7.5 Order Statistics

The **order statistics** are the items of the random sample arranged, or ordered in magnitude from the smallest to the largest. Recently, the importance of order statistics has increased owing to the more frequent use of nonparametric inferences and robust procedures. However, order statistics have always been prominent because, among other things, they are needed to determine rather simple statistics such as the sample median, the sample range, and the empirical distribution function.

In most of our discussion, we will assume that the random sample arises from a continuous-type distribution. This means, among other thing, that the probability of any two sample items being equal is zero. That is, the probability is one that the items can be ordered from smallest to largest without having two equal values. Of course, in practice, we do frequently observe ***ties***; but if the probability of this is small, the following distribution theory will hold approximately. Thus, in the discussion here, we are assuming that the probability of tie is zero.

If $X_1, X_2, \cdots, X_n$ are observations of a random sample of size n from a continuous-type distribution with distribution function $F(x)$ and p.d.f. $f(x)$, we let the random variables

$$Y_1 < Y_2 < \cdots < Y_n$$

denote the order statistics of that sample. That is,

$$\begin{aligned} Y_1 &= \text{smallest of } X_1, X_2, \cdots, X_n, \\ Y_2 &= \text{second smallest of } X_1, X_2, \cdots, X_n, \\ &\vdots \\ Y_n &= \text{largest of } X_1, X_2, \cdots, X_n. \end{aligned}$$

The probability density functions for Y_1 and Y_n can be found using the method of distribution functions. Because Y_n is the largest of $X_1, X_2, \cdots, X_n$, the event $(Y_n \le y_n)$ will occur if and only if the event $(X_i \le y_n)$ occur, for every $i = 1, 2, \ldots, n$. That is,

$$P(Y_n \le y_n) = P(X_1 \le y_n, X_2 \le y_n, \cdots, X_n \le y_n)$$

To determine the distribution of the rth order statistic, Y_r depends on the binomial distribution. Suppose that $0 < F(x) < 1$ for $a < x < b$ and $F(a) = 0$, $F(b) = 1$. (It is possible that $a = -\infty$ and/or $b = \infty$.) The event that the rth order statistic, $Y_r \le y_r$ can occur if and only if at least r of the n observations are less than or equal to y_r. That is, here the probability of "success" on each trial is $F(x)$ and we must have at least r successes. Thus,

$$\begin{aligned} G_r(y_r) = P(Y_r \le y_r) &= \sum_{k=r}^{n} \binom{n}{k} [F(y_r)]^k [1 - F(y_r)]^{n-k} \\ &= \sum_{k=r}^{n-1} \binom{n}{k} [F(y_r)]^k [1 - F(y_r)]^{n-k} + [F(y_r)]^n. \end{aligned}$$

Thus the p.d.f. of Y_r is

$$g_r(y_r) = G_r'(y_r) = \sum_{k=r}^{n-1} \binom{n}{k} (k)[F(y_r)]^{k-1} f(y_r)[1-F(y_r)]^{n-k}$$

$$+ \sum_{k=r}^{n-1} \binom{n}{k} [F(y_r)]^k (n-k)[1-F(y_r)]^{n-k-1}[-f(y_r)]$$

$$+ n[F(y_r)]^{n-1} f(y_r).$$

$$= \frac{n!}{(r-1)!(n-r)!}[F(y_r)]^{r-1} f(y_r)[1-F(y_r)]^{n-r}$$

$$+ \sum_{k=r+1}^{n-1} \frac{n!}{(k-1)!(n-k)!}[F(y_r)]^{k-1} f(y_r)[1-F(y_r)]^{n-k}$$

$$- \sum_{k=r}^{n-2} \frac{n!}{k!(n-k-1)!}[F(y_r)]^k [1-F(y_r)]^{n-k-1}[f(y_r)]$$

$$- n[F(y_r)]^{n-1}[1-F(y_r)]^0[f(y_r)] + n[F(y_r)]^{n-1} f(y_r).$$

Hence, we have that the p.d.f. of Y_r is

$$g_r(y_r) = \frac{n!}{(r-1)!(n-r)!}[F(y_r)]^{r-1}[1-F(y_r)]^{n-r} f(y_r), \quad a < y_r < b.$$

It is worth noting that the p.d.f. of the smallest order statistic is

$$g_1(y_1) = n[1-F(y_1)]^{n-1} f(y_1), \quad a < y_1 < b,$$

and the p.d.f. of the largest order statistic is

$$g_n(y_n) = n[F(y_n)]^{n-1} f(y_n), \quad a < y_n < b.$$

REMARK: There is one very satisfactory way, based on the multinomial probability, to construct heuristically the expression for the p.d.f. of Y_r. According to the definition of derivative, we have

$$g_r(y_r) = \lim_{\Delta y_r \to 0} \frac{G(y_r + \Delta y_r) - G(y_r)}{\Delta y_r} = \lim_{\Delta y_r \to 0} \frac{P(Y_r \le y_r + \Delta y_r) - P(Y_r \le y_r)}{\Delta y_r}$$

$$= \lim_{\Delta y_r \to 0} \frac{1}{\Delta y_r}\left(\frac{n!}{(r-1)!1!(n\text{-}r)!}[F(y_r)]^{r-1}[F(y_r + \Delta y_r) - F(y_r)][1-F(y_r + \Delta y_r)]^{n-r} \right)$$

$$= \frac{n!}{(r-1)!1!(n\text{-}r)!}[F(y_r)]^{r-1}\left(\lim_{\Delta y_r \to 0} \frac{[F(y_r + \Delta y_r) - F(y_r)]}{\Delta y_r} \right)\left(\lim_{\Delta y_r \to 0}[1-F(y_r + \Delta y_r)]^{n-r} \right)$$

$$= \frac{n!}{(r-1)!\,(n\text{-}r)!}[F(y_r)]^{r-1}[1-F(y_r)]^{n-r} f(y_r).$$

Another interesting heuristic argument can be given, based on the notion that the "likelihood" of an observation is assigned by the p.d.f. To have $Y_r = y_r$, one must have $r-1$ observations less than y_r, one at y_r, and $n-r$ observations greater than y_r, where $P(Y_r \le y_r) = F(y_r)$, $P(Y_r \ge y_r) = 1 - F(y_r)$, and the likelihood of an observation at y_r is $f(y_r)$. There are

$n!/[(r-1)!\,1!\,(n-r)!]$ possible orderings of the n independent observations, and $g_r(y_r)$ is given by the above multinomial expression. This is illustrated in Figure 7.1.

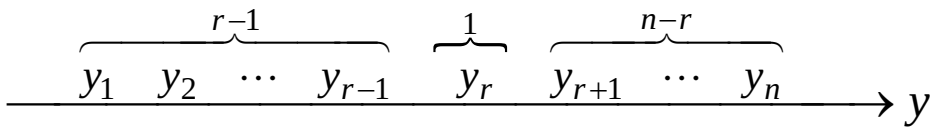

Figure 7.1 The *r*th order observation

A similar argument can be used to easily give the joint p.d.f. of any set of order statistics. For example, consider a pair of order statistics Y_i and Y_j where $i < j$. To have $Y_i = y_i$ and $Y_j = y_j$, one must have $i-1$ observations less than y_i, one at y_i, $j-i-1$ between y_i and y_j, one at y_j, and $n-j$ greater than y_j. Applying the multinomial form gives the joint p.d.f. for Y_i and Y_j as

$$g_{ij}(y_i, y_j) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}[F(y_i)]^{i-1} f(y_i)[F(y_j)-F(y_i)]^{j-i-1}[1-F(y_j)]^{n-j} f(y_j)$$

if $a < y_i < y_j < b$, and zero otherwise. This is illustrated by Figure 7.2. Furthermore, the joint p.d.f. of $Y_1, Y_2, \cdots, Y_n$ is given by

$$g(y_1, y_2, \cdots, y_n) = n!\, f(y_1) f(y_2) \cdots f(y_n)$$

if $a \le y_1 < y_2 < \cdots < y_n \le b$ and zero otherwise.

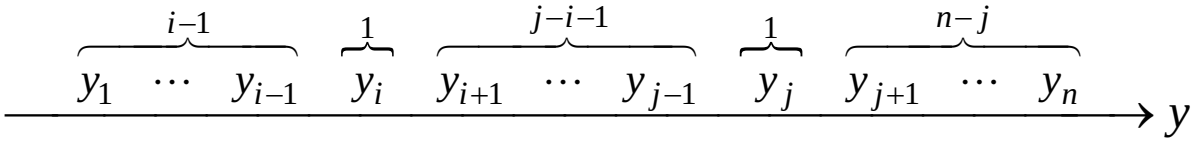

Figure 7.2 The *i*th and *j*th order observations

Example 7.5-1: Consider a random sample of size *n* from a distribution with p.d.f. and CDF given by $f(x) = 2x$ and $F(x) = x^2$; 0 < x < 1. The smallest and largest order statistics are

$$g_1(y_1) = 2n\, y_1(1 - y_1^2)^{n-1}, \qquad 0 < y_1 < 1$$

and

$$g_n(y_n) = 2n\, y_n(y_n^2)^{n-1}, \qquad 0 < y_n < 1.$$

$$E(Y_1) = \int_0^1 2n y_1^2 \left(1 - y_1^2\right)^{n-1} dy_1 \underset{\text{Let } y_1^2 = u \Rightarrow dy_1 = du/\left(2\sqrt{u}\right)}{=} n\int_0^1 u^{(3/2)-1}(1-u)^{n-1} du$$
$$= \frac{\Gamma(3/2)\Gamma(n)}{\Gamma(n+3/2)} n = \frac{\Gamma(3/2)\Gamma(n+1)}{\Gamma(n+3/2)}.$$

Define the range of the sample as $R = Y_n - Y_1$. The joint p.d.f. of Y_1 and Y_n is

$$g_{1,n}(y_1, y_n) = \frac{n!}{(n-2)!}(2y_1)\left(y_n^2 - y_1^2\right)^{n-2}(2y_n), \qquad 0 < y_1 < y_n < 1.$$

Making the transformation $R = Y_n - Y_1$, $S = Y_1$, yields the inverse transformation $y_1 = s$, $y_n = r + s$, and $|J| = 1$. Thus, the joint p.d.f. of R and S is

$$h(r,s) = \frac{n!}{(n-2)!} 4s(r+s)\left(r^2 + 2rs\right)^{n-2}, \qquad 0 < s < 1-r, \quad 0 < r < 1.$$

The marginal density of the range then is given by

$$h_1(r) = \int_0^{1-r} h(r,s)\, ds .$$

For example, for the case $n = 2$, we have

$$h_1(r) = \int_0^{1-r} 8s(r+s)\, ds = (4/3)(r+2)(1-r)^2, \qquad 0 < r < 1. \quad \blacksquare$$

Example 7.5-2: Let $Y_1 < Y_2 < \cdots < Y_7$ be the order statistics of a random sample of size $n = 7$ from a distribution with p.d.f. $f(x) = 3(1-x)^2$, $0 < x < 1$. Compute the probability that the sample median is less than $1 - \sqrt[3]{0.6}$; that is, find $P\left(Y_4 < 1 - \sqrt[3]{0.6}\right)$. We could find the p.d.f. of Y_4. However, note that the probability of a single observation being less than $1 - \sqrt[3]{0.6}$ is

$$\int_0^{1-\sqrt[3]{0.6}} 3(1-x)^2 dx = \left[-(1-x)^3\right]_0^{1-\sqrt[3]{0.6}} = 1 - \left(\sqrt[3]{0.6}\right)^3 = 0.4 .$$

Thus,

$$P\left(Y_4 < 1-\sqrt[3]{0.6}\right) = \sum_{k=4}^{7} \binom{7}{k} (0.4)^k (0.6)^{7-k} = 1 - 0.7102 = 0.2898 . \quad \blacksquare$$

Example 7.5-3: If X has a distribution function $F(x)$ of the continuous type, then $F(x)$ has a uniform distribution on the interval zero to one. If $Y_1 < Y_2 < \cdots < Y_n$ are the order statistics of a random sample $X_1, X_2, \cdots, X_n$ of size n, then

$$F(Y_1) < F(Y_2) < \cdots < F(Y_n)$$

since $F(\cdot)$ is a nondecreasing function and the probability of an equality is again zero. Note that the last display could be looked on as an ordering of the mutually independent random variables $F(Y_1) < F(Y_2) < \cdots < F(Y_n)$, each of which is $U(0,1)$. That is,

$$W_1 = F(Y_1) < W_2 = F(Y_2) < \cdots < W_n = F(Y_n)$$

can be thought of as the order statistics of a random sample of size n from that uniform distribution. Since the distribution function of $U(0,1)$ is $G(w) = w$, $0 < w < 1$, the p.d.f. of the rth order statistic $W_r = F(Y_r)$ is

$$h_r(w) = \frac{n!}{(r-1)!(n-r)!} w^{r-1}(1-w)^{n-r}, \qquad 0 < w < 1.$$

The mean, $E(W_r) = E[F(Y_r)]$ of $W_r = F(Y_r)$, is given by the integral

$$E(W_r) = \int_0^1 w \frac{n!}{(r-1)!(n-r)!} w^{r-1}(1-w)^{n-r} dw$$

$$= \left(\frac{r}{n+1}\right) \int_0^1 w \frac{(n+1)!}{r!(n-r)!} w^{r-1}(1-w)^{n-r} dw$$

$$= \frac{r}{n+1}, \qquad r = 1, 2, \cdots, n. \quad \square$$

There is an extremely interesting interpretation of $W_r = F(Y_r)$. Note that $F(Y_r)$ is the cumulated probability up to and including Y_r—or, equivalently, the area under $f(x) = F'(x)$ but less than Y_r. Hence, $F(Y_r)$ can be treated as a ***random area***. Since $F(Y_{r-1})$ is also a random area, $F(Y_r) - F(Y_{r-1})$ is the random area under $f(x)$ between Y_{r-1} and Y_r. The expected value of the random area between any two adjacent order statistics is then

$$E[F(Y_r) - F(Y_{r-1})] = \frac{r}{n+1} - \frac{r-1}{n+1} = \frac{1}{n+1}.$$

Also, it is easy to show that

$$E[F(Y_1)] = \frac{1}{n+1} \quad \text{and} \quad E[1 - F(Y_n)] = \frac{1}{n+1}.$$

That is, the order statistics $Y_1 < Y_2 < \cdots < Y_n$ partition the support of X into $n + 1$ parts and thus cerate $n + 1$ area under $f(x)$ and above the x-axis. "On the average," each of the $n + 1$ areas equal $1/(n+1)$.

If we recall that the (100p)*th* percentile π_p is such that the area under $f(x)$ to the left of π_p is p, the preceding discussion suggests that we let Y_r be an estimator of π_p, where $p = r/(n+1)$. For this reason, we define the **(100*p*)*th* percentile of the sample** as Y_r, where $r = (n+1)p$. In case $(n+1)p$ is not an integer; we use a weighted average (or an average) of the two adjacent order statistics Y_r and Y_{r+1}, where r is the greatest integer in $(n+1)p$. In particular, the sample median is

$$M = \begin{cases} Y_{(n+1)/2,} & \text{when } n \text{ is odd,} \\ \dfrac{Y_{n/2} + Y_{(n/2)+1}}{2}, & \text{when } n \text{ is even.} \end{cases}$$