Modifying the t and ANOVA F Tests When Treatment is Expected to Increase Variability Relative to Controls

Cavell Brownie, Dennis D. Boos,

and Jacqueline Hughes-Oliver

Department of Statistics, North Carolina State University

Raleigh, North Carolina 27695-8203, U.S.A.

SUMMARY

When similar experimental units are assigned randomly to two groups, one to receive a "treatment," the other to serve as a control, the homogeneity of variance assumption underlying the pooled t test is valid under the null hypothesis of no treatment effect. Thus power, and not validity, of the pooled t should be the concern in such experiments, especially if treatment may affect variability as well as mean response. We focus on studies where there is a biological basis (e.g., toxicological studies) for expecting a treatment effect on mean response to be accompanied by an *increase* in variance. To detect a change in mean response in such experimental situations, we propose a simple modification of the pooled t which is also valid under H_O: "no treatment effect," but which can have substantially better power than the pooled t. This modification can also be applied to other procedures for testing equality of means when there are 2 or more "treatment" groups in addition to the control.

1. Introduction

Most discussions on the t and analysis of variance (ANOVA) F tests for comparing means note that variance heterogeneity can seriously affect the properties of these procedures. Too often, these discussions focus on the effect of variance heterogeneity on test *validity*, and ignore the effect on *power* to detect differences between means. Test validity has been so strongly emphasized that when variance heterogeneity is observed in data, there may be no attempt to consider issues related to power when deciding on a method of analysis. One of our objectives here is to distinguish experimental situations where this overriding concern with test validity is misplaced and can result in loss of power. A second objective is to propose a modification of the t or F test to achieve better power in situations where increased variability is expected to be a consequence of treatment effect.

Two types of experimental situations have quite different implications for the homogeneity of variance assumption underlying the t and F tests, and consequently for the type I and type II error rate properties of these procedures. In the first situation (Situation A), homogeneous experimental units from a single pool are assigned at random to receive one of k "treatments." The objective is to compare mean responses for the treatments, one of which is often a control. The null hypothesis of no treatment effect corresponds to that of identical distributions for observations in each group. Assuming normality, this is

H_o: equal means and equal variances or H_o: $\mu_1=\cdots=\mu_k$ and $\sigma_1^2=\cdots=\sigma_k^2$.

Given random assignment of like units to different treatments, the homogeneity of variance assumption is therefore completely consistent with the null hypothesis of no treatment effects, and the pooled t or one-way ANOVA F are always valid (given normality) in this situation. Apparent variance heterogeneity in the resulting data does not mean that validity is a

problem, but suggests that there are treatment differences, i.e., that H_o is false. Examples of Situation A include all toxicity studies where rats or mice of the same strain, sex, and weight, are allocated randomly to a control group and to groups that will receive different concentrations of a toxin.

The second situation (Situation B) involves studies aimed at comparing responses for inherently different types of items. Examples include trials to compare yields for different plant varieties, and "observational" studies to compare attributes of individuals selected at random from k populations defined, for example, by ecotype or geographic location, or by sociological factors such as educational status, occupation type, sex, race etc. For such studies homogeneity of variances cannot usually be assumed under the null hypothesis of equal population means. The appropriate hypotheses are

$$H_o: \mu_1 = \cdots = \mu_k \quad \text{vs} \quad H_a: \mu_i \neq \mu_j \text{ for at least one i,j pair}$$
$$\text{with the } \sigma_i^2 \text{ unknown and possibly unequal.}$$

Concern about validity of the t and ANOVA F is certainly justified for this Behrens-Fisher type problem that arises in Situation B studies. It is well known that if sample sizes n_i are unequal, both procedures are liberal if the larger n_i are paired with the smaller σ_i^2, and they are conservative when the larger n_i and larger σ_i^2 are paired. For approximate validity, a procedure such as the Welch t and k-sample generalizations (e.g. Brown and Forsythe, 1974) should be used.

Consideration of test validity thus rules out use of the t and ANOVA F in Situation B but not in Situation A experiments. Situation B is not discussed further here (the interested reader is referred to Boos and Brownie, 1988). Instead we focus on Situation A and the choice

of procedure when the biology suggests alternatives H_a which involve unequal variances as well as unequal means. Specifically, we consider Situation A studies where increased variance is expected for the treatment group under H_a, for example, because of differences among individuals in their capacity to tolerate or respond to a given dose of a compound. Gad and Weil (1986, p 50-51) discuss the importance of this "variance inflation" phenomenon in toxicological studies, noting that it results in reduced sensitivity (i.e. power) of the t and ANOVA tests to detect differences in mean response. Similar considerations led Good (1979), Johnson, Verrill and Moore (1987), and Conover and Salsburg (1988) to propose tests for comparing treatment and control groups when not all treated units are expected to "respond" to treatment. Though useful, these procedures are not as easily applied as a t test and do not focus on differences in means, in contrast to the methods we suggest here.

For data where variability within the treatment groups is expected to be larger than for the controls, we propose a simple modification of the t or ANOVA F tests which involves replacing the pooled error variance with a variance computed from the controls only. This procedure is valid for Situation A nulls (given normality) and has power advantages over the pooled t and ANOVA F under alternatives where treatment affects mean response and increases variability. It is especially useful as a quick way to detect changes in mean response relative to a measure of "normal" or control variability. Computational details are given in the next section, followed by power comparisons with the pooled t and Welch t. Other applications where a similar modification can be used are described in Section 3 and illustrated with an example from a study on the toxicity of benzene (Section 4).

2. The Modified t Test

In the simplest setting $N=n_1+n_2$ experimental units are assigned randomly to two groups of

sizes n_1 and n_2. Units in the first group serve as controls, and units in the second group receive some form of "treatment." In addition, we assume that a treatment effect on mean response is expected to be accompanied by an increase in variability.

Let $X_{i1},\ldots, X_{in_i}$ denote responses in group i, i=1, 2. The $X_{i1},\ldots, X_{in_i}$ are then iid with mean μ_i and variance $\sigma_i^2<\infty$, and we wish to test equality of the μ_i in the context

$$H_o:\ \mu_1=\mu_2,\ \sigma_1^2=\sigma_2^2 \quad \text{vs} \quad H_a:\ \mu_1\neq\mu_2,\ \sigma_2^2=\rho\sigma_1^2,\ 1\leq\rho<\infty. \tag{1}$$

Probability calculations below require the X_{ij} to be normally distributed, but qualitative conclusions will hold for nonnormal data (e.g., Scheffé, 1959, Ch. 10).

Let $\bar{X}_i=\Sigma_{j=1}^{n_i} X_{ij}/n_i$ and $s_i^2=\Sigma_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2/(n_i-1)$ i=1, 2 denote the sample means and variances, and $s_p^2=\Sigma_{i=1}^2(n_i-1)s_i^2/(N-2)$ the pooled variance. Also let $t(\nu)$ denote the t distribution with ν degrees of freedom (df). To test (1), the pooled t compares $t_p=(\bar{X}_1-\bar{X}_2)/[s_p^2(1/n_1+1/n_2)]^{1/2}$ to the t(N-2) distribution. The Welch t is not designed for H_o in (1), (rather for H_o: $\mu_1=\mu_2$, $\sigma_1^2\neq\sigma_2^2$), but it is frequently applied in situations where the appropriate hypotheses are H_o and H_a in (1). This procedure compares $t_w=(\bar{X}_1-\bar{X}_2)/\ (s_1^2/n_1+s_2^2/n_2)^{1/2}$ to the t(f) distribution with estimated Satterthwaite (1941) df f, where $f^{-1}=\Sigma_{i=1}^2 c_i^2/(n_i-1)$, $c_i=(1-n_i/N)s_i^2\ /\ \Sigma(1-n_i/N)s_i^2$.

The modification we propose is to compare

$$t_m=\frac{(\bar{X}_1-\bar{X}_2)}{[s_1^2(\frac{1}{n_1}+\frac{1}{n_2})]^{1/2}}$$

to the $t(n_1-1)$ distribution, using data from the controls only to estimate variance. Like the

pooled t, this modified t is valid for H_o in (1), but for alternatives with $\rho>1$ the modified t can have better power than the pooled t. Note, however, that use of the modified t in the context (1) rather than a likelihood ratio test or, e.g., Good's (1979) method, assumes that means are of primary interest.

To indicate what sort of power gains can be expected, and to identify situations where gains are substantial, we generated power curves for the modified t and pooled t for several cases. No power calculations were carried out for the Welch t but qualitative comparisons with the Welch t can be made using existing results (e.g., Best and Rayner, 1987). For simplicity, power calculations were restricted to one-tailed tests (aimed at $\mu_1>\mu_2$) and normal distributions.

For given n_1, n_2, $\Delta=(\mu_1-\mu_2)/\sigma_1$, $\rho=\sigma_2^2/\sigma_1^2$ and significance level α, let $\beta_{m,\alpha}(n_1, n_2, \Delta, \rho)$ and $\beta_{p,\alpha}(n_1, n_2, \Delta, \rho)$ represent the power of the modified and pooled t tests, respectively. It is easy to show that

$$\beta_{m,\alpha}(n_1, n_2, \Delta, \rho)=P[T(n_1-1, \delta)> \left(\frac{n_1+n_2}{n_2+\rho n_1}\right)^{1/2} t(\alpha; n_1-1)], \qquad (2)$$

where $\delta= \Delta/(n_1^{-1}+\rho n_2^{-1})^{1/2}$, $t(\alpha; n_1-1)$ is the $1-\alpha$ percentile of the $t(n_1-1)$ distribution and $T(\nu,\delta)$ has a noncentral t distribution with ν df and noncentrality parameter δ. It is harder to evaluate exact power for the pooled t, because if $\rho>1$, the distribution of $(N-2)s_p^2/\sigma_1^2$ is not chi-square. Rather, $(N-2)s_p^2/\sigma_1^2$ is distributed as the weighted sum of independent chi-squared random variables (r.v.'s) $\chi^2(n_1-1)+\rho\chi^2(n_2-1)$, where $\chi^2(\nu)$ represents a chi-squared r.v. with ν df. This distribution can, however, be closely

approximated by that of a $c\chi^2(d)$ r.v. (e.g., Box, 1954, Theorem 3.1) where

$$c=\frac{n_1-1+\rho^2(n_2-1)}{n_1-1+\rho(n_2-1)} \text{ and } d=\frac{[n_1-1+\rho(n_2-1)]^2}{n_1-1+\rho^2(n_2-1)}.$$

With these values for c and d, an easily computed approximation to the power of the pooled t is

$$\beta_{p,\alpha}(n_1, n_2, \Delta, \rho) \doteq P[T(d, \delta) > t(\alpha;N-2)\left(\frac{n_1+n_2}{n_2+\rho n_1}\right)^{1/2}\left(\frac{c\,d}{N-2}\right)^{1/2}]. \quad (3)$$

Monte Carlo simulation of several cases showed this approximation to be sufficiently accurate for the power comparisons described below.

Using (2) and (3), with $\alpha=.05$, power was plotted against Δ as in Figures 1(a) - 1(d), with ρ defined as a continuous funtion of Δ. Based on published values for means and standard deviations in several dose response studies, (e.g., Brownie et al., 1986) the functions used were

$$\rho_1(\Delta)= 1+\Delta \quad , \Delta\geq 0 \ , \quad (4)$$

$$\rho_2(\Delta)=(1+\Delta)^2 \quad , \Delta\geq 0 \ , \quad (5)$$

and

$$\rho_3(\Delta)=1+\frac{1-p}{p}\Delta^2, \ 0\leq p\leq 1 \text{ and } \Delta=p\theta \text{ for fixed } \theta. \quad (6)$$

Equation (4) represents a situation where treatment has an increasing but small effect on variability and (5) represents a much stronger effect per change in means. Equation (6) is based on the mixture alternatives of Good (1979) and Johnson et al. (1987), where $1-p$ represents the probability of "nonresponse" to treatment and θ is the expected magnitude of response. Note that for $\Delta>0$, ρ_3 increases then decreases as suggested by Gad and Weil (1986, p. 51). Finally, four sets of sample sizes were used; $(n_1,n_2)= (20, 5), (20, 10),$

(20, 20) and (10, 10).

The resulting power curves (not all shown here) and additional calculations support some fairly general conclusions. For fixed Δ, n_1, n_2 the power gain with the modified t increases as ρ increases (compare Figures 1(a) and 1(b); 1(c) and 1(d)). Note, for example, in Figure 1(b), the substantial difference in power of the modified and pooled t procedures for $\Delta=1$ and $\rho_2(\Delta)=4$ corresponding to a standard deviation in the treatment group responses only twice that for the controls. The difference in power of the two procedures of course depends also on the sample sizes. Roughly, if $\rho>1$ then for fixed Δ and $n_1 \geq n_2$, the power advantage of the modified t increases as the proportion of observations in the treatment group increases (i.e., as n_2/N increases). This is because "inflation" of s_p^2 increases as n_2/N increases. Figures 1(b) and 1(d) show that the power difference is indeed greater for given Δ and ρ for $n_1=n_2=10$ than for $n_1= 20$, $n_2=5$.

Inspection of Figures 1(a) - 1(d) indicates that substantial power gains can be achieved with the modified t especially if $n_1 \doteq n_2$ and $\rho \geq 2$. To complete the picture, however, it is important to investigate how much loss in power might be incurred by using the modified t when $\rho \doteq 1$ (i.e., when contrary to expectation, treatment has little effect on variability). We therefore examined a "worst case" situation, assuming $\rho=1$ and n_2 large for several values of n_1. For $\rho=1$, $n_2=\infty$, power is given by

$$\beta_{m,\alpha}(n_1, \infty, \Delta, 1)=P[T(n_1-1, \delta)> t(\alpha; n_1-1)]$$

and

$$\beta_{p,\alpha}(n_1, \infty, \Delta, 1)= 1-\Phi(z_\alpha-\delta),$$

where $\delta=n_1^{1/2}(\mu_1-\mu_2)/\sigma_1$ and $\Phi(\cdot)$ is the standard normal cumulative distribution function

with $(1-\alpha)^{th}$ percentile z_α. For $\alpha=.05$, $n_2=\infty$ and $n_1=$ 5, 10, and 20, the maximum difference in power $(\beta_p-\beta_m)$ for varying δ was .14, .06, and .03, respectively. Additional exact power calculations for $n_1=n_2=5$, 10, and 20, showed that the maximum difference in power for the pooled and modified t is .10, .03, and .01, respectively. Based on these calculations, for $n_1\geq 10$ the possible power gains with the modified t seem to justify risking a loss in power of at most .03 when $n_1\doteq n_2$ or at most .06 when $n_2>n_1$.

Comparison of power of the modified t with the Welch t is also important because of the widespread use of this procedure in the context (1). If $n_1=n_2$, then $t_p=t_w$, but df for t_w are at most n_1+n_2-2, and power of the Welch t cannot exceed (but for $\rho\neq 1$ can be less than) power of the pooled t. For $n_1>n_2$ and $\rho>1$ (say $\rho>2$ or $\sigma_2^2>2\sigma_1^2$), the pooled t will tend to have considerably better power than the Welch t (e.g., Boos and Brownie, 1988). For $n_1<n_2$ and $\rho>1$ the Welch t tends to have better power than the pooled t. Thus for $n_1\geq n_2$ and $\rho>1$, as is commonly the case in Situation A toxicity studies, the Welch t will be less powerful than the pooled t and hence less powerful than the modified t.

We conclude our comparison of these three procedures under alternatives H_a in (1) by commenting on the interpretation of the calculated t values (or corresponding p values). The Welch t_w measures distance between the control and treatment means $(\bar{X}_1-\bar{X}_2)$ in terms of an estimate of $(\sigma_1^2/n_1+\sigma_2^2/n_2)^{1/2}=[\mathrm{Var}(\bar{X}_1-\bar{X}_2)]^{1/2}$, which is meaningful especially if a confidence interval for $\mu_1-\mu_2$ is of interest. The value of t_m is a measure of the distance $\bar{X}_1-\bar{X}_2$ relative to an estimate of control variability (i.e., a measure of treatment effect on mean response relative to "normal" variability). In contrast, t_p measures distance between $\bar{X}_1$ and $\bar{X}_2$ in terms of an estimate of $[(n_1\sigma_1^2+n_2\sigma_2^2)/(n_1+n_2)]^{1/2}$ which does not seem to have a useful interpretation. The corresponding p values have a similar interpretation but on a scale

measured in terms of probability units.

3. Other Applications and Discussion

The modification of the pooled t proposed in Section 2 can be applied to analogous k-sample procedures in settings similar to that of Section 2 but with more than one treatment group. Often, the k-1 treatment groups correspond to increasing doses of a suspected toxic agent, or to more than one toxin, administered separately and in combination (e.g., Brownie et al., 1986). The overall null hypothesis of no treatment effect, and an appropriate alternative, are

$$H_o: \ \mu_1=\ldots=\mu_k, \ \sigma_1^2=\ldots=\sigma_k^2$$

and (7)

$$H_a: \ \mu_i \neq \mu_j \text{ for at least one i, j pair, } \sigma_i^2=\rho_i\sigma_1^2, \ 1\leq\rho_i<\infty, \ i=2,\ldots, k.$$

For testing (7), the one-way ANOVA F test is modified by comparing

$$F_m = \frac{\sum_{i=1}^{k} n_i (\bar{X}_{i.}-\bar{X}_{..})^2/(k-1)}{s_1^2}$$

to the $F(k-1, n_1-1)$ distribution. Here s_1^2, the variance among the controls, replaces the Error Mean Square (EMS) $= \Sigma_{i=1}^{k} (n_i-1)s_i^2/(N-k)$ in the denominator of the usual one-way ANOVA statistic, and denominator df are reduced accordingly. The F_m procedure is valid under H_o in (7), and has power advantages relative to the one-way ANOVA F similar to those for the modified t relative to the pooled t.

Assuming it is of interest to detect treatment effects on mean response relative to control variability, this approach is readily applied to other test procedures. We list a few examples where the obvious modification simply substitutes s_1^2 with n_1-1 df for the ANOVA

EMS with N−k df:

(i) Tests of specific contrasts comparing treatment and control means, e.g.,

H_o: $\mu_1 = \frac{1}{k-1} \Sigma_{i=2}^{k} \mu_i$.

(ii) Dunnett's procedure for comparing each treatment with the control.

(iii) Isotonic regression for detecting specific order relations among the μ_i (Barlow et al., 1972).

Other variations of the approach might replace the EMS with a pooled variance based on some, but not all, of the experimental groups. In studies involving more than one type of control (e.g., vehicle controls and pair-fed controls) pooling all the control variances may be reasonable. In dose-response studies it may be tempting to pool variances from the lowest doses with the control variance. Note, however, that any rule for obtaining an estimate of "control" variance (with which to replace the EMS) should have a sound *a priori* justification. In fact, to prevent bias, the decision to use a modified procedure should be made *a priori*, and not because variability appears low for some subset of the k groups. We stress this point because one disadvantage of the modified procedures may be the potential for their misuse when such decisions are made after seeing the data.

We add one other caution. The modified t or F should not be used when the number of control observations is small (say $n_1 < 10$) both because of power calculations and because of a general concern for stability of the variance estimate. Offsetting these drawbacks, the proposed procedures have several advantages in addition to power. They are simple to apply (contrast Good, 1979; Johnson et al., 1987; Conover and Salsburg, 1988) and results have a useful interpretation on the original measurement scale. Variance stabilizing transformations are avoided. Thus if responses are approximately normal for the controls, but treatment

- - - Insert Table 2 - - -

Even though the effect of toxicity on variance is not marked ($\hat{\rho} \doteq 2$ pooling s_t^2 for the 4 highest benzene concentrations), in most cases we see $t_m > t_p > t_w$ with the reverse ordering of the p values. The most interesting row of the Table is for 1 ppm benzene. Erexson et al. reported that the outcome for this concentration depended on the test used (i.e. "significant" or p value $<.05$ with the t test but "nonsignificant" or p value $>.05$ with the Mann-Whitney U test). Using $\bar{s}_c^2$ with 24 df leads to a definitive outcome with the pooled t (test 2), but the modified t gives a still smaller p value (.0009 compared to .0024). The lower power of the Welch t in situations like this is reflected in its larger p value (.0246). As this example illustrates, there is little to be lost, but possibly much to be gained in terms of power, with the modified t (provided a reliable estimate of control variance is available).

REFERENCES

Barlow, R.E., Bartholomew, D.J., Brenner, J.M., and Brunk, H.D. (1972) *Statistical Inference under Order Restrictions.* New York: Wiley.

Best, D.J., and Rayner, J.C.W. (1987). On Welch's approximate solution for the Behrens-Fisher problem. *Technometrics* 29, 205-210.

Boos, D.D. and Brownie, C. (1988). Bootstrap p-values for tests of nonparametric hypotheses. Institute of Statistics Mimeo Series 1919, North Carolina State University.

Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics* 25, 290-302.

Brown, M.B., and Forsythe, A.B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics* 16, 129-132.

Brownie, C.F., Brownie, C., Noden, D., Krook, L., Haluska, M., and Aronson, A.L. (1986). Teratogenic effect of calcium edetate (Ca EDTA) in rats and the protective effect of zinc. *Toxicology and Applied Pharmacology* 82, 426-443.

Conover, W.J. and Salsburg, D.S. (1988). Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to "respond" to treatment. *Biometrics* 44, 189-196.

Erexson, G.L., Wilmer, J.L., Steinhagen, W.J. and Kligerman, A.D. (1986). Induction of cytogenetic damage in rodents after short-term inhalation of benzene. *Environmental Mutagenesis* 8, 29-40.

Gad, S. and Weil, C.S. (1986). *Statistics and Experimental Design for Toxicologists.* Caldwell, New Jersey: Telford Press.

Good, P.I. (1979). Detection of a treatment effect when not all experimental subjects will respond to treatment. *Biometrics* 35, 483-489.

Johnson, R.A., Verrill, S., and Moore, D.H. II. (1987). Two-sample rank tests for detecting changes that occur in a small proportion of the treated population. *Biometrics* 43, 641-655.

Satterthwaite, F.E. (1941). Synthesis of variance. *Psychometrika* 6, 309-316.

Scheffé, H. (1959). *The Analysis of Variance.* New York: Wiley.

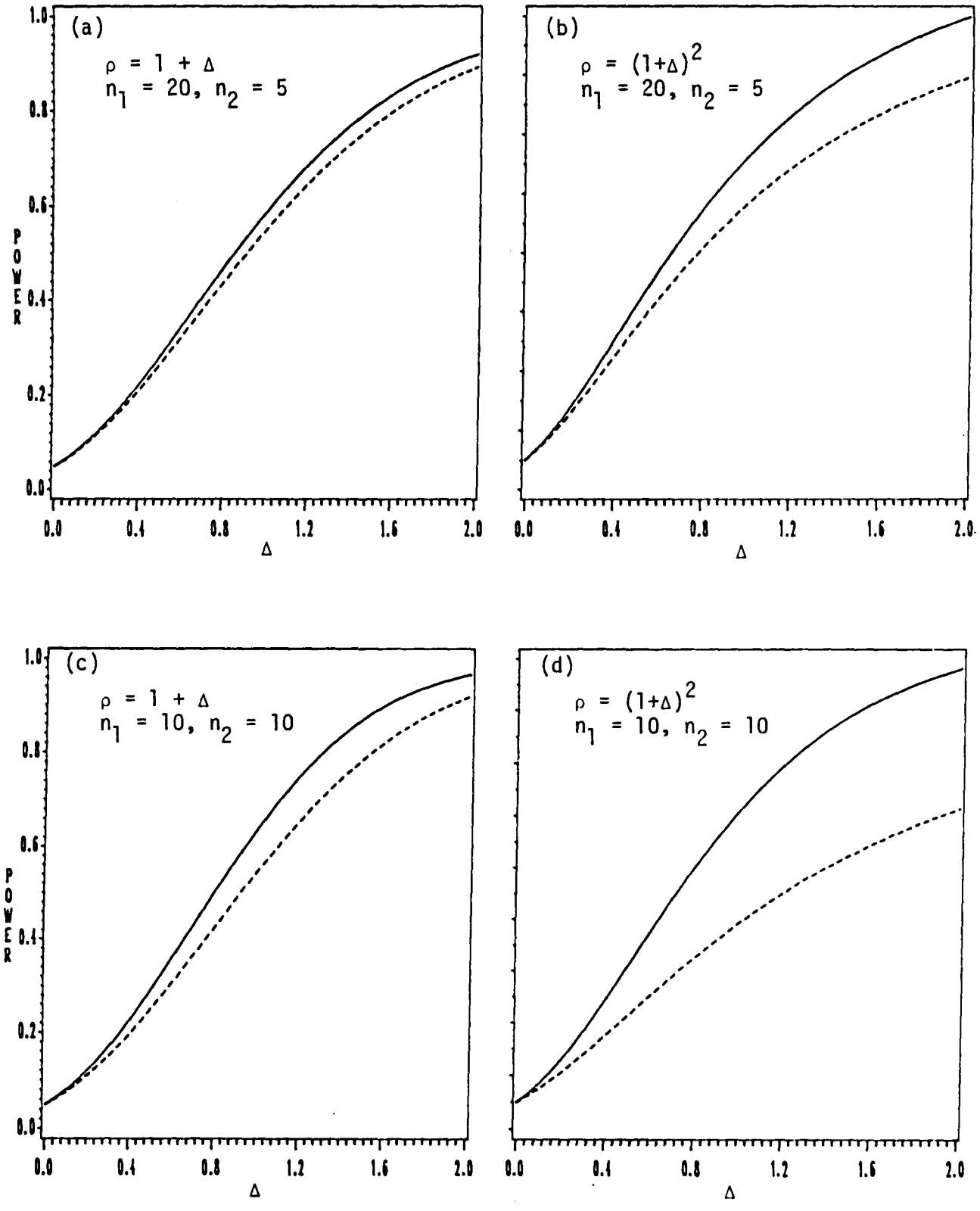

Figure 1. Power of the pooled t (----) and power of the modified t (——) for significance level α=.05, plotted against $\Delta=(\mu_1-\mu_2)/\sigma_1$ with $\rho=\sigma_2^2/\sigma_1^2$.

Table 1

Summary of data on number of SCE's per cell from benzene inhalation study on rats reported in Erexson et al. (1986).

Exposed to Benzene (5 rats per group)			Concurrent Controls (5 rats per group)		
Benzene (ppm)	$\bar{X}_t$	s_t^2	Benzene (ppm)	$\bar{X}_c$	s_c^2
0.1	8.24	.1216	0	8.08	.0616
0.3	8.23	.0827	0	8.09	.2795
1	9.09	.3011	0	8.42	.0077
3	10.51	.1371	0	8.47	.0763
10	10.43	.2270	0	8.59	.0379
30	11.14	.0629	0	8.58	.0821

pooled control variance $\bar{s}_c^2 = .09085$

Table 2

Summary of tests comparing mean SCE frequency for each exposure concentration with the concurrent control mean.

	1. Pooled t(8df)		2. Pooled t(28df)		3. Modified t(24df)		4.Welch t		
Comparison	t_p	p value*	t_p	p value	t_m	p value	t_w	p value	df
0 vs .1ppm	.836	.214	.820	.210	.839	.205	.776	.227	11.2
0 vs .3ppm	.535	.304	.760	.227	.755	.229	.773	.226	14.7
0 vs 1ppm	2.704	.013	3.056	.002	3.525	.001	2.400	.025	6.7
0 vs 3ppm	9.874	.000	10.332	.000	10.701	.000	9.554	.000	10.3
0 vs 10ppm	7.968	.000	8.731	.000	9.621	.000	7.274	.000	7.6
0 vs 30ppm	15.035	.000	13.735	.000	13.429	.000	14.60	.000	17.7

* right tailed p values for $t=(\bar{X}_t-\bar{X}_c)/[\hat{var}\ (\bar{X}_t-\bar{X}_c)]^{1/2}$