Probability and Statistics

Exercise sheet 4

Exercise 4.1 We have three dice. Two are ordinary with the numbers $1, 2, 3, 4, 5, 6$ and one is special where 6 is replaced by 7 (i.e. $1, 2, 3, 4, 5, 7$).

We first roll a ordinary die to decide which of the other two dice is chosen afterwards. If rolling the first die results in a number less than or equal to 4, we choose the second ordinary die, otherwise we choose the special die.

Then we roll the chosen die and denote its result by X_2.

(a) What is the distribution of X_2? Construct a minimal probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that one can answer this question and question (b).

(b) What is the conditional probability $\mathbb{P}[\text{first die} \geq 5 | X_2 = 5]$?

(c) Let X_1 be the result of the first die. We want to find the distribution of $X_1 + X_2$. Can we achieve this on the above defined probability space $(\Omega, \mathcal{F}, \mathbb{P})$? If not, construct a suitable probability space and find the distribution.

Solution 4.1

(a) We construct the probability space such that ω_1 describes if the result of rolling the first die is at least 5 or not. So we take $\Omega := \{0, 1\} \times \{1, 2, 3, 4, 5, f\}$, $\mathcal{F} := 2^\Omega$, $\mathbb{P} : \mathcal{F} \to [0, 1], A \mapsto \mathbb{P}[A] = \sum_{\omega \in A} p(\omega)$, where $\omega_1 = 1$ represents the situation that the first roll results in 5 or 6 and

$$p(\omega) := \begin{cases} \frac{1}{9}, & \omega_1 = 0, \\ \frac{1}{18}, & \omega_1 = 1, \end{cases} \quad \forall \omega \in \Omega.$$

(b)

$$\mathbb{P}[\text{first die} \geq 5 | X_2 = 5] = \frac{\mathbb{P}[\text{first die} \geq 5,\ X_2 = 5]}{\mathbb{P}[X_2 = 5]} = \frac{\mathbb{P}[\{(1, 5)\}]}{\mathbb{P}[X_2 = 5]} = \frac{\frac{1}{18}}{\frac{1}{6}} = \frac{1}{3}.$$

These numbers reflect the fact that there are twice as many possibilities to get $1, 2, 3, 4$ in the first roll as to get $5, 6$.

Then we define $X_2 := \omega_2$ if $\omega_2 \leq 5$, $X_2(0, f) := 6$, $X_2(1, f) := 7$. This gives

$$\mathbb{P}[X_2 = k] = \begin{cases} \frac{1}{6} & k \in \{1, 2, 3, 4, 5\}, \\ \frac{1}{9}, & k = 6, \\ \frac{1}{18}, & k = 7. \end{cases}$$

(c) No, because X_1 cannot be defined on Ω. We can construct a probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$, with $\tilde{\Omega} := \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, f\}$, $\tilde{\mathcal{F}} := 2^{\tilde{\Omega}}$ and $\mathbb{P}[A] := \frac{|A|}{36}$. This allows us to redefine

$$X_1 : \tilde{\Omega} \to \mathbb{R}, \omega \mapsto X_1(\omega) := \omega_1 \text{ and } X_2 : \tilde{\Omega} \to \mathbb{R}, \omega \mapsto X_2(\omega) := \begin{cases} \omega_2, & \omega_2 \leq 5, \\ 6, & \omega_2 = f \text{ and } \omega_1 \leq 4, \\ 7, & \omega_2 = f \text{ and } \omega_1 \geq 5. \end{cases}$$

By counting, we obtain

$$\mathbb{P}\left[X_1 + X_2 = k\right] = \begin{cases} \frac{1}{36}, & k = 2, \\ \frac{2}{36}, & k = 3, \\ \frac{3}{36}, & k = 4, \\ \frac{4}{36}, & k = 5, \\ \frac{5}{36}, & k = 6, \\ \frac{6}{36}, & k = 7, \\ \frac{5}{36}, & k = 8, \\ \frac{4}{36}, & k = 9, \\ \frac{3}{36}, & k = 10, \\ \frac{1}{36}, & k = 11, \\ \frac{1}{36}, & k = 12, \\ \frac{1}{36}, & k = 13. \end{cases}$$

Exercise 4.2 (Simpson's paradox).

We are interested in studying the probability of success of a student at an entrance exam for two departments of a university. Consider the following events:

$$A := \{\text{The student is a man}\}$$
$$A^c = \{\text{The student is a woman}\}$$
$$B := \{\text{The student applied for department I}\}$$
$$B^c = \{\text{The student applied for department II}\}$$
$$C := \{\text{The student was accepted}\}$$
$$C^c = \{\text{The student was not accepted}\}$$

We assume the following probabilities:

$$\mathbb{P}\left[A\right] = 0.73,$$
$$\mathbb{P}\left[B \mid A\right] = 0.69, \mathbb{P}\left[B \mid A^c\right] = 0.24,$$
$$\mathbb{P}\left[C \mid A \cap B\right] = 0.62, \mathbb{P}\left[C \mid A^c \cap B\right] = 0.82,$$
$$\mathbb{P}\left[C \mid A \cap B^c\right] = 0.06, \mathbb{P}\left[C \mid A^c \cap B^c\right] = 0.07.$$

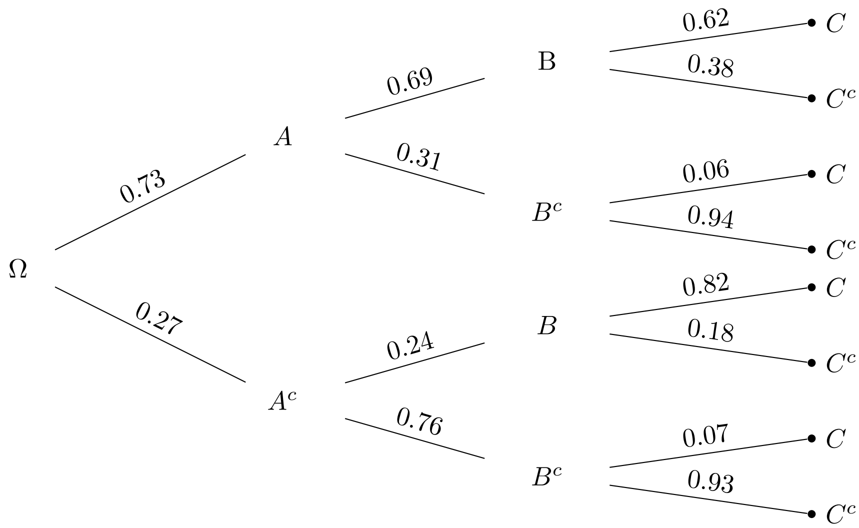
(a) Draw a tree describing the situation with the probabilities associated.

(b) From examining the probabilities in the tree, do you think that women are disadvantaged in the selection process? Why or why not?

(c) Calculate $\mathbb{P}\left[C \mid A\right]$ and $\mathbb{P}\left[C \mid A^c\right]$, i.e., the acceptance probabilities for men and women. Does this agree with your answer in (b)? Can you explain what is going on?

Solution 4.2

(a) A tree can be drawn as follows:



(b) We can see that

$$\mathbb{P}[C \mid B \cap A^c] \geq \mathbb{P}[C \mid B \cap A]$$

and

$$\mathbb{P}[C \mid B^c \cap A^c] \geq \mathbb{P}[C \mid B^c \cap A].$$

This means that in both deparments the acceptance rate are higher for women than for men. And therefore we cannot say that women are disadvantaged.

(c) We have that

$$\begin{aligned}
\mathbb{P}[C \mid A] &= \frac{\mathbb{P}[C \cap A]}{\mathbb{P}[A]} \\
&= \frac{\mathbb{P}[C \cap A \cap B] + \mathbb{P}[C \cap A \cap B^c]}{\mathbb{P}[A]} \\
&= \frac{\mathbb{P}[C \mid A \cap B]\,\mathbb{P}[A \cap B] + \mathbb{P}[C \mid A \cap B^c]\,\mathbb{P}[A \cap B^c]}{\mathbb{P}[A]} \\
&= \frac{\mathbb{P}[C \mid A \cap B]\,\mathbb{P}[B \mid A]\,\mathbb{P}[A] + \mathbb{P}[C \mid A \cap B^c]\,\mathbb{P}[B^c \mid A]\,\mathbb{P}[A]}{\mathbb{P}[A]} \\
&= 0.62 \times 0.69 + 0.06 \times 0.31 \\
&= 0.4464,
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}[C \mid A^c] &= \mathbb{P}[C \mid A^c \cap B]\,\mathbb{P}[B \mid A^c] + \mathbb{P}[C \mid A^c \cap B^c]\,\mathbb{P}[B^c \mid A^c] \\
&= 0.82 \times 0.24 + 0.07 \times 0.76 \\
&= 0.25.
\end{aligned}$$

In other words, the acceptance rates for men and women are 45% and 25%, respectively. These figures suggest now a totally different conclusion, and women seem to be really disadvantaged.

The explanation of this paradox is as follows: The higher overall rejection rate for women is not due to the gender, but to the fact that a large proportion of women apply to the department with a large rejection rate. (Why that is so is a completely different question and cannot be discussed on the basis of the information given here.)

Indeed, we can compute the acceptance rates of the two departments as

$$\begin{aligned}
\mathbb{P}[C \mid B] &= \frac{\mathbb{P}[C \cap B]}{\mathbb{P}[B]} \\
&= \frac{\mathbb{P}[C \cap B \cap A] + \mathbb{P}[C \cap B \cap A^c]}{\mathbb{P}[B \cap A] + \mathbb{P}[B \cap A^c]} \\
&= \frac{\mathbb{P}[C \mid B \cap A]\,\mathbb{P}[B \cap A] + \mathbb{P}[C \mid B \cap A^c]\,\mathbb{P}[B \cap A^c]}{\mathbb{P}[B \mid A]\,\mathbb{P}[A] + \mathbb{P}[B \mid A^c]\,\mathbb{P}[A^c]} \\
&= \frac{0.62 \times 0.69 \times 0.73 + 0.82 \times 0.24 \times 0.27}{0.69 \times 0.73 + 0.24 \times 0.27} \\
&= 0.648.
\end{aligned}$$

Similar calculations yield $\mathbb{P}[C \mid B^c] \approx 0.065$. So now the above result makes sense when we realize that $\mathbb{P}[B^c|A^c] = 76\%$ of the women apply to the highly selective department II, whereas $\mathbb{P}[B|A] = 69\%$ of the men apply to the much less selective department I.

In general, Simpson's paradox shows that correlation and causality can differ widely if an important variable (e.g. in this case the department) is left out of the consideration. In practice, one often does not know if important variables are missing. Other interesting examples of this paradox are presented very well in `https://www.statslife.org.uk/the-statistics-dictionary/2012-simpson-s-paradox-a-cautionary-tale-in-advanced-analytics`.

Exercise 4.3 (Monty Hall problem). You are on a game show, and you are given the choice of three doors. Behind one door is a car, behind the others are goats. You pick a door and the host, who knows what is behind the doors, opens another, behind which is a goat. He then asks you, "Do you want to keep your initial chosen door or do you want to switch to the other one?". Assuming that you like cars but not goats, what should you do?

(a) Construct a suitable model where you can answer this question with the help of conditional probabilities.

(b) Try to find an alternative solution (which of course must give the same answer).

Solution 4.3

(a) We number the doors as $1, 2, 3$ in a way that our first choice is door 1. We define $B_i =$"The car is behind door i.", with $i \in \{1, 2, 3\}$, $A_j =$"Moderator open door j.", with $j \in \{2, 3\}$. Let $\Omega = \{1, 2, 3\} \times \{2, 3\}$. Then for $i \in \{1, 2, 3\}$, $B_i = \{i\} \times \{2, 3\}$ and $A_2 = \{1, 3\} \times \{2\}$ and $A_3 = \{1, 2\} \times \{3\}$. We posit the following probabilities:

$\mathbb{P}[B_1] = \mathbb{P}[B_2] = \mathbb{P}[B_3] = \frac{1}{3}$, meaning that the car is placed randomly.

$\mathbb{P}[A_2|B_1] = \mathbb{P}[A_3|B_1] = \frac{1}{2}$, meaning that the moderator randomly picks a goat door if we happened to pick the car door.

$\mathbb{P}[A_2|B_2] = 0, \quad \mathbb{P}[A_3|B_2] = 1.$

$\mathbb{P}[A_2|B_3] = 1, \quad \mathbb{P}[A_3|B_3] = 0.$ Then we compute, with Bayes' formula,

$$\begin{aligned}
\mathbb{P}[B_1|A_2] &= \frac{\mathbb{P}[A_2|B_1]\,\mathbb{P}[B_1]}{\mathbb{P}[A_2|B_1]\,\mathbb{P}[B_1] + \mathbb{P}[A_2|B_2]\,\mathbb{P}[B_2] + \mathbb{P}[A_2|B_3]\,\mathbb{P}[B_3]} \\
&= \frac{\frac{1}{2}\frac{1}{3}}{\frac{1}{2}\frac{1}{3} + 0\frac{1}{3} + 1\frac{1}{3}} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}
\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}[B_1|A_3] &= \frac{\mathbb{P}[A_3|B_1]\,\mathbb{P}[B_1]}{\mathbb{P}[A_3|B_1]\,\mathbb{P}[B_1] + \mathbb{P}[A_3|B_2]\,\mathbb{P}[B_2] + \mathbb{P}[A_3|B_3]\,\mathbb{P}[B_3]}\\ &= \frac{\frac{1}{2}\frac{1}{3}}{\frac{1}{2}\frac{1}{3} + 1\frac{1}{3} + 0\frac{1}{3}} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}.\end{aligned}$$

We see that $\mathbb{P}[B_1 \mid A_2] = \mathbb{P}[B_1 \mid A_3] = \mathbb{P}[B_1]$.

But this also gives $\mathbb{P}[B_3 \mid A_2] = 1 - \mathbb{P}[B_1 \mid A_2] = \frac{2}{3}$ and $\mathbb{P}[B_2 \mid A_3] = 1 - \mathbb{P}[B_1 \mid A_3] = \frac{2}{3}$. You should pick the other door, not the initial one.You did not obtain additional information about door 1, but you obtained additional information on the last door.

(b) Take $\Omega = \{1,2,3\} \times \{1,2,3\}$, $\mathcal{F} = 2^\Omega$ and $\mathbb{P}$ the uniform distribution. For $\omega = (\omega_1, \omega_2)$, ω_1 is the number of the door with the car and ω_2 the door chosen in the first step. The decision to take is then whether we switch to another door in the second choice or not. If $w_1 = w_2$, we lose by switching; but if $\omega_1 \neq \omega_2$, we win by switching because one door is already open. So the probability of winning the car is $\frac{6}{9} = \frac{2}{3}$ if we switch, but only $\frac{3}{9} = \frac{1}{3}$ if we do not switch. So we should abandon our first choice and switch.

Exercise 4.4 Let $(S_n)_{n=0,1,\dots,N}$ be a random walk and T_0 the time of its first return to 0. Prove in detail that

$$\mathbb{P}[T_0 > 2n \mid X_1 = +1] = \mathbb{P}[T_{-1} > 2n-1]\,,$$

if $2n < N$ holds.

Solution 4.4 By definition, $\mathbb{P}[T_0 > 2n \mid X_1 = +1] = \frac{\mathbb{P}[T_0 > 2n, X_1 = +1]}{\mathbb{P}[X_1 = +1]}$, where $\mathbb{P} = \mathbb{P}_N$ is the uniform distribution on 2^Ω with $\Omega = \{-1,+1\}^N$. But we know that the uniform distributions on $\{-1,+1\}^{2n}$ and $\{-1,+1\}^{2n-1}$ also give random walks of length $2n$ and $2n-1$, respectively, so that $\mathbb{P}_N[T_{-1} > 2n-1] = \mathbb{P}_{2n-1}[T_{-1} > 2n-1]$ and similarly for P_{2n}. Now the set $B := \{\omega \in \{-1,+1\}^{2n-1} : T_{-1} > 2n-1\}$ can be bijectively mapped to the set $A := \{\omega \in \{-1,+1\}^{2n} : X_1 = +1, T_0 > 2n\}$, simply by looking at the last $2n-1$ steps of the longer trajectory. So we obtain

$$\begin{aligned}\mathbb{P}_N[T_0 > 2n, X_1 = +1] &= \mathbb{P}_{2n}[T_0 > 2n, X_1 = +1] =\\ &= 2^{-2n}|A| = \frac{1}{2}2^{-(2n-1)}|B| =\\ &= \frac{1}{2}\mathbb{P}_{2n-1}[T_{-1} > 2n-1] =\\ &= \mathbb{P}_N[X_1 = +1]\mathbb{P}_N[T_{-1} > 2n-1].\end{aligned}$$

This proves the result.

If you have feedback regarding the exercise sheets, please send a mail to Jakob Heiss.