Worksheet 20 (Solutions)

1. Throughout this worksheet, let $X_1, \ldots, X_n$ be a sequence of n i.i.d. continous random variables that have a pdf $f(x)$ and a cdf $F(x)$. Define the random variables $Y_1, \ldots, Y_n$ to be the corresponding order statistics, each with a pdf $(g_j(y))$ and cdfs $(G_j(y))$. Write down $G_n(y)$—the cdf of the maximum value—in terms of n and F. Hint: Write out the problem with probabilities before converting to the cdf.

Solution: In order for the maximum to be less than y, all values of X_j must be less than y. So, we have:

$$\begin{aligned} G_n(y) &= \mathbb{P}[Y_n \leq y] \\ &= \prod_j \mathbb{P}[X_j \leq y] \\ &= \prod_j F(y) \\ &= [F(y)]^n \end{aligned}$$

So, it's just the cdf of X_j raised to the power of n.

2. In the next few questions, we will work on the density function $g_k(y)$ for an arbitrary k. To start, fix a value y and a positive value Δ. What is the joint probability that $X_1, \ldots, X_{k-1}$ are all less than y, that $X_{k+1}, \ldots, X_n$ are all greater than $y + \Delta$, and that X_k is in the interval $[y, y + \Delta]$?

Solution: This is similar to the previous question, we just have to be more careful about the specific probabilities since they have different directions. Here are the three different components:

$$\begin{aligned} \mathbb{P}[X_1 \leq y] \times \cdots \times \mathbb{P}[X_{k-1} \leq y] &= [F(y)]^{k-1} \\ \mathbb{P}[X_{k+1} \geq y + \Delta] \times \cdots \times \mathbb{P}[X_n \geq y + \Delta] &= [1 - F(y)]^{n-k} \\ \mathbb{P}[X_k \in [y, y + \Delta]] &= F(y + \Delta) - F(y) \end{aligned}$$

Multiplying these together, we have:

$$[F(y)]^{k-1} \times [1 - F(y + \Delta)]^{n-k} \times [F(y + \Delta) - F(y)]$$

3. We are back to another counting question! The probability you have in the previous question counts only one specific configuration of the values X_j that would result in Y_k being in the interval $[y, y+\Delta]$. In general, we could have any set of $k-1$ of the n random variables be less than y, one of the random variables be in the interval $[y, y + \Delta]$, and the rest of the $n-k$ be somewhere greater than $y+\Delta$. (a) How many different configurations are there? (b) What is the probability that Y_k is in the interval $[y, y+\Delta]$?[1]

[1] Technically you are computing the probability that Y_k is in this interval and Y_{k+1} is not. The difference between these will limit to zero in the next question.

Solution: (a) We need to partition the set of n random variables into sets of sizes $k-1$, $n-k$ and 1. If you remember the formula for partitions, this is really easy. Otherwise, we can break it into a multi-stage experiment in which we select the $k-1$ variables in the first interval ($\binom{n}{k-1}$) and then from the remaining $n-k+1$ we select the $n-k$ variables in the upper interval ($\binom{n-k+1}{n-k}$). So:

$$\binom{n}{k-1} \times \binom{n-k+1}{n-k} = \frac{n!}{(k-1)!(n-k+1)!} \times \frac{(n-k+1)!}{(n-k)!(1)!}$$
$$= \frac{n!}{(n-k)!(k-1)!}.$$

Then, the probability is given by:

$$\mathbb{P}[Y_k \in [y, y+\Delta]] = \left[\frac{n!}{(n-k)!(k-1)!}\right] \times [F(y)]^{k-1} \times [1-F(y+\Delta)]^{n-k} \times [F(y+\Delta)-F(y)].$$

4. One way, if it exists, to define the pdf of a random variable Y is:

$$f_Y(y) = \lim_{\Delta \to 0}\left[\frac{1}{\Delta} \times \mathbb{P}[Y \in [y, y+\Delta]]\right] = \lim_{\Delta \to 0}\left[\frac{F_Y(y+\Delta)-F_Y(y)}{\Delta}\right]$$

Where F_Y is the cdf. This comes directly from the fundamental theorem of calculus and the definition of the relationship between the cdf and the pdf. Use this to compute the pdf $g_k(y)$ of the k-th order statistic Y_k. Your answer should be in terms of factorials using only y, k, n, F and f.

Solution: Taking our previous result and dividing by Δ gives:

$$\left[\frac{n!}{(n-k)!(k-1)!}\right] \times [F(y)]^{k-1} \times [1-F(y+\Delta)]^{n-k} \times \frac{1}{\Delta} \times [F(y+\Delta)-F(y)].$$

Taking the limit as Δ goes to zero causes the first $F(y+\Delta)$ to limit to $F(y)$ and the last terms to become $f(y)$:[2]

[2] If the limit of two functions $f(x)$ and $g(x)$ are both finite as $x \to x_0$, then the limit of $f(x) \cdot g(x)$ as x goes towards x_0 will be the limit of $f(x)$ times the limit of $f(x)$ times the limit of $g(x)$. That's why we can plug the $\Delta = 0$ into the first term while leaving the rest intact.

$$g_k(y) = \left[\frac{n!}{(n-k)!(k-1)!}\right] \times [F(y)]^{k-1} \times [1-F(y)]^{n-k} \times f(y)$$

So, finding the density of the order statistic can be basically reduced to the problem of finding the cdf. The latter usually does not have a closed form for most common distributions, but it can be easily approximated.

5. Let's apply this definition to a special case. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} U(0,1)$. For any $y \in (0,1)$, write down a formula for $F(y)$. Hint: This is easy.

Solution: The cdf $F(y) = y$ for every $y \in (0,1)$.

6. Now, write down pdf of the density function $g_k(y)$ for $y \in (0,1)$ when the X_j's come from a standard uniform distribution? We will simplify this in the next question.

Solution: Using our formula and plugging in $f(y) = 1$ and $F(y) = y$, we have:

$$g_k(y) = \left[\frac{n!}{(n-k)!(k-1)!}\right] \times y^{k-1} \times [1-y]^{n-k}.$$

7. Recall Gamma function has the property that $\Gamma(n) = (n-1)!$ for any integer n. Write your previous question in terms of the Gamma function.

Solution: We have:

$$\begin{aligned}
g_k(y) &= y^{k-1} \cdot (1-y)^{n-k} \times \left[n \cdot \binom{n-1}{k-1}\right] \\
&= y^{k-1} \cdot (1-y)^{n-k} \times \left[n \cdot \frac{(n-1)!}{(k-1)!(n-k)!}\right] \\
&= y^{k-1} \cdot (1-y)^{n-k} \times \left[\frac{(n)!}{(k-1)!(n-k)!}\right] \\
&= y^{k-1} \cdot (1-y)^{n-k} \times \left[\frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)}\right]
\end{aligned}$$

8. Set $\alpha = k$ and $\beta = n-k+1$ and plug into the solution from the previous question. What is the name for the distribution of the k-th order statistic Y_k from a set of independent random variables from the standard uniform distribution?

Solution: Plugging in, we have:

$$\begin{aligned}
g_k(y) &= y^{k-1} \cdot (1-y)^{n-k} \times \left[\frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)}\right] \\
&= y^{\alpha-1} \cdot (1-y)^{\beta-1} \times \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right]
\end{aligned}$$

And this is the density of the Beta distribution with parameters α and β. So, we finally see a justification for the form (and have a full derivation of the normalizing constant) of a Beta distribution.

9. Let's end with a even more concrete example. Let $X_1, X_2, X_3, X_4 \overset{\text{i.i.d.}}{\sim} U(0,1)$. What are the expected values of the four order statistics Y_1, Y_2, Y_3, Y_4?

Solution: The mean of a Beta distribution is $\frac{\alpha}{\alpha+\beta}$, so the mean of the order statistic is, plugging in the values from the previous question, $\frac{k}{n+1}$. With $n = 4$ we have the values $\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}$, or $0.2, 0.4, 0.6, 0.8$.