Module 3: Introduction to the Normal Distribution

Rebecca C. Steorts

Announcements

1. There was a typo regarding admissibility in Module 2.
2. Homework 3 will be graded leniantly as there are many solutions and there was also a typo. (Students will not be penalized for my typo.)
3. Exam I is Feb. 7 and Feb 5 will be a review day.
4. Exam I: Will cover material on Modules 0 – 3, labs 1 – 3, and homeworks 1 –3. I'll go into more details once we finish module 3 regarding expectations and coverage.
5. Office hours: How many students cannot make a single office hour due to a class conflict?

Exercise

Suppose $a < x < b$. Consider the notation $I_{(a,b)}(x)$, where I denotes the indicator function. We define $I_{(a,b)}(x)$ to be the following:

$$I_{(a,b)}(x) = \begin{cases} 1 & \text{if } a < x < b, \\ 0 & \text{otherwise.} \end{cases}$$

Let

$$\begin{aligned} X|\theta &\sim \text{Uniform}(0, \theta) \\ \theta &\sim \text{Pareto}(\alpha, \beta), \end{aligned}$$

where $p(\theta) = \dfrac{\alpha\beta^{\alpha}}{\theta^{\alpha+1}} I_{(\beta,\infty)}(\theta)$. Write out the likelihood $p(x \mid \theta)$. Then calculate the posterior distribution of $\theta|x$.

Solution

$$
\begin{aligned}
P(\theta \mid x) &\propto p(x \mid \theta) p(\theta) && (1)\\
&\propto \frac{1}{\theta} I(0 < x < \theta) \frac{\alpha \beta^{\alpha}}{\theta^{\alpha+1}} I(\theta > \beta) && (2)\\
&\propto \frac{\alpha \beta^{\alpha}}{\theta^{\alpha+2}} I(\theta > x) I(\theta > \beta) && (3)\\
&\propto \frac{\alpha \beta^{\alpha}}{\theta^{\alpha+2}} I(\theta > \max\{x, \beta\}). && (4)
\end{aligned}
$$

Thus,

$$\theta \mid x \sim \text{Pareto}(\alpha + 1, \max\{x, \beta\}).$$

Agenda

- The normal distribution
- The variance versus precision
- The re-parameterized normal distribution
- Common properties
- The normal-uniform model
- The normal-normal model

Normal distribution

The normal distribution $\mathcal{N}(\mu, \sigma^2)$

- with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ - (standard deviation $\sigma = \sqrt{\sigma^2}$) has p.d.f.

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x-\mu)^2 \Big)$$

  for $x \in \mathbb{R}$.

It is often more convenient to write the p.d.f. in terms of the **precision**, or inverse variance, $\lambda = 1/\sigma^2$ rather than the variance.

Re-parameterized Normal

In this parametrization, the p.d.f. is

$$\mathcal{N}(x \mid \mu, \lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\tfrac{1}{2}\lambda(x-\mu)^2\right)$$

since $\sigma^2 = 1/\lambda = \lambda^{-1}$.

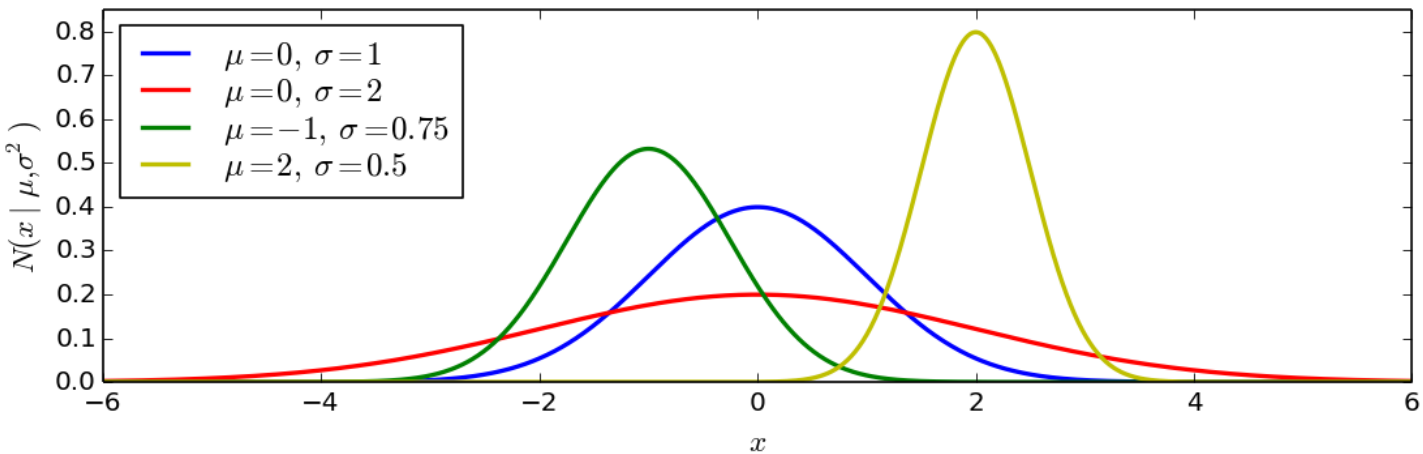

Figure 1: Normal distribution with various choices of μ and σ.

Normality?

- The central limit theorem (CLT) states that the sum of a large number of independent random variables tends to be approximately normally distributed.
- Real world data often appears approximately normal.

Normality?

- Human heights and other body measurements,
- Cumulative hydrologic measures such as annual rainfall or monthly river discharge,
- Errors in astronomical or physical observations,
- Diffusion of a substance in a liquid or gas.
- Some things are products of many independent variables (rather than sums), and in such cases the logarithm will be approximately normal since it is a sum of many independent variables

Example: stock market indices, due to the effect of compound interest.

Properties of the Normal distribution

- Mean, median, and mode are all the same (μ)
- Symmetric about the mean
- 95% probability within $\pm 1.96\sigma$ of the mean (roughly, $\pm 2\sigma$)
- If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(m, s^2)$ independently, then

$$aX + bY \sim \mathcal{N}(a\mu + bm,\ a^2\sigma^2 + b^2 s^2). \tag{5}$$

- Careful: `rnorm`, `dnorm`, `pnorm`, and `qnorm` in `R` take the mean and standard deviation σ as arguments (not mean and variance σ^2). For example, `rnorm(n,m,s)` generates n normal random variables from $\mathcal{N}(m, s^2)$.

Normal-Uniform

$$X_1, \ldots, X_n \mid \theta \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2).$$

Assume the prior on θ is constant over the real line. We can write this as $p(\theta) \propto 1$.

Derive the posterior distribution.

Solution

$$
\begin{aligned}
p(\theta \mid x_{1:n}) &\propto \mathcal{N}(\theta, \sigma^2) \times 1 && (6)\\
&\propto \left(\frac{\ell}{2\pi}\right)^{n/2} \exp\left(-\tfrac{1}{2}\ell \sum_i (x_i - \theta)^2\right)\\
&\propto \exp\left(-\tfrac{1}{2}\ell \sum_i (x_i - \bar{x} + \bar{x} - \theta)^2\right)\\
&\propto \exp\left(-\tfrac{1}{2}\ell \sum_i (x_i - \bar{x})^2\right) \exp\left(-\tfrac{1}{2}\ell \sum_i (\bar{x} - \theta)^2\right)\\
&\propto \exp\left(-\tfrac{1}{2}\ell \sum_i (\bar{x} - \theta)^2\right) && (7)\\
&= \exp\left(-\tfrac{n\ell}{2} \sum_i (\theta - \bar{x})^2\right) && (8)
\end{aligned}
$$

This implies that

$$\theta \mid x_{1:n} \sim N(\bar{x}, (n\ell)^{-1})$$

Normal-Normal

$$X_1, \ldots, X_n \mid \theta \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \lambda^{-1}).$$

Assume the precision $\lambda = 1/\sigma^2$ is known and fixed, and θ is given a $\mathcal{N}(\mu_0, \lambda_0^{-1})$ prior:

$$\boldsymbol{\theta} \sim \mathcal{N}(\mu_0, \lambda_0^{-1})$$

i.e., $p(\theta) = \mathcal{N}(\theta \mid \mu_0, \lambda_0^{-1})$. This is sometimes referred to as a **Normal–Normal** model.

Posterior derivation

We begin with the **likelihood** of the normal distribution.

For any x and ℓ,

$$\begin{aligned}\mathcal{N}(x \mid \theta, \ell^{-1}) &= \sqrt{\frac{\ell}{2\pi}} \exp\left(-\tfrac{1}{2}\ell(x-\theta)^2\right) \\ &\underset{\theta}{\propto} \exp\left(-\tfrac{1}{2}\ell(x^2 - 2x\theta + \theta^2)\right) \\ &\underset{\theta}{\propto} \exp\left(\ell x\theta - \tfrac{1}{2}\ell\theta^2\right). \end{aligned} \tag{9}$$

Note: we drop the **constant term** and we will do this often when working with the normal distribution.

Posterior derivation (continued)

We now consider the **prior** distribution on θ.

Due to the symmetry of the normal p.d.f.,

$$\mathcal{N}(\theta \mid \mu_0, \lambda_0^{-1}) = \mathcal{N}(\mu_0 \mid \theta, \lambda_0^{-1}) \underset{\theta}{\propto} \exp\left(\lambda_0 \mu_0 \theta - \tfrac{1}{2}\lambda_0 \theta^2\right), \quad (10)$$

where $x = \mu_0$ and $\ell = \lambda_0$.

Posterior derivation (continued)

Let

$$L = \lambda_0 + n\lambda \quad \text{and} \quad M = \frac{\lambda_0\mu_0 + \lambda\sum_{i=1}^{n} x_i}{\lambda_0 + n\lambda}.$$

$$\begin{aligned}
p(\theta|x_{1:n}) &\propto p(\theta)p(x_{1:n}|\theta) \\
&= \mathcal{N}(\theta \mid \mu_0, \lambda_0^{-1}) \prod_{i=1}^{n} \mathcal{N}(x_i \mid \theta, \lambda^{-1}) \\
&\overset{(a)}{\propto} \exp\left(\lambda_0\mu_0\theta - \tfrac{1}{2}\lambda_0\theta^2\right) \exp\left(\lambda(\textstyle\sum x_i)\theta - \tfrac{1}{2}n\lambda\theta^2\right) \\
&= \exp\left((\lambda_0\mu_0 + \lambda\textstyle\sum x_i)\theta - \tfrac{1}{2}(\lambda_0 + n\lambda)\theta^2\right) \\
&= \exp(LM\theta - \tfrac{1}{2}L\theta^2) \\
&\overset{(b)}{\propto} \mathcal{N}(M \mid \theta, L^{-1}) = \mathcal{N}(\theta \mid M, L^{-1}),
\end{aligned}$$

where step (a) uses Equations 9 and 10, and step (b) uses Equation 9 with $x = M$ and $\ell = L$.

Posterior derivation (continued)

Recall

$$L = \lambda_0 + n\lambda \quad \text{and} \quad M = \frac{\lambda_0 \mu_0 + \lambda \sum_{i=1}^{n} x_i}{\lambda_0 + n\lambda}.$$

It turns out that the posterior is

$$\boldsymbol{\theta} | x_{1:n} \sim \mathcal{N}(M, L^{-1})$$

i.e., $p(\theta | x_{1:n}) = \mathcal{N}(\theta \mid M, L^{-1})$.

Thus, the normal distribution is, itself, a conjugate prior for the mean of a normal distribution with known precision.

Heights of Adult Humans

- Heights tend to be normally distributed because there are many independent genetic and environmental factors which contribute additively to overall height
- This leads to a normal distribution due to the central limit theorem.

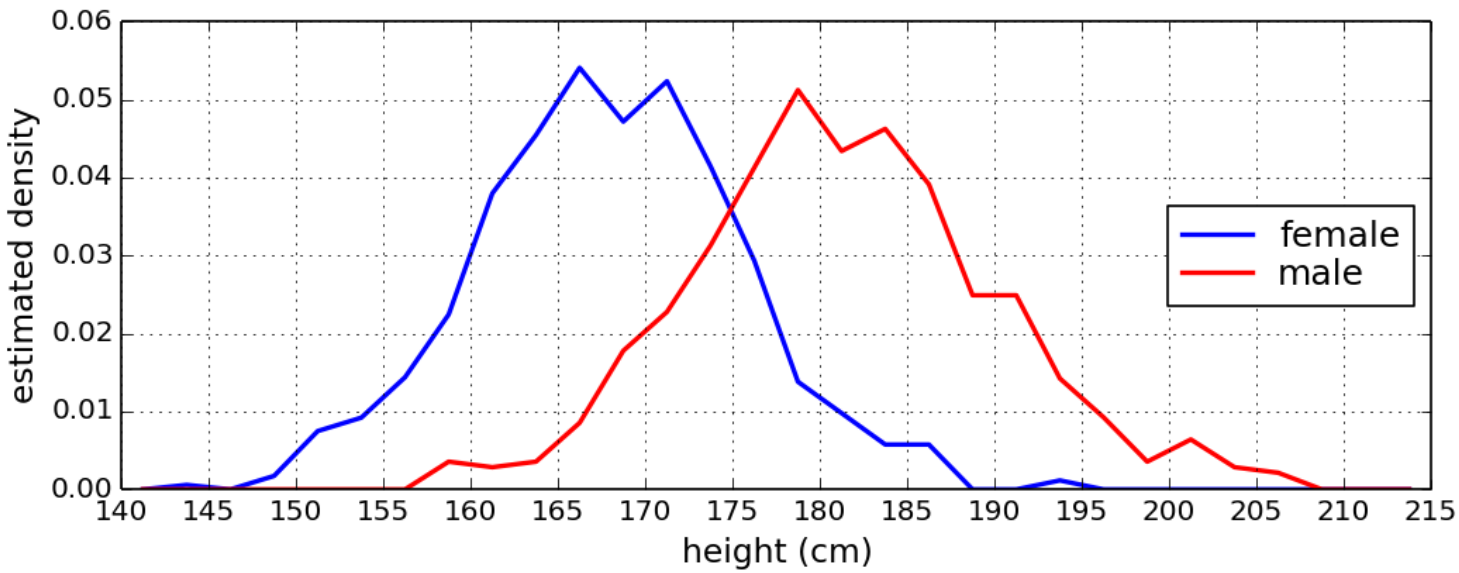

Figure 2: Estimated densities of the heights of Dutch women and Dutch men based on a sample of 695 women and 562 men.

Heights of Adult Humans

- Consider combined distribution of heights (pooling females and males together). Would this be normal?
- It is thought that such data is bimodal (having two maxima). Is it really bimodal? (See, Schilling et al. (2002))

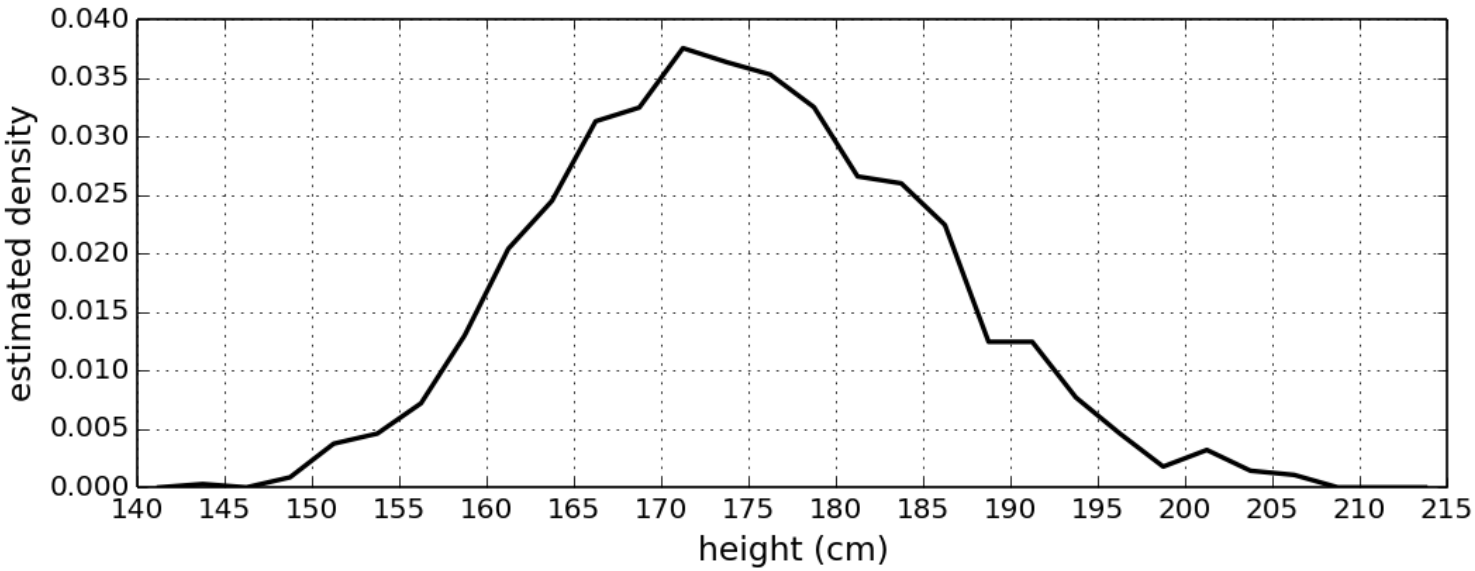

Figure 3: Estimated density for Dutch women and men together, assuming there is an equal proportion of women and men in the population.

Heights of Adult Humans, Combined

At a glance, while the heights of women and men separately do appear to be roughly normally distributed, the combined distribution does not look bimodal. How could we test whether it is bimodal in a more precise way?

Our Assumptions

- Assume female heights and male heights are each normally distributed.
- Let's assume they have the same standard deviation, and also that there is an equal proportion of women and men in the population.
- Then, it is known that the combined distribution is bimodal if and only if the difference between the means is greater than twice the standard deviation (Helguerro, 1904).

Model

In mathematical notation: Assume the female heights are

$$X_1, \ldots, X_k \overset{\text{iid}}{\sim} \mathcal{N}(\theta_f, \sigma^2),$$

where $k = 695$, the male heights are

$$Y_1, \ldots, Y_\ell \overset{\text{iid}}{\sim} \mathcal{N}(\theta_m, \sigma^2),$$

where $\ell = 562$, and the p.d.f. of the combined distribution of heights is

$$\tfrac{1}{2}\mathcal{N}(x \mid \theta_f, \sigma^2) + \tfrac{1}{2}\mathcal{N}(x \mid \theta_m, \sigma^2).$$

(This is an example of what is called a two-component ***mixture*** distribution.)

Model

Let's put independent normal priors on θ_f and θ_m:

$$p(\theta_f, \theta_m) = p(\theta_f)p(\theta_m) = \mathcal{N}(\theta_f \mid \mu_{0,f}, \sigma_0^2)\mathcal{N}(\theta_m \mid \mu_{0,m}, \sigma_0^2).$$

- Assume σ^2 is known.
- For the purposes of this example, let's use $\sigma = 8$ centimeters (about 3 inches).
- Based on common knowledge of typical human heights, let's choose the prior parameters (a.k.a. hyperparameters) as follows:

$\mu_{0,f}$	(mean of prior on female mean ht)	165 cm ($\approx$ 5 ft, 5 in)
$\mu_{0,m}$	(mean of prior on male mean ht)	178 cm ($\approx$ 5 ft, 10 in)
σ_0	(std. dev. of priors on mean ht)	15 cm ($\approx$ 6 in)

Bimodal Fact

It is known (Helguerro, 1904) that the combined distribution is bimodal if and only if

$$|\theta_f - \theta_m| > 2\sigma.$$

So, to address our question of interest ("Is human height bimodal?"), we would like to compute the posterior probability that this is the case, i.e., we want to know

$$\mathbb{P}(\text{bimodal} \mid \text{data}) = \mathbb{P}\big(|\boldsymbol{\theta}_f - \boldsymbol{\theta}_m| > 2\sigma \mid x_{1:k}, y_{1:\ell}\big).$$

Results

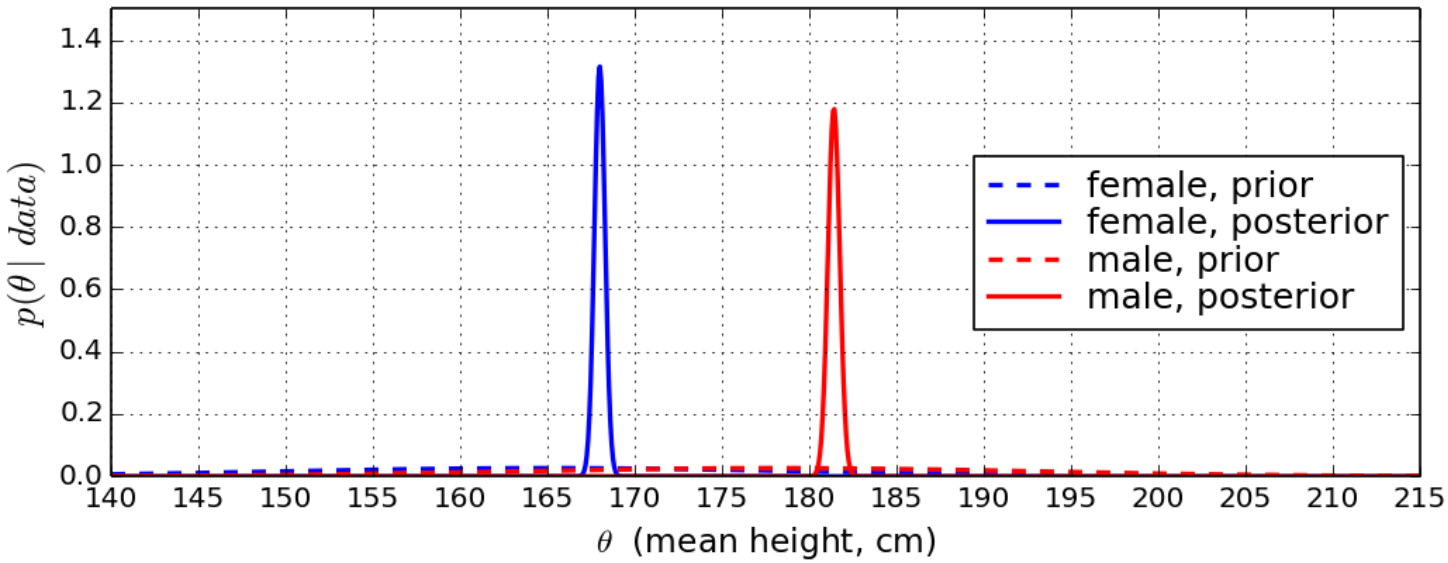

Figure 4: Priors and posteriors for the mean heights of Dutch women and men.

Results (continued)

We can compute the posteriors for θ_f and θ_m using Equation 17 for each of them, independently. Figure 4 shows the priors and posteriors.

- Sample means: $\bar{x} = 168.0$ cm (5 feet 6.1 inches) for females, and $\bar{y} = 181.4$ cm (5 feet 11.4 inches) for males.
- Posterior means: $M_f = 168.0$ cm for females, and $M_m = 181.4$ cm for males. (Essentially identical to the sample means, due to the relatively large sample size and relatively weak prior.)
- Posterior standard deviations: $1/\sqrt{L_f} = 0.30$ cm and $1/\sqrt{L_m} = 0.34$ cm.

Results (continued)

By Equation 5 (a linear combination of independent normals is normal),

$$\boldsymbol{\theta}_m - \boldsymbol{\theta}_f \mid x_{1:k}, y_{1:\ell} \sim \mathcal{N}(M_m - M_f,\, L_m^{-1} + L_f^{-1}) = \mathcal{N}(13.4, 0.45^2)$$

so we can compute $\mathbb{P}(\text{bimodal} \mid \text{data})$ using the normal c.d.f. Φ:

$$\begin{aligned}
\mathbb{P}(\text{bimodal} \mid \text{data}) &= \mathbb{P}(|\boldsymbol{\theta}_m - \boldsymbol{\theta}_f| > 2\sigma \mid x_{1:k}, y_{1:\ell}) \\
&= \Phi(-2\sigma \mid 13.4, 0.45^2) + (1 - \Phi(2\sigma \mid 13.4, 0.45^2)) \\
&= 6.1 \times 10^{-9}.
\end{aligned}$$

Intuitive interpretation: The posteriors are about 13 or 14 centimeters apart, which is under the $2\sigma = 16$ threshold for bimodality, and they are sufficiently concentrated that the posterior probability of bimodality is essentially zero.

Exercise

Suppose $a < x < b$. Consider the notation $I_{(a,b)}(x)$, where I denotes the indicator function. We define $I_{(a,b)}(x)$ to be the following:

$$I_{(a,b)}(x) = \begin{cases} 1 & \text{if } a < x < b, \\ 0 & \text{otherwise.} \end{cases}$$

Let

$$\begin{aligned} X_1 \ldots, X_n | \theta &\sim \text{Uniform}(0, \theta) \\ \theta &\sim \text{Pareto}(\alpha, \beta), \end{aligned}$$

where $p(\theta) = \dfrac{\alpha \beta^{\alpha}}{\theta^{\alpha+1}} I_{(\beta,\infty)}(\theta)$.

Write out the likelihood $p(x \mid \theta)$. Then calculate the posterior distribution of $\theta | x$.

Likelihood

$$p(x_{1:n} \mid \theta) = \prod_{i=1}^{n} \frac{1}{\theta} I(\theta > x_i) \tag{11}$$

$$= \frac{1}{\theta^n} \prod_{i=1}^{n} I(\theta > x_i) \tag{12}$$

$$= \frac{1}{\theta^n} I(\theta > \max\{x_{1:n}\}) \tag{13}$$

Posterior

$$p(\theta \mid x_{1:n}) \propto \frac{1}{\theta^n} I(\theta > \max\{x_{1:n}\}) \frac{\alpha\beta^\alpha}{\theta^{\alpha+1}} I(\theta > \beta) \tag{14}$$

$$= \frac{\alpha\beta^\alpha}{\theta^{n+\alpha+1}} I(\theta > \beta) I(\theta > \max\{x_{1:n}\}) \tag{15}$$

$$= \frac{\alpha\beta^\alpha}{\theta^{n+\alpha+1}} I(\theta > \max\{\beta, \max\{x_{1:n}\}\}) \tag{16}$$

$$= \frac{\alpha\beta^\alpha}{\theta^{n+\alpha+1}} I(\theta > \max\{\beta, x_{1:n}\}) \tag{17}$$

Announcements

- Exam I: Feb 7, in class (no make up exams)
- Exam I: closed notes
- Exam I: will answer questions regarding what is on the exam on Thursday.
- Review session will be on Tuesday Feb 5.
- Exam I: Material: TBD.

Exam I

- Closed notes, closed book.
- You will be given cover page with distributions (see Sakai under Annoucements).
- Please come on time.
- Material covers lectures (slides and written material in class), labs, and homeworks.
- Material covers: Modules 0–3.

Exam I

- To help you study, work through the practice problems from class and the ones at the end of homework 3. (Solutions on Sakai).
- 5 problems total.

Topics to Review

- Bayes Theorem (derivation)
- Derivation of posterior distributions, marginals
- How do you show the proprietry of the posterior distribution?
- Conjugate families (i.e Beta-Bernoulli, Beta-Binomial, Galenshore, etc.)
- For extra problems, please see "Some of Bayesian Methods", Chapter 1–2. All examples are worked in the text.

Topics to Review (continued)

- What is a loss function (examples)
- What is the Bayes procedure? Know how to derive it.
- Posterior risk, frequentist risk. Derivations, understand plots, know the differences of when to use these.
- Review lab concepts. Suppose I describe some data to you. What would distribution might characterize it well. What prior would be appropriate?

Topics that will not be covered on exam I

- pseudocode, R code
- integrated risk
- admissibility