Order Statistics

Suppose we have \( n \) independent, identically distributed random variables, and we are asked to calculate the expected value of the minimum or maximum of them. We can do this using order statistics. If we sort these random variables from least to greatest, the \( k \)th order statistic is the \( k \)th variable in our list. We often use the notation \( X_{(i)} \) to refer to the \( i \)th order statistic. For example, \( X_{(1)} \) (the first order statistic) is the minimum of the random variables, \( X_{(2)} \) (the second order statistic) is the second smallest, and so on. \( X_{(n)} \) (the \( n \)th order statistic) is the maximum.

Since we are discussing a set of random variables which all share the same distribution, it is useful to refer to the pdf of that distribution as \( f(x) \) and the cdf as \( F(x) \). In order to answer most questions about \( X_{(i)} \), we need to find the pdf for \( X_{(i)} \). We do this by first finding the cdf for \( X_{(i)} \) and then taking a derivative. We find the cdf using the original \( F(x) \) as well as the survival function \( 1 - F(x) = S(x) \).

**Maximum:** Finding a cdf for the maximum, or \( n \)th order statistic, \( X_{(n)} \), of a set of random variables is easiest. \( F_{X_{(n)}}(x) \) is the probability that the largest of the random variables is less than \( x \). This is equivalent to all of the random variables being less than \( x \). Thus, \( F_{X_{(n)}}(x) = (F(x))^n \). We can find the pdf by taking a derivative of this with the chain rule.

\[
\begin{align*}
\bullet \ F_{X_{(n)}}(x) &= (F(x))^n \\
\bullet \ f_{X_{(n)}}(x) &= n(F(x))^{n-1}f(x)
\end{align*}
\]

**Minimum:** Finding a cdf for the minimum, or first order statistic, \( X_{(1)} \), is similar. We really need to find the probability that at least one of the variables will be less than \( x \). It is easier to find the opposite probability, so we will find the probability that none of the variables are less than \( x \), and we will subtract from 1. Each variable has probability \( 1 - F(x) = S(x) \) to be greater than \( x \). So the probability that all of them are greater than \( x \) is \( (1 - F(x))^n = (S(x))^n \). Thus, \( F_{X_{(1)}}(x) = 1 - (1 - F(x))^n = 1 - (S(x))^n \). Notice the negative signs cancel out from the chain rule when we take the derivative.

\[
\begin{align*}
\bullet \ F_{X_{(1)}}(x) &= 1 - (1 - F(x))^n \\
\bullet \ f_{X_{(1)}}(x) &= n(1 - F(x))^{n-1}f(x)
\end{align*}
\]

**General:** Before we discuss how to find a general order statistic, just know that the majority of questions on order statistics will be about maximums. It is possible you will see minimums or other order statistics, but they are far less common. For an order statistics such as \( X_{(2)} \), you have to break the cdf into multiple cases. For example, suppose we have 4 random variables, and we want to know the cdf for \( X_{(2)} \). We must calculate the probability that at least two random variables are less than \( x \). This breaks into three cases: (1) exactly 2 are less than \( x \), (2) exactly 3 are less than \( x \), and (3) all 4 are less than \( x \). We find the probability of each case using an approach similar to a binomial. We use a probability \( F(x) \) for each random variable we want less than \( x \) and a probability \( 1 - F(x) = S(x) \) for each probability we want greater than \( x \). Additionally, we must choose which of our random variables will be less than \( x \) and which will be greater than \( x \). So to calculate the cdf for \( X_{(2)} \), we would have the following calculation.

\[
F_{X_{(2)}} = \binom{4}{2} (F(x))^2 (1 - F(x))^2 + \binom{4}{3} (F(x))^3 (1 - F(x))^1 + \binom{4}{4} (F(x))^4 (1 - F(x))^0
\]

Note: You can write this formula using survival functions, but we often follow up this step with a derivative. If you choose to use survival functions, just notice the derivative of \( S(x) \) is \(-f(x)\).

Note: I have included some extraneous terms that would evaluate to 1, such as choosing all four and using a zero exponent.

Note: It was unnecessary to include a binomial coefficient in the maximum and minimum cases, since all the random variables were acting the same (that is, either all \( F(x) \) or all \( 1 - F(x) \)). We could also use this format to calculate the minimum, but it breaks into several cases. Finding the opposite probability is much shorter.
Ex. Suppose $X_1, X_2, \text{ and } X_3$ are three independent and identically distributed continuous random variables with common probability density function $f(x) = \frac{8}{x^3}$ for $x \geq 2$. Calculate $E[X_1]$. 
Ex. Suppose $X_1, X_2,$ and $X_3$ are three independent and identically distributed continuous random variables with common probability density function $f(x) = \frac{8}{x^3}$ for $x \geq 2$. Calculate $E[X_{(3)}]$. 
Ex. Suppose $X_1$, $X_2$, and $X_3$ are three independent and identically distributed continuous random variables with common probability density function $f(x) = \frac{8}{x^3}$ for $x \geq 2$. Find the probability density function for $X_{(2)}$. 
Ex. Suppose $X_1$, $X_2$, $X_3$, $X_4$, and $X_5$ are independent and identically distributed exponential random variables with mean 4. Find the cumulative distribution function for $\max(X_1, X_2, X_3, X_4, X_5)$. 
Ex. Consider a sample of size 5 from a uniform distribution over \((0, 1)\). Compute the probability that the median is in the interval \(\left(\frac{1}{4}, \frac{3}{4}\right)\).
Ex. Claims are assumed to follow a distribution defined by \( f(x) = 3x^2 \) for \( 0 < x < 1 \). If 10 claims are observed, calculate the expected value of the 2nd largest claim.