Bayesian Data Analysis

Dr Niamh Cahill (she/her)

Bayes' Rule

Bayes' rule

Thomas Bayes' famous theorem was published in 1763.

For events A and B:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- The branch of statistics that you are probably most familiar with up to now is called *frequentist* statistics.
- Bayesian statistics uses Bayes' rule for inference and decision making, frequentist statistics does not.

Toy example

Suppose that you are interested in the probability of rain in the afternoon in your location.

Let's assume 30% of days have rain in the afternoon.

(1) $P(\text{rain}) = 0.3$

Let's assume you have additional information which is that in the morning it's cloudy.

Now assume you have an updated probability of rain which is conditional on the morning being cloudy.

(2) $P(\text{rain}|\text{cloudy}) = 0.6$

What information did we need to get from (1) to (2)?

Toy example

We added some information (data) about clouds. We need to know the probability of observing that data given what we know about rain.

(3) $P(\text{cloudy}|\text{rain}) = 0.8$

We also need to know what the marginal probability of being cloudy is.

(4) $P(\text{cloudy}) = 0.4$

Now we can put (1), (3) and (4) together using Bayes' rule

$$\underbrace{P(\text{rain}|\text{cloudy})}_{P(A|B)} = \frac{\underbrace{P(\text{cloudy}|\text{rain})}_{P(B|A)}\underbrace{P(\text{rain})}_{P(A)}}{\underbrace{P(\text{cloudy})}_{P(B)}}$$

Disease screening example

- Suppose there is a test for a disease that has a sensitivity of 80% (i.e., if 100 people that have the disease take the test then 80 of them will get a positive result).

P(+iv|disease) = 0.8

- Suppose the chance of having this disease is 2%.

P(disease) = 0.02

- Suppose we also know that the test gives false positive results 5% of the time (this relates to specificity of the test).

P(+iv|no disease) = 0.05

What is the probability you have the disease given that you receive a positive test result?

$$P(\text{disease}| + iv) = \frac{P(+ive|\text{disease})P(\text{disease})}{P(+ive)}$$

Disease screening example

Find the probability that a person has the disease given they get a positive test result.

(1) $P(+ive|\text{disease})P(\text{disease}) = 0.8 \times 0.02$

(2) $P(+ive) =$
$P(+ive|\text{disease})P(\text{disease}) + P(+ive|\text{no disease})P(\text{no disease}) = 0.065$

$$P(\text{disease}| + iv) = \frac{0.8 \times 0.02}{0.065} = 0.25$$

- So even with a positive test result for a test with an 80% "hit rate", the probability of having the disease is only 25%
- The lower probability is a consequence of the low *prior* probability of the disease and the non-negligible false positive rate.

Bayes' rule applied to parameters and data

Given a set of observed data points Y and a set of parameters θ, we write Bayes' rule as

$$\underbrace{P(\theta|Y)}_{\text{posterior}} = \frac{\underbrace{P(Y|\theta)}_{\text{likelihood}}\underbrace{P(\theta)}_{\text{prior}}}{\underbrace{P(Y)}_{\text{marginal likelihood}}}$$

Where the denominator is

$P(Y) = \sum_{\theta^*} P(Y|\theta^*)P(\theta^*)$ for discrete-valued variables, or

$P(Y) = \int P(Y|\theta^*)P(\theta^*)d\theta^*$ for continuous variables.

$P(Y)$ is often difficult to calculate (more on this later) and Baye's rule is often written more simply as a proportional statement

$$\underbrace{P(\theta|Y)}_{\text{posterior}} \propto \underbrace{P(Y|\theta)}_{\text{likelihood}}\underbrace{P(\theta)}_{\text{prior}}$$

Likelihood, Prior & Posterior

$$\underset{\text{posterior}}{P(\theta|Y)} \propto \underset{\text{likelihood}}{P(Y|\theta)}\underset{\text{prior}}{P(\theta)}$$

- $P(Y|\theta)$ which is the probability distribution of the data given the parameters is known as the *likelihood*
- $P(\theta)$ which is the probability distribution of the parameters is known as the *prior*. The prior represents what we know about the parameters before the data are observed.
- $P(\theta|Y)$ which is the probability distribution of the parameters given the data is known as the *posterior*. The posterior represents our updated knowledge about parameters after the data are observed.