Nonlinear Signal Processing
ELEG 833

Gonzalo R. Arce

Department of Electrical and Computer Engineering
University of Delaware
arce@ee.udel.edu

Fall 2008
Outline

1. Order Statistics
   - Distributions Of Order Statistics
   - Moments Of Order Statistics
     - Order Statistics From Uniform Distributions
   - Order Statistics Containing Outliers
Order Statistics

If the random variables $X_1, X_2, \ldots, X_N$ are arranged in ascending order of magnitude such that

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(N)},$$

we denote $X_{(i)}$ as the $i$th order statistic for $i = 1, \ldots, N$.

The extremes $X_{(N)}$ and $X_{(1)}$, are useful tools in the detection of outliers. The range $X_{(N)} - X_{(1)}$ is a quick estimator of the dispersion.
### Outline

1. **Order Statistics**
   - Distributions Of Order Statistics
   - Moments Of Order Statistics
     - Order Statistics From Uniform Distributions
   - Order Statistics Containing Outliers
Distributions Of Order Statistics

For continuous, independent and identically distributed (i.i.d.) samples, the density of the \( r \)th order statistic is formed as follows. First, decompose the event that \( x < X(r) \leq x + dx \) into three exclusive parts:

A) that \( r - 1 \) of the samples \( X_i \) are less than or equal to \( x \)

B) that one is between \( x \) and \( x + dx \)

C) that \( N - r \) are greater than \( x + dx \).

\[
\begin{array}{ccc}
  \hline
  r-1 & 1 & N-r \\
  x & x+dx & \\
  \hline
\end{array}
\]

(a)

\[
\begin{array}{ccc}
  \hline
  r-1 & 1 & s-r-1 & 1 \\
  x & x+dx & y & y+dy \\
  \hline
\end{array}
\]

(b)
A) The probability that \( N - r \) are greater than or equal to \( x + \, dx \) is 
\[
[1 - F(x + \, dx)]^{N-r}
\]

B) The probability that one is between \( x \) and \( x + \, dx \) is \( f_x(x) \, dx \)

C) The probability that \( r - 1 \) are less than or equal to \( x \) is \( F(x)^{r-1} \)

The probability of having more than one sample in \((x, x + \, dx]\) is on the order of \((dx)^2\) and is negligible as \(dx\) approaches zero.

Counting all enumerations of \( N \) samples in the three respective groups:

\[
f_{(r)}(x) \, dx = Pr \left[ x < X_{(r)} \leq x + \, dx \right]
= \frac{N!}{(r - 1)! \,(N - r)!} F(x)^{r-1} \left[1 - F(x)\right]^{N-r} f_x(x) \, dx.
\]
The trinomial coefficient above follows from the multinomial coefficient. Given a set of \( N \) objects, \( k_1 \) labels of type 1, \( k_2 \) labels of type 2, \( \cdots \), and \( k_m \) labels of type \( m \) where \( k_1 + k_2 + \cdots + k_m = N \), the number of ways in which we may assign the labels to the \( N \) objects is given by

\[
\frac{N!}{k_1! \, k_2! \cdots k_m!}.
\]  

(1)

The trinomial coefficient in (1) is a special case of (1) with \( k_1 = r - 1 \), \( k_2 = 1 \), and \( k_3 = N - r \).
Order Statistics

Distributions Of Order Statistics

\[\begin{array}{c|c|c}
\text{r-1} & 1 & \text{N-r} \\
\hline
x & x+dx & \\
\end{array}\]

(a)

\[\begin{array}{c|c|c|c}
\text{r-1} & 1 & \text{s-r-1} & 1 \\
\hline
x & x+dx & y & y+dy \\
\end{array}\]

(b)

**Figure:** The event \(x < X_{(r)} \leq x + dx\) and \(y < X_{(s)} \leq y + dy\) can be seen as \(r - 1\) of the samples \(X_i\) are less than \(x\), that one of the samples is between \(x\) and \(x + dx\), that \(s - r - 1\) of the samples \(X_i\) are less than \(y\) but greater than \(x\), that one of the samples is between \(y\) and \(y + dy\), and finally that \(N - s\) of the samples are greater than \(y\).
For $1 \leq r < s \leq N$ and $x \leq y$, $f_{(r,s)}(x, y)$ is obtained by decomposing the event

$$x < X(r) \leq x + dx < y < X(s) \leq y + dy$$

into five mutually exclusive parts:

A) That $r - 1$ of the samples $X_i$ are less than $x$
B) That one of the samples is between $x$ and $x + dx$
C) That $s - r - 1$ of the samples $X_i$ are less than $y$ but greater than $x + dx$
D) That one of the samples is between $y$ and $y + dy$
E) That $N - s$ of the samples are greater than $y + dy$. 
The probability of occurrence for each of the five listed parts is

A) \( F(x)^{r-1} \)

B) \( f_x(x) \, dx \)

C) \( [F(y) - F(x + dx)]^{s-r-1} \)

D) \( f_x(y) \, dy \)

E) \( [1 - F(y + dy)]^{N-s} \)
Using the multinomial counting principle to enumerate all possible occurrences in each part, and the fact that $F(x + dx) \sim F(x)$ and $F(y + dy) \sim F(y)$ as $dx \to 0$ we obtain the joint density function

$$f_{r,s}(x,y) = \frac{N!}{(r-1)! (s-r-1)! (N-s)!} F(x)^{r-1} f_x(x)$$

$$[F(y) - F(x)]^{s-r-1} f_x(y) [1 - F(y)]^{N-s}.$$ 

These density functions are only valid for continuous random variables.
For continuous and discontinuous distributions: let the i.i.d. variables $X_1, X_2, \ldots, X_N$ have a parent distribution $F(x)$, the pdf of $X_{(N)}$ is

$$F_{(N)}(x) = Pr\{X_{(N)} \leq x\} = Pr\{\text{all } X_i \leq x\} = Pr\{\text{all } X_i \leq x\} = [F(x)]^N.$$  

due to independence. Similarly, the pdf of the minimum sample $X_{(1)}$ is

$$F_{(1)}(x) = Pr\{X_{(1)} \leq x\} = 1 - Pr\{X_{(1)} > x\} = 1 - Pr\{\text{all } X_i > x\} = 1 - [1 - F(x)]^N,$$

since $X_{(1)}$ is less than, or equal to, all the samples in the set.
The distribution function for the general case is

\[
F_{(r)}(x) = Pr\{X_{(r)} \leq x\} \\
= Pr\{\text{at least } r \text{ of the } X_i \text{ are less than or equal to } x\} \\
= \sum_{i=r}^{N} Pr\{\text{exactly } i \text{ of the } X_i \text{ are less than or equal to } x\} \\
= \sum_{i=r}^{N} \binom{N}{i} [F(x)]^i[1 - F(x)]^{N-i}.
\]
The joint distribution function $F_{(r,s)}(x,y)$ of $X_{(r)}$ and $X_{(s)}$, for $1 \leq r < s \leq N$, is (for $x < y$)

$$F_{(r,s)}(x,y) = Pr\{\text{at least } r \text{ of the } X_i \leq x, \text{ at least } s \text{ of the } X_i \leq y \}$$

$$= \sum_{j=s}^{N} \sum_{i=r}^{j} Pr\{\text{exactly } i \text{ of } X_1, X_2 \ldots, X_n \text{ are at most } x$$

and exactly $j$ of $X_1, X_2 \ldots, X_n$ are at most $y\}$$

$$= \sum_{j=s}^{N} \sum_{i=r}^{j} \frac{N!}{i!(j-i)!(N-j)!} [F(x)]^i [F(y) - F(x)]^{j-i}$$

$$\times [1 - F(y)]^{N-j}. \quad (3)$$

Notice that for $x \geq y$, the ordering $X_{(r)} < x$ with $X_{(s)} \leq y$, implies that $F_{(r,s)}(x,y) = F_{(s)}(y)$. 

Gonzalo R. Arce (Department of Electrical and Computer Engineering, University of Delaware, arce@ee.udel.edu)
OUTLINE

1 ORDER STATISTICS
   • Distributions Of Order Statistics
   • Moments Of Order Statistics
     • Order Statistics From Uniform Distributions
   • Order Statistics Containing Outliers
Moments of order statistics are defined in the same fashion as moments of arbitrary random variables. Here we always assume that the sample size is $N$. The expected value of the $r$th order statistic is denoted as $\mu_{(r)}$ and is found as

\[
\mu_{(r)} = \int_{-\infty}^{\infty} x f_{(r)}(x) \, dx = \frac{N!}{(r-1)! (N-r)!} \int_{-\infty}^{\infty} x F(x)^{r-1} [1 - F(x)]^{N-r} f_x(x) \, dx.
\]
The statistical characteristics of the order-statistics $X_{(1)}, X_{(2)}, \ldots, X_{(N)}$ are not homogeneous since

$$EX_{(r)} \neq EX_{(s)}$$

for $r \neq s$, as expected since $E\{X_{(r)}\}$ should be less than $E\{X_{(r+1)}\}$. 
In general, the expectation of products of order statistics are not symmetric

\[ E(X_{(r)}X_{(r+s)}) \neq E(X_{(r)}X_{(r-s)}). \]  

(4)

This symmetry only holds in very special cases. One such case is when the parent distribution is symmetric and where \( r = (N + 1)/2 \) such that \( X_{(r)} \) is the median. The covariance of \( X_{(r)} \) and \( X_{(s)} \) is written as

\[ \text{cov} \left[ X_{(r)}X_{(s)} \right] = E \left\{ (X_{(r)} - \mu_{(r)}) (X_{(s)} - \mu_{(s)}) \right\}. \]

(5)

The covariance of order statistics satisfies \( \text{cov}[X_{(r)}X_{(s)}] \geq 0 \).
Consider $N$ samples of a standard uniform distribution with density function $f_u(u) = 1$, for $0 \leq u \leq 1$. The density function of the $r$th order-statistic $U_{(r)}$ is

$$f_{(r)}(u) = \frac{N!}{(r-1)! (N-r)!} u^{r-1} (1 - u)^{N-r}$$

in the interval $0 \leq u \leq 1$. The mode of the density function can be found at $(r - 1)/(N - 1)$.
The $k$th moment of $U_{(r)}$ is found from the above as

$$
\mu_{(r)}^{(k)} = \int_0^1 u^k f_{(r)}(u) \, du
= \frac{N!}{(r-1)! (N-r)!} \int_0^1 u^k u^{r-1} (1-u)^{N-r} \, du
= B(r+k, N-r+1)/B(r, N-r+1),
$$

where we make use of the complete beta function

$$
B(p, q) = \int_0^1 t^{p-1}(1-t)^{q-1} \, dt
$$

for $p, q > 0$. Simplifying leads to

$$
\mu_{(r)}^{(k)} = \frac{N! \ (r+k-1)!}{(N+k)! \ (r-1)!}.
$$

In particular, the first moment of the $r$th order statistic is

$$
\mu_{(r)}^{(1)} = r/(N+1).
$$
**Figure:** Density functions of $X_{(2)}$, $X_{(3)}$, $X_{(6)}$ (median), $X_{(9)}$, and $X_{(10)}$ for a set of eleven uniformly distributed samples.
Outline

1. Order Statistics
   - Distributions of Order Statistics
   - Moments of Order Statistics
     - Order Statistics from Uniform Distributions
   - Order Statistics Containing Outliers
The $N$ sample set consist of $N - 1$ i.i.d. variates $X_i, i = 1, \ldots, N - 1$, and the contaminant variable $Y$ which is also independent. Let $F(x)$ and $G(x)$ be the continuous parent distributions of $X_i$ and $Y$. Furthermore, let

$$Z_{(1):N} \leq Z_{(2):N} \leq \cdots \leq Z_{(N):N}$$

be the order statistics obtained by arranging the $N$ samples in increasing order of magnitude.
The distribution of the maxima denoted as $H_{(N)}(x)$ is

$$H_{(N)}(x) = Pr \{ \text{all of } X_1, \ldots, X_{N-1}, \text{ and } Y \leq x \} = F(x)^{N-1} \cdot G(x).$$

The distribution of the $ith$ order statistic, for $1 < i \leq N - 1$ is

$$H_{(i)}(x) = Pr \{ \text{at least } i \text{ of } X_1, X_2, \ldots, X_{N-1}, Y \leq x \}$$

$$= Pr \{ \text{exactly } i - 1 \text{ of } X_1, X_2, \ldots, X_{N-1} \leq x \text{ and } Y \leq x \} + Pr \{ \text{at least } i \text{ of } X_1, X_2, \ldots, X_{N-1} \leq x \}$$

$$= \binom{N-1}{i-1} (F(x))^{i-1} (1 - F(x))^{N-i} G(x) + F_{(i)}(N-1)(x)$$

where $F_{(i)}(N-1)(x)$ is the distribution of the $ith$ order statistic in a sample of size $N - 1$ drawn from a parent distribution $F(x)$. 
The density function of $Z_{(i):N}$ can be obtained by differentiating the above or by direct derivation which is left as an exercise:

$$h_{(i):N}(x) = \frac{(N-1)!}{(i-2)!(N-i)!} (F(x))^{i-2}(1 - F(x))^{N-i} G(x)f(x)$$

$$+ \frac{(N-1)!}{(i-1)!(N-i)!} (F(x))^{i-1}(1 - F(x))^{N-i} g(x)$$

$$+ \frac{(N-1)!}{(i-1)!(N-i-1)!} (F(x))^{i-1}(1 - F(x))^{N-i-1}(1 - G(x))f(x)$$

where the first term drops out if $i = 1$, and the last term if $N = i$. 
Densities of $Z_{(2)}$, $Z_{(6)}$ (median), and $Z_{(10)}$ for a sample set of 11, laplacian random variables. In the contaminated case, one the mean of one sample is shifted to 20.

**Figure:** Density functions of $Z_{(2)}$, $Z_{(6)}$ (median), and $Z_{(10)}$ with (solid) and without contamination (dotted).