STAT/MTHE 353: 2 – Order Statistics

T. Linder

Queen's University

Winter 2017

Order Statistics

Let $X_1, \ldots, X_n$ be jointly continuous random variables.

Definition The kth order statistic of $X_1, \ldots, X_n$ is the kth smallest value of the X_i and denoted by $X_{(k)}$. Thus

$$X_{(1)} = \min(X_1, \ldots, X_n), \qquad X_{(n)} = \max(X_1, \ldots, X_n),$$

The vector $(X_{(1)}, \ldots, X_{(n)})$ is called the order statistics of $(X_1, \ldots, X_n)$.

- We will assume that $X_1, \ldots, X_n$ are *independent and identically distributed* (i.i.d.) random variables. Thus their joint pdf is

  $$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f(x_1)f(x_1)\cdots f(x_n)$$

  where f is the *common* marginal pdf of the X_i.
- We will determine the joint pdf of $X_{(1)}, \ldots, X_{(n)}$.

If $f_{1\ldots n}$ is the joint pdf of $X_{(1)}, \ldots, X_{(n)}$, then (with $h > 0$)

$$\begin{aligned} f_{1\ldots n}(x_1, \ldots, x_n) &= \lim_{h\to 0} \frac{1}{h^n} \int_{x_i}^{x_i+h} \cdots \int_{x_n}^{x_n+h} f_{1\ldots n}(t_1, \ldots, t_n)\, dt_1 \cdots dt_n \\ &= \lim_{h\to 0} \frac{1}{h^n} P\big(x_1 \le X_{(1)} \le x_1 + h, \ldots, x_n \le X_{(n)} \le x_n + h\big) \end{aligned}$$

(true for any sufficiently well behaved pdf).

Clearly, we only need consider the case $x_1 < x_2 < \cdots < x_n$. (Why?) Let S_n denote the set of permutations of $\{1, \ldots, n\}$. Note that $|S_n| = n!$. Since $X_1, \ldots, X_n$ are jointly continuous, the $n!$ *disjoint* events

$$A_\sigma = \{X_{\sigma(1)} < X_{\sigma(2)} < \cdots < X_{\sigma(n)}\}, \quad \sigma \in S_n$$

have total probability 1, i.e.,

$$P\Big(\bigcup_{\sigma\in S_n} A_\sigma\Big) = 1.$$

Define the events

$$A = \{x_1 \le X_{(1)} \le x_1 + h, \ldots, x_n \le X_{(n)} \le x_n + h\}$$

and

$$B_\sigma = \{x_1 \le X_{\sigma(1)} \le x_1 + h, \ldots, x_n \le X_{\sigma(n)} \le x_n + h\}$$

and note that $A \cap A_\sigma = B_\sigma$ if $h > 0$ is small enough. Thus for such h

$$\begin{aligned} &P\big(x_1 \le X_{(1)} \le x_1 + h, \ldots, x_n \le X_{(n)} \le x_n + h\big) \\ &= P(A) = \sum_{\sigma\in S_n} P\big(A \cap A_\sigma\big) = \sum_{\sigma\in S_n} P\big(B_\sigma\big) \\ &= \sum_{\sigma\in S_n} \prod_{i=1}^n P\big(x_i \le X_{\sigma(i)} \le x_i + h\big) \quad \text{(since the } X_i \text{ are independent)} \\ &= \sum_{\sigma\in S_n} \prod_{i=1}^n \Big[F_{X_{\sigma(i)}}(x_i + h) - F_{X_{\sigma(i)}}(x_i)\Big] \\ &= n! \prod_{i=1}^n \Big[F(x_i + h) - F(x_i)\Big] \quad (F \text{ is the common cdf of the } X_i) \end{aligned}$$

We obtain

$$
\begin{aligned}
& f_{1\ldots n}(x_1,\ldots,x_n) \\
&\quad = \lim_{h\to 0}\frac{1}{h^n}P\big(x_1 \le X_{(1)} \le x_1+h,\ldots,x_n \le X_{(n)} \le x_n+h\big) \\
&\quad = \lim_{h\to 0}\frac{1}{h^n}n!\prod_{i=1}^{n}\Big[F(x_i+h)-F(x_i)\Big] \\
&\quad = n!\prod_{i=1}^{n}\lim_{h\to 0}\frac{F(x_i+h)-F(x_i)}{h} \\
&\quad = n!f(x_1)f(x_2)\cdots f(x_n).
\end{aligned}
$$

In conclusion, the joint pdf of the order statistics $X_{(1)},\ldots,X_{(n)}$ is given by

$$
f_{1\ldots n}(x_1,\ldots,x_n) = \begin{cases} n!f(x_1)f(x_2)\cdots f(x_n) & \text{if } x_1 < x_2 < \cdots < x_n, \\ 0 & \text{otherwise.} \end{cases}
$$

Marginal pdfs of order statistics

Notation: For $k < r$ we let $f_{k\cdots r}(x_k\cdots x_r)$ denote the marginal jpdf of $X_{(k)},\ldots,X_{(r)}$. Here we always assume $x_k < x_{k+1} < \cdots < x_r$; otherwise $f_{k\cdots r}(x_k\cdots x_r) = 0$.

"Integrating out" x_1 we get

$$
\begin{aligned}
f_{2\cdots n}(x_2,\ldots,x_n) &= \int_{-\infty}^{x_2} n!f(x_1)f(x_2)\cdots f(x_n)\,dx_1 \\
&= n!F(x_2)f(x_2)\cdots f(x_n)
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
f_{3\cdots n}(x_3,\ldots,x_n) &= \int_{-\infty}^{x_3} n!F(x_2)f(x_2)\cdots f(x_n)\,dx_2 \\
&= n!\frac{F(x_3)^2}{2}f(x_3)\cdots f(x_n)
\end{aligned}
$$

Integrating out $x_3,\ldots,x_{k-1}$, we obtain in general for all $x_k < \cdots < x_n$

$$
\boxed{f_{k\cdots n}(x_k,\ldots,x_n) = n!\frac{F(x_k)^{k-1}}{(k-1)!}f(x_k)\cdots f(x_n)}
$$

If we integrate out x_n first, we obtain

$$
\begin{aligned}
f_{1\cdots n-1}(x_1,\ldots,x_{n-1}) &= \int_{x_{n-1}}^{\infty} n!f(x_1)f(x_2)\cdots f(x_n)\,dx_n \\
&= n!f(x_1)\cdots f(x_{n-1})\big(1-F(x_{n-1})\big)
\end{aligned}
$$

Continuing with $x_{n-1},x_{n-2},\ldots,x_{r+1}$, we get for all $x_1 < \cdots < x_r$

$$
\boxed{f_{1\cdots r}(x_1,\ldots,x_r) = n!f(x_1)\cdots f(x_r)\frac{\big(1-F(x_r)\big)^{n-r}}{(n-r)!}}
$$

Integrating out $x_n,x_{n-1},\ldots,x_{r+1}$ in $f_{k\cdots n}(x_k,\ldots,x_n)$ or $x_1,x_2,\ldots,x_{k-1}$ in $f_{1\cdots r}(x_1,\ldots,x_r)$ we finally obtain for all $x_k < \cdots < x_r$

$$
\boxed{f_{k\cdots r}(x_k,\ldots,x_r) = n!\frac{F(x_k)^{k-1}}{(k-1)!}f(x_k)\cdots f(x_r)\frac{\big(1-F(x_r)\big)^{n-r}}{(n-r)!}}
$$

Setting $r = k+1$ we get for all $x_k < x_{k+1}$

$$
f_{k,k+1}(x_k,x_{k+1}) = n!\frac{F(x_k)^{k-1}}{(k-1)!}f(x_k)f(x_{k+1})\frac{\big(1-F(x_{k+1})\big)^{n-k-1}}{(n-k-1)!}
$$

From this we obtain the marginal pdf of $X_{(k)}$:

$$
\begin{aligned}
f_k(x_k) &= n!\frac{F(x_k)^{k-1}}{(k-1)!}f(x_k)\int_{x_k}^{\infty} f(x_{k+1})\frac{\big(1-F(x_{k+1})\big)^{n-k-1}}{(n-k-1)!}\,dx_{k+1} \\
&= n!\frac{F(x_k)^{k-1}}{(k-1)!}f(x_k)\left[-\frac{\big(1-F(x_{k+1})\big)^{n-k}}{(n-k)!}\right]_{x_k}^{\infty} \\
&= n!\frac{F(x_k)^{k-1}}{(k-1)!}f(x_k)\frac{\big(1-F(x_k)\big)^{n-k}}{(n-k)!}
\end{aligned}
$$

We obtained (Theorem 9.5 in text):

$$
\boxed{f_k(x) = \frac{n!}{(k-1)!(n-k)!}f(x)F(x)^{k-1}\big(1-F(x)\big)^{n-k}}
$$

Finally, if we integrate out $x_{k+1}, \ldots, x_{r-1}$ in $f_{k\cdots r}$, then we obtain for all $x_k < x_r$ (Theorem 9.6):

$$\boxed{\begin{aligned} &f_{k,r}(x_k, x_r) \\ &= n! \frac{f(x_k)F(x_k)^{k-1}}{(k-1)!} \cdot \frac{\big(F(x_r) - F(x_k)\big)^{r-k-1}}{(r-k-1)!} \cdot \frac{f(x_r)\big(1 - F(x_r)\big)^{n-r}}{(n-r)!} \end{aligned}}$$

Special cases: minimum and maximum

From f_k with $k = 1$ we get the pdf of $X_{(1)} = \min(X_1, \ldots, X_n)$:

$$\boxed{f_1(x) = nf(x)\big(1 - F(x)\big)^{n-1}}$$

For $k = n$ we get the pdf of $X_{(n)} = \max(X_1, \ldots, X_n)$:

$$\boxed{f_n(x) = nf(x)F(x)^{n-1}}$$

Example: The i.i.d. random variables $X_1, X_2, \ldots, X_n$ are often called a *random sample* from the common distribution of the X_i. The *range* of the random sample is $R = X_{(n)} - X_{(1)}$. Determine the pdf of R. Specialize to the case where $X_i \sim \text{Uniform}(0,1)$.

Solution: ...

Application: Sample median

- The *median* of the distribution of a r.v. X is any value m such that

$$P(X \le m) \ge \frac{1}{2} \quad \text{and} \quad P(X \ge m) \ge \frac{1}{2}.$$

  Equivalently, if F is the cdf of X,

$$F(m) \ge \frac{1}{2} \quad \text{and} \quad 1 - F(m^-) \ge \frac{1}{2}.$$

- The *sample median* of the random sample $X_1, \ldots, X_n$ from a continuous distribution is

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

- The sample median is often taken to be an estimate of the median, and it is sometimes preferred to the *sample mean* $\frac{1}{n}\sum_{i=1}^{n} X_i$ as an estimate of the "center of the distribution" because it is more robust to "outliers."

Example: ...