Title

---

ci — Confidence intervals for means, proportions, and counts

Syntax

Syntax for ci

> ci [*varlist*] [*if*] [*in*] [*weight*] [, *options*]

Immediate command for variable distributed as normal

> cii $\#_{\text{obs}}$ $\#_{\text{mean}}$ $\#_{\text{sd}}$ [, *ciin_option*]

Immediate command for variable distributed as binomial

> cii $\#_{\text{obs}}$ $\#_{\text{succ}}$ [, *ciib_options*]

Immediate command for variable distributed as Poisson

> cii $\#_{\text{exposure}}$ $\#_{\text{events}}$, poisson [*ciip_options*]

options	Description
Main	
<u>binomial</u>	binomial 0/1 variables; compute exact confidence intervals
<u>poisson</u>	Poisson variables; compute exact confidence intervals
<u>exposure</u>(*varname*)	exposure variable; implies poisson
<u>exact</u>	calculate exact confidence intervals; the default
<u>wald</u>	calculate Wald confidence intervals
<u>wilson</u>	calculate Wilson confidence intervals
<u>agresti</u>	calculate Agresti–Coull confidence intervals
<u>jeffreys</u>	calculate Jeffreys confidence intervals
<u>total</u>	add output for all groups combined (for use with by only)
<u>separator</u>(*#*)	draw separator line after every *#* variables; default is separator(5)
<u>level</u>(*#*)	set confidence level; default is level(95)

by is allowed with ci; see [D] **by**.

aweights and fweights are allowed, but aweights may not be specified with the binomial or poisson options; see [U] **11.1.6 weight**.

ciin_option	Description
<u>level</u>(*#*)	set confidence level; default is level(95)

1

ciib_options	Description
<u>l</u>evel(*#*)	set confidence level; default is level(95)
exact	calculate exact confidence intervals; the default
<u>wa</u>ld	calculate Wald confidence intervals
wilson	calculate Wilson confidence intervals
agresti	calculate Agresti–Coull confidence intervals
<u>j</u>effreys	calculate Jeffreys confidence intervals

ciip_options	Description
*poisson	numbers are Poisson-distributed counts
<u>l</u>evel(*#*)	set confidence level; default is level(95)

*poisson is required.

Menu

ci

Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Confidence intervals

cii for variable distributed as normal

Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Normal CI calculator

cii for variable distributed as binomial

Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Binomial CI calculator

cii for variable distributed as Poisson

Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Poisson CI calculator

Description

ci computes standard errors and confidence intervals for each of the variables in *varlist*.

cii is the immediate form of ci; see [U] **19 Immediate commands** for a general discussion of immediate commands.

In the binomial and Poisson variants of cii, the second number specified ($\#_{\text{succ}}$ or $\#_{\text{events}}$) must be an integer or between 0 and 1. If the number is between 0 and 1, Stata interprets it as the fraction of successes or events and converts it to an integer number representing the number of successes or events. The computation then proceeds as if two integers had been specified.

Options

> [Main]

binomial tells ci that the variables are 0/1 variables and that binomial confidence intervals will be calculated. (cii produces binomial confidence intervals when only two numbers are specified.)

poisson specifies that the variables (or numbers for cii) are Poisson-distributed counts; exact Poisson confidence intervals will be calculated.

exposure(*varname*) is used only with poisson. You do not need to specify poisson if you specify exposure(); poisson is assumed. *varname* contains the total exposure (typically a time or an area) during which the number of events recorded in *varlist* were observed.

exact, wald, wilson, agresti, and jeffreys specify that variables are 0/1 and specify how binomial confidence intervals are to be calculated.

exact is the default and specifies exact (also known in the literature as Clopper–Pearson [1934]) binomial confidence intervals.

wald specifies calculation of Wald confidence intervals.

wilson specifies calculation of Wilson confidence intervals.

agresti specifies calculation of Agresti–Coull confidence intervals.

jeffreys specifies calculation of Jeffreys confidence intervals.

See Brown, Cai, and DasGupta (2001) for a discussion and comparison of the different binomial confidence intervals.

total is for use with the by prefix. It requests that, in addition to output for each by-group, output be added for all groups combined.

separator(#) specifies how often separation lines should be inserted into the output. The default is separator(5), meaning that a line is drawn after every five variables. separator(10) would draw the line after every 10 variables. separator(0) suppresses the separation line.

level(#) specifies the confidence level, as a percentage, for confidence intervals. The default is level(95) or as set by set level; see [R] **level**.

Remarks and examples

stata.com

Remarks are presented under the following headings:

Ordinary confidence intervals

▷ Example 1

Without the binomial or poisson options, ci produces "ordinary" confidence intervals, meaning those that are correct if the variable is distributed normally, and *asymptotically* correct for all other distributions satisfying the conditions of the central limit theorem.

```
. use http://www.stata-press.com/data/r13/auto
(1978 Automobile Data)

. ci mpg price
```

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]	
mpg	74	21.2973	.6725511	19.9569	22.63769
price	74	6165.257	342.8719	5481.914	6848.6

The standard error of the mean of mpg is 0.67, and the 95% confidence interval is $[19.96, 22.64]$. We can obtain wider confidence intervals, 99%, by typing

```
. ci mpg price, level(99)
```

Variable	Obs	Mean	Std. Err.	[99% Conf. Interval]	
mpg	74	21.2973	.6725511	19.51849	23.07611
price	74	6165.257	342.8719	5258.405	7072.108

◁

▷ Example 2

by() breaks out the confidence intervals according to by-group; total adds an overall summary. For instance,

```
. ci mpg, by(foreign) total
```

-> foreign = Domestic

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]	
mpg	52	19.82692	.657777	18.50638	21.14747

-> foreign = Foreign

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]	
mpg	22	24.77273	1.40951	21.84149	27.70396

-> Total

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]	
mpg	74	21.2973	.6725511	19.9569	22.63769

◁

❏ Technical note

You can control the formatting of the numbers in the output by specifying a display format for the variable; see [U] **12.5 Formats: Controlling how data are displayed**. For instance,

```
. format mpg %9.2f
. ci mpg
```

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]	
mpg	74	21.30	0.67	19.96	22.64

❏

Binomial confidence intervals

▷ Example 3

We have data on employees, including a variable marking whether the employee was promoted last year.

```
. use http://www.stata-press.com/data/r13/promo
. ci promoted, binomial
```

Variable	Obs	Mean	Std. Err.	— Binomial Exact — [95% Conf. Interval]	
promoted	20	.1	.067082	.0123485	.3169827

The above interval is the default for binomial data, known equivalently as both the exact binomial and the Clopper–Pearson interval.

Nominally, the interpretation of a 95% confidence interval is that under repeated samples or experiments, 95% of the resultant intervals would contain the unknown parameter in question. However, for binomial data, the actual coverage probability, regardless of method, usually differs from that interpretation. This result occurs because of the discreteness of the binomial distribution, which produces only a finite set of outcomes, meaning that coverage probabilities are subject to discrete jumps and the exact nominal level cannot always be achieved. Therefore, the term *exact confidence interval* refers to its being derived from the binomial distribution, the distribution exactly generating the data, rather than resulting in exactly the nominal coverage.

For the Clopper–Pearson interval, the actual coverage probability is guaranteed to be greater than or equal to the nominal confidence level, here 95%. Because of the way it is calculated—see *Methods and formulas*—it may also be interpreted as follows: If the true probability of being promoted were 0.012, the chances of observing a result as extreme or more extreme than the result observed ($20 \times 0.1 = 2$ or more promotions) would be 2.5%. If the true probability of being promoted were 0.317, the chances of observing a result as extreme or more extreme than the result observed (two or fewer promotions) would be 2.5%.

◁

▷ Example 4

The Clopper–Pearson interval is desirable because it guarantees nominal coverage; however, by dropping this restriction, you may obtain accurate intervals that are not as conservative. In this vein, you might opt for the Wilson (1927) interval,

```
. ci promoted, binomial wilson
```

Variable	Obs	Mean	Std. Err.	——— Wilson ——— [95% Conf. Interval]	
promoted	20	.1	.067082	.0278665	.3010336

the Agresti–Coull (1998) interval,

```
. ci promoted, binomial agresti
```

Variable	Obs	Mean	Std. Err.	— Agresti-Coull — [95% Conf. Interval]	
promoted	20	.1	.067082	.0156562	.3132439

or the Bayesian-derived Jeffreys interval (Brown, Cai, and DasGupta 2001),

```
. ci promoted, binomial jeffreys
```

Variable	Obs	Mean	Std. Err.	——— Jeffreys ——— [95% Conf. Interval]	
promoted	20	.1	.067082	.0213725	.2838533

Picking the best interval is a matter of balancing accuracy (coverage) against precision (average interval length) and depends on sample size and success probability. Brown, Cai, and DasGupta (2001) recommend the Wilson or Jeffreys interval for small sample sizes (≤ 40) yet favor the Agresti–Coull interval for its simplicity, decent performance for sample sizes less than or equal to 40, and performance comparable to Wilson/Jeffreys for sample sizes greater than 40. They also deem the Clopper–Pearson interval to be "wastefully conservative and [. . .] not a good choice for practical use", unless of course one requires, at a minimum, the nominal coverage level.

◁

Finally, the binomial Wald confidence interval is obtained by specifying the `binomial` and `wald` options. The Wald interval is the one taught in most introductory statistics courses and for the above is simply, for level $1 - \alpha$, `Mean`$\pm z_\alpha$(`Std. Err.`), where z_α is the $1 - \alpha/2$ quantile of the standard normal. Because its overall poor performance makes it impractical, the Wald interval is available mainly for pedagogical purposes. The binomial Wald interval is also similar to the interval produced by treating binary data as normal data and using `ci` without the `binomial` option, with two exceptions. First, when `binomial` is specified, the calculation of the standard error uses denominator n rather than $n - 1$, used for normal data. Second, confidence intervals for normal data are based on the t distribution rather than the standard normal. Of course, both discrepancies vanish as sample size increases.

❑ Technical note

Let's repeat example 3, but this time with data in which there are no promotions over the observed period:

```
. use http://www.stata-press.com/data/r13/promonone
. ci promoted, binomial
```

| | | | | — Binomial Exact — |
Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]
promoted	20	0	0	0 .1684335*

(*) one-sided, 97.5% confidence interval

The confidence interval is $[0, 0.168]$, and this is the confidence interval that most books publish. It is not, however, a true 95% confidence interval because the lower tail has vanished. As Stata notes, it is a one-sided, 97.5% confidence interval. If you wanted to put 5% in the right tail, you could type `ci promoted, binomial level(90)`.

❑

❑ Technical note

`ci` with the `binomial` option ignores any variables that do not take on the values 0 and 1 exclusively. For instance, with our automobile dataset,

```
. use http://www.stata-press.com/data/r13/auto
(1978 Automobile Data)
. ci mpg foreign, binomial
```

| | | | | — Binomial Exact — |
Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]
foreign	74	.2972973	.0531331	.196584 .4148353

We also requested the confidence interval for `mpg`, but Stata ignored us. It does that so you can type `ci, binomial` and obtain correct confidence intervals for all the variables that are 0/1 in your data.

❑

Poisson confidence intervals

▷ Example 5

We have data on the number of bacterial colonies on a Petri dish. The dish has been divided into 36 small squares, and the number of colonies in each square has been counted. Each observation in our dataset represents a square on the dish. The variable `count` records the number of colonies in each square counted, which varies from 0 to 5.

```
. use http://www.stata-press.com/data/r13/petri
. ci count, poisson
```

				— Poisson	Exact —
Variable	Exposure	Mean	Std. Err.	[95% Conf.	Interval]
count	36	2.333333	.2545875	1.861158	2.888825

`ci` reports that the average number of colonies per square is 2.33. If the expected number of colonies per square were as low as 1.86, the probability of observing 2.33 or more colonies per square would be 2.5%. If the expected number were as large as 2.89, the probability of observing 2.33 or fewer colonies per square would be 2.5%.

◁

❑ Technical note

The number of "observations"—how finely the Petri dish is divided—makes no difference. The Poisson distribution is a function only of the count. In example 4, we observed a total of $2.33 \times 36 = 84$ colonies and a confidence interval of $[1.86 \times 36, 2.89 \times 36] = [67, 104]$. We would obtain the same $[67, 104]$ confidence interval if our dish were divided into, say, 49 squares, rather than 36.

For the counts, it is not even important that all the squares be of the same size. For *rates*, however, such differences do matter, but in an easy-to-calculate way. Rates are obtained from counts by dividing by exposure, which is typically a number multiplied by either time or an area. For our Petri dishes, we divide by an area to obtain a rate, but if our example were cast in terms of being infected by a disease, we might divide by person-years to obtain the rate. Rates are convenient because they are easier to compare: we might have 2.3 colonies per square inch or 0.0005 infections per person-year.

So, let's assume that we wish to obtain the number of colonies per square inch, and, moreover, that not all the "squares" on our dish are of equal size. We have a variable called `area` that records the area of each "square":

```
. ci count, exposure(area)
```

				— Poisson	Exact —
Variable	Exposure	Mean	Std. Err.	[95% Conf.	Interval]
count	3	28	3.055051	22.3339	34.66591

The rates are now in more familiar terms. In our sample, there are 28 colonies per square inch and the 95% confidence interval is $[22.3, 34.7]$. When we did not specify `exposure()`, `ci` assumed that each observation contributed 1 to exposure.

❑

❑ Technical note

As with the binomial option, if there were no colonies on our dish, ci would calculate a one-sided confidence interval:

```
. use http://www.stata-press.com/data/r13/petrinone
. ci count, poisson
```

Variable	Exposure	Mean	Std. Err.	— Poisson Exact — [95% Conf. Interval]	
count	36	0	0	0	.1024689*

(*) one-sided, 97.5% confidence interval

❑

Immediate form

▷ Example 6

We are reading a soon-to-be-published paper by a colleague. In it is a table showing the number of observations, mean, and standard deviation of 1980 median family income for the Northeast and West. We correctly think that the paper would be much improved if it included the confidence intervals. The paper claims that for 166 cities in the Northeast, the average of median family income is $19,509 with a standard deviation of $4,379:

For the Northeast:

```
. cii 166 19509 4379
```

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]	
	166	19509	339.8763	18837.93	20180.07

For the West:

```
. cii 256 22557 5003
```

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]	
	256	22557	312.6875	21941.22	23172.78

◁

▷ Example 7

We flip a coin 10 times, and it comes up heads only once. We are shocked and decide to obtain a 99% confidence interval for this coin:

```
. cii 10 1, level(99)
```

Variable	Obs	Mean	Std. Err.	— Binomial Exact — [99% Conf. Interval]	
	10	.1	.0948683	.0005011	.5442871

◁

▷ Example 8

The number of reported traffic accidents in Santa Monica over a 24-hour period is 27. We need know nothing else:

```
. cii 1 27, poisson
```

Variable	Exposure	Mean	Std. Err.	— Poisson Exact — [95% Conf. Interval]	
	1	27	5.196152	17.79317	39.28358

◁

Video examples

Immediate commands in Stata: Confidence intervals for Poisson data

Immediate commands in Stata: Confidence intervals for binomial data

Immediate commands in Stata: Confidence intervals for normal data

Stored results

ci and cii store the following in r():

Scalars
r(N)	number of observations or exposure	r(lb)	lower bound of confidence interval
r(mean)	mean	r(ub)	upper bound of confidence interval
r(se)	estimate of standard error		

Methods and formulas

Methods and formulas are presented under the following headings:

> Ordinary
> Binomial
> Poisson

Ordinary

Define n, $\overline{x}$, and s^2 as, respectively, the number of observations, (weighted) average, and (unbiased) estimated variance of the variable in question; see [R] **summarize**.

The standard error of the mean, s_μ, is defined as $\sqrt{s^2/n}$.

Let α be $1 - l/100$, where l is the significance level specified by the user. Define t_α as the two-sided t statistic corresponding to a significance level of α with $n-1$ degrees of freedom; t_α is obtained from Stata as invttail(n-1,0.5*α). The lower and upper confidence bounds are, respectively, $\overline{x} - s_\mu t_\alpha$ and $\overline{x} + s_\mu t_\alpha$.

Binomial

Given k successes of n trials, the estimated probability is $\widehat{p} = k/n$ with standard error $\sqrt{\widehat{p}(1-\widehat{p})/n}$. ci calculates the exact (Clopper–Pearson) confidence interval $[\,p_1, p_2\,]$ such that

$$\Pr(K \geq k | p = p_1) = \alpha/2$$

and

$$\Pr(K \leq k | p = p_2) = \alpha/2$$

where K is distributed as binomial(n, p). The endpoints may be obtained directly by using Stata's invbinomial() function. If $k = 0$ or $k = n$, the calculation of the appropriate tail is skipped.

The Wald interval is $\widehat{p} \pm z_\alpha \sqrt{\widehat{p}(1-\widehat{p})/n}$, where z_α is the $1 - \alpha/2$ quantile of the standard normal. The interval is obtained by inverting the acceptance region of the large-sample Wald test of $H_0 : p = p_0$ versus the two-sided alternative. That is, the confidence interval is the set of all p_0 such that

$$\left| \frac{\widehat{p} - p_0}{\sqrt{n^{-1}\widehat{p}(1-\widehat{p})}} \right| \leq z_\alpha$$

The Wilson interval is a variation on the Wald interval, using the null standard error $\sqrt{n^{-1}p_0(1-p_0)}$ in place of the estimated standard error $\sqrt{n^{-1}\widehat{p}(1-\widehat{p})}$ in the above expression. Inverting this acceptance region is more complicated yet results in the closed form

$$\frac{k + z_\alpha^2/2}{n + z_\alpha^2} \pm \frac{z_\alpha n^{1/2}}{n + z_\alpha^2/2} \left\{ \widehat{p}(1-\widehat{p}) + \frac{z_\alpha^2}{4n} \right\}^{1/2}$$

The Agresti–Coull interval is basically a Wald interval that borrows its center from the Wilson interval. Defining $\widetilde{k} = k + z_\alpha^2/2$, $\widetilde{n} = n + z_\alpha^2$, and (hence) $\widetilde{p} = \widetilde{k}/\widetilde{n}$, the Agresti–Coull interval is

$$\widetilde{p} \pm z_\alpha \sqrt{\widetilde{p}(1-\widetilde{p})/\widetilde{n}}$$

When $\alpha = 0.05$, z_α is near enough to 2 that $\widetilde{p}$ can be thought of as a typical estimate of proportion where two successes and two failures have been added to the sample (Agresti and Coull 1998). This typical estimate of proportion makes the Agresti–Coull interval an easy-to-present alternative for introductory statistics students.

The Jeffreys interval is a Bayesian interval and is based on the Jeffreys prior, which is the Beta$(1/2, 1/2)$ distribution. Assigning this prior to p results in a posterior distribution for p that is Beta with parameters $k + 1/2$ and $n - k + 1/2$. The Jeffreys interval is then taken to be the $1 - \alpha$ central posterior probability interval, namely, the $\alpha/2$ and $1 - \alpha/2$ quantiles of the Beta$(k+1/2, n-k+1/2)$ distribution. These quantiles may be obtained directly by using Stata's invibeta() function.

Poisson

Given the total cases, k, the estimate of the expected count λ is k, and its standard error is $\sqrt{k}$. ci calculates the exact confidence interval $[\,\lambda_1, \lambda_2\,]$ such that

$$\Pr(K \geq k | \lambda = \lambda_1) = \alpha/2$$

and

$$\Pr(K \leq k | \lambda = \lambda_2) = \alpha/2$$

where K is Poisson with mean λ. Solution is by Newton's method. If $k = 0$, the calculation of λ_1 is skipped. All values are then reported as rates, which are the above numbers divided by the total exposure.

Harold Jeffreys (1891–1989) was born near Durham, England, and spent more than 75 years studying and working at the University of Cambridge, principally on theoretical and observational problems in geophysics, astronomy, mathematics, and statistics. He developed a systematic Bayesian approach to inference in his monograph *Theory of Probability*.

Edwin Bidwell (E. B.) Wilson (1879–1964) majored in mathematics at Harvard and studied and taught at Yale and MIT before returning to Harvard in 1922. He worked in mathematics, physics, and statistics. His method for binomial intervals can be considered a precursor, for a particular problem, of Neyman's concept of confidence intervals.

Jerzy Neyman (1894–1981) was born in Bendery, Russia, now Moldavia. He studied and then taught at Kharkov University, moving from physics to mathematics. In 1921, Neyman moved to Poland, where he worked in statistics at Bydgoszcz and then Warsaw. Neyman received a Rockefeller Fellowship to work with Karl Pearson at University College London. There, he collaborated with Egon Pearson, Karl's son, on the theory of hypothesis testing. Life in Poland became progressively more difficult, and Neyman returned to UCL to work there from 1934 to 1938. At this time, he published on the theory of confidence intervals. He then was offered a post in California at Berkeley, where he settled. Neyman established an outstanding statistics department and remained highly active in research, including applications in astronomy, meteorology, and medicine. He was one of the great statisticians of the 20th century.

Acknowledgment

References

Agresti, A., and B. A. Coull. 1998. Approximate is better than "exact" for interval estimation of binomial proportions. *American Statistician* 52: 119–126.

Brown, L. D., T. T. Cai, and A. DasGupta. 2001. Interval estimation for a binomial proportion. *Statistical Science* 16: 101–133.

Campbell, M. J., D. Machin, and S. J. Walters. 2007. *Medical Statistics: A Textbook for the Health Sciences*. 4th ed. Chichester, UK: Wiley.

Clopper, C. J., and E. S. Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26: 404–413.

Cook, A. 1990. Sir Harold Jeffreys. 2 April 1891–18 March 1989. *Biographical Memoirs of Fellows of the Royal Society* 36: 303–333.

Gleason, J. R. 1999. sg119: Improved confidence intervals for binomial proportions. *Stata Technical Bulletin* 52: 16–18. Reprinted in *Stata Technical Bulletin Reprints*, vol. 9, pp. 208–211. College Station, TX: Stata Press.

Jeffreys, H. 1946. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, Series A* 186: 453–461.

Lindley, D. V. 2001. Harold Jeffreys. In *Statisticians of the Centuries*, ed. C. C. Heyde and E. Seneta, 402–405. New York: Springer.

Reid, C. 1982. *Neyman—from Life*. New York: Springer.

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

Seed, P. T. 2001. sg159: Confidence intervals for correlations. *Stata Technical Bulletin* 59: 27–28. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 267–269. College Station, TX: Stata Press.

Stigler, S. M. 1997. Wilson, Edwin Bidwell. In *Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present*, ed. N. L. Johnson and S. Kotz, 344–346. New York: Wiley.

Utts, J. M. 2005. *Seeing Through Statistics*. 3rd ed. Belmont, CA: Brooks/Cole.

Wang, D. 2000. sg154: Confidence intervals for the ratio of two binomial proportions by Koopman's method. *Stata Technical Bulletin* 58: 16–19. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 244–247. College Station, TX: Stata Press.

Wilson, E. B. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 22: 209–212.

Also see