BIOMETRICS INFORMATION

(You're 95% likely to need this information)

PAMPHLET NO. # 52 DATE: October 6, 1995

SUBJECT: *Post-hoc* Power Analyses for ANOVA F-tests

When an F-test for a treatment effect has a high p-value then there is little evidence for rejection of the (null) hypothesis of no treatment differences. We may not be comfortable with this result, especially if treatment differences were expected or decisions based on the test results need to be made. While the non-significant statistical test tells us that the observed differences are small relative to the background variability (as estimated by the error term) it does not tell us what ability, known as power, the test had to find differences between treatment responses. If the power is high then accepting[1] the null hypothesis is reasonable, but if it is low, a clear decision cannot be made. The power of a test depends upon the size of the treatment differences under consideration.

It is important to ask what treatment differences should be under consideration. For instance, when the power of the test is determined for the *observed F-value* it is the observed treatment differences that are under consideration. The power at this point is only of interest if they are similar in size to a practical difference. A practical difference can be chosen on the basis that treatment differences less than this value would suggest one course of action or decision while larger differences would suggest a different course of action or decision. For instance, in a trial comparing a new fertilizer to the standard one, finding a practical difference in response might imply that the new fertilizer will be used in further operations, while a smaller difference implies that the standard fertilizer will continue to be used. When studying the power of a statistical test, the discussion of the test's power should center on either:

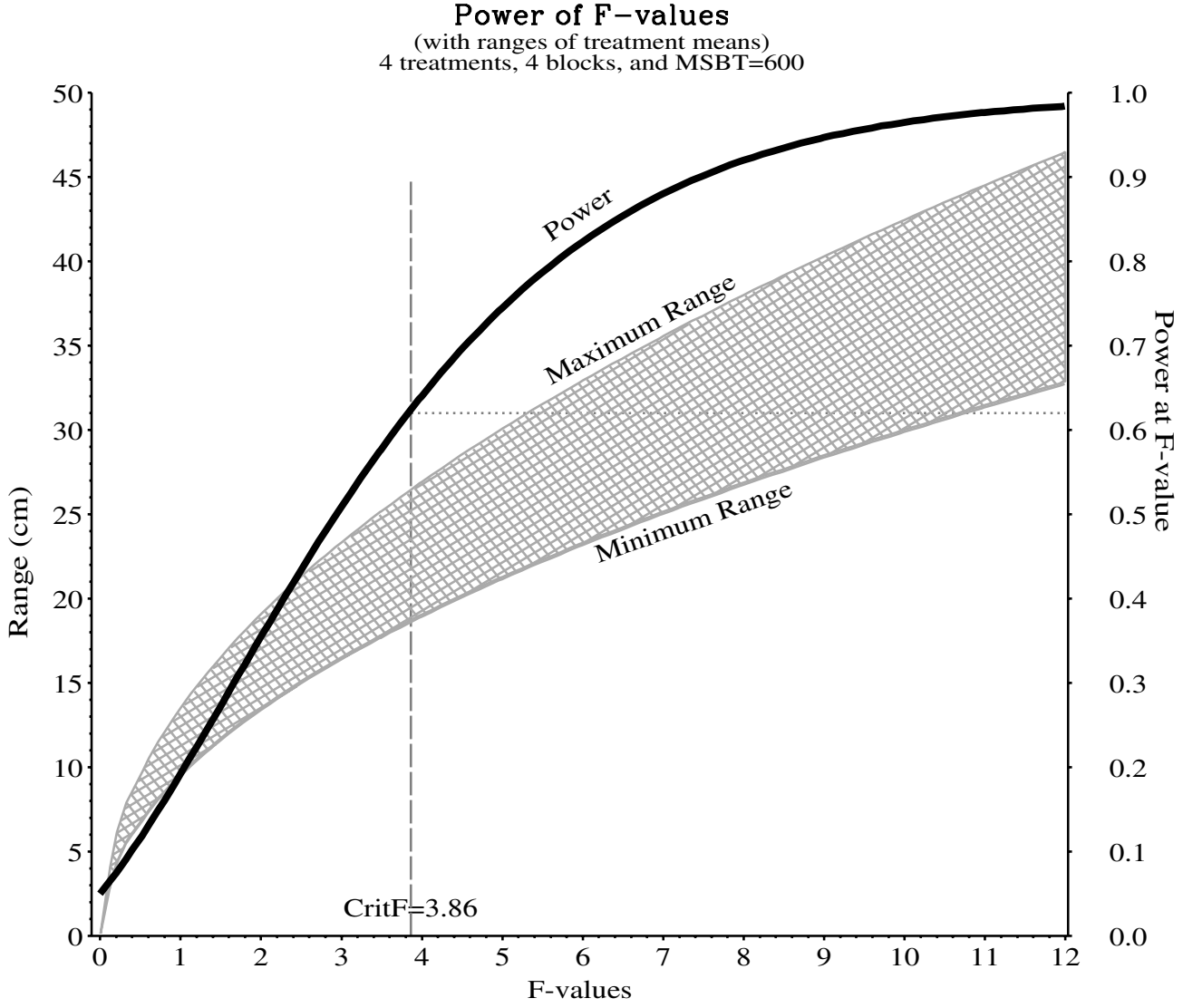
1) the power for a specified alternative hypothesis of practical consequence, *or*

2) the differences that could have been detected at a specified power (e.g. 0.90).

To explore this more fully, let us discuss a Randomized Block example with four blocks (b = 4) and four treatments (t = 4, with degrees of freedom, dfh = t-1). Each treatment level is assigned to one of four plots within each block. Plots are the experimental units and each has been subsampled (with e = 5 subsamples per plot). The degrees of freedom for the treatment test is [(t-1), (b-1)(t-1)] = 3, 9 and the critical F-value at $\alpha = 0.05$ is 3.86. Suppose that the response variable is tree height measured in cm and that a non-significant F-value of 1.389 has been observed with an Error Mean Square of 600 cm^2.

[1] Strictly speaking, hypotheses can only be rejected, not accepted. Nevertheless, to decide future actions, we must choose which hypothesis we will assume is true. It is in this sense then, that we accept an hypothesis.

Before conducting the experiment it may be useful to calculate the power for F-values that we might observe.† The power for an F-test is determined by first calculating the corresponding non-centrality parameter. Recall that the non-centrality parameter[2] is nc = SSH/σ^2 where SSH is the treatment sums of squares of the hypothesized actual means and σ^2 is the known value of the error mean square (blocks by treatment for this example). Since the F-value is calculated by F = [SSH/dfh]/σ^2 it follows that nc = F * dfh. Thus it is straightforward to calculate the theoretical power for each F-value and its degrees of freedom (see Appendix 1 for an example SAS program). The thick line on the graph below is a plot of the power (scale is on the right side) against the F-values for F-tests with 3, 9 degrees of freedom. Note that, for this example, the power for F-values less than the critical value of 3.86 have low power (less than 0.62)[3]. A barely adequate



[2] Note that there are many ways to define the non-centrality parameter. This definition is consistent with SAS.
[3] With larger degrees of freedom, some F-values less than the critial value can have adequate power.
† The following calculations assume that both SSH and σ^2 are known and not estimated from results of the experiment. Otherwise there could be a substantial bias in the calculations (Gerard et al. 1998).

power of 80% occurs with F-values greater than 5.7, while if the F-value is 7.5 or more then the power is at least 90%.

Hypothetical actual differences between treatment means can be associated with these F-values to assist us in our interpretations. The calculations are begun by noting that SSH = sample size * SSM, where SSM is the sums of squares of the hypothesized actual means and sample size is constant at n = b * e for the example. Hence $F = [SSH/dfh]/\sigma^2 = [(b*e*SSM/dfh]/\sigma^2$. This can be rearranged so that $SSM = F*dfh*\sigma^2/(b*e)$. This equation contains the error mean square, σ^2, which is usually unknown. The obvious course of action is to assume that the estimated value of $\hat{\sigma}^2$ (MSBT = 600) is the actual value of σ^2.

While SSM is a direct measure of the variability of the treatment means it would be easier to understand its meaning if we can convert it to differences between the smallest and largest means in the group, called the range. This range will differ for the same value of SSM depending upon the *pattern* of the means in the group (especially the pattern of the non-extreme means as discussed in Appendix 2). Nevertheless, we can determine a minimum and maximum range for each SSM (see Appendix 3 for a description of the calculations and Appendix 4 for an example SAS program). The large shaded area on the graph plots these ranges corresponding to the F-value on the horizontal axis.

This graph is useful for calculating either the power for a specified alternative hypothesis, or the hypothetical differences[4] that could have been detected at a specified power and σ^2. The example F-value of 1.389 corresponds to an SSM of 125 (SSM = 1.389*3*600/4*5) and a range of means between 11.2 cm and 15.3 cm. If a minimum practical difference between any two treatment levels is 35 cm then the observed differences were quite a bit smaller than the practical difference and it is not surprising that the observed F-value is not significant. Differences of 35 or more cm correspond to F-values greater than 7. The corresponding power is at least 88% so this experiment did have sufficient power to detect interesting differences. In this case, if the power of the observed F-value (1.389) had been used to measure the power of the experiment **it would have been decided that the test had had low power when, in fact, it had had sufficient power for the alternative hypothesis of interest** (which, in this case, was **not** the differences between the observed means). On the other hand, this experiment did not have sufficient power to detect hypothesized differences of 20 cm or less. Alternatively, using the second approach, suppose that a power of 90% or more had been desired. This corresponds to an F-value of about 7.5 with the range of hypothesized means between 27 and 39 cm (if $\sigma^2 = 600$). From this point of view, this experiment had sufficient power to detect differences of 39 cm or more, may have had sufficient power for

[4] Note the scale for these differences is determined by σ^2 (estimated by MSBT).

differences between 27 and 39 but did not have sufficient power to detect differences smaller than 27 cm.

In conclusion, graphs of this kind are straightforward to generate for a specific F-test within an experiment and can be used to determine the power of that test. If this power is high for treatment differences of practical consequence then it may be reasonable to accept the null hypothesis. On the other hand, if the power is low then further experimentation may be required, or a decision made on other grounds.

Contact: Wendy Bergerud
387-5676

References:

Bergerud, W.A., and V. Sit, 1992, Power Analysis Workshop Notes. B.C. Ministry of Forests.

Cohen, J., 1977, Statistical Power Analysis for the Behavioral Sciences, Academic Press, N.Y.

Gerard, Patrick D., David R. Smith, and Govinder Weerakkody. 1998. Limits of Retrospective Power Analysis. J. Wildl. Manage. 62:801-807.

Hinkelmann, K. and O. Kempthorne, 1994, Design and Analysis of Experiments: Volume I Introduction to Experimental Design, John Wiley, New York.

Keppel, G., 1982, Design and Analysis: A Researcher's Handbook, 2nd ed., Prentice-Hall, Inc., N.J.

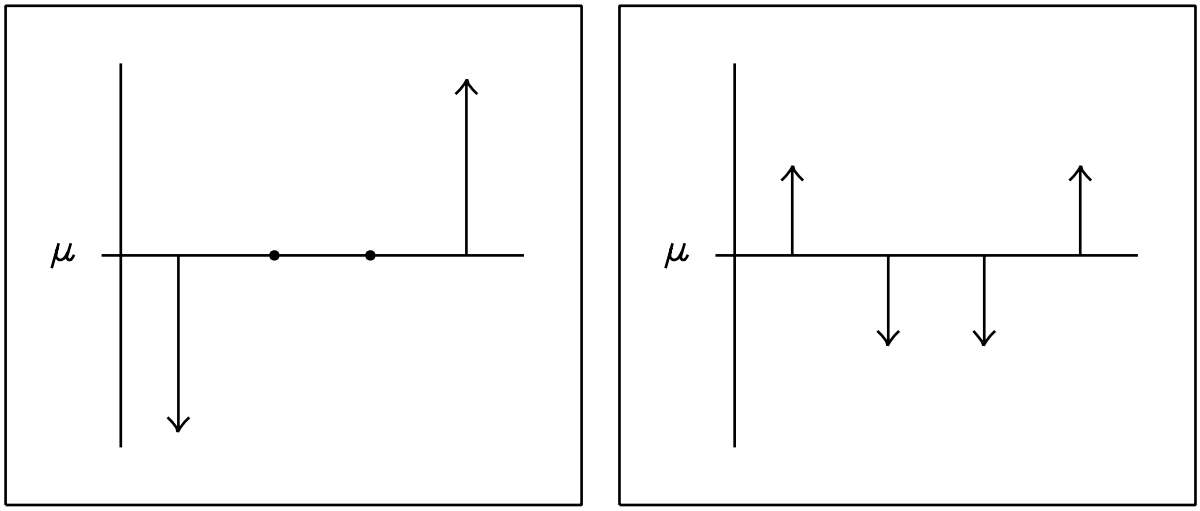
Nemec, Amanda F. Linnell, (1991), Power Analysis Handbook for the Design and Analysis of Forestry Trial, Biometrics Handbook #2, Research Branch, B.C. Ministry of Forests.

Appendix 1: SAS program to calculate the power at observed F-values

```
title 'Correspondence between F-values and their Power';
data fvalpwr;
  t = 4;  dfh = t-1;           ** Treatment or hypothesis degrees of freedom ;
  b = 4;                       ** Number of blocks;
  dfe = (b-1) * (t-1);         ** Error degrees of freedom;
  alpha = 0.05;                ** Alpha level (Type I error) for tests;
  fc = finv(1-alpha,dfh,dfe,0);     ** Critical F-value;
** A range of possible F-values for the horizontal axis;
  do Fval = 0 to 1.2 by 0.2, 1.389, 1.4 to 3 by 0.2, 3.863, 4 to 20 by 1;
    nc = Fval*dfh;             ** Associated non-centrality parameter;
    power = 1-probf(fc,dfh,dfe,nc);  ** Calculation of power;
    output;                    ** Output observation to data set;
end;  run;                     ** End Do loop and run data step;
proc print;
  by dfh dfe alpha fc notsorted; id Fval;  var nc power;
  title2 'Listing of Power for Possible F-values'; run;
proc plot;
     plot power*Fval = 'P' / overlay href = 3.86 ;
title2 'Plot of Power vs F-value';  run;
```

Appendix 2: The relationship between the pattern of means and corresponding ranges for a constant SSM (Sum of Squares of the Means):

To examine this relationship, first recall that the range is the difference between the smallest and largest means. The pattern of the other or middle means[5] determines if the range for a specific SSM will be maximum, minimum, or inbetween. If the middle means have the same value as the grand mean then the middle means contribute zero to SSM and all the variability is split between the two extremes, yielding a maximum range. On the other hand, the means can be divided into two groups with each group having one of the extreme values. The minimum range occurs when the two groups are as equal in size as possible spreading the variability out so that all the means contribute as equally as possible to the value of SSM. While the two groups will be the same size for an even number of means, this will not be the case for an odd number (see the difference in equations 2 & 3 in Appendix 3). These patterns are pictured below for four means.



Any number of other patterns are possible for a group of means, but their ranges will lie between these two extremes. The order of the means does not matter unless the power of contrasts (weighted sums of the means) is under consideration.

Appendix 3: Calculation of the size of the minimum and maximum ranges as a function of SSM.

For a set of t numbers, μ_i, the range is calculated by $\mu_{max} - \mu_{min}$ and their sum of squares by $SSM = \sum(\mu_i - \mu)^2$ where μ is the mean of the μ_i. The relationship between the range and sum of squares depends on whether t is even or odd. If even, the relationship is:

$$\frac{4}{t} * SSM \leq (\mu_{max} - \mu_{min})^2 = range^2 \leq 2 * SSM \qquad (1)$$

[5] If there are two or more means at an extreme then count only one of them as an extreme mean and count the other(s) as middle mean(s).

While if t is odd, the relationship is:

$$\frac{4t}{(t^2-1)} * SSM \le (\mu_{max} - \mu_{min})^2 = range^2 \le 2 * SSM \quad (2)$$

The upper bound on the range is constant whether t is odd or even. These equations describe the relationship between the range of a group of numbers and their sums of squares. They can be used for a set of means by simply substituting the means for the μ_i, and by recalling that SSM = F*dfh*σ^2/n (as noted on page 3).

For a numerical example, suppose that the hypothesized means have an SSM of 125, then the corresponding F = [n*SSM/(t-1)]/MSBT = (20*125/3)/(600) = 1.389. This F-value has a power of about 0.26. The difference between the largest ($\overline{Y}_{max}$) and smallest ($\overline{Y}_{min}$) means will range between the following extreme values:

$$\frac{4}{t} * SSM = 125 \le (\overline{Y}_{max} - \overline{Y}_{min})^2 = range^2 \le 2 * SSM = 250 .$$

Therefore the minimum value of the range is $\sqrt{125}$ = 11.18 cm while the maximum is $\sqrt{250}$ = 15.81 cm. These are the top and bottom values of the cross-hatched area of the graph at F = 1.389.

Derivation: First I noted that in the Power Analysis Workshop Notes (pg. 9) a standardized range of the treatment means is described as: $d = \frac{\mu_{max} - \mu_{min}}{\sigma} = \frac{range}{\sigma}$

For t treatments, where t is even, the non-centrality parameter, nc, has the following values:

$$\frac{nd^2}{2} \le nc \le \frac{tnd^2}{4}, \quad (3)$$

where n is the sample size for each mean (the number of numbers used to calculate the mean). Expression (3) can be written in terms of the sums of squares of the means, SSM, by noting that nc = n*SSM/σ^2 to get:

$$\frac{n*range^2}{2*\sigma^2} \le \frac{n*SSM}{\sigma^2} \le \frac{t*n*range^2}{4*\sigma^2}. \quad (4)$$

Both n and σ^2 drop out and, after some rearranging we obtain equation (1) above. These results are described in Keppel, pg 79 who quotes Cohen (pgs 276 to 280); and in section 6.8.2 of Hinkelmann and Kempthorne (pg 175).

Appendix 4: SAS program to calculate the minimum and maximum ranges.

```
data fvalpwr; set fvalpwr;     ** Data set created by program in Appendix 1;
** Adding the maximum and minimum ranges to the data set;
  n = 20;  ** (n = b*e for the example);   MSBT = 600;   SSM = Fval * dfh * MSBT / n ;
  Minrng = sqrt((4/(dfh+1))*SSM);  Maxrng = sqrt(2*SSM) ;
  powrplt = 60 * power;      ** Use the highest Maxrng value (e.g. 60) here to
        ** scale the printer plot to the same physical size as the plot for Minrng and
        **Maxrng.  Thus a power of .5 is plotted as 30 and 1 as 60;
label   n = 'Sample Size per Mean'    MSBT = 'Error Mean Squares'
   Minrng = 'Minimum Range'         Maxrng = 'Maximum Range'
  powrplt = 'Power plotting values'  power = 'Power' ;  run;
proc print label; by dfh n msbt dfe alpha fc notsorted;
  id Fval; title2 'Listing of Differences and Power'; run;
proc plot;
  plot Minrng*Fval = 'L'  Maxrng*Fval = 'U' powrplt*Fval = 'P'/ overlay href = 3.863;
run;
```