Lecture 5

Bayesian data analysis

Programme

- Bayes' Rule

- Simple "real-world" Bayesian problems: a demonstration of the rule's rationality

- Bayesian data analysis and its rationality

- Bayesian ANOVA example: Described in detail in readings from Kruschke textbook

- Demonstration: A Bayesian approach to our most recent analyses

Bayes' Rule

- Derivable from the laws of probability
- Premised on the idea that probability is a degree of belief inside a learner's head

The likelihood: Our degree of belief in seeing the data if *h* is true (given our assumptions about the generating mechanism behind all possible datasets – i.e., the sample space)

The prior: Our degree of belief in *h* before seeing the data

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

The probability that hypothesis *h* is true, given that we (the learners) have observed data *d*; i.e., our degree of belief in *h* after seeing the data

The marginal likelihood: The likelihood of the data under hypothesis *h* and all other possible hypotheses in the hypothesis space *H*; can also be expressed as:

$$\sum_{h \in \mathcal{H}} P(d|h)P(h)$$

or

$$\sum_{h' \in \mathcal{H}} P(d|h')P(h')$$

Bayesian inference: rational in the real world

- The taxicab problem illustrates that, in the real world, the **prior** matters!

- The game show problem illustrates that, in the real world, the **likelihood** matters!

- (The following slides describing the problems and their solutions are from Daniel Navarro's 2011 lectures on Computational Cognitive Science at *The University of Adelaide*. This is the link to the full course, and this is the link to the specific lecture used – Lecture 4.)

The taxicab problem

The taxi cab problem

- The world's simplest Bayesian problem:
  - Most (80%) of the taxis in Simpletown are green, with the rest (20%) being yellow. In a traffic accident involving a hit-and-run taxi, a witness claims the taxi was yellow. Careful testing shows that the witness can successfully identify the colour of a taxi only 75% of the time due to bad eyesight. On the balance of probabilities, should we hold Yellow Taxi Company liable?

The game show problem

- A slightly subtler inference problem:
  - You're a contestant on a game show, and there is a prize behind one of three doors, labelled A, B and C. The host asks you to pick a door, and you choose A. He then opens up door B, and shows you that there is no prize there. Finally, he asks if you would like to switch to door C, or if you want to stay with door A. What should you do? Does it matter?

Argued to also be more rational in data analysis

Reading: Kruschke Ch 11 (not provided)

- In the Bayesian version of the analyses we've been doing, the linear model and its assumptions are expressed in the **likelihood** function, which takes the form of a normal distribution (or a *t*-distribution in some cases).

- For each predictor, there is a **prior** distribution (usually of gamma form), expressing the learner's beliefs about the strength of the relationship between the predictor and the outcome variable.

- The **marginal likelihood** is usually not computed.

All four components of Bayes' Rule are probability distributions; that is:
- They have a certain shape, expressed in a function (i.e., formula).
- For probability density functions (continuous variables), the area under the shape (i.e., the integral of the function) equals 1. For probability mass functions (categorical variables) the sum of the categories' probabilities is 1.

Bayesian data analysis logic continued

- The marginal likelihood is the conjoint (summed) probability of the data points given the model. The reason we usually need specialised techniques to estimate this value is that it takes the form of a complex integral. For example, for a model with three parameters:

$$P(\theta_1, \theta_2, \phi | \mathbf{x}) = \frac{P(\mathbf{x}|\theta_1, \theta_2, \phi)P(\theta_1, \theta_2, \phi)}{\int_0^1 \int_0^1 \int_0^1 P(\mathbf{x}|\theta_1, \theta_2, \phi)P(\theta_1, \theta_2, \phi) \; d\theta_1 \; d\theta_2 \; d\phi}$$

- It follows from Bayes' Rule that the posterior is a weighted average of the prior mean and the data, with weighting corresponding to the relative precisions of the prior and likelihood. Precisions are related to standard deviations and reflect the learner's (i.e., the analyst's) uncertainty in prior beliefs, or, alternatively, model predictions. The precisions are estimated from the data through the use of hyperpriors. That is,

  - if the prior is fairly vague, and the data are numerous, the posterior will be near the parameter values that maximise the likelihood of the data.

  - "natural shrinkage" occurs, similar to the shrinkage in multilevel modelling.

Bayesian data analysis : Typical steps

1. Determine priors, likelihood, and hyperpriors (or get *R* to do this for you)
2. Estimate the posterior (and possibly the marginal likelihood) through a Markov Chain Monte Carlo process (Metropolis-Hastings, Gibbs' sampling)
   - Since there is a prior for each predictor, there is a posterior for each parameter.
3. Determine whether the parameter value predicted by the null hypothesis (e.g., a slope value of zero) falls within the 95% Highest Density Interval (HDI) of the posterior for the parameter. It does in the case below!

Regression slope value
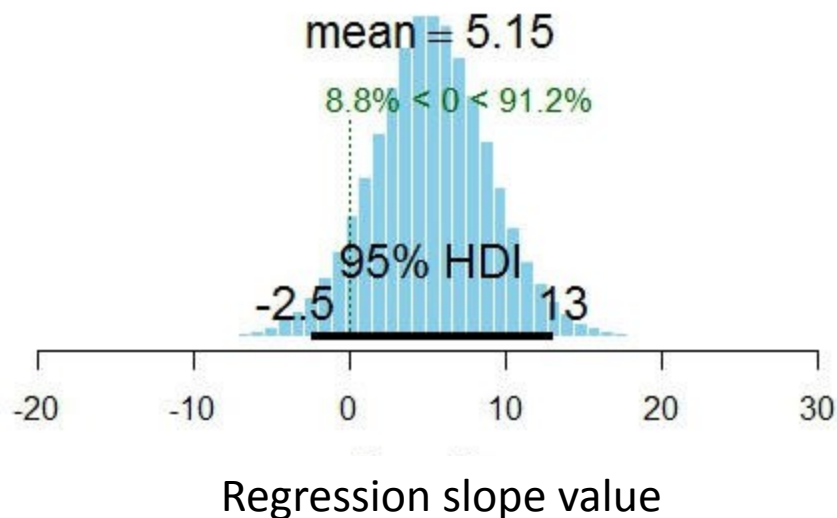
Bayesian data analysis : Advantages over the standard "Frequentist" approach

1. The use of priors allows us to incorporate findings from previous studies into the data analysis. This is particularly useful if there is a large number of preceding studies while the current study's data set is small.

2. HDI more interpretable than frequentist "confidence intervals": If a parameter value falls within the 95% HDI, it is among the most believable parameter values. Frequentist confidence intervals indicate that it is 95% likely that the parameter value could not be in the interval range if the null hypothesis were true.

3. Subtle but philosophically important point discussed at length by Kruschke: The Bayesian analyst's model assumptions are transparent and not dependent on sample size (degrees of freedom).

Bayesian ANOVA example

Likelihood function	Normal
Prior over slope parameters	Normal
(Hyper)prior over prior precisions	Folded-t
Prior over error (likelihood function precision)	Uniform

A Bayesian approach to our most recent analyses

Reading: Gelman et al 2009

Demonstration – `bayesglm` function in `arm` package. For generalized linear modelling using the Bayesian approach.

Likelihood function	Normal
Prior over slope parameters	Cauchy, *t*
(Hyper)prior over prior precisions	- (standardization of predictors and fixing of SD, which is related to the precision, to 0.5 or 1)
Prior over error (likelihood function precision)	Uniform

For mixture models (discussed in Lecture 4), you could explore the package `MCMCglmm` and the function of the same name within it.

Reading

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. (2009). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics, 2*, 1360-1383. Available online: http://www.stat.columbia.edu/~gelman/research/published/priors11.pdf

Kruschke, J. K. (2011). *Doing Bayesian Data Analysis*. Elsevier: Oxford. Chapter 4 "Bayes' Rule", Chapter 18 "Bayesian Oneway ANOVA", and Chapter 19 "Metric predicted variable with multiple nominal predictors" will be provided online