The Stata Journal (2003)
3, Number 1, pp. 47–56

Sample size calculations for main effects and interactions in case–control studies using Stata's nchi2 and npnchi2 functions

Catherine L. Saunders, D. Timothy Bishop, and Jennifer H. Barrett
Genetic Epidemiology Division, Cancer Research UK Clinical Centre, Leeds, UK

Abstract. The non-central χ^2 distribution can be used to calculate power for tests detecting departure from a null hypothesis. Required sample size can also be calculated because it is proportional to the non-centrality parameter for the distribution. We demonstrate how these calculations can be carried out in Stata using the example of calculating power and sample size for case–control studies of gene–gene and gene–environment interactions. Do-files are available for these calculations.

Keywords: st0032, gene–environment interaction, gene–gene interaction, power, sample size, study design, non-central χ^2

1 Introduction

There is increasing interest in investigating gene–environment and gene–gene interactions in the study of complex diseases (Brennan 2002). In a typical population-based case–control study of sufficient size to study main effects, there is often low power to detect interactions. Alternative study designs have therefore been considered to address this issue. A program that can quickly compare several designs, with the flexibility of being able to carry out these calculations in Stata, is therefore a useful tool for researchers. Here, we describe the application of a method based on the asymptotic distribution of the likelihood-ratio statistic to examine power for association studies.

Under certain assumptions, the distribution of the likelihood-ratio statistic is approximately a central χ^2 distribution under the null hypothesis and a non-central χ^2 distribution under the alternative hypothesis (Wilks 1938). An approximation to the non-centrality parameter can be calculated as the likelihood-ratio statistic from the analysis of an exemplary dataset (Self et al. 1992). By an exemplary dataset, we mean one in which the proportions of cases and controls in the different exposure categories take their expected values under the alternative hypothesis. Required sample size is inversely proportional to the non-centrality parameter. This method is detailed in Brown et al. (1999) and illustrations of its use in case–control studies are given by Longmate (2001). Because Stata has functions that estimate central and non-central χ^2 distributions, the implementation of these calculations in Stata is straightforward.

The steps required to calculate sample size for a gene–environment interaction (departure from multiplicative joint effects of two binary (present/absent) risk factors) are detailed below. The method can be generalized for any alternative hypothesis as long as an exemplary dataset can be defined. Similarly, although an unmatched study is shown here, the method will work for matched analyses using `clogit` (with frequency weights) rather than `blogit` as the method of analysis, as long as the expected proportions of all possible matched case–control pairs under the alternative hypothesis can still be specified.

2 Implementation

Step 1

Calculate the expected distribution of risk factors among cases and controls under the alternative hypothesis. In the applications that we consider here, this requires knowledge of the following parameters: the population frequencies of the two risk factors and their main and interaction effects, the association of the two factors in the population and disease frequency.

Step 2

Create a large exemplary dataset in Stata. Using a large sample size initially decreases the variation in the non-centrality parameter that is due to the asymptotic approximation. The following variables are needed in the dataset: `g` (0/1 for absence/presence of susceptibility genotype), `e` (0/1 for absence/presence of environmental risk factor), `gei` (0/1, which takes the value 1 only in the presence of both risk factors), `aff` (number of cases with particular exposure combination), and `tot` (total number of people in study with given exposure combinations). When considering two binary risk factors, there are four possible exposure combinations and hence four observations in the dataset.

The exemplary dataset used in this example is listed here. Risk factor frequencies were calculated under the following assumptions: disease is rare (0.1%), the relative risks for main and interaction effects equal two, and the risk factors are independent in the source populations with susceptibility genotype frequency of 10% and exposure frequency of 20%. The variable `aff` is the number of cases with each genotype/exposure combination, and the variable `tot` is the number of cases plus the number of controls with each combination. In this dataset, there are equal numbers (10^8) of cases and controls. Because integers with more than 7 digits of accuracy are being considered in such a large exemplary dataset, the data storage type is set to double, although in practice this only has a very small effect on the power and sample-size calculations.

```
. version 7.0
. use implementation, clear
. set type double
. list

     +-------------------------------------+
     | g   e   gei        aff          tot |
     |-------------------------------------|
  1. | 0   0     0   51428572    123449163 |
  2. | 0   1     0   25714286     43706564 |
  3. | 1   0     0   11428571     19425140 |
  4. | 1   1     1   11428571     13419133 |
     +-------------------------------------+
```

Step 3

Calculate the likelihood-ratio test statistic for the interaction term. This is the noncentrality parameter of the likelihood-ratio test statistic under the alternative hypothesis

```
. blogit aff tot g e gei, or
Logit estimates                                   Number of obs   =  200000000
                                                  LR chi2(3)      =  1.277e+07
                                                  Prob > chi2     =     0.0000
Log likelihood = -1.322e+08                       Pseudo R2       =     0.0461

------------------------------------------------------------------------------
    _outcome | Odds Ratio   Std. Err.      z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           g |    2.00143   .0009924  1399.29   0.000     1.999486    2.003376
           e |   2.001431   .0007155  1940.94   0.000     2.000029    2.002833
         gei |   2.007185   .0019359   722.39   0.000     2.003394    2.010983
------------------------------------------------------------------------------

. lrtest, saving(0)
. blogit aff tot g e, or
Logit estimates                                   Number of obs   =  200000000
                                                  LR chi2(2)      =  1.222e+07
                                                  Prob > chi2     =     0.0000
Log likelihood = -1.325e+08                       Pseudo R2       =     0.0441

------------------------------------------------------------------------------
    _outcome | Odds Ratio   Std. Err.      z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           g |   2.452137   .0010217  2152.69   0.000     2.450135     2.45414
           e |   2.221957   .0007319  2423.86   0.000     2.220523    2.223392
------------------------------------------------------------------------------

. lrtest
Blogit:  likelihood-ratio test                        chi2(1)     =  552415.16
                                                      Prob > chi2 =     0.0000

. return list
scalars:
                  r(p) =  0
               r(chi2) =  552415.1566385925
                 r(df) =  1
```

Step 4

To calculate study power (%) for a given sample size (500 cases and 500 controls) and significance level (0.05), use Stata's nchi2 function.

```
. display (1-nchi2(r(df), r(chi2)*500/10^8, invchi2(r(df), (1-0.05))))*100
38.29922
```

The nchi2 function gives the cumulative distribution function of a non-central chi-squared statistic. Here, the number of degrees of freedom is r(df) and the non-centrality parameter is r(chi2), scaled by the factor $500/10^8$ to account for the smaller sample size. The final expression invchi2(r(df), (1-0.05)) is the appropriate percentile of the central chi-squared distribution to achieve a 5% significance level (i.e., 3.84 in this case).

Step 5

Alternatively, to calculate the required number of cases (assumed equal to the number of controls) for a given study power (80%), use Stata's npnchi2 function.

```
. display round(10^8*(npnchi2(r(df), invchi2(r(df), (1-0.05)),
> (1-(80 /100)))/ r(chi2)), 1)
1421
```

npnchi2 provides the non-centrality parameter such that the probability of the non-central chi-squared statistic being less than the appropriate percentile (3.84 in this case) is 0.2 (i.e., 1 − power). Stata's round function is used to give a whole number for the required sample size. In this example, df = 1, since the models only differ by one parameter (the interaction parameter gei).

3 Applications

Programs that apply this method to carry out power and sample-size calculations for gene–environment interactions and gene–gene interactions and sample input files for each program are available from *http://cruk.leeds.ac.uk/katie*. These programs report power, required sample size, and the interaction odds ratio that would be obtained from the analysis of exemplary datasets. Population-based case–control studies typically have low power to detect interactions; thus, many different designs have been proposed to potentially improve power. Matching strategies including flexible matching (Sturmer and Brenner 2002) and counter-matching (Andrieu et al. 2001), plus some extensions, are considered using gei_matching. By using the siblings of cases as cases and/or controls in studies of gene–environment or gene–gene interactions, there is the potential to improve power when risk factors are rare (Andrieu and Goldstein 2000; Siegmund and Langholz 2001; Witte et al. 1999). This is because risk factor frequencies are expected to be higher, and therefore more informative, among the relatives of cases. Recent research, however, has found that a matched sibling design was in-efficient in testing for gene–gene interactions (Gauderman 2002), and calculations are provided for a range of sampling schemes involving the siblings of cases or population based subjects for gene–gene (`ggipower`) and gene–environment (`geipower`) interactions. Case–parent designs for interactions (Weinberg and Umbach 2000) are also considered in `tdt_ggipower` and `tdt_geipower`. Non-family designs have also been proposed as strategies to improve power; `second_primary` calculates power and required sample size for designs that sample people who have had more than one primary cancer (Begg and Berwick 1997), and `co_power` considers the case-only design (Piegorsch et al. 1994).

Because so many different designs have been proposed, using these programs gives a simple way of making a large number of comparisons for different parameter ranges and designs.

3.1 Syntax

```
. ggipower using input_file
. geipower using input_file
. tdt_ggipower using input_file
. tdt_geipower using input_file
. co_power using input_file
. second_primary using input_file
```

Risk factor frequencies, disease frequency, and the magnitudes of main and interaction effects affect the power of all designs to detect interactions, whereas other parameters are more design-specific. For example, the power of some of the matching designs depend on the specificity and sensitivity of surrogates for the risk factors, and the power of family-based designs depend on the association of genetic and environmental risk factors within families. Specific details of variables that are needed for calculations for each design are given in the help files and in the sample files that can be downloaded with the programs. The format of the required input files are described in the help file for each program; details for `tdt_ggipower` are given below.

4 Example

The power of different designs to detect interactions depends on the risk factor frequencies among cases and controls in the exemplary dataset. Although the power and sample-size calculations themselves are simple to carry out in Stata, these programs also carry out the calculations that are required to produce the exemplary datasets for the different designs. An example of the use of these programs is given for power and sample-size calculations for a case–parent design for gene–gene interactions. The required input dataset for `tdt_ggipower` contains values for the risk factor frequencies and effects, and the required significance level, sample size and power; details are given in Table 1. Other designs may require values for different parameters to be specified.

Table 1

Variable Name	**Details**
`pg1`	The population frequency of susceptibility genotype1
`pg2`	The population frequency of susceptibility genotype2
`inh1`	The mode of inheritance for susceptibility genotype1
`inh2`	The mode of inheritance for susceptibility genotype2
`rrg1`	The relative risk of disease in people exposed to susceptibility genotype1 but not to susceptibility genotype2, compared with those people exposed to neither factor
`rrg2`	The relative risk of disease in people exposed to susceptibility genotype2 but not to susceptibility genotype1, compared with those people exposed to neither factor
`rrint`	The interaction relative risk (such that the relative risk of disease in people exposed to both risk factors compared with no risk factors is `rrg1` × `rrg2` × `rrint`)
`pd`	The population disease frequency
`ssize`	The sample size for which power calculations are required
`power`	The power for which sample-size calculations are required
`alpha_1`	The required significance level for the interaction test

Because a person inherits one copy of every gene from their mother and one from their father, for each genotype then, a case or control can have 0, 1, or 2 copies of a ‘disease’ allele. The “mode of inheritance” variables in table 1 determine whether a person needs to have only one (dominant inheritance) or two (recessive inheritance) copies of the susceptibility allele for the genotype to be high-risk. This program also assumes that genotypes 1 and 2 are independent. This means that whether susceptibility genotype 1 is present or absent in an individual from the source population does not depend on the presence or absence of susceptibility genotype 2. In genetic terminology, this means that the two genes must be “unlinked”.

Typical output from using the program is given below.

(*Continued on next page*)

```
. tdt_ggipower using tdtggi_parameters
parameter file:              tdtggi_parameters
observations:                1
-------------------------------------------------------------------------------
PARAMETERS
-------------------------------------------------------------------------------
Power to detect interaction (%), interaction odds ratios and required sample
sizes have been calculated from an exemplary dataset for the following
population risk factor frequencies and effects:
susceptibility genotype 1 frequency (pg1):                           .1
susceptibility genotype 2 frequency (pg2):                           .2
genetic risk factor 1 main effect (rrg1):                            1
genetic risk factor 2 main effect (rrg2):                            1
interaction relative risk (rrint):                                   3
disease prevalence (pd):                                             .0001
-------------------------------------------------------------------------------
POWER
-------------------------------------------------------------------------------
Power to detect an interaction for a sample size of 500 cases
with a two-sided significance level (alpha=.05)
for a case-parent design with 500 cases
power (%) both genetic risk factors dominant inheritance:            89.69
power (%) both genetic risk factors recessive inheritance:           93.81
power (%) g1 dominant, g2 recessive:                                 91.9
-------------------------------------------------------------------------------
INTERACTION ODDS RATIOS
-------------------------------------------------------------------------------
The interaction odds ratio calculated from the exemplary dataset:
dominant inheritance:                            3
recessive inheritance:                           3
g1 dominant, g2 recessive inheritance:           3
-------------------------------------------------------------------------------
REQUIRED SAMPLE SIZE
-------------------------------------------------------------------------------
for a power of 80% and a two-sided significance level (alpha=.05) the
required number of cases (with two parental controls per case)
dominant inheritance                             378
recessive inheritance                            320
g1 dominant, g2 recessive inheritance:           348
-------------------------------------------------------------------------------
NOTES
-------------------------------------------------------------------------------
The power, required sample sizes, and calculated interaction odds ratio for the
case-parent design have been saved into file  tdtggi_parameters. Type d
for details of the output variables
```

New variables are created and results are written into the original dataset. If more than one set of parameters are considered in the input, for example, to look at required sample size over a range of risk factor frequencies, then results are saved in the original dataset, rather than being output to the results window.

It can be seen from this output that for a case–parent design, if both genotypes under study had a dominant mode of inheritance, then 378 case–parent trios would be needed. Required sample sizes under different modes of inheritance can be compared. The interaction odds ratio section of the output shows that there is no bias in the estimate of the interaction relative risk (shown in the parameter section of the output). This section is more important for designs such as the unmatched family designs considered in `ggipower` and `geipower` or the case-only design, where the estimated interaction odds ratio also depends on the level of population association of the two risk factors and may therefore lead to an increased false positive rate for this design (Albert et al. 2001). Power for a given sample size is also reported. The power, sample size, and estimated interaction odds ratios output from each program or design are in similar formats.

5 Discussion

The simplicity of power and sample size calculations carried out using Stata's `nchi2` and `npnchi2` functions are extremely useful, because they allow the power of many different designs over ranges of population parameters to be easily considered. Risk factor frequencies and the magnitudes of the main and interaction effects all have effects on the efficiencies and relative efficiencies of these designs, so it is also helpful to be able to easily compare power and sample size over different ranges. These methods are applied here to studies of interactions, which is an area in which debate about the most efficient design is very relevant. However, they are generally applicable to calculations for any likelihood-ratio test for an alternative hypothesis for which an exemplary dataset (expected risk factor frequencies) can be defined. The methods are thus simplest to apply to binary or categorical risk factors. To consider continuous risk factors the exemplary dataset would require the distribution of the risk factor to be defined and a random variable to be generated, and so in this situation, a simulation approach may be more appropriate. In order to check the accuracy of the large sample approximation method for power calculations, simulations were carried out using Stata, and the results of the two methods are compared. The same risk factor frequencies among cases and controls calculated for the exemplary dataset were used in the simulations. A wide range of parameters for each of the designs was considered. In all situations considered, the two methods reported similar power indicating the reliability of this large sample approximation. These methods and programs present a way in which the effect of a large number of case–control study designs and parameters on efficiency can be compared and provide a useful tool at the planning stage of any study.

6 Acknowledgment

This work was funded by Cancer Research UK.

7 References

Albert, P. S., D. Ratnasinghe, J. Tangrea, and S. Wacholder. 2001. Limitations of the case-only design for identifying gene–environment interactions. *American Journal of Epidemiology* 154: 687–693.

Andrieu, N. and A. M. Goldstein. 2000. A case-combined design using both population based- and related-controls: a potential alternative for increasing power in gene–environment interaction detection. *Genetic Epidemiology* 19: 235–236.

Andrieu, N., A. M. Goldstein, D. C. Thomas, and B. Langholz. 2001. Counter-matching in studies of gene–environment interaction: efficiency and feasibility. *American Journal of Epidemiology* 153: 265–274.

Begg, C. B. and M. Berwick. 1997. A note on the estimation of relative risks of rare genetic susceptibility markers. *Cancer Epidemiology Biomarkers Prevention* 6: 99–103.

Brennan, P. 2002. Gene-environment interaction and aetiology of cancer: what does it mean and how can we measure it? *Carcinogenesis* 23: 381–387.

Brown, B. W., J. Lovato, and K. Russell. 1999. Asymptotic power calculations: description, examples, computer code. *Statistics in Medicine* 18: 3137–3151.

Gauderman, W. J. 2002. Sample size requirements for association studies of gene–gene interaction. *American Journal of Epidemiology* 155: 478–484.

Longmate, J. A. 2001. Complexity and power in case–control association studies. *American Journal of Human Genetics* 68: 1229–1237.

Piegorsch, W. W., C. R. Weinberg, and J. A. Taylor. 1994. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case–control studies. *Statistics in Medicine* 13: 153–162.

Self, S. G., R. H. Mauritsen, and J. Ohara. 1992. Power calculations for likelihood ratio tests in generalized linear models. *Biometrics* 48: 31–39.

Siegmund, K. D. and B. Langholz. 2001. Stratified case sampling and the use of family controls. *Genetic Epidemiology* 20: 316–327.

Sturmer, T. and H. Brenner. 2002. Flexible matching strategies to increase power and efficiency to detect and estimate gene–environment interactions in case–control studies. *American Journal of Epidemiology* 155: 593–602.

Weinberg, C. R. and D. M. Umbach. 2000. Choosing a retrospective design to assess joint genetic and environmental contributions to risk. *American Journal of Epidemiology* 152: 197–203.

Wilks, S. S. 1938. The large sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics* 9: 60–62.

Witte, J. S., W. J. Gauderman, and D. C. Thomas. 1999. Asymptotic bias and efficiency in case–control studies of candidate genes and gene–environment interactions: basic family designs. *American Journal of Epidemiology* 149: 693–705.

About the Authors

Jenny Barrett is a genetic epidemiologist and statistician in the Cancer Research UK Genetic Epidemiology Division at the University of Leeds, UK.

Tim Bishop is Professor of Genetic Epidemiology at the University of Leeds and has a research interest into the contribution of genes to the incidence of disease in the general population.

Katie Saunders is a student in the same department and is researching case–control study designs for gene–gene and gene–environment interactions.