HW1, Theory of Inference 2016/7

Jonathan Rougier
School of Mathematics
University of Bristol UK

A general statement on homeworks. These homeworks are an opportunity for you to develop your understanding, and to practice your maths and communication skills. If you hand-in your homeworks, you will get feedback on how well you are doing. You are *strongly encouraged* to hand-in your homeworks.

It is crucial that you express your answers clearly, in well-structured sentences. Where you are writing maths, your writing must be tidy enough that there can be no ambiguity about symbols and the names of variables. The way you lay-out your maths must be logical and clear, using indentation, alignment, and other standard conventions. This is a skill you must master before you leave the university. Future employers and colleagues will rightly be critical of sloppy thinking and sloppy communicating.

I am told by students that my marking criteria are very strict. Please be absolutely clear that I am marking according to the the criteria that you will be judged by when you leave the university. Do not be put off by low marks. Come to an Office Hour to discuss how you could have done better, and study the solutions.

In the following questions, I show marks in square brackets, to give you an idea of the approximate tariff the question would carry in an exam.

1. The usual convention in statistics is to write $\mathrm{p}(x)$ for $\Pr(X = x)$, $\mathrm{p}(x \mid y)$ for $\Pr(X = x \mid Y = y)$, and so on. Under this convention, the definition of the conditional probability $\mathrm{p}(x \mid y)$ is any function that satisfies

$$\mathrm{p}(x, y) = \mathrm{p}(x \mid y)\,\mathrm{p}(y).$$

   This definition implies that $\mathrm{p}(x \mid y)$ is arbitrary when $\mathrm{p}(y) = 0$, because in this case the relation reads $0 = \mathrm{p}(x \mid y) \cdot 0$.

(a) Consider the model $\{\mathcal{X} \times \mathcal{Y}, \Omega, f_{X,Y}\}$, where I will write $p(x, y \mid \theta)$ for $f_{X,Y}(x, y; \theta)$, and so on. I will treat Ω as uncountable. Prove that if $\mathrm{p}(y) > 0$, then

$$\mathrm{p}(x \mid y) = \int_\Omega \mathrm{p}(x \mid y, \theta)\, \mathrm{p}(\theta \mid y)\, \mathrm{d}\theta.$$

This is eq. (1.11) in the handout. [10 marks]

Answer. Applying the rules of the probability calculus, and using the definition of conditonal probability given above,

$$\begin{aligned}
\mathrm{p}(x \mid y) &= \int \mathrm{p}(\theta, x \mid y)\, \mathrm{d}\theta \\
&= \int \frac{\mathrm{p}(\theta, x, y)}{\mathrm{p}(y)}\, \mathrm{d}\theta && \text{accepting that } \mathrm{p}(y) > 0 \\
&= \int \frac{\mathrm{p}(x \mid y, \theta)\, \mathrm{p}(y, \theta)}{\mathrm{p}(y)}\, \mathrm{d}\theta \\
&= \int \mathrm{p}(x \mid y, \theta)\, \mathrm{p}(\theta \mid y)\, \mathrm{d}\theta
\end{aligned}$$

as required.

Feedback. Some people argued

$$\mathrm{p}(x \mid y, \theta)\, \mathrm{p}(\theta \mid y) = \frac{\mathrm{p}(x, y, \theta)}{\mathrm{p}(y, \theta)}\, \frac{\mathrm{p}(y, \theta)}{\mathrm{p}(y)}.$$

But this is only a good strategy if $\mathrm{p}(y, \theta) > 0$, which is unnecessarily strong.

If you would like to read more about conditional probabilities, then see sections 2.1 to 2.3 of the handout *Statistical Modelling*, available at `https://people.maths.bris.ac.uk/~mazjcr/BMB/2016/BMBmodelling.pdf`.

2. Consider the case where $\mathcal{Y} = \Omega = \{-, +\}$, and where

$$f_Y(y; \theta) = \begin{array}{c|cc} & \theta = - & \theta = + \\ \hline y = - & 0.80 & 0.05 \\ y = + & 0.20 & 0.95 \end{array}$$

Suppose that Θ represents whether or not a person has a disease, and Y represents whether or not that person tests positive for the disease. Think of Y as an algorithm for predicting Θ. Such as algorithm is certified in terms of its *sensitivity*, which is is $f_Y(+; +)$, and its *specificity*, which is $f_Y(-; -)$.

(a) Identify the sensitivity and specificity of the algorithm from the values of f_Y. Suppose you have tested positive. Compute the *likelihood ratio*

$$f_Y(+;+)/f_Y(+;-).$$

What does this tell you about whether you have the disease? [10 marks]

Answer. The sensitivity is 0.95 and the specificity is 0.80. The likelihood ratio is

$$\frac{f_Y(+;+)}{f_Y(+;-)} = \frac{\text{sensitivity}}{1-\text{specificity}} = \frac{0.95}{0.2} = \frac{19}{4} = 4.75.$$

This tell me nothing about the disease, if I have tested positive, because for me $Y = +$ is known and Θ is not. The value I want is $\Pr(\theta = + \mid Y = +)$.

More details. Applying Bayes's theorem twice and taking the ratio to cancel the $\Pr(Y = y)$, we get

$$\begin{aligned}\frac{\Pr(\theta = + \mid Y = +)}{\Pr(\theta = - \mid Y = +)} &= \frac{\Pr(Y = + \mid \theta = +)\Pr(\theta = +)}{\Pr(Y = + \mid \theta = -)\Pr(\theta = -)} \\ &= \frac{f_Y(+;+)}{f_Y(+;-)} \times \frac{\Pr(\theta = +)}{\Pr(\theta = -)}.\end{aligned}$$

This is known as *Bayes's theorem in odds form*, because ratios of probabilities are termed 'odds'. So Bayes's theorem in odds form states that

$$\text{posterior odds} = \text{likelihood ratio} \times \text{prior odds}$$

where the odds are for having the disease versus not having it. So if the likelihood ratio is 4.75, then the ratio of posterior odds to prior odds is 4.75, but this does not tell me what the posterior odds are, unless I also know what the prior odds are.

It is easy to see that if $\Pr(A)/\Pr(\neg A) = o$, then $\Pr(A) = o/(o+1)$. Or, more generally, if $\Pr(A)/\Pr(\neg A) = a/b$, then $\Pr(A) = a/(a+b)$.

(b) Suppose that you now find out that the base rate of the disease in people similar to you is 5 people in every 1000. What is the probability that you have the disease, given that you have tested positive? [10 marks]

Answer. Mathematicians usually find it easiest to answer this question

using odds (see above). Hence

$$\begin{aligned}\frac{\Pr(\theta = + \mid Y = +)}{\Pr(\theta = - \mid Y = +)} &= \frac{f_Y(+;+)}{f_Y(+;-)} \times \frac{\Pr(\theta = +)}{\Pr(\theta = -)} \\ &= \frac{19}{4} \times \frac{5}{995} \\ &\approx \frac{100}{4000} = 0.025.\end{aligned}$$

Now if $\Pr(A)/\Pr(\neg A) = 25/1000$, then $\Pr(A) = 25/1025 \approx 0.025$ (see above). So the probability that I have the disease, given that I have tested positive, is approximately 0.025. (The exact answer is $19/815 \approx 0.0233$.) Because the disease is very rare, and the test is not very accurate, my probability of having the disease is still low, despite testing positive.

If you are planning to participate in any form of routine screening, or you have been to the doctor who has arranged for you to have some tests, then you should read *Reckoning with Risk* by Gerd Gigerenzer.