Wahrscheinlichkeitstheorie und Statistik

Lösungen Serie 4

Version 1.2 (26. März 2023: url in Lösung 4.2(c) repariert); Version 1.1 (24. Juni 2021: Lösung richtig verlinkt); Version 1 (13. März)

Bitte stellt Fragen in den Übungen und/oder im Forum des Moodle-Kurs und/oder (anonym) in diesem file https://docs.google.com/document/d/1AhLJZNTRsgkoC9Bmszy8kbYfTxFrEYZnSXGlSlm_0zc/edit?usp=sharing

Wir empfhelen die Aufgaben selbständig zu lösen und dann auf https://sam-up.math.ethz.ch/ hochzuladen oder selbst mit dieser Lösung zu vergleichen am besten rechtzeitig vor der Übung am **22. März**.

Aufgabe 4.1 Morgen wird es entweder ausschliesslich regnen oder schneien. Die Wahrscheinlichkeit, dass es regnen wird ist $\frac{2}{5}$ und die Wahrscheinlichkeit für Schnee liegt bei $\frac{3}{5}$. Sollte es regnen, dann ist die Wahrscheinlichkeit zu spät zur Vorlesung zu kommen $\frac{1}{5}$, während die Wahrscheinlichkeit bei Schneewetter bei $\frac{3}{5}$ liegt.Wie wahrscheinlich ist es zu spät in der Vorlesung zu erscheinen?

Lösung 4.1 Wir bezeichnen mit A das Ereignis zu spät zu kommen und mit B das Ereignis, dass es anfängt zu regnen. Da es entweder nur regnen oder schneien kann, ist B^c das Ereignis, dass es zu schneien anfängt. Wir wenden die Formel für die totale Wahrscheinlichkeit an und erhalten

$$\mathbb{P}[A] = \mathbb{P}[A|B]\,\mathbb{P}[B] + \mathbb{P}[A|B^c]\,\mathbb{P}[B^c] = \frac{1}{5}\cdot\frac{2}{5} + \frac{3}{5}\cdot\frac{3}{5} = \frac{11}{25}.$$

Aufgabe 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:

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

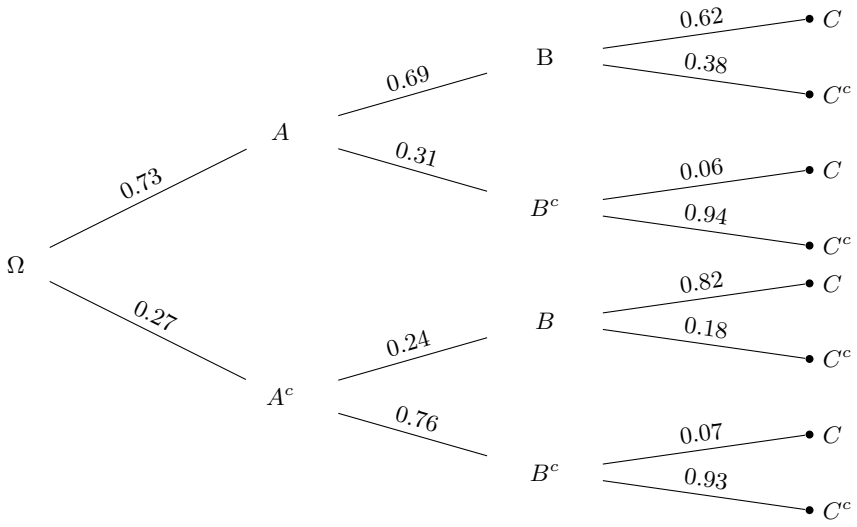
(a) Draw a tree describing the situation with the probabilities associated.

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

Lösung 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://web.archive.org/web/20200510171740/https://www.statslife.org.uk/the-statistics-dictionary/2012-simpson-s-paradox-a-cautionary-tale-in-advanced-analytics.

Aufgabe 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).

Lösung 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.

Aufgabe 4.4 Es ist sicher, dass ein betimmter Patient p eine der Krankheiten k_1, k_2 or k_3 hat. Um herauszufinden, welche dieser Krankheiten den Patienten befallen haben, werden zwei Tests hintereinander ausgeführt. Diese Tests haben als Ergebnis entweder positiv (+) oder negativ (–). In der Tabelle unten bedeutet + –, dass der erste Test positiv war und der zweite Test negativ, wobei der Patient die Krankheit auch wirklich hatte (analog definieren wir + +, – +, und – –). Wir erhalten bei 10'000 Patienten die folgende Tabelle:

Krankheit	Anzahl von Patienten mit der jeweiligen Krankheit	Testergebnisse + +	+ –	– +	– –
k_1	3'215	2'110	301	704	100
k_2	2'125	396	132	1'187	410
k_3	4'660	510	3'568	73	509
Insgesamt	10'000				

(a) Wie hoch ist die Wahrscheinlichkeit, dass der Patient p die Krankheit k_i, $i = 1,2,3$ hat, bevor der Test durchgeführt wird?

(b) Wie hoch ist die Wahrscheinlichkeit, dass der Patient p die Krankheit k_3 hat, unter der Voraussetzung, dass beide Tests positiv waren? Was gilt wenn beide negativ waren?

Lösung 4.4 Wir definieren die (paarweise disjunkten) Ereignisse $K_i = \text{‹Der Patient hat die Krankheit } k_i\text{›}$ for $i = 1,2,3$. Wir definieren mit $T_{+,-}$ das Ereignis, dass der erste Test positiv und der zweite negativ war. Analog definieren wir $T_{+,+}$, $T_{-,+}$, und $T_{-,-}$.

a) Wir berechnen:

$$\mathbb{P}(K_1) = \frac{3'215}{10'000} = 0.3215,$$
$$\mathbb{P}(K_2) = \frac{2'125}{10'000} = 0.2125,$$
$$\mathbb{P}(K_3) = \frac{4'660}{10'000} = 0.466.$$

Beachte, dass $\mathbb{P}(K_1) + \mathbb{P}(K_2) + \mathbb{P}(K_3) = 1$ gilt.

b) Mit dem Satz von Bayes erhalten wir

$$\begin{aligned}\mathbb{P}(K_3 \mid T_{+,+}) &= \frac{\mathbb{P}(T_{+,+} \mid K_3)\mathbb{P}(K_3)}{\mathbb{P}(T_{+,+})} \\ &= \frac{\mathbb{P}(T_{+,+} \mid K_3)\mathbb{P}(K_3)}{\mathbb{P}(T_{+,+} \mid K_1)\mathbb{P}(K_1) + \mathbb{P}(T_{+,+} \mid K_2)\mathbb{P}(K_2) + \mathbb{P}(T_{+,+} \mid K_3)\mathbb{P}(K_3)} \\ &= \frac{\frac{510}{4'660} \cdot 0.466}{\frac{2'110}{3'215} \cdot 0.3215 + \frac{396}{2'125} \cdot 0.2125 + \frac{510}{4'660} \cdot 0.466} \approx 0.17,\end{aligned}$$

$$\begin{aligned}\mathbb{P}(K_3 \mid T_{-,-}) &= \frac{\mathbb{P}(T_{-,-} \mid K_3)\mathbb{P}(K_3)}{\mathbb{P}(T_{-,-})} \\ &= \frac{\mathbb{P}(T_{-,-} \mid K_3)\mathbb{P}(K_3)}{\mathbb{P}(T_{-,-} \mid K_1)\mathbb{P}(K_1) + \mathbb{P}(T_{-,-} \mid K_2)\mathbb{P}(K_2) + \mathbb{P}(T_{-,-} \mid K_3)\mathbb{P}(K_3)} \\ &= \frac{\frac{509}{4'660} \cdot 0.466}{\frac{100}{3'215} \cdot 0.3215 + \frac{410}{2'125} \cdot 0.2125 + \frac{509}{4'660} \cdot 0.466} \approx 0.50.\end{aligned}$$

Weitere interessante Paradoxa zu bedingten Wahrscheinlichkeiten sind hier zu finden: https://towardsdatascience.com/the-false-positive-paradox-f86448a524bc

Wenn du Feedback zum Übungszettel hast, schreibe bitte eine Mail an Jakob Heiss.