MATH 556: MATHEMATICAL STATISTICS I

ORDER STATISTICS, SAMPLE QUANTILES AND RANKS

For n independent random variables $X_1, \ldots, X_n$, the **order statistics** $X_{(1)}, \ldots, X_{(n)}$ are defined by

$$X_{(i)} - \text{"the } i\text{th smallest value in } X_1, \ldots, X_n \text{ for } i = 1, \ldots, n\text{"}$$

It is sometimes notationally more convenient to write Y_i instead of $X_{(i)}$. For distribution with cdf F_X,

(a) $Y_1 \equiv X_{(1)} = \min\{X_1, \ldots, X_n\}$ has cdf

$$\begin{aligned}
F_{Y_1}(y) = P_{Y_1}[Y_1 \leq y] = 1 - P_{Y_1}[Y_1 > y] &\\
&= 1 - P_{X_1,\ldots,X_n}[\min\{X_1, \ldots, X_n\} > y]\\
&= 1 - P_{X_1,\ldots,X_n}\left[\bigcap_{i=1}^{n}(X_i > y)\right]\\
&= 1 - \prod_{i=1}^{n} P_{X_i}[X_i > y]\\
&= 1 - \prod_{i=1}^{n}\{1 - F_X(y)\} = 1 - \{1 - F_X(y)\}^n
\end{aligned}$$

(b) $Y_n \equiv X_{(n)} = \max\{X_1, \ldots, X_n\}$ has cdf

$$\begin{aligned}
F_{Y_n}(y) = P_{Y_n}[Y_n \leq y] &= P_{X_1,\ldots,X_n}[\max\{X_1, \ldots, X_n\} \leq y]\\
&= P_{X_1,\ldots,X_n}\left[\bigcap_{i=1}^{n}(X_i \leq y)\right]\\
&= \prod_{i=1}^{n} P_{X_i}[X_i \leq y]\\
&= \prod_{i=1}^{n}\{F_X(y)\} = \{F_X(y)\}^n
\end{aligned}$$

If the probability that two Xs are identical is zero (that is, **ties** are avoided), the **joint pdf** of order statistics $Y_1, \ldots, Y_n$ is

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = n! f_X(y_1) \ldots f_X(y_n) \qquad y_1 < \ldots < y_n$$

as there are $n!$ configurations of the xs that yield identical order statistics. From first principles, using the joint cdf and the Theorem of Total Probability, and independence of $X_1, \ldots, X_n$,

$$\begin{aligned}
F_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) &= \sum_{\rho} P_{X_1,\ldots,X_n}[X_{\rho(1)} \leq y_1, \ldots, X_{\rho(n)} \leq y_n]\\
&= \sum_{\rho}\left\{\prod_{i=1}^{n} F_{X_{\rho(i)}}(y_i)\right\} = n! F_{X_1}(y_1) \cdots F_{X_n}(y_n)
\end{aligned}$$

where the sum is over all the $n!$ permutations ρ of $\{1, \ldots, n\}$ that account for the different configurations of the original sample that can give rise to identical collections of order statistics. The result follows on differentiating with respect to $y_1, \ldots, y_n$.

For the **marginal** distribution of $Y_j = X_{(j)}$, $j = 1, \ldots, n$, note that

$$F_{Y_j}(y) = P_{Y_j}[Y_j \leq y] \equiv P_Z[Z \geq j]$$

where Z is the number (out of n) of Xs that do not exceed y; this is true as the event "$Y_j \leq y$" means that there are **at least** j of the Xs that do not exceed y. We have that $Z \sim Binomial(n, F_X(y))$, so

$$P_Z[Z \geq j] = \sum_{k=j}^{n} \binom{n}{k} \{F_X(y)\}^k \{1 - F_X(y)\}^{n-k}$$

(a) In the **discrete** case, suppose that $\mathbb{X} \equiv \{c_1, c_2, \ldots\}$, where $c_1 < c_2 < \cdots$, and suppose that

$$f_X(c_i) = p_i \qquad\qquad P_i = \sum_{k=1}^{i} p_k$$

$i = 1, 2, \ldots$. Then the marginal cdf of $Y_j = X_{(j)}$ is defined by

$$F_{Y_j}(c_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k} \qquad c_i \in \mathbb{X}$$

with the usual cdf behaviour elsewhere. The marginal pmf of $Y_j = X_{(j)}$ is

$$f_{Y_j}(c_i) = \sum_{k=j}^{n} \binom{n}{k} \left[P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k} \right] \qquad c_i \in \mathbb{X}$$

and zero otherwise.

(b) In the **continuous** case, the marginal cdf of $Y_j = X_{(j)}$ is

$$F_{Y_j}(y) = \sum_{k=j}^{n} \binom{n}{k} \{F_X(y)\}^k \{1 - F_X(y)\}^{n-k}$$

and hence by differentiation, the marginal pdf is

$$f_{Y_j}(y) = \frac{n!}{(j-1)!(n-j)!} \{F_X(y)\}^{j-1} \{1 - F_X(y)\}^{n-j} f_X(y).$$

To see this heuristically, if the jth order statistic is at y, then we have

(i) a single observation at y, which contributes $f_X(y)$;
(ii) $j - 1$ observations which have values **less than** y, which contributes $\{F_X(y)\}^{j-1}$;
(iii) $n - j$ observations which have values **greater than** y, which contributes $\{1 - F_X(y)\}^{n-j}$;

Thus the required mass/density is proportional to

$$\{F_X(y)\}^{j-1} f_X(y) \{1 - F_X(y)\}^{n-j}.$$

The combinatorial term is the number of ways of labelling the original y values to obtain this configuration of order statistics: this is

$$n \times \binom{n-1}{j-1} = \frac{n!}{(j-1)!(n-j)!}$$

we choose the single X in step (i) in n ways, and then the $j - 1$ Xs in step (ii) in $\binom{n-1}{j-1}$ ways.

Sample Quantiles: Let $0 \leq p \leq 1$. The pth **quantile** of distribution F, $x_F(p)$, is defined by

$$x_F(p) = \inf\{x \,:\, F(x) \geq p\}$$

where $\inf$ is the infimum, or greatest lower bound, that is, $x_F(p)$ is the smallest x value such that $F(x) \geq p$. The **median** is $x_F(0.5)$. The pth **sample quantile** is defined in terms of the order statistics, but there are many possible variants. In general, the pth sample quantile derived from a sample of size n can be defined

$$\widetilde{X}_n(p) = (1 - \gamma(n))X_{(k)} + \gamma(n)X_{(k+1)}$$

for some $\gamma(n)$ where $0 \leq \gamma(n) \leq 1$ is some function of n to be specified, and k is the integer such that $k/n \leq p < (k+1)/n$. One simple definition uses the kth order statistic, $\widetilde{X}_n(p) = X_{(k)}$, where $k = [np]$ is the nearest integer to np. The **sample median** is most commonly defined by

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

Ranks: In the continuous case, for $X_1, \ldots, X_n$, the **rank** of X_i, R_i is defined by

$$R_i = \sum_{j=1}^{n} \mathbb{1}_{(-\infty, X_i]}(X_j)$$

that is, the number of observations that are no greater than X_i. Thus $R_i = r \iff X_{(r)} = X_i$. If ties are possible, different rank measures may be used.