Title

> **epitab** — Tables for epidemiologists

Description

ir is used with incidence-rate (incidence-density or person-time) data. It calculates point estimates and confidence intervals for the incidence-rate ratio and difference, along with attributable or prevented fractions for the exposed and total population. iri is the immediate form of ir; see [U] **19 Immediate commands**. Also see [R] **poisson** and [ST] **stcox** for related commands.

cs is used with cohort study data with equal follow-up time per subject and sometimes with cross-sectional data. Risk is then the proportion of subjects who become cases. It calculates point estimates and confidence intervals for the risk difference, risk ratio, and (optionally) the odds ratio, along with attributable or prevented fractions for the exposed and total population. csi is the immediate form of cs; see [U] **19 Immediate commands**. Also see [R] **logistic** for related commands.

cc is used with case–control and cross-sectional data. It calculates point estimates and confidence intervals for the odds ratio, along with attributable or prevented fractions for the exposed and total population. cci is the immediate form of cc; see [U] **19 Immediate commands**. Also see [R] **logistic** for related commands.

tabodds is used with case–control and cross-sectional data. It tabulates the odds of failure against a categorical explanatory variable *expvar*. If *expvar* is specified, tabodds performs an approximate χ^2 test of homogeneity of odds and a test for linear trend of the log odds against the numerical code used for the categories of *expvar*. Both tests are based on the score statistic and its variance; see *Methods and formulas*. When *expvar* is absent, the overall odds are reported. The variable var_{case} is coded 0/1 for individual and simple frequency records and equals the number of cases for binomial frequency records.

Optionally, tabodds tabulates adjusted or unadjusted odds ratios, using either the lowest levels of *expvar* or a user-defined level as the reference group. If adjust(*varlist*) is specified, it produces odds ratios adjusted for the variables in *varlist* along with a (score) test for trend.

mhodds is used with case–control and cross-sectional data. It estimates the ratio of the odds of failure for two categories of *expvar*, controlled for specified confounding variables, $vars_{adjust}$, and tests whether this odds ratio is equal to one. When *expvar* has more than two categories but none are specified with the compare() option, mhodds assumes that *expvar* is a quantitative variable and calculates a 1-degree-of-freedom test for trend. It also calculates an approximate estimate of the log odds-ratio for a one-unit increase in *expvar*. This is a one-step Newton–Raphson approximation to the maximum likelihood estimate calculated as the ratio of the score statistic, U, to its variance, V (Clayton and Hills 1993, 103).

mcc is used with matched case–control data. It calculates McNemar's chi-squared; point estimates and confidence intervals for the difference, ratio, and relative difference of the proportion with the factor; and the odds ratio and its confidence interval. mcci is the immediate form of mcc; see [U] **19 Immediate commands**. Also see [R] **clogit** and [R] **symmetry** for related commands.

Quick start

Cohort studies

IRR and difference for the number of cases stored in `cases` for binary indicator `exposed` given time exposed `time`

```
ir cases exposed time
```

Crude and Mantel–Haenszel combined IRRs with test of homogeneity for strata defined by `svar`

```
ir cases exposed time, by(svar)
```

As above, and standardize the IRR by weighting variable `wvar1`

```
ir cases exposed time, by(svar) standard(wvar1)
```

As above, but use person-time of the unexposed group as weights

```
ir cases exposed time, by(svar) estandard
```

IRR and difference for 10 cases over 50 person-years in the exposed group and 15 cases over 100 person-years in the unexposed group

```
iri 10 15 50 100
```

Risk difference and ratio with binary indicators `case` and `exposed` using cumulative incidence data

```
cs case exposed [fweight=wvar2]
```

Add odds ratios and calculate Fisher's exact p

```
cs case exposed [fweight=wvar2], or exact
```

Internally standardized risk ratio for strata defined by `svar`

```
cs case exposed [fweight=wvar2], by(svar) istandard
```

Risk difference and ratio for 12 cases and 55 noncases among exposed subjects and 16 cases and 125 noncases among unexposed subjects

```
csi 12 16 55 125
```

Case–control studies

Odds ratios from summary data with binary indicators `case` and `exposed` and frequency weight `wvar3`

```
cc case exposed [fweight=wvar3]
```

As above, but stratify analysis by `svar` and perform Breslow–Day and Tarone's homogeneity tests

```
cc case exposed [fweight=wvar3], by(svar) bd tarone
```

Odds ratios for 37 exposed cases, 148 unexposed cases, 7 exposed controls, and 137 unexposed controls

```
cci 37 148 7 137
```

Odds of binary `event` against `catvar` using summary data with frequency weight `wvar4`

```
tabodds event catvar [fweight=wvar4]
```

As above, but report odds ratios with the fourth level of `catvar` as the reference

```
tabodds event catvar [fweight=wvar4], or base(4)
```

As above, but tabulate Mantel–Haenszel adjusted odds ratios adjusting for values of categorical variable `a`

```
tabodds event catvar [fweight=wvar4], base(4) adjust(a)
```

Graph odds and confidence intervals against categories of `catvar`

```
tabodds event catvar [fweight=wvar4], ciplot
```

Odds ratios for the effect of `catvar` on `event` controlling for categorical variable `a` using summary data with frequency weight `wvar5`

```
mhodds event catvar a [fweight=wvar5]
```

As above, but calculate odds ratios for each level of `svar`

```
mhodds event catvar a [fweight=wvar5], by(svar)
```

Maximum likelihood estimate of odds ratio for `a` equal to 4 compared with `a` equal to 1

```
mhodds event a [fweight=wvar5], compare(4,1)
```

Statistics on the difference in the proportion with the factor for exposed cases indicated in `expcase` and exposed controls indicated in `expcontrol` using summary data with frequency weight `wvar6`

```
mcc expcase expcontrol [fweight=wvar6]
```

As above, but indicate that there are 4 pairs where both cases and controls were exposed, 9 pairs where the case was exposed but the control was not, 3 pairs where the control was exposed but the case was not, and 14 pairs where neither subject was exposed

```
mcci 4 9 3 14
```

Menu

ir

Statistics > Epidemiology and related > Tables for epidemiologists > Incidence-rate ratio

iri

Statistics > Epidemiology and related > Tables for epidemiologists > Incidence-rate ratio calculator

cs

Statistics > Epidemiology and related > Tables for epidemiologists > Cohort study risk-ratio etc.

csi

Statistics > Epidemiology and related > Tables for epidemiologists > Cohort study risk-ratio etc. calculator

cc

Statistics > Epidemiology and related > Tables for epidemiologists > Case-control odds ratio

cci

Statistics > Epidemiology and related > Tables for epidemiologists > Case-control odds-ratio calculator

tabodds

Statistics > Epidemiology and related > Tables for epidemiologists > Tabulate odds of failure by category

mhodds

Statistics > Epidemiology and related > Tables for epidemiologists > Ratio of odds of failure for two categories

mcc

Statistics > Epidemiology and related > Tables for epidemiologists > Matched case-control studies

mcci

Statistics > Epidemiology and related > Tables for epidemiologists > Matched case-control calculator

Syntax

Cohort studies

ir *var*$_{\text{case}}$ *var*$_{\text{exposed}}$ *var*$_{\text{time}}$ $\big[$ *if* $\big]$ $\big[$ *in* $\big]$ $\big[$ *weight* $\big]$ $\big[$, *ir_options* $\big]$

iri #$_a$ #$_b$ #$_{N_1}$ #$_{N_2}$ $\big[$, <u>l</u>evel(#) $\big]$

cs *var*$_{\text{case}}$ *var*$_{\text{exposed}}$ $\big[$ *if* $\big]$ $\big[$ *in* $\big]$ $\big[$ *weight* $\big]$ $\big[$, *cs_options* $\big]$

csi #$_a$ #$_b$ #$_c$ #$_d$ $\big[$, *csi_options* $\big]$

Case–control studies

cc *var*$_{\text{case}}$ *var*$_{\text{exposed}}$ $\big[$ *if* $\big]$ $\big[$ *in* $\big]$ $\big[$ *weight* $\big]$ $\big[$, *cc_options* $\big]$

cci #$_a$ #$_b$ #$_c$ #$_d$ $\big[$, *cci_options* $\big]$

tabodds *var*$_{\text{case}}$ $\big[$ *expvar* $\big]$ $\big[$ *if* $\big]$ $\big[$ *in* $\big]$ $\big[$ *weight* $\big]$ $\big[$, *tabodds_options* $\big]$

mhodds *var*$_{\text{case}}$ *expvar* $\big[$ *vars*$_{\text{adjust}}$ $\big]$ $\big[$ *if* $\big]$ $\big[$ *in* $\big]$ $\big[$ *weight* $\big]$ $\big[$, *mhodds_options* $\big]$

Matched case–control studies

mcc *var*$_{\text{exposed\_case}}$ *var*$_{\text{exposed\_control}}$ $\big[$ *if* $\big]$ $\big[$ *in* $\big]$ $\big[$ *weight* $\big]$ $\big[$, <u>l</u>evel(#) $\big]$

mcci #$_a$ #$_b$ #$_c$ #$_d$ $\big[$, <u>l</u>evel(#) $\big]$

ir_options	Description
Options	
by(*varname* $\big[$, <u>miss</u>ing $\big]$)	stratify on *varname*
<u>es</u>tandard	combine external weights with within-stratum statistics
<u>i</u>standard	combine internal weights with within-stratum statistics
<u>s</u>tandard(*varname*)	combine user-specified weights with within-stratum statistics
pool	display pooled estimate
<u>nocr</u>ude	do not display crude estimate
<u>noh</u>om	do not display homogeneity test
ird	calculate standard incidence-rate difference
<u>l</u>evel(#)	set confidence level; default is level(95)

cs_options	Description
Options	
by(*varlist*[, <u>miss</u>ing])	stratify on *varlist*
<u>est</u>andard	combine external weights with within-stratum statistics
istandard	combine internal weights with within-stratum statistics
<u>st</u>andard(*varname*)	combine user-specified weights with within-stratum statistics
pool	display pooled estimate
<u>nocr</u>ude	do not display crude estimate
<u>noh</u>om	do not display homogeneity test
rd	calculate standardized risk difference
<u>b</u>inomial(*varname*)	number of subjects variable
or	report odds ratio
<u>w</u>oolf	use Woolf approximation to calculate SE and CI of the odds ratio
exact	calculate Fisher's exact *p*
<u>l</u>evel(*#*)	set confidence level; default is level(95)

csi_options	Description
or	report odds ratio
<u>w</u>oolf	use Woolf approximation to calculate SE and CI of the odds ratio
exact	calculate Fisher's exact *p*
<u>l</u>evel(*#*)	set confidence level; default is level(95)

cc_options	Description
Options	
by(*varname*[, <u>miss</u>ing])	stratify on *varname*
<u>est</u>andard	combine external weights with within-stratum statistics
istandard	combine internal weights with within-stratum statistics
<u>st</u>andard(*varname*)	combine user-specified weights with within-stratum statistics
<u>p</u>ool	display pooled estimate
<u>nocr</u>ude	do not display crude estimate
<u>noh</u>om	do not display homogeneity test
bd	perform Breslow–Day homogeneity test
tarone	perform Tarone's homogeneity test
<u>b</u>inomial(*varname*)	number of subjects variable
cornfield	use Cornfield approximation to calculate CI of the odds ratio
<u>w</u>oolf	use Woolf approximation to calculate SE and CI of the odds ratio
exact	calculate Fisher's exact *p*
<u>l</u>evel(*#*)	set confidence level; default is level(95)

cci_options	Description
<u>corn</u>field	use Cornfield approximation to calculate CI of the odds ratio
<u>woolf</u>	use Woolf approximation to calculate SE and CI of the odds ratio
<u>exact</u>	calculate Fisher's exact p
<u>level</u>(#)	set confidence level; default is level(95)

tabodds_options	Description

[Main]
<u>binomial</u>(*varname*)	number of subjects variable
<u>level</u>(#)	set confidence level; default is level(95)
or	report odds ratio
<u>adjust</u>(*varlist*)	report odds ratios adjusted for the variables in *varlist*
<u>base</u>(#)	reference group of control variable for odds ratio
<u>corn</u>field	use Cornfield approximation to calculate CI of the odds ratio
<u>woolf</u>	use Woolf approximation to calculate SE and CI of the odds ratio
graph	graph odds against categories
<u>cip</u>lot	same as graph option, except include confidence intervals

CI plot
<u>ciopts</u>(*rcap_options*)	affect rendition of the confidence bands

[Plot]
marker_options	change look of markers (color, size, etc.)
marker_label_options	add marker labels; change look or position
cline_options	affect rendition of the plotted points

Add plots
<u>addplot</u>(*plot*)	add other plots to the generated graph

Y axis, X axis, Titles, Legend, Overall
twoway_options	any options other than by() documented in [G-3] *twoway_options*

mhodds_options	Description

Options
by(*varlist* [, <u>miss</u>ing])	stratify on *varlist*
<u>binomial</u>(*varname*)	number of subjects variable
<u>compare</u>(v_1, v_2)	override categories of the control variable
<u>level</u>(#)	set confidence level; default is level(95)

fweights are allowed; see [U] **11.1.6 weight**.

Options

Options are listed in the order that they appear in the syntax tables above. The commands for which the option is valid are indicated in parentheses immediately after the option name.

Options (ir, cs, cc, and mhodds) / Main (tabodds)

by(*varname* [, <u>missing</u>]) (ir, cs, cc, and mhodds) specifies that the tables be stratified on *varname*. Missing categories in *varname* are omitted from the stratified analysis, unless option missing is specified within by(). Within-stratum statistics are shown and then combined with Mantel–Haenszel weights. If estandard, istandard, or standard() is also specified (see below), the weights specified are used in place of Mantel–Haenszel weights.

estandard, istandard, and standard(*varname*) (ir, cs, and cc) request that within-stratum statistics be combined with external, internal, or user-specified weights to produce a standardized estimate. These options are mutually exclusive and can be used only when by() is also specified. (When by() is specified without one of these options, Mantel–Haenszel weights are used.)

estandard external weights are the person-time for the unexposed (ir), the total number of unexposed (cs), or the number of unexposed controls (cc).

istandard internal weights are the person-time for the exposed (ir), the total number of exposed (cs), or the number of exposed controls (cc). istandard can be used to produce, among other things, standardized mortality ratios (SMRs).

standard(*varname*) allows user-specified weights. *varname* must contain a constant within stratum and be nonnegative. The scale of *varname* is irrelevant.

pool (ir, cs, and cc) specifies that, in a stratified analysis, the directly pooled estimate also be displayed. The pooled estimate is a weighted average of the stratum-specific estimates using inverse-variance weights, which are the inverse of the variance of the stratum-specific estimate. pool is relevant only if by() is also specified.

nocrude (ir, cs, and cc) specifies that in a stratified analysis the crude estimate—an estimate obtained without regard to strata—not be displayed. nocrude is relevant only if by() is also specified.

nohom (ir, cs, and cc) specifies that a χ^2 test of homogeneity not be included in the output of a stratified analysis. This tests whether the exposure effect is the same across strata and can be performed for any pooled estimate—directly pooled or Mantel–Haenszel. nohom is relevant only if by() is also specified.

ird (ir) may be used only with estandard, istandard, or standard(). It requests that ir calculate the standardized incidence-rate difference rather than the default incidence-rate ratio.

rd (cs) may be used only with estandard, istandard, or standard(). It requests that cs calculate the standardized risk difference rather than the default risk ratio.

bd (cc) specifies that Breslow and Day's χ^2 test of homogeneity be included in the output of a stratified analysis. This tests whether the exposure effect is the same across strata. bd is relevant only if by() is also specified.

tarone (cc) specifies that Tarone's χ^2 test of homogeneity, which is a correction to the Breslow–Day test, be included in the output of a stratified analysis. This tests whether the exposure effect is the same across strata. tarone is relevant only if by() is also specified.

binomial(*varname*) (cs, cc, tabodds, and mhodds) supplies the number of subjects (cases plus controls) for binomial frequency records. For individual and simple frequency records, this option is not used.

or (cs, csi, and tabodds), for cs and csi, reports the calculation of the odds ratio in addition to the risk ratio if by() is not specified. With by(), or specifies that a Mantel–Haenszel estimate of the combined odds ratio be made rather than the Mantel–Haenszel estimate of the risk ratio. In either case, this is the same calculation that would be made by cc and cci. Typically, cc, cci, or tabodds is preferred for calculating odds ratios. For tabodds, or specifies that odds ratios be produced; see base() for details about selecting a reference category. By default, tabodds will calculate odds.

adjust(*varlist*) (tabodds) specifies that odds ratios adjusted for the variables in *varlist* be calculated.

base(#) (tabodds) specifies that the #th category of *expvar* be used as the reference group for calculating odds ratios. If base() is not specified, the first category, corresponding to the minimum value of expvar, is used as the reference group.

cornfield (cc, cci, and tabodds) requests that the Cornfield (1956) approximation be used to calculate the confidence interval of the odds ratio. By default, cc and cci report an exact interval and tabodds reports a standard-error–based interval, with the standard error coming from the square root of the variance of the score statistic.

woolf (cs, csi, cc, cci, and tabodds) requests that the Woolf (1955) approximation, also known as the Taylor expansion, be used for calculating the standard error and confidence interval for the odds ratio. By default, cs and csi with the or option report the Cornfield (1956) interval; cc and cci report an exact interval; and tabodds reports a standard-error–based interval, with the standard error coming from the square root of the variance of the score statistic.

exact (cs, csi, cc, and cci) requests that Fisher's exact p be calculated rather than the χ^2 and its significance level. We recommend specifying exact whenever samples are small. When the least-frequent cell contains 1,000 cases or more, there will be no appreciable difference between the exact significance level and the significance level based on the χ^2, but the exact significance level will take considerably longer to calculate. exact does not affect whether exact confidence intervals are calculated. Commands always calculate exact confidence intervals where they can, unless cornfield or woolf is specified.

compare(v_1, v_2) (mhodds) indicates the categories of *expvar* to be compared; v_1 defines the numerator and v_2, the denominator. When compare() is not specified and there are only two categories, the second is compared to the first; when there are more than two categories, an approximate estimate of the odds ratio for a unit increase in *expvar*, controlled for specified confounding variables, is given.

level(#) (ir, iri, cs, csi, cc, cci, tabodds, mhodds, mcc, and mcci) specifies the confidence level, as a percentage, for confidence intervals. The default is level(95) or as set by set level; see [R] **level**.

The following options are for use only with tabodds.

⌐ Main ⌐

graph (tabodds) produces a graph of the odds against the numerical code used for the categories of *expvar*. All graph options except connect() are allowed. This option is not allowed with the or option or the adjust() option.

ciplot (tabodds) produces the same plot as the graph option, except that it also includes the confidence intervals. This option may not be used with either the or option or the adjust() option.

⌐ CI plot ⌐

ciopts(*rcap_options*) (tabodds) is allowed only with the ciplot option. It affects the rendition of the confidence bands; see [G-3] *rcap_options*.

⌐ Plot ⌐

marker_options (tabodds) affect the rendition of markers drawn at the plotted points, including their shape, size, color, and outline; see [G-3] *marker_options*.

marker_label_options (tabodds) specify if and how the markers are to be labeled; see [G-3] *marker_label_options*.

cline_options (tabodds) affect whether lines connect the plotted points and the rendition of those lines; see [G-3] *cline_options*.

⌐ Add plots ⌐

addplot(*plot*) (tabodds) provides a way to add other plots to the generated graph; see [G-3] *addplot_option*.

⌐ Y axis, X axis, Titles, Legend, Overall ⌐

twoway_options (tabodds) are any of the options documented in [G-3] *twoway_options*, excluding by(). These include options for titling the graph (see [G-3] *title_options*) and options for saving the graph to disk (see [G-3] *saving_option*).

Remarks and examples   stata.com

Remarks are presented under the following headings:

To calculate appropriate statistics and suppress inappropriate statistics, the `ir`, `cs`, `cc`, `tabodds`, `mhodds`, and `mcc` commands, along with their immediate counterparts, are organized in the way epidemiologists conceptualize data. `ir` processes incidence-rate data from prospective studies; `cs`, cohort study data with equal follow-up time (cumulative incidence); `cc`, `tabodds`, and `mhodds`, case–control or cross-sectional (prevalence) data; and `mcc`, matched case–control data. With the exception of `mcc`, these commands work with both simple and stratified tables.

Epidemiological data are often summarized in a contingency table from which various statistics are calculated. The rows of the table reflect cases and noncases or cases and person-time, and the columns reflect exposure to a *risk factor*. To an epidemiologist, *cases* and *noncases* refer to the outcomes of the process being studied. For instance, a case might be a person with cancer and a noncase might be a person without cancer.

A *factor* is something that might affect the chances of being ultimately designated a case or a noncase. Thus a case might be a cancer patient and the factor, smoking behavior. A person is said to be *exposed* or *unexposed* to the factor. Exposure can be classified as a dichotomy, smokes or does not smoke, or as multiple levels, such as number of cigarettes smoked per week.

For an introduction to epidemiological methods, see Walker (1991). For an intermediate treatment, see Clayton and Hills (1993) and Lilienfeld and Stolley (1994). For other advanced discussions, see Kleinbaum, Kupper, and Morgenstern (1982) and Rothman, Greenland, and Lash (2008). For an anthology of writings on epidemiology since World War II, see Greenland (1987). See Jewell (2004) for a text aimed at graduate students in the medical professions that uses Stata for much of the analysis. See Dohoo, Martin, and Stryhn (2010) for a graduate-level text on the principles and methods of veterinary epidemiologic research; Stata datasets and do-files are available. Also see Dohoo, Martin, and Stryhn (2012) for a text that is a revision of their veterinary epidemiology text, but examples from human epidemiology are used.

Incidence-rate data

In incidence-rate data from a prospective study, you observe the transformation of noncases into cases. Starting with a group of noncase subjects, you monitor them to determine whether they become cases (for example, stricken with cancer). You monitor two populations: those exposed and those unexposed to the factor (for example, multiple X-rays). A summary of the data is

	Exposed	Unexposed	Total
Cases	a	b	$a + b$
Person-time	N_1	N_0	$N_1 + N_0$

▷ Example 1: iri

It will be easiest to understand these commands if we start with the immediate forms. Remember, in the immediate form, we specify the data on the command line rather than specifying names of variables containing the data; see [U] **19 Immediate commands**. We have data (Boice and Monson [1977]; reported in Rothman, Greenland, and Lash [2008, 244]) on breast cancer *cases* and person-years of observation for women with tuberculosis repeatedly *exposed* to multiple X-ray fluoroscopies, and those not so exposed:

	X-ray fluoroscopy	
	Exposed	Unexposed
Breast cancer cases	41	15
Person-years	28,010	19,017

Using `iri`, the immediate form of `ir`, we specify the values in the table following the command:

```
. iri 41 15 28010 19017
                     │  Exposed    Unexposed  │     Total
          ───────────┼────────────────────────┼──────────────
               Cases │      41           15   │        56
         Person-time │   28010        19017   │     47027
          ───────────┼────────────────────────┼──────────────

       Incidence rate│  .0014638     .0007888 │    .0011908

                     │   Point estimate       │  [95% Conf. Interval]
          ───────────┼────────────────────────┼──────────────────────
      Inc. rate diff.│        .000675         │  .0000749     .0012751
      Inc. rate ratio│       1.855759         │  1.005684     3.6093  (exact)
      Attr. frac. ex.│        .4611368        │  .0056519     .722938 (exact)
      Attr. frac. pop│        .337618         │
          ───────────┴────────────────────────┴──────────────────────
                         (midp)   Pr(k>=41) =                 0.0177  (exact)
                         (midp) 2*Pr(k>=41) =                 0.0355  (exact)
```

`iri` shows the table, reports the incidence rates for the exposed and unexposed populations, and then shows the point estimates of the difference and ratio of the two incidence rates along with their confidence intervals. The incidence rate is simply the frequency with which noncases are transformed into cases.

Next `iri` reports the attributable fraction among the exposed population, an estimate of the proportion of exposed cases attributable to exposure. We estimate that 46.1% of the 41 breast cancer cases among the exposed were due to exposure. (Had the incidence-rate ratio been less than 1, `iri` would have reported the prevented fraction in the exposed population, an estimate of the net proportion of all potential cases in the exposed population that was prevented by exposure; see the following technical note.)

After that, the table shows the attributable fraction in the total population, which is the net proportion of all cases attributable to exposure. This number, of course, depends on the proportion of cases that are exposed in the base population, which here is taken to be $41/56$ and may not be relevant in all situations. We estimate that 33.8% of the 56 cases were due to exposure. We estimate that 18.9 cases were caused by exposure; that is, $0.338 \times 56 = 0.461 \times 41 = 18.9$.

At the bottom of the table, `iri` reports both one- and two-sided exact significance tests. For the one-sided test, the probability that the number of exposed cases is 41 or greater is 0.0177. This is a "midp" calculation; see *Methods and formulas* below. The two-sided test is $2 \times 0.0177 = 0.0354$.

◁

❑ Technical note

When the incidence-rate ratio is less than 1, `iri` (and `ir`, `cs`, `csi`, `cc`, and `cci`) substitutes the prevented fraction for the attributable fraction. Let's reverse the roles of exposure in the above data, treating as exposed a person who did not receive the X-ray fluoroscopy. You can think of this as a new treatment for preventing breast cancer—the suggested treatment being not to use fluoroscopy.

```
. iri 15 41 19017 28010
```

	Exposed	Unexposed	Total
Cases	15	41	56
Person-time	19017	28010	47027
Incidence rate	.0007888	.0014638	.0011908

	Point estimate	[95% Conf. Interval]	
Inc. rate diff.	-.000675	-.0012751	-.0000749
Inc. rate ratio	.5388632	.277062	.9943481 (exact)
Prev. frac. ex.	.4611368	.0056519	.722938 (exact)
Prev. frac. pop	.1864767		

(midp) Pr(k<=15) =		0.0177 (exact)
(midp) 2*Pr(k<=15) =		0.0355 (exact)

The prevented fraction among the exposed is the net proportion of all potential cases in the exposed population that were prevented by exposure. We estimate that 46.1% of potential cases among the women receiving the new "treatment" were prevented by the treatment. (Previously, we estimated that the same percentage of actual cases among women receiving the X-rays was caused by the X-rays.)

The prevented fraction for the population, which is the net proportion of all potential cases in the total population that was prevented by exposure, as with the attributable fraction, depends on the proportion of cases that are exposed in the base population—here taken as $15/56$—so it may not be relevant in all situations. We estimate that 18.6% of the potential cases were prevented by exposure.

See Greenland and Robins (1988) for a discussion of how to interpret attributable and prevented fractions.

❑

▷ Example 2: ir

ir works like iri, except that it obtains the entries in the tables by summing data. You specify three variables: the first represents the number of cases represented by this observation, the second indicates whether the observation is for subjects exposed to the factor, and the third records the total time the subjects in this observation were observed. An observation may reflect one subject or a group of subjects.

For instance, here is a 2-observation dataset for the table in the previous example:

```
. use http://www.stata-press.com/data/r14/irxmpl
. list
```

	cases	exposed	time
1.	41	0	28010
2.	15	1	19017

If we typed ir cases exposed time, we would obtain the same output that we obtained above. Another way the data might be recorded is

```
. use http://www.stata-press.com/data/r14/irxmpl2

. list
```

	cases	exposed	time
1.	20	0	14000
2.	21	0	14010
3.	15	1	19017

Here the first 2 observations will be automatically summed by `ir` because both are exposed. Finally, the data might be individual-level data:

```
. use http://www.stata-press.com/data/r14/irxmpl3

. list in 1/5
```

	cases	exposed	time
1.	1	1	10
2.	0	1	8
3.	0	0	9
4.	1	0	2
5.	0	1	1

The first observation represents a woman who got cancer, was exposed, and was observed for 10 years. The second is a woman who did not get cancer, was exposed, and was observed for 8 years, and so on.

◁

❑ Technical note

`ir` (and all the other commands) assumes that a subject was exposed if the exposed variable is nonzero and not missing, assumes the subject was not exposed if the variable is zero, and ignores the observation if the variable is missing. For `ir`, the case variable and the time variable are restricted to nonnegative integers and are summed within the exposed and unexposed groups to obtain the entries in the table.

❑

Stratified incidence-rate data

▷ Example 3: ir with stratified data

`ir` can work with stratified tables, as well as with single tables. For instance, Rothman (1986, 185) discusses data from Rothman and Monson (1973) on mortality by sex and age for patients with trigeminal neuralgia:

| | Age through 64 | | Age 65+ | |
	Males	Females	Males	Females
Deaths	14	10	76	121
Person-years	1516	1701	949	2245

Entering the data into Stata, we have the following dataset:

```
. use http://www.stata-press.com/data/r14/rm
(Rothman and Monson 1973 data)

. list
```

	age	male	deaths	pyears
1.	<65	1	14	1516
2.	<65	0	10	1701
3.	65+	1	76	949
4.	65+	0	121	2245

The stratified analysis of the incidence-rate ratio is

```
. ir deaths male pyears, by(age)
```

Age category	IRR	[95% Conf. Interval]		M-H Weight
<65	1.570844	.6489373	3.952809	4.712465 (exact)
65+	1.485862	1.100305	1.99584	35.95147 (exact)
Crude	1.099794	.831437	1.449306	(exact)
M-H combined	1.49571	1.141183	1.960377	

```
 Test of homogeneity (M-H)    chi2(1) =      0.02  Pr>chi2 = 0.8992
```

The row labeled `M-H combined` reflects the combined Mantel–Haenszel estimates.

As with the previous example, it is not important that each entry in the table correspond to 1 observation in the data — `ir` sums the time (`pyears`) and case (`deaths`) variables within the exposure (`male`) category.

The difference between the unadjusted crude estimate and the Mantel–Haenszel estimate suggests confounding by age: women in the study are older, and older patients are more likely to die. But we should not use the Mantel–Haenszel estimate without checking its homogeneity assumption. The chi-squared test of homogeneity gives a p-value of 0.8992, so we have no evidence that the exposure effect (the effect of being male) differs across age categories. We are justified in using the Mantel–Haenszel estimate.

◁

❑ Technical note

Stratification is one way to deal with confounding; that is, perhaps sex affects the incidence of trigeminal neuralgia and so does age, so the table was stratified by age in an attempt to uncover the sex effect. (We are concerned that age may confound the true association between sex and the incidence of trigeminal neuralgia because the age distributions are so different for males and females. If age affects incidence, the difference in the age distributions would induce different incidences for males and females and thus confound the true effect of sex.)

We do not, however, have to use tables to uncover effects; the estimation alternative when we have aggregate data is Poisson regression, and we can use the same data on which we ran `ir` with `poisson`. Poisson regression also works with individual-level data.

(Although `age` in the previous example appears to be a string, it is actually a numeric variable taking on values 1 and 2. We attached a value label to produce the labels `<65` and `65+` to make `ir`'s output look better; see [U] **12.6.3 Value labels**. Stata's estimation commands will ignore this labeling.)

```
. poisson deaths male age, exposure(pyears) irr

Iteration 0:   log likelihood = -10.836732
Iteration 1:   log likelihood = -10.734087
Iteration 2:   log likelihood = -10.733944
Iteration 3:   log likelihood = -10.733944

Poisson regression                              Number of obs   =           4
                                                LR chi2(2)      =      164.01
                                                Prob > chi2     =      0.0000
Log likelihood = -10.733944                     Pseudo R2       =      0.8843

------------------------------------------------------------------------------
      deaths |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   1.495096   .2060997     2.92   0.004     1.141118    1.95888
         age |   8.888775   1.934943    10.04   0.000     5.801616   13.61867
       _cons |   .0006805   .0002908   -17.07   0.000     .0002945   .0015724
   ln(pyears)|          1  (exposure)
------------------------------------------------------------------------------
```

Compare these results with the Mantel–Haenszel estimates produced by `ir`:

Source	IR Ratio	95% Conf. Int.	
Mantel–Haenszel (`ir`)	1.50	1.14	1.96
`poisson`	1.50	1.14	1.96

The results from `poisson` agree with the Mantel–Haenszel estimates to two decimal places. But `poisson` also estimates an incidence-rate ratio for age. Here the estimate is not of much interest, because the outcome variable is total mortality and we already knew that older people have a higher mortality rate. In other contexts, however, the estimate might be of greater interest.

See [R] **poisson** for an explanation of the `poisson` command.

❑

❑ Technical note

Both the model fit above and the preceding table asserted that exposure effects are the same across age categories and, if they are not, then both of the previous results are equally inappropriate. The table presented a test of homogeneity, reassuring us that the exposure effects do indeed appear to be constant. The Poisson-regression alternative can be used to reproduce that test by including interactions between the age groups and exposure:

```
. poisson deaths male age male#c.age, exposure(pyears) irr

Iteration 0:   log likelihood = -10.898799
Iteration 1:   log likelihood = -10.726225
Iteration 2:   log likelihood = -10.725904
Iteration 3:   log likelihood = -10.725904

Poisson regression                              Number of obs    =           4
                                                LR chi2(3)       =      164.03
                                                Prob > chi2      =      0.0000
Log likelihood = -10.725904                     Pseudo R2        =      0.8843
```

deaths	IRR	Std. Err.	z	P>|z|	[95% Conf. Interval]	
male	1.660688	1.396496	0.60	0.546	.3195218	8.631283
age	9.167973	3.01659	6.73	0.000	4.810583	17.47226
male#c.age						
1	.9459	.41539	-0.13	0.899	.3999832	2.236911
_cons	.0006412	.0004097	-11.51	0.000	.0001833	.0022434
ln(pyears)	1	(exposure)				

The significance level of the male#c.age effect is 0.899, the same as previously reported by ir.

Here forming the male-times-age interaction was easy because there were only two age groups. Had there been more groups, the test would have been slightly more difficult—see the following technical note.

❑

❑ Technical note

A word of caution is in order when applying poisson (or any estimation technique) to more than two age categories. Say that in our data, we had three age categories, which we will call categories 0, 1, and 2, and that they are stored in the variable agecat. We might think of the categories as corresponding to age less than 35, 35–64, and 65 and above.

With such data, we might type ir deaths male pyears, by(agecat), but we would *not* type poisson deaths male agecat, exposure(pyears) to obtain the equivalent Poisson-regression estimated results. Such a model might be reasonable, but it is not equivalent because we would be constraining the age effect in category 2 to be (multiplicatively) twice the effect in category 1.

To poisson (and all of Stata's estimation commands other than anova), agecat is simply one variable, and only one estimated coefficient is associated with it. Thus the model is

$$\text{Poisson index} = P = \beta_0 + \beta_1 \texttt{male} + \beta_2 \texttt{agecat}$$

The expected number of deaths is then e^P, and the incidence-rate ratio associated with a variable is e^β; see [R] **poisson**. Thus the value of the Poisson index when male==0 and agecat==1 is $\beta_0 + \beta_2$, and the possibilities are

	male==0	male==1
agecat==0	β_0	$\beta_0 + \beta_1$
agecat==1	$\beta_0 + \beta_2$	$\beta_0 + \beta_2 + \beta_1$
agecat==2	$\beta_0 + 2\beta_2$	$\beta_0 + 2\beta_2 + \beta_1$

The age effect for `agecat==2` is constrained to be twice the age effect for `agecat==1`—the only difference between lines 3 and 2 of the table is that β_2 is replaced with $2\beta_2$. Under certain circumstances, such a constraint might be reasonable, but it does not correspond to the assumptions made in generating the Mantel–Haenszel combined results.

To obtain results equivalent to the Mantel–Haenszel result, we must estimate a separate effect for each age group, meaning that we must replace $2\beta_2$, the constrained effect, with β_3, a new coefficient that is free to take on any value. We can achieve this by creating two new variables and using them in place of `agecat`. `agecat1` will take on the value 1 when `agecat` is 1 and 0 otherwise; `agecat2` will take on the value 1 when `agecat` is 2 and 0 otherwise:

```
. generate agecat1 = (agecat==1)
. generate agecat2 = (agecat==2)
. poisson deaths male agecat1 agecat2 [freq=pop], exposure(pyears) irr
```

In Stata, we do not have to generate these variables for ourselves. We could use factor variables:

```
. poisson deaths male i.agecat [freq=pop], exposure(pyears) irr
```

See [U] **11.4.3 Factor variables**.

To reproduce the homogeneity test with multiple age categories, we could type

```
. poisson deaths agecat##male [freq=pop], exp(pyears) irr
. testparm agecat#male
```

Poisson regression combined with factor variables generalizes to multiway tables. Suppose that there are three exposure categories. Assume exposure variable `burn` takes on the values 1, 2, and 3 for first-, second-, and third-degree burns. The table itself is estimated by typing

```
. poisson deaths i.burn i.agecat [freq=pop], exp(pyears) irr
```

and the test of homogeneity is estimated by typing

```
. poisson deaths burn##agecat [freq=pop], exp(pyears) irr
. testparm burn#agecat
```

❑

Standardized estimates with stratified incidence-rate data

The `by()` option specifies that the data are stratified and, by default, will produce a Mantel–Haenszel combined estimate of the incidence-rate ratio. With the `estandard`, `istandard`, or `standard(varname)` options, you can specify your own weights and obtain standardized estimates of the incidence-rate ratio or difference.

▷ Example 4: ir with stratified data, using standardized estimates

Rothman, Greenland, and Lash (2008, 264) report results from Doll and Hill (1966) on age-specific coronary disease deaths among British male doctors from cigarette smoking:

| Age | Smokers | | Nonsmokers | |
	Deaths	Person-years	Deaths	Person-years
35–44	32	52,407	2	18,790
45–54	104	43,248	12	10,673
55–64	206	28,612	28	5,710
65–74	186	12,663	28	2,585
75–84	102	5,317	31	1,462

We have entered these data into Stata:

```
. use http://www.stata-press.com/data/r14/dollhill3
```
```
. list
```

	agecat	smokes	deaths	pyears
1.	35–44	1	32	52,407
2.	45–54	1	104	43,248
3.	55–64	1	206	28,612
4.	65–74	1	186	12,663
5.	75–84	1	102	5,317
6.	35–44	0	2	18,790
7.	45–54	0	12	10,673
8.	55–64	0	28	5,710
9.	65–74	0	28	2,585
10.	75–84	0	31	1,462

We can obtain the Mantel–Haenszel combined estimate along with the crude estimate for ignoring stratification of the incidence-rate ratio and 90% confidence intervals by typing

```
. ir deaths smokes pyears, by(age) level(90)
```

age category	IRR	[90% Conf. Interval]		M-H Weight	
35–44	5.736638	1.704271	33.61646	1.472169	(exact)
45–54	2.138812	1.274552	3.813282	9.624747	(exact)
55–64	1.46824	1.044915	2.110422	23.34176	(exact)
65–74	1.35606	.9626026	1.953505	23.25315	(exact)
75–84	.9047304	.6375194	1.305412	24.31435	(exact)
Crude	1.719823	1.437544	2.0688		(exact)
M-H combined	1.424682	1.194375	1.699399		

```
 Test of homogeneity (M-H)    chi2(4) =     10.41  Pr>chi2 = 0.0340
```

Note the presence of heterogeneity revealed by the test; the effect of smoking is not the same across age categories. Moreover, the listed stratum-specific estimates show an effect that appears to be declining with age. (Even if the test of homogeneity is not significant, you should always examine estimates carefully when stratum-specific effects occur on both sides of 1 for ratios and 0 for differences.)

Rothman, Greenland, and Lash (2008, 269) obtain the standardized incidence-rate ratio and 90% confidence intervals, weighting each age category by the population of the exposed group, thus producing the standardized mortality ratio (SMR). This calculation can be reproduced by specifying by(age) to indicate that the table is stratified and istandard to specify that we want the internally standardized rate. We may also specify that we would like to see the pooled estimate (weighted average where the weights are based on the variance of the strata calculations):

```
. ir deaths smokes pyears, by(age) level(90) istandard pool
```

age category	IRR	[90% Conf. Interval]		Weight
35-44	5.736638	1.704271	33.61646	52407 (exact)
45-54	2.138812	1.274552	3.813282	43248 (exact)
55-64	1.46824	1.044915	2.110422	28612 (exact)
65-74	1.35606	.9626026	1.953505	12663 (exact)
75-84	.9047304	.6375194	1.305412	5317 (exact)
Crude	1.719823	1.437544	2.0688	(exact)
Pooled (direct)	1.355343	1.134356	1.619382	
I. Standardized	1.417609	1.186541	1.693676	

```
Test of homogeneity (direct) chi2(4) =     10.20  Pr>chi2 = 0.0372
```

We obtained the simple pooled results because we specified the pool option. Note the significance of the homogeneity test; it provides the motivation for standardizing the rate ratios.

If we wanted the externally standardized ratio (weights proportional to the population of the unexposed group), we would substitute estandard for istandard in the above command.

We are not limited to incidence-rate ratios; ir can also estimate incidence-rate differences. Differences may be standardized internally or externally. We will obtain the internally weighted difference (Rothman, Greenland, and Lash 2008, 266–267):

```
. ir deaths smokes pyears, by(age) level(90) istandard ird
```

age category	IRD	[90% Conf. Interval]		Weight
35-44	.0005042	.0002877	.0007206	52407
45-54	.0012804	.0006205	.0019403	43248
55-64	.0022961	.0005628	.0040294	28612
65-74	.0038567	.0000521	.0076614	12663
75-84	-.0020201	-.0090201	.00498	5317
Crude	.0018537	.001342	.0023654	
I. Standardized	.0013047	.000712	.0018974	

◁

▷ Example 5: ir with user-specified weights

In addition to calculating results by using internal or external weights, ir (and cs and cc) can calculate results for arbitrary weights. If we wanted to obtain the incidence-rate ratio weighting each age category equally, we would type

```
. generate conswgt=1
. ir deaths smokes pyears, by(age) standard(conswgt)
```

age category	IRR	[95% Conf. Interval]		Weight
35-44	5.736638	1.463557	49.40468	1 (exact)
45-54	2.138812	1.173714	4.272545	1 (exact)
55-64	1.46824	.9863624	2.264107	1 (exact)
65-74	1.35606	.9081925	2.096412	1 (exact)
75-84	.9047304	.6000757	1.399687	1 (exact)
Crude	1.719823	1.391992	2.14353	(exact)
Standardized	1.155026	.9006199	1.481295	

◁

❑ Technical note

estandard and istandard are convenience features; they do nothing different from what you could accomplish by creating the appropriate weights and using the standard() option. For instance, we could duplicate the previously shown results of istandard (example before last) by typing

```
. sort age smokes
. by age: generate wgt=pyears[_N]
. list in 1/4
```

	agecat	smokes	deaths	pyears	conswgt	wgt
1.	35–44	0	2	18,790	1	52407
2.	35–44	1	32	52,407	1	52407
3.	45–54	0	12	10,673	1	43248
4.	45–54	1	104	43,248	1	43248

```
. ir deaths smokes pyears, by(age) level(90) standard(wgt) ird
  (output omitted )
```

sort age smokes made the exposed group (smokes = 1) the last observation within each age category. by age: gen wgt=pyears[_N] created wgt equal to the last observation in each age category.

❑

Cumulative incidence data

Cumulative incidence data are "follow-up data with denominators consisting of persons rather than person-time" (Rothman 1986, 172). A group of noncases is monitored for some time, during which some become cases. Each subject is also known to be exposed or unexposed. A summary of the data is

	Exposed	Unexposed	Total
Cases	a	b	$a + b$
Noncases	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

Data of this type are generally summarized using the risk ratio, $\{a/(a+c)\}/\{b/(b+d)\}$. A ratio of 2 means that an exposed subject is twice as likely to become a case than is an unexposed subject, a ratio of one-half means half as likely, and so on. The "null" value—the number corresponding to no effect—is a ratio of 1. If cross-sectional data are analyzed in this format, the risk ratio becomes a prevalence ratio.

▷ Example 6: csi

We have data on diarrhea during a 10-day follow-up period among 30 breast-fed infants colonized with *Vibrio cholerae* 01 according to antilipopolysaccharide antibody titers in the mother's breast milk (Glass et al. [1983]; reported in Rothman, Greenland, and Lash [2008, 248]):

	Antibody level	
	High	Low
Diarrhea	7	12
No diarrhea	9	2

The `csi` command works much like the `iri` command. Our sample is small, so we will specify the `exact` option.

```
. csi 7 12 9 2, exact
```

	Exposed	Unexposed	Total
Cases	7	12	19
Noncases	9	2	11
Total	16	14	30
Risk	.4375	.8571429	.6333333

	Point estimate	[95% Conf. Interval]	
Risk difference	-.4196429	-.7240828	-.1152029
Risk ratio	.5104167	.2814332	.9257086
Prev. frac. ex.	.4895833	.0742914	.7185668
Prev. frac. pop	.2611111		

```
                          1-sided Fisher's exact P = 0.0212
                          2-sided Fisher's exact P = 0.0259
```

We find that high antibody levels reduce the risk of diarrhea (the risk falls from 0.86 to 0.44). The difference is just significant at the 2.59% two-sided level. (Had we not specified the `exact` option, a χ^2 value and its significance level would have been reported in place of Fisher's exact p. The calculated χ^2 two-sided significance level would have been 0.0173, but this calculation is inferior for small samples.)

◁

❑ Technical note

By default, `cs` and `csi` do not report the odds ratio, but they will if you specify the `or` option. If you want odds ratios, however, use the `cc` or `cci` commands—the commands appropriate for case–control data—because `cs` and `csi` calculate the attributable (prevented) fraction with the risk ratio, even if you specify `or`:

```
. csi 7 12 9 2, or exact
```

	Exposed	Unexposed	Total
Cases	7	12	19
Noncases	9	2	11
Total	16	14	30
Risk	.4375	.8571429	.6333333

	Point estimate	[95% Conf. Interval]	
Risk difference	-.4196429	-.7240828	-.1152029
Risk ratio	.5104167	.2814332	.9257086
Prev. frac. ex.	.4895833	.0742914	.7185668
Prev. frac. pop	.2611111		
Odds ratio	.1296296	.0246233	.7180882 (Cornfield)

```
                          1-sided Fisher's exact P = 0.0212
                          2-sided Fisher's exact P = 0.0259
```

❑

❑ Technical note

As with `iri` and `ir`, `csi` and `cs` report either the attributable or the prevented fraction for the exposed and total populations; see the discussion under *Incidence-rate data* above. In example 6, we estimated that 49% of potential cases in the exposed population were prevented by exposure. We also estimated that exposure accounted for a 26% reduction in cases over the entire population, but that is based on the exposure distribution of the (small) population ($16/30$) and probably is of little interest.

Fleiss, Levin, and Paik (2003, 128) report infant mortality by birthweight for 72,730 live white births in 1974 in New York City:

```
. csi 618 422 4597 67093
```

	Exposed	Unexposed	Total
Cases	618	422	1040
Noncases	4597	67093	71690
Total	5215	67515	72730
Risk	.1185043	.0062505	.0142995

	Point estimate	[95% Conf. Interval]	
Risk difference	.1122539	.1034617	.121046
Risk ratio	18.95929	16.80661	21.38769
Attr. frac. ex.	.9472554	.9404996	.9532441
Attr. frac. pop	.5628883		

```
                        chi2(1) =  4327.92  Pr>chi2 = 0.0000
```

In these data, exposed means a premature baby (birthweight $\leq 2{,}500$ g), and a case is a baby who is dead at the end of one year. We find that being premature accounts for 94.7% of deaths among the premature population. We also estimate, paraphrasing from Fleiss, Levin, and Paik (2003, 128), that 56.3% of all white infant deaths in New York City in 1974 could have been prevented if prematurity had been eliminated. (Moreover, Fleiss, Levin, and Paik put a standard error on the attributable fraction for the population. The formula is given in *Methods and formulas* but is appropriate only for the population on which the estimates are based because other populations may have different probabilities of exposure.)

❑

▷ Example 7: cs

`cs` works like `csi`, except that it obtains its information from the data. The data equivalent to typing `csi 7 12 9 2` are

```
. use http://www.stata-press.com/data/r14/csxmpl, clear
. list
```

	case	exp	pop
1.	1	1	7
2.	1	0	12
3.	0	1	9
4.	0	0	2

We could then type `cs case exp [freq=pop]`. If we had individual-level data, so that each observation reflected a patient and we had 30 observations, we would type `cs case exp`.

◁

Stratified cumulative incidence data

▷ Example 8: cs with stratified data

Rothman, Greenland, and Lash (2008, 260) reprint the following age-specific information for deaths from all causes for tolbutamide and placebo treatment groups (University Group Diabetes Program 1970):

	Age through 54		Age 55 and above	
	Tolbutamide	Placebo	Tolbutamide	Placebo
Dead	8	5	22	16
Surviving	98	115	76	69

The data corresponding to these results are

```
. use http://www.stata-press.com/data/r14/ugdp

. list
```

	age	case	exposed	pop
1.	<55	0	0	115
2.	<55	0	1	98
3.	<55	1	0	5
4.	<55	1	1	8
5.	55+	0	0	69
6.	55+	0	1	76
7.	55+	1	0	16
8.	55+	1	1	22

The order of the observations is unimportant. If we were now to type cs case exposed [freq=pop], we would obtain a summary for all the data, ignoring the stratification by age. To incorporate the stratification, we type

```
. cs case exposed [freq=pop], by(age)
```

Age category	RR	[95% Conf. Interval]		M-H Weight
<55	1.811321	.6112044	5.367898	2.345133
55+	1.192602	.6712664	2.11883	8.568306
Crude	1.435574	.8510221	2.421645	
M-H combined	1.325555	.797907	2.202132	

Test of homogeneity (M-H)   chi2(1) =   0.447   Pr>chi2 = 0.5037

Mantel–Haenszel weights are appropriate when the risks may differ according to the strata but the risk ratio is believed to be the same (homogeneous across strata). Under these assumptions, Mantel–Haenszel weights are designed to use the information efficiently. They are not intended to measure a composite risk ratio when the within-stratum risk ratios differ. Then we want a standardized ratio (see below).

The risk ratios above appear to differ markedly, but the confidence intervals are also broad because of the small sample sizes. The test of homogeneity shows that the differences can be attributed to chance; the use of the Mantel–Haenszel combined test is sensible.

◁

❑ Technical note

Stratified cumulative incidence tables are not the only way to control for confounding. Another way is logistic regression. However, logistic regression measures effects with odds ratios, not with risk ratios. So before we fit a logistic model, let's use cs to estimate the Mantel–Haenszel odds ratio:

```
. cs case exposed [freq=pop], by(age) or
     Age category |        OR        [95% Conf. Interval]    M-H Weight
------------------+--------------------------------------------------------------
              <55 |    1.877551     .6238165    5.637046      2.168142 (Cornfield)
              55+ |    1.248355     .6112772    2.547411      6.644809 (Cornfield)
------------------+--------------------------------------------------------------
            Crude |    1.510673     .8381198    2.722012
      M-H combined |    1.403149     .7625152    2.582015
------------------+--------------------------------------------------------------
Test of homogeneity (M-H)       chi2(1) =     0.347  Pr>chi2 = 0.5556

                      Test that combined OR = 1:
                                   Mantel-Haenszel chi2(1) =       1.19
                                              Pr>chi2 =     0.2750
```

The Mantel–Haenszel odds ratio is 1.40. It measures the association between death and treatment while adjusting for age. A more general way to adjust for age is logistic regression; the outcome variable is case, and it is explained by age and exposed. (As in the incidence-rate example, age may appear to be a string variable in our data—we listed the data in the previous example—but it is actually a numeric variable taking on values 0 and 1 with value labels disguising that fact; see [U] **12.6.3 Value labels**.)

```
. logistic case exposed age [freq=pop]
Logistic regression                             Number of obs    =       409
                                                LR chi2(2)       =     22.47
                                                Prob > chi2      =    0.0000
Log likelihood =  -142.6212                     Pseudo R2        =    0.0730

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     exposed |   1.404674   .4374454     1.09   0.275     .7629451    2.586175
         age |   4.216299   1.431519     4.24   0.000     2.167361    8.202223
       _cons |   .0513818   .0170762    -8.93   0.000     .0267868    .0985593
------------------------------------------------------------------------------
```

Compare these results with the Mantel–Haenszel estimates obtained with cs:

Source	Odds Ratio	95% Conf. Int.	
Mantel–Haenszel (cs)	1.40	0.76	2.58
logistic	1.40	0.76	2.59

They are virtually identical.

Logistic regression has advantages over the stratified-table approach. First, we obtained an estimate of the age effect: being 55 years or over significantly increases the odds of death. In addition to the point estimate, 4.22, we have a confidence interval for the effect: 2.17 to 8.20.

A discrete effect at age 55 is not a plausible model of aging. It would be more reasonable to assume that a 54-year-old patient has a higher probability of death, due merely to age, than does a 53-year-old patient; a 53-year-old, a higher probability than a 52-year-old patient; and so on. If we had the underlying data, where each patient's age is presumably known, we could include the actual age in the model and so better control for the age effect. This would improve our estimate of the effect of being exposed to tolbutamide.

See [R] **logistic** for an explanation of the logistic command. Also see the technical note in *Stratified incidence-rate data* concerning categorical variables, which applies to logistic regression as well as Poisson regression.

❑

Standardized estimates with stratified cumulative incidence data

As with ir, cs can produce standardized estimates, and the method is basically the same, although the options for which estimates are to be combined or standardized make it confusing. We showed above that cs can produce Mantel–Haenszel weighted estimates of the risk ratio (the default) or the odds ratio (obtained by specifying or). cs can also produce standardized estimates of the risk ratio (the default) or the risk difference (obtained by specifying rd).

▷ Example 9: cs with stratified data, using standardized estimates

To produce an estimate of the internally standardized risk ratio by using our age-specific data on deaths from all causes for tolbutamide and placebo treatment groups (example above), we type

```
. cs case exposed [freq=pop], by(age) istandard
```

Age category	RR	[95% Conf. Interval]		Weight
<55	1.811321	.6112044	5.367898	106
55+	1.192602	.6712664	2.11883	98
Crude	1.435574	.8510221	2.421645	
I. Standardized	1.312122	.7889772	2.182147	

We could obtain externally standardized estimates by substituting estandard for istandard.

To produce an estimate of the risk ratio weighting each age category equally, we could type

```
. generate wgt=1
. cs case exposed [freq=pop], by(age) standard(wgt)
```

Age category	RR	[95% Conf. Interval]		Weight
<55	1.811321	.6112044	5.367898	1
55+	1.192602	.6712664	2.11883	1
Crude	1.435574	.8510221	2.421645	
Standardized	1.304737	.7844994	2.169967	

If we instead wanted the risk difference, we would type

```
. cs case exposed [freq=pop], by(age) standard(wgt) rd
```

Age category	RD	[95% Conf. Interval]		Weight
<55	.033805	−.0278954	.0955055	1
55+	.0362545	−.0809204	.1534294	1
Crude	.0446198	−.0192936	.1085332	
Standardized	.0350298	−.0311837	.1012432	

If we wanted to weight the less-than-55 age group five times as heavily as the 55-and-over group, we would create wgt to contain 5 for the first age group and 1 for the second (or 10 for the first group and 2 for the second—the scale of the weights does not matter).

◁

Case–control data

In case–control data, you select a sample on the basis of the outcome under study; that is, cases and noncases are sampled at different rates. If you were examining the link between coffee consumption and heart attacks, for instance, you could select a sample of subjects with and without the heart problem and then examine their coffee-drinking behavior. A subject who has suffered a heart attack is called a *case* just as with cohort study data. A subject who has never suffered a heart attack, however, is called a *control* rather than merely a noncase, emphasizing that the sampling was performed with respect to the outcome.

In case–control data, all hope of identifying the risk (that is, incidence) of the outcome (heart attacks) associated with the factor (coffee drinking) vanishes, at least without information on the underlying sampling fractions, but you can examine the proportion of coffee drinkers among the two populations and reason that, if there is a difference, coffee drinking may be associated with the risk of heart attacks. Remarkably, even without the underlying sampling fractions, you can also measure the ratio of the odds of heart attacks if a subject drinks coffee to the odds if a subject does not—the so-called odds ratio.

What is lost is the ability to compare absolute rates, which is not always the same as comparing relative rates; see Fleiss, Levin, and Paik (2003, 123).

▷ Example 10: cci

cci calculates the odds ratio and the attributable risk associated with a 2×2 table. Rothman et al. (1979; reprinted in Rothman [1986, 161], and Rothman, Greenland, and Lash [2008, 251]) present case–control data on the history of chlordiazopoxide use in early pregnancy for mothers of children born with and without congenital heart defects:

| | Chlordiazopoxide use | |
	Yes	No
Case mothers	4	386
Control mothers	4	1250

```
. cci 4 386 4 1250, level(90)
                                                         Proportion
                  Exposed    Unexposed  |      Total      Exposed
-----------------------------------------+-------------------------
      Cases           4          386    |        390       0.0103
   Controls           4         1250    |       1254       0.0032
-----------------------------------------+-------------------------
      Total           8         1636    |       1644       0.0049

                  Point estimate        |   [90% Conf. Interval]
-----------------------------------------+-------------------------
 Odds ratio          3.238342           |    .7698467    13.59664 (exact)
Attr. frac. ex.         .6912           |   -.2989599    .9264524 (exact)
Attr. frac. pop      .0070892           |
-----------------------------------------+-------------------------
                     chi2(1) =       3.07  Pr>chi2 = 0.0799
```

We obtain a point estimate of the odds ratio as 3.24 and a χ^2 value, which is a test that the odds ratio is 1, significant at the 10% level.

◁

❑ Technical note

The epitab commands can calculate three different confidence intervals for the odds ratio: the exact, Woolf, and Cornfield intervals. The exact interval, illustrated in example 10, is the default. The interval is "exact" because it uses an exact sampling distribution—a distribution with no unknown parameters under the null hypothesis. An exact interval does not use a normal or chi-squared approximation. "Exact" does not describe the coverage probability; the coverage probability of a 90% exact interval is not exactly 90%. The coverage probability is actually bounded below by 90% (Agresti 2013, 606), so a 90% exact interval will always cover the odds ratio with probability at least 90% (if the model is correct).

The Woolf and Cornfield intervals, on the other hand, are approximate. They approximate the exact sampling distribution with a normal model and are not guaranteed to maintain their nominal coverage: the coverage probability of a 90% approximate interval fluctuates above and below 90%. The coverage approaches 90% only in the limit as the sample size increases. Exact intervals are conservative; approximate intervals can be conservative or anticonservative (Agresti 2013, 607).

If you wish to maintain nominal coverage, then you should use the exact interval. But you will pay a price for the coverage: the exact interval will usually be wider than the approximate intervals. Example 10 is no exception:

Method	90% Conf. Int.		Command
exact	0.77	13.60	cci
Woolf	1.01	10.40	cci, woolf
Cornfield	1.07	9.83	cci, cornfield

The exact interval is the widest of the three—so wide that it includes the null value of one—even though the chi-squared p-value of 0.0799 was significant at the 10% level. The exact interval and chi-squared test come from different models, so we should not expect them to always agree on sharp conclusions such as statistical significance.

The odds-ratio intervals are all frequentist methods, so we cannot compare them rigorously with one example. See Brown (1981), Gart and Thomas (1982), and Agresti (1999) for more rigorous comparisons. Agresti (1999) found that the Woolf interval performed well, even for small samples.

❑

Jerome Cornfield (1912–1979) was born in New York City. He majored in history at New York University and took courses in statistics at the U.S. Department of Agriculture Graduate School but otherwise had little formal training. Cornfield held positions at the Bureau of Labor Statistics, the National Cancer Institute, the National Institutes of Health, Johns Hopkins University, the University of Pittsburgh, and George Washington University. He worked on many problems in biomedical statistics, including the analysis of clinical trials, epidemiology (especially case–control studies), and Bayesian approaches.

Barnet Woolf (1902–1983) was born in London. His parents were immigrants from Lithuania. Woolf was educated at Cambridge, where he studied physiology and biochemistry, and proposed methods for linearizing plots in enzyme chemistry that were later rediscovered by others (see Haldane [1957]). His later career in London, Birmingham, Rothamsted, and Edinburgh included lasting contributions to nutrition, epidemiology, public health, genetics, and statistics. He was also active in left-wing causes and penned humorous poems, songs, and revues.

□ Technical note

By default, `cc` and `cci` report exact confidence intervals but an approximate significance test. You can replace the approximate test with Fisher's exact test by specifying the `exact` option. We recommend specifying `exact` whenever any cell count is less than 1,000.

```
. cci 4 386 4 1250, exact level(90)
                                                            Proportion
                      Exposed    Unexposed          Total     Exposed

         Cases             4          386            390      0.0103
       Controls            4         1250           1254      0.0032

         Total             8         1636           1644      0.0049

                      Point estimate            [90% Conf. Interval]

      Odds ratio           3.238342           .7698467    13.59664 (exact)
   Attr. frac. ex.             .6912          -.2989599    .9264524 (exact)
   Attr. frac. pop        .0070892

                                   1-sided Fisher's exact P = 0.0964
                                   2-sided Fisher's exact P = 0.0964
```

In this table, the one- and two-sided significance values are equal. This is not a mistake, but it does not happen often. Exact significance values are calculated by summing the probabilities for tables that have the same marginals (row and column sums) but that are less likely (given an odds ratio of 1) than the observed table. When considering each possible table, we might ask if the table is in the same or opposite tail as the observed table. If it is in the same tail, we would count the table under consideration in the one-sided test and, either way, we would count it in the two-sided test. Here all the tables more extreme than this table are in the same tail, so the one- and two-sided tests are the same.

The p-value of 0.0964 is significant at the 10% level, but the exact confidence interval is not (it includes the null odds ratio of one). It was not surprising that the exact interval disagreed with the chi-squared test; after all, they come from different models. Now the exact interval and Fisher's exact test also disagree, even though they come from the same model!

The test and interval disagree because the exact sampling distribution is asymmetric, and the test and interval handle the asymmetry differently. The two-sided test, as we have seen, sums the probabilities of all tables at least as unlikely as the observed table, and in example 10, all the unlikely tables fall in the same tail of the distribution. The other tail does not contribute to the p-value. The exact interval, on the other hand, must always use both tails of the distribution, because the interval inverts two one-sided tests, not one two-sided test (Breslow and Day 1980, 128–129).

□

□ Technical note

The reported value of the attributable or prevented fraction among the exposed is calculated using the odds ratio as a proxy for the risk ratio. This can be justified only if the outcome is rare in the population. The extrapolation to the attributable or prevented fraction for the population assumes that the control group is a random sample of the corresponding group in the underlying population.

□

▷ Example 11: cc equivalent to cci

Equivalent to typing `cci 4 386 4 1250` would be typing `cc case exposed [freq=pop]` with the following data:

```
. use http://www.stata-press.com/data/r14/ccxmpl, clear

. list
```

	case	exposed	pop
1.	1	1	4
2.	1	0	386
3.	0	1	4
4.	0	0	1250

◁

Stratified case–control data

▷ Example 12: cc with stratified data

cc can work with stratified tables. Rothman, Greenland, and Lash (2008, 276) reprint and discuss data from a case–control study on infants with congenital heart disease and Down syndrome and healthy controls, according to maternal spermicide use before conception and maternal age at delivery (Rothman 1982):

	Maternal age to 34		Maternal age 35+	
	Spermicide used	not used	Spermicide used	not used
Down syndrome	3	9	1	3
Controls	104	1059	5	86

The data corresponding to these tables are

```
. use http://www.stata-press.com/data/r14/downs

. list
```

	case	exposed	pop	age
1.	1	1	3	<35
2.	1	0	9	<35
3.	0	1	104	<35
4.	0	0	1059	<35
5.	1	1	1	35+
6.	1	0	3	35+
7.	0	1	5	35+
8.	0	0	86	35+

The stratified results for the odds ratio are

```
. cc case exposed [freq=pop], by(age) woolf
    Maternal age |        OR      [95% Conf. Interval]   M-H Weight
    -------------+-------------------------------------------------------
             <35 |   3.394231      .9048403   12.73242    .7965957 (Woolf)
             35+ |   5.733333      .5016418   65.52706    .1578947 (Woolf)
    -------------+-------------------------------------------------------
           Crude |   3.501529      1.110362   11.04208             (Woolf)
    M-H combined |   3.781172       1.18734   12.04142

Test of homogeneity (M-H)      chi2(1) =     0.14  Pr>chi2 = 0.7105
                    Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      5.81
                                            Pr>chi2 =    0.0159
```

For no particular reason, we also specified the woolf option to obtain Woolf approximations to the within-stratum confidence intervals rather than the default. Had we wanted Tarone's test of homogeneity, we would have used

```
. cc case exposed [freq=pop], by(age) tarone
    Maternal age |        OR      [95% Conf. Interval]   M-H Weight
    -------------+-------------------------------------------------------
             <35 |   3.394231      .5812415   13.87412    .7965957 (exact)
             35+ |   5.733333      .0911619   85.89602    .1578947 (exact)
    -------------+-------------------------------------------------------
           Crude |   3.501529      .8080857   11.78958             (exact)
    M-H combined |   3.781172       1.18734   12.04142

Test of homogeneity (M-H)      chi2(1) =     0.14  Pr>chi2 = 0.7105
Test of homogeneity (Tarone)   chi2(1) =     0.14  Pr>chi2 = 0.7092
                    Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      5.81
                                            Pr>chi2 =    0.0159
```

Whatever method you choose for calculating confidence intervals, Stata will report a test of homogeneity, which here is $\chi^2(1) = 0.14$ and not significant. That is, the odds of Down syndrome might vary with maternal age, but we cannot reject the hypothesis that the association between Down syndrome and spermicide is the same in the two maternal age strata. This is thus a test to reject the appropriateness of the single, Mantel–Haenszel combined odds ratio—a rejection not justified by these data.

◁

❑ Technical note

The cc command includes four tests of homogeneity: Mantel–Haenszel (the default); directly pooled, also known as the Woolf test (available with the pool option); Tarone (available with the tarone option); and Breslow–Day (available with the bd option). The preferred test is Tarone's (Tarone 1985, 94), which corrected an error in the Breslow–Day test; see Breslow (1996, 17–18) for details of the error and Tarone's correction.

The other two homogeneity tests, the Mantel–Haenszel and directly pooled, are less useful: they use the logs of the stratum-specific odds ratios, so they are undefined when any stratum has a zero cell. The epitab commands deal with the problem differently: cs omits the offending strata, while cc substitutes the Tarone test. The Tarone test does not use the stratum-specific odds ratios, so it can still be calculated when there are zero cells.

None of the tests are appropriate for finely stratified (many strata with only a few observations each) studies (Rothman, Greenland, and Lash 2008, 280). If you have fine stratification, one alternative is multilevel logistic regression; see [ME] **melogit**.

❑

❑ Technical note

As with cohort study data, an alternative to stratified tables for uncovering effects is logistic regression. From the logistic point of view, case–control data are no different from cohort study data—you must merely ignore the estimated intercept. The intercept is meaningless in case–control data because it reflects the baseline prevalence of the outcome, which you controlled by sampling.

The data we used with cc can be used directly by logistic. (The age variable, which appears to be a string, is really numeric with an associated value label; see [U] **12.6.3 Value labels**. age takes on the value 0 for the age-less-than-35 group and 1 for the 35+ group.)

```
. logistic case exposed age [freq=pop]
Logistic regression                             Number of obs   =       1,270
                                                LR chi2(2)      =        8.74
                                                Prob > chi2     =      0.0127
Log likelihood = -81.517532                     Pseudo R2       =      0.0509
```

case	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf. Interval]	
exposed	3.787779	2.241922	2.25	0.024	1.187334	12.0836
age	4.582857	2.717352	2.57	0.010	1.433594	14.65029
_cons	.0082631	.0027325	-14.50	0.000	.0043218	.0157988

We compare the results with those presented by cc in the previous example:

Source	Odds ratio	95% CI	
Mantel–Haenszel (cc)	3.78	1.19	12.04
logistic	3.79	1.19	12.08

As with the cohort study data in example 8, results are virtually identical, and all the same comments we made previously apply once again.

To demonstrate an advantage of logistic regression, let's now ask a question that would be difficult to answer on the basis of a stratified table analysis. We now know that spermicide use appears to increase the risk of having a baby with Down syndrome, and we know that the mother's age also increases the risk. Is the effect of spermicide use statistically different for mothers in the two age groups?

```
. logistic case exposed age c.age#exposed [freq=pop]
Logistic regression                             Number of obs   =       1,270
                                                LR chi2(3)      =        8.87
                                                Prob > chi2     =      0.0311
Log likelihood = -81.451332                     Pseudo R2       =      0.0516
```

case	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf. Interval]	
exposed	3.394231	2.289544	1.81	0.070	.9048403	12.73242
age	4.104651	2.774868	2.09	0.037	1.091034	15.44237
exposed#c.age						
1	1.689141	2.388785	0.37	0.711	.1056563	27.0045
_cons	.0084986	.0028449	-14.24	0.000	.0044097	.0163789

The answer is no. The odds ratio and confidence interval reported for `exposed` now measure the spermicide effect for an `age==0` (age $<$ 35) mother. The odds ratio and confidence interval reported for `c.age#exposed` are the (multiplicative) difference in the spermicide odds ratio for an `age==1` (age 35+) mother relative to a young mother. The point estimate is that the effect is larger for older mothers, suggesting grounds for future research, but the difference is not significant.

See [R] **logistic** for an explanation of the `logistic` command. Also see the technical note under *Incidence-rate data* above concerning Poisson regression, which applies equally to logistic regression. ❑

Case–control data with multiple levels of exposure

In a case–control study, subjects with the disease of interest (cases) are compared to disease-free individuals (controls) to assess the relationship between exposure to one or more risk factors and disease incidence. Often exposure is measured qualitatively at several discrete levels or measured on a continuous scale and then grouped into three or more levels. The data can be summarized as

	Exposure level				
	1	2	...	k	Total
Cases	a_1	a_2	...	a_k	M_1
Controls	c_1	c_2	...	c_k	M_0
Total	N_1	N_2	...	N_k	T

An advantage afforded by having multiple levels of exposure is the ability to examine dose–response relationships. If the association between a risk factor and a disease or outcome is real, we expect the strength of that association to increase with the level and duration of exposure. A dose–response relationship provides strong support for a direct or even causal relationship between the risk factor and the outcome. On the other hand, the lack of a dose–response is usually seen as an argument against causality.

We can use the `tabodds` command to tabulate the odds of failure or odds ratios against a categorical exposure variable. The test for trend calculated by `tabodds` can serve as a test for dose–response if the exposure variable is at least ordinal. If the exposure variable has no natural ordering, the trend test is meaningless and should be ignored. See the technical note at the end of this section for more information regarding the test for trend.

Before looking at an example, consider three possible data arrangements for case–control and prevalence studies. The most common data arrangement is individual records, where each subject in the study has his or her own record. Closely related are frequency records where identical individual records are included only once, but with a variable giving the frequency with which the record occurs. The `fweight` *weight* option is used for these data to specify the frequency variable. Data can also be arranged as binomial frequency records where each record contains a variable, D, the number of cases; another variable, N, the number of total subjects (cases plus controls); and other variables. An advantage of binomial frequency records is that large datasets can be entered succinctly into a Stata database.

▷ Example 13: tabodds

Consider the following data from the Ille-et-Vilaine study of esophageal cancer, discussed in Breslow and Day (1980, chap. 4 and app. I), corresponding to subjects age 55–64 who use from 0 to 9 g of tobacco per day:

| | Alcohol consumption (g/day) | | | | |
	0–39	40–79	80–119	120+	Total
Cases	2	9	9	5	25
Controls	47	31	9	5	92
Total	49	40	18	10	117

The study included 24 such tables, each representing one of four levels of tobacco use and one of six age categories. We can create a binomial frequency-record dataset by typing

```
. input alcohol D N agegrp tobacco
         alcohol        D        N     agegrp       tobacco
 1.            1        2       49          4             1
 2.            2        9       40          4             1
 3.            3        9       18          4             1
 4.            4        5       10          4             1
 5. end
```

where D is the number of esophageal cancer cases and N is the number of total subjects (cases plus controls) for each combination of six age groups (agegrp), four levels of alcohol consumption in g/day (alcohol), and four levels of tobacco use in g/day (tobacco).

Both the tabodds and mhodds commands can correctly handle all three data arrangements. Binomial frequency records require that the number of total subjects (cases plus controls) represented by each record N be specified with the binomial() option.

We could also enter the data as frequency-weighted data:

```
.   input alcohol case freq agegrp tobacco
         alcohol      case      freq     agegrp       tobacco
 1.            1         1         2          4             1
 2.            1         0        47          4             1
 3.            2         1         9          4             1
 4.            2         0        31          4             1
 5.            3         1         9          4             1
 6.            3         0         9          4             1
 7.            4         1         5          4             1
 8.            4         0         5          4             1
 9. end
```

If you are planning on using any of the other estimation commands, such as poisson or logistic, we recommend that you enter your data either as individual records or as frequency-weighted records and not as binomial frequency records, because the estimation commands currently do not recognize the binomial() option.

We have entered all the esophageal cancer data into Stata as a frequency-weighted record dataset as previously described. In our data, case indicates the esophageal cancer cases and controls, and freq is the number of subjects represented by each record (the weight).

We added value labels to the agegrp, alcohol, and tobacco variables in our dataset to ease interpretation in outputs, but these variables are numeric.

We are interested in the association between alcohol consumption and esophageal cancer. We first use tabodds to tabulate the odds of esophageal cancer against alcohol consumption:

```
. use http://www.stata-press.com/data/r14/bdesop, clear

. tabodds case alcohol [fweight=freq]
```

alcohol	cases	controls	odds	[95% Conf. Interval]	
0–39	29	386	0.07513	0.05151	0.10957
40–79	75	280	0.26786	0.20760	0.34560
80–119	51	87	0.58621	0.41489	0.82826
120+	45	22	2.04545	1.22843	3.40587

```
Test of homogeneity (equal odds): chi2(3)  =    158.79
                                   Pr>chi2  =     0.0000

Score test for trend of odds:      chi2(1)  =    152.97
                                   Pr>chi2  =     0.0000
```

The test of homogeneity clearly indicates that the odds of esophageal cancer differ by level of alcohol consumption, and the test for trend indicates a significant increase in odds with increasing alcohol use. This suggests a strong dose–response relation. The graph option can be used to study the shape of the relationship of the odds with alcohol consumption. Most of the heterogeneity in these data can be "explained" by the linear increase in risk of esophageal cancer with increased dosage (alcohol consumption).

We also could have requested that the odds ratios at each level of alcohol consumption be calculated by specifying the or option. For example, tabodds case alcohol [fweight=freq], or would produce odds ratios using the minimum value of alcohol—that is, alcohol = 1 (0–39)—as the reference group, and the command tabodds case alcohol [fweight=freq], or base(2) would use alcohol = 2 (40–79) as the reference group.

Although our results appear to provide strong evidence supporting an association between alcohol consumption and esophageal cancer, we need to be concerned with the possible existence of confounders, specifically age and tobacco use, in our data. We can again use tabodds to tabulate the odds of esophageal cancer against age and against tobacco use, independently:

```
. tabodds case agegrp [fweight=freq]
```

agegrp	cases	controls	odds	[95% Conf. Interval]	
25–34	1	115	0.00870	0.00121	0.06226
35–44	9	190	0.04737	0.02427	0.09244
45–54	46	167	0.27545	0.19875	0.38175
55–64	76	166	0.45783	0.34899	0.60061
65–74	55	106	0.51887	0.37463	0.71864
75+	13	31	0.41935	0.21944	0.80138

```
Test of homogeneity (equal odds): chi2(5)  =     96.94
                                   Pr>chi2  =     0.0000

Score test for trend of odds:      chi2(1)  =     83.37
                                   Pr>chi2  =     0.0000
```

```
. tabodds case tobacco [fweight=freq]
```

tobacco	cases	controls	odds	[95% Conf. Interval]	
0–9	78	447	0.17450	0.13719	0.22194
10–19	58	178	0.32584	0.24228	0.43823
20–29	33	99	0.33333	0.22479	0.49428
30+	31	51	0.60784	0.38899	0.94983

```
Test of homogeneity (equal odds): chi2(3)  =     29.33
                                   Pr>chi2  =    0.0000

Score test for trend of odds:      chi2(1)  =     26.93
                                   Pr>chi2  =    0.0000
```

We can see that there is evidence to support our concern that both age and tobacco use are potentially important confounders. Clearly, before we can make any statements regarding the association between esophageal cancer and alcohol use, we must examine and, if necessary, adjust for the effect of any confounder. We will return to this example in the following section.

◁

❑ Technical note

The score test for trend performs a test for linear trend of the log odds against the numerical code used for the exposure variable. The test depends not only on the relationship between dose level and the outcome but also on the numeric values assigned to each level or, to be more accurate, to the distance between the numeric values assigned. For example, the trend test on a dataset with four exposure levels coded 1, 2, 3, and 4 gives the same results as coding the levels 10, 20, 30, and 40 because the distance between the levels in each case is constant. In the first case, the distance is one unit, and in the second case, it is 10 units. However, if we code the exposure levels as 1, 10, 100, and 1,000, we would obtain different results because the distance between exposure levels is not constant. Thus be careful when assigning values to exposure levels. You must determine whether equally spaced numbers make sense for your data or if other more meaningful values should be used.

Remember that we are testing whether a log-linear relationship exists between the odds and the exposure variable. For your particular problem, this relationship may not be correct or even make sense, so you must be careful in interpreting the output of this trend test.

❑

Case–control data with confounders and possibly multiple levels of exposure

In the esophageal cancer data example introduced earlier, we determined that the apparent association between alcohol consumption and esophageal cancer could be confounded by age and tobacco use. You can adjust for the effect of possible confounding factors by stratifying on these factors. This is the method used by both tabodds and mhodds to adjust for other variables in the dataset. We will compare and contrast these two commands in the following example.

▷ Example 14: tabodds, adjusting for confounding factors

We begin by using tabodds to tabulate unadjusted odds ratios.

```
. tabodds case alcohol [fweight=freq], or
```

alcohol	Odds Ratio	chi2	P>chi2	[95% Conf. Interval]	
0–39	1.000000	.	.	.	.
40–79	3.565271	32.70	0.0000	2.237981	5.679744
80–119	7.802616	75.03	0.0000	4.497054	13.537932
120+	27.225705	160.41	0.0000	12.507808	59.262107

```
Test of homogeneity (equal odds): chi2(3)  =    158.79
                                   Pr>chi2  =    0.0000
Score test for trend of odds:      chi2(1)  =    152.97
                                   Pr>chi2  =    0.0000
```

The `alcohol = 1` group (0–39) was used by `tabodds` as the reference category for calculating the odds ratios. We could have selected a different group by specifying the `base()` option; however, because the lowest dosage level is most often the appropriate reference group, as it is in these data, the `base()` option is seldom used.

We use `tabodds` with the `adjust()` option to tabulate Mantel–Haenszel age-adjusted odds ratios:

```
. tabodds case alcohol [fweight=freq], adjust(age)
Mantel-Haenszel odds ratios adjusted for age
```

alcohol	Odds Ratio	chi2	P>chi2	[95% Conf. Interval]	
0–39	1.000000	.	.	.	.
40–79	4.268155	37.36	0.0000	2.570025	7.088314
80–119	8.018305	59.30	0.0000	4.266893	15.067922
120+	28.570426	139.70	0.0000	12.146409	67.202514

```
Score test for trend of odds: chi2(1)  =    135.09
                              Pr>chi2  =    0.0000
```

We observe that the age-adjusted odds ratios are just slightly higher than the unadjusted ones, so it appears that age is not as strong a confounder as it first appeared. Even after adjusting for age, the dose–response relationship, as measured by the trend test, remains strong.

We now perform the same analysis but this time adjust for tobacco use instead of age.

```
. tabodds case alcohol [fweight=freq], adjust(tobacco)
Mantel-Haenszel odds ratios adjusted for tobacco
```

alcohol	Odds Ratio	chi2	P>chi2	[95% Conf. Interval]	
0–39	1.000000	.	.	.	.
40–79	3.261178	28.53	0.0000	2.059764	5.163349
80–119	6.771638	62.54	0.0000	3.908113	11.733306
120+	19.919526	123.93	0.0000	9.443830	42.015528

```
Score test for trend of odds: chi2(1)  =    135.04
                              Pr>chi2  =    0.0000
```

Again we observe a significant dose–response relationship and not much difference between the adjusted and unadjusted odds ratios. We could also adjust for the joint effect of both age and tobacco use by specifying `adjust(tobacco age)`, but we will not bother here.

◁

A different approach to analyzing these data is to use the `mhodds` command. This command estimates the ratio of the odds of failure for two categories of an exposure variable, controlling

for any specified confounding variables, and it tests whether this odds ratio is equal to one. For multiple exposures, if two exposure levels are not specified with compare(), then mhodds assumes that exposure is quantitative and calculates a 1-degree-of-freedom test for trend. This test for trend is the same one that tabodds reports.

▷ Example 15: mhodds, controlling for confounding factors

We first use mhodds to estimate the effect of alcohol controlled for age:

```
. mhodds case alcohol agegrp [fweight=freq]
Score test for trend of odds with alcohol
controlling for agegrp
(The Odds Ratio estimate is an approximation to the odds ratio
for a one unit increase in alcohol)
```

Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]	
2.845895	135.09	0.0000	2.385749	3.394792

Because alcohol has more than two levels, mhodds estimates and reports an approximate age-adjusted odds ratio for a one-unit increase in alcohol consumption. The χ^2 value reported is identical to that reported by tabodds for the score test for trend on the previous page.

We now use mhodds to estimate the effect of alcohol controlled for age, and while we are at it, we do this by levels of tobacco consumption:

```
. mhodds case alcohol agegrp [fweight=freq], by(tobacco)
Score test for trend of odds with alcohol
controlling for agegrp
by tobacco
note: only 19 of the 24 strata formed in this analysis contribute
      information about the effect of the explanatory variable
(The Odds Ratio estimate is an approximation to the odds ratio
for a one unit increase in alcohol)
```

tobacco	Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]	
0-9	3.579667	75.95	0.0000	2.68710	4.76871
10-19	2.303580	25.77	0.0000	1.66913	3.17920
20-29	2.364135	13.27	0.0003	1.48810	3.75589
30+	2.217946	8.84	0.0029	1.31184	3.74992

Mantel–Haenszel estimate controlling for agegrp and tobacco

Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]	
2.751236	118.37	0.0000	2.292705	3.301471

```
Test of homogeneity of ORs (approx): chi2(3)  =    5.46
                                      Pr>chi2  =  0.1409
```

Again, because alcohol has more than two levels, mhodds estimates and reports an approximate Mantel–Haenszel age and tobacco-use adjusted odds ratio for a one-unit increase in alcohol consumption. The χ^2 test for trend reported with the Mantel–Haenszel estimate is again the same one that tabodds produces if adjust(agegrp tobacco) is specified.

The results from this analysis also show an effect of alcohol, controlled for age, of about $\times 2.7$, which is consistent across different levels of tobacco consumption. Similarly,

```
. mhodds case tobacco agegrp [fweight=freq], by(alcohol)

Score test for trend of odds with tobacco
controlling for agegrp
by alcohol

note: only 18 of the 24 strata formed in this analysis contribute
      information about the effect of the explanatory variable

(The Odds Ratio estimate is an approximation to the odds ratio
for a one unit increase in tobacco)
```

alcohol	Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]	
0-39	2.420650	15.61	0.0001	1.56121	3.75320
40-79	1.427713	5.75	0.0165	1.06717	1.91007
80-119	1.472218	3.38	0.0659	0.97483	2.22339
120+	1.214815	0.59	0.4432	0.73876	1.99763

Mantel-Haenszel estimate controlling for agegrp and alcohol

Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]	
1.553437	20.07	0.0000	1.281160	1.883580

```
Test of homogeneity of ORs (approx): chi2(3)   =     5.26
                                     Pr>chi2   =   0.1540
```

shows an effect of tobacco, controlled for age, of about $\times 1.5$, which is consistent across different levels of alcohol consumption.

Comparisons between particular levels of alcohol and tobacco consumption can be made by generating a new variable with levels corresponding to all combinations of alcohol and tobacco, as in

```
. egen alctob = group(alcohol tobacco)

. mhodds case alctob [fweight=freq], compare(16,1)

Maximum likelihood estimate of the odds ratio
Comparing alctob==16 vs. alctob==1
```

Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]	
93.333333	103.21	0.0000	14.766136	589.938431

which yields an odds ratio of 93 between subjects with the highest levels of alcohol and tobacco and those with the lowest levels. Similar results can be obtained simultaneously for all levels of alctob using alctob = 1 as the comparison group by specifying tabodds D alctob, bin(N) or.

◁

Standardized estimates with stratified case–control data

▷ Example 16: cc with stratified data, using standardized estimates

You obtain standardized estimates (here for the odds ratio) by using `cc` just as you obtain standardized estimates by using `ir` or `cs`. Along with the `by()` option, you specify one of `estandard`, `istandard`, or `standard(varname)`.

Case–control studies can provide standardized rate-ratio estimates when density sampling is used, or when the disease is rare (Rothman, Greenland, and Lash 2008, 269). Rothman, Greenland, and Lash (2008, 276) report the SMR for the case–control study on infants with congenital heart disease and Down syndrome. We can reproduce their estimates along with the pooled estimates by typing

```
. use http://www.stata-press.com/data/r14/downs, clear
. cc case exposed [freq=pop], by(age) istandard pool
```

Maternal age	OR	[95% Conf. Interval]		Weight
<35	3.394231	.5812415	13.87412	104 (exact)
35+	5.733333	.0911619	85.89602	5 (exact)
Crude	3.501529	.8080857	11.78958	(exact)
Pooled (direct)	3.824166	1.196437	12.22316	
I. Standardized	3.779749	1.180566	12.10141	

```
Test of homogeneity (direct)   chi2(1) =     0.14  Pr>chi2 = 0.7109
```

Using the distribution of the nonexposed subjects in the source population as the standard, we can obtain an estimate of the standardized rate ratio (SRR):

```
. cc case exposed [freq=pop], by(age) estandard
```

Maternal age	OR	[95% Conf. Interval]		Weight
<35	3.394231	.5812415	13.87412	1059 (exact)
35+	5.733333	.0911619	85.89602	86 (exact)
Crude	3.501529	.8080857	11.78958	(exact)
E. Standardized	3.979006	1.176096	13.46191	

Finally, if we wanted to weight the two age groups equally, we could type

```
. generate wgt=1
. cc case exposed [freq=pop], by(age) standard(wgt)
```

Maternal age	OR	[95% Conf. Interval]		Weight
<35	3.394231	.5812415	13.87412	1 (exact)
35+	5.733333	.0911619	85.89602	1 (exact)
Crude	3.501529	.8080857	11.78958	(exact)
Standardized	5.275104	.6233794	44.6385	

◁

Matched case–control data

Matched case–control studies are performed to gain sample-size efficiency and to control for important confounding factors. In a matched case–control design, each case is matched with a control on the basis of demographic characteristics, clinical characteristics, etc. Thus their difference with respect to the outcome must be due to something other than the matching variables. If the only difference between them was exposure to the factor, we could attribute any difference in outcome to the factor.

A summary of the data is

| | Controls | | |
Cases	Exposed	Unexposed	Total
Exposed	a	b	M_1
Unexposed	c	d	M_0
Total	N_1	N_0	$T = a + b + c + d$

Each entry in the table represents the number of case–control pairs. For instance, in a of the pairs, both members were exposed; in b of the pairs, the case was exposed but the control was not; and so on. In total, T pairs were observed.

▷ Example 17: mcci

Rothman (1986, 257) discusses data from Jick et al. (1973) on a matched case–control study of myocardial infarction and drinking six or more cups of coffee per day (persons drinking from one to five cups per day were excluded):

| | Controls | |
Cases	6+ cups	0 cups
6+ cups	8	8
0 cups	3	8

`mcci` analyzes matched case–control data:

```
. mcci 8 8 3 8
                 │       Controls
Cases            │   Exposed    Unexposed  │      Total
─────────────────┼─────────────────────────┼────────────
        Exposed  │         8            8  │         16
      Unexposed  │         3            8  │         11
─────────────────┼─────────────────────────┼────────────
          Total  │        11           16  │         27

McNemar's chi2(1) =      2.27    Prob > chi2 = 0.1317
Exact McNemar significance probability       = 0.2266

Proportion with factor
      Cases       .5925926
      Controls    .4074074      [95% Conf. Interval]
                 ──────────     ──────────────────────
      difference  .1851852       -.0822542    .4526246
      ratio       1.454545        .891101     2.374257
      rel. diff.  .3125          -.0243688    .6493688
      odds ratio  2.666667        .6400364   15.6064     (exact)
```

The point estimate states that the odds of drinking 6 or more cups of coffee per day is 2.67 times greater among the myocardial infarction patients. The confidence interval is wide, however, and the p-value of 0.1317 from McNemar's test is not statistically significant.

◁

`mcc` works like the other nonimmediate commands but does not handle stratified data. If you have stratified matched case–control data, you can use conditional logistic regression to estimate odds ratios; see [R] clogit.

Matched case–control studies can also be analyzed using `mhodds` by controlling on the variable used to identify the matched sets. For example, if the variable `set` is used to identify the matched set for each subject,

```
. mhodds fail xvar set
```

will do the job. Any attempt to control for further variables will restrict the analysis to the comparison of cases and matched controls that share the same values of these variables. In general, this would lead to the omission of many records from the analysis. Similar considerations usually apply when investigating effect modification by using the by() option. An important exception to this rule is that a variable used in matching cases to controls may appear in the by() option without loss of data.

▷ Example 18: mhodds with matched case–control data

Let's use mhodds to analyze matched case–control studies using the study of endometrial cancer and exposure to estrogen described in Breslow and Day (1980, chap. 5). In this study, there are four controls matched to each case. Cases and controls are matched on age, marital status, and time living in the community. The data collected include information on the daily dose of conjugated estrogen therapy. Breslow and Day created four levels of the dose variable and began by analyzing the 1:1 study formed by using the first control in each set. We examine the effect of exposure to estrogen:

```
. use http://www.stata-press.com/data/r14/bdendo11, clear

. describe

Contains data from http://www.stata-press.com/data/r14/bdendo11.dta
  obs:           126
 vars:            13                          3 Mar 2014 23:29
 size:         2,394

              storage   display    value
variable name   type    format     label     variable label

set             int     %8.0g                Set number
fail            byte    %8.0g                Case=1/Control=0
gall            byte    %8.0g                Gallbladder dis
hyp             byte    %8.0g                Hypertension
ob              byte    %8.0g                Obesity
est             byte    %8.0g                Estrogen
dos             byte    %8.0g                Ordinal dose
dur             byte    %8.0g                Ordinal duration
non             byte    %8.0g                Non-estrogen drug
duration        int     %8.0g                months
age             int     %8.0g                years
cest            byte    %8.0g                Conjugated est dose
agegrp          float   %9.0g                age group of set

Sorted by: set

. mhodds fail est set

Mantel-Haenszel estimate of the odds ratio
Comparing est==1 vs. est==0, controlling for set

note: only 32 of the 63 strata formed in this analysis contribute
      information about the effect of the explanatory variable
```

Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]	
9.666667	21.12	0.0000	2.944702	31.733072

For the 1:1 matched study, the Mantel–Haenszel methods are equivalent to conditional likelihood methods. The maximum conditional likelihood estimate of the odds ratio is given by the ratio of the off-diagonal frequencies in the two-way (case–control) table below. The data must be in the 1-observation-per-group format; that is, the matched case and control must appear in 1 observation (the same format as required by the mcc command; see also [R] clogit).

```
. keep fail est set

. reshape wide est, i(set) j(fail)
(note: j = 0 1)
Data                                    long    ->    wide

Number of obs.                           126    ->         63
Number of variables                        3    ->          3
j variable (2 values)                   fail    ->    (dropped)
xij variables:
                                         est    ->    est0 est1

. rename est1 case

. rename est0 control

. label variable case case

. label variable control control

. tabulate case control

                    |       control
           case     |       0           1  |      Total
      --------------+----------------------+----------
              0     |       4           3  |          7
              1     |      29          27  |         56
      --------------+----------------------+----------
          Total     |      33          30  |         63
```

The odds ratio is $29/3 = 9.67$, which agrees with the value obtained from mhodds. In the more general $1:m$ matched study, however, the Mantel–Haenszel methods are no longer equivalent to maximum conditional likelihood, although they are usually close.

To illustrate the use of the by() option in matched case–control studies, we look at the effect of exposure to estrogen, stratified by age3, which codes the sets into three age groups (55–64, 65–74, and 75+) as follows:

```
. use http://www.stata-press.com/data/r14/bdendo11, clear

. generate age3 = agegrp

. recode age3 1/2=1 3/4=2 5/6=3
(age3: 124 changes made)

. mhodds fail est set, by(age3)
Mantel-Haenszel estimate of the odds ratio
Comparing est==1 vs. est==0, controlling for set
by age3
note: only 32 of the 63 strata formed in this analysis contribute
      information about the effect of the explanatory variable
```

age3	Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]	
1	6.000000	3.57	0.0588	0.72235	49.83724
2	15.000000	12.25	0.0005	1.98141	113.55557
3	8.000000	5.44	0.0196	1.00059	63.96252

```
      Mantel-Haenszel estimate controlling for set and age3
```

Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]	
9.666667	21.12	0.0000	2.944702	31.733072

```
Test of homogeneity of ORs (approx): chi2(2)  =     0.41
                                     Pr>chi2  =   0.8128
```

There is no further loss of information when we stratify by `age3` because age was one of the matching variables.

The full set of matched controls can be used in the same way. For example, the effect of exposure to estrogen is obtained (using the full dataset) with

```
. use http://www.stata-press.com/data/r14/bdendo, clear
. mhodds fail est set
Mantel-Haenszel estimate of the odds ratio
Comparing est==1 vs. est==0, controlling for set
note: only 58 of the 63 strata formed in this analysis contribute
      information about the effect of the explanatory variable
```

Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]	
8.461538	31.16	0.0000	3.437773	20.826746

The effect of exposure to estrogen, stratified by `age3`, is obtained with

```
. generate age3 =agegrp
. recode age3 1/2=1 3/4=2 5/6=3
(age3: 310 changes made)
. mhodds fail est set, by(age3)
Mantel-Haenszel estimate of the odds ratio
Comparing est==1 vs. est==0, controlling for set
by age3
note: only 58 of the 63 strata formed in this analysis contribute
      information about the effect of the explanatory variable
```

age3	Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]	
1	3.800000	3.38	0.0660	0.82165	17.57438
2	10.666667	18.69	0.0000	2.78773	40.81376
3	13.500000	9.77	0.0018	1.59832	114.02620

Mantel-Haenszel estimate controlling for set and age3

Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]	
8.461538	31.16	0.0000	3.437773	20.826746

```
Test of homogeneity of ORs (approx): chi2(2)  =    1.41
                                     Pr>chi2  =  0.4943
```

◁

Video examples

[Immediate commands in Stata: Computing incidence-rate ratios from summary data](#) (`iri`)

[Immediate commands in Stata: Computing risk ratios from summary data](#) (`csi`)

[Odds ratios for case–control data](#) (`cc`)

[Stratified analysis of case–control data](#) (`cc`)

[Immediate commands in Stata: Computing odds ratios from summary data](#) (`cci`)

Glossary

attributable fraction. An attributable fraction is the reduction in the risk of a disease or other condition of interest when a particular risk factor is removed.

case–control studies. In case–control studies, cases meeting a fixed criterion are matched to noncases ex post to study differences in possible covariates. Relative sample sizes are usually fixed at 1:1 or 1:2 but sometimes vary once the survey is complete. In any case, sample sizes do not reflect the distribution in the underlying population.

cohort studies. In cohort studies, a group that is well defined is monitored over time to track the transition of noncases to cases. Cohort studies differ from incidence studies in that they can be retrospective as well as prospective.

confounding. In the analysis of contingency tables, factor or interaction effects are said to be confounded when the effect of one factor is combined with that of another. For example, the effect of alcohol consumption on esophageal cancer may be confounded with the effects of age, smoking, or both. In the presence of confounding, it is often useful to stratify on the confounded factors that are not of primary interest, in the above example, age and smoking.

cross-sectional or **prevalence studies**. Cross-sectional studies sample distributions of healthy and diseased subjects in the population at one point in time.

crude estimate. A crude estimate has not been adjusted for the effects of other variables. Disregarding a stratification variable, for example, yields a crude estimate.

incidence and **incidence rate**. Incidence is the number of new failures (for example, number of new cases of a disease) that occur during a specified period in a population at risk (for example, of the disease).

Incidence rate is incidence divided by the sum of the length of time each individual was exposed to the risk.

Do not confuse incidence with prevalence. Prevalence is the fraction of a population that has the disease. Incidence refers to the rate at which people contract a disease, whereas prevalence is the total number actually sick at a given time.

incidence studies, **longitudinal studies**, and **follow-up studies**. Whichever word is used, these studies monitor a population for a time to track the transition of noncases into cases. Incidence studies are prospective. Also see *cohort studies*.

matched case–control study. Also known as a retrospective study, a matched case–control study is a study in which persons with positive outcomes are each matched with one or more persons with negative outcomes but with similar characteristics.

odds and **odds ratio**. The odds in favor of an event are $o = p/(1 - p)$, where p is the probability of the event. Thus if $p = 0.2$, the odds are 0.25, and if $p = 0.8$, the odds are 4.

The log of the odds is $\ln(o) = \text{logit}(p) = \ln\{p/(1 - p)\}$, and logistic-regression models, for instance, fit $\ln(o)$ as a linear function of the covariates.

The odds ratio is a ratio of two odds: o_1/o_0. The individual odds that appear in the ratio are usually for an experimental group and a control group, or two different demographic groups.

prevented fraction. A prevented fraction is the reduction in the risk of a disease or other condition of interest caused by including a protective risk factor or public-health intervention.

prospective study. Also known as a prospective longitudinal study, a prospective study is a study based on observations over the same subjects for a given period.

risk factor. This is a variable associated with an increased or decreased risk of failure.

risk ratio. In a log-linear model, this is the ratio of probability of survival associated with a one-unit increase in a risk factor relative to that calculated without such an increase, that is, $R(x + 1)/R(x)$. Given the exponential form of the model, $R(x + 1)/R(x)$ is constant and is given by the exponentiated coefficient.

SMR. See *standardized mortality (morbidity) ratio*.

standardized mortality (morbidity) ratio. Standardized mortality (morbidity) ratio (SMR) is the observed number of deaths divided by the expected number of deaths. It is calculated using indirect standardization: you take the population of the group of interest—say, by age, sex, and other factors—and calculate the expected number of deaths in each cell (expected being defined as the number of deaths that would have been observed if those in the cell had the same mortality as some other population). You then take the ratio to compare the observed with the expected number of deaths. For instance,

Age	(1) Population of group	(2) Deaths per 100,000 in general pop.	(1)×(2) Expected # of deaths	(4) Observed deaths
25–34	95,965	105.2	100.9	92
34–44	78,280	203.6	159.4	180
44–54	52,393	428.9	224.7	242
55–64	28,914	964.6	278.9	312
Total			763.9	826

$$\text{SMR} = 826/763.9 = 1.08$$

stratified test. A stratified test is performed separately for each stratum. The stratum-specific results are then combined into an overall test statistic.

Stored results

ir and iri store the following in r():

Scalars
r(p)	one-sided p-value
r(ird)	incidence-rate difference
r(lb_ird)	lower bound of CI for ird
r(ub_ird)	upper bound of CI for ird
r(irr)	incidence-rate ratio
r(lb_irr)	lower bound of CI for irr
r(ub_irr)	upper bound of CI for irr
r(afe)	attributable (prev.) fraction among exposed
r(lb_afe)	lower bound of CI for afe
r(ub_afe)	upper bound of CI for afe
r(afp)	attributable fraction for the population
r(crude)	crude estimate (ir only)
r(lb_crude)	lower bound of CI for crude
r(ub_crude)	upper bound of CI for crude
r(pooled)	pooled estimate (ir only)
r(lb_pooled)	lower bound of CI for pooled
r(ub_pooled)	upper bound of CI for pooled
r(chi2_mh)	Mantel–Haenszel homogeneity χ^2 (ir only)
r(chi2_p)	pooled homogeneity χ^2
r(df)	degrees of freedom (ir only)

`cs` and `csi` store the following in `r()`:

Scalars

r(p)	two-sided p-value
r(rd)	risk difference
r(lb_rd)	lower bound of CI for rd
r(ub_rd)	upper bound of CI for rd
r(rr)	risk ratio
r(lb_rr)	lower bound of CI for rr
r(ub_rr)	upper bound of CI for rr
r(or)	odds ratio
r(lb_or)	lower bound of CI for or
r(ub_or)	upper bound of CI for or
r(afe)	attributable (prev.) fraction among exposed
r(lb_afe)	lower bound of CI for afe
r(ub_afe)	upper bound of CI for afe
r(afp)	attributable fraction for the population
r(crude)	crude estimate (cs only)
r(lb_crude)	lower bound of CI for crude
r(ub_crude)	upper bound of CI for crude
r(pooled)	pooled estimate (cs only)
r(lb_pooled)	lower bound of CI for pooled
r(ub_pooled)	upper bound of CI for pooled
r(chi2_mh)	Mantel–Haenszel heterogeneity χ^2 (cs only)
r(chi2_p)	pooled heterogeneity χ^2
r(df)	degrees of freedom
r(chi2)	χ^2
r(p_exact)	2-sided Fisher's exact p (exact only)
r(p1_exact)	1-sided Fisher's exact p (exact only)

`cc` and `cci` store the following in `r()`:

Scalars

r(p)	two-sided p-value
r(p1_exact)	χ^2 or one-sided exact significance
r(p_exact)	two-sided significance
r(or)	odds ratio
r(lb_or)	lower bound of CI for or
r(ub_or)	upper bound of CI for or
r(afe)	attributable (prev.) fraction among exposed
r(lb_afe)	lower bound of CI for afe
r(ub_afe)	upper bound of CI for afe
r(afp)	attributable fraction for the population
r(crude)	crude estimate (cc only)
r(lb_crude)	lower bound of CI for crude
r(ub_crude)	upper bound of CI for crude
r(pooled)	pooled estimate (cc only)
r(lb_pooled)	lower bound of CI for pooled
r(ub_pooled)	upper bound of CI for pooled
r(chi2_p)	pooled heterogeneity χ^2
r(chi2_bd)	Breslow–Day χ^2
r(df_bd)	degrees of freedom for Breslow–Day χ^2
r(chi2_t)	Tarone χ^2
r(df_t)	degrees of freedom for Tarone χ^2
r(df)	degrees of freedom
r(chi2)	χ^2

`tabodds` stores the following in `r()`:

Scalars
`r(odds)`	odds
`r(lb_odds)`	lower bound for `odds`
`r(ub_odds)`	upper bound for `odds`
`r(chi2_hom)`	χ^2 test of homogeneity
`r(p_hom)`	p-value for test of homogeneity
`r(df_hom)`	degrees of freedom for χ^2 test of homogeneity
`r(chi2_tr)`	χ^2 for score test for trend
`r(p_trend)`	p-value for score test for trend

`mhodds` stores the following in `r()`:

Scalars
`r(p)`	two-sided p-value
`r(or)`	odds ratio
`r(lb_or)`	lower bound of CI for `or`
`r(ub_or)`	upper bound of CI for `or`
`r(chi2_hom)`	χ^2 test of homogeneity
`r(df_hom)`	degrees of freedom for χ^2 test of homogeneity
`r(chi2)`	χ^2

`mcc` and `mcci` store the following in `r()`:

Scalars
`r(p_exact)`	two-sided significance
`r(or)`	odds ratio
`r(lb_or)`	lower bound of CI for `or`
`r(ub_or)`	upper bound of CI for `or`
`r(D_f)`	difference in proportion with factor
`r(lb_D_f)`	lower bound of CI for `D_f`
`r(ub_D_f)`	upper bound of CI for `D_f`
`r(R_f)`	ratio of proportion with factor
`r(lb_R_f)`	lower bound of CI for `R_f`
`r(ub_R_f)`	upper bound of CI for `R_f`
`r(RD_f)`	relative difference in proportion with factor
`r(lb_RD_f)`	lower bound of CI for `RD_f`
`r(ub_RD_f)`	upper bound of CI for `RD_f`
`r(chi2)`	χ^2

Methods and formulas

The notation for incidence-rate data is

	Exposed	Unexposed	Total
Cases	a	b	M_1
Person-time	N_1	N_0	T

The notation for 2×2 tables is

	Exposed	Unexposed	Total
Cases	a	b	M_1
Controls	c	d	M_0
Total	N_1	N_0	T

The notation for $2 \times k$ tables is

	Exposure level
	1	2	...	k	Total
Cases	a_1	a_2	...	a_k	M_1
Controls	c_1	c_2	...	c_k	M_0
Total	N_1	N_2	...	N_k	T

If the tables are stratified, all quantities are indexed by i, the stratum number.

We will refer to Fleiss, Levin, and Paik (2003); Kleinbaum, Kupper, and Morgenstern (1982); and Rothman (1986) so often that we will adopt the notation F-23 to mean Fleiss, Levin, and Paik (2003) page 23; KKM-52 to mean Kleinbaum, Kupper, and Morgenstern (1982) page 52; and R-164 to mean Rothman (1986) page 164.

We usually avoid making the continuity corrections to χ^2 statistics, following the advice of KKM-292: "... the use of a continuity correction has been the subject of considerable debate in the statistical literature .... On the basis of our evaluation of this debate and other evidence, we do *not* recommend the use of the continuity correction." Breslow and Day (1980, 133), on the other hand, argue for inclusion of the correction, but not strongly. Their summary is that for small datasets, one should use exact statistics. In practice, we believe that the adjustment makes little difference for reasonably sized datasets.

Methods and formulas are presented under the following headings:

> Unstratified incidence-rate data (ir and iri)
> Unstratified cumulative incidence data (cs and csi)
> Unstratified case–control data (cc and cci)
> Unstratified matched case–control data (mcc and mcci)
> Stratified incidence-rate data (ir with the by() option)
> Stratified cumulative incidence data (cs with the by() option)
> Stratified case–control data (cc with by() option, mhodds, tabodds)

Unstratified incidence-rate data (ir and iri)

The incidence-rate difference is defined as $I_d = a/N_1 - b/N_0$ (R-164). The standard error of the difference is $s_{I_d} \approx \sqrt{a/N_1^2 + b/N_0^2}$ (R-170), from which confidence intervals are calculated.

The incidence-rate ratio is defined as $I_r = (a/N_1)/(b/N_0)$ (R-164). Let p_l and p_u be the exact confidence interval of the binomial probability for observing a successes in M_1 trials (obtained from cii proportions; see [R] ci). The exact confidence interval for the incidence ratio is then $(p_l N_0)/\{(1-p_l)N_1\}$ to $(p_u N_0)/\{(1-p_u)N_1\}$ (R-166).

The attributable fraction among the exposed is defined as AFE $= (I_r - 1)/I_r$ for $I_r \geq 1$ (KKM-164; R-38); the confidence interval is obtained by similarly transforming the interval values of I_r. The attributable fraction for the population is AF $=$ AFE $\cdot a/M_1$ (KKM-161); no confidence interval is reported. For $I_r < 1$, the prevented fraction among the exposed is defined as PFE $= 1 - I_r$ (KKM-166; R-39); the confidence interval is obtained by similarly transforming the interval values of I_r. The prevented fraction for the population is PF $=$ PFE $\cdot N_1/T$ (KKM-165); no confidence interval is reported.

The "midp" one-sided exact significance (R-155) is calculated as the binomial probability (with $n = M_1$ and $p = N_1/T$) $\Pr(k = a)/2 + \Pr(k > a)$ if $I_r \geq 1$ and $\Pr(k = a)/2 + \Pr(k < a)$ otherwise. The two-sided significance is twice the one-sided significance (R-155). If preferred, you can obtain nonmidp exact probabilities (and, to some ways of thinking, a more reasonable definition of two-sided significance) using bitest; see [R] **bitest**.

Unstratified cumulative incidence data (cs and csi)

The risk difference is defined as $R_d = a/N_1 - b/N_0$ (R-164). Its standard error is

$$s_{R_d} \approx \left\{ \frac{ac}{N_1^3} + \frac{bd}{N_0^3} \right\}^{1/2}$$

(R-172), from which confidence intervals are calculated.

The risk ratio is defined as $R_r = (a/N_1)/(b/N_0)$ (R-165). The standard error of $\ln R_r$ is

$$s_{\ln R_r} \approx \left(\frac{c}{aN_1} + \frac{d}{bN_0} \right)^{1/2}$$

(R-173), from which confidence intervals are calculated.

For $R_r \geq 1$, the attributable fraction among the exposed is calculated as AFE $= (R_r - 1)/R_r$ (KKM-164; R-38); the confidence interval is obtained by similarly transforming the interval values for R_r. The attributable fraction for the population is calculated as AF $=$ AFE $\cdot a/M_1$ (KKM-161); no confidence interval is reported, but F-128 provides

$$\left\{ \frac{c + (a + d)\text{AF}}{bT} \right\}^{1/2}$$

as the approximate standard error of $\ln(1 - \text{AF})$.

For $R_r < 1$, the prevented fraction among the exposed is calculated as PFE $= 1 - R_r$ (KKM-166; R-39); the confidence interval is obtained by similarly transforming the interval values for R_r. The prevented fraction for the population is calculated as PF $=$ PFE $\cdot N_1/T$; no confidence interval is reported.

The odds ratio, available with the or option, is defined as $\psi = (ad)/(bc)$ (R-165). Several confidence intervals are available. The default interval for cs and csi is the Cornfield (1956) approximate interval. If we let z_α be the index from a normal distribution for an α significance level, the Cornfield interval (ψ_l, ψ_u) is calculated from

$$\psi_l = a_l(M_0 - N_1 + a_l) \Big/ \left\{ (N_1 - a_l)(M_1 - a_l) \right\}$$
$$\psi_u = a_u(M_0 - N_1 + a_u) \Big/ \left\{ (N_1 - a_u)(M_1 - a_u) \right\}$$

where a_u and a_l are determined iteratively from

$$a_{i+1} = a \pm z_\alpha \left(\frac{1}{a_i} + \frac{1}{N_1 - a_i} + \frac{1}{M_1 - a_i} + \frac{1}{M_0 - N_1 + a_i} \right)^{-1/2}$$

(Newman 2001, sec. 4.4). a_{i+1} converges to a_u using the plus sign and a_l using the minus sign. a_0 is taken as a. With small numbers, the iterative technique may fail. It is then restarted by decrementing (a_l) or incrementing (a_u) a_0. If that fails, a_0 is again decremented or incremented and iterations restarted, and so on, until a terminal condition is met ($a_0 < 0$ or $a_0 > M_1$), at which point the value is not calculated.

The Woolf odds-ratio confidence intervals are available with `cs` and `csi`. The Woolf method (Woolf 1955; R-173; Schlesselman 1982, 176), available with the `woolf` option, estimates the standard error of $\ln\psi$ by

$$s_{\ln\psi} = \left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)^{1/2}$$

from which confidence intervals are calculated. The Woolf interval cannot be calculated when there exists a zero cell. Sometimes the Woolf interval is called the "logit interval" (Breslow and Day 1980, 134).

The χ^2 statistic, reported by default, can be calculated as

$$\chi^2 = \frac{(ad - bc)^2 T}{M_1 M_0 N_1 N_0}$$

(Schlesselman 1982, 179).

Fisher's exact test, available with the `exact` option, is calculated as described in [R] **tabulate twoway**.

Unstratified case–control data (cc and cci)

`cc` and `cci` report by default the same odds ratio, ψ, that is available with the `or` option in `cs` and `csi`. But `cc` and `cci` calculate the confidence interval differently: they default to the exact odds-ratio interval, not the Cornfield interval, but you can request the Cornfield interval with the `cornfield` option. The $1 - \alpha$ exact interval $(\underline{R}, \overline{R})$ is calculated from

$$\alpha/2 = \frac{\sum_{k=a}^{\min(N_1, M_1)} \binom{N_1}{k}\binom{N_0}{M_1-k}\underline{R}^k}{\sum_{k=\max(0, M_1-N_0)}^{\min(N_1, M_1)} \binom{N_1}{k}\binom{N_0}{M_1-k}\underline{R}^k}$$

and

$$1 - \alpha/2 = \frac{\sum_{k=a+1}^{\min(N_1, M_1)} \binom{N_1}{k}\binom{N_0}{M_1-k}\overline{R}^k}{\sum_{k=\max(0, M_1-N_0)}^{\min(N_1, M_1)} \binom{N_1}{k}\binom{N_0}{M_1-k}\overline{R}^k}$$

(R-169). The equations invert two one-sided Fisher exact tests.

`cc` and `cci` also report the same tests of significance as `cs` and `csi`: the χ^2 statistic is the default, and Fisher's exact test is obtained with the `exact` option. The odds ratio, ψ, is used as an estimate of the risk ratio in calculating attributable or prevented fractions. For $\psi \geq 1$, the attributable fraction among the exposed is calculated as AFE $= (\psi - 1)/\psi$ (KKM-164); the confidence interval is obtained by similarly transforming the interval values for ψ. The attributable fraction for the population is calculated as AF $=$ AFE $\cdot \, a/M_1$ (KKM-161). No confidence interval is reported; however, F-152 provides

$$\left(\frac{a}{M_1 b} + \frac{c}{M_0 d}\right)^{1/2}$$

as the standard error of $\ln(1 - \text{AF})$.

For $\psi < 1$, the prevented fraction among the exposed is calculated as PFE $= 1 - \psi$ (KKM-166); the confidence interval is obtained by similarly transforming the interval values for ψ. The prevented fraction for the population is calculated as PF $= \{(a/M_1)\text{PFE}\}/\{(a/M_1)\text{PFE} + \psi\}$ (KKM-165); no confidence interval is reported.

Unstratified matched case–control data (mcc and mcci)

Referring to the table at the beginning of *Methods and formulas*, the columns of the 2×2 table indicate controls; the rows are cases. Each entry in the table reflects a pair of a matched case and control.

McNemar's (1947) χ^2 is defined as

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

(KKM-389).

The proportion of controls with the factor is $p_1 = N_1/T$, and the proportion of cases with the factor is $p_2 = M_1/T$.

The difference in the proportions is $P_d = p_2 - p_1$. An estimate of its standard error when the two underlying proportions are *not* hypothesized to be equal is

$$s_{P_d} \approx \frac{\{(a+d)(b+c) + 4bc\}^{1/2}}{T^{3/2}}$$

(F-378), from which confidence intervals are calculated. The confidence interval uses a continuity correction (F-378, eq. 13.15).

The ratio of the proportions is $P_r = p_2/p_1$ (R-276, R-278). The standard error of $\ln P_r$ is

$$s_{\ln P_r} \approx \left(\frac{b+c}{M_1 N_1}\right)^{1/2}$$

(R-276), from which confidence intervals are calculated.

The relative difference in the proportions is $P_e = (b-c)/(b+d)$ (F-379). Its standard error is

$$s_{P_e} \approx (b+d)^{-2} \{(b+c+d)(bc+bd+cd) - bcd\}^{1/2}$$

(F-379), from which confidence intervals are calculated.

The odds ratio is $\psi = b/c$ (F-376), and the exact Fisher confidence interval is obtained by transforming into odds ratios the exact binomial confidence interval for the binomial parameter from observing b successes in $b + c$ trials (R-264). Binomial confidence limits are obtained from cii proportions (see [R] ci) and are transformed by $p/(1-p)$.

The exact McNemar significance probability is a two-tailed exact test of $H_0: \psi = 1$. The p-value, calculated from the binomial distribution, is

$$\min\left\{1, 2\sum_{k=0}^{\min(b,c)} \binom{b+c}{k}\left(\frac{1}{2}\right)^{b+c}\right\}$$

(Agresti 2013, 416).

Quinn McNemar (1900–1986) was born in West Virginia and attended college there and in Pennsylvania. After a brief spell of high school teaching, he began graduate study of psychology at Stanford and then joined the faculty. McNemar's text *Psychological Statistics*, first published in 1949, was widely influential, and he made many substantive and methodological contributions to the application of statistics in psychology.

Stratified incidence-rate data (ir with the by() option)

Statistics presented for each stratum are calculated independently according to the formulas in *Unstratified incidence-rate data (ir and iri)* above. Within strata, the Mantel–Haenszel style weight is $W_i = b_i N_{1i}/T_i$, and the Mantel–Haenszel combined incidence-rate ratio (Rothman and Boice 1982) is

$$I_{\mathrm{mh}} = \frac{\sum_i a_i N_{0i}/T_i}{\sum_i W_i}$$

(R-196). The standard error for the log of the incidence-rate ratio was derived by Greenland and Robins (1985, 63) and appears in R-213:

$$s_{\ln I_{\mathrm{mh}}} \approx \left\{ \frac{\sum_i M_{1i} N_{1i} N_{0i}/T_i^2}{\left(\sum_i a_i N_{0i}/T_i\right)\left(\sum_i b_i N_{1i}/T_i\right)} \right\}^{1/2}$$

The confidence interval is calculated first on the log scale and then is transformed.

For standardized rates, let w_i be the user-specified weight within stratum i. The standardized rate difference (the `ird` option) and rate ratio are defined as

$$\mathrm{SRD} = \frac{\sum_i w_i(R_{1i} - R_{0i})}{\sum_i w_i}$$

$$\mathrm{SRR} = \frac{\sum_i w_i R_{1i}}{\sum_i w_i R_{0i}}$$

(R-229). The standard error of SRD is

$$s_{\mathrm{SRD}} \approx \left\{ \frac{1}{(\sum_i w_i)^2} \sum_i w_i^2 \left(\frac{a_i}{N_{1i}^2} + \frac{b_i}{N_{0i}^2} \right) \right\}^{1/2}$$

(R-231), from which confidence intervals are calculated. The standard error of ln(SRR) is

$$s_{\ln(\mathrm{SRR})} \approx \left\{ \frac{\sum_i w_i^2 a_i/N_{1i}^2}{(\sum_i w_i R_{1i})^2} + \frac{\sum_i w_i^2 b_i/N_{0i}^2}{(\sum_i w_i R_{0i})^2} \right\}^{1/2}$$

(R-231), from which confidence intervals are calculated.

Internally and externally standardized measures are calculated using $w_i = N_{1i}$ and $w_i = N_{0i}$, respectively, and are obtained with the `istandard` and `estandard` options, respectively.

Directly pooled estimates are available with the `pool` option. The directly pooled estimate is a weighted average of stratum-specific estimates; each weight, w_i, is inversely proportional to the variance of the estimate for stratum i. The variances for rate differences come from the formulas in *Unstratified incidence-rate data (ir and iri)*, while the variances of log rate-ratios are estimated by $(1/a_i + 1/b_i)$ (R-184). Ratios are averaged in the log scale before being exponentiated. The standard error of the directly pooled estimate is calculated as $1/\sqrt{\sum w_i}$, from which confidence intervals are calculated (R-183–185); the calculation for ratios again uses the log scale.

For rate differences, the χ^2 test of homogeneity is calculated as $\sum (R_{di} - \widehat{R}_d)^2/\mathrm{var}(R_{di})$, where R_{di} are the stratum-specific rate differences and $\widehat{R}_d$ is the directly pooled estimate. The number of degrees of freedom is one less than the number of strata (R-222).

For rate ratios, the same calculation is made, except that it is made on a logarithmic scale using $\ln(R_{ri})$ (R-222), and $\ln(\widehat{R}_d)$ may be the log of either the directly pooled estimate or the Mantel–Haenszel estimate.

Stratified cumulative incidence data (cs with the by() option)

Statistics presented for each stratum are calculated independently according to the formulas in *Unstratified cumulative incidence data (cs and csi)* above. The Mantel–Haenszel χ^2 test (Mantel and Haenszel 1959) is

$$\chi^2_{\text{mh}} = \frac{\left\{\sum_i (a_i - N_{1i}M_{1i}/T_i)\right\}^2}{\sum_i (N_{1i}N_{0i}M_{1i}M_{0i})/\left\{T_i^2(T_i - 1)\right\}}$$

(R-206).

For the odds ratio (available with the or option), the Mantel–Haenszel weight is $W_i = b_i c_i / T_i$, and the combined odds ratio (Mantel and Haenszel 1959) is

$$\psi_{\text{mh}} = \frac{\sum_i a_i d_i / T_i}{\sum_i W_i}$$

(R-195). The standard error (Robins, Breslow, and Greenland 1986) is

$$s_{\ln\psi_{\text{mh}}} \approx \left\{ \frac{\sum_i P_i R_i}{2\left(\sum_i R_i\right)^2} + \frac{\sum_i P_i S_i + Q_i R_i}{2\sum_i R_i \sum_i S_i} + \frac{\sum_i Q_i S_i}{2\left(\sum_i S_i\right)^2} \right\}^{1/2}$$

where

$$P_i = (a_i + d_i)/T_i$$
$$Q_i = (b_i + c_i)/T_i$$
$$R_i = a_i d_i / T_i$$
$$S_i = b_i c_i / T_i$$

(R-220).

For the risk ratio (the default), the Mantel–Haenszel-style weight is $W_i = b_i N_{1i}/T_i$, and the combined risk ratio (Rothman and Boice 1982) is

$$R_{\text{mh}} = \frac{\sum_i a_i N_{0i}/T_i}{\sum_i W_i}$$

(R-196). The standard error (Greenland and Robins 1985) is

$$s_{\ln R_{\text{mh}}} \approx \left\{ \frac{\sum_i (M_{1i}N_{1i}N_{0i} - a_i b_i T_i)/T_i^2}{\left(\sum_i a_i N_{0i}/T_i\right)\left(\sum_i b_i N_{1i}/T_i\right)} \right\}^{1/2}$$

(R-216), from which confidence intervals are calculated.

For standardized rates, let w_i be the user-specified weight within stratum i. The standardized rate difference (SRD, the rd option) and rate ratios (SRR, the default) are defined as in *Stratified incidence-rate data (ir with the by() option)*, where the individual risks are defined $R_{1i} = a_i/N_{1i}$ and $R_{0i} = b_i/N_{0i}$. The standard error of SRD is

$$s_{\text{SRD}} \approx \left[\frac{1}{\left(\sum_i w_i\right)^2} \sum_i w_i^2 \left\{ \frac{a_i(N_{1i} - a_i)}{N_{1i}^3} + \frac{b_i(N_{0i} - b_i)}{N_{0i}^3} \right\} \right]^{1/2}$$

(R-231), from which confidence intervals are calculated. The standard error of ln(SRR) is

$$s_{\ln(\text{SRR})} \approx \left\{ \frac{\sum_i w_i^2 a_i (N_{1i} - a_i)/N_{1i}^3}{(\sum_i w_i R_{1i})^2} + \frac{\sum_i w_i^2 b_i (N_{0i} - b_i)/N_{0i}^3}{(\sum_i w_i R_{0i})^2} \right\}^{1/2}$$

(R-231), from which confidence intervals are calculated.

Internally and externally standardized measures are calculated using $w_i = N_{1i}$ and $w_i = N_{0i}$, respectively, and are obtained with the istandard and estandard options, respectively.

Directly pooled estimates of the odds ratio are available when you specify both the pool and or options. The directly pooled estimate is a weighted average of stratum-specific log odds-ratios; each weight, w_i, is inversely proportional to the variance of the log odds-ratio for stratum i. The variances of the log odds-ratios are estimated by Woolf's method, described under *Unstratified cumulative incidence data (cs and csi)*. The standard error of the directly pooled log odds-ratio is calculated as $1/\sqrt{\sum w_i}$, from which confidence intervals are calculated and then exponentiated (Kahn and Sempos 1989, 113–115).

Direct pooling is also available for risk ratios and risk differences; the variance formulas may be found in *Unstratified cumulative incidence data (cs and csi)*. The directly pooled risk ratio is provided when the pool option is specified. The directly pooled risk difference is provided only when you specify the pool and rd options, and one of the estandard, istandard, and standard() options.

For risk differences, the χ^2 test of homogeneity is calculated as $\sum (R_{di} - \widehat{R}_d)^2/\text{var}(R_{di})$, where R_{di} are the stratum-specific risk differences and $\widehat{R}_d$ is the directly pooled estimate. The number of degrees of freedom is one less than the number of strata (R-222).

For risk and odds ratios, the same calculation is made, except that it is made in the log scale using $\ln(R_{ri})$ or $\ln(\psi_i)$ (R-222), and $\ln(\widehat{R}_d)$ may be the log of either the directly pooled estimate or the Mantel–Haenszel estimate.

Stratified case–control data (cc with by() option, mhodds, tabodds)

Statistics presented for each stratum are calculated independently according to the formulas in *Unstratified cumulative incidence data (cs and csi)* above. The combined odds ratio, ψ_{mh}, and the test that $\psi_{\text{mh}} = 1$ (χ_{mh}^2) are calculated as described in *Stratified cumulative incidence data (cs with the by() option)* above.

For standardized weights, let w_i be the user-specified weight within stratum i. The standardized odds ratio (the standard() option) is calculated as

$$\text{SOR} = \frac{\sum_i w_i a_i/c_i}{\sum_i w_i b_i/d_i}$$

(Greenland 1986, 473). The standard error of ln(SOR) is

$$s_{\ln(\text{SOR})} = \left\{ \frac{\sum_i (w_i a_i/c_i)^2 (\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i})}{\left(\sum_i w_i a_i/c_i\right)^2} \right\}^{1/2}$$

from which confidence intervals are calculated. The internally and externally standardized odds ratios are calculated using $w_i = c_i$ and $w_i = d_i$, respectively.

The directly pooled estimate of the odds ratio (the `pool` option) is calculated as described in *Stratified cumulative incidence data (cs with the by() option)* above.

The directly pooled and Mantel–Haenszel χ^2 tests of homogeneity are calculated as $\sum \left\{ \ln(R_{ri}) - \ln(\widehat{R}_r) \right\}^2 / \text{var}\{\ln(R_{ri})\}$, where R_{ri} are the stratum-specific odds ratios and $\widehat{R}_r$ is the pooled estimate (Mantel–Haenszel or directly pooled). The number of degrees of freedom is one less than the number of strata (R-222).

The Breslow–Day χ^2 test of homogeneity is available with the `bd` option. Let $\widehat{\psi}$ be the Mantel–Haenszel estimate of the common odds ratio, and let $A_i(\widehat{\psi})$ be the fitted count for cell a; $A_i(\widehat{\psi})$ is found by solving the quadratic equation

$$A(M_0 - N_1 + A) = (\widehat{\psi})(M_1 - A)(N_1 - A)$$

and choosing the root that makes all cells in stratum i positive. Let $\text{Var}(a_i; \widehat{\psi})$ be the estimated variance of a_i conditioned on the margins and on an odds ratio of $\widehat{\psi}$:

$$\text{Var}(a_i; \widehat{\psi}) = \left\{ \frac{1}{A_i(\widehat{\psi})} + \frac{1}{M_{1i} - A_i(\widehat{\psi})} + \frac{1}{N_{1i} - A_i(\widehat{\psi})} + \frac{1}{M_{0i} - N_{1i} + A_i(\widehat{\psi})} \right\}^{-1}$$

The Breslow–Day χ^2 statistic is then

$$\sum_i \frac{\{a_i - A_i(\widehat{\psi})\}^2}{\text{Var}(a_i; \widehat{\psi})}$$

The Tarone χ^2 test of homogeneity (the `tarone` option) is calculated as

$$\sum_i \frac{\{a_i - A_i(\widehat{\psi})\}^2}{\text{Var}(a_i; \widehat{\psi})} - \frac{\{\sum_i a_i - \sum_i A_i(\widehat{\psi})\}^2}{\sum_i \text{Var}(a_i; \widehat{\psi})}$$

Tarone (1985) provides this correction to the Breslow–Day statistic to ensure that its distribution is asymptotically chi-squared. Without the correction, the Breslow–Day statistic does not necessarily follow a chi-squared distribution because it is based on the Mantel–Haenszel estimate, $\widehat{\psi}$, which is an inefficient estimator of the common odds ratio.

When the exposure variable has multiple levels, `mhodds` calculates an approximate estimate of the log odds-ratio for a one-unit increase in exposure as the ratio of the score statistic, U, to its variance, V (Clayton and Hills 1993, 103), which are defined below. This is a one-step Newton-Raphson approximation to the maximum likelihood estimate. Within-stratum estimates are combined with Mantel–Haenszel weights.

By default, both `tabodds` and `mhodds` produce test statistics and confidence intervals based on score statistics (Clayton and Hills 1993). `tabodds` reports confidence intervals for the odds of the ith exposure level, unless the `adjust()` or `or` option is specified. The confidence interval for odds_i, $i = 1, \ldots, k$, is given by

$$\text{odds}_i \cdot \exp\left(\pm z \sqrt{1/a_i + 1/c_i} \right)$$

The score χ^2 test of homogeneity of odds is calculated as

$$\chi^2_{k-1} = \frac{T(T-1)}{M_1 M_0} \sum_{i=1}^{k} \frac{(a_i - E_i)^2}{N_i}$$

where $E_i = (M_1 N_i)/T$.

Let l_i denote the value of the exposure at the ith level. The score χ^2 test for trend of odds is calculated as

$$\chi_1^2 = \frac{U^2}{V}$$

where

$$U = \frac{M_1 M_0}{T} \left(\sum_{i=1}^{k} \frac{a_i l_i}{M_1} - \sum_{i=1}^{k} \frac{c_i l_i}{M_0} \right)$$

and

$$V = \frac{M_1 M_0}{T} \left\{ \frac{\sum_{i=1}^{k} N_i l_i^2 - (\sum_{i=1}^{k} N_i l_i)^2 / T}{T-1} \right\}$$

Acknowledgments

John Snow (1813–1858) was born in York, England. From age 14, he worked as an apprentice and assistant to surgeons in northeast England and Yorkshire. In 1836, Snow moved to London; he was admitted to the Royal College of Surgeons in 1838 and the Royal College of Physicians in 1850. He made outstanding contributions to the adoption of anesthesia and is considered one of the originators of modern epidemiology. Snow died following a stroke in 1858.

Snow calculated dosages for ether and chloroform. He personally administered chloroform to Queen Victoria for the births of her last two children, which helped obstetric anesthesia gain wider acceptance.

Snow was skeptical of the miasma theory that cholera was caused by foul air. His essay *On the Mode of Communication of Cholera* was first published in 1849 and then greatly enlarged in 1855 with the results of his very detailed investigation of the role of water supply in the epidemic of 1854 in the Soho district of London. Snow identified the source of the outbreak as the public water pump on Broad Street (now Broadwick Street), leading the local council to remove the pump handle. It was later discovered that the well had been dug very close to an old cesspit. He also mapped the clustering of cholera cases around the pump and related mortality to water sources, clearly showing higher death rates in areas supplied by the Southwark and Vauxhall Waterworks Company, which was taking water from sewage-polluted sections of the River Thames. Snow is widely regarded as a pioneer in public health, epidemiology, and medical geography.

Janet Elizabeth Lane-Claypon (1877–1967) was a pioneer in the use of cohort and case–control studies. She was born in Lincolnshire county, England, and began her studies at the London School of Medicine for Women in 1898. From 1907 to 1912, she was at the Lister Institute of Preventive Medicine, where she was a colleague of Major Greenwood. By the end of her studies, she had obtained a doctorate in both physiology and medicine.

In 1912, Lane-Claypon published one of the first retrospective cohort studies, examining the weight gain of babies fed cow's milk versus babies fed breast milk. Using statistical techniques, she determined that babies fed breast milk gained weight faster; she later employed that knowledge to become a public health advocate for breast feeding.

She also conducted one of the first case–control studies, examining risk factors associated with breast cancer. Her study included 500 women without breast cancer and 500 women with breast cancer. To obtain what was at the time a remarkably large sample, she coordinated data collection from nine different hospitals. Carefully controlling for variables including occupation and infant mortality, she determined that factors like age at first pregnancy, age at menopause, and number of children all influence the incidence of breast cancer; these factors are still considered to be among the prime determinants.

In conjunction with the Ministry of Health, in 1926 Lane-Claypon published one of the first studies to contain long-term follow-up results. In that study, she followed patients who had undergone surgery for breast cancer for up to 10 years after the operation. As is still the case today, her study showed that the sooner the cancer was treated, the better the woman's chance for long-term survival. Notably, her study was also among the first to consider survivorship bias.

References

Agresti, A. 1999. On logit confidence intervals for the odds ratio with small samples. *Biometrics* 55: 597–602.

———. 2013. *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: Wiley.

Boice, J. D., Jr., and R. R. Monson. 1977. Breast cancer in women after repeated fluoroscopic examinations of the chest. *Journal of the National Cancer Institute* 59: 823–832.

Breslow, N. E. 1996. Statistics in epidemiology: The case–control study. *Journal of the American Statistical Association* 91: 14–28.

Breslow, N. E., and N. E. Day. 1980. *Statistical Methods in Cancer Research: Vol. 1—The Analysis of Case–Control Studies*. Lyon: IARC.

Brown, C. C. 1981. The validity of approximation methods for interval estimation of the odds ratio. *American Journal of Epidemiology* 113: 474–480.

Carlin, J. B., and S. Vidmar. 2000. sbe35: Menus for epidemiological statistics. *Stata Technical Bulletin* 56: 15–16. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 86–87. College Station, TX: Stata Press.

Clayton, D. G., and M. Hills. 1993. *Statistical Models in Epidemiology*. Oxford: Oxford University Press.

———. 1995. ssa8: Analysis of case-control and prevalence studies. *Stata Technical Bulletin* 27: 26–31. Reprinted in *Stata Technical Bulletin Reprints*, vol. 5, pp. 227–233. College Station, TX: Stata Press.

Cornfield, J. 1956. A statistical problem arising from retrospective studies. In Vol. 4 of *Proceedings of the Third Berkeley Symposium*, ed. J. Neyman, 135–148. Berkeley, CA: University of California Press.

Cummings, P. 2009. Methods for estimating adjusted risk ratios. *Stata Journal* 9: 175–196.

Dohoo, I., W. Martin, and H. Stryhn. 2010. *Veterinary Epidemiologic Research*. 2nd ed. Charlottetown, Prince Edward Island: VER Inc.

———. 2012. *Methods in Epidemiologic Research*. Charlottetown, Prince Edward Island: VER Inc.

Doll, R., and A. B. Hill. 1966. Mortality of British doctors in relation to smoking: Observations on coronary thrombosis. *Journal of the National Cancer Institute, Monographs* 19: 205–268.

Dupont, W. D. 2009. *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data*. 2nd ed. Cambridge: Cambridge University Press.

Dupont, W. D., and W. D. Plummer, Jr. 1999. sbe31: Exact confidence intervals for odds ratios from case–control studies. *Stata Technical Bulletin* 52: 12–16. Reprinted in *Stata Technical Bulletin Reprints*, vol. 9, pp. 150–154. College Station, TX: Stata Press.

Fagerland, M. F. 2012. Exact and mid-p confidence intervals for the odds ratio. *Stata Journal* 12: 505–514.

Fleiss, J. L., B. Levin, and M. C. Paik. 2003. *Statistical Methods for Rates and Proportions*. 3rd ed. New York: Wiley.

Gart, J. J., and D. G. Thomas. 1982. The performance of three approximate confidence limit methods for the odds ratio. *American Journal of Epidemiology* 115: 453–470.

Gini, R., and J. Pasquini. 2006. Automatic generation of documents. *Stata Journal* 6: 22–39.

Glass, R. I., A. M. Svennerholm, B. J. Stoll, M. R. Khan, K. M. Hossain, M. I. Huq, and J. Holmgren. 1983. Protection against cholera in breast-fed children by antibodies in breast milk. *New England Journal of Medicine* 308: 1389–1392.

Gleason, J. R. 1999. sbe30: Improved confidence intervals for odds ratios. *Stata Technical Bulletin* 51: 24–27. Reprinted in *Stata Technical Bulletin Reprints*, vol. 9, pp. 146–150. College Station, TX: Stata Press.

Greenhouse, S. W., and J. B. Greenhouse. 1998. Cornfield, Jerome. In Vol. 1 of *Encyclopedia of Biostatistics*, ed. P. Armitage and T. Colton, 955–959. Chichester, UK: Wiley.

Greenland, S. 1986. Estimating variances of standardized estimators in case–control studies and sparse data. *Journal of Chronic Diseases* 39: 473–477.

Greenland, S., ed. 1987. *Evolution of Epidemiologic Ideas: Annotated Readings on Concepts and Methods*. Newton Lower Falls, MA: Epidemiology Resources.

Greenland, S., and J. M. Robins. 1985. Estimation of a common effect parameter from sparse follow-up data. *Biometrics* 41: 55–68.

——. 1988. Conceptual problems in the definition and interpretation of attributable fractions. *American Journal of Epidemiology* 128: 1185–1197.

Haldane, J. B. S. 1957. Graphical methods in enzyme chemistry. *Nature* 179: 832.

Hastorf, A. H., E. R. Hilgard, and R. R. Sears. 1988. Quinn McNemar (1900–1986). *American Psychologist* 43: 196–197.

Hempel, S. 2006. *The Medical Detective: John Snow, Cholera and the Mystery of the Broad Street Pump*. London: Granta Books.

Hill, W. G. 1984. Barnet Woolf. *Year Book, Royal Society of Edinburgh 1984* 214–219.

Jewell, N. P. 2004. *Statistics for Epidemiology*. Boca Raton, FL: Chapman & Hall/CRC.

Jick, H., O. S. Miettinen, R. K. Neff, S. Shapiro, O. P. Heinonen, and D. Slone. 1973. Coffee and myocardial infarction. *New England Journal of Medicine* 289: 63–67.

Johnson, S. 2006. *The Ghost Map: The Story of London's Most Terrifying Epidemic—and How It Changed Science, Cities, and the Modern World*. London: Penguin Books.

Kahn, H. A., and C. T. Sempos. 1989. *Statistical Methods in Epidemiology*. New York: Oxford University Press.

Kleinbaum, D. G., L. L. Kupper, and H. Morgenstern. 1982. *Epidemiologic Research: Principles and Quantitative Methods (Industrial Health and Safety)*. New York: Wiley.

Lilienfeld, D. E., and P. D. Stolley. 1994. *Foundations of Epidemiology*. 3rd ed. New York: Oxford University Press.

López-Vizcaíno, M. E., M. I. Pérez-Santiago, and L. Abraira-García. 2000a. sbe32.1: Automated outbreak detection from public health surveillance data: Errata. *Stata Technical Bulletin Reprints* 55: 2.

——. 2000b. sbe32: Automated outbreak detection from public health surveillance data. *Stata Technical Bulletin* 54: 23–25. Reprinted in *Stata Technical Bulletin Reprints*, vol. 9, pp. 154–157. College Station, TX: Stata Press.

MacMahon, B., S. Yen, D. Trichopoulos, K. Warren, and G. Nardi. 1981. Coffee and cancer of the pancreas. *New England Journal of Medicine* 304: 630–633.

Mantel, N., and W. Haenszel. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22: 719–748. Reprinted in *Evolution of Epidemiologic Ideas: Annotated Readings on Concepts and Methods*, ed. S. Greenland, pp. 112–141. Newton Lower Falls, MA: Epidemiology Resources.

Markel, H. 2013. Happy birthday, Dr Snow. *Journal of the American Medical Association* 309: 995–996.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12: 153–157.

Newman, S. C. 2001. *Biostatistical Methods in Epidemiology*. New York: Wiley.

Orsini, N., R. Bellocco, M. Bottai, A. Wolk, and S. Greenland. 2008. A tool for deterministic and probabilistic sensitivity analysis of epidemiologic studies. *Stata Journal* 8: 29–48.

Pearce, M. S., and R. Feltbower. 2000. stburlstb56.pdfsg149: Tests for seasonal data via the Edwards and Walter & Elwood tests. *Stata Technical Bulletin* 56: 47–49. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 214–217. College Station, TX: Stata Press.

Reilly, M., and A. Salim. 2000. sxd2: Computing optimal sampling designs for two-stage studies. *Stata Technical Bulletin* 58: 37–41. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 376–382. College Station, TX: Stata Press.

Robins, J. M., N. E. Breslow, and S. Greenland. 1986. Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* 42: 311–323.

Rothman, K. J. 1982. Spermicide use and Down's syndrome. *American Journal of Public Health* 72: 399–401.

——. 1986. *Modern Epidemiology*. Boston: Little, Brown.

——. 2012. *Epidemiology: An Introduction*. 2nd ed. New York: Oxford University Press.

Rothman, K. J., and J. D. Boice, Jr. 1982. *Epidemiologic Analysis with a Programmable Calculator*. Brookline, MA: Epidemiology Resources.

Rothman, K. J., D. C. Fyler, A. Goldblatt, and M. B. Kreidberg. 1979. Exogenous hormones and other drug exposures of children with congenital heart disease. *American Journal of Epidemiology* 109: 433–439.

Rothman, K. J., S. Greenland, and T. L. Lash. 2008. *Modern Epidemiology*. 3rd ed. Philadelphia: Lippincott Williams & Wilkins.

Rothman, K. J., and R. R. Monson. 1973. Survival in trigeminal neuralgia. *Journal of Chronic Diseases* 26: 303–309.

Royston, P., and A. G. Babiker. 2002. A menu-driven facility for complex sample size calculation in randomized controlled trials with a survival or a binary outcome. *Stata Journal* 2: 151–163.

Schlesselman, J. J. 1982. *Case–Control Studies: Design, Conduct, Analysis.* New York: Oxford University Press.

Selvin, S. 2011. *Statistical Tools for Epidemiologic Research.* New York: Oxford University Press.

Snow, J. 1855. *On the Mode of Communication of Cholera.* 2nd ed. London: Churchill.

Tarone, R. E. 1985. On heterogeneity tests based on efficient scores. *Biometrika* 72: 91–95.

University Group Diabetes Program. 1970. A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes, II: Mortality results. *Diabetes* 19, supplement 2: 789–830.

Vinten-Johansen, P., H. Brody, N. Paneth, S. Rachman, and M. Rip. 2003. *Cholera, Chloroform, and the Science of Medicine: A Life of John Snow.* New York: Oxford University Press.

Walker, A. M. 1991. *Observation and Inference: An Introduction to the Methods of Epidemiology.* Newton Lower Falls, MA: Epidemiology Resources.

Wang, Z. 1999. sbe27: Assessing confounding effects in epidemiological studies. *Stata Technical Bulletin* 49: 12–15. Reprinted in *Stata Technical Bulletin Reprints*, vol. 9, pp. 134–138. College Station, TX: Stata Press.

——. 2007. Two postestimation commands for assessing confounding effects in epidemiological studies. *Stata Journal* 7: 183–196.

Woodward, M. 2014. *Epidemiology: Study Design and Data Analysis.* 3rd ed. Boca Raton, FL: Chapman & Hall/CRC.

Woolf, B. 1955. On estimating the relation between blood group disease. *Annals of Human Genetics* 19: 251–253. Reprinted in *Evolution of Epidemiologic Ideas: Annotated Readings on Concepts and Methods,* ed. S. Greenland, pp. 108–110. Newton Lower Falls, MA: Epidemiology Resources.

Also see