Statistics and Machine Learning 1

Lecture 7C: Bayesian Inference

Mark Muldoon
Department of Mathematics, Alan Turing Building
University of Manchester

Week 7

Thinking probabilistically II

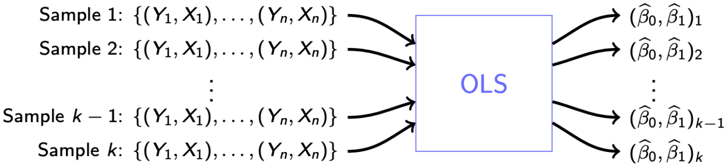
The lectures on regression began with the principle that:

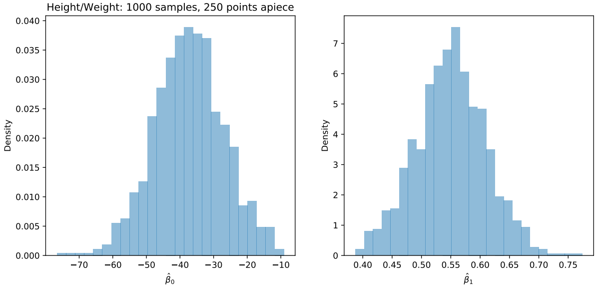
> *Statistics is the science of changing your mind under uncertainty.*

In the next few slides, I'll introduce the *Bayesian* approach to estimating the parameters of a statistical model. That is, we'll imagine some parametric model and we'll develop a systematic way to update/improve our estimates of the parameters in light of data.

Parameter estimates are already probabilistic

The paramaters of a linear regression model depend on the data, which are only a sample from the joint distribution over (X, Y):





Bayes Theorem for parameter estimation

If we use the symbol D to indicate the data and θ to indicate the parameters then $P(D \mid \theta)$ is the *likelihood*: in an earlier lecture Arek introduced the idea of maximising this as ways to estimate θ. Today, we're going to develop a different approach based on Bayes' Theorem.

Bayes' Theorem tells us

$$P(D \mid \theta) P(\theta) = P(\theta \mid D) P(D)$$

or, equivalently,

$$P(\theta \mid D) = \frac{P(D \mid \theta) P(\theta)}{P(D)}.$$

Bayesian ingredients

The key players are:

$P(D \mid \theta)$ the likelihood

$P(\theta)$ the *prior on the parameter*. It's a probability distribution over possible values of the parameter and should encapsulate anything we already know, or even believe, about θ.

$P(D)$ the *prior over the data*. This is a value from the marginal distribution obtained by integrating θ out of $P(D, \theta)$, so we can, at least in principle, compute it using

$$P(D) = \int P(D \mid \theta) P(\theta) \, d\theta.$$

$P(\theta \mid D)$ the *posterior distribution over* θ. It's essentially an updated version of the prior $P(\theta)$, informed by the observations D.

An example: opinion polling

Imagine that we've chosen N voters at random and asked them how they plan to vote in an upcoming election between Asha and Bob. Our aim is to estimate p, the proportion of voters who support Asha.

The data produced by our poll will be that some number k of the N voters in our sample support Asha. If we have chosen the N voters in our sample in an unbiased way, then k should follow a Binomial distribution, so the likelihood is:

$$P(k \mid p) = \frac{N!}{k!\,(N-k)!}\, p^k (1-p)^{N-k}.$$

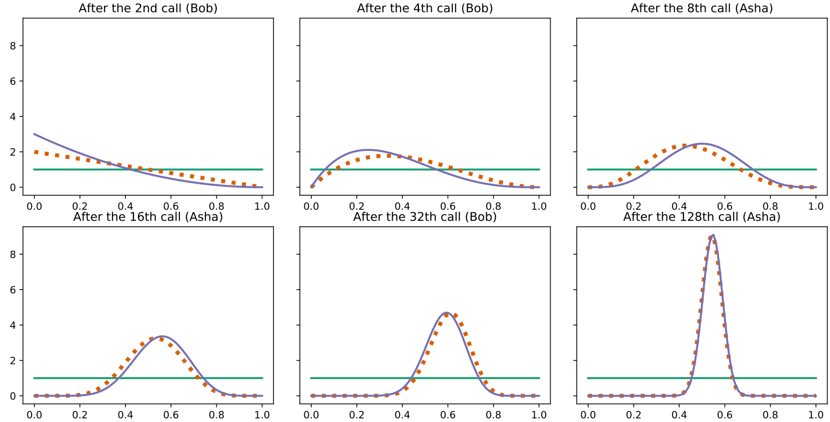
Binomial Bayes: the prior on p

The prior on p should be a probability distribution over p, which is a continuous random variable lying in the interval $0 \leq p \leq 1$, so we want densities $f(p)$.

If we're completely ignorant about the race between Asha and Bob, we might choose an *uninformative* or *flat* prior, $f(p) = 1$. This expresses our complete lack of prior knowledge by assigning equal likelihood to every possible value of p.

For reasons that will become clear in the next video, it's helpful to regard the flat prior as a special case of the Beta distribution.

Binomial Bayes: simulating Asha's poll



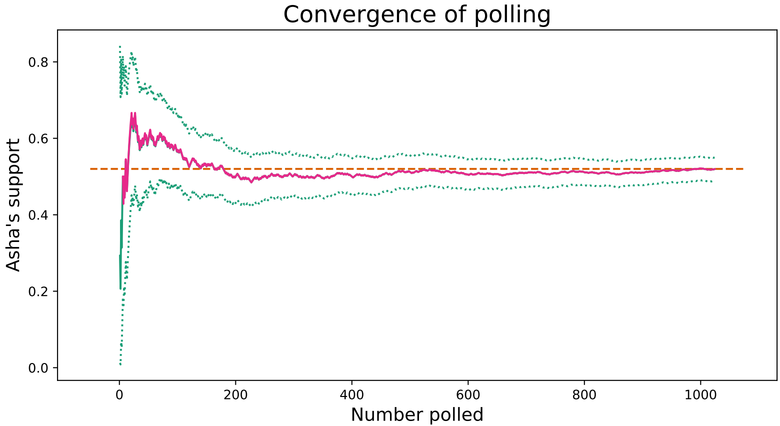
Above: the prior (green), the most recent posterior (purple) and the next-to-most recent (orange, dotted). The titles give the number of responses and the most recent voter's response.

What's the posterior good for?

The posterior is a probability density and so one can . . .

- Get an idea of how well-determined the parameter is: if the posterior is broad and flat, or multi-modal, then we probably need more data or a better experiment.
- Use it to make a point estimate for the parameter. The value of θ at which the posterior takes its maximal value is the *maximum a posteriori (MAP)* estimate of θ.
- Compute *credible regions* around the MAP estimate: they're the Bayesian analog of confidence regions.
- Make predictions by integrating over the posterior.

Binomial Bayes: credible regions



Above: the median (green, solid), 2.5% and 97.5% quantiles (green, dotted), MAP estimate (purple) the true level of Asha's support (orange, dashed). The curve for the MAP estimate is nearly indistinguishable from that for the median.

Further reading

- The polling example discussed in this lecture and the next was inspired by material in Chapter 3 of:

  S. Rogers and M. Girolami (2017), *A First Course in Machine Learning*, 2nd edition, Chapman & Hall/CRC. ISBN: 978-1-4987-3848-4.

  Available online through the University Library.
- The ecologist Richard McElreath has a very helpful set of video lectures that accompany his book

  R. McElreath (2020), *Statistical Rethinking*, 2nd edition, Chapman & Hall/CRC. ISBN: 9780367139919.

  The videos, code examples (R, Python) and the first two chapters of the book are available by following links from the book's home page.
- For a comprehensive overview of Bayesian stats, I recommend:

  A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. Rubin (2014), *Bayesian Data Analysis*, 3rd edition, Chapman & Hall/CRC. ISBN: 978-1-4398-4095-5.

  Available online.