The following material deals briefly with extensions of and alternatives to ANOVA-model CI procedures. It is a condensed version of what would have been an additional chapter in *ANOVA via CIs* had space been available. Reference is sometimes made to models and equations in *ANOVA via CIs*, using the numbering system used in the book.

Simplifying analyses of factorial data

Saturated ANOVA models can often lead to complex analyses of data from factorial experiments, primarily because of the number and nature of the interaction contrasts that must be included in an exhaustive account of variation between cell means. For example, a saturated two-factor model of data from a 3×4 design accounts for the 11 degrees of freedom between 12 cells in terms of five linearly independent main effect parameters and six additional linearly independent interaction parameters. A contrasts analysis based solely on main effect parameters would obviously be simpler, in scale and in ease of interpretation, than an analysis including at least six interaction contrasts.

If you decide (perhaps on the basis of a CI on f_{AB}) that interaction parameters can safely be ignored, then you can base the interpretation of the data on main effect contrasts. If you define these contrasts on the main effect parameters of the *saturated* two-factor model, then the analysis becomes a simplified version of the ANOVA-model analysis discussed in Chapter 5 (of *ANOVA via CIs*). Alternatively, you could choose to define contrasts on the main effect parameters of the *unsaturated* main effects model (4.2). The main effects model is not compatible with the cell means model, and for this reason *PSY* cannot construct CIs on contrasts defined on the parameters of the main effects model.

Unsaturated models In order to avoid confusion, we redefine the two-factor main effects model as

$$Y_{ijk} = \mu^* + \alpha^*_j + \beta^*_k + \varepsilon^*_{ijk} \qquad (1)$$

where $\sum_j \alpha^*_j = \sum_k \beta^*_k = 0$.

In general, none of the terms on the right hand side of (1) is identical to the corresponding term of the right hand side of the saturated model

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \varepsilon_{ijk}$$

unless the design is balanced, in which case $\alpha^*_j = \alpha_j$ and $\beta^*_k = \beta_k$. In the case of an unbalanced design, a contrast like $\alpha^*_1 - \alpha^*_2$ is usually not identical to the 'same' contrast $\alpha_1 - \alpha_2$ defined in the context of a different model. Thus the choice of model is not simply a choice between two different ways of producing interval estimates of the values of the same set of main effect contrasts (as it is with balanced designs).

The main effects model is not the only unsaturated model that can provide a basis for a contrasts analysis of data from a two-factor design with factors *A* and *B*. Two other possible models are

$$Y_{ijk} = \mu^{**} + \alpha_j^{**} + \varepsilon_{ijk}^{**} \quad (2)$$

and $$Y_{ijk} = \mu^{***} + \beta_k^{**} + \varepsilon_{ijk}^{***} . \quad (3)$$

In general, α_j^{**} is not equal to α_j^{*} (or to α_j) and β_k^{**} is not equal to β_k^{*} (or to β_k). Thus the definition of a main effect parameter and of contrasts on that parameter depend on what other parameters are included in the model.

The parameters of unsaturated models can be expressed as linear combinations of cell means, but the results are often surprising because the weights attached to the cell means in these definitions are influenced by the pattern of sample sizes (n_{jk}). Whether it is appropriate to allow differences in cell sample sizes to influence the definition of model parameters depends on what is responsible for those differences. If the pattern of differences in cell frequencies is essentially uninformative, as is the case if it is due to data missing at random, then the saturated model, which assigns the same weight to each mean, should be preferred to an unsaturated model. In general, saturated or cell means models should be adopted for the analysis of data from randomized experiments, even if there is good reason to ignore some or all of the interactions in the contrasts analysis.

The pattern of differences in cell sizes can sometimes be genuinely informative in *observationa*l studies without random assignment to cells. Suppose, for example, that the participants in a 2 × 3 between-subjects observational study are randomly sampled from a population of interest to the researcher, and that the two factors are the categorical individual-difference variables Anxiety (a_1: high anxiety, a_2: low anxiety) and Depression (b_1: low depression, b_2: medium depression, b_3: high depression). In this case the pattern of inequality in sample sizes is a reflection of the association between these variables in the population. This is not the place to consider the issues involved in choosing a model for an analysis of data from this kind of study. Various approaches have been recommended, some of which use a hierarchical approach (as in hierarchical multiple regression analysis), with a different model for each type of parameter. The issues involved are usually discussed in the context of a general treatment of multiple regression analysis or general linear model analysis, such as that provided by Cohen, Cohen, West and Aiken (2003).

Fitting an unsaturated model *SPSS MANOVA* produces an unsaturated model analysis if the effects in that model are specified on a *design* line in the syntax file. Given a 2 × 3 between-subjects design with factors *A* and *B*, the following syntax fits the main effects model and produces 95% Scheffé SCIs on contrasts on the parameters of that model.

```
manova y by A(1 2) B(1 3)
  /contrast(A)=special(1  1
                       1 -1)
  /contrast(B)=special(1  1  1
                      -1  0  1
                      -1  2 -1)
  /cinterval joint(.95) univariate(scheffe)
  /print param(estim)
  /design A B.
```

The CIs in this analysis refer to main effect contrasts defined on parameters of the main effects model (1).

The following syntax constructs a CI on the *A* main effect contrast defined by model (2).

```
manova y by A(1 2)
  /contrast(A)=special(1  1
                       1 -1)
  /cinterval joint(.95) univariate
  /print param(estim).
```

Finally, to construct Scheffé CIs on the *B* main effect contrasts defined by model (3), use the syntax:

```
manova y by B(1 3)
  /contrast(B)=special(1  1  1
                      -1  0  1
                      -1  2 -1)
  /cinterval joint(.95) univariate(scheffe)
  /print param(estim).
```

Note that a *design* line is required in the first case in order to define a two-factor model without interaction parameters.

Analysis of covariance (ANCOVA)

The most simple version of the analysis of covariance model may be written as

$$Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \varepsilon_{ij} \qquad (4)$$

where X is a covariate, a variable on which subjects are measured prior to the administration of treatments in the experiment,
the parameter β is a regression coefficient
and the remaining terms on the right hand side of (4) are analogous to the corresponding terms in the single-factor ANOVA model (3.2a).

The error variance σ_ε^2 does not include the variance in Y that is predictable from X via the regression term βX_{ij}. Therefore the error variance defined by the ANCOVA model is smaller than the error variance defined by the ANOVA model if the covariate X is correlated with the dependent variable Y. As a result, an ANCOVA can often provide more precise estimates of parameters of interest (such as the values of contrasts on effect parameters) than an ANOVA.

The model can be extended to include more than one type of effect parameter and more than one covariate. Covariates should not be included, however, simply because they are available. Each covariate accounts for one of the degrees of freedom that would otherwise make a contribution to the estimate of the error variance, so a poorly chosen set of covariates can lead to a reduction in precision of estimation.

Inferences from ANCOVA depend on stronger assumptions (concerning the nature of

the regression of *Y* on *X*) than inferences from ANOVA.

ANCOVA is often used to analyse data from quasi-experiments where assignment to treatments is not random (Cook and Campbell, 1979), and pre-existing differences between groups can be mistaken for treatment effects in an ANOVA. In this context, covariates are included in order to reduce bias (by controlling for pre-existing differences associated with the covariates) rather than to increase precision. This is a problematic application of ANCOVA (Reichardt, 1979), but there is no doubt that ANCOVA is usually superior to ANOVA in this context.

Carrying out an ANCOVA It is not possible to carry out an ANCOVA with *PSY*, because *PSY* is based on a means model, not the general linear model. In general, any GLM program that can provide an appropriate ANOVA-model (or MANOVA-model) analysis can also make provision for the inclusion of one or more covariates in the analysis. *SPSS GLM* and *SPSS MANOVA* can include covariates in a planned contrasts analysis with *t*-based CIs. If you want to include covariates in a post hoc analysis including Scheffé, T^2 or GCR CIs, you can do so with *SPSS MANOVA*. To include a covariate in an *SPSS MANOVA* analysis, simply add

```
with X
```

to the *manova* line of the syntax that would otherwise be used for an ANOVA analysis. For example, the syntax for an ANCOVA based on data in an *SPSS* file with variables *group*, *X* and *Y* would begin with the line

```
manova Y by group with X
```

Multiple covariates can be included. The selection of covariates should be made independently of the results of the analysis, because post hoc selection of covariates can introduce bias into the resulting ANCOVA.

Multivariate analysis of variance (MANOVA)

A multivariate experiment has more than one dependent variable. The most obvious approach to the analysis of data from an experiment with *p* dependent variables ($Y_1, Y_2, \cdots, Y_p$) is to carry out *p* univariate analyses, one for each dependent variable. The FWER from such an analysis can be controlled with Bonferroni-adjusted critical constants. If the analysis is carried out with a FWER per dependent variable of α/p, then the FWER for inferences on all contrasts in the analysis cannot exceed α. All of the analyses discussed in *ANOVA via CIs* can be modified in this way. Consider, for example, a single-factor experiment with $J = 4$ groups, $n = 20$ subjects per group, and $p = 3$ dependent variables. If the experimenter wishes to control the experimentwise error rate for post hoc contrasts on all three dependent variables, then the appropriate Bonferroni-adjusted Scheffé critical constant for CIs is

$$\sqrt{\nu_1 F_{\alpha/p;\,\nu_1,\nu_2}} = \sqrt{3F_{.05/3;3,76}} = 3.299.$$

Analyses of this kind can be implemented with *PSY* or *SPSS* by changing the value of α [or $100(1 - \alpha)$] from the default value to α/p [or $100(1 - \alpha/p)$]. To run the Bonferroni-adjusted Scheffé analysis in *PSY*, construct a separate input file for each dependent variable. For each of the p sub-analysis (three in this case), select *post hoc* on the Analysis Options screen and change the *Confidence level %* value from 95 to $100[1 - .05/p]$ (98.333 in this case).

An alternative solution to the multiplicity problem is to adopt a MANOVA model for the analysis of the data. In Chapter 7 we saw how the MANOVA-based *GCR* procedure can be used to construct SCIs on interaction contrasts from a mixed (between × within) two-factor design. A similar procedure can be used to construct SCIs on contrasts from a single-factor between-subjects multivariate experiment. The *GCR* CC is

$$\sqrt{\frac{\nu_E \, \theta_{\alpha;\, s,m,n}}{1-\theta_{\alpha;\, s,m,n}}}$$

where the *GCR* (θ) parameters are

$$s = \min(\nu_1,\ p)$$
$$m = \left(|\nu_1 - p| - 1\right)/2$$
and $$n = \left(\nu_E - p - 1\right)/2 .$$

The *GCR* CC for the current example is

$$\sqrt{\frac{76\,\theta_{.05;\, 3,-0.5,36}}{1-\theta_{.05;\, 3,-0.5,36}}} = 3.798 ,$$

which is 15.1% larger than the Bonferroni-adjusted Scheffé CC. This example illustrates an important general point: MANOVA-based SCIs are always wider than Bonferroni-adjusted Scheffé intervals in analyses where the only contrasts of interest are contrasts on individual dependent variables. MANOVA is not appropriate for analyses of this kind.

When is MANOVA appropriate? Any contrast in a multivariate experiment can be expressed as a contrast on a linear combination of the dependent variables. The MANOVA-based *GCR* procedure controls the FWER (the EWER in the case of single factor designs) for all contrasts on all linear combinations of dependent variables, including the maximal contrast on the *first discriminant function*, the (necessarily post hoc) linear combination of dependent variables with the largest ANOVA F statistic. A MANOVA is appropriate when the experimenter wishes to allow for the possibility that the linear combinations (such as discriminant functions) that emerge from the analysis might provide a more informative account of differences between groups than any of the individual dependent variables. Experimenters who have no interest in emergent linear combinations should not consider a MANOVA-model analysis.

It is sometimes suggested that a statistically significant overall MANOVA test is required to justify a set of 'follow-up' univariate ANOVAs on individual dependent variables. The logic underlying this suggestion is essentially the same as the (discredited) logic underlying the use of 'protected' *t* tests on all comparisons following a significant *F* test in a univariate analysis. As Gleitzman (1996) has demonstrated, neither procedure controls the familywise Type I error rate. It should also be noted that this type of sequential analysis leads only to (invalid) follow-up tests. It does not allow for the construction of CIs.

Assumptions

The inferential procedures discussed in *ANOVA via CIs* are based on models including random error components (such as $\varepsilon_{ijk} = Y_{ij} - \mu_j$ in the case of the means model) that allow for discrepancies between expected values of the dependent variable (given the parameters of the model) and the values actually observed. In the case of between-subjects designs, the justification for claims about the error rates and confidence levels associated with these procedures depend on the following assumptions about error:

- the error components associated with each pair of *Y* values are statistically independent.
- the variances of the *j* error distributions ($\sigma^2_{\varepsilon_j}$) are homogeneous
- within each of the *j* populations the error components are normally distributed.

The first of these assumptions is often violated in practice, and it is reasonable to suppose that the second and third assumptions are almost always false. How well, then, do ANOVA-model analyses perform when assumptions about error distributions are violated to the extent likely to be encountered in practice? If ANOVA-model analyses are suspect in the presence of a given degree of variance heterogeneity or non-normality, are alternative procedures available that are likely to perform better?

The assumption of independence

Consider an experiment with three treatments for depression, where each treatment is administered by five experienced therapists, but each therapist administers only one treatment. Suppose that 150 subjects are randomly assigned to the therapists, so that each therapist treats 10 subjects.

A single-factor fixed-effects ANOVA model for the data from this experiment would ignore therapists (just as was done for the Depression study discussed in Chapters 2 and 3 of *ANOVA via CIs*). In effect, an analysis based on this model would assume that differences in treatment outcomes might be influenced by differences between treatments, but not by differences between therapists. If this assumption is false, because some therapists administer treatments more effectively than others, then *Y* values (and

associated error components) obtained by different subjects assigned to the same therapist would not be statistically independent, and this failure of the independence assumption would bias the estimates of effect parameters in the single-factor ANOVA model.

We can obtain some idea of the consequences of this threat to the validity of inferences about treatment effects by considering an alternative model of the data that takes into account the role of therapists in the design. Although the alternative model has a Therapist factor as well as a Treatment factor, it is not a factorial model because the design does not produce data for all possible combinations of treatments and therapists. Rather, the Therapist factor is *nested within treatments*, meaning that each therapist is observed at only one level of the Treatment factor. Furthermore, we will suppose that Therapist factor is to be treated as *random* (as distinct from *fixed*), meaning that the therapists administering a particular treatment are regarded as a random sample from a population of qualified therapists, and that conclusions from the analysis are to refer to the population from which the therapists are sampled. The resulting model can be written as:

$$Y_{ijk} = \mu + \alpha_j + b_{k(j)} + \varepsilon_{i(jk)} \ , \qquad (5)$$

where Y_{ijk} is the dependent variable score obtained by subject *i* given treatment *j* by therapist *k* ($i = 1, \cdots, 10$; $j = 1, \cdots, 3$; $k = 1, \cdots, 5$);

$b_{k(j)}$ is the effect of therapist *k* within treatment *j*;

$\varepsilon_{i(jk)}$ is the error due to variation between subjects within treatment-therapist combinations.

The therapist random effect parameter $b_{k(j)}$ is a value of a random variable with variance $\sigma^2_{b_{k(j)}}$. The model includes two random variables, the second being $\varepsilon_{i(jk)}$ with variance $\sigma^2_{\varepsilon_{i(jk)}}$.

The single-factor ANOVA model does not distinguish between therapist variance and error variance. The proportion of the error variance defined by the single-factor model that is attributable to differences between therapists is

$$\rho = \frac{\sigma^2_{b_{k(j)}}}{\sigma^2_{b_{k(j)}} + \sigma^2_{\varepsilon_{i(jk)}}} \ , \qquad (6)$$

an *intraclass correlation coefficient* (ICC). An analysis that ignores therapists (that is, an analysis based on the single-factor ANOVA model) depends on the assumption that $\rho = 0$. If $\rho > 0$, the standard errors of treatment contrast values are systematically underestimated, with the result that noncoverage error rates for raw CIs on treatment contrasts are inflated. In addition, point estimates of standardized treatment effects are inflated. Given particularly unfavourable combinations of *J*, *K*, and *n*, the bias in the

estimation of standardized treatment effects resulting from even modest values of ρ can be substantial. Wampold and Serlin (2000) have shown by Monte Carlo methods that if $J = K = 2$, $n = 10$ and $\mu_1 - \mu_2 = 0$ (two treatments, two therapists per treatment, 10 subjects per therapist, no treatment effect), then the expected value of $\hat{\omega}^2$ (an estimate of effect size) is about .067 when $\rho = 3$. When $J = 2$, an ω^2 value of .067 is equivalent to a standardized mean difference of $|\mu_1 - \mu_2|/\sigma_\varepsilon = 0.536$, a medium effect according to Cohen's guidelines. That is, if there is no difference between the two treatment means and 30% of the 'error' variance defined by the single-factor ANOVA model is due to differences between therapists, then a point estimate of the standardized effect size can be expected to indicate a medium effect (in one direction or the other). As might be expected, the Type I error rate from a .05-level test of $H_0: \mu_1 - \mu_2 = 0$ (about .339 according to Wampold and Serlin's Monte Carlo data) is seriously inflated in this situation. The noncoverage error rate for raw CIs on $\mu_1 - \mu_2$ would be similarly inflated.

Note that the problems outlined in the previous paragraph arise because of a mismatch between the structure of the experimental design and the model chosen as the basis for the analysis. The error components defined by (5) can be statistically independent in the presence of Therapist effects, but the error components defined by the inappropriate single-factor ANOVA model cannot.

Dealing with statistically dependent observations.

In general, the best way of analysing data from a hierarchical design (where levels of one factor are nested within levels of another factor) is to base the analysis on a model [such as (5)] that reflects the hierarchical structure of the design. Analyses informed by hierarchical models are usually *multilevel* analyses, where error terms (for tests) and standard errors (for CIs) vary across levels. For example, an analysis of treatment effects informed by (5) would use an error term based on variation between therapists within treatments, whereas an analysis of therapist effects (which may or may not be of interest to the experimenter) would use an error term based on variation between subjects within therapists.

If you are thinking of adopting a model for an analysis that appears to ignore statistical dependencies that may exist in the data, then you should think about the possibility of adopting a different model that takes the dependencies into account. If the experiment is well designed, some form of multilevel analysis may be appropriate. For an introduction to multilevel models, see Cohen, Cohen, West and Aiken (2003), Kreft and DeLeeuw (1998), or Maxwell and Delaney (2004).

The assumption of variance homogeneity

If the variance of *Y* values in population *j* is $\sigma^2_{Y_j} = \sigma^2_{\varepsilon_j}$, the variance homogeneity assumption states that

$$\sigma^2_{\varepsilon_1} = \sigma^2_{\varepsilon_2} = \cdots = \sigma^2_{\varepsilon_J} = \sigma^2_{\varepsilon} \ .$$

If this assumption is false because error variances $\sigma^2_{\varepsilon_j}$ vary across populations, then the standard error of the estimated value of the contrast ψ_g is

$$\sigma_{\hat{\psi}_g} = \sqrt{\sum_j \frac{c_j^2 \sigma_j^2}{n_j}} \ . \tag{7}$$

Variance heterogeneity implies that, even if sample sizes are equal, standard errors can vary across different {*m*, *r*} contrasts of the same degree of complexity (and therefore the same value of $\sum c^2$). As a consequence, the procedure used in *ANOVA via CIs* to estimate standard errors of between- subjects contrasts (2.12) will systematically understimate the standard errors of some contrasts and overestimate the standard errors of others. As a consequence, individual CIs will be too wide on some contrasts, and too narrow on others.

This problem can be corrected by estimating contrast standard errors from

$$\hat{\sigma}_{\hat{\psi}_g} = \sqrt{\sum_j \frac{c_j^2 s_j^2}{n_j}} \ . \tag{8}$$

Given this estimated standard error, the CC required for the construction of an approximate 100(1 – α)% individual raw CI on ψ_g is $t_{\alpha/2;\nu'}$, where

$$\nu' = \frac{\hat{\sigma}^4_{\hat{\psi}_g}}{\sum_j \left[\frac{c_j^4 s_j^4}{n_j^2 (n_j - 1)} \right]} \ . \tag{9}$$

This CI procedure (developed by Brown and Forsythe, 1974) is not supported by *PSY*, *SPSS MANOVA*, or any of the other programs mentioned elsewhere in *ANOVA via CIs*. *SPSS ONEWAY*, however, can provide the required values of $\hat{\psi}_g$, $\hat{\sigma}_{\hat{\psi}_g}$ and ν' . Given these values, hand calculations of CI limits from

$$\psi_g \in \hat{\psi}_g \pm \text{CC} \times \hat{\sigma}_{\hat{\psi}_g} \ ,$$

are not onerous unless the number of contrasts is large.

To illustrate the Brown-Forsythe *t* (*BF-t*) procedure, we will reanalyse the Depression data using the same set of planned orthogonal contrasts that were used to illustrate individual CIs in Chapter 2. The *SPSS* syntax

```
oneway y by group
  /contrast = .333333 .333333 .333333 -1
  /contrast = .5 .5 -1 0
  /contrast = 1 -1 0 0.
```

produces output including the following:

Contrast Tests

	Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
Assume equal variances	1	8.717[a]	2.1627	4.031	76	.000
	2	2.800	2.2938	1.221	76	.226
	3	.200	2.6487	.076	76	.940
Does not assume equal variances	1	8.717[a]	2.0072	4.343	37.409	.000
	2	2.800	2.3214	1.206	40.249	.235
	3	.200	2.7962	.072	37.903	.943

a. The sum of the contrast coefficients is not zero.

The values in the first section of the output (appropriate when variance homogeneity is assumed) are compatible with the CI analysis discussed in Chapter 2, and repeated in Table 1(a). *BF-t* tests (which do not assume homogeneous variances) are reported in the second section of the *SPSS* output. Note that ν' values (*df* values in the output) are not integers. These values were rounded to the nearest integer to determine the CCs for the calculation of the CI limits shown in Table 1(b).

Table 1 *Critical constants and 95% individual confidence intervals from Y scores (a) assuming variance homogeneity and (b) not assuming variance homogeneity*

(a) Variance homogeneity assumed

Contrast	*CC*	*Raw CI*		*Standardized CI*	
		LL	*UL*	*LL*	*UL*
ψ_1	$t_{.025;76} = 1.992$	4.409	13.024	0.526	1.555
ψ_2	$t_{.025;76} = 1.992$	–1.769	7.369	–0.211	0.880
ψ_3	$t_{.025;76} = 1.992$	–5.075	5.475	–0.606	0.654

(b) Variance homogeneity not assumed

Contrast	*CC*	*Raw CI*		*Standardized CI*	
		LL	*UL*	*LL*	*UL*
ψ_1	$t_{.025;37} = 2.026$	4.710	12.724	0.562	1.519
ψ_2	$t_{.025;40} = 2.021$	–1.892	7.492	–0.226	0.894
ψ_3	$t_{.025;38} = 2.024$	–5.461	5.861	–0.652	0.700

The *BF* CIs in Table 1(b) are very similar to those constructed on the assumption of homogeneous population variances. This is not surprising, given the fact that $s_{\max} / s_{\min}$ is only 1.22. If sample variances differ substantially, however, then the *BF*-*t* procedure can produce CIs that look very different from standard *t*-based intervals.

Most of the methods recommended in *ANOVA via CIs* for the construction of CIs on between-subjects contrasts (and interaction contrasts with a between-subjects component) can be modified along the lines suggested by Brown and Forsythe (1974). For example, the Scheffé procedure can be modified by replacing the standard CC of $\sqrt{\nu_1 F_{\alpha;\nu_1,\nu_2}}$ with $\sqrt{\nu_1 F_{\alpha;\nu_1,\nu'}}$, and replacing the usual contrast standard error with (8). The modified Scheffé procedure, however, does not have the same relationship to a similarly modified ANOVA *F* test that the standard Scheffé procedure has to the standard ANOVA *F* test. A similar comment could be made about Brown-Forsythe modifications of any other procedure allowing for post hoc analysis. Harris (1994, p.159) recommends that when there is reason to be concerned about variance heterogeneity, a Bonferroni-adjusted version of the *BF*-*t* procedure (perhaps allowing for inferences on all {*m*, *r*} contrasts) should be used in preference to Brown and Forsythe's (1974) modification of the Scheffé procedure. None of the modified CI procedures produces exact CIs. It is not clear whether any of them produce inflated noncoverage error rates.

If variance homogeneity is not assumed, the usual definition of standardized CIs does not apply. Standardization can be based on some kind of average of the population standard deviations, on the square root of the average of the population variances (the standardization implicitly used in the example), or on variability in a particular experimental condition (such as a control condition).

The assumption of normal distributions

It is reasonable to suppose that the assumption of normality of error distributions (implying that *Y* scores are normally distributed within populations) is always false. Indeed, the distributions of dependent variables used in psychological research are often skewed, bimodal, multimodal, or lumpy, sometimes with a discrete mass of scores with

a value of zero (Micceri, 1989, Sawilowsky and Blair, 1992). Monte Carlo studies have shown that the two-group *t* test and the ANOVA *F* test generally perform well (in the sense of producing error rates close to nominal values) when *Y* values are sampled from populations with mildly or moderately skewed distributions (Glass, Peckham and Sanders, 1972) and from distributions likely to be encountered in practice (Sawilowsky and Blair, 1992). This conclusion applies to experiments where the population distributions for all treatments have the same shape. It probably does not apply to experiments where the distribution shape varies substantially across treatments, or where non-normality is combined with variance heterogeneity.

Substantial departures from normality can produce serious discrepancies between nominal and actual error rates. Consider the problem of constructing a 90% CI on the mean of a variable *Y* whose distribution is *lognormal* rather than normal. [The lognormal distribution of $Y = e^Z$ (where *Z* has a standard normal distribution) is heavily skewed, as shown in Figure 1.] If $n = 20$, the actual noncoverage error rate is about .161 when the nominal error rate is $\alpha = .10$ (Westfall and Young, 1993). More importantly, the distribution of errors is asymmetrical. The probability of an overestimation error, where all of the values in the interval are greater than the population mean, is much lower than the nominal value (about .008, rather than $\alpha/2 = .05$), whereas the probability of an underestimation error, where all of the values in the interval are smaller than the population mean, is unacceptably large (about .153 rather than .05). If the variable *Y* is a within-subjects contrast variable with a population value of zero, then the *t*-based CI (or test) procedure will produce far too many erroneous inferences implying that $\psi < 0$. If the population value of the contrast is positive, the procedure will produce too many directional inferences in the wrong direction (Type III errors) and too few directional inferences in the right direction (that is, poor power).

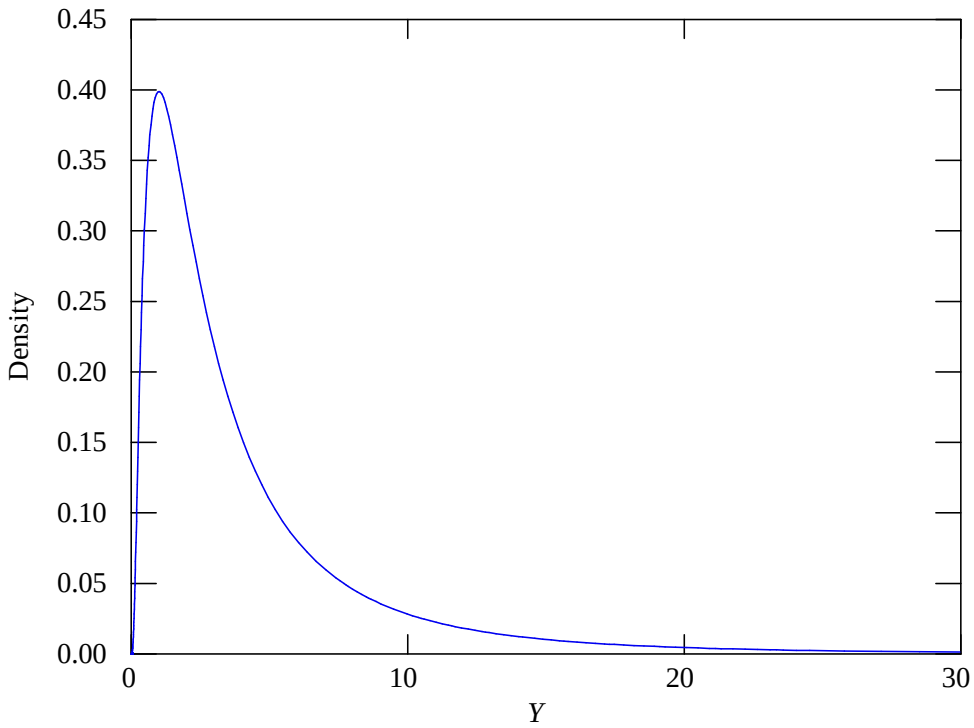

Figure 1 *A lognormal distribution*

These results seem to be at odds with the claim that ANOVA methods work reasonably well with non-normal distributions. It should be noted that the outcome can be quite different when two or more groups are involved in the analysis. In a between-groups experiment with approximately equal sample sizes where the dependent variable distribution is lognormal with the same variance in all populations, ANOVA procedures work reasonably well (Westfall and Young, 1993, p.60). The same would be true of within-subjects experiments with lognormal distributions of dependent variable scores (as distinct from contrast variable scores). Problems similar to those outlined above are likely to arise in between-subjects experiments, however, if the magnitude or direction of skew varies across populations. Problems associated with non-normality can be magnified in analyses including multiple inferences. Westfall and Young (1993, p.62) have shown that if $n = 20$, Sidak-adjusted simultaneous inferences on the means of ten independent lognormal distributions with a nominal FWER of .10 produce an actual FWER slightly greater than .50, and that almost all of the noncoverage errors are underestimation errors. These results cannot be generalized directly to multiple comparisons (as distinct from multiple inferences on individual means), but they do illustrate the fact that problems associated with a single inference can be magnified considerably in a multiple inference context.

Skew is not the only kind of departure from normality that has the potential to produce problems for ANOVA procedures. When these procedures are applied to symmetrical distributions with thick tails (usually called *heavy-tailed* distributions), noncoverage error rates can be expected to fall below the nominal value, thereby decreasing the power to detect nonzero contrast values (Wilcox, 1998).

Robust alternatives to ANOVA procedures

Least squares (LS) estimation procedures (used in ANOVA and other general linear model analyses) are known to be sensitive to *outliers* (extreme values) that occur relatively frequently in samples from various non-normal distributions, including heavy-tailed and skewed distributions. Outliers distort LS estimates of standard errors and CI CCs. *Robust* estimation procedures provide better estimates of standard errors or better CCs when ANOVA-model distributional assumptions fail, thereby providing a basis for CI procedures with better control over error rates There is a large, complex and rapidly changing literature on robust estimation procedures, and it is not possible to provide an adequate summary here. For our purposes it will be sufficient to provide a brief indication of some of the methods that have so far been developed in this area.

Bootstrap procedures A bootstrap procedure (Efron and Tibshirani, 1993) for CI construction uses a computer-intensive *resampling* procedure to determine CI limits. The data set to be analysed defines the population from which the resampling takes place. Some bootstrap procedures determine CI limits directly, while others use expressions like (2.16) with empirically determined CCs in place of theoretically derived values (Westfall and Young, 1993).

Suppose, for example, that an experimenter wishes to abandon the assumption of normally distributed error (while retaining the assumption of variance homogeneity) when constructing a set of 95% SCIs in a post hoc analysis of data from a single-factor

between-subjects experiment. The CC for a Scheffé analysis (which assumes normally distributed error) would be $\sqrt{(J-1)F_{.05;\,J-1,\,N-J}}$. In a bootstrap analysis the set of N within-group deviation scores $(Y_{ij} - M_j)$ is used as a 'population' from which a large number (say 10,000) of random samples of size N (obtained by sampling *with replacement*) is drawn. An ANOVA F value is calculated from each sample, and the 95th percentile of this distribution ($F_{.05;\,\text{boot}}$) is used to define the bootstrap CC_{boot} of $\sqrt{(J-1)F_{.05;\,\text{boot}}}$. The experimenter can then use *PSY* (or some other program) to carry out a post hoc analysis of the data with a user-supplied CC of CC_{boot}.

Justification for the claim that bootstrap CI procedures control the FWER depend partly on bootstrap estimation theory and partly on evidence from Monte Carlo studies. Some (but very little) discrepancy between the nominal and actual FWER will result from the use of a finite number of bootstrap samples. Even if simulation error is not an issue, some bootstrap CI procedures do not work well with small data sets (Westfall and Young, 1993). Nevertheless, there is good evidence that the bootstrap approach can provide acceptable control over error rates in a number of situations where classical ANOVA procedures do not. In practice, however, these computer-intensive procedures are currently inaccessible to many researchers.

Trimmed means procedures A *trimmed mean* is the mean of the values in a distribution that remain when a certain proportion of extreme values is removed from each tail of the distribution. Many robust estimation procedures are based on 20% trimmed means (means of the middle 60% of the relevant distribution), because removing the upper 20% and lower 20% of values from a distribution eliminates or substantially reduces the influence of outliers without doing too much damage to precision of estimation when sampling from normal distributions. A sample trimmed mean M_t is an unbiased estimate of the corresponding population trimmed mean μ_t, which is not identical to the population mean μ unless the distribution is symmetrical. In general, therefore, an analysis based on trimmed means is appropriate only if the experimenter wishes to make inferences about population trimmed means rather than population means. A population trimmed mean is closer than the population mean to most of the values in a substantially skewed distribution, and for that reason it may be preferred as a location parameter (Wilcox, 1998).

The standard error of a sample trimmed mean is slightly larger than the standard error of the sample mean when the sample is drawn from a normal distribution. For example, if $n = 10$, the standard error of the 20% trimmed mean is 0.337 and the standard error of the mean is 0.316. That is, the sample mean is a slightly more precise estimator of the population mean of a normal distribution than is the 20% trimmed mean. (The population 20% trimmed mean of a normal distribution is also the population mean.) When samples are drawn from a lognormal distribution, however, the standard error of the 20% trimmed mean (when $n = 10$) is only slightly larger (0.421), whereas the standard error of the mean is much larger (0.684). That is, the precision with which the sample 20% trimmed mean estimates the population 20% trimmed mean of a lognormal distribution is considerably greater than the precision with which the sample mean estimates the population mean of the same distribution. In practice, the standard error of

a sample trimmed mean is estimated by a procedure that relaces the extreme values discarded by the trimmed mean with less extreme values.

A number of SCI procedures based on trimmed means, sometimes in conjunction with bootstrapping, have been proposed as robust alternatives to ANOVA-model tests. For descriptions of some of these procedures, see Wilcox (2003).

Robust estimation procedures will almost certainly be widely used by researchers when their performance is more thoroughly investigated and when they can be easily implemented with user-friendly software. At the time of writing, however, there remains a great deal to be discovered about the conditions under which particular procedures work well. For example, when Keselman, Lix and Kowalchuk (1998) examined the performance of a number of stepwise multiple comparison procedures for trimmed means, they discovered that although the trimmed means MCPs provide more power than corresponding LS procedures when distributions are highly skewed, the reverse is the case in the presence of moderate skew. Although it is not clear whether this finding has direct implications for CI procedures, it does illustrate the need for very thorough investigations of the comparative performance of robust and LS procedures in a wide variety of conditions.

References

With the exception of the additional references listed below, references are given in *ANOVA via CIs.*

Additional references:

Brown, M.B. & Forsythe, A.B. (1974). The ANOVA and multiple comparisons for data with heterogeneous variances. *Biometrics, 30*, 719-724.

Efron, B. & Tibshirani,R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Gleitzman, M. (1996). Familywise error rates of protected test procedures in multivariate analysis of variance. Unpublished PhD thesis, University of New South Wales, Sydney, Australia.

Keselman, H.J., Lix, L.M. and Kowalchuk, R.K. (1998). Multiple comparison procedures for trimmed means. *Psychological Methods, 3*, 123-141.

Kreft, I. & DeLeeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

Maxwell, S.E. & Delaney, H.D. (2004). *Designing experiments and analyzing data: A model comparison perspective (2nd Ed.)*. Mahlwah, NJ: Lawrence Erlbaum.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Sawilowsky, S.S. & Blair, R.C. (1992). A more realistic look at the robustness and Type II eror probabilities of the *t* test to departures from population normality. *Psychological Bulletin, 111,* 352-360.

Wampold, B.E. & Serlin, R. (2000). The consequences of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods,* 5, 425-433.

Westfall, P.H. & Young, S.S. (1993). *Resampling based multiple testing: Examples and methods for p-value adjustment*. New York: Wiley.

Wilcox, R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300-314.

Wilcox, R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Academic Press.