

## QUANTILES AND SAMPLE QUANTILES

There is no consensus among statisticians on how to define sample quantiles, except for the sample median. I looked at four textbooks and found four different, conflicting definitions. I will give one for purposes of this course.

First, to define quantiles in general, let  $F$  be the distribution function of a random variable  $X$ ,  $F(x) = \Pr(X \leq x)$  for all  $x$ . Any distribution function is continuous from the right. It has a limit from the left,  $F(x-) = \lim_{u \uparrow x} F(u)$  for any  $x$ , which equals  $F(x)$  for a continuous distribution function, but  $F(x) - F(x-) = \Pr(X = x)$  which is larger than 0 for some  $x$  if  $X$  has a discrete distribution.

For  $0 < p < 1$ , a  $p$ th quantile of  $F$  or  $X$  is an  $x$  such that  $F(x-) \leq p \leq F(x)$ . If there is only one such  $x$ , it is called *the*  $p$ th quantile of  $F$  or  $X$ . If there is an interval  $J$  (possibly a half-line or the whole line) such that  $F$  is strictly increasing and continuous on  $J$ , 0 to the left of  $J$  and 1 to the right of  $J$ , then all  $p$ th quantiles are unique and equal to  $F^{-1}(p)$ .

In general,  $p$ th quantiles may not be unique. Then there is an interval  $[a, b]$  with  $a < b$  such that  $F(x) < p$  for  $x < a$ ,  $F(x) = p$  for  $a \leq x < b$ , and  $F(x) > p$  for  $x > b$ . In such a case we define *the*  $p$ th quantile of  $F$  as the midpoint  $(a + b)/2$ .

Given a finite sample  $X_1, \dots, X_n$  of observations, the *empirical distribution function*  $F_n$  is defined by  $F_n(x) = \frac{1}{n} \sum_{j=1}^n 1_{X_j \leq x}$ , where  $1_{X_j \leq x}$  is defined as 1 if  $X_j \leq x$  and 0 otherwise. This is indeed a distribution function. Then sample quantiles are defined as quantiles for  $F_n$ . The sample median is the sample  $1/2$  quantile, from which it follows that if  $n = 2k + 1$  is odd, the sample median is  $X_{(k+1)}$ , whereas if  $n = 2k$  is even, the sample median is  $(X_{(k)} + X_{(k+1)})/2$ .

Suppose the  $n$  observations are all different. Then their order statistics satisfy

$$X_{(1)} < X_{(2)} < \dots < X_{(n)}.$$

For each  $j = 1, \dots, n$ , we have  $F_n(X_{(j)}) = j/n > F_n(X_{(j)}-) = (j-1)/n$ . It follows that for any  $p$  such that  $(j-1)/n < p < j/n$ , the  $p$ th sample quantile is  $X_{(j)}$ . Whereas, if  $p = j/n$  for some  $j = 1, \dots, n-1$ , then the  $p$ th sample quantile is  $(X_{(j)} + X_{(j+1)})/2$ . If  $n = 2j$  is even, we get again the formula for the sample median in that case.