Probability and Statistics

Exercise sheet 2

Exercise 2.1 Birthday problem (without Stirling's approximation).

Consider again n people assumed to have the same probability of being born on any day of the year calendar. Assume also that this calendar consists of 365 days. Write

$$A_n = \{\text{at least 2 people in the group have the same birthday}\}.$$

The aim of this exercise is to find bounds for $P(A_n^c)$ for any fixed $2 \leq n \leq 365 = N$.

(a) Show that for any two given people i and j $(1 \leq i \neq j \leq n)$ from the group,

$$P(i \text{ and } j \text{ have the same birthday}) = \frac{1}{N}.$$

(b) Using (a), and the known inequality

$$\log(1+x) \leq x, \quad x \in (-1, \infty)$$

show that

$$1 - \frac{n(n-1)}{2N} \leq P(A_n^c) \leq \exp\left(-\frac{n(n-1)}{2N}\right).$$

(c) Find that $n_m = 23$ is the smallest n so that $P(A_n) \geq \frac{1}{2}$.

Solution 2.1

(a) Fix $(i,j) \in \{1,...,n\}^2 : i \neq j$ and consider

$$B_{ij} := \{i \text{ and } j \text{ are born on the same day}\}.$$

It is not difficult to see that counting the sample points in B_{ij} corresponds exactly to counting the number of $(n-1)$-tuplets $(w_1, ..., w_{n-1})$ with each $w_k \in \{1, ..., N\}$. Thus $\text{card}(B_{ij}) = N^{n-1}$ so that

$$P(B_{ij}) = \frac{N^{n-1}}{N^n} = \frac{1}{N}.$$

(b) We have that

$$P(A_n^c) = \frac{N(N-1)...(N-n+1)}{N^n} = \prod_{j=0}^{n-1}\left(1 - \frac{j}{N}\right).$$

Hence,

$$\begin{aligned}\log(P(A_n^c)) &= \sum_{j=0}^{n-1} \log\left(1 - \frac{j}{N}\right) \\ &\leq -\frac{1}{N}\sum_{j=0}^{n-1} j \\ &= -\frac{1}{N}\frac{(n-1)n}{2}.\end{aligned}$$

This shows the upper bound

$$P(A_n^c) \leq \exp\left(-\frac{n(n-1)}{2N}\right).$$

To show the lower bound, note that

$$A_n = \bigcup_{1\leq i<j\leq n} B_{ij}.$$

where B_{ij} is the same event defined before in (a).
Hence,

$$\begin{aligned}P(A_n) &= P\left(\bigcup_{1\leq i<j\leq n} B_{ij}\right) \\ &\leq \sum_{1\leq i<j\leq n} P(B_{ij}) \\ &= \frac{1}{N}\left(\sum_{1\leq i<j\leq n} 1\right) \\ &= \binom{n}{2} \times \frac{1}{N} \\ &= \frac{n(n-1)}{2N}.\end{aligned}$$

The latter is also equivalent to

$$1 - \frac{n(n-1)}{2N} \leq P(A_n^c).$$

(c)

$$P(A_n) \geq \frac{1}{2} \Leftrightarrow P(A^c) \leq \frac{1}{2}.$$

Now, if $\exp\left(-\frac{n(n-1)}{2N}\right) \leq \frac{1}{2}$, then $P(A_n^c) \leq \frac{1}{2}$.

$$\begin{aligned}\exp\left(-\frac{n(n-1)}{2N}\right) \leq \frac{1}{2} &\Leftrightarrow \frac{n(n-1)}{2N} \geq \log(2) \\ &\Leftrightarrow n^2 - n - 2N\log(2) \geq 0 \\ &\Leftrightarrow n \geq \frac{1+\sqrt{1+8N\log(2)}}{2} =: n_1\end{aligned}$$

With $N = 365$, we find that $n_1 \approx 22.99$ and hence $n \geq 23 \Rightarrow P(A_n) \geq \frac{1}{2}$.
On the other hand, if $P(A_n) \geq \frac{1}{2}$ then $P(A_n^c) \leq \frac{1}{2}$ and we should have

$$\begin{aligned}1 - \frac{n(n-1)}{2N} \leq \frac{1}{2} &\Leftrightarrow \frac{n(n-1)}{2N} \geq \frac{1}{2}\\ &\Leftrightarrow n^2 - n - N \geq 0\\ &\Leftrightarrow n \geq \frac{1+\sqrt{1+4N}}{2} =: n_2\end{aligned}$$

with $n_2 = 19.61$.
Hence $P(A_n) \geq \frac{1}{2} \Rightarrow n \geq 20$.
Now

$$\begin{aligned}P(A_{20}) &= 0.411\\ P(A_{21}) &= 0.443\\ P(A_{22}) &= 0.475\\ P(A_{23}) &= 0.507\end{aligned}$$

and $n_m = 23$.

Exercise 2.2 We have a box which contains 3 different coins. Each of these coins has a different probability to show heads after it is tossed. Call these probabilities $p_j, j = 1, 2, 3$. We know that

$$p_1 = \frac{1}{4}, \quad p_2 = \frac{1}{2}, \quad p_3 = \frac{3}{4}.$$

(a) We select a coin from the box completely at random. When this coin is tossed, it shows heads. What is the conditional probability that coin j was selected?

(b) The same coin is tossed again. What is the conditional probability of obtaining heads again?

(c) [1] Prove the following result:
Let $A_1, ..., A_k$ be a partition of Ω, and B, C events in Ω with $P(B \cap C) > 0$ and $P(A_i \cap B) > 0$ for $i = 1, ..., k$. Then,

$$P(A_j \mid B \cap C) = \frac{P(A_j \mid B)P(C \mid A_j \cap B)}{\sum_{i=1}^k P(A_i \mid B)P(C \mid A_i \cap B)}.$$

(This result is called the conditional Bayes' theorem.)

(d) If the same coin shows again heads at the second toss, what is the conditional probability that coin j was selected?

Solution 2.2

(a) Define the events

$$A_j := \{\text{coin } j \text{ is selected}\}, \ (j = 1, 2, 3)$$

[1] This part was updated a posteriori.

$$H_1 := \{\text{we obtain heads at the first toss}\},$$

$$H_2 := \{\text{we obtain heads at the second toss}\}.$$

Then

$$\begin{aligned} P(A_j \mid H_1) &= \frac{P(A_j \cap H_1)}{P(H_1)} \\ &= \frac{P(H_1 \mid A_j)P(A_j)}{\sum_{i=1}^3 P(H_1 \mid A_i)P(A_i)}. \end{aligned}$$

Now,

$$P(A_1) = P(A_2) = P(A_3) = \frac{1}{3};$$

$$P(H_1 \mid A_j) = p_j = \begin{cases} \frac{1}{4}, & j = 1, \\ \frac{1}{2}, & j = 2, \\ \frac{3}{4}, & j = 3. \end{cases}$$

Thus

$$\begin{aligned} P(A_j \mid H_1) &= \frac{p_j/3}{\sum_{i=1}^3 p_i/3} \\ &= \frac{p_j}{\frac{1}{4} + \frac{1}{2} + \frac{3}{4}} \\ &= \frac{2p_j}{3} \\ &= \begin{cases} \frac{1}{6}, & j = 1 \\ \frac{1}{3}, & j = 2 \\ \frac{1}{2} & j = 3 \end{cases} \end{aligned}$$

(b) We want to calculate $P(H_2 \mid H_1)$:

$$P(H_2 \mid H_1) = \frac{P(H_1 \cap H_2)}{P(H_1)}$$

with

$$\begin{aligned} P(H_1 \cap H_2) &= \sum_{i=1}^3 P(A_i \cap H_1 \cap H_2) \\ &= \sum_{i=1}^3 P(H_1 \cap H_2 \mid A_i)P(A_i) \\ &= \frac{1}{3} \sum_{i=1}^3 P(H_1 \mid A_i)P(H_2 \mid A_i) \end{aligned}$$

because given that we know that we are tossing a particular coin (j, say), the event of obtaining heads in the first toss and heads in the second toss are independent. Also, they have the same (conditional) probability p_j. Hence

$$\begin{aligned} P(H_2 \mid H_1) &= \frac{\frac{1}{3}\sum_{i=1}^{3} p_i^2}{\frac{1}{3}\sum_{i=1}^{3} p_i} \\ &= \frac{\frac{1}{16}+\frac{1}{4}+\frac{9}{16}}{\frac{3}{2}} \\ &= \frac{7}{12}. \end{aligned}$$

(using $P(H_1) = \sum_{i=1}^{3} P(H_1 \mid A_i)\frac{1}{3} = \frac{1}{3}\sum_{i=1}^{3} p_i$).

(c) We have that

$$P(A_j \mid B \cap C) = \frac{P(A_j \cap B \cap C)}{P(B \cap C)}$$

with

$$P(A_j \cap B \cap C) = P(C \mid A_j \cap B)P(A_j \cap B) = P(C \mid A_j \cap B)P(A_j \mid B)P(B)$$

and

$$P(B \cap C) = \sum_{i=1}^{k} P(A_i \cap B \cap C) = \sum_{i=1}^{k} P(C \mid A_i \cap B)P(A_i \mid B)P(B),$$

therefore

$$P(A_j \mid B \cap C) = \frac{P(A_j \mid B)P(C \mid A_j \cap B)}{\sum_{i=1}^{k} P(A_i \mid B)P(C \mid A_i \cap B)}.$$

(d) We want to compute $P(A_j \mid H_1 \cap H_2)$. Using (c) we can write

$$P(A_j \mid H_1 \cap H_2) = \frac{P(A_j \mid H_1)P(H_2 \mid A_j \cap H_1)}{\sum_{i=1}^{3} P(A_i \mid H_1)P(H_2 \mid A_i \cap H_1)}$$

where $P(A_j \mid H_1)$ was calculated in (a), and $P(H_2 \mid A_j \cap H_1) = p_j$ (knowing that the first toss of A_j resulted in heads does not affect the probability of obtaining heads in any other toss).

To be more convinced,

$$\begin{aligned} P(H_2 \mid A_j \cap H_1) &= \frac{P(H_2 \cap H_1 \cap A_j)}{P(A_j \cap H_1)} \\ &= \frac{P(H_2 \mid A_j)P(H_1 \mid A_j)}{P(H_1 \mid A_j)} \\ &= P(H_2 \mid A_j) \\ &= p_j. \end{aligned}$$

Therefore,

$$
\begin{aligned}
P(A_j \mid H_1 \cap H_2) &= \frac{P(A_j \mid H_1)p_j}{\sum_{i=1}^{3} P(A_i \mid H_1)p_i} \\
&= \begin{cases} \frac{\frac{1}{6}\times\frac{1}{4}}{\frac{1}{6}\times\frac{1}{4}+\frac{1}{3}\times\frac{1}{2}+\frac{1}{2}\times\frac{3}{4}}, & j=1 \\ \frac{\frac{1}{3}\times\frac{1}{2}}{\frac{1}{6}\times\frac{1}{4}+\frac{1}{3}\times\frac{1}{2}+\frac{1}{2}\times\frac{3}{4}}, & j=2 \\ \frac{\frac{1}{2}\times\frac{3}{4}}{\frac{1}{6}\times\frac{1}{4}+\frac{1}{3}\times\frac{1}{2}+\frac{1}{2}\times\frac{3}{4}}, & j=3 \end{cases} \\
&= \begin{cases} \frac{1}{14}, & j=1 \\ \frac{2}{7}, & j=2 \\ \frac{9}{14}. & j=3 \end{cases}
\end{aligned}
$$

Exercise 2.3 (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:

$$P(A) = 0.73,$$
$$P(B \mid A) = 0.69, P(B \mid A^c) = 0.24,$$
$$P(C \mid A \cap B) = 0.62, P(C \mid A^c \cap B) = 0.82,$$
$$P(C \mid A \cap B^c) = 0.06, P(C \mid A^c \cap B^c) = 0.07.$$

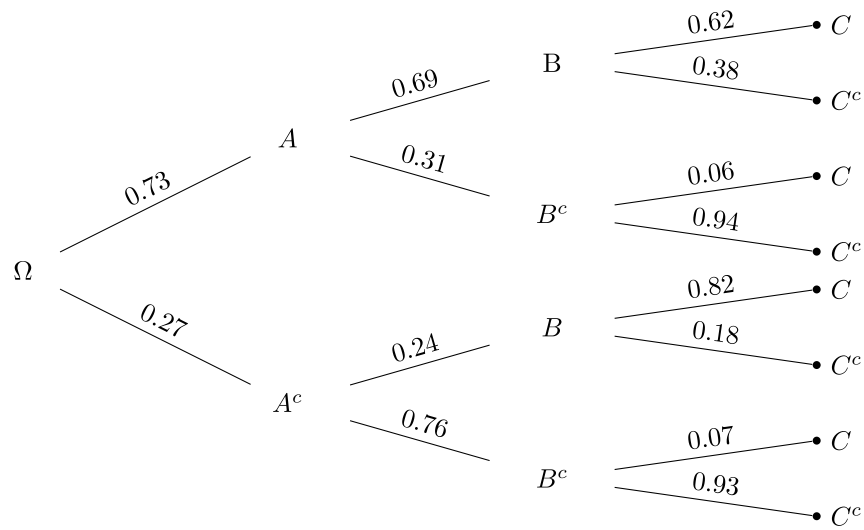
(a) Draw a tree describing the situation with the probabilities associated.

(b) From examining the probabilities in the tree, do you think that in the selection process the women are disadvantaged?

(c) Calculate $P(C \mid A)$ and $P(C \mid A^c)$. Do you agree with your answer in 3.2? What is going on?

Solution 2.3

(a) A tree can be drawn as follows:



(b) We can see that

$$P(C \mid B \cap A^c) \geq P(C \mid B \cap A)$$

and

$$P(C \mid B^c \cap A^c) \geq P(C \mid B^c \cap A)$$

and therefore we can't really say that women are disadvantaged.

(c) We have that

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

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

These figures suggest now a totally different conclusion, and women seem now to be really disadvantaged.

The higher rejection rate 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.

Indeed,

$$\begin{aligned}
P(C \mid B) &= \frac{P(C \cap B)}{P(B)} \\
&= \frac{P(C \cap B \cap A) + P(C \cap B \cap A^c)}{P(B \cap A) + P(B \cap A^c)} \\
&= \frac{P(C \mid B \cap A)P(B \cap A) + P(C \mid B \cap A^c)P(B \cap A^c)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)} \\
&= \frac{0.62 \times 0.69 \times 0.73 + 0.82 \times 0.24 \times 0.27}{0.69 \times 0.73 + 0.24 \times 0.27} \\
&= 0.648
\end{aligned}$$

Similar calculations yield $P(C \mid B^c) \approx 0.065$, so now this makes sense considering that 76% of women apply to department II vs. 69% of men who apply to department I.