# MATH50010 - Probability for Statistics
## Unseen Problem 5

1. Consider a rod of unit length. The rod is broken at two points, whose locations can be modelled as independent, uniformly distributed random variables.

   (a) What is the density function of the *ordered* breakpoints $(x_{(1)}, x_{(2)})$, where $x_{(1)} < x_{(2)}$?

   (b) What is the probability that the three segments of the rod fit together to form a triangle?

   *(a) First observe that the joint density of any $X, Y \sim \textsc{Uniform}(0,1)$ and independent is $f_{X,Y}(x,y) = 1$ for $(x,y) \in (0,1)^2$. Then,*

$$\Pr(X < Y) = \int_0^1 \int_0^x f_{X,Y}(x,y)\, dy\, dx = \int_0^1 x\, dx = [x^2/2]_0^1 = 1/2.$$

   *Let $A \subseteq \{(x,y) \in (0,1)^2 : x < y\}$ be a region of the space where $x < y$. Then,*

$$\begin{aligned}\Pr((X,Y) \in A | X < Y) &= \frac{\Pr(\{(X,Y) \in A\} \cap \{X < Y\})}{\Pr(X < Y)} = \frac{\Pr((X,Y) \in A)}{1/2} \\ &= 2\Pr((X,Y) \in A)\end{aligned}$$

   *since $\{(X,Y) \in A\} \cap \{X < Y\} = \{(X,Y) \in A\}$ by definition of $A$. Consequently, $\Pr(X \le x, Y \le y | X < Y) = 2\Pr(X \le x, Y \le y) = 2F_{X,Y}(x,y)$ so differentiating and using the density of independent uniform random variables, we can write down the density for $(x_{(1)}, x_{(2)})$ in any region where $0 < x_{(1)} < x_{(2)} < 1$ as*

$$f(x_{(1)}, x_{(2)}) = \begin{cases} 2 & 0 < x_{(1)} < x_{(2)} < 1, \\ 0 & \textit{otherwise} \end{cases}$$

   *(b) The three segments form a triangle precisely when no segment is larger than the sum of the other two (think of the triangle inequality). Equivalently, if no segment has length $\geq \frac{1}{2}$. This is the same as the event*

$$\{X_{(1)} < \frac{1}{2}\} \cap \{X_{(2)} - X_{(1)} < \frac{1}{2}\} \cap \{1 - X_{(2)} < \frac{1}{2}\}.$$

   *Compute this from the joint density above, by conditioning of the value of $x_{(1)}$ :*

$$\int_0^{\frac{1}{2}} \int_{\frac{1}{2}}^{x_{(1)}+\frac{1}{2}} 2\, dx_{(2)} dx_{(1)} = \int_0^{\frac{1}{2}} 2x_{(1)}\, dx_{(1)} = \left(\frac{1}{2}\right)^2 = \frac{1}{4}.$$

   *Nice geometric interpretations of this well-known result are possible. Can you find one?*

2. Assume that the interval $[0,1]$ is determinstically partitioned into $n$ disjoint sub-intervals with lengths $p_1, p_2, \ldots, p_n$, the *entropy* of this partition is defined to be

$$h = -\sum_{i=1}^{n} p_i \log p_i.$$

Let $X_1, X_2, \ldots$ be independent $\textsc{Uniform}[0,1]$ random variables and let $Z_m(i)$ be the number of the $X_1, \ldots, X_m$ which lie in the $i$th interval of the partition. Show that

$$R_m = \prod_{i=1}^{n} p_i^{Z_m(i)}$$

satisfies $\frac{1}{m} \log R_m \xrightarrow{P} -h$ as $m \to \infty$.

*Let $I_{i,j}$ be the indicator function of the event that $X_j$ lies in the $i$th interval, and note that $\mathrm{E}[I_{i,j}] = p_i$ since interval $i$ has length $p_i$. Then,*

$$\log R_m = \sum_{i=1}^{n} Z_m(i) \log p{+}_i = \sum_{i=1}^{n} \sum_{j=1}^{m} I_{i,j} \log p_i$$

*and define $Y_j = \sum_{i=1}^{n} I_{i,j} \log p_i$ for $1 \le j \le m$. Then we can write $\log R_m = \sum_{j=1}^{m} Y_j$. Observe that the mean and variance of $Y_j$ are*

$$\mathrm{E}[Y_j] = \sum_{i=1}^{n} p_i \log p_i = -h$$

$$\mathrm{Var}(Y_j) = \mathrm{E}[Y_j^2] - \mathrm{E}[Y_j]^2 \le \mathrm{E}[Y_j^2] = \mathrm{E}\left[\left(\sum_{i=1}^{n} I_{i,j} \log p_i\right)^2\right] = \sum_{i,k=1}^{n} \mathrm{E}[I_{i,j} I_{k,j} \log p_i \log p_k]$$

$$= \sum_{i=1}^{n} \log^2 p_i \mathrm{E}[I_{i,j}^2] = \sum_{i=1}^{n} p_i \log^2 p_i < \infty$$

*since by definition of $I_{i,j}$, only one of $I_{i,j}$ and $I_{k,j}$ can be non-zero for $i \ne k$, so if $i \ne k$, $I_{i,j} I_{k,j} = 0$. Moreover, the $Y_j$ themselves are independent and identically distributed since the $X_j$'s are i.i.d. and each $Y_j$ is a deterministic function of $X_j$ only. Therefore, we can apply the weak law of large numbers (Proposition 4.22) to show that*

$$\frac{1}{m} \sum_{i=1}^{m} Y_i = \frac{1}{m} \log R_m \xrightarrow{P} \mathrm{E}[Y_j] = -h.$$