Appendix 5

Computing the Power of Statistical Tests

There are two types of errors that one can make when performing a statistical test. A **false positive** (a **Type I error**) occurs when the null hypothesis is rejected when in fact it is correct. We control for this by setting the significance level α of a test (the probability of a false positive) to be small. The other source of error is a **false negative** (a **Type II error**), *failing* to reject the null hypothesis when in fact it is false. The **power** of a test is defined to be the probability that the null hypothesis is rejected when it is indeed false. Hence if β is the probability of a false negative, the power is $1 - \beta$ (Figure A5.1).

Before embarking on a potentially very costly experiment, the investigator would like to be certain that the design ensures sufficiently high power given the objectives of the proposed study. Indeed, many experiments that report negative results may in fact have significant biological effects that are swamped out by the high sampling error generated by insufficient sample size. Power depends not

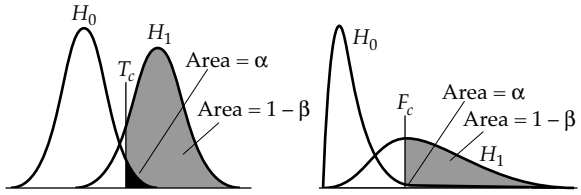

Figure A5.1 Power $(1-\beta)$ and significance (α) of a normally (**left**) and F-distributed test statistics (**right**). The distributions of the test statistic under the null hypothesis (H_0) and at a particular parameter value under the alternative hypothesis (H_1) are shown. Under the null hypothesis, the probability that the test statistic exceeds a critical value (T_c, F_c) is α, and we define this as the significance of the test. The power of the test for a particular value of the alternative hypothesis is the probability that the test statistic exceeds the critical value, and we denote this by $1 - \beta$. Hence, β is the probability that the test statistic is below the critical value (a false negative).

only on sample size and the actual values of the unknown distribution parameters being estimated, but also on the assumed level of significance α of the test. Here we consider some of the basics of computing the power of a given design for tests whose statistics are normally and F distributed. Both appear often throughout this book.

POWER OF NORMALLY DISTRIBUTED TEST STATISTICS

Assume the test statistic T is normally distributed under the null hypothesis, with $T \sim \mathrm{N}(\mu_0, \sigma_0^2)$. For power calculations, assume the null hypothesis is false and that the test statistic actually has mean μ_1 and variance σ_1^2 but remains normally distributed, so that $T \sim \mathrm{N}(\mu_1, \sigma_1^2)$. To obtain expressions for the required sample size for a given power, it is convenient to first write the sample variance as a function of sample size n, with $\sigma_i^2 = f_i^2/n$ for $i = 0, 1$. (Often $\sigma_0^2 = \sigma_1^2$, in which case $f_0 = f_1$.) Finally, let $z_{(\alpha)}$ satisfy

$$\Pr(U \leq z_{(\alpha)}) = \alpha \tag{A5.1a}$$

where U is a unit normal random variable. For example, $\Pr(U \leq 1.65) = 0.95$, so that $z_{(0.95)} = 1.65$. Two identities involving z_α will prove useful in the following discussions. First, note from Figure A5.2 that

$$\Pr(U > z_{(1-\alpha)}) = \alpha \tag{A5.1b}$$

so that, for example, $\Pr(U > 1.65) = 0.05$. Second, from the symmetry of the normal distribution it can be shown that

$$z_{(\alpha)} = -z_{(1-\alpha)} \tag{A5.1c}$$

We now have the all the necessary definitions in hand to consider the power of normally distributed tests. Hypothesis tests generally fall into two categories: **one-sided** and **two-sided**, and we examine these in turn before considering specific applications.

One-sided Tests

Some hypotheses are naturally one-sided. For example, we may wish to test whether heritability h^2 is significantly different from zero. In this case, true values of $h^2 < 0$ cannot occur, so that the alternative to the null hypothesis $h^2 = 0$ is $h^2 > 0$.

Consider the null hypothesis that the test statistic has mean $\mu = \mu_o$ versus the alternative hypothesis that $\mu = \mu_1$, where (for example) $\mu_1 > \mu_0$. Suppose that

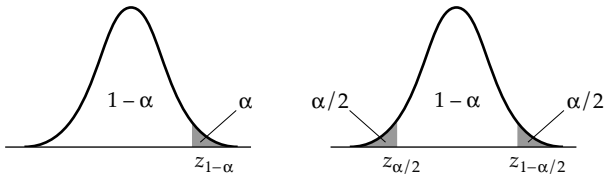

Figure A5.2 Areas under the normal curve for one-sided (left) and two-sided (right) tests of significance at level α.

under the null hypothesis, the test statistic $T \sim \mathrm{N}(\mu_o, \sigma_o^2)$, so that $(T - \mu_o)/\sigma_o$ is distributed as a unit normal, implying from Equation A5.1b that

$$\Pr\left(\frac{T-\mu_o}{\sigma_o} > z_{(1-\alpha)}\right) = \Pr\left(T > \mu_o + \sigma_o\, z_{(1-\alpha)}\right) = \alpha \tag{A5.2a}$$

Thus, comparing the observed statistic T with the α-level critical value, $T_c(\alpha)$, where

$$T_c(\alpha) = \mu_o + \sigma_o\, z_{(1-\alpha)} \tag{A5.2b}$$

gives the one-sided test of significance with probability α of a false positive. We reject the null hypothesis when $T > T_c(\alpha)$.

Suppose that the alternative hypothesis is correct, so that $T \sim \mathrm{N}(\mu_1, \sigma^2)$. What is the probability that the resulting test statistic is significant (and hence a false negative is avoided)? Since it is now $(T - \mu_1)/\sigma_1$ that follows a unit normal, the probability that the test statistic exceeds the critical value $T_c(\alpha)$ given by Equation A5.2b is

$$\begin{aligned}\Pr\left(T > T_c(\alpha)\right) &= \Pr\left(\frac{T-\mu_1}{\sigma_1} > \frac{T_c(\alpha)-\mu_1}{\sigma_1}\right) \\ &= \Pr\left(U > \frac{\mu_o-\mu_1}{\sigma_1} + z_{(1-\alpha)}\,\frac{\sigma_o}{\sigma_1}\right)\end{aligned} \tag{A5.3}$$

where U is a unit normal random variable. Noting that $\Pr(U > z_{(\beta)}) = 1 - \beta$ (Equation A5.1b), from Equation A5.3 it immediately follows that for the test to have power $1 - \beta$ requires

$$\frac{\mu_o-\mu_1}{\sigma_1} + z_{(1-\alpha)}\,\frac{\sigma_o}{\sigma_1} = z_{(\beta)}$$

Using our notation for writing the variance of the test statistic as a function f divided by the sample size n (i.e., $\sigma_i^2 = f_i^2/n$), this becomes

$$\sqrt{n}\,\frac{(\mu_o-\mu_1)}{f_1} + z_{(1-\alpha)}\,\frac{f_o}{f_1} = z_{(\beta)}$$

Since $-z_{(\beta)} = z_{(1-\beta)}$ (Equation A5.1c), this rearranges to give

$$n = \frac{f_1^2}{(\mu_1 - \mu_o)^2} \left(z_{(1-\alpha)} \frac{f_o}{f_1} + z_{(1-\beta)} \right)^2 \tag{A5.4a}$$

$$= \left(\frac{z_{(1-\beta)} f_1 + z_{(1-\alpha)} f_0}{\mu_1 - \mu_0} \right)^2 \tag{A5.4b}$$

$$\simeq \frac{f_1^2}{(\mu_1 - \mu_o)^2} \left(z_{(1-\alpha)} + z_{(1-\beta)} \right)^2 \tag{A5.4c}$$

Equation A5.4c holds when the sample variances under the null hypothesis and true parameter values are approximately equal ($f_0 \simeq f_1$), as is often the case.

Two-sided Tests

Other hypotheses are naturally **two-sided**. Here, the null hypothesis is $\mu = \mu_o$ versus the alternative of $\mu \neq \mu_o$. A glance at Figure A5.2 shows that under the null hypothesis,

$$\Pr\left(z_{(\alpha/2)} \leq \frac{T - \mu_o}{\sigma_o} \leq z_{(1-\alpha/2)} \right) = 1 - \alpha \tag{A5.5a}$$

Again recalling Equation A5.1c, this rearranges to become

$$\Pr\left(-\sigma_o \, z_{(1-\alpha/2)} \leq T - \mu_o \leq \sigma_o \, z_{(1-\alpha/2)} \right) = 1 - \alpha$$

giving

$$| \, T - \mu_o \, | > \sigma_o \, z_{(1-\alpha/2)} \tag{A5.5b}$$

as a two-sided test of significance with probability α of a false positive under the null hypothesis. Under the alternative hypothesis that $T \sim \mathrm{N}(\mu_1, \sigma_1^2)$, the power of this test is

$$\Pr[\, T - \mu_0 < -\sigma_o \, z_{(1-\alpha/2)} \,] + \Pr[\, T - \mu_0 > \sigma_o \, z_{(1-\alpha/2)} \,]$$

Since $(T - \mu_1)/\sigma_1 \sim U$, this is equivalent to

$$\Pr\left(U < \frac{\mu_o - \mu_1}{\sigma_1} - \frac{\sigma_o}{\sigma_1} \, z_{(1-\alpha/2)} \right) + \Pr\left(U > \frac{\mu_o - \mu_1}{\sigma_1} + \frac{\sigma_o}{\sigma_1} \, z_{(1-\alpha/2)} \right) \tag{A5.6}$$

Using the same logic leading to Equation A5.4, the sample size n required to give power $1 - \beta$ is found to be

$$n = \frac{f_1^2}{(\mu_o - \mu_1)^2} \left(z_{(1-\alpha/2)} \frac{f_o}{f_1} + z_{(1-\beta)} \right)^2 \tag{A5.7a}$$

$$\simeq \frac{f_1^2}{(\mu_o - \mu_1)^2} \left(z_{(1-\alpha/2)} + z_{(1-\beta)} \right)^2 \tag{A5.7b}$$

Thus, the two-sided expressions are identical to the one-sided expressions with α replaced by $\alpha/2$.

One popular approach when considering possible experimental designs is to compute the sample size required to give the expected confidence interval a preset length, so that the interval given by $\mu_1 \pm z_\alpha f_1/\sqrt{n}$ does not include μ_0. Here z_α is the appropriate value to give an α-level one- or two-sided test (depending on the hypothesis being tested). Solving for n,

$$(\mu_1 - \mu_o)^2 \geq \frac{z_\alpha^2 f_1^2}{n}, \qquad \text{or} \qquad n \geq \frac{z_\alpha^2 f_1^2}{(\mu_1 - \mu_o)^2}$$

This approach is often used in place of a direct power calculation, but just what is its actual power? From Equations A5.4c and A5.7b, this expression yields the n which gives a test whose power satisfies $z_{(1-\beta)} = 0$ or $\beta = 0.5$. Thus, designs using this sample size have only 50% power, i.e., they are expected to detect an effect only half the time.

POWER OF *F*-RATIO TESTS

The analysis of variance (ANOVA) is widely used throughout this book, e.g., in estimating variance components under balanced experimental designs (Chapters 18–24) and in detecting QTLs via marker-trait associations (Chapters 14–16). ANOVA designs are typically based upon sums of squares (SS_x) and their associated mean squares ($MS_x = SS_x/n_x$, with n_x the associated degrees of freedom). ANOVA test statistics are generally given by ratios of mean squares, $F = MS_x/MS_y$. Under normality assumptions, sums of squares are χ^2-distributed, while under the null hypothesis F is distributed as a (central) F-ratio distribution with n_x (numerator) and n_y (denominator) degrees of freedom, and we denote this as $F \sim F_{n_x,n_y}$. Thus, a test of level α for the hypothesis of no additional effects on level x (compared to level y) is whether $F > F_{n_x,n_y,[1-\alpha]}$, where $F_{n_x,n_y,[\alpha]}$ satisfies

$$\Pr\left[F_{n_x,n_y} \leq F_{n_x,n_y,[\alpha]} \right] = \alpha \tag{A5.13a}$$

Two useful identities related to this definition are

$$\Pr\left[F_{n_x,n_y} > F_{n_x,n_y,[1-\alpha]} \right] = \alpha \tag{A5.13b}$$

and

$$F_{n_x,n_y,[\alpha]} = \frac{1}{F_{n_y,n_x,[1-\alpha]}} \tag{A5.13c}$$

The power of the test is the probability that the test statistic F exceeds this critical value,

$$\Pr\left[F > F_{n_x,n_y,[1-\alpha]} \right] \tag{A5.13d}$$

In order to compute this probability, we require the distribution of the test statistic F under the alternative hypothesis. Here, some sums of squares follow a noncentral χ^2 distribution, and the resulting F statistic involving these follows a noncentral F distribution. Hence, before considering power, we describe the properties of central and noncentral χ^2 and F distributions, proofs of which can be found in standard texts such as Scheffé (1959), Searle (1971), and Johnson and Kotz (1970b).

Power of Fixed-effects ANOVA Designs

We now have the necessary machinery in hand to perform power calculations for ANOVA designs. We consider fixed-effects designs first, using the simple one-way ANOVA to illustrate the basic approach. Letting y_{ij} denote the value of the jth individual with treatment i, the model is $y_{ij} = \mu + \tau_i + e_{ij}$. We assume N fixed factors (the treatment effects $\tau_1, \cdots, \tau_N$); that the residuals are independent and normally distributed, with $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$; and that the design is balanced with n observations for each treatment. As discussed in Chapter 18, the resulting treatment (or model) and error sums of squares are given by

$$\mathrm{SS}_t = n \sum_{i=1}^{N} (\,\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2, \qquad \mathrm{SS}_e = \sum_{j=1}^{n} \sum_{i=1}^{N} (y_{ij} - \overline{y}_{i\cdot})^2 \tag{A5.18a}$$

where $\overline{y}_{i\cdot}$ is the sample mean for factor i and $\overline{y}_{\cdot\cdot}$ is the overall sample mean. As we will see shortly, these sums of squares have degrees of freedom $N - 1$ and $N(n-1)$, respectively. The test for a significant treatment effect (at least one $\tau_i \neq 0$) is the F-ratio statistic

$$F = \frac{\mathrm{MS}_t}{\mathrm{MS}_e} = \frac{\mathrm{SS}_t/(N-1)}{\mathrm{SS}_e/N(n-1)} \tag{A5.18b}$$

The distribution of F depends on the distribution of the sums of squares. Since $\overline{y}_{i\cdot} \sim \mathrm{N}(\tau_i, \sigma_e^2/n)$, Equation A5.15e gives

$$\mathrm{SS}_t \sim n \cdot (\sigma_e^2/n) \cdot \chi^2_{N-1,\lambda} \qquad \text{where} \qquad \lambda = \sum_{i=1}^{N} \frac{{\tau_i}^2}{\sigma_e^2/n} \tag{A5.19}$$

The noncentrality parameter λ can also be expressed in terms of amount of the total variance attributable to model (treatment) effects, σ_τ^2. Since $\overline{\tau} = 0$ (by construction), it follows that

$$\sigma_\tau^2 = \frac{1}{N-1} \sum_{i=1}^{N} {\tau_i}^2 \tag{A5.20a}$$

and hence,

$$\lambda = n\,(N-1)\,\frac{\sigma_\tau^2}{\sigma_e^2} \tag{A5.20b}$$

Finally, noting that $y_{ij} - \overline{y}_{i\cdot} = e_{ij} - \overline{e}_{i\cdot}$ and that $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$, we have

$$\mathrm{SS}_e = \sum_{i=1}^{N} \left[\sum_{j=1}^{n} (e_{ij} - \overline{e}_{i\cdot})^2 \right] \sim \sum_{i=1}^{N} \sigma_e^2 \cdot \chi_{n-1}^2 \sim \sigma_e^2 \cdot \chi_{N(n-1)}^2 \tag{A5.21}$$

where the last two identities follow from Equations A5.14c and A5.14a, respectively. Thus the F-ratio test statistic is distributed as

$$F \sim \left(\frac{\sigma_e^2}{\sigma_e^2} \right) \left[\frac{\chi_{N-1,\lambda}^2/(N-1)}{\chi_{N(n-1)}^2/N(n-1)} \right] \sim F_{N-1,N(n-1),\lambda} \tag{A5.22}$$

where the noncentrality parameter is given by Equation A5.20b.

Under the hypothesis of no treatment effects, $\lambda = 0$ and $F \sim F_{(N-1),N(n-1)}$. A test with significance level α is thus given by whether $f > F_{(N-1),N(n-1),[1-\alpha]}$. When at least one treatment effect is nonzero, the test statistic f follows a noncentral F distribution, and the power of the test is given by

$$\Pr\left[F_{N-1,N(n-1),\lambda} > F_{N-1,N(n-1),[1-\alpha]} \right] \tag{A5.23}$$

To find the sample size required to give this test power $1 - \beta$, we equate Equation A5.23 to $1 - \beta$ and solve for n given the fixed values of α, β, N, and $\sum \tau_i^2/\sigma_e^2$ (or, equivalently, σ_τ^2/σ_e^2).

Example 3. Consider a fixed-effect design with four factors ($N = 4$), and suppose that $\sigma_\tau^2/\sigma_e^2 = 1/3$, so that the treatment effects account for 25% of the total variance ($\sigma_\tau^2 + \sigma_e^2$). If each treatment has a sample size of $n = 5$, what is the power to detect a significant treatment effect using a test of $\alpha = 0.05$? Here $N(n-1) = 16$, and we find from *F* distribution tables that

$$\Pr(F_{3,16} > 3.24) = 0.05$$

so that the critical value for the test is $F_{3,16,[0.95]} = 3.24$. Likewise, the noncentrality parameter is

$$\lambda = (N-1)\, n\, (\sigma_\tau^2/\sigma_e^2) = 3 \cdot 5 \cdot (1/3) = 5$$

and from noncentral *F* tables, the power is found to be

$$\Pr(F_{3,16,5} > 3.24) = 0.353$$

In the absence of noncentral *F* tables or programs, one could instead use the normal approximation given by Equation A5.17b. Just how good is this? Here $F_o = 3.24$, and substituting the degrees of freedom and $\lambda = 5$ yields $A = 1.625$ and $B = 0.6075$, giving

$$z_o = \frac{\sqrt{31 \cdot 0.6075} - \sqrt{2(3+5) - 1.625}}{\sqrt{1.625 + 0.6075}} = .3669$$

and hence

$$\Pr(F_{3,16,5} > 3.24) \simeq \Pr(U > 0.3669) = 0.357$$

showing that this approximation works quite well.

What sample size n is required to give this test 90% power? Here $\alpha = 0.05$, $N = 4$, and $\sigma_\tau^2/\sigma_e^2 = 1/3$ are fixed values. For $n = 15$, we find that $N(n-1) = 56$, $\lambda = 15$, and (from central *F* tables) $F_{3,56,[0.95]} = 2.769$. Substituting these values into Equation A5.23 and using noncentral *F* tables gives

$$\Pr(F_{3,56,15} > 2.769) = 0.8957$$

showing that a test with this sample size has a power of 89.6%. Increasing *n* to 16, $N(n-1) = 60$, $\lambda = 16$, and $F_{3,60,[0.95]} = 2.758$, giving

$$\Pr(F_{3,60,16} > 2.758) = 0.9167$$

for a power of 91.7%.

Power of a Random-effects ANOVA Design

Under a random-effects model, the simple one-way ANOVA model becomes $y_{ij} = \mu + t_i + e_{ij}$, where t is now a random variable drawn from a normal distribution with variance σ_t^2, so that $t_i \sim \mathrm{N}(0, \sigma_t^2)$. Residuals are assumed independent with $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$ as before. Assuming a balanced design with n replicates for each of the N t_i values, the treatment and error sums of squares are also given by Equation A5.18a. As with the fixed-effects model, $\mathrm{SS}_e \sim \sigma_e^2 \cdot \chi^2_{N(n-1)}$. However, the distribution of the treatment sum of squares remains a central χ^2, even when $\sigma_t^2 > 0$. To see this, note that $\overline{y}_{i\cdot} \sim \mathrm{N}(0, \sigma^2)$ with $\sigma^2 = \sigma_t^2 + \sigma_e^2/n$. Applying Equation A5.14c, the same logic leading to Equation A5.19 now gives

$$\mathrm{SS}_t \sim n \cdot \sigma^2 \cdot \chi^2_{N-1} = (n\,\sigma_t^2 + \sigma_e^2) \cdot \chi^2_{N-1} \tag{A5.28}$$

Hence, the F-ratio statistic is distributed as

$$\begin{aligned} F = \frac{\mathrm{SS}_t/(N-1)}{\mathrm{SS}_e/N(n-1)} &\sim \frac{(n\,\sigma_t^2 + \sigma_e^2) \cdot \chi^2_{N-1}/(N-1)}{\sigma_e^2 \cdot \chi^2_{N(n-1)}/N(n-1)} \\ &\sim \left(1 + n\,\frac{\sigma_t^2}{\sigma_e^2}\right) F_{(N-1),N(n-1)} \end{aligned} \tag{A5.29}$$

Note that the term in parentheses is just $E(\mathrm{MS}_t)/E(\mathrm{MS}_e)$. Hence

$$\frac{F}{1 + n\,(\sigma_t^2/\sigma_e^2)} \sim F_{N-1,N(n-1)} \tag{A5.30a}$$

giving the power of this test as

$$\Pr\left[F_{N-1,N(n-1)} > \frac{F_{N-1,N(n-1),[1-\alpha]}}{1 + n\,(\sigma_t^2/\sigma_e^2)} \right] \tag{A5.30b}$$

Since, by definition (Equation A5.13b),

$$\Pr\left[\, F_{N-1,N(n-1)} > F_{N-1,N(n-1),[\beta]} \,\right] = 1 - \beta$$

to have power $1 - \beta$ requires that

$$F_{N-1,N(n-1),[1-\alpha]} = \left(1 + n\,\frac{\sigma_t^2}{\sigma_e^2}\right) F_{N-1,N(n-1),[\beta]} \tag{A5.31}$$

Under more general random-effect models, if the test statistic is $F = \text{MS}_x/\text{MS}_y$, then

$$\frac{F}{R} \sim F_{n_x,n_y}, \qquad \text{where} \qquad R = \frac{E(\text{MS}_x)}{E(\text{MS}_y)} \tag{A5.32a}$$

and to obtain power $1 - \beta$ requires choosing sample sizes such that

$$F_{n_x,n_y,[1-\alpha]} = \left[\,\frac{E(\text{MS}_x)}{E(\text{MS}_y)}\,\right] F_{n_x,n_y,[\beta]} \tag{A5.32b}$$

Example 5. Consider a random-effects model with design parameters similar to Example 3, with $\sigma_t^2/\sigma_e^2 = 1/3$, $N = 4$, and $n = 5$. What is the power of this design for a test with $\alpha = 0.05$? Here $F_{3,16,[0.95]} = 3.24$ and $1 + n\,(\sigma_t^2/\sigma_e^2) = 1 + 5/3 = 8/3$. From Equation A5.30b, the power becomes

$$\Pr\left(F_{3,16} > \frac{3.24}{8/3} = 1.215\right) = 0.103$$

What value of n is required to give 90% power? With increasing values of n, we find that for $n = 38$, $N(n-1) = 148$, $1 + n\,(\sigma_t^2/\sigma_e^2) = 13.67$, and $F_{3,148,[0.95]} = 2.66$, giving

$$\Pr\left(F_{3,148} > \frac{2.66}{13.67} = 0.195\right) = 0.90$$
