Manuscript for Chapter 8 (pp. 297-344) of book edited by Edwards LK (1993), *Applied Analysis of Variance in the Behavioral Sciences,* New York: Marcel Dekker. Translated to pdf format: 4 July 1998.

8
Unified Power Analysis for t-Tests through Multivariate Hypotheses

Ralph G. O'Brien[1]
Cleveland Clinic Foundation
Cleveland, OH

Keith E. Muller[2]
University of North Carolina
Chapel Hill, NC

8.1 INTRODUCTION

Determining adequate and efficient sample sizes is often critical in designing worthy studies. Yet too many studies have sample sizes that are too small to ensure enough statistical power to confirm meaningful effects. Freiman, Chalmers, Smith, and Kuebler (1979) concluded this about clinical trials in medicine. Sedlmeier and Gigerenzer (1989) reached a similar judgment about studies in psychology. The message in both articles is cogent to all fields that rely on statistical inference. Perhaps such articles are having positive effects, for we see signs that researchers are now paying more attention to power. For example, reviewers of research proposals now often require that sound power analyses be done before they will recommend funding or access to facilities and subject populations.

Going through the process of determining and justifying the sample size also has an important ancillary effect: it catalyzes the synergism between science and statistics at the study's conception. The statistician who performs a thorough power analysis is more likely to scrutinize the proposed design, assess issues regarding data management, and develop a sound plan for the data analysis. Such involvement can improve the proposal in a number of ways, thus increasing its chance for approval, funding, scientific success, and publication.

In this chapter, we present a strategy for performing power analyses that is applicable to the broad range of methods subsumed by the classical normal-theory univariate or multivariate general linear models. First, we introduce the requisite concepts of statistical power using concepts from the familiar t-test to compare two independent group means. Second, we proceed to the comparison of two correlated means (matched-pairs problem) and on to the one-way analysis of variance (ANOVA) with contrasts for a completely randomized design. Third, we develop power analysis for the univariate general linear model, thus providing a broad range of applications. We illustrate this with an analysis of covariance (ANCOVA) problem that has unequal distributions of the covariate's values

[1]Supported in part by grants from the US National Institutes of Health (GCRC: RR00082) and the University of Florida Division of Sponsored Research, which funded Zhanying Bai and Yonghwan Um in their wriitng of one portion of the OneWyPow.sas freeware module. Dan Bowling helped in many ways. E-mail: robrien@bio.ri.ccf.org.

[2]Supported in part by grants from the US National Institutes of Health (NICHD: P30-HD03110-22, NCI: P01 CA47982-04, GCRC: RR00046). E-mail: muller@bios.unc.edu

among the groups as well as heterogeneous slopes. Fourth, we broaden the range still further by outlining an approximation for determining power under the multivariate general linear model. This is illustrated with a repeated-measures problem solved by using the multivariate analysis of variance (MANOVA) approach. Fifth and finally, we outline power analysis strategies developed for other types of methods, especially for tests to compare two independent proportions.

With writings on sample-size choice and power analysis for many methods now so plentiful, why do we offer yet another one? Like Kraemer and Thiemann (1987), we present an approach that unifies many seemingly diverse methods. We think our method is intuitive, because we develop *the strong parallels between ordinary data analysis and power analysis.* To make the methods easier to use, we distribute modules of statements to direct the popular SAS® System (1990) to perform and table (or graph) sets of power computations. Our ultimate goal is to show how a single approach covers a broad class of tests.

Rather than restricting attention to the power of the traditional tests (e.g., overall main effects and interactions), our methods allow one to easily examine statistical hypotheses that are more tailored to specific research questions. Departing from most writings on statistical power, we take unbalanced designs to be the norm rather than the exception. Many effective research designs use unequal sample sizes, as when certain types of subjects are easier than others to recruit or when certain treatments are more expensive per subject to apply. Thus, researchers and statisticians must decide how the total sample size will be allocated among the different groups of cases, with a balanced allocation being a special case.

We avoid oversimplifying the concept of effect size, as researchers often do when they employ rules of thumb, such as Cohen's (1988, 1992) “small,” medium,” and “large” categorizations. A tiny effect size for one research question and study could be a huge effect size in another. Researchers often claim that their studies promise “medium” effect sizes, but they have no objective grounds to justify such a claim. Our scheme forces researchers to give specific conjectures or estimates for the relevant statistical parameters, such as the population means and standard deviations for an ANOVA problem. The conjectures are then used directly to calculate effect sizes, which determine statistical power for a proposed sample size. Our detailed examples illustrate how straightforward it is to do these things. Sample-size analysis is not harder to do than data analysis; as we shall see, the two problems are very similar.

The best hypothesis-driven research proposals include quite definite plans for data analyses, plans that merge the scientific hypotheses with the research design and the data. A good sample-size analysis must be congruent with a good plan for the data analysis. There are hundreds of other common statistical procedures besides normal-theory linear models, and there are thousands of uncommon methods and an unlimited number of “customized” ones that are developed for unique applications. While we can cover power for many of the common methods for statistical inference, one chapter cannot be exhaustive of all known or possible sample-size methods.

One final general point is in order. Many treatments on statistical power choose to

address questions like "If I perform this particular hypothesis test and my conjectures are true about the expected data, what sample size do I need to achieve a power of .90?" But most studies have multiple hypotheses to test, and thus this kind of sample-size analysis determines separate sample sizes for the different hypothesis tests within a single study. This is confusing. Instead, our approach addresses questions like "If I study a total of N subjects, what are the powers for the different hypothesis tests I plan to perform?" This is the way of most studies—one total sample size, several hypotheses to be examined. In those limited cases where there really is but one key hypothesis, one is still limited by resources on how large the sample size can be. The question here becomes: "I can study, at most, N subjects. Will this give me sufficient power?" If that value for N gives "too much" power, one can spend a few minutes investigating lower values to find something more efficient.

At the end of each major section, we explore the art of power analysis by presenting realistic, detailed examples inspired by actual studies conducted within the interface of behavioral and medical research. All of these case studies are fictionalized to some degree to make them more useful in this chapter. We use bogus names for the researchers and drugs, as well as for many of the scientific terms and measurements.

8.2 UNIVARIATE t-TESTS AND THE ONE-WAY ANOVA

8.2.1 Comparing Two Independent Means

Common t-tests probably are used more frequently than any other statistical method. In comparing the population means (μ) from two independent groups, we are formally comparing a null hypothesis, $H_0: \mu_1 = \mu_2$, with a research ("alternative") hypothesis that is either nondirectional, $H_A: \mu_1 \neq \mu_2$, or directional, $H_A: \mu_1 > \mu_2$ or $H_A: \mu_1 < \mu_2$. We "assess the improbability of H_0" by measuring the observed statistical difference between the sample means with

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}}, \tag{8.1}$$

where $\hat{\mu}_j$ is the sample mean for group j, n_j is the sample size for group j, and $\hat{\sigma}$ is the pooled sample standard deviation. Using $N = n_1 + n_2$, we can reexpress the t statistic as

$$t = N^{1/2} \left\{(w_1 w_2)^{1/2} \left[(\hat{\mu}_1 - \hat{\mu}_2)/\hat{\sigma}\right]\right\}, \tag{8.2}$$

where $w_j = n_j/N$ is the proportion of cases in group j. Obviously, $w_2 = 1 - w_1$. We have grouped the terms to show the distinct components of this t statistic. We call the term $(\hat{\mu}_1 - \hat{\mu}_2)/\hat{\sigma} = \hat{\psi}$ the estimate of *nature's effect size*. $\hat{\psi}$ is used often in meta analysis to measure the relative difference between the sample means. The structure of the design defines the term $(w_1 w_2)^{1/2}$; thus $(w_1 w_2)^{1/2}\hat{\psi}$ is an estimate of the effect size specific to that structure. Factoring in $N^{1/2}$ gives us t. We will soon see how the t and its components translate directly into their counterparts in power analysis.

If H_0 is true, and the observations are independent, and they are distributed as Normal

random variables having the same variance in both groups, then t is a *central* t random variable with N – 2 degrees of freedom for error (df_E), denoted here as $t(df_E)$. Let t_α be the upper-tail critical value, satisfying: $\alpha = Pr[\ t(df_E) \geq t_\alpha\ ]$. Thus, α is the Type I error rate. Directional tests are based on t_α (or $-t_\alpha$, depending on how H_A is defined). Nondirectional tests can use $t_{\alpha/2}$ and $-t_{\alpha/2}$. This is identical to using $F = t^2$ and $F_\alpha = t^2_{\alpha/2}$, which is a form that better unifies our discussion throughout. While strict Normality usually does not hold, the practical consequences of the Central Limit Theorem allow us to suppose that t is still distributed as $t(df_E)$ for many non-Normal situations.

8.2.1.1 Directional Research Hypothesis

First we discuss the power of the directional test, and we define it using H_A: $\mu_1 > \mu_2$. The power is simply the rejection rate, $Pr[t \geq t_\alpha]$, when H_0 is false. (When H_0 is true, the rejection rate is simply α.) Power is dependent on the *noncentrality* value,

$$\delta = N^{1/2}\left\{(w_1 w_2)^{1/2}\ [(\mu_1 - \mu_2)/\sigma]\right\}, \tag{8.3}$$

which is positive for H_A: $\mu_1 > \mu_2$. Note that δ is merely the t value that would result if one had "exemplary" data, that is, data in which $\hat{\mu}_1 \equiv \mu_1$, $\hat{\mu}_2 \equiv \mu_2$, and $\hat{\sigma} \equiv \sigma$. Thus δ measures the statistical difference between two population means as realized with n_1 and n_2 observations.

It is helpful to see δ constructed as diagrammed in Figure 8.1. Begin with nature's effect size

$$\psi = (\mu_1 - \mu_2)/\sigma, \tag{8.4}$$

which depends on parameters related only to *what* is being studied, not *how* it is going to be studied. Because μ_1, μ_2, and σ are almost always unknown in practice, making conjectures or estimates for these values comes after considering the coordinating theory of the problem, reviewing the literature, and, sometimes, collecting pilot data. The *primary noncentrality* is

$$\delta^* = (w_1 w_2)^{1/2}\psi. \tag{8.5}$$

This $(w_1 w_2)^{1/2}$ factor depends only on the structure of design. Note that δ^* is maximized at $w_1 = w_2 = .5$. Factoring in the total sample size gives the noncentrality,

$$\delta = N^{1/2}\delta^*. \tag{8.6}$$

This description shows how noncentrality, which determines power, is a function of three distinct components: what nature hides from us (ψ), the structure of the design (w_1 and w_2), and the size of the sample (N).

Under the same distributional assumptions outlined above, t is distributed as a *noncentral* t random variable with $df_E = N - 2$ and noncentrality δ, which is denoted as $t(df_E, \delta)$. The central t is simply $t(df_E, 0)$. The power is

$$\Omega = Pr[\ t(df_E, \delta) \geq t_\alpha\ ]. \tag{8.7}$$

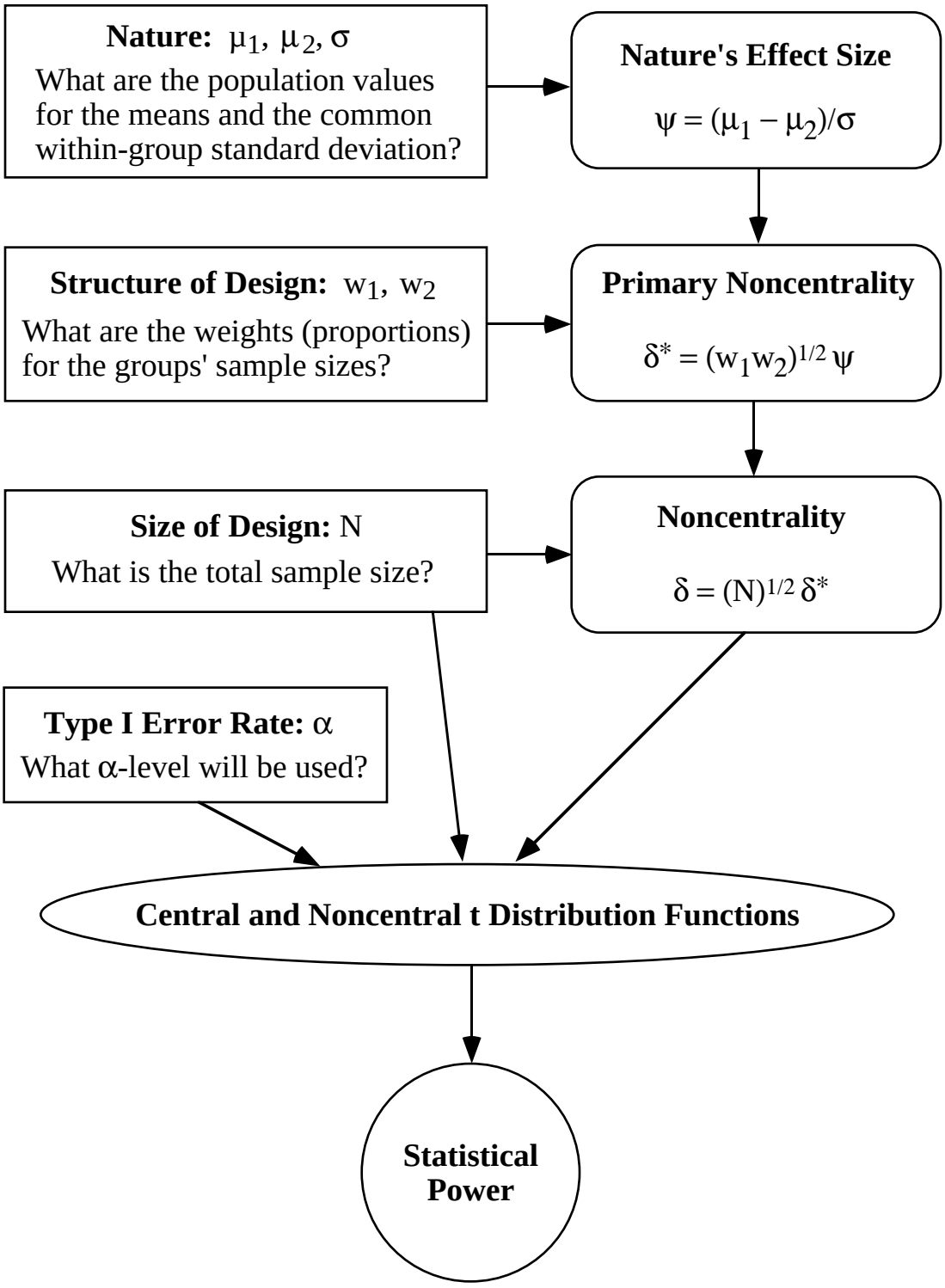

Figure 8.1. Building the noncentrality for the two-group t statistic.

Power values can be found using a suitable computing routine, table, or graph. For example, one can use the SAS statements:

```
t_alpha = TINV(1 - alpha, dfE, 0);
power = 1 - PROBT( t_alpha, dfE, delta);
```

Note that TINV and PROBT are defined with respect to the cumulative distribution function, hence the need for "1 - alpha" and "1 - PROBT()" to get values with respect to the upper tail.

The mean of a noncentral t random variable is approximated well by

$$E\,[t(df_E, \delta)] \approx N^{1/2}\delta^*\{(4df_E - 1)/(4df_E - 4)\}, \qquad \text{for } df_E > 1, \tag{8.8}$$

a result immediately deducible from the work on meta analysis by Hedges (1981). Note that the mean increases as either N or δ^* increase. This always increases the power of the test.

8.2.1.2 Nondirectional Research Hypothesis

When the research hypothesis is H_A: $\mu_1 \neq \mu_2$, it is more straightforward to use the statistic $F = t^2$, so that the split rejection region of the central t distribution is unified into the upper tail of the central F. In general, F statistics in ANOVA and linear models can be denoted as being central (if H_0 is true) or noncentral F random variables with noncentrality λ (with df_H and df_E degrees of freedom in the numerator and denominator, respectively). We denote this as $F(df_H, df_E, \lambda)$; $\lambda = 0$ defines the familiar central F. For the two-group problem, we test H_0 by taking $F = t^2$ to be $F(1, N - 2, 0)$. In general, we define F_α to be the upper-tail critical value, satisfying: $\alpha = Pr[\, F(df_H, df_E, 0) \geq F_\alpha \,]$.

For $F = t^2$ the noncentrality and primary noncentrality are

$$\lambda = \delta^2 \qquad \text{and} \qquad \lambda^* = \lambda/N = (\delta^*)^2. \tag{8.9}$$

The power for the nondirectional ("2-tailed") t-test is then

$$\Omega = Pr[\, F(df_H, df_E, \lambda) \geq F_\alpha \,], \tag{8.10}$$

with $df_H = 1$ and $df_E = N - 2$. We can compute this using the SAS functions FINV and PROBF, as in:

```
F_alpha = FINV(1-alpha, dfH, dfE, 0);
power = 1 - PROBF( F_alpha, dfH, dfE, lambda);
```

It helps in understanding the noncentral F to know that the expected value of any F random variable is

$$E\,[F] = (1 + \lambda/df_H)[df_E/(df_E - 2)]. \tag{8.11}$$

As $df_E/(df_E - 2)$ is usually close to 1.0, we have

$$E\,[\mathrm{F}] \approx 1 + \lambda/\mathrm{df}_{\mathrm{H}}. \tag{8.12}$$

Thus with increasing λ, the distribution of F is shifted to the right, making it more likely that F will exceed F_α, thus increasing the power of the test.

Example 1: Two-Group t-test

Dr. Seth Alalgia is planning a study as to whether yet another form of biofeedback therapy can reduce, at least in the short term, the frequency and severity of vascular headaches (cf. Blanchard et al., 1990). Specifically, he plans to conduct a double-blind, randomized trial in which patients will receive either an enhanced thermal (ET) biofeedback therapy or a sham placebo (SP) that simply gives non-contingent feedback to the patient. Each patient will be studied from a Monday to a Friday. On Monday morning, patients will be admitted to the university's General Clinical Research Center (GCRC) to begin identical, standard medical therapy consisting of 5μg/kg/day of litosamine. On Monday and Tuesday evenings, patients will complete an extensive questionnaire that produces the pretreatment score on the Vascular Headache Index (PreVHI). On Wednesday morning, patients will be randomized to ET or SP groups, using the minimization scheme of Pocock and Simon (1975) to minimize group differences on PreVHI, gender, and age strata. The main behavioral therapy session will be done on the Wednesday afternoon, with a followup session early Thursday morning. Questionnaire data from Thursday and Friday will produce the PostVHI measure.

The main outcome measure is to be the relative change in the VHI values,

$$\mathrm{RC} = (\mathrm{PreVHI} - \mathrm{PostVHI})/\mathrm{PreVHI}.$$

PreVHI will serve as a covariate for RC, but this action is expected only to reduce error variance; the dynamic blocking minimizes group differences on the PreVHI, and no PreVHI × Group interaction is expected. Thus for our purposes here, we do not need to consider that PreVHI is to be a covariate, except when specifying σ^2. The costs for recruiting and studying ET and SP subjects are the same, so that a balanced design ($w_1 = w_2 = .50$) is optimal.

Dr. Alalgia has no pilot data. Based on his knowledge and experience, he conjectures that the ET group's true mean on RC is $\mu_{ET} = -.30$, while the SP group's is $\mu_{SP} = -.15$ (due to placebo effects). He postulates that the common within-group standard deviation is about $\sigma = .125$, because if the data are Normally distributed, about 95% of the patients' RC scores should lie within $.25 = 2\sigma$ of the group mean. (This logic helps researchers conjecture values for the standard deviation, a task difficult to do.) In continuing discussions with his statistician, Dr. Alalgia agrees that σ could possibly be

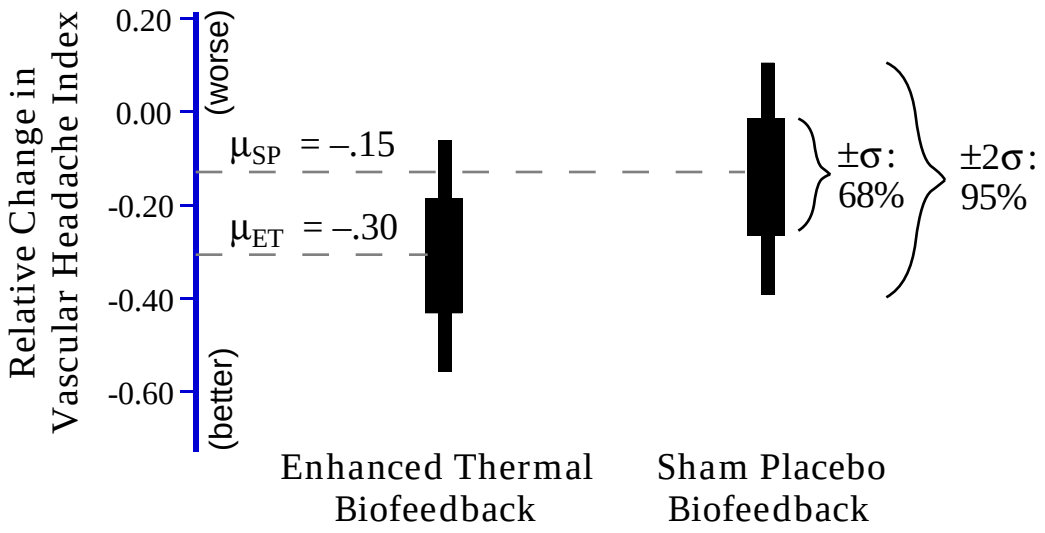

Figure 8.2. Dr. Alalgia's scenario for the two-group design.

50% higher, at $\sigma = .1875$. The $\sigma = .125$ scenario is depicted in Figure 8.2.

Dr. Alalgia estimates that he has enough resources to study N = 20 patients. The GCRC's Scientific Review Committee hopes that a valid study can be conducted by using fewer than the 100 hospital bed-days requested (20 patients $\times$ 5 days/patient). This study is an early clinical trial of this new "enhanced" behavioral treatment, and research in biofeedback therapy has a history of conflicting findings. Thus Dr. Alalgia and his statistician feel that a directional test is unwarranted at this time; only the nondirectional test will be considered. Finally, he will concentrate on the powers for $\alpha = .05$ but is still curious to see the powers under $\alpha = .01$. Under these conditions, will N = 20 provide acceptable power? Would fewer suffice?

Listing 8.1 gives input and output related to executing OneWyPow, a module of standard SAS® data step statements. (See Appendix.) Here we use OneWyPow to get powers for Dr. Alalgia's scenario by all combinations of $N \in \{14, 20, 26, 32\}$ with $\sigma \in \{.1250, .1875\}$. The $\alpha = .05$ results for the nondirectional test (labeled "2-tailed t") indicate that

Listing 8.1. Computations for Dr. Alalgia's Two-Group Design

Input

```
options ls=72 nosource2;
title1 "Dr. Alalgia: Enhanced Thermal vs Sham-Placebo Biofeedback";
%include OneWyPow;
cards;
mu -.30 -.15 .
weight .50 .50 .
sigma .125 .1875 .
alpha .05 .01 .
Ntotal 14 20 26 32 .
end
%include FPowTab1;
```

Selected Output

Effect: Two-Group Test,
DF Hypothesis: 1,
AND Primary SSHe: 0.005625

		Std Dev					
		0.125			0.1875		
		Total N			Total N		
		14	20	26	32	14	20	26	32
		Power	Power	Power	Power	Power	Power	Power	Power
Test Type	Alpha								
2-tailed t	0.05	.541	.718	.835	.907	.281	.395	.499	.591
	0.01	.264	.445	.607	.735	.101	.172	.250	.331
1-tailed t	0.05	.681	.825	.908	.953	.408	.530	.632	.714
	0.01	.370	.561	.712	.819	.160	.251	.344	.434

N = 20 may be too few subjects to run: $\Omega = .72$ for $\sigma = .125$, and $\Omega = .40$ for $\sigma = .1875$. Even N = 26 gives powers of only .84 and .50 under the same conjectures. Dr. Alalgia is discouraged about these results and understands that the GCRC cannot afford to provide more than 100 bed-days for his project. He decides to restructure the design. See Example 2.

Even researchers with limited experience in statistical planning readily understand the rationale and implications of a power analysis like that just presented. The results in Listing 8.1 show concretely how power is increased by (1) increasing the α level, (2) decreasing the variance, (3) increasing the total sample size, and (4) using a directional test.

8.2.1.3 Using Data To Specify δ^*, λ^*, ψ, and ψ^2

[Note: This section is somewhat difficult. Readers may elect to skip to Section 8.2.2.]

General Comments and Directional Test

The population group means, μ_1 and μ_2, are the focus of this t-test. The population standard deviation of the observations within the groups, σ, is usually a nuisance parameter—in both a mathematical and an everyday sense. In practice, one may be forced to make a best guess or, preferably, a range of plausible guesses for them. Often it is prudent to set $(\mu_1 - \mu_2)$ or ψ at some level that represents the lowest effect that would be of scientific or practical (clinical) interest. Although some researchers are are uncomfortable with the subjectivity of making such choices, a well-done sample-size analysis usually convinces them that statistical planning is valuable.

Often, just one part of ψ is unable to be fixed by a solid scientific conjecture or other argument. This unfixed part may be μ_1, μ_2, $(\mu_1 - \mu_2)$, or σ. If one has good preliminary (pilot) data, one can get a point estimate or confidence interval for the unfixed part, then use the other, fixed parts to get an estimate or interval for δ^*. The standard methods can be used to get confidence intervals for μ_1, μ_2, or $(\mu_1 - \mu_2)$. The problem comes when the unfixed part is σ^2. For example, μ_1 may be a control-group mean that is "well known" and thus fixed; μ_2 may be set by a "minimum practical (clinical) effect" argument at, say, $.80\mu_1$; and σ may be the lone unfixed parameter. If pilot data are available, an estimate for σ can be obtained using the usual pooled estimate, $\hat{\sigma}$. If those data are Normal, the traditional confidence limits for σ^2 can be obtained and used to define a range of σ^2 values for specifying ψ. This interval would be based on taking $(N - 2)\hat{\sigma}^2/\sigma^2$ to have a chi-square distribution with $N - 2$ degrees of freedom.

Unless the data are strictly Normal, the traditional χ^2-based confidence intervals for σ^2 cannot be trusted to have their nominal confidence levels, even for infinitely large sample sizes. Somewhat better confidence intervals for σ^2 may be obtained by adapting the transform developed by O'Brien (1979, 1981) for testing variances. Simply convert the raw observations, y_{ij}, to

$$r_{ij} = \frac{(n_j - 1.5)n_j(y_{ij} - \hat{\mu}_j)^2 - .5\hat{\sigma}_j^2(n_j - 1)}{(n_j - 1)(n_j - 2)}, \qquad (8.13)$$

and use them to construct a one-mean confidence interval based on $\bar{r}$. This can be done in most statistics packages. The method works because $\bar{r}_j = \hat{\sigma}_j{}^2$, so the interval based on $\bar{r}$ is really an interval for σ^2. The method gives proper intervals when the sample size is large. By conducting Monte Carlo studies, Melton (1986) showed that while it is considerably more robust than the Normal-theory method, it cannot be recommended for small N. We are not aware of acceptably robust methods for forming confidence limits on σ^2 when the

sample size is small.

We can see from (8.8) that given an observed t value based on suitable pilot data, a nearly unbiased estimate of the primary noncentrality is

$$\hat{\delta}^* = (t/N^{1/2})[(4df_E - 4)/(4df_E - 1)]. \quad (8.14)$$

Thus for the two-group problem, we have the nearly unbiased estimator,

$$\hat{\psi} = \hat{\delta}^*/(w_1 w_2)^{1/2} = [(\hat{\mu}_1 - \hat{\mu}_2)/\hat{\sigma}]\,[(4df_E - 4)/(4df_E - 1)]. \quad (8.15)$$

Confidence limits for ψ can also be formed. It may be particularly worthwhile to find the lower limit, $\hat{\psi}_\gamma$, of a *one-tailed*, $1 - \gamma$ interval (no upper limit) by finding the noncentrality, δ_γ, that solves

$$\Pr[\, t(df_E, \hat{\delta}_\gamma) \geq t \,] = \gamma, \quad (8.16)$$

where t is the value observed with pilot data. This can be done using the SAS statement:

```
delhatgm = TNONCT(t, dfE, 1 - gamma);
```

$\hat{\delta}_\gamma$ may then be converted to the two forms of effect size,

$$\hat{\delta}^*_\gamma = \hat{\delta}_\gamma/N^{1/2} \quad (8.17)$$

and

$$\hat{\psi}_\gamma = \hat{\delta}^*_\gamma/(w_1 w_2)^{1/2}. \quad (8.18)$$

Using this logic allows one to say with $1 - \gamma$ confidence that the nature's effect size is *at least* $\hat{\psi}_\gamma$. Setting $\gamma = .50$ gives a median estimator that competes with the mean estimator, $\hat{\psi}$.

Either $\hat{\psi}$, $\hat{\psi}_{.50}$, or $\hat{\psi}_\gamma$ can be combined with particular N and $(w_1 w_2)^{1/2}$ values to form noncentralities and corresponding estimates of power. Suppose that a pilot study with $n_1 = 6$ and $n_2 = 4$ produces t = 1.50 (p = .086). The above formulas yield estimates for nature's effect size of $\hat{\psi} = .875$, $\hat{\psi}_{.50} = .937$, and $\hat{\psi}_{.20} = .359$. For a full study having N = 50 and a balanced design, the corresponding powers are $\Omega = .920$, $\Omega_{.50} = .948$, and $\Omega_{.20} = .347$ for the .05-level, directional test. Thus, the researcher would see that although the mean-type and median-type estimates of ψ give powers greater than .90, there is a .20 chance that the power may be less than .35. This is hardly reassuring. If it seems erroneously low, realize that $\gamma = .20$ is close to p = .086, and $\hat{\psi}_{.086} = 0.00$, which corresponds to a "power" of .05. If $\gamma < p$ is used, then $\hat{\psi}_\gamma < 0$, and $\Omega < \alpha$, a result that would perplex most people.

Most power analyses are *prospective* in that they assess power for a study yet to be done, as we just discussed. *Retrospective power analyses* assess the power of a study already completed. It is often requested after a nonsignificant outcome: "What *was* my power in this (nonsignificant) study?" "How large should my N have been to have ensured acceptable power?" "What if I had had a more balanced design?" The estimation methods discussed here can help address these questions also.

Suppose that a study was completed with $n_1 = 17$, $n_2 = 15$, $\hat{\mu}_1 - \hat{\mu}_2 = 4.30$, and $\hat{\sigma} = 9.25$, which give t = 1.31 (p = .10). The researcher asks what the power was for a value of

$\mu_1 - \mu_2 = 8.00$, which would have been the minimum difference worth seeing. Fixing $\mu_1 - \mu_2$ at 8.00 and using $\hat{\sigma} = 9.25$, the mean estimate of nature's effect size is $\hat{\psi} = 8/9.25 = .865$, which translates to a power of .771 using $\alpha = .05$ and the above sample sizes. A more conservative assessment begins by finding the lower .10 critical value for the $\chi^2(17 + 15 - 2 = 30)$ distribution, which is 20.6. Under Normal-theory, $\Pr[\sigma < 9.25(30/20.6)^{1/2} = 11.2] = .90$ This gives a lower 90% one-sided confidence limit of $\hat{\psi}_{.10} = 8/11.16 = .717$, which translates to a power of .630. Technically, these are powers for an exact replication of the study. If the study were to be replicated with twice the total sample size, these two values of $\hat{\psi}$ translate to powers of $\Omega = .958$ and $\Omega_{.10} = .875$.

If faced with the power results we have just seen, some researchers would decide that the "ongoing" study does not have enough subjects yet, so they will double the total sample size by *adding* another 32 subjects. This strategy will inflate the Type I error rate unless the whole sequence of *interim analyses* is planned at the outset using special statistical methods (cf. Fleming, Harrington, and O'Brien, 1984).

As the first example (with $n_1 = 6$ and $n_2 = 4$) shows, confidence intervals for ψ can be quite wide when based on small pilot studies, providing values that researchers find unreasonable when planning a new study. It is useful to use something like $\hat{\psi}_{.33}$ to estimate a lower limit for the effect size, understanding, of course, that these offer lower levels of confidence than do more standard values, such as $\hat{\psi}_{.05}$, $\hat{\psi}_{.10}$, and $\hat{\psi}_{.20}$. From the frequentist's perspective, using $\hat{\psi}_{.05}$ is more likely to lead to choosing an N that provides adequate power. But researchers' knowledge of the subject matter should not be discounted entirely in favor of pilot results. Bayesian strategies may one day be commonly available to mix one's subjective beliefs about ψ with estimates of it from pilot data.

Nondirectional Test

Based on (8.11), it can be shown that

$$\hat{\lambda}^* = \{[(df_E - 2)/df_E]df_H F - df_H\}/N \tag{8.19}$$

is an unbiased estimate of λ^*. Unfortunately, if $F < df_E/(df_E - 2)$, then $\hat{\lambda}^* < 0$. The adjusted estimator,

$$\hat{\lambda}^*_{adj} = \text{Max}[\, 0, \hat{\lambda}^*] \tag{8.20}$$

can still be positively biased for small λ.

Adapting a method introduced by Venables (1975), the lower limit of a one-tailed $1 - \gamma$ interval for λ^* can be taken to be $\hat{\lambda}^*_\gamma = \hat{\lambda}_\gamma/N$, where $\hat{\lambda}_\gamma$ solves

$$\Pr[\, F(df_H, df_E, \hat{\lambda}_\gamma) \geq F \,] = \gamma. \tag{8.21}$$

The SAS statement that will do this is:

```
lamhatgm = FNONCT(F, dfH, dfE, 1 - gamma);
```

A solution for $\hat{\lambda}_\gamma$ is possible only if γ exceeds the p-value for F, i.e., $F > F_{\alpha=\gamma}$. As above, we may set $\gamma = .50$ for a median-type estimator or use $\gamma = .20$ or .33 to get some other

useful lower bound on λ. We remark once again that with small N, setting γ too low often results in a lower bound for λ^* that is of little practical use.

When applied to the two-group problem, any of the estimators for λ^* can be converted to estimators for nature's effect size squared, $\psi^2 = (\mu_1 - \mu_2)^2/\sigma^2$, using

$$\widehat{\psi}^2 = \lambda^*/(w_1 w_2). \tag{8.22}$$

This can then be the basis for estimating noncentralities for different $w_1 w_2$ values than were used in the pilot or previous study.

While these methods for estimating λ for a nondirectional test share the practical weaknesses of their counterparts for estimating δ for a directional test, they still may have utility when ample pilot or previous data are available. Unfortunately, there are no excellent methods for using small pilot studies to provide firm, objective estimates for effect sizes and noncentralities. Because there is usually little or no pilot data available for prospective power analyses, the methods suggested in this section rarely play a leading role in developing and justifying specific estimates. *In practice, most good estimates for effect sizes are conjectures formed in the mind of thoughtful researchers, experienced in their science, and careful in their use of pilot data or related results found in the literature.*

8.2.2 Comparing Two Correlated Means (Matched-Pairs)

The common matched-pairs t-test applies to studies that have only one group of cases, but each case provides two measurements, y_1 and y_2, whose means will be compared to each other. The analysis focuses on the difference, $y_d = y_1 - y_2$, which we will take to have a mean of μ_d and a standard deviation of σ_d. Now we work with a null hypothesis, H_0: $\mu_d = 0$ (no difference), and a research ("alternative") hypothesis that is either nondirectional (H_A: $\mu_d \neq 0$) or directional (say, H_A: $\mu_d > 0$). The statistic is

$$t = N^{1/2}(\widehat{\mu}_d/\widehat{\sigma}_d), \tag{8.23}$$

where N is the number of *pairs*. We take t to be distributed as a $t(df_E, \delta)$ random variable with $df_E = N - 1$ and

$$\delta = N^{1/2}(\mu_d/\sigma_d) \qquad \text{and} \qquad \delta^* = \psi = (\mu_d/\sigma_d). \tag{8.24}$$

Here again, t and δ have the same form. There is no way to vary the sample-size structure of a one-group design, thus the primary noncentrality and nature's effect size are identical, involving only μ_d and σ_d.

If it is difficult to specify σ_d directly, it may be specified indirectly by conjecturing values for σ_1 and σ_2, the standard deviations of y_1 and y_2, as well as ρ, the ordinary product-moment correlation between y_1 and y_2. Then,

$$\sigma_d = (\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\ \sigma_2)^{1/2}. \tag{8.25}$$

Once δ^* is determined, power is computed in the same manner as given in the previous section, except that we have $df_E = N - 1$ pairs, instead of $df_E = N - 2$ individual subjects.

Listing 8.2. Computations for Dr. Alalgia's Matched-Pairs Design

Input

```
options ls=72 nosource2;
%include OneWyPow;
title1 "Dr. Alalgia: Yoked-Pairs Design for Biofeedback Study";
title2 "mu_diff = .15";
cards;
mu .15 .
sigma .137 .205 .
alpha .05 .01 .
Ntotal 10 14 17 20 .
end
%include FPowTab1;

%include OneWyPow;
title2 'mu_diff = .10';
cards;
mu .10 .
sigma .137 .205 .
alpha .05 .01 .
Ntotal 10 14 17 20 .
end
%include FPowTab1;
```

Selected Output

mu_diff = .15

Effect: One-Group Test,
DF Hypothesis: 1,
AND Primary SSHe: 0.0225

		Std Dev						
		0.137			0.205			
		Total N			Total N			
		10	14	17	20	10	14	17	20
		Power	Power	Power	Power	Power	Power	Power	Power
Test Type	Alpha								
2-tailed t	0.05	.868	.966	.988	.996	.542	.716	.808	.873
	0.01	.598	.838	.927	.970	.251	.427	.551	.659
1-tailed t	0.05	.938	.987	.996	.999	.688	.828	.893	.934
	0.01	.727	.908	.963	.986	.362	.550	.667	.761

Example 2: Matched-Pairs t-test

Dr. Seth Alalgia now considers a new design to study the efficacy of the enhanced thermal (ET) biofeedback therapy relative to the sham-placebo (SP) therapy in the relief of vascular headache. (See Example 1.) Patients will be run in pairs that are matched on several factors, including a screening version of the Vascular Headache Index (VHI). They will no longer be admitted to the hospital's research unit during their pretreatment phase, but will be seen as outpatients on Monday and Tuesday to complete the PreVHI measure and begin their standardized medical treatment with litosamine. The pair will be admitted on Wednesday, already randomly split into the two treatment groups. The ET subject will be treated first, getting legitimate biofeedback. Then the SP subject will be treated as a "yoked control," getting the ET subject's session as non-contingent biofeedback. The same relative change score, RC, will serve the outcome measure, so that our paired differences are $d = RC_{SP} - RC_{ET}$. The paired difference on the PreVHI may serve as a covariate to lower the error variance in d. But as in Example 1, covariates need not concern us here other than to influence our specification of σ_d. It is hoped that by matching and yoking pairs of patients, the error variance can be markedly reduced. Note also that now only three GCRC bed-days per subject will be used, rather than five, as before.

Dr. Alalgia believes that this study carries with it the same general treatment effects and variability as for the two-group study. Thus, he takes $\mu_d = .15$. As some of the reviewers consider this to be overly optimistic, he decides to look at $\mu_d = .10$ as well. Dr. Alalgia believes that the correlation between RC_{SP} and RC_{ET} is at least .40; thus it is decided to base the power analysis on both $\sigma_d = .125[2 - 2(.40)]^{1/2} \approx .137$ and $\sigma_d = .1875[2 - 2(.40)]^{1/2} \approx .205$. He is prepared to study up to 17 pairs of patients, which will require $17 \times 6 = 102$ bed-days. Is this enough? Too many?

Listing 8.2 gives the input to OneWyPow and some of its output. Under Dr. Alalgia's scenario ($\mu_d = .15$), the $\alpha = .05$, nondirectional test with N = 17 has power of .99 and .81 for the two standard deviations used. The corresponding powers under $\mu_d = .10$ (not shown) are .81 and .47. Dr. Alalgia decides to go forward with N = 17 and feels he can now easily defend this choice before the Scientific Review Committee.

8.2.3 One-Way Design: Overall Test and Planned Contrasts

The one-way analysis of variance is an extension of the two-group t-test. To compare the means of J independent groups, we work with two types of hypothesis tests, overall ("between-groups") tests and planned contrasts. The overall hypothesis is $H_0: \mu_1 = \mu_2 = ... = \mu_J$ (all populations means are equal), and the research ("alternative") hypothesis, $H_A: \mu_j \neq \mu_{j'}$, for some $j \neq j'$ (at least two means are different). But note that the overall research hypothesis is a nonspecific one. Often the researcher is more interested in one or more planned contrasts, which can be specified using $H_0: c_1\mu_1 + c_2\mu_2 + ... + c_J\mu_J = \theta_0 = 0$. The research hypothesis is either nondirectional, $H_A: \theta_0 \neq 0$, or it is directional, which we take to be $H_A: \theta_0 > 0$. The contrast weights, c_j, are usually defined such that $c_1 + c_2 + ... + c_J = 0$. Contrasts are formed to represent the specific hypotheses of interest. For example in a four-group design, the contrast $H_0: (.5)\mu_1 + (.5)\mu_2 + (-1)\mu_3 + (0)\mu_4 = 0$ compares the

average of the first two means to the third mean. Example 3 illustrates several others. Planned contrasts are useful because they provide a more sharply focused analysis than do overall tests. This usually makes tests of planned contrasts easier to interpret and more powerful. In those rare cases in which the overall test captures the scientific question under study, the power of that test should naturally be of concern. On the other hand, if the main purpose is to test specific planned contrasts, one should choose a sample size based on those contrasts and pay little, if any, attention to the power of the overall test. For more on this, see Rosenthal and Rosnow (1985).

OneWyPow computes power for both overall tests and contrasts in the one-way design.

8.2.3.1 Overall Test in the One-Way ANOVA

The overall test of the equality of J independent means is performed using

$$F = N\left[\frac{\sum_{j=1}^{J} w_j(\hat{\mu}_j - \hat{\bar{\mu}})^2}{\hat{\sigma}^2(J-1)}\right], \tag{8.26}$$

where $w_j = n_j/N$ is the proportion of cases in group j,

$$\hat{\bar{\mu}} = (w_1\hat{\mu}_1 + w_2\hat{\mu}_2 + ... + w_J\hat{\mu}_J) \tag{8.27}$$

is the weighted average of the sample means, and $\hat{\sigma}^2$ is the weighted (pooled) average of the sample variances. Under the classic Normal-theory assumptions, the observed overall F statistic is an F random variable with $df_H = J - 1$ degrees of freedom for the hypothesis (numerator) and $df_E = N - J$ degrees of freedom for error (denominator). Its noncentrality is $\lambda = N\lambda^*$, where

$$\lambda^* = \frac{\sum_{j=1}^{J} w_j(\mu_j - \bar{\mu})^2}{\sigma^2}, \tag{8.28}$$

with $\bar{\mu} = w_1\mu + w_2\mu_2 + ... + w_J\mu_J$. Note that $\lambda/(J-1)$ is isomorphic to the overall F.

8.2.3.2 ANOVA Contrasts

It will simplify matters at times if we express $H_0: c_1\mu_1 + c_2\mu_2 + ... + c_J\mu_J = 0$ by only writing out the contrast coefficients in a matrix with one row, $\mathbf{C} = [\, c_1 \;\; c_2 \;\; ... \;\; c_J \,]$. Because $df_H = 1$, the choice must be made between a nondirectional or directional test. For the directional test, we form the statistic

$$t = N^{1/2} \frac{\sum_{j=1}^{J} c_j \hat{\mu}_j}{\hat{\sigma} \left(\sum_{j=1}^{J} c_j^2 / w_j \right)^{1/2}} \quad , \tag{8.29}$$

and use it as we used the two-group t-test described in Section 8.2.1.1 except that we now have $df_E = N - J$. The primary noncentrality is

$$\delta^* = \frac{\sum_{j=1}^{J} c_j \mu}{\sigma \left(\sum_{j=1}^{J} c_j^2 / w_j \right)^{1/2}} \quad , \tag{8.30}$$

and the noncentrality is $\delta = N^{1/2} \delta^*$. ψ cannot be defined, as the w_j are linked to the c_j and cannot be factored out of δ^*.

The material in Section 8.2.1.1 covers some methods to use pilot data to help estimate δ, and these extend to the t-tests for contrasts.

For the nondirectional case, the test involves $F = t^2$, which is taken to be $F(1, N - J, \lambda)$, where $\lambda = \delta^2$. Methods for using pilot data to help estimate λ follow from the material in Section 8.2.1.2.

Example 3: One-Way ANOVA with Contrasts

Dr. Mindy Bowdy studies relationships between personality characteristics and immune functioning (cf. Jemmott, et al., 1990). For her next study, up to 100 (but preferably 60) male college students will give a sample of blood and complete a set of psychological instruments that are used to partition the sample into four types of people (D, O, L, and F):

Dominator: one who has a strong need to impact others; argumentative, assertive, overly competitive.

Ordinary: one who is not a Dominator, a Loner, or a Friendly.

Loner: one who has a minimal need for friendships; no desire to impact others.

Friendly: one who has a very strong need to create and nurture friendships; accepts this as an ideal and an end in its own right.

Other studies have found Friendlies to have better immune functioning than Ordinaries, who have better functioning than Dominators. Dr. Bowdy believes that the immune functioning of Loners may only be slightly better than that of Ordinaries. The specific question now is: “How do the DOLF groups differ with respect to Q-type killer cell activity (QKCA), as measured by the percentage of Q843 human leukemia cells lysed at a

50:1 effector-to-target ratio?"

Dr. Bowdy conjectures that the mean lysis rates are $\mu_D = .35$ (worst), $\mu_O = .50$, $\mu_L = .52$, $\mu_F = .60$ (best). Data from previous studies working with QKCA and Q843 suggest that the within-group standard deviation (σ) in this population is between .16 and .19. Furthermore, Dr. Bowdy expects that subjects will be partitioned according to the following proportions: $w_D = .20$, $w_O = .50$, $w_L = .10$, $w_F = .20$.

There are many strategies available to compare the groups' means in this study; good scientists and statisticians might disagree on what is best. Dr. Bowdy recognizes that the overall test of whether the four groups have equal lysis rates is weakened by the small difference between Ordinaries and Loners. She will assess it anyway, as peer reviewers and editors will expect to see it. Several contrasts are meaningful to Dr. Bowdy. The largest nature's effect size, ψ, for any two-group comparison is "Dominators vs. Friendlies," H_A: $\mu_F - \mu_D > 0$, i.e., the directional test of $\mathbf{C} = [\,-1 \;\; 0 \;\; 0 \;\; 1]$. She also wants directional contrasts of "Friendlies vs. Ordinaries and Loners," $\mathbf{C} = [\,0 \;\; -5/6 \;\; -1/6 \;\; 1]$, where the 5/6 and 1/6 are used because Ordinaries are expected to outnumber Loners by 5:1. Similarly, we have "Dominators vs. Ordinaries and Loners," $\mathbf{C} = [\,-1 \;\; 5/6 \;\; 1/6 \;\; 0]$, directional. This family of three contrasts will be protected using a Bonferroni-adjusted level of $\alpha = .05/3$ for each. Though "Loners vs. Ordinaries" probably will have little power, Dr. Bowdy will test it with a nondirectional test of $\mathbf{C} = [\,0 \;\; 1 \;\; -1 \;\; 0]$ at $\alpha = .05$.

Listing 8.3.1 gives the input and Listing 8.3.2 gives some of the key output related to using OneWyPow. These results demonstrate the benefits of looking at more than just overall tests when selecting a sample size. Here the overall test for the four-group design is fairly powerful; even if σ is large (.19), N = 80 gives a .05-based power of .89. It can be shown that the noncentrality for the "Friendlies vs. Dominators" comparison captures over 95% of noncentrality for the overall test. Examining the powers for "Dominators vs. Ordinaries and Loners", we find that most of the power in this design is due to the conjectured immune inferiority of the Dominators. The immune superiority of the Friendlies is not likely to be supported by a significant "Friendlies vs. Ordinaries and Loners" test. Reasonably assurance that this will be significant seems to require far more than a total of 100 subjects, especially when we use Bonferroni protection.

Finally, we see almost no power for the .05-level comparison of the Ordinaries vs. Loners. It might be that we could create a more powerful overall test, call it "Almost Overall," by pooling the means over the Ordinaries and Loners. This is can be done by aggregating the between-groups variance defined by $\mathbf{C} = [\,1 \;\; -5/6 \;\; -1/6 \;\; 0]$ and $\mathbf{C} = [\,0 \;\; -5/6 \;\; -1/6 \;\; 1]$ giving a two degree of-freedom contrast. The theory for this is reviewed in Section 8.3, but the programming of OneWyPow is straightforward, as shown in Listing 8.3.1. Because there is so little difference between the conjectured means for the Ordinaries and Loners, pooling them creates a test that is a little more powerful than the ordinary overall test.

Listing 8.3. Computations for Dr. Bowdy's Four-Group Design

Listing 8.3.1. Input

```
options ls=72 nosource2;
%include OneWyPow;
title1 "Mindy Bowdy : Four-Group Design";
*Order of groups: D O L F;
cards;
mu .35 .50 .52 .60 .
weight .20 .50 .10 .20 .
sigma  .16 .19 .
alpha .05 .0167 .
Ntotal 60 80 100 .
contrasts
"Friendlies vs Ordin & Loners" 0 -.83 -.17 1 .
"Dominators vs Ordin & Loners" -1 .83 .17 0 .
"Friendlies vs Dominators" -1 0 0 1 .
"Ordinaries vs Loners" 0 1 -1 0 .
"Almost Overall (2 DF)" 1 -.83 -.17  0 .
>                       0 -.83 -.17  1 .
end
%include FPowTab2;
run;
```

Listing 8.3.2. Selected Output

ALPHA 0.05

		Std Dev					
		0.16			0.19		
		Total N			Total N		
		60	80	100	60	80	100
		Power	Power	Power	Power	Power	Power
Overall test	Regular F	.899	.970	.992	.763	.887	.951
Ordinaries vs Loners	2-tailed t	.059	.062	.065	.056	.058	.060
	1-tailed t	.086	.093	.099	.079	.084	.090
Almost Overall (2 DF)	Regular F	.933	.982	.996	.821	.923	.969

ALPHA 0.0167

		Std Dev					
		0.16			0.19		
		Total N			Total N		
		60	80	100	60	80	100
		Power	Power	Power	Power	Power	Power
Friendlies vs Ordin & Loners	2-tailed t	.265	.366	.464	.182	.253	.325
	1-tailed t	.362	.473	.573	.263	.347	.428
Dominators vs Ordin & Loners	2-tailed t	.659	.806	.897	.487	.637	.754
	1-tailed t	.755	.874	.938	.597	.735	.832
Friendlies vs Dominators	2-tailed t	.909	.974	.993	.772	.896	.956
	1-tailed t	.948	.987	.997	.849	.938	.976

8.3 UNIVARIATE GENERAL LINEAR (FIXED-EFFECTS) MODEL

All of the tests we have thus far considered, as well as many others we have not considered, are special cases of tests within univariate general linear modeling (GLM). Understanding noncentrality in GLM testing, and knowing how to easily perform the computations, creates a wonderfully broad spectrum of power analyses.

Consider the standard model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{8.31}$$

where $\mathbf{y}$ is the $N \times 1$ vector of the dependent variable, $\mathbf{X}$ is the $N \times r$ model matrix of nonrandom, known predictor values, $\boldsymbol{\beta}$ is the $r \times 1$ vector of the fixed, unknown coefficients, and $\boldsymbol{\varepsilon}$ is the $N \times 1$ vector of true residuals. Without loss of generality, we require the columns of $\mathbf{X}$ to be linearly independent, i.e., rank($\mathbf{X}$) = r. For tests on $\boldsymbol{\beta}$, we take the elements of $\boldsymbol{\varepsilon}$ to be independent $N(0, \sigma^2)$ random variables. The usual estimates are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{8.32}$$

and

$$\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(N - r)\ . \tag{8.33}$$

We focus here on the general linear hypothesis,

$$H_0: \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\theta}_0, \tag{8.34}$$

where $\mathbf{C}$ is $df_H \times r$ with rank($\mathbf{C}$) = $df_H \le r$. $\boldsymbol{\theta}_0$ is a vector of constants appropriate to the research question. It is usually chosen to be $\mathbf{0}$. For $df_H > 1$, $H_A: \mathbf{C}\boldsymbol{\beta} \neq \boldsymbol{\theta}_0$ is the only alternative we consider. For $df_H = 1$, the directional alternative might be appropriate, which is defined here to be $H_A: \mathbf{C}\boldsymbol{\beta} > \boldsymbol{\theta}_0$. The test statistic is $F = SSH/(df_H\hat{\sigma}^2)$, where

$$SSH = (\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0)^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0) \tag{8.35}$$

is the sums of squares for the hypothesis. F is distributed $F(df_H, df_E, \lambda)$, where $df_E = (N - r)$ and

$$\lambda = (\mathbf{C}\boldsymbol{\beta} - \boldsymbol{\theta}_0)^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}(\mathbf{C}\boldsymbol{\beta} - \boldsymbol{\theta}_0)/\sigma^2. \tag{8.36}$$

This general structure encompasses t-tests, fixed-effects ANOVA and ANCOVA, and ordinary multiple regression.

λ may be expressed in a way that displays its distinct components. Let $\dot{\mathbf{X}}$ be the $q \times r$ essence model matrix formed by assembling the q unique rows of $\mathbf{X}$. in other words, $\dot{\mathbf{X}}$ is the collection of the unique design points (e.g. groups) for the proposed study. Let $\mathbf{W}$ be the $q \times q$ diagonal matrix having elements w_j, the proportion of the total sample size associated with the j^{th} row of $\dot{\mathbf{X}}$. Thus, N$\mathbf{W}$ holds the q sample sizes. It can be shown that

$$\lambda = N\{(\mathbf{C}\boldsymbol{\beta} - \boldsymbol{\theta}_0)^T[\mathbf{C}(\dot{\mathbf{X}}^T\mathbf{W}\dot{\mathbf{X}})^{-1}\mathbf{C}^T]^{-1}(\mathbf{C}\boldsymbol{\beta} - \boldsymbol{\theta}_0)/\sigma^2\}. \tag{8.37}$$

We see that the primary noncentrality, $\lambda^* = \lambda/N$, is based on the design points to be used

($\dot{\mathbf{X}}$), the sample size weightings for those points (**W**), the conjectured estimates for the unknown effects ($\boldsymbol{\beta}$) and the variance (σ^2), and the specification of the hypothesis (**C**, $\boldsymbol{\theta}_0$).

One particularly straightforward and useful application of the GLM involves the use of the cell means model, $y_{ij} = \mu_j + e_{ij}$, for the J-group ANOVA. Here $\dot{\mathbf{X}} = \mathbf{I}$, the $J \times J$ identity matrix, and $\boldsymbol{\beta} = \boldsymbol{\mu}$, the vector of population means. Taking $\boldsymbol{\theta}_0 = \mathbf{0}$, we have

$$\lambda = N(\mathbf{C}\boldsymbol{\mu})^T[\mathbf{C}\mathbf{W}^{-1}\mathbf{C}^T]^{-1}(\mathbf{C}\boldsymbol{\mu})/\sigma^2. \tag{8.38}$$

For the two-group t-test, **I** is 2×2, $\boldsymbol{\mu}^T = [\mu_1 \ \mu_2]$, **W** = diag[$w_1$ w_2], and **C** = [1 –1]. One can use (8.38) to match the square of (8.3). For the three-group problem, **I** is 3×3, $\boldsymbol{\mu}^T = [\mu_1 \ \mu_2 \ \mu_3]$, and **W** = diag[$w_1$ w_2 w_3]. The overall test can then use

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

This can be extended easily to J groups. To derive the λ for contrasts in a J-group ANOVA [see (8.30)], use $\mathbf{C} = [c_1 \ c_2 \ \ldots \ c_J]$. It can also be used to specify the effects for factorial designs. For example, a 2×3 factorial design can usually be handled by taking **I** to be 6×6, $\boldsymbol{\mu}^T = [\mu_{11} \ \mu_{12} \ \mu_{13} \ \mu_{21} \ \mu_{22} \ \mu_{23}]$, and **W** = diag[$w_{11}$ w_{12} w_{13} w_{21} w_{22} w_{23}]. The interaction test requires

$$\mathbf{C} = \begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 \end{bmatrix},$$

and the main effects can be defined as is appropriate for the research design and question.

Other constructions for $\dot{\mathbf{X}}$ will handle designs with missing cells, nested factors, fixed blocking factors, and multiple covariates. Therefore, power analysis can be done for any hypothesis that can be handled by the fixed-effects features of any linear models program. Thus, there is rarely a justification for limiting the power analysis to a minor part of a large design, such as the comparison of just two of the groups contained within a $2 \times 3 \times 3$ factorial.

Exemplary Data: SSH_e, SSH_e^*

Note that if we use an *exemplary data set* of N_e cases conforming to $\mathbf{y} \equiv \mathbf{X}\boldsymbol{\beta}$, we obtain $\hat{\boldsymbol{\beta}} \equiv \boldsymbol{\beta}$. This makes $SSH \equiv \lambda\sigma^2$, which is called the exemplary sums of squares hypothesis, or SSH_e. Defining $SSH_e^* = SSH_e/N_e$ to be the primary SSH_e, we have

$$\lambda = N \cdot SSH_e^*/\sigma^2. \tag{8.39}$$

In comparing means using t-tests or the ANOVA, the exemplary data must be designed to produce sample means that are identical to the conjectured population means: $\hat{\mu}_j \equiv \mu_j$. For the two-group case this yields

$$SSH_e = N_e w_1 w_2 (\mu_1 - \mu_2)^2 = N_e \sigma^2 \lambda^*. \tag{8.40}$$

SSH_e^* subsumes the complex parts of computing the noncentrality parameter. By first computing SSH_e using standard software for data analysis, an array of λ values can be formed by combining SSH_e^* with various values of N and σ^2. SSH_e^* is a common concept that unifies the notion of λ across all cases of the general linear model. As a computing

scheme, it is very helpful in handling complex linear models, such as in the unbalanced ANCOVA design illustrated below. The module PowSetUp has been designed to create tables of power probabilities based on SSH_e, N_e, N, σ, and α values.

Random X Variables

As noted well by Gatsonis and Sampson (1989), if **X** involves random variables (the correlational model), then the noncentral F results given here do not hold strictly, unless one is willing to take the *conditional* viewpoint, that is, "given this **X**." Fortunately, the conditional results discussed in this chapter do provide reasonable approximations to methods developed under the assumption that the columns of **X** are multivariate normal. The practical discrepancy between the two approaches disappears as the sample size increases. Because the values for the population parameters are conjectures or estimates, strict numerical accuracy of the power computations is usually not critical. Those who find this view too cavalier may use the excellent tables or software provided by Gatsonis and Sampson, which have their own limitation due to their dependence on multivariate normality.

Example 4: Univariate GLM for ANCOVA

Dr. Mindy Bowdy completed the four-group study (Example 3). She now wants to extend the work to focus on whether greater psychological stress is associated with poorer immune function, as Cohen, Tyrrell, and Smith (1991) reported by correlating higher stress with an increased risk of catching a cold. Dr. Bowdy's proposed design is similar to her D-O-L-F four-group design, except that the Ordinaries and Loners are now combined into a single group called Regulars. Thus we now have three groups, and we still expect the Dominators to have lower QKCA-Q843 lysis rates than the Regulars, who will have lower rates than the Friendlies. This study, however, will focus on the Life Events Stress Index (LESI), which takes on values:

–2: low stress
–1: below average stress
0: average stress
+1: above average stress
+2: high stress

Regulars and Friendlies are conjectured to have uniform distributions for the LESI, whereas the Dominators are conjectured to have LESI values higher than normal. The main research questions involve the relationships between stress and immunity, as measured by the LESI and the lysis rates. Dr. Bowdy conjectures that the Friendlies have no relationship between the LESI and lysis rates, the Regulars have a weak negative relationship, and the Dominators have a stronger negative relationship. She understands that addressing this question probably requires several hundred subjects.

Listing 8.4. Computations for Dr. Bowdy's ANCOVA Design

Listing 8.4.1. SAS statements to compute SSH_e values using regular data analysis on exemplary data.

```
options ls =72;
title "Mindy Bowdy: ANCOVA design";
data;
/*
Three groups, one covariate, unequal slopes;
* Order of exemplary data:
  D = Dominator
  R = Ordinary or Loner
  F = Friendly
*/
input lysis0 DRF $ LESI n;

*create exemplary data;
if DRF = 'D' then beta = -.03;
if DRF = 'R' then beta = -.01;
if DRF = 'F' then beta =  .00;
lysis = lysis0 + beta*LESI;

*  lysis at                      ;
*   LESI=0  DRF  LESI    n       ;   cards;
    .3350    D    -2    02
    .3350    D    -1    03
    .3350    D     0    04
    .3350    D     1    05
    .3350    D     2    06
    .5033    R    -2    12
    .5033    R    -1    12
    .5033    R     0    12
    .5033    R     1    12
    .5033    R     2    12
    .6000    F    -2    04
    .6000    F    -1    04
    .6000    F     0    04
    .6000    F     1    04
    .6000    F     2    04
* analyze exemplary data to get SSHe values;
proc glm order=data; class DRF; freq n;
model lysis = DRF DRF*LESI/noint solution;
contrast 'DRF main|LESI=0' DRF 1 -1 0, DRF 0 1 -1;
contrast 'Means:DvsR|LESI=0' DRF 1 -1 0;
contrast 'Means:FvsR|LESI=0' DRF 0 1 -1;
contrast 'LESI main | DRF' DRF*LESI 1 1 1;
contrast 'DFR*LESI' DRF*LESI 1 -1 0, DRF*LESI 0 1 -1;
contrast 'Slopes: D vs R' DRF*LESI 1 -1 0;
contrast 'Slopes: F vs R' DRF*LESI 0 1 -1;
```

Listing 8.4.2. Selected portions from SAS output giving SSH_e values.

```
Dependent Variable: LYSIS
Frequency:          N


Source                    DF

Uncorrected Total        100
```

This shows N_e value.

```
Contrast                  DF    Contrast SS

DRF main|LESI=0            2     0.6722149
Means:DvsR|LESI=0          1     0.3837566
Means:FvsR|LESI=0          1     0.1402634
LESI main | DRF            1     0.0258462
DRF*LESI                   2     0.0175385
Slopes: D vs R             1     0.0108387
Slopes: F vs R             1     0.0030000
```

SSH_e values

Copy this section into PowSetUp.

Listing 8.4.3. SAS statements to obtain power tables.

```
options ls=72 nosource2;
title1 "Mindy Bowdy: DRF groups and LESI stress measure";
%include PowSetUp;
cards;
Ne 100
alpha .05 .01 .
sigma .12 .15 .
Ntotal  200 300 500 .
numparms 6
effects
 DRF main | LESI = 0           2     0.6722149
 Means: D vs R | LESI = 0      1     0.3837566
 Means: F vs R | LESI = 0      1     0.1402634
 LESI main | DRF               1     0.0258462
 DFR*LESI                      2     0.0175385
 LESI slopes: D vs R           1     0.0108387
 LESI slopes: F vs R           1     0.0030000
end
*;
%include FPowTab2;
```

This section was copied from SAS GLM output, then the effects' titles were modified slightly.

Listing 8.4.4. Output from PowSetUp and FPowTab2 SAS modules.

ALPHA 0.05

		Std Dev					
		0.12			0.15		
		Total N			Total N		
		200	300	500	200	300	500
		Power	Power	Power	Power	Power	Power
DRF main | LESI = 0	Regular F	.999	.999	.999	.999	.999	.999
Means: D vs R | LESI = 0	2-tailed t	.999	.999	.999	.999	.999	.999
	1-tailed t	.999	.999	.999	.999	.999	.999
Means: F vs R | LESI = 0	2-tailed t	.992	.999	.999	.940	.991	.999
	1-tailed t	.997	.999	.999	.970	.996	.999
LESI main | DRF	2-tailed t	.470	.638	.848	.326	.456	.667
	1-tailed t	.596	.749	.911	.447	.582	.773
DRF*LESI	2-tailed t	.264	.380	.588	.182	.256	.404
LESI slopes: D vs R	2-tailed t	.231	.322	.491	.164	.224	.341
	1-tailed t	.336	.442	.615	.252	.328	.462
LESI slopes: F vs R	2-tailed t	.098	.124	.175	.081	.097	.129
	1-tailed t	.158	.196	.266	.129	.155	.203

An essence matrix for this design is

$$\ddot{\mathbf{X}} = \begin{bmatrix} 1 & -2 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 \end{bmatrix}, \quad \begin{array}{l} \text{("D" LESI = –2)} \\ \text{("D" LESI = –1)} \\ \text{("D" LESI = 0)} \\ \text{("D" LESI = +1)} \\ \text{("D" LESI = +2)} \\ \\ \text{("R" LESI = –2)} \\ \text{("R" LESI = –1)} \\ \text{("R" LESI = 0)} \\ \text{("R" LESI = +1)} \\ \text{("R" LESI = +2)} \\ \\ \text{("F" LESI = –2)} \\ \text{("F" LESI = –1)} \\ \text{("F" LESI = 0)} \\ \text{("F" LESI = +1)} \\ \text{("F" LESI = +2)} \end{array}$$

which has corresponding regression coefficients

$$\boldsymbol{\beta}^T = [\mu_D \ \beta_D \quad \mu_R \ \beta_R \quad \mu_F \ \beta_F].$$

This defines three distinct simple regression models of the form

$$E[\text{lysis}_{ij}] = \mu_j + \beta_j \text{ LESI, where } j \in \{\text{“D” “R” “F”}\}. \tag{8.41}$$

Dr. Bowdy specifies the regression coefficients to be

$$\boldsymbol{\beta}^T = [.3350 \quad -.0300 \quad .5033 \quad -.010 \quad .6000 \quad .000].$$

She takes σ to be .12 or .15. Dr. Bowdy cannot control how the total sample size will disperse among the 15 combinations of DRF and LESI values, but this is conjectured to follow the weights

$$\mathbf{w}^T = \text{diag}(\mathbf{W}) = [.02 \ .03 \ .04 \ .05 \ .06 \ .12 \ .12 \ .12 \ .12 \ .12 \ .04 \ .04 \ .04 \ .04 \ .04].$$

Listing 8.4.1 gives basic SAS statements showing how an exemplary data set can be constructed and then analyzed with a standard linear models routine, PROC GLM. Note how the 'n' variable and the FREQ statement define the weights, which are 100*$\mathbf{w}$. Thus, the first data line effectively creates 2 observations, the second creates 3, etc. Thus N_e = 100 cases are produced, all with lysis values equal to their respective expected values. We could have used PROC GLM's WEIGHT statement to take the non-integer weights given by $\mathbf{w}$ (here, .02 .03 ...), thus giving N_e = 1.00, but not all routines will handle this. If, as in some routines, PROC GLM had been unable to handle data giving a sum of squares residuals of 0.0, we could have simply replaced the last data line, which is

```
.6000   F   2   04
```

with the two lines

```
.5000   F   2   02
.7000   F   2   02
```

This forces two residuals to be +0.10 and two to be –0.10.

The MODEL statement defines a model matrix corresponding to $\dot{\dot{\mathbf{X}}}$, with the first three columns defining the three intercepts and labeled DRF, and the last three defining the three slopes and labeled DRF*LESI. Knowing this, the CONTRAST statements shown here are quickly discernible, especially after reading the effect titles. The main point here is that power computations are practical for any test within a fixed-effects general linear models framework. Just begin by defining an exemplary data set and analyzing it as if it were the real data.

Listing 8.4.2 gives the key parts of output produced by PROC GLM's analysis of this exemplary data. The uncorrected total degrees of freedom simply gives the value for N_e; the Contrast SS "statistics" give the SSH_e values. Listing 8.4.3 shows how another SAS module, PowSetUp (see Appendix), turns these SSH values into power probabilities. Most input lines are self-explanatory; NUMPARMS gives the rank of **X** and the lines following EFFECTS are those copied verbatim from PROC GLM's output. Running the statements in Listing 8.4.3 produces the power tables, one of which is Listing 8.4.4.

These results show that to address whether stress is related to (QKCA) immune function will require at least 500 subjects, and even this number is likely to fail to show that the DRF groups have different LESI slopes (DRF*LESI effect). It is unclear at this point whether Dr. Bowdy should go ahead with the study at N = 500, seek to increase that total sample size, or redesign the study to create something with more powerful tests of hypotheses.

8.4 THE MULTIVARIATE GENERAL LINEAR MODEL

Many of the common methods in multivariate analysis can be developed within the framework of the multivariate general linear model, including Hotelling's T^2 Test, multivariate analysis of variance and covariance (MANOVA and MANCOVA) , linear discriminant analysis, profile analysis, and univariate and MANOVA-based repeated measures analyses. To our knowledge, a *general and practical* method for computing the power for such tests was not available until Muller and Peterson (1984) showed that the scheme just described for the univariate GLM extends to a sound method for the multivariate GLM. We briefly summarize the theory here and present a straightforward example. Note how our development here parallels that of the last section. In fact, all of univariate results given heretofore in this chapter can be shown to be special cases of the multivariate results given in this section. While the conceptual jump from the univariate to the multivariate power problem is not a particularly difficult one, actually conducting a multivariate power analysis can be a most complex exercise due to the need to make conjectures about many parameters. For more on this, see Muller, LaVange, Ramey, and

Ramey (1992).

Consider the standard model

$$\mathbf{Y} = \mathbf{XB} + \boldsymbol{\mathcal{E}}, \tag{8.42}$$

in which $\mathbf{Y}$ is the $N \times p$ matrix of the dependent variables, $\mathbf{X}$ is the $N \times r$ model matrix (full column rank, as in the univariate case), $\mathbf{B}$ is an $r \times p$ matrix of the fixed, unknown coefficients, and $\boldsymbol{\mathcal{E}}$ is the $N \times p$ matrix with independent row vectors of true residuals having covariance matrix $\boldsymbol{\Sigma}$. The usual estimates are

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{8.43}$$

and

$$\hat{\boldsymbol{\Sigma}} = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})/(N - r). \tag{8.44}$$

We focus here on the general linear hypothesis,

$$H_0: \mathbf{CBU} = \boldsymbol{\Theta}_0 , \tag{8.45}$$

where $\mathbf{C}$ is $df_C \times r$ with full rank($\mathbf{C}$) = $df_C \le r$, identical in form and function to that of the univariate case. The $\mathbf{U}$ matrix is $p \times df_U$ with full rank($\mathbf{U}$) = $df_U \le p$. The degrees of freedom for H_0 is $df_H = df_C df_U$. $\boldsymbol{\Theta}_0$ is almost always chosen to be $\mathbf{0}$. Just as $\mathbf{C}$ controls contrasts on the rows of $\mathbf{B}$, $\mathbf{U}$ controls contrasts on its columns. The net effect of $\mathbf{U}$ is to create a set of df_U "contrast variables," $\tilde{\mathbf{Y}} = \mathbf{YU}$, giving the model

$$\tilde{\mathbf{Y}} = \mathbf{XBU} + \boldsymbol{\mathcal{E}}\mathbf{U} = \mathbf{X}\tilde{\mathbf{B}} + \tilde{\boldsymbol{\mathcal{E}}}, \tag{8.46}$$

This now involves the covariance matrix $\mathbf{U}^T\boldsymbol{\Sigma}\mathbf{U} = \tilde{\boldsymbol{\Sigma}}$ and the hypothesis H_0: $\mathbf{C}\tilde{\mathbf{B}} = \boldsymbol{\Theta}_0$. Thus if $df_U = 1$, the problem becomes a univariate GLM on $\tilde{\mathbf{y}} = \mathbf{YU}$ and the methods of the last section apply directly. Such a reduction in dimensionality is often done in data analysis. The matched-pairs t-test discussed in Section 8.2.2 is a basic example of this, using $p = 2$, $\mathbf{U}^T = [1 \ \ -1]$. Ordinary MANOVAs require $\mathbf{U} = \mathbf{I}$, along with the same $\mathbf{X}$ and $\mathbf{C}$ matrices used for their univariate counterparts. The cell means model, for example, becomes the cell mean-vectors model. For a three-group case, $\mathbf{C} = [1 \ \ -.5 \ \ -.5]$ tests the first mean vector versus the average of the other two mean vectors. For a full treatment on how forming contrast variables can handle repeated measures, see O'Brien and Kaiser (1985).

All of the common test statistics for H_0 are based on the assumption that the rows of $\boldsymbol{\mathcal{E}}\mathbf{U}$ are independent, multivariate Normal vectors. SSH from the univariate model now generalizes to

$$\begin{aligned} \mathbf{H} &= (\mathbf{C}\hat{\mathbf{B}}\mathbf{U} - \boldsymbol{\Theta}_0)^T[\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T]^{-1}(\mathbf{C}\hat{\mathbf{B}}\mathbf{U} - \boldsymbol{\Theta}_0) \\ &= N(\mathbf{C}\hat{\mathbf{B}}\mathbf{U} - \boldsymbol{\Theta}_0)^T[\mathbf{C}(\dot{\mathbf{X}}^T\mathbf{W}\dot{\mathbf{X}})^{-1}\mathbf{C}^T]^{-1}(\mathbf{C}\hat{\mathbf{B}}\mathbf{U} - \boldsymbol{\Theta}_0) \\ &= N\mathbf{H}^*, \end{aligned} \tag{8.47}$$

the $df_U \times df_U$ sums of squares and cross products matrix for the hypothesis. $\dot{\mathbf{X}}$ and $\mathbf{W}$

were defined and discussed in Section 8.3. $\hat{\sigma}^2$ generalizes to $\mathbf{U}^T\hat{\mathbf{\Sigma}}\mathbf{U} = \mathbf{E}/(N - r)$, where

$$\mathbf{E} = \mathbf{U}^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})\mathbf{U}. \tag{8.48}$$

The $s = \min(df_C, df_U)$ eigenvalues of $\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}$, $\hat{\rho}_k^2$, are the (generalized) squared canonical correlations. Many authors choose to work with the roots of $\mathbf{H}\mathbf{E}^{-1}$, $\hat{\rho}_k^2/(1 - \hat{\rho}_k^2)$; mathematically, the choice is arbitrary. Many multivariate texts, including Seber's (1984), describe the Wilks, Hotelling-Lawley, Pillai-Bartlett, and Roy methods for forming test statistics from these eigenvalues and the transforms that convert the first three to approximate F statistics. Wilks' likelihood ratio statistic is the determinant of $\mathbf{E}(\mathbf{H} + \mathbf{E})^{-1}$, or equivalently,

$$W = \prod_{k=1}^{s} (1 - \hat{\rho}_k^2). \tag{8.49}$$

Rao's transformation converts this to an $F(df_H, df_E)$ statistic,

$$F_W = \frac{(1 - W^{1/g})/df_H}{(W^{1/g})/df_E}, \tag{8.50}$$

where

$$g = \begin{cases} 1 & df_H \le 3 \\ [(df_H^2 - 4)/(df_C^2 + df_U^2 - 5)]^{1/2} & df_H \ge 4 \end{cases} \tag{8.51}$$

and

$$df_E = g[N - r - (df_U - df_C + 1)/2] - (df_H - 2)/2. \tag{8.52}$$

Muller and Peterson outlined a good approximation (exact if $df_U = 1$) for the noncentrality of F_W. Replacing $\hat{\mathbf{B}}$ with $\mathbf{B}$ in (8.47) gives us $\mathbf{H}_e = N\mathbf{H}_e^*$, the multivariate version of $SSH_e = N \cdot SSH_e^*$, the exemplary data form used in Section 8.3. Compute the eigenvalues of $\mathbf{H}_e^*\{\mathbf{H}_e^* + [(N - r)/N]\mathbf{U}^T\mathbf{\Sigma}\mathbf{U}\}^{-1}$, which are denoted ρ_{ek}^2. Use ρ_{ek}^2 in place of $\hat{\rho}_k^2$ in computing F_W to get F_{We}. Then F_W is distributed approximately as $F(df_H, df_E, \lambda)$, where

$$\lambda = df_H F_{We}. \tag{8.53}$$

Due to the $(N - r)/N$ term in the eigenproblem leading to λ, λ/N is not invariant to N, as it is in the univariate case. Thus the primary noncentrality cannot be defined. O'Brien and Shieh (under review) have proposed a modification to this method that does define a primary noncentrality and that may give more accurate power probabilities.

Muller and Peterson applied this idea to get power approximations for the F statistics associated with the Hotelling-Lawley and Pillai-Bartlett statistics as well. They summarized arguments supporting the method's numerical accuracy. Barton and Cramer (1989) reported Monte Carlo results that also supported the accuracy of the approximation. A general, practical power approximation for Roy's Greatest Root statistic does not exist, for here even its null distribution is difficult to characterize.

Example 5: Multivariate GLM for a Cross-Over Design

Dr. Katie Kohlimein is planning an experiment to assess how mental stress and an adrenaline-like hormone affect the metabolism of cholesterol in people age 30-45 with normal cholesterol levels. Gender differences are of key interest here, because of the much greater incidence of coronary heart disease in men (cf. Stoney, Matthews, McDonald, and Johnson, 1988). The study will compare the following conditions:

Control Day (C): Subject will have an eight-hour admission to the General Clinical Research Center (GCRC), having blood and urine sampled every hour.

Mental Stress Day (S): Same as C, but in addition performs a series of stressful mental tasks for two hours.

Dosophrine Day (D): Same as C, but in addition is infused with 30 mg/kg/hour dosophrine continuously for two hours. Dosophrine is similar to adrenaline (epinephrine).

Subjects will go through each condition, one per month in a randomly determined order, thus forming a three-period cross-over design. Each admission will be preceded by two weeks on a low-fat, low-cholesterol diet. At Hour 0 the subject will be given a "meal" very high in cholesterol. Baseline measurements will be taken at Hours 1 and 2 and averaged. Experimental manipulations will occur during Hours 3 and 4. Measurements at Hours 5 and 6 will be averaged to give short-term, post-treatment values. The outcome measure of interest for this power analysis will ΔLDL56, the change in low-density lipoprotein cholesterol (LDL-C) blood levels from Hours 1-2 to Hours 5-6. Equal numbers of men and women will be studied and equal numbers of subjects will get the six orders of of treatment (CSD, CDS, ..., DSC). Thus, the total sample size will be a factor of 12. Due to the expensive labwork, a tight budget calls for N = 36 (18 of each gender), but it may be possible to run N = 48. Cross-over effects will be checked, but are considered unlikely. The main analysis will employ a MANOVA-based repeated measures structure, with one two-level between-subjects factor, Gender, and one three-level within-subjects (cross-over) factor, Treatment.

Dr. Kohlimein makes conjectures for the mean ΔLDL56 values, which can be arranged to form the matrix,

$$\mathbf{B} = \begin{array}{c} \begin{matrix} \text{C} & \text{S} & \text{D} \end{matrix} \\ \begin{bmatrix} 3 & 12 & 8 \\ 1 & 5 & 7 \end{bmatrix} \end{array} \begin{matrix} \\ \text{male} \\ \text{female} \end{matrix} .$$

She further specifies that the within-group variances are 25, 64, and 36 across the C, S, and D conditions. She also believes that the C/S and C/D correlations are .400 and the S/D correlation is .625 (5/8), thus forming the covariance matrix,

Listing 8.5. Computations for Dr. Kohlimein's Cross-Over Study.

Listing 8.5.1. SAS PROC IML statements to use MVpower module.

```
options nosource2;
title "Katie Kohlimein: Power for 2 (between) x 3 (within)";
proc iml;
%include MVpower;
Opt_off = {UNIGG};
*Define matrices *;
*****************************
* BETA ROW1: Males            *
*      ROW2: Females          *
*****************************;
* COLUMNS:  Control  Psych Stress  Dosophrine;
beta   = {    3          12             8     ,
              1           5             7     };

sigma = {    25          16            12     ,
             16          64            30     ,
             12          30            36     };

essenceX=I(2);    *creates 2 x 2 identity matrix;
repN={12 18 24};  *creates N sizes of 24, 36, and 48;
round=3;

Utreat =  { 1  0,
           -1  1,
            0 -1};

title2 "Gender main effect";
C = {1 -1};
U = {1,
     1,
     1}/3;
run power;

title2 "Treatment main effect";
U = Utreat;
C = {.5 .5};
run power;

title2 "Gender*Treatment interaction";
C = {1 -1};
U = Utreat;
run power;
```

Listing 8.5.2. Selected Output

Katie Kohlimein: Power for 2 (between) x 3 (within)

Gender main effect

CASE	ALPHA	TOTAL_N	WLK_PWR
1	0.05	24	0.326
2	0.05	36	0.467
3	0.05	48	0.589

Treatment main effect

CASE	ALPHA	TOTAL_N	WLK_PWR
1	0.05	24	0.983
2	0.05	36	0.999
3	0.05	48	0.999

Gender*Treatment interaction

CASE	ALPHA	TOTAL_N	WLK_PWR
1	0.05	24	0.461
2	0.05	36	0.671
3	0.05	48	0.814

$$\boldsymbol{\Sigma} = \begin{array}{c} \begin{matrix} C & S & D \end{matrix} \\ \begin{bmatrix} 25 & 16 & 12 \\ 16 & 64 & 30 \\ 12 & 30 & 36 \end{bmatrix} \end{array} \begin{matrix} C \\ S \\ D \end{matrix} .$$

Define $\mathbf{C}_A$ = [.5 .5], $\mathbf{C}_G$ = [1 -1],

$$\mathbf{U}_A = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}, \text{ and } \mathbf{U}_T = \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix},$$

where A = "average", G = "Gender," and T = "Treatment." Then the null hypotheses that specify the Gender main, Treatment main, and Gender × Treatment interaction effects are H_G: $\mathbf{C}_G\mathbf{B}\mathbf{U}_A = \mathbf{0}$, H_T: $\mathbf{C}_A\mathbf{B}\mathbf{U}_T = \mathbf{0}$, and $H_{G\times T}$: $\mathbf{C}_G\mathbf{B}\mathbf{U}_T = \mathbf{0}$.

We compute the approximate λ and associated power using MVpower, a module of SAS IML statements (see Appendix). The input and selected output are given in Listing 8.5. In MVpower, $\mathbf{X}$ can be formed by repeating the "essenceX" matrix, $\dot{\mathbf{X}}$, "RepN" times. For example, if RepN is 3, then

$$\mathbf{X} = \begin{bmatrix} \dot{\mathbf{X}} \\ \dot{\mathbf{X}} \\ \dot{\mathbf{X}} \end{bmatrix}.$$

Because this is just a balanced two-group design (between-subjects), we set $\dot{\mathbf{X}}$ to be a 2×2 identity matrix. With RepN set at 12, 18, or 24, we create an $\mathbf{X}$ matrix for a cell-means model having N of 24, 36, or 48. The rest of the input should be easy to follow. MVpower has far more generality and complexity than we deal with here.

With $\Omega > .99$, the test of the Treatment main effect has superb power at N = 36. But the Gender main and Gender*Treatment interaction effects are the key components of Dr. Kohlimein's research question, and thus she is troubled that N = 36 yields powers of only 0.47 and 0.67 under this scenario. Even with N = 48, the powers of the two tests are only .59 and .81. The latter sounds powerful, but represents a Type II error rate of .19. Of course, this is a rather limited power analysis at this point. Other reasonable scenarios for $\mathbf{B}$ and $\boldsymbol{\Sigma}$ should be studied as well. This might include the specification of tighter, planned contrasts having $df_U = 1$.

8.5 OTHER METHODS

This chapter focuses almost exclusively on performing sample-size analyses for situations in which the research questions call for studies designed to reject null hypotheses to be tested under fixed-effects linear models, both univariate and multivariate. What about analyzing sample size in other situations? In this section, we summarize some methods and cite key references that are helpful in other common situations. This includes methods appropriate for handling tests for simple random effects models and univariate repeated

measures analysis. One large class of interest is the analysis of categorical variables, from comparing two independent proportions, to more complex situations, such as comparing or modeling many proportions (logit analysis) or modeling contingency tables (log-linear models). Survival analyses are useful when the outcome measure is "time to event A" and the observations will be right censored because the subject is lost to followup or the study will end before the some subjects have experienced "event A." Finally, we note that sample size considerations should sometimes focus on the widths of key confidence intervals, as when studies are being designed to show that treatments are indeed very similar rather than different.

8.5.1 Univariate Repeated Measures

Most texts on univariate experimental design discuss the so-called traditional univariate repeated measures analysis, also known as within-subjects designs and split-plot designs. Maxwell and Delaney (1990) were particularly thorough in comparing the univariate approach with the multivariate approach, which is outlined above in Example 5. The univariate method makes Subjects a random-effects blocking factor and thus becomes a mixed-model ANOVA. Its key assumption is that the vectors of repeated residuals from each subject are distributed according to a covariance matrix that has a structure called sphericity. Because sphericity is commonly violated, the Geisser-Greenhouse (1958) and Huynh-Feldt (1970) corrections to the mixed-model solution are often used. Muller and Barton (1989, 1991) and Muller, LaVange, Ramey, and Ramey (1992) presented a scheme to compute approximate powers for these methods, but this work is beyond the scope of this chapter. The MVpower module contains special features to handle the univariate repeated measures problem.

8.5.2 Random-Effects ANOVAs

For ordinary random-effects designs, the non-null distributions of the common F-statistics are entirely different from their counterparts for fixed-effects designs—even when the null distributions are identical. Consider a one-way, random-effects design with J levels in a factor "A," in which we test the variance component, $H_0: \sigma_A^2 = 0$ versus $H_0: \sigma_A^2 > 0$. Most texts note that the appropriate F-test (Normal-theory) is identical to that for the fixed-effects case, given here as (8.26). Scheffé (1959, p. 226) shows that if there are n cases in each level, F is distributed as $\{1 + n\sigma_A^2/\sigma^2\}F(J - 1, nJ - J, \lambda = 0)$. The term $\{1 + n\sigma_A^2/\sigma^2\}$ is the ratio of the numerator and denominator expected means squares (EMS). When $H_0: \sigma_A^2 = 0$ is true, $\{1 + n\sigma_A^2/\sigma^2\} = 1.0$, making the null case identical to that for fixed effects. When $\sigma_A^2 > 0$, the non-null sampling distribution is just $\{1 + n\sigma_A^2/\sigma^2\}$ times a *central* F distribution. Thus, the power is

$$\Omega = \Pr[\, F(J - 1, nJ - J, \lambda = 0) \geq F_\alpha/\{1 + n\sigma_A^2/\sigma^2\}]. \qquad (8.54)$$

The SAS code to do this is simply

```
EMSratio = 1 + n*VarCompA/(sigma**2);
F_alpha = FINV(1-alpha, J – 1, n*J – J, 0);
```

```
power = 1 – PROBF( F_alpha/EMSratio, J – 1, J*n – J, 0);
```

Prihoda (1983) generalized this idea somewhat and offered FORTRAN shareware.

8.5.3 Comparing Two Independent Proportions

Many studies require comparing two independent proportions, i.e., testing $H_0: \pi_1 = \pi_2$, using sample proportions, p_1 and p_2, estimated from $n_1 = w_1 N$ and $n_2 = w_2 N$ cases randomly drawn from (or assigned to) two groups. This is sometimes called the "2×2 comparative trial," because it can be put into the context of a 2×2 contingency table in which one tests the independence of Group (control vs. experimental) and Outcome (success vs. failure). Just getting good p-values is such a complex problem that over twenty-five different methods have been suggested for the generic 2×2 problem. Some of these have several possible sample-size/power solutions. Here we give two power approximations based on the common t statistic and another for Fisher's exact test. The OneWyPow module performs the computations for all three.

Fisher's exact test is a *conditional* test in that it supposes that both margins of the 2×2 table are fixed. The methods based on the t statistic give *unconditional* tests. Debate on the merits of unconditional versus conditional tests has continued for decades with eminent statisticians taking both points of view. For the 2×2 comparative trial, however, we agree with Upton (1982) that the exact conditional test is "inappropriate," with its wide popularity coming largely from its designation of "exact." Suissa and Shuster (1985) stressed that unconditional tests are easier to interpret and explain to most nonstatisticians, who often do not understand the implications of making conditional inferences. Accordingly, Suissa and Shuster presented an *exact unconditional* test, which though computationally intensive, should one day become a common tool. For now, however, Fisher's exact conditional test is still very popular.

Unconditional tests

The standard error of $(p_1 - p_2)$ is

$$\text{s.e.}[(p_1 - p_2)] = \left[\frac{w_2 \pi_1 (1 - \pi_1) + w_1 \pi_2 (1 - \pi_2)}{N w_1 w_2}\right]^{1/2}. \tag{8.55}$$

This naturally leads to the *unpooled* t statistic,

$$t_u = N^{1/2} \left\{ \frac{[w_1 w_2]^{1/2} (p_1 - p_2)}{\{N[w_2 p_1 (1 - p_1) + w_1 p_2 (1 - p_2)]/(N - 2)\}^{1/2}} \right\}. \tag{8.56}$$

Suissa and Shuster presented exact critical values for $Z_u \equiv [N/(N - 2)]^{1/2} t_u$. When transformed to t_u, they agree well with those from the $t(N - 2)$ distribution. Accordingly, we take t_u be a $t(N - 2)$ variate with noncentrality parameter

$$\delta_u = N^{1/2} \left\{ \frac{[w_1 w_2]^{1/2} (\pi_1 - \pi_2)}{[w_2 \pi_1 (1 - \pi_1) + w_1 \pi_2 (1 - \pi_2)]^{1/2}} \right\}. \tag{8.57}$$

After studying this problem with respect to the robustness of the Type I error rates for balanced and unbalanced designs, D'Agostino, Chase, and Belanger (1988) recommended that the ordinary two-group t statistic be applied to data coded Y = 0 or 1, thus making $\hat{\mu}_j = p_j$. This is the *pooled* t statistic,

$$t_p = N^{1/2} \left\{ \frac{[w_1 w_2]^{1/2} (p_1 - p_2)}{\{N[w_1 p_1 (1 - p_1) + w_2 p_2 (1 - p_2)]/(N - 2)\}^{1/2}} \right\}. \tag{8.58}$$

As per Dozier and Muller (1993), t_p is taken to be a $t(N - 2)$ random variable with noncentrality

$$\delta_p = N^{1/2} \left\{ \frac{[w_1 w_2]^{1/2} (\pi_1 - \pi_2)}{[w_1 \pi_1 (1 - \pi_1) + w_2 \pi_2 (1 - \pi_2)]^{1/2}} \right\}, \tag{8.59}$$

which produced acceptable power estimates in their simulations.

Both noncentralities keep the form $\delta = N^{1/2}\delta^*$, where δ^*, the primary noncentrality, is not dependent on N. For balanced designs, $t_u = t_p$ and and $\delta_u = \delta_p$. Letting

$$A = \frac{w_1 \pi_1 (1 - \pi_1) + w_2 \pi_2 (1 - \pi_2)}{w_2 \pi_1 (1 - \pi_1) + w_1 \pi_2 (1 - \pi_2)}, \tag{8.60}$$

we see that $\delta_u = A^{1/2}\delta_p$. t_u will have greater (less) approximate noncentrality than t_p if and only if $A > 1$ ($A < 1$). By taking the first and second derivatives of δ_u and δ_p with respect to w_1, one can show that the optimal w_1 values for t_u and t_p are

$$\tilde{w}_{1u} = \left[1 + \left(\frac{\pi_2 (1 - \pi_2)}{\pi_1 (1 - \pi_1)} \right)^{1/2} \right]^{-1} \quad \text{and} \quad \tilde{w}_{1p} = \left[1 + \left(\frac{\pi_1 (1 - \pi_1)}{\pi_2 (1 - \pi_2)} \right)^{1/2} \right]^{-1}, \tag{8.61}$$

respectively. These results show that the unpooled test increases in power when the lower w_j weights correspond to the groups with π_j closer to 0 or 1, whereas the pooled test works in the opposite manner.

The Conditional Test

The Fisher (or Fisher-Irwin) exact test computes each p-value by defining a specific hypergeometric distribution based on fixing both margins of the table. The power approximation presented here stems from the sequence of work by Casagrande, Pike, and Smith (1978), Fleiss, Tytun, and Ury (1980), and Diegert and Diegert (1981). It is based around Yates' (1934) statistic, which can be written

$$Z_y = N^{1/2} \left\{ \frac{[w_1 w_2]^{1/2}[(p_1 - p_2) - (2Nw_1 w_2)^{-1}]}{[\bar{p}(1-\bar{p})]^{1/2}} \right\}, \qquad (8.62)$$

where $\bar{p} = w_1 p_1 + w_2 p_2$. Z_y is taken to be a standard Normal variate, Z. For the directional test, which by our convention is $H_A: \pi_1 > \pi_2$, Z_y is significant at level α if $Z_y > Z_\alpha$. The nondirectional test simply uses $Z_{\alpha/2}$. Upton (1982) concluded that Z_y gives "extremely conservative" Type I error rates when considered as an unconditional test, but that it is "practically identical" to Fisher's exact conditional test. Suissa and Shuster (1985) gave exact results showing that Fisher's test requires larger sample sizes than does their exact version of t_u.

To compute approximate power for the directional test, let $r = w_2/w_1$; $\bar{\pi} = w_1\pi_1 + w_2\pi_2$; $h = (r + 1)/(r\,|\pi_1 - \pi_2|)$; $m = (r + 1)^2 h^2/N^2 - h + N/(r + 1)$. Compute

$$Z_\Omega = \frac{-(\pi_1 - \pi_2)(m\,r)^{1/2} + Z_\alpha[(r+1)\bar{\pi}(1-\bar{\pi})]^{1/2}}{[r\,\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]^{1/2}}, \qquad (8.63)$$

then find the power, $\Omega_1 = \Pr[\, Z \geq Z_\Omega \,]$. For the nondirectional test, use $Z_{\alpha/2}$ instead of Z_α. This method undergirds the sample-size tables in Fleiss (1981), which are very handy but are limited to balanced designs. Thomas and Conlon (1992) presented an efficient algorithm to compute powers directly from the hypergeometric distribution. In particular, it should be used in cases involving a small expected frequency in one of the cells of the 2 $\times$ 2 table.

Example 6: Comparing Two Independent Proportions

One hundred forty early-Alzheimer's patients are to be recruited for a randomized, placebo controlled, double-blinded clinical trial to test whether DS110891, a synthetic form of glutamate, improves their memory and learning, at least temporarily. All patients will continue to receive their standard care. Based on several cognitive measurements, each patient will be scored "improved" or "not improved" after three months on study. It is conjectured that 40% of the DS110891 patients versus 20% of the placebo patients will improve. To optimize the power of the t_u statistic, $w_1 = 55\%$ will be randomized to receive DS110891. Because A = 1.04, t_u will be slightly more powerful than t_p.

Listing 8.6 gives input and output related to the use of OneWyPow. With N = 140, the power for the .05-based, directional test is .842 for t_u and .829 for t_p. Z_y, which corresponds to Fisher's exact conditional test, has an approximate power of .780. For a balanced design, the power is .838 for both t_u and t_p and .786 for Z_y. Suissa and Shuster reported that with N = 136 and $\alpha = .048$, the power of their exact unconditional test is exactly .804 for this case. OneWyPow returns a value for t_u of .824 for these specifications.

Listing 8.6. Computations for Comparing Two Independent Proportions

SAS Input

```
options ls=72 nosource2;
%include OneWyPow;
title1 "DS110891 trial with Alzheimer's Patients";
cards;
pi .40 .20 .
weight .55 .45 .
alpha .01 .05 .
Ntotal 100 140 200 .
end
%include PowTab2;
```

SAS Output

		ALPHA				
		0.01			0.05	
		Total N			Total N	
		100	140	200	100	140	200
		Pow-er	Pow-er	Pow-er	Pow-er	Pow-er	Pow-er
Unpooled Approx. Uncond. Test	2-tailed t	.357	.521	.718	.605	.752	.886
	1-tailed t	.456	.620	.797	.721	.842	.936
Pooled Approx. Uncond. Test	2-tailed t	.341	.500	.696	.588	.735	.873
	1-tailed t	.439	.600	.779	.706	.829	.928
Approx. Conditional Test	2-tailed Z	.264	.420	.633	.504	.671	.837
	1-tailed Z	.355	.524	.726	.633	.780	.905

Other Tests on Proportions: Log-Linear Models

Comparing two proportions is only one of the myriad of ways to analyze categorical data. The large class of methods known as log-linear models (which include logit models)

provides the familiarity and flexibility of a linear models framework. These methods commonly employ log-likelihood-ratio χ^2 test statistics, and O'Brien (1986) discussed how the strategy of using exemplary data extends well to these important methods. This strategy was summarized in Agresti's (1990) extensive text on categorical data analysis. The PowSetUp module is designed to create power tables using results obtained from an ordinary log-linear analysis of exemplary data.

8.5.4 Survival Analysis

There is a growing literature on methods for estimating power and determining sample-size for studies involving survival analysis, where the dependent measure is time until "death." Commonly, subjects are lost to followup after a known time on study or are still "living" at the time of data analysis. Such data are not missing, they are *censored*. We know that a given subject "lived" at least Y days, and we should use that information as fully as possible. Performing power analyses for such studies is a most complex problem. Analytical approximations have been developed for straightforward situations; see the summaries by Lachin (1981) and Donner (1984). Shuster (1990) developed an extensive set of tables useful for planning clinical trials. Goldman and Hillman (1992) presented a scheme that uses the "analysis" of exemplary data. Computer simulations are sometimes used to assess complex situations (Halpern and Brown, 1987). Simulation is an option that is always available when the situation is too complex or unique to warrant the effort required to develop "nice" analytical solutions.

8.5.5 "Accepting the Null" Using Appropriate Confidence Intervals

Finally, we mention that many studies are designed to show either that two or more treatments or groups are similar or that no relationship exists between or among variables. In medical research, bioequivalence studies are designed to test whether a new therapy has nearly the same average efficacy as existing standard therapies, albeit the new one offers fewer side effects, lower costs, or more reliability across patients. Thus, the researcher is expecting to "accept" null hypotheses about efficacy but reject null hypotheses about side effects, costs, and reliability. Because appropriate confidence intervals or regions often play the central role in such analyses, the sample size analysis should ensure that those intervals or regions will be sufficiently small. Beal (1989) has developed a method and provided tables to handle situations involving the one-group and two-group t-test.

8.6 CONCLUSION

It is important and practical to perform sound, in-depth power analyses for proposed studies. Power analysis has many parallels with data analysis. Our common test statistics have non-null distributions that can be easily characterized using terms and formulas that are familiar to the data analyst. Any general method for data analysis should have a parallel general method for power analysis. This chapter shows how this holds for tests falling under the univariate and multivariate fixed-effects linear models, with Normal errors. We presented realistic examples for the two-group t-test, a matched-pairs t-test, the one-way

ANOVA with contrasts, an analysis of covariance, and a repeated-measures analysis. The concepts developed here apply to all of the special cases of the univariate and multivariate general linear models with fixed effects, and to many other methods. In particular, we treated the comparison of two proportions in some depth. We briefly discussed how it is possible to use existing data to get more objective estimates for the effect sizes. Usually, however, the number of subjects in most pilot studies is too small to make this effort "safe" and worthwhile.

Computing systems need to be developed so that researchers can conduct full sensitivity analyses of the power over a range of reasonable conjectures for the critical population parameters. Whenever possible, power analysis and data analysis should share the same software systems. Wright and O'Brien (1988) discussed how options in SAS PROC GLM could be developed to make it perform both steps of the two-step process used in Example 4. Some of these notions have now been implemented in JMP®, a Macintosh application from the The SAS Institute (1991).

References

Agresti, A. (1990). *Categorical Data Analysis.* John Wiley, New York.

Barton, C. N. and Cramer, E. C. (1989). Hypothesis Testing in Multivariate Linear Models with Randomly Missing Data. *Comm. Stat.—Simul. and Comp.,* **18**: 875-895.

Beal, S. L. (1989). Sample Size Determination for Confidence Intervals on the Population Means and on the Difference Between Two Population Means. *Biometrics,* **45**: 969-977.

Blanchard, E. B., Appelbaum, K. A., Radnitz, C. L., Morrill, B., Michultka, D., Kirsch, C., Guarnieri, P., Hillhouse, J., Evans, D. D., Jaccard, J., and Barron, K. D. (1990). A Controlled Evaluation of Thermal Biofeedback and Thermal Biofeedback Combined with Cognitive Therapy in the Treatment of Vascular Headache. *J. Cons. Clin. Psych.,* **2**: 216-224.

Casagrande, J. T., Pike, M. C., and Smith, P. G. (1978). An Improved Approximate Formula for Calculating Sample Sizes for Comparing Two Binomial Distributions. *Biometrics*, **34**: 483-486.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum, Hillsdale, New Jersey.

Cohen, J. (1992). A Power Primer. *Psych. Bull.,* **112**: 155-159.

Cohen, S., Tyrrell, D. A. J., and Smith, A. P. (1991). Psychological Stress and Susceptibility to the Common Cold. *N. Engl. J. Med.,* **325**: 606-612.

D'Agostino, R. B., Chase, W., and Belanger, A. (1988). The Appropriateness of Some Common Procedures for Testing the Equality of Two Independent Binomial Populations. *Amer. Statist.,* **42**: 198-202.

Diegert, C. and Diegert, K. V. (1981). Note on Inversion of Casagrande-Pike-Smith Approximate Sample-Size Formula for Fisher-Irwin Test on 2x2 Tables. *Biometrics*, **37**: 595.

Donner, A. (1984). Approaches to Sample Size Estimation in the Design of Clinical Trials-A Review. *Statis. Med.,* **3**: 199-214.

Dozier, W. G., and Muller, K. E. (1993). Small-Sample Power of Uncorrected and Satterthwaite Corrected t Tests for Comparing Two Proportions. *Comm. Stat.—Simul. and Comp.,* **22**: xxx-xxx.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions* (2nd ed). John Wiley, New York.

Fleiss, J. L., Tytun, A., and Ury, H. K. (1980). A Simple Approximation for Calculating Sample Sizes for Comparing Independent Proportions. *Biometrics,* **36**: 343-346.

Fleming, T. R., Harrington, D. P., and O'Brien, P. C. (1984). Designs for Group Sequential Tests. *Controlled Clinical Trials,* **5**: 348-361.

Freiman, J. A., Chalmers, T. C., Smith, H., and Kuebler, R. R. (1979). The Importance of Beta, the Type II Error, and Sample Size in the Design and Interpretation of the Randomized Clinical Trial. *N. Engl. J. Med.,* 299: 690-694. [Updated and reprinted as Chapter 19 in Bailar, J. C. III and Mosteller, F. (1992). *Medical Uses of Statistics.* NEJM Books, Boston.

Gatsonis, C. and Sampson, A. R. (1989). Multiple Correlation: Exact Power and Sample Size Calculations. *Psych. Bull.,* **106**: 516-524.

Geisser, S. and Greenhouse, S. W. (1958). An Extension of Box's Results on the Use of the F Distribution in Multivariate Analysis. *Ann. Math. Statist.,* **29**: 885-891.

Goldman, A. I., and Hillman, D. W. (1992). Exemplary Data: Sample Size and Power in the Design of Event-Time Clinical Trials. *Controlled Clinical Trials,* **13**: 256-271.

Goldstein, R. (1989). Power and Sample Size via MS/PC-DOS Computers. *Amer. Statist.,* **43**: 253-260.

Halpern, J. and Brown, B. W. (1987). Designing Clinical Trials with Arbitrary Specification of Survival Functions and for the Log Rank and Generalized Wilcoxon Test. *Controlled Clinical Trials,* **8**: 177-189.

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *J. Educ. Statist.,* **6**: 107-128.

Huynh, H. and Feldt, L. S. (1970). Conditions Under Which Mean Square Ratios in Repeated Measurements Have Exact F-Distribution. *J. Amer. Statist. Assoc.,* **65**: 1582-1589.

Jemmott III, J. B., Hellman, C., McClelland, D. C., Locke, S. E., Kraus, L., Williams, R. M., and Valeri, C. R. (1990). Motivational Syndromes Associated with Natural Killer Cell Activity. *J. Behav. Med.,* **13**: 53-73.

Kraemer, H., and Thiemann, S. (1987), *How Many Subjects?* Sage, Beverly Hills, California.

Lachin, J. M. (1981). Introduction to Sample Size Determination and Power Analysis for Clinical Trials. *Controlled Clinical Trials,* **2**: 93-113.

Maxwell, S. E. and Delaney, H. D. (1990). *Designing Experiments and Analyzing Data: A Model Comparison Perspective.* Wadsworth, Belmont, California.

Melton, K. (1986). A Procedure for Initiating Process Control. Unpublished Ph.D. dissertation, University of Tennessee, Knoxville.

Muller, K. E. and Barton, C. N. (1989). Approximate Power for Repeated Measures ANOVA Lacking Sphericity. *J. Amer. Statist. Assoc.,* **84**: 549-555.

Muller, K. E. and Barton, C. N. (1991). Correction to 'Approximate Power for Repeated Measures ANOVA Lacking Sphericity.' *J. Amer. Statist. Assoc.,* **86**: 255-256.

Muller, K. E., LaVange, L. M., Ramey, S. L., and Ramey, C. T. (1992). Power Calculations for General Linear Multivariate Models including Repeated Measures Applications. *J. of the Amer. Statist. Assoc.,* 87: 1209-1226.

Muller, K. E., and Peterson, B. L. (1984). Practical Methods for Computing Power in Testing the Multivariate General Linear Hypothesis. *Comp. Statist. Data Analy.,* **2**: 143-158.

O'Brien, R. G. (1979). A General ANOVA Method for Robust Tests of Additive Models for Variances. *J. Amer. Statist. Assoc.,* **74**: 877-881.

O'Brien, R. G. (1981). A Simple Test for Variance Effects in Experimental Designs. *Psych. Bull.,* **89**: 570-574.

O'Brien, R. G. (1986). Using the SAS System to Perform Power Analyses for Log-Linear Models. *Proceedings of the Eleventh SAS Users Group International Conference,* SAS Institute, Cary, North Carolina, pp. 778-784.

O'Brien, R. G. (1988). Review of PowerPack 2.22. *Amer. Statist.,* **42**: 266-270.

O'Brien, R.G. and Kaiser, M. K. (1985). MANOVA Method for Analyzing Repeated Measures Designs: An Extensive Primer. *Psych. Bull.,* **97**: 316-333.

O'Brien, R. G. and Shieh, G. (under review). Pragmatic, Unifying Algorithm Gives Power Probabilities for Common F Tests of the Multivariate General Linear Hypothesis.

Pocock, S.J. and Simon, R. (1975). Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trial. *Biometrics,* **31**: 103-115.

Prihoda, T. J. (1983). Convenient Power Analyses for Complex Analysis of Variance Models. *Proceedings of the ASA Statistical Computing Section,* pp. 267-271.

Rosenthal, R. and Rosnow, R. L. (1985). *Contrast Analysis: Focused Comparisons in the Analysis of Variance.* Cambridge University Press, Cambridge, U. K.

SAS Institute, Inc. (1990). *SAS® Language Reference, V. 6.* SAS Institute, Cary, North Carolina.

SAS Institute, Inc. (1991). *JMP® Users Guide, V. 2.0.* SAS Institute, Cary, North Carolina.

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

Seber, G. A. F. (1984). *Multivariate Observations.* John Wiley, New York.

Sedlmeier, P., and Gigerenzer, G. (1989). Do Studies of Statistical Power Have an Effect on the Power of Studies? *Psych. Bull.,* **105**: 309-316.

Shuster, J. J. (1990). *Handbook of Sample Size Guidelines for Clinical Trials.* CRC Press, Boca Raton, FL.

Stoney, C. M., Matthews, K. A., McDonald, R. H., and Johnson, C. A. (1988). Sex Differences in Lipid, Lipoprotein, Cardiovascular, and Neuroendocrine Responses to Acute Stress. *Psychophys,* **25**: 645-656.

Suissa, S. and Shuster, J. J. (1985). Exact Unconditional Sample Sizes for the 2×2 Comparative Trial. *J. Roy. Statist. Soc. A,***148**: 317-327.

Thomas, R. G. and Conlon, M. (1992). Sample Size Determination Based on Fisher's Exact Test for Use in 2 x 2 Comparative Trials with Small Event Rates. *Controlled Clinical Trials,* **13:** 134-147.

Upton, G. J. G. (1982). A Comparison of Alternative Tests for the 2×2 Comparative Trial. *J. Roy. Statist. Soc. A,* **145**: 86-105.

Venables, W. (1975). Calculation of Confidence Intervals for Noncentrality Parameters. *J. Roy. Statist. Soc. B,* **37**: 406-412.

Wright, S. P. and O'Brien, R. G. (1988). Power Analysis in an Enhanced GLM Procedure: What It Might Look Like. *Proceedings of the Thirteenth SAS Users Group International Conference,* SAS Institute, Cary, North Carolina, pp. 1097-1102.

Yates, F. (1934). Contingency Tables Involving Small Numbers and the χ^2 Test. *J. Roy. Stat. Soc. Supplement,* **1**: 217-235.

Appendix: Getting the Freeware
(updated July 1998)

All computations in this chapter where performed within The SAS® System using freeware modules (files containing SAS statements) that are processed via %INCLUDE statements in SAS input.

The OneWyPow, PowSetUp, and several PowTab modules handled all univariate methods (Examples 1-4 and 6). Their use and output is displayed in Listings 8.1-8.4 and 8.6. The modules run within the "base" SAS environment. All of this functionality, and far more, has been incorporated into UnifyPow (O'Brien, 1998), a freeware SAS module/macro that is easier to use than the older programs described above. The UnifyPow project continues.

The MVpower module performed the computations for Example 5. See Listing 8.5. MVpower is a module of PROC IML statements. PROC IML is not part of "base" SAS.

UnifyPow

Developer: Ralph G. O'Brien
Department of Biostatistics and Epidemiology/ P88
Cleveland Clinic Foundation
Cleveland, OH 44195
Voice: 216-445-9451; email: robrien@bio.ri.ccf.org
Homepage: http://www.bio.ri.ccf.org/Resume/Pages/robrien.html

Distribution site: http://www.bio.ri.ccf.org/power.html

UnifyPow is freeware. It runs within the "base" SAS environment and thus should run on any platform that runs the SAS System, including MS Windows, UNIX, Macintosh, IBM mainframes, and others. The files distributed contain all source statements and instructions on installation and use.

UnifyPow HAS NO WARRANTY WHATSOEVER. The developer would appreciate hearing about problems and suggestions for improvements regarding the software. He would also like to hear when they worked well for you. Please correspond via electronic mail (preferred) or telephone. Time rarely allows him to respond to "consulting" questions.

MVpower

Co-developers: Lynette L. Keyes

Keith E. Muller
Department of Biostatistics
CB #7400, School of Public Health
3105C McGavran-Greenberg Hall
The University of North Carolina
Chapel Hill, North Carolina 27599
Email: muller@bios.unc.edu
Website: http://www.bios.unc.edu/~muller

MVpower.sas is a module of SAS PROC IML statements that perform power analyses for the General Linear Multivariate Model and for the univariate approach to repeated measures, including the Geisser-Greenhouse and Huynh-Feldt solutions. The power approximation methods of Muller and Peterson (1984) are used for the general multivariate case. Also, at the time of this printing, Muller was planning to soon include the algorithm described in O'Brien and Shieh (under review). Exact results are provided for all univariate models and for some multivariate models. The methods of Muller and Barton (1989, 1991) undergird the multivariate and univariate approaches to the power approximations for repeated measures. Key restrictions include the assumptions of Gaussian (Normal) errors, fixed predictor values, a common design for all responses, and no missing data.

MVpower handles a very broad class of univariate and multivariate ANOVA and regression problems, because users specify directly the **X**, **B**, **Σ**, **C**, and **U** matrices that form the models and hypotheses outlined in Section 8.4. All matrices must be full rank. Little or no experience with PROC IML is required to use MVpower. PROC IML is a supplement to the "basic" distribution of SAS. If it is not yet available to you, you may order it from the SAS Institute.

As of July 1998, this freeware was available at Dr. Muller's website given above.

Reference

O'Brien RG (1998), "A Tour of UnifyPow: A SAS Module/Macro for Sample-Size Analysis," Proceedings of the 23nd SAS Users Group International Conference, Cary, NC, SAS Institute, 1346-1355. [This too is available from the UnifyPow website.]