3.6 Sample Quantiles/Order Statistics

Definition 8 Let $X_1, \ldots, X_n$ be n random variables.

The *order statistics* corresponding to $X_1, \ldots, X_n$, denoted by $X_{1:n}, \ldots, X_{n:n}$, are defined as:

$$\begin{aligned} X_{1:n} &= \min(X_1, \ldots, X_n) \\ &\vdots \\ X_{j:n} &= j\text{th smallest of } X_1, \ldots, X_n \\ &\vdots \\ X_{n:n} &= \max(X_1, \ldots, X_n) \end{aligned}$$ ////

The order statistics are clearly *statistics* inasmuch as they are functions of the sample $X_1, \ldots, X_n$; and, they are *ordered*. We shall see that the distribution(s) of order statistic(s) have relatively simple formulas when even the mean may not. In fact, using the cdf technique, we saw in Chapter V that

$$\begin{aligned} F_{X_{n:n}}(x) &= F^n(x), \quad \text{and} \\ F_{X_{1:n}}(x) &= 1 - (1 - F(x))^n \end{aligned}$$

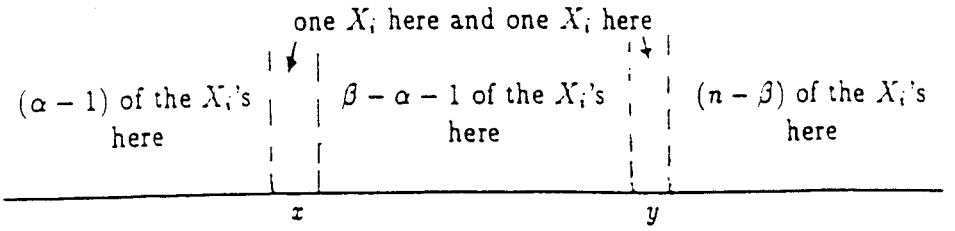
when $X_1, \ldots, X_n \overset{iid}{\sim} F(\cdot)$. Using that same technique and model, note the following result.

Result: If $X_1, \ldots, X_n \overset{iid}{\sim} F(\cdot)$ and $X_{1:n} \leq \cdots \leq X_{n:n}$ are the corresponding order statistics, then $F_{X_{\alpha:n}}(z) = P[X_{\alpha:n} \leq z] = P\left[\sum_{i=1}^{n} I_{(-\infty,z]}(X_i) \geq \alpha\right] = \sum_{j=\alpha}^{n} \binom{n}{j} F^j(z)(1 - F(z))^{n-j}$

using $\sum_{i=1}^{n} I_{(-\infty,z]}(X_i) \sim$ bin $(n, F(z))$ and noting the equivalence of events $\{X_{\alpha:n} \leq z\}$ and $\left\{\sum_{i=1}^{n} I_{(-\infty,z]}(X_i) \geq \alpha\right\}$. (The αth order statistic is less than or equal z if and only if at least α of the X_i's are less than or equal z, and $\sum_{i=1}^{n} I_{(-\infty,z]}(X_i)$ = the number of the X_i's which are less than or equal z.) ////

This result gives a simple formula for the cdf of an arbitrary order statistic under the iid model $X_1, \ldots, X_n \overset{iid}{\sim} F(\cdot)$ for arbitrary $F(\cdot)$.

Under the iid model $X_1, \ldots, X_n \overset{iid}{\sim} F(\cdot)$, where $F(\cdot)$ is absolutely continuous with probability density function $f(\cdot)$, there is a simple formula for the joint density of an arbitrary number of order statistics. Such joint density need not be memorized since it is readily derived using the "draw-the-line" trick and the multinomial distribution. Let us illustrate with two order statistics, say $X_{\alpha:n}$ and $X_{\beta:n}$ where $\alpha < \beta$. Need



for $P[X_{\alpha:n} \approx x;\ X_{\beta:n} \approx y]$, and using the multinomial distribution with five categories, we get

$$f_{X_{\alpha:n},X_{\beta:n}}(x,y) =$$

$$\frac{n!}{[(\alpha-1)!][1!][(\beta-\alpha-1)!][1!][(n-\beta)!]} F^{\alpha-1}(x)\ f(x)\ (F(y)-F(x))^{\beta-\alpha-1}\ f(y)(1-F(y))^{n-\beta}$$

for $x < y$.

Similarly,

$$\begin{aligned} f_{X_{\alpha:n}}(x) &= \tfrac{n!}{(\alpha-1)!(n-\alpha)!} F^{\alpha-1}(x)(1-F(x))^{n-\alpha} f(x) \\ &\vdots \\ f_{X_{1:n},\ldots,X_{n:n}}(x_1,\ldots,x_n) &= n!\textstyle\prod_{i=1}^{n} f(x_i) \text{for } x_1 < x_2 < \cdots < x_n \end{aligned}$$

Formula-wise these probability density functions are quite nice, yet they involve both the probability density function and the cumulative distribution function of the population sampled.

The title of this section suggests a relationship between order statistics and sample quantiles, which we have yet to define. Recall that the qth population quantile, denoted by ξ_q, is defined as the smallest number ξ satisfying $F(\xi) \geq q$ for population cdf $F(\cdot)$. Since the sample cdf $F_n(\cdot)$ is an estimator of the population cdf $F(\cdot)$, we define the qth sample quantile to be the qth quantile of $F_n(\cdot)$.

Definition 9 **qth sample quantile** Let $X_1,\ldots,X_n$ be a sample, where each X_i has a common cdf. The *qth sample quantile*, denoted by Z_q, is defined as the qth quantile of $F_n(\cdot)$. That is, Z_q is $X_{j:n}$ for $\frac{j-1}{n} < q \leq \frac{j}{n}$. ////

According to this definition a sample quantile is always an order statistic. An important quantile is the qth quantile for $q = 1/2$, which is called the *median*. According to our definition note that the *sample median* is the middle order statistic for odd sample size n and is $X_{n/2:n}$ for n even. This is true of n odd, say $n = 2k+1$, since $\frac{k}{2k+1} < \frac{1}{2} = q \leq \frac{k+1}{2k+1}$ implying that the sample median is the $(k+1)$st order statistic, which is the middle order statistic. (For n even the sample median is sometimes defined to be the average of the middle two order statistics.)

EXAMPLE 4 Assume $X_1,\ldots,X_n \overset{iid}{\sim}$ unif $(\alpha-\beta,\alpha+\beta)$, $\beta > 0$, and let $X_{1:n},\ldots,X_{n:n}$ be the corresponding order statistics. The n order statistics "tend" to divide the interval $(\alpha-\beta,\alpha+\beta)$ into $n+1$ non-overlapping bits of equal length, hence

$$\mathcal{E}\,[X_{j:n}] = \alpha - \beta + \frac{j}{n+1}(2\beta).$$

(See the Problems.) Note that the qth population quantile is $(\alpha - \beta) + q(2\beta)$ in this case, so the expected value of the qth sample quantile is close to its population counterpart. Now for simplicity, let $\alpha = \beta = 1/2$ and write $U_1, \ldots, U_n \overset{iid}{\sim}$ unif $(0,1)$. Note $f_{U_{j:n}}(u) = \frac{n!}{(j-1)!(n-j)!}\, u^{j-1}(1-u)^{n-j} I_{(0,1)}(u)$, that is, $U_{j:n} \sim$ beta $(j, n-j+1)$. Note that $U_{1:n}, U_{2:n} - U_{1:n}, \ldots, 1 - U_{n:n}$ are the $n+1$ "spacings" of these order statistics. It can be shown that these "spacings" are identically distributed. Do you think they are independent? ////

EXAMPLE 5 Assume $X_1, \ldots, X_n \overset{iid}{\sim} \exp(\beta), \beta > 0$, with $X_{1:n}, \ldots, X_{n:n}$ the corresponding order statistics. Recall that $X_{1:n} \sim \exp(\beta/n)$ and so $\mathcal{E}[X_{1:n}] = \beta/n$. Now, using the "lack-of-memory" property of the exponential, $X_{2:n} - X_{1:n} \sim \exp(\beta/(n-1)), \ldots, X_{j:n} - X_{j-1:n} \sim \exp(\beta/(n-j+1)), \ldots, X_{n:n} - X_{n-1:n} \sim \exp(\beta)$. That is, based on "lack-of-memory,"

$$X_{j:n} = E_1 + E_2 + \cdots + E_j, \qquad \text{for } j = 1, \ldots, n,$$

where $E_1, \ldots, E_j$ are independent and $E_1 \sim \exp(\beta/n), E_2 \sim \exp(\beta/(n-1)), \ldots, E_j \sim \exp((\beta/(n-j+1))$. This trick of writing the jth order statistic as the sum of j independent exponentials makes it particularly easy to obtain the mean and variance of $X_{j:n}$, to wit,

$$\begin{aligned}
\mathcal{E}[X_{j:n}] &= \mathcal{E}[E_1] + \mathcal{E}[E_2] + \cdots + \mathcal{E}[E_j] \\
&= \tfrac{\beta}{n} + \tfrac{\beta}{n-1} + \cdots + \tfrac{\beta}{n-j+1} \\
&= \beta\left[\tfrac{1}{n} + \tfrac{1}{n-1} + \cdots + \tfrac{1}{n-j+1}\right], \quad \text{and} \\
\operatorname{var}[X_{j:n}] &= \beta^2\left[\tfrac{1}{n^2} + \tfrac{1}{(n-1)^2} + \cdots + \tfrac{1}{(n-j+1)^2}\right].
\end{aligned}$$

Note that, for $j = n$, $\mathcal{E}\,[X_{n:n}] = \beta\left[1 + \frac{1}{2} + \cdots + \frac{1}{n}\right]$, and since $1 + \frac{1}{2} + \cdots + \frac{1}{n}$ behaves like $\ell n\, n$ for large n, $\mathcal{E}[X_{n:n}]$ grows like $\beta \ell n\, n$ for large n. Note that for this exponential distribution the qth population quantile is $\xi_q = \beta[-\ell n(1-q)]$. Since $\mathcal{E}\,[X_{j:n}] = \beta\left[\sum_{i=1}^{n}(1/i) - \sum_{i=1}^{n-j}(1/i)\right]$ for large j and n, $\mathcal{E}\,[X_{j:n}] \approx \beta[\ell n\, n - \ell n(n-j)] = \beta\left[-\ell n(1 - \frac{j}{n})\right]$ and for $q \approx j/n$ the expected value of a sample quantile is near its population counterpart. ////

Definition 10 Let $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$ be the order statistics of sample $X_1, \ldots, X_n$. The *sample range* is defined to be $X_{n:n} - X_{1:n}$ and the *sample midrange* is defined to be $(X_{1:n} + X_{n:n})/2$. ////

If $X_1, \ldots, X_n$ are assumed to be iid continuous random variables with cdf $F(\cdot)$ and pdf $f(\cdot)$, then the joint distribution of sample range and sample midrange can be routinely obtained from the joint density of $X_{1:n}$ and $X_{n:n}$ using the Jacobian technique.

We have

$$f_{X_{1:n},X_{n:n}}(x,y) = n(n-1)[F(y) - F(x)]^{n-2} f(x) f(y)\ I_{(x,\infty)}(y).$$

Make the transformation from $(X_{1:n}, X_{n:n})$ to (R,T) where $R = X_{n:n} - X_{1:n}$ and $T = (X_{1:n} + X_{n:n})/2$. We have $r = y - x$ and $t = (x+y)/2$, or, inversely, $x = t - r/2$ and $y = t + r/2$; hence

$$J = \begin{vmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial t} \end{vmatrix} = \begin{vmatrix} -1/2 & 1 \\ 1/2 & 1 \end{vmatrix} = -1, \qquad \text{and,}$$

$$f_{R,T}(r,t) =$$

$$n(n-1)[F(t+r/2) - F(t-r/2)]^{n-2} f(t-r/2) f(t+r/2)\ I_{(0,\infty)}(r)$$

which implies

$$f_R(r) = \int_{-\infty}^{\infty} f_{R,T}(r,t)dt, \qquad \text{and}$$

$$f_T(t) = \int_0^{\infty} f_{R,T}(r,t)dr.$$

EXAMPLE 6 Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{unif}\,(\alpha-\beta, \alpha+\beta)$, where $\beta > 0$. Here $f(x) = (1/2\beta)\ I_{(\alpha-\beta,\alpha+\beta)}(x)$ and $F(x) = \{[x - (\alpha-\beta)]/2\beta\}\ I_{(\alpha-\beta,\alpha+\beta)}(x) + I_{[\alpha+\beta,\infty)}(x)$.

Note that

$$\begin{aligned} f_{R,T}(r,t) &= n(n-1)[r/2\beta]^{n-2}(1/2\beta)^2\ I_{(\alpha-\beta,\alpha+\beta)}(t - r/2) I_{(\alpha-\beta,\alpha+\beta)}(t + r/2) I_{(0,\infty)}(r) \\ &= [n(n-1)r^{n-2}/(2\beta)^n]\ I_{(\alpha-\beta+r/2,\alpha+\beta-r/2)}(t) I_{(0,2\beta)}(r); \quad \text{and,} \\ f_R(r) &= \int f_{R,T}(r,t)dt \\ &= n(n-1)r^{n-2}/(2\beta)^n \int_{\alpha-\beta+r/2}^{\alpha+\beta-r/2} dt\ I_{(0,2\beta)}(r) \\ &= n(n-1)r^{n-2}(2\beta - r)/(2\beta)^n\ I_{(0,2\beta)}(r) \\ &= n(n-1)(r/2\beta)^{n-2}(1 - r/2\beta)(1/2\beta)\ I_{(0,1)}(r/2\beta). \end{aligned}$$

Note that $f_R(r)$ does not depend on α.

Also,

$$\begin{aligned}\mathcal{E}[R] &= 2\beta\int_0^{2\beta} n(n-1)(r/2\beta)^{n-1}(1-r/2\beta)dr/2\beta \\ &= 2\beta n(n-1)\int_0^1 u^{n-1}(1-u)du \\ &= 2\beta n(n-1)B(n,2) \\ &= [(n-1)/(n+1)](2\beta).\end{aligned}$$

Finally,

$$\begin{aligned}f_T(t) &= \int_0^\infty f_{R,T}(r,t)dr \\ &= [n(n-1)/(2\beta)^n]\int_0^{\min[2t-2(\alpha-\beta),2(\alpha+\beta)-2t]} r^{n-2}dr\ I_{(\alpha-\beta,\alpha+\beta)}(t) \\ &= [n(n-1)/(2\beta)^n]\left\{\int_0^{2t-2(\alpha-\beta)} r^{n-2}dr\ I_{(\alpha-\beta,\alpha)}(t) + \int_0^{2(\alpha+\beta)-2t} r^{n-2}dr\ I_{(\alpha,\alpha+\beta)}(t)\right\} \\ &= [n/2\beta]\left\{\left(\tfrac{t-\alpha}{\beta}+1\right)^{n-1} I_{(\alpha-\beta,\alpha)}(t) + \left(1-\tfrac{t-\alpha}{\beta}\right)^{n-1} I_{[\alpha,\alpha+\beta)}(t)\right\}.\end{aligned}$$

Note that $\mathcal{E}[(T-\alpha)/\beta] = (n/2)\left\{\int_{-1}^0 u(u+1)^{n-1}du + \int_0^1 u(1-u)^{n-1}du\right\} = 0$, so $\mathcal{E}[T] = \alpha$, as anticipated due to the symmetry of the uniform density.

If n is odd, say $n = 2k+1$, then $X_{k+1:2k+1}$ is the sample median with density $f_{X_{k+1:2k+1}}(x) = [(2k+1)!/(k!)^2][x-(\alpha-\beta)]^k[2\beta-x+(\alpha-\beta)]^k(1/2\beta)^{2k+1}I_{(\alpha-\beta,\alpha+\beta)}(x)$.

So,

$$\begin{aligned}\mathcal{E}\left[\frac{X_{k+1:2k+1}-\alpha}{\beta}\right] &= \frac{(2k+1)!}{2^{2k+1}(k!)^2}\int_{-1}^1 u(u+1)^k(1-u)^k du \\ &= \frac{(2k+1)!}{2^{2k+1}(k!)^2}\int_{-1}^1 u(1-u^2)^k du \\ &= 0\end{aligned}$$

which means $\mathcal{E}[X_{k+1:2k+1}] = \alpha$. So the three statistics, sample mean, sample midrange, and sample median all have expectation α (the population mean) and hence are competing estimators of α. Which has the smaller variance? (See the Problems.) ////