Title

stata.com

centile — Report centile and confidence interval

Description

`centile` estimates specified centiles and calculates confidence intervals. If no *varlist* is specified, `centile` calculates centiles for all the variables in the dataset. If no centiles are specified, medians are reported.

By default, `centile` uses a binomial method for obtaining confidence intervals that makes no assumptions about the underlying distribution of the variable.

Quick start

50th percentile with 95% confidence intervals for `v1` and `v2`

```
centile v1 v2
```

For all variables in the dataset

```
centile
```

25th, 50th, and 75th percentiles of `v1`

```
centile v1, centile(25 50 75)
```

10th, 20th, 30th, ..., 90th percentiles of `v1`

```
centile v1, centile(10(10)90)
```

Force confidence limits to fall on sample values

```
centile v1 v2, cci
```

Confidence intervals based on standard errors for a normal-distribution quantile

```
centile v1 v2, normal
```

Centile and confidence intervals based on mean and standard deviation

```
centile v1 v2, meansd
```

Replace data in memory with centiles for groups defined by categorical variable `cvar`

```
statsby, by(cvar) clear: centile v1, centile(25 50 75)
```

Menu

Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Centiles with CIs

Syntax

```
centile [varlist] [if] [in] [, options]
```

options	Description
Main	
`centile(`*numlist*`)`	report specified centiles; default is `centile(50)`
Options	
`cci`	binomial exact; conservative confidence interval
`normal`	normal, based on observed centiles
`meansd`	normal, based on mean and standard deviation
`level(#)`	set confidence level; default is `level(95)`

`by`, `collect`, and `statsby` are allowed; see **[U] 11.1.10 Prefix commands**.

Options

Main

`centile(`*numlist*`)` specifies the centiles to be reported. The default is to display the 50th centile. Specifying `centile(5)` requests that the fifth centile be reported. Specifying `centile(5 50 95)` requests that the 5th, 50th, and 95th centiles be reported. Specifying `centile(10(10)90)` requests that the 10th, 20th, . . . , 90th centiles be reported; see **[U] 11.1.8 numlist**.

Options

`cci` (conservative confidence interval) forces the confidence limits to fall exactly on sample values. Confidence intervals displayed with the `cci` option are slightly wider than those with the default (`nocci`) option.

`normal` causes the confidence interval to be calculated by using a formula for the standard error of a normal-distribution quantile given by Kendall and Stuart (1969, 237). The `normal` option is useful when you want empirical centiles—that is, centiles based on sample order statistics rather than on the mean and standard deviation—and are willing to assume normality.

`meansd` causes the centile and confidence interval to be calculated based on the sample mean and standard deviation, and it assumes normality.

`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

The qth centile of a continuous random variable, X, is defined as the value of C_q, which fulfills the condition $\Pr(X \leq C_q) = q/100$. The value of q must be in the range $0 < q < 100$, though q is not necessarily an integer. By default, `centile` estimates C_q for the variables in *varlist* and for the values of q given in `centile(`*numlist*`)`. It makes no assumptions about the distribution of X and, if necessary, uses linear interpolation between neighboring sample values. Extreme centiles (for example, the 99th centile in samples smaller than 100) are fixed at the minimum or maximum sample value. An "exact" confidence interval for C_q is also given, using the binomial-based method described below in *Methods and formulas* and in Conover (1999, 143–148). Again linear interpolation is used to improve the accuracy of the estimated confidence limits, but extremes are fixed at the minimum or maximum sample value.

You can prevent `centile` from interpolating when calculating binomial-based confidence intervals by specifying `cci`. The resulting intervals are generally wider than with the default; that is, the coverage (confidence level) tends to be greater than the nominal value (given as usual by `level(#)`, by default 95%).

If the data are believed to be normally distributed (a common case), there are two alternative methods for estimating centiles. If `normal` is specified, C_q is calculated, as just described, but its confidence interval is based on a formula for the standard error (se) of a normal-distribution quantile given by Kendall and Stuart (1969, 237). If `meansd` is alternatively specified, C_q is estimated as $\overline{x} + z_q \times s$, where $\overline{x}$ and s are the sample mean and standard deviation, respectively, and z_q is the qth centile of the standard normal distribution (for example, $z_{95} = 1.645$). The confidence interval is derived from the se of the estimate of C_q.

▷ Example 1

Using `auto.dta`, we estimate the 5th, 50th, and 95th centiles of the `price` variable:

```
. use https://www.stata-press.com/data/r17/auto
(1978 automobile data)
. format price %8.2fc
. centile price, centile(5 50 95)
```

Variable	Obs	Percentile	Centile	Binom. interp. [95% conf.	interval]
price	74	5	3,727.75	3,291.23	3,914.16
		50	5,006.50	4,593.57	5,717.90
		95	13,498.00	11,061.53	15,865.30

`summarize` produces somewhat different results from `centile`; see *Methods and formulas*.

```
. summarize price, detail
```

Price

Percentiles		Smallest		
1%	3291	3291		
5%	3748	3299		
10%	3895	3667	Obs	74
25%	4195	3748	Sum of wgt.	74
50%	5006.5		Mean	6165.257
		Largest	Std. dev.	2949.496
75%	6342	13466		
90%	11385	13594	Variance	8699526
95%	13466	14500	Skewness	1.653434
99%	15906	15906	Kurtosis	4.819188

The confidence limits produced by using the `cci` option are slightly wider than those produced without this option:

```
. centile price, c(5 50 95) cci
```

Variable	Obs	Percentile	Centile	Binomial exact [95% conf.	interval]
price	74	5	3,727.75	3,291.00	3,955.00
		50	5,006.50	4,589.00	5,719.00
		95	13,498.00	10,372.00	15,906.00

If we are willing to assume that price is normally distributed, we could include either the normal or the meansd option:

```
. centile price, c(5 50 95) normal
```

Variable	Obs	Percentile	Centile	— Normal, based on observed centiles — [95% conf. interval]	
price	74	5	3,727.75	3,211.19	4,244.31
		50	5,006.50	4,096.68	5,916.32
		95	13,498.00	5,426.81	21,569.19

```
. centile price, c(5 50 95) meansd
```

Variable	Obs	Percentile	Centile	— Normal, based on mean and std. dev.— [95% conf. interval]	
price	74	5	1,313.77	278.93	2,348.61
		50	6,165.26	5,493.24	6,837.27
		95	11,016.75	9,981.90	12,051.59

With the normal option, the centile estimates are, by definition, the same as before. The confidence intervals for the 5th and 50th centiles are similar to the previous ones, but the interval for the 95th centile is different. The results using the meansd option also differ from both previous sets of estimates.

We can use sktest (see [R] **sktest**) to check the correctness of the normality assumption:

```
. sktest price
Skewness and kurtosis tests for normality
```

Variable	Obs	Pr(skewness)	Pr(kurtosis)	Joint test Adj chi2(2)	Prob>chi2
price	74	0.0000	0.0127	21.77	0.0000

sktest reveals that price is definitely not normally distributed, so the normal assumption is not reasonable, and the normal and meansd options are not appropriate for these data. We should rely on the results from the default choice, which does not assume normality. If the data are normally distributed, however, the precision of the estimated centiles and their confidence intervals will be ordered (best) meansd > normal > [default] (worst). The normal option is useful when we really do want empirical centiles (that is, centiles based on sample order statistics rather than on the mean and standard deviation) but are willing to assume normality.

◁

Stored results

centile stores the following in r():

Scalars

r(N)	number of observations
r(n_cent)	number of centiles requested
r(c_#)	value of # centile
r(lb_#)	#-requested centile lower confidence bound
r(ub_#)	#-requested centile upper confidence bound

Macros

r(centiles)	centiles requested

Methods and formulas

Methods and formulas are presented under the following headings:

Default case

The calculation is based on the method of Mood and Graybill (1963, 408). Let $x_1 \leq x_2 \leq \cdots \leq x_n$ be a sample of size n arranged in ascending order. Denote the estimated qth centile of the x's as c_q. We require that $0 < q < 100$. Let $R = (n+1)q/100$ have integer part r and fractional part f; that is, $r = \text{int}(R)$ and $f = R - r$. (If R is itself an integer, then $r = R$ and $f = 0$.) Note that $0 \leq r \leq n$. For convenience, define $x_0 = x_1$ and $x_{n+1} = x_n$. C_q is estimated by

$$c_q = x_r + f \times (x_{r+1} - x_r)$$

that is, c_q is a weighted average of x_r and x_{r+1}. Loosely speaking, a (conservative) $p\%$ confidence interval for C_q involves finding the observations ranked t and u, which correspond, respectively, to the $\alpha = (100-p)/200$ and $1-\alpha$ quantiles of a binomial distribution with parameters n and $q/100$, that is, $\text{B}(n, q/100)$. More precisely, define the ith value ($i = 0, \ldots, n$) of the cumulative binomial distribution function as $F_i = \Pr(S \leq i)$, where S has distribution $\text{B}(n, q/100)$. For convenience, let $F_{-1} = 0$ and $F_{n+1} = 1$. t is found such that $F_t \leq \alpha$ and $F_{t+1} > \alpha$, and u is found such that $1 - F_u \leq \alpha$ and $1 - F_{u-1} > \alpha$.

With the `cci` option in force, the (conservative) confidence interval is (x_{t+1}, x_{u+1}), and its actual coverage probability is $F_u - F_t$.

The default case uses linear interpolation on the F_i as follows. Let

$$\begin{aligned} g &= (\alpha - F_t)/(F_{t+1} - F_t) \\ h &= \{\alpha - (1 - F_u)\}/\{(1 - F_{u-1}) - (1 - F_u)\} \\ &= (\alpha - 1 + F_u)/(F_u - F_{u-1}) \end{aligned}$$

The interpolated lower and upper confidence limits (c_{qL}, c_{qU}) for C_q are

$$\begin{aligned} c_{qL} &= x_{t+1} + g \times (x_{t+2} - x_{t+1}) \\ c_{qU} &= x_{u+1} - h \times (x_{u+1} - x_u) \end{aligned}$$

Suppose that we want a 95% confidence interval for the median of a sample of size 13. $n = 13$, $q = 50$, $p = 95$, $\alpha = 0.025$, $R = 14 \times 50/100 = 7$, and $f = 0$. Therefore, the median is the 7th observation. Some example data, x_i, and the values of F_i are as follows:

i	F_i	$1 - F_i$	x_i	i	F_i	$1 - F_i$	x_i
0	0.0001	0.9999	–	7	0.7095	0.2905	33
1	0.0017	0.9983	5	8	0.8666	0.1334	37
2	0.0112	0.9888	7	9	0.9539	0.0461	45
3	0.0461	0.9539	10	10	0.9888	0.0112	59
4	0.1334	0.8666	15	11	0.9983	0.0017	77
5	0.2905	0.7095	23	12	0.9999	0.0001	104
6	0.5000	0.5000	28	13	1.0000	0.0000	211

The median is $x_7 = 33$. Also, $F_2 \leq 0.025$ and $F_3 > 0.025$, so $t = 2$; $1 - F_{10} \leq 0.025$ and $1 - F_9 > 0.025$, so $u = 10$. The conservative confidence interval is therefore

$$(c_{50L}, c_{50U}) = (x_{t+1}, x_{u+1}) = (x_3, x_{11}) = (10, 77)$$

with actual coverage $F_{10} - F_2 = 0.9888 - 0.0112 = 0.9776$ (97.8% confidence). For the interpolation calculation, we have

$$g = (0.025 - 0.0112)/(0.0461 - 0.0112) = 0.395$$
$$h = (0.025 - 1 + 0.9888)/(0.9888 - 0.9539) = 0.395$$

So,

$$c_{50L} = x_3 + 0.395 \times (x_4 - x_3) = 10 + 0.395 \times 5 = 11.98$$
$$c_{50U} = x_{11} - 0.395 \times (x_{11} - x_{10}) = 77 - 0.395 \times 18 = 69.89$$

Normal case

The value of c_q is as above. Its se is given by the formula

$$s_q = \sqrt{q(100-q)} \Big/ \left\{ 100\sqrt{n} Z(c_q; \overline{x}, s) \right\}$$

where $\overline{x}$ and s are the mean and standard deviation of the x_i, and

$$Z(Y; \mu, \sigma) = \left(1/\sqrt{2\pi\sigma^2}\right) e^{-(Y-\mu)^2/2\sigma^2}$$

is the density function of a normally distributed variable Y with mean μ and standard deviation σ. The confidence interval for C_q is $(c_q - z_{100(1-\alpha)} s_q, c_q + z_{100(1-\alpha)} s_q)$.

meansd case

The value of c_q is $\overline{x} + z_q \times s$. Its se is given by the formula

$$s_q^\star = s\sqrt{1/n + z_q^2/(2n-2)}$$

The confidence interval for C_q is $(c_q - z_{100(1-\alpha)} \times s_q^\star, c_q + z_{100(1-\alpha)} \times s_q^\star)$.

Acknowledgment

`centile` was written by Patrick Royston of the MRC Clinical Trials Unit, London, and coauthor of the Stata Press book *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*.

References

Conover, W. J. 1999. *Practical Nonparametric Statistics*. 3rd ed. New York: Wiley.

Kendall, M. G., and A. Stuart. 1969. *The Advanced Theory of Statistics, Vol. 1: Distribution Theory*. 3rd ed. London: Griffin.

Mood, A. M., and F. A. Graybill. 1963. *Introduction to the Theory of Statistics*. 2nd ed. New York: McGraw–Hill.

Stuart, A., and J. K. Ord. 1994. *Kendall's Advanced Theory of Statistics: Distribution Theory, Vol I*. 6th ed. London: Arnold.

Also see

[R] **ci** — Confidence intervals for means, proportions, and variances

[R] **summarize** — Summary statistics

[D] **pctile** — Create variable containing percentiles