The Sample and Population Variation Methods

More ways to select sample size

This slide show is a free open source document.
See the last slide for copyright information.

These methods are non-standard

- The sample variation method is another way to select sample size. It’s not based on power.
- The population variation method is a way to select sample size on the basis of power, with the same flavour as the sample variation method.
- The population variation method is standard only in Psychology.
- *Statistical power analysis for the behavioral sciences* by Jacob Cohen.

Sample Size by Power

- Statisticians usually recommend that sample size be based on a power analysis.
- Statistical power is the probability of rejecting the null hypothesis when the null hypothesis is false.
- Power depends on sample size and how wrong H_0 is (effect size).

How to select sample size by power

- Pick an effect size you’d like to be able to detect. It should be just over the boundary of interesting and meaningful.
- Pick a desired power – a probability with which you’d like to be able to detect the effect by rejecting the null hypothesis.
- Start with a fairly small n and calculate the power. Increase the sample size until the desired power is reached.

The problem

- For regression and analysis of variance, effects must be expressed in units of σ.
- This can be tough in practice.
- Scientists who understand the subject matter and want to select a sample size mostly do not think in terms of Greek letters.
- Statisticians can think in terms of Greek letters, but they have no idea what effect size is important.
- Regression on observational data presents additional problems.
- The sample variation method is quick and easy.

F test is based upon

$$a = \frac{R_F^2 - R_R^2}{1 - R_R^2}$$

Increase in explained variation expressed as a fraction
of the variation that the reduced model does *not* explain.

$$F = \left(\frac{n-p}{s}\right)\left(\frac{a}{1-a}\right)$$

- For any given sample size, the bigger *a* is, the bigger *F* becomes.
- For any a ≠0, *F* increases as a function of *n.*
- So you can get a large *F* from strong results and a small sample, or from weak results and a large sample.

$$F = \left(\frac{n-p}{s}\right)\left(\frac{a}{1-a}\right)$$

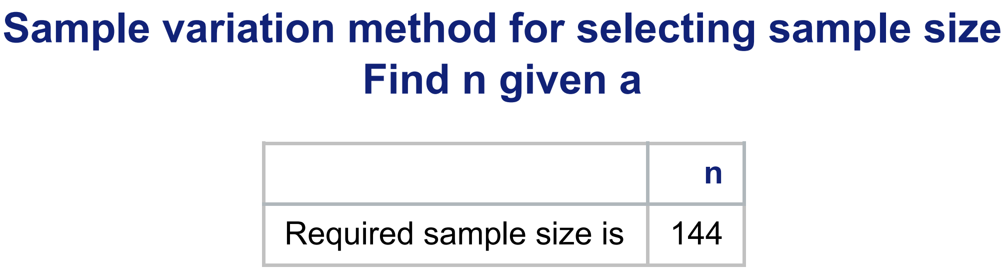
The sample variation method is to choose a value of *a* that is just large enough to be interesting, and increase n, calculating *F* and its p-value each time until p < 0.05; then stop. The final value of n is the smallest sample size for which an effect explaining that much of the remaining variation will be significant. With that sample size, the effect will be significant if and only if it explains *a* or more of the remaining variation.

That's all there is to it. You tell me a proportion of remaining variation that you want to be statistically significant, and I'll tell you a sample size.

Example

Suppose we are planning a 2x3x4 analysis of covariance, with two covariates, and factors named A, B and C. We are setting it up as a regression model, with one dummy variable for A, 2 dummy variables for B, and 3 for C. Interactions are represented by product terms, and there are 2 products for the AxB interaction, 3 for AxC, 6 for BxC, and 1*2*3 = 6 for AxBxC. The regression coefficients for these plus two for the covariates and one for the intercept give us p = 26. The null hypothesis is that of no BxC interaction, so r = 6. The "other effects in the model" for which we are "controlling" are represented by 2 covariates and 17 dummy variables and products of dummy variables.

```
proc iml;
     title2 ’Find n given a’;
     alpha = 0.05;  /* Significance level.                      */
     s = 6;         /* Numerator df = # Expl vars tested.       */
     p = 26;        /* There are p beta parameters.             */
     a = .10  ;     /* Proportion of remaining variation after */
                   /* controlling for all other variables.    */
   /* Initializing ... */  pval = 1; n = p;
   do until (pval <= alpha);
      n = n+1 ;
      F = (n-p)/s * a/(1-a);
      df2 = n-p;
      pval = 1-probf(F,s,df2);
   end;
   print "Required sample size is" n;
```



Sample variation method for selecting sample size
Find n given a

	n
Required sample size is	144

Sometimes it’s helpful to know what proportion of remaining variation you need for significance with a given sample size.

Potato Example

In the potato data, there are 3 potatoes per treatment combination in a Temperature (2 levels) by Bacteria type (3 levels) by Oxygen level (3 levels) design. What proportion of remaining variation is required for the main effect of bacteria type to be significant?

- Effect coding: How many dummy variables?
- Main effects: Temp = 1, Bact=2, Ox=2 5
- 2-factor: TxB = 1*2, T*O = 1*2, B*O = 2*2 8
- 3-factor: TxBxO = 1*2*2 4
- Plus the intercept 1

```
proc iml;
   title2 'Find a given n';
   alpha = 0.05;  /* Significance level.                */
   s = 2;         /* Numerator df = # Expl vars tested. */
   p = 18;        /* There are p beta parameters.       */
   n = 54  ;      /* Sample size                        */

   /* Initializing ... */  a = 0; df2 = n-p;
   do until (pval <= alpha);
      a = a + .001 ;
      F = (n-p)/s * a/(1-a);
      pval = 1-probf(F,s,df2);
     end;
     print "Required proportion of remaining variation is" a;
```

```
Sample variation method for selecting sample size
                 Find a given n

                                                      a
Required proportion of remaining variation is     0.154
```

The Population Variation Method

- This is a method of sample size selection for multiple regression due to Cohen (1988).
- It combines elements of classical power analysis and the sample variation method.
- Cohen does not call it the “Population Variation Method;” he calls it “Statistical Power Analysis.”
- For most research psychologists, statistical power analysis *is* the population variation method, period.

The idea

- Looking closely at the formula for the non-centrality parameter of the non-central F distribution, Cohen decides that it is based on something interprets as a population version of the quantity we are denoting by *a*.

- One thinks of it as the proportion of remaining variation that is explained by the effect in question *in the population*. He calls it ``effect size."

Cohen’s Population Variation Method

$$\begin{aligned}
\phi &= \frac{(\mathbf{C}\boldsymbol{\beta}-\mathbf{h})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\boldsymbol{\beta}-\mathbf{h})}{\sigma^2} \\
F^* &= \frac{(\mathbf{C}\widehat{\boldsymbol{\beta}}-\mathbf{h})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}}-\mathbf{h})}{r\,MSE} \\
&= \left(\frac{n-p}{r}\right)\left(\frac{a}{1-a}\right)
\end{aligned}$$

$$\phi/n = \frac{(\mathbf{C}\boldsymbol{\beta}-\mathbf{h})'(\mathbf{C}(\mathbf{X}'\mathbf{X}/n)^{-1}\mathbf{C}')^{-1}(\mathbf{C}\boldsymbol{\beta}-\mathbf{h})}{\sigma^2}$$

Let $n \rightarrow \infty$ and call the result “population effect size.” Write it as $\frac{a}{1-a}$. Call a the proportion of remaining variation in the *population*.

Population Variation Method

- It’s a way to choose an effect size without having to guess true beta (mu) or sigma values
- To get a non-centrality parameter for power analysis, Cohen multiplies by n-p instead of n.
- That’s because he thinks of phi as r times a population F statistic.

```
poprsq <- function(r,q,a,wantpow=0.80,alpha=0.05)
# Cohen's Popularion R-squared Method
#   r   Number of IVs in full model
#   q   Numerator df = number of linear constraints being tested
#   a   Population proportion of remaining variation explained.
#       This is Cohen's "effect size."
#   wantpow     Desired power (default = 0.80)
#   alpha       Significance level (default = 0.05)
    {
    pow <- 0 ; nn <- r+1 ; oneminus <- 1 - alpha
    while(pow < wantpow)
        {
        nn <- nn+1
        phi <- (nn-r) * a/(1-a)
        ddf <- nn-r
        pow <- 1 - pf(qf(oneminus,q,ddf),q,ddf,phi)
        }#End while
     poprsq <- nn
     poprsq
     } # End of function poprsq
```

```
> samprsq1(r=26,q=6,a=0.10)
[1] 144
> poprsq(r=26,q=6,a=0.10)
[1] 155
>
> samprsq1(r=26,q=6,a=0.05)
[1] 270
> poprsq(r=26,q=6,a=0.05)
[1] 292
```

Copyright Information