Open in Colab

Joe Bormeh Faryean

Definition

↳ 1 cell hidden

Tutorial and Exercises

To use this tutorial, read the text and then try to generate code to solve the exercises. Answers will be posted to GitHub after the class they are due.

The learning objective is to gain insights into thinking about inference from a "Frequentist" versus a "Bayesian" perspective. In brief, because a Frequentist does not consider the probability of an event or state of the world or hypothesis, only their frequency of occurrance, it is not possible to ask questions of the form "what is the probabilty that hypothesis x is true?" Instead, one can only consider questions of the form, "what is the probabilty that I would have obtained my data, given that hypothesis x is true?" In contrast, Bayesians consider the probabilities of such things (often called the strength of belief), but doing so can require making assumptions that can be difficult to prove.

Let's start with a simple example, taken from:

https://en.wikipedia.org/wiki/Base_rate_fallacy#Example_1:_HIV

"Imagine running an HIV test on A SAMPLE of 1000 persons ..."

"The test has a false positive rate of 5% (0.05)..." i.e., the probability that someone who takes the test gets a POSITIVE result despite the fact that the person does NOT have HIV

"...and no false negative rate." i.e., The probability that someone who takes the test gets a NEGATIVE result despite the fact that the person DOES have HIV.

Answers to the exercises below will be found here after the due date.

Double-click (or enter) to edit

Exercise #1: If someone gets a positive test, is it "statistically significant" at the p<0.05 level? Why or why not?

No, it is not because p<0.05 is not the same as p=0.05. This is an arbitrary definition.

Exercise #2: What is the probability that if someone gets a positive test, that person is infected?

Assume there is 30 true positive result among 1000 samples, then calculate the number of false positive among the remaining 970 samples, which gives us 48.5. 30 + 48.5 is the total number of positives test. 30/78.5 = 0.38 will give the probability of true positive results.

Double-click (or enter) to edit

Following on Exercise #2, let's do the same thing, but this time we will try different values for the proportion of the population that is actually infected. What you should notice is that the **PROPORTION INFECTED GIVEN A POSITIVE TEST** depends (a lot!) on the **OVERALL RATE OF INFECTION**. Put another way, to determine the probabilty of a hypothesis, given your data (e.g., proportion infected given a positive test), you have to know the probability that the hypothesis was true without any data.

Why is this the case? It is a simple consequence of the definition of a conditional probability, formulated as Bayes' Rule. Specifically, the joint probability of two events, call them A and B, is defined as:

$$p(A\ and\ B) = p(A) \times p(B \mid A)$$
$$p(B\ and\ A) = p(B) \times p(A \mid B)$$

Now, calling A the Hypothesis and B the Data, then rearranging, we get:

$$p(Hypothesis \mid Data) = \frac{p(Data \mid Hypothesis) \times p(Hypothesis)}{p(Data)}$$

So you cannot calculate the probability of the hypothesis, given the data (i.e., the Bayesian posterior), without knowing the probability of the hypothesis independent of any data (i.e., the prior).

For this exercise, assume a range of priors (infection rates) from 0 to 1 in steps of 0.1.

Double-click (or enter) to edit

Let's assume p(Data|Hypothesis) = 0.95, p(Hypothesis) = range (0 to 1 in steps of 0.1), p(Data) = 0.5

The result of each assumed prior is shown below:

Double-click (or enter) to edit

```
import numpy as np

for prior in np.linspace(0, 1, 10):
  print("p(Data|Hypothesis)", (((0.95)*(prior))/0.5))
```

```
p(Data|Hypothesis) 0.0
p(Data|Hypothesis) 0.21111111111111108
p(Data|Hypothesis) 0.42222222222222217
p(Data|Hypothesis) 0.6333333333333333
p(Data|Hypothesis) 0.8444444444444443
p(Data|Hypothesis) 1.0555555555555556
p(Data|Hypothesis) 1.2666666666666666
p(Data|Hypothesis) 1.4777777777777774
p(Data|Hypothesis) 1.6888888888888887
p(Data|Hypothesis) 1.9
```

Credits