MA 413 - Statistics for Data Science

Solutions to Exercise 1

1. (a) Clearly, for any $n \geq 1$ we have,

$$C_{n+1} = \cup_{j=n+1}^{\infty} B_j \subset B_n \cup \left[\cup_{j=n+1}^{\infty} B_j\right] = \cup_{j=n}^{\infty} B_j = C_n$$

$$D_n = \cap_{j=n}^{\infty} B_j = B_n \cap \left[\cap_{j=n+1}^{\infty} B_j\right] \subset \cap_{j=n+1}^{\infty} B_j = D_{n+1}$$

Therefore, $C_n \supset C_{n+1}$ and $D_n \subset D_{n+1}$, and the conclusion follows.

(b) ($\Longrightarrow$) If $\omega \in \cap_{j=1}^{\infty} C_j$ then $\omega \in C_j = \cup_{k=j}^{\infty} B_k$ for all $j \geq 1$. Clearly, if $\omega \in B_j$ for only a finite number of j then there will be a largest m such that $\omega \in B_m$. Then $\omega \notin \cup_{k=m+1}^{\infty} B_k = C_{m+1}$, thus contradicting $\omega \in \cap_{j=1}^{\infty} C_j$. It follows that $\omega \in B_j$ for an infinite number of $j \geq 1$.

($\Longleftarrow$) Now, if $\omega \in B_j$ for an infinite number of j, then for any $k \geq 1$ there exists $m \geq k$ such that $\omega \in B_m$. So, $\omega \in \cup_{j=k}^{\infty} B_j = C_k$ for all $k \geq 1$. And thus, $\omega \in \cap_{k=1}^{\infty} C_k$. Hence proved.

(c) ($\Longrightarrow$) If $\omega \in \cup_{j=1}^{\infty} D_j$, then there exists $m \geq 1$ such that $\omega \in D_m = \cap_{j=m}^{\infty} B_j$. Thus, $\omega \in B_j$ for all $j \geq m$. This implies that $\omega \in B_j$ for all but finitely many j.

($\Longleftarrow$) If $\omega \in B_j$ for all but finitely many j, then there exists a largest $m \geq 1$ such that $\omega \notin B_m$. Then $\omega \in B_j$ for all $j \geq m+1$. Thus, $\omega \in \cap_{j=m+1}^{\infty} B_j = D_{m+1} \subset \cup_{j=1}^{\infty} D_j$. Hence proved.

2. **Note:** This question is too difficult for this course. But for the sake of completeness I am including the solution here. Be assured that questions which are this difficult will not appear in the exams.

Clearly, Ω contains all finite length sequences of $\{\texttt{H}, \texttt{T}\}$ which contain two consecutive Hs only at the end. This is not particularly helpful. We can try writing down some of the strings:

$$\texttt{HH}, \texttt{THH}, \texttt{HTHH}, \texttt{TTHH}, \texttt{TTTHH}, \texttt{THTHH}, \texttt{HTTHH}, \cdots$$

The trick is to consider the set of strings which begin with T. Let $\mathcal{T}_k$ denote the set of length-k strings which begin with T and end with HH. And similarly, we can define the set we are actually interested in, as $\mathcal{O}_k$, the set of length-k strings which end with HH.

Now, for $k \geq 1$, any member of $\mathcal{O}_{k+2}$ can be constructed as follows: If the first letter is H then it must be followed by a member of $\mathcal{T}_{k+1}$. If it is T then it is a member of $\mathcal{T}_{k+2}$. We can write this as:

$$\mathcal{O}_{k+2} = \texttt{H}\#\mathcal{T}_{k+1} \cup \mathcal{T}_{k+2}$$

Since $\mathcal{O}_2 = \{\texttt{HH}\}$ and $\mathcal{O}_1 = \emptyset$, we have completely specified $\mathcal{O}_k$ in terms of $\mathcal{T}_k$. But clearly, $\mathcal{T}_{k+1} = \texttt{T}\#\mathcal{O}_k$. It follows that,

$$\mathcal{O}_{k+2} = \texttt{HT}\#\mathcal{O}_k \cup \texttt{T}\#\mathcal{O}_{k+1}$$

Clearly, $\Omega = \cup_{j=1}^{\infty}\mathcal{O}_j$. And the probability we are interested in is

$$\Pr(\omega \in \mathcal{O}_k) = \frac{\frac{1}{2^k}|\mathcal{O}_k|}{\sum_{j=1}^{\infty}\frac{1}{2^j}|\mathcal{O}_j|} = \frac{1}{2^k}|\mathcal{O}_k|$$

since, $\sum_{j=1}^{\infty}\frac{1}{2^j}|\mathcal{O}_j| = 1$ as we are sure to hit two heads in a row. And since, $|\mathcal{O}_1| = 0$,$|\mathcal{O}_2| = 1$ and $|\mathcal{O}_{k+2}| = |\mathcal{O}_{k+1}| + |\mathcal{O}_k|$ for $k \geq 1$, we can recognize that this is the famous Fibonacci sequence.

Remark. $|A|$ denotes the number of elements in the set A.

3. We proceed as follows,

$$\begin{aligned}
\Pr(A^c \cap B^c) &= \Pr\left([A \cup B]^c\right) \\
&= 1 - \Pr\left(A \cup B\right) \\
&= 1 - [\Pr(A) + \Pr(B) - \Pr(A \cap B)] \\
&= 1 - \Pr(A) - \Pr(B) + \Pr(A)\Pr(B) \\
&= [1 - \Pr(A)]\,[1 - \Pr(B)] \\
&= \Pr(A^c)\Pr(B^c)
\end{aligned}$$

Remark. You might want to keep in mind this simple formula:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

4. We shall begin by assuming the opposite: for all $i = 2, \ldots, m$ we have $\Pr(A_i|B) \leq \Pr(A_i)$. Then by adding all of these inequalities and $\Pr(A_1|B) < \Pr(A_1)$ we have,

$$\sum_{j=1}^{m}\Pr(A_j|B) < \sum_{j=1}^{m}\Pr(A_j)$$

but since $\{A_i\}$ form a partition of Ω, and as a result $\{A_i \cap B\}$ form a partition of B, we can write

$$\sum_{j=1}^{m}\Pr(A_j) = \Pr(\Omega) = 1$$

and

$$\sum_{j=1}^{m}\Pr(A_j|B) = \frac{1}{\Pr(B)}\sum_{j=1}^{m}\Pr(A_j \cap B) = \frac{1}{\Pr(B)} \cdot \Pr(B) = 1$$

From the previous inequality we derive a contradiction: $1 < 1$. So, it follows that there exists some $j = 2, \ldots, m$ such that $\Pr(A_j|B) > \Pr(A_j)$.

5. Here is the R code,

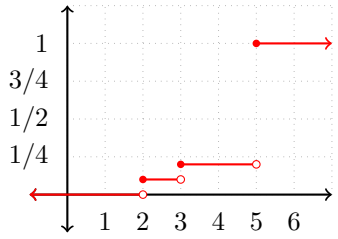
```
N <- 1000; # Number of samples.
p <- 0.2; # Probability of heads.
n <- 1000; # Number of coin tosses.
b <- 100; # Number of bins in the histogram.

x <- numeric(N);

for(i in 1:N) {
y <- rbinom(n=n,prob=p,size=1);
x[i] <- sum(y);
}
hist(x, breaks = b);
```

Remark. This phenomena that we are observing are: the law of large numbers and that when $n \to \infty$ while np remains constant, the distribution $\text{Bin}(n, p)$ approaches the Poisson distribution.

6. (a) Here is the plot,



(b) Clearly, $\Pr(1.5 < Y < 4.2) = F(4.2) - F(1.5) = 2/10 - 0 = 0.2$.

7. (a) The c.d.f. is given by $F_Y(y) = \int_{-\infty}^{y} f_Y(u)\, du$. So, we calculate

$$F_Y(y) = \begin{cases} 0 & -\infty < y < 1 \\ y/4 & 0 \le y < 1 \\ 1/4 & 1 \le y < 3 \\ 1/4 + (3/8)(y-3) & 3 \le y < 5 \\ 1 & 5 \le y < \infty \end{cases}$$

(b) Using transformation formula, we get

$$\begin{aligned} f_Z(z) &= f_Y(1/z) \cdot \left| \frac{d}{dz}\,[1/z] \right| \\ &= \begin{cases} \frac{1}{4z^2} & 1 < z < \infty \\ \frac{3}{8z^2} & 1/5 < z < 1/3 \\ 0 & \text{o/w} \end{cases} \end{aligned}$$

8. By definition the c.d.f. of Y, say F_Y is given by $F_Y(y) = \Pr(Y \leq y)$. So,

$$\begin{aligned} F_Y(y) &= \Pr(\max\{X, 0\} \leq y) \\ &= \begin{cases} 0 & y < 0 \\ F(y) & 0 \leq y < \infty \end{cases} \\ &= F(y)\mathbf{1}_{\{y \geq 0\}} \end{aligned}$$

9. Since, $X \sim \text{Exp}(\eta)$ we have, $f_X(x) = \eta e^{-\eta x}$ for $x \geq 0$ and 0 otherwise. We can evaluate $F(x) = \int_{-\infty}^{x} f_X(u)\ du$ to get,

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\eta x} & x \geq 0 \end{cases}$$

Strictly speaking, the inverse does not exist. But if we restrict F to $[0, \infty)$, the inverse exists and is given by $y = 1 - e^{-\eta F^{-1}(y)}$. Thus, $F^{-1}(y) = -\frac{1}{\eta}\log(1 - y)$.

10. We calculate as follows,

$$\begin{aligned} \Pr(X < \tfrac{1}{2} | Y \leq \tfrac{1}{2}) &= \frac{\Pr(-\infty < X < \frac{1}{2}, -\infty < Y \leq \frac{1}{2})}{\Pr(-\infty < X < \infty, -\infty < Y \leq \frac{1}{2})} \\ &= \frac{\int_{-\infty}^{1/2}\int_{-\infty}^{1/2} f_{X,Y}(x,y)\ dx\ dy}{\int_{-\infty}^{1/2}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx\ dy} \\ &= \frac{C \cdot \int_0^{1/2}\int_0^{1/2}(x + y^3)\ dx\ dy}{C \cdot \int_0^{1/2}\int_0^{1}(x + y^3)\ dx\ dy} \\ &= \frac{\frac{x^2 y}{2} + \frac{xy^4}{4}|_{x=1/2, y=1/2}}{\frac{x^2 y}{2} + \frac{xy^4}{4}|_{x=1, y=1/2}} \\ &= 9/34 \end{aligned}$$

11. One way to approach this problem is to use the convolution formula, since U is a sum of two independent and continuous random variables.

$$\begin{aligned} f_U(u) = f_{X+(-Y)}(u) &= \int_{-\infty}^{\infty} f_X(u - v) f_{-Y}(v)\ dv \\ &= \int_{-\infty}^{\infty} \mathbf{1}_{\{u - v \in [0,1]\}}\mathbf{1}_{\{v \in [-1,0]\}}\ dv \\ &= \int_{-\infty}^{\infty} \mathbf{1}_{\{v \in [u-1, u]\}}\mathbf{1}_{\{v \in [-1,0]\}}\ dv \\ &= \begin{cases} 1 - |u| & u \in [-1, 1] \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

For V, we can do the same by noticing that $V = e^{\log X - \log Y}$. We define $X' = \log(X)$, $Y' = -\log(Y)$ and $Z = X' + Y'$. Then,

$$f_{X'}(x) = f_X(e^x)\left|\tfrac{d}{dx}\left[e^x\right]\right| = e^x \mathbf{1}_{\{x \le 0\}}$$

$$f_{Y'}(y) = f_Y(e^{-y})\left|\tfrac{d}{dy}\left[e^{-y}\right]\right| = e^{-y}\mathbf{1}_{\{y \ge 0\}}$$

Using the convolution formula,

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_{X'}(z-t) f_{Y'}(t)\ dt \\ &= \int_{-\infty}^{\infty} e^{z-t}\mathbf{1}_{\{z-t\le 0\}} \cdot e^{-t}\mathbf{1}_{\{t\ge 0\}}\ dt \\ &= e^z \int_{\max\{z,0\}}^{\infty} e^{-2t}\ dt \\ &= \frac{1}{2}e^z(e^{-2\max\{z,0\}}) = \frac{1}{2}e^{-|z|} \end{aligned}$$

Notice that $V = e^Z$. Therefore,

$$\begin{aligned} f_V(v) &= f_Z(\log v)\left|\frac{d}{dv}\log v\right| = \frac{1}{2v}e^{-|\log v|} \\ &= \begin{cases} \frac{1}{2} & 0 \le v < 1 \\ \frac{1}{2v^2} & 1 \le v < \infty \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

So, we are done.

Remark. Here we could use the convolution formula because U and V could be written as sums of two continuous and independent random variables whose p.d.f.s we could derive. When this is not the case, one must use other methods such as multidimensional transformations or simply evaluating that c.d.f. of the random variable of interest and differentiating it.

Try using these methods to solve the above problem.