BIOMETRICS INFORMATION

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

PAMPHLET NO. # 51 DATE: September 29, 1995

SUBJECT: Programs for Power Analysis/Sample Size Calculations for CR and RB Designs with Subsampling

This pamphlet describes how to program the calculations for the graphs presented in pamphlet #49. The steps required are described in general terms for use with any statistical software and then a working SAS program is presented. First, recall that decisions have to be made regarding the value or range of values for the following quantities (see both BI # 49 and # 50 for discussion of concepts and notation):

1. The alternate hypothesis, H_A. For discussion purposes, we will continue with the alternate hypothesis that the four treatments have values of $\mu_i = 10, 15, 20, 25$, which can be characterized by the Sums of Squares of the Means: **SSM = 125**. The hypothesis sums of squares (SSH) is equal to the sample size * SSM[1]. The degrees of freedom for the hypothesis, dfh, is equal to number of treatments minus one, i.e. dfh = t - 1 = 3 for the previous examples;
2. Our choice for α = prob(rejecting H_o if it is true) will be the usual value of $\boldsymbol{\alpha = 0.05}$.
   We don't need to choose a value for $1 - \beta$ = prob(rejecting H_o when H_A is true), since we will be plotting this as a function of the other variables;
3. A range of values for the number of subsamples, e to be used, for example, **e = 1, 2, 3, 4, 5, 7, 10, 15, 20, 25, 30, 35 and 40**; and the corresponding component of variation, σ^2_e will be estimated by **MSE = 100, 500, and 1000**;
4. A range of values for the number of plots, p (for CRD), or number of blocks, b (for RBD), will be used, namely, **b** or **p = 2 to 16 by 2 and 20;** and the corresponding component of variation, σ^2_p or $\sigma^2_{p(b)} + \sigma^2_{BT}$ which will be estimated, respectively, by **VarP** or **VarBT = 100 and 500.**

The steps in the calculations for one F-test are described below. The values at each step are recalculated for each change in numbers of subsamples and plots or blocks. Variable names used in the SAS program are in brackets.

1. The Critical F-value (`fc`) for the test is calculated. In SAS, this is done using the `finv` function (the inverse of the cumulative probability of the F-distribution). The necessary parameters are:

[1] for balanced ANOVA's only. Note that the sample size is the total number of numbers used to calculate a treatment mean. For both designs this is the number of subsamples, e, times the number of plots, p, or blocks, b.

BRITISH COLUMBIA
Ministry of Forests
Research Program

i) the numerator or hypothesis degrees of freedom (`dfh`) given the number of treatments, t, which is constant for any particular graph;
ii) the degrees of freedom of the error term (denominator) that is used to test the treatment (`dfp` or `dfbt`). These change with every change in plot or block numbers; and
iii) the chosen value for α (`alpha`).

2. The non-centrality parameter[2] (`nc`) is determined by the ratio of the recalculated values for:
   i) the Treatment Sums of Squares given the Alternate Hypothesis (`ssh`), which changes with each change in both plot or block and subsample numbers and
   ii) the Denominator Mean Square (`msp` or `msbt`), which changes with each change in subsample numbers.

3. The power (`power`) is calculated to take into account the above changes to degrees of freedom and the non-centrality parameter. In SAS, the `probf` (cumulative probability of the F-distribution) function is used for this. It calculates the probability of observing the critical F-value calculated in the first step, given its degrees of freedom and the non-centrality parameter calculated in the second step.

The following SAS program performs these calculations. The code required to include the diamonds, representing a total subsample size of 320, on the graphs is not included. The programs use the following notational changes: `bp` = `p` or `b;` `msbp` = `msp` or `msbt;` `varbp` = `varp` or `varbt`; and `dfbp` = `dfp` or `dfbt.`

```
title 'Power Analysis for either the CR or the RB Design with subsampling';

data power;
** Establishing the values that are constant for this set of graphs:    ;
   dfh = 3;                    ** hypothesis degrees of freedom (= t-1);
 alpha = 0.05;                 ** alpha level for the power plots;
   ssm = 125;                  ** 'size' measure of the alternate hypothesis;

* Do Loops for those variables with a range of values to consider;
  do   mse = 100, 500, 1000;   ** possible values for the Mean Square Error;
  do varbp = 100, 500;  ** possible values for the Plot or B*T component of variation;
  do    bp = 2 to 16 by 2, 20; ** Number of plots or blocks to consider;

* Statements for variables which only change with changing plot or block number;
  * Degrees of Freedom for the Denominator Mean Squares for the Treatment Test ;
  * Only one of the following dfbp assignment statements should be used;
   dfbp = (dfh + 1) * (bp-1);  ** For CR design the Error Mean Square is P(T);
*  dfbp =  dfh  * (bp-1);      ** For RB design the Error Mean Square is B*T;
    fc  = finv(1-alpha,dfh,dfbp,0);  ** Critical F-value for Treatment Test;
```

[2] Recall that the non-centrality parameter measures the size of the alternative hypothesis and can be defined in many ways. For simplicity, I use a definition consistent with SAS. Before using other software, check their definition.

```
* Another Do Loop for the range of subsamples numbers to consider;
  do e = 1 to 5, 7, 10 to 40 by 5;
    ssh   = bp*e*ssm;          ** Hypothesis Sums of Squares for the Trmt Test;
    msbp  = mse + e*varbp;     ** Denominator Mean Square for the Treatment Test;
    nc    = ssh/msbp;          ** Non-centrality parameter for the Trmt Test;
    power = 1-probf(fc,dfh,dfbp,nc);  **  Calculating the power;
    output;                    ** Output this observation to the data set;
  end; end; end; end;          ** Ending all the Do Loops;

* Providing labels for the variables to improve the appearance of the output;
label mse = 'MSE'  varbp = 'VarP or VarBT'
    alpha = 'Alpha'  ssm = 'SSM' ;
run;                           ** Run and create this data set;

* Print the observations created by the data step above;
proc print data = power label;
  by mse varbp alpha ssm notsorted;    id bp e;
title2 'Listing of the Data';
run;

* Creating Printer Plots;
proc plot data = power uniform;
  by mse varbp alpha ssm notsorted;
  plot power*e = bp;
title2 'Plot of the Power for various numbers of plots or blocks and subsamples';
run;
```

This program provides all the output necessary to produce useful printer plots of the power graphs. In most cases, these will be satisfactory for planning suitable sample sizes for proposed studies. Nevertheless, the following program uses SAS/Graph and the data output above to produce report quality graphs.

```
* Creating a optional dataset to add the line labels: bp=2 etc.;
data annote;   set power;
* Defining the lengths of SAS variables for the annotate dataset;
  length function $ 8 text $ 7;
* Establishing variables constant for all observations;
    * The xsys and ysys variables define the axes values as the
    coordinate system for definition of x, y. Function defines
    the action to be taken there: to put a label at x, y ;
  xsys = '2'; ysys = '2'; function = 'label';
    * Size is the size of the text and Position defines that
    the text will end above the x, y coordinate;
  size = 3.0;    position = '1';
* Creating labels for end of line (at greatest subsample no.);
  if e = 40 then do;
* Variables x and y define where the label goes and the text
  is the label to be put at that point;
     x = e; y = power;  text = 'bp='||put(bp,2.0);
     output;  end;    * outputting the labels;
run;
```

optional

```
*  See SAS/Graph manuals for more explanation of the following code:;
goptions reset=all;
* Establishing features of the lines for each number of plots/blocks ;
symbol1  l=1  i=join c=black;  symbol2  l=4  i=join c=black;
symbol3  l=5  i=join c=black;  symbol4  l=1  i=join c=black;
symbol5  l=4  i=join c=black;  symbol6  l=5  i=join c=black;
symbol7  l=1  i=join c=black;  symbol8  l=4  i=join c=black;
symbol9  l=5  i=join c=black;

* Axis descriptions                            ;
axis1  order = 0 to 1 by 0.2  label = (a=90 j=c h=4 'Power')      offset = (0) ;
axis2  order = 0 to 40 by 5   label = (j=c h=4 'Sub-sample Size') offset= (1,0);

* goptions statement to define size of plots etc.     ;
goptions hsize = 7 in vsize = 8.0 in rotate = portrait  horigin = 0 in
         hpos  = 140  vpos  = 160    htitle = 2.8 htext = 2.8 chartype = 5;
* goptions statement to create plots in PostScript on disk     ;
goptions device = ps noprompt gsfmode = replace gsfname = graph;

* Note that the following code stores each graph in its own file.  It is possible
  to produce several graphs at once in one gplot procedure by including:
    by mse varbp notsorted;
* The title statements would also need to be changed and the annotate data set
  may not work as you would expect;

filename graph 'crdpwr1.ps';
proc gplot data = power annotate = annote;
title h=5 'MSE = 100, VarP = 100, Alpha = 0.05, SSM = 125';
  where mse = 100 and varbp = 100;
  plot power * e = bp / nolegend vaxis = axis1 haxis = axis2 ;
run;  quit;
filename graph 'crdpwr2.ps';
proc gplot data = power annotate = annote;
title h=5 'MSE = 100, VarP = 500, Alpha = 0.05, SSM = 125';
  where mse = 100 and varbp = 500;
  plot power * e = bp / nolegend vaxis = axis1 haxis = axis2 ;
run;  quit;
```

and etc. for the rest of the graphs.

```
* Note: the repeated redefinition of the same filename in Version 6.10 is only possible
by quitting the previous proc plot (not necessary in Version 6.04!).  Otherwise the
filename is still in use and cannot be redefined;

*  This program ran fine in Version 6.04 but Version 6.10 is does not behave as
expected.  For instance, the Where statement works on the annotate dataset as expected
in 6.04 but is apparently ignored by 6.10, so that all the labels come out on all the
graphs (hardly useful)!  Separate annotate datasets had to be created for each graph.
```

Contact: Wendy Bergerud
387-5676