

## Chapter 208

# Confidence Intervals for the Odds Ratio of Two Proportions using an Unmatched Case-Control Design

---

### Introduction

This routine calculates the group sample sizes necessary to achieve a specified confidence interval width of the odds ratio (OR) from data collected using an unmatched case-control design. Such a design is used for retrospective studies.

OR is calculated as

$$OR = \frac{p_1(1 - p_2)}{p_2(1 - p_1)}$$

where  $p_1$  and  $p_2$  are the *probabilities of exposure* among cases (disease) and controls (no disease), respectively.

This routine is partially based on O'Neill (1988). This paper provides a useful overview of the vaccine efficacy studies, the assumptions that are made, and discussion of when this design is useful. We highly recommend it.

---

### Technical Details

This section will review the details of calculating sample sizes for the odds ratio of two proportions in case-control designs.

---

### Comparing Two Proportions

Suppose you have two populations from which dichotomous (binary) responses will be recorded. The probability of obtaining an event of interest (risk factor exposure) in population 1 (the case group) is  $p_1$  and in population 2 (the control group) is  $p_2$ . The risk factor is a dichotomous variable that represents a variable (vaccination, smoking status, etc.) that is thought to be related to the outcome (disease or no-disease). The corresponding failure probabilities are given by  $q_1 = 1 - p_1$  and  $q_2 = 1 - p_2$ .

The assumption is made that the responses from each group follow a binomial distribution. This means that the event probability  $p_i$  is the same for all subjects within a population and that the responses from one subject to the next are independent of one another.

## Confidence Intervals for the Odds Ratio of Two Proportions using an Unmatched Case-Control Design

Random samples of  $n_1$  and  $n_2$  individuals are obtained from these two populations. The data from these samples can be displayed in a 2-by-2 contingency table as follows

	<b>Cases</b>	<b>Controls</b>	<b>Total</b>
<b>Vaccinees</b>	$x_{11}, a$	$x_{12}, b$	$m_1$
<b>Controls</b>	$x_{21}, c$	$x_{22}, d$	$m_2$
<b>Totals</b>	$n_1, m$	$n_2, n$	$N$

The binomial proportions are estimated from these data using the formulae

$$\hat{p}_1 = \frac{x_{11}}{n_1} \text{ and } \hat{p}_2 = \frac{x_{12}}{n_2}$$

In this procedure, our attention will focus on the using the odds ratio to compare the two binomial proportions. The odds ratio is given by  $\psi = (p_1/q_1)/(p_2/q_2)$  which gives the relative change in the odds of exposure to the risk factor among the cases to the odds of the controls.

Although the odds ratio is more complicated than the simple risk ratio of the disease attack rates, it is used in this case because a direct estimate of the ratio of the attack rates is not available in a case-control design. In this case, it is hoped that the odds ratio will provide a reasonable approximation of the risk ratio of the attack rates.

---

## Confidence Intervals for the Odds Ratio

Many methods have been devised for computing confidence intervals for the odds ratio of two proportions. Eight of these methods are available in this procedure. They are

1. Exact (Conditional)
2. Score (Farrington and Manning)
3. Score (Miettinen and Nurminen)
4. Fleiss
5. Logarithm
6. Mantel-Haenszel
7. Simple
8. Simple + 1/2

## Confidence Intervals for the Odds Ratio of Two Proportions using an Unmatched Case-Control Design

### Conditional Exact

The conditional exact confidence interval of the odds ratio is calculated using the noncentral hypergeometric distribution as given in Sahai and Khurshid (1995). That is, a  $100(1 - \alpha)\%$  confidence interval is found by searching for  $\psi_L$  and  $\psi_U$  such that

$$\frac{\sum_{k=x}^{k_2} \binom{n_1}{k} \binom{n_2}{m_1 - k} (\psi_L)^k}{\sum_{k=k_1}^{k_2} \binom{n_1}{k} \binom{n_2}{m_1 - k} (\psi_L)^k} = \frac{\alpha}{2}$$

and

$$\frac{\sum_{k=k_1}^x \binom{n_1}{k} \binom{n_2}{m_1 - k} (\psi_U)^k}{\sum_{k=k_1}^{k_2} \binom{n_1}{k} \binom{n_2}{m_1 - k} (\psi_U)^k} = \frac{\alpha}{2}$$

where

$$k_1 = \max(0, m_1 - n_1) \text{ and } k_2 = \min(n_1, m_1)$$

### Farrington and Manning's Score

Farrington and Manning (1990) developed a test statistic similar to that of Miettinen and Nurminen but with the factor  $N/(N-1)$  removed.

The formula for computing this test statistic is

$$z_{FMO} = \frac{\frac{(\hat{p}_1 - \tilde{p}_1)}{\tilde{p}_1 \tilde{q}_1} - \frac{(\hat{p}_2 - \tilde{p}_2)}{\tilde{p}_2 \tilde{q}_2}}{\sqrt{\left( \frac{1}{n_1 \tilde{p}_1 \tilde{q}_1} + \frac{1}{n_2 \tilde{p}_2 \tilde{q}_2} \right)}}$$

where the estimates  $\tilde{p}_1$  and  $\tilde{p}_2$  are computed as in the corresponding test of Miettinen and Nurminen (1985) as

$$\tilde{p}_1 = \frac{\tilde{p}_2 \psi_0}{1 + \tilde{p}_2 (\psi_0 - 1)}$$

$$\tilde{p}_2 = \frac{-B + \sqrt{B^2 - 4AC}}{2A}$$

$$A = n_2 (\psi_0 - 1)$$

$$B = n_1 \psi_0 + n_2 - m_1 (\psi_0 - 1)$$

$$C = -m_1$$

## Confidence Intervals for the Odds Ratio of Two Proportions using an Unmatched Case-Control Design

Farrington and Manning (1990) proposed inverting their score test to find the confidence interval. The lower limit is found by solving

$$z_{FMO} = |z_{\alpha/2}|$$

and the upper limit is the solution of

$$z_{FMO} = -|z_{\alpha/2}|$$

### Miettinen and Nurminen's Score

Miettinen and Nurminen (1985) proposed a test statistic for testing whether the odds ratio is equal to a specified value  $\psi_0$ . Because the approach they used with the difference and ratio does not easily extend to the odds ratio, they used a score statistic approach for the odds ratio. The regular MLE's are  $\hat{p}_1$  and  $\hat{p}_2$ . The constrained MLE's are  $\tilde{p}_1$  and  $\tilde{p}_2$ , These estimates are constrained so that  $\tilde{\psi} = \psi_0$ . A correction factor of  $N/(N-1)$  is applied to make the variance estimate less biased. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing the test statistic is

$$z_{MNO} = \frac{\frac{(\hat{p}_1 - \tilde{p}_1)}{\tilde{p}_1 \tilde{q}_1} - \frac{(\hat{p}_2 - \tilde{p}_2)}{\tilde{p}_2 \tilde{q}_2}}{\sqrt{\left(\frac{1}{n_1 \tilde{p}_1 \tilde{q}_1} + \frac{1}{n_2 \tilde{p}_2 \tilde{q}_2}\right) \left(\frac{N}{N-1}\right)}}$$

where

$$\tilde{p}_1 = \frac{\tilde{p}_2 \psi_0}{1 + \tilde{p}_2 (\psi_0 - 1)}$$

$$\tilde{p}_2 = \frac{-B + \sqrt{B^2 - 4AC}}{2A}$$

$$A = n_2 (\psi_0 - 1)$$

$$B = n_1 \psi_0 + n_2 - m_1 (\psi_0 - 1)$$

$$C = -m_1$$

Miettinen and Nurminen (1985) proposed inverting their score test to find the confidence interval. The lower limit is found by solving

$$z_{MNO} = |z_{\alpha/2}|$$

and the upper limit is the solution of

$$z_{MNO} = -|z_{\alpha/2}|$$

## Confidence Intervals for the Odds Ratio of Two Proportions using an Unmatched Case-Control Design

## Iterated Method of Fleiss

Fleiss (1981) presents an improved confidence interval for the odds ratio. This method forms the confidence interval as all those values of the odds ratio which would not be rejected by a chi-square hypothesis test. Fleiss gives the following details about how to construct this confidence interval. To compute the lower limit, do the following.

1. For a trial value of  $\psi$ , compute the quantities  $X$ ,  $Y$ ,  $W$ ,  $F$ ,  $U$ , and  $V$  using the formulas

$$X = \psi(m + s) + (n - s)$$

$$Y = \sqrt{X^2 - 4ms\psi(\psi - 1)}$$

$$A = \frac{X - Y}{2(\psi - 1)}$$

$$B = s - A$$

$$C = m - A$$

$$D = f - m + A$$

$$W = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}$$

$$F = \left(a - A - \frac{1}{2}\right)^2 W - z_{\alpha/2}^2$$

$$T = \frac{1}{2(\psi - 1)^2} \left( Y - n_{..} - \frac{\psi - 1}{Y} [X(m + s) - 2ms(2\psi - 1)] \right)$$

$$U = \frac{1}{B^2} + \frac{1}{C^2} - \frac{1}{A^2} - \frac{1}{D^2}$$

$$V = T \left[ \left(a - A - \frac{1}{2}\right)^2 U - 2W \left(a - A - \frac{1}{2}\right) \right]$$

Finally, use the updating equation below to calculate a new value for the odds ratio using the updating equation

$$\psi^{(k+1)} = \psi^{(k)} - \frac{F}{V}$$

2. Continue iterating until the value of  $F$  is arbitrarily close to zero.

The upper limit is found by substituting  $+\frac{1}{2}$  for  $-\frac{1}{2}$  in the formulas for  $F$  and  $V$ .

Confidence limits for the *relative risk* can be calculated using the expected counts  $A$ ,  $B$ ,  $C$ , and  $D$  from the last iteration of the above procedure. The lower limit of the relative risk

$$\phi_{lower} = \frac{A_{lower}n}{B_{lower}m}$$

$$\phi_{upper} = \frac{A_{upper}n}{B_{upper}m}$$

## Confidence Intervals for the Odds Ratio of Two Proportions using an Unmatched Case-Control Design

### Mantel-Haenszel

The common estimate of the logarithm of the odds ratio is used to create this estimator. That is

$$\ln(\hat{\psi}) = \ln\left(\frac{ad}{bc}\right)$$

The standard error of this estimator is estimated using the Robins, Breslow, Greenland (1986) estimator which performs well in most situations. The standard error is given by

$$se(\ln(\hat{\psi})) = \sqrt{\frac{A}{2C} + \frac{AD + BC}{2CD} + \frac{B}{2D}}$$

where

$$A = \frac{a + d}{N}$$

$$B = \frac{b + c}{N}$$

$$C = \frac{ad}{N}$$

$$D = \frac{bc}{N}$$

The confidence limits are calculated as

$$\hat{\psi}_{lower} = \exp(\ln(\hat{\psi}) - z_{1-\alpha/2} se(\ln(\hat{\psi})))$$

$$\hat{\psi}_{upper} = \exp(\ln(\hat{\psi}) + z_{1-\alpha/2} se(\ln(\hat{\psi})))$$

### Simple, Simple + 1/2, and Logarithm

The simple estimate of the odds ratio uses the formula

$$\begin{aligned}\hat{\psi} &= \frac{\hat{p}_1 \hat{q}_2}{\hat{p}_2 \hat{q}_1} \\ &= \frac{ad}{bc}\end{aligned}$$

The standard error of this estimator is estimated by

$$se(\hat{\psi}) = \hat{\psi} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Problems occur if any one of the quantities  $a$ ,  $b$ ,  $c$ , or  $d$  are zero. To correct this problem, many authors recommend adding one-half to each cell count so that a zero cannot occur. Now, the formulas become

$$\hat{\psi}' = \frac{(a + 0.5)(d + 0.5)}{(b + 0.5)(c + 0.5)}$$

and

$$se(\hat{\psi}') = \hat{\psi}' \sqrt{\frac{1}{a + 0.5} + \frac{1}{b + 0.5} + \frac{1}{c + 0.5} + \frac{1}{d + 0.5}}$$

## Confidence Intervals for the Odds Ratio of Two Proportions using an Unmatched Case-Control Design

The distribution of these direct estimates of the odds ratio do not converge to normality as fast as does their logarithm, so the logarithm of the odds ratio is used to form confidence intervals. The formula for the standard error of the log odds ratio is

$$L' = \ln(\hat{\psi}')$$

and

$$se(L') = \sqrt{\frac{1}{a+0.5} + \frac{1}{b+0.5} + \frac{1}{c+0.5} + \frac{1}{d+0.5}}$$

A  $100(1 - \alpha)\%$  confidence interval for the log odds ratio is formed using the standard normal distribution as follows

$$\hat{\psi}_{lower} = \exp(L' - z_{1-\alpha/2} se(L'))$$

$$\hat{\psi}_{upper} = \exp(L' + z_{1-\alpha/2} se(L'))$$

See Fleiss et al (2003) for more details.

## Sample Size Estimation

Sample size estimation is straight forward. For each method, anticipated (planned) values of the two probabilities are substituted for the estimated values given in the above formulas, along with the sample sizes of each of the groups and the confidence level. The width between the two limits becomes the measure of precision. The more narrow the confidence interval, the more precise it is.

To find an appropriate sample size for a given set of parameters, a binary search is conducted for the smallest sample size that meets the width requirements.

## Confidence Level

The confidence level,  $1 - \alpha$ , has the following interpretation. If thousands of random samples of size  $n_1$  and  $n_2$  are drawn from populations 1 and 2, respectively, and a confidence interval for the true difference/ratio/odds ratio of proportions is calculated for each pair of samples, the proportion of those intervals that will include the true difference/ratio/odds ratio of proportions is  $1 - \alpha$ .

## Example 1 – Calculating Sample Size

Suppose a study is planned in which the researcher wishes to construct a two-sided 95% confidence interval for the odds ratio such that the width of the interval is no wider than 0.2. Additional widths of 0.15 and 0.25 are also to be investigated. The confidence interval method to be used is the Mantel-Haenszel method. The confidence level is set to 0.95. OR is set at 0.7 and 0.8. P2 is 0.06. A case-control design is planned in which there will be four controls chosen for each case.

The goal is to determine the necessary sample size.

### Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the procedure window. You may then make the appropriate entries as listed below, or open **Example 1** by going to the **File** menu and choosing **Open Example Template**.

<u>Option</u>	<u>Value</u>
<b>Design Tab</b>	
Solve For .....	<b>Sample Size</b>
Method.....	<b>Mantel-Haenszel</b>
Interval Type .....	<b>Two-Sided</b>
Confidence Level (1 - $\alpha$ ) .....	<b>0.95</b>
Group Allocation .....	<b>Enter R = N2/N1, solve for N1 and N2</b>
R (Group Sample Size Ratio) .....	<b>4</b>
W (Confidence Interval Width).....	<b>0.15 0.2 0.25</b>
Odds Ratio Input Type.....	<b>Enter OR and P2</b>
OR (Odds Ratio) .....	<b>0.7 0.8</b>
P2 (Probability of Exposure in Controls).....	<b>0.06</b>

### Annotated Output

Click the Calculate button to perform the calculations and generate the following output.

### Numeric Results

#### Numeric Results

Interval Type: Two-Sided  
Confidence Interval Method: Mantel-Haenszel

Conf Level	Case N1	Control N2	N	Target R	Target C.I. Width W <sub>T</sub>	Actual C.I. Width W <sub>A</sub>	Probability of Exposure Case P1	Probability of Exposure Control P2	Odds Ratio OR	Lower Conf Limit of OR LCL	Upper Conf Limit of OR UCL
0.95	9694	38776	48470	4	0.15	0.15000	0.04277	0.06	0.7	0.62901	0.77901
0.95	5469	21876	27345	4	0.20	0.20000	0.04277	0.06	0.7	0.60711	0.80711
0.95	3514	14056	17570	4	0.25	0.24997	0.04277	0.06	0.7	0.58608	0.83606
0.95	11427	45708	57135	4	0.15	0.15000	0.04858	0.06	0.8	0.72851	0.87851
0.95	6442	25768	32210	4	0.20	0.20000	0.04858	0.06	0.8	0.70623	0.90623
0.95	4135	16540	20675	4	0.25	0.24999	0.04858	0.06	0.8	0.68471	0.93470

#### References

- O'Neill, Robert T. 1988. 'On Sample Sizes to Estimate the Protective Efficacy of a Vaccine'. *Statistics in Medicine*, Volume 7, Pages 1279-1288.
- Farrington, C. P. and Manning, G. 1990. 'Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-Zero Risk Difference or Non-Unity Relative Risk.' *Statistics in Medicine*, Vol. 9, pages 1447-1454.



## Confidence Intervals for the Odds Ratio of Two Proportions using an Unmatched Case-Control Design

Fleiss, J. L., Levin, B., Paik, M.C. 2003. Statistical Methods for Rates and Proportions. Third Edition. John Wiley & Sons. New York.  
 Miettinen, O.S. and Nurminen, M. 1985. 'Comparative analysis of two rates.' Statistics in Medicine 4: 213-226.  
 Robbins, Breslow, and Greenland. 1986. 'Estimators of the Mantel-Haenszel Variance Consistent in both Sparse Data and Large-Strata Limiting Models.' Biometrics, Volume 42, 311-323.  
 Sahai, H. and Khurshid, A. 1995. Statistics in Epidemiology. CRC Press. Boca Raton, Florida.  
 Schlesselman, James. 1982. Case-Control Studies: Design, Conduct, Analysis. Oxford University Press. New York.

### Report Definitions

Confidence level is the proportion of confidence intervals (constructed with this same confidence level, sample size, etc.) that would contain the true value of OR.

N1 is the number of subjects sampled from the population of cases.

N2 is the number of subjects sampled from the population of controls.

N is the total sample size, N1 + N2.

Target R is the desired sampling ratio.  $R = N2 / N1$  or  $N2 = R(N1)$ . Often, R is labelled 'C' in the case-control literature.

W<sub>T</sub> is the target width of the confidence interval of OR.

W<sub>A</sub> is the actual width of the confidence interval of OR that was computed by the procedure.

P1 is the probability of having been vaccinated (vaccine prevalence) in the case group.

P2 is the probability of having been vaccinated (vaccine prevalence) in the control group.

OR is the Odds Ratio of the exposure probability of cases and the exposure probability of controls.

LCL is the lower confidence limit of OR.

UCL is the upper confidence limit of OR.

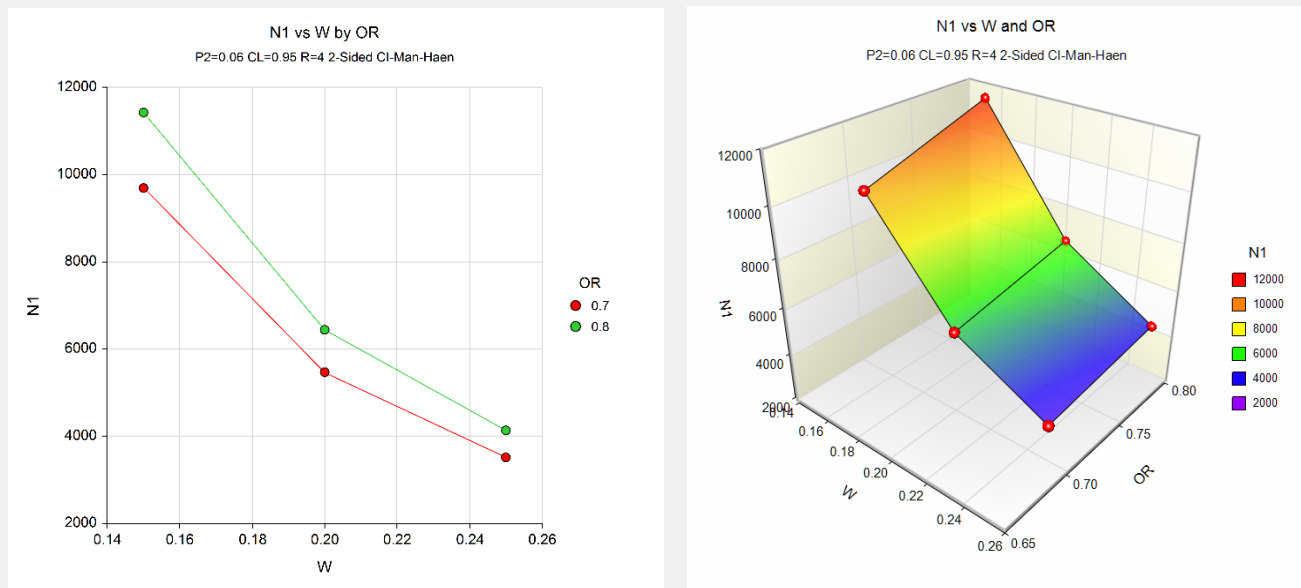
### Summary Statements

Sample sizes of 9694 from the cases and 38776 from the controls a width of 0.15 using a two-sided confidence interval of OR based on the Mantel-Haenszel method. The odds ratio is assumed to be 0.7. The confidence level of the interval is 0.95. The probability of exposure among the cases is 0.04277. The probability of exposure among the controls is 0.06.

This report shows the calculated sample sizes for each of the scenarios.

## Chart Section

### Chart Section



These plots show the group sample size versus the confidence interval width for the two OR values.

## Example 2 – Validation using O’Neill (1988)

O’Neill (1988) page 1286 gives an example of a sample size calculation for a confidence interval for OR when the confidence level is 95%, OR = 0.2 (VE is 0.8), P2 is 0.2, and the desired width is 0.24. The confidence interval method is Mantel-Haenszel. The design calls for 4 controls for each case. The sample size in from the cases is 280.

### Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the procedure window. You may then make the appropriate entries as listed below, or open **Example 2** by going to the **File** menu and choosing **Open Example Template**.

<u>Option</u>	<u>Value</u>
<b>Design Tab</b>	
Solve For .....	<b>Sample Size</b>
Method.....	<b>Mantel-Haenszel</b>
Interval Type .....	<b>Two-Sided</b>
Confidence Level (1 - $\alpha$ ) .....	<b>0.95</b>
Group Allocation .....	<b>Enter R = N2/N1, solve for N1 and N2</b>
R (Group Sample Size Ratio) .....	<b>4</b>
W (Confidence Interval Width).....	<b>0.24</b>
Odds Ratio Input Type.....	<b>Enter OR and P2</b>
OR (Odds Ratio) .....	<b>0.2</b>
P2 (Probability of Exposure in Controls).....	<b>0.2</b>

### Output

Click the Calculate button to perform the calculations and generate the following output.

### Numeric Results

Numeric Results												
Interval Type:		Two-Sided										
Confidence Interval Method:		Mantel-Haenszel										
Conf Level	Case	Control	N	Target R	Target C.I. Width	Actual C.I. Width	Probability of Exposure		Odds Ratio	Lower Conf Limit of OR	Upper Conf Limit of OR	
1 - $\alpha$	N1	N2		R	W $\tau$	W $A$	Case P1	Control P2	OR	LCL	UCL	
0.95	281	1124	1405	4	0.24	0.23969	0.04762	0.2	0.2	0.11331	0.353	

PASS has also computed an N1 of 281 which is just slightly above the 280 in the article. This difference arises because of rounding. The procedure is validated.