MTH794P Probability and Statistics for Data Analytics — Problem Sheet 3

Kavit Tolia
October 16, 2025

Problem 1.
I have three coins in my pocket. Two of them are ordinary fair coins, the third has Heads on both sides.

(a) I take a random coin from my pocket and toss it. What is the probability that it comes up Heads?

(b) Given that it did come up Heads, what is the conditional probability that it is the double-headed coin?

(c) Given that it did come up Heads, what is the conditional probability that a second toss of the same coin also comes up Heads?

(d) Given that two tosses of the same coin both come up Heads, what is the conditional probability that it is the double-headed coin? [Before doing this think do you expect the answer to be larger or smaller than the answer to part (b)? Why?]

(e) Suppose that 100 tosses of the same coin show a Head every time. Without doing any more calculations, say roughly what you expect the conditional probability that it is the double-headed coin to be? How would you explain this in non-mathematical terms?

Solution.

(a) Let $\mathbb{P}(H)$ be the probability of the coin showing Heads, $\mathbb{P}(F)$ be the probability that it is a fair coin and $\mathbb{P}(B)$ be the probability that it is a biased coin. Then,

$$\mathbb{P}(H) = \mathbb{P}(H \mid F)\,\mathbb{P}(F) + \mathbb{P}(H \mid B)\,\mathbb{P}(B) = \left(\frac{1}{2}\right)\left(\frac{2}{3}\right) + (1)\left(\frac{1}{3}\right) = \frac{2}{3}$$

(b) Now we want to work out $\mathbb{P}(B|H)$:

$$\mathbb{P}(B \mid H) = \frac{\mathbb{P}(H \mid B)\mathbb{P}(B)}{\mathbb{P}(H)} = \frac{(1)\left(\frac{1}{3}\right)}{\frac{2}{3}} = \frac{1}{2}$$

(c) Now let's say H_i is the i^{th} toss, then we want to calculate $\mathbb{P}(H_2 = H \mid H_1 = H)$:

$$\mathbb{P}(H_2 = H \mid H_1 = H) = \frac{\mathbb{P}(H_2 = H \cap H_1 = H)}{\mathbb{P}(H_1 = H)} = \frac{\frac{1}{2}}{\frac{2}{3}} = \frac{3}{4}$$

(d) I expect this probability to be higher than (b), as I'm more certain of the fact that it is now a baised coin.

$$\mathbb{P}(B \mid HH) = \frac{\mathbb{P}(HH \mid B)\mathbb{P}(B)}{\mathbb{P}(HH)} = \frac{(1)\left(\frac{1}{3}\right)}{\frac{1}{2}} = \frac{2}{3}$$

(e) I would expect this conditional probability to be close to 1, as it is highly unlikely that the coin is a fair coin. If it was a fair coin, the chances of getting 100 Heads in a row is close to impossible. The only plausible explanation is that it is the double-headed coin.

Problem 2.

A certain medical condition affect 1% of the population. A new AI tool for detecting this condition from a scan has been developed. It has a 95% success rate at correctly detecting that a person with the condition has it, and only a 2% chance of incorrectly deciding that a healthy person has the condition. A randomly chosen person from the population undertakes this test and the test shows positive.

(a) What is the probability that they do have the condition?

(b) A politician suggests that this test could be used for a national screening programme. What insight into the possible disadvantages of doing this does the calculation of part (a) provide? How could these disadvantages be mitigated?

Solution.

(a) Let's define some outcomes. Let H denote a person being healthy and P denote a positive test result. Then,

$$\mathbb{P}(H^c \mid P) = \frac{\mathbb{P}(P \mid H^c)\mathbb{P}(H^c)}{\mathbb{P}(P)} = \frac{(95\%)\,(1\%)}{(95\%)\,(1\%) + (2\%)\,(99\%)} = 32.5\%$$

(b) There is a major disadvantage here in that you have less than 1 in 3 chance of a person actually having the condition if they tested positive. This seems odd given the test's high accuracy. But this happens due to the low rate of the condition in the general population.

We can mitigate this by either:

- We can use the test as a first screen, followed by more robust tests for those who are positive.
- Only screen high-risk groups, where we know the prevalence to be higher, significantly increasing the probability that a positive test is indeed someone with the condition.

Problem 3.

Suppose that A and B are events with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$ and that $\mathbb{P}(A \mid B) > \mathbb{P}(A)$.

(a) What can you say about $\mathbb{P}(B \mid A)$?

(b) How do you explain the relationship between events A and B which satisfy this property in non-mathematical language?

Solution.

(a) Let us write this out:

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)} > \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(A)} = \mathbb{P}(B)$$

So, we have $\mathbb{P}(B \mid A) > \mathbb{P}(B)$

(b) This tells us that the occurance of one event makes the other more likely to happen. In laymen's terms, A and B are positively linked, as they occur together more often.

Problem 4.

(a) Suppose that T is a discrete random variable measuring the number of days until some event happens. We say that T has a Geometric(p) distribution if it takes values in 1, 2, 3, ... and the pmf is $\mathbb{P}(T = k) = p\,(1-p)^{k-1}$.

(i) What is $\mathbb{P}(T > k)$?

(ii) Show that for any $a, k \in \mathbb{N}$, we have $\mathbb{P}(T > a + k \mid T > a) = \mathbb{P}(T > k)$.

(b) Suppose that the lifetime of a component is a continuous random variable V which follows an Exponential distribution with parameter λ.

(i) What is $\mathbb{P}(V > k)$?

(ii) Show that for any $a, k \in \mathbb{R}$, we have $\mathbb{P}(V > a + k \mid V > a) = \mathbb{P}(V > k)$.

(c) Why do you think the results in parts (a)(ii) and (b)(ii) are sometimes called the *memoryless property* of the Geometric and Exponential distributions?

Solution.

(a) (i) Given we have the pmf of T defined, we can work out the cdf:

$$F_T(k) = \mathbb{P}(T \leq k) = \sum_{i=1}^{k} p\,(1-p)^{i-1} = p\sum_{j=0}^{k-1}(1-p)^j = \frac{p\,(1-(1-p)^k)}{1-(1-p)} = 1-(1-p)^k$$

Using this cdf, we can then work out $\mathbb{P}(T > k)$:

$$\mathbb{P}(T > k) = 1 - \mathbb{P}(T \leq k) = 1 - F_T(k) = 1 - \Big(1-(1-p)^k\Big) = (1-p)^k$$

(ii) We can show this as follows:

$$\mathbb{P}(T > a + k \mid T > a) = \frac{\mathbb{P}(T > a + k \cap T > a)}{\mathbb{P}(T > a)} = \frac{\mathbb{P}(T > a + k)}{\mathbb{P}(T > a)} = \frac{(1-p)^{a+k}}{(1-p)^a} = (1-p)^k$$

Which is the same as $\mathbb{P}(T > k)$.

(b) (i) The cdf of an Exponential distribution with parameter λ is given by $1 - e^{-\lambda k}$ for $k \geq 0$ and 0 otherwise. Given this, we then have:

$$\mathbb{P}(V > k) = 1 - \mathbb{P}(V \leq k) = 1 - F_V(k) = 1 - (1 - e^{-\lambda k}) = e^{-\lambda k} \text{ for } k \geq 0 \text{ and 1 otherwise}$$

(ii) The proof below is for cases when $a \geq 0$ and $k \geq 0$. If either $a < 0$ or $k < 0$, then the conditional probability is trivial.

$$\mathbb{P}(V > a + k \mid V > a) = \frac{\mathbb{P}(V > a + k \cap V > a)}{\mathbb{P}(V > a)} = \frac{\mathbb{P}(V > a + k)}{\mathbb{P}(V > a)} = \frac{e^{-\lambda(a+k)}}{e^{-\lambda a}} = e^{-\lambda k}$$

Which is the same as $\mathbb{P}(V > k)$.

(iii) This is called the *memoryless property* because the probability of waiting an additional amount of time does not depend on how much time has already passed. So, the distribution has no "memory" of the past and the process "restarts" at each point in time.

Problem 5.

Read this article from mathematical epidemiologist Adam Kucharski's blog:
`https://kucharski.substack.com/p/small-hallucinations-big-problems`

(a) Using what you learn in this module, redo this analysis in the case that the probability of the unusual event we are looking for is p (rather than the 1 in 1000 that the article uses).

(b) Do you think the author did a good job at explaining the mathematics to a non-mathematical audience?

(c) Look at the Student Forum on the module QMPlus page and make a comment (which could be about your answer to parts (a) or (b) of this question or something else) on the discussion thread about this article.

Solution.

(a) If the probability of the unusual event is p, then let us define E as the event occurring and F as it being flagged by the LLM. Then,

$$\mathbb{P}(E \mid F) = \frac{\mathbb{P}(F \mid E)\,\mathbb{P}(E)}{\mathbb{P}(F \mid E)\,\mathbb{P}(E) + \mathbb{P}(F \mid E^c)\,\mathbb{P}(E^c)} = \frac{0.99p}{0.99p + 0.01(1-p)}$$

We can use this formula to understand how the rarity of an event impacts the probability of a flag being correct:

p (%)	0.1	1.0	10.0	25.0	50.0	75.0	90.0	99.0	99.9
$\mathbb{P}(E \mid F)$ (%)	9.016	50.000	91.667	97.059	99.000	99.664	99.888	99.990	99.999

Table 1: How p impacts the flag being correct.

Two things worth mentioning from this table:

- Unless p exceeds 1%, most flags will be false alarms! And this is with an optimistic hallucination rate of 1%.
- We would need a p of greater than 50% to have a 99% confidence that a flag was real.

One way to get a high confidence with a low p would be to have multiple independent LLMs being used for flagging an event. However, this relies on the assumption that their hallucinations are uncorrelated. This might not be the case if they share similar architecture or training data.

(b) Yes, I think the author did a very good job at translating Bayesian statistics into non-mathematical language. They explained the scenario clearly, provided their reasoning for every step and highlighted the pitfalls of using just accuracy as a measure of model performance.

(c) Done!