Order Statistics

Statistics 110

Summer 2006

Order Statistics and Extrema

Another problem of interest has to do with ordering of values. Lets assume that $X_1, X_2, \ldots, X_n$ are an iid sample from a distribution with density f and CDF F.

Let $U = \max(X_1, X_2, \ldots, X_n)$ and $V = \min(X_1, X_2, \ldots, X_n)$

What are the distributions of these two RVs?

When determining these its easier to deal with the CDFs than the densities.

Let

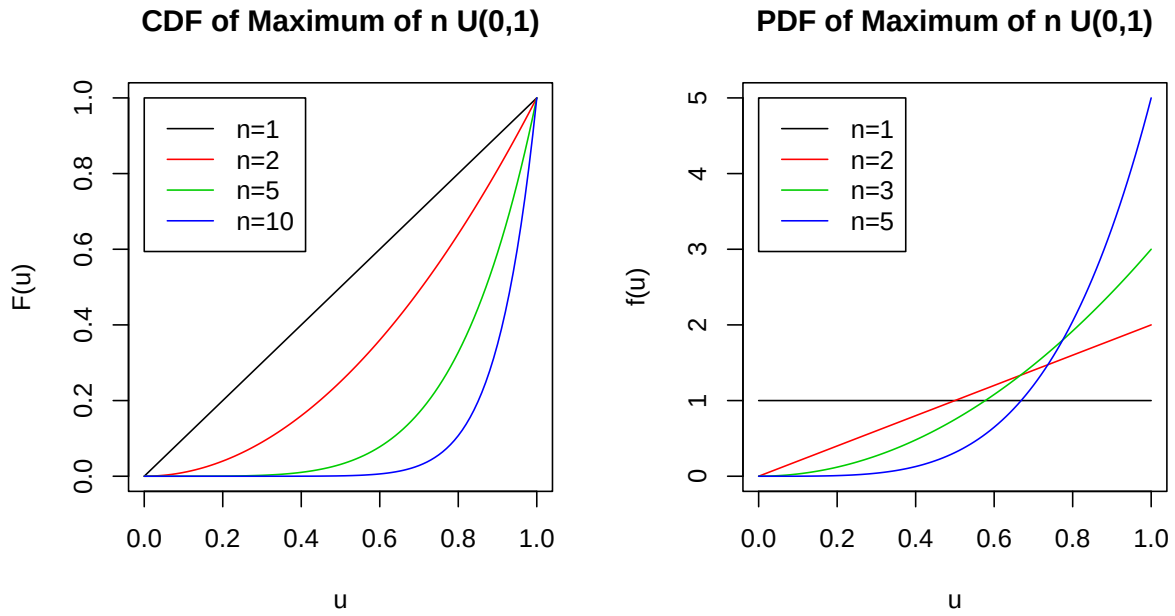
$$\begin{aligned} A_u &= \{U \leq u\} \qquad (\text{the maximum is at most } u) \\ &= \{X_1 \leq u\} \cap \{X_2 \leq u\} \cap \ldots \cap \{X_n \leq u\} \end{aligned}$$

So

$$F_U(u) = P[U \leq u] = P[A_u] = \prod_{i=1}^{n} P[X_i \leq u] = [F(u)]^n$$

Therefore the density is

$$f_U(u) = \frac{d}{du}[F(u)]^n = nf(u)[F(u)]^{n-1}$$



Both of these plots imply, not surprisingly, the more items you take the maximum of, the bigger the maximum tends to be.

Similarly for the minimum

$$\begin{aligned} B_v &= \{V \geq v\} \qquad \text{(the minimum is at least } u\text{)} \\ &= \{X_1 \geq v\} \cap \{X_2 \geq v\} \cap \ldots \cap \{X_n \geq v\} \end{aligned}$$

So

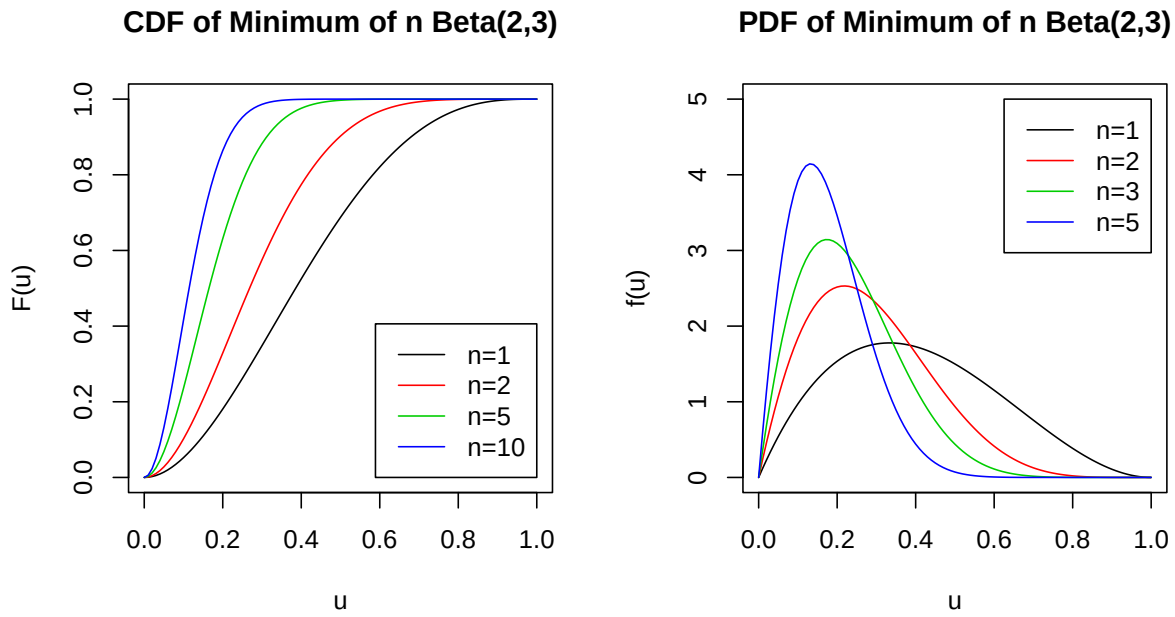
$$P[V \geq v] = P[B_v] = \prod_{i=1}^{n} P[X_i \geq v] = [1 - F(v)]^n$$

giving a CDF of

$$F_V(v) = 1 - [1 - F(v)]^n$$

and a density of

$$f_V(v) = \frac{d}{dv}[1 - F(v)]^n = nf(v)[1 - F(v)]^{n-1}$$



Similarly, the more items you take the minimum of, the smaller the minimum will tend to be.

If the distributions aren't identically distributed, but are still independent, the CDFs for the maximum and minimum are

$$F_U(u) = P[U \leq u] = P[A_u] = \prod_{i=1}^{n} P[X_i \leq u] = \prod_{i=1}^{n} [F_i(u)]$$

$$P[V \geq v] = P[B_v] = \prod_{i=1}^{n} P[X_i \geq v] = \prod_{i=1}^{n} [1 - F_i(v)]$$

giving a CDF of

$$F_V(v) = 1 - \prod_{i=1}^{n} [1 - F_i(v)]$$

where F_i is the CDF for X_i.

You can determine the densities in this case, but they aren't particularly nice. For example, the density of the maximum is

$$f_U(u) = F_U(u) \sum_{i=1}^{n} \frac{f_i(u)}{F_i(u)}$$

Let

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$$

be the sorted values (known as order statistics). Note that $X_{(i)}$ is not necessarily X_i. In fact they will be the same with probability of $\frac{1}{n}$.

So the sample maximum is $U = X_{(n)}$ and the sample minimum is $V = X_{(1)}$. If n is odd, (say $n = 2m + 1$), then $X_{(m+1)}$ is the sample median.

What are the distribution of these order statistics in general? For this let us assume that the X's are iid again.

Let

$$\begin{aligned} C_{k,z} &= \{X_{(k)} \leq z\} \\ &= \{\text{Exactly } k \text{ of the } X_i \leq z\} \cup \{\text{Exactly } k+1 \text{ of the } X_i \leq z\} \\ &\quad \cup \ldots \cup \{\text{Exactly } n \text{ of the } X_i \leq z\} \end{aligned}$$

Thus the CDF of $X_{(k)}$ is

$$\begin{aligned} F_{X_{(k)}}(z) &= P[X_{(k)} \leq z] = P[C_{k,z}] \\ &= \sum_{j=k}^{n} P[\text{Exactly } j \text{ of the } X_i \leq z] \\ &= \sum_{j=k}^{n} \binom{n}{j} [F(z)]^j (1 - F(z))^{n-j} \end{aligned}$$

One way to get the density in this case is to differentiate this. However the differential argument is easier is this case

Define the event

$$\begin{aligned} D_{k,x} &= \{x \leq X_{(k)} \leq x + \Delta_x\} \\ &= \{\text{Exactly } k-1 \text{ of the } X_i \leq x\} \\ &\quad \cap \ \{\text{Exactly 1 of the } X_i \text{ in } [x, x+\Delta_x]\} \\ &\quad \cap \ \{\text{Exactly } n-k \text{ of the } X_i \geq x + \Delta_x\} \end{aligned}$$

$$P[D_{k,x}] \approx \binom{n}{k-1\ 1\ n-k} [F(x)]^{k-1} f(x) \Delta_x (1 - F(x + \Delta_x))^{n-k}$$

So the density of $X_{(k)}$ is given by

$$\lim_{\Delta_x \to 0} \frac{P[D_{k,x}]}{\Delta_x} = \frac{n!}{(k-1)!(n-k)!} f(x) [F(x)]^{k-1} (1 - F(x))^{n-k}$$

By a similar arguments, the joint distribution of order statistics can be determined. For example, the joint density of $X_{(j)}$ and $X_{(k)}$, with $j < k$ is

$$f_{X_{(j)},X_{(k)}}(x,y) = \frac{n!}{(j-1)!(k-j-1)!(n-k)!} f(x)f(y) \times [F(x)]^{j-1}[F(y)-F(x)]^{k-j-1}[1-F(y)]^{n-k}; x < y$$

For example, the joint density of the maximum (u) and minimum (v) is

$$f_{X_{(1)},X_{(n)}}(u,v) = n(n-1)f(u)f(v)[F(u)-F(v)]^{n-2}; \quad u < v$$

From this we can determine the distribution of the range ($R = X_{(n)} - X_{(1)}$) and the midpoint ($M = \frac{X_{(n)}+X_{(1)}}{2}$)

For example, the marginal density of the range is

$$f_R(r) = \int_{-\infty}^{\infty} n(n-1)f(v+r)f(v)[F(v+r) - F(v)]^{n-2}dv$$

while the marginal density of the midpoint is

$$f_M(m) = \int_{-\infty}^{\infty} 2n(n-1)f(u)f(2m-u)[F(2m-u) - F(u)]^{n-2}du$$

The joint density of all of the order statistics is

$$f_{X_{(1)},X_{(2)},\ldots,X_{(n)}}(y_1, y_2, \ldots, y_n) = n!f(y_1)f(y_2)\ldots f(y_n); y_1 < y_2 < \ldots < y_n$$

If the distribution of the X's aren't iid, i.e. they aren't identically distributed, aren't independent, or both, determining the densities of the order statistics is difficult. You need to determine the density of each possible order separately. For example to determine the density of the sample median of 3 observations you need to consider the 6 possible orders of observations

$$\begin{array}{lll} x_1 < x_2 < x_3 & x_2 < x_1 < x_3 & x_3 < x_1 < x_2 \\ x_1 < x_3 < x_2 & x_2 < x_3 < x_1 & x_3 < x_2 < x_1 \end{array}$$

This then leads to the density of the median (in the independence case)

$$\begin{aligned} f(z) = {} & F_1(z)f_2(z)(1-F_3(z)) + F_1(z)f_3(z)(1-F_2(z)) \\ & + F_2(z)f_1(z)(1-F_3(z)) + F_2(z)f_3(z)(1-F_1(z)) \\ & + F_3(z)f_1(z)(1-F_2(z)) + F_3(z)f_2(z)(1-F_1(z)) \end{aligned}$$

Needless to say, this can get ugly looking very quickly.