Lecture 15: Order Statistics

Mathematical Statistics I, MATH 60061/70061

Thursday October 28, 2021

Reference: Casella & Berger, 5.4

Order statistics

Sample values such as the smallest, largest, or middle observation from a random sample can provide useful summary information.

- The highest snowfall recorded during the last 50 years
- The lowest winter temperature recorded during the last 50 years
- The median price of houses sold during the previous month

These are examples of **order statistics**.

Order statistics

For random variables $X_1, X_2, \ldots, X_n$, the **order statistics** are the random variables $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$, where

$$
\begin{aligned}
&X_{(1)} = \min(X_1, \ldots, X_n), \\
&X_{(2)} \text{ is the second-smallest of } X_1, \ldots, X_n, \\
&\vdots \\
&X_{(n-1)} \text{ is the second-largest of } X_1, \ldots, X_n, \\
&X_{(n)} = \max(X_1, \ldots, X_n).
\end{aligned}
$$

By definition, $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$. We call $X_{(j)}$ the j**th order statistic**. If n is odd, $X_{((n+1)/2)}$ is called the **sample median** of $X_1, \ldots, X_n$.

Functions of order statistics

- Both sample mean and variance are functions of order statistics, since

$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} X_{(i)} \quad \text{and} \quad \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_{(i)}^2$$

- The **sample range** $R = X_{(n)} - X_{(1)}$, is a measure of the *dispersion* in the sample.
- The **sample median**, the 50th sample percentile, is a measure of *location* and is defined by

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

- The **sample lower quartile** is the 25th sample percentile, and the **upper quartile** is the 75th sample percentile.
- The **sample mid-range** is defined as $V = (X_{(1)} + X_{(n)})/2$.

Properties of order statistics

- The order statistics $X_{(1)}, \ldots, X_{(n)}$ are random variables , and each $X_{(j)}$ is a function of $X_1, \ldots, X_n$.
- Even if the original random variables are independent, the order statistics are *dependent*: if we know that $X_{(1)} = 100$, then $X_{(n)}$ must be at least 100.
- The transformation to order statistics is *not invertible*: starting with $\min(X, Y) = 3$ and $\max(X, Y) = 5$, we can't tell whether the original values of X and Y are 3 and 5, respectively, or 5 and 3. Therefore the change of variables formula from $\mathbb{R}^n$ to $\mathbb{R}^n$ does not apply.

If $X_1, \ldots, X_n$ is a random sample of *discrete* random variables, then the calculation of probabilities for the order statistics is mainly a *counting* task.

Distribution of order statistics of discrete RVs

Let $X_1, \ldots, X_n$ be a random sample from a discrete distribution with PMF $f(x_i) = p_i$, where $x_1 < x_2 < \cdots$ are the possible values of X in ascending order. Define

$$
\begin{aligned}
P_0 &= 0 \\
P_1 &= p_1 \\
P_2 &= p_1 + p_2 \\
&\vdots \\
P_i &= p_1 + p_2 + \cdots + p_i \\
&\vdots
\end{aligned}
$$

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

$$P(X_{(j)} \leq x_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k},$$

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

Proof: For any fixed i, let Y be a random variable that counts the number of $X_1, \dots, X_n$ that are less than or equal to x_i.

For each of $X_1, \dots, X_n$, 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 and is distributed as $\mathrm{Bin}(n, P_i)$.

The event $\{X_{(j)} \le x_i\}$ is equivalent to the event $\{Y \ge j\}$, so the result follows from

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

and

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

CDFs of the maximum and the minimum

Suppose $X_1, \dots, X_n$ are i.i.d. and continuous, with CDF F and PDF f. The CDF of $X_{(n)}$ is

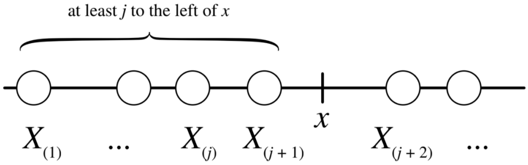
$$\begin{aligned} F_{X_{(n)}}(x) &= P(\max(X_1, \dots, X_n) \leq x) \\ &= P(X_1 \leq x, \dots, X_n \leq x) \\ &= P(X_1 \leq x) \dots P(X_n \leq x) \\ &= (F(x))^n, \end{aligned}$$

where F is the CDF of the individual X_i's. Similarly, the CDF of $X_{(1)}$ is

$$\begin{aligned} F_{X_{(1)}}(x) &= 1 - P(\min(X_1, \dots, X_n) > x) \\ &= 1 - P(X_1 > x, \dots, X_n > x) \\ &= 1 - (1 - F(x))^n. \end{aligned}$$

Distribution of order statistics of continuous RVs

For the event $X_{(j)} \leq x$ to occur, we need at least j of the X_i's to fall to the left of x:

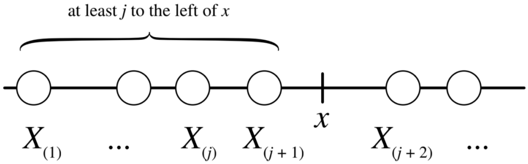


Let N be the number of X_i's that land to the left of x.

- Each X_i lands to the left of x with probability $F(x)$, independently.
- There are n independent Bernoulli trials with probability $F(x)$ of success (landing to the left of x), so $N \sim \mathrm{Bin}(n, F(x))$.

Distribution of order statistics of continuous RVs

For the event $X_{(j)} \leq x$ to occur, we need at least j of the X_i's to fall to the left of x:

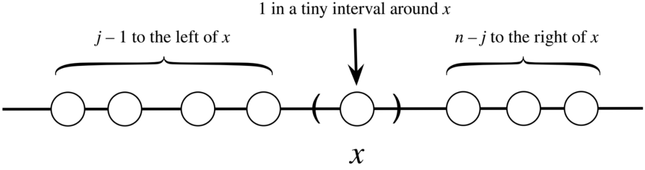


The CDF of the jth order statistic $X_{(j)}$ is

$$\begin{aligned}
F_{X_{(j)}}(x) &= P(X_{(j)} \leq x) \\
&= P(N \geq j) \\
&= \sum_{k=j}^{n} \binom{n}{k} F(x)^k (1 - F(x))^{n-k}.
\end{aligned}$$

Distribution of order statistics of continuous RVs

Consider $f_{X_{(j)}}(x)dx$, the probability that the jth order statistic falls into an infinitesimal interval of length dx around x:

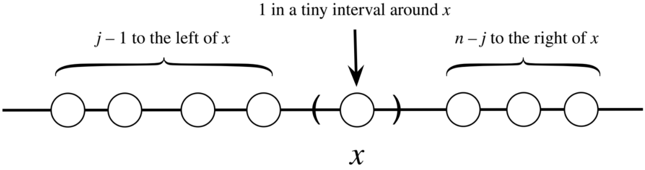


- One of the X_i's falls into the infinitesimal interval around x
- Exactly $j-1$ of the X_i's fall to the left of x
- The remaining $n-j$ fall to the right of x

$$f_{X_{(j)}}(x)dx = nf(x)dx\binom{n-1}{j-1}F(x)^{j-1}(1-F(x))^{n-j}$$

Distribution of order statistics of continuous RVs

Consider $f_{X_{(j)}}(x)dx$, the probability that the jth order statistic falls into an infinitesimal interval of length dx around x:



Let $X_1, \dots, X_n$ are i.i.d. continuous R.V.s. with CDF F and PDF f. The the marginal PDF of the jth order statistic $X_{(j)}$ is

$$\begin{aligned} f_{X_{(j)}}(x) &= n\binom{n-1}{j-1} f(x)F(x)^{j-1}(1-F(x))^{n-j} \\ &= \frac{n!}{(j-1)!(n-j)!} f(x)F(x)^{j-1}(1-F(x))^{n-j}. \end{aligned}$$

Order statistics of Uniforms

Let $X_1, \ldots, X_n$ are i.i.d. $\mathrm{Unif}(0,1)$. Then for $0 \leq x \leq 1$, $f(x) = 1$ and $F(x) = x$, so the PDF of $X_{(j)}$ is

$$\begin{aligned} f_{X_{(j)}}(x) &= \frac{n!}{(j-1)!(n-j)!} f(x) F(x)^{j-1} (1 - F(x))^{n-j} \\ &= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} x^{j-1} (1-x)^{(n-j+1)-1}. \end{aligned}$$

This is the $\mathrm{Beta}(j, n-j+1)$ PDF. So $X_{(j)} \sim \mathrm{Beta}(j, n-j+1)$, and from this we know

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

Joint PDF of two order statistics

Let $X_{(1)}, \dots, X_{(n)}$ be the order statistics of a random sample $X_1, \dots, X_n$ from a continuous population with CDF F and PDF f. Then the joint PDF of $X_{(i)}$ and $X_{(j)}$, $1 \le i < j \le n$, is

$$\begin{aligned} f_{X_{(i)},X_{(j)}}(u, v) = & \frac{n!}{(i-1)!(j-i-1)!(n-j)!} f(u) f(v) \\ & \times [F(u)]^{i-1} [F(v) - F(u)]^{j-i-1} [1 - F(v)]^{n-j} \end{aligned}$$

for $-\infty < x < y < \infty$.

Distribution of the midrange and range

Let $X_1, \ldots, X_n$ be a random sample from $\mathrm{Unif}(0, a)$, $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics, $R = X_{(n)} - X_{(1)}$ be the range, and $V = (X_{(1)} + X_{(n)})/2$ be the midrange.

We want to find the joint PDF of R and V as well as the marginal PDFs of R and V.

Distribution of the midrange and range

Let $X_1, \ldots, X_n$ be a random sample from $\mathrm{Unif}(0, a)$, $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics, $R = X_{(n)} - X_{(1)}$ be the range, and $V = (X_{(1)} + X_{(n)})/2$ be the midrange.

We want to find the joint PDF of R and V as well as the marginal PDFs of R and V.

The joint PDF of $X_{(1)}$ and $X_{(n)}$ is

$$f_{X_{(1)},X_{(n)}}(z, y) = \frac{n(n-1)}{a^2}\left(\frac{y}{a} - \frac{z}{a}\right)^{n-2} = \frac{n(n-1)(y-z)^{n-2}}{a^n},$$

for $0 < z < y < a$.

Since $R = X_{(n)} - X_{(1)}$ and $V = (X_{(1)} + X_{(n)})/2$, we obtain $X_{(1)} = V - R/2$ and $X_{(n)} = V + R/2$, and the Jacobian for this transformation is -1.

Joint PDF of the midrange and range

The transformation from $(X_{(1)}, X_{(n)})$ to (R, V) maps

$\{(z, y) : 0 < z < y < a\} \to \{(r, v) : 0 < r < a, r/2 < v < a{-}r/2\}$.

For a fixed r,

- the smallest value of v is $r/2$ (when $z = 0$ and $y = r$), and
- the largest value of v is $a - r/2$ (when $z = a - r$ and $y = a$).

Thus, the joint PDF of R and V is

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

for $0 < r < a,\ r/2 < v < a - r/2$.

Marginal PDFs of the midrange and range

The marginal PDF of the range R is

$$
\begin{aligned}
f_R(r) &= \int_{r/2}^{a-r/2} \frac{n(n-1)r^{n-2}}{a^n} dv \\
&= \frac{n(n-1)r^{n-2}(a-r)}{a^n}, \quad 0 < r < a
\end{aligned}
$$

The marginal PDF of the midrange V is

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

and

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