556: MATHEMATICAL STATISTICS I

ORDER STATISTICS AND SAMPLE QUANTILES

For n random variables $X_1, \ldots, X_n$, the **order statistics**, $Y_1, \ldots, Y_n$, are defined by

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

for $i = 1, \ldots, n$. For example

$$Y_1 = X_{(1)} = \min\{X_1, \ldots, X_n\} \qquad\qquad Y_n = X_{(n)} = \max\{X_1, \ldots, X_n\}.$$

Now let $0 \leq p \leq 1$. Recall that the pth **quantile** of a distribution F is denoted by $x_F(p)$ 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

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

For random sample $X_1, \ldots, X_n$ from population with pmf/pdf f_X and cdf F_X,

(a) $Y_1 = X_{(1)}$ has cdf $F_{Y_1}(y) = 1 - \{1 - F_X(y)\}^n$;

(b) $Y_n = X_{(n)}$ has cdf $F_{Y_n}(y) = \{F_X(y)\}^n$

For the marginal cdf for Y_1,

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

For Y_n,

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

The pmf/pdf can be computed from the cdf.

- **Joint distribution:** For random sample $X_1, \ldots, X_n$ from population with pdf f_X, the joint pdf of order statistics $Y_1, \ldots, Y_n$

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

  There are $n!$ configurations of the xs that yield identical order statistics, and the result follows by the Theorem of Total Probability.

- **Marginal distribution:** For random sample $X_1, \ldots, X_n$ from population with pmf/pdf f_X and cdf F_X,

  (a) In the **discrete** case, suppose that $\mathbb{X} \equiv \{x_1, x_2, \ldots\}$, where $x_1 < x_2 < \cdots$, and suppose that

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

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

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

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

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

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

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

  and the marginal pdf is

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

  To see this in the continuous case, if the jth order statistic is at x, then we have

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

  Thus the required mass/density is proportional to

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

  The normalizing constant is the number of ways of labelling the original x values to obtain this configuration of order statistics: this is

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

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