STAT 512: Statistical Inference **Autumn 2020**

Lecture 9: Order Statistics from a Continuous Univariate Distribution

Instructor: Yen-Chi Chen

Let $X_1, \cdots, X_n$ be IID continuous R.V.'s with a PDF $p_X(x)$ and a CDF $F_X(x)$. Since they are continuous R.V.s, we assume that they all take distinct values. The order statistics $Y_1 < Y_2 < \cdots < Y_n$ are the ordered version of these n random variables such that Y_j is the j-th smallest values among $\{X_1, \cdots, X_n\}$. Thus,

$$
\begin{aligned}
Y_1 &= \min\{X_1, \cdots, X_n\}\\
Y_n &= \max\{X_1, \cdots, X_n\}.
\end{aligned}
$$

Sometimes, we use the notation $X_{(j)} = Y_j$. An interesting note: the mapping

$$(X_1, \cdots, X_n) \to (Y_1, \cdots, Y_n)$$

is not 1-1 but (n!)-1. This is due to the fact that any of the $n!$ permutation among $X_1, \cdots, X_n$ will lead to the same order statistics.

The PDF of Y_i's are associated with the PDF of X_i's:

- **Distribution of Y_j.** Using the fact that

$$p_{Y_j}(y)dy \approx P(y \leq Y_j \leq y + dy)$$

and the event

$$\{y \leq Y_j \leq y + dy\}$$

is approximately the same as

$\{(i-1)$ of X_i's are below y and $(n-i)$ of X_i's are above $y + dy$ and one X_i falls within $[y, y + dy]\}$.

So we conclude

$$
\begin{aligned}
p_{Y_j}(y)dy &\approx P(y \leq Y_j \leq y + dy)\\
&= \binom{n}{j-1} F_X(y)^{j-1} \binom{n-j+1}{n-j} (1 - F_X(y))^{n-j} p_X(y)dy\\
&= \frac{n!}{(j-1)!(n-j)!} F_X(y)^{j-1}(1 - F_X(y))^{n-j} p_X(y)dy.
\end{aligned}
$$

Thus,

$$p_{Y_j}(y) \approx \frac{n!}{(j-1)!(n-j)!} F_X(y)^{j-1}(1 - F_X(y))^{n-j} p_X(y).$$

- **Distribution of Y_j, Y_ℓ.** WLOG, we assume $j < \ell$. Using the fact that

$$p_{Y_j,Y_\ell}(y, z)dydz \approx P(y \leq Y_j \leq y + dy, z \leq Y_\ell \leq z + dz)$$

and the event

$$\{y \leq Y_j \leq y + dy, z \leq Y_\ell \leq z + dz\}$$

is approximated by the event that

1. $(i-1)$ X_i's are below y,
2. one X_i is within $[y, y+dy]$,
3. $\ell - j - 1$ X_i's are between $(y+dy, z)$,
4. one X_i is within $[z, z+dz]$,
5. the remaining $n-\ell$ X_i's are above $z+dx$.

The probability of the above event is about

$$\frac{n!}{(j-1)!1(\ell-j-1)!1(n-\ell)!}F_X(y)^{j-1}p_X(y)dy(F_X(z)-F_X(y))^{(\ell-j-1)}p_X(z)dz(1-F_X(z))^{n-\ell}.$$

Thus,

$$p_{Y_j,Y_\ell}(y,z) \approx \frac{n!}{(j-1)!1(\ell-j-1)!1(n-\ell)!}F_X(y)^{j-1}p_X(y)(F_X(z)-F_X(y))^{(\ell-j-1)}p_X(z)(1-F_X(z))^{n-\ell}.$$

You can generalize this method to the joint distribution of more order statistics.

- **Distribution of** $(Y_1, \cdots, Y_n)$**.** On this extreme end, you can apply the same procedure and you will end up with
$$p(y_1, \cdots, y_n) = n! p_X(y_1) \cdots p_X(y_n).$$

9.1 Case study: uniform distribution

Consider the case where $X_1, \cdots, X_n$ are IID from $\mathsf{Uni}[0,1]$. Then $p_X(x) = 1$ and $F_X(x) = x$ when $x \in [0,1]$. Thus,

$$p_{Y_j}(y) = \frac{n!}{(j-1)!(n-j)!}y^{j-1}(1-y)^{n-j},$$

which is the PDF of $\mathsf{Beta}(j, n-j+1)$.

Here is an interest note about the variance. The variance of Y_j is

$$\mathsf{Var}(Y_j) = \mathsf{Var}(X_j) = \frac{j(n-j+1)}{(n+1)^2(n+2)},$$

which is maximized when $j = \frac{n+1}{2}$ assuming n is an odd number. The value $j = \frac{n+1}{2}$ corresponds to the 'median' of $\{X_1, \cdots, X_n\}$. Thus, the median has the highest variability. In this case,

$$\mathsf{Var}(Y_{\frac{n+1}{2}}) = \frac{1}{4(n+2)} = O(n^{-1}).$$

On the other hand, the maximal or minimal value has the lowest variance:

$$\mathsf{Var}(Y_1) = \mathsf{Var}(Y_n) = \frac{n}{(n+1)^2(n+2)} = O(n^{-2}).$$

Now we consider another way to look at the order statistics. Let $W_1, \cdots, W_n, W_{n+1}$ be the 'spacing' between consecutive order statistics:

$$\begin{aligned}
W_1 &= Y_1 - 0 \\
W_2 &= Y_2 - Y_1 \\
W_3 &= Y_3 - Y_2 \\
&\vdots \\
W_n &= Y_n - Y_{n-1} \\
W_{n+1} &= 1 - Y_n.
\end{aligned}$$

It is easy to see that $W_i \in [0,1]$ and $W_1 + W_2 + \cdots + W_{n+1} = 1$. Also, we can reparametrize Y_j via W_i's:

$$Y_j = W_1 + W_2 + \cdots + W_j.$$

Since X_i's are uniform over $[0,1]$, the joint PDF of $Y_1, \cdots, Y_n$ is

$$p(y_1, \cdots, y_n) = n!$$

whenever $0 < y_1 < \cdots < y_n < 1$. By the Jacobian method with the fact that $\mathsf{det}(\frac{dY}{dW}) = 1$ (think about why), we conclude that

$$p(w_1, \cdots, w_n) = n!$$

whenever $w_i \in [0,1]$ and $w_1 + \cdots + w_n < 1$. One can easily see that $p(w_1, \cdots, w_n)$ is invariant under the permutation of $W_1, \cdots, W_n$ (i.e., they are *exchangeable*), so the marginal distribution of W_i is the same as the marginal distribution of W_j for all $i, j = 1, \cdots, n$. Because $W_1 = Y_1$ follows from $\mathsf{Beta}(1, n)$, we conclude that W_j is a $\mathsf{Beta}(1, n)$ random variable.

Note that W_i and W_j are dependent ($i \neq j$)! Due to the exchangeability property, the joint distribution (W_i, W_j) is the same as the joint distribution of W_1, W_2, so

$$\begin{aligned}
\mathsf{Cov}(W_i, W_j) &= \mathsf{Cov}(W_1, W_2) \\
&= \frac{1}{2}\left(\mathsf{Var}(W_1 + W_2) - \mathsf{Var}(W_1) - \mathsf{Var}(W_2)\right) \\
&= \frac{1}{2}\left(\mathsf{Var}(Y_2) - 2\mathsf{Var}(Y_1)\right) \\
&= \frac{1}{2}\left(\frac{2(n-1)}{(n+1)^2(n+2)} - 2\frac{n)}{(n+1)^2(n+2)}\right) \\
&= \frac{-1}{(n+1)^2(n+2)} < 0.
\end{aligned}$$