Chapter 1 - Probability and Inference

Conor Houghton

reading Gelman

(2022-01-20) bayesianreadinggroup.github.io

What is Bayesian inference?

Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities such as predictions for new observations.

What is Bayesian inference?

Bayesian inference is the process of:

- fitting a probability model to a set of data
- estimating probability distributions on model parameters
- estimating unobserved quantities

The three steps of Bayesian data analysis

1. setting up a **full probability model**: a joint probability distribution for all observable and unobservable quantities in a problem. The model should be consistent with knowledge about the underlying scientific problem and the data collection process.

The three steps of Bayesian data analysis

2. **conditioning on observed data**: calculating and interpreting the appropriate posterior distribution - the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data.

The three steps of Bayesian data analysis

3. **evaluating** the fit of the model and the implications of the resulting posterior distribution: how well does the model fit the data, are the substantive conclusions reasonable.

Bayes rule and some notation

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

where

$$p(y) = \sum_{\theta} p(y|\theta)p(\theta)$$

and, often, use

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

where y are the data and θ parameterises the model.

Bayes rule and the three steps

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

1. setting up a **full model**: $p(y|\theta)$
2. **conditioning on the observed data**:
   - the distribution of model parameters: $p(\theta|y)$
   - unobserved value $\tilde{y}$: $p(\tilde{y}|y) = \sum_\theta p(\tilde{y}|\theta)p(\theta|y)$
3. **evaluating** the fit of the model means checking the predictions from step 2.

Example of Bayesian reasoning - dodgy character

One night in a bar in Las Vegas you meet a dodgy character who tells you that there are two types of slot machine in Shannon's Palace, one that pays out 10% of the time, the other 20%. One sort of machine is green, the other red. *Unfortunately the dodgy character is too drunk to remember which is which.*

I stole this problem from `courses.smp.uq.edu.au/MATH3104/`

Example of Bayesian reasoning - an experiment

The next day you flip a coin and select red to try, you find a red machine and put in a coin. You lose. *Assuming the dodgy character was telling the truth*, what is the chance the red machine is the good one. If you had won instead of losing, what would the chance be?

Example of Bayesian reasoning - the model

The **model** is: "Assuming the dodgy character was telling the truth"

$$\begin{aligned} p(L|R) &= 0.8 \\ p(L|!R) &= 0.9 \end{aligned}$$

where R is the event of the good machine being the red one and L is losing.

Example of Bayesian reasoning - the reasoning

$$p(R|L) = \frac{p(L|R)p(R)}{p(L)}$$

where $p(R)$ represents your knowledge before the experiment; you haven't a clue so $p(R) = 0.5$.

Example of Bayesian reasoning - the answer

$$p(R|L) = \frac{p(L|R)p(R)}{p(L)}$$

with

$$p(L) = p(L|R)p(R) + p(L|!R)p(!R) = 0.85$$

and so we know everything we need to know and

$$p(R|L) \approx 0.47$$

Example of Bayesian reasoning - around again

Say you nonetheless tried red again and lost again

$$p(R|L) = \frac{p(L|R)p(R)}{p(L)}$$

with $p(R) = 0.47$ and $p(!R) = 0.53$ we'd get

$$p(R|L) \approx 0.44$$

In fact

$$\begin{aligned} p(R|LL) &\approx 0.44 \\ p(R|LW) = p(R|WL) &\approx 0.64 \\ p(R|WW) &= 0.8 \end{aligned}$$

and, by the way, $p(R|W) = 0.66$; see `shannons_palace.jl`.

Example of Bayesian reasoning - decision making

To go beyond this we'd probably like to modify our behaviour according to the result: we would probably want to know more about the estimated probabilities, like, how do we estimate the quality of the estimates, surely that's getting better as we do repeated experiments? What about the model itsself, when do we decide the information from the dodgy character was false? That's what we'll learn about, of course, but first, the traditional diversion in thinking about what probabilities are!

Probabilities - what do they mean?

A fair coin!

We would say $p(\text{harps}) = 0.5$.

Probabilities - what do they mean?

An unknown unfair coin!

or

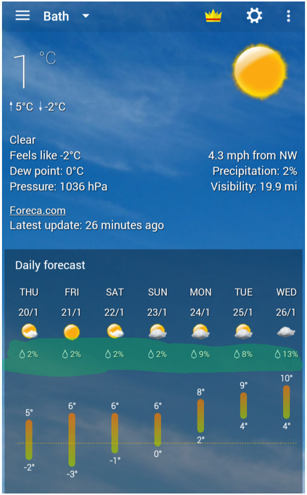
What if we were told that the coin was not fair but either both sides are harps or both sides are heads. Would we then say $p(\text{harps}) = 0.5$; perhaps, but in this case our response to an experiment would be very different.

Probabilities - what do they mean?

The usual rain example:

Probabilities - what do they mean?

The usual rain example:



Probabilities - what are they?

Mathematically

> *Probabilities are numerical quantities, defined on a set of 'outcomes', that are nonnegative, additive over mutually exclusive outcomes, and sum to one over all possible mutually exclusive outcomes.*

$$\text{outcome} \mapsto p(\text{outcome})$$

with some sensible properties.

Probabilities - what are they good for?

In **Bayesian statistics**, *probability is used as the fundamental measure or yardstick of uncertainty.*

Probabilities - what are they good for?

In **Bayesian statistics***, probability is used as the fundamental measure or yardstick of uncertainty.*

Within this paradigm, it is equally legitimate to discuss the probability of 'rain tomorrow' or of a Brazilian victory in the soccer World Cup as it is to discuss the probability that a coin toss will land heads.

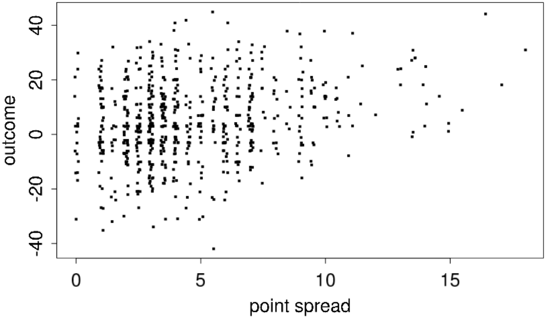
Probabilities - what are they good for?

Bayesian methods *enable statements to be made about the partial knowledge available (based on data) concerning some situation or 'state of nature' - unobservable or as yet unobserved - in a systematic way, using probability as the yardstick. The guiding principle is that the state of knowledge about anything unknown is described by a probability distribution.*

Football example

A *point-spread* is a prediction from a pundit or betting firm of a score difference for a football game, the idea being the result is equally likely to be higher or lower than the point spread. In betting the vig is provided by the point spread with payout coming for higher or lower.

Football example 1981 / 1983 / 1984



The horizontal axis is the point spread, vertical axis is actual score for favourite minus underdog; some jitter has been added.

figure from the book

Football example - what do we want to know

Examples of questions we'd like to ask are

- probability of favourite wins Pr(favourite wins)
- Pr(favorite wins | PS is 3.5)
- Pr(favorite wins by more than the PS)
- Pr(favorite wins by more than the PS | PS is 3.5)

and so on; PS is point-spread.

Football example - counting things

The ‘intuitive’ approach is to count outcomes

- Pr(favourite wins)=410.5/655=0.63
- Pr(favorite wins | PS is 3.5)=46/59=0.61
- Pr(favorite wins by more than the PS)=308/655=0.47
- Pr(favorite wins by more than the PS | PS is 3.5)=32/59=0.54

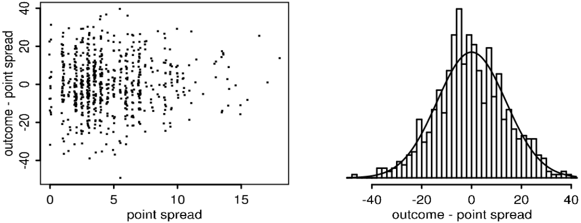
where the zero point-spread games are ignored and a draw is a 0.5 win. This all looks sensible and useful.

Football example - intuitive approach stops working

- Pr(favourite wins|PS is 8.5)=5/5=1
- Pr(favourite wins|PS is 9)=13/20=0.65

This is obviously wrong or mysterious, we could just give up and say the high PS region is undersampled, however, we could make a model, or more correctly, a different model.

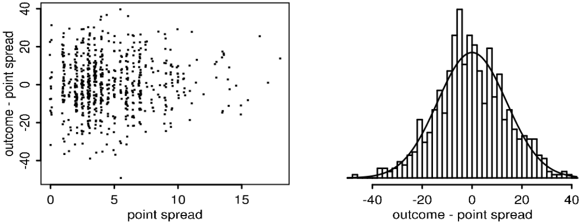
Football example 1981 / 1983 / 1984



The black line on the right is $\text{Normal}(0, 14^2)$.

figure from the book

Football example - model



$$\text{score} = \text{PS} + \xi$$

where

$$\xi \sim \text{Normal}(0, 14^2)$$

and the basic assumption is that the score minus PS is independent of PS.

figure from the book

Football example - intuitive approach stops working

- Pr(favourite wins|PS is 3.5)=0.60
- Pr(favourite wins|PS is 8.5)=0.73
- Pr(favourite wins|PS is 9)=0.74

which is sensible and maybe even useful, but ignores the **errors** in the model: score minus PS is probably neither precisely normal nor precisely independent of PS and it **erases** the potential discovery of something special about games with PS of 8.5.