Ann Ist Super Sanità 2004;40(2):165-171

Detecting assumption violations in mixed-model analysis of variance

Flavia CHIAROTTI

Dipartimento di Biologia Cellulare e Neuroscienze, Istituto Superiore di Sanità, Rome, Italy

Summary. - Parametric analysis of variance (ANOVA) is frequently used to analyse experimental data, yet for the results to be considered as accurate, certain assumptions must be respected: the normality of the distribution of the sampled data, the homogeneity of variance among the groups being compared (i.e., homoscedasticity), and, in certain cases, sphericity. The present work focuses on the methods for detecting violations of these assumptions and provides an example of the application of these methods.

Key words: normality, homoscedasticity, sphericity, analysis of variance, mixed-model.

Riassunto *(Verifica della violazione degli assunti in un modello misto di analisi della varianza)*. - L'analisi della varianza (ANOVA) parametrica è uno dei metodi più frequentemente utilizzati per l'analisi di dati sperimentali. I risultati dell'ANOVA parametrica sono corretti se gli assunti su cui l'ANOVA si basa sono rispettati. Tali assunti comprendono la normalità della distribuzione dei dati, l'omogeneità di varianza (omoschedasticità) nei gruppi a confronto e, nel caso di specifici disegni sperimentali, la sfericità. Gli assunti dell'ANOVA vengono descritti insieme ai principali metodi (sia descrittivi sia inferenziali) per la loro verifica. Un esempio illustra l'applicazione delle procedure descritte. Infine, vengono forniti alcuni suggerimenti su una procedura a più passi per l'analisi dei dati con l'ANOVA parametrica.

Parole chiave: normalità, omoschedasticità, sfericità, analisi della varianza, modello misto.

Introduction

The results of parametric analysis of variance (ANOVA), one of the most commonly used methods for analysing experimental data, can be considered as accurate if the assumptions on which the analysis is based are respected, specifically: the normality of the distribution of the sampled data, the homogeneity of variance among the groups being compared (i.e., homoscedasticity), and, for certain experimental designs, sphericity. In the present work, ANOVA assumptions are described, together with methods for detecting violations of these assumptions (i.e., descriptive statistics and significance tests) and an example of the application of these methods. Suggestions on the multi-step procedure for analysing data with parametric ANOVA are also provided.

Experimental designs

There are three main types of experimental designs: a) completely randomised designs; b) randomised block designs; and c) split-plot designs.

Completely randomised designs

In completely randomised designs (CRD), a random sample of units is extracted from a population and then randomly divided into two or more subgroups, each of which is assigned to a different treatment (one-factor CRD) or to a combination of treatments (factorial CRD). Treatment factor is usually a fixed effect factor. The sample size in the different subgroups can be equal (balanced design) or not (unbalanced design), and although balancing is not required, it is desirable.

Randomised block designs

In randomised block designs (RBD), a random sample of blocks is extracted from a population, with each block consisting of more than one unit. The units within each block are then randomly assigned to different treatments (one-factor RBD) or to combinations of treatments (factorial RBD). Treatment factor is usually a fixed effect factor, whereas blocking factor is a random effect factor. The sample size in the

Indirizzo per la corrispondenza (Address for correspondence): Flavia Chiarotti, Dipartimento di Biologia Cellulare e Neuroscienze, Istituto Superiore di Sanità, V.le Regina Elena 299, 00161 Rome. Italy. E-mail: flavia.chiarotti@iss.it.

different subgroups is necessarily equal (balanced design), unless there are missing values in the experiment.

A particular type of RBD is the repeated measure design (RMD), in which a random sample of units is extracted from a population and each unit is repeatedly assigned to different treatments (one-factor RMD) or to combinations of treatments (factorial RMD), with the treatment administered in a randomised order within each unit. In some cases, such as when the effect of time on the response variable is to be evaluated, the units can receive the same treatment, which is evaluated at different times.

Split-plot designs

Split-plot designs are a combination of CRD and RBD. A random sample of blocks, each consisting of more than one unit, is extracted from a population. The blocks are randomly divided into two or more subgroups, each assigned to a different treatment or to a combination of treatments (between-subject factor/s). The units within each block are randomly assigned to different treatments or to combinations of treatments (within-subject factor/s) or are evaluated at different times (repeated measures). Between-subject, within-subject, and repeated-measure factors are usually fixed effect factors, whereas blocks and units are random effect factors.

For split-plot designs, the sample size in the final cells can be quite small (usually ranging from 6 to 12 units). When the experiment is designed as a complete split-litter, only one unit is included in each final cell.

The small sample size in the final cells can reduce the robustness of parametric ANOVA because of departures from the assumptions of normality and homoscedasticity. This can produce levels of significance that are far from the nominal ones, either furnishing significant results when the null hypothesis H_0 is true (Type I error = loss of validity) or no significant results when the null hypothesis H_0 is false (Type II error = loss of efficiency). The small final sample size can also make it difficult to detect violations of the assumptions.

Normality

The assumption of normality implies that the distribution of the variable to be analysed by ANOVA is normal in the population from which units (or blocks of units) are sampled. It is also assumed that the treatments administered to the units (and/or blocks of units) in the different subgroups do not affect the shape of the distribution of the variable and that only the mean value of the variable changes among subgroups.

Unimodal frequency distributions are characterised by two parameters: skewness and kurtosis. Skewness describes the asymmetry of a distribution. It is equal to 0 for unimodal symmetrical distributions, such as the normal distribution, whereas it is negative for unimodal distributions with a longer left tail (towards lower values of the variable) and positive for distributions with a longer right tail (towards higher values). Kurtosis describes the steepness of the unimodal frequency curve towards the mode. As proposed by Karl Pearson in 1906, it is usually measured by the moment-ratio $\beta_2 = \mu_4/\mu_2^2$. The normal distribution is characterised by a kurtosis index equal to 3, which is adopted as a standard for other distributions. However, for the sake of simplicity, the kurtosis index β_2 is usually centred by subtracting 3, so that it is equal to 0 in normal distributions. Distributions with a positive, 0, and negative kurtosis index are referred to as, respectively, “leptokurtic” (heavy-tailedness), “mesokurtic”, and “platykurtic” (light-tailedness).

ANOVA is considered to be robust against slight violations of the normality assumption. For example, in CRD the F test is not seriously affected by either moderate skewness, unless the design is unbalanced, or by light- or heavy-tailedness, unless the sample sizes in the final cells are very small (less than 5) or kurtosis is extreme (less than -1 or greater than 2). For these reasons, and because it is difficult to determine the shape, skewness, and kurtosis of the distribution of the variables in advance, violations of the normality assumption should be assessed before applying ANOVA to experimental data. To do so, descriptive statistics, diagnostic plots, and significance tests can be used.

Descriptive statistics include the above-mentioned skewness and kurtosis indexes, each of which can be approximately tested for deviations from normality by dividing it by its standard error. The ratio can be roughly read as a standardised score derived from a normal distribution, with absolute values exceeding 2 expected to be rare in normal samples. These indexes can also be tested using significance tests (see below). Diagnostic plots (i.e., boxplots and normal probability plots) can also provide information on normality. A boxplot is a graph that summarises the distribution of a set of data, providing information on the median and the mean (represented by, respectively, a straight line and a symbol within the box), the first and third quartiles (bottom and top lines of the box, respectively), and outliers (symbols outside of the box). Whiskers can be present and can cover the entire range of data or a defined percentile range. If the distribution is normal, the mean and median are equal and in the middle of the box, outliers are not present, and the whiskers are symmetrical with respect to the box. The presence of outliers on only one side of the

box suggests that skewness is present in the distribution of data, whereas if appearing on both sides they suggest the presence of heavy-tailedness. Skewness is also likely when the mean and median do not coincide and are not in the middle of the box, or when the whiskers are not symmetrical.

Normal probability plots include Q-Q plots and P-P plots. In Q-Q plots, the quantiles of the variable in the sample (y) are plotted against the expected quantiles of normal distribution (x). In P-P plots, the normal cumulative distribution function of the standardised variable is plotted against the empirical cumulative distribution function of the variable. Q-Q plots highlight non-normality in the tails of the distribution, whereas P-P plots highlight non-normality in the central part of the distribution. For both plots, if the data in the sample come from a normal distribution, the points on the plot should form a relatively straight line.

Significance tests provide a significance level that, if low (typically less than 0.05), indicates a departure from normality. The departure can be generic (Kormogorov-Smirnov, Shapiro-Wilk, Shapiro-Francia, D'Agostino-Pearson, Stephens) or specific (D'Agostino's test for skewness, Anscombe-Glynn test for kurtosis). The results of significance tests for normality can be affected by the presence of outliers, which can cause a significant result to be obtained even if the remainder of the units come from a population with a normal distribution. The results can also depend on the sample size: very large samples are likely to be detected as non-normal even for unimportant departures from normality, whereas small samples (less than 10 units) are unlikely to be detected as non-normal. The Shapiro-Wilk test is the most frequently used test for normality in common statistical packages, such as BMDP, Stata, SPSS, and SAS. It can be performed when the sample size is $n \leq 2000$. The lower limit depends on the specific software; although critical values for the Shapiro-Wilk test have been derived from $n \geq 4$, some software programs require a sample size of at least 7 to perform the test. Nonetheless, as mentioned, the test is unlikely to detect non-normal distributions with small sample sizes, and unfortunately, in animal experiments the final sample size usually ranges from 5 to 12. Moreover, the number of final cells can be very large. As an example of how to detect violations of the normality assumption using the Shapiro-Wilk test, let us consider an experiment performed to assess the effect of different treatments on oedema induced by carrageen in rats. The treatments consisted of 3 homeopathic remedies, each administered at 2 different dosages, and 1 negative and 1 positive control treatment, for a total of 8 treatments. The rats were housed 8 per cage for 7 days from arrival to testing. The 8 rats in each cage were randomly assigned to the 8 different treatments. At 1, 3, 5, 7, and 24 hours from carrageen subplantar injection, paw volume was measured using a plethysmometer and compared to the paw volume at 0 hours (baseline). All rats in the same cage remained in the experimental room when their cagemates were tested. The experiment was replicated 3 times, using 48 rats (6 cages with 8 rats each) for each replication. The number of final cells was 120 [3 (replications) x 8 (cages) x 5 (repeated measures) = 120], with each final cell consisting of 6 rats.

To detect violations of the normality assumption, a normality test can be performed on each of the data subgroups, which, for the above experiment, entails performing 120 tests (one test for each subgroup, each of which consists of 6 units). In Table 1, the significance levels of the 120 Shapiro-Wilk normality tests are presented. Below are described three different empirical methods that can be used to summarise these significance levels.

1) As shown in Table 1, significant results were obtained for only 6 samples (p value ≤ 0.05) (significance suggesting that the samples were probably non-normal). These results were obtained on days 2 and 3 only and for 5 of the 8 different treatments. However, 2 of the 6 significant results were obtained for one treatment (treatment G), for which there was a significant non-normal distribution on day 2 at 24 hours and on day 3 at 5 hours. Four of the 8 p values that were nearly significant ($0.05 < p \leq 0.10$) were also obtained for treatment G, although the remaining 9 p values for this treatment were all far from significant (range: 0.4369 to 0.9975). It thus seems that the effect of treatment G probably did not modify the distribution of data, which, in the other treatment groups, in nearly all cases did not deviate from normality.

2) In statistical inference, the p value represents the probability of a result being equal to or more extreme than the one obtained in the sample, under the null hypothesis H_0: in other words, the p value represents the probability of making an error (Type I) in considering H_0 to be false when it is actually true. This implies that, in repeated testing, H_0 being true, approximately 5% of the tests will be expected to provide a p value of ≤ 0.05 merely by chance. This is exactly what happened in the example provided. As seen in Table 2, which shows the frequency distribution of the p values presented in Table 1, the observed cumulative relative frequency of the p values in the consecutive classes is very close to, if not identical, the expected one. In particular, the relative frequency of p values ≤ 0.05 is exactly 5%, suggesting that chance played a role in determining the observed significances.

3) The above-described empirical methods do not solve the problem of insufficient power due to the small sample size in each final cell, which could explain why

Table 1. - Significance levels of the Shapiro-Wilk normality test. Significant values ($p \leq 0.05$) are indicated in bold; nearly significant values ($0.05 < p \leq 0.10$) are indicated in bold and italics

Subgroup	p values		
	Day 1	Day 2	Day 3
A_t01-t00	0.5498	***0.0992***	0.4818
A_t03-t00	0.1694	0.9892	0.5166
A_t05-t00	0.1848	0.4589	0.7861
A_t07-t00	0.6367	0.9669	0.8672
A_t24-t00	0.4289	0.2718	0.2282
B_t01-t00	0.4299	0.1382	0.6328
B_t03-t00	0.7622	0.5971	0.9222
B_t05-t00	0.6193	0.4170	0.3298
B_t07-t00	0.8094	0.3936	0.3002
B_t24-t00	0.2423	0.8423	0.3932
C_t01-t00	0.8875	0.5608	0.6771
C_t03-t00	0.6728	0.3869	0.8114
C_t05-t00	0.6644	0.3457	0.7514
C_t07-t00	0.9774	0.6782	0.8865
C_t24-t00	0.4571	0.4084	0.7801
D_t01-t00	0.3714	0.5940	0.9073
D_t03-t00	0.4757	0.2862	0.6421
D_t05-t00	0.8250	**0.0151**	0.9558
D_t07-t00	0.6040	0.6314	0.7548
D_t24-t00	0.7497	0.5056	0.1646
E_t01-t00	0.7400	0.8901	0.2227
E_t03-t00	***0.0976***	0.8547	**0.0299**
E_t05-t00	0.9843	0.3214	***0.0614***
E_t07-t00	0.9691	0.7586	0.9465
E_t24-t00	0.4832	0.4447	0.3209
F_t01-t00	0.4451	0.7715	**0.0235**
F_t03-t00	0.1402	0.3783	0.1896
F_t05-t00	0.4138	0.5910	***0.0689***
F_t07-t00	0.2744	0.3833	0.2821
F_t24-t00	0.3725	0.4278	0.9423
G_t01-t00	0.4369	***0.0780***	0.6240
G_t03-t00	0.7849	0.4582	***0.0577***
G_t05-t00	0.9975	0.5526	**0.0498**
G_t07-t00	0.5948	***0.0805***	***0.0664***
G_t24-t00	0.8405	**0.0187**	0.5858
H_t01-t00	0.9194	0.5804	0.9195
H_t03-t00	0.9906	0.2504	0.7188
H_t05-t00	0.6783	0.1219	0.4589
H_t07-t00	0.2486	0.1942	0.1890
H_t24-t00	0.9563	0.4297	**0.0283**

most of the 120 p values were not significant and, consequently, why non-significance was found for the overall sample. To overcome this problem, the sample size should be increased by collapsing the original data in the subgroups and performing the normality test on the overall group of units. If all of the ANOVA assumptions are respected, then the data in each subgroup will be distributed normally, with the variance across subgroups being equal. However, if the factor or factors being studied are effective, then the subgroups will have different means. For these reasons, when data are collapsed without performing a previous transformation, the resulting overall distribution may be non-normal, even if all subgroup distributions are normal. In fact, a mixture of normal distributions with different means but equal variance is, in most cases, substantially skewed, with the extent of skewness depending on the size of the differences among the means.

To overcome this obstacle, the data in each subgroup can be centred on the subgroup mean. The residuals in each subgroup would then have a mean equal to 0 and unaltered variance and distribution. Therefore, the data can then be collapsed, which would produce an overall group of units with a mean of 0 and a normal distribution, if the original subgroup distributions are all normal.

In the above example, the original data in the 120 subgroups were centred on the subgroup means and then collapsed. The Shapiro-Wilk normality test performed on the overall group of 720 units produced a non-significant result ($W = 0.9840$, $p = 0.2550$), confirming the substantial normality of the data, as already revealed by the normality tests performed on the 120 subgroups (Tables 1 and 2).

To determine the power of the above-mentioned procedure, simulation studies will need to be conducted. In fact, no information is available on the capability of the Shapiro-Wilk test, applied on collapsed data, to detect non-trivial violations of the normality assumption (which could affect the results of ANOVA).

Finally, even if all subgroup distributions are normal but variances are not homogeneous across subgroups, collapsing the residuals will result in a heavy-tailed distribution, that is, a distribution with greater than normal kurtosis. It is thus necessary to determine whether or not the homoscedasticity assumption has been respected, as described below.

Homoscedasticity

"Homoscedasticity" refers to the homogeneity of variance among the independent groups being compared. In CRD and mixed-model designs, homoscedasticity must be evaluated by comparing variances among the groups based on between-subject

Table 2. - Frequency distribution of the p values of the 120 Shapiro-Wilk normality tests

Interval	p values		
	Absolute frequency	Cumulative absolute frequency	Cumulative relative frequency
$0.00 \le p \le 0.05$	6	6	0.0500
$0.05 < p \le 0.10$	8	14	0.1167
$0.10 < p \le 0.20$	9	23	0.1917
$0.20 < p \le 0.30$	9	32	0.2667
$0.30 < p \le 0.40$	12	44	0.3667
$0.40 < p \le 0.50$	17	61	0.5083
$0.50 < p \le 0.60$	11	72	0.6000
$0.60 < p \le 0.70$	12	84	0.7000
$0.70 < p \le 0.80$	11	95	0.7917
$0.80 < p \le 0.90$	10	105	0.8750
$0.90 < p \le 1.00$	15	120	1.0000

factor levels. In mixed-model designs, the data in each block (corresponding to within-subject factor levels) or in each unit (corresponding to repeated measures) should first be averaged, so as to reduce the mixed-model design to a CRD. Homoscedasticity can then be verified using boxplots and significance tests.

As mentioned, boxplots summarise the distribution of a set of data. If boxplots corresponding to the independent subgroups are graphed side-by-side, the differences in their dimensions or in the length of their whiskers and the presence of outliers in some boxplots but not others can reveal heteroscedasticity.

Significance tests provide a significance level that, if low (typically less than 0.05), indicates a departure from homoscedasticity. The most frequently used tests for assessing homoscedasticity are Bartlett's test and Levene's test, which are also commonly provided in statistical packages. Bartlett's test is more sensitive than ANOVA itself to violations of the normality assumption and thus should not be used to verify the homoscedasticity assumption for ANOVA, unless normality has been demonstrated, in which case Bartlett's test would be more powerful than other tests. Levene's test consists of performing ANOVA (according to the same model used for analysing the original data) on the absolute deviations of each original data from its cell mean (or median, in later proposed versions of the test). It is much more robust than Bartlett's test yet should be applied with caution when the sample sizes are small.

In the example, the data were blocked by day and cage to take into account the possible effects of the day of testing (replication) and the housing cage. Each final cell (derived from combinations of replication, cage, and treatment) consisted of only one animal; thus, in the full model analysis, it was not possible to evaluate homoscedasticity. However, when disregarding day and cage, the final cells based on treatment levels consisted of 18 animals each. The Levene's test, performed to compare variances among the 8 treatment groups, both on the mean of the repeated measures and on each repeated measure, was never significant, confirming that there was substantial homoscedasticity among the groups (Table 3).

Sphericity

When using the parametric ANOVA in RBD or mixed-model designs, the F statistic which tests the significance of the main effect of the within-subject factor (or repeated-measure factor) and its interactions with other factors in the model, no longer follows the theoretical F distribution but is instead positively distorted. This means that the results of ANOVA may be significant more frequently than they should be (increase in the probability of a Type I error). Such a distortion does not affect the ANOVA results if the variance-covariance matrix of the data follows the sphericity pattern, which indicates that the sphericity assumption is respected. The sphericity assumption corresponds to the homoscedasticity assumption for CRD. Specifically, when calculating the differences between paired observations in the RBD, the sphericity assumption states that the variances of these differences (i.e., the σ_d^2 in a paired t-test) are equal across all groups in the sampled population.

A simpler yet stricter condition is that referred to as "compound symmetry", which is met when, in the variance-covariance matrix of the sampled population, both all of the variances and all of the covariances are equal, although the covariances are not necessarily equal to the variances. This means that the data collected on statistical units under different conditions (corresponding to different levels of the within-subject factor or to different repeated measures) must be equally related to each other with the same correlation coefficient ρ. Compound symmetry is sufficient, yet not necessary, for ensuring the validity of the F ratio under the general null hypothesis of no treatment effect. In other words, if compound symmetry is satisfied, then sphericity is also satisfied, but if compound symmetry is not satisfied, then sphericity must still be evaluated.

Although the Mauchly's test, which is a chi-square test, can be used to assess the sphericity assumption, it results in too many Type II errors when sample sizes are small (i.e., it does not detect sphericity violations) and too many Type I errors when sample sizes are large (i.e., producing significant results even for small and unimportant violations). Other approaches that can be used to address violations of the sphericity assumption are: a correction in standard ANOVA, and the multivariate analysis of variance (MANOVA).

Table 3. - Standard deviations for the 8 treatment groups and Levene's test results (n = 18 in each treatment group)

Variate	Standard deviation of treatment group								Results of Levene's test	
	A	B	C	D	E	F	G	H	F (7,136)	p value
Mean*(*)*	7.203	10.359	9.684	11.642	15.955	12.974	15.147	13.643	1.42	0.2026
t01-t00	13.516	11.198	11.403	12.285	17.271	15.403	16.313	16.723	0.65	0.7119
t03-t00	14.770	11.220	11.399	18.563	19.218	18.098	22.560	16.421	1.23	0.2909
t05-t00	10.260	19.590	16.732	18.230	19.491	16.180	20.536	18.348	1.18	0.3167
t07-t00	10.429	14.933	13.469	17.164	20.719	20.111	17.972	15.839	1.42	0.2012
t24-t00	7.461	10.205	10.496	10.331	15.217	10.733	12.935	12.911	1.34	0.2345

(*) The mean represents the average of differences between paw volume at time 0 (baseline) and at 1, 3, 5, 7, and 24 hours, calculated for each rat.

Correction in standard ANOVA

As mentioned, a violation of the sphericity assumption produces a positive distortion of the F test, so that the significance level of the F statistic is higher than it should be. To reduce the significance level, it would be necessary to reduce the F statistic. However, the significance level can be more simply reduced by reducing the associated degrees of freedom, with the extent of the reduction depending on the extent of the sphericity violation.

The most frequently used corrections for the degrees of freedom are those developed by Greenhouse and Geisser and by Huynh and Feldt. In both cases, a coefficient denoted by "ε" must be calculated. The upper bound for ε is 1, in case of perfect sphericity, whereas the lower bound is 1/(k-1), where k is the number of levels of the within-subject factor (or of the repeated measures): the worse the sphericity violation, the smaller the value of ε.

The Greenhouse-Geisser correction is more conservative than the Huynh-Feldt correction. The choice of the most appropriate correction depends on the relative importance of Type I and Type II errors in the experiment (in terms of cost, effects, and consequences). If Type I error is more important, then the Greenhouse-Geisser correction is preferable, whereas the Huynh-Feldt correction should be used if Type II error is more important.

A correction in standard ANOVA seems to work well for modest violations of the assumption, or when the sample size is small.

Multivariate analysis of variance (MANOVA)

MANOVA is a parametric method developed to test the difference among groups with respect to more than one outcome variable. The different observations made on each statistical unit in RBD can be treated as different outcome variables and analysed by MANOVA. In general, MANOVA is less powerful than repeated-measure ANOVA, even when applying corrections for the degrees of freedom, and thus should be avoided. MANOVA should instead be used when marked violations of the sphericity assumption can be hypothesised ($\varepsilon < 0.70$) and the sample size is sufficiently large ($n > 10 + k$).

In summary, to determine whether the sphericity assumption is violated and which method must be applied (correction for degrees of freedom or MANOVA), the significance levels obtained using standard ANOVA, corrected ANOVA, and MANOVA can be compared. If these levels are fairly similar, then any of the methods can be used, whereas if the differences are large, then the corrected ANOVA or MANOVA must be used, depending on the extent of violation, the sample size, and the relative importance of Type I and Type II errors.

Conclusions

Parametric ANOVA provides accurate results if the assumptions on which the analysis is based (normality, homoscedasticity, sphericity) are respected, whereas violations in these assumptions can be damaging, with the degree of damage depending on the type and extent of the violation and the sample size. Thus the assumptions must be evaluated before drawing any conclusion based on the results of ANOVA. Descriptive statistics, diagnostic plots, or significance tests can be used, depending on the available software, the complexity of the experimental design, and the sample size.

Normality must be evaluated in each final cell of the experimental design. The significance levels of normality tests (most commonly the Shapiro-Wilk test) can be listed in increasing order and their cumulative frequency can be determined. Quantiles can then be

compared with the expected quantiles under the null hypothesis of the normality of the distribution in the sampled population. Moreover, when the sample size in the final cells is small, the information provided by the separate tests should be summarised. To this end, the data in each final cell can be centred on the cell's mean and the normality test can be performed on the overall group of centred data.

Homoscedasticity must be evaluated by comparing variances in the groups based on between-subject factor levels. In the case of mixed-model designs, data on each block (corresponding to within-subject factor levels) or on each unit (corresponding to repeated measures) must first be averaged, so as to reduce the mixed-model design to a CRD. The most frequently used significance test is the Levene's test, which is available in most statistical packages.

Finally, sphericity must be evaluated when using the parametric ANOVA in RBD or mixed-model designs. To do so, it is suggested that standard ANOVA, the Greenhouse-Geisser or Huynh-Feldt ANOVA correction, or MANOVA be used. If the significance levels are fairly similar, any of these methods can be used, whereas if there is a large discrepancy, the corrected ANOVA or MANOVA must be used, the choice depending on the extent of violation, on the sample size, and on the relative importance of Type I and Type II errors. The Greenhouse-Geisser and Huynh-Feldt corrections are usually simpler to use and easier to interpret than MANOVA, and they are available in most statistical packages.

In conclusion, if one or more assumptions are violated, then parametric ANOVA performed on raw data is not the best method of statistical analysis, and other approaches must be followed. Specifically, the data can be transformed so as to meet the assumptions of the parametric tests, or statistical methods that do not rely on such stringent assumptions, such as non-parametric tests, can be adopted. Although beyond the scope of this paper, indications regarding the most important non-parametric tests in experimental designs can be found in the textbooks listed as recommended reading below.

Acknowledgements

The study was supported by the Italian Ministry of Health, Project no. 01/C "Biostatistical and ethological approaches for the promotion of welfare of laboratory animals and of the quality of experimental data", to Maria Puopolo.

Submitted on invitation.
Accepted on 3 March 2004.

RECOMMENDED READING

Experimental designs and data analysis

Chiarotti F, Sparks T. Problems of data analysis in behavioural ecotoxicology. In: Dell'Omo G (Ed.). *Behavioural ecotoxicology*. Chichester: John Wiley & Sons; 2002. p. 273-89.

Cochran WG, Cox GM. *Experimental designs*. 2. ed. USA: John Wiley & Sons; 1992.

Edwards AL. *Experimental design in psychological research*. 5. ed. New York: Harper & Row; 1985.

Winer BJ. *Statistical principles in experimental design*. 2. ed. Tokyo: McGraw-Hill; 1971.

Non-parametric statistical methods

Brunner E, Domhof S, Langer F. *Nonparametric analysis of longitudinal data in factorial experiments*. USA: John Wiley & Sons; 2002.

Marascuilo LA, McSweeney M. *Nonparametric and distribution-free methods for the social sciences*. Monterey CA: Brooks/Cole; 1977.

Software manuals

Dixon WJ. *BMDP Statistical software manual. Release 7*. Berkeley CA: University of California Press; 1992.

SAS Institute. *SAS Manual*. Cary NC: SAS Institute; 1997.

StataCorp. *Stata statistical software. Release 7.0*. College Station TX: Stata Corporation; 2001.

SPSS Inc. *SPSS Base 10.0 for Windows User's Guide*. Chicago IL: SPSS Inc.; 1999.

Documents available from Internet

Baguley T. *An introduction to sphericity*. Available from: http://www-staff.lboro.ac.uk/~hutsb/Spheric.htm. Last visited 26/5/2003.

NIST/SEMATECH. *e-Handbook of statistical methods*. Available from: http://www.itl.nist.gov/div898/handbook/. Last visited: 26/5/2003.

Northwestern University. *PROPHET StatGuide*. Available from: http://www.basic.nwu.edu/statguidefiles/list.html; last visited 26/5/2003.

Sytsma S. *The basics of experimental design [A quick and non-technical guide]*. Available from: http://www.sytsma.com/phad530/expdesig.html. Last visited 26/5/2003.