5.4 Order Statistics

Sample values such as the smallest, largest, or middle observation from a random sample can provide additional summary information. For example, the highest flood waters or the lowest winter temperature recorded during the last 50 years might be useful data when planning for future emergencies. The median price of houses sold during the previous month might be useful for estimating the cost of living. These are all examples of *order statistics*.

Definition 5.4.1 The *order statistics* of a random sample $X_1, \ldots, X_n$ are the sample values placed in ascending order. They are denoted by $X_{(1)}, \ldots, X_{(n)}$.

The order statistics are random variables that satisfy $X_{(1)} \le \cdots \le X_{(n)}$. In particular,

$$X_{(1)} = \min_{1 \le i \le n} X_i,$$

$$X_{(2)} = \text{second smallest } X_i,$$

$$\vdots$$

$$X_{(n)} = \max_{1 \le i \le n} X_i.$$

Since they are random variables, we can discuss the probabilities that they take on various values. To calculate these probabilities we need the pdfs or pmfs of the order statistics. The formulas for the pdfs of the order statistics of a random sample from a continuous population will be the main topic later in this section, but first, we will mention some statistics that are easily defined in terms of the order statistics.

The *sample range*, $R = X_{(n)} - X_{(1)}$, is the distance between the smallest and largest observations. It is a measure of the dispersion in the sample and should reflect the dispersion in the population.

The *sample median*, which we will denote by M, is a number such that approximately one-half of the observations are less than M and one-half are greater. In terms of the order statistics, M is defined by

$$M = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \left(X_{(n/2)} + X_{(n/2+1)}\right)/2 & \text{if } n \text{ is even.} \end{cases} \tag{5.4.1}$$

The median is a measure of location that might be considered an alternative to the sample mean. One advantage of the sample median over the sample mean is that it is less affected by extreme observations. (See Section 10.2 for details.)

Although related, the mean and median usually measure different things. For example, in recent baseball salary negotiations a major point of contention was the owners' contributions to the players' pension fund. The owners' view could be paraphrased as, "The average baseball player's annual salary is \$433,659 so, with that kind of money, the current pension is adequate." But the players' view was, "Over half of the players make less than \$250,000 annually and, because of the short professional life of most players, need the security of a larger pension." (These figures are for the 1988 season, not the year of the dispute.) Both figures were correct, but the owners were discussing the mean while the players were discussing the median. About a dozen players with salaries over \$2 million can raise the average salary to \$433,659 while the majority of the players make less than \$250,000, including all rookies who make \$62,500. When discussing salaries, prices, or any variable with a few extreme values, the median gives a better indication of "typical" values than the mean. Other statistics that can be defined in terms of order statistics and are less sensitive to extreme values (such as the *α-trimmed* mean discussed in Exercise 10.20) are discussed in texts such as Tukey (1977).

For any number p between 0 and 1, the $(100p)$th sample percentile is the observation such that approximately np of the observations are less than this observation and $n(1-p)$ of the observations are greater. The 50th sample percentile ($p = .5$) is the sample median. For other values of p, we can more precisely define the sample percentiles in terms of the order statistics in the following way.

Definition 5.4.2 The notation $\{b\}$, when appearing in a subscript, is defined to be *the number b rounded to the nearest integer in the usual way.* More precisely, if i is an integer and $i - .5 \le b < i + .5$, then $\{b\} = i$.

The $(100p)$th sample percentile is $X_{(\{np\})}$ if $\frac{1}{2n} < p < .5$ and $X_{(n+1-\{n(1-p)\})}$ if $.5 < p < 1 - \frac{1}{2n}$. For example, if $n = 12$ and the 65th percentile is wanted, we note that $12 \times (1 - .65) = 4.2$ and $12 + 1 - 4 = 9$. Thus the 65th percentile is $X_{(9)}$. There is a restriction on the range of p because the size of the sample limits the range of sample percentiles.

The cases $p < .5$ and $p > .5$ are defined separately so that the sample percentiles exhibit the following symmetry. If the $(100p)$th sample percentile is the ith smallest observation, then the $(100(1-p))$th sample percentile should be the ith largest observation and the above definition achieves this. For example, if $n = 11$, the 30th sample percentile is $X_{(3)}$ and the 70th sample percentile is $X_{(9)}$.

In addition to the median, two other sample percentiles are commonly identified. These are the *lower quartile* (25th percentile) and *upper quartile* (75th percentile). A measure of dispersion that is sometimes used is the *interquartile range*, the distance between the lower and upper quartiles.

Since the order statistics are functions of the sample, probabilities concerning order statistics can be computed in terms of probabilities for the sample. If $X_1, \ldots, X_n$ are iid discrete random variables, then the calculation of probabilities for the order statistics is mainly a counting task. These formulas are derived in Theorem 5.4.3. If $X_1, \ldots, X_n$ are a random sample from a continuous population, then convenient expressions for the pdf of one or more order statistics are derived in Theorems 5.4.4 and 5.4.6. These can then be used to derive the distribution of functions of the order statistics.

Theorem 5.4.3 *Let $X_1, \ldots, X_n$ be a random sample from a discrete distribution with pmf $f_X(x_i) = p_i$, where $x_1 < x_2 < \cdots$ are the possible values of X in ascending*

order. Define

$$P_0 = 0$$
$$P_1 = p_1$$
$$P_2 = p_1 + p_2$$
$$\vdots$$
$$P_i = p_1 + p_2 + \cdots + p_i$$
$$\vdots$$

Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics from the sample. Then

$$P(X_{(j)} \le x_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k} \tag{5.4.2}$$

and

$$P(X_{(j)} = x_i) = \sum_{k=j}^{n} \binom{n}{k} \left[P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k}\right]. \tag{5.4.3}$$

Proof: Fix i, and let Y be a random variable that counts the number of $X_1, \ldots, X_n$ that are less than or equal to x_i. For each of $X_1, \ldots, X_n$, call the event $\{X_j \le x_i\}$ a "success" and $\{X_j > x_i\}$ a "failure." Then Y is the number of successes in n trials. The probability of a success is the same value, namely $P_i = P(X_j \le x_i)$, for each trial, since $X_1, \ldots, X_n$ are identically distributed. The success or failure of the jth trial is independent of the outcome of any other trial, since X_j is independent of the other X_js. Thus, $Y \sim$ binomial(n, P_i).

The event $\{X_{(j)} \le x_i\}$ is equivalent to the event $\{Y \ge j\}$; that is, at least j of the sample values are less than or equal to x_i. Equation (5.4.2) expresses this binomial probability.

$$P(X_{(j)} \le x_i) = P(Y \ge j).$$

Equation (5.4.3) simply expresses the difference.

$$P(X_{(j)} = x_i) = P(X_{(j)} \le x_i) - P(X_{(j)} \le x_{i-1}).$$

The case $i = 1$ is exceptional in that $P(X_{(j)} = x_1) = P(X_{(j)} \le x_1)$. The definition of $P_0 = 0$ takes care of this exception in (5.4.3). □

If $X_1, \ldots, X_n$ are a random sample from a continuous population, then the situation is simplified slightly by the fact that the probability is 0 that any two X_js are equal, freeing us from worrying about ties. Thus $P(X_{(1)} < X_{(2)} < \cdots < X_{(n)}) = 1$ and the sample space for $(X_{(1)}, \ldots, X_{(n)})$ is $\{(x_1, \ldots, x_n) : x_1 < x_2 < \cdots < x_n\}$. In Theorems 5.4.4 and 5.4.6 we derive the pdf for one and the joint pdf for two order statistics, again using binomial arguments.

Theorem 5.4.4 *Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of a random sample, $X_1, \ldots, X_n$, from a continuous population with cdf $F_X(x)$ and pdf $f_X(x)$. Then the pdf of $X_{(j)}$ is*

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x) [F_X(x)]^{j-1} [1 - F_X(x)]^{n-j}. \tag{5.4.4}$$

Proof: We first find the cdf of $X_{(j)}$ and then differentiate it to obtain the pdf. As in Theorem 5.4.3, let Y be a random variable that counts the number of $X_1, \ldots, X_n$ less than or equal to x. Then, defining a "success" as the event $\{X_j \le x\}$, we see that $Y \sim$ binomial$(n, F_X(x))$. (Note that we can write $P_i = F_X(x_i)$ in Theorem 5.4.3. Also, although $X_1, \ldots, X_n$ are continuous random variables, the counting variable Y is discrete.) Thus,

$$F_{X_{(j)}}(x) = P(Y \ge j) = \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k},$$

and the pdf of $X_{(j)}$ is

$$\begin{aligned}
f_{X_{(j)}}(x) &= \frac{d}{dx} F_{X_{(j)}}(x) \\
&= \sum_{k=j}^{n} \binom{n}{k} \Big(k [F_X(x)]^{k-1} [1 - F_X(x)]^{n-k} f_X(x) \\
&\quad - (n-k)[F_X(x)]^k [1 - F_X(x)]^{n-k-1} f_X(x)\Big) && \text{(chain rule)} \\
&= \binom{n}{j} j f_X(x) [F_X(x)]^{j-1} [1 - F_X(x)]^{n-j} \\
&\quad + \sum_{k=j+1}^{n} \binom{n}{k} k [F_X(x)]^{k-1} [1 - F_X(x)]^{n-k} f_X(x) \\
&\quad - \sum_{k=j}^{n-1} \binom{n}{k} (n-k) [F_X(x)]^k [1 - F_X(x)]^{n-k-1} f_X(x) && \binom{k = n \text{ term}}{\text{is 0}} \\
&= \frac{n!}{(j-1)!(n-j)!} f_X(x) [F_X(x)]^{j-1} [1 - F_X(x)]^{n-j} \\
&\quad + \sum_{k=j}^{n-1} \binom{n}{k+1} (k+1) [F_X(x)]^k [1 - F_X(x)]^{n-k-1} f_X(x) && \binom{\text{change}}{\text{dummy}} \\
&\quad - \sum_{k=j}^{n-1} \binom{n}{k} (n-k) [F_X(x)]^k [1 - F_X(x)]^{n-k-1} f_X(x).
\end{aligned} \tag{5.4.5}$$

Noting that

$$\binom{n}{k+1} (k+1) = \frac{n!}{k!(n-k-1)!} = \binom{n}{k} (n-k), \tag{5.4.6}$$

we see that the last two sums in (5.4.5) cancel. Thus, the pdf $f_{X_{(j)}}(x)$ is given by the expression in (5.4.4). □

Example 5.4.5 (Uniform order statistic pdf) Let $X_1, \ldots, X_n$ be iid uniform(0, 1), so $f_X(x) = 1$ for $x \in (0,1)$ and $F_X(x) = x$ for $x \in (0,1)$. Using (5.4.4), we see that the pdf of the jth order statistic is

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} x^{j-1}(1-x)^{n-j} \quad \text{for } x \in (0,1)$$
$$= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} x^{j-1}(1-x)^{(n-j+1)-1}.$$

Thus, the jth order statistic from a uniform(0, 1) sample has a beta$(j, n-j+1)$ distribution. From this we can deduce that

$$\mathrm{E}X_{(j)} = \frac{j}{n+1} \quad \text{and} \quad \mathrm{Var}\, X_{(j)} = \frac{j(n-j+1)}{(n+1)^2(n+2)}. \qquad \|$$

The joint distribution of two or more order statistics can be used to derive the distribution of some of the statistics mentioned at the beginning of this section. The joint pdf of any two order statistics is given in the following theorem, whose proof is left to Exercise 5.26.

Theorem 5.4.6 *Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of a random sample, $X_1, \ldots, X_n$, from a continuous population with cdf $F_X(x)$ and pdf $f_X(x)$. Then the joint pdf of $X_{(i)}$ and $X_{(j)}$, $1 \le i < j \le n$, is*

$$(5.4.7) \quad f_{X_{(i)},X_{(j)}}(u,v) = \frac{n!}{(i-1)!(j-1-i)!(n-j)!} f_X(u) f_X(v) [F_X(u)]^{i-1} \times [F_X(v) - F_X(u)]^{j-1-i} [1 - F_X(v)]^{n-j}$$

for $-\infty < u < v < \infty$.

The joint pdf of three or more order statistics could be derived using similar but even more involved arguments. Perhaps the other most useful pdf is $f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n)$, the joint pdf of all the order statistics, which is given by

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = \begin{cases} n! f_X(x_1) \cdot \cdots \cdot f_X(x_n) & -\infty < x_1 < \cdots < x_n < \infty \\ 0 & \text{otherwise.} \end{cases}$$

The $n!$ naturally comes into this formula because, for any set of values $x_1, \ldots, x_n$, there are $n!$ equally likely assignments for these values to $X_1, \ldots, X_n$ that all yield the same values for the order statistics. This joint pdf and the techniques from Chapter 4 can be used to derive marginal and conditional distributions and distributions of other functions of the order statistics. (See Exercises 5.27 and 5.28.)

We now use the joint pdf (5.4.7) to derive the distribution of some of the functions mentioned at the beginning of this section.

xample 5.4.7 (Distribution of the midrange and range) Let $X_1, \ldots, X_n$ be d uniform$(0, a)$ and let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics. The range was arlier defined as $R = X_{(n)} - X_{(1)}$. The *midrange*, a measure of location like the ample median or the sample mean, is defined by $V = (X_{(1)} + X_{(n)})/2$. We will derive he joint pdf of R and V from the joint pdf of $X_{(1)}$ and $X_{(n)}$. From (5.4.7) we have hat

$$f_{X_{(1)},X_{(n)}}(x_1, x_n) = \frac{n(n-1)}{a^2}\left(\frac{x_n}{a} - \frac{x_1}{a}\right)^{n-2} = \frac{n(n-1)(x_n - x_1)^{n-2}}{a^n}, \quad 0 < x_1 < x_n < a.$$

olving for $X_{(1)}$ and $X_{(n)}$, we obtain $X_{(1)} = V - R/2$ and $X_{(n)} = V + R/2$. The Jacobian for this transformation is -1. The transformation from $(X_{(1)}, X_{(n)})$ to (R, V) maps $\{(x_1, x_n) : 0 < x_1 < x_n < a\}$ onto the set $\{(r, v) : 0 < r < a, r/2 < v < a - r/2\}$. o see this, note that obviously $0 < r < a$ and for a fixed value of r, v ranges from $r/2$ corresponding to $x_1 = 0, x_n = r$) to $a - r/2$ (corresponding to $x_1 = a - r, x_n = a$). hus, the joint pdf of (R, V) is

$$f_{R,V}(r, v) = \frac{n(n-1)r^{n-2}}{a^n}, \quad 0 < r < a, \quad r/2 < v < a - r/2.$$

he marginal pdf of R is thus

$$(5.4.8) \quad f_R(r) = \int_{r/2}^{a-r/2} \frac{n(n-1)r^{n-2}}{a^n}\, dv = \frac{n(n-1)r^{n-2}(a-r)}{a^n}, \quad 0 < r < a.$$

f $a = 1$, we see that R has a beta$(n-1, 2)$ distribution. Or, for arbitrary a, it is easy o deduce from (5.4.8) that R/a has a beta distribution. Note that the constant a is a scale parameter.

The set where $f_{R,V}(r, v) > 0$ is shown in Figure 5.4.1, where we see that the range of integration of r depends on whether $v > a/2$ or $v \le a/2$. Thus, the marginal pdf of V is given by

$$f_V(v) = \int_0^{2v} \frac{n(n-1)r^{n-2}}{a^n} dr = \frac{n(2v)^{n-1}}{a^n}, \quad 0 < v \le a/2,$$

and

$$f_V(v) = \int_0^{2(a-v)} \frac{n(n-1)r^{n-2}}{a^n} dr = \frac{n[2(a-v)]^{n-1}}{a^n}, \quad a/2 < v \le a.$$

his pdf is symmetric about $a/2$ and has a peak at $a/2$. $\|$