# Statistics for Astronomers
# Solutions to Homework #1

Prof. Sundar Srinivasan

September 29, 2020

1. (a) For Question 1, I assumed implicitly that each answer is equally likely and that each individual answer has a probability of $\dfrac{1}{\text{total number of answers}}$. These numbers are the result of applying the Principle of Indifference.

   (b) For Question 2, I am updating the prior probability of 20% with evidence (that the professor's lectures alluded to this answer in particular), resulting in a higher posterior probability. This is a result of applying the Bayesian interpretation of probability to the problem.

2. The table below shows the breakdown by number of the possible scenarios, with the ones relevant to this problem shown in bold:

| Spectral type | Classified O-rich | Classified C-rich | Total |
|---|---|---|---|
| O-rich | **782** | **20** | 802 |
| C-rich | **3** | 79 | 82 |
| | | | 882 |

   The total number of sources that are either classified as O-rich or misclassified is 782 + 20 + 3 = 805, out of a total of 884 objects. The third of these numbers is for the situation where the sources are classified as O-rich **and** misclassified. The relevant probability is therefore $\dfrac{805}{884} \approx 0.91$.

3. From the table, total number of FIR(RGB) objects in the Magellanic Clouds = 1262 + 303 = 1565. Of these, FIR(RGB) objects in the LMC = 1262.
   Therefore, $P(\text{LMC}|\text{FIR(RGB)}) = \dfrac{1262}{1565} \approx 0.81$.

4. (a) We define the following six events:
   $\mathbf{D}$ = "person infected"; $P(D) = 0.0025$ (given in problem).
   $\mathbf{D}^c$ = "person not infected"; $P(D^c) = 1 - P(D) = 0.9975$.
   $+|\mathbf{D}$ = "test is positive given person is infected"; $P(+|D) = 0.995$ (given in problem).
   $+|\mathbf{D^c}$ = "test is positive given person isn't infected"; $P(+|D^c) = 0.072$ (given in problem).
   This event is called a **false positive** or a Type I error[1].

[1] The rejection of a TRUE null hypothesis is a Type I error ("false positive"), and the failure to reject a FALSE null hypothesis is a Type II error ("false negative"). In this problem, the null hypothesis is that $H_0$ = "person is not infected".
Type I error: $H_0$ is true (*i.e.*, person is not diseased), but the ELISA test is positive for infection.
Type II error: $H_0$ is false (*i.e.*, person is infected), but the ELISA test is negative for infection.
Both types of errors can be accounted for in the Bayesian framework, as demonstrated in this problem.

$-|\mathbf{D}$ = "test is negative given person is infected"; $P(-|D) = 1 - P(+|D) = 0.005$.
This is a **false negative** or a Type II error.
$-|\mathbf{D^c}$ = "test is negative given person isn't infected"; $P(-|D^c) = 1 - P(+|D) = 0.005$.

The problem asks us to compute $P(D|+)$. Using Bayes' Theorem,

$$P(D|+) = \frac{P(+|D)}{P(+)} \times P(D) = \frac{\text{Accuracy of test}}{\text{Total probability of testing positive}} \times \text{Probability of being infected}$$

Using the Law of Total Probability,
$P(+) = P(+|D) \times P(D) + P(+|D^c) \times P(D^c) = 0.995 \times 0.0025 + 0.072 \times 0.9975 \approx 0.074$

Therefore,
$P(D|+) = \frac{0.995}{0.074} \times 0.0025 \approx 0.034 \approx 3\%$

Even though the accuracy of the test is quite high, testing positive does not necessarily mean a high probability of being infected, because of (a) the very low incidence of the disease in the population (very low prior), which reduces the numerator, and (b) the non-zero false positive rate of the test, which increases the denominator.

(b) If the same person is administered a second ELISA test, the accuracy remains the same, as does the total probability of testing positive. The only thing that changes is that we have to update our prior – the probability of being infected increases to 0.034, as the person is no longer from the general population but one who has tested positive. With this updated prior,
$P(D|+) = \frac{0.995}{0.995 \times 0.034 + 0.072 \times (1 - 0.034)} \times 0.034 \approx 33\%$.

5. For brevity, we denote the array [0, 1, 2, 3, 4] as $\vec{C}$, and store the probabilities $P(H_1|C_i)$ in an array:
$P(H_1|\vec{C}) = \frac{1}{4} \times [0, 1, 2, 3, 4]$
$\sum_{i=1}^{5} P(H_1|C_i) = \frac{1}{4} \times 10$
The probability $P(C_i)$ of randomly selecting coin $C_i$ is independent of $i$: $P(C_i) = \frac{1}{5}\ \forall\ i$

(a) $$\begin{aligned} P(C_i|H_1) &= \frac{P(H_1|C_i)}{P(H_1)} \times P(C_i) \text{ (Using Bayes' Theorem)} \\ &= \frac{P(H_1|C_i)}{\sum_{j=1}^{N} P(H_1|C_j) \times P(C_j)} \times P(C_i) \text{ (Law of Total Probability)} \end{aligned}$$
So that, in vector form,
$P(\vec{C}|H_1) = \frac{1}{10} \times [0, 1, 2, 3, 4]$
So, for instance, the probability that the coin #4 was selected, given that the first toss resulted in a head, is equal to $\frac{3}{10}$.

(b) We need to compute $P(H_2|H_1)$, which we first rewrite as
$P(H_2|H_1) = \frac{P(H_2 \cap H_1)}{P(H_1)}$.

As before, the denominator can be written using the Law of Total Probability in terms of conditional probabilities involving the coins $C_i$:

$P(H_1) = \sum_{i=1}^{N} P(H_1|C_i) \times P(C_i)$.

A version of the Law of Total Probability can also be used to rewrite the numerator:

$P(H_2 \cap H_1) = \sum_{i=1}^{N} P(H_2 \cap H_1|C_i) \times P(C_i)$.

The term $P(H_2 \cap H_1|C_i)$ represents the probability of getting two heads once coin $C_i$ is selected. For each coin $C_i$, the outcomes of successive tosses are independent; therefore,

$P(H_2 \cap H_1|C_i) = P(H_2|C_i) \times P(H_1|C_i) = P(H|C_i)^2$.

In vector form, we write

$P(H_2 \cap H_1|\vec{C}) = \dfrac{1}{16} \times [0, 1, 2^2, 3^2, 4^2]$.

Using the fact that $\displaystyle\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$, we get $P(H_2|H_1) = \dfrac{30/16}{10/4} = \dfrac{3}{4}$.

This probability for two consecutive heads seems quite high. It is, however, consistent with the fact that we have two coins with $P(H) > 0.5$ and only one with $0 < P(H) < 0.5$. In fact, one of the coins has $P(H) = 1$. This artificial example biases the probability of success towards higher values. The following code snippet prints out probabilities close to our theoretical answer above:

```
import numpy as np
from scipy.stats import bernoulli
trials = 1000
c = np.arange(5) #the coins
#pick a coin
k = np.random.choice(c, size = trials)
p = 0.25 * k #probability of success once coin k is chosen
tosses = bernoulli.rvs(p, size = (2, len(p)))
numerator = len(np.where(tosses.sum(axis = 0) == 2)[0])
denominator = len(np.where(tosses[0, :]  == 1)[0])
prob = numerator / denominator
print("The probability P(H2|H1) = ".format(np.round(prob, decimals = 3)))
```