Methods of Psychological Research Online 2001, Vol. 6, No. 1
Internet: http://www.mpr-online.de
Institute for Science Education

Asserting the smallness of effects in ANOVA

Bruno Lecoutre and Gérard Derzko[1]

Abstract

Statistical inference procedures dedicated to asserting the smallness of effects are commonly used in the field of bioequivalence studies in pharmacology. They are however still virtually ignored in psychology. One possible reason is that experimental investigations generally involve complex designs for which solutions have not been developed in detail. The focus here is on the generalization of these procedures to all the situations where the usual ANOVA F tests apply. Smallness confidence interval procedures, both for raw effects, such as contrasts between means and their several df extensions, and for standardized effect size measures similar to Cohen's d and f, are considered. They are illustrated and compared with alternative Bayesian procedures previously studied. From a practical viewpoint, the computations require no more than the observed effect size, the usual F ratio, and percent points of statistical distributions.

Keywords: Effect sizes; bioequivalence; individual bioequivalence; interval estimates; two one-sided tests procedure; Bayesian methods.

[1]Author's addresses: Bruno Lecoutre, ERIS, Laboratoire de Mathématiques Raphaël Salem, UMR 6085, C.N.R.S. et Université de Rouen, Mathématiques, Site Colbert, 76821 Mont-Saint-Aignan Cedex, France (bruno.lecoutre@univ-rouen.fr; http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/Eris.htm) and Gérard Derzko, SANOFI~SYNTHELABO Recherche, 374 rue du Professeur Joseph Blayac, 34184 Montpellier Cedex, France (gerard.derzko@sanofi-synthelabo.com).

Statistical programs: Windows programs that perform all the procedures exposed here are freely available at the Internet address http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/pac.htm or upon request to the first author.

Acknowledgments: We are particularly indebted to Jacques Poitevineau for helpful comments and suggestions. Our special thanks go to Victoria Bishop for improving our English. The remaining mistakes are ours.

Introduction

"*Never use the unfortunate expression 'accept the null hypothesis'.*" This is one of the recommendations of the American Psychological Association (APA) Task Force on Statistical Inference (Wilkinson and Task Force on Statistical Inference, 1999). If this recommendation is adopted in the next edition of the APA publication manual, it could put an end to one of the most often denounced abuses of Null Hypothesis Significance Testing. Indeed usual Null Hypothesis Significance Testing is essentially asymmetric. If the result is *significant*, one rejects the null hypothesis in favor of the *alternative*. If the result is *nonsignificant*, one *fails* to reject the null hypothesis, but *this is no evidence for the null hypothesis*. In particular, if the experiment is sufficiently insensitive, a descriptively large departure from the null hypothesis may be nonsignificant.

However in some cases the investigator finds that the observed difference between groups (for instance) is quite small and he/she would like to report something other than a failure to reject the null hypothesis. It is important to alert other investigators to the fact that a treatment has nothing more than a small effect. Another relevant issue is to demonstrate the fit of a theoretical model. It is essential to go beyond the conclusion that "the model is not rejected by the data" and to really demonstrate a "good-enough" fit. The aim of this paper is to provide a principled way to make such statements, and more precisely to propose reasonable effective procedures in order to assert the smallness of effects in ANOVA.

In spite of continuous warnings, a *nonsignificant* result is often improperly interpreted as evidence *in favor of the null hypothesis*. For example, Harcum (1990) noticed that incorrect null conclusions drawn from nonsignificant tests are not uncommon, even in prestigious journals. Reporting a measure of effect size (ES) is the first necessary remedy to such abuses. It can effectively prevent researchers from some unjustified conclusions in the conflicting cases where a nonsignificant result is associated with a large observed ES. However the cases of small observed ESs can be perceived by researchers as *favorable* to a conclusion of no effect, whilst at the same time they can in no way be considered as sufficient proof (Lecoutre, 2000; Lecoutre, Lecoutre & Poitevineau, 2001). Power studies (Binder, 1963; Cohen, 1962, 1977) can also be viewed as a handrail to avoid hasty generalizations. However referring to statistical papers that discuss and compare procedures (for instance Schuirmann, 1987), a more and more widespread opinion is that "for interpretation of observed results, the concept of power has no place, and confidence intervals, likelihood, or Bayesian methods should be used instead" (Goodman & Berlin, 1994). This agrees with the APA's recommendations: "once the study is analyzed, confidence intervals replace calculated power in describing results" (Wilkinson and Task Force on Statistical Inference, 1999).

In a short note, Bartko (1991) supplied psychologists with some references in biostatistics, where the issue of assserting smallness has been specifically investigated in the framework of

pharmacological bioequivalence studies. The typical aim of these studies is to demonstrate that a generic version of a drug is "neither better nor worse" than the standard drug. In order to evaluate equivalence between two experimental treatments, a small, positive value Δ is used to define an "equivalence region" $[-\Delta\,,\Delta]$ for the true difference $\delta = \mu_1 - \mu_2$ between the two treatment means. It is assumed that a difference less than Δ in either the negative or positive direction can be considered to be negligible according to the particular context of the experiment. If it is stated otherwise, a difference greater than Δ cannot be ignored. The logic of Null Hypothesis Significance Testing requires that the hypothesis to be demonstrated should be the alternative hypothesis. Consequently an appropriate test procedure for assessing equivalence must deal with the composite null hypothesis $H_0 : |\mu_1 - \mu_2| \geq \Delta$ (which is to be rejected) when opposed with the alternative $H_1 : |\mu_1 - \mu_2| < \Delta$ ("equivalence").

The seemingly natural solution consists in using the absolute value of the usual t test statistic. Then the test is to reject H_0 (to assert equivalence) if $|t|$ is small enough, formally if $|t|$ is smaller than the $(1-\alpha)\%$ lower point of its sampling distribution given $|\mu_1 - \mu_2| = \Delta$ (*i.e.* the absolute value of a *noncentral t* distribution). When the error variance is known, this test is the uniformly most powerful test for testing H_0 against the alternative H_1. Unfortunately this test has several undesirable properties which have been often alluded to (see Appendix A) which discourage its use. Therefore it will not be considered further in this article.

Rogers, Howard and Vessey (1993) described an alternative test, known as the "two one-sided tests procedure" to evaluate equivalence. This test is significant (equivalence is asserted) if the two one-sided tests $H'_0 : \mu_1 - \mu_2 = \Delta$ against $H'_1 : \mu_1 - \mu_2 < \Delta$ and $H''_0 : \mu_1 - \mu_2 = -\Delta$ against $H''_1 : \mu_1 - \mu_2 > -\Delta$ are simultaneously significant at level α. It can be easily verified that the two one-sided tests procedure is operationally identical to the procedure of asserting equivalence if the usual $100(1-2\alpha)\%$ (not $100(1-\alpha)\%$) confidence interval for $\mu_1 - \mu_2$ (centered on the observed difference $D = M_1 - M_2$) is contained within the equivalence region (Westlake, 1981; Deheuvels, 1984). This solution has gained increasing popularity and has become the reference method for the Food and Drug Administration (FDA, 1992) in the assessment of the bioequivalence of two drugs [2].

Following the APA's Task Force recommendations, here we shall systematically adopt a presentation in terms of interval estimates. Interval estimates include the decisional test procedure, but are more informative (see Brandstätter, 1999). A $100(1-\alpha\%)$ "equivalence (or smallness) confidence interval" for $\delta = \mu_1 - \mu_2$, *centered on zero* (and not on the observed

[2] It must be noted that some recent solutions intended to improve the two one-sided tests procedure (in particular, Berger & Hsu, 1996; Brown, Hwang & Munk, 1997; Munk, 2000) have the same undesirable properties as the uniformly most powerful test and are also unacceptable. Lately this has been denounced by Perlman and Wu (1999) who showed that "the Neyman-Pearson theory *desideratum* of a more (or most) powerful size α test may be scientifically inappropriate".

difference), can be deduced from the two one-sided tests procedure. We shall extend this solution on order to assert the smallness of any degree of freedom fixed effect in ANOVA. Both inferences on raw effects, such as contrasts between means, and on standardized *ES* measures similar to Cohen's indices, will be considered. A remarkable result is that all the computations require no more than the observed ES index, the usual F ratio, and percent points of a statistical distribution. In particular confidence intervals for raw effects involve the usual Student distribution.

Bayesian methods have also been proposed for bioequivalence studies (in particular Mandallaz & Mau, 1981; Selwyn, Dempster & Hall, 1981; Selwyn & Hall, 1984). The Bayesian solution to evaluate equivalence between two experimental treatments is simply obtained by stating the *posterior* distribution of δ and then computing the probability of any statement of interest (*i.e.*, for asserting equivalence, $Pr(-\Delta < \delta < \Delta \,|\, data)$). The posterior distribution expresses uncertainty about δ, conditionally on data. This is supposing a given distribution *prior* to the data. According to Jeffreys (1961), the *noninformative* approach aims at assigning the prior probability when nothing is known about the value of the parameter. In practice, noninformative prior distributions do not favor any particular value and consequently they let the data "speak for themselves" (Box & Tiao, 1973, page 2). The noninformative Bayesian approach provides *standard* (reference) methods appropriate for situations involving scientific reporting, which "should be recognized as being at least as objective as any other statistical techniques" (Berger, 1985, page 110). Furthermore Bayesian methods appear to be very well suited to asserting the magnitude of effects (see e.g., Rouanet & Lecoutre, 1983; Spiegelhalter, Freedman & Parmar, 1994; Lecoutre, Derzko & Grouin, 1995; Rouanet, 1996; Rindskopf, 1997, 1998; Lecoutre & Charron, 2000; Rouanet *et al.*, 2000; Lecoutre, Mabika & Derzko, 2001) [3].

Moreover in recent years the idea that individual bioequivalence, and not only average bioequivalence, should be demonstrated has gained increasing attention. The most useful working definition of the individual equivalence of two treatments for a population of subjects, can be stated as follows: the difference associated with one administration of each of these treatments comes from a distribution which is such as the proportion π of "small" differences, *i.e.* in the equivalence region $[-\Delta\,,\Delta]$, is "sufficiently high", *i.e.* at least equal to a proportion P (Schall & Luus, 1993; Schall, 1995). Of course this requires a within-subject design.

The paper is organized as follows. We shall introduce numerical examples involving designs which become gradually more complicated. They will serve to define some basic

[3] The use of Bayesian methods in Psychology had been advocated long before their present revival: see in particular Rozeboom, 1960; Edwards, Lindman and Savage, 1963; Edwards, 1965; Bakan, 1966; Wilson, Miller and Lower, 1967; Phillips, 1973; Novick and Jackson, 1974; Winkler 1974; Rouanet, Lépine and Pelnard-Considère, 1976.

notions about *ES* measures. We shall define and illustrate confidence interval procedures suitable for asserting the smallness of the difference between two means. Then we shall extend these procedures to ANOVA fixed effects. We shall also examine the complementary issues of asserting largeness and mediumness of effects. Then the illustrative example of a repeated measures design serves to compare the smallness confidence intervals with the standard Bayesian solution. We shall also discuss the interpretation of standardized effects for within-subject sources of variation in relation to individual equivalence.

1 Examples

1.1 Preliminary Remarks

First of all we must remember that the size of an effect must not be mistaken for its relevance. Small effects can be quite important in some fields, while large effects may have trivial practical implications. Therefore different research domains and different experimental designs require different criteria. Also it is quite reasonable to assume that criteria will evolve as progress is made in the domain under study. For the standardized difference (Cohen's d) here we shall adopt the operational criteria proposed by Cohen (1977, pages 24-27). According to these criteria the values 0.20, 0.50 and 0.80 delimit *small*, *medium* and *large* effect sizes. However these criteria are only used as benchmarks for illustrative purposes. As stated by Cohen himself, "more is to be gained that lost by supplying a common conventional frame of reference which is recommended for use only when no better basis for estimating the *ES* index is available".

Notational conventions will be introduced for the needs of generalization. Following the usual rule in statistics, we use Greek lower-case letters for parameters and corresponding Roman upper-case letters for descriptive statistics. Then, it is natural to denote the true difference between two means by δ, and hence the observed difference by D and the observed standard deviation by S. More generally, these notations will be used for any contrast between means.

1.2 Example 1: Comparison of Two Independent Group Means

This first elementary example, introduced by Rogers, Howard and Vessey (1993, Table 2 page 559), aims to demonstrate equivalence between alcoholic subjects and drug-abusive subjects on the Minnesota Multiphasic Personality Inventory (from Cannon *et al.*, 1990). The authors defined the equivalence (or smallness) region for the raw difference $\delta = \mu_1 - \mu_2$ as $\pm 10\%$ of the observed alcohol group mean M_1. Three examples of scores corresponding to different issues will be considered here. For the Lie scale ($D = 49.3 - 48.9 = +0.4$, $\frac{D}{S} = \frac{\pm 0.4}{7.52} = +0.05$,

$t[254] = +0.335$), the observed differences are small for the two criteria. Can it be inferred that the *true* raw and standardized differences are also small, *i.e.* respectively $|\delta| < 4.93$ and $|\frac{\delta}{\sigma}| < 0.20$? For the Correction scale ($D = 47.4 - 49.5 = -2.1$, $\frac{D}{S} = \frac{-2.1}{6.75} = -0.31$, $t[254] = -1.957$), only the raw observed difference is small but the usual t test is significant at one-sided level 0.05. For the Depression scale, ($D = 76.7 - 68.7 = +8.0$, $\frac{D}{S} = \frac{+8.0}{14.86} = +0.54$, $t[254] = +3.389$), the observed differences are nonsmall for the two criteria and the usual t test is significant. This last outcome refers to asserting largeness.

1.3 Example 2: Trend Analysis

The issue of accepting (or not accepting) smallness is of particular interest when testing the goodness of fit with a model. Consider here the example of trend-analysis procedures reported by Kirk (1982, pages 149-161). This is a study on the effects of sleep deprivation on hand-steadiness. 32 subjects were randomly assigned to one of the four groups characterized by the level of sleep deprivation (12, 24, 36, and 48 hours respectively). Consider the question of whether the trend of the dependent variable in the population is (approximately) linear. The observed means for each of the four groups are 2.75, 3.50, 6.25 and 9.00. The corresponding predicted means are 2.15, 4.30, 6.45 and 8.60. The linear component of the trend is significant ($F[1, 28] = 126.30$). In the conventional Null Hypothesis Significance Testing approach, an overall F test for the higher-order (quadratic and cubic) trend components is performed. This test is nonsignificant at level $\alpha = 0.05$ ($F[2, 28] = 3.28$), but this is no evidence for the null hypothesis that the trend does not depart from linearity.

1.3.1 Effect Size Measures

Cohen's f for comparing means (Cohen, 1977, page 274) is intended to give an index of "the degree of departure from no effect" that generalizes the standardized difference between two means. It is the standard deviation (σ_{m} in the author notations) of the population means μ_g divided by the common standard deviation of the populations σ. Thus here, for the overall comparison of the $\mathrm{G} = 4$ groups (with equal group sizes)

$$f = \frac{\sigma_{\mathrm{m}}}{\sigma} \quad \text{where } \sigma_{\mathrm{m}} = \sqrt{\frac{\sum_{g=1}^{\mathrm{G}} (\mu_g - \mu_.)^2}{\mathrm{G}}}$$

The values 0.10, 0.25 and 0.40 delimit small, medium and large effect sizes (pages 285-288). Cohen justified these criteria by the fact that they are equivalent to his comparable definitions for a difference of two means (in this case the standard deviation of means is equal to half of the difference). Extending Cohen's definition for main effects, a raw effect size measure of departure from polynomial adjustment is the standard deviation of residuals.

The true residuals are the G differences between the cell means μ_g and the predicted means $\tilde{\mu}_g$. By definition their mean is zero and consequently their variance is the mean of their squares. In other words, the standard deviation of residuals is their quadratic mean

$$\sqrt{\frac{\sum_{g=1}^{G}(\mu_g - \tilde{\mu}_g)^2}{G}}$$

Again dividing this index by σ gives a standardized effect size measure.

1.3.2 The Shortcomings of Usual *F* Tests

It is becoming a common practice to report the observed value of Cohen's indices [4]. For instance here, the departure from linear trend can be characterized by the observed raw *ES* index $\sqrt{\frac{0.60^2+0.80^2+0.20^2+0.40^2}{4}} = 0.55$. σ is estimated by the within-group standard deviation S, *i.e.* here the square root of the error mean square, hence the standardized index $\frac{0.55}{1.21} = 0.45$. This can hardly be considered as a small departure. Nevertheless, from the nonsignificant F ratio, Kirk stated that "because $F < F_{.05;2,28}$, it is concluded that the trend does not depart from linearity" (page 156). Appropriate inferential procedures are intended to clarify this seemingly conflicting situation.

A further question will be whether a higher-degree equation can provide a more acceptable fit. Worth noticing here is a paradoxical result of the use of the traditional null hypothesis significance test for goodness of fit: adding a cubic component to the linear trend leads to a significant result, since the F test for departure from this pattern is in this case equivalent to the F test of the quadratic trend ($F[1, 28] = 5.46$, $p < 0.05$). This result demonstrates to what extent binary decisions based on significance levels, as advocated by Chow (1996), can be inconsistent.

1.4 Example 3: A Repeated Measures Design

The experiment (Holender & Bertelson, 1975) was designed to investigate the model of processing stages in reaction time. Two repeated factors were involved: Factor A (signal frequency) with two levels, frequent ($a1$) and rare ($a2$), and Factor B (foreperiod duration), also with two levels, short ($b1$) and long ($b2$). The main research hypothesis was a null (or approximately null) interaction effect between factors A and B (*model of additive stages*). The analyzed data were, for each of the 12 subjects, the averages of several reaction time records in each of the four conditions.

[4] It has been objected that these observed indices are not unbiased estimates (e.g., Richardson, 1996). Nothing prevents us from using more efficient estimates, but this issue is presumably of secondary interest if interval estimates are used.

For illustrative purposes, a between-subject factor G, not present in the original experiment, was added. The data are reported in Table 1. They have been previously analyzed with standard Bayesian methods in Rouanet and Lecoutre (1983) and Rouanet (1996). Let us consider the $A.B$ interaction effect and the main effect of factor G as typical respective examples of one *df* and several *df* sources of variations.

1.4.1 Example of One *df* Effect: $A.B$ Interaction

The $A.B$ raw effect size will be naturally measured by the difference between mean differences, hence the observed value $D = (353.25 - 403.67) - (388.92 - 437.25) = -2.08$. The appropriate σ for standardization is the square-root of the sampling variance of individual interaction effects. It is estimated by the within-group standard deviation of the 12 observed individual effects (which can be termed "relevant data for the $A.B$ effect") reported in Table 1, *i.e.* $S = 33.28$, hence the observed standardized measure $\frac{D}{S} = -0.06$.

Table 1: Example 3 (A repeated measures design): Reaction time data. Reaction times are in ms. G = added between subject factor; A = signal frequency; B = foreperiod duration; $A.B$ = interaction.

					Relevant data	
	$a1b1$	$a2b1$	$a1b2$	$a2b2$	$A.B$	G
	387	435	416	473	+9	427.75
group	321	336	343	368	+10	342.00
$g1$	333	362	358	390	+3	360.75
	344	430	352	393	-45	379.75
mean	346.25	390.75	367.25	406.00	-5.75	377.56
	368	432	432	504	+8	434.00
group	357	367	394	411	+7	382.25
$g2$	336	346	340	421	+71	360.75
	387	454	438	496	-9	443.75
mean	362.00	399.75	401.00	458.00	+19.25	405.19
	345	408	417	479	-1	412.25
group	358	389	372	407	+4	381.50
$g3$	317	375	341	392	-7	356.25
	386	510	464	513	-75	468.25
mean	351.50	420.50	398.50	447.75	-19.75	404.56
mean	353.3	403.7	388.9	437.3	-2.08	395.77
	within-group standard deviation				33.28	41.99

The F ratio is nonsignificant ($F[1,9] = 0.05$, $p = 0.83$): the model of additive stages is not rejected. Since the observed effect can be assessed as small, the situation can be perceived as *favorable* to the acceptance of the additive model. However we shall see that this impressionistic judgment is not fully justified (Section 4.1).

1.4.2 Example of Several *df* Effect: Factor *G*

For factor G with $G = 3$ levels, the observed numerator of Cohen's f is the standard deviation of the three group means 377.56, 405.19 and 404.56, *i.e.* 12.88. The appropriate σ for standardization is the square-root of the sampling variance of individual means (averaged over the four conditions). It is estimated by the within-group standard deviation of the 12 observed individual means (relevant data for the G effect) reported in Table 1, *i.e.* $S = 41.99$, hence the observed standardized index $\frac{12.88}{41.99} = 0.31$. The F ratio is nonsignificant ($F[2,9] = 0.56$, $p = 0.59$). However, according to Cohen's criteria, the smallness of the standardized effect cannot be asserted.

Note that Rouanet (1996) defined the raw effect size measure as the *corrected* standard deviation of the group means, *i.e.* in his notations $s_G = 15.77$. This is an alternative to Cohen's uncorrected solution [5].

2 Confidence Intervals for Asserting Smallness

Firstly, we shall consider asserting the smallness of the difference between two normal means from independent groups with the usual assumption of a common variance σ^2. Subsequently, we shall see that confidence intervals for asserting the smallness of any *df* effects in ANOVA are straightforward generalizations of the solutions for this basic situation.

2.1 Asserting the Smallness of the Difference Between Two Means

2.1.1 Raw Difference

According to our notational conventions, $\delta = \mu_1 - \mu_2$ and $D = M_1 - M_2$ are respectively the true (population) and observed mean differences.

[5] Another alternative is to consider the quadratic mean of the pairwise differences between groups. This solution has been extensively developed in statistical software packages that include all the procedures presented in this paper (Lecoutre & Poitevineau, 1992; Lecoutre, 1996). From a formal viewpoint, all these *ES* indices are proportional and therefore equivalent. The choice of a particular solution is essentially a matter of convention and easiness for interpreting and communicating results.

Usual confidence interval centered on D. The usual $100(1-\alpha)\%$ confidence interval (UCI) for δ is

$$\left[D - S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\, t_{\nu;1-\frac{\alpha}{2}}\ ,\ D + S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\, t_{\nu;1-\frac{\alpha}{2}}\right] \tag{1}$$

where $t_{\nu;1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-percentile of the standard Student t distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom.

Smallness confidence interval. A $100(1-\alpha)\%$ "smallness confidence interval" (SCI) for δ (centered on zero) is

$$[-\delta_{SCI}\,,\delta_{SCI}] \text{ where } \delta_{SCI} = |D| + S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\, t_{\nu;1-\alpha} \tag{2}$$

It can be verified that δ_{SCI} is the greater in absolute value of the limits of the $100(1-2\alpha)\%$ UCI. Note the computational analogy between δ_{SCI} and the upper bound of the usual confidence interval: the observed difference D is simply replaced with its absolute value $|D|$.

Given the smallness region $[-\Delta\,,\Delta]$, smallness can be asserted (with confidence level $(1-\alpha)\%$) if the SCI is contained within $[-\Delta\,,\Delta]$, *i.e.* if $\delta_{SCI} < \Delta$. This is equivalent to accepting smallness at level α with the two one-sided tests procedure.

Improved smallness confidence interval. The $(1-\alpha)\%$ smallness confidence interval above has a coverage which is larger than $1-\alpha$, *i.e.* in the long run more than $(1-\alpha)\%$ of the SCI contains δ. Furthermore, the procedure does not distinguish between the case where the usual t test of the null hypothesis $\delta = 0$ is nonsignificant and the case where it is significant. In order to remedy this situation the procedure can be improved as follows: when the usual t test for the null hypothesis $\delta = 0$ is nonsignificant at the one-sided level α consider the $100(1-2\alpha)\%$ usual confidence interval for δ, and when this test is significant (*i.e.* when the $100(1-2\alpha)\%$ UCI does not contain zero) replace either the lower limit with 0 if $D > 0$ or the upper limit with 0 if $D < 0$. This results in the following (nonsymmetrical) confidence interval

$$\left[\min\left(0\,,D - S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\, t_{\nu;1-\alpha}\right)\ ,\ \max\left(0\,,D + S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\, t_{\nu;1-\alpha}\right)\right] \tag{3}$$

This improved SCI interval has been derived by several authors (see e.g., Bofinger, 1992; Hsu *et al.*, 1994). It is contained in the symmetrical SCI interval and explicitly takes the outcome of the usual one-sided t test into account. Moreover the improved interval has coverage probabilities $1-\alpha$ for $\delta \neq 0$ (and 1 at $\delta = 0$). Nevertheless it must be stressed that the same decision of asserting/not asserting smallness will be concluded from both intervals.

2.1.2 Numerical Applications

Practical computations. A remarkable property is that the smallness confidence interval (as well as the usual one) can be immediately computed from the observed difference D and the

usual t test statistic. This results from the fact that $S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$, *i.e.* the square root of the sampling error variance, is the denominator of the t statistic. Therefore, it can be computed as $\frac{D}{t}$ (provided that $D \neq 0$). Hence the bounds of the *SCI* are given by

$$\delta_{SCI} = |D| + \frac{D}{t}t_{\nu;1-\alpha} = |D|\Big(1 + \frac{t_{\nu;1-\alpha}}{|t|}\Big) \tag{4}$$

Example 1, Lie scale. $D = 49.3 - 48.9 = +0.4$, $t = +0.335$, $\Delta = 4.93$ (10% of the observed alcohol group mean), $\alpha = 0.05$.

95% *UCI* for δ	$[-1.95\,,+2.75]$	$(t_{254;0.975} = +1.969)$
90% *UCI* for δ	$[-1.57\,,+2.37]$	$(t_{254;0.95} = +1.651)$
95% *SCI* for δ	$[-2.37\,,+2.37]$	

The 95% *SCI* is contained within the smallness region $[-4.93\,,+4.93]$: the smallness of the true difference can be asserted for the criterion.

Example 1, Correction scale. $D = 47.4 - 49.5 = -2.1$, $t = -1.957$, $\Delta = 4.74$, $\alpha = 0.05$. The 95% *SCI* is $[-3.87\,,+3.87]$: smallness can also be asserted. In this case the usual t test is significant at one-sided level 0.05, or equivalently the 90% *UCI* $[-3.87\,,-0.33]$ does not contain zero. Using the improved procedure, the 95% *SCI* is $[-3.87\,,0]$. It explicitly combines the smallness procedure with the outcome of the usual one-sided t test: we can simultaneously conclude that not only is δ small but also that δ is negative with confidence level 95%. Note that the 95% improved *SCI* is included within the 95% usual confidence interval $[-4.21\,,+0.01]$. This result will be discussed later on.

Example 1, Depression scale. $D = 76.7-68.7 = +8.0$, $t = +3.389$, $\Delta = 7.67$. In this example smallness cannot be asserted since the observed difference is larger than Δ. The improved *SCI* for δ is $[0\,,+11.9]$. However it does not tackle the issue of asserting the largeness of the effect.

2.1.3 Standardized Difference

The above solution can be readily extended to the standardized difference $\frac{\delta}{\sigma}$ (Cohen's d). If the smallness region for $|\frac{\delta}{\sigma}|$ is $[-\Delta^*\,,\Delta^*]$, the principle of the two-one sided tests procedure can be applied to test $H_0 : |\frac{\delta}{\sigma}| \geq \Delta^*$ (which is to be rejected) against the alternative $|\frac{\delta}{\sigma}| < \Delta^*$. This procedure is equivalent to computing the usual (equal tails) $100(1-2\alpha)\%$ confidence interval for $\frac{\delta}{\sigma}$ and to asserting equivalence if this interval is contained within the smallness region. This involves no other conceptual difficulty, but only additional technical work. Indeed the statistical distribution, namely the *noncentral* t distribution (instead of the usual t), is more complicated. However this distribution is already familiar to power analysts and computer programs are now available.

$100(1-\alpha)\%$ *(equal tails) usual confidence interval.* This interval is $[\delta^*_{UCI_1}, \delta^*_{UCI_2}]$, where $\delta^*_{UCI_1}$ and $\delta^*_{UCI_2}$ are defined by

$$Pr\left(t'_\nu\left(\frac{\delta^*_{UCI_1}}{\frac{D}{S}}t\right) < t\right) = \frac{\alpha}{2} \quad \text{and} \quad Pr\left(t'_\nu\left(\frac{\delta^*_{UCI_2}}{\frac{D}{S}}t\right) > t\right) = \frac{\alpha}{2} \tag{5}$$

$100(1-\alpha)\%$ *smallness confidence interval.* This interval is $[-\delta^*_{SCI}, \delta^*_{SCI}]$, where δ^*_{SCI} is defined by

$$Pr\left(t'_\nu\left(\frac{\delta^*_{SCI}}{\frac{D}{S}}t\right) < |t|\right) = \alpha \tag{6}$$

As for the raw difference, δ^*_{SCI} is the greater in absolute value of the limits of the $100(1-2\alpha)\%$ *UCI*. The procedure can be improved in the same way, resulting in the following (nonsymmetrical) confidence interval

$$\left[\min(0, \delta^*_{UCI_1}), \max(0, \delta^*_{UCI_1})\right] \tag{7}$$

which takes the outcome of the one-sided t test for the null hypothesis $\delta = 0$ into account.

2.1.4 Numerical Applications

Example 1, Lie scale. $\frac{D}{S} = +0.053$, $t = +0.335$, $\Delta^* = 0.20$, $\alpha = 0.05$.

$$\begin{aligned} 95\%\ UCI \text{ for } \tfrac{\delta}{\sigma} \quad & [-0.26\,, +0.36] \\ 90\%\ UCI \text{ for } \tfrac{\delta}{\sigma} \quad & [-0.21\,, +0.31] \\ 95\%\ SCI \text{ for } \tfrac{\delta}{\sigma} \quad & [-0.31\,, +0.31] \end{aligned}$$

The 95% *SCI* is not contained within the smallness region $[-0.20\,, +0.20]$: the smallness of the true standardized difference cannot be asserted for the criterion. Exact computations require the cumulative distribution function of the noncentral t distribution. For instance δ^*_{SCI} is defined by $Pr(t'_{254}(6.295\delta^*_{SCI}) < 0.335) = 0.05$. By successive approximations over the noncentrality parameter it can be found that $Pr(t'_{254}(1.980) < 0.335) = 0.05$, hence $\delta^*_{SCI} = 1.980/6.295 = 0.314$ [6]. Of course, when the number of *df* is sufficiently large, a normal approximation can be used. Alternatively in this case, a good approximation of the confidence bounds for $\frac{\delta}{\sigma}$ can be found by dividing the bounds computed for δ by the observed value S. For instance here, with three decimal places these approximations give 0.315 instead of 0.314 for the bound of the *SCI*.

[6] As a more straightforward result, δ^*_{SCI} is the percentile of a distribution introduced in the fiducial framework by Fisher (1990, pages 126-127). However computer programs for this distribution, called lambda-prime in Lecoutre, 1999, are less common.

2.2 Asserting the Smallness of Effects in ANOVA: General Setting

The elementary situation of two independent groups is nothing but a particular case of a simple one-way layout and thus can alternatively be analyzed by means of the corresponding ANOVA. Here the following relations hold with the usual ANOVA parameters σ^2_{effect} and σ^2_{error}: $\delta^2 = 2\sigma^2_{\text{effect}}$ and $\sigma^2 = \sigma^2_{\text{error}}$. This outlines the general setting for asserting the smallness of fixed effects in ANOVA. The raw *ES* index can be any quantity proportional to σ_{effect} and the appropriate standardization variance is proportional to the ANOVA error parameter σ^2_{error}. All the procedures exposed hereafter apply to these general definitions. Note that for the raw *ES* index the proportionality property is essential to ensure that a null effect size is actually equivalent to the null hypothesis $\sigma^2_{\text{effect}} = 0$ involved in the F test. In the case of unbalanced designs, this property also ensures that all the problems linked to the choice of appropriate weights are taken into account. In particular constructing an appropriate *ES* index is technically neither more nor less complicated than computing the usual ANOVA sum of squares.

In the particular case of a one *df* source of variation, the raw *ES* index is the effect of a single contrast between means, therefore a directional (signed) quantity. Consequently, the same notations introduced for a difference of means, *i.e.* δ and D, can be used. Of course, in the case of several *df* effects, the *ES* index is a nondirectional (unsigned) quantity (like $|\delta|$) and new notations are needed. As general notations, we use λ and L respectively for the true and observed *ES* indices. A smallness region for λ will be defined as $[0\,,\Lambda]$. With these definitions, smallness confidence intervals for a ν_1 *df* effect are straightforward generalizations of the formulas for the difference between two means.

It follows that asserting the smallness of raw effects for fixed sources of variation from an ANOVA table only involves one real additional difficulty. This is to select appropriate, easily interpretable, *ES* measures δ or λ. For several *df* sources, we have considered Cohen's indices, mainly because they are familiar to psychologists. However procedures also apply to alternative definitions.

2.3 Smallness Confidence Intervals for ANOVA Fixed Effects

All the procedures are straightforward generalizations of the two one-sided tests procedure. A justification is given in Appendix B. Assuming $L \neq 0$, formulas only involve the observed effect size measure and the usual ANOVA F ratio (with ν_1 and ν_2 *df*). They include the particular case of a one *df* effect, setting $\lambda = |\delta|$ and $L = |D|$.

2.3.1 Raw Effects

$100(1-\alpha)\%$ *usual confidence interval for* δ $(\nu_1 = 1)$.

$$\left[D - \frac{|D|}{\sqrt{F}} t_{\nu_2;1-\frac{\alpha}{2}} \, , \; D + \frac{|D|}{\sqrt{F}} t_{\nu_2;1-\frac{\alpha}{2}}\right] \tag{8}$$

Note that the generalization of the usual confidence interval to the $\nu_1 > 1$ *df* case involves another parameter other than λ. This is beyond the scope of this paper.

$100(1-\alpha)\%$ *smallness confidence interval for* λ.

$$\left[0 \, , \; L\Big(1 + \frac{t_{\nu_2;1-\alpha}}{\sqrt{\nu_1 F}}\Big)\right] \tag{9}$$

2.3.2 Standardized Effects

$100(1-\alpha)\%$ *(equal tails) usual confidence interval for* $\frac{\delta}{\sigma}$ $(\nu_1 = 1)$. This interval is $[\delta^*_{UCI_1} \, , \delta^*_{UCI_2}]$ where $\delta^*_{UCI_1}$ and $\delta^*_{UCI_2}$ are defined by

$$Pr\Big(t'_{\nu_2}\Big(\frac{\delta^*_{UCI_1}}{\frac{D}{S}}\sqrt{F}\Big) < \sqrt{F}\Big) = \frac{1}{2}\alpha \quad \text{and} \quad Pr\Big(t'_{\nu}\Big(\frac{\delta^*_{UCI_2}}{\frac{D}{S}}\sqrt{F}\Big) > \sqrt{F}\Big) = \frac{1}{2}\alpha \tag{10}$$

$100(1-\alpha)\%$ *smallness confidence interval for* $\frac{\lambda}{\sigma}$. This interval is $[0\,, \lambda^*_{SCI}]$ where λ^*_{SCI} is defined by

$$Pr\Big(t'_{\nu_2}\Big(\frac{\lambda^*_{SCI}}{\frac{L}{S}}\sqrt{\nu_1 F}\Big) < \sqrt{\nu_1 F}\Big) = \alpha \tag{11}$$

2.3.3 Illustration: Example 2 (Trend Analysis)

Kirk's conclusion that "the trend does not depart from linearity" is clearly not supported by the data, even if "does not depart" is understood as "departs slightly". This is evidenced by the *SCI* for the *ES* indices λ (the quadratic mean of residuals) and $\frac{\lambda}{\sigma}$ reported in Table 2.

Table 2: Example 2 (Trend analysis): 95% smallness confidence intervals for raw and standardized effect size indices (observed means: 2.75, 3.50, 6.25, 9.00; $S = 1.21$).

Source Departure from	-$Lin(G)$ linear trend	-($Lin(G)$+$Qua(G)$) linear+quadratic trends	-($Lin(G)$+$Cub(G)$) linear+cubic trends
F ratio	3.278 [2,28]	1.093 [1,28]	5.463 [1,28]
Predicted means	2.15 4.30 6.45 8.60	2.65 3.80 5.95 9.10	2.25 4.00 6.75 8.50
Raw index	$L = 0.548$	$L = 0.224$	$L = 0.500$
95% *SCI* for λ	$[0\,, 0.91]$	$[0\,, 0.59]$	$[0\,, 0.86]$
Standardized index	$\frac{L}{S} = 0.453$	$\frac{L}{S} = 0.185$	$\frac{L}{S} = 0.413$
95% *SCI* for $\frac{\lambda}{\sigma}$	$[0\,, 0.76]$	$[0\,, 0.48]$	$[0\,, 0.71]$

Neither does the nonsignificant F allow us to conclude that the trend departs from linearity. This is a typical example of inconclusive data. A further question is whether a higher-degree equation can provide an acceptable fit. From Table 2, a two-degree polynomial, including linear and quadratic components, is shown to give a clearly better fit, with more acceptable upper limits for λ and $\frac{\lambda}{\sigma}$. Nevertheless, a more definite conclusion would need more data.

2.3.4 Numerical Applications: Reanalyzing Data

Computations for fixed ANOVA effects are exactly the same as for the elementary situation of the difference between two means. Furthermore, the procedures can be automatically applied to data previously analyzed with usual t or F tests, and especially to published results. Of course precise calculations require both the observed ES index and the tests statistic being to be known (or recomputed) with sufficient accuracy. For instance, in Example 3, for the $A.B$ effect ($D = -2.08333$, $F[2, 9] = 0.04703$, $t_{9;0.95} = 1.83311$), the 95% SCI interval for δ with two decimal places is $[-19.69\,, +19.69]$. This means we have to know F with at least five decimal places. When only rounded values are available, a cautious attitude that should be adopted is to consider the "worst interval". For instance, let us assume that we only know $D = -2.08$ and $F = 0.05$ with two decimal places. Then the most unfavorable case (with respect to smallness) is $D = -2.085$ and $F = 0.045$. This gives the 95% SCI $[-20.10\,, +20.10]$. Note that in the most favorable case ($D = -2.075$ and $F = 0.055$), the SCI should be $[-18.29\,, +18.29]$. Here the worst interval is only slightly larger than the exact one. However in some cases, especially for very small values of F, there can be far more discrepancies in results.

It is also possible for only the p-value to be known. In this case, the test statistic can be derived with a detailed table (or by program). For instance in the above example, let us assume that we do not know the F ratio but only $p = 0.83$. Considering the most unfavorable case $p = 0.835$, we obtain $F = 0.0460$, and for $D = -2.085$ the worst 95% SCI $[-19.90\,, +19.90]$ [7].

3 Not Accepting Smallness Is Not Accepting Nonsmallness and Nonsmallness Is Not Largeness

Of course warnings about casual acceptance of null hypotheses apply here: not accepting smallness is not accepting nonsmallness. Other procedures must be used for asserting nonsmallness. From a logical viewpoint, a "nonsmallness region" for δ, as the complement of

[7] Moreover reanalysis is again feasible when no descriptive statistics about means are available. However, in the case of complex designs, this generally needs some sophistication.

the smallness region, corresponds to $|\delta| \geq \Delta$. But, from a methodological viewpoint, such a nonsmallness region gives very poor information, since it means that δ can be large in a positive as well as a negative direction. This implies that *largeness* must be distinguished from nonsmallness and requires the direction of the effect to be explicitly taken into account.

Here we only consider the case of a one *df* effect. Procedures for the case of several *df* are conceptually much more difficult and also beyond the scope of this article. The usual confidence interval could be used as a routine procedure. However other types of intervals presumably have a role to play in addressing precise questions about the magnitude of effects. In particular, one-sided intervals can be used to assert *largeness* either in a positive or negative direction. Intervals centered on a non-null prespecified value allow us to assert *mediumness*. Methods will be illustrated by numerical applications.

3.1 Asserting Largeness

Example 1, Depression scale. $D = 76.7 - 68.7 = +8.0$, $\frac{D}{S} = +0.538$, $t = +3.389$. Smallness cannot be asserted for the smallness region defined by $\pm 10\%$ of the observed alcohol group mean (7.67), but can it be inferred that δ is larger than $\Delta = +7.67$? The answer is negative, since the one-sided 95% confidence interval $[+4.10\,, +\infty[$ (deduced from the usual 90% *CI* $[+4.10\,, +11.90]$) contains +7.67. Note that there are situations where it is of interest to assert "nonlargeness in a direction". In clinical trials, this issue is known as one-sided or "clinical" equivalence. A typical clinical equivalence study is designed to demonstrate that a new drug is "as least as good" as the standard (*i.e.* "not worse" than it). As an example in psychology, Lecoutre (1992) studied elementary problem solving situations about probability. She aimed at demonstrating that students with a thorough background in the theory of probability did not succeed any better than students with no background, *i.e.* that the corresponding difference δ was less than a "nonlarge" positive value Δ. Clearly, one-sided intervals are also appropriate for this case.

3.2 Asserting Mediumness

For illustrative purposes, let us consider the alternative issue of asserting that δ is *medium*, say for instance between 5% (+3.835) and 15% (+11.505) of the observed alcohol group mean. From a formal viewpoint, this amounts to demonstrating that δ is close to the mid-value $\tilde{\delta} = +7.67$, *i.e.* to asserting the smallness of $\delta - \tilde{\delta}$. First a 95% *SCI* for $\delta - \tilde{\delta}$ (centered on zero) is computed by substituting $|D - \tilde{\delta}| = 0.33$ for $|D|$. We find here $[-4.23\,, +4.23]$. Then the required 95% "mediumness confidence interval" (*MCI*) for δ $[+3.44\,, +11.90]$ (centered on +7.67) is deduced by adding $\tilde{\delta}$ to the bounds. This interval is not contained within the mediumness region $[+3.835\,, +11.505]$: mediumness cannot be asserted. Note that in the present case the 95% *MCI* for δ is included within the 95% usual confidence interval

$[+3.35\,,+12.65]$; this result will be discussed later on.

Note again that the *MCI* is appropriate for asserting nonsmallness with asymmetrical regions. Indeed the smallness region can be generalized to $[\Delta'\,,\Delta'']$, where $\Delta' < 0$ and $\Delta'' > 0$, so that a greater effect in one direction than the other can be allowed. This amounts to considering an interval for δ centered on $\tilde{\delta} = \frac{1}{2}(\Delta' + \Delta'')$.

Standardized effects. Similar procedures apply to standardized effects. For instance, for the depression scale ($\frac{D}{S} = +0.538$), can it be asserted that $\frac{\delta}{\sigma}$ is relatively close to $+0.50$, say between $+0.20$ and $+0.80$? The 95% *MCI* for $\frac{\delta}{\sigma}$ centered on $+0.50$ is $[+0.198\,,+0.802]$. It is not included within the mediumness region $[+0.20\,,+0.80]$: from a strict decisional viewpoint, mediumness cannot be asserted with confidence level 95%.

4 Illustration: A Repeated Measures Design (Example 3)

This example will serve to show the feasibility of the procedures in the case of realistic complex designs [8] and to compare the frequentist and Bayesian solutions.

Standard Bayesian procedures for asserting the smallness (and more generally the magnitude) of effects in ANOVA have been developed in detail in Rouanet and Lecoutre (1983). The reader can also refer to Rouanet (1966) for an introductory presentation dedicated to psychologists. We will use the numerical results given in these two articles to compare the frequentist (confidence intervals) and Bayesian solutions. Technical results are briefly recapitulated in Appendix C.

4.1 Asserting the Magnitude of Effects

The main results, both for raw and standardized effects, are summarized in Table 3 and compared with the standard Bayesian solution. In order to facilitate the comparison with Rouanet's (1996) Table 4, her we adopt his definition of effect size measures (involving the corrected standard deviation instead of the uncorrected one considered in Cohen's f). For the same reason, *SCI* as well as the Bayesian smallness credibility intervals [9] have been

[8] It could be objected that the analysis of this example is simplified because there is only one *df* within-subject sources. It must be noted that all the procedures described here can be used for several *df* within-subject sources with the same validity conditions as for univariate ANOVA F ratios (assuming circularity conditions on covariance matrices). Otherwise, procedures for asserting the smallness of MANOVA effects have been developed, especially in the Bayesian framework (Lecoutre & Poitevineau, 1992; Schervish, 1995; Rouanet *et al.*, 2000), but they are beyond the scope of this paper.

[9] The difference in terminology (credibility *vs* confidence) between Bayesian and frequentist procedures is common. It emphasizes the difference in interpretation. In the frequentist framework we condition

computed for 90% levels.

Table 3: Example 3 (Repeated measures design): Frequentist and Bayesian procedures for asserting the importance of effects (using Rouanet's definitions of *ES* measures).

Asserting smallness

		Procedures:	Frequentist		Bayesian
Source	Observed index	F ratio	95% *SCI*	90% *SCI*	90% Bayesian *SCI*
Raw effect index λ					
G	$L = 15.77$	0.56 [2,9]	[0 , 43.0]	[0 , 36.3]	[0 , 42.8]
$G.A$	$L = 8.92$	0.64 [2,9]	[0 , 23.3]	[0 , 19.8]	[0 , 23.1]
$G.B$	$L = 15.40$	2.19 [2,9]	[0 , 28.9]	[0 , 25.6]	[0 , 27.5]
$A.B$	$\lvert D\rvert = 2.08$	0.05 [1,9]	[0 , 19.7]	[0 , 15.4]	[0 , 17.9]
$G.A.B$	$L = 19.76$	1.41 [2,9]	[0 , 41.3]	[0 , 36.0]	[0 , 39.7]
Standardized effect index $\frac{\lambda}{\sigma}$					
G	$\frac{L}{S} = 0.38$	0.56 [2,9]	[0 , 0.96]	[0 , 0.83]	[0 , 0.94]
$G.A$	$\frac{L}{S} = 0.40$	0.64 [2,9]	[0 , 0.99]	[0 , 0.86]	[0 , 0.96]
$G.B$	$\frac{L}{S} = 0.74$	2.19 [2,9]	[0 , 1.37]	[0 , 1.22]	[0 , 1.29]
$A.B$	$\frac{\lvert D\rvert}{S} = 0.06$	0.05 [1,9]	[0 , 0.54]	[0 , 0.43]	[0 , 0.49]
$G.A.B$	$\frac{L}{S} = 0.59$	1.41 [2,9]	[0 , 1.20]	[0 , 1.06]	[0 , 1.14]

Asserting largeness

		Procedures:	Frequentist		Bayesian
Source	Observed index	F ratio	90% *UCI*	90% *LCI*	90% Bayesian *LCI*
Raw effect index δ					
A	$D = +49.38$	59.06 [1,9]	[+37.6 , +61.2]	[+40.5 , $+\infty$[	[+40.5 , $+\infty$[
B	$D = +34.63$	33.25 [1,9]	[+23.6 , +45.6]	[+26.3 , $+\infty$[	[+26.3 , $+\infty$[
Standardized effect index $\frac{\delta}{\sigma}$					
A	$\frac{D}{S} = +2.22$	59.06 [1,9]	[+1.22 , +3.16]	[+1.41 , $+\infty$[	[+1.41 , $+\infty$[
B	$\frac{D}{S} = +1.66$	33.25 [1,9]	[+0.84 , +2.43]	[+1.01 , $+\infty$[	[+1.01 , $+\infty$[

It can be verified that the frequentist and Bayesian procedures for asserting smallness give distinct solutions. However one can also see that the 90% Bayesian smallness credibility

on the parameter, so the interpretation must be that the confidence interval is a *random* interval and that $1-\alpha$ refers to a long run repetition of the same experiment: in the long run $(1-\alpha)$% of computed confidence intervals will contain the "true value" of the parameter; each interval in isolation has either a 0 or 100% probability of containing it. In the Bayesian framework we condition on the data, so we can refer directly to the probability of the parameter being in the *fixed* interval associated with the current experiment.

bounds are always between the 90% and 95% frequentist bounds. This can be contrasted with the fact that the bounds of the usual intervals centered on the observed effects (and thus the largeness bounds) coincide. This issue is developed in Appendix C.

We agree with Rouanet's conclusions that "largeness is asserted" for the main effects of factors A and B, and that "smallness is not asserted" for the other sources. However the interpretation of the standardized indices $\frac{\delta}{\sigma}$ and $\frac{\lambda}{\sigma}$ for within-subject sources of variation deserves special attention in relation to individual effects. This point is discussed for the one *df* case in the following Section which can be viewed as a starting point for further research.

4.2 Individual Smallness: Interpreting Standardized Effects for One *df* Within-Subject Sources of Variation

The basic situation concerns the difference between paired observations. In this case δ and σ are respectively the mean and standard deviation of the parent population distribution of the n individual differences D_i. Assuming normality, the proportion of population values larger than Δ is a parameter π_Δ, determined by δ and σ and defined as

$$\pi_\Delta = Pr\Big(N(\delta, \sigma^2) > \Delta\Big) = Pr\left(N(0,1) < \frac{\delta - \Delta}{\sigma}\right) \tag{12}$$

Consequently a statement about $\frac{\delta}{\sigma}$ corresponds to the particular case $\Delta = 0$ and is basically a statement about π_0. For instance the inequality $|\frac{\delta}{\sigma}| < 0.20$ means that the proportion π_0 of positive (or negative) population values is between 42.1% ($z_{0.421} = -0.20$) and 57.9% ($z_{0.579} = +0.20$). Also $\frac{\delta}{\sigma} > +0.80$ means that the proportion of positive population values is more than 78.8% ($z_{0.788} = +0.80$), or equivalently that the proportion of negative population values is less than 21.2%. More generally, the inequality $\frac{\delta-\Delta}{\sigma} > +0.80$ means that the proportion of population values larger than Δ is more than 78.8%. In particular, if Δ is a large positive value, this gives a direct and elegant way both to assert average largeness and to handle the issue of individual differences. Confidence intervals about π_Δ are directly derived from intervals about $\frac{\delta-\Delta}{\sigma}$ computed by substituting $\frac{D-\Delta}{S}$ for $\frac{D}{S}$ in formulas for standardized effects.

The procedures can again be extended to the case of the proportion of population values included within two limits Δ' and Δ'' ($\Delta' < \Delta''$), *i.e.* $\pi_{[\Delta',\Delta'']} = \pi_{\Delta''} - \pi_{\Delta'}$. In particular, if Δ is a small positive value, a high proportion $\pi_{[-\Delta,\Delta]}$ ensures, not only the smallness of the average effect, but also the smallness of a large proportion of population effects. Here we touch on the issue of individual bioequivalence. Unfortunately, inferences about the parameter $\pi_{[-\Delta',\Delta'']}$ cannot be derived in a simple manner and require numerical integration methods. Alternatively, separate inferences about $\pi_{\Delta'}$ and $\pi_{-\Delta''}$ with a Bonferroni adjustment can be made.

In the case where subjects are nested within several groups, with different mean group effects (and possibly different variances), a separate analysis can be reported for each group. Also an overall statement about the averaged (over groups) population proportion can be computed by numerical methods.

Numerical applications. Let us return to the original design with a single group of 12 subjects. For factor A ($S = 21.52$, $\frac{D}{S} = +2.29$, $F[1, 11] = 63.17$), we find the 90% LCIs $[+1.53\ ,\ +\infty[$ for $\frac{\delta}{\sigma}$, and consequently $[93.7\%\,, 100\%]$ for the proportion of positive population differences π_0. Given for instance $\Delta = +25$, we similarly find the intervals $[+0.630\,, +\infty[$ for $\frac{\delta - 25}{\sigma}$ and $[73.6\%\,, 100\%]$ for π_{+25} (the proportion of population differences larger than $+25$). For any α and any given proportion P, we can also determine the larger value Δ which is compatible with the hypothesis that π_Δ is at least $100P\%$. For instance, setting $\alpha = 0.10$ and $P = 0.75$, we find by successive approximations $\Delta = +23.8$ which gives the 90% LCI $[75\%\,, 100\%]$ for $\pi_{+23.8}$ (the proportion of population differences larger than $+23.8$). This is a comprehensive statement for asserting largeness at an individual level. The larger the proportion P and the value Δ, the more the individual largeness conclusion is supported (for a specified α).

For the interaction $A.B$ ($S = 34.495$, $\frac{D}{S} = -0.060$, $F[1, 11] = 0.044$, the 90% SCI $[0\,, 0.429]$ for $|\frac{\delta}{\sigma}|$ corresponds to the interval (symmetrical around 50%) $[37.1\%\,, 62.9\%]$ for π_0. Given a smallness region $[-\Delta\,, \Delta]$ for δ, separate confidence intervals can be computed for $\pi_{-\Delta}$ (the proportion of population interaction effects larger than $-\Delta$) and $1 - \pi_\Delta$ (the proportion of population interaction effects smaller than Δ). For instance here we find the 90% UCI $[54.9\%\,, 88.1\%]$ for π_{-25} and $[58.8\%\,, 90.6\%]$ for $1 - \pi_{+25}$.

In the Bayesian framework, the posterior distribution of $\pi_{[-\Delta,+\Delta]} = \pi_{-\Delta} - \pi_\Delta$ (the proportion of population interaction effects smaller than Δ in absolute value) can be derived in a straightforward manner. The procedure is more sophisticated, but credibility intervals can be computed by numerical integration or simulation. For instance we find that $Pr(37.6\% < \pi_{[-25,+25]} < 62.0\%) = 0.90$. Alternatively, given γ and P, we can search for Δ such as $Pr(\pi_{[-\Delta,+\Delta]} > P) = \gamma$. For instance we find here $Pr(\pi_{[-58.5,+58.5]} > 75\%) = 0.90$: the proportion of population interaction effects smaller than 58.5 in absolute value is at least 75% with the probability 0.90. This is a comprehensive statement for asserting smallness at an individual level. The larger γ and P and the smaller Δ, the more the individual smallness conclusion is supported.

Conclusion

The aim of this paper was to propose effective solutions in order to assert the smallness of effects in ANOVA. Reasonable procedures are available both in the frequentist and Bayesian frameworks and in most applications give convergent results. It follows that suitable interval

estimates can easily be reported as a routine part of an ANOVA.

Theoretical discussions about the frequentist and Bayesian frameworks are outside the scope of this article. The interested reader can refer to Rouanet *et al.* (2000) and Lecoutre, Lecoutre & Poitevineau (2001). One must only have in mind the fundamental opposition between the two approaches. While the frequentist solution is conditional on parameters and involves all potential data (that have not occurred), the Bayesian approach is conditional on observed data.

Consequently the honest use of frequentist confidence intervals requires the selection of the procedure independently of data. The user of frequentist confidence intervals based on the two one-sided tests procedure must be aware of the following issue. In the case where he/she observes a null effect size ($D = 0$), if he/she did not plan to assert the smallness of this effect, he/she computes the usual $100(1-\alpha)\%$ confidence interval, centered here on zero. However, if he/she planned to assert smallness, this same computed interval is considered as a $100(1 - \frac{1}{2}\alpha)\%$ confidence interval. More generally, if he/she computes a $100(1 - \alpha)\%$ confidence interval for δ centered on a prespecified value δ_1, from an observed value D close to δ_1, this latter interval can be included within the usual $100(1 - \alpha)\%$ confidence interval. Illustrations of this fact have been carried out for the Correction and Depression scales in Example 1.

On the contrary, the Bayesian methodology offers more flexibility in allowing the most appropriate type of intervals to be selected *in view of the data*. In particular, this makes it ideally suited to extend exploratory data analysis. Consequently it is not surprising that the standard Bayesian smallness credibility interval can be larger than the corresponding frequentist confidence interval.

As a matter of fact, the issue of asserting the smallness of an effect brings a theoretical clear-cut distinction between the frequentist and Bayesian approaches. In practice however this distinction must be played down, since in most applications the results of the procedures are convergent. In fact, from a pragmatic viewpoint, the discrepancy between the frequentist and standard Bayesian procedures for asserting smallness looks like the one-sided *vs* two-sided levels controversy. It must be expected that it will not lead to the same endless academic discussions, but rather that the opportunity for users to have readily applicable general procedures in the two frameworks available will stimulate a positive debate.

Concerning more precisely the role of usual null hypothesis significance testing (NHST) in experimental research, a fundamental property is that all the procedures for asserting the magnitude of an effect in ANOVA can be carried out from the value of the F ratio. Ironically, reporting F ratios (or equivalently p-values) with sufficient accuracy then appears to be valuable for subsequent analysis about the magnitude of effects. Be that as it may, even the ones who think that NHST should be banned in publications (which would be without

doubt a shock therapy, see Shrout, 1997), should acknowledge that F ratios and p-values remain useful for computations. However one has to keep in mind that the shift from NHST to statements about the magnitude of effects implies formal reasoning, which is certainly far from intuitive. *Conditionally on the observed effect*, the F ratio is an estimate of the experimental accuracy. Hence, the higher the test statistic, *i.e.* the more significant the result, the closer the true effect and the observed effect must be. This is clearly demonstrated by the fact that the $100(1-\alpha)\%$ usual confidence interval for the deviation $\delta - D$ is $[\,-(\frac{|D|}{\sqrt{F}}t_{\nu;1-\frac{1}{2}\alpha}\ ,\ (\frac{|D|}{\sqrt{F}}t_{\nu;1-\frac{1}{2}\alpha}]$. Here the words "conditionally on the observed effect" are essential, since they remind us that the F ratio *alone* only allows us to test the point null hypothesis of no effect, but does not authorize any conclusion about the magnitude of the effect. Consequently NHST by itself is of course inadequate for the purposes of asserting the magnitude of effects. Nevertheless, when jointly considered with the observed effect size, it allows us to construct interval estimates. It follows that a "very significant" result generally allows the descriptive result to be extended. Thus, depending on the observed effect, this can lead to assert largeness, mediumness or smallness. On the contrary, a "very nonsignificant" result (F close to zero) will induce a smallness conclusion only if the observed effect is very small. In practice, more often than not, it will correspond to a statement of ignorance. Of course, the special case of a significance level equal to one only means that the observed effect is null and provides no information about the experimental accuracy.

References

[1] Anderson, S. & Hauck, W.W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics, Theory and Methods*, *12*, 2663–2692.

[2] Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 423–437.

[3] Bartko, J.J. (1991). Proving the null hypothesis. *American Psychologist*, *46*, 1089.

[4] Berger, G.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New-York: Springer Verlag.

[5] Berger, R.L. & Hsu, J.C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets (with comments). *Statistical Science*, *11*, 283–319.

[6] Binder, A. (1963). Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, *70*, 107–115.

[7] Bofinger, E. (1992). Expanded confidence intervals, one-sided tests and equivalence testing. *Journal of Biopharmaceutical Statistics*, *2*, 181–188.

[8] Bondy, W.A. (1969). A test of an experimental hypothesis of negligible difference between means. *The American Statistician*, *23*, 28–30.

[9] Box, G.E.P. & Tiao, J.W. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wesley.

[10] Brandstätter, E. (1999). Confidence intervals as an alternative to significance testing. *Methods in Psychological Research Online*, *Vol. 4 No. 2.* http://www. mpr-online.de.

[11] Brown, L.D., Hwang, J.T.G. & Munk, A. (1997). An unbiased test for the bioequivalence problem. *The Annals of Statistics*, *25*, 2345–2367.

[12] Cannon, D.S., Bell, W.E., Fowler, D.R., Penk, W.E., Finkelstein, A.S. (1990). MMPI differences between alcoholics and drug abusers: Effects of age and race. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *2*, 51–55.

[13] Chow, S.L. (1996). *Statistical significance: rationale, validity and utility.* London: Sage.

[14] Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

[15] Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*, 2^{nd} edition. New-York: Academic Press.

[16] Cohen, J. (1994). The earth is round ($p \leq .05$). *American Psychologist*, *49*, 997–1003. With replies in *American Psychologist*, 1995, *50*, 1098–1103.

[17] Deheuvels, P. (1984). How to analyze bio-equivalence studies? The right use of confidence intervals. *Journal of Organizational Behaviour and Statistics*, *1*, 1–15.

[18] Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin*, *63*, 400–402.

[19] Edwards, W., Lindman, H. & Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

[20] Efron, B. (1998). R.A. Fisher in the 21st century (with discussion). *Statistical Science*, *13*, 95–122.

[21] FDA (1992). *Draft recommendation on statistical procedures for bioequivalence studies using the standard two-treatment crossover design.* Center for Drug Evaluation and Research, Food and Drug Administration, Rockville, MD.

[22] Fisher, R.A. (1962). Some examples of Bayes's method of the experimental determination of probabilities a priori. *Journal of the Royal Statistical Society B*, *24*, 118–124.

[23] Fisher, R.A. (1990). *Statistical methods, experimental design, and scientific inference* (Re-issue). Oxford: Oxford University Press.

[24] Fowler, R.L. (1984). Approximating probability levels for testing null hypotheses with noncentral F distributions. *Educational and Psychological Measurement*, *44*, 275–281.

[25] Fowler, R.L. (1985). Testing for substantive significance in applied research by specifying nonzero effect null hypotheses. *Journal of Applied Statistics*, *70*, 215–218.

[26] Frick, R.W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*, 379–390.

[27] Goodman, S.N. & Berlin, J.A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, *121*, 200–206.

[28] Harcum, E.R. (1990). Methodological versus empirical literature: Two views on casual acceptance of the null hypothesis. *American Psychologist*, *45*, 404–405.

[29] Hodges, J.L. & Lehmann, E.L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society, Series B*, *16*, 261–268.

[30] Holender, D. & Bertelson, P. (1975). Selective preparation and time uncertainty. *Acta Psychologica*, *39*, 193–203.

[31] Hsu, J.C., Hwang, J.T.G., Liu H.-K. & Ruberg, S.J. (1994). Confidence intervals associated with tests for bioequivalence. *Biometrika*, *81*, 103–114.

[32] Jeffreys, H. (1961). *Theory of probability* (3^{rd} edition; 1^{st} edition: 1939). Oxford: Clarendon.

[33] Kirk, R.E. (1982). *Experimental design* (2^{nd} edition). Belmont: Brook-Cole.

[34] Lecoutre, B. (1985). How to derive Bayes-fiducial conclusions from usual significance tests. *Cahiers de Psychologie Cognitive*, *5*, 553–563.

[35] Lecoutre, B. (1996). *Traitement statistique des données expérimentales: Des pratiques traditionnelles aux pratiques bayésiennes* (with Bayesian Windows programs by B. Lecoutre and J. Poitevineau, freely available at the Internet address http://www.univ-rouen.fr/LMRS /Persopage/Lecoutre/pac.htm). Montreuil (France): CISIA.

[36] Lecoutre, B. (1999). Two useful distributions for Bayesian predictive procedures under normal models. *Journal of Statistical Planning and Inference*, *77*, 93–105.

[37] Lecoutre, B. & Charron, C. (2000). Bayesian procedures for prediction analysis of implication hypotheses in 2×2 contingency tables. *Journal of Educational and Behavioral Statistics*, *25*, 185–201.

[38] Lecoutre, B., Derzko, G. & Grouin, J.-M. (1995). Bayesian predictive approach for inference about proportions. *Statistics in Medicine*, *14*, 1057–1063.

[39] Lecoutre, B., Guigues, J.-L. & Poitevineau, J. (1992). Distribution of quadratic forms of multivariate Student variables. *Applied Statistics*, *41*, 617–627.

[40] Lecoutre, B., Lecoutre, M.-P. & Poitevineau, J. (2001). Uses, abuses and misuses of significance tests in the scientific community: won't the Bayesian choice be unavoidable? *International Statistical Review*, *69*, to appear.

[41] Lecoutre, B., Mabika, B. & Derzko, G. (2001). Assessment and monitoring in clinical trials when survival curves have distinct shapes in two groups: a Bayesian approach with Weibull modeling illustrated in higher risk patients of EMIAT. *Statistics in Medicine*, to appear.

[42] Lecoutre, B. & Poitevineau, J. (1992). PAC (*Programme d'Analyse des Comparaisons*): *Manuel de référence*. Montreuil, France: CISIA-CERESTA (a limited version of this Windows program for analysis of variance is freely available at Internet address http://www.univ-rouen.fr/LMRS/Persopage/Lecoutre/pac.htm).

[43] Lecoutre, M.-P. (1992). Cognitive models and problem spaces in "purely random" situations. *Educational Studies in Mathematics*, *23*, 557–568.

[44] Lecoutre, M.-P. (2000). And... what about the researcher's point of view? *In* H. Rouanet, J.-M. Bernard, M.-C. Bert, B. Lecoutre, M.-P. Lecoutre, & B. Le Roux, *New ways in statistical methodology: From significance tests to Bayesian inference*, 2^{nd} edition (pp. 65–95). Bern, Switzerland: Peter Lang.

[45] Lehmann, E.L. (1959). *Testing statistical hypothesis* (2^{nd} edition). Wiley: New York.

[46] Mandallaz, D. & Mau, J. (1981). Comparison of different methods for decision-making in bioequivalence assessment. *Biometrics*, *37*, 213–222.

[47] Munk, A. (2000). An unbiased test for the average equivalence problem: The small sample case. *Journal of Statistical Planning And Inference*, *87*, 69–86.

[48] Murphy, K.R. & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology*, *84*, 234–248

[49] Novick, M.R. & Jackson, P.H. (1974). *Statistical methods for educational and psychological research*. NewYork: McGraw-Hill.

[50] Patel, H.I. & Gupta, G.D. (1984). A problem of equivalence in clinical trials. *Biometrical Journal*, *33*, 1225–1230.

[51] Perlman, M.D. & Wu, L. (1999). The emperor's new tests. *Statistical Science*, *14*, 355-369.

[52] Phillips, L.D. (1973). *Bayesian statistics for social scientists.* London: Nelson.

[53] Richardson, J.T.E. (1996). Measures of effect size. *Behavior Research Methods, Instruments, and Computers*, *28*, 12–22.

[54] Rindskopf, D. (1997). Testing "small", not null, hypotheses: Classical and Bayesian approaches. *In* L.L. Harlow, S.A. Mulaik et J.H. Steiger (Eds.), *What if there were no significance tests* (pp. 319–332). Hillsdale, NJ: Erlbaum.

[55] Rindskopf, D. (1998). Null-hypothesis tests are not completely stupid, but Bayesian statistics are better. *Behavioral and Brain Sciences*, *21*, 215–216.

[56] Rocke, D.M. (1984). On testing for bioequivalence. *Biometrics*, *40*, 220–225.

[57] Rogers, J.L., Howard, K.I. & Vessey, J. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*, 553–565.

[58] Rouanet, H. (1996). Bayesian procedures for assessing importance of effects. *Psychological Bulletin*, *119*, 149–158.

[59] Rouanet, H., Bernard, J.-M., Bert, M.-C., Lecoutre, B., Lecoutre, M.-P. & Le Roux, B. (2000). *New ways in statistical methodology: From significance tests to Bayesian inference* (2^{nd} edition). Bern, CH: Peter Lang.

[60] Rouanet, H. & Lecoutre, B. (1983). Specific inference in ANOVA: From significance tests to Bayesian procedures. *British Journal of Mathematical and Statistical Psychology*, *36*, 252–268.

[61] Rouanet, H., Lépine, D. & Pelnard-Considère, J. (1976). Bayes-fiducial procedures as practical substitutes for misplaced significance testing: An application to educational data. *In* D.N.M. De Gruijter & L.J.T. Van Der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 33–50). New York: Wiley.

[62] Rouanet, H., Le Roux, B., Bernard, J.-M. & Lecoutre, B. (2000). The analysis of structured multidimensional data: From Euclidean clouds to Bayesian inference. *In* H. Rouanet, J.-M. Bernard, M.-C. Bert, B. Lecoutre, M.-P. Lecoutre & B. Le Roux, *New ways in statistical methodology: From significance tests to Bayesian inference*, 2^{nd} edition, (pp. 227–254). Bern, CH: Peter Lang.

[63] Rozeboom, W.W (1960). The fallacy of the hypothetico-deductive significance test. *Psychological Bulletin*, *57*, 416–428.

[64] Schall, R. (1995). Assessment of individual and population bioequivalence using the probability that bioavailabilities are similar. *Biometrics*, *51*, 615–626.

[65] Schall, R. & Luus, H.G. (1993). On population and individual equivalence. *Statistics in Medicine*, *12*, 1109–1124.

[66] Schervish, M.J. (1995). *Theory of statistics.* New York: Springer Verlag.

[67] Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*, 657–680.

[68] Selwyn, W.J., Dempster, A.P. & Hall, N.R. (1981). A Bayesian approach to bioequivalence for the 2×2 changeover design. *Biometrics*, *37*, 11–21.

[69] Selwyn, W.J. & Hall, N.R. (1984). On Bayesian methods for bioequivalence. *Biometrics*, *40*, 1103–1108.

[70] Selwyn, W.J., Hall N.R. & Dempster, A.P. (1985). Letter to the Editor. *Biometrics*, *41*, 561.

[71] Serlin, R.C. & Lapsley, D.K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*, 77–83.

[72] Serlin, R.C. & Lapsley, D.K. (1993). Rational appraisal of psychological research and the good-enough principle. *In* G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences. Vol 1: Methodological issues* (pp. 199–228). Hillsdale, NJ: Erlbaum.

[73] Shrout, P.E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science*, *8*, 1–2.

[74] Spiegelhalter, D.J., Freedman, L.S. & Parmar M.K.B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society A*, *157*, 357–416.

[75] Wang, Y.H. (2000). Fiducial Intervals: What are they? *The American Statistician*, *54*, 105–111.

[76] Wellek, S. & Michaelis, J. (1991). Elements of significance testing with equivalence problems. *Methods of Information in Medicine*, *30*, 194–198.

[77] Westlake, W.J. (1981). Response to bioequivalence testing: A need to rethink (reader reaction response). *Biometrics*, *37*, 591–593.

[78] Wilkinson, L. and Task Force on Statistical Inference, APA Board of Scientific Affairs (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist*, *54*, 594–604.

[79] Wilson, W.R., Miller, H.L. & Lower, J.S. (1967). Much ado about the null hypothesis. *Psychological Bulletin*, *67*, 188–196.

[80] Winkler, R.L. (1974). Statistical analysis: Theory versus practice. *In* C.-A.S. Staël Von Holstein (Ed.), *The concept of probability in psychological experiments* (pp. 127–140). Dordrecht, Holland: D. Reidel Publishing Company.

Appendix A: An Unacceptable Procedure

It is often argued that the two-one sided tests procedure (and this holds for its generalization to several *df* effects as well) is biased (conservative), because the probability of rejecting $H_0 : |\mu_1 - \mu_2| \geq \Delta$ is less than α for every $\mu_1 - \mu_2$ in H_0. Similarly the long-run coverage probability of the $100(1-\alpha)\%$ associated (non improved) smallness confidence interval is larger than $1-\alpha$. Nevertheless, from a formal viewpoint (following Lehmann, 1959), the procedure is a size-α test, and the main criticism of statisticians regarding the procedure is rather its lack of power. Faced with this non optimality of the two one-sided tests procedure, a challenge for statisticians has been to derive more efficient solutions. Unfortunately, in spite of the apparent simplicity of the problem, it is a formidable task.

So the most powerful invariant test described in the introduction has been recurrently proposed in various contexts (e.g., Hodges & Lehman, 1954; Bondy, 1969; Anderson & Hauck, 1983; Patel & Gupta, 1984; Rocke, 1984; Fowler, 1984, 1985; Wellek & Michaelis, 1991; Serlin & Lapsley, 1985; 1993). Unfortunately this test has undesirable properties, previously discussed in particular by Selwyn, Hall and Dempster (1985) and Schuirman (1987).

(i) When the observed difference $D = M_1 - M_2$ is null, the observed significance level is null, leading to the automatic rejection of H_0, whatever Δ, the sample size and σ.

(ii) The critical rejection region varies in a nonmonotonic way as a function of the sampling error variance. Eventually the rejection region includes values D that lie *outside* the equivalence region.

(iii) Moreover this procedure can be generalized for testing a null hypothesis for $\delta = \mu_1 - \mu_2$ such as H_0 : "$\delta \leq \Delta'$ or $\delta \geq \Delta''$" ($\Delta'' > \Delta'$). Unfortunately contradictory conclusions can be obtained. For instance assuming a known sampling error variance equal to 1, when $-0.020 < D \leq +0.071$, the hypothesis "$\delta \leq -0.50$ *or* $\delta \geq +0.50$" is rejected at level $\alpha = 0.05$, but the hypothesis "$\delta < -0.70$ or $\delta > +0.51$" (*i.e.* "$\delta + 0.095 < -0.605$ *or* $\delta + 0.095 > +0.605$"), even though implied by the former, is not rejected at this same level (Schervish, 1995, page 252) [10]. This incoherence makes the procedure unacceptable [11].

To explain these undesirable properties, it can be noted that the observed level of this test is the difference between the two p-values associated with the usual one-sided tests of the point null hypotheses $\delta = -\Delta$ and $\delta = \Delta$, instead of their maximum for the two one-sided tests procedure and of their sum for the standard Bayesian solution. These undesirable properties are even more pronounced in the case of multiple *df* effects. The two one-sided tests procedure, as well as our generalization to multiple *df* effects, have been developed

[10] The respective rejection regions are $-0.071 \leq D \leq +0.071$ and $-0.170 \leq D \leq -0.020$.

[11] Note that a similar procedure has been proposed for testing the reverse null hypothesis that $H_0 : |\delta| < \Delta$ (and more generally $\lambda < \Lambda$), *i.e.* for asserting the "nondirectional largeness" of an effect (Fowler, 1985; Murphy & Myors, 1999). Unfortunately this test exhibits the same kind of incoherence (Schervish, 1995, p. 257).

precisely in order to avoid undesirable properties.

In conclusion, as Perlman and Wu (1999, page 362) states, the two one-sided tests procedure "already makes perfectly appropriate inferences: if s [S in our notations] and/or $|x|$ [$|D|$] is too large, it correctly declares that the evidence is insufficient to reject H_0 in favor of H_1. If this is deemed unsatisfactory, then the solution is very simple: more observations are needed, not Better New Tests."

Appendix B: Generalization of the Two One-sided Tests Procedure

Cohen's f and Partial Contrasts

Let us consider the overall comparison of G independent group means in a balanced design with equal cell counts $\overline{n}$. The square of Cohen's f numerator is given by the formula

$$\sigma_{\mathrm{m}}^2 = \frac{1}{\mathrm{G}} \sum_{g=1}^{\mathrm{G}} (\mu_g - \mu_.)^2 = \sum_{g=1}^{\mathrm{G}} \frac{\mu_g - \mu_.}{\mathrm{G}} \mu_g \tag{13}$$

Consequently σ_{m} is the contrast between population means $\delta^* = \sum_{g=1}^{\mathrm{G}} \kappa_g^* \mu_g$ with coefficients $\kappa_g^* = \frac{\mu_g - \mu_.}{\mathrm{G}\sigma_{\mathrm{m}}}$. Let us define the corresponding contrast between sample means with coefficients $c_g^* = \frac{M_g - M_.}{\sqrt{(\mathrm{G}/\overline{n}) SS_G}}$, where $SS_G = \overline{n} \sum_{g=1}^{\mathrm{G}} (M_g - M_.)^2 = \sum_{g=1}^{\mathrm{G}} \overline{n}(M_g - M.)M_g$ is the overall sum of squares. It follows that $D^* = \sum_{g=1}^{\mathrm{G}} c_g^* M_g$ is equal to $\sqrt{SS_G/(\mathrm{G}\overline{n})}$ and has sampling error variance $b^{*2}\sigma^2$ where $b^{*2} = \sum_{g=1}^{\mathrm{G}} \frac{c_g^{*2}}{\overline{n}} = \frac{1}{\mathrm{G}\overline{n}}$.

Each possible contrast between means with coefficients c_g has an associated partial sum of squares $\overline{n}(\sum_{g=1}^{\mathrm{G}} c_g M_g)^2 / (\sum_{g=1}^{\mathrm{G}} c_g^2)$ inferior or equal to SS_G. Hence, if $\sum_{g=1}^{\mathrm{G}} c_g^2 = \frac{1}{\mathrm{G}}$, it can be deduced that $D = \sum_{g=1}^{\mathrm{G}} c_g M_g \leq D^*$ just as $\delta = \sum_{g=1}^{\mathrm{G}} c_g \mu_g \leq \delta^* = \sigma_{\mathrm{m}}$. In particular the true pairwise differences between groups, which are contrasts such as $\sum_{g=1}^{\mathrm{G}} c_g^2 = 2$, are inferior or equal to $\sigma_{\mathrm{m}}\sqrt{2\mathrm{G}}$ in absolute value.

Smallness Test for m df Effects

Here we give an intuitive justification for the generalization of the two one-sided tests procedure ($TOST$) to m df effects. Adopting the general definition of λ as any index proportional to σ_{effect}, L^2 is proportional to the mean square MS_{effect}. In the same way, S^2 the usual estimate of the standardization variance is proportional to MS_{error}. Then, generalizing the notations introduced for a difference between means, the usual F test of the null hypothesis $\lambda = 0$ can be written

$$F = \frac{MS_{\mathrm{effect}}}{MS_{\mathrm{error}}} = \left(\frac{L}{bS}\right)^2 \tag{14}$$

A fundamental property is that there is a unique contrast $D^* = \sum_{g=1}^{G} c_g^* M_g$ within the effect source such as $D^* = L$ and its associated partial sum of squares is equal to $SS_{\text{effect}} = mMS_{\text{effect}}$. This generalizes the case of the overall comparison of G independent group means above. The usual F test of the null hypothesis $\delta^* = 0$ for this contrast can be written

$$F^* = mF = \frac{mMS_{\text{effect}}}{MS_{\text{error}}} = \left(\frac{D^*}{b^* S}\right)^2 \text{ where } b^* = \frac{b}{\sqrt{m}} \tag{15}$$

Each possible contrast between observed means with coefficients c_g such as $\sum_{g=1}^{G} \frac{{c_g}^2}{n} = {b^*}^2$ (*i.e.* with sampling error variance ${b^*}^2\sigma^2$) has an associated sum of squares inferior or equal to SS_{effect}, and consequently is inferior or equal to L in absolute value.

For similar reasons the overall test of the null hypothesis $\lambda \geq \Lambda$ is equivalent to a simultaneous test of the hypothesis that all possible contrasts between population means within the effect source, associated with the constant b^*, are superior or equal to Λ in absolute value. Then the null hypothesis $H_0 : \lambda \geq \Lambda$ is rejected at level α if, for each of these possible contrasts, the hypothesis $|\delta| \geq \Lambda$ is rejected by the *TOST* at level α. Thus the procedure simply amounts to applying the *TOST* to the contrast with coefficients c_g^* defined above (such as $D^* = L$). The same formulas as for a difference between two means apply to this contrast. Therefore H_0 is rejected at level α if $L < \Lambda$ and $t_{min}^{[\Lambda]} \geq t_{q;1-\alpha}$ where

$$t_{min}^{[\Lambda]} = \frac{\left|\Lambda - |D^*|\right|}{b^* S} = \frac{\left|\Lambda - L\right|}{bS\sqrt{m}} \tag{16}$$

A $100(1-\alpha)\%$ *SCI* for λ can be defined as the set of Λ so that this smallness test is nonsignificant at level α. As for the case of a difference between means, the procedure can be improved by distinguishing between the case where the t test of the null hypothesis $\delta^* = 0$ is significant at one-sided level α (*i.e.* when $mF \geq t_{q;1-\alpha}^2$ or equivalently $\lambda_{SCI} \leq 2L$) and the case where it is nonsignificant. However, unlike the smallness test, this improved confidence interval procedure cannot be expressed as an inference about λ and needs some sophistication. This is not surprising since λ does not allow us to take into account the direction of the effect.

All these procedures can be extended to standardized effects.

Appendix C: Bayesian Solutions for Asserting Smallness

Standard Bayesian procedures express uncertainty about the true effect size by a probability statement, only taking into account the information provided by the analyzed data ("what the data have to say") [12]. Just like frequentist procedures, they can be derived from observed ES indices and t or F ratios (Lecoutre, 1985, 1996).

[12] We call also these procedures *fiducial Bayesian* (see Lecoutre, Lecoutre & Poitevineau, 2001). This explicitly refers to Fisher's fiducial method (Fisher, 1990) and makes up for the inequity of some criticisms

Standard Bayesian Procedures for the Difference Between two Means

It will be sufficient here to consider yet again the basic situation of the inference about the difference δ of two normal means. Assuming the usual noninformative prior, the posterior distribution of δ is a scaled (or generalized) t distribution, with ν *df* (see e.g., Box & Tiao, 1973). It is centered on the observed difference D (instead of 0 for the usual t) and has scale factor $\frac{D}{t}$ (instead of 1). We write this as follows

$$\delta \mid data \sim t_{\nu}\left(D\ ,\ \left(\frac{D}{t}\right)^2\right) \text{ (assuming } D \neq 0) \tag{17}$$

and we read "given data, δ is distributed as a t, with center D and scale $\frac{D}{t}$". The normal distribution with center D and standard deviation $\frac{D}{t}$ can be used as approximation when ν is large.

This standard posterior distribution allows us to interpret the $100(1-\alpha)\%$ usual confidence interval in probabilistic terms: from a Bayesian viewpoint, given data "δ is contained within the confidence set with probability $1-\alpha$".

The Theoretical Irreconcilability of the Frequentist and Bayesian Solutions

The Bayesian solution for asserting the smallness of a difference, and more generally of a one *df* ANOVA raw fixed effect, automatically follows. Given the limit Δ, the posterior probability that $|\delta|$ is less than Δ can be computed. Alternatively, given a probability γ, a "Bayesian Smallness Credibility Interval" (*BSCI*) for δ can be derived by solving $Pr(|\delta| < \Delta) = \gamma$. The involved probability is

$$Pr(|\delta| < \Delta \mid data) = Pr\left(t_{\nu}\left(d\ ,\ \left(\frac{d}{t}\right)^2\right) < \Delta\right) - Pr\left(t_{\nu}\left(d\ ,\ \left(\frac{d}{t}\right)^2\right) < -\Delta\right) \tag{18}$$

It can be verified that the standard Bayesian probability $Pr(|\delta| > \Delta \mid data)$ is the *sum* of the observed p-values of the two-one sided tests respectively associated with the null hypotheses $\delta = -\Delta$ and $\delta = \Delta$. Since the two one-sided tests procedure only involves the greater of these two p-values, the two solutions are basically distinct. Consequently, unlike the usual confidence interval, the Bayesian probability that δ is contained within the $100(1-\alpha)\%$ smallness confidence interval $[-\delta_{SCI}\,, \delta_{SCI}]$ is no longer $1-\alpha$. In fact $1-\alpha$ is the Bayesian probability that δ is smaller than δ_{SCI} (if $d \geq 0$) or greater than $-\delta_{SCI}$ (if $d \leq 0$). On the one hand, from the Bayesian viewpoint, the interval $[-\delta_{SCI}\,, \delta_{SCI}]$ has too weak a

(e.g., Cohen, 1994 and Frick, 1996). It must be acknowledged that Fisher was constantly concerned with considering a method that only expressed evidence from the data in terms of probability about parameters. Also his work on the Bayesian method in his last years (Fisher 1962) should not be ignored. The fiducial distribution is admittedly considered by modern statisticians to be a blunder, but it could be speculated with Efron (1998) that "maybe Fisher's biggest blunder will become a big hit in the 21^{st} century" (see also Wang, 2000).

posterior probability (less than $1-\alpha$). On the other hand, from the frequentist viewpoint, the Bayesian solution must be discarded since the Bayesian bound δ_{BSCI} is greater than δ_{SCI} [13]. Consequently, if $\delta_{SCI} < \Delta < \delta_{BSCI}$, smallness cannot be asserted with the Bayesian procedure, as done with the frequentist procedure.

It appears that the issue of asserting the smallness of an effect brings a theoretical clear-cut distinction between frequentist and Bayesian approaches. In practice, it must be acknowledged that this distinction is not that dramatic, since the Bayesian probability that δ is contained within the $100(1-\alpha)\%$ smallness *SCI* is between $1-2\alpha$ and $1-\alpha$. So in most applications the results of the two procedures are convergent.

Standard Bayesian Procedures for ANOVA Fixed Effects

Asserting the smallness of any ν_1 *df* ANOVA fixed effect involves no other conceptual difficulties but only additional technical work. The involved Bayesian probability is given by $Pr(\lambda < \Lambda)$, or equivalently by $Pr(\lambda^2 < \Lambda^2)$. The standard Bayesian posterior distribution of λ^2 is a probability distribution that we call *psi-square*. The relevant probability is computed as

$$Pr(\lambda < \Lambda \mid data) = Pr\left(\psi^2_{\nu_1,\nu_2}(\nu_1 F) < \left(\frac{\Lambda}{L}\right)^2 F\right) \tag{19}$$

In the same way, for standardized effects, the relevant probability is given by a *lambda-square* distribution

$$Pr(\frac{\lambda}{\sigma} < \Lambda^* \mid data) = Pr\left(\Lambda^2_{\nu_1,\nu_2}(\nu_1 F) < \left(\frac{\Lambda^*}{\frac{L}{S}}\right)^2 F\right) \tag{20}$$

Algorithms and computer programs which implement the *psi-square* and *lambda-square* distributions [14] are available (Lecoutre, Guigues & Poitevineau, 1992; Lecoutre, 1996, 1999). The noncentral *chi-square* distribution with ν_1 *df* and noncentrality parameter $\nu_1 F$ can be used as approximation when ν_2 is large.

[13] And not the reverse, as suggested by Rouanet (1996) in his note 9.

[14] In Schervisch (1995) these two distributions are considered under the names of alternate F and alternate chi-square distributions.