Paper 202-26

Probabilistic Hindsight: A SAS Macro for Retrospective Statistical Power Analysis

Kristine Y. Hogarty and Jeffrey D. Kromrey

Department of Educational Measurement and Research, University of South Florida

ABSTRACT

Statistical power analysis is useful from both prospective and retrospective viewpoints. Prospectively, statistical power analysis is used in the planning of research to estimate sample size requirements. In contrast, retrospective power analysis provides an estimate of the statistical power of a hypothesis test after an investigation has been conducted. Several techniques for retrospective power analysis have been suggested in the statistical literature, including both point and interval estimates of power. This paper presents a SAS macro that calculates three point estimators (a plug-in estimate, an *unbiased* estimate and a median unbiased estimate) and a confidence band estimate of the power of a hypothesis test using sample estimates of the noncentrality parameter of the test statistic. The macro was written to evaluate the power of an *F* test (obtained through an analysis of variance, multiple regression analysis, or any of several multivariate analyses). Inputs to the macro include the obtained value and degrees of freedom of *F*, the total sample size, and the nominal alpha level of the hypothesis test. Macro outputs include the four estimates of statistical power as well as the sample effect size. The paper provides a demonstration of the SAS/IML® code and examples of the application of the code in simulation studies.

INTRODUCTION

In many ways, retrospective power analysis is more complicated than prospective power analysis. Recent literature suggests that retrospective power analysis is conceptualized in two very different forms. Characteristic of one approach, Zumbo and Hubley (1998) and Ottenbacher and Maas (1999) present Bayesian power estimation techniques directed at determining the probability of the null hypothesis being false, given that the null has been rejected, that is *Pr (H_o=false|rejected H_o)*. While this probability is of importance in applied research, it's practical applications appear to be limited because of the unknown proportions of true and false null hypotheses in any field of inquiry (Zumbo & Hubley, 1998). This approach also introduces a different formal definition of "power" than is typically considered in inferential statistics (i.e., power usually represents *Pr (H_o will be rejected|H_o=false)* which is equal to 1 – β). These two probabilities are typically very different. Because this conceptualization of retrospective power is not practical, it will not be further addressed here.

The second approach to retrospective power analysis (Gerard, Smith & Weerakkody, 1998; Steiger & Fouladi, 1997; Brewer & Sindelar, 1987) aims to estimate the statistical power of a hypothesis test after the test has been conducted. That is, information obtained from a particular study may be used to estimate the population effect size, which in turn may be used (in concert with the study's sample size and nominal alpha level) to estimate the power under which the research was conducted. This approach to retrospective power analysis appears to satisfy a practical need in applied research and retains the familiar formal definition of power (i.e., 1 – β). As applied researchers, we have been urged to consider the effect sizes associated with our data (e.g., Kirk, 1996; Harlow, Muliak & Steiger, 1997), in conjunction with the reject/fail-to-reject decisions of our hypothesis tests. The second approach to retrospective power analysis simply extends our use of sample effect sizes to provide estimates of power. However, the estimation of statistical power based on a sample effect size is characterized by considerable controversy.

ESTIMATION PROCEDURES

Several techniques for the second approach to retrospective power analysis have been suggested in the literature. Gerard, Smith and Weerakkody (1998) describe three estimates of noncentrality that lead to point estimates of retrospective power: a "plug-in estimator" (λ_p), an "unbiased estimator" (λ_{ub}), and a "median unbiased" or "percentile estimator" (λ_{50}).

The plug-in estimator simply represents the use of the sample noncentrality parameter (λ_p) as if it were the same as the population parameter. For the F distribution, the sample noncentrality parameter is given by

$$\lambda_p = v_1 F$$

where v_1 = numerator degrees of freedom for the sample F, and
F = obtained sample *F* statistic.

The obtained sample noncentrality parameter is then used to estimate the statistical power of the test

$$Power = \Pr(F_{v_1, v_2, \lambda_p} \geq F_{v_1, v_2, 1-\alpha})$$

where F_{v_1, v_2, λ_p} = the noncentral F distribution with v_1 and v_2 degrees of freedom and a noncentrality parameter λ_p, and

$F_{v_1, v_2, 1-\alpha}$ = the $(1-\alpha)$ percentile of central *F*-distribution (i.e., the critical value of *F* with v_1 and v_2 degrees of freedom).

The sample noncentrality parameter is a function of both the sample effect size (Cohen, 1988) and the sample size. The sample effect size is a descriptive statistic used to represent the strength of relationship between the independent and dependent variables. In the context of the analysis of variance, Cohen's effect size f is related to λ_p by the equations

$$f = \sqrt{\frac{\lambda_p}{N}}$$

and

$$\lambda_p = Nf^2$$

where *N* = total sample size.

The use of λ_p is known to produce biased estimates of power with a distinct positive bias in conditions of low power (Johnson et al., 1995). Johnson et al. suggested an alternative estimator (λ_{ub}) intended to reduced the bias inherent in λ_p. This *unbiased* estimator of noncentrality is given by

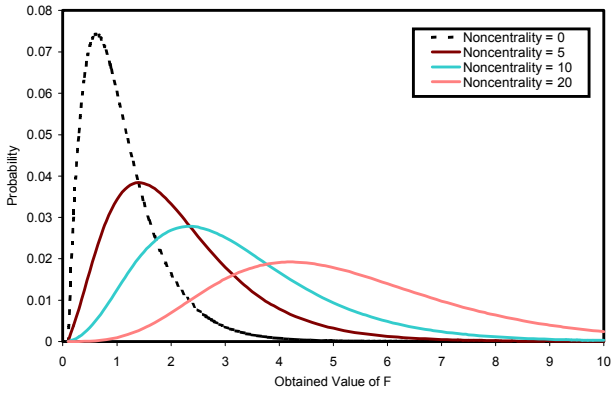
$$\lambda_{ub} = \frac{v_1 (v_2 - 2) F}{v_2} - v_1$$

Although λ_{ub} may provide an unbiased estimate of the population noncentrality, estimates of power derived from unbiased noncentrality estimates are not necessarily unbiased themselves, because power is a nonlinear function of noncentrality (Gerard et al., 1988).

A third point estimate of noncentrality was suggested by Taylor and Muller (1996). This approach (λ_{50}) is reported to underestimate noncentrality 50% of the time and overestimate it 50% of the time (hence, Gerard et al., 1998, refer to the method as "median unbiased"). This method makes use of the cumulative distribution function of F and seeks the value of noncentrality for which the obtained value of F in a particular study (i.e., with a given v_1 and v_2) is expected 50% of the time. Because analytical formulae for solving this problem are not available, the value of noncentrality must be obtained by numerical methods (see, for example, Press, Teukolsky, Vetterling & Flannery, 1992).

An illustration of this method of noncentrality estimation is provided in Figures 1 and 2. Assume that a research analysis yields a sample *F* statistic of 5.20, with 5 and 39 degrees of freedom. The sampling distribution of *F* is graphed in Figure 1 for the null case (in which population noncentrality = 0) and for three non-null conditions (population noncentrality > 0). Although a value of F as large as 5.20 is quite rare under the null hypothesis, this value becomes more common as noncentrality increases.

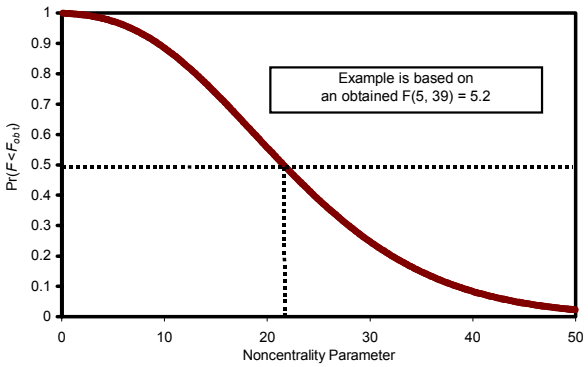
Figure 1
Central and Noncentral F Distributions df = 5, 39



Given the obtained value of *F* from the sample, one can compute the proportion of each sampling distribution that is less than this obtained F value. Extending this thinking to an infinite number of noncentrality values (rather than the four illustrated in Figure 1), one can plot the proportion of the *F* distribution that is less than the obtained *F* value, $\Pr(F < F_{obt})$ as a function of noncentrality. This graph

is provided in Figure 2. The median unbiased estimate of noncentrality (λ_{50}) is that noncentrality value for which $Pr(F < F_{obt}) = .50$.

Figure 2
Probability of $F<F_{obt}$ as a Function of Population Noncentrality



In contrast to the point estimates suggested by Gerard et al. (1998), Steiger and Fouladi (1997) presented an interval estimation approach based on the earlier work of Hedges and Olkin (1985). This approach provides confidence bands on the noncentrality parameter (noncentrality interval estimates) which subsequently may be used to obtain confidence bands on statistical power. Using logic analogous to that used to obtain the λ_{50} point estimate, the approach involves the inversion of percentiles from noncentral sampling distributions to obtain confidence bands around the noncentrality parameter. That is, instead of seeking the value of noncentrality expected 50% of the time, a 95% confidence band is obtained by seeking the value of noncentrality $\left(\hat{\lambda}\right)$ for which $\Pr(F_{v_1,v_2,\hat{\lambda}} < F_{obt})$ = .025 and the value for which $\Pr(F_{v_1,v_2,\hat{\lambda}} < F_{obt})$ = .975.

This provides a confidence band for noncentrality, the endpoints of which are transformed into the endpoints of a 95% confidence band for statistical power. A simulation study by Kromrey and Hogarty (2000) suggested estimation problems with these confidence bands when power values were near unity. As a correction, these authors recommended the use of one-sided confidence intervals (e.g., "I am 95% certain that the power is greater than .978") for such conditions.

AN EXAMPLE

A SAS macro was designed to compute the three point estimates of power and both two-sided and one-sided confidence bands for interval estimates. The macro was designed to be executed as a stand alone program (with manual input of an obtained value of *F*, sample size, degrees of freedom and alpha level) or modified to run as part of a larger SAS job stream.

MACRO RETR_PWR

The macro RETR_PWR calculates a plug-in estimate, an *unbiased* estimate and a median unbiased estimate of retrospective power. This macro also computes both two-sided and one-sided confidence bands for estimating the power of a hypothesis test using sample estimates of the noncentrality parameter of the test statistic.

The inputs to the macro include the obtained value of *F*, the numerator and denominator degrees of freedom, the total sample size and the nominal alpha level of the hypothesis test. The critical value of *F* is obtained using the FINV function in conjunction with the degrees of freedom and alpha level input to the macro. Power estimates are subsequently obtained using the PROBF function. Error trapping is incorporated into the macro by checking for negative values of noncentrality (which are invalid arguments to the PROBF function) and resetting these to zero. In addition, very large values of noncentrality are not valid arguments for the PROBF function. To overcome this limitation, the smallest value of noncentrality that provides power of approximately 1.00, given the sample degrees of freedom and alpha level is obtained (the scalar MIN_NCC in the macro). Noncentrality estimates that exceed this value are reset to MIN_NCC before the PROBF function is called.

The plug-in and *unbiased* estimates of noncentrality are obtained using the formulas presented earlier. The median unbiased point estimate and the interval estimates require computation of percentiles from the sampling distributions of *F* (as illustrated in Figure 2) that are obtained from the FIND_NC subroutine.

The subroutine FIND_NC computes the noncentrality parameter corresponding to a requested percentile that is supplied as an argument (e.g. see Figure 2). These estimates are used in the calculation of the median unbiased point estimate of power and all of the interval power estimates. The algorithm uses a simple bracketing approach that computes a noncentrality value larger than the target noncentrality sought (the scalar LOW in the subroutine) and a noncentrality value that is smaller than the target noncentrality (the scalar HIGH). A noncentrality of zero is used for the smaller value. The interval defined by these points is successively

halved until the desired precision in the estimate is obtained (precision is set in the scalar SMALL). For conditions with very large power (values of noncentrality that are far to the right in Figure 2), large changes in noncentrality yield only small changes in the cumulative proportions. Such sample conditions do not converge with this algorithm. For these conditions, the convergence criterion SMALL is increased after 1500 halving iterations and increased a second time after an additional 1500 iterations. The output of the macro (obtained via the FILE PRINT command) is presented in Table 1.

```
%macro retr_pwr(obt_F,df_num,df_denom,total_N,alpha);

proc iml;

start FIND_NC(F_obt,u,v,ncc,pctl);
  OK = 0;
  nc = 0;
  target = pctl;
  do until (OK = 1);
            cumprob = PROBF(F_obt,u,v,nc);
            if cumprob<target then OK = 1;
            if cumprob>target then nc = nc + 3.0;
  end;
  low = nc;
  high = 0;

  change = 1;
  loop = 0;
  small = .00000000001;
  do until (change<small);
         half = (high + low)/2;
         cum_h = PROBF(F_obt,u,v,half);
         if cum_h < pctl then do;
             low =  half;
        end;
        if cum_h > pctl then do;
             high = half;
        end;
        change = abs(high - low);
        loop = loop + 1;
        if loop > 1500 then small = .000000001;
        if loop > 3000 then small = .01;
  end;
  ncc = (high + low)/2;
finish;

* +-----------------------------------------------------------------+
            Estimates of Noncentrality Parameters
* +-----------------------------------------------------------------+;

* +-----------------------------------------------------------------+
                          Point Estimates
* +-----------------------------------------------------------------+;

  nc_p = &df_num#&obt_F;
  nc_ub= (&df_num/&df_denom)#(&df_denom-2)#&obt_F -
             &df_num;

  run FIND_NC(&obt_F,&df_num,&df_denom,nc_50,0.50);

* +-----------------------------------------------------------------+
                  Endpoints of Interval Estimates
* +-----------------------------------------------------------------+;

  run FIND_NC(&obt_F,&df_num,&df_denom,nc_25,0.25);
  run FIND_NC(&obt_F,&df_num,&df_denom,nc_75,0.75);

  run FIND_NC(&obt_F,&df_num,&df_denom,nc_10,0.10);
  run FIND_NC(&obt_F,&df_num,&df_denom,nc_90,0.90);

  run FIND_NC(&obt_F,&df_num,&df_denom,nc_025,
                    0.025);
  run FIND_NC(&obt_F,&df_num,&df_denom,nc_975,
                    0.975);

  run FIND_NC(&obt_F,&df_num,&df_denom,nc_20,0.20);
  run FIND_NC(&obt_F,&df_num,&df_denom,nc_80,0.80);

  run FIND_NC(&obt_F,&df_num,&df_denom,nc_05,0.05);
  run FIND_NC(&obt_F,&df_num,&df_denom,nc_95,0.95);

* +-----------------------------------------------------------------+
    Convert Noncentrality Estimates to Power Estimates
  +-----------------------------------------------------------------+;
   alpha1 = 1 - &alpha;

* +-----------------------------------------------------------------+
                        Critical Value of F
  +-----------------------------------------------------------------+;

   fc=finv(alpha1,&df_num,&df_denom);

* +-----------------------------------------------------------------+
        Replace Negative Noncentrality Values with Zero
  +-----------------------------------------------------------------+;

   if nc_p    < 0 then nc_p    = 0;
   if nc_ub   < 0 then nc_ub   = 0;
   if nc_50   < 0 then nc_50   = 0;
   if nc_25   < 0 then nc_25   = 0;
   if nc_75   < 0 then nc_75   = 0;
   if nc_10   < 0 then nc_10   = 0;
   if nc_90   < 0 then nc_90   = 0;
   if nc_025 < 0 then nc_025 = 0;
   if nc_975 < 0 then nc_975 = 0;

   if nc_20  < 0 then nc_20  = 0;
   if nc_80  < 0 then nc_80  = 0;
   if nc_05  < 0 then nc_05  = 0;
   if nc_95  < 0 then nc_95  = 0;

* +-----------------------------------------------------------------+
        Find Lowest Noncentrality Parameter that Yields
        Power=1.00
  +-----------------------------------------------------------------+;
     min_ncc = 0;
     OK=0;
     do until (OK = 1);
       min_1 = 1 - PROBF(fc,&df_num,&df_denom,
                              min_ncc);
       if min_1 < 1.00 then min_ncc = min_ncc + 1;
       if min_1 > 0.9999999999 then OK = 1;
     end;

    * +-----------------------------------------------------------------+
```

```
                Power using Plug-in Estimator
  +-------------------------------------------------------------------+;
  if nc_p < min_ncc then do;
     p_p= 1 - PROBF(fc,&df_num,&df_denom,nc_p);
  end;
  else p_p = 1.00;

* +-------------------------------------------------------------------+
                Power using Unbiased Estimator
  +-------------------------------------------------------------------+;
  if nc_ub < min_ncc then do;
     p_ub= 1 - PROBF(fc,&df_num,&df_denom,nc_ub);
  end;
  else p_ub = 1.00;

* +-------------------------------------------------------------------+
                Power using Median Estimator
  +-------------------------------------------------------------------+;
  if nc_50 < min_ncc then do;
     p_50= 1 - PROBF(fc,&df_num,&df_denom,nc_50);
  end;
  else p_50 = 1.00;

* +-------------------------------------------------------------------+
          Two-sided Confidence Bands for Power
  +-------------------------------------------------------------------+;
  if nc_25 < min_ncc then do;
     p_25= 1 - PROBF(fc,&df_num,&df_denom,nc_25);
  end;
  else p_25 = 1.00;
  if nc_75 < min_ncc then do;
     p_75= 1 - PROBF(fc,&df_num,&df_denom,nc_75);
  end;
  else p_75 = 1.00;

  if nc_10 < min_ncc then do;
     p_10= 1 - PROBF(fc,&df_num,&df_denom,nc_10);
  end;
  else p_10 = 1.00;
  if nc_90 < min_ncc then do;
     p_90= 1 - PROBF(fc,&df_num,&df_denom,nc_90);
  end;
  else p_90 = 1.00;

  if nc_025 < min_ncc then do;
     p_025= 1 - PROBF(fc,&df_num,&df_denom,nc_025);
  end;
  else p_025 = 1.00;
  if nc_975 < min_ncc then do;
     p_975= 1 - PROBF(fc,&df_num,&df_denom,nc_975);
  end;
  else p_975 = 1.00;

* +-------------------------------------------------------------------+
          One-sided Confidence Bands for Power
  +-------------------------------------------------------------------+;
  if nc_20 < min_ncc then do;
     p_20= 1 - PROBF(fc,&df_num,&df_denom,nc_20);
  end;
  else p_20 = 1.00;
  if nc_80 < min_ncc then do;
     p_80= 1 - PROBF(fc,&df_num,&df_denom,nc_80);
  end;
  else p_80 = 1.00;
     if nc_05 < min_ncc then do;
        p_05= 1 - PROBF(fc,&df_num,&df_denom,nc_05);
     end;
     else p_05 = 1.00;
     if nc_95 < min_ncc then do;
        p_95= 1 - PROBF(fc,&df_num,&df_denom,nc_95);
     end;
     else p_95 = 1.00;


   * +-------------------------------------------------------------------+
                      Cohen f Effect Size
     +-------------------------------------------------------------------+;

           Cohen_f = SQRT(nc_p/&total_N);

* +-------------------------------------------------------------------+
                Print Resulting Power Estimates
   +-------------------------------------------------------------------+;
file print;
 put @1 '------------------------------' /
     @1 'Retrospective Power Estimates' /
     @1 '------------------------------' //
     @1 'Sample Statistics'/
     @1 '-----------------'/
     @5 'Obtained F' @40 &obt_F Best7. /
     @5 'Numerator df' @40 &df_num Best7. /
     @5 'Demoninator df' @40 &df_denom Best7. /
     @5 'Alpha Level' @40 &alpha Best7. /
     @5 'Critical F'  @40 fc Best7. /
     @5 'Cohen''s f' @40 Cohen_f Best7. //
     @1 'Point Estimates' /
     @1 '---------------' /
     @5 'Plug-in' @40 p_p 7.5 /
     @5 'Unbiased' @40 p_ub 7.5 /
     @5 'Median Unbiased' @40 p_50 7.5 //
     @36 'Limits' /
     @31 '----------------' /
     @1 'Two-sided Confidence Bands' @32 'Lower
Upper' /
     @1 '-------------------------' @31 '-------  -------' /
     @5 '50%' @31 p_75  7.5 @40 p_25  7.5 /
     @5 '80%' @31 p_90  7.5 @40 p_10  7.5 /
     @5 '95%' @31 p_975 7.5 @40 p_025 7.5 //
     @1 'One-sided Confidence Bands: Lower' @41 'Limit' /
     @1 '---------------------------------' @40 '-------' /
     @5 '80%' @40 p_80  7.5 /
     @5 '95%' @40 p_95  7.5 //
     @1 'One-sided Confidence Bands: Upper' @41 'Limit' /
     @1 '---------------------------------' @40 '-------' /
     @5 '80%' @40 p_20  7.5 /
     @5 '95%' @40 p_05  7.5;
quit;
%mend retr_pwr;
```

OUTPUT

The macro is called by supplying an obtained value of *F* with its degrees of freedom, the total sample size and the alpha level for the hypothesis test. For example, a four-group experiment with a total sample size of 40 provides an *F* statistic with 3 and 36 degrees of freedom. If the obtained value of *F*

from the sample is 2.578 and the hypothesis is tested using an alpha level of .05, the following statements will invoke the macro and produce the output illustrated in Table 1.

```
%RETR_PWR(2.578,3,36,40,.05)
run;
```

Table 1

Retrospective Power Estimates

Sample Statistics		
Obtained F		2.578
Numerator df		3
Denominator df		36
Alpha level		.05
Critical F		2.86627
Cohen's f		0.43972
Point Estimates		
Plug-in		0.58559
Unbiased		0.34739
Median Unbiased		0.43970
	Limits	
Two-sided Confidence Bands	Lower	Upper
50%	0.20442	0.69879
80%	0.07534	0.86986
95%	0.05000	0.96417
One-sided Confidence Band: Lower		Limit
80%		0.16023
95%		0.05000
One-sided Confidence Band: Upper		
80%		0.75411
95%		0.93157

The point estimates of the power of this *F* test range from 0.34739 for the *unbiased* estimate, to 0.58559 for the plug-in estimate. The median unbiased power estimate is 0.43970. For this sample, Cohen's ***f*** is estimated to be 0.43972, indicating a relatively large effect. The two-sided confidence bands for this sample illustrate the extent of uncertainty in retrospective power estimation. The 95% confidence interval indicates that the researcher can be 95% sure that the actual power of the test is between 0.05 (the alpha level of the test) and 0.96417, an interval nearly as wide as the possible range of statistical power. The use of lower levels of confidence provides somewhat narrower bands. For example, with these data the researcher can be 80% sure that the power of the test is between 0.07534 and 0.86986; and can be 50% sure that the actual power of the test is between 0.20442 and 0.69879.

For many research applications, one-sided confidence bands may be preferred. For example, using a lower tail confidence interval, the data illustrated in Table 1 indicate that the researcher may be 80% certain that the power of the test is greater than 0.16023. Conversely, using an upper tail confidence interval, the researcher may be 80% certain that the power of the *F* test is less than 0.75411.

A second example, (see Table 2), provides retrospective power estimates under conditions of relatively high power. In this case, a three-group ANOVA with a total sample size of 1697 provides an *F* statistic with 2 and 1694 degrees of freedom. If the obtained value of *F* from the sample is 10.0 and the hypothesis is tested using an alpha level of .05, the following statements will invoke the macro and produce the output illustrated in Table 2.

```
%retr_pwr(10.0,2,1694,1697,.05)
run;
```

In this example, the point estimates of the power of this *F* test range from 0.97408 for the *unbiased* estimate, to 0.98507 for the plug-in estimate. The median unbiased power estimate is 0.98027. Cohen's ***f*** was calculated to be 0.10856, suggesting a medium effect size. The two-sided confidence bands for this sample illustrate the notable reduction of uncertainty in retrospective power estimation with large samples. The 95% confidence interval indicates that the researcher can be 95% sure that the actual power of the test is between 0.54656 and 0.99997, an interval somewhat smaller than half the possible range of statistical power. Similarly, the 80% two-sided confidence band limits (0.78425 and 0.99958) indicate a band of moderate width, and a one-sided 80% confidence band suggests that the researcher may be 80% certain that the power of the hypothesis test is greater than 0.88928. The smaller confidence interval widths obtained from this larger sample are illustrative of the type of condition in which retrospective power analysis may be an

informative and useful procedure for the evaluation of research results.

Table 2

Retrospective Power Estimates

Sample Statistics		
Obtained F		10.0
Numerator df		2
Denominator df		1694
Alpha level		.05
Critical F		3.001
Cohen's f		0.10856
Point Estimates		
Plug-in		0.98507
Unbiased		0.97408
Median Unbiased		0.98027
	Limits	
Two-sided Confidence Bands	Lower	Upper
50%	0.91753	0.99684
80%	0.78425	0.99958
95%	0.54656	0.99997
One-sided Confidence Band: Lower		Limit
80%		0.88928
95%		0.66548
One-sided Confidence Band: Upper		
80%		0.99812
95%		0.99989

The macro as provided can be used to obtain retrospective power estimates for *F* tests computed from analysis of variance and regression analysis, as well as the many multivariate test statistics that yield exact or approximate *F* statistics. The code is easily modified to obtain different degrees of confidence in the interval estimates (e.g., 70% or 90% confidence bands if these are desired). Such a modification can be made by changing the percentile values used in the call to the FIND_NC subroutine. In addition, the strategy employed in this macro can be applied to other test statistics, such as the chi-square test, by using the noncentrality formulas for the desired test (see, for example, Johnson et al., 1995). Finally, the macro can be employed in conjunction with SAS procedures such as PROC GLM, by creating an output data set from GLM and using the *F* statistics and degrees of freedom in the output data set as arguments for the macro. This linking will provide analysis of variance test results and retrospective power estimates in the same program without requiring a separate manual input of the obtained *F* and degrees of freedom.

For example, the following code creates an output data set (OUT_F) from PROC GLM and assigns the obtained value of *F*, the degrees of freedom, and the sample size to SAS macro variables (FF, DF1, and DF2, TOTAL_N) using the SYMPUT function. These macro variables are then used, in conjunction with the desired alpha level, as inputs to the macro RETR_PWR.

```
proc glm outstat=out_F;
  class x;
  model y = x / ss3 ;

data two;
  set out_F;
  if _type_ ='ERROR' then dfdenom = df;
  retain dfdenom;
  if _type_ = 'SS3';
  n = df + dfdenom + 1;
  call symput('FF',trim(left(F)));
  call symput('df1',trim(left(df)));
  call symput('df2',trim(left(dfdenom)));
  call symput('Total_N',trim(left(n)));

%RETR_PWR(&FF,&df1,&df2,&Total_N,.05)

run;
```

When this strategy is employed a minor modification of the code within the macro RETR_PWR is necessary. Because macro variables are used as inputs to RETR_PWR, the references to the input variables within RETR_PWR must be represented with a double ampersand (i.e., &&obt_F instead of &obt_F).

DISCUSSION

In retrospective power estimation, as with any application of statistical estimation, researchers should be concerned with both the accuracy and precision of the estimates employed. For point estimates of retrospective power, accuracy is reflected in the closeness of the average sample estimate to the true power of the hypothesis test (i.e., the statistical bias of the estimate) and

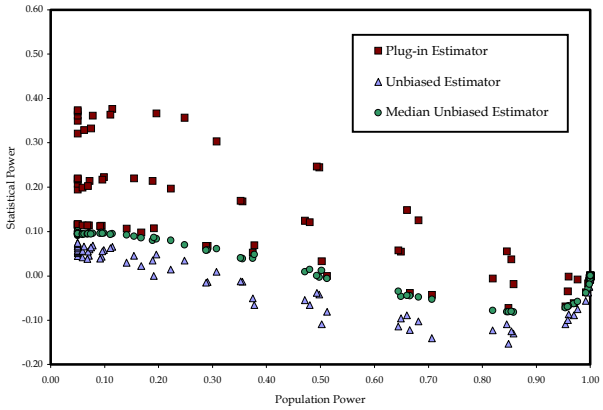
precision is reflected in the standard error of the power estimates (i.e., the standard deviation of the power estimates across repeated samples of the same size from the same population). For interval estimation of retrospective power, the accuracy of the estimates is indicated by the proportion of confidence intervals that contain the true population power (e.g., 95% confidence intervals should contain the true power value 95% of the time) and precision is indexed by the width of the resulting confidence interval (with tighter confidence bands representing more precise estimates). Because analytic solutions to questions of accuracy and precision of retrospective power estimates are not available, Monte Carlo methods may be used to study the behavior of these statistics.

A series of empirical investigations of the behavior of these retrospective power estimates under a variety of conditions typically encountered in educational research were conducted by Kromrey and Hogarty (2000, in press). The results of these studies suggest that none of the estimation techniques were effective across all conditions examined. For the point estimates, the *unbiased* and median unbiased estimators showed improved performance relative to the plug-in estimator, but these procedures were not completely free from bias except under large sample sizes and large effect sizes (as the statistical power approaches unity).

Figure 3 presents the statistical bias observed by Kromrey and Hogarty (2000, in press) in the three point estimates of retrospective power. Five thousand samples of various sizes, ranging from 5 to 100 per cell, were randomly generated in these investigations for each of a variety of population effect size conditions. Additionally, both balanced and unbalanced conditions were examined for both single factor and factorial ANOVA designs. For each sample, the point estimates of retrospective power estimates were obtained and were compared to the true power of the *F* test. The figure illustrates the differences between each retrospective power estimate and the true power as a function of true power. As seen in the figure, under conditions of low power, the plug-in estimate evidences substantial positive bias (i.e., a tendency to overestimate the true power). The *unbiased* and median unbiased power estimates evidence substantially less bias under these low power conditions. However, under conditions of moderate to high power, these estimates showed a negative bias (i.e., a tendency to underestimate the true power). It is interesting to note that the same pattern was evidenced for both balanced and unbalanced designs, but for unbalanced designs the bias estimates were, in general, slightly larger.

Figure 3

Statistical Bias in Point Estimates of Retrospective Power

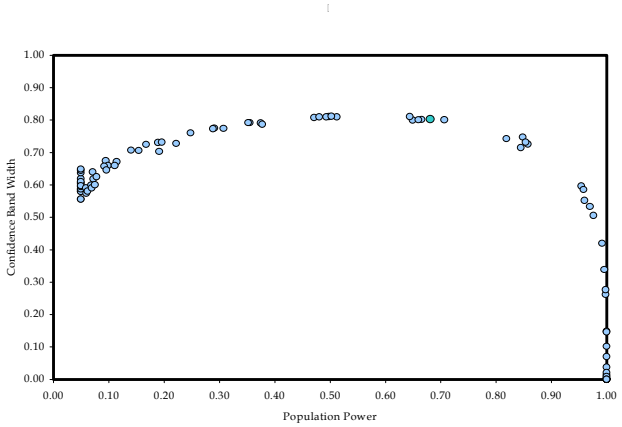


The confidence band approach suggested by Steiger and Fouladi (1997) provided excellent coverage of the parameter in many conditions (i.e., the 95% confidence bands contained the true power approximately 95% of the time). This method appears to be a wise choice (because it is unbiased). However, the width of the resulting confidence bands that provide such excellent coverage are typically so broad that they provided little information about the true power of the study. Only with relatively large samples and large effect sizes did the band width become small enough to be useful for research applications. For researchers who have the luxury of working with very large samples, these bands appear to be the best approach to retrospective power analyses.

An example of the confidence band width results from Kromrey and Hogarty (2000, in press) is provided in Figure 4. This figure illustrates the average width of 95% confidence bands obtained from 5000 samples generated under a variety of sample sizes and population effect sizes. The average confidence band widths are plotted in the figure as a function of population power. As seen in this figure, 95% confidence bands obtained from conditions in which the true population power is low or moderate range in width from 0.50 (half the range of possible power values) to nearly .80. The information provided by such large confidence intervals is quite limited. Under conditions of high power, however, the band width becomes smaller, making the confidence bands more informative for the interpretation of research.

Figure 4

Interval Estimate of Retrospective Power:
95% Confidence Band Width



The broad confidence bands obtained with this method indicate the extent to which the use of sample statistics to estimate power is characterized by a great deal of uncertainty. The use of smaller levels of confidence (e.g., 80% bands or 50% bands instead of 95% confidence intervals) and one-sided confidence intervals can improve the usefulness of this method while continuing to indicate the degree of uncertainty that is inherent in retrospective statistical power estimation.

Although prospective power analysis is of critical importance in the planning of empirical investigations, retrospective power analysis is important for both the interpretation of research results and the planning of subsequent studies, hence it is a logical extension of the substantive interpretation of sample effect sizes. However, retrospective power analysis has received little attention in the research methods literature. The results of earlier investigations (Kromrey & Hogarty, 2000, in press) suggest that the currently available methods for retrospective power analysis evidence severe limitations (except for studies with large sample sizes). Such results highlight the importance of the caveats that should be employed when researchers use retrospective power estimates. Additionally, these results suggest that improved methods of estimation appear to be necessary to supply researchers with an important tool that can be trusted to provide unbiased and precise estimates of retrospective power across conditions typically encountered in applied research.

This macro is provided to facilitate researchers' calculation, interpretation and use of estimates of retrospective power. The macro was designed to be executed as a stand alone program or in conjunction with SAS procedures such as PROC GLM. The program is also easily modified to obtain different degrees of confidence for both the two-sided and one-sided interval estimates. Lastly, estimates of retrospective power for factorial designs may be obtained by invoking the macro for each of the specific effects of interest.

REFERENCES

Brewer, J. K. & Sindelar, P. T. (1987). Adequate sample size: A priori and post hoc considerations. *Journal of Special Education*, *21*, 74 - 84.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Erlbaum.

Gerard, P. D., Smith, D. R. & Weerakkody, G. (1998). Limits of retrospective power analysis. *Journal of Wildlife Management*, *62*, 801 - 807.

Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Johnson, N. L., Kotz, S. & Balakrishnan, N. (1995). *Continuous univariate distributions, Volume 2* (2nd Ed.). New York: Wiley.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759.

Kromrey, J. D. & Hogarty, K. Y. (2000, April). *Retrospective statistical power analysis: An empirical investigation of point and interval estimation techniques*. Paper presented at the annual meeting of the American Educational Research Association: New Orleans.

Kromrey, J. D. & Hogarty, K. Y. (in press). Retrospective power analysis: An empirical comparison of point and interval estimates. *American Statistical Association: Proceedings of the Section on Government Statistics and Section on Social Statistics.*

Ottenbacher, K. J. & Maas, F. (1998). How to detect effects: Statistical power and evidence-based practice in occupational therapy research. *American Journal of Occupational Therapy*, *53*, 181 - 188.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical recipes in FORTRAN: The art of scientific computing* (2nd Ed.). New York: Cambridge.

Steiger, J. H. & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation

of statistical models. In L. L. Harlow, S. A. Muliak & J. H. Steiger (Eds.), *What if there were no significance tests?* Mahweh, NJ: Erlbaum.

Taylor, D. J. & Muller, K. E. (1996). Bias in linear model power and sample size cacluation due to estimating noncentrality. *Communications in Statistics: Theory and Methods*, *25*, 1595-1610.

Zumbo, B. D. & Hubley, A. M. (1998). A note on misconceptions concerning prospective and retrospective power. *The Statistician*, *47*, 385 - 388.

CONTACT INFORMATION

The authors can be contacted at the University of South Florida, Department of Educational Measurement and Research, 4202 E. Fowler Avenue, EDU 162, Tampa, FL 33620. They may be contacted by telephone at (813) 974-3220 or contacted by e-mail at khogarty@luna.cas.usf.edu or kromrey@typhoon.coedu.usf.edu.