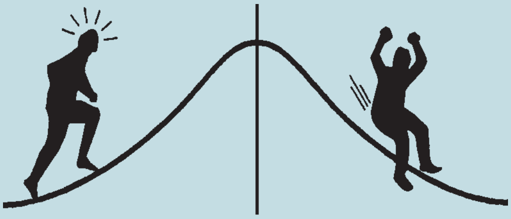


BIOMETRICS
INFORMATION

HANDBOOK NO. 2 OCTOBER 1991

Power Analysis Handbook for the Design and Analysis of Forestry Trials

Biometrics Information Handbook Series

BC Ministry of Forests

Power Analysis Handbook for the Design and Analysis of Forestry Trials

by
Amanda F. Linnell Nemec
International Statistics and Research Corporation
P.O. Box 496
Brentwood Bay, B.C.
V0S 1A0

October 1991

Series Editor
Wendy Bergerud
Forest Science Research Branch
B.C. Ministry of Forests
31 Bastion Square
Victoria, B.C.
V8W 3E7

BC
Ministry of Forests

Canadian Cataloguing in Publication Data

Nemec, Amanda F. Linnell (Amanda Frances Linnell)
Power analysis handbook for the design and analysis of forestry trials

(Biometrics information handbook series, ISSN 1183-9759 ; no. 2)

Includes bibliographical references: p.
ISBN 0-7718-9112-1

1. Forests and forestry - Statistical methods. 2. Forests and forestry - Research. I. Bergerud, W. A. (Wendy Ann), 1954- . II. British Columbia Ministry of Forests. III. Title. IV. Series.

SD387.S73N45 1991 634.9072 C91-092299-3

Published by the
Forest Science Research Branch
Ministry of Forests
31 Bastion Square
Victoria, B.C. V8W 3E7

Copies of this and other Ministry of Forests titles
are available from Crown Publications Inc.,
546 Yates Street, Victoria, B.C. V8W 1K8.

TABLE OF CONTENTS

TABLES

FIGURES

1 INTRODUCTION

In the past, the emphasis of statistics in forestry, and other applied fields, has been on an assessment of statistical significance, or the probability that the null hypothesis will be rejected when it is true (i.e., the probability of committing a ‘‘Type I error’’). However, there is growing awareness (e.g., see Peterman 1990a, 1990b and Toft and Shea 1983) that researchers should also be concerned with the possibility that statistical methods may fail to reject a false null hypothesis (i.e., a ‘‘Type II error’’ might be committed). The statistical theory and methods by which this important issue can be examined are referred to as power analysis.

This handbook is intended as an introduction to power analysis. It contains basic definitions and a review of power theory for t-tests and ANOVA F-tests, both of which are widely used in the analysis of forestry data. Examples, with step-by-step instructions and SAS programs for performing the necessary calculations, are provided. The handbook also contains a discussion of the two primary applications of power analysis: experimental design (e.g., sample size calculations); and the interpretation of the results of statistical analyses using *post hoc* power analysis. Although this handbook does not cover tests for categorical data (e.g., tests for proportions or log-linear methods), the same general principles apply to those methods as well. A detailed exposition of power analysis for a broad collection of tests can be found in Cohen (1977).

2 POWER ANALYSIS

2.1 What is a Power Analysis?

The statistical analysis of forestry data usually involves a test of some (null) hypothesis that is central to the investigation. For example, in an assessment of the regeneration performance of a plantation, a one-sample t-test might be used to test the assumption that the plantation meets a required height standard. In another instance, a one-way analysis of variance F-test might be used to test an *a priori* belief that the diameter growth of a certain species of tree is the same regardless of which of several herbicides is applied.

Since experimental data are invariably subject to random error, there is always some uncertainty about any decision to reject or retain a null hypothesis on the basis of a statistical test. There are two reasons for this uncertainty. First is the possibility that the data might, by chance, be so unusual that the null hypothesis is rejected even though it is true. For example, a plantation that is performing satisfactorily might be declared sub-standard simply because the sample happened to include a disproportionate number of small trees. Fortunately, this uncertainty is readily quantified by calculating a P value or by quoting the significance level of the test.

The other source of uncertainty is often not even considered. It is the possibility that, given the available data, a statistical test may fail to reject a false null hypothesis. For example, the inherent variability between the treatment plots in a herbicide trial might be so large that, unless the number of replications is large, a real and potentially important difference between the herbicides might not be detected. In contrast to the first, this second possibility is more difficult to quantify since it requires an investigation of the power of the test.

The power of a statistical test is the probability that, when the null hypothesis is false, the test will reject that hypothesis. A powerful test is one that has a high success rate in detecting even small departures from the null hypothesis. In general, the power of a test depends on the adopted level of significance, the inherent variability of the data, the degree to which the true state of nature departs from the null hypothesis, and the sample size. Computation of this probability for one or more combinations of these factors is referred to as a power analysis.

Power is obviously an important consideration in the design of forestry trials. Among other things, power calculations can help ensure that the sample size is large enough that the smallest effect that is of biological or economic importance will be detected with a reasonable degree of certainty. A power analysis can also be helpful in interpreting the results of a statistical analysis. For example, if a power calculation shows that the sample size is so small that only very large effects were expected to be identified in the first place, then failure to reject the null hypothesis does not provide very convincing evidence that the null hypothesis is true. On the other hand, if the test is known to be sensitive to small departures from the null hypothesis, a non-significant result might reasonably be interpreted as confirmation of the null hypothesis.

2.2 Basic Definitions and Theory

To carry out a power analysis, it is necessary to have a good understanding of the basic concepts of classical hypothesis testing. In any hypothesis testing situation, two competing hypotheses must be considered: **the null hypothesis**, **H_0**, and the **alternative hypothesis**, **H_a**. The null hypothesis is the one that is retained unless the data provide convincing evidence to the contrary (i.e., in support of H_a). For this reason, H_0 often represents the conservative point of view. For example, in a trial to assess the effectiveness of a new fertilizer, the null hypothesis might be the hypothesis that there is no difference in the height growth of trees receiving the fertilizer and the growth of a control group of untreated but otherwise comparable trees. Only if there is convincing evidence that the null hypothesis is false will it be concluded that the fertilizer has some effect.

The alternative hypothesis encompasses all departures from the null hypothesis that are of interest to the researcher. In the previous fertilizer example, H_a might be the **one-sided hypothesis** that the expected growth of the fertilizer group is greater than that of the control. However, if the possibility that the fertilizer might actually inhibit growth is also of concern, it might be more appropriate to adopt as the alternative the **two-sided hypothesis** that the difference between the means of the two groups is either positive or negative.

After the data have been collected, the strength of the evidence against H_0 and in favour of H_a is assessed with a statistical test, such as a t-test, F-test, or chi-squared test. In classical hypothesis testing, the null hypothesis is either rejected or not rejected depending on whether the observed value of a suitable **test statistic** (e.g., sample mean, F-ratio, etc.) falls into a predetermined **rejection region**. The rejection region corresponds to those values of the test statistic that are so extreme that they are unlikely to have occurred by chance under H_0, but are not unexpected if H_a is true. Depending on H_a, this will usually correspond to either a one-sided or two-sided region.

Whenever a null hypothesis is tested, the decision to reject or not to reject H_0 is either correct or incorrect. If the decision is incorrect, then one of two possible errors has been committed, namely, a **Type I error** or a **Type II error**. A Type I error is the rejection of a null hypothesis that is actually true, and a Type II error is the failure to reject a false null hypothesis.

When H_0 is true, the correct decision is to retain H_0 and the only possible error is a Type I error. The maximum probability of committing a Type I error is denoted α and is referred to as the **significance level** of the test. The corresponding probability of making the correct decision is $1-\alpha$, which is sometimes called the **confidence level**. A researcher is usually free to choose α and therefore can control the probability of making a Type I error. This is particularly advantageous when the Type I error is more serious than the Type II error.

When H_a is true, the two relevant probabilities are the probability of committing a Type II error, which is usually denoted β, and the probability of making the correct decision to reject H_0 when it is false. The latter is $1-\beta$ and is defined to be the **power** of the test. The two types of errors and the two correct decisions that can be made in any hypothesis testing situation are summarized in Table 1, along with their associated probabilities.

2.3 Power Calculations

Computation of both power and the significance level of a test requires an evaluation of the probability that the null hypothesis will be rejected. The only difference between the two calculations is that power is computed under the assumption that H_a is true, while the significance level is computed under the assumption that H_0 is true. In general, the former calculation is more difficult than the latter.

TABLE 1. Errors in hypothesis testing

	Researcher's decision	
State of nature	**Reject H_0**	**Do not reject H_0**
H_0 true	Type I error Probability = α (Significance level)	Correct decision Probability = $1-\alpha$ (Confidence level)
H_0 false (H_a true)	Correct decision Probability = $1-\beta$ (Power)	Type II error Probability = β

To illustrate the relationship between power and significance level, consider a hypothetical regeneration performance assessment survey, in which the objective is to determine whether or not a plantation satisfies the requirement that the trees have a mean height of at least $\mu_0 = 200$ cm. The null hypothesis to be tested is H_0: $\mu \geq \mu_0$ (the plantation shows satisfactory performance) versus the one-sided alternative H_a: $\mu < \mu_0$ (the plantation shows sub-standard performance). In the calculations that follow, it will be assumed that the heights of the trees are normally distributed with unknown mean μ and known standard deviation $\sigma = 80$ cm. (In practice, σ is unknown and must be estimated from the data. This affects the computational details but does not change the general conclusions that may be drawn from this example.)

To test H_0, suppose a simple random sample of $n = 50$ trees is selected from the plantation and the heights of the trees are measured. Let $\bar{y}$ be the sample mean and let the test statistic z be defined as follows:

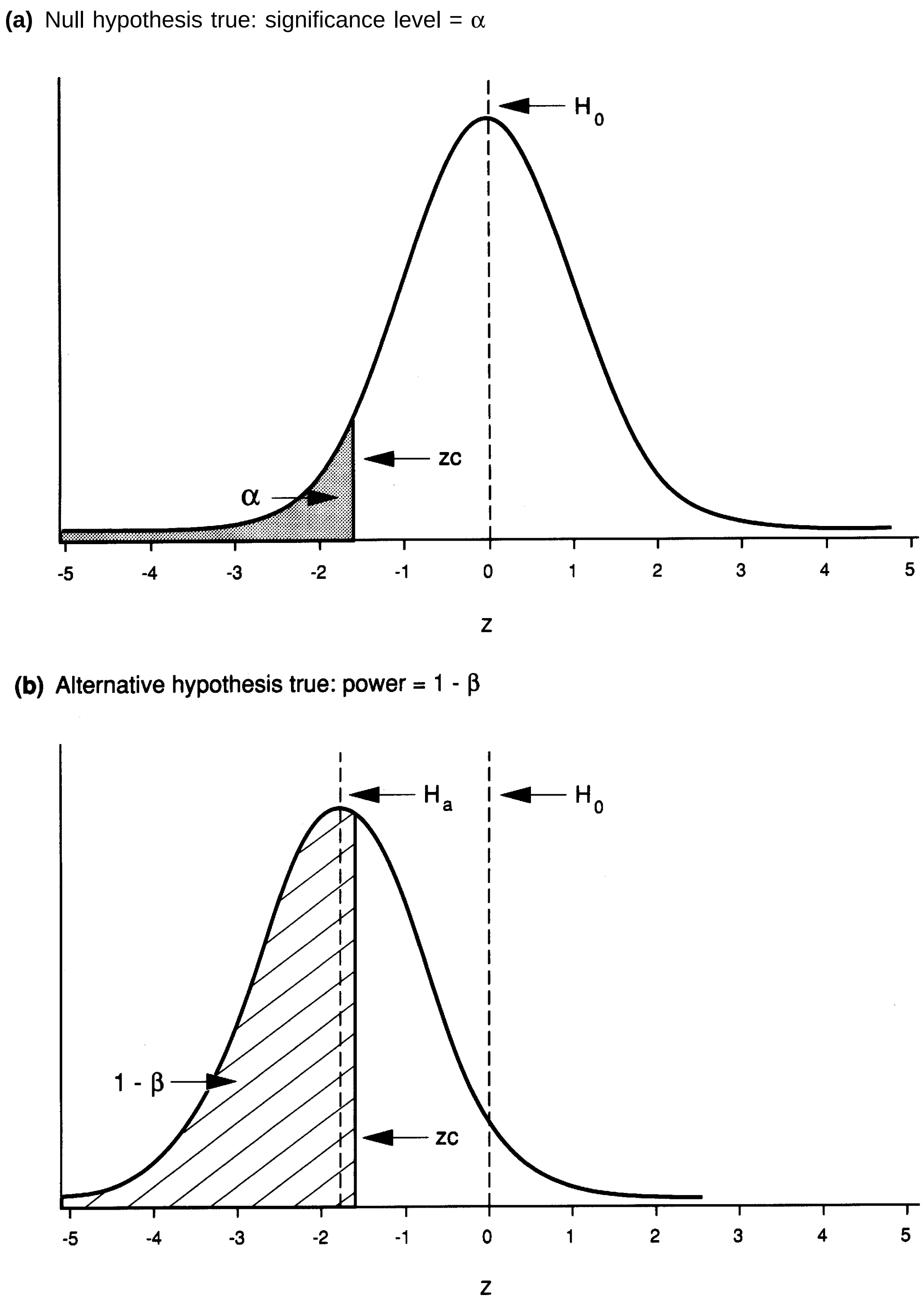
$$z = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$$

Since the distribution of the heights of the individual trees is assumed to be normal, $\bar{y}$ has a normal distribution with mean μ and standard error (i.e., standard deviation) $\sigma/\sqrt{n}$. Therefore, if $\mu = \mu_0$ (H_0 is true), z has a standard normal distribution (i.e., normal distribution with a mean of 0 and a standard deviation of 1) as shown in Figure 1a (the vertical line locates the mean of the distribution when H_0 is true). On the other hand, if $\mu = \mu_a < \mu_0$ (H_a is true), the mean of z is shifted to the left as shown in Figure 1b. Thus, H_0 should be rejected in favour of H_a whenever z falls into the one-sided rejection region $z \leq zc$, where zc is a **critical value** to be determined.

The significance level of a test is computed under the assumption that H_0 is true. Therefore, the significance level of the preceding test is the probability that $z \leq zc$, given that $\mu = \mu_0$, which will be written as follows:

$$\alpha = Prob\ [\, z \leq zc \mid \mu = \mu_o \,]$$

which is read as "the probability that z is less than or equal to zc given that $\mu = \mu_0$." This probability is the shaded area in Figure 1a, which is bounded on the right by zc. It is evident from the figure that the significance level can be increased by shifting zc to the right, or decreased by shifting zc to the left, and that the value required to achieve a particular significance level α is $zc = z_\alpha$, where z_α is the lower $\alpha \times 100$ percentile of the standard normal distribution. For example, if $\alpha = 0.05$, then the appropriate critical value is $zc = -1.64$.


(a) Null hypothesis true: significance level = α

(b) Alternative hypothesis true: power = 1 - β

FIGURE 1. One-sample test (σ known): (a) significance level = α; (b) power = $1-\beta$.

The power of a statistical test is analogous to the significance level, with the relevant probability computed under the assumption that H_a, rather than H_0, is true. Therefore, the power ($1{-}\beta$) of the preceding test is:

$$1{-}\beta = Prob\ [\, z \le zc \,|\, \mu = \mu_a < \mu_0 \,]$$

This probability corresponds to the shaded area in Figure 1b (cf. the shaded area of Figure 1a). An equivalent expression for the power can be obtained by subtracting $\frac{\mu_a - \mu_0}{\sigma / \sqrt{n}}$ (which is the mean of z under H_α) from both sides of the inequality, that is:

$$1{-}\beta = Prob\ [\, z - \frac{\mu_a - \mu_0}{\sigma / \sqrt{n}} \le zc - \frac{\mu_a - \mu_0}{\sigma / \sqrt{n}} \,|\, \mu = \mu_a \,]$$

Since the random variable on the left side of the inequality has a standard normal distribution when $\mu = \mu_a$, the power can be evaluated by referring the number on the right side, to a table of the cumulative standard normal distribution. For example, substituting $z_{0.05} = -1.64$, $\mu_0 = 200$ cm, $\mu_a = 180$ cm, $\sigma = 80$ cm, and $n = 50$ gives:

$$zc - \frac{\mu_a - \mu_0}{\sigma / \sqrt{n}} = -1.64 - \left(\frac{180 - 200}{80 \ / \ \sqrt{50}} \right) = 0.1278$$

which corresponds to $1{-}\beta = 0.55$. In other words, if the mean height of the entire plantation is actually 20 cm less than the minimum requirement of 200 cm, there is a 55% chance that the test will identify the plantation as sub-standard and a 45% chance that it will not.

2.3.1 Factors that determine power

In general, the power of a test depends on three main factors:

i. the **significance level** of the test α;

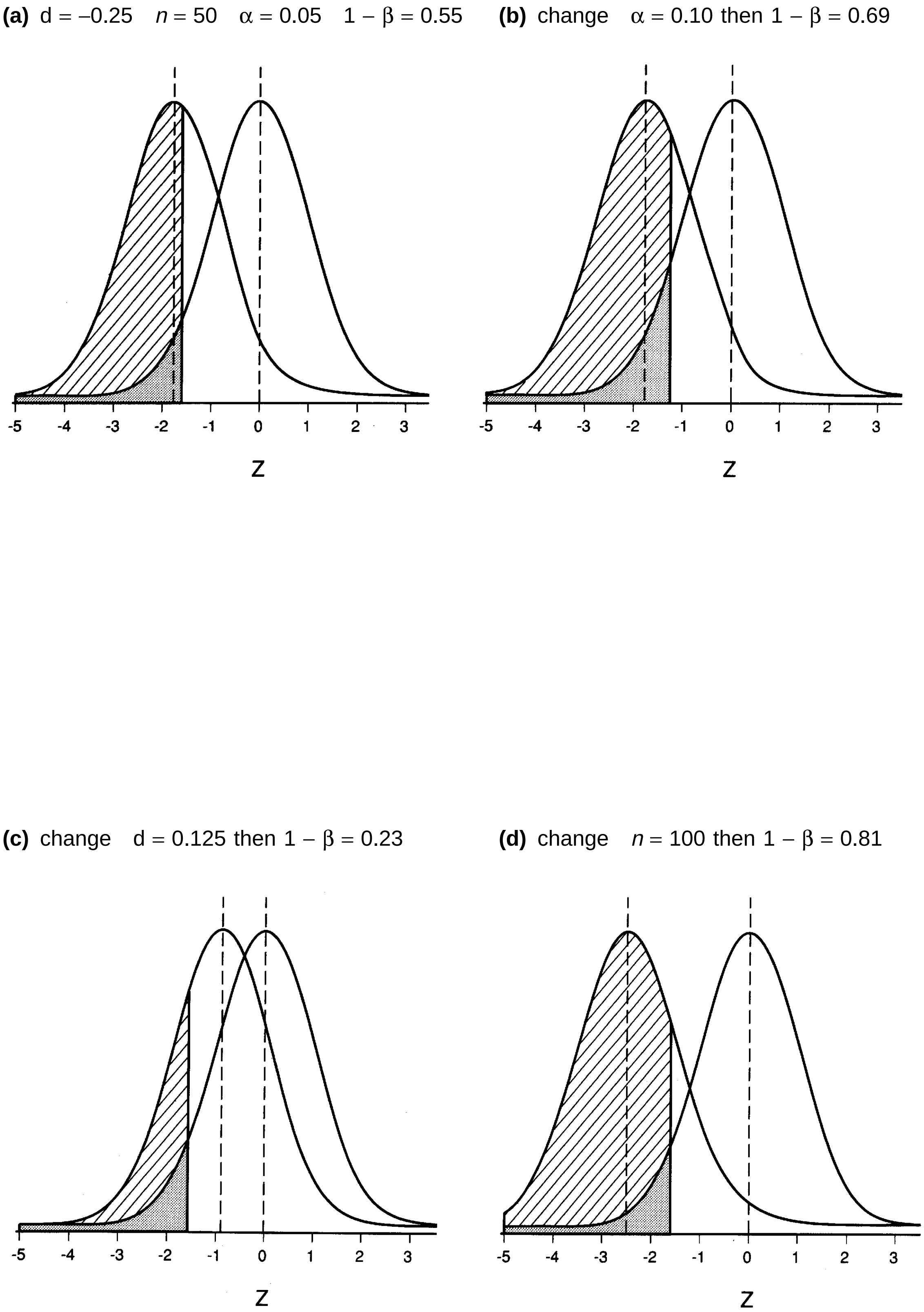
ii. the **effect size** d (which is a measure of the degree to which the null hypothesis is false); and

iii. the **sample size** n.

To see how each of these factors influences power, consider the example described in the previous section. Table 2 lists the power for five cases. Case 1, which corresponds to $\alpha = 0.05$, $\mu_a = 180$ cm, $\sigma = 80$ cm, and $n = 50$, will serve as a standard for the comparison of Cases 2–5, which are the result of varying α, μ_a, σ and n, one at a time. The significance level (shaded area) and power (line filled area) for Cases 1, 2, 3, (4) and 5 are depicted in Figures 2a–d, respectively. As in Figure 1, the vertical reference lines are used to locate the mean of the test statistic, z, under H_0: $\mu = \mu_0$ and H_a: $\mu = \mu_a$.

TABLE 2. Effect of varying α, d, σ and n on the power ($1{-}\beta$) of a simple one-sample test of H_0: $\mu \ge$ 200 cm versus H_a: $\mu <$ 200 cm with σ known

Case	α	μ_a (cm)	σ (cm)	d	n	$1{-}\beta$
1	0.05	180	80	–0.250	50	0.55
2	0.10	180	80	–0.250	50	0.69
3	0.05	190	80	–0.125	50	0.23
4	0.05	180	160	–0.125	50	0.23
5	0.05	180	80	–0.250	100	0.81



(a) d = –0.25   *n* = 50   α = 0.05   1 – β = 0.55

(b) change   α = 0.10 then 1 – β = 0.69

(c) change   d = 0.125 then 1 – β = 0.23

(d) change   *n* = 100 then 1 – β = 0.81

FIGURE 2. Comparison of power for one-sample test with σ known, and the effect on power of changing (b) α, (c) d, and (d) n.

It has already been noted that the significance level of the test under consideration can be varied by adjusting the critical value *zc*. Figure 2b (Case 2) shows that by increasing *zc* from –1.65 to –1.28, thereby increasing α from 0.05 to 0.10, the power can be increased from 0.55 to 0.69 (cf. Figure 2a). In fact, regardless of the test, greater power can always be achieved by increasing the significance level. In other words, the risk of a Type II error can always be decreased by incurring an increased risk of a Type I error, and vice versa.

The power of a test also depends on the degree to which the null hypothesis is false. In the present example, if μ_a is increased from 180 to 190 cm (Case 3), the power of the test decreases from 0.55 to 0.23 (cf. Figures 2a and 2c). However, since the same decrease in power occurs when σ is increased from 80 to 160 cm, and μ_a is held fixed at 180 cm (Case 4), it is apparent that what is actually important is the standardized difference between the two means (column 5 of Table 2):

$$d = \frac{\mu_a - \mu_0}{\sigma}$$

For this reason, power is often expressed as a function of *d*, which is called the **effect size**. The idea of using a dimensionless index to quantify departures from the null hypothesis, such that power increases as the magnitude of the index increases, can be generalized to such other tests as one-sample and two-sample t-tests and ANOVA F-tests (see Cohen 1977).

The final factor that determines power is the sample size (Case 5). Comparison of Figures 2a and 2d shows that, for a given effect size *d* and significance level α, power increases as the sample size increases. Thus, the one way to achieve greater power without an accompanying increase in the significance level is to collect more data.

2.3.2 Power of t-test

One sample t-test

In the example discussed in the previous section, the standard deviation σ was assumed to be known. In practice, σ is unknown and a t-test must be used. Given a simple random sample of size *n* from a normal distribution, with mean μ and standard deviation σ, the one-sample *t*-statistic is:

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

where $\bar{y}$ is the sample mean, as before, and σ has been replaced by the sample standard deviation *s.* When $\mu = \mu_0$, *t* has a **central *t*-distribution** (or simply a *t*-distribution) with *n–1* degrees of freedom. Therefore, the test that rejects H_0: $\mu \geq \mu_0$ in favour of H_a: $\mu < \mu_0$ when $t \leq tc$ has significance level $\alpha.$ Here the critical value $tc = t_{\alpha(n-1)}$ is the number that cuts off probability α in the lower tail of a *t*-distribution with *n–1* degrees of freedom.

To calculate the power of the t-test, it is necessary to determine the distribution of the *t*-statistic under H_a. If $\mu = \mu_a < \mu_0$, the distribution of *t* is shifted to the left, relative to its distribution under H_0, in the same way that the distribution of *z* was shifted under H_a (Figure 1). However, this time the result is a **noncentral *t*-distribution** with *n-1* degrees of freedom and **noncentrality parameter**,

$$\delta = \frac{(\mu_a - \mu_0)}{\sigma}\sqrt{n} = d\sqrt{n}$$

Consequently, the power of the t-test is:

$$1-\beta = Prob\ [\ t \leq tc \mid \delta\ ]$$

which is equal to the cumulative distribution function of the appropriate noncentral *t*-distribution evaluated at $tc = t_{\alpha(n-1)}$.

To calculate the critical value and the corresponding power of a t-test, tables of the central and noncentral *t*-distribution, or a computer program, are required. In SAS (SAS Institute 1985), the critical value *tc* can be computed as TC = TINV(ALPHA,DF,0), where ALPHA is the significance level α, DF = *n–1* is the degrees of freedom, and the final argument, which is the noncentrality parameter δ, is used to select the central *t*-distribution (i.e., $\delta = 0$). The corresponding cumulative probability for the noncentral *t*-distribution can be computed as P = PROBT(TC,DF,NC), where TC and DF have already been defined and NC is the noncentrality parameter δ. By adjusting the arguments, these two functions can be used to calculate the critical value and the power of both one-sided and two-sided *t*-tests, as shown in Table 3.

TABLE 3. Power of t-test: arguments for SAS functions TINV and PROBT

H_0	H_α	Critical value(s)	Power
$\mu \leq \mu_0$	$\mu > \mu_0$	TC = TINV (1–ALPHA,DF,0)	1–PROBT (TC,DF,NC)
$\mu \geq \mu_0$	$\mu < \mu_0$	TC = TINV (ALPHA,DF,0)	PROBT (TC,DF,NC)

$\mu = \mu_0$	$\mu \neq \mu_0$	TC1 = TINV (1–ALPHA/2,DF,0)
		TC2 = –TC1
			1–PROBT (TC1,DF,ABS(NC))
			+ PROBT (TC2,DF,ABS(NC))

Two-sample t-test

The computation of power for a two-sample t-test is similar to that for the one-sample case. Suppose a random sample of size n_1 is drawn from a normal population, which has mean μ_1 and standard deviation σ, and a second independent sample of size n_2 is drawn from another normal population, which has mean μ_2 and the same standard deviation σ. The two-sample *t* statistic for comparing the two means μ_1 and μ_2 is:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s_p^2 = \frac{(n_1 - 1)\ s_1^2 + (n_2 - 1)\ s_2^2}{n_1 + n_2 - 2}$$

is the pooled sample variance and $\bar{y}_1$, $\bar{y}_2$, s_1^2, s_2^2 are the respective sample means and variances. When $\mu_1 = \mu_2$, the two-sample *t*-statistic has a central *t*-distribution with n_1+n_2-2 degrees of freedom. Otherwise, it has a noncentral *t*-distribution with the same number of degrees of freedom and noncentrality parameter:

$$\delta = d\ \sqrt{\frac{n_1 n_2}{2\ (n_1 + n_2)}}$$

In this case, the effect size is $d = \frac{\mu_1 - \mu_2}{\sigma}$. Therefore, the power of the two-sample t-test can be calculated in the same way as the power of the one-sample test (see Table 3), with an appropriate adjustment to the degrees of freedom and the noncentrality parameter.

2.3.3 Power of ANOVA F-test

Analysis of variance (ANOVA) F-tests for **fixed effects** are often employed in the statistical analysis of forestry data. The general form of the null hypothesis (H_0) for the fixed effects case is that the group means are equal, or that all contrasts* involving these means are zero. For example, the null hypothesis for comparing *a* treatment groups in a randomized block design is H_0: $\mu_{1j} = \mu_{2j} = \ldots = \mu_{aj}$ for all blocks j (here μ_{ij} is the mean response for treatment group i in block j). The alternative hypothesis (H_a) for a fixed effects ANOVA is that at least one of the means specified in H_0 differs from the rest, or that at least one contrast differs from zero.

The *F*-ratio for testing H_0 against H_a has the following general form:

$$F = \frac{SS_H / df_H}{SS_E / df_E}$$

where SS_H is the sum of squares associated with the null hypothesis, SS_E is the applicable error sum of squares, and df_H and df_E are the respective degrees of freedom. Under H_0, the numerator and denominator of the *F* ratio are both unbiased estimates of the error variance, σ^2_E, and *F* has a **central *F*-distribution** with df_H degrees of freedom in the numerator and df_E degrees of freedom in the denominator (this is the usual *F*-distribution). However, when H_a holds, the distribution of *F* is shifted towards larger values (since, in that case, the numerator has an expected value that exceeds the error variance) and the result is a **noncentral *F*-distribution** with df_H and df_E degrees of freedom in the numerator and denominator, respectively, and a noncentrality parameter λ,† which can be written as follows (see O'Brien 1987):

$$\lambda = \frac{\mathrm{SS}_{Ha}}{\sigma^2_E}$$

Where SS_{Ha} is the between sums of squares of the observations replaced by their expected values under H_a.

The ANOVA F-test rejects H_0 when $F \geq FC$, where the critical value *FC* is the lower $(1-\alpha) \times 100$ percentile (or upper $\alpha \times 100$ percentile) of the central *F*-distribution, with the appropriate degrees of freedom for the numerator and denominator. Knowledge of the distribution of *F* under H_0 and H_a is sufficient to verify that this test has significance level α and that the power is:

$$1-\beta = Prob\ [F \geq FC \mid \lambda]$$

which is one minus the cumulative distribution of the appropriate noncentral *F* evaluated at *FC*. Thus, computation of the power of an F-test involves the following steps:

i. determine the degrees of freedom for the numerator and the denominator of the *F*-ratio, df_H and df_E;

* Recall that a contrast is a weighted sum of the means where the weights sum to zero.

† The definition of the noncentrality parameter varies throughout the literature. Check definitions carefully when using published tables and commercial computer programs.

ii. fix α and calculate the corresponding critical value, FC = $F_{1-\alpha}(df_H, df_E)$;

iii. calculate the noncentrality parameter, λ; and

iv. compute the power using the appropriate noncentral *F*-distribution.

As in the case of the t-test, statistical tables (of the central and noncentral *F*-distribution) or a computer program are needed to carry out the calculations in steps (ii) and (iv). The SAS function PROBF, which is the cumulative distribution function of the noncentral *F*-distribution, and FINV, which is its inverse, can be used for this purpose. In particular, if the following variables have been assigned values

ALPHA = significance level, α
DFH = degrees of freedom for numerator, df_H
DFE = degrees of freedom for denominator, df_E
NC = noncentrality parameter, λ

then FC = FINV(1–ALPHA,DFH,DFE,0) is the required critical value and POWER = 1–PROBF (FC,DFH,DFE,NC) is the corresponding power of the F-test. Notice that since the critical value is calculated under the assumption that H_0 is true, the noncentrality parameter (last argument for both functions) must be set equal to zero when FINV is used to calculate the critical value. This corresponds to the usual (i.e., central) *F*-distribution.

The noncentrality parameter can easily be computed with SAS. O'Brien (1987) and Sanders (1989) point out that, if PROC GLM is used to carry out an ANOVA of a set of means representing a specific departure from H_0 (i.e., a data set comprising a hypothesized mean for each combination of factors), then the resultant sum of squares for the factor, or contrast, under investigation (multiplied by a suitable scale factor, if necessary) is SS_{Ha}. They also describe how FINV and PROBF can be used, in conjunction with PROC GLM and PROC TABULATE, to tabulate the results of a series of power calculations. O'Brien has written an SAS program, **FPOWTAB**, for this purpose. Instructions for its use are given in the Appendix 1, along with an example. Additional examples illustrating the use of FINV and PROBF are in Section 3, and in Sanders (1989) and O'Brien (1987).

One-way ANOVA (equal sample sizes, fixed effect)

The one-way ANOVA model with fixed effects (and equal sample sizes), which is described in detail in many statistics textbooks (e.g., Keppel 1973; Snedecor and Cochran 1973; Devore 1987), is:

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

where μ_i (sometimes written as $\mu + \alpha_i$) is the mean for treatment group i. The measurements y_{ij} {i=1,2..,*a*; j=1,2,...,*n*} are obtained by randomly assigning each of the *a* treatments to *n* treatment units (t.u.'s).

Application of the procedure outlined above is relatively straightforward for this case. The *F*-ratio for testing H_0: $\mu_1=\mu_2=\ldots=\mu_a$ (no treatment effect) has $df_H = a–1$ degrees of freedom in the numerator and $df_E = a(n–1)$ degrees of freedom in the denominator, and is given by:

$$F = \frac{SS_A / (a-1)}{SS_E / a\,(n-1)}$$

where

$$SS_A = n \sum_{i=1}^{a} (\bar{y}_i - \bar{y})^2$$

is the treatment sum of squares and SS_E is the usual residual sum of squares ($\bar{y}_i$ is the sample mean for treatment group i and $\bar{y}$ is the overall mean).

To calculate the noncentrality parameter for a particular set of population means ($\mu_1,\mu_2,\ldots,\mu_a$), the treatment sum of squares SS_A is evaluated by substituting $y_{ij} = \mu_i$. In that case $\bar{y}_i = \mu_i$ and $\bar{y} = \bar{\mu}$ (mean of $\mu_1,\mu_2,\ldots,\mu_a$), which gives:

$$SS_A = n \sum_{i=1}^{a} (\mu_i - \bar{\mu})^2$$

Therefore, the noncentrality paramater for the one-way ANOVA is:

$$\lambda = \frac{n \sum_{i=1}^{a} (\mu_i - \bar{\mu})^2}{\sigma^2}$$

By applying the results of Cohen (1977), one can show that, when a is odd, λ is bounded above and below as follows:

$$\frac{nd^2}{2} \leq \lambda \leq \frac{nd^2}{4a}(a^2 - 1)$$

where $d = \frac{\mu_{max} - \mu_{min}}{\sigma}$ is the **effect size index** (or the standardized range of the treatment means) and μ_{min} and μ_{max} are, respectively, the minimum and maximum of $\mu_1,\mu_2,\ldots,\mu_a$. The effect size index is a generalization of the effect size for the two-sample t-test. When a is even, the lower bound is unchanged but the upper bound is $d^2na/4$. Since the power of an F-test is an increasing function of the noncentrality parameter, the preceding bounds imply that the minimum and maximum power occur when $\lambda = \frac{nd^2}{2}$ and when $\lambda = \frac{nd^2}{4a}(a^2-1)$, respectively, with the appropriate substitution when a is even. The application of these bounds is demonstrated in Example 3.3 (Section 3).

Two-way ANOVA (equal sample sizes, fixed effects)

The two-way ANOVA model with fixed effects is:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

Here the observations y_{ijk}, which are assumed to have a common variance, σ^2, comprise ab independent random samples each of size n, one for each possible combination of the a levels of factor A and the b levels of factor B. If the means μ_{ij} are written as $\mu + \alpha_i + \beta_j + \alpha\beta_{ij,}$ with α_i and β_j denoting the main effects of A and B respectively, and $\alpha\beta_{ij}$ denoting the interaction of the two factors, there are three null hypotheses of interest: (1) H_{AB}: $\alpha\beta_{ij} = 0$ for all i and j (no interaction between A and B); (2) H_A: $\alpha_i = 0$ for all i (no A effect); and (3) H_B: $\beta_j = 0$ for all j (no B effect). Refer to Devore (1987), Snedecor and Cochran (1973), and Keppel (1973) for a complete description of the two-way model and the corresponding ANOVA. The power of the associated F-tests can be computed by applying the same general procedure as was used in the one-way case. Table 4 summarizes the pertinent information.

TABLE 4. Power of F-tests for two-way ANOVA

Null hypothesis	**Noncentrality parameter λ**	**Numerator df df_H**	**Denominator df df_E**
H_{AB}: $\alpha\beta_{ij} = 0$	$n \sum \sum (\alpha\beta_{ij})^2/\sigma^2$	$(a-1)\ (b-1)$	$ab\ (n-1)$
H_A: $\alpha_i = 0$	$nb \sum \alpha_i^2/\sigma^2$	$a-1$	$ab\ (n-1)$
H_B: $\beta_j = 0$	$na \sum \beta_j^2/\sigma^2$	$b-1$	$ab\ (n-1)$

The only real difference from the one-way case is that the specification of the alternative is slightly more complicated because there are two main effects and possible interactions to consider.

Since the interpretation of the main effects (α_i, β_j) is complicated by the presence of an interaction, H_{AB} is tested first. If H_{AB} is not rejected, it is generally assumed that there are no interactions and, on this basis, H_A and H_B are tested and the results interpreted accordingly. However, as noted previously, failure to reject a hypothesis cannot be taken as evidence that the hypothesis is true, unless it can first be shown that the test is powerful against all relevant alternatives. For this reason, it is particularly important to assess the power of the F-test for an interaction. This requires some thought as to the types of interactions that are likely to be of concern, so that a suitable set of population means can be constructed for the evaluation of λ. If the test is found to be sufficiently powerful, the power of the tests for the two main effects can be examined with the same methods as were applied in the one-way case.

2.4 Applications of Power Analysis

The two principal applications of power analysis are: (1) in experimental design and (2) in *post hoc* power analysis. In the former, a power analysis is performed before the experiment is conducted, and so the adopted parameter values (i.e., the error variance, effect size, and noncentrality parameter) are necessarily based on prior knowledge and are considered fixed (although power is usually computed for a range of values). In the latter, the power analysis is performed after the data have been collected, with the parameter values replaced by estimates based on the sample. Because these estimates are subject to sampling error, the derived values for the power are also subject to error (which is difficult to quantify). For this reason, *post hoc* evaluations of power should be viewed with some caution (refer to Keppel 1973 and Korn 1990 for discussions of this point).

2.4.1 Experimental design

Several important factors must be considered when planning a forestry trial, including: the study objectives; the method of randomization (e.g., completely randomized design versus randomized blocks); sample size and optimum allocation of the sampling effort to the (primary) experimental units and (secondary) sampling units (e.g., number of plots, subplots and trees within subplots); use of adequate controls; quality control in the data collection and data entry; and the data analysis. Since power is an objective criterion by which competing designs can be judged, it should feature prominently in the resolution of at least some of these issues, most notably selection of a suitable sample size.

Sample size selection

Selection of the sample size is the most common way in which a power analysis is employed at the design stage of a study. As discussed previously, sample size is only one of the parameters that determines power. The other parameters are the significance level, the degree to which the null hypothesis is false, and the error variance. Consequently, all sample size calculations depend on the adopted values of these parameters. It is therefore generally advisable to carry out sample size calculations for several representative alternative hypotheses and several combinations of the significance level (e.g., $\alpha = 0.01, 0.05$, and 0.10) and error variance (e.g., σ_E estimated from a pilot study, $\sigma_E/2$, $2\sigma_E$, etc.). In some cases, it may be convenient to express the alternative in terms of a dimensionless index of effect size, which can be transformed back to the original units of measurements if an estimate of the error variance is available.

Alternative hypothesis

Apart from its role in determining sample size, a power analysis is valuable because it serves to focus attention on the alternative hypothesis. In particular, to perform a power analysis, the investigator must consider carefully the types of effects that are of interest and attempt to quantify those effects. This in itself is a useful exercise, which can lead to improvements in the design.

2.4.2 *Post hoc* power analysis

A power analysis provides valuable insight into the interpretation of the results of a statistical analysis, particularly when the null hypothesis is not rejected. Failure to reject a null hypothesis is an ambiguous result: it neither confirms nor rules out the null hypothesis. The only way to resolve this issue is to determine the power. In the absence of any other information, a *post hoc* power analysis is often used for this purpose. In addition, a *post hoc* power analysis can provide valuable information for planning future trials.

3 EXAMPLES

3.1 One-Sample t-test

Suppose, in the regeneration performance assessment example described in Section 2.3, that σ is unknown and a one-sample t-test, based on a sample of $n = 50$ trees, is used to test H_0: $\mu \geq 200$ cm versus H_a: $\mu < 200$ cm, at the $\alpha = 0.05$ level of significance. To investigate how the power varies as a function of the effect size $d = (\mu_a - 200)/\sigma$ (which, in this case, is the amount by which the plantation mean falls below the standard of 200 cm), SAS was used to calculate the power for d = –0.05, –0.10, …, –0.50 as follows (see Table 3):

```
DATA TTEST;
    TC = TINV (.05,49,0);
    DO D = −.05 TO −.5 BY −.05;
        NC = SQRT (50) *D;
        POWER = PROBT (TC,49,NC);
        OUTPUT;
    END;
LABEL D=′Effect Size′  POWER=′Power′
        NC=′Noncentrality Parameter′
        TC=′Critical Value′ ;
PROC PRINT NOOBS LABEL;
TITLE ′EXAMPLE 3.1: One-Sample t-test (n=50)′ ;
    BY TC;
    VAR D NC POWER; FORMAT TC 6.3;
RUN;
```

The output from the program is listed on the next page:

EXAMPLE 3.1. One-Sample t-test (n=50)

```
     ──────────────────── Critical Value=–1.677 ────────────────────

              Effect          Noncentrality
               Size             Parameter            Power

              –0.05             –0.35355            0.09746
              –0.10             –0.70711            0.17170
              –0.15             –1.06066            0.27464
              –0.20             –1.41421            0.40122
              –0.25             –1.76777            0.53921  ← σ/4
              –0.30             –2.12132            0.67257
              –0.35             –2.47487            0.78687
              –0.40             –2.82843            0.87372
              –0.45             –3.18198            0.93224
              –0.50             –3.53553            0.96721  ← σ/2
```

This power analysis demonstrates that, with a sample size of $n = 50$, the t-test (with $\alpha = 0.05$) has a greater than 50% chance of classifying the plantation's performance as sub-standard if the mean height falls more than $\sigma/4$ below the standard of 200 cm (i.e., $d = -0.25$). If the mean falls more than $\sigma/2$ below the standard (i.e., $d = -0.50$), there is a greater than 95% chance that its performance will be declared sub-standard.

3.2 Two-Sample t-test

In an experiment to compare two containers for growing conifer seedlings, n one-year-old seedlings will be randomly assigned to each of the two types of containers (for a total sample size of $2n$). The height of the seedlings will be measured at the time of planting and annually for the next 5 years. Before conducting the experiment, the researcher wants to know whether $n = 50$ seedlings per container type is sufficient to have an 80% chance of detecting, at the $\alpha = 0.05$ level of significance, a difference between the two groups of $|\mu_1 - \mu_2| = 1$ cm in the average annual height growth. Based on the results of similar studies, the estimated standard deviation of the annual height growth is $\sigma = 2$ cm for both groups.

The size of the effect that the researcher wishes to detect is $d = 0.5$ and the corresponding noncentrality parameter is:

$$\delta = d \sqrt{\frac{n \times n}{2\,(n+n)}} = \frac{\sqrt{n}}{4}$$

To investigate how the power of the two-sided, two-sample t-test varies as a function of the sample size, the power can be computed for $n = 10, 20, \ldots, 150$ using the following SAS program (see Table 3):

```
DATA TTEST2;
   DO N = 10 TO 150 BY 10;
      NC = SQRT(N)/4;
      TC1 = TINV(.975,2*N–2,0);
      TC2 = –TC1;
      POWER = 1–PROBT (TC1,2*N–2,NC)+PROBT(TC2,2*N–2,NC);
      OUTPUT;
   END; LABEL POWER =′Power′;
   KEEP N POWER;
PROC PRINT DATA=TTEST2 NOOBS;
   TITLE′EXAMPLE 3.2: Two-Sample t-test (n1=n2=n, d=0.5)′;
RUN;
```

The results show that $n = 50$ is not large enough to meet the power requirement. A sample size between 120 and 130 is needed. To obtain a precise estimate the calculations must be repeated with $n = 121,122,\ldots,130$.

EXAMPLE 3.2. Two-Sample t-test (n1=n2=n, d=0.5)

N	**Power**
10	0.11635
20	0.19328
30	0.27032
40	0.34543
50	0.41712
60	0.48440
70	0.54664
80	0.60355
90	0.65504
100	0.70122
110	0.74231
120	0.77863
130	0.81053
140	0.83839
150	0.86259

3.3 F-test for One-Way ANOVA

In a study to compare five vegetation management treatments — three herbicides (“Veg-X,” “Noweed,” “Super H”), manual removal, and a control — a completely randomized design is employed, in which each treatment is assigned to three plots. The plots contain 20 subplots and the response variable of interest is the average (over subplots) height of the vegetation after 5 years.

Suppose the means in Table 5 represent the minimum effect that would be considered of practical importance. To calculate the probability that the relevant F-test (with $df_A = 5–1 = 4$ and $df_E = 5(3–1) = 10$) will detect such an effect, the noncentrality parameter must first be calculated.

TABLE 5. Minimum expected increments in overall height of vegetation (cm)

Control	Manual	Veg-X	Noweed	Super H
600	500	500	400	400

The following SAS program calculates the numerator of the noncentrality parameter (SS_{Ha}) by creating a data set in which the height increments (DHT) are equal to the above group means, and then using this data set as input to PROC GLM:

```
DATA MEANS;
    N=3;
    DO TREAT=1 TO 5;
        INPUT DHT @@;
        OUTPUT;
    END;
CARDS;
600 500 500 400 400
PROC GLM DATA=MEANS;
    CLASS TREAT;
    WEIGHT N;
    MODEL DHT=TREAT;
RUN;
```

Notice that each mean is weighted by n (= 3) since the actual data set would have three plot averages per treatment. The resultant treatment sum of squares is $SS_{\mathrm{Ha}} = 84000$. The same result could also have been obtained by running the analysis without the weighting factor ($n = 1$) and then multiplying the treatment sum of squares by three. Since this is true in general, SS_{Ha} is usually most conveniently calculated by first computing the sum of squares for the minimum number of ‘‘observations’’ needed to represent the proportional allocation of the total sample to the various combinations of factors, and then multiplying the relevant sum of squares by the appropriate value of n.

To complete the calculation of λ, an estimate of the common standard deviation of the plot averages, σ, must be provided. Suppose $\sigma = 200$ cm is the best available estimate. Then the noncentrality parameter is $\lambda = 84000/200^2 = 2.1$ and the probability that the F-test, with $\alpha = 0.10$, will detect the differences between the means given in Table 5 is only 0.22. This number can be calculated using SAS as follows:

```
DATA POWER;
    FC = FINV(.9,4,10,0);
    POWER = 1–PROBF(FC,4,10,2.1);
    LABEL FC=′Critical Value′  POWER=′Power′;
PROC PRINT NOOBS LABEL;
RUN;
```

```
 Critical       Power
  Value

2.60534        0.22378
```

In the preceding calculation, it was necessary to supply H_a values for the individual treatment means (see Table 5). An alternative approach is to specify a standardized range for the means ($d = [\mu_{\max} - \mu_{\min}]/\sigma$), which is the maximum difference (under H_a) between any pair of means divided by the standard deviation. The minimum and maximum power can then be computed by calculating the maximum and minimum values for the noncentrality parameter, as described in Section 2.3.3. In particular, substituting $a = 5$ and $n = 3$ into the appropriate expressions gives minimum and maximum λ values of $1.5d^2$ and $3.6d^2$, respectively.

A SAS program for calculating the corresponding minimum and maximum power when $d = 0, 0.05, \ldots, 2$, and for plotting the results, is given below. The output is shown in Figure 3. If the SAS/GRAPH procedure GPLOT is not available, the PLOT procedure can be substituted with appropriate changes to the program.

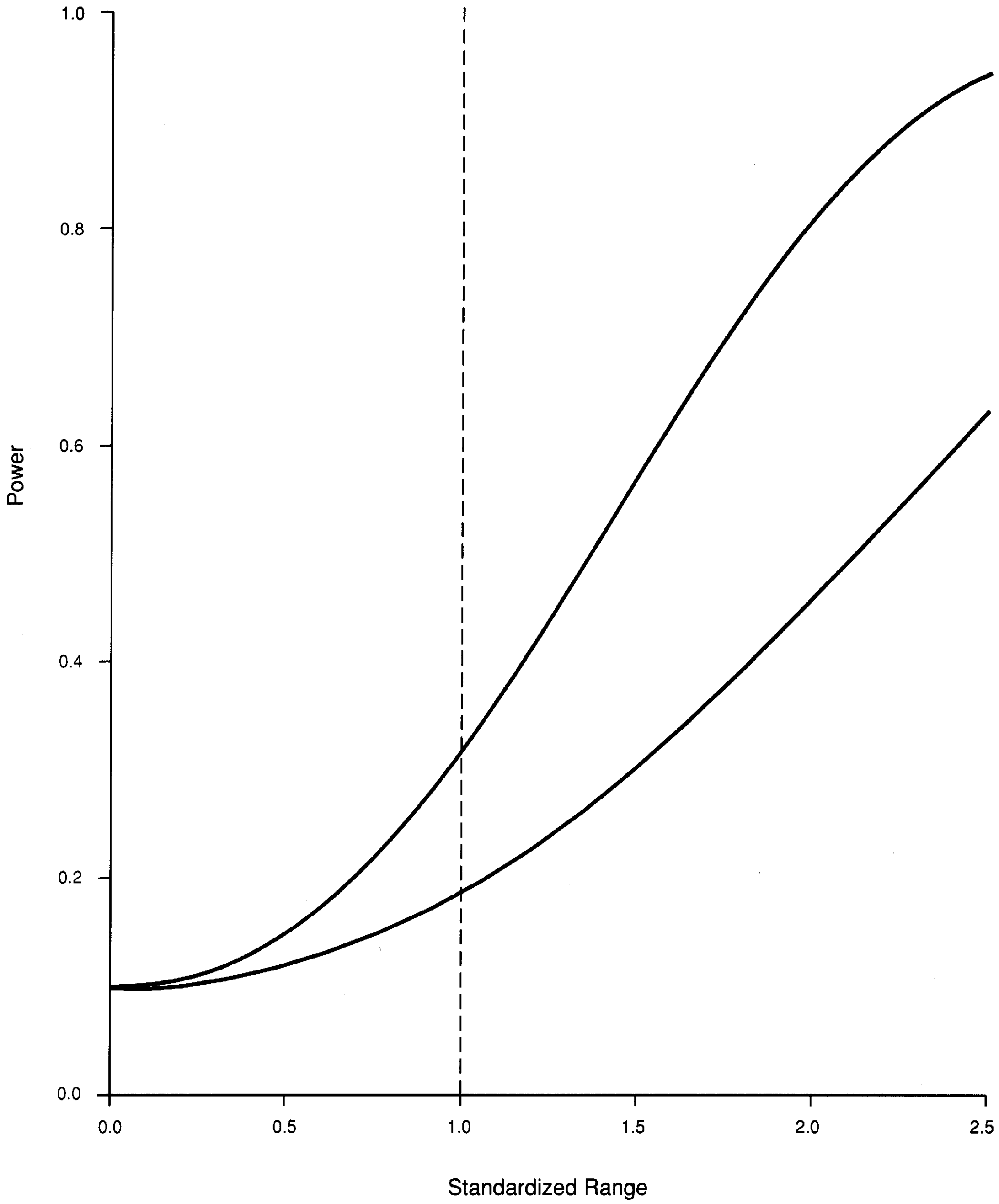


FIGURE 3. Maximum (upper curve) and minimum (lower curve) power versus standardized range of means for one-way ANOVA F-test (see Example 3.3).

```
DATA FTEST1;
   FC = FINV(.9,4,10,0);
   DO D=0 TO 2.5 BY .05;
      NCMIN = 1.5*D*D;
      NCMAX = 3.6*D*D;
      POWMIN = 1–PROBF(FC,4,10,NCMIN);
      POWMAX = 1–PROBF(FC,4,10,NCMAX);
      OUTPUT;
   END;
   KEEP D POWMIN POWMAX;
   TITLE J=C F=SIMPLEX H=.2 IN′EXAMPLE 3.3: F–TEST FOR ONE-WAY ANOVA′ ;
   SYMBOL1 I=JOIN V=NONE;
   SYMBOL2 I=JOIN V=NONE;
   AXIS1 LABEL=(J=C F=SIMPLEX H=.2 IN′STANDARDIZED RANGE′ )
         ORDER=0 TO 2.5 BY .5 OFFSET=(0)
         VALUE=(F=SIMPLEX H=.15 IN)
         MINOR=NONE;
   AXIS 2 LABEL=(J=C F=SIMPLEX H=.2 IN A=90′POWER′ )
         ORDER=0 TO 1 BY .2 OFFSET=(0)
         VALUE=(F=SIMPLEX H=.15 IN)
         MINOR=NONE;
PROC GPLOT;
   PLOT (POWMIN POWMAX)*D/OVERLAY HAXIS=AXIS1 VAXIS=AXIS2 HREF=1 LH=2;
RUN;
```

According to these results, there is not much chance of detecting a maximum difference between the means of 1σ (i.e., reading from the graph with the standardized range equal to one [vertical line], the minimum power is 18% [lower curve] and the maximum power is 32% [upper curve]). Only if the range is about 2σ is there likely to be a reasonable chance that the F-test will reject the hypothesis that there is no difference between the five treatments.

3.4 Comparison of Completely Randomized and Randomized Block Designs

For the trial described in Example 3.3, suppose that instead of using a completely randomized design, the 15 plots are divided into three homogeneous blocks, each containing five plots to which the treatments are randomly assigned. The ANOVA tables for the two designs are compared in Table 6. Notice that the number of degrees of freedom for the treatment sum of squares SS_A is the same for both designs. In fact, the computational formula for SS_A is the same for both designs. However, this does not carry over to the error term and, as a result, the power of the F-test is different for the two designs.

TABLE 6. Comparison of ANOVA tables

(a) Completely randomized design

Source	Degrees of freedom	Sum of squares	*F*-ratio
Treatment	$a-1 = 4$	SS_A	$\frac{SS_A/(a-1)}{SS_{CR}/a(n-1)}$
Error	$a\ (n-1) = 10$	SS_{CR}	
Total	$an - 1 = 14$		

TABLE 6. (continued)

(b) Randomized block design

Source	Degrees of freedom	Sum of squares	***F*-ratio**
Treatment	$a-1 = 4$	SS_A	$\dfrac{SS_A/(a-1)}{SS_{RB}/(a-1)\ (n-1)}$
Block	$n-1 = 2$	SS_B	
Error	$(a-1)(n-1) = 8$	SS_{RB}	
Total	$an - 1 = 14$		

To compare the power of the two designs, let σ^2_{CR} and σ^2_{RB} denote, respectively, the error variance for the completely randomized design and the error variance for the randomized block design. As before, assume that σ_{CR} = 200 cm and that Table 5 represents the minimum treatment effect to be detected. The noncentrality parameter for the completely randomized design, which was calculated in Example 3.3, is $\lambda_{CR} = 2.1$. Since the numerator is the same for both designs, the noncentrality parameter for the randomized block design can be written as $\lambda_{RB} = r\lambda_{CR}$, where $r = \sigma^2_{CR}/\sigma^2_{RB}$. Therefore, the two designs can be compared by varying r and calculating the corresponding power of the F-test. The computations with $\alpha = 0.10$ can be carried out as follows:

```
DATA CRRB;
   DO RVAR=1 TO 10;
      FCCR=FINV(.9,4,10,0);
      FCRB=FINV(.9,4,8,0);
      POWCR=1–PROBF(FCCR,4,10,2.1);
      POWRB=1–PROBF(FCRB,4,8,RVAR*2.1);
      RPOW = POWRB/POWCR;
      OUTPUT;
   END;
   KEEP RVAR POWCR POWRB RPOW;
   LABEL RVAR =′Ratio of Variances (CR/RB)′
         POWCR=′Power for CR′
         POWRB=′Power for RB′
         RPOW =′Power Ratio (CR/CB)′ ;
PROC PRINT NOOBS LABEL;
TITLE1 ′EXAMPLE 3.4: Comparison of Completely Random Design (CR) and Randomized Block Design (RB)′ ;
RUN;
```

The results show that, if blocking reduces the error variance by 50% (i.e., $r = 2$), there is a 49% increase in power (against the effects given in Table 5 and assuming that σ_{CR} = 200 cm and $\alpha = 0.10$). If blocking reduces the error variance by 66%, there is a 100% increase in power, and so on. If there is no reduction in σ^2_{CR}, then there is a slight (approximately 4%) decrease in power for the randomized block design, which is due to the loss of degrees of freedom in the denominator of the *F*-ratio. Hence, a randomized block design is to be preferred only if it is likely that $\sigma^2_{RB} < \sigma^2_{CR}$. That is, the blocking factor accounts for a considerable part of the variability between plots that is not due to treatment.

EXAMPLE 3.4. Comparison of Completely Random Design (CR) and Randomized Block Design (RB)

Ratio of Variances (CR/RB)	Power for CR	Power for RB	Power Ratio (CR/CB)
1	0.22378	0.21417	0.95708
2	0.22378	0.33325	1.48920
3	0.22378	0.44749	1.99974
4	0.22378	0.55115	2.46294
5	0.22378	0.64145	2.86647
6	0.22378	0.71770	3.20725
7	0.22378	0.78053	3.48801
8	0.22378	0.83125	3.71468
9	0.22378	0.87152	3.89462
10	0.22378	0.90303	4.03543

3.5 F-tests for Two-Way ANOVA

In a study to establish the optimum rate of application for a new fertilizer, six rates were used: 0 (control), 100, 200, 300, 400, and 500 kg/ha. Each rate was randomly assigned to 40 plots, of which 20 contained a single fir tree and 20 contained a single spruce tree. All trees were comparable with respect to age and, for a given species, with respect to initial height, diameter, and condition. The response variable of interest is the 3-year increment in diameter. A two-way ANOVA (species × level of fertilizer) will be used to analyse the results of the trial.

From previous studies, the diameter growth is thought to vary quadratically with rate, but it is uncertain whether or not the optimum rate differs for the two species. If there is a difference in the optimum rate, which implies that there is an interaction between species and rate, then it is thought to be of the order ± 100 kg N/ha. This situation (Case 1) is represented by the hypothetical means given in Table 7. These are plotted in Figure 4a. If there is no interaction (i.e., the species and rate effects are additive), then it is appropriate to compare the overall response of the two species. It will also be useful to test the significance of the linear and quadratic terms for the rate of application. Table 8 gives a set of means that might be expected in the additive case (Case 2). These are plotted in Figure 4b. Notice that in this case, the curves for the two species are parallel (cf. Figure 4a).

TABLE 7. Expected increments in diameter (mm) for Case 1

	Rate of application of fertilizer (100 kg N/ha)					
Species	**0**	**1**	**2**	**3**	**4**	**5**
Fir	2.50	4.00	5.00	5.50	5.50	5.00
Spruce	3.50	5.05	6.20	6.95	7.30	7.25

TABLE 8. Expected increments in diameter (mm) for Case 2

	Rate of application of fertilizer (100 kg N/ha)					
Species	**0**	**1**	**2**	**3**	**4**	**5**
Fir	2.50	4.00	5.00	5.50	5.50	5.00
Spruce	3.50	5.00	6.00	6.50	6.50	6.00

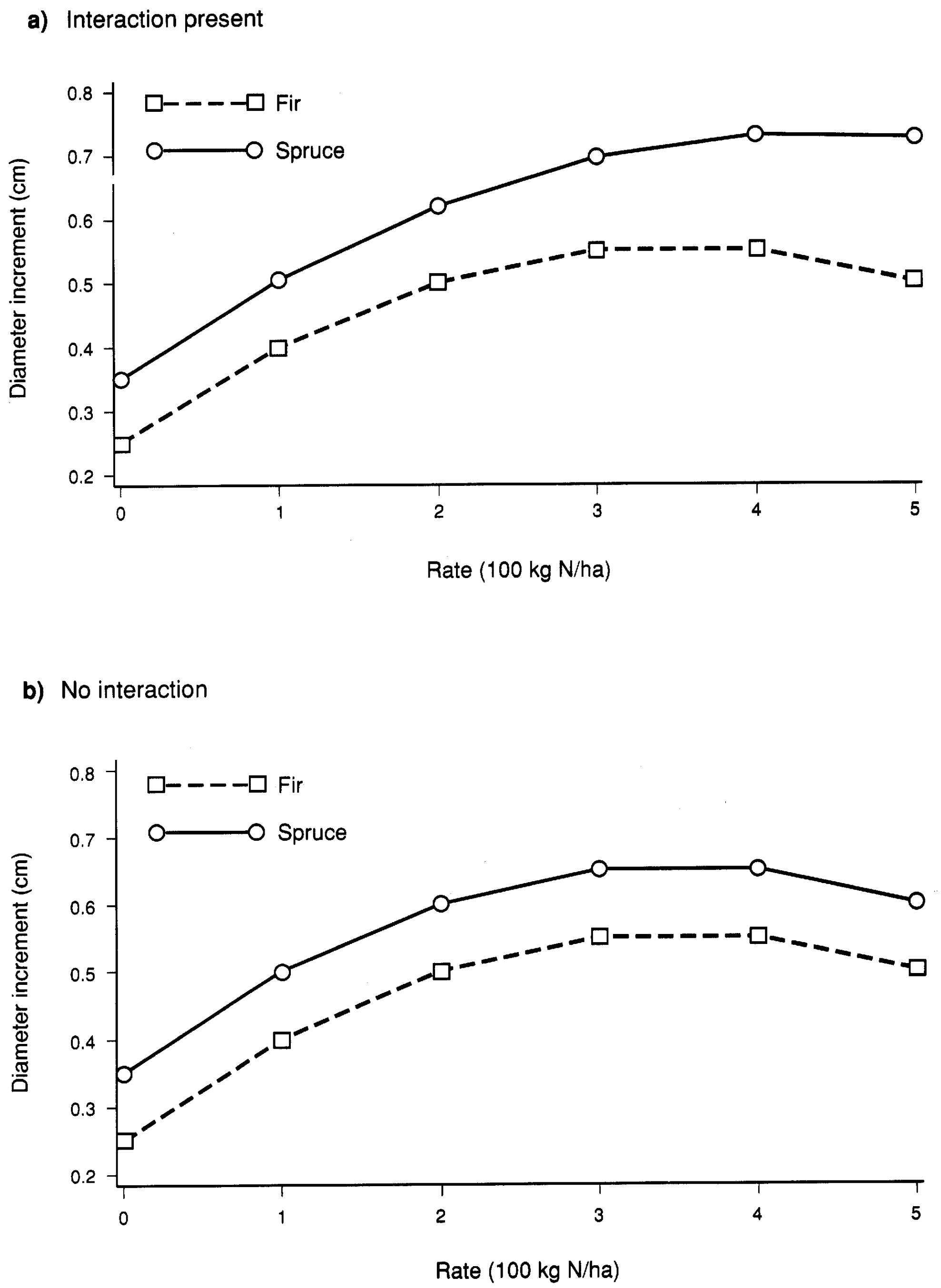


FIGURE 4. Diameter increment versus rate of application of fertilizer for fir and spruce: (a) interaction present; (b) no interaction (see Example 3.5).

To calculate the power of the relevant F-tests for these two cases, an estimate of σ^2 must be supplied and SS_{Ha} must be evaluated for each hypothesis of interest (see Table 4). These values are then substituted into the appropriate formula for calculating the critical value and power. The steps are illustrated below for Case 2 (no interaction), with $\sigma = 4$ mm and $\alpha = 0.05$.

- Step (i): Create a data set containing hypothetical means (Table 7).

```
DATA CASE2;
    N=20;
    DO SPECIES=1 TO 2;
        DO RATE=0 TO 5;
            INPUT DDIAM @;
            OUTPUT;
        END;
    END;
    CARDS;
    .25 .40 .50 .55 .55 .50   (entered in cm)
    .35 .50 .60 .65 .65 .60
    RUN;
```

- Step (ii): Calculate SS_{Ha} for all hypotheses of interest, including tests of contrasts, by carrying out an ANOVA of the data set created in Step (i).

```
PROC GLM DATA=CASE2;
    CLASS SPECIES RATE;
    MODEL DDIAM=SPECIES RATE SPECIES*RATE;
    WEIGHT N;
    CONTRAST ′Quadratic′  RATE   5 –1 –4 –4 –1 5;
    CONTRAST ′Linear′       RATE –5 –3 –1   1   3 5;
RUN;
```

The resultant hypothesis sums of squares (SS_{Ha}) will be denoted SSS for the species main effect and SSR for the rate main effect in Case 1; SSQ and SSL will denote the corresponding values for testing for a quadratic and a linear rate effect.

- Step (iii): Calculate the power of each test, with noncentrality parameter (for PROBF) SS_{Ha} from Step (ii) divided by σ^2.

```
DATA POWER;
    FC1=FINV(.95,1,228,0);
    FC2=FINV(.95,5,228,0);
    INPUT SSS SSR SSQ SSL;
    PSPEC  =  1–PROBF(FC1,1,228,6.25*SSS);
    PRATE  =  1–PROBF(FC2,5,228,6.25*SSR);
    PQUAD =  1–PROBF(FC1,1,228,6.25*SSQ);
    PLIN     =  1–PROBF(FC1,1,228,6.25*SSL);
CARDS;
0.6000 2.6833 0.9333 1.7500
RUN;
```

- Step (iv): Tabulate or print the results in a convenient form.

```
        PROC PRINT; RUN;
```

The results for Cases 1 and 2 are summarized in Table 9. These show that the F-test for an interaction is unlikely to detect an interaction as small as the one represented by Case 1. The test for the species main effect (Case 2) has only a 49% chance of detecting an overall difference between the diameter growth of the two species of 1.0 mm. The power of the remaining tests, except perhaps the test for a quadratic term in the rate effect, is within acceptable limits.

TABLE 9. Power of F-test for Cases 1 and 2 ($\alpha = 0.05$, $\sigma = 4$ mm)

H_0	H_a	SS_{Ha}	df_H	Power
H_{AB}: no interaction	Case 1	0.1187	5	0.0841
H_A: no species effect	Case 2	0.6000	1	0.4874
H_B: no rate effect	Case 2	2.6833	5	0.8980
H_L: no linear rate effect	Case 2	1.7500	1	0.9088
H_Q: no quadratic rate effect	Case 2	0.9333	1	0.6719

4 SUMMARY

Power analysis is a useful tool for designing and analysing forestry trials. The basic definitions and methods can, in theory, be applied to any hypothesis testing situation, thereby providing a common framework for the interpretation of the results of statistical analyses. Although, power computations are often cumbersome and time-consuming, they are relatively easy to perform for t-tests and a wide variety of multi-factor ANOVA F-tests. In SAS, calculation of the required noncentrality parameters is facilitated by PROC GLM, and the functions TINV, FINV, PROBT and PROBF can be used to evaluate the necessary critical values and cumulative probabilities.

APPENDIX 1: Instructions for using FPOWTAB

FPOWTAB (O'Brien 1988) is a SAS program (a FORTRAN version is also available) for tabulating the power of the F-test for various combinations of the input values: SS_{Ha}, α, n, df_H and df_E. A copy of this program can be obtained from R.G. O'Brien or from the Biometrics Section, Forest Science Research Branch, B.C. Ministry of Forests. The steps for running the program are:

Step (i): Calculate SS_{Ha} for each set of means of interest (H_a true), using a convenient sample size. For example, for a 2 × 3 factorial design with equal sample sizes, a sample size of six (one observation per cell) could be used. However, if the design is unbalanced so that two cells contain twice as many observations as the rest, a minimum sample size of $1 \times 4 + 2 \times 2 = 8$ is required to represent the proportional allocation of the total sample to the individual cells.

In FPOWTAB, each set of means is referred to as a **scenario** and the sample size used to calculate SS_{Ha} is called the **basis total sample** (BASTOTN). The computation of SS_{Ha} using PROC GLM is illustrated in Examples 3.3 and 3.5 of Section 3.

Step (ii): Create an input file with the following records:

Record 1: TITLEVAR = main title (maximum = 78 characters)

Record 2: BASETOTN = basis total sample size
RANKX = total number of nonredundant parameters in model (i.e., RANKX = total sample size − df_E)

Record 3: NUM_SCN = number of scenarios (maximum = 5)
SCNARIOV(i) = title of scenario i, i = 1 to NUM_SCN

Record 4: NUM_ALPH = number of levels of significance α (maximum = 3)
ALPHAV(i) = α_i, i = 1 to NUM_ALPH

Record 5: NUM_SD = number of σ_E values (maximum = 3)
STDDEVV(i) = $\sigma_{E,i}$, i = 1 to NUM_SD

Record 6: NUM_N = total number of sample sizes (maximum = 5)
TOTALNV(i) = ith **total** sample size, i = 1 to NUM_N

Record 7+: One record for each effect to be tested. Each record has the following form:

EFF_TITL = title of effect (maximum = 78 characters)
DF_HYPTH = df_H (degrees of freedom for numerator)
SSHPOPV(i) = SS_{Ha}, i = 1 to NUM_SCN

Note: All titles must be followed by at least two blank characters and records can be more than one line.

Step (iii): Change the OPTIONS and INFILE statements as required (see program documentation). Run FPOWTAB.

Example A.1: Using FPOWTAB to tabulate power of F-tests

To illustrate the use of FPOWTAB, the calculations in Example 3.5 will be repeated for $\alpha = 0.01, 0.10$; $\sigma = 3$ and 5 mm; and $n = 10$ and 30. The steps are as follows:

Step (i): Calculate SS_{Ha} for the two scenarios of interest: Case 1 and Case 2 (see Tables 7 and 8). This has already been done in Example 3.5 (Table 9).

Step (ii): Create an input data file. The contents of the file are listed on the next page. The missing values in the last five records are included because only the interaction is to be tested for Case 1 (first SSHPOPV value) and only the main effects (second SSHPOPV value) are to be tested for Case 2.

```
Record  1:     USING FPOWTAB TO TABULATE POWER OF F-TESTS
Record  2:     240 12
Record  3:     2
               CASE 1:  INTERACTION BETWEEN SPECIES AND RATE
               CASE 2:  NO INTERACTION BETWEEN SPECIES AND RATE
Record  4:     2 0.01 0.10
Record  5:     2 .3 .5
Record  6:     2 120 360
Record  7:     INTERACTION   5 0.1187  .
Record  8:     SPECIES MAIN EFFECT   1   .  0.600
Record  9:     RATE MAIN EFFECT   5  .  2.6833
Record 10:     LINEAR EFFECT OF RATE   1  .  1.7500
Record 11:     QUADRATIC EFFECT OF RATE   1  .  0.9333
```

Step (iii): Run FPOWTAB. Part of the output is listed below. Notice that the results for a one-sided t-test are also printed. The reason is that an F-test, with one degree of freedom in the numerator and significance level α, is equivalent to a two-sided t-test with significance level $\alpha/2$. For example, the one-sided t-test with significance level $\alpha = 0.10$ is equivalent to an F-test with significance level $\alpha = 0.20$.

USING FPOWTAB TO TABULATE POWER OF F-TESTS

EFFECT: SPECIES MAIN EFFECT,
DEGREES OF FREEDOM HYPOTHESIS: 1,
SCENARIO: CASE 2: NO INTERACTION BETWEEN SPECIES AND RATE,
POWERS COMPUTED FROM SSH(POPULATION): 0.6,
USING THE BASIS TOTAL SAMPLE SIZE: 240,
AND TOTAL NONREDUNDANT PARAMETERS IN MODEL: 12

		STD DEV			
		0.3		0.5	
		TOTAL N		TOTAL N	
		120	360	120	360
		POWER	POWER	POWER	POWER
TEST TYPE	ALPHA				
REGULAR F	0.01	.22	.72	.07	.25
	0.1	.57	.93	.29	.60
1-TAILED T	0.01	.30	.79	.11	.33
	0.1	.70	.97	.42	.73

REFERENCES

Cohen, J. 1977. Statistical power for the behavioural sciences. Revised edition. Academic Press, Orlando, Fla.

Devore, J.L. 1987. Probability and statistics for engineering and the sciences. 2nd edition. Brooks/Cole, Belmont, Ca.

Keppel, G. 1973. Design and analysis: a researcher's handbook. Prentice-Hall, Englewood Cliffs, N.J.

Korn, E.L. 1990. Projecting power from a previous study: maximum likelihood estimation. The American Statistician 44:290–292.

O'Brien, R.G. 1987. Teaching power analysis using regular statistical software. *In* Proc. 2nd International Conf. on Teaching Statistics. Aug. 1986, Univ. Victoria, Victoria, B.C., pp. 204–211.

______. 1988. FPOWTAB (SAS Version). Univ. Tenn. Statistics Dept. and Computing Cent., Knoxville, Tenn.

Peterman, R.M. 1990a. Statistical power analysis can improve fisheries research and management. Can. J. Fish. Aquat. Sci. 47:1–15.

______. 1990b. The importance of reporting statistical power: the forest decline and acidic deposition example. Ecology 71:2024–2027.

Sanders, W.L. 1989. Use of PROC GLM of SAS (or a similar linear model computing tool) in research planning. HortScience 24:40–45.

SAS Institute Inc. 1985. SAS user's guide: basics. Version 5 edition. SAS Institute Inc., Cary, N.C.

Snedecor, G.W. and W.G. Cochran. 1973. Statistical methods. Iowa State Univ. Press, Ames, Iowa.

Toft, C.A. and P.J. Shea. 1983. Detecting community-wide patterns: estimating power strengthens statistical inference. Am. Nat. 122:618–625.