Introduction to Bayesian inference

What are Bayesian methods?

- Bayesian methods are data analysis tools derived from the principles of Bayesian inference and provide
  - parameter estimates with good statistical properties;
  - parsimonious descriptions of observed data;
  - predictions for missing data and forecasts of future data; and
  - a computational framework for model estimation, selection, and validation.

Building blocks of Bayesian inference

- Generally (unless otherwise stated), in this course, we will use the following notation. Let
  - $\mathcal{Y}$ be the sample space;
  - y be the observed data;
  - Θ be the parameter space; and
  - θ be the parameter of interest.
- More to come later.

Bayes' theorem - basic conditional probability

- Let's take a step back and quickly review the basic form of Bayes' theorem.
- Suppose there are some events A and B having probabilities $\Pr(A)$ and $\Pr(B)$.
- Bayes' rule gives the relationship between the marginal probabilities of A and B and the conditional probabilities.
- In particular, the basic form of Bayes' rule or Bayes' theorem is

$$\Pr(A|B) = \frac{\Pr(A \text{ and } B)}{\Pr(B)} = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

$\Pr(A)$ = marginal probability of event A, $\Pr(B|A)$ = conditional probability of event B given event A, and so on.

Building blocks of Bayesian inference

- Now, to a slightly more complicated version of Bayes' rule. First,

  1. For each $\theta \in \Theta$, specify a prior distribution $p(\theta)$ or $\pi(\theta)$, describing our beliefs about θ being the true population parameter.

  2. For each $\theta \in \Theta$ and $y \in \mathcal{Y}$, specify a sampling distribution $p(y|\theta)$, describing our belief that the data we see y is the outcome of a study with true parameter θ. $p(y|\theta)$ gets us the likelihood $L(y;\theta)$

  3. After observing the data y, for each $\theta \in \Theta$, update the prior distribution to a posterior distribution $p(\theta|y)$, describing our "updated" belief about θ being the true population parameter.

- Now, how do we get from Step 1 to 3? Bayes' rule!

$$p(\theta|y) = \frac{p(\theta)L(y;\theta)}{\int_\Theta p(\tilde{\theta})L(y;\tilde{\theta})\mathrm{d}\tilde{\theta}} = \frac{p(\theta)L(y;\theta)}{L(y)}$$

We will use this over and over throughout the course!

Notes on prior distributions

Many types of priors may be of interest. These may

- represent our own beliefs;
- represent beliefs of a variety of people with differing prior opinions; or
- assign probability more or less evenly over a large region of the parameter space.
- and so on...

Notes on prior distributions

- Subjective Bayes: a prior should accurately quantify some individual's beliefs about θ.
- Objective Bayes: the prior should be chosen to produce a procedure with "good" operating characteristics without including subjective prior knowledge.
- Weakly informative: prior centered in a plausible region but not overly-informative, as there is a tendency to be over confident about one's beliefs.

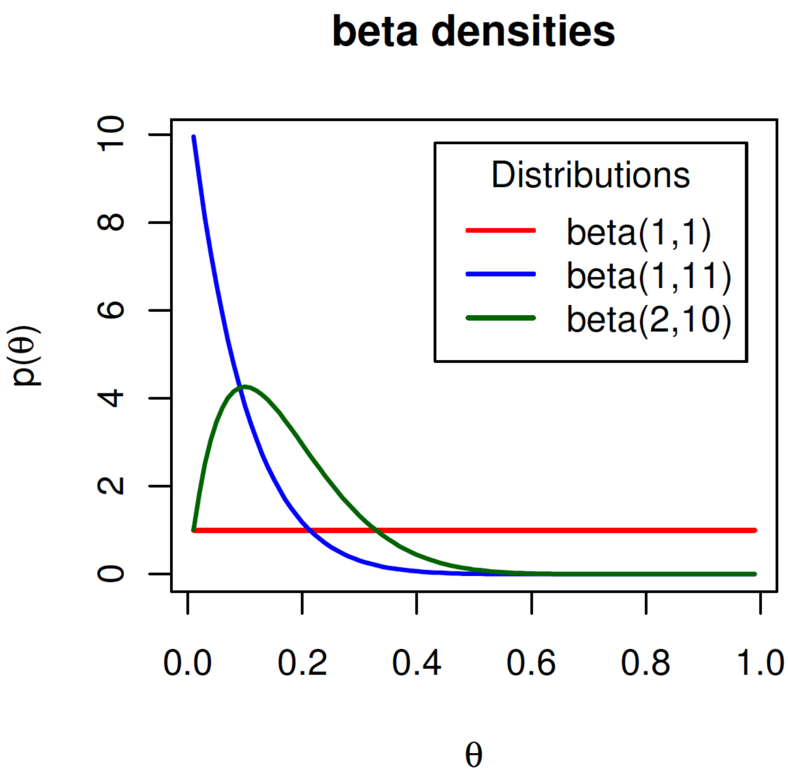
Notes on prior distributions

- The prior quantifies your initial uncertainty in θ before you observe new data (new information) - this may be necessarily subjective & summarize experience in a field or prior research.
- Even if the prior is not "perfect", placing higher probability in a ballpark of the truth leads to better performance.
- Hence, it is very seldom the case that a weakly informative prior is not preferred over no prior.
- One (very important) role of the prior is to stabilize estimates in the presence of limited data.

Simple example - estimating a population proportion

- Suppose $\theta \in (0, 1)$ is the population proportion of individuals with diabetes in the US.
- A prior distribution for θ would correspond to some distribution that distributes probability across $(0, 1)$.
- A very precise prior corresponding to abundant prior knowledge would be concentrated tightly in a small sub-interval of $(0, 1)$.
- A vague prior may be distributed widely across $(0, 1)$ - e.g., a uniform distribution would be the common choice here.

Some possible prior densities



Beta prior densities

- These three priors correspond to beta(1,1) [also, Unif(0,1)], beta(1,11) and beta(2,10) densities.
- beta(a,b) is a probability density function (pdf) on (0,1),

$$\pi(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1},$$

  where $B(a,b)$ = beta function = normalizing constant ensuring the kernel integrates to one. Note: some texts write $\text{beta}(\alpha,\beta)$ instead.

- The beta(a,b) distribution has expectation $\mathbb{E} = a/(a+b)$ and the density becomes more and more concentrated as $a+b$ = prior "sample size" increases.
- We will look more carefully into the beta-binomial model next week but for now, I'll illustrate how this prior gets updated as data becomes available.