EPSE 581C: Bayesian Methods

Ed Kroc

University of British Columbia

ed.kroc@ubc.ca

September 30, 2019

Last time

- More on joint, conditional, and marginal random variables
- Single parameter models

Today

- Summarizing the posterior information
- Multiparameter models

Bayes' Theorem

Bayes' Theorem

Let X be a continuous random variable; i.e. can take on any real number. Let Y be any random variable (discrete or continuous). Then the distribution of X given Y is determined by the conditional PDF:

$$f_{X|Y=y}(x \mid Y = y) = \frac{f_{Y|X=x}(y \mid X = x) f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X=t}(y \mid X = t) f_X(t)\ dt}$$

or, more simply:

$$f_{X|Y}(x \mid y) = \frac{f_{Y|X}(y \mid x) f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y \mid t) f_X(t)\ dt}$$

Estimation and inference

We now have all the machinery we need to perform estimation and inference from a Bayesian perspective. Let's start to see how this works!

- Bernoulli phenomenon of interest: $Y \sim Ber(\theta_Y)$. E.g. Y could denote the presence or absence of a disease in our study population, and so θ_Y denotes the *incidence* of the disease in our study population.
- Naturally, we might like to *estimate* θ_Y by using some sample data from the study population.
- Classical approach:
  (1) Collect sample of size n: $Y_1, Y_2, \ldots, Y_n$
  (2) Calculate sample mean/proportion (same as ML estimator): $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
  (3) Estimate standard error to quantify uncertainty in sample mean estimate: $\widehat{SE}(\widehat{\theta}) = \frac{s}{\sqrt{n}}$, where s is sample standard deviation

Estimation and inference

- Classical approach:
  (1) Collect sample of size n: $Y_1, Y_2, \ldots, Y_n$
  (2) Calculate sample mean/proportion (same as ML estimator): $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
  (3) Estimate standard error to quantify uncertainty in sample mean estimate: $\widehat{SE}(\widehat{\theta}) = \frac{s}{\sqrt{n}}$, where s is sample standard deviation

- Recall what this means:
  - Our best estimate for θ_Y is the sample mean/proportion: $\widehat{\theta}$, and
  - if we were to resample and recalculate the sample mean many, many times, then about 95% of the resulting 95% confidence intervals would contain the *true* population proportion θ_Y.
  - In particular, the CI for *our single sample* may or may not contain the true proportion θ_Y.

Estimation and inference

- Bayesian approach:
  (1) Collect sample of size n: $Y_1, Y_2, \ldots, Y_n$.
  (2) Compute the *likelihood* of observing these data.
  (3) Choose a *prior* distribution for our parameter of interest: incidence proportion.
  (4) Compute the *posterior* distribution for the incidence proportion.
  (5) Summarize the posterior distribution with a "best" estimate and associated uncertainty.

Computing the likelihood

Sample: $Y_1, \ldots, Y_n \sim \text{Ber}(\theta_Y)$

(2) Compute the *likelihood* of observing these data: $\Pr(Y_1, \ldots, Y_n \mid \theta)$

- For each sample point (randomly sampled), we have $\Pr(Y_i = 0 \mid \theta) = 1 - \theta$ and $\Pr(Y_i = 1 \mid \theta) = \theta$.
- In general, we can write

$$\Pr(Y_i = y_i \mid \theta) = \theta^{y_i}(1-\theta)^{1-y_i}$$

- Since we have a *random sample*, all data are observed *independently*; thus,

$$\begin{aligned}\Pr(Y_1 = y_1, \ldots, Y_n = y_n \mid \theta) &= \prod_{i=1}^{n} \Pr(Y_i = y_i \mid \theta) \\ &= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}\end{aligned}$$

Computing the likelihood

Notice that this likelihood can be rewritten as follows:

$$\begin{aligned}\Pr(Y_1 = y_1, \ldots, Y_n = y_n \mid \theta) &= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^{\sum_{i=1}^{n} y_i}(1-\theta)^{n-\sum_{i=1}^{n} y_i} \\ &= \theta^{n\bar{Y}}(1-\theta)^{n-n\bar{Y}}\end{aligned}$$

- Classical maximum likelihood estimation would say that our best estimate of θ_Y is the value of θ that *maximizes* this likelihood function.
- Using calculus (taking a derivative), easy to show that this MLE is $\bar{Y}$.
- But for the Bayesian approach, we want to know how reasonable *any* value of θ (*any hypothesis*) is as an estimate for θ_Y.

Choosing a prior

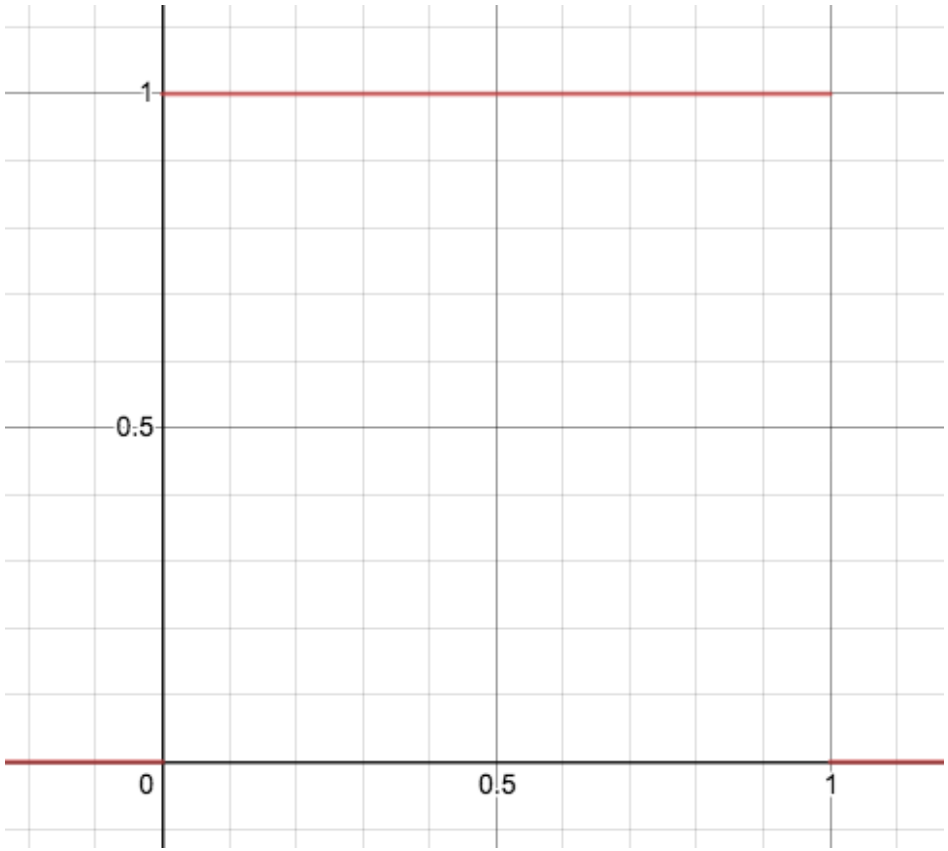
(3) Choose a *prior* distribution for our parameter of interest: incidence proportion, θ.

- Our parameter of interest is a *proportion* in [0,1], so any *prior* distribution should tell us how likely it is that θ_Y lies *anywhere* in this interval.
- Critically, we have many *choices* for the prior.
- Notice: our likelihood quantified a *discrete* (joint) random variable: $Y_1, \ldots, Y_n$, but our prior quantifies a *continuous* random variable: θ.
- Of course, our posterior will also quantify a *continuous* random variable: θ.

Choosing a prior

- Prior 1: $\pi_1(\theta)$ is *uniform* on [0,1].

$$\pi_1(\theta) = \begin{cases} 1, & 0 \leqslant \theta \leqslant 1 \\ 0, & \text{o.w.} \end{cases}$$



Choosing a prior

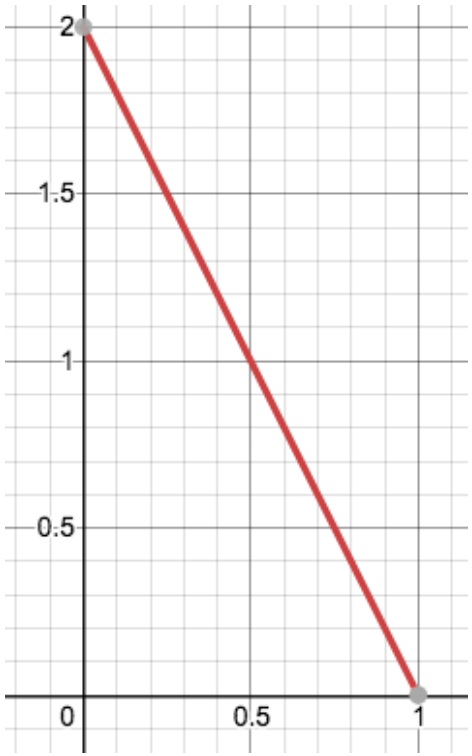
- Prior 2: $\pi_2(\theta)$ is *Beta*(3,3) on [0,1].

$$\pi_2(\theta) = \begin{cases} 30\theta^2(1-\theta)^2, & 0 \leqslant \theta \leqslant 1 \\ 0, & \text{o.w.} \end{cases}$$

Choosing a prior

- Prior 3: $\pi_3(\theta)$ is *skewed* on [0,1].

$$\pi_3(\theta) = \begin{cases} 2(1-\theta), & 0 \leqslant \theta \leqslant 1 \\ 0, & \text{o.w.} \end{cases}$$



Computing the posterior

Posterior 1:

$$f_1(\theta \mid \boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \theta)\pi_1(\theta)}{\int_{-\infty}^{\infty} f(\boldsymbol{y} \mid t)\pi_1(t)\ dt}$$

$$= \frac{\theta^{n\bar{Y}}(1-\theta)^{n-n\bar{Y}}}{\int_0^1 t^{n\bar{Y}}(1-t)^{n-n\bar{Y}}\ dt}, \quad 0 \leqslant \theta \leqslant 1$$

$$= c_1\theta^{n\bar{Y}}(1-\theta)^{n-n\bar{Y}},$$

where c_1 is the normalizing constant (inverse of normalizing factor); just there to make sure the posterior probability distribution will *integrate to 1 over all possible hypotheses/parameters* θ.

Computing the posterior

Posterior 2:

$$f_2(\theta \mid \boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \theta)\pi_2(\theta)}{\int_{-\infty}^{\infty} f(\boldsymbol{y} \mid t)\pi_2(t)\ dt}$$

$$= \frac{\theta^{n\bar{Y}}(1-\theta)^{n-n\bar{Y}} \cdot 30\theta^2(1-\theta)^2}{\int_0^1 t^{n\bar{Y}}(1-t)^{n-n\bar{Y}} \cdot 30t^2(1-t)^2\ dt}, \quad 0 \leqslant \theta \leqslant 1$$

$$= \frac{\theta^{n\bar{Y}+2}(1-\theta)^{n-n\bar{Y}+2}}{\int_0^1 t^{n\bar{Y}+2}(1-t)^{n-n\bar{Y}+2}\ dt}$$

$$= c_2\theta^{n\bar{Y}+2}(1-\theta)^{n-n\bar{Y}+2},$$

where c_2 is the normalizing constant.

Computing the posterior

Posterior 3:

$$f_3(\theta \mid \boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \theta)\pi_3(\theta)}{\int_{-\infty}^{\infty} f(\boldsymbol{y} \mid t)\pi_3(t)\ dt}$$

$$= \frac{\theta^{n\bar{Y}}(1-\theta)^{n-n\bar{Y}} \cdot 2(1-\theta)}{\int_0^1 t^{n\bar{Y}}(1-t)^{n-n\bar{Y}} \cdot 2(1-t)\ dt}, \quad 0 \leqslant \theta \leqslant 1$$

$$= \frac{\theta^{n\bar{Y}}(1-\theta)^{n-n\bar{Y}+1}}{\int_0^1 t^{n\bar{Y}}(1-t)^{n-n\bar{Y}+1}\ dt}$$

$$= c_3\theta^{n\bar{Y}}(1-\theta)^{n-n\bar{Y}+1},$$

where c_3 is the normalizing constant.

Conjugate distributions

- Notice that all three of these posteriors (generated by three different priors) all look quite similar.
- This is no coincidence: in fact, all of the priors we have considered are *Beta distributions*, which always give rise to *Beta distributed* posteriors for Binomial/Bernoulli (0 or 1) data.
- Terminology: we say that Beta distributions are *conjugate priors* for Bernoulli/Binomial likelihoods (data).
- In general, if the posterior distribution is in the same family as the prior distribution, then the prior and posterior are call *conjugate distributions*, and the prior is called a *conjugate prior* for the likelihood.

Beta distributions (random variables)

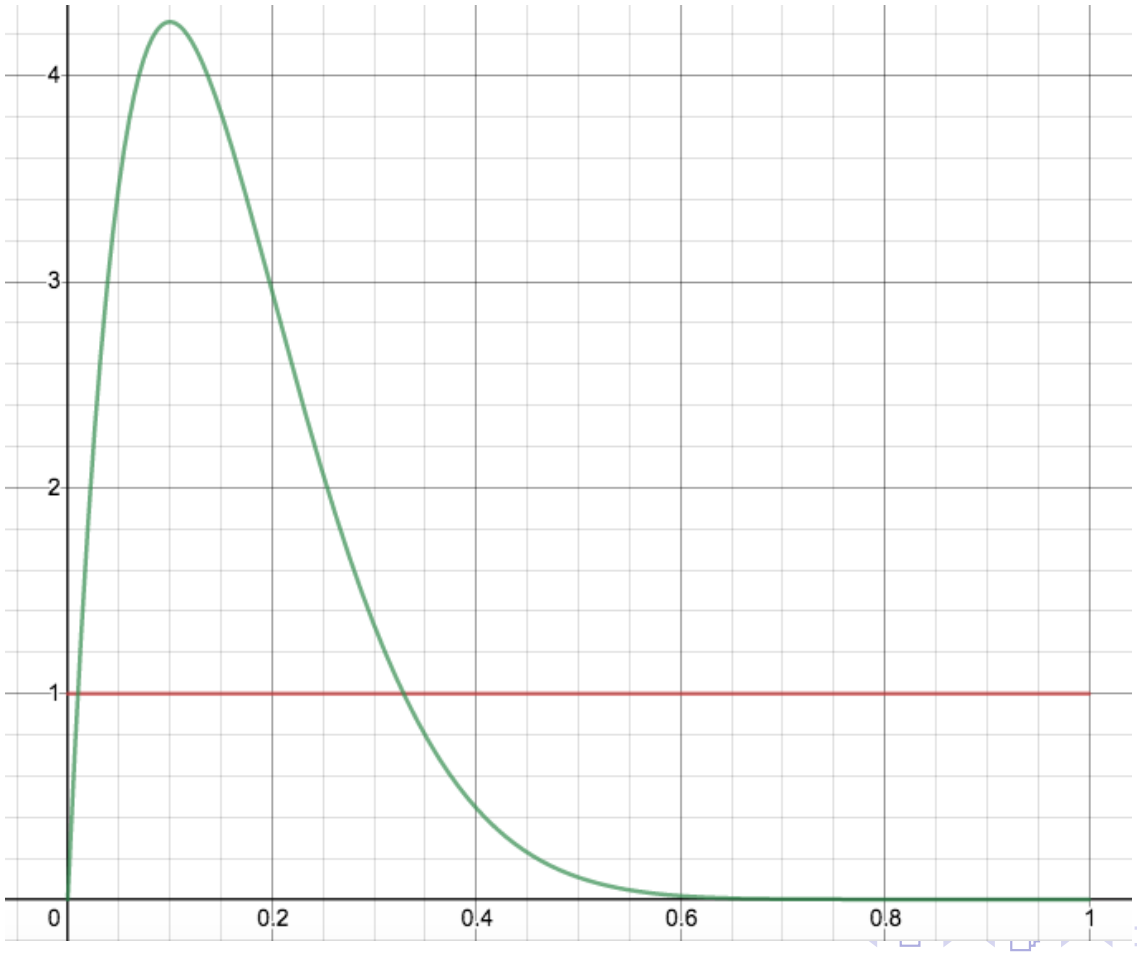
- Beta distributions are great for encoding probabilities on the interval [0,1].
- A *Beta* distribution is specified by two shape parameters: $\alpha, \beta > 0$.
- In general, $X \sim \mathit{Beta}(\alpha, \beta)$ means that the PDF of X is:

$$f_X(x) = c \cdot x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \leqslant x \leqslant 1.$$

- The normalizing constant c can be calculated exactly using calculus.
- Notice that all our priors (and posteriors) so far are Beta distributions for different values of α and β.
- `https://homepage.divms.uiowa.edu/~mbognar/applets/beta.html`

Bayesian estimation and inference

- Suppose $n = 10$ and we observe 1 patient with disease: $n\bar{Y} = 1$
- Prior 1: $\pi_1(\theta) = 1, \quad 0 \leqslant \theta \leqslant 1$
- Posterior 1: $f_1(\theta \mid \boldsymbol{y}) = 110 \cdot \theta(1-\theta)^9, \quad 0 \leqslant \theta \leqslant 1$

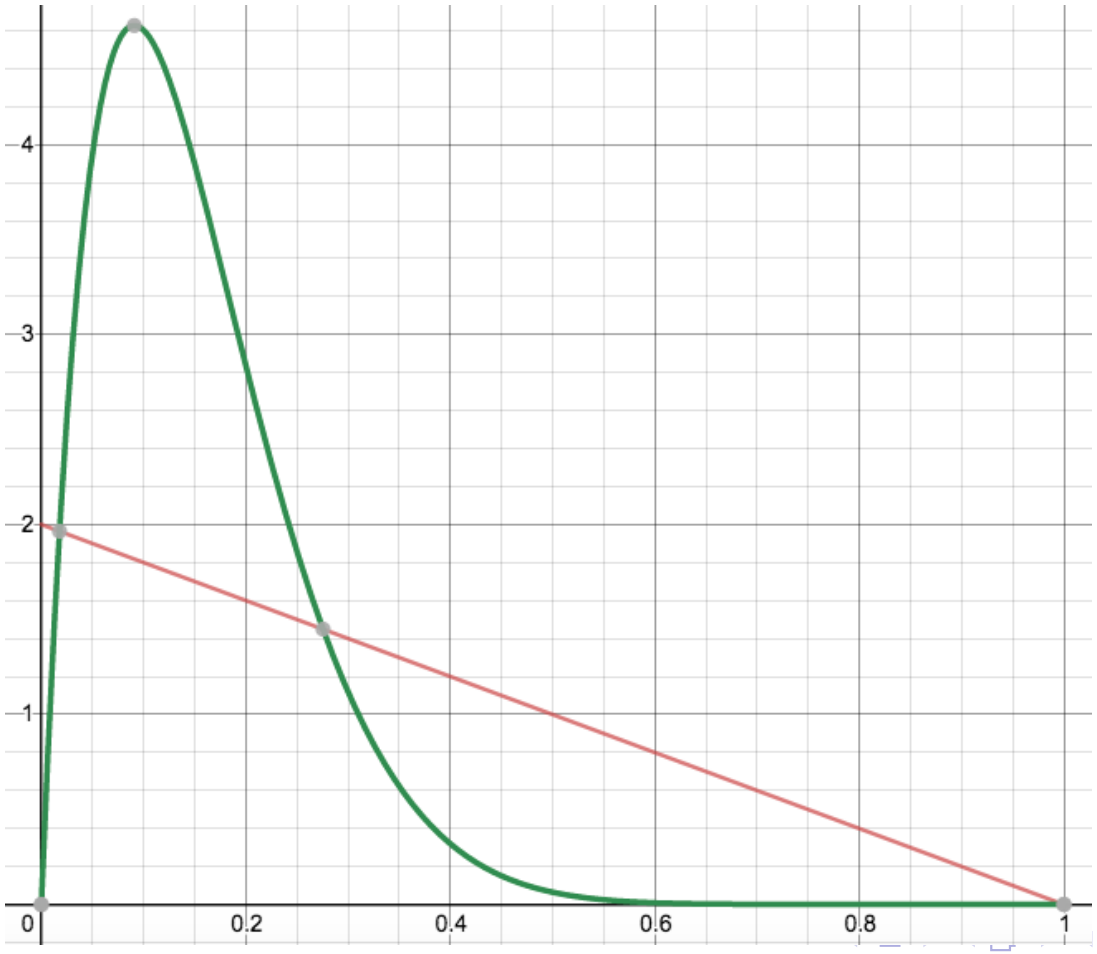


Bayesian estimation and inference

- Suppose $n = 10$ and we observe 1 patient with disease: $n\bar{Y} = 1$
- Prior 2: $\pi_2(\theta) = 30 \cdot \theta^2(1-\theta)^2, \quad 0 \leqslant \theta \leqslant 1$
- Posterior 2: $f_2(\theta \mid \boldsymbol{y}) = 5460 \cdot \theta^3(1-\theta)^{11}, \quad 0 \leqslant \theta \leqslant 1$

Bayesian estimation and inference

- Suppose $n = 10$ and we observe 1 patient with disease: $n\bar{Y} = 1$
- Prior 3: $\pi_3(\theta) = 2(1-\theta), \quad 0 \leqslant \theta \leqslant 1$
- Posterior 3: $f_3(\theta \mid \boldsymbol{y}) = 132 \cdot \theta(1-\theta)^{10}, \quad 0 \leqslant \theta \leqslant 1$

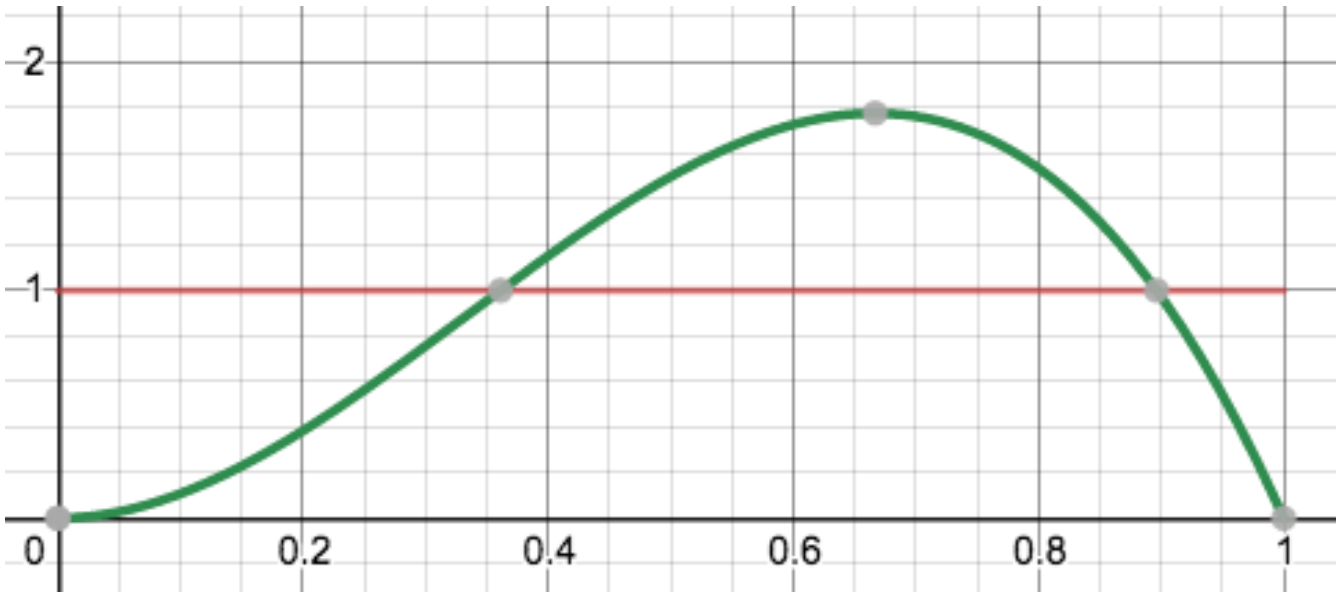


Bayesian estimation and inference

- Notice that all posterior distributions look very similar regardless of the shape of the prior.
- This is a reflection of the fact that the shape of the prior will matter little as long as we have "enough" data.
- Let's consider how the posteriors would look if we had a sample of size 3 only, and then if we had a sample of size 100.

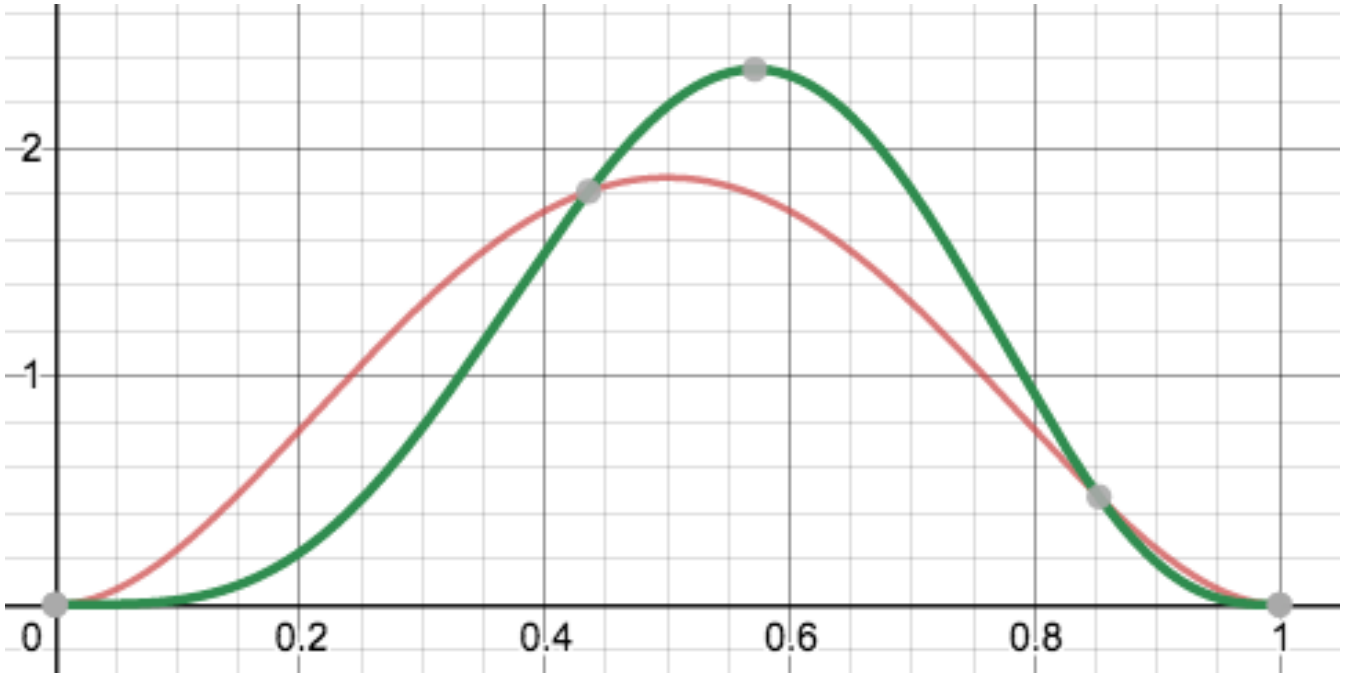
Bayesian estimation and inference

- Suppose $n = 3$ and we observe 2 patients with disease: $n\bar{Y} = 2$
- Prior 1: $\pi_1(\theta) = 1, \quad 0 \leqslant \theta \leqslant 1$
- Posterior 1: $f_1(\theta \mid \boldsymbol{y}) = 12 \cdot \theta^2(1-\theta), \quad 0 \leqslant \theta \leqslant 1$



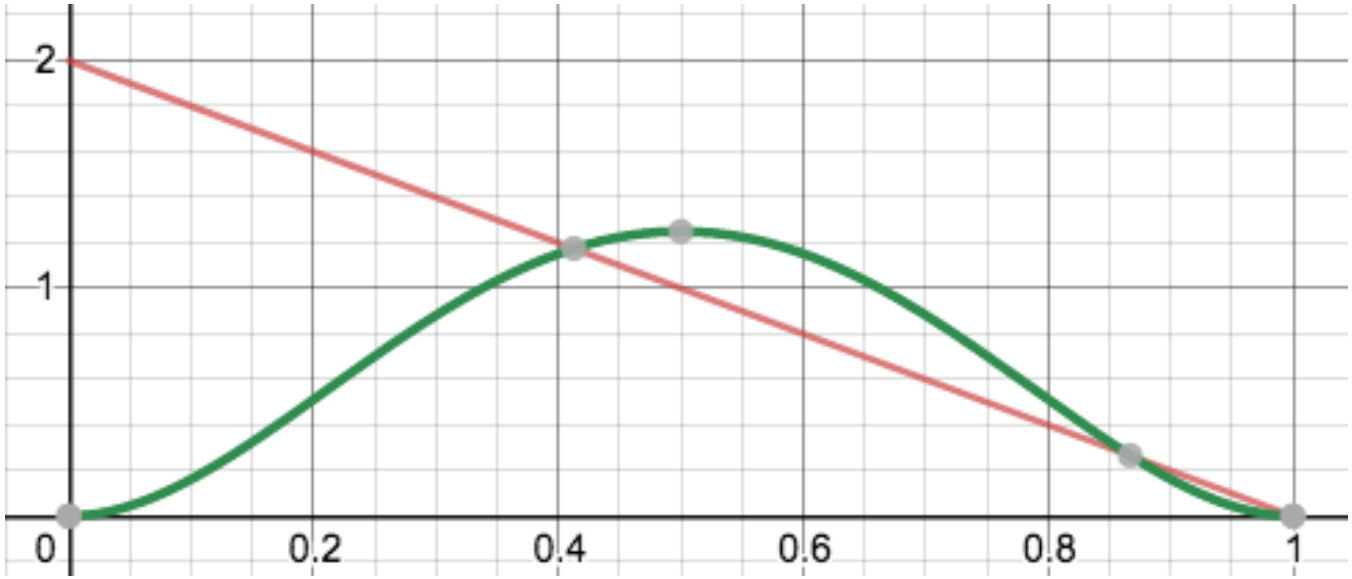
Bayesian estimation and inference

- Suppose $n = 3$ and we observe 2 patients with disease: $n\bar{Y} = 2$
- Prior 2: $\pi_2(\theta) = 30 \cdot \theta^2(1-\theta)^2, \quad 0 \leqslant \theta \leqslant 1$
- Posterior 2: $f_2(\theta \mid \boldsymbol{y}) = 280 \cdot \theta^4(1-\theta)^3, \quad 0 \leqslant \theta \leqslant 1$



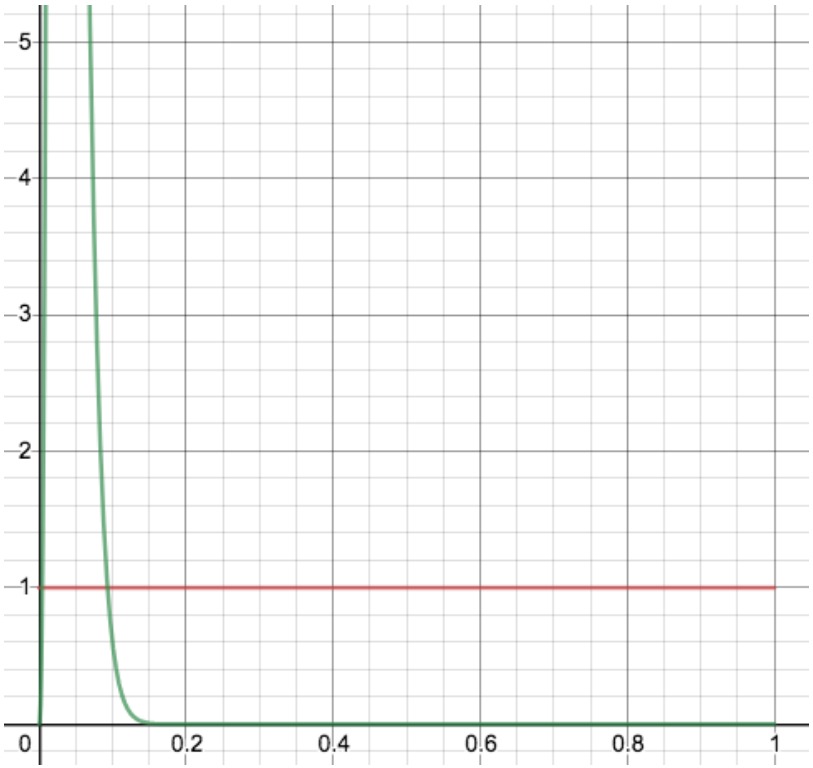
Bayesian estimation and inference

- Suppose $n = 3$ and we observe 2 patients with disease: $n\bar{Y} = 2$
- Prior 3: $\pi_3(\theta) = 2(1-\theta), \quad 0 \leqslant \theta \leqslant 1$
- Posterior 3: $f_3(\theta \mid \boldsymbol{y}) = 20 \cdot \theta^2(1-\theta)^2, \quad 0 \leqslant \theta \leqslant 1$



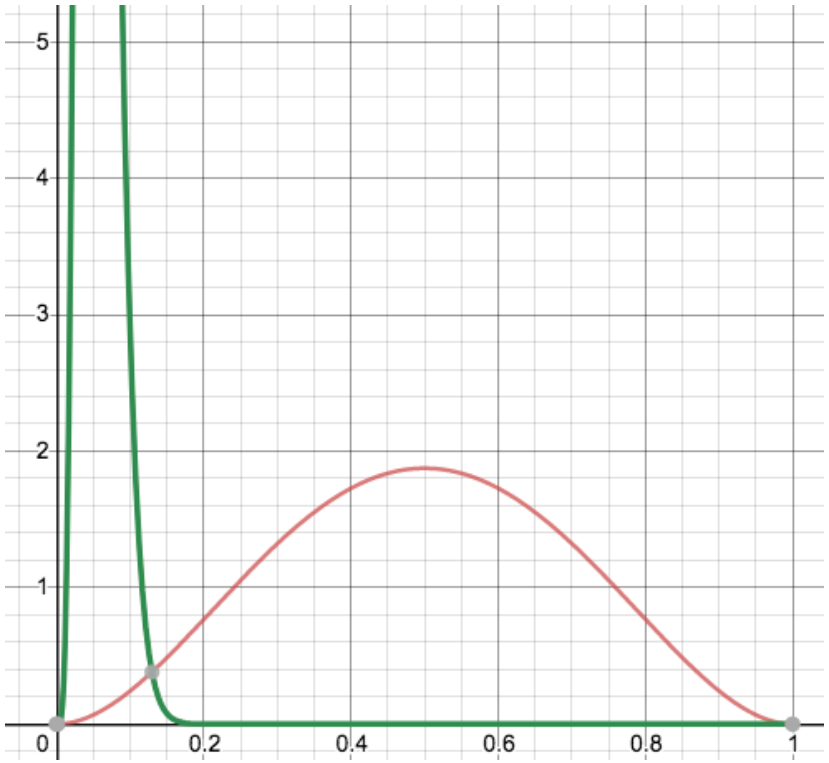
Bayesian estimation and inference

- Suppose $n = 100$ and we observe 3 patients with disease: $n\bar{Y} = 3$
- Prior 1: $\pi_1(\theta) = 1, \quad 0 \leqslant \theta \leqslant 1$
- Posterior 1: $f_1(\theta \mid \boldsymbol{y}) = c_1 \cdot \theta^3(1-\theta)^{97}, \quad 0 \leqslant \theta \leqslant 1$



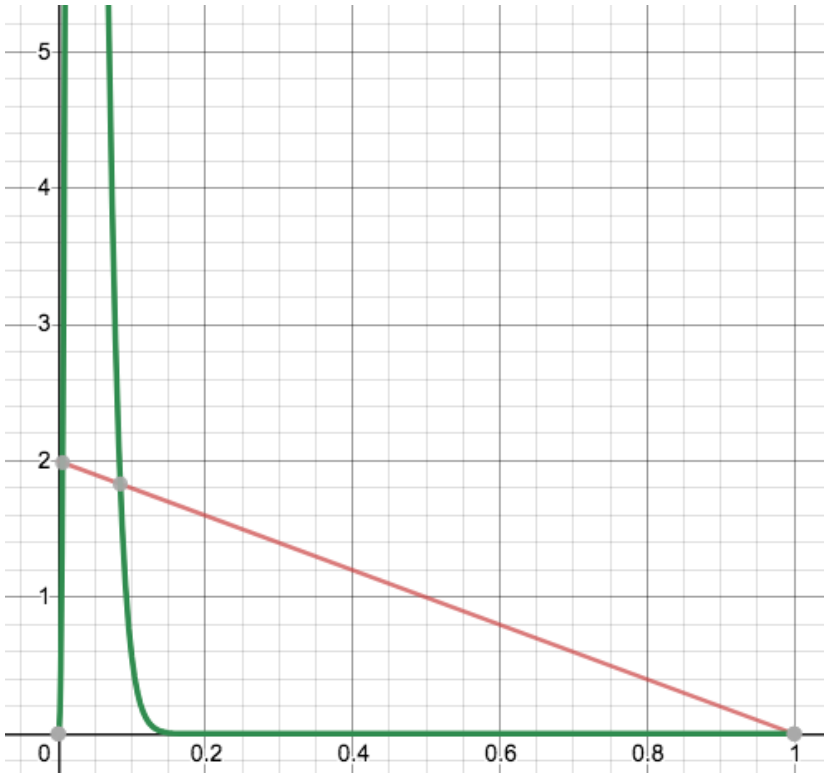
Bayesian estimation and inference

- Suppose $n = 100$ and we observe 3 patients with disease: $n\bar{Y} = 3$
- Prior 2: $\pi_2(\theta) = 30\theta^2(1-\theta)^2, \quad 0 \leqslant \theta \leqslant 1$
- Posterior 2: $f_2(\theta \mid \boldsymbol{y}) = c_2 \cdot \theta^5(1-\theta)^{99}, \quad 0 \leqslant \theta \leqslant 1$



Bayesian estimation and inference

- Suppose $n = 100$ and we observe 3 patients with disease: $n\bar{Y} = 3$
- Prior 3: $\pi_3(\theta) = 2(1-\theta), \quad 0 \leqslant \theta \leqslant 1$
- Posterior 3: $f_3(\theta \mid \boldsymbol{y}) = c_3 \cdot \theta^3(1-\theta)^{98}, \quad 0 \leqslant \theta \leqslant 1$



Bayesian estimation and inference

- While we should *always* examine the entire posterior distribution, it is often convenient to summarize the information it contains:
  - Posterior mean: $\mathbb{E}(\theta \mid \boldsymbol{y})$
  - Posterior mode: θ such that $f(\theta \mid \boldsymbol{y})$ is maximized
  - Posterior standard deviation: $SD(\theta \mid \boldsymbol{y})$
  - Posterior IQR: 25th and 75th percentile of $f(\theta \mid \boldsymbol{y})$
  - Credible/credibility intervals:
    - $\alpha\%$ mean-centred interval: interval containing $\alpha\%$ of the posterior density with the mean at the centre of the interval.
    - $\alpha\%$ highest posterior density interval: interval containing $\alpha\%$ of the posterior density with the *highest* density values (must contain mode).
    - $\alpha\%$ equal-tailed interval: probability of falling below the interval is $(1-\alpha)/2$ and probability of falling above the interval is $(1-\alpha)/2$ (must contain the median).

Expectations and Variances of Random Variables

Recall that the *expectation* of a random variable is its *mean*, or *average* outcome.

Definition

The **expectation** of X (discrete) is given by

$$\mathbb{E}(X) = \sum_x x \cdot \Pr(X = x).$$

The **expectation** of X (continuous) is given by

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\ dx,$$

where $f_X(x)$ is the PDF of X. The expectation of X is also referred to as the **expected value** of X or the **mean** of X.

Expectations and Variances of Random Variables

Similarly, the *variance* of a random variable is its *average squared dispersion from the mean*.

Definition

The **variance** of X is given by

$$\text{Var}(X) := \mathbb{E}[(X - \mathbb{E}(X))^2].$$

For X discrete, this is

$$\text{Var}(X) = \sum_x (x - \mathbb{E}(X))^2 \text{Pr}(X = x).$$

For X continuous, this is

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mathbb{E}(X))^2 f_X(x)\ dx.$$

Bayesian estimation and inference

- Posterior mean:

$$\mathbb{E}(\theta \mid \boldsymbol{y}) = \int_{-\infty}^{\infty} \theta \cdot f(\theta \mid \boldsymbol{y}) \; d\theta.$$

- Posterior variance:

$$\text{Var}(\theta \mid \boldsymbol{y}) = \int_{-\infty}^{\infty} (\theta - \mathbb{E}(\theta \mid \boldsymbol{y}))^2 \cdot f(\theta \mid \boldsymbol{y}) \; d\theta.$$

- Posterior standard deviation:

$$SD(\theta \mid \boldsymbol{y}) = \sqrt{\text{Var}(\theta \mid \boldsymbol{y})}.$$

- All calculable using techniques of calculus, or can be numerically approximated (more later).

Bayesian estimation and inference

- $\alpha\%$ mean-centred credible interval: $a \in \mathbb{R}$ such that

$$\alpha = \int_{\mathbb{E}(\theta \mid \boldsymbol{y})-a}^{\mathbb{E}(\theta \mid \boldsymbol{y})+a} f(\theta \mid \boldsymbol{y})\, d\theta.$$

  Then

$$[\mathbb{E}(\theta \mid \boldsymbol{y}) - a,\ \mathbb{E}(\theta \mid \boldsymbol{y}) + a]$$

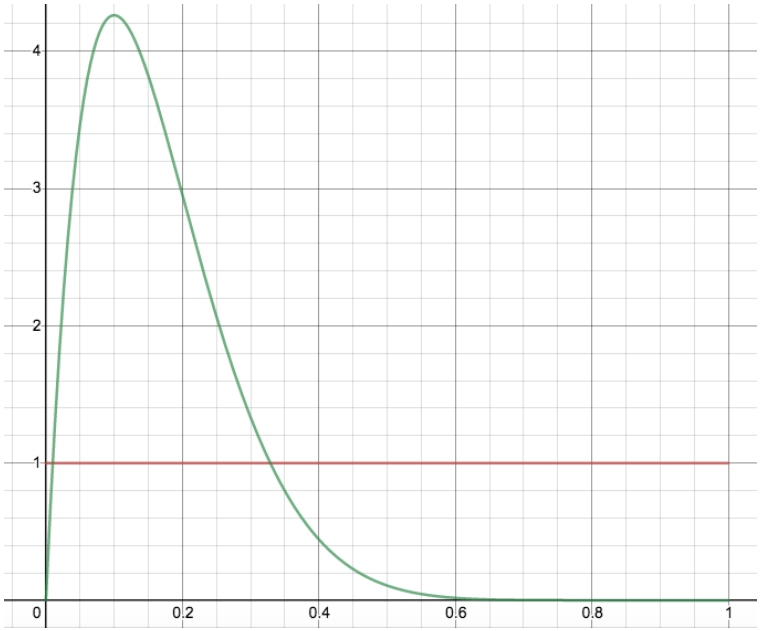
  is the $\alpha\%$ mean-centred credible interval.

- $\alpha\%$ highest posterior density interval: the *shortest* interval $[a, b]$ such that

$$\alpha = \int_a^b f(\theta \mid \boldsymbol{y})\, d\theta.$$

Bayesian estimation and inference

- Suppose $n = 10$ and we observe 1 patient with disease: $\bar{Y} = 0.1$
- Prior 1: $\pi_1(\theta) = 1, \quad 0 \leqslant \theta \leqslant 1$
- Posterior 1: $f_1(\theta \mid \boldsymbol{y}) = 110 \cdot \theta(1-\theta)^9, \quad 0 \leqslant \theta \leqslant 1$
- Posterior mean: $\mathbb{E}(\theta \mid \boldsymbol{y}) = 0.2$
- Posterior mode: 0.1
- Posterior standard deviation: $SD(\theta \mid \boldsymbol{y}) = 0.1033$



Bayesian estimation and inference

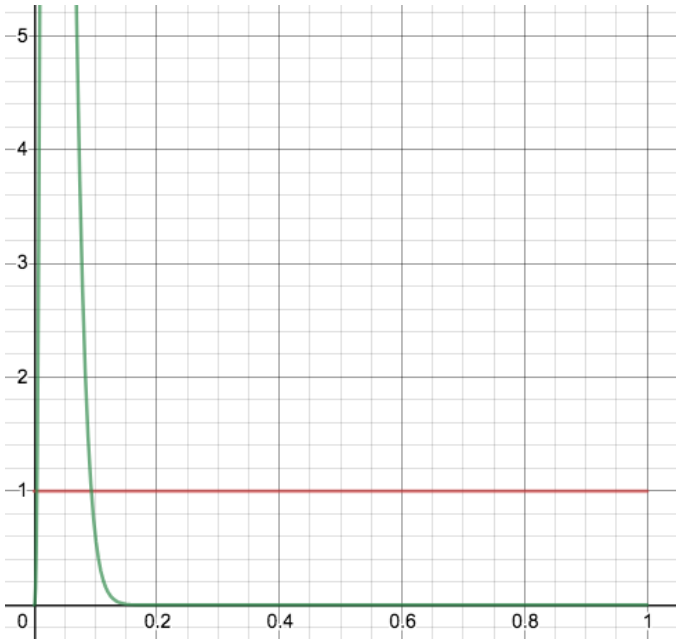
- Suppose $n = 10$ and we observe 1 patient with disease: $\bar{Y} = 0.1$
- Prior 2: $\pi_2(\theta) = 30 \cdot \theta^2(1-\theta)^2, \quad 0 \leqslant \theta \leqslant 1$
- Posterior 2: $f_2(\theta \mid \boldsymbol{y}) = 5460 \cdot \theta^3(1-\theta)^{11}, \quad 0 \leqslant \theta \leqslant 1$
- Posterior mean: $\mathbb{E}(\theta \mid \boldsymbol{y}) = 0.25$
- Posterior mode: 0.21
- Posterior standard deviation: $SD(\theta \mid \boldsymbol{y}) = 0.1050$

Bayesian estimation and inference

- Suppose $n = 10$ and we observe 1 patient with disease: $\bar{Y} = 0.1$
- Prior 3: $\pi_3(\theta) = 2(1-\theta), \quad 0 \leqslant \theta \leqslant 1$
- Posterior 3: $f_3(\theta \mid \boldsymbol{y}) = 132 \cdot \theta(1-\theta)^{10}, \quad 0 \leqslant \theta \leqslant 1$
- Posterior mean: $\mathbb{E}(\theta \mid \boldsymbol{y}) = 0.15$
- Posterior mode: 0.09
- Posterior standard deviation: $SD(\theta \mid \boldsymbol{y}) = 0.0964$

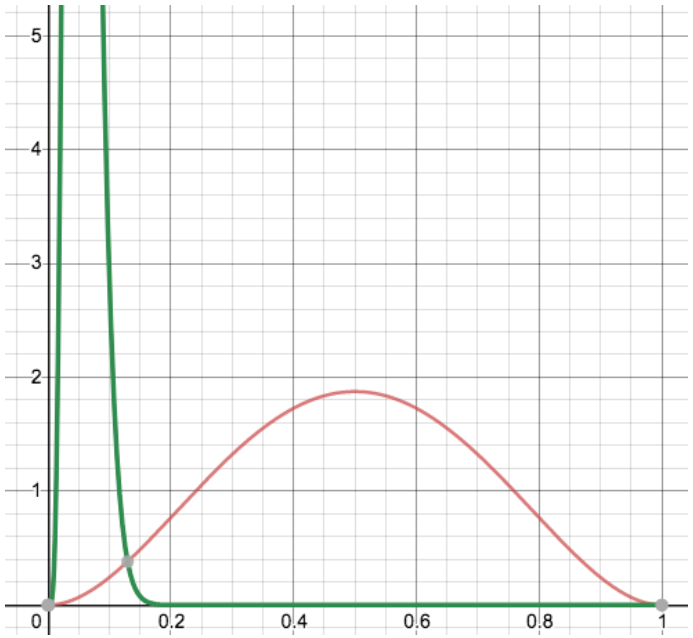
Bayesian estimation and inference

- Suppose $n = 100$ and we observe 3 patients with disease: $\bar{Y} = 0.03$
- Prior 1: $\pi_1(\theta) = 1, \quad 0 \leqslant \theta \leqslant 1$
- Posterior 1: $f_1(\theta \mid \boldsymbol{y}) = c_1 \cdot \theta^3(1-\theta)^{97}, \quad 0 \leqslant \theta \leqslant 1$
- Posterior mean: $\mathbb{E}(\theta \mid \boldsymbol{y}) = 0.039$
- Posterior mode: 0.03
- Posterior standard deviation: $SD(\theta \mid \boldsymbol{y}) = 0.0191$



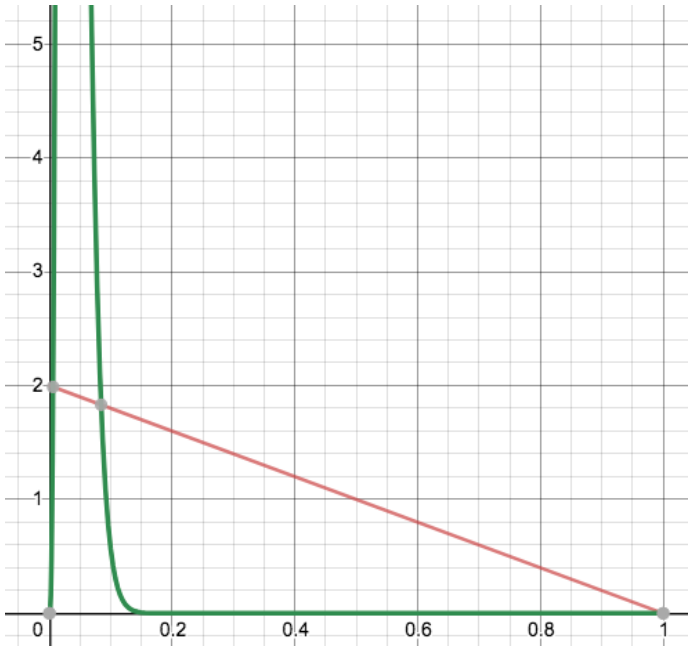
Bayesian estimation and inference

- Suppose $n = 100$ and we observe 3 patients with disease: $\bar{Y} = 0.03$
- Prior 2: $\pi_2(\theta) = 30\theta^2(1-\theta)^2, \quad 0 \leqslant \theta \leqslant 1$
- Posterior 2: $f_2(\theta \mid \boldsymbol{y}) = c_2 \cdot \theta^5(1-\theta)^{99}, \quad 0 \leqslant \theta \leqslant 1$
- Posterior mean: $\mathbb{E}(\theta \mid \boldsymbol{y}) = 0.057$
- Posterior mode: 0.048
- Posterior standard deviation: $SD(\theta \mid \boldsymbol{y}) = 0.0223$



Bayesian estimation and inference

- Suppose $n = 100$ and we observe 3 patients with disease: $\bar{Y} = 0.03$
- Prior 3: $\pi_3(\theta) = 2(1-\theta), \quad 0 \leqslant \theta \leqslant 1$
- Posterior 3: $f_3(\theta \mid \boldsymbol{y}) = c_3 \cdot \theta^3(1-\theta)^{98}, \quad 0 \leqslant \theta \leqslant 1$
- Posterior mean: $\mathbb{E}(\theta \mid \boldsymbol{y}) = 0.039$
- Posterior mode: 0.03
- Posterior standard deviation: $SD(\theta \mid \boldsymbol{y}) = 0.0189$



More on the Beta-Binomial model

- The models we have been considering all fall under the Beta-Binomial class; i.e. we observe Binomial/Bernoulli data (Y equals 0 or 1) and assume a Beta prior for $\theta = \Pr(Y = 1)$, which combined generate Beta posteriors for θ.
- Note: for Binomial/Bernoulli data, θ completely characterizes the random phenomenon of interest; i.e. generates the likelihood (PMF) for any sample.
- Beta distributions are a very natural, flexible, and convenient choice for the priors, but they are not the only ones.
- I.e. there are lots of other PDFs on [0,1] that can *not* be written as Beta distribution: e.g. the "tent" function.
- Nonetheless, we *usually* assume some kind of Beta prior for convenience.

Bayesian inference vs. hypothesis testing

- In classical Bayesian methodology, there is *no such thing as hypothesis testing*.
- Some modern Bayesian methodologists have introduced a notion a hypothesis testing (which we will return to later), but in general: *avoid it*.
- No need to "test" hypotheses when we can explicitly quantify the believability of *any* set of hypotheses directly using the posterior distribution.
- Bayesian inference is all about *estimation* of various properties of the posterior distribution:
  - The posterior mean/mode/median all give "best estimates" (can be formalized).
  - The posterior standard deviation and/or credible intervals all give measures of uncertainty about these estimates.

Multiparameter models

Most real-world situations are going to require *multiparameter* models; i.e. there will be more than one unknown parameter that characterizes the random phenomenon (distribution) of interest.

- Normal data: $Y \sim N(\mu, \sigma^2)$
- Simple regression: $Y = \beta_0 + \beta_X X + \varepsilon$
- Pretty much any other kind of statistical modelling!

Multiparameter model: Normal data

First consider the simple problem of estimating the mean height in a population.

- Reasonable to assume a normal model for the data: $H \sim N(\mu, \sigma^2)$
- Here, μ is the parameter of inferential interest, but need to deal with σ^2 too.
- If we aren't actually interested in the value of σ^2, it becomes what we call a *nuisance parameter*; i.e. have to account for it, but aren't actually interested in its value.

Normal data: likelihood function

Sample $H_1, \ldots, H_n \sim N(\mu, \sigma^2)$.

- Here, the random variable H is *continuous*, so we describe its likelihood function via a PDF (density function):

$$f_H(h) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(h-\mu)^2}{2\sigma^2}}$$

- Again, assumign random sampling so that all observations are independent, the likelihood function is a product of point-likelihoods:

$$\begin{aligned} f(\boldsymbol{h} \mid \mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(h_i-\mu)^2}{2\sigma^2}} \\ &= (\sqrt{2\pi}\sigma)^{-n} \exp\left(-\frac{\sum_{i=1}^{n}(h_i-\mu)^2}{2\sigma^2}\right) \end{aligned}$$

Normal data: choosing priors

Just as before, we now need to choose a prior distribution for the unknown parameters.

- But here, we have two parameters: thus, need two priors (or one *joint* prior)!
- Values for mean μ and values for variance σ^2 can theoretically only assume *nonnegative* values in the context of this problem (measuring heights in cm).
- There are *many* reasonable prior distributions one could define, but we will consider two of the most common here:

$$\mu \sim N(0, 100^2)$$
$$\sigma \sim U(0, 50)$$

Normal data: choosing priors

- Notice that our particular priors

$$\mu \sim N(100, 100^2)$$
$$\sigma \sim U(0, 50)$$

  are very *uninformative* in some ways (i.e. very diffuse), but very *informative* in other ways (e.g. standard deviation cannot be too large)

- Also, notice that defining $\mu \sim N(100, 100)$ might seem a bit silly here: Normal r.v.'s can assume *any* real value, but we know we can't have *negative* heights!

- Is this a problem? Not for this context, since the data will easily pull the posterior away from the negative region of the prior. (Recall too that we chose to model the *likelihood* as a normal r.v. too.]

- But could be a problem in more complicated scenarios.

Normal data: calculating the posterior

- With the likelihood function and the priors, we could now write down the posterior function:

$$f(\mu, \sigma \mid \boldsymbol{h}) \propto f(\boldsymbol{h} \mid \mu, \sigma)\pi(\mu, \sigma).$$

- Note that we defined *independent* priors for our two parameters, so

$$\pi(\mu, \sigma) = \pi(\mu)\pi(\sigma).$$

- Note: since we started with two parameters, we end up with a posterior that quantifies both parameters; i.e. we have a *joint* posterior density function.

Normal data: calculating the posterior

- But we really only care about making inferences on the mean μ; i.e. the variance σ^2 is a *nuisance* parameter.
- How do we get rid of a variable we don't care about? Marginalize it out!

$$\begin{aligned} f(\mu \mid \boldsymbol{h}) &= \int_{-\infty}^{\infty} f(\mu, \sigma \mid \boldsymbol{h}) \; d\sigma \\ &\propto \int_{-\infty}^{\infty} f(\boldsymbol{h} \mid \mu, \sigma)\pi(\mu)\pi(\sigma) \; d\sigma \\ &= \pi(\mu) \int_{-\infty}^{\infty} f(\boldsymbol{h} \mid \mu, \sigma)\pi(\sigma) \; d\sigma \end{aligned}$$

  and all this we have (though not necessarily easy to calculate analytically).
- Thus, we can get our usual estimates (e.g. posterior mean, mode, etc.) on the *marginal posterior* of interest.

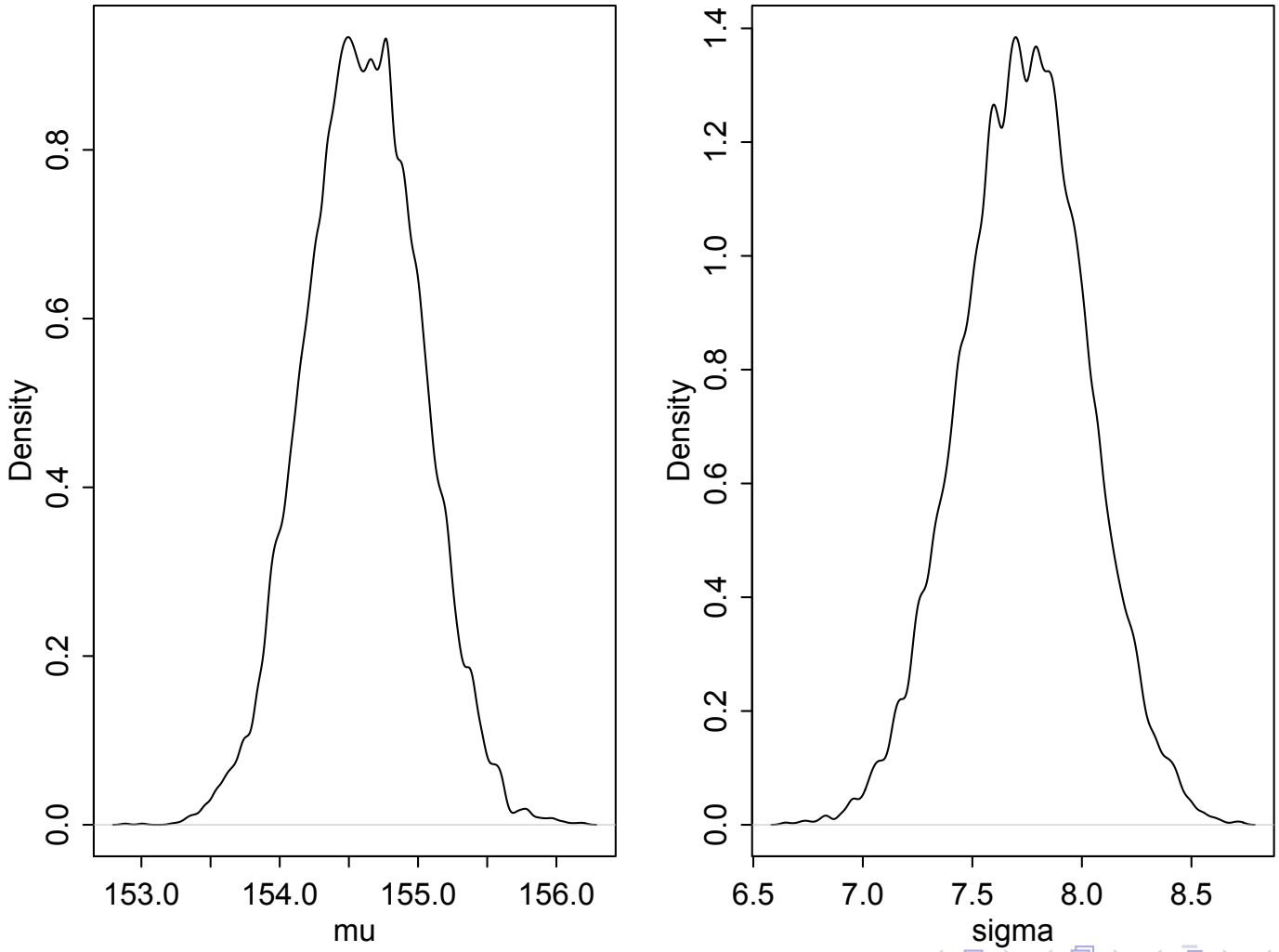
Normal data: numerical estimates

- Here, we'll use the height data given by McElreath (Chapter 4)
- R code (straight from McElreath book):

```
library(rethinking)
data(Howell1)
d <- Howell1
d2 <- d[d$age >=18, ]
flist <- alist(
        height ~ dnorm(mu,sigma),
        mu ~ dnorm(100,100),
        sigma ~ dunif(0,50)
)
m1 <- map(flist,data=d2)
precis(m1,prob=.95)
post <- extract.samples(m1,n=10000)
dens(post)
```

Normal data: numerical estimates

Marginal posterior densities (approximated by simulation):



Normal data: numerical estimates

Default model output: (marginal) posterior means and standard deviations, 95% mean-centred credible intervals:

	Mean	StdDev	2.5%	97.5%
mu	154.60	0.41	153.79	155.4
sigma	7.73	0.29	7.16	8.3

- Try playing around with the prior specifications and see what happens to the posteriors (there are 352 data points here, so a decent sample size).
- Also, compare to the classical MLEs.

A simple regression model

Can now extend these ideas to model, say, mean *height* as a linear function of *weight* (same McElreath dataset as before).

- Typical regression model would be:

$$H = \beta_0 + \beta_1 W + \varepsilon,$$

  where $\varepsilon \sim N(0, \sigma^2)$ is classically assumed.

- We will use the same model (there is no such thing as "Bayesian modelling", only Bayesian estimation or inference).

- How many parameters will this model have? How many are nuisance parameters?

Regression model: compact specification

The typical way to express this regression model (or any model) in a Bayesian context is to explicitly list the likelihood and priors compactly:

$$
\begin{aligned}
H &\sim N(\beta_0 + \beta_1 W, \sigma) \\
\beta_0 &\sim N(0, 200) \\
\beta_1 &\sim N(0, 20) \\
\sigma &\sim U(0, 50)
\end{aligned}
$$

- Notice that this tells us everything we need to know to construct the posterior distribution for any or all of the model parameters.
- Note: here, I'm using $N(\mu, \sigma)$ notation to define Normal r.v.s.

Regression model: long specification

Sample data: $X_1 = (H_1, W_1), \ldots, X_n = (H_n, W_n)$.

- Explicit likelihood function:

$$f(\boldsymbol{x} \mid \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(h_i - \beta_0 - \beta_1 w_i)^2}{2\sigma^2}}$$

$$= (\sqrt{2\pi}\sigma)^{-n} \exp\left(-\frac{\sum_{i=1}^{n} (h_i - \beta_0 - \beta_1 w_i)^2}{2\sigma^2} \right)$$

- Explicit priors:

$$\pi(\beta_0) = \frac{1}{\sqrt{2\pi} \cdot 200} e^{-\frac{\beta_0^2}{2 \cdot 200^2}}, \quad \pi(\beta_1) = \frac{1}{\sqrt{2\pi} \cdot 20} e^{-\frac{\beta_1^2}{2 \cdot 20^2}}$$

$$\pi(\sigma) = \frac{1}{50}, \quad 0 \leqslant \sigma \leqslant 50$$

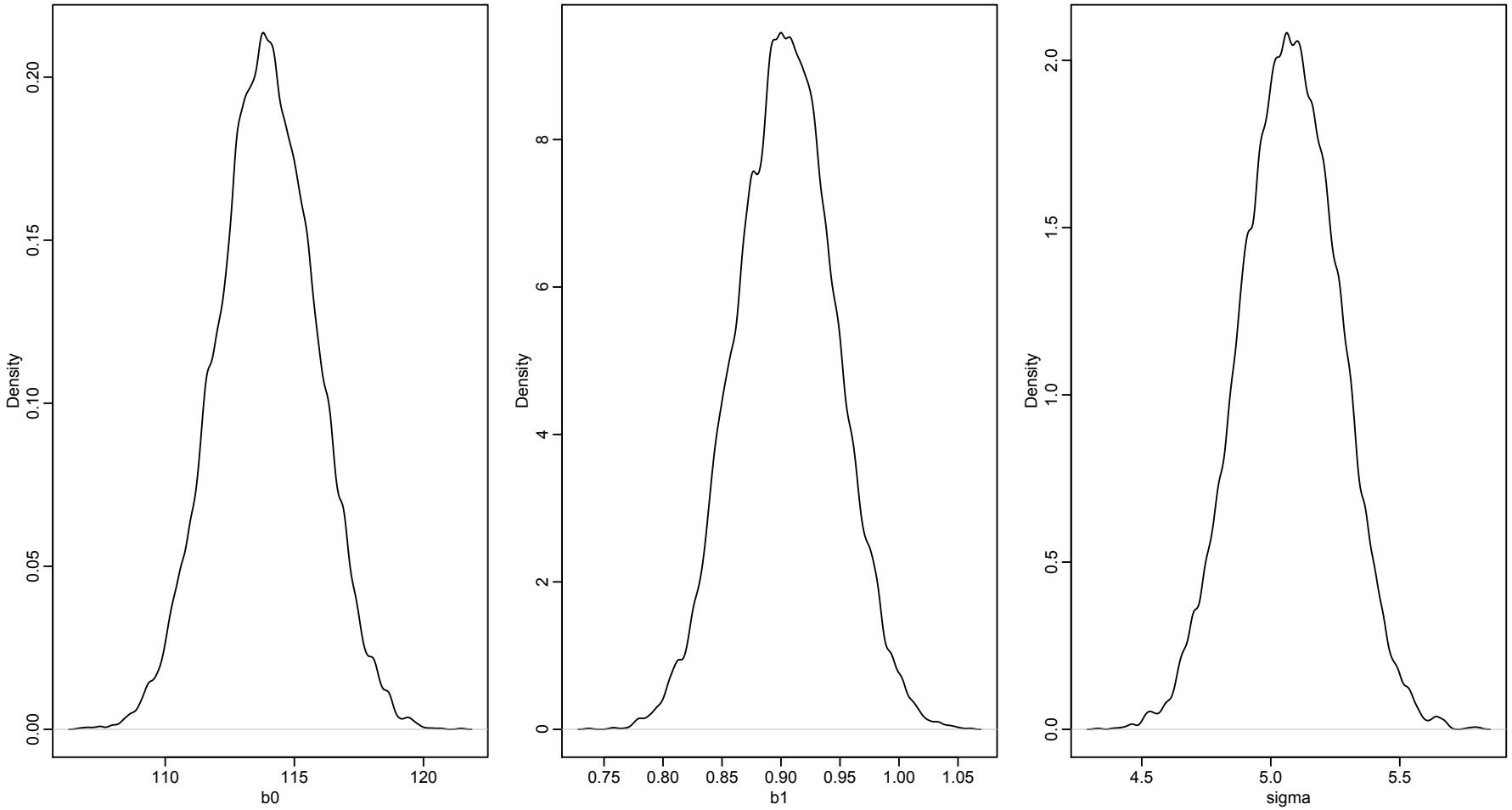
Regression model: numerical estimates

- Using same dataset as before:

```
library(rethinking)
flist <- alist(
        height ~ dnorm(b0 + b1*weight,sigma),
        b0 ~ dnorm(0,200),
        b1 ~ dnorm(0,20),
        sigma ~ dunif(0,50)
)
m2 <- map(flist,data=d2)
precis(m2,prob=.95)
post <- extract.samples(m2,n=10000)
dens(post)
```

Regression model: numerical estimates

Marginal posterior densities (approximated by simulation):



Regression model: numerical estimates

Default model output: (marginal) posterior means and standard deviations, 95% mean-centred credible intervals:

	Mean	StdDev	2.5%	97.5%
b0	113.87	1.91	110.13	117.60
b1	0.91	0.04	0.82	0.99
sigma	5.07	0.19	4.70	5.45

- Try playing around with the prior specifications and see what happens to the posteriors.
- Notice how the marginal posteriors are all *linked*; i.e. change one prior, you change *all* posterior estimates (potentially).
- Also, compare to the classical MLEs.