTNESS OF BOX'S M

Box's $M$ statistic is extremely sensitive to low levels of heteroscedasticity irrespective of the number of groups, dimensionality, normality, or equality of sample size (Fig 8, see also Hopkins & Clay, 1963; Korin, 1972; Olson, 1974). The test rejects the requirement of heteroscedasticity at levels of variance heterogeneity that have no serious effect on the behaviour of some of the MANOVA statistics it is designed to protect (cf. Fig 8 with Figs 1–4). In this respect it is not a useful test. Moreover, like its univariate analogue (Bartlett's test), Box's $M$ is highly sensitive to non-normality and cannot, therefore, distinguish between

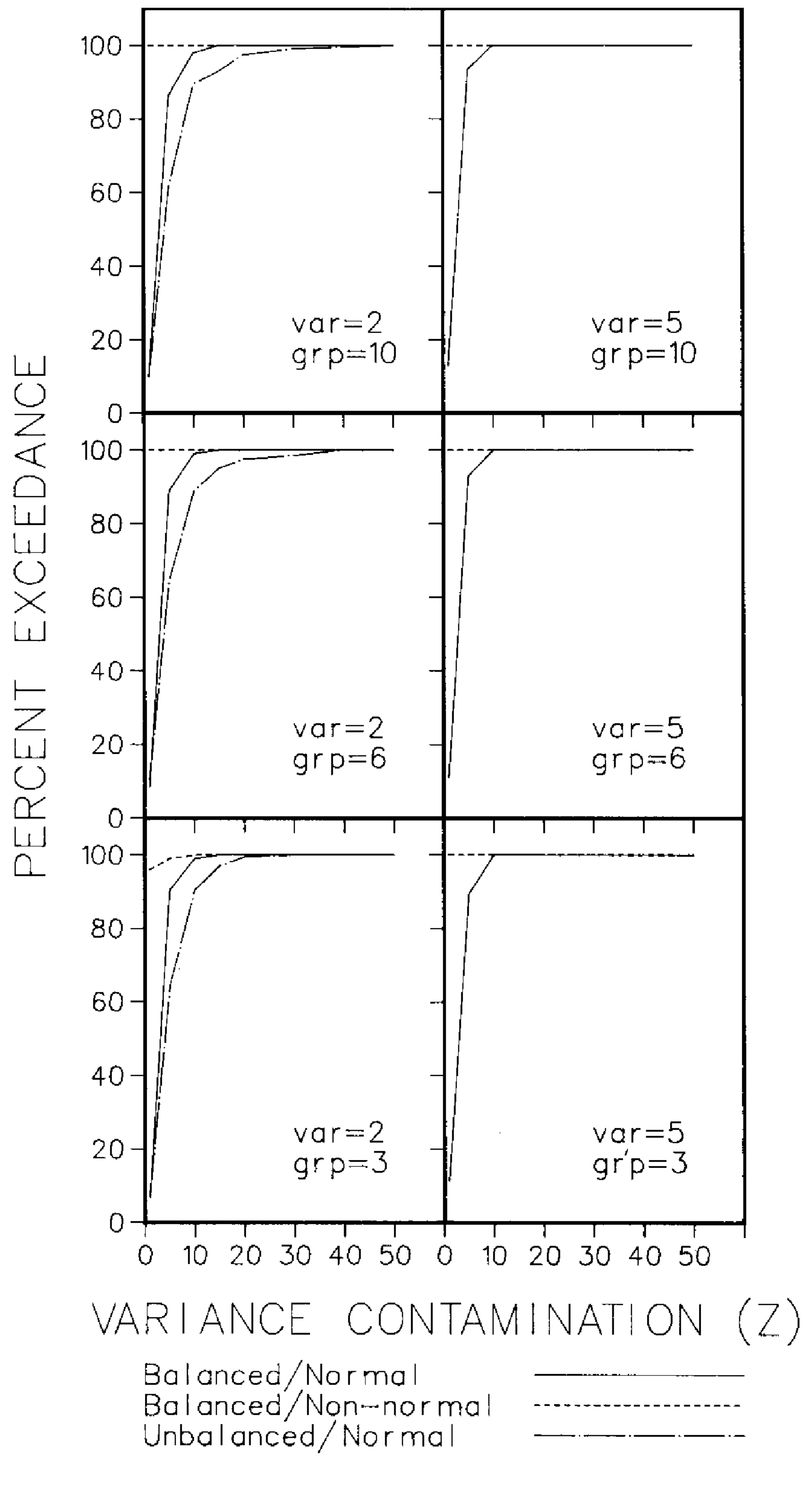


Fig 8.—Sensitivity of Box's *M* statistic to heteroscedasticity under different conditions of normality and balance in the data, and numbers of groups (grp) and variables within groups (var). In all cases variance heterogeneity was introduced into Group 1, and $n$ = 10 except in unbalanced designs, in which case $n$ = 5,10,10... for Groups 1,2,3,... respectively. No plots are given for unbalanced data when $n < p$ (right hand side of figure) in any group because in these cases dispersion matrices are overdefined and their determinants are zero.

non-normality and heteroscedasticity (Fig 8). Other lesser-known tests for equality of covariance matrices have also proven to be extremely sensitive to non-normality (Mardia, 1971). These undesirable properties of *M* are magnified with increasing sample size. Thus, the test is most useful for indicating when data are multinormal and homoscedastic.

## RECOMMENDATIONS FOR USING MANOVA

It is clear that under certain conditions the power and rates of Type I error of the MANOVA statistics *R*, *T*, *W* and *V* are not robust to some violations of the multivariate general linear model, and that there are considerable differences among the statistics in their response to violations. These results suggest several recommendations in designing and analysing multivariate experiments. Specifically, they provide some answers to the questions: (1) is there a 'most reliable' MANOVA statistic? (2) what are important features to include in multivariate experimental design? and (3) does Box's *M* statistic provide a satisfactory test of the assumption of heteroscedasticity with a power commensurate with the robustness of the MANOVA statistics? They also raise pertinent questions: (1) is there an alternative statistic to test reliably whether data meet the assumptions of heteroscedasticity and multinormality? (2) given that if assumptions are violated it is often desirable to reduce dimensionality, how might this be achieved? (3) can transformations be found to redescribe data so that test statistics will perform reliably? (4) are there procedures for multiple range tests for planned and unplanned multiple comparisons after MANOVA?, and (5) are there alternatives to parametric MANOVA when data are unsuitable for parametric analysis?

### WHICH MANOVA STATISTIC TO USE?

Choice of test statistic necessarily reflects opinions of the relative evils of Type I and Type II errors. We argue that in most cases the first priority is to ensure rates of Type I error do not deviate greatly from the nominal level, after which choice can be based on maximum power (i.e. minimising Type II errors). This conservative policy minimises the danger of claiming significance too often, which we view as hazardous, but offers less protection against inability to detect real differences, which is inconvenient. An exception to this argument arises in environmental impact studies where power is crucial (particularly in that $df_{\text{error}}$ is often small in impact studies).

By these criteria, Pillai's *V* statistic emerges clearly as the choice for general use. For balanced designs *V* shows considerable robustness to moderate levels of variance heterogeneity and is relatively insensitive to violations of the assumption of normality. In these respects it parallels closely the univariate *F*-statistic (Scheffe, 1959; Glass *et al.*, 1972). Also, *V* is robust in demonstrating the smallest increase in exceedance rates with dimensionality (*p*) for a given amount of heterogeneity, and is the only statistic where for large *k*, increasing *p* reduces the degree of liberality. In many circumstances *V* can also be recommended on the basis of its power. *V* has good power over a wide range of conditions and is the most powerful of the four statistics when non-centrality is diffuse, i.e. when group centroids are scattered and do not align along a single main axis. In marine ecological data from benthic studies, non-centrality is often diffuse in that it occurs in a large proportion of dimensions across several groups, particularly when there are strong seasonal or treatment effects (e.g. Field, 1971; Johnson & Mann, 1986, 1988). In these cases *V* is the clear choice.

Olson (1974, 1976, 1979) also advocates the *V* statistic for general use, but this recommendation is not supported by all researchers. Stevens (1979) based his recommendations on the power of the statistics, arguing that because differences in robustness of the *V*, *W* and *T* statistics are small for 'typical' amounts of variance heterogeneity, the *W* or *T* statistics are preferable to *V* if non-centrality is concentrated in a single group or small number of groups (*R* is not usually considered suitable given its excessive liberality). Recall that the ranking of power of the tests is $R > T > W > V$ for concentrated non-centrality and $V > W > T > R$ for diffuse non-centrality, at least when the amount of non-centrality is defined by a standardised measure (Schatzoff, 1966; Lee, 1971; Olson, 1974, 1976; Stevens, 1979). However, the assertion of equivalent robustness to even low levels of heteroscedasticity is correct only for very large samples (cf. Figs 1, 2 and 4 and see Schatzoff, 1966). For small samples *V* is clearly more robust than *W* and *T*. From his empirical results, Olson (1976) suggested that the three can be considered equivalent only when $(df_{\text{error}}/10p) \geq df_{\text{hypothesis}}$, or equivalently when $(df_{\text{error}}/df_{\text{hypothesis}}) \geq 10p$. Even if sample size is large, another problem is to assess whether non-centrality is concentrated or diffuse. Clearly it is inappropriate to base this assessment on significance tests of different dimensions of the sample to be tested by MANOVA. A straightforward approach can be based on the fact that if non-centrality is concentrated then differences among group centroids align along a single dimension in Mahalanobis' space. This may occur, for example, when sampling along a uniform ecological gradient. The distribution of centroids in Mahalanobis' space can be examined by canonical discriminant analysis (CDA), which is a standard inclusion in most multivariate software packages (CDA is equivalent to a graphical form of MANOVA, both being based on Mahalanobis' distances between centroids; see section on Multiple Comparisons p. 209). Thus, unless assumptions of MANOVA are not violated, *W* or *T* should be used in preference to *V* only when sample size is sufficiently large and it can be ascertained that differences among mean vectors are concentrated in very few dimensions.

In the special circumstance in which non-centrality is concentrated and there are no violations, then *R* is the statistic of choice given its superior power. It should also be noted that if non-centrality is concentrated such that centroids align in a unidimensional arrangement, then dimensionality of the data can be reduced greatly (see section on Reducing Dimensionality, p. 202). If the number of dimensions can be reduced effectively to $\leq 2$ and if the number of groups is small ($\leq 6$), then *R* is robust to moderate levels of variance heterogeneity (Figs 1 and 2). Thus, for small to moderate departures from homogeneity, if non-centrality is concentrated and dimensionality small then *R* would be the appropriate choice, particularly if power was an issue.

We conclude by commenting that the choice of an appropriate statistic has been a contentious issue in the literature and is likely to continue to be a topic of spirited and divisive debate depending largely on judgements on the rather slippery issue of the relative seriousness of Type I and Type II errors. However, we contend strongly that the recommendations given here provide for reliable results.

## RECOMMENDATIONS FOR DESIGN

Our recommendations for design are based on the view that violations are likely to occur in multidimensional ecological data sets, and that some kinds of violations affect adversely the behaviour of all test statistics. Violations are likely to occur for two reasons; first, as the number of variables increases so do the number of ways that assumptions can be violated, and second, regarding marine ecological data in particular, species abundances often manifest large temporal and spatial variability, and there are often large differences in the abundances of coexisting species at any point in time or space. In suggesting guidelines for multivariate experiments, we assume that design requirements general to all inferential statistics are satisfied, e.g. that observations are distributed independently.

The most critical implication of our simulation results for design is that sample sizes be identical (or nearly identical for very large *n*). When sample sizes are unequal, the rates of Type I error of Pillai's *V* (and the others) are not robust to even low levels of heteroscedasticity (Fig 3). This behaviour is similar to that of the *F*-test of univariate ANOVA in which heteroscedasticity affects severely the likelihood of Type I error when group sizes are unequal (Scheffe, 1959). If a small number of replicate measurements are missing from some groups and inequality among covariance matrices is indicated, in most cases it is preferable and conservative to make a small sacrifice in power and engineer a balanced design by randomly discarding replicates from those groups in which $n > n_{\min}$.

Reducing dimensionality also contributes to minimising the deleterious effects of heteroscedasticity on rates of Type I error, particularly if the number of treatment groups is small (see p. 202). This can be facilitated by careful planning in the design stage and a precise statement of the hypothesis. Researchers must not yield to the temptation to include variables of peripheral importance. Similarly, although not so critical, the number of groups should be minimised insofar as the hypothesis permits.

Predictably, if assumptions are met, the power of all MANOVA statistics increases with sample size, and for a given absolute increase in sample size the improvement in power is greatest for small *n*. Clearly, it is best to take as many replicates as is practical.

## TESTING WHETHER DATA MEET ASSUMPTIONS

Because MANOVA statistics are not robust to all violations, it is necessary to know when, and which, violations occur, and by what magnitude. A test for violations is only useful if it does not detect violations to which the MANOVA statistic of choice is robust. By this standard the *M* statistic can be rejected as too sensitive (indeed, its sensitivity is such that calibrating it to be commensurate with the robustness of the MANOVA statistics is likely to be problematical). Furthermore, *M* is sensitive to both non-normality and heteroscedasticity, so it is not possible to know which assumptions are violated.

We suggest an alternative test proposed by Hawkins (1981) that tests for non-normality and heteroscedasticity simultaneously. For each data point, a quantity $A_{ij}$ is computed. The behaviour of these $A_{ij}$s will provide information on the deviation of the data from normality and heteroscedasticity. The $A_{ij}$s are computed using the pooled covariance matrix and the deviation of the observation from the within-group mean. Details of the computation are as follows; first compute

$$V_{\mathrm{ij}} = (\mathbf{X}_{\mathrm{ij}}-\mathbf{X}_{\mathrm{i.}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{X}_{\mathrm{ij}}-\mathbf{X}_{\mathrm{i.}})$$

where $\mathbf{X}_{\mathrm{i.}}$ is the mean vector for the observations in group $i$ for each of the $k$ groups, and $\mathbf{S}$ is the pooled covariance matrix. Note that $N$ is $\sum_{i=1}^{k} n_i$, the total number of observations. These quantities are generally easy to compute in any standard statistics package. The next step is to compute

$$F_{ij} = \frac{(N-k-p)\,n_i V_{ij}}{p((n_i-1)\ (N-k)\ -n_i V_{ij})}$$

If the data are normal and homoscedastic, it can be shown that $F_{\mathrm{ij}}$ follows an $F$ distribution with $(p,N\text{-}p\text{-}k)$ degrees of freedom (recall that $p$ represents the dimension of the data). Finally

$$A_{\mathrm{ij}} = Pr\,[F>F_{\mathrm{ij}}]$$

denotes the tail of $F_{\mathrm{ij}}$ under this distribution. Once the $V_{\mathrm{ij}}$s have been calculated, the computation of $A_{\mathrm{ij}}$s is routine. If the data are normal and heteroscedastic, then the $A_{\mathrm{ij}}$s will be distributed uniformly over the interval (0,1). Hawkins (1981) proposes a test statistic based on the $A_{\mathrm{ij}}$s which tests each group separately as well as overall. For group $i$ the test statistic $W_{\mathrm{i}}$ is computed as follows:

Order the $A_{\mathrm{ij}}$s as $A_{i(1)} \leq A_{i(2)} \leq \ldots A_{i(n_i)}$ . Now

$$W_i = n_i - n_i^{-1} \sum_{j=1}^{n_i} (2j-1)\ (log A_{i(j)} + \log\ (1-A_{i(n_i-j+1)}))$$

These statistics can be computed for each group and the value compared against the critical values for the Anderson-Darling statistic (see Anderson & Darling, 1954). A test statistic for the overall data set can be computed by repeating the procedure on the $N$ $A_{\mathrm{ij}}$s.

At this point there is still the difficulty, as with the $M$ statistic, that rejecting the null hypothesis of normal and homoscedastic data does not necessarily mean that the MANOVA procedures are invalid, i.e. the statistic needs to be calibrated to be appropriate to the robustness of the MANOVA $V$ statistic. As a first step in addressing this problem we recommend that the $W_{\mathrm{i}}$s be computed. If none is rejected then proceed with the MANOVA with assurance. If some are rejected, it is necessary to examine the $A_{\mathrm{ij}}$s to ascertain the type of departure from the null hypothesis, and whether the violation is harmful in terms of the validity of the MANOVA. If the data are non-normal but homoscedastic, the $A_{\mathrm{ij}}$s from each group will have the same non-uniform distribution. If the data are longer tailed than normal (leptokurtic) there will be an excess of large and small values of $A_{\mathrm{ij}}$s giving rise to a U-shaped distribution. A distribution which is shorter-tailed than the normal (platykurtic) gives rise to a distribution of $A_{\mathrm{ij}}$s which has a peak in the middle of the interval (0,1). If the data are normal but heteroscedastic, the $A_{\mathrm{ij}}$s will cluster near 0 for the groups with small variance and near 1 for groups with large variance.

The simulation results in the previous section indicate that heteroscedasticity has a much worse effect on level (Type I error) than does non-normality. With this in mind, a simple test to detect deviations from the assumptions which are harmful to MANOVA can be based on the range of the medians of the $A_{ij}$s. Specifically, if the overall test based on the $W_i$s indicates problems, then compute the median of the $A_{ij}$s for each group $i$ (to give $k$ medians). Now compute the range of the medians (i.e. the maximum-minimum). If the range is large then there is an indication of harmful deviations.

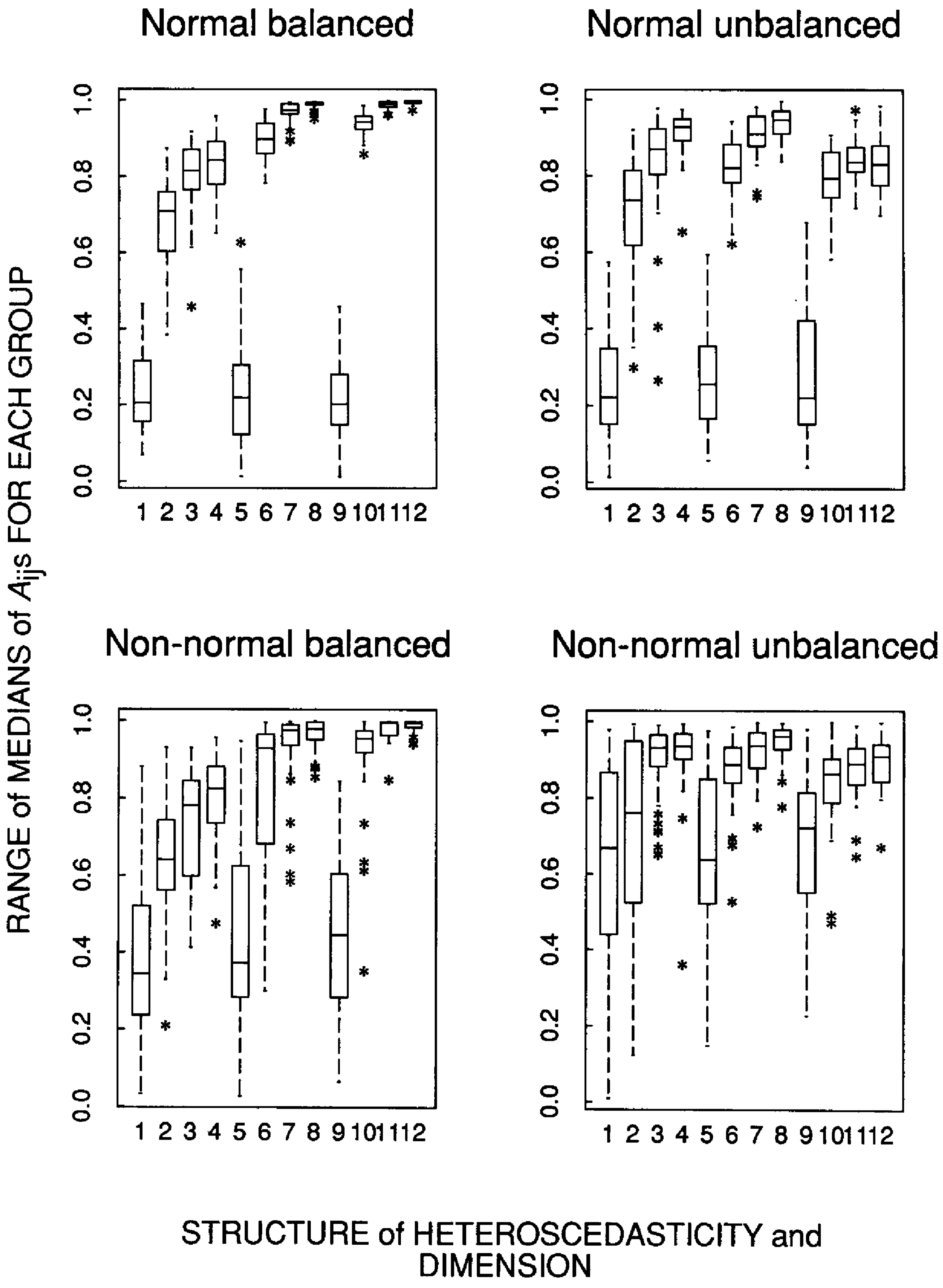
Since the distribution of the range of the medians is difficult to work out, we carried out a simulation to calibrate the range for the situations considered in this paper. The results are given as boxplots (Fig 9) for each of twelve combinations of dimension ($p = 2,5,10$) and heteroscedasticity ($z = 1,10,30,50$). In the case of balanced data, the range exceeds 0.85 for problematic cases of $p = 5$ and $z = 30$ or 50 (labelled 7 and 8, respectively), and of $p = 10$ with $z = 10,30$ or 50 (labelled 10,11,12, respectively). For unbalanced data where any heteroscedasticity is harmful, the range exceeds 0.50 for situations in which the level is liberal.

In summary, we recommend that the test statistic $W_i$ be computed for each group and then for the overall data set. If they do not exceed the critical value of the Anderson-Darling statistic, then proceed with the MANOVA. Otherwise, compute the range of the medians as outlined above. If the data are balanced, a range in excess of 0.85 is a certain indicator of problems with the MANOVA, while if the data are unbalanced, a range in excess of 0.5 indicates problems. It should be emphasised that these cut-off values are based on a small simulation study and there is a need for further work to obtain more precise critical values.

## REDUCING DIMENSIONALITY

Minimising the number of variables ($p$) is important in so far as it reduces the sensitivity of the $R$, $W$ and $T$ statistics to violations under all conditions, and of the $V$ criterion when (approximately) $p > k$. Other reasons to reduce dimensionality are that it reduces the likelihood of a violation occurring at all, provides greater power (especially important if the original number of variables is high relative to $df_{\text{error}}$), and reduces costs of computation.

Fig 9.—Results of simulations to calibrate range of medians of $A_{ij}$s of Hawkins' (1981) test for heteroscedasticity and non-normality so that power of the test is commensurate with robustness of the $V$ statistic. Each boxplot shows range from first to third quartile, and median (central horizontal line); dotted lines show extremes, not including outliers (by arbitrary definition) which are given as stars. Results are for combinations of normal/non-normal, balanced/unbalanced, and homoscedastic/heteroscedastic data. For balanced data, the range exceeds 0.85 for problematic cases with respect to level for $p = 5$ and $z \geq 30$, and for $p = 10$ with $z \geq 10$. For unbalanced data where any heteroscedasticity is harmful, the range exceeds 0.50 for situations in which Type I error rates are liberal. The 12 combinations of heteroscedasticity ($z$) and dimensionality ($p$) structures are: (1) $p = 2, z = 1$; (2) $p = 2, z = 10$; (3) $p = 2, z = 30$; (4) $p = 2, z = 50$; (5) $p = 5, z = 1$; (6) $p = 5, z = 10$; (7) $p = 5, z = 30$; (8) $p = 5, z = 50$; (9) $p = 10, z = 1$; (10) $p = 10, z = 10$; (11) $p = 10, z = 30$; (12) $p = 10, z = 50$. In all cases number of groups $k = 3$; sample size for balanced data $n = (10,10,10)$, and for unbalanced data $n = (5,10,10)$. Each of the 12 structures was generated from 50 runs, giving a total of 600 for each of the 4 combinations of normality and balance.


Normal balanced
Normal unbalanced
Non-normal balanced
Non-normal unbalanced
RANGE of MEDIANS of $A_{ij}$s FOR EACH GROUP
0.0
0.2
0.4
0.6
0.8
1.0
1 2 3 4 5 6 7 8 9 10 11 12
STRUCTURE of HETEROSCEDASTICITY and DIMENSION

We emphasised earlier the importance of including only those variables which are necessary to the hypothesis. Once a set of variables has been chosen for analysis, dimensionality can often be reduced effectively by any of several ordination procedures (e.g. Green, 1979; Field *et al.*, 1982; Gauch, 1982; Legendre & Legendre, 1983; Pielou, 1984; Austin, 1985; Minchin, 1987). These procedures describe the relationships between the $n \times k$ objects (= replicates × groups in a balanced design) in reduced space and the MANOVA can be conducted on the scores or co-ordinates of the objects in the reduced number of dimensions.

Several ordination procedures are applied widely in reducing dimensionality of multidimensional ecological data, and the question arises as to which method is most appropriate prior to MANOVA. The choice is sensibly based on how the different techniques distort the data to represent relationships among objects in reduced space. Principal components analysis (PCA) uses metric Euclidean space and thus, in the context of preceding MANOVA, can be recommended as a straightforward and (potentially) easily interpreted analysis since in most ecological studies the data space is Euclidean. However, PCA does not always provide a parsimonious description of ecological data and is particularly ineffective in reducing dimensionality when data contain pronounced non-linearities. In these circumstances non-metric or hybrid multidimensional scaling (MDS) are suitable alternatives (both PCA and MDS are discussed in more detail below). Ordinations in Mahalanobis' space, while theoretically possible, are inappropriate (in the sense of preceding MANOVA) since distinctions between treatment groups are likely to be increased artificially. For example, canonical discriminant analysis (= CDA), like MANOVA, is based on Mahalanobis' distances and maximises among group variation relative to within group variation. Thus, CDA is an ordination technique analogous to a graphical MANOVA (see section on Multiple Comparisons p. 209), so that conducting a MANOVA on CDA scores is to undertake a MANOVA on MANOVA output (that has already attempted to maximise among groups differences), which clearly is both unwise and misleading.

### *Principal components analysis*

PCA yields $p$ new uncorrelated variables, the principal components, that are linear combinations of (usually) linear functions of the original $p$ variables. Because principal components (PCs) are ordered in terms of decreasing fractions of the total variance described by the $p$ original variables, the MANOVA can be conducted on $m < p$ principal components that account for a given amount of the total variance of the original $p$ dimensions. Other advantages to using PCA preliminary to MANOVA are that PCs are likely to be more normally distributed than the original variables by virtue of the central limit theorem (Morrison, 1976), and because they are uncorrelated, if multinormality is satisfied, they will also be independent (see Anderson, 1984) which means that separate transformations can be used on each PC (see p. 208).

The PCA can be conducted on either the covariance matrix (or equivalently, the total sums of squares and cross products matrix) or correlation matrix that describes the relationships between the $n \times k$ objects (in a balanced design) in the original $p$-dimensional space. The choice is important and depends both on the hypothesis and nature of the data. If the response variables are not measured in the same units then linear combinations of variables are meaningless and covariances difficult to interpret, so the correlation matrix must be used.˙For example, depending on their life-form, the abundances of different species may be measured either in terms of biomass, percent cover, or as numbers of individuals.

When all variables are of the same units, e.g. as for some sets of species abundance data, the decision to use the covariance or correlation matrix in the PCA depends on the hypothesis. The correlation matrix is equivalent to a covariance matrix obtained from data that have been standardised so that each variable has a mean of zero and unit standard deviation (i.e. in terms of the untransformed observation $Y_{ij}$, the standardised observation $X_{ij} = (Y_{ij}\text{-}\mu_i)/\sigma_i$, where $\mu_i$ and $\sigma_i$ are the mean and standard deviation of the $i^{th}$ variable respectively). In ecological data, the largest mean scores (e.g. most abundant species) tend to have the highest variances, therefore standardising by equalising the variances prevents variables with large values from dominating the analysis. Thus, if the response of rare and common species are considered to be of equal importance, it is appropriate to conduct the PCA on the correlation matrix. It follows that a danger of using the correlation matrix is that, for example, rare species whose presence or absence is determined largely by chance, can overemphasise a spurious result. In these circumstances, there is good reason to exclude very rare species from the analysis (e.g. Johnson & Mann, 1988). Noy-Meir *et al.* (1975) and Pielou (1984) discuss in greater detail the hazards and advantages of standardising data. Domination by variables with large values can also be prevented by other kinds of transformations (rescaling) prior to conducting a PCA on the covariance matrix (see below). These transformations preserve the rankings of means and therefore may be more useful than the correlation matrix in achieving parity in the contribution of each variable. If changes in absolute abundance are of interest, or if it is felt the responses of common species should receive more weight than less abundant ones, the covariance matrix should be used. In PCAs of species abundance data using covariance matrices, the most abundant species tend to have higher variances, and therefore dominate the first principal components (e.g. Johnson & Mann, 1986). In this situation, fewer PCs calculated from a covariance matrix will account for an equivalent amount of information as a greater number of PCs obtained from a correlation matrix from the same set of original variables.

There are several problems associated with using PCA to reduce dimensionality prior to MANOVA, so the practice should not be followed automatically. The most critical is that the combination of variables that account for the largest amount of variance (i.e. the first PCs) need not be the combination that best describes variance among groups, which after all, is the focus of interest in MANOVA. The danger in discarding PCs associated with the smallest eigenvalues, which, for example, may equate with species having relatively small variances, is that one may be discarding species containing information useful in emphasising differences in species abundances among groups. It is, therefore, a vital precaution to check plots of PC scores, especially of the smallest ones, against the different levels of classes or factors. If clear differences in PC scores are evident among groups, those PCs should be included in the MANOVA. However, the problem is to decide how large these differences should be, and this decision is largely subjective.

A second point is that some data sets will contain prominent non-linearities, so that a small number of PCs that are additive combinations of linear functions of original variables will not represent a parsimonious description of variance structure and thus will not be efficient in reducing dimensionality. In these cases it is not wrong to use PCs, it is simply unwise, because the PCA will be ineffective in reducing the number of dimensions. However, PCs do not have to be additive linear functions of original variables, and it may be that additive quadratic, or higher order polynomial, functions best describe the largest amount of variance in the smallest number of dimensions. Gnanadesikan (1977) gives a clear account of generalised polynomial PCA, of which linear functions of the original variables is a special case. The method is straightforward and involves adding columns and rows to the covariance (or correlation) matrix that describe covariance (or correlation) structure of the quadratic (or higher order) terms, and then conducting a regular PCA on the expanded matrix. For example, in the simple case of $p = 2$ dimensions in a quadratic PCA, instead of describing covariance structure of the 2 variables ($y_1$ and $y_2$) in a $2 \times 2$ covariance matrix as in linear PCA, it is expanded to a $5 \times 5$ matrix to include covariance structure of the quadratic terms $y_1y_2$, $y_1^2$, and $y_2^2$. Although software for PCA is usually tailored for linear PCA, higher order polynomial PCA can be performed readily using standard packages for eigenanalysis. However, polynomial PCA is fraught with limiting practical considerations. First, it may be clear that linear PCA is not appropriate to use, but there is no ready method to determine what higher order polynomial PCA is most appropriate. Second, as $p$ and/or the degree of the polynomial increases, so does the dimensionality of the eigenanalysis and the number of replicates $n$ required for a nontrivial eigenvector solution, e.g. in the case of $p = 5$, for a quadratic PCA the eigenanalysis is 20-dimensional and it is required that $n \geq 20$, and for a cubic PCA, the eigenanalysis is 55-dimensional, and $n$ must exceed 55.

Another problem arising in using PCs is interpretation. The results of MANOVA on PCs, particularly if significance is indicated, are not particularly meaningful other than to indicate an overall 'treatment' effect unless each PC, or at very least the most important PCs, can be given clear interpretations. This is difficult when many of the original variables of a given PC have high loadings, or when the PCA is nonlinear.

### *Multidimensional scaling*

Either non-metric or hybrid MDS (N-MDS and H-MDS, respectively) can reduce effectively the dimensionality of multidimensional data that contain certain forms of non-linearity. Non-metric MDS has gained widespread acceptance among quantitative ecologists as a powerful and robust ordination technique (robust not in the classical statistical sense but in relating abundances of species to underlying ecological gradients; e.g. Fasham, 1977; Field *et al.*, 1987; Minchin, 1987). The method of hybrid MDS (hybrid in the sense of combining both non-metric and metric criteria; see Faith *et al.*, 1987) is slightly more robust in most circumstances (P. Minchin, pers. comm.) and has much to recommend it as a useful approach to handling non-linearity. However, for most practical purposes, N- and H-MDS yield similar results (P. Minchin, D. Faith and L. Belbin, unpubl. data). Both techniques are conceptually simple (but computationally complex) in that they seek to arrange objects, usually in 2- or 3-dimensional space, such that the distances between pairs of objects reflects, in some defined sense, their dissimilarities. MDS differs from PCA in that it only requires measures of dissimilarity among objects and does not require that objects can be positioned in multidimensional space (whether euclidean or otherwise). Dissimilarity can be defined by any of a host of available measures that are suited to different kinds of data (see Clifford & Stephenson, 1975; Field *et al.*, 1982; Krebs, 1989). However, different distance measures have widely disparate robustness (Faith *et al.*, 1987). Thus, effectiveness of MDS procedures to reduce dimensionality in a manner suitable for input to MANOVA depends more on the choice of dissimilarity measure than on whether N- or H-MDS is used.

An advantage to using N- or H-MDS (other than that of circumventing problems of non-linearity) is that a measure of dissimilarity can be chosen that best suits the data, e.g. the Bray-Curtis coefficient is a robust measure of dissimilarity (Faith *et al.*, 1987) that is usefully applied to species abundance data in which some species do not occur in most treatments (i.e. this measure is not affected by large numbers of zero counts), or to binary data (Field *et al.*, 1982). Note that flexibility in choosing a dissimilarity measure is not the exclusive domain of non-metric methods; any dissimilarity measure can also be used in principal coordinates analysis (PCoA), which is effectively a generalisation of PCA, i.e. a metric linear model. However, using measures of dissimilarity other than correlation or covariance structure also introduces a problem in that it becomes difficult to know exactly how data are 'distorted' in reduced space, which therefore obscures detailed interpretation of MANOVA.

Disadvantages to using MDS procedures are that, unlike PCA, the new axes are not interpretable in terms of the original variables, it is difficult to ascertain how well the configuration in reduced space represents the arrangement of objects in higher dimensional space (in situations where data can be defined in higher dimensional space) and the dimensionality of the reduced space is entirely an arbitrary decision (specified by the experimenter). Choosing too few dimensions can distort the arrangement of objects but choosing too many will ensure that some dimensions largely account for 'noise' in the data (P. Minchin, pers. comm.). A pragmatic approach is to compare results of MANOVA on objects described in 2- and 3-dimensional N-MDS space. Minor problems with MDS are that the new axes (which have an arbitrary scale) do not necessarily align with major trends in the configuration space and are not uncorrelated (so cannot be transformed independently). These problems are minor in that they can be rectified easily by rotating the MDS axes to align with the major trends using PCA, i.e. by conducting a PCA on the covariance matrix that describes the final configuration matrix from the N- or H-MDS. Rotation in this way is usually an option, and can therefore be conducted routinely, in commercial software packages.

## TRANSFORMATIONS

As in the univariate case, multivariate data can be transformed to minimise or prevent violations, obviating considerations of robustness. Monotone transformations are changes of scale undertaken to improve the efficiency and reliability of tests (for general discussion on transformations see for example Green, 1979; Steel & Torrie, 1981; Legendre & Legendre, 1983).

If violations are indicated, a useful first step is to check scatter plots of residuals that may identify outliers which may be corrected or discarded for *a priori* reasons. If violations are still evident it is prudent to search first for a transformation to stabilise variances. There are two reasons for this; first, violating the requirement of covariance homogeneity is far more serious than violating that of multinormality. Second, transformations that stabilise variances will often simultaneously improve normality of marginal distributions (multinormality requires normality of both marginal and conditional distributions but in practice often only the unidimensional marginal distributions are examined; see Legendre & Legendre, 1983).

Transformations may be made on original variables, or on PCs if these are to be used in MANOVA. The greatest problem likely to be encountered in transforming original ecological variables is that there are nearly always correlations among them (e.g. in species abundances), so the same single transformation must be applied to all variables even though it may not be suitable for some. Methods used for univariate data can be followed to find a suitable overall transformation (if it exists) to stabilise variances, viz. by examining the relationship among the $p \times k$ standard deviations ($\sigma$) and means ($\mu$) (i.e. $\sigma$'s and $\mu$'s of all variables from all groups). If dependence of variance on mean follows any one of the family of general relationships described by $\sigma \propto \mu^k$, then the appropriate variance stabilising transformation, in terms of the untransformed variate $Y$, is $Y^{1-k}$ (Draper & Smith, 1981; see their p. 238 for particular values of $k$).

When conducting MANOVA on PCs, data may be transformed before or after calculating the PCs, or both. In most cases it is preferable to conduct the PCA on the transformed original variables if an overall transformation is suitable. Because ecological variables with larger means tend to have higher variances, if the hypothesis dictates that variables with large scores should not dominate the analysis it is usually best to transform the original variables anyway. The transformations $X = \log(Y+1)$ (Green, 1979) or $X = \sqrt{\sqrt{Y}} = Y^{0.25}$ (Field *et al.*, 1982) are often applicable to species abundance data to prevent responses of abundant species from swamping those of rarer ones. However, although transformations may prevent abundant species from dominating the analysis, they may not make means independent of variances, in which case the PCs can also be transformed. Transforming before and after PCA complicates detailed interpretation of significant overall MANOVA results.

If the hypothesis is structured such that the experimenter is more interested in responses of common than of rare species then it is appropriate to transform the PCs and not the original variables. In this case a PCA on the covariance matrix of untransformed variables will weight the first PCs toward abundant species. An advantage of transforming the PCs is that since they are geometrically orthogonal (uncorrelated) and often approximately normal (because of the central limit theorem), they are often close to independence (Anderson, 1984), and it is therefore reasonable to use different transformations on each one. Special problems associated with transforming PCs are that the relationship between the group standard deviations and means of the PCs ($\sigma_{pc}$ and $\mu_{pc}$, respectively) can be symmetrical about $\mu_{pc} = 0$, which prevents transformation, and interpretation of significant MANOVA results can be difficult.

Finally, it must be realised that heteroscedasticity may exist in a form where variances and covariances are not dependent on means, in which case variance stabilising transformations cannot be found. For example, there may be negligible differences in means among groups, but large differences in variances and covariances. Although the statistical analysis may be difficult in this situation, ecologists should not despair, since variances often contain as much ecological information as do means. Identifying the offending group(s) and examining the data carefully may yield an ecologically meaningful interpretation.

## MULTIPLE COMPARISONS

MANOVA is properly viewed as a two-step procedure if the null hypothesis is rejected in an overall test. The second step is to conduct a multiple range test or simultaneous test procedure (STP) to determine the nature of the differences among groups. Several tests are used widely for comparing among treatments after univariate ANOVA (see Day & Quinn, 1989, for discussion), but equivalent multivariate tests are not so available or studied. A major problem in moving from the univariate to the multivariate case is that the number of possible comparisons grows dramatically. For example, for a 1-way MANOVA with 6 groups and 3 response variables, there are 398 component hypotheses of equality in subgroups of two or more groups on one or more variables (see example in Gabriel, 1968).

There are two classes of multiple comparisons available, those based on Scheffe-type methods which provide simultaneous tests or confidence intervals for all possible contrasts, and those based on Bonferroni methods which can be used when there is a fixed number of multiple comparisons to be tested. Several factors make Bonferroni techniques desirable; first, they are simple to construct in that individual tests for a particular comparison are carried out with an adjusted level of significance $\alpha_{\mathrm{adj}} = \alpha/L$, where $L$ is the total number of contrasts to be tested. Second, provided that $L$ is small or moderate in size, the confidence intervals from a Bonferroni method will typically be smaller than those obtained from a Scheffe-method, implying that Bonferroni methods are more powerful for detecting differences. Miller's text (1981, chapter 5) has a useful discussion of the thorny issue of the choice of a family of statistical statements for which we want to control the error rate, and simultaneous statistical inference.

A useful strategy for testing differences among groups on variables together or individually, is to limit the number of comparisons of interest and then use a Bonferroni procedure (see Day & Quinn, 1989, for discussion of the advantages of limiting the number of planned comparisons). If the contrasts of interest are restricted to pairwise comparisons on individual variables, we recommend using a good univariate STP, e.g. Tukey's studentised range (see Day & Quinn, 1989), with level $\alpha/p$, where $p$ is the number of dimensions. Alternatively, to test any contrast among groups over a combination of variables, a 2-sample t-test with adjusted $\alpha$ and the overall estimate of within-group variance ($\mathrm{SS}_{\mathrm{residuals}}/df$) can be used (see Miller, 1981, chapter 5, for details). In some situations the experimenter simply wants to know which groups differ without specifying the variables on which they differ, and in this case the desired contrasts are the pairwise comparisons on all $p$-variables. Hotelling's $T^2$-test for comparing two multivariate means can be used with a significance level $2\alpha/k(k\text{-}1)$, where $k$ is the number of groups, and with the overall estimate of within-group variation.

Among the Scheffe-type methods, which are recommended when all possible contrasts are of interest (usually unplanned), the consensus is that a STP based on Roy's largest root ($R$) is the most powerful (e.g. Wijsman, 1979; Bird &

Hadzi-Pavlovic, 1983). In fact it is the only procedure among the usual tests with the property that rejecting the null hypothesis guarantees that there is at least one contrast that will be rejected (Bird & Hadzi-Pavlovic, 1983). However, because of its lack of robustness, *R* cannot be recommended for a STP. A reasonable alternative procedure suggested by Bird & Hadzi-Pavlovic (1983) is to use the robust Pillai's statistic (*V*) for the overall test, and then to use *R* for the follow-up tests after a significant *V*-test. However, although this will give protection in the overall test, we have some concern that if we reject with *V*, then *R* may be excessively liberal in identifying differences among groups. Their reported Monte Carlo results provide some evidence that for this method the Type I error rate is reasonably controlled with small loss of power (see also Gabriel (1968) for details of follow-up tests using *R*).

In the context of discerning the nature of differences among groups, we recommend following up a significant MANOVA result with a canonical discriminant analysis (CDA; sometimes referred to as canonical variates analysis or multiple discriminant analysis, e.g. see Legendre & Legendre, 1983; Harris, 1985). In one sense CDA can be considered equivalent to graphical MANOVA (both procedures are based on eigenanalysis of the matrix $\mathbf{HE}^{-1}$, i.e. the matrix of variation between multivariate group means scaled by the within-group variation). CDA is an ordination technique that displays differences among multivariate group means in a reduced space in which the between-group variation is maximised relative to within-group variation. Plotting of group means (centroids) and their 95% confidence ellipsoids on the first 2–3 canonical axes is the multivariate equivalent of displaying univariate means and associated confidence intervals. In the same way that graphical display of univariate means and confidence intervals often indicates clearly the relationships among group means, plots of relative positions of multivariate group means with confidence ellipsoids in reduced canonical variates space can reveal similar information. Robustness of CDA when using confidence ellipsoids will be equivalent to robustness of MANOVA under the same conditions.

## ALTERNATIVES TO PARAMETRIC MANOVA

Data not appropriate for analysis by parametric MANOVA are those for which (1) levels of variance heterogeneity remain high despite preceding MANOVA with techniques to reduce dimensionality (e.g. PCA) and/or attempts to find transformations that stabilise variances, and/or (2) sample sizes are unequal and there is any amount of variance heterogeneity. In these situations alternatives to parametric MANOVA include robust non-parametric methods and non-inferential multivariate methods such as ordination and classification. Here is not the place to describe these methods in detail, but we point the reader to some helpful literature.

There has been some development of non-parametric MANOVA procedures. Mantel & Valand (1970) offer a test statistic suitable for two or more groups, but only for the 1-way case. However, whereas the permutation distribution of their statistic is conceptually simple, its computation is likely to be too complex in most problems. The method proposed by Bhargava (1972) is also limited to the 1-way case. Moreover, his procedure of comparing one group to the sum off all remaining groups is counter to our recommendations about maintaining a balanced sample size.

More promising are multiresponse permutation procedures (Biondini *et al.*, 1988), which are likely to emerge as a useful and powerful alternative to parametric MANOVA. This technique is distribution-free, does not require a linear data structure and has *P*-values which depend on the observed data and not on a hypothesised error distribution. However, it should be noted that for moderate to large data sets there will be substantial computational effort in obtaining *P*-values. For example, if there are 5 groups with 10 observations per group, the test statistic has to be computed for $50!/(5 \times (10!)) = 6.8 \times 10^{57}$ cases.

Also encouraging is the recent work by Rousseeuw & Leroy (1987). In making a MANOVA procedure less sensitive to deviations from assumptions, a key step is to use a robust covariance matrix estimate as a basis for the test rather than the classical covariance estimate used in the tests *V*, *W*, *T* and *R*. Rousseeuw & Leroy (1987) give a computationally feasible robust estimate. This robust covariance estimate can then be used to construct analogues of any of the classical tests. It will be most interesting to see the results of a Monte Carlo study undertaken for these tests.

Ordination and classification techniques are descriptive techniques that (generally) cannot be used for hypothesis testing (see Harris, 1985; also review of James & McCulloch, 1990 for introduction to relevant literature). They redescribe multidimensional data enabling relationships among groups to be viewed in lower (usually 2—3) dimensional space (see section Reducing Dimensionality, p. 202) or in the form of dendrograms. Two families of techniques applicable to data not suitable for parametric analysis are cluster analyses and nonmetric- or hybrid-multidimensional scaling (e.g. see Gnanadesikan, 1977; Field *et al.*, 1982; Gauch, 1982; Pielou, 1984; Austin, 1985; Faith *et al.*, 1987; Minchin, 1987). Both are based on indices of similarity or dissimilarity, of which there are many to choose from (e.g. Clifford & Stephenson, 1975; Krebs, 1989). The Bray-Curtis dissimilarity coefficient (Bray & Curtis, 1957) has much to recommend it for use with species abundance or presence/absence data (Field *et al.*, 1982; Faith *et al.*, 1987).

## SUMMARY OF RECOMMENDATIONS

(1) In designing multivariate experiments with a view to testing of hypotheses it is critical that sample sizes are equal, or almost equal if *n* is large, since in unbalanced designs none of the statistics is robust to variance heterogeneity.

(2) The number of response variables, and to a lesser extent the number of groups, in the design should be kept to a minimum in so far as the hypothesis permits. In many situations the deleterious effects of variance heterogeneity are exacerbated by increasing *p* and to a lesser extent *k*. Also, reducing dimensionality improves power in that it increases $df_{\text{error}}$ for a given *n*.

(3) Box's *M* statistic is rejected as a test of homogeneity of covariance matrices as it is much more sensitive to variance heterogeneity than is the *V* statistic that it is designed to protect, and is unacceptably sensitive to non-normality. The test statistic ($W_i$) of Hawkins (1981) is a practical alternative that tests for heteroscedasticity and non-normality simultaneously. Our simulations enabled calibrating the statistic such that its power to identify problematic data is commensurate with the robustness of *V* for the situations considered in this paper.

(4) If violations are indicated, data should be examined to determine whether it is possible to reduce dimensionality further by principal components analysis (PCA) or non-metric- or hybrid-multidimensional scaling (N- or H-MDS). PCA can be conducted on either the covariance matrix (or equivalently the total SSCP matrix) or correlation matrix describing the relationship between all objects in $p$-dimensional space, depending on the hypothesis and nature of the data. It is often appropriate to conduct the MANOVA on $m < p$ principal components (PCs) that account for most of the total variance. Other advantages of using PCA prior to MANOVA are that normality is usually improved, and that different transformations can be used on each PC. PCA should not be used if it does not provide a parsimonious description of variance structure. If linear PCA is not effective in reducing dimensionality, higher order polynomial PCA may provide an efficient description of data, or alternatively, N- or H-MDS routines may be useful. Since the axes of N- or H-MDS space are unlikely to align with major trends in the distribution of objects in the reduced space, rotation of the MDS axes using PCA (i.e. to align them with major trends) should be undertaken routinely prior to MANOVA.

(5) Multivariate data can often be transformed to improve agreement with assumptions. Wherever possible, transformations that stabilise variances should be used. Since it cannot be assumed that ecological variables are uncorrelated the same transformation should be applied to all original variables. If PCs are used in MANOVA, data can be transformed before or after calculation of PCs, or both. It is preferable to rescale data before PCA if a suitable overall transformation exists. For species abundance data, original variables can first be transformed prior to PCA to prevent abundant species from dominating the analysis and if necessary the PCs can then be transformed to stabilise variances.

(6) On the basis that Type I errors are more serious than Type II errors the $V$ statistic is recommended for general use. For Type I errors, the violation of the assumption of homoscedasticity is more serious than is non-normality, and the $V$ statistic is clearly the most robust to variance heterogeneity in terms of controlling level. Its rates of Type I error are acceptable at moderate levels of heteroscedasticity providing sample sizes are equal and it is the most powerful of the four statistics when non-centrality is diffuse (i.e. when group centroids are not described by a single major trend). Non-normality (kurtosis) has a much smaller impact on Type I error rates in making all tests slightly more conservative, but it reduces the power of all statistics considerably. Loss of power is dramatic if assumptions of normality and homoscedasticity are violated simultaneously.

For large samples (approximately when $(df_{\text{error}}/df_{\text{hypothesis}}) \geq 10p$) the robustness of $W$ and $T$ compares to $V$, and if non-centrality is concentrated (i.e. group centroids align along a single main axis), then $T$ and $W$ are more powerful than $V$ and would be the preferred choice. In the special circumstance in which non-centrality is concentrated and there are no violations, or if departures from homogeneity are small and dimensionality of the data can be reduced to two dependent variables, then $R$ is the statistic of choice given its superior power for this distribution of non-centrality.

(7) MANOVA is a two-step procedure if the overall test indicates significance. The second step is to discern the nature of the differences among means. The preferred approach is to use Bonferroni-type methods in which the level of

significance $\alpha$ is adjusted to $\alpha/L$, where $L$ (the total number of comparisons) is limited to the fewest possible contrasts of interest. If all possible comparisons are required, an alternative is to use the $V$ statistic in the overall test, and the $R$ statistic in a follow-up simultaneous test procedure. Interpretation of a significant result in MANOVA can be aided greatly by canonical discriminant analysis which is analogous to a graphical MANOVA.

(8) Classical parametric MANOVA should not be used with data in which high levels of variance heterogeneity cannot be overcome by transformation and/or reduction of dimensionality, or in which sample sizes are unequal and assumptions are not satisfied. Viable alternatives are to use a robust covariance matrix estimate rather than the classical covariance estimates used in the $V$, $W$, $T$ and $R$ tests, or permutation techniques. Non-metric non-inferential multivariate procedures such as some ordination and classification methods may facilitate useful interpretation.

## ACKNOWLEDGEMENTS

We are indebted to Carla Caverhill and Randy Currie for their enthusiasm in writing the programs and assistance in writing the plotting routines, to Brian McCardle for extensive but constructive criticism and discussion of the manuscript, and to Mike Dale, Dan Faith and particularly Peter Minchin for discussion about ordination procedures. The simulations were undertaken while CRJ was in receipt of a Visiting Fellowship from the Department of Fisheries and Oceans Canada at the Bedford Institute of Oceanography, Dartmouth, Nova Scotia. While reviewing the literature CRJ was supported by an AIMS Postdoctoral Fellowship at the Australian Institute of Marine Science, Townsville, Queensland (AIMS contribution No. 538).

## REFERENCES

Anderson, T. W., 1984. *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, 2nd edition, 675 pp.

Anderson, T. W. & Darling, D. A., 1954. A test of goodness of fit. *Journal of the American Statistical Association*, **49**, 765–769.

Austin, M. P., 1985. Continuum concept, ordination methods, and niche theory. *Annual Review of Ecology and Systematics*, **16**, 39–61.

Barker, H. R. & Barker, B. M., 1984. *Multivariate Analysis of Variance (MANOVA) A Practical Guide to its Use in Scientific Decision Making*. University of Alabama Press, Alabama, 129 pp.

Bhargava, R. P., 1972. A test for equality of means of multivariate normal distributions when covariances are unequal. *Calcutta Statistical Association Bulletin*, **20**, 153–156.

Biondini, M. E., Mielke, P. W. & Redente, E. F., 1988. Permutation techniques based on Euclidean analysis spaces: a new and powerful statistical method for ecological research. *Coenoses*, **3**, 155–174.

Bird, K. D. & Hadzi-Pavlovic, D., 1983. Simultaneous test procedures and the choice of a test statistic in MANOVA. *Psychological Bulletin*, **93**, 167–178.

Box, G. E. P., 1949. A general distribution theory for a class of likelihood criteria. *Biometrika*, **36**, 317–346.

Bray, J. R. & Curtis, J. T., 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, **27**, 325–349.

Clifford, H. T. & Stephenson, W., 1975. *An Introduction to Numerical Classification*. Academic Press, New York, 229 pp.

Cooley, W. W. & Lohnes, P. R., 1971. *Multivariate Data Analysis*. Wiley, New York, 364 pp.

Day, R. W. & Quinn, G. P., 1989. Comparisons of treatments after an analysis of variance in ecology. *Ecological Monographs*, **59**, 433–463.

Draper, N. R. & Smith, H., 1981. *Applied Regression Analysis*. Wiley, New York, 709 pp.

Faith, D. P., Minchin, P. R. & Belbin, L., 1987. Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio*, **69**, 57–68.

Fasham, M. J. R., 1977. A comparison of nonmetric multidimensional scaling, principal components and reciprocal averaging for the ordination of simulated coenoclines, and coenoplanes. *Ecology*, **58**, 551–561.

Field, J. G., 1971. A numerical analysis of changes in the soft-bottom fauna along a transect across False Bay, South Africa, *Journal of Experimental Marine Biology and Ecology*, **7**, 215–253.

Field, J. G., Clarke, K. R. & Warwick, R. M., 1982. A practical strategy for analysing multispecies distribution patterns. *Marine Ecology Progress Series*, **8**, 37–52.

Gabriel, K. R., 1968. Simultaneous test procedures in multivariate analysis of variance. *Biometrika*, **55**, 489–504.

Gauch, H. G., 1982. *Multivariate Analysis in Community Ecology*. Cambridge University Press, Cambridge, 298 pp.

Glass, G. V., Peckham, P. D. & Sanders, J. R., 1972. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, **42**, 237–288.

Gnanadesikan, R., 1977. *Methods for Statistical Data Analysis of Multivariate Observations*. Wiley, New York, 311 pp.

Green, R. H., 1979. *Sampling Design and Statistical Methods for Environmental Biologists*. Wiley, New York, 257 pp.

Hair, J. F., Anderson, R. E. & Tatham, R. L., 1987. *Multivariate Data Analysis*. Macmillan, New York, 2nd edition, 449 pp.

Harris, R. J., 1985. *A Primer of Multivariate Statistics*. Academic Press, Orlando, 2nd edition, 576 pp.

Hawkins, D. M., 1981. A new test for multivariate normality and homoscedasticity. *Technometrics*, **23**, 105–110.

Hopkins, J. W. & Clay, P. P. F., 1963. Some empirical distributions of bivariate $T^2$ and homoscedasticity criterion $M$ under unequal variance and leptokurtosis. *Journal of the American Statistical Association*, **58**, 1048–1053.

Hull, C. H. & Nie, N. H., 1981. *SPSS Update 7–9*. McGraw-Hill, New York, 402 pp.

Hurlbert, S. H., 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, **54**, 187–211.

Ito, K. & Schull, W. J., 1964. On the robustness of the $T_0^2$ test in multivariate analysis of variance when variance-covariance matrices are not equal. *Biometrika*, **51**, 71–82.

Ito, P. K., 1980. Robustness of ANOVA and MANOVA test procedures. In, *Handbook of Statistics, Volume 1*, edited by P. R. Krishnaiah, North-Holland, New York, pp. 199–236.

James, F. C. & McCulloch, C. E., 1990. Multivariate analysis in ecology and systematics: panacea or Pandora's box. *Annual Review of Ecology and Systematics*, **21**, 129–166.

Johnson, C. R. & Mann, K. H., 1986. The crustose coralline alga, *Phymatolithon* Foslie, inhibits the overgrowth of seaweeds without relying on herbivores. *Journal of Experimental Marine Biology and Ecology*, **96,** 127–146.

Johnson, C. R. & Mann, K. H., 1988. Diversity, patterns of adaptation, and stability of Nova Scotian kelp beds. *Ecological Monographs*, **58,** 129–154.

Korin, B. P., 1972. Some comments on the homoscedasticity criterion $M$ and the multivariate analysis of variance tests $T^2$, $W$ and $R$. *Biometrika*, **59,** 215–216.

Krebs, C. J., 1989. *Ecological Methodology*. Harper & Row, New York, 654 pp.

Krzanowski, W. J., 1988. *Principles of Multivariate Analysis: A User's Perspective*. Clarendon Press, Oxford, 563 pp.

Lee, Y. S., 1971. Asymptotic formulae for the distribution of a multivariate test statistic: power comparisons of certain multivariate tests. *Biometrika*, **58,** 647–651.

Legendre, L. & Legendre, P., 1983. *Numerical Ecology*. Elsevier, Amsterdam, 419 pp.

Mantel, N. & Valand, R. S., 1970. A technique of nonparametric multivariate analysis. *Biometrics*, **26,** 547–558.

Mardia, K. V., 1971. The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika*, **58,** 105–121.

Marriot, F. H. C., 1974. *The Interpretation of Multiple Observations*. Academic Press, London, 117 pp.

Miller, R. G., 1981. *Simultaneous Statistical Inference*. Springer-Verlag, New York, 299 pp.

Minchin, P. R., 1987. An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio*, **69,** 89–107.

Morrison, D. F., 1976. *Multivariate Statistical Methods*. McGraw-Hill, New York, 415 pp.

Noy-Meir, I., Walker, D. & Williams, W. T., 1975. Data transformations in ecological ordination. II. On the meaning of data standardization. *Journal of Ecology*, **63,** 779–800.

Olson, C. L., 1974. Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, **69,** 894–908.

Olson, C. L., 1976. On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, **83,** 579–586.

Olson, C. L., 1979. Practical considerations in choosing a MANOVA test statistic: a rejoinder to Stevens. *Psychological Bulletin*, **86,** 1350–1352.

Pielou, E. C., 1984. *The Interpretation of Ecological Data*. Wiley, New York, 263 pp.

Pillai, K. C. S. & Jayachandran, K., 1967. Power comparisons of tests of two multivariate hypotheses based on four criteria. *Biometrika*, **54,** 195–210.

Press, S. J., 1972. *Applied Multivariate Analysis*. Holt, Rinehart & Winston, New York, 521 pp.

Rousseeuw, P. J. & Leroy, A. M., 1987. *Robust Regression and Outlier Detection*. Wiley, New York, 329 pp.

Schatzoff, M., 1966. Sensitivity comparisons among tests of the general linear hypothesis. *Journal of the American Statistical Association*, **61,** 415–435.

Scheffe, H., 1959. *The Analysis of Variance*. Wiley, New York, 477 pp.

Sokal, R. R. & Rohlf, F. J., 1981. *Biometry*. Freeman, San Francisco, 2nd edition, 859 pp.

Srivastava, A. B. L., 1959. Effect of nonnormality on the power of the analysis of variance test. *Biometrika*, **46,** 114–122.

Srivastava, M. S. & Carter, E. M., 1983. *An Introduction to Applied Multivariate Statistics*. North-Holland, New York, 394 pp.

Steel, R. G. D. & Torrie, J. H., 1981. *Principles and Procedures of Statistics*. McGraw Hill, Singapore, 2nd edition, 633 pp.

Stevens, J., 1979. Comment on Olson: choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, **86,** 355–360.

Stevens, J., 1986. *Applied Multivariate Statistics for the Social Sciences*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 515 pp.

Stewart-Oaten, A., Murdoch, W. M. & Parker, K. R., 1986. Environmental impact assessment: 'pseudoreplication' in time? *Ecology*, **67**, 929–940.

Tabachnick, B. G. & Fidell, L. S., 1989. *Using Multivariate Statistics*. Harper & Row, New York, 2nd edn, 746 pp.

Tatsuoka, M. M., 1988. *Multivariate Analysis*. Macmillan, New York, 2nd edition, 479 pp.

Tiku, M. L., 1971. Power function of the *F*-test under nonnormal situations. *Journal of the American Statistical Association*, **66**, 913–916.

Underwood, A. J., 1981. Techniques of analysis of variance in experimental marine biology and ecology. *Oceanography and Marine Biology: an Annual Review*, **19**, 513–605.

Underwood, A. J., 1990. Experiments in ecology and management: their logics, functions and interpretations. *Australian Journal of Ecolology*, **15**, 365–389.

Wijsman, R., 1979. Constructing all smallest simultaneous confidence sets in a given class with applications to MANOVA. *Annals of Statistics*, **7**, 1003–1018.

# APPENDIX 1

## GLOSSARY OF TERMS

*Balance*; data are balanced when sample sizes are equal in all groups. In an *unbalanced* design the number of samples for each group is unequal.

*Centroid*; the point in multivariate space defined by the group mean vector, i.e. coordinates are the mean values of all variables in that group.

*Covariance matrix*; square and symmetric matrix giving variances on the diagonal and covariances on the off-diagonals.

*Dimensionality*; the number of response (= dependent) variables.

*Dispersion matrix*; see covariance matrix.

*Exceedance*; estimate of the probability (= percentage or proportion of occasions) that a test statistic exceeds the nominal significance level ($\alpha$) to indicate a significant difference.

*Fixed effects model*; ANOVA model in which the different levels of a factor are fixed treatments determined by the experimenter, i.e. the concern is with differences among means that can be ascribed to factors whose levels ('categories') are fixed (cf. random effects model).

*Heteroscedasticity*; heterogenous variance structure, e.g. inequality of covariance matrices among groups (cf. homoscedasticity).

*Homoscedasticity*; homogeneous variance structure, e.g. equality of covariance matrices among groups (cf. heteroscedasticity).

*Kurtosis*; symmetrical deviation from the normal distribution; a *leptokurtic* distribution is symmetrical but has more values around the mean and the tails of the distribution than the normal; a *platykurtic* distribution is symmetrical but has more values between the mean and the tails than the normal (cf. skewness).

*Level*; the rate of Type I error (see Type I error, significance level).

*Model I*; see fixed effects model.

*Model II*; see random effects model.

*Noncentrality*; differences among group mean vectors; non-centrality is said to be *concentrated* when group centroids are arranged along a single dimension in multivariate space and *diffuse* when the centroids are spread almost equally in all dimensions.

*Normality*; data are distributed normally if their frequency distribution is described by a normal or Gaussian distribution; data are *non-normal* if their frequency distribution does not fit a normal curve. Note that multinormality (normality of multivariate data) requires normality of both marginal (= individual univariate) distributions and conditional distributions. *Non-normal* data can be described by *skewness* and *kurtosis*.

*Power*; the probability of detecting real deviations from the null hypothesis, i.e. real differences among group means. It is defined as 1-ß where ß = rate of Type II error.

*Random effects model*; ANOVA model in which different levels of a factor are viewed as a random sample of a population of all possible levels of the factor (cf. fixed effects model).

*Repeated measures*; repeated measurement of the same experimental individuals through time, i.e. measurements through time are not independent.

*Significance level*; rate of Type I error set by and acceptable to the experimenter, symbolised by $\alpha$.

*Skewness*; asymmetric deviation from the normal distribution; i.e. one tail of the distribution is extended further than the other (cf. *kurtosis*).

*Type I error*; rejecting the null hypothesis when it is true, i.e. incorrectly claiming a significant test result. Usually expressed as a probability symbolised by $\alpha$. See significance level (cf. Type II error).

*Type II error*; accepting the null hypothesis when it is false, i.e. incorrectly claiming a non-significant test result. Usually expressed as a probability symbolised by ß (cf. Type I error).

# APPENDIX 2

## THE MANOVA TEST CRITERIA

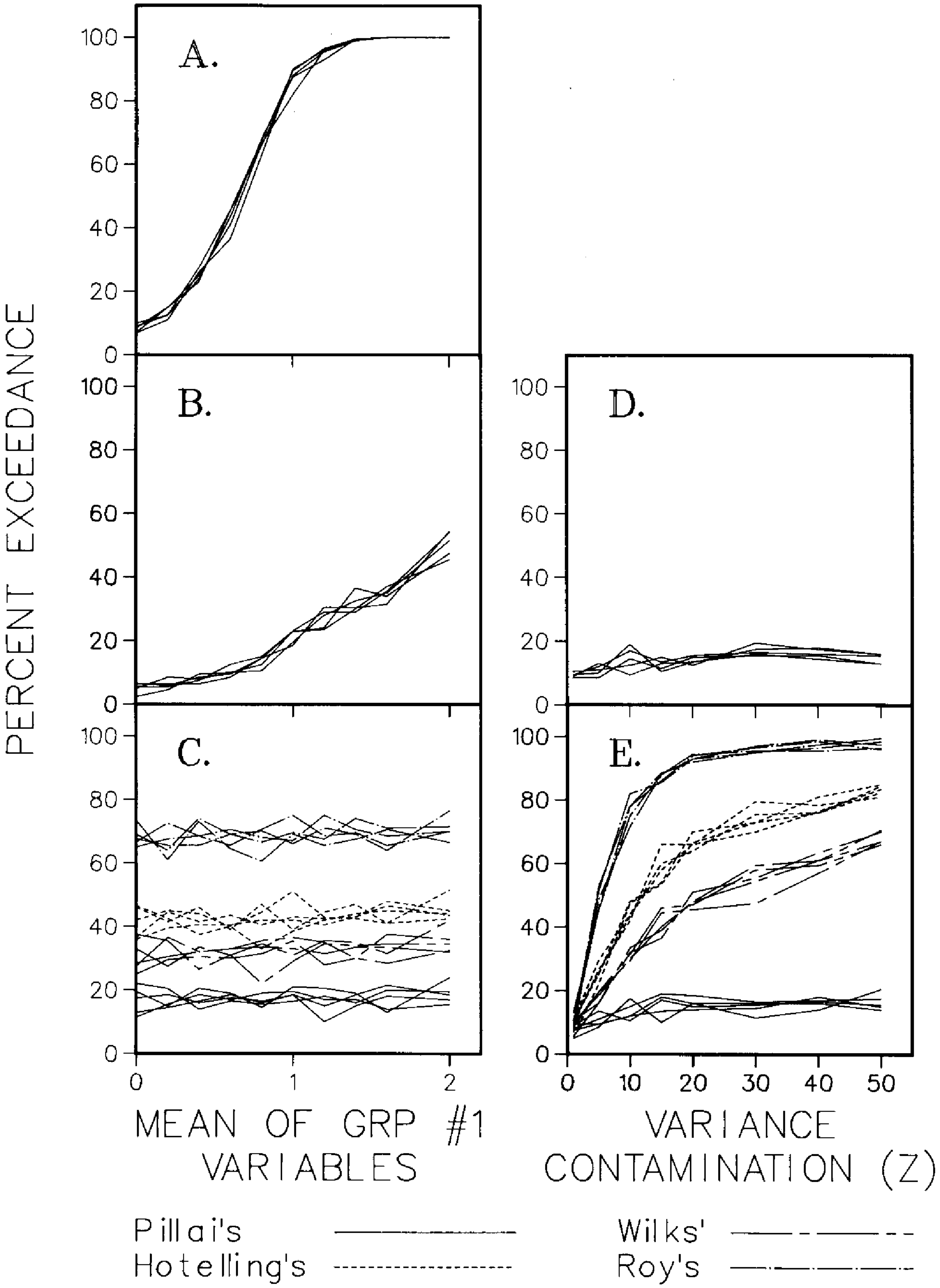
The four statistics are functions of the non-zero eigenvalues ($\lambda_i$) of $\mathbf{HE}^{-1}$, where **H** and **E** are the $p \times p$ sums-of-squares-and-cross-products matrices for hypothesis (among groups) and error (within groups) respectively (i.e. the multivariate equivalents of the hypothesis and error mean squares in the univariate case). The test statistics are defined as (for further information see Olson, 1974; Ito, 1980; Hull & Nie, 1981, and references):

Pillai's trace $V$ = trace of $\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}$ = sum of $\lambda_i/(1+\lambda_i)$
Hotelling's trace $T$ = trace of $\mathbf{HE}^{-1}$ = sum of $\lambda_i$
Wilks' lambda $W$ = determinant of $\mathbf{E}(\mathbf{H}+\mathbf{E})^{-1}$ = product of $\lambda_i/(1+\lambda_i)$
Roy's largest root $R$ = largest eigenvalue of $\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1} = \lambda_p/(1+\lambda_p)$, where $\lambda_p$ is the largest eigenvalue of $\mathbf{HE}^{-1}$.

Fig 10.—Demonstration that Monte Carlo simulations provide unambiguous results. Data are of five repeated simulations for each of three examples of power curves (A–C) and two of rates of Type I error (D and E). They show that simulations were reliable, i.e. that independent replicate simulations, each of 200 runs, gave quantitatively similar and qualitatively identical results. When the different statistics gave similar results, only values for Pillai's criterion are given (A, B and D). The simulation conditions were as follows: A, Power curve; data are balanced, normal and homoscedastic; $n = 10$, $k = 3$, $p = 2$. B, Power curve; data are balanced, non-normal and homoscedastic; $n = 10$, $k = 3$, $p = 2$. C, Power curve; data are balanced, non-normal and heteroscedastic; $n = 10$, $k = 6$, $p = 5$. D, Rates of Type I error; data are balanced and normal; $n = 10$, $k = 3$, $p = 2$. E, Rates of Type I error; data are balanced and normal; $n = 10$, $k = 10$, $p = 10$; where $n$ = sample size, $k$ = number of groups, $p$ = number of variables.

# APPENDIX 3

## RELIABILITY OF MONTE CARLO SIMULATIONS



# APPENDIX 4

## PROOF THAT MANOVA STATISTICS *V*, *T*, *W* AND *R* ARE AFFINE INVARIANT

The proof that MANOVA statistics *V*, *T*, *W* and *R* are affine invariant (i.e. behave identically whether dispersion matrices are in raw or canonical form) justifies use of diagonal matrices in the Monte Carlo simulations.

Let **Y** denote the original observations and **Z** the transformed observations where $\mathbf{Z} = \mathbf{MY}$ for some choice of **M**. If $\mathbf{\Sigma}_y$ is the covariance matrix of **Y** then **M** is chosen so that $\mathbf{M\Sigma}_y\mathbf{M}^T$ is diagonal. The covariance matrix $\mathbf{\Sigma}_z$ is given by $\mathbf{M\Sigma}_Y\mathbf{M}^T$.

The next step is to show that the test criteria do not change if we transform **Y** to **Z**. Let **H** and **E** be the sum-of-squares-and-cross-products matrices for hypothesis (between groups) and error (within groups) respectively, defined as:

$$\mathbf{H}_Y = \sum_{i=1}^{k} n_i\,(\mathbf{Y}_i - \mathbf{Y})(\mathbf{Y}_i - \mathbf{Y})^T$$

$$\mathbf{H}_Z = \sum_{i=1}^{k} n_i\,(\mathbf{Z}_i - \mathbf{Z})(\mathbf{Z}_i - \mathbf{Z})^T$$

$$\mathbf{E}_Y = \sum_{i=1}^{k} \sum_{j=1}^{n} n_i\,(\mathbf{Y}_{ij} - \mathbf{Y}_i)(\mathbf{Y}_{ij} - \mathbf{Y}_i)^T$$

$$\mathbf{E}_Z = \sum_{i=1}^{k} \sum_{j=1}^{n} n_i\,(\mathbf{Z}_{ij} - \mathbf{Z}_i)(\mathbf{Z}_{ij} - \mathbf{Z}_i)^T$$

Since the test criteria are based on functions of the eigenvalues of $\mathbf{HE}^{-1}$, to show equivalence it suffices to show that $\mathbf{H}_Y\mathbf{E}_Y{}^{-1}$ and $\mathbf{H}_Z\mathbf{E}_Z{}^{-1}$ have the same eigenvalues and that $\mathbf{H}_Y(\mathbf{H}_Y + \mathbf{E}_Y)^{-1}$ and $\mathbf{H}_Z(\mathbf{H}_Z + \mathbf{E}_Z)^{-1}$ have the same eigenvalues. Now

$$\mathbf{H}_Z = \sum_{i=1}^{k} n_i\,(\mathbf{MY}_i - \mathbf{MY})(\mathbf{MY}_i - \mathbf{MY})^T$$

$$= \mathbf{M} \sum_{i=1}^{k} n_i\,(\mathbf{Y}_i - \mathbf{Y})(\mathbf{Y}_i - \mathbf{Y})^T$$

$$= \mathbf{MH}_Y\mathbf{M}^T$$

and it can be similarly shown that

$$\mathbf{E}_Z = \mathbf{ME}_Y\mathbf{M}^T$$

and

$$\mathbf{H}_Z + \mathbf{E}_Z = \mathbf{M}(\mathbf{H}_Y + \mathbf{E}_Y)\mathbf{M}^T$$

and therefore that

$$\mathbf{H}_Z\mathbf{E}_Z{}^{-1} = \mathbf{MH}_Y\mathbf{M}^T\mathbf{M}^{T-1}\mathbf{E}_Y{}^{-1}\mathbf{M}^{-1} = \mathbf{MH}_Y\mathbf{E}_Y{}^{-1}\mathbf{M}^{-1}$$

and

$$\mathbf{H}_Z(\mathbf{H}_Z + \mathbf{E}_Z)^{-1} = \mathbf{MH}_Y(\mathbf{H}_Y + \mathbf{E}_Y)^{-1}\mathbf{M}^{-1}$$

Thus, to complete the verification we need only to show that a matrix **A** and a matrix $\mathbf{BAB}^{-1}$ have the same eigenvalues. The eigenvalues of **A** are found by solving the determinantal equation $|\mathbf{A} - \lambda\mathbf{I}| = 0$, and similarly, those of $\mathbf{BAB}^{-1}$ by solving the equation

$|\mathbf{BAB}^{-1}-\lambda\mathbf{I}| = 0$

or equivalently $|\mathbf{B}|^{-1}|\mathbf{BAB}^{-1}-\lambda\mathbf{I}| = 0$

and therefore $|\mathbf{AB}^{-1}-\lambda\mathbf{B}^{-1}| = 0$

or equivalently $|\mathbf{AB}^{-1}-\lambda\mathbf{B}^{-1}|\ |\mathbf{B}| = 0$

and therefore $|\mathbf{A}-\lambda\mathbf{I}| = 0$, which implies that **A** and $\mathbf{BAB}^{-1}$ have the same eigenvalues, and therefore that the test statistics give the same value for **Y** and **Z**.

If we wish to model the situation of the mixture of two normals as we have done in the paper, it suffices to use a normal mixture of $N(\mathbf{0},\mathbf{I})$ and $N(\mathbf{0},\mathbf{D}\lambda)$. To see this note that if **Y** is a mixture of $N(\mathbf{0},\mathbf{V}_1)$ and $N(\mathbf{0},\mathbf{V}_2)$ there exists a matrix **M**, $|\mathbf{M}| = 0$, such that $\mathbf{MV}_1\mathbf{M}^{\mathrm{T}} = \mathbf{I}$, and $\mathbf{MV}_2\mathbf{M}^{\mathrm{T}} = \mathbf{D}\lambda$, where $\mathbf{D}\lambda$ is a diagonal matrix whose elements $\lambda_i$ are the latent roots of $|\mathbf{V}_2-\lambda\mathbf{V}_1| = 0$ (Press, 1972), or equivalently of $|\mathbf{V}_2\mathbf{V}_1{}^{-1}-\lambda\mathbf{I}| = 0$, i.e. are the eigenvalues of $\mathbf{V}_2\mathbf{V}_1{}^{-1}$. Note that although the above discusses the multivariate normal, the argument works for a mixture from any distribution since the results are based only on the covariance matrix.