quantile( ) — Sample quantiles+

+This function is part of StataNow.

Description

`quantile(`*X*, *p*, *method*`)` returns the quantiles specified in vector p for data matrix X. The elements of vector p must be real numbers between 0 and 1 (inclusive).

method specifies the quantile calculation method to use. The available methods are based on Hyndman and Fan (1996) and Cunnane (1978) and include both methods that rely on discontinuous functions and methods that rely on continuous piecewise linear functions.

Syntax

real matrix `quantile(`*real matrix X* , *real colvector p* [, *string scalar method*]`)`

method specifies the type of quantile estimator to use and may be one of `"inv_cdf"`, `"avg_inv_cdf"`, `"closest"`, `"parzen"`, `"hazen"`, `"weibull"`, `"gumbel"`, `"tukey"`, `"blom"`, `"beard"`, `"benard"`, `"cooper"`, `"california"`, or `"gringorten"`. See *Methods and formulas* for more details. The default is `"tukey"`. The `"avg_inv_cdf"` method is the same method used by the Stata `pctile` and `_pctile` commands. The `"weibull"` method is the method used by `pctile` and `_pctile` with the `altdef` option.

Remarks and examples

We begin by creating a small dataset for demonstration:

```
. quietly set obs 5
. mata: x = (10 \ 5 \ 5 \ 1 \ 20)
. mata: st_store(., st_addvar("double", "v1"), x)
```

This creates the `v1` variable containing the values 1, 5, 5, 10, and 20.

▷ Example 1: Correspondence with `_pctile`'s default method

Here is the default method for computing percentiles by using the Stata `_pctile` command:

```
. _pctile v1, percentiles(20 50 60 95)
. return list
scalars:
                  r(r1) =  3
                  r(r2) =  5
                  r(r3) =  7.5
                  r(r4) =  20
```

Method "avg_inv_cdf" of function quantile() produces identical results:

```
. mata: quantile(x, (0.2 \ 0.5 \ 0.6 \ 0.95), "avg_inv_cdf")
         1
    ┌───────┐
  1 │     3 │
  2 │     5 │
  3 │   7.5 │
  4 │    20 │
    └───────┘
```

◁

▷ Example 2: Correspondence with _pctile's alternative method

_pctile provides an alternative definition via the altdef option. This corresponds to quantile()'s "weibull" method:

```
. _pctile v1, percentiles(20 50 60 95) altdef
. return list
scalars:
                  r(r1) =  1.8
                  r(r2) =  5
                  r(r3) =  8
                  r(r4) =  20
. mata: quantile(x, (0.2 \ 0.5 \ 0.6 \ 0.95), "weibull")
         1
    ┌───────┐
  1 │   1.8 │
  2 │     5 │
  3 │     8 │
  4 │    20 │
    └───────┘
```

◁

Methods and formulas

The theoretical quantile of a distribution with cumulative distribution function F is defined as

$$Q(p) = F^{-1}(p) = \inf\{x : F(x) \geq p\} \qquad 0 < p < 1$$

For a sample of observations $X_1, X_2, \ldots, X_n$ with order statistics $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, the sample quantile is computed as

$$\widehat{Q}(p) = (1-\gamma)X_{(j)} + \gamma X_{(j+1)}$$

where j and γ are determined by the specific method chosen.

Methods and formulas are presented under the following headings:

Discontinuous methods
Piecewise linear continuous methods

Discontinuous methods

For these methods, $j = \lfloor pn + m \rfloor$, where m is a method-specific constant that satisfies $(j - m)/n \leq p < (j - m + 1)/n$ and γ equals a method-specific function of $g = pn + m - j$.

The table below presents the values for m and γ for each of these methods. We refer to Hyndman and Fan (1996) as H&F for brevity:

Method	m	γ determination	Reference
`"inv_cdf"`	0	$\gamma = 1$ if $g > 0$; $\gamma = 0$ otherwise	H&F def. 1
`"avg_inv_cdf"`	0	$\gamma = 1$ if $g > 0$; $\gamma = 0.5$ otherwise	H&F def. 2
`"closest"`	$-1/2$	$\gamma = 0$ if $g = 0$ and j is even; $\gamma = 1$ otherwise	H&F def. 3

Piecewise linear continuous methods

These methods are based on the concept of plotting positions introduced by Blom (1958). Each ordered observation $X_{(k)}$ is assigned a "virtual" plotting position p_k defined by

$$p_k = \frac{k - \alpha}{n - \alpha - \beta + 1}$$

where $k = 1, 2, \ldots, n$ indexes the ordered observations and α and β are parameters that characterize each specific method.

Quantiles are estimated by interpolating between the points $(p_k, X_{(k)})$. This is equivalent to setting $m = \alpha + p(1 - \alpha - \beta)$ and $\gamma = g$.

Different choices of α and β lead to different quantile estimation methods. Cunnane (1978) reviewed unbiased plotting positions for various theoretical probability distributions and proposed using $\alpha = \beta$ for several distributions.

The following table presents the parameter values and their proponents:

Method	α	β	Reference
`"parzen"`	0	1	H&F def. 4; Parzen (1979)
`"hazen"`	1/2	1/2	H&F def. 5; Hazen (1914)
`"weibull"`	0	0	H&F def. 6; Weibull (1939)
`"gumbel"`	1	1	H&F def. 7; Gumbel (1939)
`"tukey"`	1/3	1/3	H&F def. 8; Tukey (1992)
`"blom"`	3/8	3/8	H&F def. 9; Blom (1958)
`"california"`	1	0	State of California, Department of Public Works (1923)
`"beard"`	0.31	0.31	Beard (1943)
`"benard"`	0.3	0.3	Benard and Bos-Levenbach (1953)
`"cooper"`	0.4075	0.4075	Cooper (2005)
`"gringorten"`	0.44	0.44	Gringorten (1963)

Conformability

`quantile(`X, p, *method*`)`:

X:	$n \times k$
p:	$r \times 1$
method:	1×1
result:	$r \times k$

Diagnostics

`quantile(`X, p, *method*`)` aborts with error if

- any element of p is not in the range $[0, 1]$,
- any element of p is missing,
- *method* is not one of the recognized method strings (case sensitive), or
- X has zero rows (is void).

References

Beard, L. R. 1943. Statistical analysis in hydrology. *Transactions of the American Society of Civil Engineers* 108: 1110–1121. https://doi.org/10.1061/TACEAT.0005568.

Benard, A., and E. C. Bos-Levenbach. 1953. Het uitzetten van waarnemingen op waarschijnlijkheids-papier. *Statistica Neerandica* 7: 163–173. https://doi.org/10.1111/j.1467-9574.1953.tb00821.x.

Blom, G. 1958. *Statistical Estimates and Transformed Beta-Variables*. New York: Wiley.

Cooper, R. M. 2005. *Estimation of Peak Discharges for Rural, Unregulated Streams in Western Oregon: Scientific Investigations Report 2005–5116*. Denver, CO: US Geological Survey Information Services.

Cunnane, C. 1978. Unbiased plotting positions—a review. *Journal of Hydrology* 37: 205–222. https://doi.org/10.1016/0022-1694(78)90017-3.

Gringorten, I. I. 1963. A plotting rule for extreme probability paper. *Journal of Geophysical Research* 68: 813–814. https://doi.org/10.1029/JZ068i003p00813.

Gumbel, E. J. 1939. La probabilité des hypothèses. *Comptes Rendus de l'Académie des Sciences (Paris)* 209: 645–647.

Hazen, A. 1914. Storage to be provided in impounding municipal water supply. *Transactions of the American Society of Civil Engineers* 77: 1539–1640. https://doi.org/10.1061/taceat.0002563.

Hyndman, R. J., and Y. Fan. 1996. Sample quantiles in statistical packages. *American Statistician* 50: 361–365. https://doi.org/10.2307/2684934.

Parzen, E. 1979. Nonparametric statistical data modeling. *Journal of the American Statistical Association* 74: 105–121. https://doi.org/10.2307/2286734.

State of California, Department of Public Works. 1923. "Flow in California streams". Report to the Legislature of 1923 on the Water Resources of California Bulletin no. 5.

Tukey, J. W. 1992. "The future of data analysis". In *Methodology and Distribution*. Breakthroughs in Statistics, edited by S. Kotz and N. L. Johnson, vol. 2: 404–452. New York: Springer. https://doi.org/10.1007/978-1-4612-4380-9_31.

Weibull, W. 1939. "The phenomenon of rupture in solids". In *Ingeniörs Vetenskaps Akademien Handlingar*, vol. 153. Stockholm: Generalstabens Litografiska Anstalts Förlag.

Also see

[M-4] **Statistical** — Statistical functions

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