Definition

- For any $X_1, \ldots, X_n$ the **order statistics** are the ordered sample values denoted $X_1 \leq X_2 \cdots X_n$.
- $X_1 = \min_{1 \leq i \leq n} X_i$ and $X_n = \max_{1 \leq i \leq n} X_i$
- If $n = 2m + 1$ the **median** is $X_{(m+1)}$; if $n = 2m$ the median is $X_{(m)}$
- For $X_1, \ldots, X_n$ with a common density function $f(x)$ the joint density function of $X_1, \ldots, X_n$ is

$$f(y_1, \ldots, y_n) = n! f(y_1) \cdots f(y_n) I_{y_1 < y_2 < \cdots < y_n}$$
Example

Let $U_1, \ldots, U_n$ be $Unif[0, 1]$

Then, $f(u_1, \ldots, u_n) = n! I_{0 < u_1 < u_2 < \cdots < u_n < 1}$
How to obtain a marginal distribution

- You want to obtain a marginal distribution of $X_{(1)}$
- In the uniform example, the correct domain of integration is
  \[ u_1 < u_2 < \cdots < u_n < 1 \]
- The marginal density of $U_{(1)}$ is
  \[
f_1(u_{(1)}) = n! \int_{u_1}^1 \int_{u_2}^1 \cdots \int_{u_{n-1}}^1 du_n du_{n-1} \cdots du_3 du_2 = n!(1-u_1)^{n-1}\]
  for any $0 < u_1 < 1$
- The marginal density of $U_{(n)}$ is
  \[
f_n(u_{(n)}) = n! \int_{0}^{u_n} \int_{0}^{u_{n-1}} \cdots \int_{0}^{u_2} du_1 du_2 \cdots du_{n-1} = nu_n^{n-1}\]
  for any $0 < u_n < 1$
Two special cases

- Note that the max and min are two special cases where the distribution can be obtained much easier.

- E.g. for the maximum

\[
F_{U(n)}(u) = P(U_{(n)} \leq u) = \prod_{i=1}^{n} P(X_i \leq u) = u^n
\]

- Thus, the density function is

\[
f_n(u) = nu^{n-1}
\]

for any \(0 < u < 1\)

- Likewise, for the minimum, the survival function is

\[
F_1(u) = P(U_{(1)} \geq u) = (1 - u)^n
\]

- The density is, then,

\[
f_1(u) = [1 - (1 - u)^n]' = n(1 - u)^{n-1}
\]

for \(0 < u < 1\)
Two general formulas

- If the support of the density is \((a, b)\) we have for any \(x \in (a, b)\)

\[
f_k(y) = \frac{n!}{(k - 1)!(n - k)!}[F(y_k)]^{k-1}[1 - F(y_k)]^{n-k}f(y_k)
\]

and 0 otherwise

- For any two order statistics, their joint density is

\[
f_{r,s} = \frac{n!}{(r - 1)!(n - s)!(s - r - 1)!}F^{r-1}(u)(1 - F(\nu))^{n-s}
\]

\[
(F(\nu) - F(s))^{s-r-1}f(u)f(\nu)
\]

for any \(a < u < \nu < b\) and 0 otherwise
Moments of the uniform order statistics

- For $U_1, \ldots, U_n$

  \[ E(U_{(1)}) = \frac{1}{n+1}, \quad E(U_{(n)}) = \frac{n}{n+1} \]

- \[ \text{Var}(U_{(1)}) = \text{Var}(U_{(n)}) = \frac{n}{(n+1)^2(n+2)} \]

- Also, $1 - U_{(n)}$ has the same distribution as $U_{(1)}$

- Finally,

  \[ \text{Cov}(U_{(1)}, U_{(n)}) = \frac{1}{(n+1)^2(n+2)} > 0 \]
Let $X \sim F(x)$; for any $0 < p < 1$ define the quantile 
\[ \xi_p = F^{-1}(p) \]

Example: if $p = 0.5$, $\xi_{0.5}$ is the median of $X$

This quantile is the population quantity and needs to be estimated...

Assume the sample $X_1, \ldots, X_n$ and let $k$ be the greatest integer less than or equal to $p(n + 1)$: 
\[ k = \lfloor p(n + 1) \rfloor \]

Seems sensible to estimate $\xi_p$ with $X_{(k)}$:
\[ \hat{\xi}_p = X_{(k)} \]

$X_{(k)}$ is called the $p$th sample quantile or the $100p$th percentile of the sample
Is this a really sensible way of estimation

- Since the area under the pdf $f(x)$ to the left of $X(k)$ is $F(X(k))$

$$E[F(X(k))] = \int_a^b F(X(k))g_k(X(k)) \, dX(k)$$

- Using the marginal pdf expression for the $k$th order statistics, one can find out that

$$E[F(X(k))] = \frac{k}{n + 1}$$
A five number summary and a boxplot

- A **five number summary** consists of the minimum, first and third quartiles, the median, and the maximum of the sample.
- Its graphical form is the **boxplot**.
- The box encloses the middle 50% of the data and a line segment is typically used to indicate the median.
- Of course, extreme order statistics are very sensitive to outlying points...
First, let \( h = 1.5(Q_3 - Q_1) \)

Define the lower fence (LF) as

\[
LF = Q_1 - h
\]

and the upper fence (UF) as

\[
UF = Q_3 + h
\]

Points that are outside the fences are called potential outliers and denoted as “0” on the boxplot.

The whiskers then protrude from the side of the box to so-called adjacent points that are the points within fences but closest to them.
Exact confidence intervals for quantiles

- By definition $\xi_p$ is a solution of $F(\xi_p) = p$ for any $0 < p < 1$
- Define the integer $k = \lfloor p(n + 1) \rfloor$ and let $Y_1 = X(1), \ldots, Y_n = X(n)$
- Clearly, $Y_k$ is a point estimator of $\xi_p$...
- For any $i < \lfloor (n + 1)p \rfloor < j$, the event $Y_i < \xi_p < Y_j$ is equivalent to obtaining between $i$ and $j$ successes in $n$ independent trials with probability of success $P(X < \xi_p) = F(\xi_p) = p$
- Thus,
  \[ P(Y_i < \xi_p < Y_j) = \sum_{l=i}^{n-1} \binom{n}{l} p^l (1 - p)^{n-l} \]
- Specific values of $y_i$ and $y_j$ make up a $100\gamma\%$ confidence interval for $\xi_p$ where $\gamma = P(Y_i < \xi_p < Y_j)$