QUANTILES: Feb. 4, 2005, R. Dudley, 18.465 notes

Let X be a real random variable with distribution function F, so that $F(x) = P(X \leq x)$ for all x. Let $0 < p < 1$. Then a number x is called a *pth quantile* of F, or of X, if $F(x) = p$, or more generally if $F(x) \geq p$ and $P(X \geq x) \geq 1 - p$. The definition with $F(x) = p$ applies to all continuous distributions. The more general definition is needed for discrete distributions where there may be no x with $F(x) = p$.

If a pth quantile x is uniquely determined, as it is if F is strictly increasing in a neighborhood of x, it is called *the* pth quantile of F or X and can be written as x_p. For a lot of distributions used in statistics such as χ^2 and F distributions, specific quantiles are tabulated such as the 0.95, 0.975, and 0.99 quantiles.

A median is a 1/2 quantile. If it is not unique, there is an interval of medians and *the* median is defined as the midpoint of this interval.

Now let's consider how to define pth quantiles ξ_p of a finite sample $X_1, ..., X_n$. A rough definition is that a fraction p of the observations should be less than (or equal) ξ_p and a fraction $1 - p$ should be larger than (or equal) to ξ_p. If np is not an integer then we seem to be talking about a non-integer count of number of observations which is not well-defined.

There is a generally agreed-on definition of the 1/2 sample quantile, the sample median, namely if $n = 2k + 1$ odd, it's the middle order statistic $X_{(k+1)}$. If $n = 2k$ even, then it's $(X_{(k)} + X_{(k+1)})/2$. But it seems that for $p \neq 1/2$ there is no such agreed definition. The next most often mentioned quantiles for finite samples are the quartiles, where $p = 1/4$ (lower quartile) and $p = 3/4$ (upper quartile). Possible summary statistics for a class's exam scores are to give the median and the upper and lower quartiles.

Other quantiles sometimes mentioned are percentiles, often used about scores for an individual on a standardized exam. The pth quantile is the same as the $100p$th percentile.

I looked at several statistics books searching for precise definitions of sample quantiles. Many books have no words beginning with q in their subject indices. Other books including Randles and Wolfe (our text) mention quantiles only for probability distributions, not for samples.

I found precise definitions of sample pth quantiles for $p \neq 1/2$ in four books. The four definitions were all different. I will list them, but there will not be regular problems assigned about these, just maybe some extra-credit problem(s). So, don't memorize them or pay very much attention to them. Just notice that from the rough definition, we'd expect ξ_p to be something like $X_{(np)}$, but np is often not an integer. To formulate the definitions, here is some notation. Let $\lfloor x \rfloor$, the integer part of x, be the largest integer $\leq x$. Let $\{x\}$, the fractional part of x, be $x - \lfloor x \rfloor$. Let $\lceil x \rceil$ be the smallest integer $\geq x$. Let $r(x)$ be x rounded to the nearest integer, rounded up if $\{x\} = 1/2$.

Here are the definitions in alphabetical order by first author of the textbook. The pth quantile of a sample of n numbers with order statistics $X_{(1)} \leq ... \leq X_{(n)}$ is:

1. $X_{(r(np))}$ if $p < 1/2$, $X_{(n+1-r(n(1-p)))}$ if $p > 1/2$, the sample median if $p = 1/2$ (Casella and Berger, *Statistical Inference*, 1990).

2. $X_{(\lfloor (n+1)p \rfloor)} + \{(n+1)p\} \left(X_{(\lceil (n+1)p \rceil)} - X_{(\lfloor (n+1)p \rfloor)} \right)$: R. Hogg and E. Tanis, *Probability and Statistical Inference*, Sixth Ed.

3. $X_{(\lceil np \rceil)}$ if np is not an integer, or if it is, $(X_{(np)} + X_{(np+1)})/2$: R. A. Johnson, *Miller and Freund's Probability and Statistics for Engineers* 5th ed., 1994.

4. $X_{(r((n+1)p))}$, given just for $p = 1/4$ or $3/4$ (would be undefined if $(n+1)p < 1/2$ or $\geq n + (1/2)$): Mendenhall and Sincich, *Statistics for Engineering and the Sciences*.

Some of the apparent complexity of some definitions is motivated by a consideration of symmetry: if all X_i are replaced by $-X_i$, reversing the order of the order statistics while also changing their signs, one would like ξ_p for the $-X_i$ to be $-\xi_{1-p}$ for the X_i.

Since there is no agreement on a precise definition of sample quantiles other than the sample median, one can just keep in mind the rough definition.