Andrew Gard - equitable.equations@gmail.com

Order Statistics

Order statistics are just the values obtained when a random sample is arranged from smallest to largest.

Order statistics are just the values obtained when a random sample is arranged from smallest to largest. More specifically, if $X_1, \ldots, X_n$ are the observations, then the order statistics $Y_1, \ldots, Y_n$ are defined like this:

$$\begin{aligned} Y_1 &= \text{smallest of the } X_i \\ Y_2 &= \text{second smallest of the } X_i \\ &\vdots \\ Y_n &= \text{largest of the } X_i \end{aligned}$$

Order statistics are just the values obtained when a random sample is arranged from smallest to largest. More specifically, if $X_1, \ldots, X_n$ are the observations, then the order statistics $Y_1, \ldots, Y_n$ are defined like this:

$$\begin{aligned} Y_1 &= \text{smallest of the } X_i \\ Y_2 &= \text{second smallest of the } X_i \\ &\vdots \\ Y_n &= \text{largest of the } X_i \end{aligned}$$

In particular, Y_1 the sample minimum and Y_n is the sample maximum. If n is odd, then $Y_{\frac{1}{2}(n+1)}$ is the sample median.

Order statistics are just the values obtained when a random sample is arranged from smallest to largest. More specifically, if $X_1, \ldots, X_n$ are the observations, then the order statistics $Y_1, \ldots, Y_n$ are defined like this:

$$\begin{aligned} Y_1 &= \text{smallest of the } X_i \\ Y_2 &= \text{second smallest of the } X_i \\ &\vdots \\ Y_n &= \text{largest of the } X_i \end{aligned}$$

In particular, Y_1 the sample minimum and Y_n is the sample maximum. If n is odd, then $Y_{\frac{1}{2}(n+1)}$ is the sample median.

Order statistics can be used to perform nonparametric statistical inference, for instance when an assumption of normality wouldn't be justified.

Example. Let $X_1, \dots, X_5$ be a simple random sample from $U(0,1)$, the uniform distribution on the interval $(0,1)$. Compute the cumulative distribution functions for the order statistics $Y_1, \dots, Y_5$.

Example. Let $X_1, \ldots, X_5$ be a simple random sample from $U(0,1)$, the uniform distribution on the interval $(0,1)$. Compute the cumulative distribution functions for the order statistics $Y_1, \ldots, Y_5$.

The simplest to compute is Y_5, so we'll start with that.

Example. Let $X_1, \ldots, X_5$ be a simple random sample from $U(0,1)$, the uniform distribution on the interval $(0,1)$. Compute the cumulative distribution functions for the order statistics $Y_1, \ldots, Y_5$.

The simplest to compute is Y_5, so we'll start with that. The key observation is that for any $y \in (0,1)$, $Y_5 \leq y$ exactly when none of the five X_i exceed y.

Example. Let $X_1, \ldots, X_5$ be a simple random sample from $U(0,1)$, the uniform distribution on the interval $(0,1)$. Compute the cumulative distribution functions for the order statistics $Y_1, \ldots, Y_5$.

The simplest to compute is Y_5, so we'll start with that. The key observation is that for any $y \in (0,1)$, $Y_5 \le y$ exactly when none of the five X_i exceed y. That is,

$$G_5(y) \;=\; P(Y_5 \le y) \;=\; P(\max\{X_i\} \le y)$$

Example. Let $X_1, \ldots, X_5$ be a simple random sample from $U(0,1)$, the uniform distribution on the interval $(0,1)$. Compute the cumulative distribution functions for the order statistics $Y_1, \ldots, Y_5$.

The simplest to compute is Y_5, so we'll start with that. The key observation is that for any $y \in (0,1)$, $Y_5 \leq y$ exactly when none of the five X_i exceed y. That is,

$$\begin{aligned} G_5(y) &= P(Y_5 \leq y) = P(\max\{X_i\} \leq y) \\ &= P(\text{all } X_i \leq y) \end{aligned}$$

Example. Let $X_1, \ldots, X_5$ be a simple random sample from $U(0,1)$, the uniform distribution on the interval $(0,1)$. Compute the cumulative distribution functions for the order statistics $Y_1, \ldots, Y_5$.

The simplest to compute is Y_5, so we'll start with that. The key observation is that for any $y \in (0,1)$, $Y_5 \leq y$ exactly when none of the five X_i exceed y. That is,

$$\begin{aligned} G_5(y) &= P(Y_5 \leq y) = P(\max\{X_i\} \leq y) \\ &= P(\text{all } X_i \leq y) \\ &= [F(y)]^5 \end{aligned}$$

Example. Let $X_1, \ldots, X_5$ be a simple random sample from $U(0,1)$, the uniform distribution on the interval $(0,1)$. Compute the cumulative distribution functions for the order statistics $Y_1, \ldots, Y_5$.

The simplest to compute is Y_5, so we'll start with that. The key observation is that for any $y \in (0,1)$, $Y_5 \leq y$ exactly when none of the five X_i exceed y. That is,

$$\begin{aligned} G_5(y) &= P(Y_5 \leq y) = P(\max\{X_i\} \leq y) \\ &= P(\text{all } X_i \leq y) \\ &= [F(y)]^5 \\ &= y^5 \end{aligned}$$

We use similar logic for Y_4. In order for Y_4 to be less than or equal to y, at least 4 of the 5 X_i must be less than y.

We use similar logic for Y_4. In order for Y_4 to be less than or equal to y, at least 4 of the 5 X_i must be less than y. That is,

$$G_4(y) \;=\; P(Y_4 \le y) \;=\; P(\text{at least } 4 \; X_i \le y)$$

We use similar logic for Y_4. In order for Y_4 to be less than or equal to y, at least 4 of the 5 X_i must be less than y. That is,

$$\begin{aligned} G_4(y) &= P(Y_4 \leq y) = P(\text{at least } 4 \ X_i \leq y) \\ &= P(\text{exactly } 4 \ X_i \leq y) + P(\text{exactly } 5 \ X_i \leq y) \end{aligned}$$

We use similar logic for Y_4. In order for Y_4 to be less than or equal to y, at least 4 of the 5 X_i must be less than y. That is,

$$\begin{aligned}
G_4(y) &= P(Y_4 \leq y) = P(\text{at least } 4\ X_i \leq y) \\
&= P(\text{exactly } 4\ X_i \leq y) + P(\text{exactly } 5\ X_i \leq y) \\
&= 5[P(X \leq y)]^4 P(X > y] + [P(X \leq y)]^5
\end{aligned}$$

We use similar logic for Y_4. In order for Y_4 to be less than or equal to y, at least 4 of the 5 X_i must be less than y. That is,

$$\begin{aligned}
G_4(y) &= P(Y_4 \le y) \;=\; P(\text{at least } 4 \; X_i \le y) \\
&= P(\text{exactly } 4 \; X_i \le y) + P(\text{exactly } 5 \; X_i \le y) \\
&= 5[P(X \le y)]^4 P(X > y] + [P(X \le y)]^5 \\
&= 5[F(y)]^4[1 - F(y)] + [F(y)]^5
\end{aligned}$$

We use similar logic for Y_4. In order for Y_4 to be less than or equal to y, at least 4 of the 5 X_i must be less than y. That is,

$$\begin{aligned}
G_4(y) &= P(Y_4 \leq y) = P(\text{at least } 4\ X_i \leq y) \\
&= P(\text{exactly } 4\ X_i \leq y) + P(\text{exactly } 5\ X_i \leq y) \\
&= 5[P(X \leq y)]^4 P(X > y] + [P(X \leq y)]^5 \\
&= 5[F(y)]^4[1 - F(y)] + [F(y)]^5 \\
&= 5y^4(1 - y) + y^5
\end{aligned}$$

The computations for the other three order statistics are similar.

The computations for the other three order statistics are similar. Overall, we get

$$
\begin{aligned}
G_5(y) &= y^5 \\
G_4(y) &= 5y^4(1-y) + y^5 \\
G_3(y) &= 10y^3(1-y)^2 + 5y^4(1-y) + y^5 \\
G_2(y) &= 10y^2(1-y)^3 + 10y^3(1-y)^2 + 5y^4(1-y) + y^5 \\
G_1(y) &= 5y(1-y)^4 + 10y^2(1-y)^3 + 10y^3(1-y)^2 + 5y^4(1-y) + y^5
\end{aligned}
$$

The computations for the other three order statistics are similar. Overall, we get

$$
\begin{aligned}
G_5(y) &= y^5 \\
G_4(y) &= 5y^4(1-y) + y^5 \\
G_3(y) &= 10y^3(1-y)^2 + 5y^4(1-y) + y^5 \\
G_2(y) &= 10y^2(1-y)^3 + 10y^3(1-y)^2 + 5y^4(1-y) + y^5 \\
G_1(y) &= 5y(1-y)^4 + 10y^2(1-y)^3 + 10y^3(1-y)^2 + 5y^4(1-y) + y^5
\end{aligned}
$$

For any given $y \in (0,1)$, $G_r(y)$ gives the probability that the r^{th} smallest value in a sample of size 5 will be less than or equal to y.

As usual, we can obtain the corresponding probability density functions of by differentiating.

As usual, we can obtain the corresponding probability density functions of by differentiating. After simplifying, we get

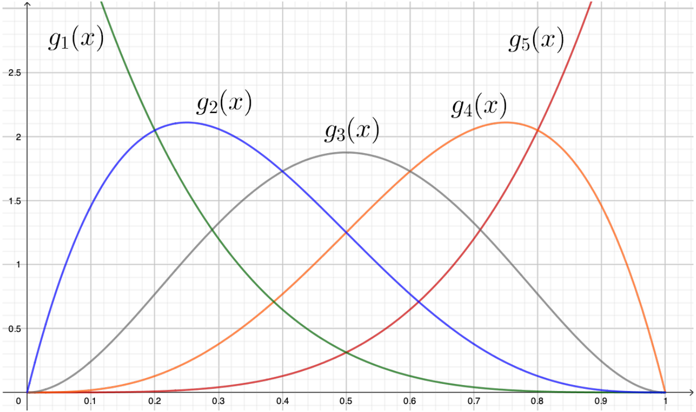
$$\begin{aligned} g_5(y) &= 5y^4 \\ g_4(y) &= 20y^3(1-y) \\ g_3(y) &= 30y^2(1-y)^2 \\ g_2(y) &= 20y(1-y)^3 \\ g_1(y) &= 5(1-y)^4 \end{aligned}$$

As usual, we can obtain the corresponding probability density functions of by differentiating. After simplifying, we get

$$\begin{aligned} g_5(y) &= 5y^4 \\ g_4(y) &= 20y^3(1-y) \\ g_3(y) &= 30y^2(1-y)^2 \\ g_2(y) &= 20y(1-y)^3 \\ g_1(y) &= 5(1-y)^4 \end{aligned}$$

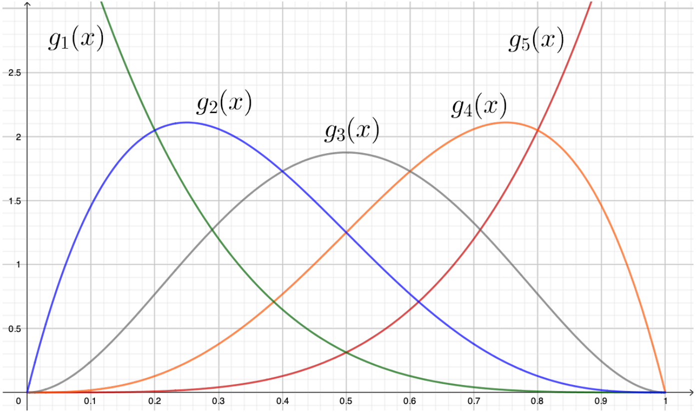
All of these are defined on the support of the original random variable, namely $(0, 1)$.

The graphs of these pdfs are typical:



As we'd expect, the probability densities are concentrated further to the right for higher order statistics.

The graphs of these pdfs are typical:



As we'd expect, the probability densities are concentrated further to the right for higher order statistics. In light of this, the cdfs will satisfy the relation $F_1(y) < F_2(y) < \cdots < F_5(y)$.

The calculations in this example are completely typical, so let's generalize them.

The calculations in this example are completely typical, so let's generalize them.

Theorem. When sampling from a continuous distribution, the cdf of the r^{th} order statistic Y_r is

$$G_r(y) = \sum_{k=r}^{n} \binom{n}{k} [F(y)]^k [1 - F(y)]^{n-k}$$

where F is the cdf of the distribution being sampled from and n is the size of the sample.

The calculations in this example are completely typical, so let's generalize them.

Theorem. When sampling from a continuous distribution, the cdf of the r^{th} order statistic Y_r is

$$G_r(y) = \sum_{k=r}^{n} \binom{n}{k} [F(y)]^k [1 - F(y)]^{n-k}$$

where F is the cdf of the distribution being sampled from and n is the size of the sample. Differentiating and simplifying gives the pdf,

$$g_r(y) = \frac{n!}{(r-1)!(n-r)!} [F(y)]^{r-1} [1 - F(y)]^{n-r} f(y)$$

The calculations in this example are completely typical, so let's generalize them.

Theorem. When sampling from a continuous distribution, the cdf of the r^{th} order statistic Y_r is

$$G_r(y) = \sum_{k=r}^{n} \binom{n}{k} [F(y)]^k [1 - F(y)]^{n-k}$$

where F is the cdf of the distribution being sampled from and n is the size of the sample. Differentiating and simplifying gives the pdf,

$$g_r(y) = \frac{n!}{(r-1)!(n-r)!} [F(y)]^{r-1} [1 - F(y)]^{n-r} f(y)$$

Memorizing these formulas isn't the best idea. Generally, it's easier to compute cdfs directly using the process in the example.

Example 2. Suppose $X_1, \ldots X_7$ is a random sample from a continuous probability distribution with pdf $f(x) = 2x$ on $(0, 1)$. What is the probability that the median is less than .5?

Example 2. Suppose $X_1, \dots X_7$ is a random sample from a continuous probability distribution with pdf $f(x) = 2x$ on $(0, 1)$. What is the probability that the median is less than .5?

The cdf of the sample median Y_4 is given by

$$G_4(y) = P(\text{at least } 4\ X_i \le y)$$

Example 2. Suppose $X_1, \ldots X_7$ is a random sample from a continuous probability distribution with pdf $f(x) = 2x$ on $(0, 1)$. What is the probability that the median is less than .5?

The cdf of the sample median Y_4 is given by

$$
\begin{aligned}
G_4(y) &= P(\text{at least } 4 \; X_i \leq y) \\
&= \binom{7}{4}[F(y)]^4[1 - F(y)]^3 + \binom{7}{5}[F(y)]^5[1 - F(y)]^2 + \\
&\quad \binom{7}{6}[F(y)]^6[1 - F(y)] + [F(y)]^7
\end{aligned}
$$

Example 2. Suppose $X_1, \ldots X_7$ is a random sample from a continuous probability distribution with pdf $f(x) = 2x$ on $(0, 1)$. What is the probability that the median is less than .5?

The cdf of the sample median Y_4 is given by

$$
\begin{aligned}
G_4(y) &= P(\text{at least } 4 \; X_i \leq y) \\
&= \binom{7}{4}[F(y)]^4[1 - F(y)]^3 + \binom{7}{5}[F(y)]^5[1 - F(y)]^2 + \\
&\quad \binom{7}{6}[F(y)]^6[1 - F(y)] + [F(y)]^7
\end{aligned}
$$

The cdf of this distribution is $F(x) = x^2$. All we ned to do is substitute.

The algebra isn't pretty, but it isn't complicated either..

$$
\begin{aligned}
G_4(y) &= \binom{7}{4}[F(y)]^4[1-F(y)]^3 + \binom{7}{5}[F(y)]^5[1-F(y)]^2 + \\
&\quad \binom{7}{6}[F(y)]^6[1-F(y)] + [F(y)]^7
\end{aligned}
$$

The algebra isn't pretty, but it isn't complicated either..

$$\begin{aligned}
G_4(y) &= \binom{7}{4}[F(y)]^4[1-F(y)]^3 + \binom{7}{5}[F(y)]^5[1-F(y)]^2 + \\
&\quad \binom{7}{6}[F(y)]^6[1-F(y)] + [F(y)]^7 \\
&= 35y^8(1-y^2)^3 + 21y^{10}(1-y^2)^2 + 7y^{12}(1-y^2) + y^{14}
\end{aligned}$$

The algebra isn't pretty, but it isn't complicated either..

$$
\begin{aligned}
G_4(y) &= \binom{7}{4}[F(y)]^4[1-F(y)]^3 + \binom{7}{5}[F(y)]^5[1-F(y)]^2 + \\
&\quad \binom{7}{6}[F(y)]^6[1-F(y)] + [F(y)]^7 \\
&= 35y^8(1-y^2)^3 + 21y^{10}(1-y^2)^2 + 7y^{12}(1-y^2) + y^{14} \\
G_4(.5) &= 35(.5)^8(1-(.5)^2)^3 + 21(.5)^{10}(1-(.5)^2)^2 + \\
&\quad 7(.5)^{12}(1-(.5)^2) + (.5)^{14}
\end{aligned}
$$

The algebra isn't pretty, but it isn't complicated either..

$$
\begin{aligned}
G_4(y) &= \binom{7}{4}[F(y)]^4[1-F(y)]^3+\binom{7}{5}[F(y)]^5[1-F(y)]^2+ \\
&\quad \binom{7}{6}[F(y)]^6[1-F(y)]+[F(y)]^7 \\
&= 35y^8(1-y^2)^3+21y^{10}(1-y^2)^2+7y^{12}(1-y^2)+y^{14} \\
G_4(.5) &= 35(.5)^8(1-(.5)^2)^3+21(.5)^{10}(1-(.5)^2)^2+ \\
&\quad 7(.5)^{12}(1-(.5)^2)+(.5)^{14} \\
&\approx .071
\end{aligned}
$$

There is a 7.1% chance that the median will be less than .5 in a sample of size 7 from this distribution.