Introduction to Bayesian Statistics for the Social Sciences

Peter D. Hoff

Contents

List of Figures

Preface

This book serves as a set of lecture notes for CSSS-Stat 564, a graduate level course that I have been teaching at the University of Washington. The primary audience for this course includes graduate students in the social and behavioral sciences who (a) did well in their department's graduate introductory stats sequence, and (b) are interested in statistics. Students lacking one of (a) or (b) can succeed, but lacking both is generally fatal. Graduate students in many other departments take this class. In particular, first- and second-year statistics graduate students have found this course to be a useful introduction to statistical modeling.

Chapter 1

Introduction and examples

1.1 Beliefs

Let

- d=data be a variable which ranges over the possible outcomes of a survey/experiment/study *having an uncertain outcome*.
- Let θ be a variable that ranges over some set of hypotheses/states of nature/parameter values that *you are uncertain about*.

Bayesian inference proceeds as follows:

1. For each possible value of the parameter, define a *belief measure* $B(\theta)$ which describes roughly your "belief that θ is the true value of the parameter."
    - $B(\theta_a) > B(\theta_b)$ means you would prefer to bet that θ_a is the truth rather than bet θ_b is the truth.
2. For each possible value of the parameter and each possible outcome of the study define a belief measure $B(d : \theta)$ which describes roughly your "belief that d will be the outcome of your study, if θ is true."
    - $B(d_1|\theta_a) > B(d_2|\theta_a)$ means that if you knew for sure that θ_a were true, then you would prefer to bet on d_1 occurring than d_2.
    - $B(d_1|\theta_a) > B(d_1|\theta_b)$ means that if you were forced to bet on d_1, you would prefer that θ_a be true than θ_b.

3. Run the survey/experiment/study and obtain the data d_{obs}.
4. Update your belief measure: combine your prior beliefs $B(\theta)$ with the information in the data $B(d_{\text{obs}} : \theta)$ to obtain $B(\theta : d_{\text{obs}})$.

BAYESIAN INFERENCE refers to the use of

- probabilities to represent beliefs;
- Bayes' rule to update beliefs.

Bayesian inference doesn't tell you what to believe, it just tells you how your beliefs should change in light of new data.

1.2 Examples:

1.2.1 One-sample inference

100 people sampled from a given county are asked

"do you support policy Z?"

- Let θ = % of people in the county who would say they support the policy if asked.
- Let Y = number of people sampled who support the policy.
- $Y|\theta \sim$ binomial(100,θ), approximately. (why approximately?)

$Y_{\text{obs}} = 54$. A quantity of interest: $P(\theta > .5)$.

1.2.2 Two-sample inference

- Let θ_1, θ_2 be the support in two different counties.
- Let

$$\begin{aligned} \beta &= \log \frac{\text{odds in county 2}}{\text{odds in county 1}} \\ &= \log \text{odds in county 1} - \log \text{odds in county 2} \\ &= \log \frac{\theta_2}{1-\theta_2} - \log \frac{\theta_1}{1-\theta_1} \\ &= \text{the log-odds ratio.} \end{aligned}$$

$Y_{1,\text{obs}} = 56, Y_{2,\text{obs}} = 47$. A quantity of interest: $P(\beta > 0) = P(\theta_1 > \theta_2)$.

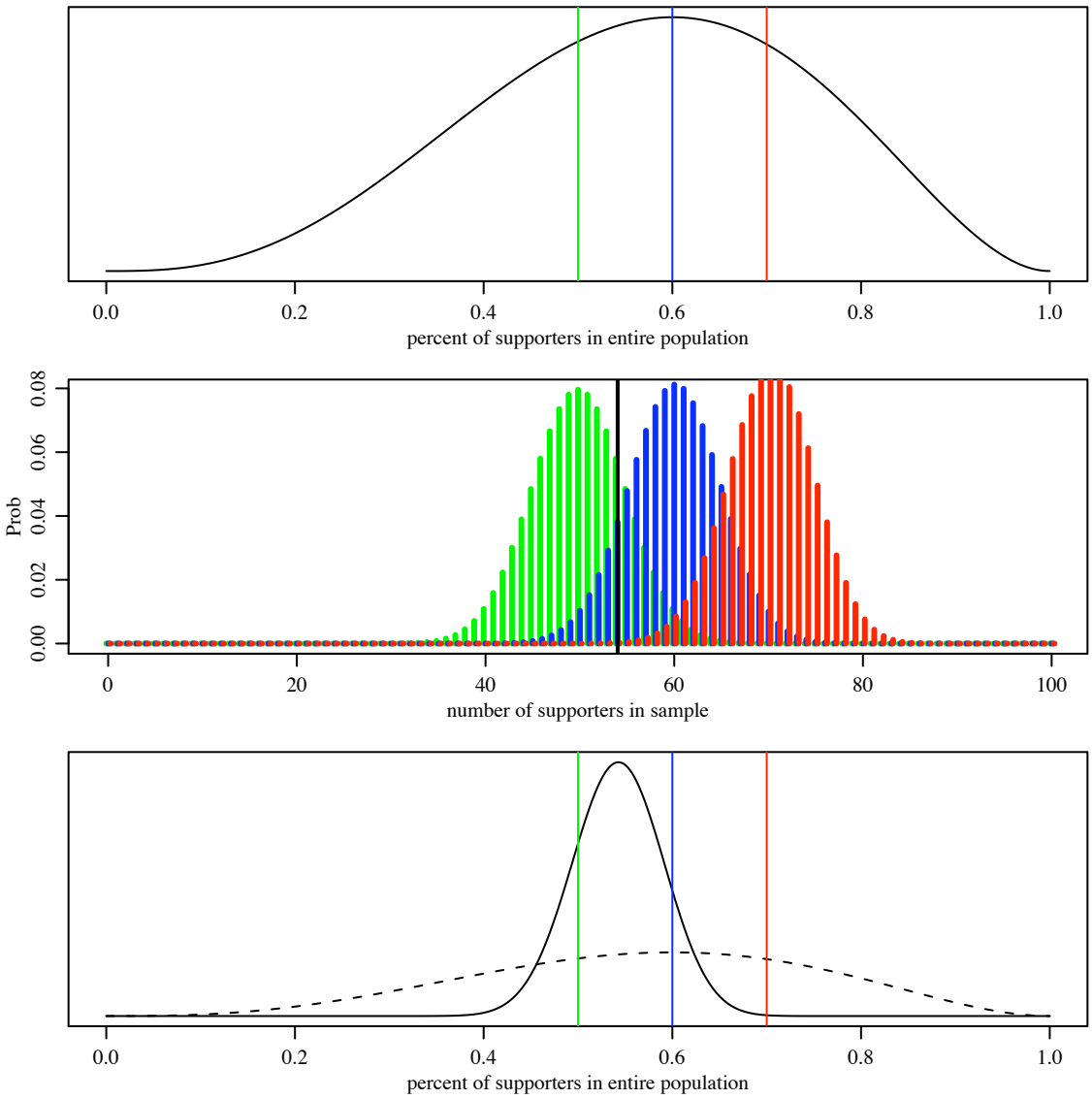


Figure 1.1: One-sample binomial model: The first panel displays a beta(4,3) prior distribution. The second panel displays the binomial sampling model for the survey outcome Y, for $\theta \in \{.5, .6, .7\}$. The third panel plots the prior distribution and the posterior distribution given $Y_{\text{obs}} = 54$.

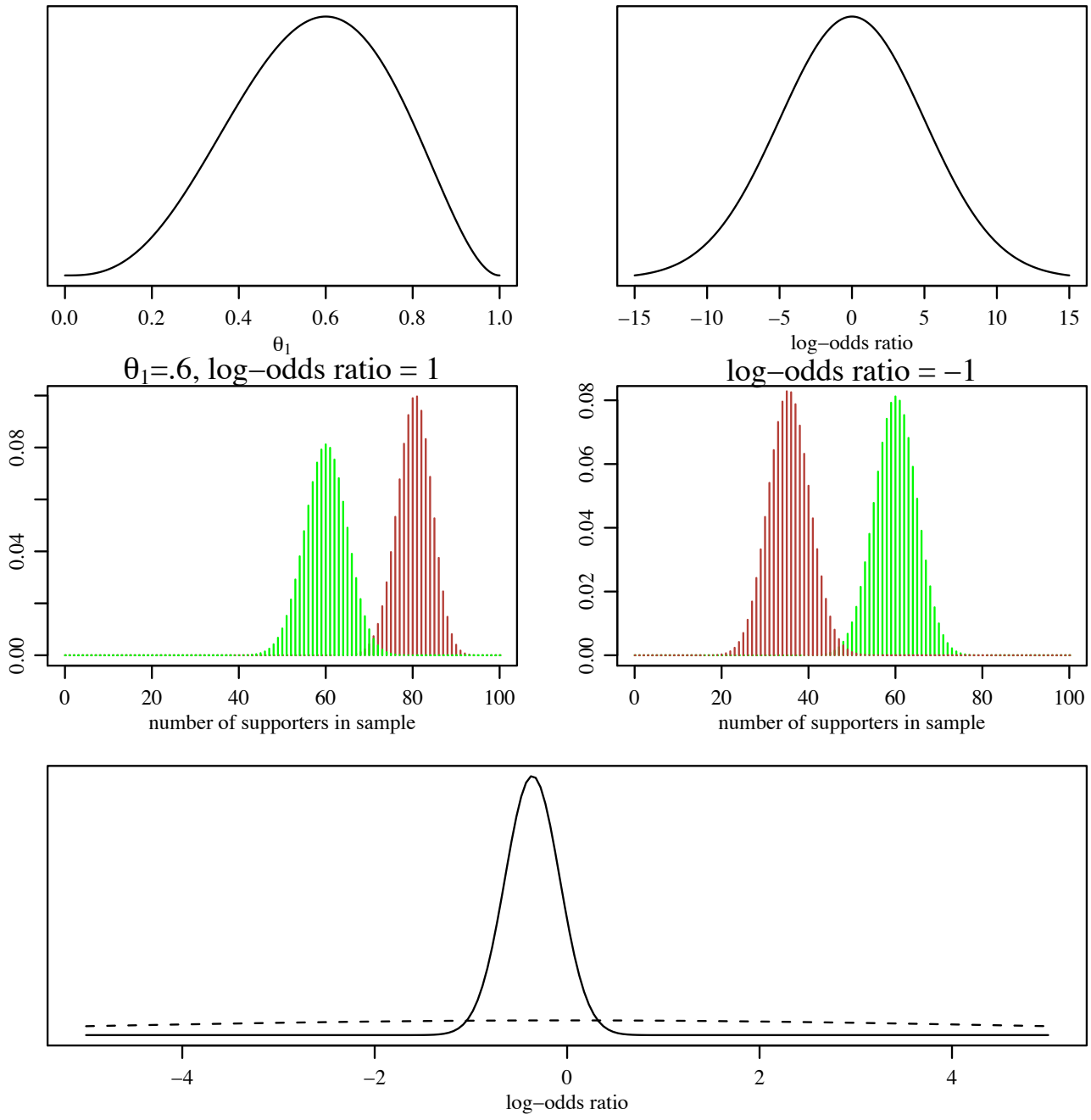


Figure 1.2: Two-sample binomial model: The first panel displays a beta(4,3) prior distribution for θ_1 , and the second shows the prior for the log-odds ratio. The second panel show the sampling models for Y_1 and Y_2 under two different sets of parameters. The third panel displays the prior and posterior distribution of the log-odds ratio given $Y_{1,\text{obs}} = 56, Y_{2,\text{obs}} = 47$.

1.2.3 Regression

Suppose each person in the sample is also asked their SES$\in (-5, -4, \ldots, 4, 5)$. Let

$$\beta = \log \frac{P(\text{support}|x+1)}{P(\text{don't support}|x+1)} - \log \frac{P(\text{support}|x)}{P(\text{don't support}|x)}$$

A quantity of interest: $P(\beta > 0)$.

1.2.4 Hierarchical data

Suppose a sample of size $n \in \{25, 50, 100\}$ was made from each of 35 counties.

Some quantities of interest:

- $P(\theta_i > \theta_j)$
- $P(\sum_{i=1}^{35} \theta_i n_i > .5 \sum_{i=1}^{35} n_i)$
- $P(\theta_{36} > .5)$

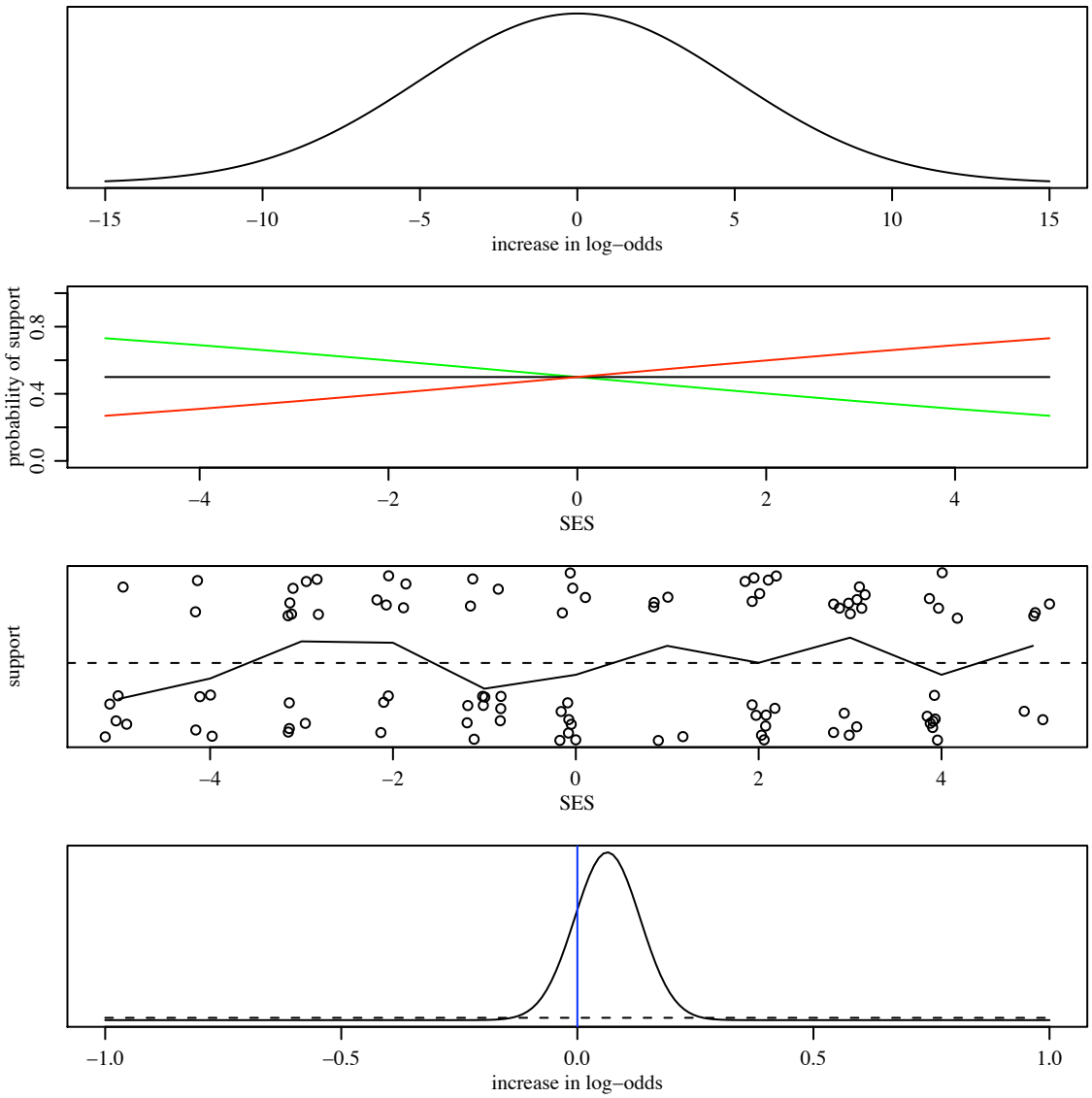


Figure 1.3: Logistic regression: The first panel is a normal prior distribution for the odds-ratio. The second panel show the increase in the probability of support as a function of SES, for $\beta_1 \in \{-.2, 0, .2\}$. The third panel shows the observed data (jittered). The fourth panel shows the posterior distribution of β_1 given the observed data.

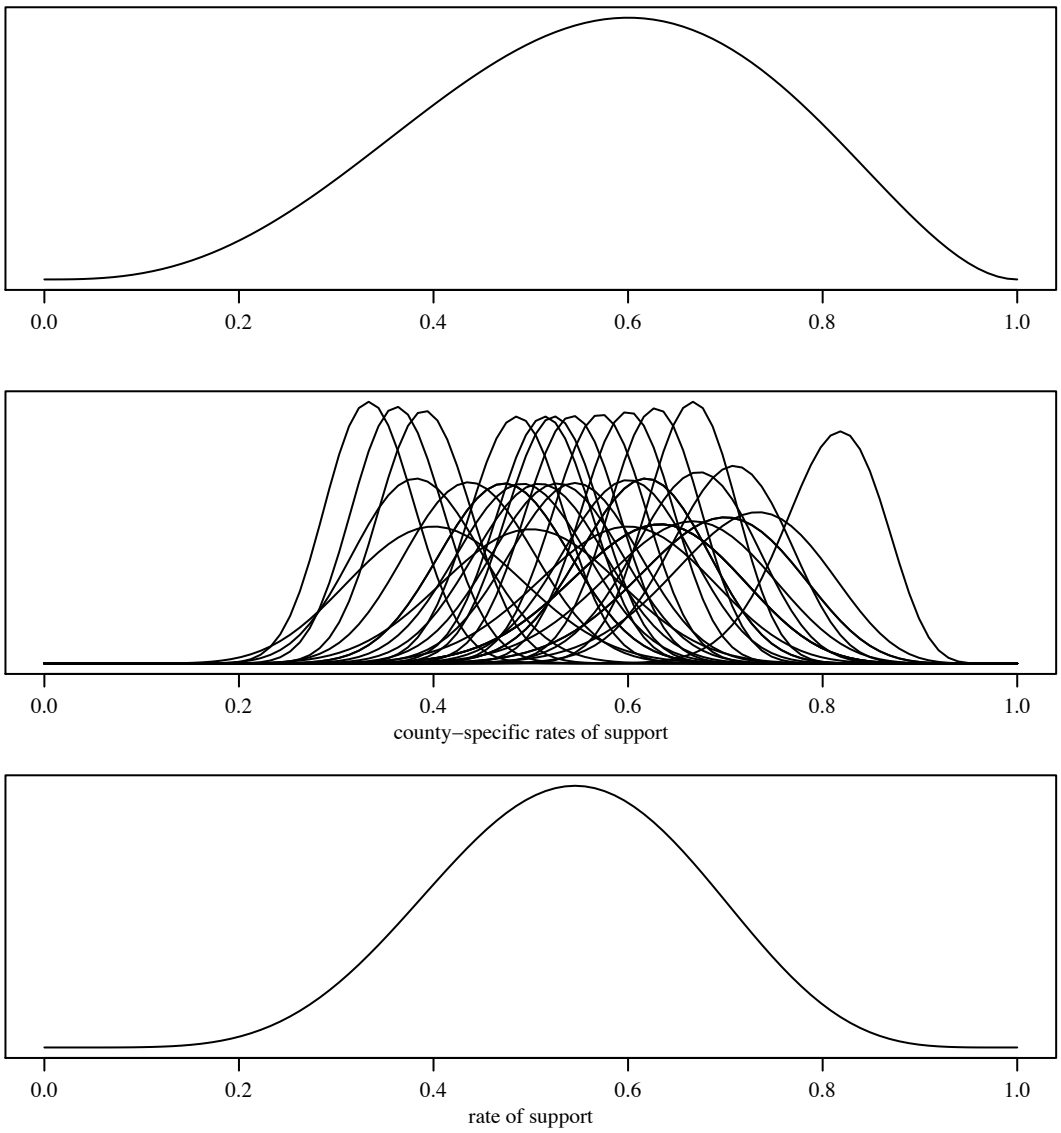


Figure 1.4: Hierarchical binary data: The first panel is a beta(4,3) prior distribution, used for each county. The second panel gives the posterior distributions, given a dataset. The third panel gives the posterior predictive sampling for θ_{36}

Chapter 2

Review of probability

2.1 Axioms of probability

Let E, F and H be (possibly overlapping) statements. Let $P(E|H)$= "a measure of your belief that E is true given that H is true," i.e. a belief measure. It may be useful for $P(\cdot|\cdot)$ to have the following properties:

P1 : $P(E|H) \geq 0$ for all statements E.

P2 : $P(H|H) = 1$ for all statements H.

P3 : $P(E \cup F|H) = P(E|H) + P(F|H)$ if $E \cap F = \emptyset$.

P4 : $P(E \cap F|H) = P(F|H)P(E|F \cap H)$.

Why?

- **P1** and **P2**: Clearly we want $p_{\min} \leq P(E|H) \leq P(H|H) \leq p_{\max}$ for all sets E and H. Setting $p_{\min} = 0$ and $p_{\max} = 1$ are arbitrary but convenient choices: It allows for a correspondence between empirical proportions and beliefs about empirical proportions.
- **P3**: Clearly we want $P(E \cup F|H) \geq P(E|H)$ *and* $P(E \cup F|H) \geq P(F|H)$. It also makes sense to have $P(E \cup F|H) = P(F \cup E|H)$. With the inclusion of **P1**, **P2** and an arbitrary scaling assumption, Jaynes (Chapter 2) shows that **P3** must hold.
- **P4**: Can also be derived from thinking logically about what properties a belief measure should have. For now think:

"My belief that E and F are true is a function of my belief that F is true my belief that E is true given F."

We call $P(A|B)$ "the conditional probability of A given B" for any sets A and B.

2.2 Partitions of events and Bayes' rule

Definition 1 (Partition of events) *A collection of events* $\{H_1, \ldots, H_K\}$ *is a partition if*

- $\{H_1, \ldots, H_K\}$ *are disjoint, meaning* $H_i \cap H_j = \emptyset$ *if* $i \neq j$*, and*
- *At least one of the* H_k*'s must be true, meaning* $P(\cup_{k=1}^K H_k) = 1$.

In other words, $\{H_1, \ldots, H_K\}$ is a partition if one and only one H_k can be true.

Examples:

- GSS religion categories
  - { Protestant, Catholic, Jewish, None, Other }
  - { Christian, non-Christian}
  - { monotheist, multitheist, atheist}
-

Using partitions From the axioms of probability, one can show

1. Total probability: $\sum_{k=1}^K P(H_k) = 1$;
2. Marginal probability: $P(E) = \sum_{k=1}^K P(E \cap H_k)$;
3. Bayes rule:

$$\begin{aligned} P(H_j|E) &= \frac{P(E|H_j)P(H_j)}{P(E)} \\ &= \frac{P(E|H_j)P(H_j)}{\sum_{k=1}^K P(E|H_k)P(H_k)} \end{aligned}$$

In particular, let $H_1, \ldots, H_K$ be various disjoint hypothesis, and let E be an observed event or dataset.

$$\begin{aligned} \frac{P(H_i|E)}{P(H_j|E)} &= \frac{P(E|H_i)P(H_i)}{P(E|H_j)P(H_j)} \\ &= \frac{P(E|H_i)}{P(E|H_j)} \times \frac{P(H_i)}{P(H_j)} \\ &= \text{“Bayes factor” } \times \text{ “prior beliefs”} \end{aligned}$$

Recall interpretation of $P(E|H_i) > P(E|H_j)$ from beginning. The Bayes factor determines how our beliefs get updated in light of the data.

KEEP IN MIND: Bayes rule

- does not determine what beliefs one should have in light of the data,
- does determine how pre-existing beliefs should be *updated* in light of the data.

2.3 Independence

Definition 2 (Independence) *Two events E and F are conditionally independent given H if $P(E \cap F|H) = P(E|H)P(F|H)$.*

Interpretation via conditional probability: By Axiom **P3**, the following is always true:

$$P(E \cap F|H) = P(F|H)P(E|H \cap F).$$

If E,F are conditionally independent, then we must have

$$\begin{aligned} P(F|H)P(E|H \cap F) &\overset{always}{=} P(E \cap F|H) \overset{ind}{=} P(E|H)P(F|H) \\ P(F|H)P(E|H \cap F) &= P(E|H)P(F|H) \\ P(E|H \cap F) &= P(E|H) \end{aligned}$$

So conditional independence implies $P(E|H \cap F) = P(E|H)$. In other words, given that we know H, knowing F adds no information about E.

2.4 Random variables

For our purposes, a random variable is the quantitative outcome of a survey/experiment/study. Let Y be such an outcome, and let $\mathcal{Y}$ be the set of all possible outcomes.

2.4.1 Discrete random variables

$\mathcal{Y} = \{y_1, y_2, \ldots\}$. The set of possible outcomes is *countable*.

Examples:

- Y = years of education;
-

Probability distribution/density: How do we represent our beliefs about Y? For each $y \in \mathcal{Y}$, define $p(y) = P(\{Y = y\})$. Then the function $p(\cdot) : \mathcal{Y} \to [0, 1]$ is called a probability density for Y, and has the following properties:

- $0 \leq p(y) \leq 1$ for all $y \in \mathcal{Y}$;
- $\sum_{y \in \mathcal{Y}} p(y) = 1$;
- if A and B are disjoint subsets of $\mathcal{Y}$, then

$$\begin{aligned} P(Y \in A \text{ or } Y \in B) \equiv P(Y \in A \cup B) &= P(Y \in A) + P(Y \in B) \\ &= \sum_{y \in A} p(y) + \sum_{y \in B} p(y) \end{aligned}$$

Example: Binomial distribution. In a sample of n units, let Y = the number having a given characteristic. Let $\mathcal{Y} = \{0, 1, \ldots, n\}$. See previous and next chapter for figures and details.

Example: Poisson distribution. Let $\mathcal{Y} = \{0, 1, 2, \ldots\}$. Our beliefs about $Y \in \mathcal{Y}$ follow the *Poisson model with mean* θ if

$$P(Y = y|\theta) = \theta^y e^{-\theta}/y!$$

For example, if $\theta = 2.1$,

$$\begin{array}{rll} P(Y = 0|\theta = 2.1) = & (2.1)^0 e^{-2.1}/(0!) = & .12 \\ P(Y = 1|\theta = 2.1) = & (2.1)^1 e^{-2.1}/(1!) = & .26 \\ P(Y = 2|\theta = 2.1) = & (2.1)^2 e^{-2.1}/(2!) = & .27 \\ P(Y = 3|\theta = 2.1) = & (2.1)^3 e^{-2.1}/(3!) = & .19 \\ \vdots & \vdots & \vdots \end{array}$$

Maximum entropy justification: The Poisson(θ) distribution is the "least informative/most spread-out" (maximum entropy) distribution on $\mathcal{Y}$ having mean θ.

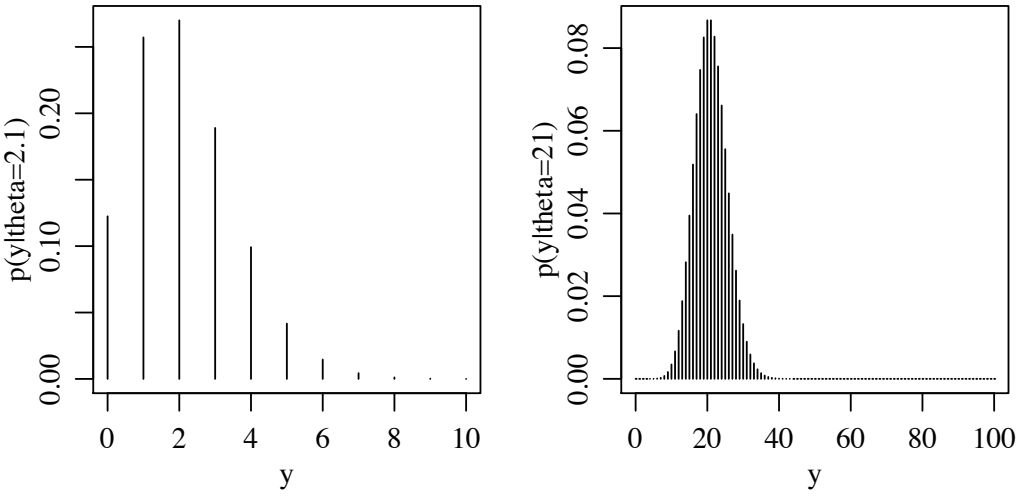


Figure 2.1: Poisson distributions with means of 2.1 and 21.

2.4.2 Continuous random variables:

Suppose to a rough approximation $\mathcal{Y} = \mathbb{R}$, the set of all real numbers (could this really ever be the case?)

Then we can't define something like $P(Y \leq 5) = \sum_{y \leq 5} p(y)$ because the sum doesn't make sense. Instead, we can start with the cumulative distribution function, or CDF:

$$F(y) = P(Y \leq y)$$

Note that $F(\infty) = 1$, $F(-\infty) = 0$, and $F(b) \leq F(a)$ if $b < a$ (why?). From the CDF, a wide variety of probability assessments can be derived:

- $P(Y > a) = 1 - F(y)$
- $P(b < Y \leq a) = F(a) - F(B)$

A theorem in real analysis that says there exists a positive function $p(y)$ such that

$$F(a) = \int_{-\infty}^{a} p(y)\, dy$$

This function is called the probability density function of Y. It behaves similarly to pdf in the case of discrete random variables:

- $0 \leq p(y)$ for all $y \in \mathcal{Y}$ (why not ≤ 1?)
- $\int_{y \in \mathbb{R}} p(y)\ dy = 1$;
- if A and B are disjoint subsets of $\mathcal{Y}$, then

$$\begin{aligned} P(Y \in A \text{ or } Y \in B) \equiv P(Y \in A \cup B) &= P(Y \in A) + P(Y \in B) \\ &= \int_{y \in A} p(y)\ dy + \int_{y \in B} p(y)\ dy \end{aligned}$$

Example: The normal distribution Suppose we knew we were sampling from a population on $\mathcal{Y} = (-\infty, \infty)$, and that the mean of the population was μ and the standard deviation was σ. The most diffuse distribution with such properties is the normal(μ, σ) distribution, with CDF

$$P(Y \leq y | \mu, \sigma) = F(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \left(\frac{y - \mu}{\sigma} \right)^2 \right\}\ dy$$

Evidently,

$$p(y | \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \left(\frac{y - \mu}{\sigma} \right)^2 \right\}$$

Letting $\mu = 10.9, \sigma = .8$ gives the CDF and density in Figure 2.2. This mean and standard deviation make the median value of e^Y equal to about 54 thousand, the median income in King County.

2.5 Joint Distributions

Discrete distributions Let

- $\mathcal{Y}_1, \mathcal{Y}_2$ be two countable sample spaces.
- $Y_1 \in \mathcal{Y}_1$, $Y_2 \in \mathcal{Y}_2$ be two random variables.

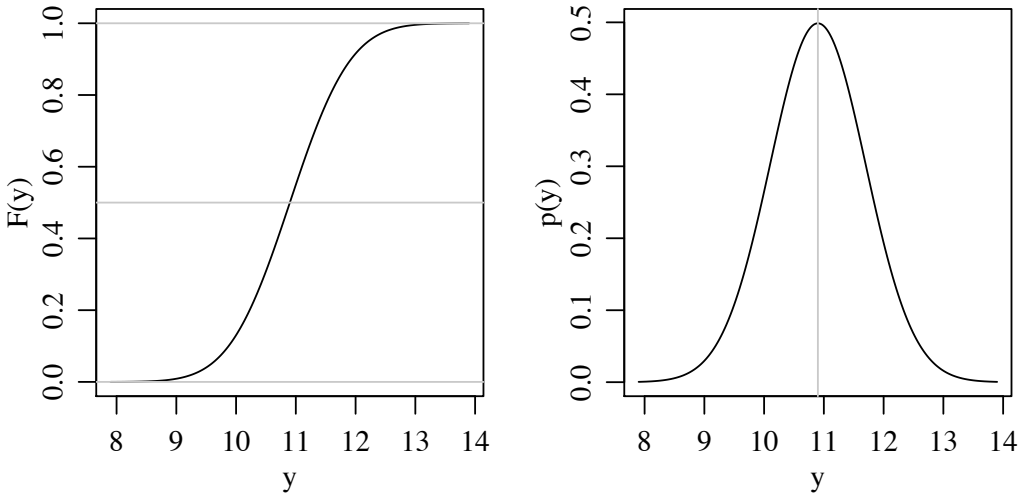


Figure 2.2: Normal distribution with mean 10.9 and standard deviation 0.8

Beliefs about Y_1, Y_2 are represented with probabilities, i.e. $P(\{Y_1 \in A\} \cap \{Y_2 \in B\})$. The joint pdf is defined as

$$p_{Y_1,Y_2}(y_1, y_2) = P(\{Y_1 = y_1\} \cap \{Y_2 = y_2\}), \text{ for } y_1 \in \mathcal{Y}_1,\ y_2 \in \mathcal{Y}_2.$$

The **marginal density** of y_1 can be computed from

$$\begin{aligned} p_{Y_1}(y_1) &\equiv P(\{Y_1 = y_1\}) \\ &= \sum_{y_2 \in \mathcal{Y}_2} P(\{Y_1 = y_1\} \cap \{Y_2 = y_2\}) \\ &\equiv \sum_{y_2 \in \mathcal{Y}_2} p_{Y_1,Y_2}(y_1, y_2) \end{aligned}$$

The **conditional density** of y_1 given y_2 can be computed from

$$\begin{aligned} p_{Y_1|Y_2}(y_1|y_2) &= \frac{P(\{Y_1 = y_1\} \cap \{Y_2 = y_2\})}{P(Y_2 = y_2)} \\ &= \frac{p_{Y_1,Y_2}(y_1, y_2)}{p_{Y_2}(y_2)} \end{aligned}$$

Convince yourself that

- $\{p_{Y_1}, p_{Y_2|Y_1}\}$ can be derived from p_{Y_1,Y_2};

- $\{p_{Y_2}, p_{Y_1|Y_2}\}$ can be derived from p_{Y_1,Y_2};
- p_{Y_1,Y_2} can be derived from $\{p_{Y_1}, p_{Y_2|Y_1}\}$;
- p_{Y_1,Y_2} can be derived from $\{p_{Y_2}, p_{Y_1|Y_2}\}$;

and

- p_{Y_1,Y_2} cannot be derived from $\{p_{Y_1}, p_{Y_2}\}$

Example: Social Mobility (Data from Logan, 1983)

		son's occ				
		farm	operatives	craftsmen	sales	professional
father's occ	farm	0.018	0.035	0.031	0.008	0.018
	operatives	0.002	0.112	0.064	0.032	0.069
	craftsmen	0.001	0.066	0.094	0.032	0.084
	sales	0.001	0.018	0.019	0.01	0.051
	professional	0.001	0.029	0.032	0.043	0.13

Let Y_1 = fathers occupation, Y_2 = son's.

$$\begin{aligned} P(Y_2 = \text{professional}|Y_1 = \text{farm}) &= \frac{P(Y_2 = \text{professional} \cap Y_1 = \text{farm})}{P(Y_1 = \text{farm})} \\ &= \frac{.018}{.018 + .035 + .031 + .008 + .018} \\ &= .164 \end{aligned}$$

Continuous joint distributions: Given a joint CDF $F_{Y_1,Y_2}(a,b) \equiv P(\{Y_1 \leq a\} \cap \{Y_2 \leq b\})$, there is a function p_{Y_1,Y_2} such that

$$F_{Y_1,Y_2}(a,b) = \int_{-\infty}^{a} \int_{-\infty}^{b} p_{Y_1,Y_2}(y_1,y_2)\ dy_1 dy_2$$

p_{Y_1,Y_2} is the joint density of Y_1, Y_2. As in the discrete case, we have

- $p_{Y_1}(y_1) = \int_{-\infty}^{\infty} p_{Y_1,Y_2}(y_2,y_2)\ dy_2$;
- $p_{Y_2|Y_1}(y_2|y_1) = p_{Y_1,Y_2}(y_1,y_2)/p_{Y_1}(y_1)$.

Convince yourself that $p_{Y_2|Y_1}(y_2|y_1)$ is an actual probability density, i.e. it satisfies the conditions described above.

Mixed continuous and discrete: Let Y_1 be continuous, Y_2 discrete. (Examples ...). Suppose we define

- a marginal density p_{Y_1} from $F(y_1) = P(Y_1 \leq y_1)$ as above;
- a conditional density $p_{Y_2|Y_1}(y_2|y_1)$ from $P(Y_2 = y_2|Y_1 = y_1)$ as above.

Then the joint density of Y_1 and Y_2 is

$$p_{Y_1,Y_2}(y_1, y_2) = p_{Y_1}(y_1) \times p_{Y_2|Y_1}(y_2|y_1),$$

and is such that

$$P(Y_1 \in A, Y_2 \in B) = \sum_{y_2 \in B} \int_{y_1 \in A} p_{Y_1,Y_2}(y_1, y_2)\ dy_1$$

Example:

- θ = nation-wide presidential approval rating.
- Y = number of people in a sample of size 100 who indicate approval of the president's job performance.

Consider how to construct a reasonable $p(y, \theta)$.

Bayes Theorem: Let θ and Y be two random variables. Often it is natural to construct the joint belief model $p(y, \theta)$ from

- $p(\theta)$, beliefs about θ;
- $p(y|\theta)$, beliefs about Y, for each value of θ.

Having observed y, we need to compute our updated beliefs about θ:

$$p(\theta|y) = p(\theta, y)/p(y) = p(\theta)p(y|\theta)/p(y)$$

The probability (density) of θ_1 relative to θ_2, conditional on $Y = y$, is

$$\begin{aligned} \frac{p(\theta_1|y)}{p(\theta_2|y)} &= \frac{p(\theta_1)p(y|\theta_1)/p(y)}{p(\theta_2)p(y|\theta_2)/p(y)} \\ &= \frac{p(\theta_1)p(y|\theta_1)}{p(\theta_2)p(y|\theta_2)} \end{aligned}$$

This means that, to evaluate the relative probabilities of θ_1 and θ_2, we don't need to compute $p(y)$. Another way to think about it is that, as a function of θ,

$$p(\theta|y) \propto p(\theta)p(y|\theta).$$

The constant of proportionality is $p(y)$ which *could* be computed as

$$p(y) = \int_\Theta p(y, \theta)\ d\theta = \int_\Theta p(y|\theta)p(\theta)\ d\theta$$

giving

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{\int_\theta p(\theta)p(y|\theta)\ d\theta}.$$

But, as we will see, the numerator is the important part.

2.6 Independent random variables

Because we will use it later, consider the following setup:

- $Y_1, \ldots, Y_n$ are all random variables on $\mathcal{Y}$;
- Let θ be some piece of information (vague, I know...).

$Y_1, \ldots, Y_n$ are conditionally independent given θ if for every set of n subsets $\{A_1, \ldots, A_n\}$ of $\mathcal{Y}$, we have

$$P(Y_1 \in A_1, \ldots, Y_n \in A_n|\theta) = P(Y_1 \in A_1|\theta) \times \cdots P(Y_n \in A_n|\theta)$$

Note this corresponds with our previous definition of independent events. As before, if independence holds,

$$P(Y_i \in A_i|\theta, Y_j \in A_j) = P(Y_i \in A_i|\theta),$$

so independence can be interpreted as meaning that, given θ, knowing Y_j gives no additional information about Y_i.

If independence holds then the joint density is

$$p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} p_{Y_i}(y_i|\theta),$$

i.e. the product of the marginal densities. If all marginal densities are the same, we write

$$p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta),$$

and say the Y_i's are **conditionally independent and identically distributed** (i.i.d.).

2.7 Means, medians modes, variances, covariances, interquartile ranges.

Hopefully the following is very familiar to you. If not, please consult your favorite introductory stats textbook.

The mean of an unknown quantity X is given by

- $E(X) = \sum_{x \in \mathcal{X}} xp(x)$ if X is discrete;
- $E(X) = \int_{x \in \mathcal{X}} xp(x)\ dx$ if X is continuous;

The mean is the center of mass of the distribution. However, it is generally not equal to either of

- "the most likely value of X" (the **mode**);
- 'the value of X in the middle of of the distribution (the **median**).

Some justifications for reporting/studying the mean include

- The mean of $Y_1, \ldots, Y_n$ is a scaled version of the total.
- Suppose you are forced to guess what the value of Y is, and you are penalized an amount $(Y - Y_{guess})^2$. Then guessing $E(Y)$ minimizes your expected value of your penalty.

Discuss: Y_i=income of person i in King county.

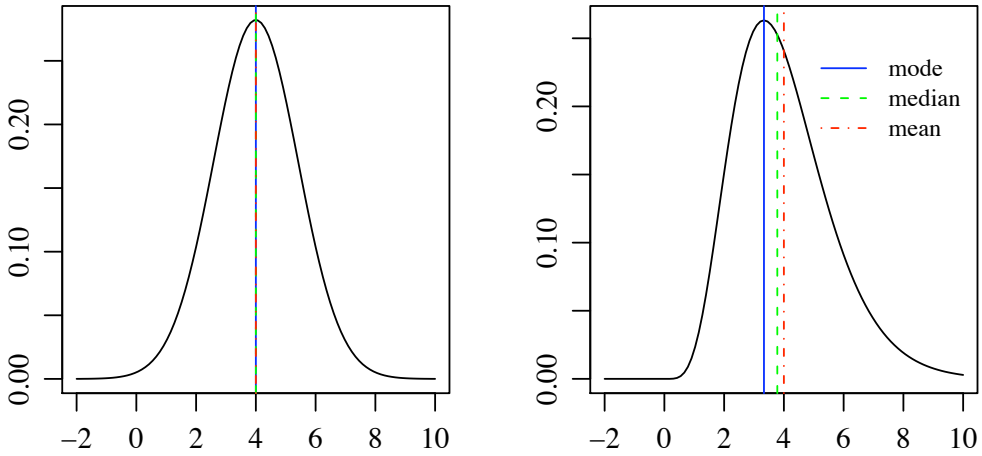


Figure 2.3: Mode, median and mean

2.8 Putting it all together: Exchangeability

Example: Recidivism rates. Suppose 10 juvenile offenders are randomly sampled for a study. Let

$$Y_i = \begin{cases} 1 & \text{if subject } i \text{ re-offends within one year,} \\ 0 & \text{otherwise.} \end{cases}$$

Now lets consider the structure of our joint beliefs about $Y_1, \ldots, Y_{10}$: Let $P(Y_1 = y_1, \ldots, Y_{10} = y_{10}) = p(y_1, \ldots, y_{10})$.

Exchangeability

$$\begin{aligned} p(1,0,0,1,0,1,1,0,1,1) &= ? \\ p(1,0,1,0,1,1,0,1,1,0) &= ? \end{aligned}$$

Definition 3 (Exchangeable) *Let $p(y_1, \ldots, y_n)$ be the joint density of $Y_1, \ldots, Y_n$. If $p(y_1, \ldots, y_n) = p(y_{\pi 1}, \ldots, y_{\pi n})$ for all permutations π of $\{1, \ldots, n\}$ then $Y_1, \ldots, Y_n$ are* **exchangeable**.

Independence/dependence: Let

$$\begin{aligned} P(Y_{10} = 1) &= a \\ P(Y_{10} = 1 | Y_1 = 0, Y_2 = 0, \ldots, Y_8 = 0, Y_9 = 0) &= b \end{aligned}$$

Should we have $a < b$, $a = b$, or $a > b$? If $a \neq b$ then Y_{10} is NOT independent of $Y_1, \ldots, Y_9$.

Conditional independence: Suppose another researcher had sampled 1000 juvenile offenders from the same population and found the recidivism rate to be θ.

$$\begin{aligned} P(Y_{10} = 1|\theta) &\stackrel{?}{\approx} \theta \\ P(Y_{10} = 1|Y_1, \ldots, Y_9, \theta) &\stackrel{?}{\approx} \theta \\ P(Y_9 = 1|Y_1, \ldots, Y_8, Y_{10}, \theta) &\stackrel{?}{\approx} \theta \end{aligned}$$

We might then view the Y_i's as conditionally independent and identically distributed given θ. That is, given θ, the Y_i's are i.i.d. . In this case,

$$\begin{aligned} P(Y_i = y_i|\theta, Y_j, j \neq i) &= \theta^{y_i}(1-\theta)^{1-y_i} \\ P(Y_1 = y_1, \ldots, Y_n = y_n|\theta) &= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^{\sum y_i}(1-\theta)^{n-\sum y_i} \end{aligned}$$

If θ is uncertain to us, we describe our beliefs with $p(\theta)$. The marginal joint distribution of $Y_1, \ldots, Y_n$ is then :

$$p(y_1, \ldots, y_n) = \int_0^1 p(y_1, \ldots, y_n|\theta)p(\theta)\ d\theta = \int_0^1 \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}p(\theta)\ d\theta$$

Check exchangeability:

- $p(1,0,0,1,1,0,1,1,0,1) = \int \theta^6(1-\theta)^4 p(\theta)\ d\theta$
- $p(1,1,0,0,1,0,0,1,1,1) = \int \theta^6(1-\theta)^4 p(\theta)\ d\theta$

It looks like $Y_1, \ldots, Y_n$ are exchangeable under this model/set of beliefs.

Claim: If $\theta \sim p(\theta)$ and $Y_1, \ldots, Y_n|\theta$ are conditionally i.i.d. given θ, then marginally/unconditional on θ, $Y_1, \ldots, Y_n$ are exchangeable.

Proof: For general densities,

$$\begin{aligned}
p(y_1, \ldots, y_n) &= \int p(y_1, \ldots, y_n|\theta)p(\theta)\ d\theta && \text{(def of marginal probability)} \\
&= \int \left\{ \prod_{i=1}^{n} p(y_i|\theta) \right\} p(\theta)\ d\theta && \text{(by cond ind)} \\
&= \int \left\{ \prod_{i=1}^{n} p(y_{\pi i}|\theta) \right\} p(\theta)\ d\theta && \text{(product doesn't depend on order)} \\
&= p(y_{\pi_1}, \ldots y_{\pi_n}) && \text{(def of marginal probability)}
\end{aligned}$$

Roughly speaking, $Y_1, \ldots, Y_n$ are exchangeable if the labels carry no information about the outcomes.

de Finetti's theorem: We've seen that

$$\left.\begin{array}{l} Y_1, \ldots, Y_n|\theta \text{ i.i.d} \\ \theta \sim p(\theta) \end{array}\right\} \Rightarrow Y_1, \ldots, Y_n \text{ are exchangeable.}$$

What about an arrow in the other direction?

Let $\{Y_1, \ldots\}$ be a potentially infinite sequence of random variables all having a common sample space $\mathcal{Y}$.

Theorem 1 (de Finetti) *Let $Y_i \in \mathcal{Y}$ for all i. Suppose that, for any n,* **your model** *for $Y_1, \ldots, Y_n$ is exchangeable:*

$$p(y_1, \ldots, y_n) = p(y_{\pi 1}, \ldots, y_{\pi_n})$$

for all permutations π of $1, \ldots, n$. Then **your model** *can be written as*

$$p(y_1, \ldots, y_n) = \int \left\{ \prod_{1}^{n} p(y_i|\theta) \right\} p(\theta)\ d\theta$$

for some parameter θ, some (prior) distribution on θ, and some (sampling) distribution $p(y|\theta)$.

$p(\theta)$ represents **your belief** about the outcomes of $Y_1, Y_2, \ldots$, induced by your belief model $p(y_1, y_2, \ldots)$. More precisely,

- $p(\theta)$ represents your belief about $\lim_{n\to\infty} \sum Y_i/n$ in the binary case.
- $p(\theta)$ represents your belief about $\lim_{n\to\infty} \sum(Y_i \leq c)/n$ for each c in the general case.

So to summarize,

$$\left.\begin{array}{l} Y_1, \ldots, Y_n|\theta \text{ i.i.d} \\ \theta \sim p(\theta) \end{array}\right\} \Leftrightarrow Y_1, \ldots, Y_n \text{ are exchangeable for all } n$$

The condition "$Y_1, \ldots, Y_n$ are exchangeable for all n " is reasonable if

- $Y_1, \ldots, Y_n$ are outcomes of a repeatable experiment, or
- $Y_1, \ldots, Y_n$ are sampled from a potentially infinite population without replacement, or
- $Y_1, \ldots, Y_n$ are sampled from a finite population with replacement;

de Finetti's theorem holds approximately when

- $Y_1, \ldots, Y_n$ are sampled from a finite population without replacement where $n << N$.

Chapter 3

The binomial model

3.1 Recidivism example

Suppose a study will sample $n = 43$ juvenile offenders from a large population and follow them for one year. If our beliefs about the juveniles are exchangeable, then by de Finetti's theorem our joint beliefs about $Y_1, \ldots, Y_{43}$ can be decomposed into

- our beliefs about $\theta = \lim_{n\to\infty} \sum_{i=1}^{n} Y_i/n$;
- the belief that, given θ, the Y_i's are i.i.d. binary with mean θ.

so our probability for any potential outcome $\{y_1, \ldots, y_{43}\}$, conditional on θ, is given by

$$P(y_1, \ldots, y_{43}|\theta) = \theta^{\sum_{i=1}^{43} y_i}(1-\theta)^{43-\sum_{i=1}^{43} y_i}$$

A uniform prior: Suppose our prior density for θ is such that

$$P(a \le \theta \le b) = P(a + c \le \theta \le b + c), \qquad 0 \le a < b < b + c \le 1$$

i.e. all intervals of the same length have the same probability. Then our density for θ is the uniform density:

$$p(\theta) = 1 \text{ for all } \theta \in [0, 1].$$

In this case, Bayes' rule says:

$$\begin{aligned} p(\theta|y_1,\ldots,y_{43}) &= \frac{p(\theta)p(y_1,\ldots,y_{43}|\theta)}{p(y_1,\ldots,y_{43})} \\ &\propto p(\theta)p(y_1,\ldots,y_{43}|\theta) \\ &= 1 \times p(y_1,\ldots,y_{43}|\theta) \end{aligned}$$

So the posterior density of θ, as a function of θ, has the same shape (is proportional to)

$$p(y_1,\ldots,y_{43}|\theta) = \theta^{\sum_{i=1}^{43} y_i}(1-\theta)^{43-\sum_{i=1}^{43} y_i}$$

Data:

- 43 individuals surveyed;
- 15 individuals reoffend (34.9%);
- 28 individuals do not reoffend (65.1%);

$$\begin{aligned} p(y_1,\ldots,y_{43}|\theta) &= \theta^{15}(1-\theta)^{28} \text{ , so} \\ p(\theta|y_1,\ldots,y_{43}) &= \theta^{15}(1-\theta)^{28} \times p(\theta)/p(y_1,\ldots,y_{43}) \\ &= \theta^{15}(1-\theta)^{28} \times 1/p(y_1,\ldots,y_{43}) \end{aligned}$$

It turns out we can calculate the "normalizing constant" $1/p(y_1,\ldots,y_{43})$ using the following result from calculus:

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\, d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

The value of $\Gamma(x)$ for any number $x > 0$ can be looked up in a table, or with R using the `gamma()` function. How does this help us find $p(\theta|y_1,\ldots,y_{43})$?

We know the following:

- $p(\theta|y_1,\ldots,y_{43}) = \theta^{15}(1-\theta)^{28}/p(y_1,\ldots,y_{43})$;
- $\int_0^1 p(\theta|y_1,\ldots,y_{43})\, d\theta = 1$.

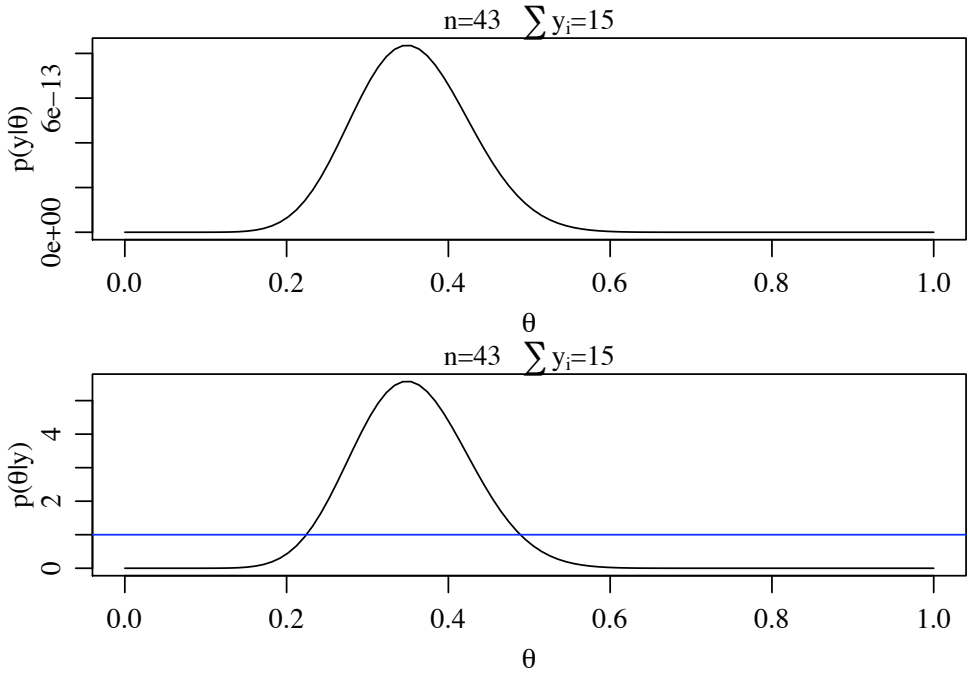


Figure 3.1: Binomial likelihood and posterior. Note that in the case of a uniform prior distribution, the posterior is proportional to the likelihood.

Therefore,

$$\begin{aligned} 1 &= \int_0^1 p(\theta|y_1,\ldots,y_{43})\ d\theta \\ &= \int_0^1 \theta^{15}(1-\theta)^{28}/p(y_1,\ldots,y_{43})\ d\theta \\ &= \frac{1}{p(y_1,\ldots,y_{43})}\int_0^1 \theta^{15}(1-\theta)^{28}\ d\theta \\ &= \frac{1}{p(y_1,\ldots,y_{43})}\frac{\Gamma(16)\Gamma(29)}{\Gamma(45)}, \quad \text{and so} \\ p(y_1,\ldots,y_{43}) &= \frac{\Gamma(16)\Gamma(29)}{\Gamma(45)} \end{aligned}$$

And finally,

$$\begin{aligned} p(\theta|y_1,\ldots,y_{43}) &= \frac{\Gamma(45)}{\Gamma(16)\Gamma(29)}\theta^{15}(1-\theta)^{28}\ \text{, which we often write} \\ &= \frac{\Gamma(45)}{\Gamma(16)\Gamma(29)}\theta^{16-1}(1-\theta)^{29-1} \end{aligned}$$

This density for θ is called a beta distribution with parameters $a = 16$ and $b = 29$, which can be calculated/plotted/sampled from in R, using the functions `dbeta(), rbeta()`.

Beta distribution: θ has a beta(a, b) distribution if

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

in which case

- $\text{Mode}(\theta) = \frac{a-1}{(a-1)+(b-1)}$
- $\text{E}(\theta) = \frac{a}{a+b}$
- $\text{V}(\theta) = \frac{ab}{(a+b+1)(a+b)^2} = E(\theta) \times E(1-\theta) \times \frac{1}{a+b+1}$

Some posterior quantities for these data: Given $\sum_{i=1}^{43} Y_i = 15$,

- $\text{Mode}(\theta|Y_i,\ldots,Y_{43}) = 0.349$;
- $\text{E}(\theta|Y_1,\ldots,Y_{43}) = 0.356$;
- $\text{SD}(\theta|Y_1,\ldots,Y_{43}) = 0.071$;

3.2 Inference for exchangeable binary data

3.2.1 Inference under a uniform prior

If $Y_1, \ldots, Y_n|\theta$ are i.i.d. binary(θ), we showed that

$$p(\theta|y_1, \ldots, y_n) = \theta^{\sum y_i}(1-\theta)^{n-\sum y_i} \times p(\theta)/p(y_1, \ldots, y_n)$$

and so, if we were to compare the relative probabilities of two θ-values, we see that

$$\begin{aligned} \frac{p(\theta_a|y_1, \ldots, y_n)}{p(\theta_b|y_1, \ldots, y_n)} &= \frac{\theta_a^{\sum y_i}(1-\theta_a)^{n-\sum y_i} \times p(\theta_a)/p(y_1, \ldots, y_n)}{\theta_b^{\sum y_i}(1-\theta_b)^{n-\sum y_i} \times p(\theta_b)/p(y_1, \ldots, y_n)} \\ &= \left(\frac{\theta_a}{\theta_b}\right)^{\sum y_i} \left(\frac{1-\theta_a}{1-\theta_b}\right)^{n-\sum y_i} \frac{p(\theta_a)}{p(\theta_b)} \end{aligned}$$

This shows that

the relative probability of θ_a to θ_b depends on $Y_1, \ldots, Y_n$ only through $\sum_{i=1}^n Y_i \Leftrightarrow$

$\sum_{i=1}^n Y_i$ contains all the information about θ in the data $\Leftrightarrow$

$\sum_{i=1}^n Y_i$ is a *sufficient statistic* for $\theta, p(y_1, \ldots, y_n|\theta)$.

Since $\sum Y_i$ is sufficient, we only need to model the relationship between θ and $\sum Y_i$.

Binomial distribution: If $Y_1, \ldots, Y_n|\theta$ are i.i.d. binary(θ) then $Y = \sum_{i=1}^n Y_i$ has a binomial (n, θ) distribution, given by

$$P(Y = y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad y \in \{0, 1, \ldots, n\}$$

- $E(Y|\theta) = n\theta$;
- $V(Y|\theta) = n\theta(1-\theta)$;

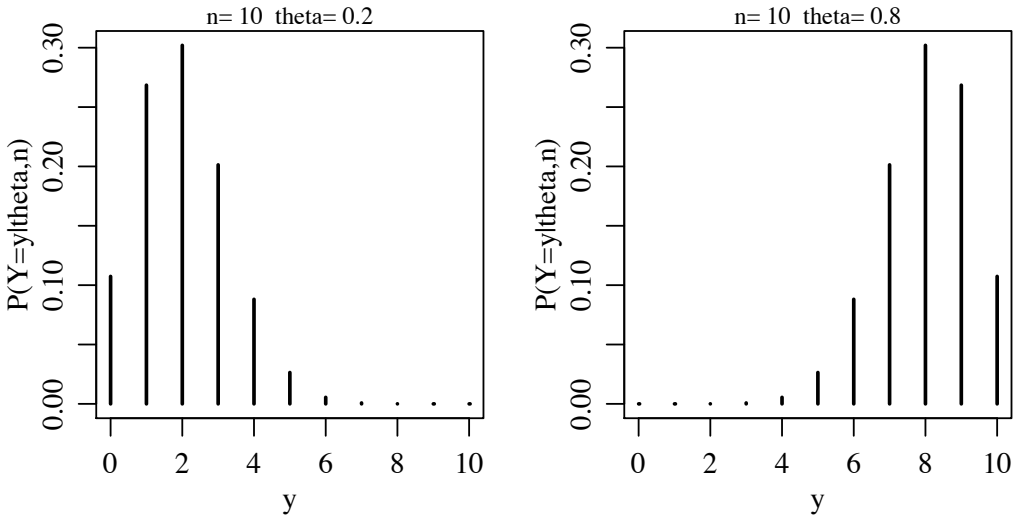


Figure 3.2: Binomial distribution, $n = 10$

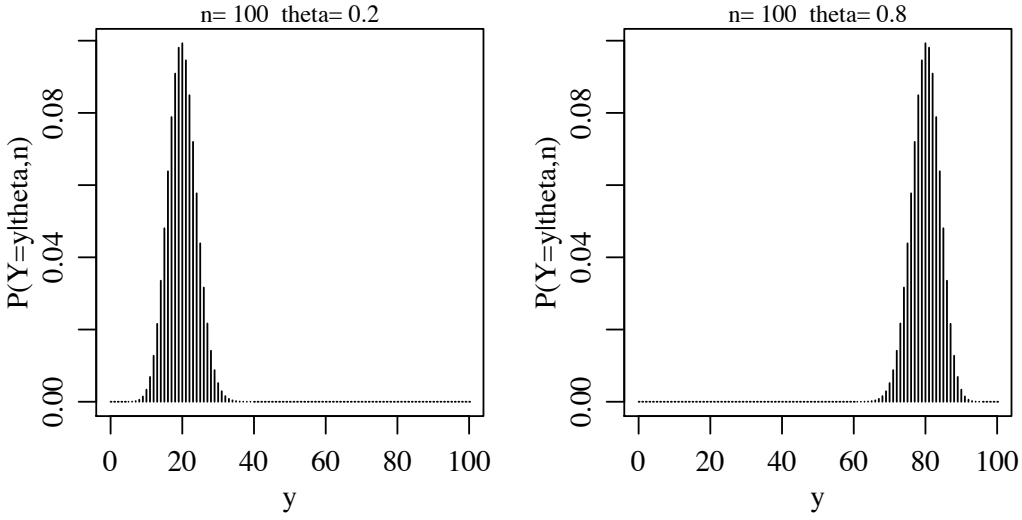


Figure 3.3: Binomial distribution, $n = 100$

Posterior inference: Having observed $Y = y$, what is the posterior of θ?

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &= \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}p(\theta)}{p(y)} \\ &= c(y)\theta^y(1-\theta)^{n-y}p(\theta) \end{aligned}$$

Lets start with $p(\theta) = 1$, the uniform distribution. Again, we can find out what $c(y)$ is using our calculus trick:

$$\begin{aligned} 1 &= \int_0^1 c(y)\theta^y(1-\theta)^{n-y}\ d\theta \\ &= c(y)\int_0^1 \theta^y(1-\theta)^{n-y}\ d\theta \\ &= c(y)\frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)} \end{aligned}$$

So

$$\begin{aligned} p(\theta|y) &= \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^y(1-\theta)^{n-y} \\ &= \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^{(y+1)-1}(1-\theta)^{(n-y+1)-1} \\ &= \text{beta}(y+1, n-y+1) \end{aligned}$$

Recall recidivism example:

$$n = 43, Y \equiv \sum Y_i = 15 \Rightarrow \theta|\{Y = 15\} \sim \text{beta}(16, 29)$$

and so in this model

$$P(\theta \in A|Y_1, \ldots, Y_n) = P(\theta \in A|\sum_{i=1}^n Y_i).$$

3.2.2 Posterior distributions under beta priors

The uniform prior has $p(\theta) = 1$ for all $\theta \in [0, 1]$. Note that this can be thought of as a beta prior with parameters $a = 1, b = 1$:

$$p(\theta) = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}\theta^{1-1}(1-\theta)^{1-1} = \frac{1}{1 \times 1} 1 \times 1 = 1$$

(Note: $\Gamma(x+1) = x! = x \times (x-1) \cdots \times 1$ if x is a positive integer. $\Gamma(1) = 1$ by convention). In the previous section, we saw that if

- $\theta \sim$ beta$(1, 1)$ (uniform) , and
- $Y \sim$ binomial(n, θ), then
- $\theta | Y = y \sim$ beta$(1 + y, 1 + n - y)$.

So to get the posterior, we simply added the number of 1's to the a parameter and the number of 0's to the b parameters. Does this result hold for arbitrary beta priors?

Let

- $\theta \sim$ beta(a, b);
- $Y | \theta \sim$ binomial(n, θ).

Then having observed $Y = y$,

$$\begin{aligned} p(\theta|y) &= \frac{p(\theta)p(y|\theta)}{p(y)} \\ &= \frac{1}{p(y)} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \times \binom{n}{y}\theta^y(1\theta)^{n-y} \\ &= c(n, y, a, b) \times \theta^{a+y-1}(1-\theta)^{b+n-y-1} \\ &= \text{beta}(a + y, b + n - y) \end{aligned}$$

$$\text{beta prior} + \text{binomial sampling} \Rightarrow \text{beta posterior}$$

We say the class of beta priors is *conjugate* for the binomial sampling model.

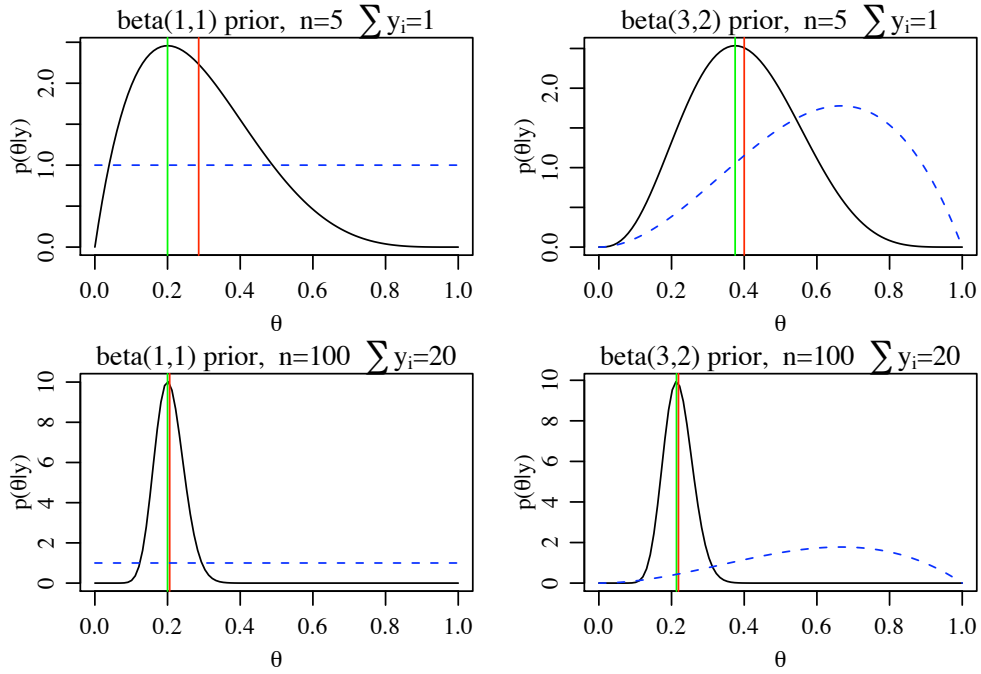


Figure 3.4: Binomial posterior distributions under two different sample sizes and two different prior distributions.

Definition 4 (Conjugate) *A class $\mathcal{P}$ of prior distribution for θ is called* conjugate *for a sampling model $p(y|\theta)$ if*

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|y) \in \mathcal{P}$$

Conjugate priors

- make posterior calculation easy;
- might not actually represent prior information (although mixtures of conjugate priors are very flexible).

3.3 Data combination

If $\theta|\{Y = y\} \sim \text{beta}(a + y, b + n - y)$, then

- $\text{E}(\theta|y) = \frac{a+y}{a+b+n}$;

- $\mathrm{V}(\theta|y) = \frac{E(\theta|y)E(1-\theta|y)}{a+b+n+1}$;
- $\mathrm{Mode}(\theta|y) = \frac{a+y-1}{a+b+n-2}$.

The posterior mean combines prior information and data information in a natural way:

$$\begin{aligned} E(\theta|y) = \frac{a+y}{a+b+n} &= \frac{a}{a+b+n} + \frac{y}{a+b+n} \\ &= \frac{a+b}{a+b+n}\frac{a}{a+b} + \frac{n}{a+b+n}\frac{y}{n} \\ &= \frac{a+b}{a+b+n} \times \text{prior mean} + \frac{n}{a+b+n} \times \text{data mean} \end{aligned}$$

The parameters a and b are sometimes thought of as prior data:

a="prior number of 1's"

b="prior number of 0's "

$a+b$="prior sample size"

For large n, the data dominates the posterior: If $n >> a+b$, then

$$\frac{a+b}{a+b+n} \approx 0 \ , \ E(\theta|y) \approx \frac{y}{n} \ , \ V(\theta|y) \approx \frac{1}{n}\frac{y}{n}\left(1-\frac{y}{n}\right)$$

3.4 Prediction

An important feature of Bayesian inference is the existence of a predictive distribution for new observations. Reverting for the moment to our notation for binary data, let $y_1, \ldots, y_n$ be a binary sample of size n:

$$\begin{aligned} p(y_{n+1}|y_1, \ldots, y_n) &= \int p(y_{n+1}=1|\theta, y_1, \ldots, y_n) p(\theta|y_1, \ldots, y_n)\ d\theta \\ &= \int \theta p(\theta|y_1, \ldots, y_n)\ d\theta \\ &= E(\theta|y_1, \ldots, y_n) = \frac{a+\sum_{i=1}^n y_i}{a+b+n} \end{aligned}$$

Example: Uniform prior $\Rightarrow \theta \sim \text{beta}(1,1) \approx$ "a prior 0 and a prior 1"

- $p(y_{n+1} = 1|y_1, \ldots, y_n) = E(\theta|y) = \frac{2}{2+n}\frac{1}{2} + \frac{2}{2+n}\frac{y}{n}$
- $\text{Mode}(\theta|y) = \frac{y}{n}$.

Think about whether or not it makes sense for your predictive probability to be different from the mode (consider the cases that $y = 0$ or $y = n$).

3.5 Confidence regions

Goal: identify regions of the parameter space having high posterior probability.

Definition 5 (Frequentist interval) *A (frequentist) 95% confidence interval for a parameter θ is any random interval $[l(Y), h(Y)]$ such that, before you run the experiment/survey*

$$P(l(Y) < \theta < h(Y)|\theta) = .95$$

Once you observe $Y = y$, and you plug the data into the formula for the confidence interval,

$$P(l(y) < \theta < h(y)|\theta) = \begin{cases} 0 & \text{if } \theta \notin [l(y), h(y)] \\ 1 & \text{if } \theta \in [l(y), h(y)] \end{cases}$$

So (when being strict about the frequentist notion of probability), the actual numerical value of the interval doesn't tell you about the value of θ.

Definition 6 (Bayesian interval) *A (Bayesian) 95% confidence interval for a parameter θ consists of any two numbers $[l(y), h(y)]$, based on the observed data $Y = y$, such that*

$$P(l(y) < \theta < h(y)|Y = y) = .95$$

In the Bayesian setting

- Once the survey is done $Y = y$ is fixed and known $\Rightarrow$ condition on it;
- θ is unknown $\Rightarrow$ describe information using probabilities.

3.5.1 Quantile-based interval

To make a $100 \times (1-\alpha)\%$ quantile-based confidence interval, find numbers $\theta_{\alpha/2} < \theta_{1-\alpha/2}$ such that

- $P(\theta < \theta_\alpha | y) = \alpha/2$;
- $P(\theta > \theta_{1-\alpha/2} | y) = \alpha/2$.

$\theta_{\alpha/2}$, $\theta_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ posterior quantiles of θ, and so

$$P(\theta_{\alpha/2} < \theta < \theta_{1-\alpha/2} | y) = 1 - \alpha.$$

Example: uniform prior, $n = 10, y = 2$.

```
> a<-1  ; b<-1   #prior
> n<-10 ; y<-2   #data

>theta.support<-seq(0,1,length=100)
>plot(theta.support, dbeta(theta.support, a+y, b+n-y), type="l",xlab="theta",
      ylab="p(theta|y)"  )
>qbeta( c(.025,.975), a+y,b+n-y)
[1] 0.06021773 0.51775585
>abline(v=qbeta( c(.025,.975), a+y,b+n-y)
```

So we are 95% sure that $\theta \in [0.06, 0.52]$. (Note-the frequentist interval is $[0.025, 0.56]$).

3.5.2 Highest posterior density interval

Note there are θ-values *outside* the quantile-based interval that have higher probability than some points *inside* he interval. This suggests a more restrictive type of interval:

Definition 7 (HPD interval) *A $100 \times (1-\alpha)\%$ HPD interval consists of θ-values between two points θ_l and θ_h such that*

- $P(\theta_l < \theta < \theta_h | y) = 1 - \alpha$;
- *If $\theta_1 \in [\theta_l, \theta_h]$, and $\theta_2 \notin [\theta_l, \theta_h]$, then $P(\theta_1|y) > P(\theta_2|y)$.*

All points in an hpd interval have a higher posterior density than points outside the interval. More precisely, we should talk about hpd *regions*. An HPD region might not be an interval if the posterior density is multi-modal.

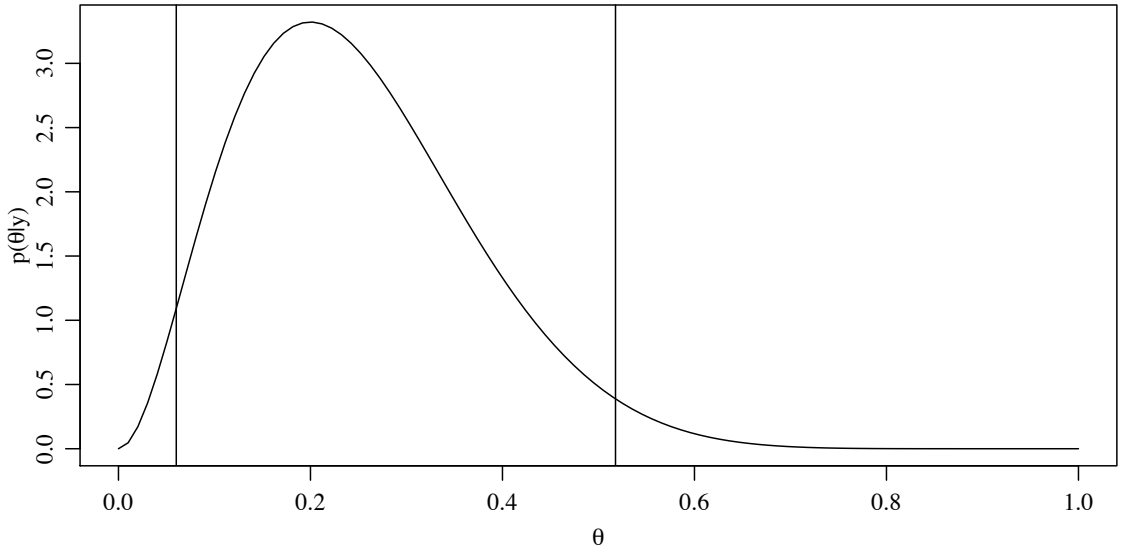


Figure 3.5: 95% Quantile-based confidence interval

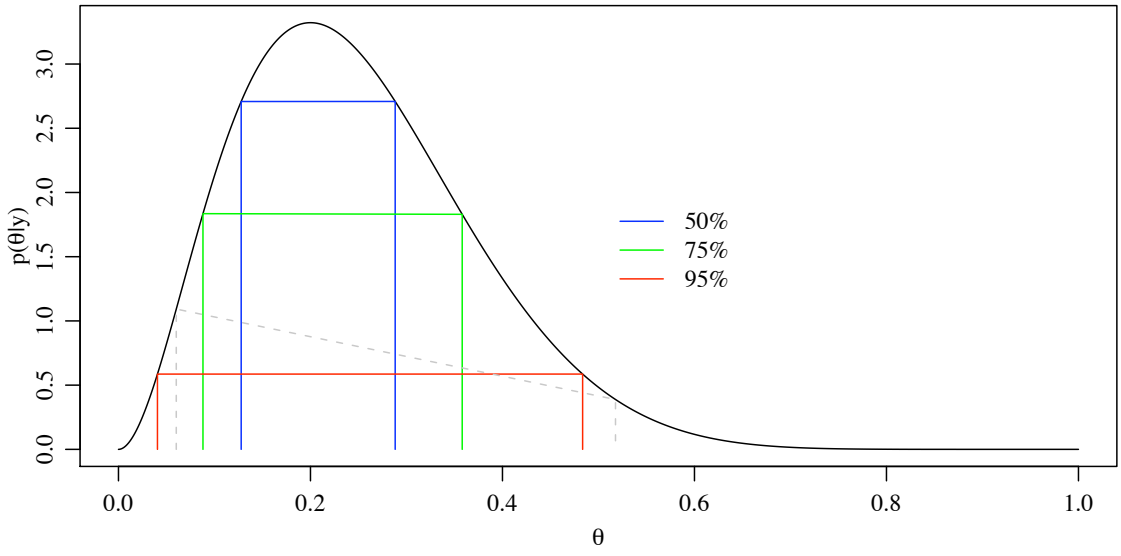


Figure 3.6: Highest posterior density regions of varying probability content. The dashed line is the quantile-based interval.

Chapter 4

The Poisson model

4.1 The Poisson model for Counts

Suppose $\mathcal{Y} = \{0, 1, 2, \ldots\}$. Examples:

- number of friends;
- number of children;
- ...

Perhaps the simplest probability distribution on $\mathcal{Y}$ is the Poisson distribution.

Poisson distribution: A random variable Y has a Poisson distribution with mean θ if

$$P(Y = y|\theta) = \theta^y e^{-\theta}/y!$$

In this case,

- $E(Y|\theta) = \theta$;
- $V(Y|\theta) = \theta$.

Sometimes people say the Poisson distribution has a mean-variance relationship.

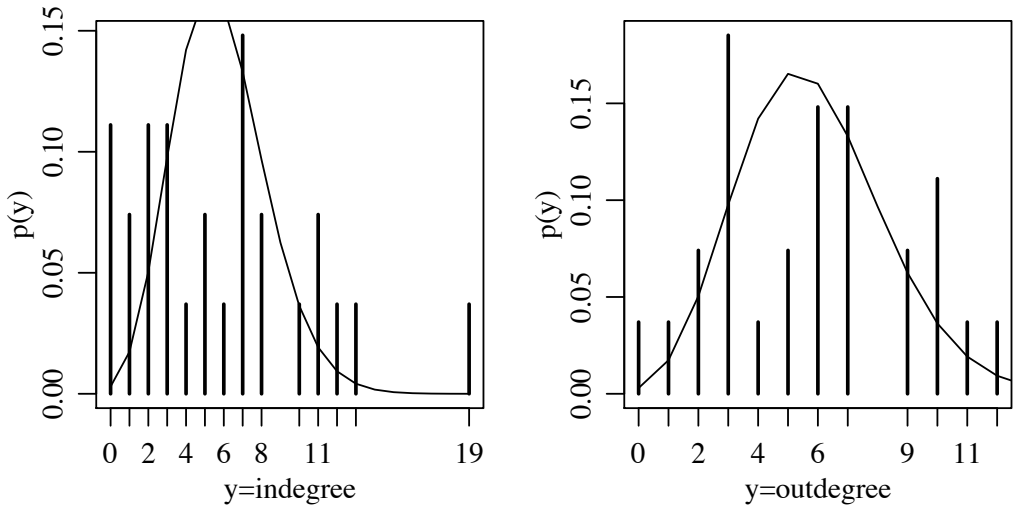


Figure 4.1: Indegrees and outdegrees of classroom friendships data. Connected lines are Poisson distributions with the same means.

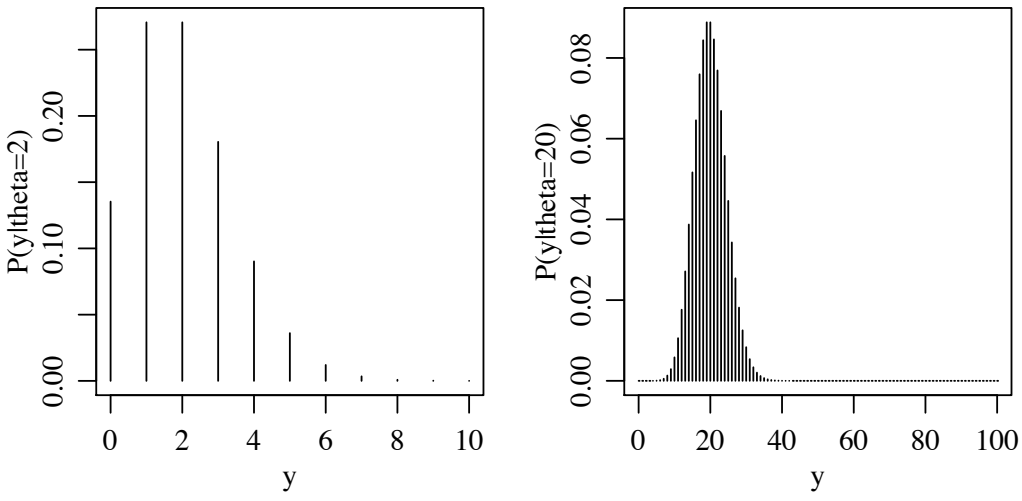


Figure 4.2: Poisson distributions.

4.1.1 Posterior inference:

Suppose we sample $Y_1, \ldots, Y_n \sim$ i.i.d. Pois(θ) ,

$$\begin{aligned} p(Y_1 = y_1, \ldots, Y_n = y_n|\theta) &= \prod_{i=1}^{n} p(y_i|\theta) \\ &= \prod_{i=1}^{n} \frac{1}{y_i!}\theta^{y_i}e^{-\theta} \\ &= c(\boldsymbol{y})\theta^{\sum y_i}e^{-n\theta} \end{aligned}$$

Comparing two values of θ a posteriori

$$\begin{aligned} \frac{p(\theta_a|y_1, \ldots, y_n)}{p(\theta_b|y_1, \ldots, y_n)} &= \frac{c(\boldsymbol{y})}{c(\boldsymbol{y})}\frac{e^{-n\theta_a}}{e^{-n\theta_b}}\frac{\theta_a^{\sum y_i}}{\theta_b^{\sum y_i}}\frac{p(\theta_a)}{p(\theta_b)} \\ &= \frac{e^{-n\theta_a}}{e^{-n\theta_b}}\frac{\theta_a^{\sum y_i}}{\theta_b^{\sum y_i}}\frac{p(\theta_a)}{p(\theta_b)} \end{aligned}$$

As before, $\sum_{i=1}^n Y_i$ contains all the information about θ in the data $\Rightarrow \sum_{i=1}^n Y_i$ is a *sufficient statistic*. Also, $\sum_{i=1}^n Y_i \sim$ Poisson($n\theta$).

Conjugate prior: Notice $p(\boldsymbol{y}|\theta)$, as a function of θ, is of the form $\theta^{c_1}e^{-c_2\theta}$. $\Rightarrow$ a conjugate family of priors are of the form

$$p(\theta) \propto \theta^{a-1}e^{-b\theta}$$

Such prior densities make up the family of gamma distributions.

Gamma distribution $\theta \sim \text{gamma}\,(a, b) \Leftrightarrow p(\theta) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}$, for $\theta > 0 \quad (a, b > 0)$.

- $E(\theta) = a/b$;
- $V(\theta) = a/b^2$;
- $\text{Mode}(\theta) = \begin{cases} (a-1)/b & \text{if } a > 1 \\ 0 & \text{if } a \leq 1 \end{cases}$

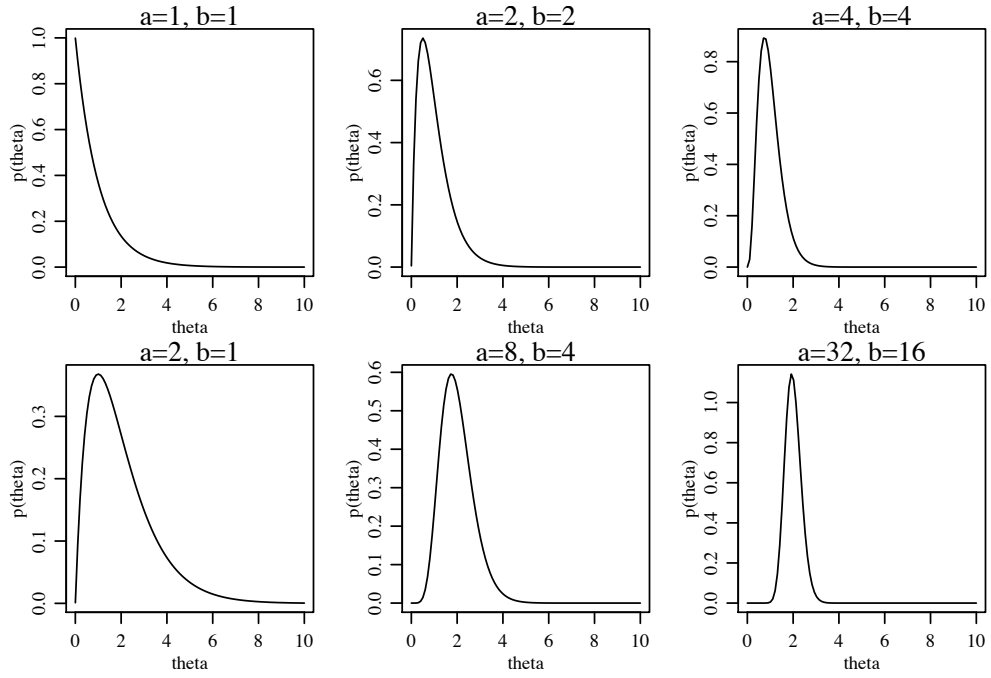


Figure 4.3: Gamma distributions

Posterior distribution of $\theta|\boldsymbol{y}$: Suppose $p(\theta)$= gamma(a, b). Then

$$\begin{aligned} p(\theta|y_1,\ldots,y_n) &= p(\theta)\times p(y_1,\ldots,y_n|\theta)/p(y_1,\ldots,y_n) \\ &= c(\boldsymbol{y},a,b)\times\left\{\theta^{a-1}e^{-b\theta}\right\}\times\left\{\theta^{\sum y_i}e^{-n\theta}\right\} \\ &= c(\boldsymbol{y},a,b)\times\theta^{a+\sum y_i-1}e^{-(b+n)\theta} \end{aligned}$$

This is evidently a gamma distribution.

$$\left.\begin{array}{c}\theta\sim\text{gamma}(a,b)\\ Y_1,\ldots,Y_n|\theta\sim\text{Poisson}(\theta)\end{array}\right\}\Rightarrow\theta|Y_1,\ldots,Y_n\sim\text{gamma}(a+\sum_{i=1}^{n}Y_i,b+n)$$

Notes:

1. Combining information: Again, posterior mean is a combination of prior and data information:

$$\begin{aligned} E(\theta|\mathbf{y}) &= \frac{a+\sum y_i}{b+n} \\ &= \frac{b}{b+n}\frac{a}{b}+\frac{n}{b+n}\frac{\sum y_i}{n} \end{aligned}$$

   - $b\approx$ number of prior observations;
   - $a\approx$ sum of counts from b prior observations.

2. Behavior for large n: As n increases, data dominates the prior:

$$n>>b\Rightarrow E(\theta|\mathbf{y})\approx\sum y_i/n$$

3. Posterior prediction:

$$\begin{aligned} p(y_{n+1}|y_1,\ldots,y_n) &= \int_0^\infty p(y_{n+1}|\theta,y_1,\ldots,y_n)p(\theta|y_1,\ldots,y_n)\;d\theta \\ &= \int p(y_{n+1}|\theta)p(\theta|y_1,\ldots,y_n)\;d\theta \\ &= \int\left\{\frac{1}{y_{n+1}!}\theta^{y_{n+1}}e^{-\theta}\right\}\left\{\frac{(b+n)^{a+\sum y_i}}{\Gamma(a+\sum y_i)}\theta^{a+\sum y_i-1}e^{-(b+n)\theta}\right\}\;d\theta \\ &= \frac{(b+n)^{a+\sum y_i}}{\Gamma(y_{n+1}+1)\Gamma(a+\sum y_i)}\int_0^\infty\theta^{a+\sum y_i+y_{n+1}-1}e^{-(b+n+1)\theta}\;d\theta \end{aligned}$$

How can we evaluate this integral? Lets use what we know about the gamma density:

$$1 = \int_0^\infty \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \, d\theta \quad \text{for any values } a, b > 0$$

This means that

$$\int_0^\infty \theta^{a-1} e^{-b\theta} \, d\theta = \frac{\Gamma(a)}{b^a} \quad \text{for any values } a, b > 0$$

Now substitute in $a + \sum y_i + y_{n+1}$ instead of a, and $b + n + 1$ instead of b, to get

$$\int_0^\infty \theta^{a+\sum y_i + y_{n+1} - 1} e^{-(b+n+1)\theta} \, d\theta = \frac{\Gamma(a + \sum y_i + y_{n+1})}{(b+n+1)^{a+\sum y_i + y_{n+1}}}$$

After simplifying some of the algebra, this gives

$$p(y_{n+1}|y_1, \ldots, y_n) = \frac{\Gamma(a + \sum y_i + y_{n+1})}{\Gamma(y_{n+1}+1)\Gamma(a + \sum y_i)} \left(\frac{b+n}{b+n+1}\right)^{a+\sum y_i} \left(\frac{1}{b+n+1}\right)^{y_{n+1}}$$

for $y_{n+1} \in \{0, 1, 2, \ldots\}$. This is a negative binomial distribution with parameters $(a + \sum y_i, b + n)$.

$$\begin{aligned} E(Y_{n+1}|y_1, \ldots, y_n) &= \frac{a + \sum y_i}{b+n} = E(\theta|y_1, \ldots, y_n) \\ V(Y_{n+1}|y_1, \ldots, y_n) &= \frac{a + \sum y_i}{b+n} \frac{b+n+1}{b+n} = V(\theta|y_1, \ldots y_n) \times (b+n+1) \\ &= E(\theta|y_1, \ldots, y_n) \times \frac{b+n+1}{b+n} \end{aligned}$$

Try to give a heuristic interpretation of the last line.

Numerical Example: In each of two cities, a sample of IV drug users are asked how many needle sharing partners they have. City 1 has a needle exchange program, city 2 does not.

- $n_1 = 50,\ \sum_{i=1}^{n_1} Y_{i,1} = 102\ \bar{Y}_1 = 2.04$;
- $n_2 = 46,\ \sum_{i=1}^{n_1} Y_{i,1} = 104\ \bar{Y}_2 = 2.26$.

In the case where where $\theta_1 \sim$ gamma$(a = 2, b = 1)$, $\theta_2 \sim$ gamma$(a = 2, b = 1)$, we have

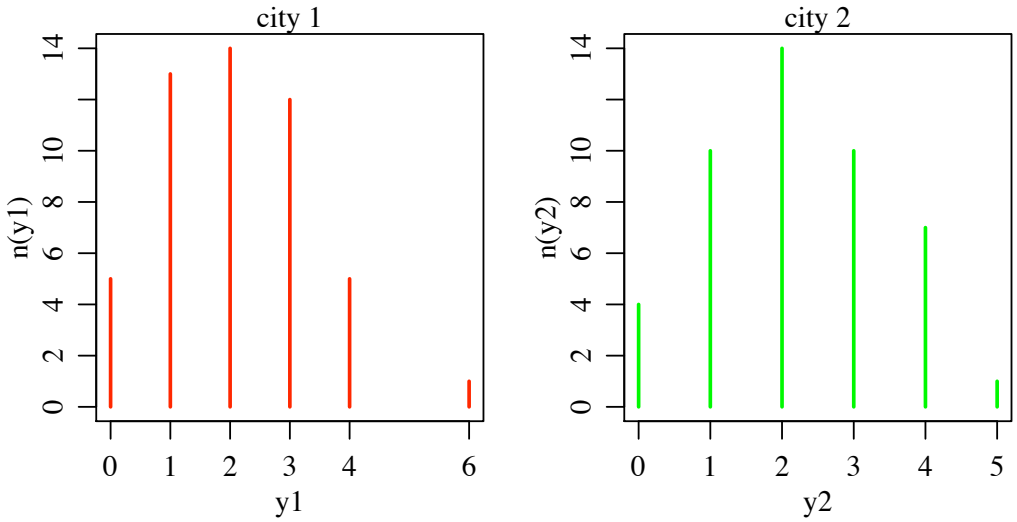


Figure 4.4: Number of needle-sharing partners

- $(\theta_1|n_1 = 50, \sum Y_{i,1} = 102) \sim$ gamma(2+102, 1+50) = gamma(104, 51).
- $(\theta_2|n_2 = 46, \sum Y_{i,2} = 104) \sim$ gamma(2+104, 1+46) = gamma(106, 48) .

```
a<-2 ; b<-1
n1<-length(y1) ; s1<-sum(y1)
n2<-length(y2) ; s2<-sum(y2)

(a+s1-1)/(b+n1)
[1] 2.019608
qgamma( c(.025,.975),a+s1,b+n1)
[1] 1.666187 2.449362

(a+s2-1)/(b+n2)
[1] 2.234043
qgamma( c(.025,.975),a+s2,b+n2)
[1] 1.846471 2.704445

theta1.samp<-rgamma(1000,a+s1,b+n1)
theta2.samp<-rgamma(1000,a+s2,b+n2)
mean(theta1.samp<theta2.samp)
[1] 0.772
```

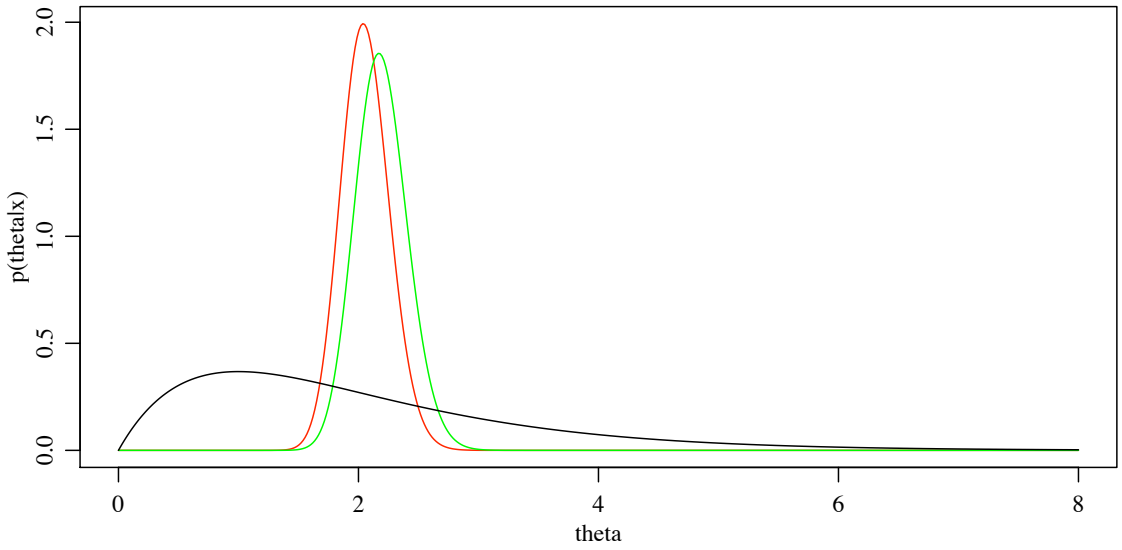


Figure 4.5: Posterior distributions of needle-sharing rates

```
y<-seq(0:10)
dnbinom(y, size=(a+n1), mu=(a+s1)/(b+n1))
 [1] 2.654981e-01 2.654981e-01 1.803383e-01 9.357177e-02 3.954732e-02
 [6] 1.417734e-02 4.432807e-03 1.233659e-03 3.103544e-04 7.144006e-05
[11] 1.519480e-05
```

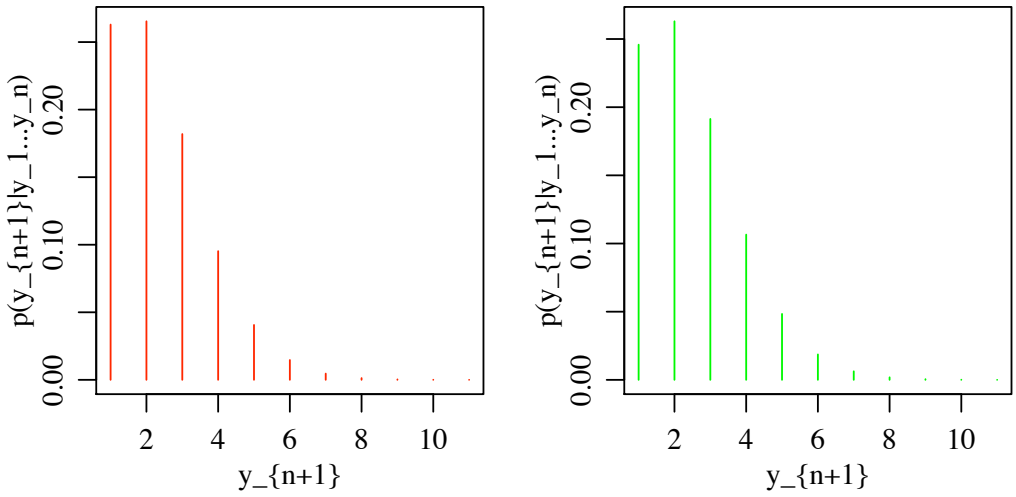


Figure 4.6: Posterior predictive distributions for number of partners

Chapter 5

Monte Carlo approximation

Goal: Given a posterior $p(\theta|\mathbf{y})$ that we can sample from, approximate

- $P(\theta \in A|\mathbf{y})$ for arbitrary sets A,
- $E(\theta|\mathbf{y})$,
- $V(\theta|\mathbf{y})$,
- quantiles of $p(\theta|\mathbf{y})$

and perhaps more importantly, do the same for any function $g(\theta)$.

5.1 Monte Carlo method

Suppose $p(\theta|\mathbf{y})$ is known and we can sample from it, for example:

binomial model/beta prior: $\theta|y_1, \ldots, y_n \sim$ beta $(a+\sum y_i, b+n-\sum y_i)$, generate samples using the `rbeta` command;

Poisson model/gamma prior: $\theta|y_1, \ldots, y_n \sim$ gamma $(a+\sum y_i, b+n)$, generate samples using the `rgamma` command.

Sample

$$\left.\begin{array}{rcl} \theta_{(1)} & \sim & p(\theta|\mathbf{y}) \\ \theta_{(2)} & \sim & p(\theta|\mathbf{y}) \\ \vdots & & \\ \theta_{(m)} & \sim & p(\theta|\mathbf{y}) \end{array}\right\} \text{ independently}$$

Then the *empirical distribution* of $\theta_{(1)}, \ldots, \theta_{(m)}$ approximates $p(\theta|\mathbf{y})$. The larger m is, the better the approximation:

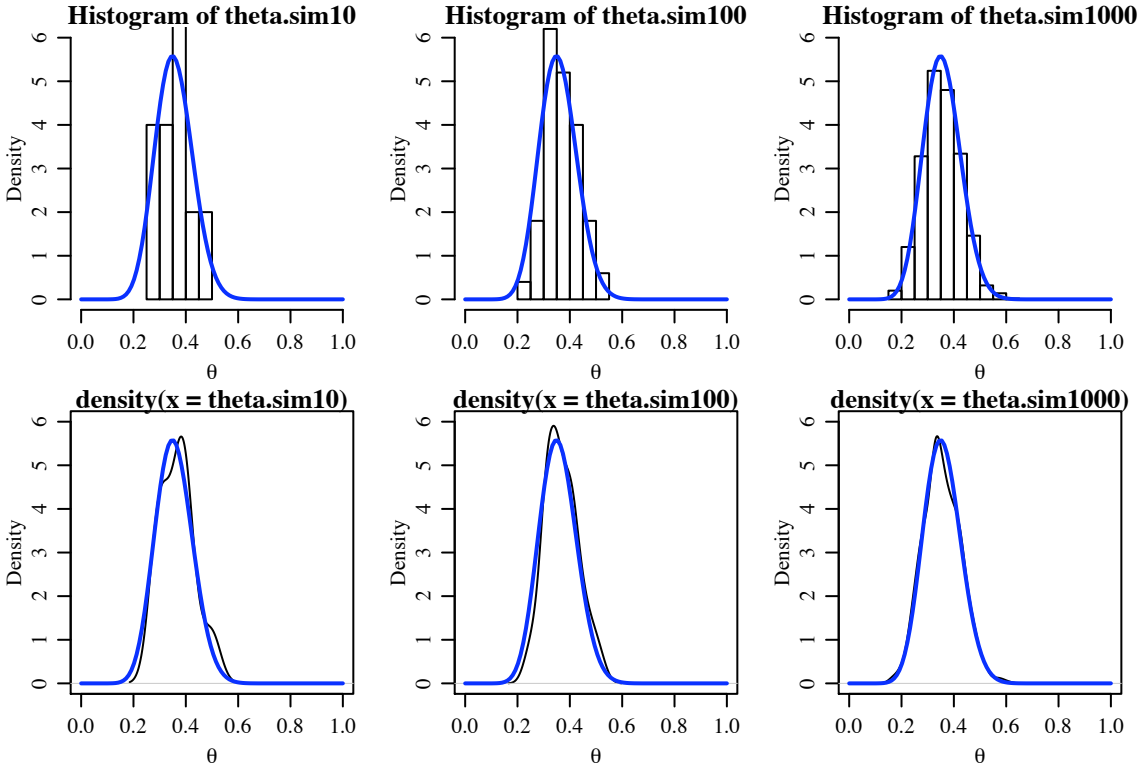


Figure 5.1: Monte Carlo approximations to a beta(16, 29) distribution

Additionally, let $g(\theta)$ be (just about) any function. Then

$$\frac{1}{m}\sum_{j=1}^{m} g(\theta_{(j)}) \rightarrow E[g(\theta)|y_1, \ldots, y_n] = \int g(\theta)p(\theta|y_1, \ldots, y_n)\, d\theta \text{ as } m \rightarrow \infty.$$

In particular,

- empirical distribution $\{\theta_{(1)}, \ldots, \theta_{(m)}\} \rightarrow p(\theta|\mathbf{y})$; as $m \rightarrow \infty$;
- $\sum \theta_{(j)}/m \rightarrow E(\theta|\mathbf{y})$ as $m \rightarrow \infty$;
- $\sum(\theta_{(j)} - \bar{\theta})^2/m \rightarrow V(\theta|\mathbf{y})$ as $m \rightarrow \infty$;
- $\#(\theta_{(j)} \leq c)/m \rightarrow P(\theta \leq c|\mathbf{y})$ as $m \rightarrow \infty$;

- median $\{\theta_{(1)}, \ldots, \theta_{(m)}\} \to \theta_{1/2}$ as $m \to \infty$ (recall $P(\theta \leq \theta_{1/2}|\mathbf{y}) = 1/2$) ;
- α-percentile of $\{\theta_{(1)}, \ldots, \theta_{(m)}\} \to \theta_\alpha$ as $m \to \infty$.

5.2 Recidivism Example

Recidivism example: $\theta|\{n = 43, \sum Y_i = 15\} \sim$ beta $(a + 15, b + 28)$

```
a<-1 ; b<-1
n<-43; n1<-15 ; n0<-28

theta.sim10<-rbeta(10,a+n1,b+n0)
theta.sim100<-rbeta(100,a+n1,b+n0)
theta.sim1000<-rbeta(1000,a+n1,b+n0)

#########

> (a+n1)/(a+b+n1+n0)
[1] 0.3555556

> mean(theta.sim10)
[1] 0.3731655
> mean(theta.sim100)
[1] 0.3588319
> mean(theta.sim1000)
[1] 0.3574751

#########

> pbeta(.3,a+n1,b+n0)
[1] 0.2219175

> mean( theta.sim10<.3)
[1] 0.2
> mean( theta.sim100<.3)
[1] 0.18
> mean( theta.sim1000<.3)
[1] 0.218

#########

> qbeta(c(.025,.975),a+n1,b+n0)
[1] 0.2240801 0.4991950
```

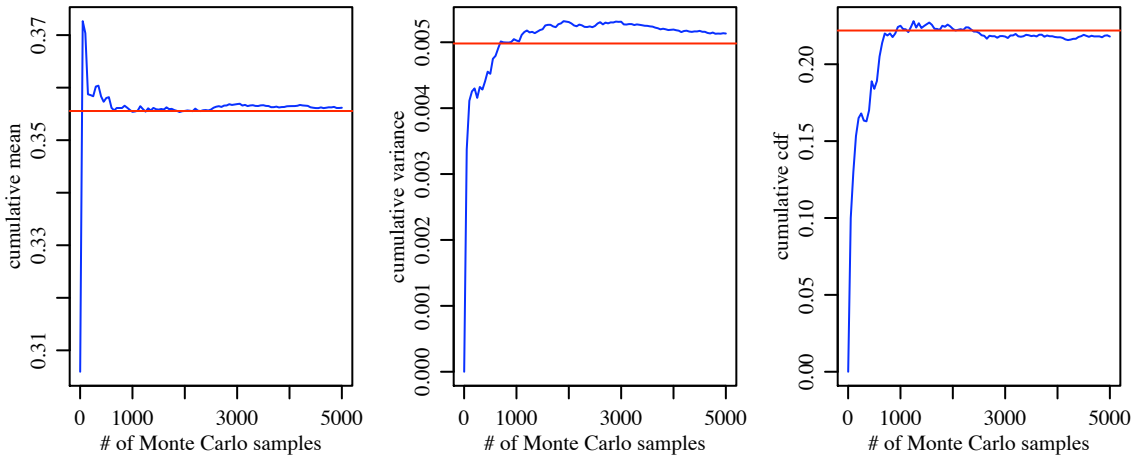


Figure 5.2: Estimates of the mean, variance and CDF of a beta distribution as a function of the number of Monte Carlo samples. Horizontal red lines are the true values.

```
> quantile( theta.sim10, c(.025,.975))
     2.5%     97.5%
0.2405577 0.4639730
> quantile( theta.sim100, c(.025,.975))
     2.5%     97.5%
0.2384364 0.5094468
> quantile( theta.sim1000, c(.025,.975))
     2.5%     97.5%
0.2266413 0.5031118

#########
```

5.3 Posterior inference on $g(\theta)$

Let $g(\theta)$ be some computable function of θ. In the recidivism example, we might be interested in

$$\text{log odds of recidivism} = \log \frac{\theta}{1-\theta} = \gamma$$

Questions:

- What is $p(\gamma|y_1,\ldots,y_n)$?
- What is $E(\gamma|y_1,\ldots,y_n)$?
- What is $p(\gamma<1|y_1,\ldots,y_n)$?

Monte Carlo method: Sample

$$\left.\begin{array}{rcll} \theta_{(1)} & \sim & p(\theta|\mathbf{y}), & \text{compute } \gamma_{(1)}=g(\theta_{(1)}) \\ \theta_{(2)} & \sim & p(\theta|\mathbf{y}), & \text{compute } \gamma_{(2)}=g(\theta_{(2)}) \\ \vdots & & & \\ \theta_{(m)} & \sim & p(\theta|\mathbf{y}), & \text{compute } \gamma_{(m)}=g(\theta_{(m)}) \end{array}\right\} \text{ independently}$$

Then $\{\gamma_{(1)},\ldots,\gamma_{(m)}\}$ constitutes m independent samples from $p(\gamma|\boldsymbol{y})$, and so

- empirical distribution $\{\gamma_{(1)},\ldots,\gamma_{(m)}\} \rightarrow p(\gamma|\mathbf{y})$; as $m\rightarrow\infty$;
- $\sum\gamma_{(i)}/m \rightarrow E(\gamma|\mathbf{y})$ as $m\rightarrow\infty$;
- $\sum(\gamma_{(i)}-\bar{\gamma})^2/m \rightarrow V(\gamma|\mathbf{y})$ as $m\rightarrow\infty$;
- etc.

```
> theta.prior.sim<-rbeta(5000,a,b)
> gamma.prior.sim<- log( theta.prior.sim/(1-theta.prior.sim) )

> mean(gamma.prior.sim)
[1] -0.03344748
> mean(gamma.prior.sim<0)
[1] 0.4986
> mean(theta.prior.sim<1/2)
[1] 0.4986


> theta.post.sim<-rbeta(5000,a+n1,b+n0)
> gamma.post.sim<- log( theta.post.sim/(1-theta.post.sim) )

> mean(gamma.post.sim)
[1] -0.6050782
> mean(gamma.post.sim<0)
[1] 0.9726
> mean(theta.post.sim<1/2)
[1] 0.9726
```

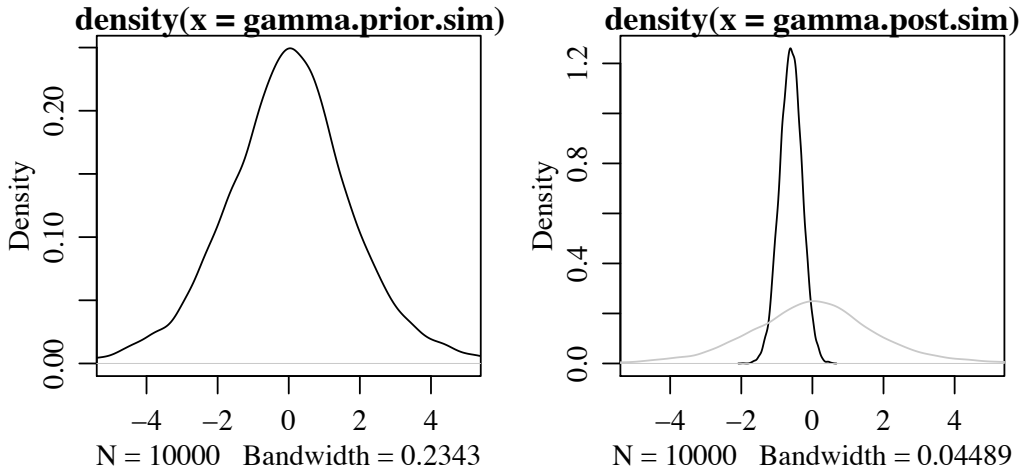


Figure 5.3: Monte Carlo approximations to the prior and posterior distributions of the log-odds.

5.4 Sampling from posterior predictive distributions

Suppose we can sample from

$p(\theta|y_1, \ldots, y_n)$ and

$p(y|\theta)$,

but $p(y_{n+1}|y_1, \ldots, y_n)$ is complicated. Recall,

$$p(y_{n+1}|y_1, \ldots, y_n) = \int p(y_{n+1}|\theta)p(\theta|y_1, \ldots, y_n)\ d\theta$$

so $p(y_{n+1}|y_1, \ldots, y_n)$ is the posterior expectation of $p(y_{n+1}|\theta)$.

One approach to sampling $y_{n+1}|y_1, \ldots, y_n$:

$$\left.\begin{array}{lcl} \theta_{(1)} \sim p(\theta|\mathbf{y}) & , & y_{n+1,(1)} \sim p(y_{n+1}|\theta_{(1)}) \\ \theta_{(2)} \sim p(\theta|\mathbf{y}) & , & y_{n+1,(2)} \sim p(y_{n+1}|\theta_{(2)}) \\ & \vdots & \\ \theta_{(m)} \sim p(\theta|\mathbf{y}) & , & y_{n+1,(m)} \sim p(y_{n+1}|\theta_{(m)}) \end{array}\right\} \text{ independently}$$

then

empirical distribution of $y_{n+1,(1)}, \ldots, y_{n+1,(m)} \approx p(y_{n+1}|y_1, \ldots, y_n)$

Self-test question: Consider two random variables X, Y and two sampling schemes:

(a) Sample $Y_a \sim p(y)$, the marginal distribution of Y.

(b) Sample $X \sim p(x)$ and then $Y_b \sim p(y|X)$.

Show that Y_a and Y_b have the same distribution, i.e. $\Pr(Y_a \leq y) = \Pr(Y_b \leq y)$ for any number y.

Example (Poisson distribution): Suppose

- $\theta \sim \text{gamma}(a, b)$;
- $y_1 \ldots, y_n|\theta \sim \text{Poisson}(\theta)$, so
- $\theta|y_1, \ldots, y_n \sim \text{gamma}(a + \sum y_i, b + n)$.

Sample

$$\left.\begin{array}{lcl} \theta_{(1)} \sim \text{gamma}(a + \sum y_i, b + n) & , & y_{n+1,(1)} \sim \text{Poisson}(\theta_{(1)}) \\ \theta_{(2)} \sim \text{gamma}(a + \sum y_i, b + n) & , & y_{n+1,(2)} \sim \text{Poisson}(\theta_{(2)}) \\ & \vdots & \\ \theta_{(m)} \sim \text{gamma}(a + \sum y_i, b + n) & , & y_{n+1,(m)} \sim \text{Poisson}(\theta_{(m)}) \end{array}\right\} \text{independently}$$

```
a<-2 ; b<-1
n<-10 ;  sy<-17

theta.sim<-rgamma(1000,a+sy, b+n)
ynew.sim<-rpois(1000,theta.sim)
```

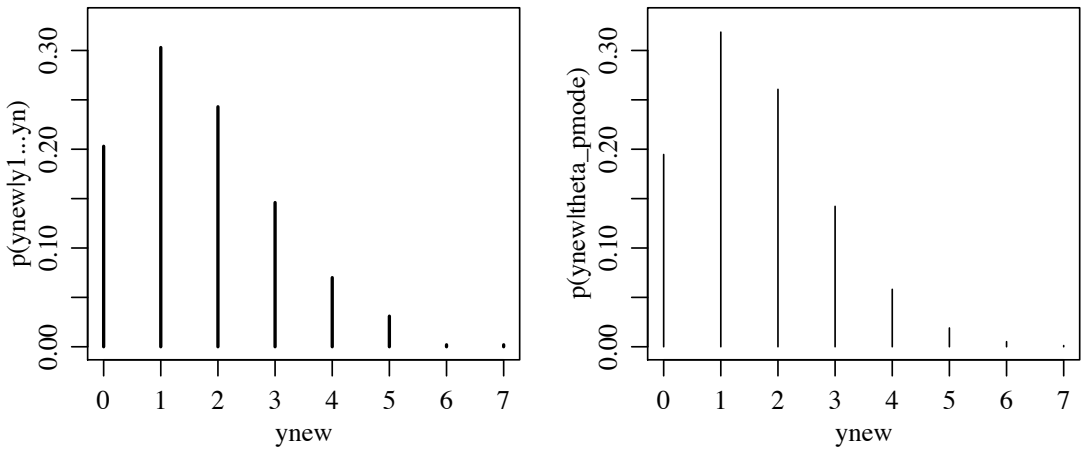


Figure 5.4: The first panel gives the posterior predictive distribution for a conjugate Poisson model. The second panel is the sampling distribution conditioning on $\theta = \text{mode}(\theta|\mathbf{y})$. Notice it is slightly less dispersed.

5.5 Posterior inference on two parameters

Suppose we have two populations, for which

- (a) $y_{1,a}, \ldots, y_{n_a,a}|\theta_a \sim$ i.i.d. $p(y|\theta_a)$;
- (b) $y_{1,b}, \ldots, y_{n_b,b}|\theta_b \sim$ i.i.d. $p(y|\theta_b)$;

and that a priori θ_a and θ_b are independent: $p(\theta_a, \theta_b) = p(\theta_a)p(\theta_b)$.

One way to calculate $p(\theta_a > \theta_b|\mathbf{y}_a, \mathbf{y}_b)$ is via Monte Carlo sampling: Sample

$$\left.\begin{array}{llll} \theta_{(1),a} & \sim & p(\theta_a|\mathbf{y}_a), & \theta_{(1),b} \sim p(\theta_b|\mathbf{y}_b) \\ \theta_{(2),a} & \sim & p(\theta_a|\mathbf{y}_a), & \theta_{(1),b} \sim p(\theta_b|\mathbf{y}_b) \\ \vdots & & & \\ \theta_{(m),a} & \sim & p(\theta_a|\mathbf{y}_a), & \theta_{(m),b} \sim p(\theta_b|\mathbf{y}_b) \end{array}\right\} \text{ independently}$$

Then

$$\frac{1}{m}\sum_{j=1}^{m} 1(\theta_{(j),a} > \theta_{(j),b}) \approx p(\theta_a > \theta_b|\mathbf{y}_a, \mathbf{y}_b)$$

Actually, we don't need the parameters to be independent here. If they were dependent, we would just have to obtain our Monte Carlo samples from their joint distribution. We will see examples of this shortly.

Chapter 6

The normal model

The **normal distribution** is a good model for *sums* of things.

Example: Let $Y_i = CO$ concentration in neighborhood i.

$$Y_1 = a + b \times \text{traffic}_1 + c \times \text{industry}_1 + d \times \text{greenspace}_1 + \cdots$$

$$Y_2 = a + b \times \text{traffic}_2 + c \times \text{industry}_2 + d \times \text{greenspace}_2 + \cdots$$

$$\vdots$$

$$Y_n = a + b \times \text{traffic}_n + c \times \text{industry}_n + d \times \text{greenspace}_n + \cdots$$

if each type of effect varies somewhat independently of the others, and their are many *additive* effects, then the *distribution* of $Y_1, \ldots, Y_n$ will look approximately normal.

Lots of independent, additive effects $\Rightarrow$ population looks normal.

Normal density: $Y \sim \text{normal}\ (\theta, \sigma) \Leftrightarrow$

$$p(y|\theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{y-\theta}{\sigma})^2}$$

Some useful information...

- About 95% of the probability lies within two standard deviations of the mean (more precisely, 1.96).

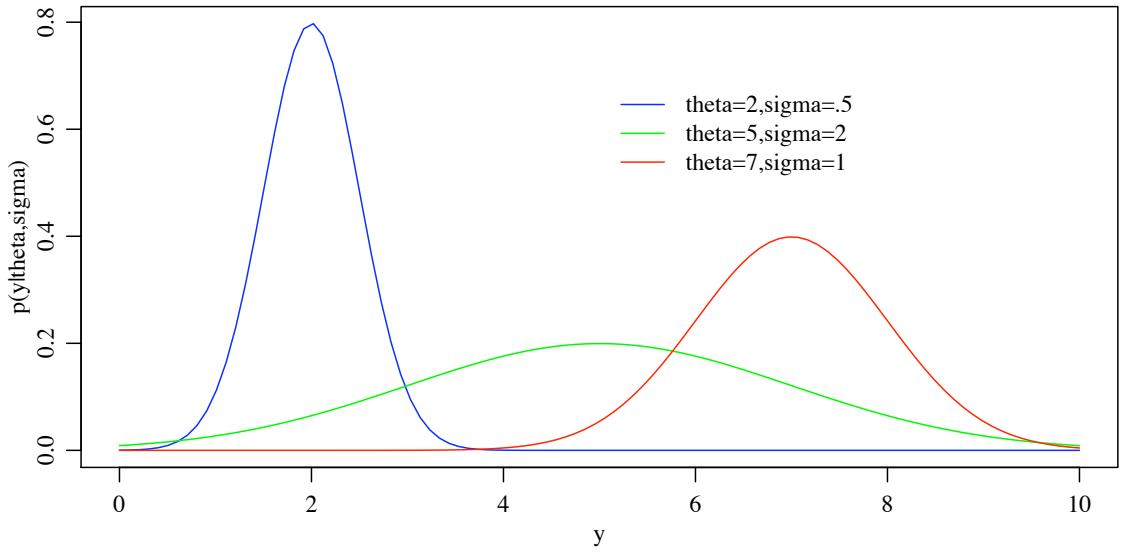


Figure 6.1: Some normal densities.

- The normal distribution with mean θ and sd σ is the most diffuse distribution of all distributions having mean μ and sd σ (it is the *maximum entropy* distribution).
- The `dnorm, rnorm, pnorm, qnorm` commands in R take σ as their argument, not σ^2:

```
> dnorm
function (x, mean = 0, sd = 1, log = FALSE)
.Internal(dnorm(x, mean, sd, log))
```

Suppose our model is $\{Y_1, \ldots, Y_n|\theta, \sigma\} \sim$ i.i.d. normal (θ, σ). Then

$$\begin{aligned} p(y_1, \ldots, y_n|\theta, \sigma) &= \prod_{i=1}^{n} p(y_i|\theta, \sigma) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{y_i-\theta}{\sigma})^2} \\ &= (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2}\sum(\frac{y_i-\theta}{\sigma})^2\} \end{aligned}$$

Note: $p(y_1, \ldots, y_n|\theta, \sigma)$ depends on $y_1, \ldots, y_n$ through

$$\sum_{i=1}^{n}(\frac{y_i - \theta}{\sigma})^2 = \frac{1}{\sigma^2}\sum y_i^2 - 2\frac{\theta}{\sigma^2}\sum y_i + n\frac{\theta^2}{\sigma^2}$$

We can show that $\sum y_i^2, \sum y_i$ are sufficient statistics. Knowing the values of these statistics is equivalent to knowing

- $\bar{y} = \sum y_i/n$;
- $s^2 = \sum(y_i - \bar{y})^2/(n-1)$.

so $\bar{y}, s^2$ are also sufficient statistics.

6.1 Inference for the mean, conditional on the variance:

Conjugate prior for θ :

$$p(y_1, \ldots, y_n|\theta, \sigma) = c(\sigma)e^{-\frac{1}{2\sigma^2}\sum(y_i-\theta)^2}$$

A conjugate family of densities for θ will have the form

$$p(\theta) \propto e^{c_1(\theta - c_1)^2}$$

which must be the normal family of densities.

Try it out: Let $\theta \sim$ normal (μ_0, τ_0):

$$\begin{aligned} p(\theta|y_1, \ldots, y_n, \sigma) &= p(\theta|\sigma)p(y_1, \ldots, y_n|\theta, \sigma)/p(y_1, \ldots, y_n|\sigma) \\ &\propto p(\theta|\sigma)p(y_1, \ldots, y_n|\theta, \sigma) \\ &\propto \exp\{-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\}\exp\{-\frac{1}{2\sigma^2}\sum(y_i - \theta)^2\} \\ &\propto \exp\{-\frac{1}{2\tau_n^2}(\theta - \mu_n)^2\} \end{aligned}$$

where

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \qquad \tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

$$\Rightarrow \theta|\{y_1, \ldots, y_n, \sigma\} \sim \text{normal } (\mu_n, \tau_n^2).$$

Notes:

1. Let

   - $\tilde{\sigma}^2 = 1/\sigma^2$ = sampling precision, i.e. how close to θ the y_i's are;
   - $\tilde{\tau}_0^2 = 1/\tau_0^2$ = prior precision;
   - $\tilde{\tau}_n^2 = 1/\tau_n^2$ = posterior precision;

   Note that
   $$\tilde{\tau}_n^2 = \tilde{\tau}_0^2 + n\tilde{\sigma}^2,$$
   i.e. posterior precision = prior precision + data precision.

2. Combining information:
   $$\mu_n = \frac{\tilde{\tau}_0^2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\mu_0 + \frac{n\tilde{\sigma}_2}{\tilde{\tau}_0^2 + n\tilde{\sigma}^2}\bar{y}$$
   so we can think of

   - μ_0 = mean of previous observations;
   - τ_0^2 = variance of the *mean* of previous observations.

3. Posterior prediction: Consider predicting a new observation y_{n+1} :

   $y_{n+1}|\theta, \sigma \sim$ normal $(\theta, \sigma) \Leftrightarrow y_{n+1} = \theta + \epsilon_{n+1}, \quad \epsilon_{n+1}|\sigma \sim$ normal$(0, \sigma)$

   Lets first compute the posterior mean and variance of y_{n+1} :

   $$\begin{aligned} E[y_{n+1}|y_1, \ldots, y_n, \sigma] &= E[\theta + \epsilon_{n+1}|y_1, \ldots, y_n, \sigma] \\ &= E[\theta|y_1, \ldots, y_n, \sigma] + E[\epsilon_{n+1}|y_1, \ldots, y_n, \sigma] \\ &= \mu_n + 0 = \mu_n \end{aligned}$$

   $$\begin{aligned} V[y_{n+1}|y_1, \ldots, y_n, \sigma] &= V[\theta + \epsilon_{n+1}|y_1, \ldots, y_n, \sigma] \\ &= V[\theta|y_1, \ldots, y_n, \sigma] + V[\epsilon_{n+1}|y_1, \ldots, y_n, \sigma] \\ &= \tau_n^2 + \sigma^2 \end{aligned}$$

   Also, since both θ and ϵ_{n+1}, conditional on $y_1, \ldots, y_n$, are normally distributed, so is $y_{n+1} = \theta + \epsilon_{n+1}$. So
   $$y_{n+1}|y_1, \ldots, y_n \sim \text{ normal}(\mu_n, \sqrt{\tau_n^2 + \sigma^2}).$$

Example: log income Suppose we are interested θ, the mean log-income of a subpopulation of the U.S.. Based on national data, we might use $\mu \sim$ normal $(\mu_0 = 10.7, \tau_0 = 0.5)$. Based on a sample of size $n = 50$, we observe $\bar{y} = 9.9, s^2 = 0.4$. Then $\mu|y_1, \ldots, y_{50}, \sigma \sim$ normal (μ_n, τ_n), where

$$\begin{aligned} \mu_n &= \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \\ &= \frac{4 \times 10.7 + \frac{50}{\sigma^2}9.9}{4 + \frac{50}{\sigma^2}} \end{aligned}$$

$$\tau_n^2 = \frac{1}{4 + \frac{50}{\sigma^2}}$$

If $\sigma^2 = s^2 = 0.4$, then $\mu|y_1, \ldots, y_{50}, \sigma = .632 \sim$ normal $(9.92, 0.088)$.

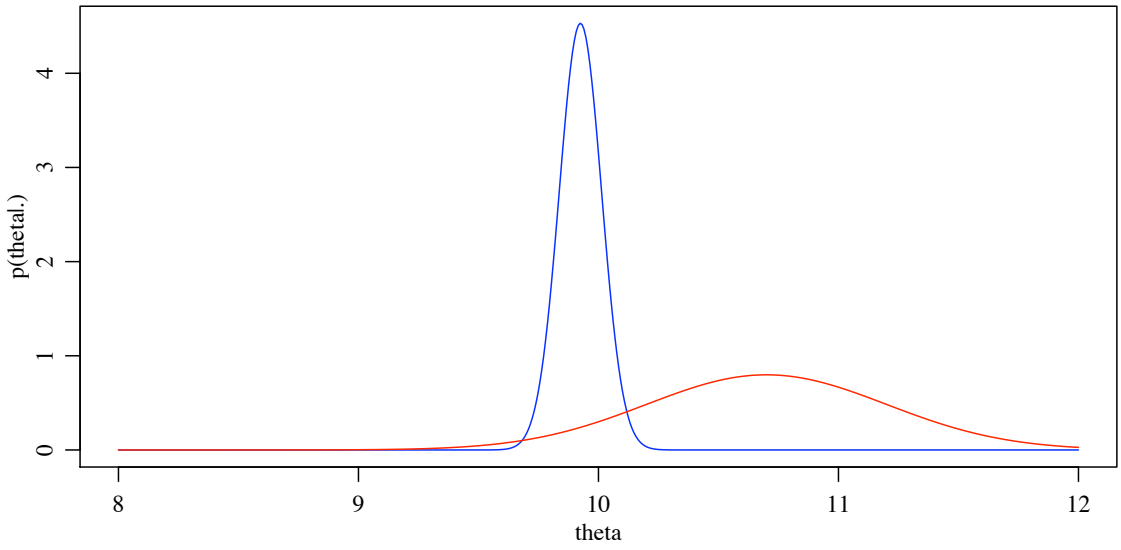


Figure 6.2: Prior and posterior distributions for the mean in the income example.

Also, the predictive distribution for a person sampled from this subpopulation is

$$y_{new} \sim \text{ normal } (9.92, \sqrt{0.088^2 + .632^2}) = \text{ normal } (9.92, .639)$$

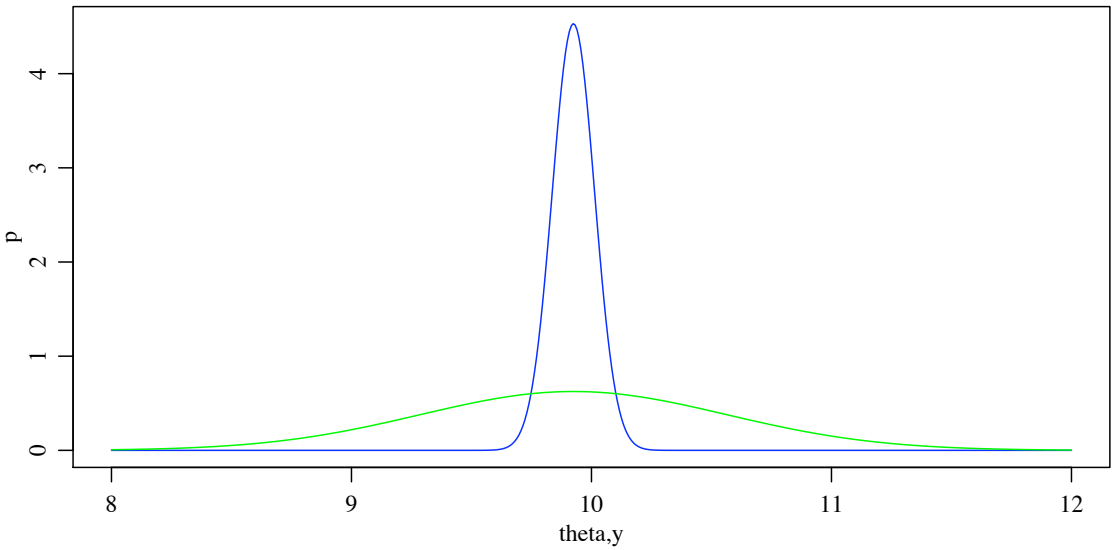


Figure 6.3: Posterior predictive distribution and distribution of the mean. Note the difference in variance.

6.2 Joint inference for the mean and variance

$$p(\theta, \sigma|y_1, \ldots, y_n) = p(y_1, \ldots, y_n|\theta, \sigma)p(\theta, \sigma)/p(y_1, \ldots, y_n)$$

So, what kind of *joint* priors make sense for θ, σ? We can always write

$$p(\theta, \sigma) = p(\theta|\sigma)p(\sigma)$$

Recall, for fixed σ, a conjugate prior for θ was normal (μ_0, τ_0). Consider

$$p(\theta|\sigma) = \text{ normal } (\mu_0, \tau_0 = \sigma/\sqrt{\kappa_0}),$$

i.e. $\tau_0^2 = \sigma^2/\kappa_0$, so

- $\mu_0 \approx$ "mean from prior observations " (as before);
- $\tau_0^2 = \sigma^2/\kappa_0 \approx$ "variance of mean from prior observations";
- $\kappa_0 \approx$ prior sample size.

Now, for σ^2 we need a prior that has support on $(0, \infty)$. A conjugate prior for σ^2 is the *inverse-gamma* prior:

$$\begin{array}{rcccc} \text{variance} & = & \sigma^2 & \sim & \text{inverse-gamma}(a, b) \\ \text{precision} & = & 1/\sigma^2 & \sim & \text{gamma}(a, b) \end{array}$$

For interpretability, we will parameterize this as

$$1/\sigma^2 \sim \text{ gamma } (\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2)$$

Under this parameterization,

- $\text{E}(\sigma^2) = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2-1}$;
- $\text{Mode}(\sigma^2) = \sigma_0^2 \frac{\nu_0/2}{\nu_0/2+1}$, so $\text{Mode}(\sigma^2) < \sigma_0^2 < \text{E}(\sigma^2)$;
- $\text{Var}(\sigma^2)$ is decreasing in ν_0.

As we will see in a moment, we can interpret the prior parameters (σ_0^2, ν_0) as the sample variance and sample size of prior observations.

Conjugate posterior inference for the normal model: Let

$$\begin{array}{rcl} 1/\sigma^2 & \sim & \text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2) \\ \theta|\sigma & \sim & \text{normal}(\mu_0, \sqrt{\sigma^2/\kappa_0}) \\ Y_1, \ldots, Y_n|\theta, \sigma & \sim & \text{i.i.d. normal } (\theta, \sigma) \end{array}$$

Then

- $\theta|y_1, \ldots, y_n, \sigma \sim \text{normal } (\mu_n, \sigma/\sqrt{\kappa_0 + n})$, where

$$\mu_n = \frac{(\kappa_0/\sigma^2)\mu_0 + (n/\sigma^2)\bar{y}}{\kappa_0/\sigma^2 + n/\sigma^2} = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n}$$

- $1/\sigma^2|y_1, \ldots, y_n \sim \text{gamma } (\nu_n/2, \nu_n\sigma_n^2/2)$, where
  - $\nu_n = \nu_0 + n$;
  - $\sigma_n^2 = \frac{1}{\nu_n}[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y} - \mu_0)^2]$.

Recall $s^2 = \sum(y_i - \bar{y})^2/(n-1)$.

Example (Commute times): 140 employed people in King county were asked about the amount of time they spend commuting during the last week. Y_i = average daily commute time, in minutes. Data from a previous survey of 100 people gave $\bar{y}_{\text{prior}} = 27.4$ and $s_{\text{prior}} = 13.2$ ($s^2 = 174.24$).

Prior: $p(\mu, \sigma^2) = \text{normal}\ (\mu_0, \sigma/\sqrt{\kappa_0}) \times \text{inverse-gamma}\ (\nu_0/2, \sigma_0^2\nu_0/2)$

- $\mu_0 = 27.4; \kappa_0 = 5\ (< 100)$;
- $\sigma_0^2 = 174.24; \nu_0 = 5$;

Data:

- $\bar{y} = 31.84$;
- $s = 13.226,\ s^2 = 174.93.$

Posterior inference: Need to find μ_n, σ_n^2 :

$$\mu_n = \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n} = \frac{5 \times 27.4 + 140 \times 31.84}{5 + 140} = 31.69$$

$$\sigma_n^2 = \frac{1}{\nu_n}[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2] = \frac{871.2 + 24315.53 + 95.236}{145} = 174.36$$

So

- $\theta|\{y_1, \ldots, y_n, \sigma\} \sim \text{normal}\ (31.69, \sigma/\sqrt{145})$;
- $1/\sigma^2|\{y_1, \ldots, y_n\} \sim \text{gamma}\ (145/2, 145 \times 174.36/2)$.

In R:

```
### prior
mu0<-27.4
k0<-5

s20<-174.24
nu0<-5

### data
n<-140
ybar<-31.84
```

```
s2<-174.93

### posterior inference
mun<- (k0*mu0 + n*ybar)/(k0+n)
s2n<- (nu0*s20 +(n-1)*s2 +k0*n*(ybar-mu0)^2/(k0+n))/(nu0+n)

> mun
[1] 31.68690
> s2n
[1] 174.3561
```

Usually we are primarily interested in θ, and would like to calculate

$$\mathrm{E}(\theta|\mathbf{y}), \quad \mathrm{sd}(\theta|\mathbf{y}), \quad \Pr(\theta_1 < \theta_2|\mathbf{y}_1, \mathbf{y}_2), \quad \text{etc.}$$

But all these quantities involve the *marginal* distribution of θ given $\mathbf{y}$. All we know (so far) is that the *conditional* distribution of θ given $\mathbf{y}$ and σ is normal. If we could generate *marginal* samples of θ, from $p(\theta|\mathbf{y})$, then we could use the Monte Carlo method to approximate these quantities.

Monte Carlo sampling: Sample

$$\begin{array}{cc} \sigma^2_{(1)} \sim \text{inverse gamma}(\nu_n/2, \sigma^2_n\nu_n/2) & \theta_{(1)} \sim \text{normal}(\mu_n, \sigma_{(1)}/\sqrt{\kappa_0+n}) \\ \vdots & \vdots \\ \sigma^2_{(m)} \sim \text{inverse gamma}(\nu_n/2, \sigma^2_n\nu_n/2) & \theta_{(m)} \sim \text{normal}(\mu_n, \sigma_{(m)}/\sqrt{\kappa_0+n}) \end{array}$$

Then

- $\{(\sigma_{(1)}, \theta_{(1)}), \ldots, (\sigma_{(m)}, \theta_{(m)})\}$ are samples from the joint posterior distribution of $p(\theta, \sigma|y_1, \ldots, y_n)$;
- $\{\theta_{(1)}, \ldots, \theta_{(m)}\}$ are samples from the marginal posterior distribution of $p(\theta|y_1, \ldots, y_n)$.

To see this, refer to the self-help exercise from the last chapter.

In R:

```
### Monte Carlo sampling
s2.postsample<-1/rgamma(5000,  (nu0+n)/2, s2n*(nu0+n)/2 )
theta.postsample<-rnorm(5000, mun, sqrt(s2.postsample/(k0+n)))
```

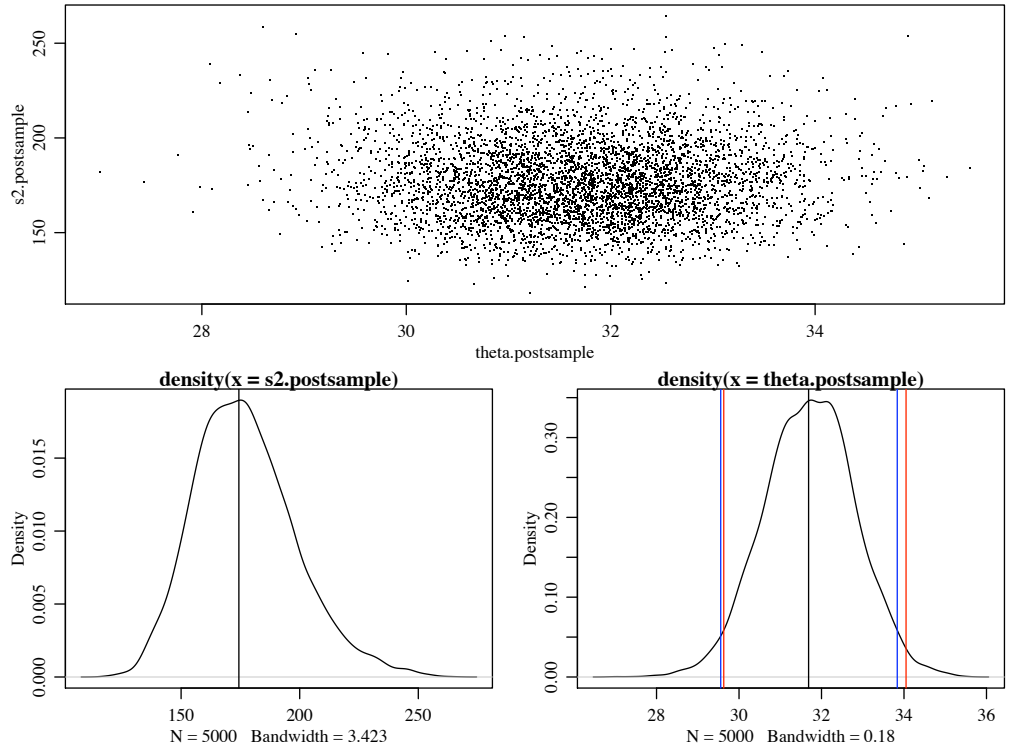


Figure 6.4: Monte Carlo estimates of the joint and marginal distributions of the mean and variance.

```
quantile( theta.postsample,c(.025,.975))

    2.5%    97.5%
29.59142 33.84020

### t-test based confidence interval
ybar+qt( c(.025,.975), n-1) *sqrt(s2/n)

[1] 29.62989 34.05011
```

Alternatively: It can be shown that

$$\begin{aligned} p(\theta|y_1,\ldots,y_n) &= \int p(\theta|\sigma,y_1,\ldots,y_n)p(\sigma|y_1,\ldots,y_n)\ d\sigma \\ &= \text{t-distribution}_{\nu_0+n}(\mu_n, \frac{\sigma_n^2}{\kappa_0+n}) \end{aligned}$$

6.3 A note on the bias-variance tradeoff

Consider for the moment estimating the mean θ of a population with normal sampling and prior distributions, in the case where σ^2 is known. Using a normal(μ_0, τ_0) prior for θ, the posterior mean of θ is

$$\hat{\theta}_b = \mu_n = \frac{n/\sigma^2}{n/\sigma^2 + 1/\tau_0^2}\bar{y} + \frac{1/\tau_0^2}{n/\sigma^2 + 1/\tau_0^2}\mu_0 = w\bar{y} + (1-w)\mu_0$$

Lets compare this to $\hat{\theta}_e = \bar{y}$: The *sampling properties* of these estimators describe their properties under hypothetically repeatable surveys/experiments.

$$\begin{aligned} E(\hat{\theta}_e|\theta = \theta_0) &= \theta_0 \ , \ \hat{\theta}_e \text{ is "unbiased"} \\ E(\hat{\theta}_b|\theta = \theta_0) &= w\theta_0 + (1-w)\mu_0 \ , \ \hat{\theta}_b \text{ is "biased"} \end{aligned}$$

but

$$\begin{aligned} V(\hat{\theta}_e|\theta = \theta_0, \sigma) &= \frac{\sigma^2}{n} \\ V(\hat{\theta}_b|\theta = \theta_0, \sigma) &= w^2 \times \frac{\sigma^2}{n} < \frac{\sigma^2}{n} \end{aligned}$$

so θ_b has lower variability. A better way to compare the two estimates is with expected (mean) square error (MSE) : How far away from θ_0 do you expect the estimate to be, given that $\theta = \theta_0$?

$$\begin{aligned} MSE[\hat{\theta}_e] = E[(\hat{\theta}_e - \theta)^2|\theta = \theta_0] &= \frac{\sigma^2}{n} \\ MSE[\hat{\theta}_b] = E[(\hat{\theta}_b - \theta)^2|\theta = \theta_0] &= E[\{w(\bar{y} - \theta_0) + (1-w)(\mu_0 - \theta_0)\}^2|\theta = \theta_0] \\ &= w^2 \times \frac{\sigma^2}{n} + (1-w)^2(\mu_0 - \theta_0)^2 \end{aligned}$$

With some algebra, you can show that $MSE[\hat{\theta}_b] < MSE[\hat{\theta}_e]$ if

$$(\mu_0 - \theta_0)^2 < \frac{\sigma^2}{n}\frac{1+w}{1-w}$$

or in the case of $\hat{\theta}_b$,

$$(\mu_0 - \theta_0)^2 < \frac{\sigma^2}{n} + 2\tau_0^2$$

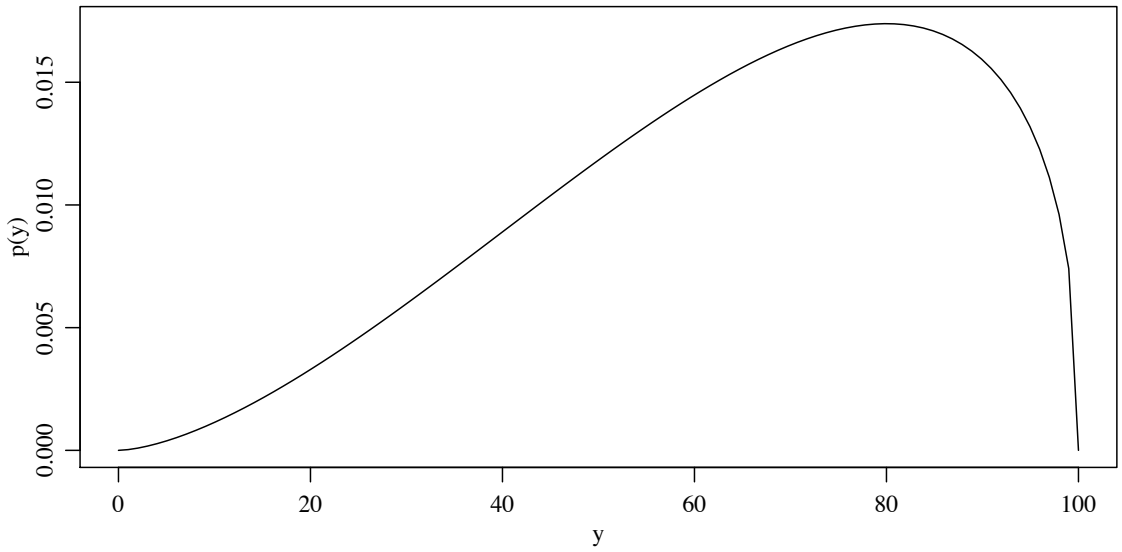


Figure 6.5: Distribution of standardized test scores

Some say that, if you know anything at all about the population under study, you should be able to find values of μ_0 and τ_0^2 such that this holds, and thus reduce the average (squared) distance from your estimate to the truth.

Example (Educational Testing): Scores on a certain standardized test range between 0 and 100. Developers of the test want to estimate the average score of students within the state, based on sample of size 10.

For this problem, we *know* that $0 \leq \theta_0 \leq 100$ and so if we pick $\mu_0 = 50$, then

$$(\mu_0 - \theta_0)^2 < 50^2 \text{ for any possible value of } \theta_0$$

Thus we are guaranteed that the estimator $\hat{\theta}_b$ will beat $\bar{y}$ in terms of MSE if

$$50^2 < \frac{\sigma^2}{n} + 2\tau_0^2$$

Even with $\sigma^2 = 0$, which it probably isn't, this result will hold with $\tau_0^2 = 50^2/2$. So lets consider setting $\tau_0 = 50/\sqrt{2} = 35.36$.

Simulation Study: Suppose the state-level distribution is as follows: This

distribution, unknown to the testers, has a mean of 65 and a standard deviation of $\sigma = 21.33$.

Consider simulating the testing experiment as follows:

- sample $y_1, \ldots, y_{10}$ from this distribution.
- compute $\hat{\theta}_e = \bar{y}$;
- compute $\hat{\theta}_b = w\bar{y} + (1-w)50$, where $w = \frac{n/s^2}{n/s^2+1/\tau_0^2}$.

```
a<- 4* 65/100
b<- 4*(100-65)/100

n<-10 ; mu0<- 50 ; t20<-50^2/2
mu.est<-matrix(nrow=nsim,ncol=2)

nsim<-100000
set.seed(1)
for(ns in 1:nsim)  {
y<-100*rbeta(n,a,b)
w<- t20/(t20+ var(y)/n)
mu.hat.e<- mean(y)
mu.hat.b<- mean(y) + (1-w)*(mu0 -mean(y))
mu.est[ns,]<- c(mu.hat.e,mu.hat.b)
                    }
```

```
> mean( (mu.est[,1] - m0)^2 )
[1] 44.97217
> mean( (mu.est[,2] - m0)^2 )
[1] 43.32431

> mean(mu.est[,1])
[1] 64.98616
> mean(mu.est[,2])
[1] 64.49509

> var(mu.est[,1])
[1] 44.97243
> var(mu.est[,2])
[1] 43.0698
```

$\text{MSE}(\hat{\theta}_b) = 43.32 < 44.97 = \text{MSE}(\hat{\theta}_e)$.

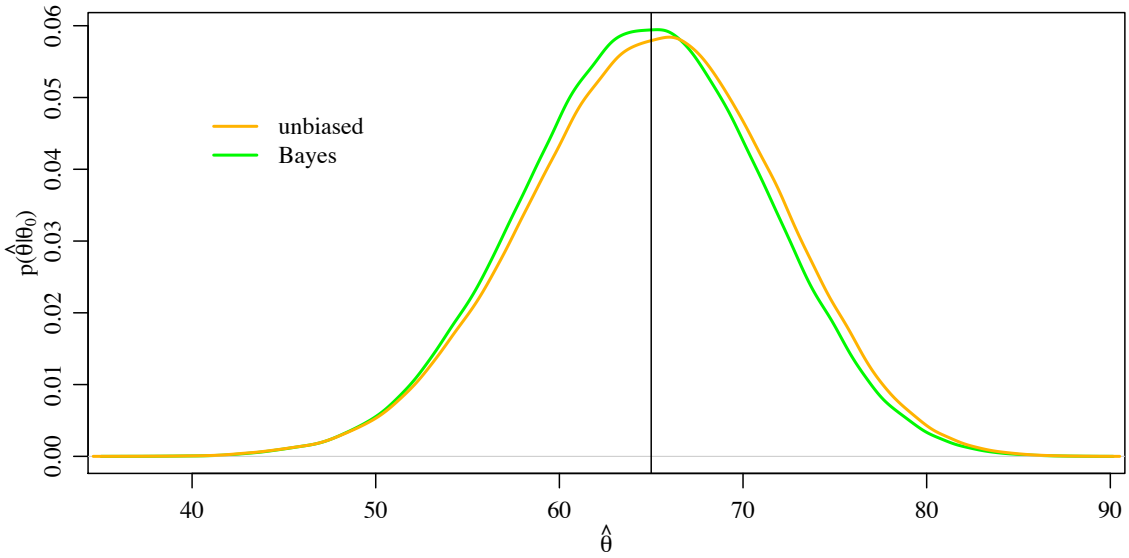


Figure 6.6: Sampling distributions of the unbiased and (plug-in) Bayes estimates in the educational testing example.

Is this surprising? Is the result due to the non-normality of the underlying data, or that we used a plug-in estimate of the variance?

Normal example: Consider the same setup, but where the population is actually normally distributed (i.e. the model is correct), with mean $\theta_0 = 65$ and standard deviation $\sigma = 21.33$. Suppose we use the conjugate prior for θ and σ:

$$p(\theta|\sigma) = \text{normal}(\mu_0, \sigma/\sqrt{\kappa_0})$$

with $\mu_0 = 50$ and $\kappa_0 = 1$, i.e. we are off in our information about the mean θ_0 by roughly a standard deviation of *the sampling distribution*.

Perform the following simulation study:

- sample $y_1, \ldots, y_{10}$ from this distribution.
- compute $\hat{\theta}_e = \bar{y}$;
- compute $\hat{\theta}_b = w\bar{y} + (1-w)50$, where $w = \frac{n}{n+\kappa_0} = 0.91$.

```
for(ns in 1:nsim)  {
```

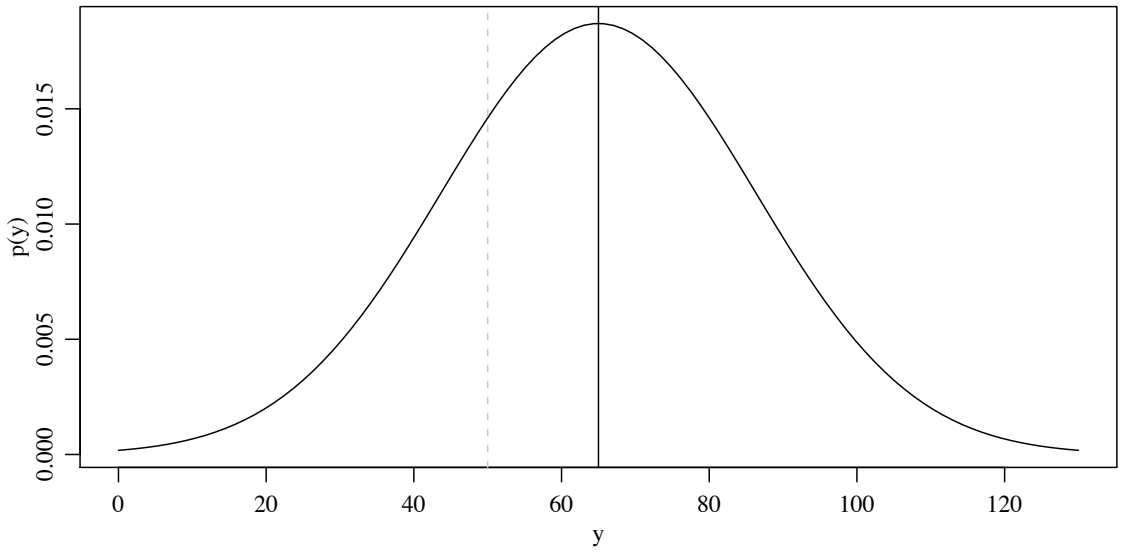


Figure 6.7: Normal sampling distribution for the educational testing example

```
y<-rnorm(n,m0,s0)
w<- n/(k0+ n)
mu.hat.e<- mean(y)
mu.hat.b<- mean(y) + (1-w)*(mu0 -mean(y))
mu.est[ns,]<- c(mu.hat.e,mu.hat.b)
                    }

> mean( (mu.est[,1] - m0)^2 )
[1] 45.32402
> mean( (mu.est[,2] - m0)^2 )
[1] 39.31489
```

$\text{MSE}(\hat{\theta}_b) = 39.31 < 45.32 = \text{MSE}(\hat{\theta}_e)$.

Recall: for the normal model with the conjugate prior, you can calculate what the MSE is

- $\text{MSE}(\hat{\theta}_e) = \frac{\sigma^2}{n}$
- $\text{MSE}(\hat{\theta}_b) = w^2 \times \frac{\sigma^2}{n} + (1-w)^2(\mu_0 - \theta_0)^2$.

with $w = \frac{n}{n+\kappa_0}$, and so the Bayes estimate will have lower MSE than the

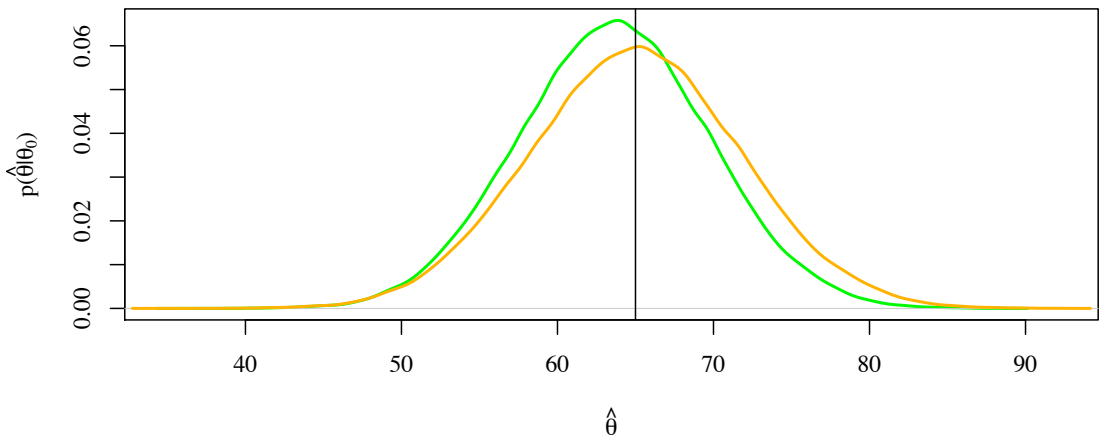


Figure 6.8: sampling distributions of the unbiased and conjugate Bayes estimates for the educational testing example

unbiased estimate if

$$(\mu_0 - \theta_0)^2 < \frac{\sigma^2}{n}\frac{1+w}{1-w} = \frac{\sigma^2}{n}\left(1 + 2\frac{n}{\kappa_0}\right) = \frac{\sigma^2}{n} + 2\frac{\sigma^2}{\kappa_0}$$

Discussion:

In one-sample problems like the ones discussed above, we have shown that Bayes estimates will beat $\bar{y}$ in terms of MSE if the amount by which the prior mean is off ($[\theta_0 - \mu_0]^2$) is bounded by some function of the prior precision (κ_0 or τ_0^2). In most applications we can't know in advance if this condition will be met, so how is Bayesian analysis useful?

Example: Let θ be the mean number of children per household in a residential neighborhood of Seattle having N households. Let $\sum y_1, \ldots, y_n$ be the number of children in each household from a random sample of n households , with $n \ll N$. Recall that $\bar{y} = \sum_{i=1}^{n} y_i/n$ is an unbiased estimator of θ.

Suppose $\bar{y} = 0$. Then your unbiased estimate of θ is $\hat{\theta} = 0$=no children in any household. This is your estimator for any $n \geq 1$. Is $\hat{\theta}$ a good estimator?

- Yes: It is unbiased.
- No: It gives values that quite possiby don't represent your beliefs.

What would a research who obtained $\bar{y} = 0$ do?

1. If the goal is simply to report the results of the survey, then they would just report $\bar{y}$,n and N.
2. If actual decisions have to be made about where to put schools/services/bus routes/etc. what should the researcher do? Possibilities include
   (a) presuming that $\theta = \frac{1}{N}\sum_{i=1}^{N} y_i = 0$ (i.e. using the unbiased estimate).
   (b) presuming θ is equal to some city-wide average;
   (c) presuming that θ is somewhat smaller than a city wide average.

In practice, people often make data-analysis and decisions in ad-hoc and arbitrary ways, and don't describe the motivations for their approaches. One of the goals of Bayesian inference is to remove this arbitrariness, and make inference *transparent*, although subjective. Furthermore, subjectivity can to some degree be removed by identifying what the posterior information or optimal decision is under any a variety of prior distributions. The statistician can always honestly say "*if* $p(\theta)$ is your prior *and* $p(y|\theta)$ is your sampling model, *then* your posterior is $p(\theta|y)$." If someone asks about a different prior, then the researcher can always provide the corresponding posterior.

6.4 Improper priors

What if you want to "use the Bayesian machinery" to calculate things like $P(\theta_a < \theta_b|\mathbf{y}_a, \mathbf{y}_b)$ but want to "be objective" by not using prior information.

Recall, we have referred to κ_0 and ν_0 as "prior sample sizes". What happens if we let $\kappa_0, \nu_0 \to 0$?

$$\begin{aligned}\mu_n &= \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n} \\ \sigma_n^2 &= \frac{1}{\nu_0 + n}[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2]\end{aligned}$$

So as $\kappa_0, \nu_0 \to 0$,

- $\mu_n = \bar{y}$;
- $\sigma_n^2 = \frac{n-1}{n}s^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$.

This has led some to advocate the following "posterior distribution":

- $1/\sigma^2|y_1, \ldots, y_n \sim$ gamma $(\frac{n}{2}, \frac{n}{2}\frac{1}{n}\sum(y_i - \bar{y})^2)$;
- $\theta|\sigma, y_1, \ldots, y_n \sim$ normal $(\bar{y}, \frac{\sigma^2}{n})$.

You can marginalize over σ to show that

$$\frac{\theta - \bar{y}}{s/\sqrt{n}}|y_1, \ldots, y_n \sim t_{n-1}.$$

Compare to the t-statistic :

$$\frac{\bar{y} - \theta}{s/\sqrt{n}}|\theta \sim t_{n-1}.$$

There are no priors that will lead to the above posterior distributions, so such an analysis is not Bayesian. Sometimes, taking limits like this leads to sensible answers which are approximately equal to Bayesian answers using vague prior information. Sometimes taking such limits leads to really bad answers. Each non-Bayesian estimator or procedure must be evaluated on a case-by-case basis.

6.5 Semi-conjugate prior distributions

In the previous section we modeled our uncertainty about θ as depending on σ:

$$p(\theta|\sigma) = \text{ normal } (\mu_0, \sigma/\sqrt{\kappa_0})$$

However, sometimes we might want to specify our uncertainty about θ independently of σ. For example:

$$p(\theta, \sigma) = p(\theta) \times p(\sigma)$$

- $\theta \sim$ normal (μ_0, τ_0) ;
- $1/\sigma^2 \sim$ gamma $(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2)$;

A few pages ago we showed that $\theta|\{\sigma, y_1, \ldots, y_n\} \sim \text{normal}(\mu_n, \tau_n)$ with

- $\mu_n = \frac{\mu_0/\tau_0^2 + n\bar{y}/\sigma^2}{1/\tau_0^2 + n/\sigma^2}$;
- $\tau_n^2 = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)^{-1}$.

In this case, the marginal density of $1/\sigma^2$ is *not* a gamma distribution (try it out). But you can show that

$$1/\sigma^2|\theta, y_1, \ldots, y_n \sim \text{ gamma}((\nu_0 + n)/2, (\nu_0\sigma_0^2 + \sum(y_i - \theta)^2)/2)$$

So we have nice forms for / can sample from

- $p(\theta|\sigma, y_1, \ldots, y_n)$
- $p(\sigma|\theta, y_1, \ldots, y_n)$

But how can we sample from the joint posterior?

Chapter 7

Posterior approximation 1

7.1 Gibbs sampling

In the conjugate normal model we could sample from the joint posterior distribution of (θ, σ) given $y_1, \ldots, y_n$ by sampling

$$\sigma | y_1, \ldots, y_n \text{ and then } \theta | \sigma, y_1, \ldots, y_n$$

Repeating this over and over generated i.i.d. samples from $p(\theta, \sigma | y_1, \ldots, y_n)$.

Unfortunately, for the semiconjugate prior in which $\theta \sim \text{normal}(\mu_0, \tau_0)$, $1/\sigma^2 \sim \text{gamma}(\nu_0/2, \nu_0\sigma_0^2/2)$, we don't have an easy way to draw samples from $\sigma | y_1, \ldots, y_n$ - we need to know θ to do this. However, the *full conditional distributions* of θ and σ are available and easy to sample from:

- $\theta | \sigma, y_1, \ldots, y_n \sim \text{normal } (\mu_n, \tau_n)$;
- $1/\sigma^2 | \theta, y_1, \ldots, y_n \sim \text{gamma } (\nu_n/2, \nu_n\sigma_n^2/2)$.

where τ_n depends on σ, and σ_n^2 depends on θ. These are called full conditionals because they give the conditional distribution of the parameter given everything else.

Consider the following algorithm: Given $\phi_m = \{\theta_m, \sigma_m\}$,

1. sample $\theta_{m+1} | y_1, \ldots, y_n, \sigma_m$, i.e. plug σ_m into the formula for τ_n;
2. sample $\sigma_{m+1} | y_1, \ldots, y_n, \theta_{m+1}$, i.e. plug θ_{m+1} into the formula for σ_n^2;
3. let $\phi_{m+1} = \{\theta_{m+1}, \sigma_{m+1}\}$.

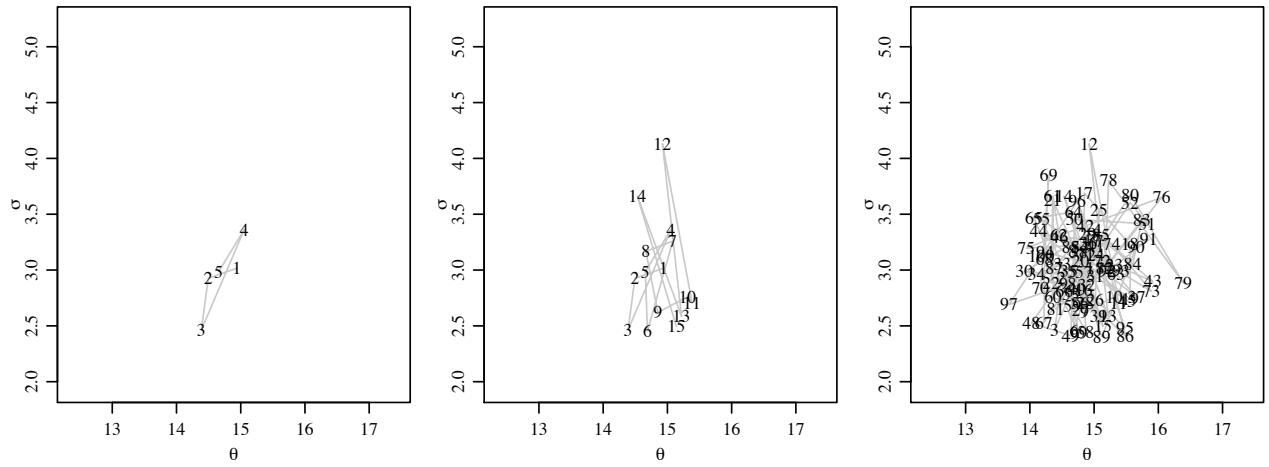


Figure 7.1: The first 5, 15 and 100 iterations of a Gibbs sampler.

This algorithm is called the **Gibbs sampler** , and generates a **dependent** sequence of our parameters $\phi_1, \phi_2, \ldots$.

```
### starting value
nscan<-1000
PHI<-matrix(nrow=nscan,ncol=2)
PHI[1,]<-phi<-c( mean.y, sqrt(var.y))

for(nsim in 1:(nscan-1)) {

### generate a new theta value from its full conditional
mn<- ( m0/t0^2 + n*mean.y/phi[2]^2) / ( 1/t0^2 + n/phi[2]^2 )
tn<- sqrt( 1/( 1/t0^2 + n/phi[2]^2 )  )
phi[1]<-rnorm(1, mn, tn )

### generate a new sigma value from its full conditional
a<- (nu0+n)/2
b<- (nu0*s20 + (n-1)*var.y + n*(mean.y-phi[1])^2 ) /2
phi[2]<-sqrt( 1/rgamma(1,a,b) )

PHI[nsim+1,]<-phi         }
```

Figure 7.2 plots of 500 such samples, plotting $\{\theta, 1/\sigma^2\}$ and $\{\theta, \sigma\}$. Figure 7.3 gives density estimates of the θ-samples and $1/\sigma^2$-samples, looked at separately (marginally).

Finally, lets find some empirical quantiles of our Gibbs samples:

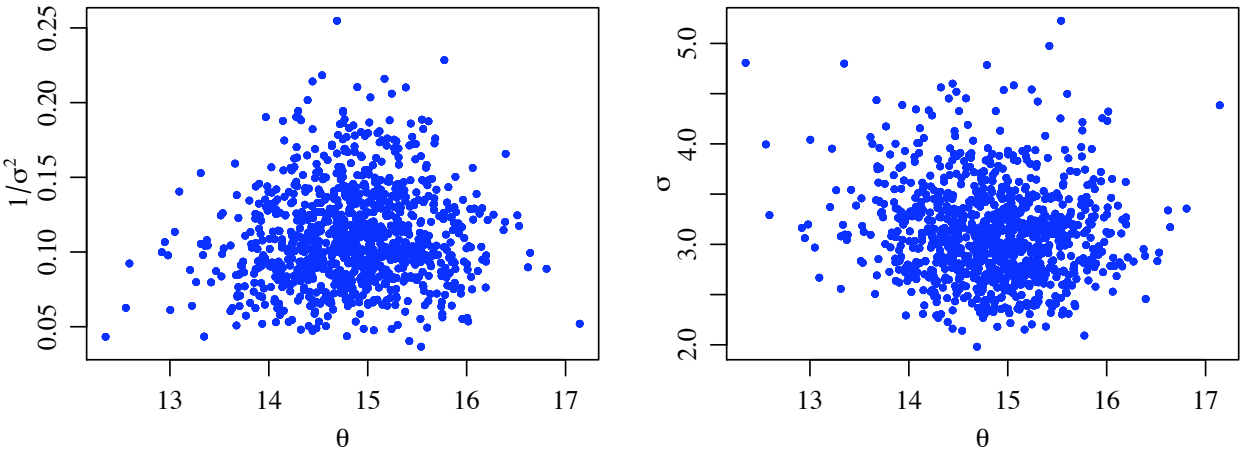


Figure 7.2: 1000 samples from the Gibbs sampler, giving a Monte Carlo approximation to the joint posterior.

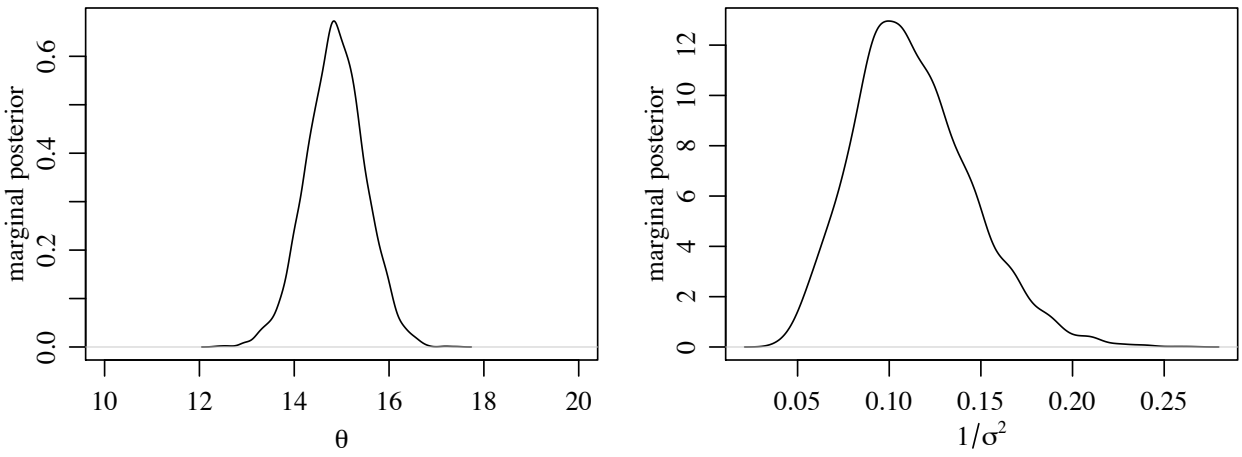


Figure 7.3: Monte Carlo estimates of the marginal densities of θ and $1/\sigma^2$.

```
> quantile(PHI[,1],c(.025,.975))
    2.5%    97.5%
13.64993 16.06671
> quantile(PHI[,2],c(.025,.975))
    2.5%    97.5%
2.330209 4.132244
```

More formally, the Gibbs sampler:

Suppose you have a vector of parameters $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_k\}$, and your information about $\boldsymbol{\phi}$ is measured with $p(\boldsymbol{\phi}) = p(\phi_1, \ldots, \phi_k)$.

Example: $\boldsymbol{\phi} = \{\theta, \sigma\}$, and the probability measure of interest is $p(\theta, \sigma|\boldsymbol{y})$.

The **Gibbs sampler** is implemented as follows:

A. Pick a starting point $\boldsymbol{\phi}_{(0)} = \{\phi_{1,(0)}, \ldots, \phi_{k,(0)}\}$

B. For $m = 0, \ldots, M-1$, generate $\boldsymbol{\phi}_{(m+1)}$ as follows:

1. sample $\phi_{1,(m+1)} \sim p(\phi_1|\phi_{2,(m)}, \phi_{3,(m)}, \ldots, \phi_{k,(m)})$;
2. sample $\phi_{2,(m+1)} \sim p(\phi_2|\phi_{1,(m+1)}, \phi_{3,(m)}, \ldots, \phi_{k,(m)})$;

$\vdots$

k. sample $\phi_{k,(m+1)} \sim p(\phi_k|\phi_{1,(m+1)}, \phi_{2,(m+1)}, \ldots, \phi_{k-1,(m+1)})$.

Notes:

- This generates a *dependent sequence* of vectors :

$$\begin{aligned} \boldsymbol{\phi}_{(1)} &= \{\phi_{1,(1)}, \ldots, \phi_{k,(1)}\} \\ \boldsymbol{\phi}_{(2)} &= \{\phi_{1,(2)}, \ldots, \phi_{k,(2)}\} \\ \vdots &= \vdots \\ \boldsymbol{\phi}_{(M)} &= \{\phi_{1,(M)}, \ldots, \phi_{k,(M)}\}. \end{aligned}$$

In this sequence, $\boldsymbol{\phi}_{(m+1)}$ depends on $\boldsymbol{\phi}_{(0)}, \ldots, \boldsymbol{\phi}_{(m)}$ only through $\boldsymbol{\phi}_{(m)}$, i.e. $\boldsymbol{\phi}_{(m+1)}$ is conditionally independent of $\boldsymbol{\phi}_{(0)}, \ldots, \boldsymbol{\phi}_{(m-1)}$ given $\boldsymbol{\phi}_{(m)}$. This is called the Markov property, and so the sequence is called a *Markov chain*.

- Under some conditions, no matter what $\boldsymbol{\phi}_0$ is,

$$\Pr(\boldsymbol{\phi}_{(M)} \in A) \rightarrow \int_A p(\boldsymbol{\phi})\, d\boldsymbol{\phi} \quad \text{as } M \rightarrow \infty.$$

  In other words, the *sampling distribution* of $\boldsymbol{\phi}_{(M)}$ approaches the *target distribution* as $M \rightarrow \infty$. Of course, some choices of $\boldsymbol{\phi}_0$ will "get you there sooner" than others.

- Perhaps most importantly,

$$\frac{1}{M}\sum_{m=1}^{M} g(\boldsymbol{\phi}_m) \rightarrow E[g(\boldsymbol{\phi})] = \int g(\boldsymbol{\phi})p(\boldsymbol{\phi})\, d\boldsymbol{\phi} \quad \text{as } M \rightarrow \infty$$

  This means we can approximate $E[g(\boldsymbol{\phi})]$ with the sample average of $g(\boldsymbol{\phi}_{(1)}), \ldots, g(\boldsymbol{\phi}_{(M)})$, just as in Monte Carlo approximation. For this reason, we call such approximations *Markov chain Monte Carlo* approximations, and the procedure an MCMC algorithm.

We will discuss practical aspects of MCMC in the context of specific models.

7.2 Grid-based approximation

Recall the semi-conjugate model described above, where $\phi = \{\theta, 1/\sigma^2\}$ are the parameters of interest. What do we know about $p(\phi|\boldsymbol{y})$?

$$\begin{aligned} p(\phi|\boldsymbol{y}) &= \frac{p(\phi)p(\boldsymbol{y}|\phi)}{p(\boldsymbol{y})} \\ &= p(\boldsymbol{y})^{-1} \times p(\theta)p(1/\sigma^2)p(\boldsymbol{y}|\theta, 1/\sigma^2) \\ &= p(\boldsymbol{y})^{-1} \times \text{normal}(\theta|\mu_0, \tau_0^2) \times \text{gamma}(1/\sigma^2|\frac{\nu_0}{2}, \sigma_0^2\frac{\nu_0}{2}) \times \prod_{i=1}^{n} \text{normal}(y_i|\theta, \sigma) \end{aligned}$$

How can we make use of this?

Grid based approximation method:

1. define a grid of points $\{\theta_1, \ldots, \theta_K\} \times \{1/\sigma_1^2, \ldots, 1/\sigma_L^2\}$;

2. compute $p(\theta_k)p(1/\sigma_l^2)p(\boldsymbol{y}|\theta_k,\sigma_l)$ for each pair $\{\theta_k, 1/\sigma_l^2\}$;

3. for each grid point $\{\theta, 1/\sigma^2\}$, let

$$\tilde{p}(\theta, 1/\sigma^2|\boldsymbol{y}) = \frac{p(\theta)p(1/\sigma^2)p(\boldsymbol{y}|\theta,\sigma)}{\sum_{k=1}^{K}\sum_{l=1}^{L} p(\theta_k)p(1/\sigma_l^2)p(\boldsymbol{y}|\theta_k,\sigma_l)}$$

$\tilde{p}(\theta, 1/\sigma^2|\boldsymbol{y})$ defined on $\{\theta_1,\ldots,\theta_K\}\times\{1/\sigma_1^2,\ldots,1/\sigma_L^2\}$, is a *discrete approximation* to $p(\theta, 1/\sigma^2|\boldsymbol{y})$.

Notes:

1. For points $\theta, 1/\sigma^2$ on the grid, $\tilde{p}(\theta, 1/\sigma^2|\boldsymbol{y})$ can be written

$$\tilde{p}(\theta, 1/\sigma^2|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta,\sigma)\tilde{p}(\theta, 1/\sigma^2)}{\tilde{p}(\boldsymbol{y})}, \text{ where}$$

   - $\tilde{p}(\theta, 1/\sigma^2) = p(\theta, 1/\sigma^2)/\sum_{k,l} p(\theta_k, 1/\sigma_l^2)$, a discrete prior based on $p(\theta, 1/\sigma^2)$;
   - $\tilde{p}(\boldsymbol{y}) = \sum_{k,l} p(\boldsymbol{y}|\theta_k,\sigma_l)\tilde{p}(\theta_k, 1/\sigma_l^2)$, the marginal/predictive distribution of $\boldsymbol{y}$, based on the discrete prior $\tilde{p}(\theta, 1/\sigma^2)$;

2. Relative probabilities: Consider a pair of points on the grid $\{\theta, 1/\sigma^2\}_a$ and $\{\theta, 1/\sigma^2\}_b$.

$$\frac{\tilde{p}(\theta_a, 1/\sigma_a^2|\boldsymbol{y})}{\tilde{p}(\theta_b, 1/\sigma_b^2|\boldsymbol{y})} = \frac{p(\theta_a)p(1/\sigma_a^2)p(\boldsymbol{y}|\theta_a,\sigma_a)}{p(\theta_b)p(1/\sigma_b^2)p(\boldsymbol{y}|\theta_b,\sigma_b)} = \frac{p(\theta_a, 1/\sigma_a^2|\boldsymbol{y})}{p(\theta_b, 1/\sigma_b^2|\boldsymbol{y})}$$

3. Posterior inference

   - $\tilde{E}(g(\theta,\sigma)|\boldsymbol{y}) = \sum_{k,l} g(\theta_k,\sigma_l)\tilde{p}(\theta_k, 1/\sigma_l^2|\boldsymbol{y})$;
   - $\tilde{p}(\theta|\boldsymbol{y}) = \sum_l \tilde{p}(\theta, 1/\sigma_l^2|\boldsymbol{y})$;

4. Sampling from $\tilde{p}$ is easy using the `sample` command.

5. WARNING: The grid must be chosen carefully. A grid that is too small or in the wrong part of the parameter space can lead to poor approximations of $p(\theta, 1/\sigma^2)$.

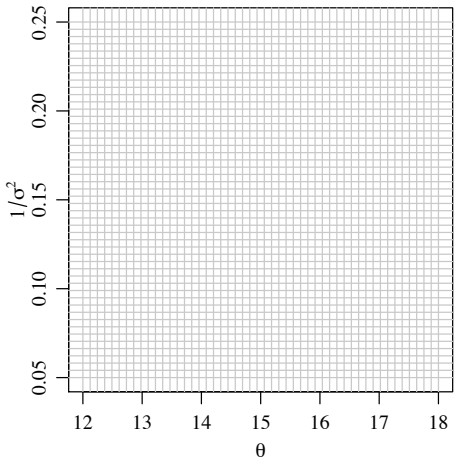


Figure 7.4: A grid on which we will make a discrete approximation to the posterior.

Example: Let

- $\theta \sim \text{normal}(\mu_0, \tau_0)$;
- $1/\sigma^2 \sim \text{gamma}(\nu_0/2, \sigma_0^2\nu_0/2)$;

with $(\mu_0, \tau_0) = (10, 5)$ and $(\sigma_0^2, \nu_0) = (2, 1)$.

Constructing $\tilde{p}$: First, construct a grid of evenly spaced θ and $1/\sigma^2$ values:

```
K<-50 ; L<-50
theta.grid<-seq( 12,18, length=K)
prec.grid<-seq(.05,.25, length=L)
post.grid<-matrix(nrow=K,ncol=L)
```

Now, compute the value of $\tilde{p}$ at each grid point:

```
for(k in 1:K) {
for(l in 1:L) {
    post.grid[k,l]<- dnorm(theta.grid[k],mu0,tau0) *
                     dgamma(prec.grid[l], nu0/2, s20*nu0/2 ) *
                     prod(dnorm(y,theta.grid[k],1/sqrt(prec.grid[l])))  }}

post.grid<-post.grid/sum(post.grid)
```

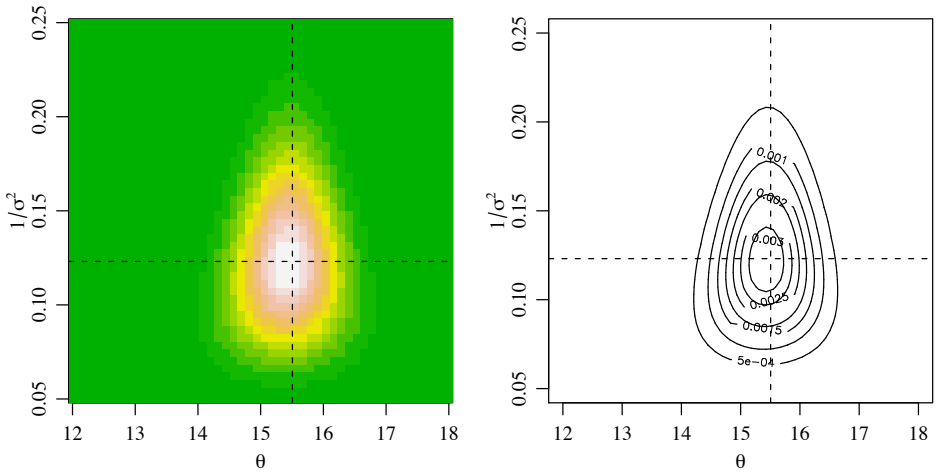


Figure 7.5: Two pictures of the discretized posterior distribution.

We can get nice pictures of the joint posterior with the `image` and `contour` commands.

```
image(theta.grid,prec.grid,post.grid)
contour(theta.grid,prec.grid,post.grid)
```

Discrete marginal distributions are obtained by calculating the margins of the joint posterior matrix:

```
posttheta.grid<- theta.grid*0
for(k in 1:K) { posttheta.grid[k]<- sum( post.grid[k,] )  }

postprec.grid<- prec.grid*0
for(l in 1:L) { postprec.grid[l]<- sum( post.grid[,l] )  }
```

Based on these plots, what do you think about the grid choice?

We can take i.i.d. samples directly from the discrete posterior using `sample`:

```
> kl<- cbind( rep(1:K,times=rep(L,K) ), rep( 1:L,times=K) )
> kl[1:5,]

      [,1] [,2]
 [1,]    1    1
 [2,]    1    2
 [3,]    1    3
```

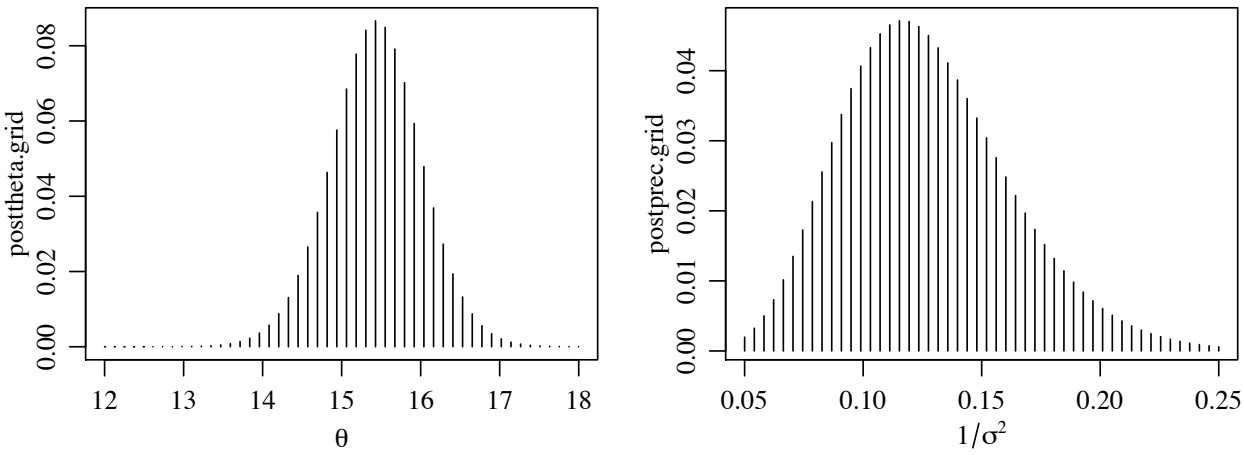


Figure 7.6: Discretized versions of the marginal posterior distributions.

```
 [4,]    1    4
 [5,]    1    5

> kl.samp<-sample( 1:(K*L), 500, prob=c(t(post.grid)), replace=T)
> post.sample<- cbind( theta.grid[ kl[kl.samp,1]], prec.grid[ kl[kl.samp,2] ] )

> quantile( post.sample[,1], c(.025, .975))
    2.5%    97.5%
13.67347 16.12245

> quantile( 1/sqrt(post.sample[,2]), c(.025, .975))
    2.5%    97.5%
2.357850 4.198321
```

We can also do the sampling with a marginal and a conditional:

```
l.prec<-sample(1:L, 500, prob=postprec.grid,replace=T)
k.theta<-rep(0,500)
for(ns in 1:500) { k.theta[ns]<-sample(1:K, 1, prob=post.grid[,l]) }
post.sample<- cbind( theta.grid[k.theta], prec.grid[l.prec] )
```

Difficulties with grid-based approximations:

- The grid needs to be fine, and cover a large range. A coarse, small grid can give biased results. In particular, one needs to be careful about estimating quantiles or anything based on the tails of the distribution.

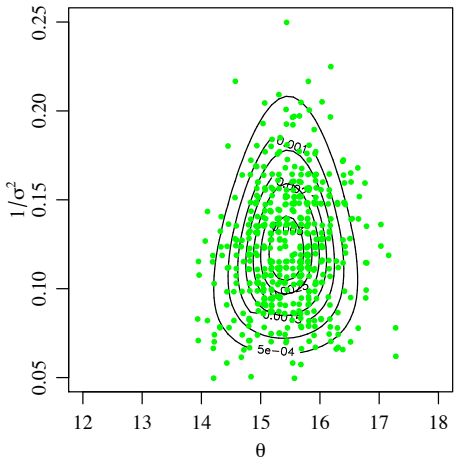


Figure 7.7: Samples from the discrete posterior distribution.

- If there are d parameters, need a d-dimensional array to store the posterior. With 5 parameters and 100 grid points each this is $100^5 = 10^{10}$ = 10 billion points total.

7.3 A refresher on sampling

Suppose you have two urns, labeled H and T. Within each urn there are blue, green and red balls. Let $X \in \{B, G, R\}$, $Y \in \{H, T\}$. Suppose there is the following joint distribution on X, Y:

	H	T	
B	p_{BH}	p_{BT}	p_B
G	p_{GH}	p_{GT}	p_G
R	p_{RH}	p_{RT}	p_R
	p_H	p_T	

Suppose you do the following:

1. sample $Y \in \{H, T\}$ from $P(Y)$ (i.e. flip a coin)
2. sample X from $P(X|Y = H)$, i.e.
   - If Y is H, sample X from $P(X|Y = H)$ (sample a ball from urn H);

- If Y is T, sample X from $P(X|Y = T)$ (sample a ball from urn T).

Joint sampling via a marginal and conditional: Question: what is the probability that your sample, X, Y, is equal to B, T? Well, you performed two physically independent actions here - flipping a coin, then drawing a ball from an urn. The probability that you flipped an H is $P(Y = H)$, and the probability that, given you are sampling from urn H, you draw a B, is $P(X = B|Y = H)$. Thus hopefully is seems logical that $P(X = B, Y = H) = P(Y = H)p(X = B|Y = H)$. Thus to sample from a joint distribution, I can sample from the marginal of one variable then sample from the conditional of the other.

Marginal sampling from another marginal and a conditional: Question: what is the probability that the ball you draw from the above procedure is B? Again, elementary probability tells us that it is $P(X = B)$.

Is the ball you drew a sample from the marginal of X or the conditional of X given Y? It is both a sample from the marginal, and it is also a sample from the conditional distribution of X given Y for *the particular value of Y that occurred in the procedure*. Here's one way of thinking about it:

Marginally: If I were to repeat the procedure over and over, the fraction of times I draw B, G, R balls is p_B, p_G, p_R. Thus the procedure is a way to draw samples from the marginal.

Conditionally: If I were to repeat the procedure over and over, and look *only at the balls I drew when $Y = H$* , then the fraction of times B, G, R appear in my sample will be $p_{BH}/p_H, p_{GH}/p_H, p_{RH}/p_H$, the conditional probabilities.

Now lets get back to parameters and data:

Sampling from a joint distribution: Suppose we have k parameters in our model, represented by the vector $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_k\}$. The posterior distribution of $\boldsymbol{\phi}$ given $\boldsymbol{y}$ can be factored as

$$p(\boldsymbol{\phi}|\boldsymbol{y}) = p(\phi_1|\boldsymbol{y}) \times p(\phi_2|\boldsymbol{y}, \phi_1) \times p(\phi_3|\boldsymbol{y}, \phi_1, \phi_2) \times \cdots \times p(\phi_k|\boldsymbol{y}, \phi_1, \ldots, \phi_{k-1})$$

In fact, the joint distribution can be factored in any order (i.e. the labels on the ϕ's do not matter). Similarly, $\boldsymbol{\phi}$ can be sampled from its distribution given $\boldsymbol{y}$ in a factored fashion:

1. sample $\phi_1 \sim p(\phi_1|\boldsymbol{y})$;
2. sample $\phi_2 \sim p(\phi_2|\boldsymbol{y}, \phi_1)$;
3. sample $\phi_3 \sim p(\phi_3|\boldsymbol{y}, \phi_1, \phi_2)$;

   $\vdots$

k. sample $\phi_k \sim p(\phi_k|\boldsymbol{y}, \phi_1, \ldots, \phi_{k-1})$;

Again, the labels on the ϕ's don't matter. All that is required is that you can actually sample from these distributions.

Example: Conjugate normal model.

$$\begin{aligned} p(\theta, 1/\sigma^2|\boldsymbol{y}) &= p(1/\sigma^2|\boldsymbol{y}) \times p(\theta|\boldsymbol{y}, 1/\sigma^2) \\ &= \text{gamma}(\nu_n/2, \sigma_n^2\nu_n/2) \times \text{normal}(\mu_n, \sigma/\sqrt{\kappa_n}) \end{aligned}$$

To sample $\{\theta, 1/\sigma^2\}$ from its joint posterior, you can sample $1/\sigma^2|\boldsymbol{y}$, then $\theta|\boldsymbol{y}, 1/\sigma^2$. Alternatively, we could have written:

$$p(\theta, 1/\sigma^2|\boldsymbol{y}) = p(\theta|\boldsymbol{y}) \times p(1/\sigma^2|\boldsymbol{y}, \theta)$$

but $p(\theta|\boldsymbol{y})$ is a bit more complicated than $p(\theta|\boldsymbol{y}, 1/\sigma^2)$. We have chosen the order of the factorization for computational convenience.

Marginal sampling: Suppose we are interested in $\phi_1|\boldsymbol{y}$, but its hard to draw samples from this distribution. Perhaps its easier to draw samples from $\phi_1|\{\boldsymbol{y}, \phi_2\}$. How does this help? Look at the equation for the marginal distribution of $\phi_1|\boldsymbol{y}$:

$$p(\phi_1|\boldsymbol{y}) = \int p(\phi_1|\boldsymbol{y}, \phi_2)p(\phi_2|\boldsymbol{y})\ d\phi_2$$

We see that the marginal distribution $p(\phi_1|\boldsymbol{y})$ is a *weighted average* of the full conditional $p(\phi_1|\boldsymbol{y}, \phi_2)$, averaged over values of ϕ_2. This suggests the following way to sample a value of $\phi_1|\boldsymbol{y}$:

1. sample $\phi_2|\boldsymbol{y}$
2. sample $\phi_1|\boldsymbol{y}, \phi_1$.

The resulting value of ϕ_1 is indeed a sample from $p(\phi|\boldsymbol{y})$. It is also a sample from $p(\phi_1|\boldsymbol{y}, \phi_2)$ *for this particular* ϕ_2 .

7.4 Data, posterior analysis and Monte Carlo methods

Here are the steps in a Bayesian statistical analysis:

Model specification Specify a model for your data: $p(\boldsymbol{y}|\phi)$ should represent the sampling distribution of your data, given a specific set of parameters ϕ.

Prior elicitation Specify a prior distribution: $p(\phi)$ should ideally represent someones (potential) prior information about the population parameter ϕ.

At this point, the posterior $p(\phi|\boldsymbol{y})$ is "determined." It is given by

$$p(\phi|\boldsymbol{y}) = \frac{p(\phi)p(\boldsymbol{y}|\phi)}{p(\boldsymbol{y})} = \frac{p(\phi)p(\boldsymbol{y}|\phi)}{\int p(\phi)p(\boldsymbol{y}|\phi)\ d\phi}$$

And so in a sense there is no more modeling. All that is left is

Examination of the posterior distribution Compute posterior means medians, modes, probabilities and confidence regions, all derived from $p(\phi|\boldsymbol{y})$.

Now sometimes, $p(\phi|\boldsymbol{y})$ is complicated, hard to write down, etc. The basic method for "looking at" $p(\phi|\boldsymbol{y})$ in this case is by studying Monte-Carlo samples from $p(\phi|\boldsymbol{y})$. So, Monte Carlo sampling algorithms (and MCMC sampling)

- are not models, nor do they generate "more information" than is in $\boldsymbol{y}$ and $p(\phi)$;
- they are simply "ways of looking at" $p(\phi|\boldsymbol{y})$.

For example, if we have Monte Carlo samples $\phi_{(1)}, \ldots, \phi_{(S)}$ that are approximate draws from $p(\phi|y)$, then these samples help describe $p(\phi|\boldsymbol{y})$:

- $\frac{1}{S}\sum \phi_{(s)} \approx \int \phi p(\phi|y)\ d\phi$
- $\frac{1}{S}\sum 1(\phi_{(s)} \leq c) \approx P(\phi \leq c|y) = \int_{-\infty}^{c} p(\phi|y)\ d\phi.$

and so on. So keep in mind, these Monte Carlo procedures are simply ways to approximate/look at $p(\phi|y)$.

Chapter 8

Hierarchical modeling

Suppose you are gathering data on

- people within several neighborhoods, or
- counties within several states, or
- children with several schools, or in general
- units within several groups.

Then your data is called *hierarchical* or *multi-level* data. The simplest type of multi-level data is two-level data, in which one level consists of *groups* and the other level consists of *units within groups*. In this case we denote $y_{i,j}$ as the observation on the ith unit in group j.

8.1 ANOVA

Hypothesis testing and model selection in the context of ANOVA assumes that the *within-group variation* is well-described by a normal distribution: For group j, $j = 1, \ldots, J$

$$y_{1,j}, \ldots, y_{n_j,j} | \theta_j, \sigma \sim \text{ i.i.d. normal}(\theta_j, \sigma).$$

This means that we are representing the within-group variability of the observations in group j with a normal distribution. Letting $\bar{y}_{\cdot,j} = \frac{1}{n_j} \sum y_{i,j}$, this means that

$$\bar{y}_{\cdot,j} | \theta_j, \sigma \sim \text{normal}(\theta_j, \sigma_j),$$

where $\sigma_j = \sigma / \sqrt{n_j}$.

Classical ANOVA: θ_j's are estimated under one of two different assumptions:

- Groups are unrelated: $\hat{\theta}_j = \bar{y}_{\cdot,j}$;
- Groups are very similar: $\hat{\theta}_{j_1} = \hat{\theta}_{j_2} = \bar{y}_{\cdot,\cdot} = \frac{\sum\sum y_{i,j}}{\sum n_j} = \frac{\sum \bar{y}_{\cdot,j}/\sigma_j^2}{\sum 1/\sigma_j^2}$.

Which assumption to use? The classical decision procedure is made via analysis of variance (ANOVA). In the simple case of constant sample size, in which $n_j = n$ for all j, and letting τ^2 be the variance of $\theta_1, \ldots, \theta_J$, we have

source of variation	sum of squares	mean square	E[mean square]
between group variation	SSB=$n\sum_j(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2$	MSB=SSB/$J-1$	$\sigma^2 + n\tau^2$
within group variation	SSW=$\sum_i\sum_j(\bar{y}_{i,j} - \bar{y}_{\cdot}j)^2$	MSW=SSW/$J(n-1)$	σ^2

Recall how ANOVA typically works:

- if variability of $\bar{y}_{\cdot 1}, \ldots, \bar{y}_{\cdot J}$ is *large* compared to within group variance, i.e. MSB $\gg$ MSW,
  - this is evidence that $\tau^2 > 0$;
  - reject $H_0 : \tau^2 = 0$, estimate $\hat{\theta}_j = \bar{y}_{\cdot j}$.
- if variability of $\bar{y}_{\cdot 1}, \ldots, \bar{y}_{\cdot J}$ is *about the same as* within group variance, i.e. MSB $\approx$ MSW,
  - this is evidence that $\tau^2 \approx 0$;
  - accept $H_0 : \tau^2 = 0$, estimate $\hat{\theta}_j = \bar{y}_{\cdot\cdot}$.

The result is one of two extremely different sets of estimates: no sharing of information across groups *or* complete sharing of info across groups. Consider instead an estimate of the type:

$$\hat{\theta}_j = \lambda_{1j}\bar{y}_{\cdot j} + \lambda_{2j}\bar{y}_{\cdot\cdot} + \lambda_{3j}\mu_0$$

where $\lambda_{1j} + \lambda_{2j} + \lambda_{3j} = 1$, where the λ's might depend on

- sample size,
- variance within groups σ^2,
- variance between groups τ^2.

Such weighted estimates are obtained via hierarchical models.

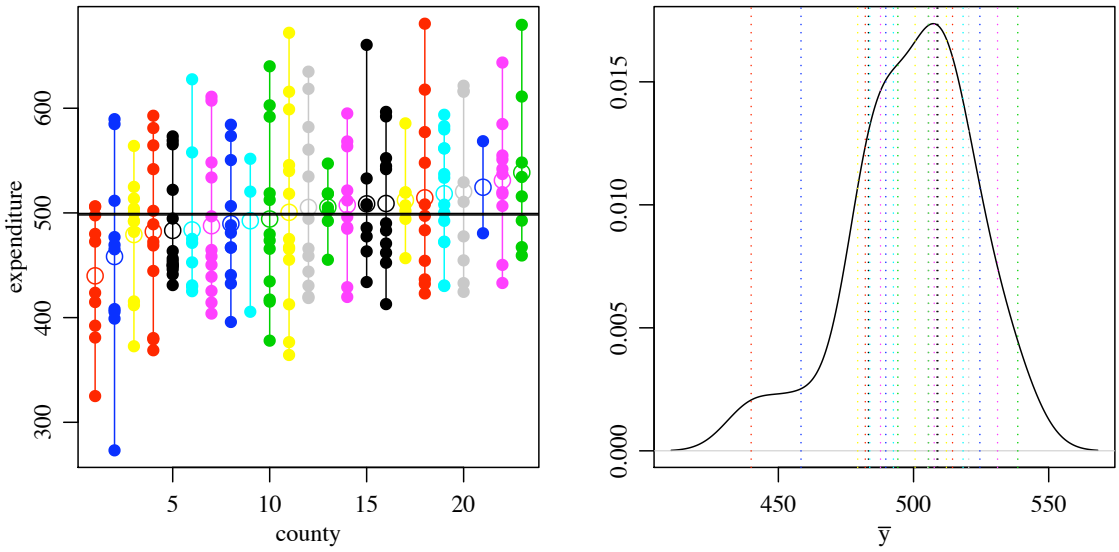


Figure 8.1: Expenditure data from 23 counties with between 2 and 14 respondents per county. The right panel gives the empirical distribution of the county-specific sample means.

8.2 Exchangeability

Example (household expenditures) : Let $y_{i,j}$ = be the amount spent on household expenditures in two weeks, for household i, $i = 1, \ldots, n_j$ in county j, $j = 1, \ldots, m$. Consider a model for $p(y_{1,j}, \ldots, y_{n_j,j})$:

- Does $y_{2,j}, \ldots, y_{n_j,j}$ give you any information about $y_{1,j}$? If so, we probably want $p(y_{1,j}) \neq p(y_{1,j}|y_{2,j}, \ldots, y_{n_j,j})$ i.e. some dependence;
- In random sampling , exchangeability *within* a county might be appropriate: $p(y_{1,j}, \ldots, y_{n_j,j}) = p(y_{\pi 1,j}, \ldots, y_{\pi n_j,j})$.

By de Finetti's theorem, we can then (approximately) model the data within county j as conditionally i.i.d. given some county-specific parameter ϕ_j.

$$y_{\pi 1,j}, \ldots, y_{\pi n_j,j}|\phi_j \sim \text{ i.i.d. } p(y|\phi_j)$$

How do we represent our uncertainty about $\phi_1, \ldots, \phi_J$? Again, if the group labels don't convey information (by themselves), but knowing $\phi_2, \ldots, \phi_J$

gives information about ϕ_1, then exchangeability would be appropriate. Applying de Finetti's theorem again gives

$$\phi_1, \ldots, \phi_J | \psi \sim \text{ i.i.d. } p(\phi|\psi)$$

for some parameter ψ and sampling distribution $p(\phi|\psi)$.

So to summarize:

$$\begin{aligned} y_{1,j}, \ldots, y_{n_j,j} \text{ exchangeable} &\Leftrightarrow y_{1,j}, \ldots, y_{n_j,j} | \phi_j \sim \text{i.i.d. } p(y|\phi_j) \\ \phi_1, \ldots, \phi_J \text{ exchangeable} &\Leftrightarrow \phi_1, \ldots, \phi_J | \psi \sim \text{i.i.d. } p(\phi|\psi) \end{aligned}$$

for some parameters $\psi, \phi_1, \ldots, \phi_J$.

8.3 The Hierarchical normal model

The hierarchical normal model corresponds to particular choices of the sampling distributions $p(y|\phi_j)$ and $p(\phi|\psi)$:

- $\phi_j = \{\theta_j, \sigma\}$, $p(y|\phi_j) =$ normal (θ_j, σ), representing sampling variability of data within a group;
- $\psi = \{\mu, \tau\}$, $p(\theta|\mu, \tau) =$ normal (μ, τ) , representing sampling variability of the group-specific means.

It might help to visualize it as Figure 8.2. Such a model is called a hierarchical model because there is a hierarchy of sampling variability. Some people call it a multi-level model because there are multiple levels of sampling variability.

The unknown but fixed parameters in this model are μ, τ and σ. For now we use standard semiconjugate priors for these parameters. Under these priors, the specification of the full probability model for all quantities is given as follows:

$$\begin{aligned} \text{Within group sampling distribution:} \quad & y_{1,j}, \ldots, y_{n_j,j} | \theta_j, \sigma \sim \text{ i.i.d. normal } (\theta_j, \sigma) \\ \text{Between group sampling distribution:} \quad & \theta_1, \ldots, \theta_J | \mu, \tau \sim \text{ i.i.d. normal } (\mu, \tau) \\ \text{Prior distributions:} \quad & 1/\sigma^2 \sim \text{ gamma } (\nu_0/2, \nu_0\sigma_0^2/2) \\ & 1/\tau^2 \sim \text{ gamma } (\eta_0/2, \eta_0\tau_0^2/2) \\ & \mu \sim \text{ normal } (\mu_0, \tau/\sqrt{\kappa_0}) \end{aligned}$$

We have probability distributions for lots of things, $y_{i,j}$'s θ_j's, μ , τ, and σ. Keep in mind that these distributions represent

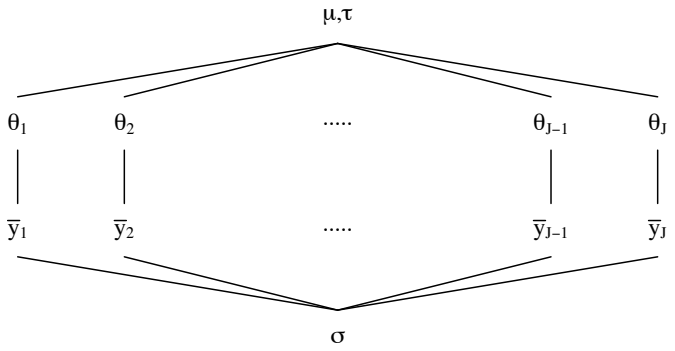


Figure 8.2: The graphical representation of the basic hierarchical model

- Sampling variation in the case of $y_{i,j}$'s and the θ_j's;
- Prior/posterior information about fixed quantities in the case of μ, τ, σ.

8.4 Posterior inference

Goal: Draw samples of $\{\theta_1, \ldots, \theta_J, \mu, \tau, \sigma\}$ conditional on $\boldsymbol{y}$.

Not surprisingly, it is difficult to draw independent samples from the joint posterior (although not too difficult, see me for details). We will use the Gibbs sampler to make dependent, approximate draws from this target distribution. To do this, we will need the full conditionals of each parameter. Useful for this will be the following factorization:

$$\begin{aligned} p(\theta_1, \ldots, \theta_J, \mu, \tau, \sigma|\boldsymbol{y}) &\propto p(\sigma)p(\tau|\sigma)p(\mu|\tau,\sigma)p(\theta_1, \ldots, \theta_J|\mu,\tau,\sigma)p(\boldsymbol{y}|\theta_1, \ldots, \theta_J, \mu, \tau, \sigma) \\ &= p(\sigma)p(\tau)p(\mu|\tau)p(\theta_1, \ldots, \theta_J|\mu,\tau)p(\boldsymbol{y}|\theta_1, \ldots, \theta_J, \sigma) \\ &= p(\sigma)p(\tau)p(\mu|\tau)\left\{\prod_{j=1}^{J} p(\theta_j|\mu,\tau)\right\}\left\{\prod_{j=1}^{J}\prod_{i=1}^{n_j} p(y_{i,j}|\theta_j,\sigma)\right\} \end{aligned}$$

Also helpful in finding full conditional distributions will be the following result: If $p(x, y, z)$ is the joint distribution of three random variables x, y, z, and can be written as $p(x, y, z) = f(x, z) \times g(y, z) \times h(z)$ for some functions f, g, h, then x and y are conditionally independent given z.

8.4.1 Full conditional of θ_j:

The full conditional of θ_j depends on $\{\mu, \tau, \sigma\}$ and data from group j. Conditional on these parameters, it is independent of the other θ's and data from other groups.

The parts of the above equation that involve θ_j are

$$p(\theta_j|\mu, \tau, \sigma, \boldsymbol{y}_j) \propto p(\theta_j|\mu, \tau) \times \prod_{i=1}^{n_j} p(y_{i,j}|\theta_j, \sigma)$$

This is exactly the same form as in the one-sample normal problem, so $\theta_j|\mu, \tau, \sigma, \boldsymbol{y}_j \sim$ normal $(\hat{\theta}_j, \sigma_{n_j})$, where

- $\hat{\theta}_j = \frac{\bar{y}_{\cdot j} n_j/\sigma^2 + \mu/\tau^2}{n_j/\sigma_j^2 + 1/\tau^2}$;
- $\sigma_{n_j}^2 = (n_j/\sigma^2 + 1/\tau^2)^{-1}$.

8.4.2 Full conditional of σ^2

Similar to the one sample normal model, except now we have information about σ from J populations:

$$\begin{aligned} p(\sigma|\theta_1, \ldots, \theta_J, \boldsymbol{y}) &\propto p(\sigma) \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(y_{i,j}|\theta_j, \sigma) \\ &\propto (\sigma^2)^{-\nu_0/2+1} e^{-\frac{\nu_0\sigma_0^2}{2\sigma^2}} (\sigma^2)^{-\sum n_j/2} e^{-\frac{\sum\sum(y_{i,j}-\theta_j)^2}{2\sigma^2}} \end{aligned}$$

Collecting terms,

$$1/\sigma^2|\boldsymbol{\theta}, \boldsymbol{y} \sim \text{inverse} - \text{gamma}[\frac{1}{2}(\nu_0 + \sum n_j), \frac{1}{2}(\nu_0\sigma_0^2 + \sum\sum(y_{i,j} - \theta_j)^2)]$$

Note $\sum\sum(y_{i,j} - \theta_j)^2$ is the within group sum of squares, conditional on the within-group means.

8.4.3 Full conditional of μ, τ

The part of the posterior that depends on μ, τ is

$$p(\tau)p(\mu|\tau)\prod_{j=1}^{J} p(\theta_j|\mu,\tau)$$

Note this is exactly a one sample normal problem with a conjugate prior. It is just a sample of θ's instead of a sample of y's.

$$\mu|\boldsymbol{\theta},\tau \sim \text{normal}(\mu_J, \tau/\sqrt{\kappa_0+J}) \text{ where } \mu_J = \frac{J\bar{\theta}+\kappa_0\mu_0}{J+\kappa_0}$$

$$1/\tau^2|\boldsymbol{\theta},\mu \sim \text{gamma}(\frac{\eta_0+J+1}{2}, \frac{\eta_0\tau_0^2+\kappa_0(\mu-\mu_0)^2+\sum(\theta_j-\mu)^2}{2})$$

or, unconditional on μ,

$$1/\tau^2|\boldsymbol{\theta} \sim \text{gamma}(\frac{\eta_0+J}{2}, \frac{\eta_0\tau_0^2+\frac{\kappa_0 J}{\kappa_0+J}(\bar{\theta}-\mu_0)^2+\sum(\theta_j-\bar{\theta})^2}{2})$$

.

8.5 Example: IQ scores in the Netherlands

Students ($\approx$ age 11) in $J = 132$ classes (from 131 schools) in the Netherlands were given verbal IQ exams. Let $y_{i,j}$ = score of student i in school j. (data from Snijders and Bosker, 1999). Initial goal: examine heterogeneity of verbal IQ scores within and across schools.

Some questions we may or may not be able to address:

- What is the variation across schools/within schools?
- To what extent do some schools have higher mean scores than others?
- Are some schools "under-performing"?

- Normal model for the school level means might be ok, except for the heavy left tail.

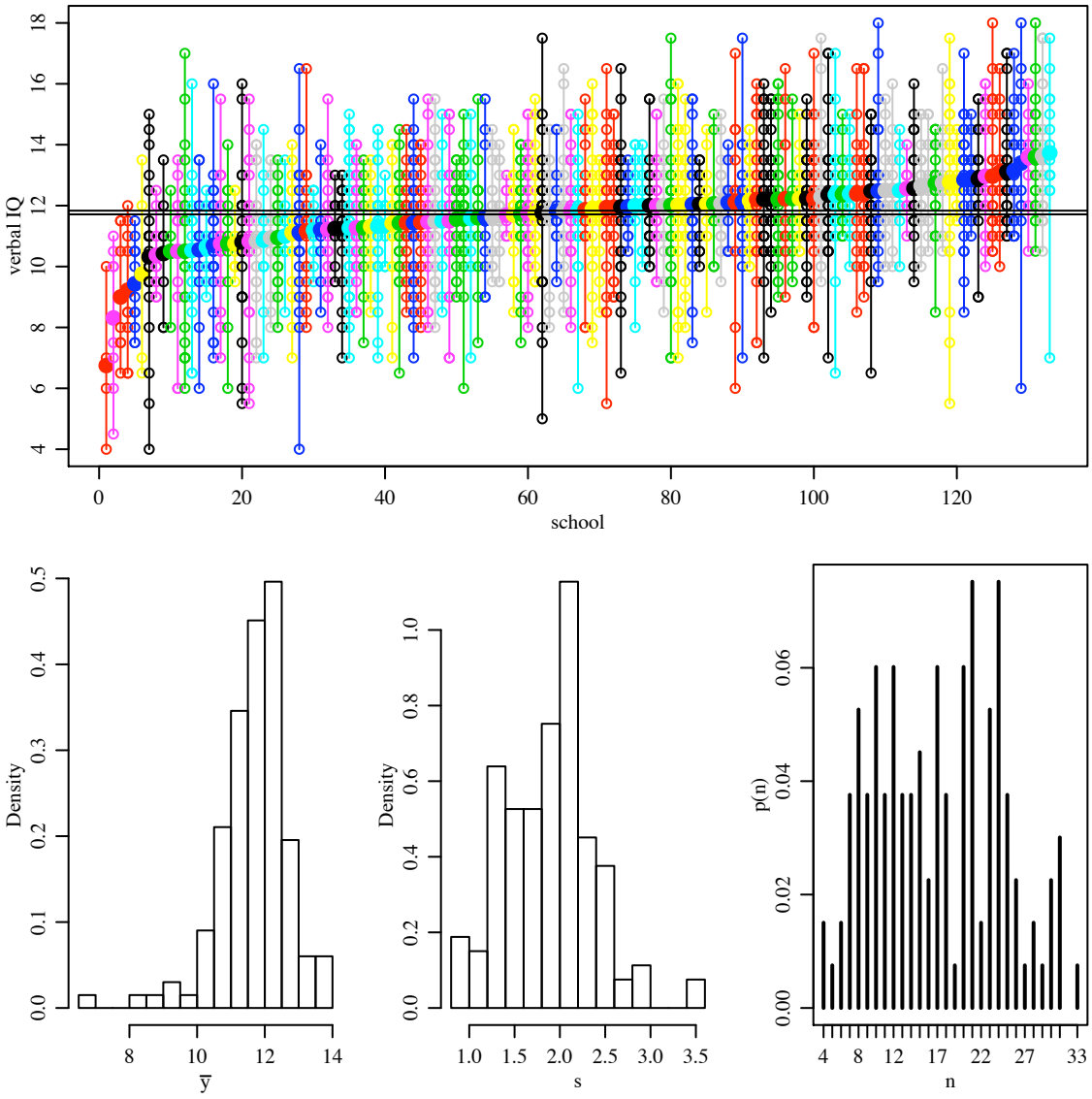


Figure 8.3: A graphical representation of the Netherlands schools data

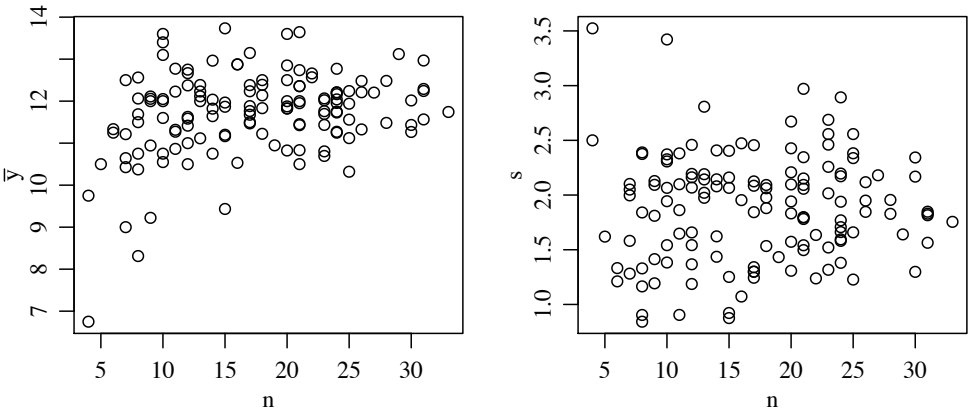


Figure 8.4: Relationship between sample size and within-group sample mean and variance

- Sample standard deviations are not too different.
- High variability in sample size.
- Lowest mean scores are from low sample size schools. Highest mean scores are from moderate sample size schools.

Bayesian analysis of variance. Recall the hierarchical normal model:

$$\begin{aligned}
\text{Within group sampling distribution:} \quad & y_{1,j}, \ldots, y_{n_j,j} | \theta_j, \sigma \sim \text{ i.i.d. normal } (\theta_j, \sigma) \\
\text{Between group sampling distribution:} \quad & \theta_1, \ldots, \theta_J | \mu, \tau \sim \text{ i.i.d. normal } (\mu, \tau) \\
\text{Prior distributions:} \quad & 1/\sigma^2 \sim \text{ gamma } (\nu_0/2, \nu_0\sigma_0^2/2) \\
& 1/\tau^2 \sim \text{ gamma } (\eta_0/2, \eta_0\tau_0^2/2) \\
& \mu \sim \text{ normal } (\mu_0, \tau/\sqrt{\kappa_0})
\end{aligned}$$

We'll use something similar to the so-called "unit information" priors (Kass and Wasserman, 1995):

- $\sigma_0^2 =$ sample mean $\{s_1^2, \ldots, s_{133}^2\} = 3.81$, $\nu_0 = 2$
- $\tau_0^2 =$ sample variance $\{\bar{y}_1, \ldots, \bar{y}_{133}\} = 1.02$, $\eta_0 = 2$
- $\mu_0 =$ sample mean $\{\bar{y}_1, \ldots, \bar{y}_{133}\} = 11.71$, $\kappa_0 = 1$;

Some people say this type of prior distribution (one that is very diffuse but uses the data) is cheating. Some say that it is practical and produces good results. It does seem to be cheating "a little," and is "not really" Bayesian. In any case, there is a vast literature on the subject which you are welcome to read.

Gibbs sampling:

```
nmc<-5000
THETA<-matrix( nrow=nmc,ncol=J)
MTS<-matrix( nrow=nmc,ncol=3)

for(ns in 1:nmc) {

# sample new values of thetas
for(j in 1:J) {
theta.hat.j<- (ybar[j]*n[j]/sigma^2 + mu/tau^2)/(n[j]/sigma^2+1/tau^2)
sigma.j<- 1/sqrt(n[j]/sigma^2+1/tau^2)
theta[j]<-rnorm(1,theta.hat.j,sigma.j)
               }

# sample new value of sigma
nun<- nu0+sum(n)
ss<- nu0*s20
for(j in 1:J) { ss<-ss+ sum( (Y[[j]]-theta[j])^2 ) }
sigma<-1/sqrt(rgamma(1,nun/2,ss/2))

# sample a new value of mu
mu<-rnorm(1, (J*mean(theta) + k0*mu0 )/(J+k0), tau/sqrt(J+k0) )

# sample a new value of tau
etaJ<-eta0+1+J
ss<- eta0*t20 + k0*(mu-mu0)^2 + sum( (theta-mu)^2 )
tau<-1/sqrt(rgamma(1,etaJ/2,ss/2))

# store results
THETA[ns,]<-theta
MTS[ns,]<-c(mu,tau,sigma)
                       }
```

We now have an `nmc`$\times$ `J` matrix of θ values and an `nmc`$\times$ `3` matrix of parameter values, representing approximate, correlated draws from the posterior distribution. How approximate? How correlated?

Basic MCMC diagnostics:

- **what we'd like:** independent samples from the posterior distribution.
- **what we have:** correlated samples from a process which converges to the posterior distribution.

How does this reality affect our estimation? Suppose for the moment the Markov chain has "converged," and we want to approximate the integral (as opposed to estimate the parameter) $\hat{\mu} = E[\mu|\mathbf{y}]$ from our MCMC samples $\{\mu_1, \mu_2, \ldots, \mu_{\text{nmc}}\}$ If our samples were *independent*, then

$$E[(\bar{\mu} - \hat{\mu})^2] = \frac{\text{Var}(\mu|\mathbf{y})}{\text{nmc}}$$

The square root of this quantity is called the *Monte Carlo standard error*. Note that we can make this is as small as we want by taking more Monte Carlo samples. You should think of this Monte Carlo error as completely distinct from data analysis, model building or statistical inference. It is just a measure of how good your approximation of the integral $\int \mu p(d\mu|\mathbf{y})$ is.

Unfortunately our Monte Carlo samples are dependent - we generated $\{\mu, \sigma, \tau\}_{t+1}$ from $\{\mu, \sigma, \tau\}_t$. In general, we will have positive correlation of the Monte Carlo samples, and so we will require nmc to be larger than if we had independent samples in order to achieve the same Monte Carlo error. How correlated are the Monte Carlo samples for our Markov chain? Correlation is across lags is shown in Figure 8.6. This low level of correlation is about as good as it gets.

The other issue was "convergence of the Markov chain." Assessing convergence is fraught with epistemological problems. In general, you can't know if your chain has converged. But sometimes you can know if your chain has not converged, so we at least check for this latter case. One thing to check for is *stationarity*, or that samples taken in one part of the chain have a similar distribution to samples taken in other parts. Based in Figure 8.5, the Markov chain seems to have achieved stationarity. Later on we will see examples where autocorrelation is high and it takes a long time to get to stationarity. In these cases you might need to run the MCMC sampler for a very, very long time.

Back to inference: Now lets make plots of approximate posterior densities and confidence regions based on these samples:

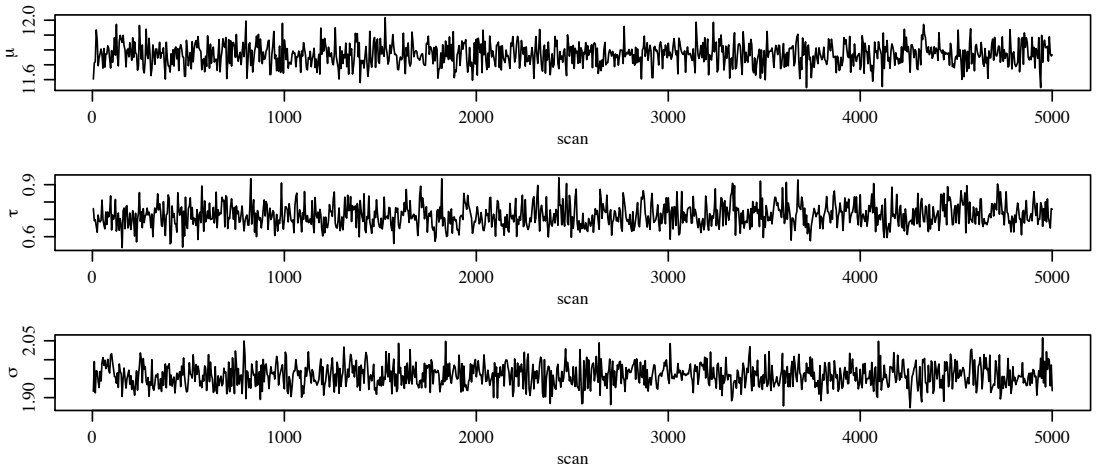


Figure 8.5: MCMC diagnostics: times series plots of the parameters

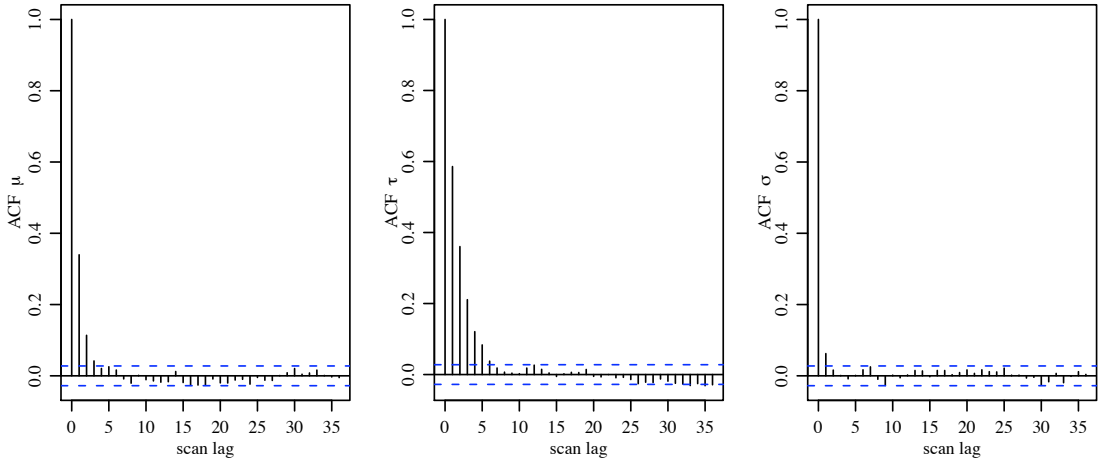


Figure 8.6: MCMC diagnostics: Autocorrelation functions

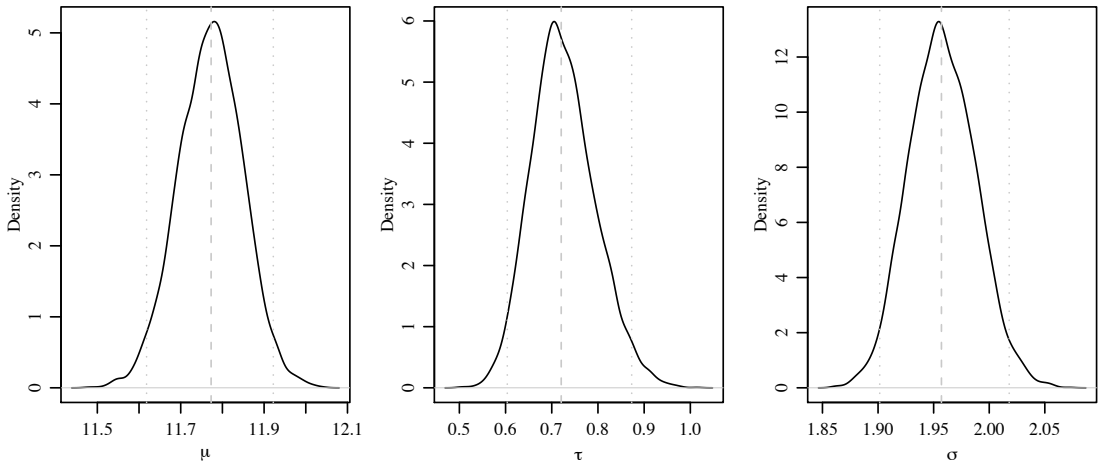


Figure 8.7: Marginal posterior distributions

```
plot(density(MTS[,1]),xlab=expression(mu),main="")
abline( v=quantile(MTS[,1],c(.025,.5,.975)),col="gray",lty=c(3,2,3) )
```

Recall that μ and τ are supposed to represent the distribution of $\theta_1, \ldots, \theta_{133}$. Lets compare the estimates of the θ's to their estimated distribution.

```
mu.hat<-mean(MTS[,1])
tau.hat<-mean(MTS[,2])
```

```
> mu.hat
[1] 11.77091
> tau.hat
[1] 0.7328663
```

```
theta.hat<-apply(THETA,2,mean)
x<-seq(mu.hat-3*tau.hat,mu.hat+3*tau.hat,length=100)
hist(theta.hat,prob=T,main="",xlab=expression(theta))
lines(x, dnorm(x, mu.hat, tau.hat) )
```

It looks like the normal distribution isn't quite capturing this left tail behavior.

One of the motivations behind hierarchical modeling is that information can be shared across groups. Recall, conditional on the data, μ, τ and σ,

$$E(\theta_j|\boldsymbol{y}_j, \mu, \tau, \sigma) = \frac{\bar{y}_j n_j/\sigma^2 + \mu/\tau^2}{n_j/\sigma^2 + 1/\tau^2}$$

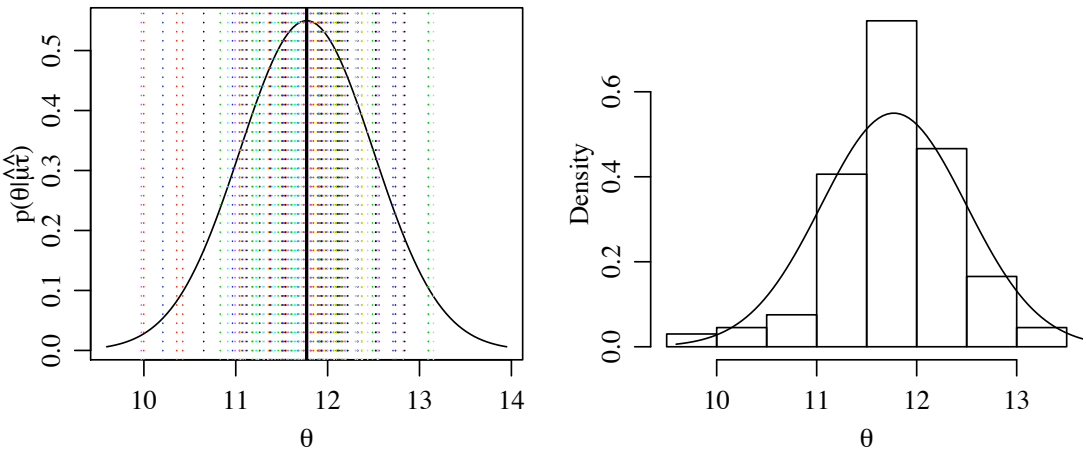


Figure 8.8: Between group variance in estimated means

and so the estimate θ_j is pulled a bit from $\bar{y}_j$ towards μ, the amount depending on n_j. This effect is called "shrinkage": Extreme values of $\bar{y}_j$ have slightly less extreme θ_j. The amount of shrinkage decreases with decreasing sample size: The larger the sample size for a group, the more information we have for that group, and the less information we need to "borrow" from the rest of the population. Also, suppose there are no differences between groups. Then groups with larger sample sizes will have $\bar{y}_j$'s closer to the grand mean already.

How would one go about ranking schools, if one had to? Consider comparing ranks based on $\bar{y}_j$ to those from θ_j.

```
ybar.order<-order(ybar)
> ybar.order[1:20]
 [1]  50  22   2  90   4  63   1 126  57   3  30 107  37 132 133  76  14  51  95  49

theta.order<-order(theta.hat)
> theta.order[1:20]
> theta.order[1:20]
> theta.order[1:20]
 [1]  22  50   4  90   2   1 107  37  76  30 126  49  63 132  40  54  14  57  59
[20]  51

> mean(THETA[,63]>THETA[,126])
[1] 0.5068
> mean(THETA[,63]>THETA[,107])
```

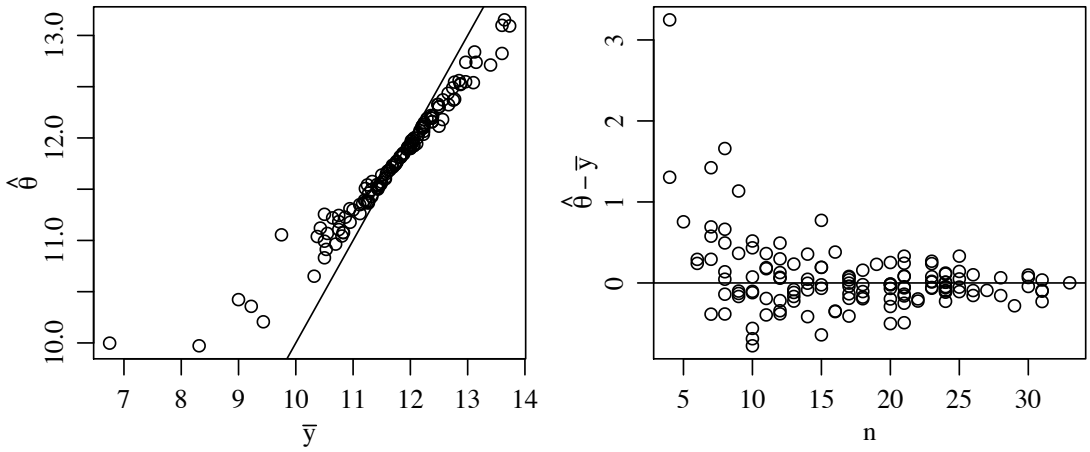


Figure 8.9: Shrinkage as a function of sample size

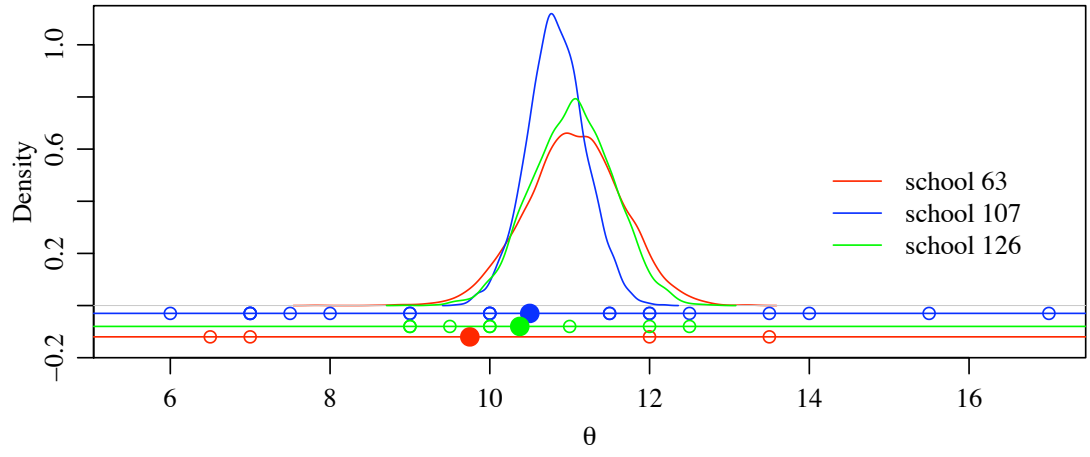


Figure 8.10: Data and posterior distributions for three schools

```
[1] 0.6372
> mean(THETA[,126]>THETA[,107])
[1] 0.6342
```

8.6 A bit more on shrinkage

Assume for the moment that σ^2 and τ^2 are known (or their ratio is known). Then from the full conditional distributions we see that

$$\begin{aligned} E[\theta_j|\boldsymbol{y}] &= \omega\bar{y}_{\cdot j} + (1-\omega)E[\mu|\boldsymbol{y}] \\ E[\mu\boldsymbol{y}] &= \alpha\frac{1}{J}\sum E[\theta_j|\boldsymbol{y}] + (1-\alpha)\mu_0 \end{aligned}$$

writing $E[\theta_j|\mathbf{y}]$ as $\hat{\theta}_j$, and plugging the second equation into the first we get

$$\begin{aligned} \hat{\theta}_j &= \omega\bar{y}_{\cdot,j} + (1-\omega)\alpha\mu_0 + (1-\omega)(1-\alpha)\bar{\theta}_\cdot, \text{ or in matrix form} \\ \{\mathbf{I} - (1-\omega)\alpha\mathbf{1}\mathbf{1}'/J\}\hat{\boldsymbol{\theta}} &= \omega\bar{\boldsymbol{y}} + (1-\omega)(1-\alpha)\mu_0\mathbf{1} \end{aligned}$$

Solving for $\hat{\theta}_j$ is just solving this linear system. If we use "empirical Bayes" and set $\mu_0 = \bar{y}_{\cdot\cdot}$, we get

$$\hat{\theta}_j = \tilde{\omega}\bar{y}_{\cdot j} + (1-\tilde{\omega})\bar{y}_{\cdot\cdot}$$

where the weights depend on n_j, J and the ratio of σ^2 to τ^2. "Shrinkage" estimators such as these have a variety of justifications and their properties have been explored in non-Bayesian contexts (see Stein 1956, Lindley 1962, Draper and Van Nostrand 1972). In many cases they perform better than setting $\hat{\theta}_j = \bar{y}_{\cdot j}$. The Bayesian analogy hopefully provides some intuition as to why.

Chapter 9

The multivariate normal model

Before we get into regression we should learn a bit about the multivariate normal distribution.

Multivariate normal model: Model for multivariate data sampled from a population.

Goal: Inference on means, variances, covariances, correlations.

Example: Test-retest scores for students.

$$\mathbf{y}_i = \begin{pmatrix} y_{i,1} \\ y_{i,2} \end{pmatrix} = \begin{pmatrix} \text{score on initial test} \\ \text{score on retest} \end{pmatrix}$$

Questions:

- What is the population mean?

$$E(\mathbf{y}|\boldsymbol{\mu}) = \begin{pmatrix} E(y_{i,1}|\boldsymbol{\mu}) \\ E(y_{i,2}|\boldsymbol{\mu}) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

- Is $\mu_2 > \mu_1$?
- Are $y_{i,1}, y_{i,2}$ correlated?

Contrast with regression: Let $y_i = y_{i,2}$, $x_i = y_{i,1}$. A regression model might be $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, which treats x_i as fixed.

- Regression analysis: models the conditional distribution of one variable given the rest, $p(y|\beta, x, \sigma)$.
- Multivariate analysis: models the joint distribution of the variables, $p(\boldsymbol{y}|\mu, \Sigma)$.

9.1 The multivariate normal density

We say a population of vectors of data $\boldsymbol{y}_1, \ldots, y_n$ have a multivariate normal distribution if the sampling density is given by

$$p(\boldsymbol{y}|\boldsymbol{\mu}, \Sigma) = |2\pi\Sigma|^{-1/2} \exp\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\}$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & & \vdots \\ \sigma_{1m} & \cdots & \cdots & \sigma_m^2 \end{pmatrix}$$

- $\mathrm{E}(y_k|\boldsymbol{\mu}, \Sigma) = \mu_k$;
- $\mathrm{V}(y_k|\boldsymbol{\mu}, \Sigma) = \sigma_k^2$;
- $\mathrm{Cov}(y_k, y_l|\boldsymbol{\mu}, \Sigma) = \sigma_{kl} = \sigma_{lk}$.

Figure 9.1 gives a picture for $\boldsymbol{\mu} = (75, 76)'$, $\sigma_1^2 = \sigma_2^2 = 5$, $\sigma_{12} = 3$, so $\mathrm{Cor}(y_1, y_2) = 0.6$. Given independent samples of $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ from a multivariate normal population, how can we make inference on $\boldsymbol{\mu}, \Sigma$? Bayesian inference proceeds as usual,

$$p(\boldsymbol{\mu}, \Sigma|\boldsymbol{y}) = p(\boldsymbol{\mu}, \Sigma) \times p(\boldsymbol{y}|\boldsymbol{\mu}, \Sigma)/p(\boldsymbol{y})$$

9.2 Conjugate priors and posterior inference

Conjugate priors: Can be viewed as a generalization of the univariate normal case.

$$p(\boldsymbol{\mu}, \Sigma) = p(\boldsymbol{\mu}|\Sigma)p(\Sigma)$$

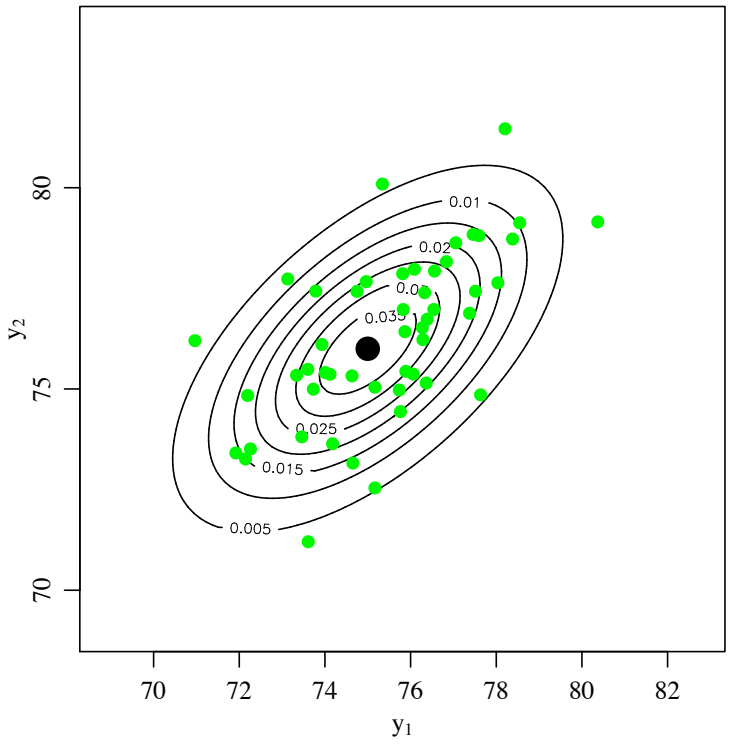


Figure 9.1: Multivariate normal samples and density

- $p(\boldsymbol{\mu}|\Sigma)$ = multivariate normal $(\boldsymbol{\mu}_0, \frac{1}{\kappa_0}\Sigma)$;
- $p(\Sigma)$ =?

Positive definiteness Just as a variance σ^2 must be positive, a variance-covariance matrix Σ must be *positive definite*, meaning

$$\boldsymbol{x}'\Sigma\boldsymbol{x} > 0 \text{ for all vectors } \boldsymbol{x}$$

Positive definiteness guarantees that $\sigma_k^2 > 0$ for all k, and that all correlations are between -1 and 1. Also, Σ must be symmetric, which means that $\sigma_{kl} = \sigma_{lk}$. So to do Bayesian analysis for the multivariate normal model, we must come up with a way of specifying a prior distribution on the space of $p \times p$ positive definite matrices.

Empirical covariance matrices: The *empirical sum of squares* of a collection of multivariate vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ will be denoted $\hat{S}$ and is given by

$$\hat{S} = \sum_{i=1}^{n} (\boldsymbol{z}_i - \bar{\boldsymbol{z}})(\boldsymbol{z}_i - \bar{\boldsymbol{z}})^T$$

Note that $(\boldsymbol{z}_i - \bar{\boldsymbol{z}})$ is a p-dimensional vector and so $(\boldsymbol{z}_i - \bar{\boldsymbol{z}})(\boldsymbol{z}_i - \bar{\boldsymbol{z}})^T$ is a (rank one) $p \times p$ matrix, and so the sum is also a $p \times p$ matrix. If we divide the matrix by $n-1$, we get the standard (unbiased) estimators of a population covariance matrix:

- $\frac{1}{n-1}\hat{S}_{j,j} = \frac{1}{n-1}\sum_{i=1}^{n}(z_{i,j} - \bar{z}_j)^2 \equiv s_{j,j} \equiv s_j^2$.
- $\frac{1}{n-1}\hat{S}_{j_1,j_2} = \frac{1}{n-1}\sum_{i=1}^{n}(z_{i,j_1} - \bar{z}_{j_1})(z_{i,j_2} - \bar{z}_{j_2}) \equiv s_{j_1,j_2}$

If $n > p$ and the bmz_i's are linearly independent, then $\hat{S}$ will be positive definite and symmetric. This suggests the following construction of a "random" covariance matrix:

A constructive definition of the Wishart distribution:

1. Sample $\boldsymbol{z}_1, \ldots, z_{\nu_0} \sim$ i.i.d. multivariate normal$(\boldsymbol{0}, T_0)$;
2. Set $\Omega = \sum_{i=1}^{\nu_0} \boldsymbol{z}_i \boldsymbol{z}_i^T$;

The distribution of Ω is called a *Wishart distribution* with parameters (ν_0, T_0).

- If $\nu_0 > p$ then Ω is positive definite.
- $E[\Omega] = \nu_0 T_0$.

The Wishart distribution is a multivariate version of the gamma distribution. Recall in the univariate normal model, our prior for the *precision* $1/\sigma^2$ had a gamma distribution, and our prior for the *variance* was an inverse-gamma distribution. Similarly, the Wishart distribution is a conjugate prior for the *precision matrix*, and the inverse Wishart distribution is our prior for the *covariance matrix*. To sample from this distribution, we need to add one more step to the above:

3. Set $\Sigma = \Omega^{-1}$.

To summarize,

- Precision matrix:
  - $\Omega = \Sigma^{-1} \sim$ Wishart (T_0, ν_0)
  - $E(\Sigma^{-1}) = \nu_0 T_0$;
- Covariance matrix:
  - $\Sigma = \Omega^{-1} \sim$ inverse-Wishart (T_0, ν_0)
  - $E(\Sigma) = \frac{1}{\nu_0 - m - 1} T_0^{-1}$;

So if we want Σ to be "near" a matrix Σ_0, we would use an inverse-Wishart prior with parameters ν_0 and $T_0 = S_0^{-1} = (\nu_0 - m - 1)\Sigma_0^{-1}$, and then

$$E(\Sigma) = \frac{1}{\nu_0 - m - 1} T_0^{-1} = \frac{\nu_0 - m - 1}{\nu_0 - m - 1} \Sigma_0 = \Sigma_0$$

The larger ν_0 is, the more concentrated $p(\Sigma)$ is around Σ_0. The matrix S_0 can be thought of as "prior sum of squares", and ν_0 the degrees of freedom. We will write $\Sigma^{-1} \sim$ Wishart (S_0^{-1}, ν_0).

Posterior inference Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n | \boldsymbol{\mu}, \Sigma$ be i.i.d. multivariate normal$(\boldsymbol{\mu}, \Sigma)$. Then we can write

$$p(\boldsymbol{\mu}, \Sigma | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = p(\boldsymbol{\mu} | \Sigma, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) p(\Sigma | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$$

where

- $p(\boldsymbol{\mu}|\Sigma, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) =$ multivariate normal $(\boldsymbol{\mu}_n, \frac{1}{\kappa_0+n}\Sigma)$, with

$$\boldsymbol{\mu}_n = \frac{n}{\kappa_0 + n}\bar{\boldsymbol{y}} + \frac{\kappa_0}{\kappa_0 + n}\boldsymbol{\mu}_0$$

- $p(\Sigma^{-1}|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) =$ Wishart $(S_n^{-1}, \nu_0 + n)$, with

$$S_n = S_0 + \sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})' + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{\boldsymbol{y}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{y}} - \boldsymbol{\mu}_0)'$$

So the "posterior sum of squares" is the prior sum of squares plus the data sum of squares plus the sum of squares correction for the mean.

Sampling from the posterior distribution: Everything is conjugate here, so we can make independent Monte Carlo samples from the posterior distribution.

$$\begin{array}{lll} \text{sample } \Sigma_{(1)}^{-1} \sim \text{Wishart}(S_n^{-1}\nu_0 + n), & \text{compute } \Sigma_{(1)}, & \text{sample } \boldsymbol{\mu}_{(1)} \sim \text{mvn}(\boldsymbol{\mu}_n, \frac{1}{\kappa_0+n}\Sigma_{(1)}) \\ \vdots & \vdots & \vdots \\ \text{sample } \Sigma_{(k)}^{-1} \sim \text{Wishart}(S_n^{-1}\nu_0 + n), & \text{compute } \Sigma_{(k)}, & \text{sample } \boldsymbol{\mu}_{(k)} \sim \text{mvn}(\boldsymbol{\mu}_n, \frac{1}{\kappa_0+n}\Sigma_{(k)}) \end{array}$$

Note: marginally, $\boldsymbol{\mu}$ has a multivariate t-distribution.

Example: Test-retest

```
mu0<-c(75,75)
k0<-1
S0<-matrix( c(3,2,2,3),byrow=T,nrow=2,ncol=2)
v0<-3

ybar<-apply(Y,2,mean)
mun<- n*ybar/(n+k0) + k0*mu0/(n+k0)
Sn<- S0 +  (n-1)*var(Y) + k0*n*(ybar-mu0)%*%t(ybar-mu0)/(k0+n)

Ysim<-MU<-matrix(nrow=1000,ncol=2)
for(nmc in 1:1000) {

Sigma<-solve(rwish( solve(Sn),v0+n) )
mu<- rmvnorm( mun, Sigma/(k0+n) )
```

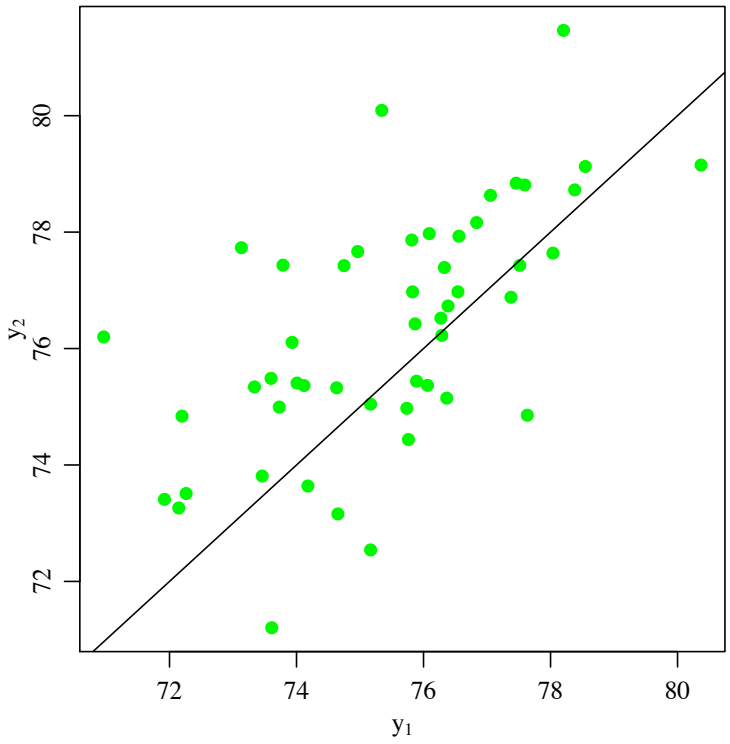


Figure 9.2: Test-retest data

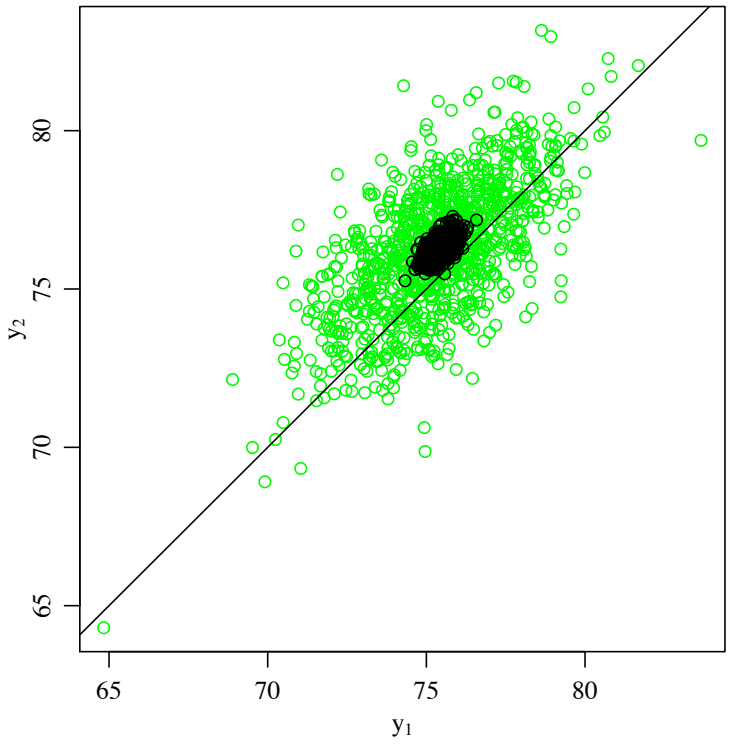


Figure 9.3: Posterior distributions for $\boldsymbol{\mu}$ and $\boldsymbol{y}_{\text{new}}$ for the test-retest data

```
ysim<-rmvnorm( mu, Sigma )
MU[nmc,]<-mu
Ysim[nmc,]<-ysim
                    }

> mean(MU[,2]>MU[,1])
[1] 0.999

> mean(Ysim[,2]>Ysim[,1])
[1] 0.692
```

9.3 A comment on admissibility

Suppose we want to estimate $\boldsymbol{\mu}$, the mean of a multivariate normal distribution. A data-based estimate of $\boldsymbol{\mu}$ is written $\hat{\boldsymbol{\mu}}(\boldsymbol{y})$. The *risk* of the estimate

$\hat{\boldsymbol{\mu}}(\boldsymbol{y})$ is defined as

$$R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = E[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^2 | \boldsymbol{\mu}]$$

One estimator $\hat{\boldsymbol{\mu}}_a$ is said to dominate another $\hat{\boldsymbol{\mu}}_b$ if

$$R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_a) \leq R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}_b)$$

for all $\boldsymbol{\mu}$ with inequality for at least one value of $\boldsymbol{\mu}$. An estimator $\hat{\boldsymbol{\mu}}$ is admissible if it is not dominated by another estimator. Some things to note:

- Any constant estimator is admissible. For example, the estimator $\hat{\boldsymbol{\mu}} = \mathbf{0}$ is admissible because its risk is zero if $\boldsymbol{\mu} = \mathbf{0}$ were true, and thus cannot be dominated. This suggests that admissibility by itself is not that strong of a criterion for estimation.

- If an estimator $\hat{\boldsymbol{\mu}}_a$ is not admissible, that means there exists another estimator $\hat{\boldsymbol{\mu}}_a$ out there with lower risk *for every value of* $\boldsymbol{\mu}$. So if an estimator is not admissible you would be better off using something else.

One obvious estimator of $\boldsymbol{\mu}$ is of course $\bar{\boldsymbol{y}}$. Surprisingly, Stein (1956) showed that in fact $\bar{\boldsymbol{y}}$ is inadmissible if $m > 2$. In contrast, the posterior mean $\hat{\boldsymbol{\mu}}_{\text{bayes}} = E(\boldsymbol{\mu}|\boldsymbol{y})$ is admissible for any proper prior distribution.

More generally, every Bayes estimator is admissible, and every admissible estimator is either a Bayes estimator under some proper prior distribution or the limit of such estimators. Therefore if you are doing estimation and you care about risk then you should only use Bayes estimators. Of course, there are situations where one doesn't care about risk (data description, for example).

Chapter 10

Regression

10.1 Goals and modeling assumptions

Regression models:

- y = variable of interest (avoid the term "dependent variable," perhaps use "response" if necessary).
- $\boldsymbol{x} = (x_1, \ldots, x_p)$ covariates (avoid "independent variables", perhaps use "explanatory variables").

Question: How does y vary as a function of $\boldsymbol{x}$? Data consist of responses and covariates for n experimental units:

$$\begin{array}{ll} y_1 & (x_{1,1}, \ldots, x_{1,p}) = \boldsymbol{x}_1 \\ \vdots & \vdots \\ y_n & (x_{n,1}, \ldots, x_{n,p}) = \boldsymbol{x}_n \end{array}$$

A regression model for y as a function of $\boldsymbol{x}$ relates the mean of y to a linear combination of the covariates:

$$\begin{aligned} E[y_i|\boldsymbol{\beta}, \boldsymbol{x}_i] &= \beta_1 x_{i,1} + \cdots \beta_p x_{i,p} \\ &= \boldsymbol{x}_i \boldsymbol{\beta}, \end{aligned}$$

where $\boldsymbol{x}_i$ is a $1 \times p$ row-vector of covariates.

What about the distribution of the y_i's around their means? Think about exchangeability:

$$(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n) \stackrel{d}{=} (y_{\pi 1}, \boldsymbol{x}_{\pi 1}), \ldots, (y_{\pi n}, \boldsymbol{x}_{\pi n})$$

This is sometimes called partial exchangeability. If partial exchangeability holds, and $\boldsymbol{x}_{i_1} = \boldsymbol{x}_{i_2} = \cdots = \boldsymbol{x}_{i_p} = \boldsymbol{x}$, then

$$y_{i_1}|\boldsymbol{x}_{i_1} \stackrel{d}{=} \cdots \stackrel{d}{=} y_{i_p}|\boldsymbol{x}_{i_p}$$

i.e. the conditional distributions for all observations y_i having the same value of $\boldsymbol{x}$ should be all the same. If we let $\epsilon_i = y_i - \boldsymbol{x}_i\beta$, partial exchangeability means that we can model all ϵ_i's *having the same* $\boldsymbol{x}_i$ *value* as conditionally independent and identically distributed.

Typical regression analyses make much more drastic assumptions, such as normal errors with constant variance.

$$\begin{aligned} y_1, \ldots, y_n|\beta, \sigma, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n &\sim \text{independent random variables} \\ y_i|\beta, \sigma, \boldsymbol{x}_i &\sim \text{normal}(\boldsymbol{x}_i\beta, \sigma) \\ &\Leftrightarrow \\ \epsilon_1, \ldots, \epsilon_n|\beta, \sigma, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n &\sim \text{i.i.d. normal}(0, \sigma) \\ y_i &= \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i \end{aligned}$$

These will be our modeling assumptions for now.

Example (1985 CPS data on wages and education): How much is a year of schooling worth? Demographic data on $n = 534$ employed individuals was gathered. Figures 10.1 and 10.2 show the dependence of hourly wage on education and sex. Discuss the modeling assumptions in the context of the data and the plots.

10.2 Conditional modeling

Let

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix}$$

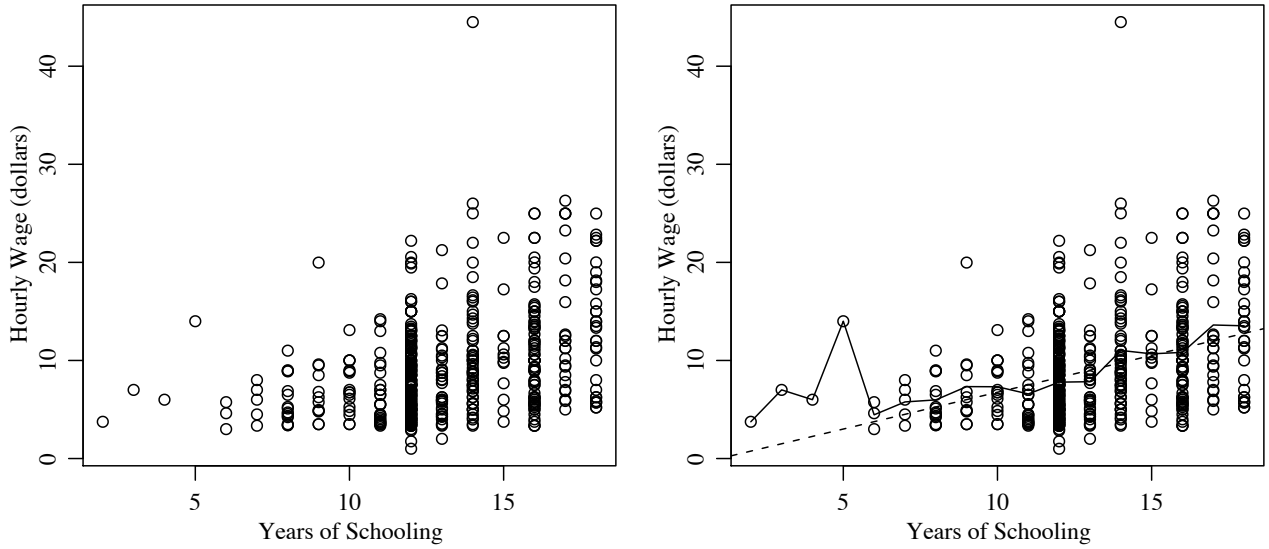


Figure 10.1: Wage data versus years of schooling: The second panel plots the mean wage for each level of education.

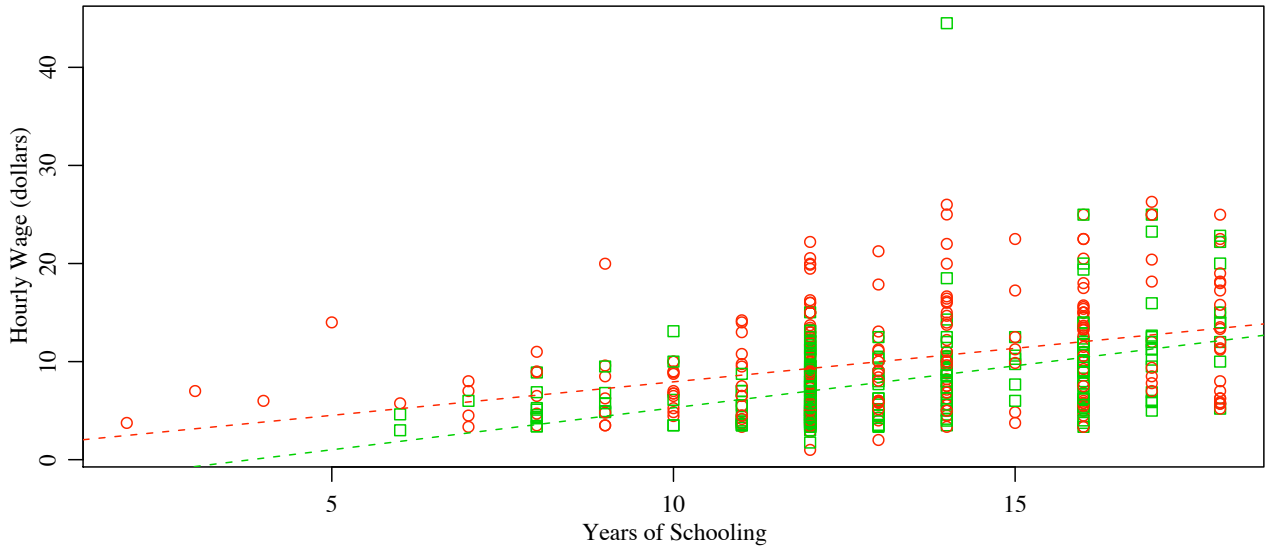


Figure 10.2: Wage data with OLS regression lines for each sex.

$\mathbf{X}$ is an $n \times p$ *design matrix*. Observed data is $\boldsymbol{y}$ and $\mathbf{X}$, but typically we make inference on β using $p(y|\mathbf{X}, \boldsymbol{\beta})$.

- Data: $\boldsymbol{y}$ and $\mathbf{X}$;
- Inference: $p(\boldsymbol{\beta}|\boldsymbol{y}, \mathbf{X}) \propto p(\boldsymbol{\beta})p(\mathbf{X}|\boldsymbol{\beta})p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta})$

Generally we model the posterior distribution as proportional to $p(\boldsymbol{\beta})p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta})$.

Justification of conditional modeling:

1. assume $p(\mathbf{X}|\boldsymbol{\beta}, \psi) = p(\mathbf{X}|\psi)$: $\mathbf{X}$ is independent of $\boldsymbol{\beta}$ given some parameter ψ. The parameter ψ describes the sampling distribution of $\mathbf{X}$.
2. Assume $p(\boldsymbol{\beta}, \psi) = p(\boldsymbol{\beta})p(\psi)$: This says that the population from which $\mathbf{X}$ was sampled doesn't give information about the relationship between $\boldsymbol{y}$ and $\mathbf{X}$.

In such a case,

$$\begin{aligned} p(\psi, \boldsymbol{\beta}|\mathbf{X}, \boldsymbol{y}) &\propto p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta}, \psi)p(\mathbf{X}|\boldsymbol{\beta}, \psi)p(\boldsymbol{\beta}, \psi) \\ &= [p(\psi)p(\mathbf{X}|\psi)] \times [p(\boldsymbol{\beta})p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta})] \end{aligned}$$

and so

$$p(\beta|\mathbf{X}, \boldsymbol{y}) \propto p(\beta)p(\boldsymbol{y}|\mathbf{X}, \beta)$$

In this case we say the *design* of the study/experiment/survey is *ignorable*. Put more simply, we need $p(\mathbf{X}|\beta) = p(\mathbf{X})$, i.e. there is no information about β in $\mathbf{X}$. Sometimes $p(\mathbf{X})$ is known explicitly:

- randomization scheme in controlled experiments;
- sampling method in a survey.

10.3 Estimation of regression coefficients

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} |\mathbf{X}, \boldsymbol{\beta}, \sigma \sim \text{multivariate normal}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{x}_1 \rightarrow \\ \boldsymbol{x}_2 \rightarrow \\ \vdots \\ \boldsymbol{x}_n \rightarrow \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \beta_1 x_{1,1} + \cdots + \beta_p x_{1,p} \\ \vdots \\ \beta_1 x_{n,1} + \cdots + \beta_p x_{n,p} \end{pmatrix} = \begin{pmatrix} E(y_1|\boldsymbol{\beta}, \boldsymbol{x}_1) \\ \vdots \\ E(y_n|\boldsymbol{\beta}, \boldsymbol{x}_n) \end{pmatrix}$$

Classical least-squares estimation: Find the value $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ that minimizes $SSR(\boldsymbol{\beta})$, where

$$SSR(\boldsymbol{\beta}) = \sum (y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2 = (\boldsymbol{y} - \mathbf{X}\beta)'(y - \mathbf{X}\beta).$$

$SSR(\boldsymbol{\beta})$ is minimized if $(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$ is orthogonal to $\mathbf{X}$, i.e.

$$\begin{aligned} \mathbf{X}'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0} &\Leftrightarrow \mathbf{X}'\boldsymbol{y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &\Leftrightarrow \hat{\boldsymbol{\beta}}_{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} \end{aligned}$$

Unbiased variance estimate:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-p}\sum (y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_{ols})^2 \\ V(\hat{\boldsymbol{\beta}}_{ols}|\sigma^2) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \approx \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

Maximum likelihood estimation: Find $\boldsymbol{\beta}, \sigma^2$ to maximize $p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma)$:

$$\hat{\boldsymbol{\beta}}_{mle} = \hat{\boldsymbol{\beta}}_{ols}, \qquad \hat{\sigma}^2_{mle} = \frac{1}{n}\sum (y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_{mle})^2$$

Example (Least squares regression in R):

```
> X
      educ
[1,] 1    8
[2,] 1    9
[3,] 1   12
[4,] 1   12
[5,] 1   12
.
.
.
```

```
> y
[1] 5.10 4.95 6.67 4.00 7.50 . . .


> t(X)%*%y
          [,1]
      4818.85
educ 65471.33

> sum(X[,1]*y)
[1] 4818.85
> sum(X[,2]*y)
[1] 65471.33

> t(X)%*%X
           educ
      534  6952
educ 6952 94152


> solve(t(X)%*%X)
                            educ
      0.048360851 -0.0035708709
educ -0.003570871  0.0002742873

> beta.ols<-solve(t(X)%*%X)%*%t(X)%*%y
> beta.ols
           [,1]
     -0.7459797
educ  0.7504608

> s2.ols<-sum( (y- X%*%beta.ols)^2 )/(n-2)
> s2.ols
[1] 22.60039

> s2.ols*solve(t(X)%*%X)
                         educ
      1.09297419 -0.08070308
educ -0.08070308  0.00619900

> sqrt(diag(s2.ols*solve(t(X)%*%X)))
                 educ
1.04545406 0.07873373
```

```
## compare to

> summary(lm(y~X[,2]))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.74598    1.04545  -0.714    0.476
X[, 2]       0.75046    0.07873   9.532   <2e-16 ***
```

Bayesian analysis: For now we will use a semi-conjugate prior.

- $\boldsymbol{\beta} \sim$ multivariate normal $(\boldsymbol{\beta}_0, \Sigma_0)$,
- $\sigma^2 \sim$ inverse-gamma$(\nu_0/2, \nu_0\sigma_0^2/2)$.

Given $y_1, \ldots, y_n$, $y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \sigma \sim$ normal $(\boldsymbol{x}_i\boldsymbol{\beta}, \sigma)$, we have

- $\boldsymbol{\beta}|\boldsymbol{y}, \sigma, \mathbf{X} \sim$ multivariate normal $(\boldsymbol{\beta}_n, \Sigma_n)$, where
  - $\Sigma_n^{-1} = \Sigma_0^{-1} + \frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}$
  - $\boldsymbol{\beta}_n = \Sigma_n(\Sigma_0^{-1}\boldsymbol{\beta}_0 + \frac{1}{\sigma^2}\mathbf{X}'y) = (\Sigma_0^{-1} + \frac{1}{\sigma^2}\mathbf{X}'\mathbf{X})^{-1}(\Sigma_0^{-1}\boldsymbol{\beta}_0 + \frac{1}{\sigma^2}\mathbf{X}'y)$
- $\sigma^2|\boldsymbol{\beta}, \boldsymbol{y}, \mathbf{X} \sim$ inverse-gamma$(\frac{1}{2}(\nu_0 + n), \frac{1}{2}[\nu_0\sigma_0^2 + (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})])$.

So we have nice forms for $(\boldsymbol{\beta}|\boldsymbol{y}, \sigma, \mathbf{X})$ and $(\sigma|\boldsymbol{y}, \boldsymbol{\beta}, \mathbf{X})$. It looks like a job for Gibbs sampling.

Improper prior: Note that if Σ_0 is "large" $\Leftrightarrow$ prior precision small, then

$$\Sigma_n \approx \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \quad \boldsymbol{\beta}_n = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\frac{1}{\sigma^2}\mathbf{X}'y = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y = \hat{\boldsymbol{\beta}}_{mle}$$

Some people go even further: suppose $p(\boldsymbol{\beta}, \sigma) \propto 1/\sigma$ (note that this is not a proper probability distribution). It can be shown that

- $\tilde{p}(\boldsymbol{\beta}|\sigma, \boldsymbol{y}, \mathbf{X}) \propto p(\boldsymbol{\beta}, \sigma, \boldsymbol{y}, \mathbf{X}) \propto \text{mvn}(\hat{\beta}_{\text{mle}}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$
- $\tilde{p}(\sigma^2|\boldsymbol{\beta}, \boldsymbol{y}, \mathbf{X}) \propto p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{y}, \mathbf{X}) \propto \text{inverse} - \text{gamma}[\frac{1}{2}(n-p), \frac{1}{2}(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{mle}})^T(\boldsymbol{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{mle}})] = \text{inverse} - \text{gamma}[\frac{1}{2}(n-p), \frac{1}{2}SSR(\hat{\boldsymbol{\beta}}_{\text{mle}})]$

This resulting joint probability distribution is proper, and will approximate the posterior distribution under valid priors in cases where the prior is diffuse and the sample size is not too small. Samples from these distributions can be obtained using regular Monte Carlo methods.

10.3.1 Derivation of full conditionals

We will now outline the derivation of these full conditionals. First consider $p(\boldsymbol{\beta}|\sigma, \boldsymbol{y}, \mathbf{X})$:

$$p(\boldsymbol{\beta}|\sigma, \boldsymbol{y}, \mathbf{X}) \propto p(\boldsymbol{\beta}|\mathbf{X}, \sigma)p(\boldsymbol{y}|\boldsymbol{\beta}, \mathbf{X}, \sigma) = p(\boldsymbol{\beta})p(\boldsymbol{y}|\boldsymbol{\beta}, \mathbf{X}, \sigma)$$

- Prior: $p(\boldsymbol{\beta}) = |2\pi\Sigma|^{-1} \exp\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\Sigma^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\}$
- Data: $p(\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})/\sigma^2\}$

The parts that matter are the parts in the exponents.

$$\text{Prior:} \quad (\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\Sigma^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = \boldsymbol{\beta}'\Sigma^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\Sigma^{-1}\boldsymbol{\beta}_0 + \boldsymbol{\beta}_0'\Sigma^{-1}\boldsymbol{\beta}_0$$

Note this **Fact:** A random vector $\boldsymbol{z}$ is multivariate normal with mean $\boldsymbol{m}$ and variance matrix V if and only if $p(\boldsymbol{z}|\boldsymbol{m}, V) \propto \exp\{-\frac{1}{2}(\boldsymbol{z}'V^{-1}\boldsymbol{z} - 2\boldsymbol{z}'V^{-1}\boldsymbol{m})\}$.

$$\text{Data:} \quad (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})/\sigma^2 = \boldsymbol{y}'\boldsymbol{y}'/\sigma^2 - 2\boldsymbol{\beta}'\mathbf{X}'\boldsymbol{y}/\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/\sigma^2$$

$$\text{Prior+Data:} \quad \boldsymbol{\beta}'(\Sigma^{-1} + \mathbf{X}'\mathbf{X}/\sigma^2)\boldsymbol{\beta} - 2\boldsymbol{\beta}'(\Sigma^{-1}\boldsymbol{\beta}_0 + \mathbf{X}'\boldsymbol{y}/\sigma^2) + \text{stuff that doesn't depend on } \boldsymbol{\beta}$$

From our **Fact**, we know that this means that the full conditional distribution of $\boldsymbol{\beta}$ is normal with mean $\boldsymbol{m}$ and variance V where

- $V^{-1} = \Sigma^{-1} + \mathbf{X}'\mathbf{X}/\sigma^2$, i.e. $V = (\Sigma^{-1} + \mathbf{X}'\mathbf{X}/\sigma^2)^{-1}$;
- $V^{-1}\boldsymbol{m} = \Sigma^{-1}\boldsymbol{\beta}_0 + \mathbf{X}'\boldsymbol{y}/\sigma^2$. Solving for $\boldsymbol{m}$ gives $\boldsymbol{m} = V(\Sigma^{-1}\boldsymbol{\beta}_0 + \frac{1}{\sigma^2}\mathbf{X}'\boldsymbol{y})$.

Now call $\boldsymbol{\beta}_n = \boldsymbol{m}$, and $\Sigma_n = V$.

Lets derive the full conditional of σ^2:

$$\begin{aligned} p(\sigma^2|\boldsymbol{\beta}, \boldsymbol{y}, \mathbf{X}) &\propto p(\sigma^2)p(\boldsymbol{y}|\sigma, \boldsymbol{\beta}, \mathbf{X}) \\ &\propto \left\{(\sigma^2)^{\frac{\nu_0}{2}+1} e^{-\frac{\nu_0}{2}\frac{\sigma_0^2}{\sigma^2}}\right\} \times \left\{(\sigma^2)^{\frac{n}{2}} e^{\frac{-1}{2\sigma^2}\sum(y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2}\right\} \\ &= (\sigma^2)^{-\frac{\nu_0+n}{2}} \exp\{-\frac{1}{2}[\nu_0\sigma_0^2 + \sum(y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2]/\sigma^2\} \end{aligned}$$

which is proportional to an inverse gamma distribution:

$$\sigma^2|\boldsymbol{\beta}, \boldsymbol{y}, \mathbf{X} \sim \text{inverse-gamma}[\frac{1}{2}(\nu_0 + n), \frac{1}{2}(\nu_0\sigma_0^2 + \sum(y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2]$$

So we have nice distributions for $\boldsymbol{\beta}|\sigma, \boldsymbol{y}, \mathbf{X}$ and $\sigma|\boldsymbol{\beta}, \boldsymbol{y}, \mathbf{X}$.

10.3.2 Posterior sampling

Q: How can we approximate posterior probabilities/expectations/quantiles?

A: By generating samples from the posterior distribution.

Gibbs sampling: Given starting values $\boldsymbol{\beta}_{(0)}$ and $\sigma_{(0)}$, iterate the following:

- sample $\boldsymbol{\beta}_{(l+1)} \sim p(\boldsymbol{\beta}|\sigma, \boldsymbol{y}, \mathbf{X})$;
- sample $\sigma_{(l+1)} \sim p(\sigma|\boldsymbol{\beta}, \boldsymbol{y}, \mathbf{X})$.

Here is R-code in to do regression using the conjugate normal model.

```
# Given :
# beta.0, Sigma.0 : prior distribution parameters for beta
# nu.0, sigma.0   : prior distribution parameters for sigma
# y : a vector of responses
# X : a matrix of covariates
# nscan : number of scans of Markov chain to run

### some convenient quantities
n<-length(y)
k<-length(beta.0)
iSigma.0<-solve(Sigma.0)
XtX<-t(X)%*%X

### store mcmc samples in these objects
beta.post<-matrix(nrow=nscan,ncol=k)
sigma.post<-rep(NA,nscan)

### starting value
sigma<-sd(lm(y~0+X)$residuals)

### MCMC algorithm
for(scan in 1:nscan) {

#update beta
Sigma.n<-solve(  iSigma.0 + XtX/sigma^2 )
beta.n<-Sigma.n%*%( iSigma.0%*%beta.0 + t(X)%*%y/sigma^2 )
beta<-rmvnorm( beta.n,Sigma.n)

#update sigma
nu.n<-nu.0+n
ss.n<-nu.0*sigma.0^2 + sum(  (y-X%*%beta)^2 )
```

```
sigma<-1/sqrt( rgamma(1,nu.n/2, ss.n/2) )

#save results of this scan
beta.post[scan,]<-beta
sigma.post[scan]<-sigma
                        }
```

10.4 Example: Stephens/Wallerstein debate

Data studying cross-national variation in union density. The population under study include 20 countries with a history of democracy.

- y_i = percentage of workforce that are union members;
- $x_{i,1} = 1$ (intercept);
- $x_{i,2}$ = index of left-wing government;
- $x_{i,3}$ = log civilian labor force (in thousands) ;
- $x_{i,4}$ = economic concentration.

Model:

$$y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \epsilon_i$$

Western and Jackman (APSR 1994) do a Bayesian analysis, trying to incorporate the different researchers' prior opinions.

Prior beliefs:

- Both researchers agree left governments assist union growth.
- Wallerstein believes in a negative labor force size effect.
- Stephens believes that economic concentration increases union density.

Prior distributions:

- Intercept : $\beta_1 \sim$ normal $(0, 10^6)$;
- Left government: $\beta_2 \sim$ normal $(.3, .15)$;

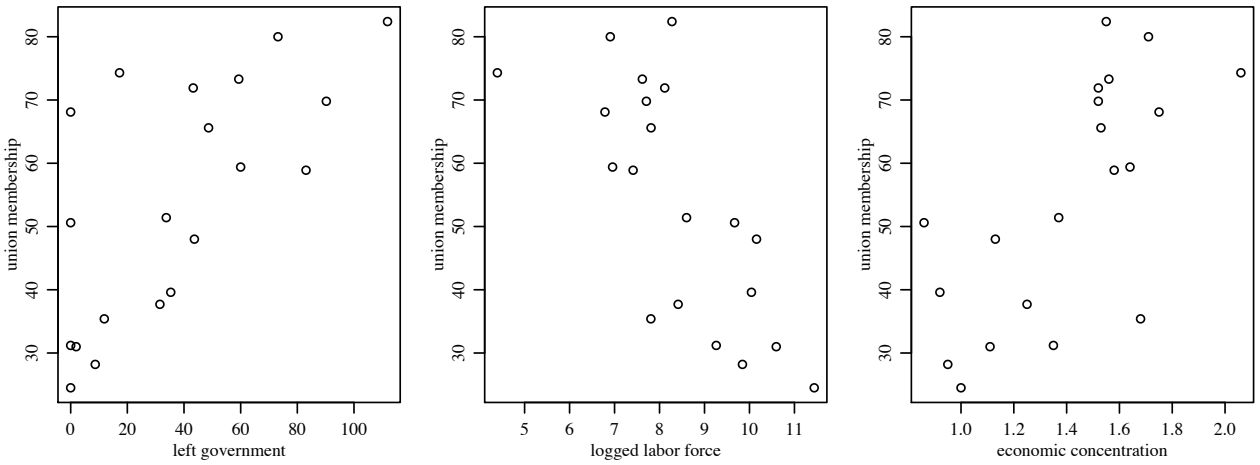


Figure 10.3: Union membership versus country-specific characteristics.

- Logged labor force : β_3

  Wallerstein: $\beta_3 \sim$ normal $(-5, 2.5)$;

  Stephens: $\beta_3 \sim$ normal $(0, 10^6)$;

- Economic concentration: β_4

  Wallerstein: $\beta_4 \sim$ normal $(0, 10^6)$;

  Stephens: $\beta_4 \sim$ normal $(10, 5)$;

- $1/\sigma^2 \sim$ inverse-gamma $(1, 10)$.

Prior opinions are confirmed by the data: Plots indicate a positive relationship between unions and left government and economic concentration, and a negative relationship between unions and labor force. Lets look at a non-Bayesian least-squares analysis:

```
fit<-lm(y~0+X)
summary(fit,correlation=T)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
Xint           97.59081   57.48063   1.698  0.10891
Xleft.gov       0.27027    0.07553   3.578  0.00251 **
Xlog.lab.force -6.45959    3.79381  -1.703  0.10797
Xecon.conc      0.35118   19.25292   0.018  0.98567
```

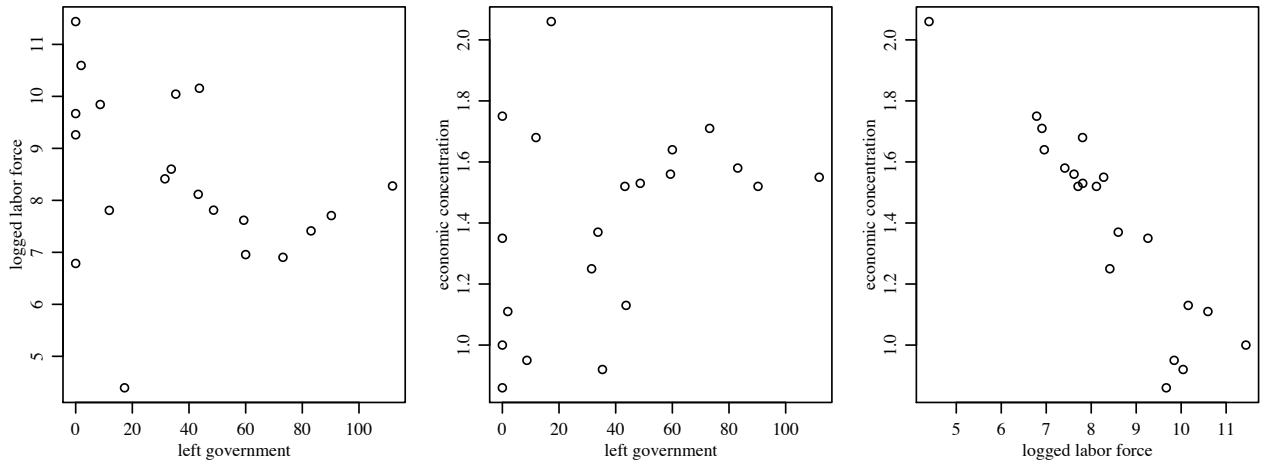


Figure 10.4: Correlation among covariates.

```
Correlation of Coefficients:
               Xint  Xleft.gov Xlog.lab.force
Xleft.gov       0.01
Xlog.lab.force -0.98  0.01
Xecon.conc     -0.97 -0.14      0.91
```

The coefficients β_{labor} and β_{econ} are highly correlated. This is a because the corresponding x-variables are highly correlated. As a result, it is hard to tell which variable is "having an effect". In fact, if we take out either variable:

```
> summary( lm(y~0+X[,c(1,2,3)]))

                         Estimate Std. Error t value Pr(>|t|)
X[, c(1, 2, 3)]int          98.60565   14.01150   7.037 2.00e-06 ***
X[, c(1, 2, 3)]left.gov      0.27045    0.07259   3.726 0.001680 **
X[, c(1, 2, 3)]log.lab.force -6.52271    1.50934  -4.322 0.000463 ***


> summary( lm(y~0+X[,c(1,2,4)]))

                         Estimate Std. Error t value Pr(>|t|)
X[, c(1, 2, 4)]int        1.41244   11.22037   0.126  0.90130
X[, c(1, 2, 4)]left.gov   0.27187    0.07963   3.414  0.00330 **
X[, c(1, 2, 4)]econ.conc 30.24928    8.32462   3.634  0.00205 **
```

Some quick model checks:

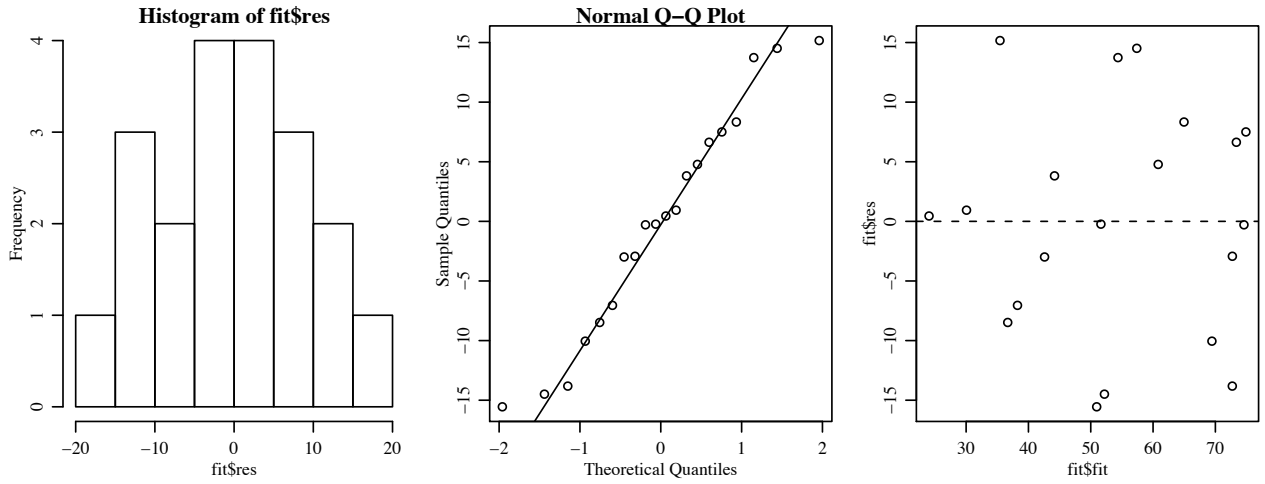


Figure 10.5: Regression diagnostics for the unionization data

```
fit<-lm(y~0+X)
hist(fit$res)
qqnorm(fit$res) ; qqline(fit$res)
plot(fit$fit,fit$res) ; abline(h=0,lty=2)
```

How do the two different Bayesian analyses do?

```
nu.0<-1
sigma.0<-10

# Wallerstein prior
beta.0w<-c(0, .3, -5, 0 )
Sigma.0w<-diag( c(10^6,.15,2.5,10^6)^2 )

# Stephens prior
beta.0s<-c(0,.3,0,10)
Sigma.0s<-diag( c(10^6,.15,10^6,5)^2 )

beta.0<-beta.0w
Sigma.0<-Sigma.0w
source("gibbs.regression.r")
beta.w<-beta.post

beta.0<-beta.0s
Sigma.0<-Sigma.0s
source("gibbs.regression.r")
beta.s<-beta.post
```

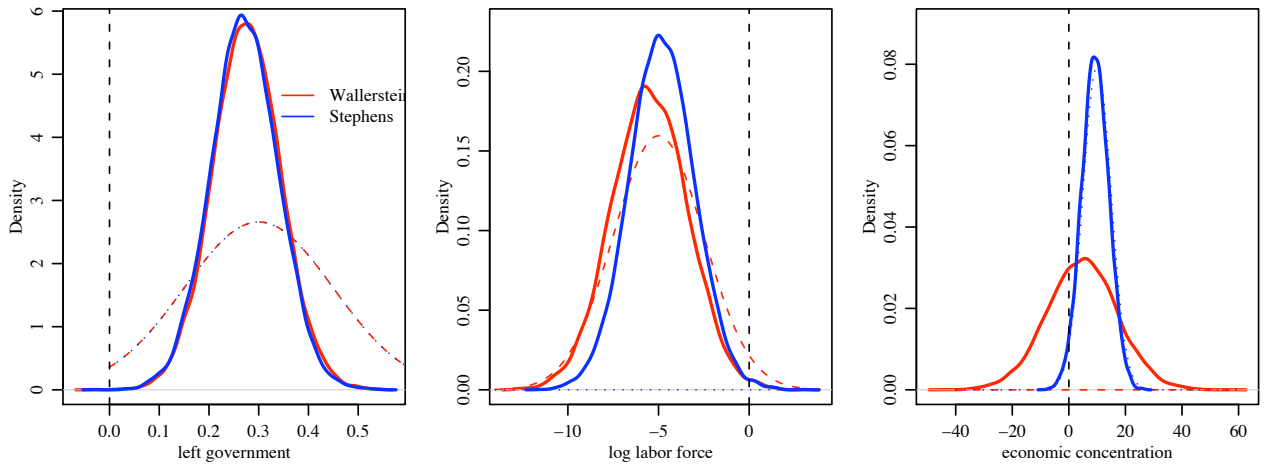


Figure 10.6: Posterior distributions under Stephens' and Wallerstein's prior distributions

For discussion: "Regression analysis in comparative research suffers from two distinct problems of statistical inference" (Western and Jackman, 94).

- Data are all observations from a population - inference based on long-run behavior of repeatable sampling is not appropriate.
- Small, collinear datasets yield imprecise estimates of the effects of explanatory variables.

"We provide a Bayesian approach to statistical inference that provides a unified solution to these two problems." What do you think?

10.5 Model selection and model averaging

Consider the following data analysis procedure: Given an $n \times p$ matrix $\mathbf{X}$ of predictor variables and an n-dimensional response vector $\boldsymbol{y}$,

1. Fit the regression model $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

From our intro stats class on regression analysis, we know that lots of covariates (large p), or a few highly correlated ones, can lead to unstable parameter

estimates (meaning they have high sampling variability). Maybe we should get rid of some of our covariates. Consider the following screening procedure:

2. Remove all covariates having p-values > 0.25, leaving a new covariate matrix $\tilde{\mathbf{X}}$;

3. Fit the regression model $\boldsymbol{y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

Lets actually try this, in the case where the data are generated as follows:

0. Let $\mathbf{X}$ be an $n \times p$ matrix of numbers sampled independently from a standard normal distribution. Let $\boldsymbol{y}$ be a vector of n number sampled from a standard normal distribution, independent of each other and of $\mathbf{X}$.

```
set.seed(1)
p<-30
n<-100
X<-matrix( rnorm(n*p), nrow=n,ncol=p)
y<-rnorm(n)

fit1<-lm(y~X)
p1<-summary(fit1)$coef[-1,4]
s<- (1:p)[ summary(fit1)$coef[-1,4] < .25 ]
Xs<-X[,s]
fit2<-lm(y~Xs)
summary(fit2)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.14676    0.11223   1.308   0.1942
Xs1         -0.19909    0.11572  -1.720   0.0887 .
Xs2          0.15251    0.11385   1.340   0.1837
Xs3          0.23208    0.10630   2.183   0.0315 *
Xs4          0.18894    0.11303   1.672   0.0980 .
Xs5         -0.17755    0.09668  -1.836   0.0695 .
Xs6          0.21679    0.10835   2.001   0.0483 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.098 on 93 degrees of freedom
Multiple R-Squared: 0.1549,     Adjusted R-squared: 0.1003
F-statistic:  2.84 on 6 and 93 DF,  p-value: 0.01387
```

Note that we used our data twice:

- first to select a model;
- second to estimate parameters.

Our procedure was ad-hoc, and is not justified by usual considerations of testing error. However, it is a fairly common practice. What does Bayesian inference have to say about model selection?

Bayesian model selection: Consider a response variable y, and a set of potential regressors $\{x_1, \ldots, x_5\}$. Selecting regressors 1,2, and 4 and fitting

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \epsilon$$

is equivalent to fitting the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon,$$

but setting β_3 and β_5 equal to zero. So the question of "which regressors to include" can be rephrased as "which coefficients are zero".

Bayesian approach: "I have some prior belief, say p that regression coefficient β_k is zero. If it not zero, then my belief about it is approximated by a normal distribution with mean $\beta_{0,k}$ and variance σ_k^2."

Formally,

$$p(\beta_k) = p \times \delta_0(\beta_k) + (1-p) \times \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\frac{-1}{2\tau_k^2}(\beta_k - \beta_{0,k})^2}$$

Such a prior commonly used in practice is

- $p = \frac{1}{2}$;
- $\beta_{0k} = 0$;
- $\tau_{0k}^2 =$ big.

This prior distribution sort of looks like the one in Figure 10.7, but not quite.

This is a real prior, and Bayesian inference proceeds as usual:

$$p(\boldsymbol{\beta}, \sigma | \boldsymbol{y}, X) = \frac{p(\boldsymbol{\beta}, \sigma) p(\boldsymbol{y} | \boldsymbol{\beta}, \sigma, X)}{p(y|X)}$$

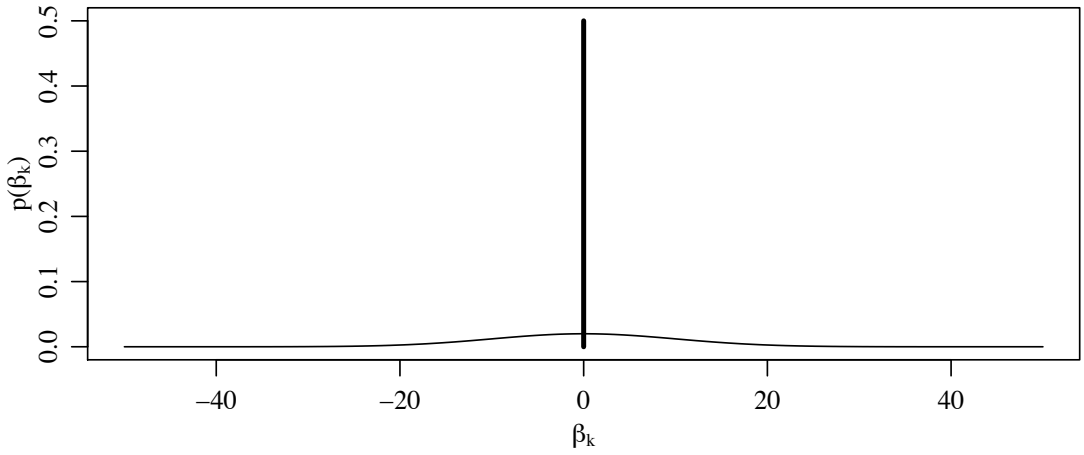


Figure 10.7: Point-mass mixture prior distribution

Full conditionals for $\boldsymbol{\beta}$'s and σ are available, and posterior inference proceeds by MCMC approximation of the posterior. Not too surprisingly,

$$p(\beta_k|\beta_{-k}, \sigma, \boldsymbol{y}, X) = \tilde{p} \times \delta_0(\beta_k) + (1 - \tilde{p}) \times \text{normal}(\beta_k|\beta_{n,k}, \tau_{n,k})$$

So conditional on the other quantities, β_k is zero with some probability and normally distributed otherwise. This means that different β's may be "turned on" from scan to scan:

scan	model										
1	β_0	+	$\beta_1 x_1$	+	0	+	$\beta_3 x_3$	+	$\beta_4 x_4$	+	0
2	β_0	+	$\beta_1 x_1$	+	$\beta_2 x_2$	+	$\beta_3 x_3$	+	$\beta_4 x_4$	+	0
3	β_0	+	$\beta_1 x_1$	+	0	+	0	+	$\beta_4 x_4$	+	0
⋮											
nscan	β_0	+	$\beta_1 x_1$	+	$\beta_3 x_3$	+	0	+	$\beta_4 x_4$	+	0

At the end, we may compute things like

- $\Pr(\beta_k = 0|\boldsymbol{y}) \approx \texttt{mean}(\texttt{beta.post}[, \texttt{k}] == 0)$
- $E(\beta_k|\boldsymbol{y}) \approx \texttt{mean}(\texttt{beta.post}[, \texttt{k}])$.

Note that at each scan, we have (in a sense) a *potentially different model*. Since we are averaging over scans to make posterior inference, this approach

is sometimes called *Bayesian model averaging*. Adrian and some of his former students have been kind enough to provide an R-add on package to do Bayesian model averaging (sort of - they do an approximate version). You can download it, install it, and try it out for free.

Lets try it out in the context of the simulated data.

```
library(BMA)
bfit<-bicreg(X,y)

 Posterior probabilities(%):
  X1   X2   X3   X4   X5   X6   X7   X8   X9  X10  X11  X12  X13  X14  X15  X16
 5.4  6.9  0.0  0.0  6.0 19.9  0.0  0.0  0.0  0.0  0.0 10.2  0.0  0.0  0.0 43.8
 X17  X18  X19  X20  X21  X22  X23  X24  X25  X26  X27  X28  X29  X30
11.0 11.5  0.0  0.0  0.0  0.0  0.0  0.0  6.8  0.0  0.0  4.5  0.0  3.4

 Coefficient posterior expected values:
(Intercept)          X1          X2          X3          X4          X5
   0.143884    0.008332    0.010704    0.000000    0.000000    0.007612
         X6          X7          X8          X9         X10         X11
  -0.039870    0.000000    0.000000    0.000000    0.000000    0.000000
        X12         X13         X14         X15         X16         X17
   0.016605    0.000000    0.000000    0.000000    0.091714    0.019202
        X18         X19         X20         X21         X22         X23
  -0.017162    0.000000    0.000000    0.000000    0.000000    0.000000
        X24         X25         X26         X27         X28         X29
   0.000000    0.008830    0.000000    0.000000    0.006873    0.000000
        X30
  -0.004274
```

Now lets see what happens when there is one covariate (x_1) with a real effect ($\beta_1 = .25$):

```
fit2<-lm(y~Xs)
summary(fit2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.13405    0.11310   1.185  0.23898
Xs1          0.36979    0.12674   2.918  0.00443 **
Xs2         -0.18721    0.11646  -1.607  0.11139
Xs3          0.16328    0.11449   1.426  0.15719
Xs4          0.23494    0.10641   2.208  0.02973 *
Xs5          0.16785    0.11528   1.456  0.14878
Xs6         -0.17355    0.09683  -1.792  0.07636 .
Xs7          0.21982    0.10846   2.027  0.04557 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

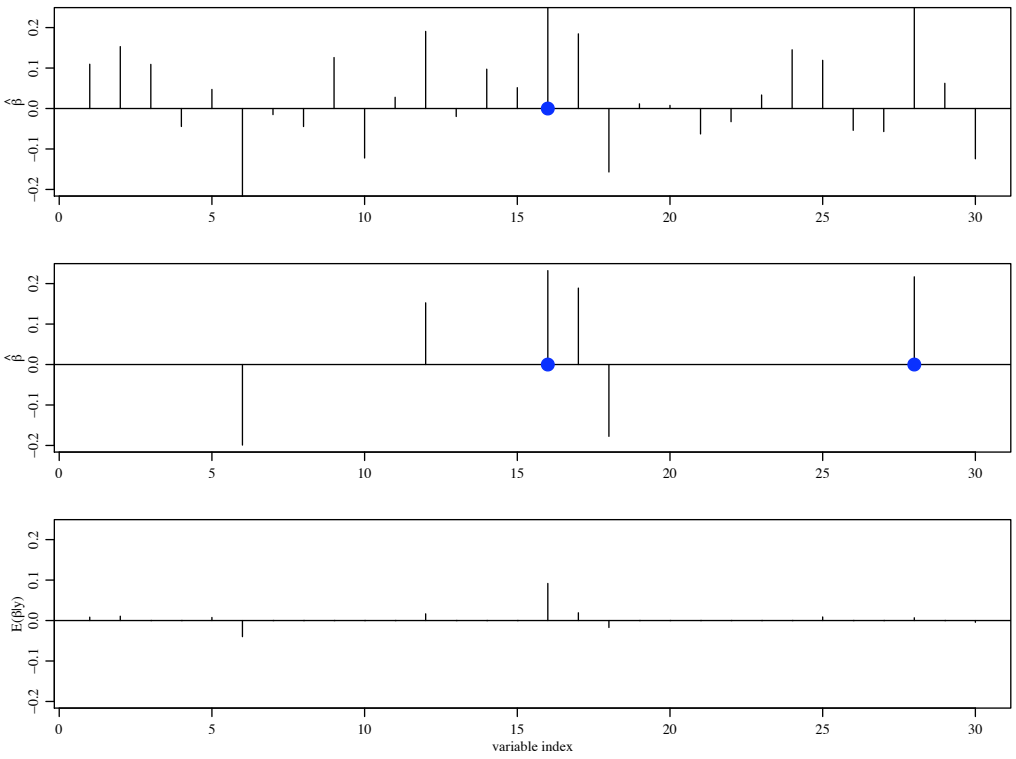


Figure 10.8: Results of different model selection techniques

```
Residual standard error: 1.099 on 92 degrees of freedom
Multiple R-Squared: 0.2268,     Adjusted R-squared: 0.168
F-statistic: 3.856 on 7 and 92 DF,  p-value: 0.001012


 Posterior probabilities(%):
  X1   X2   X3   X4   X5   X6   X7   X8   X9  X10  X11  X12  X13  X14  X15  X16
97.6  7.1  0.0  0.0  5.1 15.5  0.0  0.0  0.0  0.0  0.0 11.4  0.0  5.1  0.0 47.0
 X17  X18  X19  X20  X21  X22  X23  X24  X25  X26  X27  X28  X29  X30
 8.0 12.0  0.0  0.0  0.0  0.0  0.0  0.0  2.4  0.0  0.0  8.8  0.0  5.3

 Coefficient posterior expected values:
(Intercept)           X1           X2           X3           X4           X5
   0.131265     0.389910     0.011002     0.000000     0.000000     0.005714
         X6           X7           X8           X9          X10          X11
  -0.029401     0.000000     0.000000     0.000000     0.000000     0.000000
        X12          X13          X14          X15          X16          X17
   0.019919     0.000000     0.007542     0.000000     0.100318     0.012236
        X18          X19          X20          X21          X22          X23
  -0.017398     0.000000     0.000000     0.000000     0.000000     0.000000
        X24          X25          X26          X27          X28          X29
   0.000000     0.003246     0.000000     0.000000     0.012529     0.000000
        X30
  -0.006394
```

And just for fun...

```
y<-scan("y.union")
X<-as.matrix(read.table("X.union",header=T))
X<-X[ ,-1]
swbfit<-bicreg(X,y)

 Posterior probabilities(%):
    left.gov log.lab.force     econ.conc
       100.0          86.6          29.2

 Coefficient posterior expected values:
  (Intercept)       left.gov  log.lab.force      econ.conc
      85.4319         0.2706        -5.6394         4.1056
```

Discussion of model selection versus averaging, and prediction

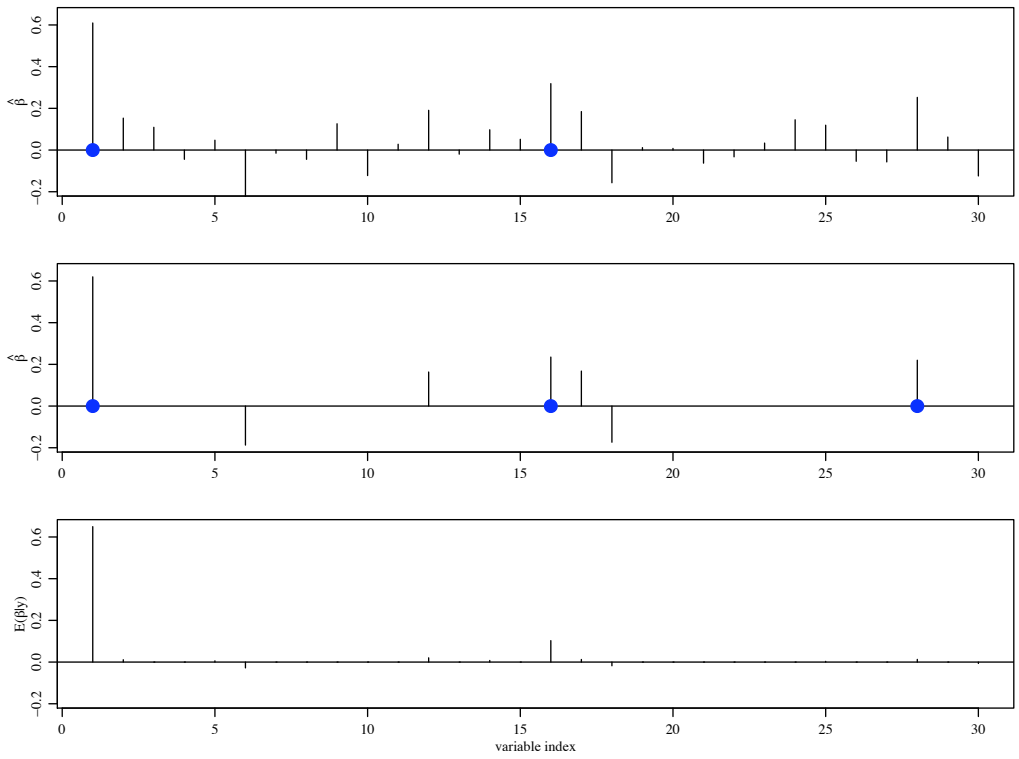


Figure 10.9: Results of different model selection techniques

10.6 Regression with general covariance structures

Our ordinary linear regression model had the form

$$\begin{aligned} y_i &= \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon_i \\ \epsilon_1, \ldots, \epsilon_n &\sim \text{i.i.d. normal}(0, \sigma) \end{aligned}$$

The matrix form for this model was

$$\begin{aligned} \boldsymbol{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim \text{multivariate normal}(\mathbf{0}, \Sigma) \\ \Sigma &= \sigma^2 I \end{aligned}$$

These formulas encapsulate two strong modeling assumptions:

- $E(y_i|\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i\boldsymbol{\beta}$, i.e. the expectation of y is a linear function of the x's;
- ϵ's are normally distributed ;
- ϵ's are independent and identically distributed.

Today we'll discuss deviations from this last assumption.

10.6.1 Some useful covariance structures

If there are n observations, then Σ is an $n \times n$ symmetric matrix $\rightarrow (n+1)n/2$ parameters specify. In general, this is too many, and we will need to impose some structure on Σ.

Unequal variances

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdots \\ 0 & \sigma_2^2 & 0 & \cdots \\ 0 & 0 & \sigma_3^2 & \\ \vdots & \vdots & & \end{pmatrix}$$

Here, Σ is parameterized by n parameters - this is typically still too many. Further structure:

- Weighted regression : $\sigma_i^2 = \sigma^2/w_i$ If weights are known, then number of parameters is just 1. Useful if y_i is an average of data from a single source i, and $w_i = \#$ of observations.

  Example:

  - y_i=average income of a sample of people from county i,
  - w_i = number of people in sample.

- Covariate specific variances: $\sigma_i^2 = \sigma^2(x_i)$.

  Example:

$$x_i = \begin{cases} 1 \text{ if } i \text{ is in group } A \\ 0 \text{ if } i \text{ is in group } B \end{cases} \qquad \sigma^2(x_i) = \begin{cases} \sigma_A^2 \text{ if } i \text{ is in group } A \\ \sigma_B^2 \text{ if } i \text{ is in group } B \end{cases}$$

Can use inverse-gamma priors for σ_a^2, σ_b^2, and can extend to more than two groups. Number of parameters=number of groups.

Spatial or temporal correlation: Observations "nearby" in time or space will be correlated.

1. temporal : y_i = observation at time t_i, $t_1, \ldots, t_n$ are evenly spaced time points.

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots \\ \rho & 1 & \rho & \cdots \\ \rho^2 & \rho & 1 & \\ \vdots & \vdots & & \end{pmatrix}$$

   $\sigma^2 > 0, -1 < \rho < 1 \Rightarrow$ 2-parameter model.

2. spatial : write $\Sigma = \{\sigma_{i,j}\}$

   - let $d_{i,j}$ = observed distance between units i and j (geographic or otherwise).
   - let $\sigma_{i,j} = \sigma^2 \rho^{d_{i,j}}$, so $V(y_{i,j}) = \sigma^2$, $C(y_i, y_j) = \sigma^2 \rho^{d_{i,j}}$.

Blockstructures and repeated measurements: Suppose $\boldsymbol{y}_i = \begin{pmatrix} y_{i,1} \\ \vdots \\ y_{i,m} \end{pmatrix}$ are all gathered from unit i (repeated measures, longitudinal or generically multivariate data).

write

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_m & 0 & 0 & \cdots \\ 0 & \Sigma_m & 0 & \cdots \\ 0 & 0 & \Sigma_m & \\ \vdots & \vdots & & \end{pmatrix}$$

Were Σ is a potentially arbitrary $m \times m$ matrix. Now there are $(m+1)m/2$ parameters to estimate, instead of $(nm)(nm+1)/2$.

Example (Twin IQ studies):

$$\boldsymbol{y}_i = \begin{pmatrix} y_{i,1} \\ y_{i,2} \end{pmatrix} \quad \Sigma_2 = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

Relation to HLM: i = family, j = twin.

$$y_{i,j} = \mathbf{x}_{i,j}\boldsymbol{\beta} + a_i + \epsilon_{i,j}$$

- $\epsilon_{i,j} \sim$ i.i.d. normal$(0, \sigma_\epsilon)$;
- $a_1, \ldots, a_J \sim$ i.i.d. normal$(0, \sigma_a)$.

Then

- $\text{Var}(y_{i,j}|\boldsymbol{\beta}, a_i) = \sigma_\epsilon^2$;
- $\text{Var}(y_{i,j}|\boldsymbol{\beta}) = \sigma_\epsilon^2 + \sigma_a^2$;
- $\text{Cov}(y_{i1,j}, y_{i2,j}|\boldsymbol{\beta}, a_i) = 0$;
- $\text{Cov}(y_{i1,j}, y_{i2,j}|\boldsymbol{\beta}) = \sigma_a^2$;
- $\text{Cor}(y_{i1,j}, y_{i2,j}|\boldsymbol{\beta}) = \frac{\sigma_a^2}{\sigma_a^2+\sigma_\epsilon^2} = \rho$.

So the following models are equivalent:

1. General regression model:

$$y_{i,j} = \mathbf{x}_{i,j}\boldsymbol{\beta} + \epsilon_{i,j} \quad \text{Cov}(\epsilon_{i,1}, \epsilon_{i,2}) = \Sigma_2 = \begin{pmatrix} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{pmatrix}$$

2. HLM:

$$\begin{aligned} y_{i,j} &= \mathbf{x}_{i,j}\beta + a_i + \epsilon_{i,j} \\ \{\epsilon_{i,j}\} &\sim \text{ i.i.d. normal}(0, \sigma_\epsilon) \\ a_1, \ldots, a_J &\sim \text{ i.i.d. normal}(0, \sigma_a) \end{aligned}$$

The equivalence is via

- $\sigma_1^2 = \sigma_a^2 + \sigma_\epsilon^2$;
- $\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_\epsilon^2}$.

Note this only works for $\rho > 0$.

10.6.2 Estimation

A general estimation strategy is via MCMC: Given $\Sigma^{(l)}$ at the lth stage of the Markov chain,

- sample $\boldsymbol{\beta}^{(l+1)}$ from $p(\boldsymbol{\beta}|\boldsymbol{y}, \Sigma^{(l)})$;
- sample $\Sigma^{(l+1)}$ from $p(\Sigma|\boldsymbol{y}, \boldsymbol{\beta}^{(l+1)})$ if possible, otherwise ...

Full conditional of $\boldsymbol{\beta}$: We know the full conditional in the OLS case (when $\Sigma = \sigma^2 I$). In the general case, we simply convert the problem to an OLS problem, then convert back.

Results from matrix theory:

- if $\text{Cov}(\boldsymbol{y}) = \Sigma_y$, then $\text{Cov}(A\boldsymbol{y}) = A\Sigma_y A'$, where A is an $n \times n$ matrix.
- if Σ_y is a proper covariance matrix (positive definite) then there is an invertible matrix $\Sigma^{1/2}$ such that $\Sigma = \Sigma^{1/2}\Sigma^{1/2'}$.

Application:

$$\begin{aligned} \text{Cov}(\boldsymbol{y}) &= \Sigma_y = \Sigma^{1/2}\Sigma^{1/2'} \\ \text{Cov}(\Sigma^{-1/2}\boldsymbol{y}) &= \Sigma^{-1/2}\Sigma^{1/2}\Sigma^{1/2'}\Sigma^{-1/2'} \\ &= I \end{aligned}$$

Therefore if $\boldsymbol{y} \sim$ multivariate normal $(\mathbf{X}\boldsymbol{\beta}, \Sigma)$, then given Σ,

$$\Sigma^{-1/2}\boldsymbol{y} \sim \text{ multivariate normal } (\Sigma^{-1/2}\mathbf{X}\boldsymbol{\beta}, I)$$

Letting

- $\boldsymbol{y}^* = \Sigma^{-1/2}\boldsymbol{y}$;
- $\mathbf{X}^* = \Sigma^{-1/2}\mathbf{X}$,

we have

$$\boldsymbol{y}^* \sim \text{ multivariate normal } (\mathbf{X}^*\boldsymbol{\beta}, I)$$

We know how to sample from the full conditional of $\boldsymbol{\beta}$ in this case. This leads to the following full conditionals:

$\boldsymbol{\beta}|\boldsymbol{y}, \mathbf{X}, \Sigma \sim$ multivariate normal$(\boldsymbol{\beta}_n, V_n)$, where

- $V_n^{-1} = \Sigma_\beta^{-1} + \mathbf{X}^{*'}\mathbf{X}^* \Rightarrow V_n = (\Sigma_\beta^{-1} + \mathbf{X}'\Sigma_y^{-1}\mathbf{X})^{-1}$
- $\boldsymbol{\beta}_n = V_n(\Sigma_\beta^{-1}\boldsymbol{\beta}_0 + \mathbf{X}^{*'}\boldsymbol{y}^*) = V_n(\Sigma_\beta^{-1}\beta_0 + \mathbf{X}'\Sigma_y^{-1}\boldsymbol{y})$

What about Σ_y? In some cases, conjugate priors are available:

- Group specific variances : $1/\sigma_j^2 \sim \text{gamma}(a, b)$;
- Unrestricted blockstructures: $\Sigma_m^{-1} \sim \text{Wishart}(\Sigma_0, \nu_0)$.

In these cases, you can sample from $p(\Sigma|\boldsymbol{y}, \boldsymbol{\beta}^{(l+1)})$ and thus can do Gibbs sampling. In other cases, such as in the spatial correlation model, we don't have conjugate priors. How can we approximate the posterior distribution when we don't have conjugate priors? Can we use grid-based methods?

Chapter 11

Posterior approximation 2

11.1 The Metropolis algorithm

Motivation: For $i = 1, \ldots, n$, let

- $\boldsymbol{x}_i$ = vector of demographic variables for couple i;
- y_i = indicator of divorce after 5 years.

Consider the logistic regression model $P(y_i = 1|\boldsymbol{\beta}, \boldsymbol{x}_i) = \frac{e^{\beta x_i}}{1+e^{\beta x_i}}$ This is a type of *generalized linear model*, or glm. A glm has the property that some function of $E(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})$ can be written as $\boldsymbol{\beta}\boldsymbol{x}_i$. In this case,

$$\text{log-odds}[E(y_i|\boldsymbol{x}_i, \beta)] = \log \frac{E(y_i|\boldsymbol{x}_i, \beta)}{1 - E(y_i|\boldsymbol{x}_i, \beta)} = \boldsymbol{\beta}\boldsymbol{x}_i$$

Given a prior distribution for β, the posterior information is expressed

$$p(\beta|\boldsymbol{y}, \mathbf{X}) = \frac{p(\boldsymbol{\beta})p(\boldsymbol{y}|\boldsymbol{\beta}, \mathbf{X})}{p(\boldsymbol{y}|\mathbf{X})}$$

As in the case of ordinary regression, a natural type of prior for β might be a multivariate normal distribution with mean $\boldsymbol{\beta}_0$ and prior covariance matrix Σ_β. However, under this prior, the posterior of $\boldsymbol{\beta}$ is not multivariate normal. In fact, there is not really a conjugate prior for such models (except in the normal regression case). One possibility is to use a grid-based approximation, but this will be problematic when the dimension of $\boldsymbol{\beta}$ is large.

11.1.1 Heuristic derivation of the algorithm:

Suppose $\theta \sim p(\theta),\ y|\theta \sim p(y|\theta)$. Monte Carlo estimation of $E[g(\theta)|y]$ proceeds by sampling $\theta_{(1)}, \ldots, \theta_{(S)} \sim p(\theta|y)$, and then

$$E[g(\theta)|y] \approx \frac{1}{S}\sum_{s=1}^{M} g(\theta_{(s)})$$

But what if

- $p(\theta|y) = p(\theta)p(y|\theta)/p(y)$ is not a standard distribution, or
- $p(y)$ is not available?

Idea: Construct a sequence of θ's whose distribution approximates $p(\theta|y)$. Roughly speaking, we need

$$\frac{\#\{\theta\text{'s in sample} = \theta_a\}}{\#\{\theta\text{'s in sample} = \theta_b\}} \approx \frac{p(\theta_a|y)}{p(\theta_b|y)}$$

How to construct such a sequence: Given $\{\theta_1, \ldots, \theta_s\}$ construct θ_{s+1} as follows: Consider a possible value θ^* near θ_s. Is θ^* a "more probable sample" from $p(\theta|y)$ than θ_s? Evaluate this via

$$r = \frac{p(\theta^*|y)}{p(\theta_s|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y)}\frac{p(y)}{p(y|\theta_s)p(\theta_s)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta_s)p(\theta_s)}$$

Note that $p(y|\theta^*)p(\theta^*)$ is computable without any integration.

- if $r > 1$:
  - **idea:** if θ_s is already in our sample, then we should include θ^* as it has a higher probability.
  - **procedure:** accept θ^* into our sequence, i.e. set $\theta_{s+1} = \theta^*$.
- if $r < 1$:
  - **idea:** θ_s is already in our sample - "to what extent" should θ^* be in the sample?

$$r = \frac{p(\theta^*|y)}{p(\theta_s|y)} \Rightarrow$$

for each θ_s in the sample, we should have about r θ^*'s, i.e. we want

$$\frac{\#\{\theta\text{'s in sample} = \theta_a\}}{\#\{\theta\text{'s in sample} = \theta_b\}} = r$$

– **procedure:**
let $\theta_{s+1} = \theta^*$ with probability r ;
let $\theta_{s+1} = \theta_s$ with probability $1 - r$.

This is the basic idea of the famous **Metropolis algorithm:** Given θ_s, generate θ_{s+1} as follows:

1. Sample a **proposal** θ^* from a **symmetric proposal distribution** $J(\theta^*|\theta_s)$.

   Symmetric means $J(\theta_a|\theta_b) = J(\theta_b|\theta_a)$, i.e. the probability of proposing $\theta^* = \theta_a$ given that $\theta_s = \theta_b$ is equal to the probability of proposing $\theta^* = \theta_b$ given that $\theta_s = \theta_a$.

   Usually J is very simple, and gives values near θ_s with high probability. Examples:

   - $J(\theta^*|\theta_s) = \text{uniform}(\theta_s - \delta, \theta_s + \delta)$;
   - $J(\theta^*|\theta_s) = \text{normal}(\theta_s, \delta)$.

2. Compute the **acceptance ratio**

$$r = \frac{p(\theta^*|y)}{p(\theta_s|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta_s)p(\theta_s)}$$

3. Let

$$\theta_{s+1} = \begin{cases} \theta^* & \text{with probability } r \wedge 1 \\ \theta_s & \text{with probability } 1 - r \wedge 1 \end{cases}$$

4. Repeat.

Note that step 3 can be accomplished by sampling $u \sim \text{uniform}(0, 1)$, and setting $\theta_{s+1} = \theta^*$ is $u < r$ and setting $\theta_{s+1} = \theta_s$ otherwise.

Example: Normal distribution with known variance In this toy problem we have

$$\begin{aligned} \theta &\sim \text{normal}(0, \tau) \\ y|\theta &\sim \text{normal}(\theta, \sigma) \end{aligned}$$

With σ known, we know the posterior distribution of θ is $\theta|y \sim$ normal (μ_n, τ_n) where

- $\mu_n = \bar{y}\frac{n/\sigma^2}{n\sigma^2+1/\tau^2} + \mu\frac{1/\tau^2}{n/\sigma^2+1/\tau^2}$
- $\tau_n^{-2} = n/\sigma^2 + 1/\tau^2$

Here is how coding it in R would work:

```
# generate some simulated data
sd.y<-1
mu.y<-10
n<-20
y<-rnorm(n,mu.y,sd.y)

#prior for theta
mu.theta<-0
sd.theta<-10

nscan<-2000                     #number of scans of Markov chain
THETA<-rep(NA,nscan)            #store output in here
theta<-0                        #starting/current value
sd.mh<-.33                      #sd for proposal distribution

for(s in 1:nscan) {

#propose new theta
theta.star<-rnorm(1,theta,sd.mh)

#compute acceptance probability
lr<-
  sum(dnorm(y,theta.star,sd.y,log=T))+
  dnorm(theta.star,mu.theta,sd.theta,log=T)-
  sum(dnorm(y,theta,sd.y,log=T))-
  dnorm(theta,mu.theta,sd.theta,log=T)

#accept theta.start with appropriate probability
if(log(runif(1))<lr) { theta<-theta.star  }
```

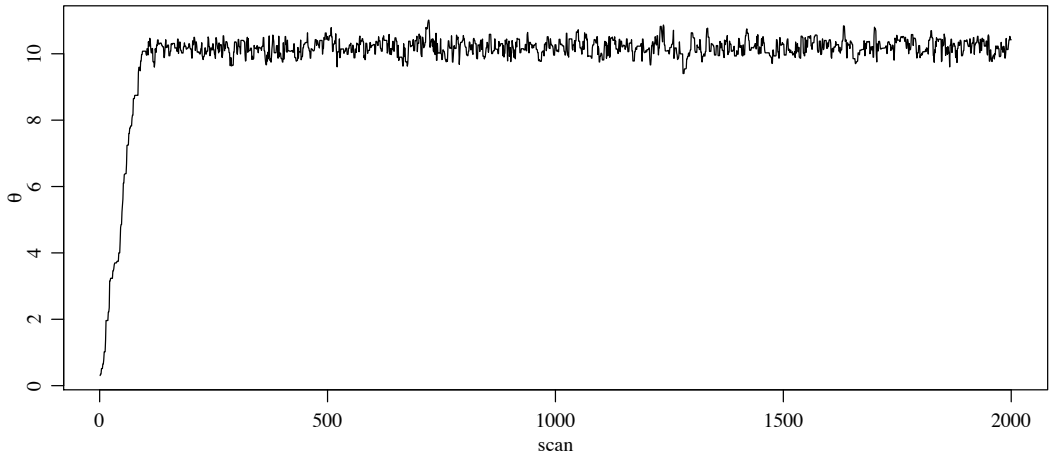


Figure 11.1: Samples from the Metropolis algorithm

```
#store the current value of theta
THETA[s]<-theta

                    }
```

Figure 1 plots these samples as a time series. Figure 2 examples how well these samples approximate the true posterior. Figure 3 examines Markov chains constructed using different proposal distributions.

11.1.2 Output of the algorithm

$\{\theta_1, \ldots, \theta_S\}$ is a *dependent* sequence of θ-values, in which the *relative frequencies* of the θ's in the sequence match the *relative probabilities* of $p(\theta|y)$, i.e.

$$\frac{\#\{\theta\text{'s in sample} = \theta_a\}}{\#\{\theta\text{'s in sample} = \theta_b\}} \approx \frac{p(\theta_a|y)}{p(\theta_b|y)}$$

or more precisely/importantly

$$\frac{\#\{\theta\text{'s in sample} < \theta_a\}}{\#\{\theta\text{'s in sample}\}} \approx p(\theta < \theta_a|y)$$

Notes:

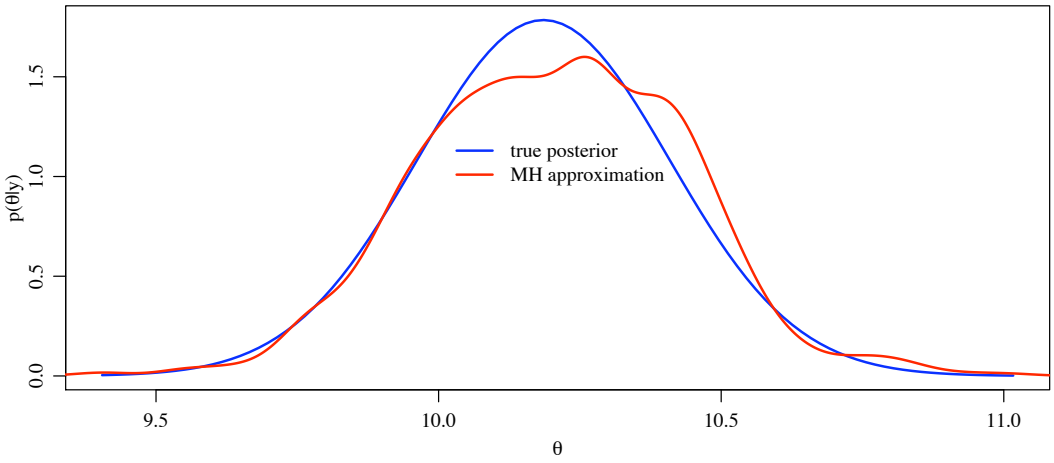


Figure 11.2: Comparison of true posterior to an MCMC approximation

- The conditional distribution of θ_{s+1} given $\{\theta_1, \ldots, \theta_s\}$ depends only on θ_s. Recall this is the Markov property, and so the sequence $\{\theta_1, \theta_2, \ldots\}$ is called a Markov chain.
- Under some mild conditions, the **marginal** sampling distribution of θ_s is approximately $p(\theta|y)$ for large s, i.e.

$$\lim_{s\to\infty} p_s(\theta|y) = p(\theta|y)$$

  regardless of θ_1.
- Under the same conditions, the empirical distribution of $\{\theta_1, \ldots, \theta_s\}$ approximates $p(\theta|y)$.

$$\lim_{s\to\infty} \frac{\#\{\theta_1, \ldots, \theta_s\} < c}{s} = p(\theta < c|y)$$

Practically speaking: Goal: generate samples of $\theta \sim p(\theta|y)$ so we can see what $p(\theta|y)$ looks like. We know that, in theory,

$$p_s(\theta_{(s)}) \to p(\theta|y) \quad \text{as } s \to \infty$$

but we can't run the chain forever. Instead, in practice we

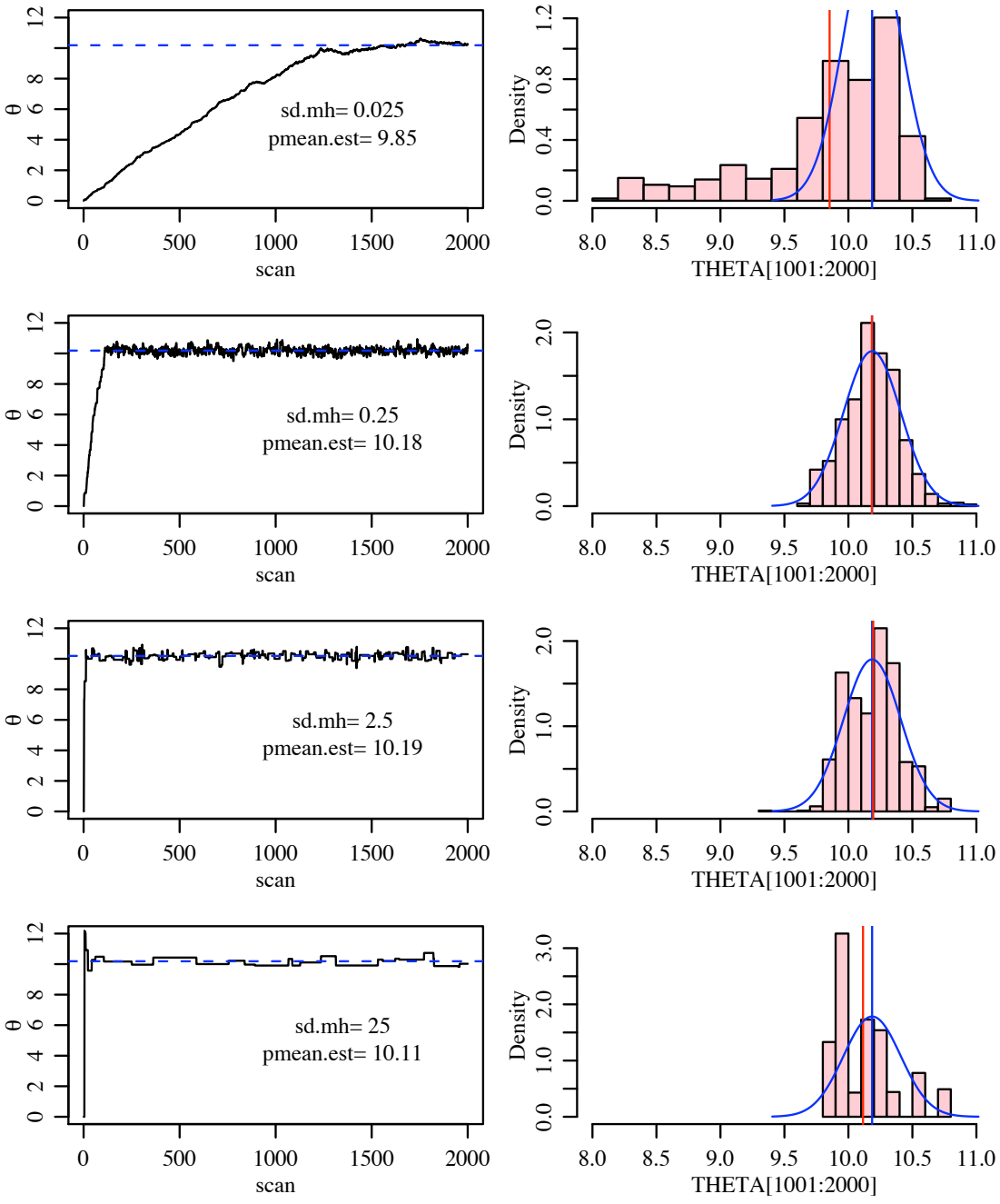


Figure 11.3: MCMC estimation under different proposal distributions

1. run algorithm until some scan B , where it looks like $p_B(\theta)$ has converged (B stands for "burn in");
2. hope $p_B(\theta) \approx p(\theta|y)$;
3. Run the algorithm S more times, generating $\theta_{(B+1)}, \ldots, \theta_{(B+S)}$. These samples are approximately (marginally) distributed as $p(\theta|y)$, but are correlated.
4. Approximate $p(\theta|y)$ with the empirical distribution of $\theta_{(B+1)}, \ldots, \theta_{(B+S)}$.

Problem with correlation: Recall the following basic concept: Suppose we want to estimate the mean μ of some population of $y's$. Generally speaking,

info from n independent samples of $y >$ info from n dependent samples of y.

The same holds when trying to approximate things about θ , like $E(\theta)$:

info about $E(\theta|y)$ in n independent samples of $\theta >$ info about $E(\theta|y)$ in n dependent samples of θ.

The more correlated our Markov chain is, the "less" information we get per scan. Fortunately in Monte Carlo sampling, we can take as many samples from the Markov chain as we have patience for. If the chain is highly correlated, we will need to take many scans to get good approximations to $p(\theta)$.

For an efficient Markov chain, we want

- Fast convergence of $p_k(\theta) \to p(\theta|y)$;
- low correlation between $\theta_{(s)}, \theta_{(s+1)}, \ldots$

Efficiency is determined by the

- choice of starting value;
- choice of proposal distribution.

These things can be monitored by examining the *acceptance rate* of the Markov chain, and the *autocorrelation function*. Lets look at these things with the normal model with known variance. We'll use a normal proposal distribution with standard deviations ranging in 0.025, 0.25, 2.5, 25.

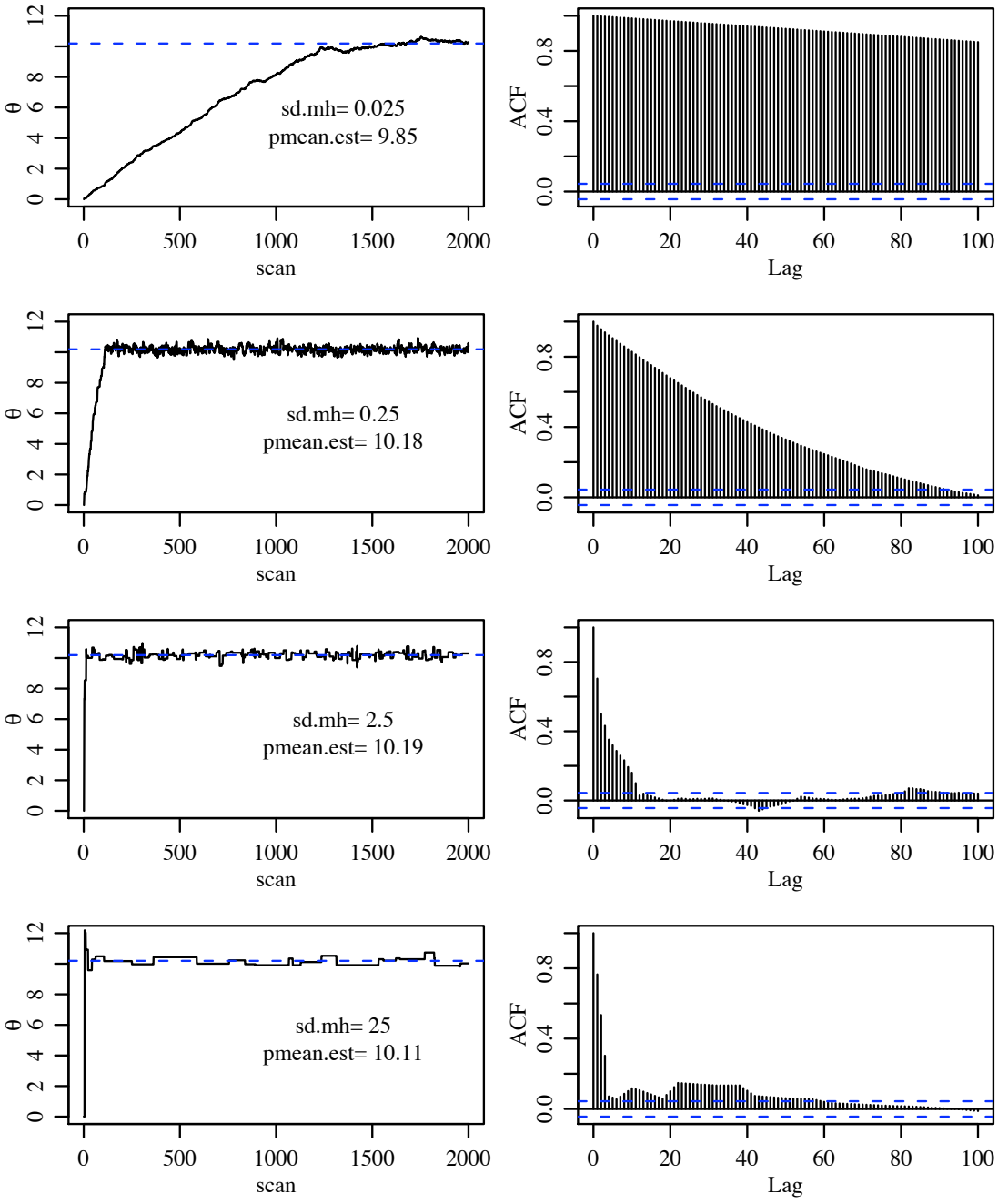


Figure 11.4: MCMC estimation under different proposal distributions

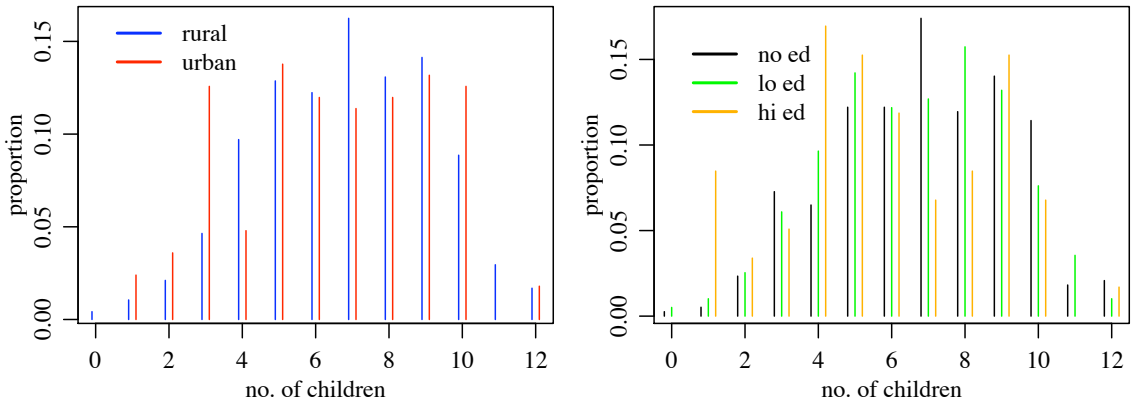


Figure 11.5: Fijian women data

```
> w<- (n/sd.y^2)/( n/sd.y^2 + 1/sd.theta^2)
> mean(y)*w + (1-w)*mu.theta

[1] 10.18543
```

sd.J	acc	mean(THETA)	mean(THETA[1001:2000])
0.025	0.762	7.143	9.853
0.250	0.642	9.931	10.183
2.500	0.108	10.166	10.194
25.000	0.014	10.119	10.114

11.1.3 Example: Poisson regression

Survey of Fijian women who had been married more than 20 years.

- y_i = number of children birthed;
- $x_{i,1}$ = rural residence (0 or 1);
- $x_{i,2}$ = primary school education (0 or 1);
- $x_{i,3}$ = education beyond primary school (0 or 1).

for $i = 1, \ldots, 641$.

We will model $y_i|\boldsymbol{x}_i$ as Poisson with mean

$$E(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = e^{\beta_0+\beta_1 x_{i,1}+\beta_2 x_{i,2}+\beta_3 x_{i,3}}$$

Thus, the mean response is log-linear in the covariates:

$$\log E(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}$$

This is called a **generalized linear model**, because the some function of the mean response is of the form $\boldsymbol{\beta}'\boldsymbol{x}_i$. The function that relates the expectation to the **linear predictor** $\boldsymbol{\beta}'\boldsymbol{x}_i$ is called the **link function**. More specifically, this model is the standard *Poisson regression model with a log-link*.

We will analyze these data with a normal(0,10) prior on each β coefficient, using the Metropolis algorithm. Below is how such an analysis might proceed:

```
# ML estimation
fit.mle<-glm(dat$children~dat$rural+dat$lowed+dat$uped, family=poisson)
summary(fit.mle)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.90331    0.03531  53.910   <2e-16 ***
dat$rural    0.03471    0.03594   0.966   0.3342
dat$lowed   -0.02668    0.03425  -0.779   0.4360
dat$uped    -0.15397    0.05787  -2.661   0.0078 **

# prior
mn.beta<-rep(0,3)
sd.beta<-rep(10,3)

# starting values and other things for the MCMC
beta<-fit.mle$coef
sd.prop<-summary(fit.mle)$coef[,2]/5
nscan<-10000
BETA<-matrix(0,nrow=nscan,ncol=4)
ac<-0

for(s in 1:nscan) {

#propose a new beta
beta.p<- rnorm(4, beta, sd.prop )

lhr<- sum(dpois(y,exp(X%*%beta.p),log=T)) -
      sum(dpois(y,exp(X%*%beta),log=T)) +
      sum(dnorm(beta.p,mn.beta,sd.beta,log=T)) -
      sum(dnorm(beta,mn.beta,sd.beta,log=T))

if( log(runif(1))< lhr ) { beta<-beta.p ; ac<-ac+1 }
```

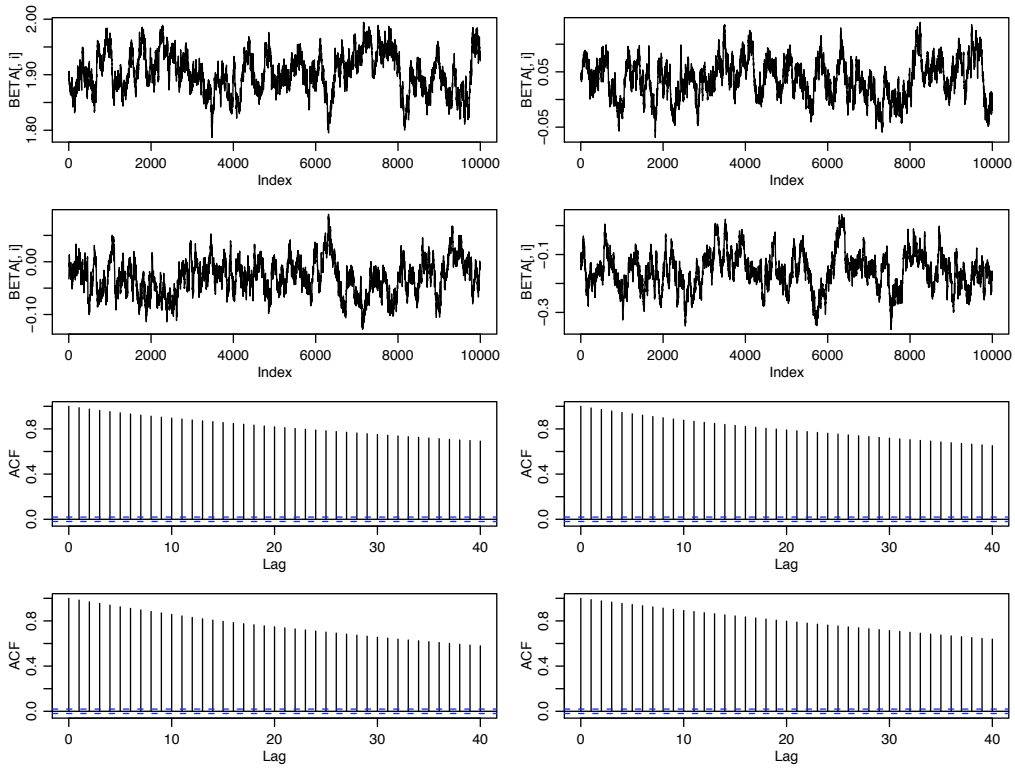


Figure 11.6: First MCMC estimation for Fijian analysis

```
BETA[s,]<-beta
                    }
ac/nscan
0.7559
```

Note that the acceptance rate is quite high. This is because we are proposing $\boldsymbol{\beta}^*$'s very close to $\boldsymbol{\beta}$, i.e. perhaps or proposal distribution is not diffuse enough. Figure 6 displays some standard MCMC diagnostic plots. As can be seen, the autocorrelation is very high, and the chain "mixes slowly". Can we improve the situation by using a different proposal distribution? Lets try

```
sd.prop<-summary(fit.mle)$coef[,2]
```

MCMC results with this proposal distribution are shown in Figure 7. We have the same number of scans, and the parameters are in the same region

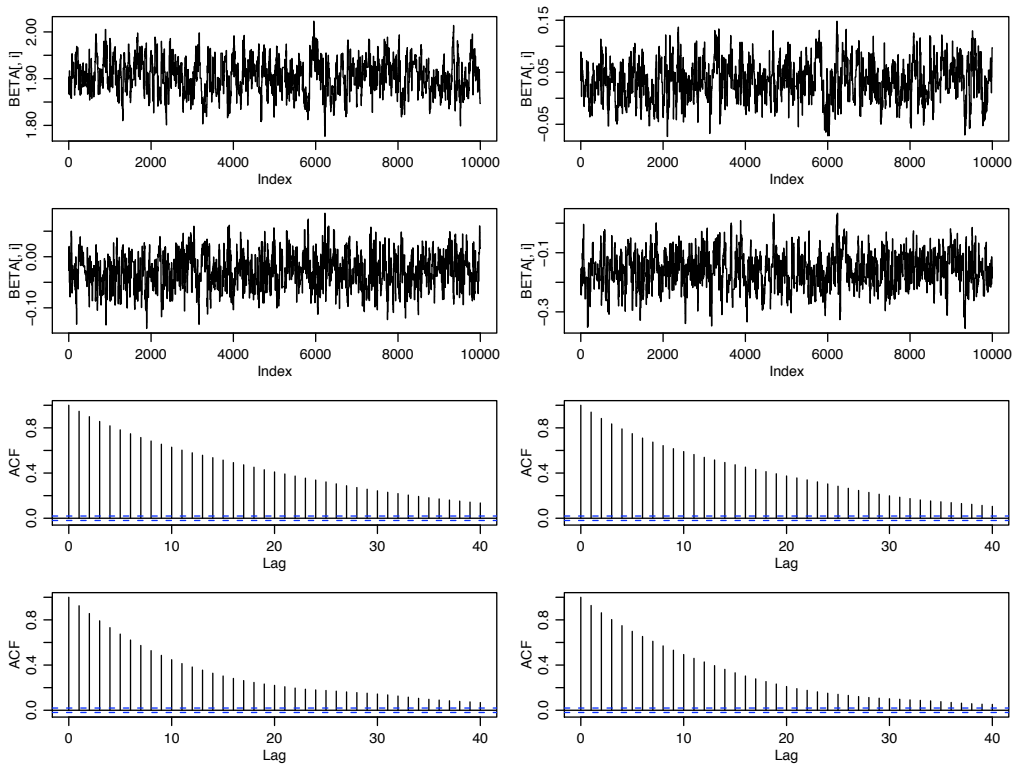


Figure 11.7: Second MCMC estimation for Fijian analysis

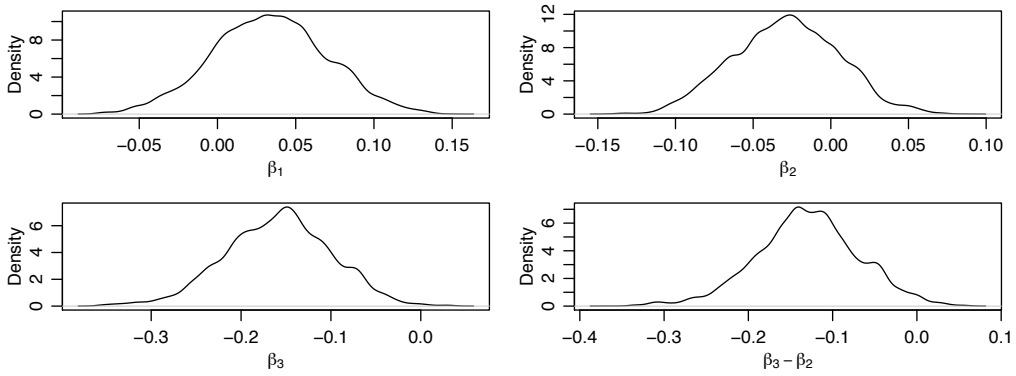


Figure 11.8: Density estimates from second MCMC estimation

of the parameter space, but the autocorrelation is much lower (even though the acceptance rate has dropped to 0.2094). This is an improvement. What do density estimates look like in this case?

Figure 8 shows that the Monte Carlo approximations to the densities are still rather bumpy: We don't quite have enough information from the MCMC samples to get good estimates of the density. We can remedy this situation by

- getting more samples, or better yet,
- getting more independent samples.

Consider saving only every 25th scan:

```
nscan<-50000
odens<-25
BETA<-matrix(0,nscan/odens,4)

for(s in 1:nscan) {
.
.
.
if( s%% odens ==0 ) { BETA[s/odens,]<-beta  }
                    }
```

So we only have saved 50000/25=2000 samples (compared to 10000 before), but they are nearly independent. Thus we have 2000 nearly independent samples from the posterior with which to make density estimates, which is a reasonable number. See the difference in the approximations to the posterior densities in Figure 10.

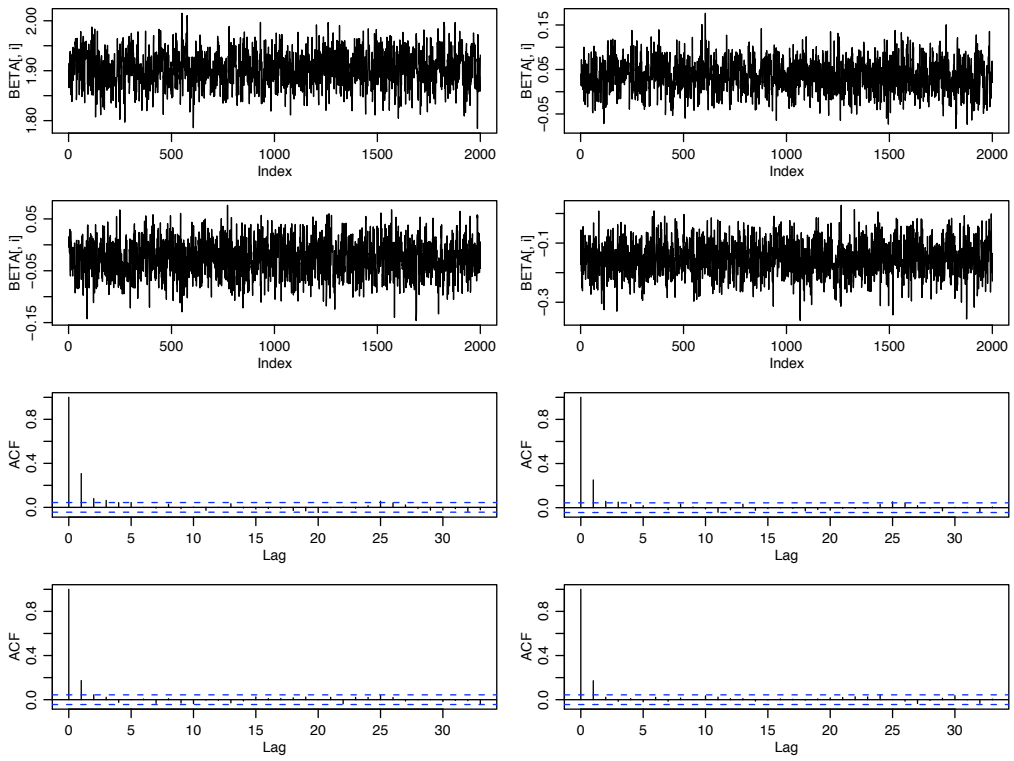


Figure 11.9: Every 25th scan of a length 50,000 scan Markov chain

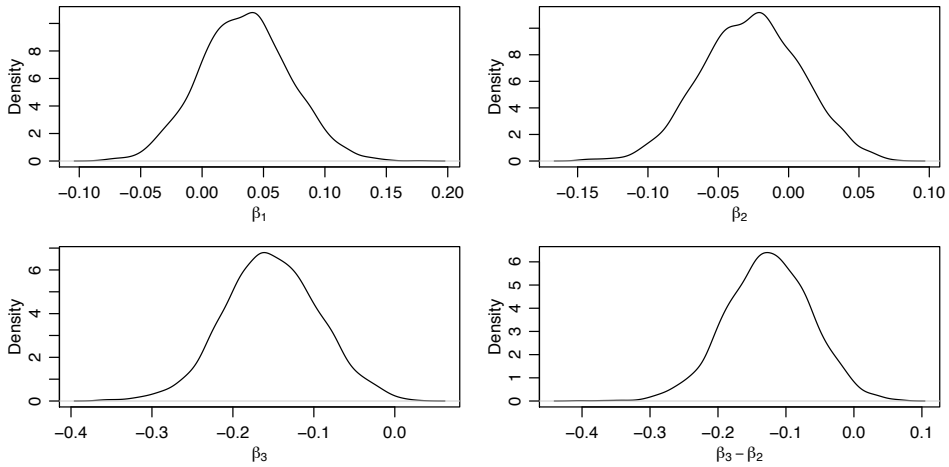


Figure 11.10: Density estimates from second MCMC estimation

11.1.4 Why does the Metropolis algorithm work?

Theorem: If $\{\theta_{(1)}, \ldots, \theta_{(s)}\}$ is a Markov chain that is irreducible , aperiodic and non-transient, then there is a unique probability distribution $\tilde{p}$ such that

- $p_s(\cdot) \to \tilde{p}(\cdot)$;
- $\frac{1}{s}\sum g(\theta_{(s)}) \to \int g(\theta)\tilde{p}(\theta)\, d\theta$

as $s \to \infty$. $\tilde{p}$ is called the **stationary distribution** of the Markov chain. It is called the stationary distribution because it has the following property:

- if $\theta_{(s)} \sim \tilde{p}$;
- and $\theta_{(s+1)}$ is generated from the Markov chain, i.e. $\theta_{(s+1)} \sim p_{(s+1)}(\theta_{(s+1)}|\theta_{(s)})$,
- then $p_{(s+1)}(\theta_{(s+1)}) = \tilde{p}(\theta_{(s+1)})$

in other words, if you sample $\theta_{(s)}$ from $\tilde{p}$, and then generate $\theta_{(s+1)}$ conditional on $\theta_{(s)}$ from the Markov chain, then *unconditional* on $\theta_{(s)}$, the distribution of $\theta_{(s+1)}$ is $\tilde{p}$, the stationary distribution. Thus " *the Markov chain leaves the stationary distribution unchanged*" The big question is whether or not $\tilde{p}(\theta) = p(\theta|y)$.

"Proof" that $\tilde{p}(\theta) = p(\theta|y)$: We will show that $p(\theta|y)$ is the stationary distribution of the Markov chain generated by the Metropolis algorithm. Suppose $\theta_{(s)}$ is sampled from the target distribution $p(\theta|y)$, and then $\theta_{(s+1)}$ is generated from $\theta_{(s)}$ using the Metropolis algorithm.

Let θ_a, θ_b be two values of θ such that $p(\theta_a|y) \leq p(\theta_b|y)$.

Then

$$\begin{aligned} P(\theta_{(s)} = \theta_a, \theta_{(s+1)} = \theta_b) &= P(\theta_{(s)} = \theta_a) \times P(\theta_{(s+1)} = \theta_b|\theta_{(s)} = \theta_a) \\ &= p(\theta_a|y) \times J(\theta_b|\theta_a) \times 1 \end{aligned}$$

because the acceptance ratio $r > 1$. On the other hand,

$$\begin{aligned} P(\theta_{(s)} = \theta_b, \theta_{(s+1)} = \theta_a) &= P(\theta_{(s)} = \theta_b) \times P(\theta_{(s+1)} = \theta_a|\theta_{(s)} = \theta_b) \\ &= p(\theta_b|y) \times J(\theta_a|\theta_b) \times \frac{p(\theta_a|y)}{p(\theta_b|y)} \\ &= p(\theta_a|y) \times J(\theta_a|\theta_b) \end{aligned}$$

Recall J is symmetric, so $J(\theta_b|\theta_a) = J(\theta_a|\theta_b)$ so for all values θ_a and θ_b of θ,

$$P(\theta_{(s)} = \theta_a, \theta_{(s+1)} = \theta_b) = P(\theta_{(s)} = \theta_b, \theta_{(s+1)} = \theta_a)$$

Now sum/integrate both sides, allowing θ_b to range over all possible values of θ.

$$P(\theta_{(s)} = \theta_a) = P(\theta_{(s+1)} = \theta_a)$$

But $\theta_{(s)}$ was sampled from $p(\theta|y)$, so both sides of the above equation are $p(\theta_a|y)$.

11.2 The Metropolis-Hastings algorithm

Let $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_p\}$ be a potentially vector-valued parameter. We've just discussed that the following algorithm generates approximate samples from $p(\boldsymbol{\theta}|y)$:

Metropolis Algorithm: Given $\{\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(s)}\}$, generate $\boldsymbol{\theta}_{(s+1)}$ by

1. sampling $\boldsymbol{\theta}^*$ from $J(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{(s)})$;
2. compute $r = \frac{p(\boldsymbol{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{p(\boldsymbol{y}|\boldsymbol{\theta}_{(s)})p(\boldsymbol{\theta}_{(s)})}$;
3. sample $u \sim \text{uniform}[0, 1]$;
4. set $\boldsymbol{\theta}_{(s+1)} = \boldsymbol{\theta}^*$ if $u < r$; otherwise, set $\boldsymbol{\theta}_{(s+1)} = \boldsymbol{\theta}_{(s)}$.

In some multiparameter problems, we could also use the following algorithm:

Gibbs sampling: Given $\{\boldsymbol{\theta}_{(1)}, \ldots, \boldsymbol{\theta}_{(s)}\}$, generate $\boldsymbol{\theta}_{(s+1)}$ by

1. sample $\theta_{(s+1),1}|\theta_{(s),2}, \ldots, \theta_{(s),p}, \boldsymbol{y}$;
2. sample $\theta_{(s+1),2}|\theta_{(s+1),1}, \ldots, \theta_{(s),p}, \boldsymbol{y}$;

   $\vdots$

p. sample $\theta_{(s+1),p}|\theta_{(s+1),1}, \ldots, \theta_{(s+1),p-1}, \boldsymbol{y}$;

	Proposal	Acceptance
Metropolis	$J(\theta^*\|\theta_s)$	$p(\theta^*\|\boldsymbol{y})/p(\theta_s\|\boldsymbol{y})$
Gibbs	$p(\theta_k\|\boldsymbol{\theta}_{-k},\boldsymbol{y})$	1

It turns out these are both special cases of a general Markov chain Monte Carlo procedure:

Metropolis-Hastings Algorithm: Given $\{\boldsymbol{\theta}_{(1)},\ldots,\boldsymbol{\theta}_{(s)}\}$, generate $\boldsymbol{\theta}_{(s+1)}$ by

1. sampling $\boldsymbol{\theta}^*$ from $J_s(\boldsymbol{\theta}^*)$;
2. compute $r = \frac{p(\boldsymbol{\theta}^*)p(\boldsymbol{y}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}_{(s)})p(\boldsymbol{y}|\boldsymbol{\theta}_{(s)})}\frac{J_s(\boldsymbol{\theta}_{(s)})}{J_s(\boldsymbol{\theta}^*)}$;
3. sample $u \sim$ uniform$[0,1]$;
4. set $\boldsymbol{\theta}_{(s+1)} = \boldsymbol{\theta}^*$ if $u < r$; otherwise, set $\boldsymbol{\theta}_{(s+1)} = \boldsymbol{\theta}_{(s)}$.

Here, J_s is not necessarily symmetric and can change as s changes. The only requirement is that J_s depends on $\{\boldsymbol{\theta}_{(1)},\ldots,\boldsymbol{\theta}_{(s)}\}$ only through $\boldsymbol{\theta}_{(s)}$. This ensures that the sequence is a **Markovchain**.

Example (Gibbs as a special case of MH): Consider the one-sample normal problem:

Model: $y_1,\ldots,y_n|\mu,\sigma \sim$ i.i.d. normal(μ,σ), i.e.

$$\begin{aligned} p(y_1,\ldots,y_n|\mu,\sigma) &= \prod_{i=1}^n p(y_i|\mu,\sigma) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2}\} \\ &= \texttt{prod(dnorm(y, mu, sigma))} \end{aligned}$$

Prior: $\mu \sim$ normal(μ_0,τ_0), $1/\sigma^2 \sim$ inverse-gamma$(\nu_0/2, \nu_0\sigma_0^2/2)$. i.e.

$$p(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\tau_0^2}}\exp\{-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\tau_0^2}\}\times\frac{1}{\Gamma(\nu_0/2)}\left(\frac{\nu_0\sigma_0^2}{2}\right)^{\nu_0/2}(\sigma^2)^{-\nu_0/2-1}\exp\{\frac{1}{2}\nu_0\sigma_0^2/\sigma^2\}$$

Posterior:

$$p(\mu, \sigma|\boldsymbol{y}) = \frac{p(\mu, \sigma)p(\boldsymbol{y}|\mu, \sigma)}{\int p(\mu, \sigma)p(\boldsymbol{y}|\mu, \sigma)\ d\mu d\sigma}$$

We can examine/look at/study the posterior distribution via samples from the posterior.

Gibbs sampling: Given $\mu_{(s)}, \sigma_{(s)}$,

1. sample $\mu_{(s+1)}$ from $p(\mu|\boldsymbol{y}, \sigma_{(s)})$;
2. sample $\sigma_{(s+1)}$ from $p(\sigma|\boldsymbol{y}, \mu_{(s)})$.

We will now show this is equivalent to a Metropolis-Hastings procedure: Let $\theta_{(s)} = \{\mu_{(s)}, \sigma_{(s)}\}$:

- If s is odd,
  - set $\sigma^* = \sigma_{(s)}$;
  - sample $\mu^* \sim p(\mu|y, \sigma_{(s)})$;

  So $J(\theta^*|\theta_{(s)}) = p(\mu^*|y, \sigma_{(s)})$ The acceptance probability is

$$\begin{aligned} r &= \frac{p(\theta^*|y)}{p(\theta_{(s)}|y)} \frac{J(\theta_{(s)}|y)}{J(\theta^*|y)} \\ &= \frac{p(\mu^*, \sigma_{(s)}|y)}{p(\mu_{(s)}, \sigma_{(s)}|y)} \frac{p(\mu_{(s)}|y, \sigma_{(s)})}{p(\mu^*|y, \sigma_{(s)})} \\ &= \frac{p(\mu^*|y, \sigma_{(s)})}{p(\mu_{(s)}|y, \sigma_{(s)})} \frac{p(\mu_{(s)}|y, \sigma_{(s)})}{p(\mu^*|y, \sigma_{(s)})} \\ &= 1 \end{aligned}$$

- If s is even,
  - set $\mu^* = \mu_{(s)}$;

– sample $\sigma^* \sim p(\sigma|y, \mu_{(s)})$; So $J(\theta^*|\theta_{(s)}) = p(\mu^*|y, \sigma_{(s)})$ The acceptance probability is

$$\begin{aligned} r &= \frac{p(\theta^*|y)}{p(\theta_{(s)}|y)} \frac{J(\theta_{(s)}|y)}{J(\theta^*|y)} \\ &= \frac{p(\sigma^*, \mu_{(s)}|y)}{p(\sigma_{(s)}, \mu_{(s)}|y)} \frac{p(\sigma_{(s)}|y, \mu_{(s)})}{p(\sigma^*|y, \mu_{(s)})} \\ &= \frac{p(\sigma^*|y, \mu_{(s)})}{p(\sigma_{(s)}|y, \mu_{(s)})} \frac{p(\sigma_{(s)}|y, \mu_{(s)})}{p(\sigma^*|y, \mu_{(s)})} \\ &= 1 \end{aligned}$$

So we see that Gibbs sampling is a Metropolis-Hastings procedure in which the proposal distributions are the full conditionals. Using the full conditionals leads to acceptance probabilities of 1.

11.3 Recap of the goals of MCMC

Here are the steps in a Bayesian statistical analysis:

Model specification. Specify a model for your data: $p(\boldsymbol{y}|\theta)$ should represent the sampling distribution of your data, given a specific set of parameters θ.

Prior elicitation. Specify a prior distribution: $p(\theta)$ should ideally represent someones (potential) prior information about the population parameter θ.

At this point, the posterior $p(\theta|\boldsymbol{y})$ is "determined." It is given by

$$p(\theta|\boldsymbol{y}) = \frac{p(\theta)p(\boldsymbol{y}|\theta)}{p(\boldsymbol{y})} = \frac{p(\theta)p(\boldsymbol{y}|\theta)}{\int p(\theta)p(\boldsymbol{y}|\theta)\ d\theta}$$

And so in a sense there is no more modeling. All that is left is

Examination of the posterior distribution. Compute posterior means medians, modes, probabilities and confidence regions, all derived from $p(\theta|\boldsymbol{y})$.

Now sometimes, $p(\theta|\boldsymbol{y})$ is complicated, hard to write down, etc. Our method for "looking at" $p(\theta|\boldsymbol{y})$ in this case has been by studying Monte-Carlo samples from $p(\theta|\boldsymbol{y})$. Monte Carlo samples should not be confused with data samples $y_1, \ldots, y_n$ from some population of interest: Monte Carlo sampling and MCMC algorithms

- are not models, nor do they generate "more information" than is in $\boldsymbol{y}$ and $p(\theta)$;
- they are simply "ways of looking at" $p(\theta|\boldsymbol{y})$.

For example, if we have MCMC samples $\theta_{(1)}, \ldots, \theta_{(S)}$ that are approximate draws from $p(\theta|\boldsymbol{y})$, then these samples help describe $p(\theta|\boldsymbol{y})$:

- $\frac{1}{S}\sum \theta_{(s)} \approx \int \theta p(\theta|\boldsymbol{y})\ d\theta$
- $\frac{1}{S}\sum 1(\theta_{(s)} \leq c) \approx P(\theta \leq c|\boldsymbol{y}) = \int_{-\infty}^{c} p(\theta|\boldsymbol{y})\ d\theta$.

and so on. So keep in mind, these MCMC procedures are simply ways to approximate/look at $p(\theta|\boldsymbol{y})$.

11.3.1 Recap children example

Survey of Fijian women who had been married more than 20 years.

- y_i = number of children birthed;
- $x_{i,1}$ = rural residence (0 or 1);
- $x_{i,2}$ = primary school education (0 or 1);
- $x_{i,3}$ = education beyond primary school (0 or 1).

for $i = 1, \ldots, 641$.

Model specification:

1. Goal: relate education and residency to the number of children a woman has in her lifetime.

2. Model: Response y_i is a count. A convenient model for y_i is the Poisson distribution. We estimate the multiplicative (log-linear) effects of x_i via the relationship

$$E(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = \exp\{\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}\}$$

Thus the $\boldsymbol{\beta}$'s represent the multiplicative effects of the x's.

Prior elicitation: Based on previous studies or other knowledge specify a prior $p(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$. A common choice is a multivariate normal distribution, but any probability distribution can be used.

At this point $p(\boldsymbol{\beta}|\boldsymbol{y})$ is determined. There is, in a sense no further input from the data or prior. The researcher may be interested in specific aspects of $p(\boldsymbol{\beta}|y)$, such as

- A 95% posterior confidence region for β_1, the effect of rural residence.
- $\Pr(\beta_3 < 0|y)$ the probability that upper level education is negatively related to the number of children on average.
- $P(Y_{\text{new}} = y|x_{\text{new}})$ the predictive distribution of the number of children for a woman with covariates x_{new}.

These are all things that can be derived from $p(\boldsymbol{\beta}|\boldsymbol{y})$.

Examination of the posterior distribution:

1. We approximate the above aspects of $p(\boldsymbol{\beta}|y)$ with samples of the $\boldsymbol{\beta}$'s from $p(\boldsymbol{\beta}|\boldsymbol{y})$.
2. Since our model/prior is not conjugate, we cannot generate independent samples of the $\boldsymbol{\beta}$ directly from $p(\boldsymbol{\beta}|\boldsymbol{y})$. Instead we generate MCMC samples to describe $p(\boldsymbol{\beta}|\boldsymbol{y})$.

Chapter 12

Generalized linear mixed effects models

12.1 Generalized linear models

Goal: Relate response y_i to explanatory variables $x_{i,1}, \ldots, x_{i,p}$.

A generalized linear model is a model where some function of the mean is linear in the explanatory variables:

$$g(E[Y_i|\beta, \boldsymbol{x}_i]) = \beta_0 + \beta_1 x_{i,1} + \cdots \beta_p x_{i,p}$$

- $g(\cdot)$ is called the *link* function;
- $\eta_i = \beta' \boldsymbol{x}_i$ is called the *linear predictor*.

Examples: Let $\mu_i = E(Y_i|\beta, \boldsymbol{x}_i)$

- Logistic regression

$$\begin{aligned} \mu_i = \Pr(Y_i = 1) &= \frac{e^{\beta' x_i}}{1 + e^{\beta' x_i}} \\ \log \frac{\mu_i}{1 - \mu_i} &= \beta' x_i \end{aligned}$$

so $g(\mu_i) = \log \frac{\mu_i}{1-\mu_i}$ is the logit link.

- Poisson regression

$$\begin{aligned} \Pr(Y_i = y_i|\mu_i) &= \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \\ \mu_i &= e^{\beta' \boldsymbol{x}_i} \\ \log \mu_i &= \beta' \boldsymbol{x}_i \end{aligned}$$

  so $g(\mu_i) = \log \mu_i$ is the log link.

- Normal regression

$$\mu_i = \beta' \boldsymbol{x}_i$$

  so $g(\mu_i) = \mu_i$ is the identity link.

12.2 Mixed effects models

Now suppose we have multilevel sampling: $Y_{i,j}$ is the response from the

- ith individual in the
- jth group,

$i = 1, \ldots, n_j$, $j = 1, \ldots, J$. How should we model these data? Consider two standard approaches in the non-Bayesian realm.

No between-group differences:

$$g(\mu_{i,j}) = \beta_0 + \beta_1 x_{i,j,1} + \cdots \beta_p x_{i,j,p}$$

No between-group similarities:

$$g(\mu_{i,j}) = \beta_{0,j} + \beta_{1,j} x_{i,j,1} + \cdots \beta_{p,j} x_{i,j,p}$$

Of course, we could have some components of β vary across groups, and others not.

Within group variation

$$\begin{aligned} \mu_{i,j} &= E(y_{i,j}|\boldsymbol{\beta}_j, \boldsymbol{x}_{i,j}) = g^{-1}(\boldsymbol{\beta}_j' \boldsymbol{x}_{i,j}) \\ p(y_{1,1}, \ldots, y_{n_J,J}|\mu_{1,1}, \ldots, \mu_{n_J,J}) &= \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(y_{i,j}|\mu_{i,j}) \end{aligned}$$

Between group variation

$$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J \sim \text{i.i.d. multivariate normal}(\boldsymbol{\beta}, \Sigma)$$

where $\boldsymbol{\beta} = \{\beta_0, \beta_1, \ldots, \beta_p\}$ and Σ is a $(p+1)\times(p+1)$ covariance matrix. For now, lets just use $\Sigma = \text{diag}(\sigma_0^2, \sigma_1^2, \ldots, \sigma_p^2)$.

12.2.1 Variability of $\beta_1, \ldots, \beta_J$

Consider the following data analysis approaches:

Fix $\boldsymbol{\beta}$ and Σ : If you say you know $\boldsymbol{\beta}$ and Σ, then the distribution

$$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J \sim \text{i.i.d. multivariate normal}(\boldsymbol{\beta}, \Sigma)$$

can be viewed as a *prior*. Each $\boldsymbol{\beta}_j$ is estimated separately, and there is no information shared across groups.

$$p(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{-j}, \boldsymbol{\beta}, \Sigma) = p(\boldsymbol{\beta}_j|\boldsymbol{\beta}, \Sigma)$$

This can be viewed as the Bayesian version of the *fixed effects model*, with group-specific parameters. Your posterior estimate for $\boldsymbol{\beta}_j$ is not affected by the data from groups other than j.

Fix $\Sigma = \text{diag}(\mathbf{0})$, estimate $\boldsymbol{\beta}$: If we let $\Sigma = \text{diag}(\mathbf{0})$ then the between-group variability is zero, and $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_J = \boldsymbol{\beta}$. A multivariate normal prior on $\boldsymbol{\beta}$ then reduces this case to the standard regression problem $g(E(y_{i,j}|\boldsymbol{\beta}, \boldsymbol{x}_{i,j})) = \boldsymbol{\beta}'\boldsymbol{x}_{i,j}$, with no between group variance (i.e. no within group correlation).

Estimate Σ and $\boldsymbol{\beta}$: In this case,

$$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J \sim \text{i.i.d. multivariate normal}(\boldsymbol{\beta}, \Sigma)$$

represents not a prior, but an *unknown sampling distribution* for the $\boldsymbol{\beta}_j$'s. The $\boldsymbol{\beta}_j$'s share information, because although

$$\begin{aligned} p(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{-j}, \boldsymbol{\beta}, \Sigma) &= p(\boldsymbol{\beta}_j|\boldsymbol{\beta}, \Sigma), \text{ we have} \\ p(\boldsymbol{\beta}_j|\boldsymbol{\beta}_{-j}) &\neq p(\boldsymbol{\beta}_j) \end{aligned}$$

This is the *random effects* model, and your posterior estimate for $\boldsymbol{\beta}_j$ *is* influenced by data from groups other than j.

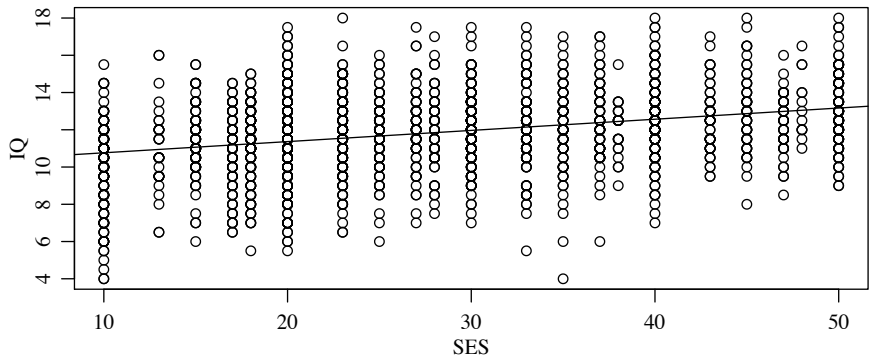
To summarize,

- Fixed effects:
  - $\boldsymbol{\beta}$, Σ fixed;
  - $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J \sim$ i.i.d. multivariate normal$(\boldsymbol{\beta}, \Sigma)$ is a prior;
  - $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$ are independent of each other in the posterior.
- Random effects:
  - $\boldsymbol{\beta}$, Σ are unknown, have a prior, and are estimated.
  - $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J \sim$ i.i.d. multivariate normal$(\boldsymbol{\beta}, \Sigma)$ is a sampling distribution;
  - $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$ are dependent in the posterior.

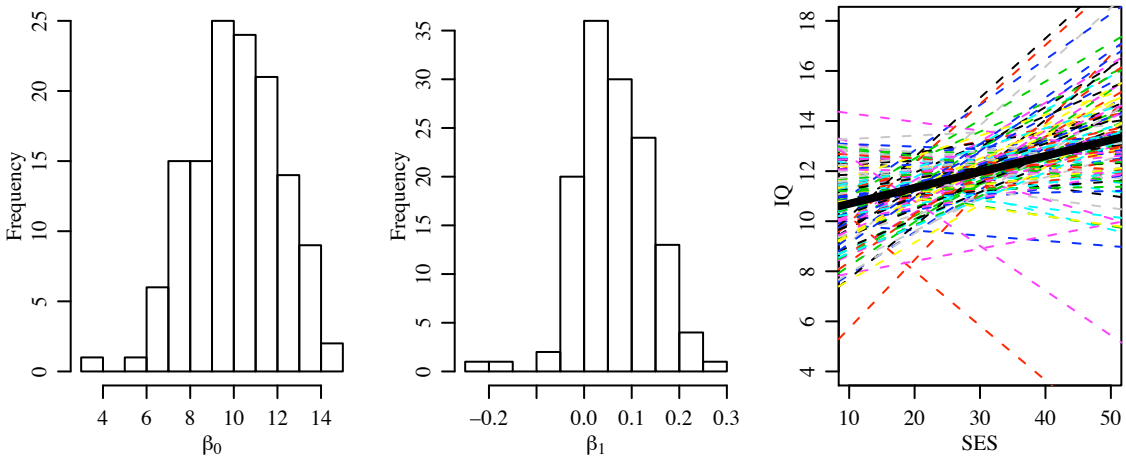
Compare to the situation in a non-Bayesian setting

- Fixed effects:
  - $\boldsymbol{\beta}$, Σ not estimated;
  - $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$ estimated separately.
- Random effects:
  - $\boldsymbol{\beta}$, Σ estimated;
  - $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$ not really estimated.

12.3 Example (Eighth graders in the Netherlands):



```
> summary(lm(dat$IQ~dat$SES))

Call:
lm(formula = dat$IQ ~ dat$SES)

Residuals:
     Min       1Q   Median       3Q      Max
-8.26595 -1.16511  0.03447  1.21473  6.45506

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.163000   0.112430   90.39   <2e-16 ***
dat$SES      0.060084   0.003764   15.96   <2e-16 ***


B0B1<-NULL
for(j in 1:J) {
B0B1<-rbind(B0B1, lm( dat$IQ[class==j]~dat$SES[class==j] )$coef )
              }
```



12.3.1 Priors and posteriors for linear random effects models

The model we are considering is

Between group variance:

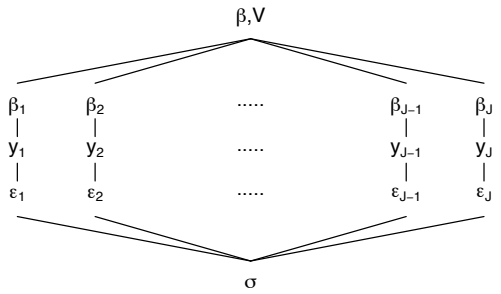
$$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J | \boldsymbol{\beta}, V \sim \text{i.i.d. multivariate normal}(\boldsymbol{\beta}, V)$$

Within group variance:

$$y_{i,j} \sim \text{normal } (\boldsymbol{\beta}_j' \boldsymbol{x}_{i,j}, \sigma) \quad \text{independently across individuals and groups}$$

The unknown quantities we would like to have information about are

$$\{\boldsymbol{\beta}, V\},\ \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J\},\ \sigma$$



We have specified sampling distributions for $\boldsymbol{y}_j$'s and the $\boldsymbol{\beta}_j$'s. The priors we need to specify are for $\boldsymbol{\beta}, V$ and σ^2. We will use the following for now:

- $1/\sigma^2 \sim$ inverse gamma $(\nu_0/2, \nu_0\sigma_0^2/2)$;
- $\boldsymbol{\beta} \sim$ multivariate normal $(\boldsymbol{\beta}_0, \Lambda)$;
- $V^{-1} \sim$ Wishart (S_0^{-1}, η_0)

Your job in the final homework will be to show that the following are the full conditionals of the parameters (or find the full conditionals in a similar setup).

- $\boldsymbol{\beta}_j | \ldots \sim$ multivariate normal $(\hat{\boldsymbol{\beta}}_j, \hat{V}_j)$;
- $\sigma^2 | \ldots \sim$ inverse gamma $(\frac{\hat{\nu}}{2}, \frac{\hat{\nu}\hat{\sigma}^2}{2})$;
- $\boldsymbol{\beta} | \ldots \sim$ multivariate normal $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$, where
  - $\hat{\Lambda} = (\Lambda_0^{-1} + JV^{-1})^{-1}$;

  - $\hat{\boldsymbol{\beta}} = \hat{\Lambda}(\Lambda_0^{-1}\boldsymbol{\beta}_0 + JV^{-1}\bar{\boldsymbol{\beta}})$;

- $V^{-1}|\ldots \sim$ inverse Wishart $(S_n^{-1}, \eta_0 + J)$, where
  - $S_n = S_0 + \sum_{j=1}^{J}(\boldsymbol{\beta}_j - \boldsymbol{\beta})(\boldsymbol{\beta}_j - \boldsymbol{\beta})'$.

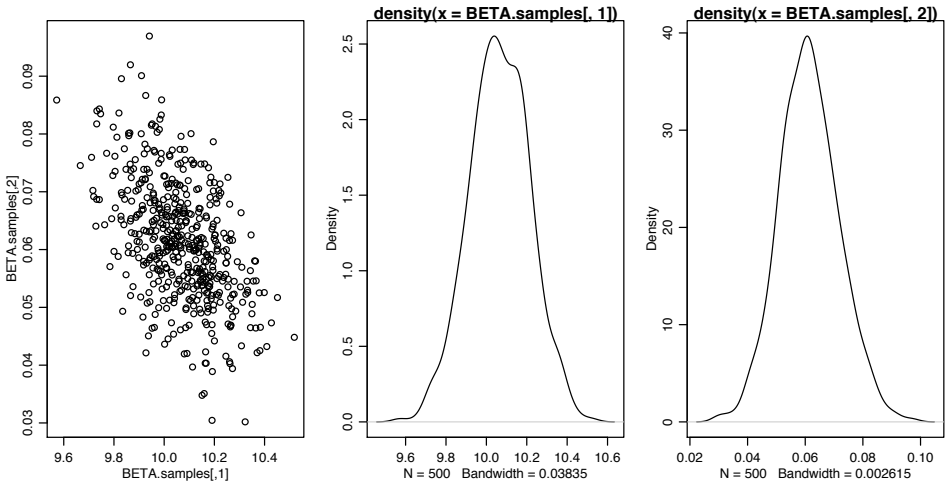
In the above $\hat{\boldsymbol{\beta}}$ is the vector mean of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J$. Since in the linear regression case we have all the full conditionals, we can approximate the posterior with Gibbs sampling.

12.3.2 Back to the example

```
#####  priors
nu0<-3
s20<-2
beta0<-rep(0,2)
Gam0<-diag(rep(100,2))
S0<-diag(rep(1,2))
eta0<-3
#####

##### starting values
beta<-rep(0,2)
s2<-1
V<-diag(rep(100,2))
#####

##### things to store parameters in
BETA.J<-matrix(nrow=J,ncol=2)
BETA.J.samples<- list()
for(j in 1:J) { BETA.J.samples[[j]]<- matrix(0,0,2) }
S2.samples<-NULL
BETA.samples<-NULL

##### output density
odens<-10

##### MCMC
for(ns in 1:5000) {

## sample new beta.j's and compute the SSE for them
SSE<-0
for(j in 1:J) {
```

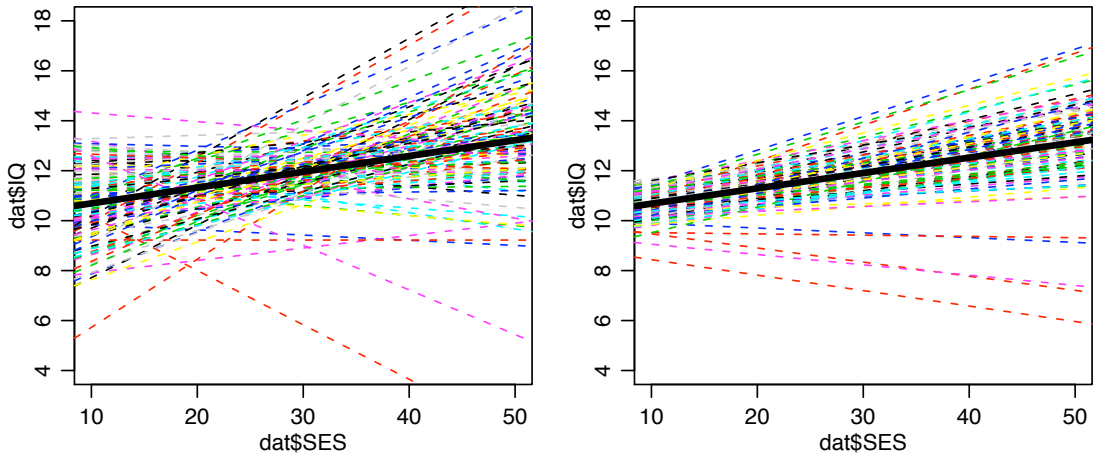
```
yj<-dat$IQ[class==j]
xj<-cbind( rep(1,length(yj)), dat$SES[class==j] )
V.j<- XXXXX
beta.hat.j<-  XXXXX
BETA.J[j,]<-rmvnorm( beta.hat.j, V.j)
SSE<-SSE+ sum((yj - BETA.J[j,]%*%t(xj) )^2)
               }
##

## sample new s2
s2<-1/rgamma( 1, XXXXX/2 , XXXXX/2 )

## sample new population mean of the beta.j's
beta.j.bar<-apply(BETA.J,2,mean)
Gam.hat<- XXXXX
beta.hat<- XXXXX
beta<-rmvnorm( beta.hat, Gam.hat)

## sample new population variance/covariance  of the beta.j's
Sn<-S0
for(j in 1:J) { Sn<-Sn+ (BETA.J[j,]-beta)%*% t( (BETA.J[j,]-beta) ) }
V<-solve( rwish( solve(Sn), eta0+J) )

## done with scan, now store output
if(ns %% odens ==0 ) {
for(j in 1:J) { BETA.J.samples[[j]]<-rbind(BETA.J.samples[[j]],BETA.J[j,] ) }
S2.samples<-c(S2.samples,s2)
BETA.samples<-rbind(BETA.samples,t(beta))
cat(ns,beta,s2,"\n")
                     }
                    }

## done with MCMC, compute posterior means of the beta.j's

BETA.J.pm<-BETA.J*0
for(j in 1:J) { BETA.J.pm[j,]<-apply( BETA.J.samples[[j]],2,mean) }
```

Lets look at the marginal posterior distribution of $\boldsymbol{\beta}$:

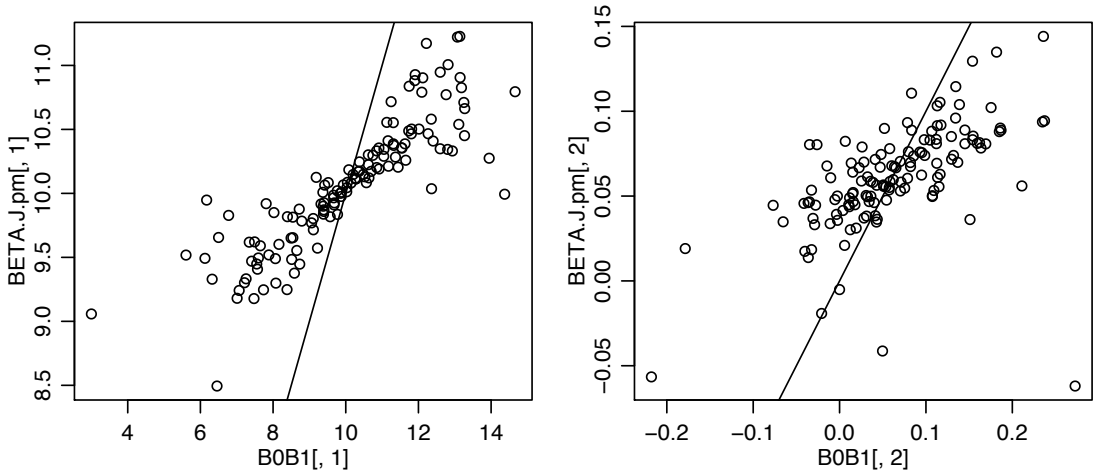


These posterior distributions represent uncertainty in the population mean of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J, \boldsymbol{\beta}_{J+1}, \ldots$. It does not represent the sampling variability of the $\boldsymbol{\beta}_j$'s .

Now lets compare the posterior means of the $\boldsymbol{\beta}_j$'s to the estimates obtained from doing a regression for each school separately:



Another way to look at the shrinkage:



Based on these plots, what do you think the estimate of V is?

12.3.3 MCMC for GLME's

Example: Logistic regression.

$$\begin{aligned} \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J &\sim \text{i.i.d. multivariate normal}(\boldsymbol{\beta}, V) \\ P(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_j | \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J, \boldsymbol{X}) &= \prod_{j=1}^{J} \prod_{i=1}^{n_j} \left(\frac{e^{\boldsymbol{\beta}_j' \boldsymbol{x}_{i,j}}}{1 + e^{\boldsymbol{\beta}_j' \boldsymbol{x}_{i,j}}} \right)^{y_{i,j}} \left(\frac{1}{1 + e^{\boldsymbol{\beta}_j' \boldsymbol{x}_{i,j}}} \right)^{1 - y_{i,j}} \end{aligned}$$

Priors:

- $\boldsymbol{\beta} \sim$ multivariate normal $(\boldsymbol{\beta}_0, \Lambda)$;
- $V^{-1} \sim$ Wishart (S_0^{-1}, η_0)

Full conditionals? For $\boldsymbol{\beta}$ and V,

- $\boldsymbol{\beta} | \ldots \sim$ multivariate normal $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$, where
  - $\hat{\Lambda} = (\Lambda_0^{-1} + JV^{-1})^{-1}$;
  - $\hat{\boldsymbol{\beta}} = \hat{\Lambda}(\Lambda_0^{-1} \boldsymbol{\beta}_0 + JV^{-1} \bar{\boldsymbol{\beta}})$;
- $V^{-1} | \ldots \sim$ inverse Wishart $(S_n^{-1}, \eta_0 + J)$, where
  - $S_n = S_0 + \sum_{j=1}^{J} (\boldsymbol{\beta}_j - \boldsymbol{\beta})(\boldsymbol{\beta}_j - \boldsymbol{\beta})'$.

Full conditionals are not available for the $\boldsymbol{\beta}_j$'s.

Metropolis-Hastings algorithm:

1. For $j = 1, \ldots, J$
   (a) Propose a $\boldsymbol{\beta}_j^*$ from a symmetric proposal distribution around $\boldsymbol{\beta}_j$.
   (b) Accept or reject with the appropriate probability.
2. Sample $\boldsymbol{\beta}$ from its full conditional;
3. Sample V from its full conditional.

This combines Metropolis steps with Gibbs steps. Since both types of steps are types of Metropolis-Hastings steps, this can be seen as a proper Metropolis-Hastings algorithm.