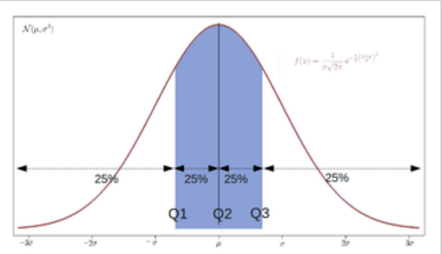
Quantile

From Wikipedia, the free encyclopedia



Probability density of a normal distribution, with quartiles shown. The area below the red curve is the same in the intervals (-∞,*Q1*), (*Q1*,*Q2*), (*Q2*,*Q3*), and (*Q3*,+∞).

In statistics and the theory of probability, **quantiles** are cutpoints dividing the range of a probability distribution into contiguous intervals with equal probabilities, or dividing the observations in a sample in the same way. There is one less quantile than the number of groups created. Thus quartiles are the three cut points that will divide a dataset into four equal-size groups (cf. depicted example). Common quantiles have special names: for instance quartile, decile (creating 10 groups: see below for more). The groups created are termed halves, thirds, quarters, etc., though sometimes the terms for the quantile are used for the groups created, rather than for the cut points.

q-**Quantiles** are values that partition a finite set of values into q subsets of (nearly) equal sizes. There are $q-1$ of the q-quantiles, one for each integer k satisfying $0 < k < q$. In some cases the value of a quantile may not be uniquely determined, as can be the case for the median (2-quantile) of a uniform probability distribution on a set of even size. Quantiles can also be applied to continuous distributions, providing a way to generalize rank statistics to continuous variables. When the cumulative distribution function of a random variable is known, the q-quantiles are the application of the quantile function (the inverse function of the cumulative distribution function) to the values $\{1/q, 2/q, \dots, (q-1)/q\}$.

Contents

Specialized quantiles

Some q-quantiles have special names:

- The only 2-quantile is called the median
- The 3-quantiles are called tertiles or terciles → T
- The 4-quantiles are called quartiles → Q; the difference between upper and lower quartiles is also called the interquartile range, **midspread** or **middle fifty** → IQR = $Q_3 - Q_1$
- The 5-quantiles are called quintiles → QU
- The 6-quantiles are called sextiles → S
- The 7-quantiles are called septiles

- The 8-quantiles are called octiles → O
- The 10-quantiles are called deciles → D
- The 12-quantiles are called duo-deciles → Dd
- The 16-quantiles are called hexadeciles → H
- The 20-quantiles are called ventiles or vigintiles→ V
- The 33-quantiles are called trigintatreciles → TT
- The 100-quantiles are called percentiles → P
- The 1000-quantiles are called permilles → Pr

Quantiles of a population

As in the computation of, for example, standard deviation, the estimation of a quantile depends upon whether one is operating with a statistical population or with a sample drawn from it. For a population, of discrete values or for a continuous population density, the k-th q-quantile is the data value where the cumulative distribution function crosses k/q. That is, x is a k-th q-quantile for a variable X if

$$\Pr[X < x] \le k/q \text{ or, equivalently, } \Pr[X \ge x] \ge 1 - k/q$$

and

$$\Pr[X \le x] \ge k/q \text{ or, equivalently, } \Pr[X > x] \le 1 - k/q.$$

For a finite population of N equally probable values indexed $1, \ldots, N$ from lowest to highest, the k-th q-quantile of this population can equivalently be computed via the value of $I_p = N\,k/q$. If I_p is not an integer, then round up to the next integer to get the appropriate index; the corresponding data value is the k-th q-quantile. On the other hand, if I_p is an integer then any number from the data value at that index to the data value of the next can be taken as the quantile, and it is conventional (though arbitrary) to take the average of those two values (see Estimating the quantiles).

If, instead of using integers k and q, the “p-quantile” is based on a real number p with $0 < p < 1$ then p replaces k/q in the above formulae. Some software programs (including Microsoft Excel) regard the minimum and maximum as the 0th and 100th percentile, respectively; however, such terminology is an extension beyond traditional statistics definitions.

Examples

The following two examples use the Nearest Rank definition of quantile with rounding. For an explanation of this definition, see percentiles.

Even-sized population

Consider an ordered population of 10 data values $\{3, 6, 7, 8, 8, 10, 13, 15, 16, 20\}$. What are the 4-quantiles (the "quartiles") of this dataset?

Quartile	Calculation	Result
Zeroth quartile	Although not universally accepted, one can also speak of the zeroth quartile. This is the minimum value of the set, so the zeroth quartile in this example would be 3.	3
First quartile	The rank of the first quartile is 10×(1/4) = 2.5, which rounds up to 3, meaning that 3 is the rank in the population (from least to greatest values) at which approximately 1/4 of the values are less than the value of the first quartile. The third value in the population is 7.	7
Second quartile	The rank of the second quartile (same as the median) is 10×(2/4) = 5, which is an integer, while the number of values (10) is an even number, so the average of both the fifth and sixth values is taken—that is (8+10)/2 = 9, though any value from 8 through to 10 could be taken to be the median.	9
Third quartile	The rank of the third quartile is 10×(3/4) = 7.5, which rounds up to 8. The eighth value in the population is 15.	15
Fourth quartile	Although not universally accepted, one can also speak of the fourth quartile. This is the maximum value of the set, so the fourth quartile in this example would be 20. Under the Nearest Rank definition of quantile, the rank of the fourth quartile is the rank of the biggest number, so the rank of the fourth quartile would be 10.	20

So the first, second and third 4-quantiles (the "quartiles") of the dataset {3, 6, 7, 8, 8, 10, 13, 15, 16, 20} are {7, 9, 15}. If also required, the zeroth quartile is 3 and the fourth quartile is 20.

Odd-sized population

Consider an ordered population of 11 data values {3, 6, 7, 8, 8, 9, 10, 13, 15, 16, 20}. What are the 4-quantiles (the "quartiles") of this dataset?

Quartile	Calculation	Result
Zeroth quartile	Although not universally accepted, one can also speak of the zeroth quartile. This is the minimum value of the set, so the zeroth quartile in this example would be 3.	3
First quartile	The first quartile is determined by 11×(1/4) = 2.75, which rounds up to 3, meaning that 3 is the rank in the population (from least to greatest values) at which approximately 1/4 of the values are less than the value of the first quartile. The third value in the population is 7.	7
Second quartile	The second quartile value (same as the median) is determined by 11×(2/4) = 5.5, which rounds up to 6. Therefore, 6 is the rank in the population (from least to greatest values) at which approximately 2/4 of the values are less than the value of the second quartile (or median). The sixth value in the population is 9.	9
Third quartile	The third quartile value for the original example above is determined by 11×(3/4) = 8.25, which rounds up to 9. The ninth value in the population is 15.	15
Fourth quartile	Although not universally accepted, one can also speak of the fourth quartile. This is the maximum value of the set, so the fourth quartile in this example would be 20. Under the Nearest Rank definition of quantile, the rank of the fourth quartile is the rank of the biggest number, so the rank of the fourth quartile would be 11.	20

So the first, second and third 4-quantiles (the "quartiles") of the dataset {3, 6, 7, 8, 8, 9, 10, 13, 15, 16, 20} are {7, 9, 15}. If also required, the zeroth quartile is 3 and the fourth quartile is 20.

Estimating quantiles from a sample

When one has a sample drawn from an unknown population, the cumulative distribution function and quantile function of the underlying population are not known and the task becomes that of estimating the quantiles. There are several methods.[1] Mathematica,[2] Matlab,[3] R[4] and GNU Octave[5] programming languages include nine sample quantile methods. SAS includes five sample quantile methods, SciPy[6] and Maple[7] both include eight, EViews[8] includes the six piecewise linear functions, STATA includes two, and Microsoft Excel includes one. Mathematica supports an arbitrary parameter for methods that allows for other, non-standard, methods.

In effect, the methods compute Q_p, the estimate for the k-th q-quantile, where $p = k/q$, from a sample of size N by computing a real valued index h. When h is an integer, the h-th smallest of the N values, x_h, is the quantile estimate. Otherwise a rounding or interpolation scheme is used to compute the quantile estimate from h, $x_{\lfloor h \rfloor}$, and $x_{\lceil h \rceil}$. (For notation, see floor and ceiling functions).

The estimate types and interpolation schemes used include:

Type	***h***	$\boldsymbol{Q_p}$	**Notes**
R-1, SAS-3, Maple-1	$Np + 1/2$	$x_{\lceil h - 1/2 \rceil}$	Inverse of empirical distribution function. When $p = 0$, use x_1.
R-2, SAS-5, Maple-2	$Np + 1/2$	$(x_{\lceil h - 1/2 \rceil} + x_{\lceil h + 1/2 \rceil}) / 2$	The same as R-1, but with averaging at discontinuities. When $p = 0$, use x_1. When $p = 1$, use x_N.
R-3, SAS-2	Np	$x_{\lfloor h \rceil}$	The observation numbered closest to Np. Here, $\lfloor h \rceil$ indicates rounding to the nearest integer, choosing the even integer in the case of a tie. When $p \leq (1/2) / N$, use x_1.
R-4, SAS-1, SciPy-(0,1), Maple-3	Np	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)\,(x_{\lfloor h \rfloor + 1} - x_{\lfloor h \rfloor})$	Linear interpolation of the empirical distribution function. When $p < 1 / N$, use x_1. When $p = 1$, use x_N.
R-5, SciPy-(.5,.5), Maple-4	$Np + 1/2$	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)\,(x_{\lfloor h \rfloor + 1} - x_{\lfloor h \rfloor})$	Piecewise linear function where the knots are the values midway through the steps of the empirical distribution function. When $p < (1/2) / N$, use x_1. When $p \geq (N - 1/2) / N$, use x_N.
R-6, SAS-4, SciPy-(0,0), Maple-5	$(N + 1)p$	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)\,(x_{\lfloor h \rfloor + 1} - x_{\lfloor h \rfloor})$	Linear interpolation of the expectations for the order statistics for the uniform distribution on [0,1]. That is, it is the linear interpolation between points (p_h, x_h), where $p_h = h/(N{+}1)$ is the probability that the last of $(N{+}1)$ randomly drawn values will not exceed the h-th smallest of the first N randomly drawn values. When $p < 1 / (N{+}1)$, use x_1. When $p \geq N / (N + 1)$, use x_N.
R-7, Excel, SciPy-(1,1), Maple-6	$(N - 1)p + 1$	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)\,(x_{\lfloor h \rfloor + 1} - x_{\lfloor h \rfloor})$	Linear interpolation of the modes for the order statistics for the uniform distribution on [0,1]. When $p = 1$, use x_N.
R-8, SciPy-(1/3,1/3), Maple-7	$(N + 1/3)p + 1/3$	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)\,(x_{\lfloor h \rfloor + 1} - x_{\lfloor h \rfloor})$	Linear interpolation of the approximate medians for order statistics. When $p < (2/3) / (N + 1/3)$, use x_1. When $p \geq (N - 1/3) / (N + 1/3)$, use x_N.
R-9, SciPy-(3/8,3/8), Maple-8	$(N + 1/4)p + 3/8$	$x_{\lfloor h \rfloor} + (h - \lfloor h \rfloor)\,(x_{\lfloor h \rfloor + 1} - x_{\lfloor h \rfloor})$	The resulting quantile estimates are approximately unbiased for the expected order statistics if x is normally distributed. When $p < (5/8) / (N + 1/4)$, use x_1. When $p \geq (N - 3/8) / (N + 1/4)$, use x_N.

Notes:

- R-1 through R-3 are piecewise constant, with discontinuities.
- R-4 and following are piecewise linear, without discontinuities, but differ in how h is computed.
- R-3 and R-4 are not symmetric in that they do not give $h = (N + 1) / 2$ when $p = 1/2$.

The standard error of a quantile estimate can in general be estimated via the bootstrap. The Maritz-Jarrett method can also be used.[9]

Discussion

Standardized test results are commonly misinterpreted as a student scoring "in the 80th percentile," for example, as if the 80th percentile is an interval to score "in," which it is not; one can score "at" some percentile, or between two percentiles, but not "in" some percentile. Perhaps by this example it is meant that the student scores between the 80th and 81st percentiles, or "in" the group of students whose score placed them at the 80th percentile.

If a distribution is symmetric, then the median is the mean (so long as the latter exists). But, in general, the median and the mean can differ. For instance, with a random variable that has an exponential distribution, any particular sample of this random variable will have roughly a 63% chance of being less than the mean. This is because the exponential distribution has a long tail for positive values but is zero for negative numbers.

Quantiles are useful measures because they are less susceptible than means to long-tailed distributions and outliers. Empirically, if the data being analyzed are not actually distributed according to an assumed distribution, or if there are other potential sources for outliers that are far removed from the mean, then quantiles may be more useful descriptive statistics than means and other moment-related statistics.

Closely related is the subject of least absolute deviations, a method of regression[10] that is more robust to outliers than is least squares, in which the sum of the absolute value of the observed errors is used in place of the squared error. The connection is that the mean is the single estimate of a distribution that minimizes expected squared error while the median minimizes expected absolute error. Least absolute deviations shares the ability to be relatively insensitive to large deviations in outlying observations, although even better methods of robust regression are available.

The quantiles of a random variable are preserved under increasing transformations, in the sense that, for example, if m is the median of a random variable X, then 2^m is the median of 2^X, unless an arbitrary choice has been made from a range of values to specify a particular quantile. (See quantile estimation, above, for examples of such interpolation.) Quantiles can also be used in cases where only ordinal data are available.

See also

- Flashsort – sort by first bucketing by quantile
- Interquartile range
- Descriptive statistics
- Quartile
- Q-Q plot
- Quantile function
- Quantile normalization
- Quantile regression
- Quantization
- Summary statistics
- Tolerance interval ("confidence intervals for the pth quantile"[11])

Notes

References

1. Hyndman, R.J.; Fan, Y. (November 1996). "Sample Quantiles in Statistical Packages". *American Statistician*. American Statistical Association. **50** (4): 361–365. doi:10.2307/2684934 (https://doi.org/10.2307%2F2684934). JSTOR 2684934 (https://www.jstor.org/stable/2684934).
2. Mathematica Documentation (http://reference.wolfram.com/language/ref/Quantile.html) See 'Details' section
3. MATLAB implementation of the various estimation methods (http://www.mathworks.co.uk/matlabcentral/fileexchange/46555-quantile-calculation)
4. Frohne, I.; Hyndman, R.J. (2009). *Sample Quantiles* (http://stat.ethz.ch/R-manual/R-devel/library/stats/html/quantile.html). R Project. ISBN 3-900051-07-0.
5. "Function Reference: quantile - Octave-Forge - SourceForge" (http://octave.sourceforge.net/octave/function/quantile.html). Retrieved 6 September 2013.
6. http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mstats.mquantiles.html
7. http://www.maplesoft.com/support/help/maple/view.aspx?path=Statistics%2FQuantile
8. "Archived copy" (https://web.archive.org/web/20160416123322/http://www.eviews.com/help/EViews%209%20Help/graphs.020.09.html). Archived from the original (http://www.eviews.com/help/EViews%209%20Help/graphs.020.09.html#ww140852) on April 16, 2016. Retrieved April 4, 2016.
9. Rand R. Wilcox. Introduction to robust estimation and hypothesis testing. ISBN 0-12-751542-9
10. Ijsmi, Editor (2017-03-26). "Application of Quantile regression in clinical research: An overview with the help of R and SAS statistical package" (http://www.ijsmi.com/Journal/index.php/IJSMI/article/view/5). *International Journal of Statistics and Medical Informatics*. **2** (1): 1–6. doi:10.3000/ijsmi.v2i1.5 (https://doi.org/10.3000%2Fijsmi.v2i1.5).
11. Stephen B. Vardeman (1992). "What about the Other Intervals?". *The American Statistician*. **46** (3): 193–197. doi:10.2307/2685212 (https://doi.org/10.2307%2F2685212). JSTOR 2685212 (https://www.jstor.org/stable/2685212).

Further reading

- R.J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 1980.

Wikimedia Commons has media related to ***Quantiles***.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Quantile&oldid=772749192"

Categories: Summary statistics