ORDER STATISTICS, QUANTILES, AND SAMPLE QUANTILES

1. Order statistics

Let $X_1, \ldots, X_n$ be n real-valued observations. One can always arrange them in order to get the *order statistics* $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$. Since $X_{(k)}$ actually depends on n, not only on k, a more precise notation for it is $X_{k:n}$.

1.1. **Uniform order statistics — beta distributions.** To find the distribution of an order statistic, let $X_1, \ldots, X_n$ be i.i.d. with a distribution function F. For any x, the probability that $X_{k:n} \le x$ is the probability that k or more of the X_j with $j \le n$ are $\le x$. Expressed in terms of binomial probabilities this gives

$$\Pr(X_{k:n} \le x) = E(k, n, F(x)), \tag{1}$$

the probability that in n independent trials with probability $F(x)$ of success on each, there are k or more successes.

Gamma and beta probabilities are treated in the file gammabeta.pdf on the course website. For any $a > 0$ and $b > 0$, the beta function $B(a, b)$ is defined as $\int_0^1 x^{a-1}(1-x)^{b-1}dx$, with $0 < B(a, b) < +\infty$. Recall that for $a > 0$, the gamma function is defined by $\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx$. For any integer $k \ge 1$ we have $\Gamma(k) = (k-1)!$. We have $\Gamma(a+1) = a\Gamma(a)$ for all $a > 0$. Letting $a \downarrow 0$, we have $\Gamma(a+1) \to 1$, so $\Gamma(a) = \Gamma(a+1)/a \to +\infty$.

The identity $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ relating beta and gamma functions holds for all $a, b > 0$ (Theorem 1.5.5 of gammabeta.pdf). The beta(a, b) density is defined by $\beta_{a,b}(x) = x^{a-1}(1-x)^{b-1}/B(a, b)$ for $0 < x < 1$ and 0 elsewhere. The corresponding distribution function is written as $I_x(a, b) = \int_0^x \beta_{a,b}(u)du$. There is an identity relating the binomial and beta distributions: for $0 \le p \le 1$ and $k \ge 1$,

$$E(k, n, p) \equiv I_p(k, n-k+1) \tag{2}$$

(gammabeta.pdf, Theorem 1.5.12). This is proved by differentiating with respect to p, which gives a telescoping sum on the left. For $k = 0$, simply $E(0, n, p) \equiv 1$. Combining with (1) gives

$$\Pr(X_{k:n} \le x) = I_{F(x)}(k, n-k+1). \tag{3}$$

Date: 18.650, Nov. 13, 2015.

Order statistics are defined for $k = 1, ..., n$, so $k \geq 1$ as desired in (2). The equation (3) simplifies if F is the $U[0,1]$ distribution function, $F(x) = 0$ for $x < 0$, $F(x) = x$ for $0 \leq x \leq 1$, and $F(x) = 1$ for $x > 1$. Let $U_{k:n}$ be the kth order statistic from n i.i.d. $U[0,1]$ random variables. Then

$$\Pr(U_{k:n} \leq x) = I_x(k, n-k+1). \tag{4}$$

This gives not only the distribution function but also the density, namely $\beta_{k,n-k+1}$, of $U_{k:n}$. It will be seen in Corollary 1(b) that it's possible, starting with $U[0,1]$ order statistics, to get them for other, general distributions.

If X has a $\beta_{a,b}$ distribution, it's known (Rice, Appendix A, Common Distributions) that $EX = a/(a+b)$, which can be seen as follows:

$$EX = \frac{B(a+1,b)}{B(a,b)} = \frac{\Gamma(a+1)\Gamma(b)\Gamma(a+b)}{\Gamma(a+b+1)\Gamma(a)\Gamma(b)} = \frac{a}{a+b}.$$

From this and (4) it follows that

$$EU_{j:n} = \frac{j}{n+1}, \quad j = 1, \ldots, n. \tag{5}$$

2. Quantiles

Let X be a real random variable with distribution function F, so that $P(X \leq x) = F(x)$ for all x. Let's define the left limit of F at x by $F(x-) := \lim_{y \uparrow x} F(y)$, which equals $F(x)$ for a continuous F but is less than $F(x)$ if x is a possible value of X with a discrete distribution. 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 $F(x-) \leq p \leq F(x)$. The definition with $F(x) = p$ applies to all continuous distribution functions F. The more general definition is needed for discrete distributions where there may be no x with $F(x) = p$.

We have $F(x-) = P(X < x)$ and $P(X \geq x) = 1 - F(x-)$. So, x is a pth quantile of F or X if and only if $F(x) \geq p$ and $P(X \geq x) \geq 1 - p$.

A pth quantile x of F is uniquely determined if F is strictly increasing in a neighborhood of x, or if $F(x-) < p < F(x)$. Then it is called *the* pth quantile of F or X and can be written as x_p. If $F(x-) < F(x)$, then x is a pth quantile of F for all p such that $F(x-) \leq p \leq F(x)$.

For a lot of continuous distributions used in statistics such as χ^2 and F distributions, specific quantiles such as the 0.95, 0.975, and 0.99 quantiles are tabulated. If F is continuous and is strictly increasing on the interval J (possibly a half-line or the whole line) of x for which $0 < F(x) < 1$, then F has an inverse function F^{-1} from $(0,1)$ onto J, such that $F^{-1}(p) = x_p$ and $F(F^{-1}(p)) \equiv p$ for $0 < p < 1$.

A more general distribution function F may not have an inverse as just defined but there are substitutes for it defined as follows. For $0 < p < 1$ let $F^{\leftarrow}(p) := \inf\{x : \ F(x) \geq p\}$. Then $a_p := F^{\leftarrow}(p)$ is always a pth quantile of F. Similarly let $F^{\rightarrow}(p) := b_p := \sup\{x : \ F(x) \leq p\}$.

The following statements are not proved here. Part of the proof is left as a problem. The set of pth quantiles of F is the finite closed interval $[a_p, b_p]$, which often reduces to a point. For a discrete distribution such as a binomial distribution, there are selected values of p such that $a_p < b_p$. For example, for the distribution function F of the binomial$(n, 1/2)$ distribution, each value $p = B(k, n, 1/2)$ for $k = 1, ..., n-1$ is attained on the interval $[k, k+1)$, $a_p = k$ and $b_p = k+1$.

Another way to choose a pth quantile is, when it is not unique, to take the *midpoint pth quantile* as the midpoint $(a_p + b_p)/2$ of the interval of pth quantiles. This definition is used in defining *the median* of F, or of a random variable X with distribution function F, namely, as the midpoint $1/2$ quantile of F.

The following will let us define random variables with any distribution, given one with a $U[0, 1]$ distribution.

Theorem 1. *Let F be any probability distribution function and V a random variable having a $U[0, 1]$ distribution. Then $F^{\leftarrow}(V)$ has distribution function F.*

Proof. First, it will be shown that for any real x and any y with $0 < y < 1$, $y \leq F(x)$ if and only if $F^{\leftarrow}(y) \leq x$. For "only if," suppose $y \leq F(x)$. Then $F^{\leftarrow}(y) \leq x$ by definition of $F^{\leftarrow}$. For "if," suppose $F^{\leftarrow}(y) \leq x$. Then there exists a sequence x_n decreasing down to some $u \leq x$ with $F(x_n) \geq y$. By right-continuity of distribution functions, $F(u) \geq y$, and so $F(x) \geq F(u) \geq y$ as desired.

Then, for any x, $\Pr(F^{\leftarrow}(V) \leq x) = \Pr(V \leq F(x)) = F(x)$ as V has a $U[0, 1]$ distribution, so the theorem is proved. □

If one can generate a $U[0, 1]$ random variable V (as any reasonable computer system can) and evaluate $F^{\leftarrow}$, then one can generate a random variable with distribution function F as $F^{\leftarrow}(V)$.

Examples. Let F be the standard exponential distribution function: $F(x) = 0$ for $x \leq 0$ and $F(x) = 1 - e^{-x}$ for $0 < x < \infty$. For $0 < V < 1$, $F^{\leftarrow}(V) = F^{-1}(V)$ is the unique X such that $1 - e^{-X} = V$, or $e^{-X} = 1 - V$, so $X = -\ln(1 - V)$. Noting that $U := 1 - V$ also has a $U[0, 1]$ distribution, we are taking $X = -\ln(U)$ to get X standard exponential.

For the standard normal distribution function Φ, there is no simple closed form expression for Φ itself, nor for Φ^{-1}, although it can be computed (as in R, qnorm(p)). There are competing ways to generate a $N(0,1)$ variable which may be preferred. Anyhow, R can generate n i.i.d. $N(\mu, \sigma^2)$ random variables by rnorm(n, μ, σ).

Corollary 1. *Let F be any distribution function and let $V_1, ..., V_n$ be i.i.d. $U[0,1]$. Then (a) $X_j := F^{\leftarrow}(V_j)$ are i.i.d. F, and (b) $X_{j:n} = F^{\leftarrow}(V_{j:n})$ for $j = 1, \ldots, n$.*

Proof. Part (a) follows directly from Theorem 1. Then (b) follows since $F^{\leftarrow}$ is a nondecreasing function. □

For any $X_1, \ldots, X_n$ i.i.d. with a distribution function F, and $k = 1, \ldots, n$, $X_{k:n}$ is equal in distribution to $F^{\leftarrow}(V_{k:n})$. But since $F^{\leftarrow}$ is a nonlinear function on $0 < v < 1$ (unless F is a $U[a,b]$ distribution function), it may not be easy to evaluate $EX_{k:n}$. For example, if F is the $N(0,1)$ distribution function Φ and $Z_1, ..., Z_n$ are i.i.d. $N(0,1)$, then the expectations $EZ_{k:n}$ can be computed but not so easily. (Such expectations are used in the Shapiro–Wilk test of normality.)

The least pth quantile $F^{\leftarrow}(p)$ is useful for some theoretical purposes as in Theorem 1. But for practical purposes, and even in theory for defining the median, the midpoint pth quantile seems preferable.

3. Empirical distribution functions and sample quantiles

For any observations $X_1, \ldots, X_n$, the *empirical distribution function* is defined by $F_n(x) = \frac{1}{n}\sum_{j=1}^{n} 1_{X_j \le x}$ where $1_{X_j \le x} = 1$ if $X_j \le x$ and 0 otherwise. In other words, it's the fraction of X_j for $j \le n$ that are $\le x$.

Here F_n is important in nonparametric statistics. Up to now we've had estimation of finite-dimensional parameters θ. Here a general distribution function F, an object in an infinite-dimensional space, is estimated by F_n. A simple hypothesis H_0 that $X_1, ..., X_n$ are i.i.d. with distribution function F can be tested by evaluating $K_n := \sup_x |(F_n - F)(x)|$. If H_0 is true and F is continuous, then $\sqrt{n}K_n$ has a specific limiting distribution as $n \to \infty$, so H_0 can be tested (Kolmogorov's test, 18.465).

If X_j are all distinct, as they will be with probability 1 if they are i.i.d. from a continuous distribution, then $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$. We will have $F_n(x) = 0$ for $x < X_{(1)}$, $F_n(x) = j/n$ for $X_{(j)} \le x < X_{(j+1)}$ and $j = 1, ..., n-1$, and $F_n(x) = 1$ for $x \ge X_{(n)}$. These relations actually hold for any X_j, some of which may be tied. If $X_{(j)} = X_{(j+1)}$, then

there are no x with $X_{(j)} \le x < X_{(j+1)}$, and F_n never takes the value j/n.

Now let's consider how to define pth quantiles ξ_p, for $0 < p < 1$ as always, of a finite sample $X_1, ..., X_n$. Suppose we define ξ_p as a pth quantile of F_n. Then if np is not an integer, there will be just one value of $j = 1, ..., n$ such that $(j-1)/n < p < j/n$ and F_n has a unique pth quantile, namely $X_{(j)}$. We can write $j = \lceil np \rceil$ where $\lceil x \rceil$ is the smallest integer $\ge x$. Whereas, if np is an integer j, then F_n has an interval of pth quantiles with endpoints $X_{(j)}$ and $X_{(j+1)}$. In this course, the pth quantile ξ_p of the finite sample will be defined as the midpoint pth quantile of F_n. For $p = 1/2$ this agrees with the generally accepted definition of sample median: if $n = 2k+1$ is odd, the sample median is the middle order statistic $X_{(k+1)}$. If $n = 2k$ even, then it's $(X_{(k)} + X_{(k+1)})/2$. For $p \ne 1/2$ textbook authors give a variety of different definitions of sample pth quantile (see the Appendix), but taking the midpoint pth quantile of F_n seems to me to be the most justified.

The following symmetry property is desirable in a definition of sample quantiles ξ_p: if all X_i are replaced by $-X_i$, reversing the order of the order statistics while also changing their signs, one would like, for $0 < p < 1$,

$$\xi_p(\{-X_i\}_{i=1}^n) = -\xi_{1-p}(\{X_i\}_{i=1}^n). \tag{6}$$

This symmetry property does not hold for ξ_p equal to the least pth quantile $F_n^{\leftarrow}(p)$.

Proposition 1. *The symmetry (6) holds for the midpoint empirical quantiles.*

Proof. If we take $Y_j = -X_j$ for $j = 1, ..., n$ then the order statistics are clearly $Y_{(j)} \equiv -X_{(n+1-j)}$. For $0 < p < 1$, if np is not an integer, we can check that $\lceil np \rceil + \lceil n(1-p) \rceil = n+1$ because the sum of two integers on the left must be an integer, larger than n and less than $n+2$ because if x is not an integer then $0 < \lceil x \rceil - x < 1$. Thus the pth quantile of the Y_i is

$$Y_{(\lceil np \rceil)} = -X_{(n+1-\lceil np \rceil)} = -X_{(\lceil n(1-p) \rceil)} = -\xi_{1-p}$$

for the X_i as desired. If $np = j$, an integer, with $1 \le j \le n-1$, then $n(1-p) = n-j$ is also an integer in the same range. The pth midpoint empirical quantile of the Y_i is

$$\frac{Y_{(j)} + Y_{(j+1)}}{2} = -\frac{X_{(n+1-j)} + X_{(n-j)}}{2} = -\frac{X_{(n(1-p)+1)} + X_{(n(1-p))}}{2} = -\xi_{1-p}$$

for the X_i. So the symmetry (6) does hold for the midpoint empirical quantiles for all p with $0 < p < 1$. □

4. Robust estimates of location and scale

We know that if $X_1, ..., X_n$ are i.i.d. with a distribution having a finite mean μ and variance σ^2, and $\overline{X}$ is their sample mean, then $\sqrt{n}(\overline{X} - \mu)$ has a distribution converging as $n \to \infty$ to $N(0, \sigma^2)$, by the central limit theorem. But, suppose X_j do not have a finite mean. Then the sample may contain outliers, observations far from most of the others. These can have a large influence on $\overline{X}$. Even if there is a finite mean, if the variance is infinite, the distribution of $\overline{X}$ is not so well controlled. The sample variance is also very sensitive to outliers. The median provides a measure of location, as the mean does, but the sample median is very little sensitive to outliers. To get a measure of scale, analogous to the standard deviation σ, but which is well defined and can be estimated even when the variance is infinite, one can use the *interquartile range* (IQR), the difference between the 3/4 and 1/4 quantiles. The sample IQR may be divided by 1.35, which is the IQR for a $N(0, 1)$ distribution, so as to give an estimate of σ for normal distributions (Rice, §10.5 pp. 401-402).

5. **The large-sample distribution of sample medians**

Under some conditions, sample medians also become asymptotically normal for large n. It's convenient to consider $n = 2k + 1$ odd where k is a positive integer. Then for n i.i.d. random variables $X_1, ..., X_n$ with order statistics $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$, the sample median is $X_{(k+1)}$.

In the special case where X_j have $U[0, 1]$ distribution, we know by (4) that the sample median $X_{(k+1)}$ has a beta$(k+1, k+1)$ distribution, namely it has a density $f_k(x) = x^k(1 - x)^k/B_k$ for $0 \le x \le 1$ and 0 elsewhere, where B_k is the beta function

$$B_k \;=\; B(k + 1, k + 1) \;=\; \Gamma(k + 1)^2/\Gamma(2k + 2) \;=\; k!^2/(2k + 1)!$$

and Γ is the gamma function. Also, given the sample median for $U[0, 1]$ variables we can get sample medians for any other distribution F by Corollary 1 as the sample median $X_{(k+1)} = F^{\leftarrow}(U_{(k+1)})$. For a continuous distribution function F that is well-behaved, meaning it is strictly increasing when $0 < F(x) < 1$, we have $F^{\leftarrow} = F^{-1}$ on $0 < p < 1$. So we can find the asymptotic distribution (asymptotic meaning as $n \to \infty$) of the sample median via the delta-method (treated in the handout deltamethod-.....pdf) and the following fact:

Proposition 2. *Let Y_k have a beta$(k+1, k+1)$ distribution. Then $\sqrt{k}\left(Y_k - \frac{1}{2}\right)$ converges in distribution to $N(0, 1/8)$ as $k \to \infty$.*

Proof. $V_k = Y_k - \frac{1}{2}$ has density for $|v| \le 1/2$

$$\left(\frac{1}{2} + v\right)^k \left(\frac{1}{2} - v\right)^k / B_k \;=\; \left(\frac{1}{4} - v^2\right)^k / B_k,$$

so $W_k := \sqrt{k} V_k$ has the density for $|w| \le \sqrt{k}/2$

$$k^{-1/2} 4^{-k} B_k^{-1} \left(1 - \frac{4w^2}{k}\right)^k \;\sim\; k^{-1/2} 4^{-k} B_k^{-1} \exp(-4w^2)$$

as $k \to \infty$. This is now factored into a constant C_k depending only on k times $\exp(-4w^2)$, and because we had a probability density, and the form of the density with respect to w is that of a $N(0, 1/8)$ density, C_k must converge to the correct normalizing constant for it, namely $2/\sqrt{\pi}$. So, the convergence in distribution follows as stated. □

To apply the delta-method, recall the following fact from beginning calculus on the derivative of an inverse function.

Fact 1. *Let F be defined, strictly increasing, and continuous on an open interval U containing a point x_0 and have a derivative $F'(x_0) > 0$. Let $y_0 = F(x_0)$. Then F has an inverse function F^{-1} defined on the interval $V = \{F(x) : \; x \in U\}$, so that $F^{-1}(F(x)) = x$ for all $x \in U$, and F^{-1} has a derivative at y_0 given by $(F^{-1})'(y_0) = 1/F'(x_0)$.*

Fact 1 can seem obvious in the usual (Leibniz) notation:

$$\frac{dx}{dy} = 1 / \left(\frac{dy}{dx}\right),$$

where the derivative on the right is evaluated at some $x = x_0$ and the one on the left at $y_0 = F(x_0)$.

6. Appendix: textbook definitions of sample quantiles

I found precise definitions of sample pth quantiles for $p \neq 1/2$ in six beginning statistics textbooks. One (Rice) gave two definitions. The seven definitions were all different. I will list them.

Next to the sample median, perhaps the most often mentioned sample quantiles are the quartiles, where $p = 1/4$ (lower quartile) and $p = 3/4$ (upper quartile).

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.

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 more 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 $r(x)$ be x rounded to the nearest integer, rounded up if $\{x\} = 1/2$. Order statistics $X_{(j)}$ are defined only for $j = 1, 2, ..., n$ (not for $j = 0$ or $n + 1$).

Here are the definitions in alphabetical order by first author of the textbook. By the way: James Berger (1) is a leading (Bayesian) statistician.

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). (The latest edition seems to be the second, from 2002. I have not seen it.) This is undefined if $p < 1/(2n)$ or $p > 1 - 1/(2n)$ (such extreme quantiles are not very important; if one is interested in extremes one can consider just $X_{(1)}$ and $X_{(n)}$).

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., 2001. This is undefined if $p < 1/(n+1)$ or $p > n/(n+1)$ (again, those are extreme values of p). This gives a piecewise linear, continuous, nondecreasing function of p, defined for $1/(n+1) \leq p \leq n/(n+1)$, equal to $X_{(j)}$ for $p = j/(n+1)$, $j = 1, ..., n$. Recall that these numbers $j/(n+1)$ appeared in (5) as $EU_{j:n}$. (In the Seventh Ed., 2006, the definitions seem to be the same but I could not be sure in a short time.)

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. Only this definition, of those to be mentioned, defines

quantiles as midpoint empirical quantiles, as in the present course. Another seemingly related book, Irwin Miller and Marylees Miller, *John E. Freund's Mathematical Statistics*, 6th Ed., 1999, like several other texts I looked at, has no words beginning with "q" in its subject index.

4. $X_{(r((n+1)p))}$, given only for quartiles, $p = 1/4$ or $3/4$; jth percentile, defined as $X_{(r((n+1)j/100))}$, presumably for $j = 1, ..., 99$ (would be undefined if $(n+1)j/100 < 1/2$, specifically if $j = 1$, $n \leq 48$, or if $(n+1)j/100 \geq n + (1/2)$, specifically if $n \leq 49$, $j = 99$): Mendenhall and Sincich, *Statistics for Engineering and the Sciences*, 5th ed., 2007, p. 39. If 48 or fewer individuals are ranked, it indeed arguably makes no sense to say that the highest-ranked individual was in the top 1% or the lowest-ranked in the bottom 1%.

5, 6. Rice (Third Ed., p. 387) defines ξ_p only for special values of $p = p_j$ with $\xi_{p_j} = X_{(j)}$. One choice is $p_j = j/(n+1)$ for $j = 1, ..., n$, and the other is $p_j = (j - \frac{1}{2})/n$. Both choices satisfy $0 < p_1 < \cdots < p_n < 1$ with the points equally spaced, in other words $p_j - p_{j-1}$ doesn't depend on j. The first definition with $p_j = j/(n+1)$ agrees with definition (2). for those values of p_j. Conversely, definition (2) is the piecewise linear interpolation of the Rice values. (Here $j/(n+1)$ can equal $1/2$, for a median, only if n is odd, and can equal $1/4$ or $3/4$ for quartiles only if $n+1$ is divisible by 4; $(2j-1)/(2n)$ can equal $1/2$ only for n odd, and can equal $1/4$ only if n is even but not divisible by 4; definition (2) gives any p with $1/(n+1) \leq p \leq n/(n+1)$ and so in particular $1/2, 1/4$, and $3/4$ as long as $n \geq 3$.)

7. R. E. Walpole and R. H. Meyers, *Probability and Statistics for Engineers and Scientists*, Fifth Ed., 1993, p. 210, gives a definition of the same type as Rice does but with $p_j = (j - \frac{3}{8})/(n + \frac{1}{4})$. (This can never equal $1/4$ or $3/4$ to define quartiles.) There is an 8th Edition from 2007, which I have not seen.

Different definitions of sample quantiles may have been made with different purposes in mind. There is a consensus only about the sample median.