

power mcc — Power analysis for matched case–control studies

Description	Quick start	Menu	Syntax
Options	Remarks and examples	Stored results	Methods and formulas
References	Also see		

Description

`power mcc` computes sample size, power, or effect size (the minimum detectable odds ratio) for a test of association between a risk factor and a disease in 1: M matched case–control studies.

Quick start

Number of cases for a test with exposure probability for controls of 0.2 and odds ratio for exposure of 1.4 from a 1:1 matched design using default power of 0.8 and significance level $\alpha = 0.05$

```
power mcc .2, oratio(1.4)
```

As above, but for a 1:2 matched design and compute the ratio of the number of cases for a 1:2 matched design relative to a 1:1 matched design

```
power mcc .2, oratio(1.4) m(2) compare
```

Number of cases when correlation of exposure between matched pairs is 0.3

```
power mcc .2, oratio(1.4) corr(.3)
```

Power for 500 cases and a 1:1 matched design

```
power mcc .2, oratio(1.4) n(500)
```

Plot of power against the number of cases for 450, 475, 500, 525, and 550 cases

```
power mcc .2, oratio(1.4) n(450(25)550) graph
```

Minimum detectable odds ratio with 80% power and 500 cases using a 1:1 matched design

```
power mcc .2, power(.8) n(500)
```

As above, but for an upper one-sided test

```
power mcc .2, power(.8) n(500) onesided direction(upper)
```

Same as above

```
power mcc .2, power(.8) n(500) onesided
```

Menu

Statistics > Power and sample size

Syntax

Compute sample size

```
power mcc  $p_0$ , oratio(numlist) [power(numlist) options]
```

Compute power

```
power mcc  $p_0$ , oratio(numlist) n(numlist) [options]
```

Compute target odds ratio

```
power mcc  $p_0$ , power(numlist) n(numlist) [options]
```

where p_0 is the probability of exposure among control patients. p_0 must satisfy the condition $0 < p_0 < 1$ and may be specified either as one number or as a list of values in parentheses; see [U] **11.1.8 numlist**.

<i>options</i>	Description
Main	
* <u>alpha</u> (<i>numlist</i>)	significance level; default is <code>alpha(0.05)</code>
* <u>power</u> (<i>numlist</i>)	power; default is <code>power(0.8)</code>
* <u>beta</u> (<i>numlist</i>)	probability of type II error; default is <code>beta(0.2)</code>
* <u>n</u> (<i>numlist</i>)	sample size; required to compute power or effect size
<u>nfractional</u>	allow fractional sample size
* <u>oratio</u> (<i>numlist</i>)	odds ratio of exposure in cases relative to controls; required to compute power or sample size
* <u>m</u> (<i>numlist</i>)	number of matched controls per case; default is <code>m(1)</code>
<u>compare</u>	ratio of the required number of cases for a 1: <i>M</i> design relative to a paired 1:1 design
* <u>corr</u> (<i>numlist</i>)	correlation of exposure between cases and controls; default is <code>corr(0)</code>
<u>direction</u> (<u>upper</u> <u>lower</u>)	direction of the effect for effect-size determination; default is <code>direction(upper)</code> , which means that the postulated value of the parameter is larger than the hypothesized value
<u>onesided</u>	one-sided test; default is two sided
<u>parallel</u>	treat number lists in starred options or in command arguments as parallel when multiple values per option or argument are specified (do not enumerate all possible combinations of values)
Table	
<u>[no]</u> <u>table</u> [(<i>tablespec</i>)]	suppress table or display results as a table; see [PSS] power, table
<u>saving</u> (<i>filename</i> [, replace])	save the table data to <i>filename</i> ; use <code>replace</code> to overwrite existing <i>filename</i>
Graph	
<u>graph</u> [(<i>graphopts</i>)]	graph results; see [PSS] power, graph
Iteration	
<u>init</u> (#)	initial value for the sample size or the effect size
<u>iterate</u> (#)	maximum number of iterations; default is <code>iterate(500)</code>
<u>tolerance</u> (#)	parameter tolerance; default is <code>tolerance(1e-12)</code>
<u>ftolerance</u> (#)	function tolerance; default is <code>ftolerance(1e-12)</code>
<u>[no]</u> <u>log</u>	suppress or display iteration log
<u>[no]</u> <u>dots</u>	suppress or display iterations as dots
<u>notitle</u>	suppress the title

*Specifying a list of values in at least two starred options, or at least two command arguments, or at least one starred option and one argument results in computations for all possible combinations of the values; see [U] [11.1.8 numlist](#). Also see the `parallel` option.

`notitle` does not appear in the dialog box.

where *tablespec* is

```
column [ :label ] [ column [ :label ] [ ... ] ] [ , tableopts ]
```

column is one of the columns defined [below](#), and *label* is a column label (may contain quotes and compound quotes).

<i>column</i>	Description	Symbol
<code>alpha</code>	significance level	α
<code>power</code>	power	$1 - \beta$
<code>beta</code>	type II error probability	β
<code>N</code>	number of cases	N
<code>delta</code>	effect size	δ
<code>M</code>	number of matched controls	M
<code>F_M</code>	ratio of the number of cases with M controls relative to one control	F_M
<code>p0</code>	probability of exposure among controls	p_0
<code>p1</code>	probability of exposure among cases	p_1
<code>oratio</code>	odds ratio	θ
<code>corr</code>	correlation of exposure between cases and controls	ρ
<code>target</code>	target parameter; synonym for <code>oratio</code>	
<code>_all</code>	display all supported columns	

Column `beta` is shown in the default table in place of column `power` if option `beta()` is specified.

Column `F_M` is shown in the default table only if option `compare` is specified.

Column `p1` is not shown in the default table.

Options

Main

`alpha()`, `power()`, `beta()`, `n()`, `nfractional`; see [\[PSS\] power](#). The sample size in `n()` is the number of matched case–control sets or, equivalently, the number of cases. The `nfractional` option is allowed only for sample-size determination.

`oratio(numlist)` specifies the odds ratio of exposure in cases relative to controls. This option is required for `power` or sample-size determination and may not be specified for effect-size determination.

`m(numlist)` specifies the number of matched controls per case. Only positive integers are allowed. The default is `m(1)`, which implies a paired design.

`compare` specifies that the ratio, F_M , of the required number of cases for a 1: M design relative to a paired 1:1 design be computed. `compare` can be specified only when computing sample size and when a value of 2 or greater is specified in option `m()`.

`corr(numlist)` specifies the correlation coefficient for exposure ρ between matched cases and controls. `corr()` must contain numbers between -1 and 1 . The default is `corr(0)`, meaning no correlation between matched cases and controls. This assumption may not be realistic in practice; see [example 3](#) for discussion.

`direction()`, `onesided`, `parallel`; see [\[PSS\] power](#).

Table

`table`, `table()`, `notable`; see [\[PSS\] power, table](#).

`saving()`; see [\[PSS\] power](#).

Graph

`graph`, `graph()`; see [PSS] [power](#), [graph](#). Also see the *column* table for a list of symbols used by the graphs.

Iteration

`init(#)` specifies the initial value for the estimated sample size or effect size when an iterative search is required. When computing the sample size for the two-sided test, the closed-form sample-size computation for the one-sided test is used. The initial estimate for computing the minimum detectable odds ratio is obtained from a bisection search.

`iterate()`, `tolerance()`, `ftolerance()`, `log`, `nolog`, `dots`, `nodots`; see [PSS] [power](#).

The following option is available with `power mcc` but is not shown in the dialog box:

`notitle`; see [PSS] [power](#).

Remarks and examples

[stata.com](http://www.stata.com)

Remarks are presented under the following headings:

Introduction

Using power mcc

Computing sample size

Computing power

Computing target odds ratio

Testing hypotheses in matched case–control studies

This entry describes the `power mcc` command and the methodology for power and sample size analysis for $1:M$ matched case–control studies. See [PSS] [intro](#) for a general introduction to power and sample-size analysis and [PSS] [power](#) for a general introduction to the `power` command using hypothesis tests.

Introduction

Matched case–control studies investigate the relationship between disease and exposure, controlling for the effect of confounding variables. Cases are observations that have the outcome of interest; controls are observations that do not. Cases are matched to the controls on the basis of similar values of the variable or variables thought to confound the relationship between exposure and disease.

Matched case–control studies are used to investigate a variety of outcomes. A pediatrician might be interested in the relationship between low birthweight (case) and mother’s smoking status during pregnancy (exposure), where case and control mothers are matched on the basis of age, race, alcohol consumption, and history of hypertension (confounding variables). An oncologist might want to know if women with endometrial cancer are more likely to have taken estrogen, where cases and controls are matched on age, marital status, and time living in the community. A psychologist might design a study to see if suicide is more prevalent among patients who used a particular antidepressant, where cases and controls are matched on age, sex, race, severity of depression symptoms, and history of head injury.

This entry describes power and sample-size analysis for inference about correlated binary outcomes. In a $1:M$ matched case–control study, we first randomly select cases from a population of cases and observe their exposure status. The population is then stratified by the confounding (matching) variables, and for each selected case, M matched controls are randomly selected from the same

stratum as the selected case. This creates a series of 2×2 contingency tables within the stratum defined by the matching covariates. Each contingency table summarizes the probability of observing each exposure–outcome combination within a given stratum, and the probabilities are assumed to be equal across strata (or, equivalently, odds ratios are assumed to be equal across strata) so that for any study we need to analyze only one table.

Case	Control		Total
	Exposed	Unexposed	
Exposed	p_{11}	p_{10}	p_1
Unexposed	p_{01}	p_{00}	q_1
Total	p_0	q_0	1

The concordant probabilities lie on the diagonal; p_{11} is the probability that an exposed case subject is matched to an exposed control subject, and p_{00} is the probability that an unexposed case subject is matched to an unexposed control subject. The target parameter is the odds ratio θ of developing the disease in exposed and unexposed subjects who have equal values of matching variables. It can be calculated from the discordant probabilities as p_{10}/p_{01} .

The probability of exposure for controls, p_0 , is the probability that the sampled control subject is exposed and is simply the sum of p_{11} and p_{01} . The probability of exposure for cases, p_1 , is the probability that the sampled case subject is exposed and is simply the sum of p_{11} and p_{10} .

The two-sided hypothesis test for association between disease and exposure can be formally stated in terms of the odds ratio as $H_0: \theta = 1$ versus $H_a: \theta \neq 1$. We can equivalently state the hypothesis test in terms of marginal homogeneity: $H_0: p_0 = p_1$ versus $H_a: p_0 \neq p_1$.

In a matched case–control study, n cases are sampled and then matched to M controls. When $M = 1$, a 1:1 matched design or paired design, there are n matched pairs and $n \times 2$ total subjects. When $M > 1$, there are n matched sets and $n \times (M + 1)$ total subjects. Unlike a study without matching, a matched case–control study does not have $n \times 2$ independent observations [or $n \times (M + 1)$ for matched designs with multiples controls].

All calculations performed by `power mcc` treat n as the relevant sample size. Throughout the remainder of this entry, when we refer to the sample size, we mean n , the number of cases and thus the number of matched pairs (or sets).

Using `power mcc`

`power mcc` computes sample size, power, or effect size (the minimum detectable odds ratio) for 1: M matched case–control studies, in which one case is matched to M controls. All computations are performed for a two-sided hypothesis test where, by default, the significance level is set to 0.05. You may change the significance level by specifying the `alpha()` option. You can specify the `onesided` option to request a one-sided test.

To compute sample size, you must specify the probability of exposure for the control group p_0 ; the odds ratio for exposure θ in option `oratio()`; and, optionally, the power of the test in the `power()` option. The default power is set to 0.8. The sample-size estimate returned is the number of matched pairs or, if option `m()` was specified, the number of matched sets. This is equivalent to the number of cases. Hereafter, we simply refer to the number of cases.

To compute power, you must specify the sample size in option `n()`, the probability of exposure for the control group p_0 , and the odds ratio in option `oratio()`.

To compute the minimum detectable odds ratio, you must specify the sample size in option `n()`; the power in option `power()`; the probability of exposure for the control group p_0 ; and, optionally, the direction of the effect. The direction is upper by default, `direction(upper)`, which means that the probability of exposure among cases is assumed to be larger than the specified control-group value. You can change the direction to be lower, which means that the probability of exposure among cases is assumed to be lower than the specified control-group value, by specifying the `direction(lower)` option. `power mcc` defines the effect size as the target odds ratio.

By default, all computations assume a 1:1 or paired design, in which one case is matched to one control. You may specify the `m()` option to accommodate multiple matches per case.

The correlation between the matched case–control subjects is set to 0 by default but may be changed by specifying option `corr()`.

For sample-size determination, you can specify the `compare` option to compute the ratio of the required number of cases when using M matched controls rather than one.

Sample-size determination for the two-sided test and effect-size determination for $M > 1$ require iteration. The default initial sample-size value is set to the closed-form one-sided sample size. The initial value for the effect size is computed using a bisection algorithm. You can use the `init()` option to specify your own value. See [PSS] `power` for a description of other options that control the iteration process.

In the following sections, we describe the use of `power mcc` accompanied by examples for computing sample size, power, and the minimum detectable odds ratio.

Computing sample size

To compute sample size, you must specify the probability of exposure among control patients p_0 after the command name; the odds ratio θ in option `oratio()`; and, optionally, the power in option `power()`.

► Example 1: Sample size for a 1:1 matched case–control study

Consider a study comparing the odds of developing lung cancer among smokers compared with the odds among nonsmokers. Suppose that previous studies matching smokers and nonsmokers on the basis of age, gender, race, and alcohol consumption found the following proportions:

Lung Cancer (Case)	No Lung Cancer (Control)		Total
	Smoker	Nonsmoker	
Smoker	0.180	0.144	0.324
Nonsmoker	0.040	0.636	0.676
Total	0.220	0.780	1

If we wish to plan a new case–control study, we might assume that these proportions represent population probabilities and, therefore, let $p_0 = 0.22$. Under the assumption of no correlation of exposure in matched pairs, $\theta = (0.324 \times 0.78)/(0.22 \times 0.676) = 1.7$.

We would like to determine the number of case–control pairs that we will need to achieve 80% power to detect an odds ratio of 1.7 with a 5%-level two-sided test.

To compute the required sample size, we specify 0.22 as the probability of exposure for the control group after the command name and specify 1.7 as the odds ratio in option `oratio()`. We omit options `alpha(0.05)` and `power(0.8)` because the specified values are their defaults.

```

. power mcc .22, oratio(1.7)
Performing iteration ...
Estimated sample size for a matched case-control study
Asymptotic z test, 1:1 matched design
Ho: OR = 1 versus Ha: OR != 1
Study parameters:
      alpha =    0.0500
      power =    0.8000
      delta =    1.7000
      p0 =     0.2200
      oratio =    1.7000
      corr =    0.0000
      M =        1
Estimated sample size:
      N cases =     285

```

Our calculation indicates that we will need a sample of 285 cases to detect an odds ratio of 1.7 with 80% power using a 5%-level test.



► Example 2: Sample size for a 1: M matched case–control study

Multiple controls are often matched with one case to increase the efficiency of the study. Continuing with [example 1](#), we note that we have access to many more control participants than case participants.

We specify option `m(2)` to recalculate our sample size assuming that we will match two controls with each case (a 1:2 matched design). We also specify the `compare` option to calculate the ratio of the number of required cases relative to the 1:1 paired design.

```

. power mcc .22, oratio(1.7) m(2) compare
Performing iteration ...
Estimated sample size for a matched case-control study
Asymptotic z test, 1:2 matched design
Ho: OR = 1 versus Ha: OR != 1
Study parameters:
      alpha =    0.0500
      power =    0.8000
      delta =    1.7000
      p0 =     0.2200
      oratio =    1.7000
      corr =    0.0000
      M =        2
Estimated sample size:
      N cases =     210
      F_M =     0.7368

```

We obtain a new sample-size estimate of 210 cases, and the reduction in the number of cases relative to the paired design is approximately 26% ($1 - F_M = 0.2632$).



▷ Example 3: Sample size with correlated exposures

Matching based on confounders will typically lead to correlation of the exposure between cases and controls. For example, smoking status is known to be correlated with alcohol consumption. Matching on alcohol consumption might, therefore, result in the cases and controls being more similar with regard to smoking status. Ignoring this correlation will lead to underestimation of the required sample size or overestimation of power.

The correlation coefficient ρ for exposure between cases and controls can be computed from the probabilities in our contingency table; see (1) in *Methods and formulas*.

Returning to [example 1](#), we compute the correlation of exposure:

$$\rho = (0.180 \times 0.636 - 0.144 \times 0.040) / \sqrt{0.324 \times 0.676 \times 0.220 \times 0.780} = 0.56$$

We then specify its value in the `corr()` option.

```
. power mcc .22, oratio(1.7) corr(.56)
Performing iteration ...
Estimated sample size for a matched case-control study
Asymptotic z test, 1:1 matched design
Ho: OR = 1 versus Ha: OR != 1
Study parameters:
    alpha =    0.0500
    power =    0.8000
    delta =    1.7000
    p0 =      0.2200
    oratio =    1.7000
    corr =    0.5600
    M =        1
Estimated sample size:
    N cases =      703
```

With such a high level of correlation between exposure and the matching variables, our sample size increases dramatically to 703 cases.

This examples demonstrates the importance of taking into account the correlation of exposure between matched cases and controls. For this reason, [Dupont \(1988\)](#) recommends using the value of, say, 0.2 in computations instead of making an independence assumption, which is unlikely to hold in practice.

◀

▷ Example 4: Multiple values of study parameters

Continuing with [example 3](#), suppose that we believe that the previously gathered data may provide a good estimate of exposure probability for controls p_0 and the odds ratio θ , but that the individual cell proportions are not precise enough estimates of the population probabilities to produce a good estimate of ρ .

In this case, we may want to use a range of plausible values and consider how it affects the sample size for our study. We can specify a range of correlations between 0.4 and 0.6 with a step size of 0.05 using standard *numlist* (see [\[U\] 11.1.8 numlist](#)) notation in option `corr()`.

```
. power mcc .22, oratio(1.7) corr(.4(.05).6)
Performing iteration ...
Estimated sample size for a matched case-control study
Asymptotic z test, 1:1 matched design
Ho: OR = 1 versus Ha: OR != 1
```

alpha	power	N	delta	M	p0	oratio	corr
.05	.8	503	1.7	1	.22	1.7	.4
.05	.8	553	1.7	1	.22	1.7	.45
.05	.8	613	1.7	1	.22	1.7	.5
.05	.8	687	1.7	1	.22	1.7	.55
.05	.8	779	1.7	1	.22	1.7	.6

For a given power, the required sample size increases as the correlation ρ increases. In this example, the choice of the correlation has a large effect on the required sample size, which suggests that we should carefully consider the choice of matching variables.

For multiple values of parameters, the results are automatically displayed in a table, as we see above. For more examples of tables, see [PSS] [power, table](#). If you wish to produce a power plot, see [PSS] [power, graph](#).

◀

Computing power

To compute power, you must specify the number of cases in option `n()`, the exposure probability among controls p_0 following the command name, and the odds ratio in option `oratio()`.

▶ Example 5: Power for matched case–control studies

Returning to [example 1](#), we discover that we are able to recruit 300 cases for our study. To compute the corresponding power, we specify 300 as the sample size in option `n()`.

```
. power mcc .22, oratio(1.7) n(300)
Estimated power for a matched case-control study
Asymptotic z test, 1:1 matched design
Ho: OR = 1 versus Ha: OR != 1
Study parameters:
    alpha =    0.0500
    N cases =     300
    delta =    1.7000
    p0 =     0.2200
    oratio =    1.7000
    corr =    0.0000
    M =         1
Estimated power:
    power =    0.8204
```

As expected, with a larger sample size, this example has a higher power (about 82%) than [example 1](#).

◀

► Example 6: Power for a one-sided test

Continuing with [example 5](#), suppose that we are interested in testing whether the odds ratio is greater than 1 because we hypothesize that a history of smoking will lead to an increased incidence of lung cancer. In this case, we can specify the `onesided` option to calculate power for a one-sided test.

```
. power mcc .22, oratio(1.7) n(300) onesided
Estimated power for a matched case-control study
Asymptotic z test, 1:1 matched design
Ho: OR = 1 versus Ha: OR > 1
Study parameters:
    alpha =    0.0500
  N cases =     300
    delta =    1.7000
     p0 =    0.2200
    oratio =    1.7000
     corr =    0.0000
      M =      1
Estimated power:
    power =    0.8931
```

As expected, the power of the one-sided is higher (89%) than the power of the corresponding two-sided test.

◀

Computing target odds ratio

Sometimes, we may be interested in determining the smallest effect that will yield a statistically significant result for a prespecified sample size and power. In this case, power, sample size, and the exposure probability among controls must be specified. The effect size in `power mcc` is expressed as an odds ratio of exposure in cases relative to controls.

► Example 7: Minimum detectable odds ratio with 1:1 matching

Continuing with [example 5](#), we now would like to calculate the minimum detectable odds ratio that we can identify with the study design we have planned and knowing that we will be able to recruit 300 cases. We again specify 300 cases in option `n()` and also specify 80% power by using option `power(0.8)`:

```
. power mcc .22, n(300) power(.8)
Performing iteration ...
Estimated odds ratio for a matched case-control study
Asymptotic z test, 1:1 matched design
Ho: OR = 1 versus Ha: OR != 1
Study parameters:
    alpha =    0.0500
    power =    0.8000
  N cases =     300
     p0 =    0.2200
    corr =    0.0000
      M =      1
Estimated effect size and odds ratio:
    delta =    1.6783
    odds ratio =    1.6783
```

Our minimum detectable odds ratio is about 1.68 for this study design.

◀

► **Example 8: Minimum detectable odds ratio with 1: M matching**

We can specify other options to tailor our estimates of the minimum detectable odds ratio to accommodate other study design parameters. Continuing with [example 7](#), we may wish to calculate the minimum detectable odds ratio if we adopt a 1:2 matched design.

```
. power mcc .22, n(300) power(.8) m(2)
Performing iteration ...
Estimated odds ratio for a matched case-control study
Asymptotic z test, 1:2 matched design
Ho: OR = 1 versus Ha: OR != 1
Study parameters:
      alpha =    0.0500
      power =    0.8000
      N cases =     300
      p0 =    0.2200
      corr =    0.0000
      M =      2
Estimated effect size and odds ratio:
      delta =    1.5656
      odds ratio = 1.5656
```

Our minimum detectable odds ratio estimate decreases to 1.57 from 1.68 when the number of cases is held constant at 300.

Comparing results with [example 7](#), we see that increasing M while holding power and sample size constant decreases the minimum detectable odds ratio. Consequently, increasing M while holding sample size and the odds ratio constant increases power.

◀

Testing hypotheses in matched case–control studies

Matched case–control data can be organized in two ways: long and wide format. In long format, each row corresponds to a person; this is the format used by `clogit` and `mhodds` (see [\[R\] clogit](#) and [\[R\] epitab](#)). In wide format, each row corresponds to a matched pair or set; this is the format used by `mcc` (see [\[R\] epitab](#)).

► **Example 9: Analysis of matched case–control data in long format**

[Hosmer, Lemeshow, and Sturdivant \(2013\)](#) describe a study in which low birthweight infants were matched with normal weight infants on the basis of the age of the mother. The mothers were then asked whether or not they smoked during pregnancy. We will list a subset of these data to illustrate long format.

```
. use http://www.stata-press.com/data/r14/lowbirth2
(Appplied Logistic Regression, Hosmer & Lemeshow)
. list pairid low smoke in 1/6, sepby(pairid)
```

	pairid	low	smoke
1.	1	0	0
2.	1	1	1
3.	2	0	0
4.	2	1	0
5.	3	0	0
6.	3	1	0

The first column is the case–control identifier for each pair of infants. The second column identifies each infant as a case (`low==1`, indicating low birthweight) or a control (`low==0`). The third column identifies each infant as exposed (`smoke==1`, indicating that the mother smoked) or not exposed (`smoke==0`). We can estimate the odds ratio and test the null hypothesis that it equals one by using `clogit`.

We must use option `group()` to specify the identifier for our case–control matched pairs. We also specify option `nolog` to suppress the iteration log and `or` to view the result as an odds ratio.

```
. clogit low smoke, group(pairid) nolog or
Conditional (fixed-effects) logistic regression
```

```

                                     Number of obs   =       112
                                     LR chi2(1)       =         6.79
                                     Prob > chi2      =       0.0091
                                     Pseudo R2       =       0.0875
Log likelihood = -35.419282
```

	low	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
	smoke	2.75	1.135369	2.45	0.014	1.224347 6.176763

The estimated odds ratio is 2.75. We reject the null hypothesis that mothers who smoke and mothers who do not smoke have equal odds of giving birth to an infant with low birthweight at the 5%-level ($p = 0.014$).

◀

▷ Example 10: Analysis of paired case–control data in wide format

Continuing with [example 9](#), because the [Hosmer, Lemeshow, and Sturdivant \(2013\)](#) data are for a study with matched pairs, we could also conduct a classic McNemar’s test. In Stata, McNemar’s test is calculated by the `mcc` command; see [\[R\] epitab](#). The `mcc` command, however, requires that the data be in wide form. For details, see the [technical note](#) in [\[R\] clogit](#).

We can reshape our low birthweight data using the `reshape` command.

```
. keep low smoke pairid
. reshape wide smoke, i(pairid) j(low 0 1)
Data                                long  ->  wide
-----
Number of obs.                      112  ->   56
Number of variables                  3    ->   3
j variable (2 values)               low   ->  (dropped)
xij variables:
                                smoke  ->  smoke0 smoke1
```

```
. list pairid smoke1 smoke0 in 1/3
```

	pairid	smoke1	smoke0
1.	1	1	0
2.	2	0	0
3.	3	0	0

The variable `smoke1` indicates whether or not the case mother smoked, and the variable `smoke0` indicates whether or not the control mother smoked. We can now use `mcc` to estimate the overall odds ratio.

```
. mcc smoke1 smoke0
```

Cases	Controls		Total
	Exposed	Unexposed	
Exposed	8	22	30
Unexposed	8	18	26
Total	16	40	56

```
McNemar's chi2(1) = 6.53 Prob > chi2 = 0.0106
```

```
Exact McNemar significance probability = 0.0161
```

```
Proportion with factor
```

```
Cases .5357143
```

```
Controls .2857143
```

```
[95% Conf. Interval]
```

```
difference .25 .0519726 .4480274
```

```
ratio 1.875 1.148685 3.060565
```

```
rel. diff. .35 .1336258 .5663742
```

```
odds ratio 2.75 1.179154 7.143667 (exact)
```

Again, the estimated odds ratio θ is 2.75, and McNemar's χ^2 is 6.53. As with the analysis using `clogit`, we reject the null hypothesis of equal odds at the 5% significance level ($p = 0.0106$). The confidence interval for the odds ratio calculated by `clogit` and `mcc` differ slightly due to different estimation methods.

We could also calculate the same test statistic based on symmetry and marginal homogeneity; see [\[R\] symmetry](#) for further details.

Stored results

power mcc stores the following in `r()`:

Scalars

<code>r(alpha)</code>	significance level
<code>r(power)</code>	power
<code>r(beta)</code>	probability of a type II error
<code>r(delta)</code>	effect size
<code>r(N)</code>	sample size
<code>r(nfractional)</code>	1 if <code>nfractional</code> is specified; 0 otherwise
<code>r(onesided)</code>	1 for a one-sided test; 0 otherwise
<code>r(p0)</code>	probability of exposure among controls
<code>r(M)</code>	number of matched controls per case
<code>r(F_M)</code>	ratio of the number of cases relative to the 1:1 paired design
<code>r(oratio)</code>	odds ratio
<code>r(corr)</code>	correlation of exposure between matched cases and controls
<code>r(init)</code>	initial value for the sample size or effect size
<code>r(maxiter)</code>	maximum number of iterations
<code>r(iter)</code>	number of iterations performed
<code>r(tolerance)</code>	requested parameter tolerance
<code>r(deltax)</code>	final parameter tolerance achieved
<code>r(ftolerance)</code>	requested distance of the objective function from zero
<code>r(function)</code>	final distance of the objective function from zero
<code>r(converged)</code>	1 if iteration algorithm converged; 0 otherwise
<code>r(separator)</code>	number of lines between separator lines in the table
<code>r(divider)</code>	1 if <code>divider</code> is requested in the table; 0 otherwise

Macros

<code>r(type)</code>	test
<code>r(method)</code>	mcc
<code>r(direction)</code>	upper or lower
<code>r(columns)</code>	displayed table columns
<code>r(labels)</code>	table column labels
<code>r(widths)</code>	table column widths
<code>r(formats)</code>	table column formats

Matrix

<code>r(pss_table)</code>	table of results
---------------------------	------------------

Methods and formulas

Consider a 1: M matched case–control study, where n cases with a disease are matched to M controls without the disease on the basis of similar values of confounding variables such as age, gender, etc. Some patients in the study have prior exposure to a certain risk factor of interest. Below we provide power and sample-size formulas based on Dupont (1988). Also see Breslow and Day (1980, sec. 5.3) for background on the analysis of 1: M matched data.

Referring to the table in the [Introduction](#), let p_0 denote the probability of exposure among the control patients. Under the null hypothesis, the odds ratio θ of developing the disease in exposed and unexposed subjects is constant for equal values of the confounding variables.

The value of p_{ij} for each cell represents the joint probability of exposure status of the matched case–control pair for $i, j = 0, 1$. (With M matches, the first control is used.) Let p_0 and p_1 represent the probability of exposure of the control and case patient, respectively. Let ρ denote the correlation coefficient for exposure in matched pairs of case–control patients. Let $q_0 = 1 - p_0$ and $q_1 = 1 - p_1$, then the odds ratio is given by $\theta = p_1q_0/p_0q_1$ if $\rho = 0$ and $\theta = p_{10}/p_{01}$ if $\rho \neq 0$.

The correlation coefficient can be expressed as

$$\rho = \frac{p_{11}p_{00} - p_{10}p_{01}}{\sqrt{p_1q_1p_0q_0}} \quad (1)$$

The individual cell probabilities can be expressed in terms of exposure probabilities as

$$\begin{aligned} p_{11} &= p_1p_0 + \rho\sqrt{p_1q_1p_0q_0} \\ p_{10} &= p_1q_0 - \rho\sqrt{p_1q_1p_0q_0} \\ p_{01} &= q_1p_0 - \rho\sqrt{p_1q_1p_0q_0} \\ p_{00} &= q_1q_0 + \rho\sqrt{p_1q_1p_0q_0} \end{aligned}$$

Let p_{0+} and p_{0-} denote the probability that a control patient is exposed given the corresponding matched case is or is not exposed, respectively. Then

$$\begin{aligned} p_{0+} &= \frac{p_{11}}{p_1} \\ p_{0-} &= \frac{p_{01}}{q_1} \\ q_{0+} &= 1 - p_{0+} \\ q_{0-} &= 1 - p_{0-} \end{aligned}$$

The probability of observing m exposed subjects among a case matched with M controls is given by

$$t_m = p_1 \binom{M}{m-1} p_{0+}^{m-1} q_{0+}^{M-m+1} + q_1 \binom{M}{m} p_{0-}^m q_{0-}^{M-m}$$

for $m = 1, \dots, M$.

Let $n_{i,j}$ denote the number of matched sets such that $n_{1,j}$ is the number of exposed cases matched with j of M exposed controls and $n_{0,j}$ is the number of nonexposed cases matched with j of M exposed controls. Then, the number of matched sets in which m subjects were exposed is

$$T_m = n_{1,m-1} + n_{0,m}$$

Define a matched set as a discordant matched set if there is a least one discordant pair between the case and the M controls. Then, the number of discordant matched sets in which the case patients were exposed is

$$y = \sum_{m=1}^M n_{1,m-1}$$

Denote E_θ and s_θ the conditional mean and standard deviation of y , respectively, given $T_m = E(T_m) = nt_m$, for $m = 1, \dots, M$. Then $E_\theta = ne_\theta$ and $s_\theta = \sqrt{n\nu_\theta}$, where

$$e_\theta = \sum_{m=1}^M \frac{mt_m\theta}{m\theta + M - m + 1} \quad \text{and} \quad \nu_\theta = \sum_{m=1}^M \frac{mt_m\theta(M - m + 1)}{(m\theta + M - m + 1)^2}$$

(Breslow and Day 1980, eq. 5.19).

Let $L_\alpha = (E_1 - E_\theta - z_{1-\alpha}s_1)/s_\theta$ and $U_\alpha = (E_1 - E_\theta + z_{1-\alpha}s_1)/s_\theta$, where $z_{1-\alpha}$ is the quantile of a standard normal distribution such that $P(Z \geq z_{1-\alpha}) = \alpha$ and $\Phi(\cdot)$ is the cumulative of a standard normal distribution.

Then, the power of the test is given by

$$1 - \beta = \begin{cases} \Phi(L_{\alpha/2}) + 1 - \Phi(U_{\alpha/2}) & \text{for a two-sided test} \\ \Phi(L_\alpha) & \text{for a lower one-sided test} \\ 1 - \Phi(U_\alpha) & \text{for an upper one-sided test} \end{cases} \quad (2)$$

The sample size for a one-sided test can be obtained from the inverse of the power computation, $n = (z_{1-\beta}\sqrt{V_\theta} + z_{1-\alpha}\sqrt{V_1})^2/(e_1 - e_\theta)^2$. The sample size for a two-sided test and the minimum detectable odds ratio θ must be computed iteratively from (2). The starting value for the sample-size computation is the sample-size estimate for the one-sided test. For the minimum detectable odds ratio, the starting value is obtained from a bisection algorithm.

Let F_M denote the ratio of the sample sizes for a study with a 1: M matched design relative to a study with a 1:1 matched design. If n_1 is the sample size required for a study with 1 control matched to 1 case and n_M is the sample size required for a study with M controls matched to 1 case, then $F_M = n_M/n_1$.

References

- Breslow, N. E., and N. E. Day. 1980. *Statistical Methods in Cancer Research: Vol. 1—The Analysis of Case–Control Studies*. Lyon: IARC.
- Dupont, W. D. 1988. Power calculations for matched case–control studies. *Biometrics* 44: 1157–1168.
- Hosmer, D. W., Jr., S. A. Lemeshow, and R. X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: Wiley.

Also see

- [PSS] **power** — Power and sample-size analysis for hypothesis tests
- [PSS] **power pairedproportions** — Power analysis for a two-sample paired-proportions test
- [PSS] **power, graph** — Graph results from the power command
- [PSS] **power, table** — Produce table of results from the power command
- [PSS] **Glossary**
- [R] **clomit** — Conditional (fixed-effects) logistic regression
- [R] **epitab** — Tables for epidemiologists
- [R] **symmetry** — Symmetry and marginal homogeneity tests