Bayesian Theory

Ken Newman with updates by Finn Lindgren

2025/26

Contents

Course Information

Personnel & Meeting Times

- Finn Lindgren, Course Organizer and Primary Lecturer
- Additional Tutors: ...
- *Lectures*:
  Week 1-11: Tuesday 10:00–11:50 Nucleus Building, 1.15 Larch Lecture Theatre
- *Workshops*:Tuesdays of Weeks 1, 3, 5, 7, 9, 11
  - Workshop Session 1, 12:10–13:00, Murchison 1.19: ...
  - Workshop Session 2, 13:10–14:00, Murchison 1.19: ...
  - Workshop Session 3, 14:10–15:00, Murchison 1.19: ...
- *Office Hours* TBD.

Primary Materials

With the exception of the Reich & Ghosh book, the materials below will be made available on Learn in the Course Materials section. The materials are organised by week (Weeks 1 through 11).

- Book: **Bayesian Statistical Methods** (2019) by B.J. Reich and S.K. Ghosh, CRC Press. (Online access.)
- Course lecture notes organised by week.
- Recordings of This Year's Lectures
- Homework and (non-assessed) Exercises that focus on particular aspects of lecture notes and/or videos.
- Workshop assignments on odd numbered weeks.

Assessment

Coursework is worth 20% and the final examination is worth 80% of the course mark. The coursework consists of four online quizzes. Each quiz is then worth 5%. The quizzes will be made available at 09:00(morning) on Tuesday of weeks 4, 6, 8, and 10 and will be due at 16:00(afternoon) on Monday of the following week. A non-assessed trial quiz will be available from week 2.

Quiz	Week	Available	Due
Trial	2	23 Sep	29 Sept
1	4	7 Oct	13 Oct
2	6	21 Oct	27 Oct
3	8	4 Nov	10 Nov
4	10	18 Nov	24 Nov

Exams from previous years will be made available on Learn around week 8 during the semester.

Other References

- Applied Bayesian Statistics, With R and OpenBUGS Examples. 2013. M.K. Cowles. Springer. Online access.
- Bayesian Methods for Data Analysis, 3rd Ed. 2009. B. Carlin and RT. Louis. CRC Press. Online access.
- Bayesian Data Analysis, 3rd Ed. 2013. A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. CRC Press. Online access.
- 2017/18 course notes by King and Ross. These notes overlap considerably with the current course notes and it may be helpful to read different explanations. On Learn in the Course Materials Section.
- Bayesian statistics: an introduction 4th Ed. 2012. P.M. Lee. Wiley. Online access.
- Bayesian Analysis for Population Ecology. 2010. King, Morgan, Gimenez, and Brooks. CRC Press. Online Access.
- Bayesian Models. 2015. Hobbs and Hooten. Princeton University Press. Online access.
- Bringing Bayesian Models to Life. 2019. Hooten and Hefley. CRC Press. (Not online.)
- Markov chain Monte Carlo: Stochastic simulation for Bayesian Inference, 2nd Ed. Gamerman and Lopes. (2006) CRC Press. Online access.
- Computational Statistics, 2nd Ed. 2012. G. Givens and J. Hoeting. Wiley. Online access.

Lectures and Syllabus (preliminary)

1. L1. Probability Review, Bayes Theorem (Discrete & Continuous cases), Bayesian Statistical Inference, Prior and Posterior Prediction.
2. L2. Priors Part 1: Subjective vs Objective, Conjugate Priors, Mixture Priors.
3. L3. Priors Part 2: Specifying Hyperparameters, Normal Distribution Priors.
4. L4. Priors Part 3: Jeffreys' Prior, Eliciting Priors, Analysing Priors.
5. L5. Summarising Posteriors: Loss Functions & Bayesian Estimators, Credible Intervals. Hypothesis Testing: Simple vs Composite Hypotheses and Posteriors, Bayes Factors.
6. L6. Bayesian Computation: Numerical Integration, Normal Approximation and Bayesian Central Limit Theorem.
7. L7: Bayesian Computation: Introduction to Monte Carlo Integration, Direct Sampling, Inverse Probability Integral Transform, Rejection Sampling, Importance Sampling, Sampling Importance Resampling.
8. L8: Bayesian Computation: Introduction to Markov Chains, Metropolis-Hastings (Part 1).
9. L9: Bayesian Computation: Metropolis-Hastings (cont), Multiple Parameters, MCMC Diagnostics (Part 1).
10. L10: Bayesian Computation: MCMC Diagnostics (Part 2), Gibbs Sampling.
11. L11: Review for examination.

1 Bayesian Theory–Introduction and Motivation

Reading

1. Read all of this Chapter 1.
2. Examine Distributional Summaries shown in Appendices A and B of King & Ross lecture notes. (On Learn: Course Material $->$ Other Readings $->$ 2017-2018 BT: King & Ross LN.)
3. (Optional) Chapters 1 and 2 of *Applied Bayesian Statistics With R and OpenBUGS Examples* by Mary Kathryn Cowles. The full text is available online via the University of Edinburgh Library webpage.
4. (Optional) Read Sections 1.1-1.3 (pp 5-12) in the King & Ross lecture notes.

1.1 Human Needs & Statistics

A larger perspective, starting with a human-centred focus.

1. What are basic human needs? Maslow's hierarchy (much of the following is taken from Wikipedia).

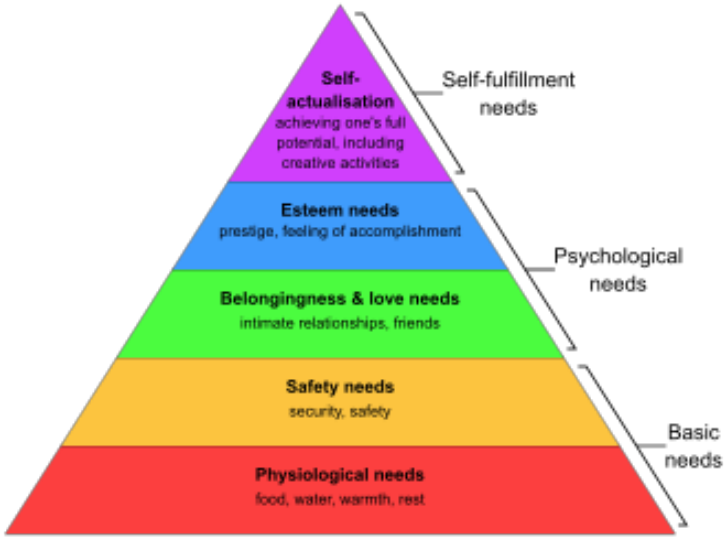


Figure 1.1: Maslow's Hierarchy of Needs. Taken from Wikipedia

The 1st four levels are sometimes called Deficiency Needs— humans must have them met to "survive". The top level includes Growth Needs.

Statistics plays a role in determining whether or not, and how well, some of these needs are met. Focusing on 3 basic needs (very incompletely):

- *Physiological needs.* Includes Water, Food, Air, Heat, Shelter.
- *Safety needs.* Includes Health, Personal Security, Safety, Emotional Security, Financial Security.
- *Belongingness & love needs.* Includes Family, Friendship, Intimacy, Trust, . . .

2. Expanding on the above 3 needs.
   - Physiological: Do people have enough water for drinking? For bathing, washing? Is the water safe, clean, unpolluted—free of toxic chemicals and harmful organisms?

     Are people getting enough food? Are they getting healthy, nutritious, life enhancing food? How many are dying from hunger, starving, malnourished? Is the food high enough quality to keep them healthy, thriving?

     Is the air clean enough to breathe? How polluted by smoke, particulate matter, toxic substances?

Do people have adequate shelter, housing, a roof and protection from wind, rain, snow, cold, able to stay warm?

Are people getting enough Medical care? Safe from infectious diseases (plagues, epidemics, pandemics)?

- Safety: What are the risks of death from different causes? Are people in war zones? What are the levels of crime, violence, bodily harm, domestic violence? How stressful is their life, their degree of financial security, their emotional security?

  What are the educational opportunities that can lead to jobs, financial security? (Higher level: cognitive needs.)

- Belongingness & love needs: What are the family structures? What community, social, spiritual organizations exist? Teams, work groups, clubs?

3. Data: What is known?

   Worldometer, https://www.worldometers.info/.

   - Population & Demographics: In 2022: 7.9 Billion people, 99 Million born this year, 41 Million die this year.

     Contrast: 1950: 2.5 Billion. Over 300% increase in 72 years!

   - *Physiological*:
     - Food: 864 Million undernourished
     - Food: 818 Million obese
     - Food: 5.5 Million starve to death this year
     - Water: 780 Million no access to clean water
     - Water: 0.6 Million die from water related diseases this year
     - Health: 9.2 Million die from Communicable diseases this year
     - Health: 5.4 Million $<$ children age 5 die this year
     - Health: 5.8 Million die from Cancer this year
     - Health: 1.2 Million die from HIV/AIDS this year
     - Health: 0.3 Million die from Malaria this year
     - Air: 25 Trillion tons CO2 emitted this year
     - Air/Land: 3.7 Million ha forest land lost this year
     - Air/Water: 6.9 Million tons Toxic Chemicals released this year
     - Shelter: 150 Million homeless; https://make-the-shift.org/homelessnessaction/
   - *Safety*

- Safety: 2 Billion people in conflict areas; https://www.voanews.com/a/un-chief-2-billion-people-live-in-conflict-areas-today/6509020.html
- Safety: crime rates per 100,000 people Venezuela (83.76); Papua New Guinea (80.79); South Africa (76.86); Afghanistan (76.31); . . . UK (46.1) . . . https://worldpopulationreview.com/country-rankings/crime-rate-by-country

- *Belongingness & love.* Measures of income inequality, happiness, social mobility.

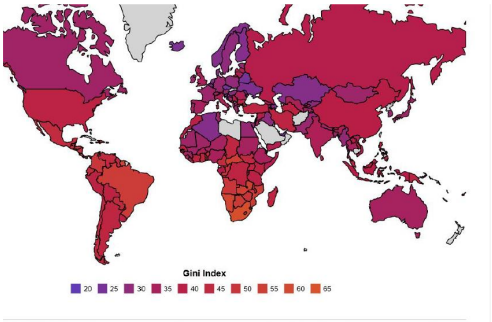


Figure 1.2: Income Inequality (Gini Index)

4. How do we know? **Statistics**: in particular Sampling and Censuses.

   How to collect data properly: Statistical Sampling Theory & Methods.

5. Using the Data: Models with Inputs & Outputs.

$$y = f(x_1, \ldots, x_k, \Theta) \tag{1.1}$$

   For example, The Happiness Index:

   Cantril Ladder $\sim$ Normal $\left(\beta_0 + 0.36 \text{ log GDP} + 2.42 \text{ Social support} + \ldots - 0.074 \text{ Corruption}, \sigma^2\right.$

   - How to fit the model? Estimate β's.
   - Interpreting the model? What is most important?
   - Making predictions with the model.
   - Assessing effects of "actions": change in environment, change in management, change in policy

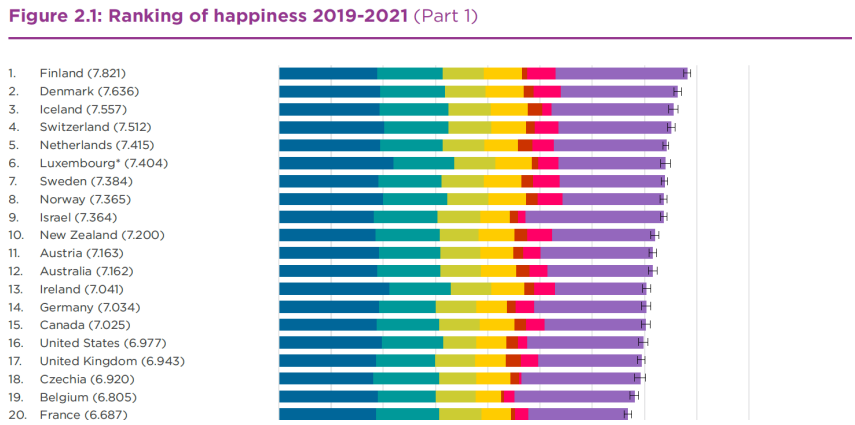


Figure 1.3: Happiness Index. `https://worldhappiness.report/faq/`

1.2 Scientific Method and Bayesian Statistics

The scientific method can be viewed as a framework for learning about the world: nature, biological and physical processes, human beings and activities. Cowles (2013) (taken from Berry (1996)) describes the scientific method as the following five step process:

1. Define the question or the problem to be addressed.

   For example: water temperatures in freshwater streams with bare banks (no vegetation) can sometimes get too hot for aquatic life. If vegetation is planted along the banks (habitat restoration), how much might temperatures be lowered, and how much might aquatic life survival increase?

2. Assess the relevant available information. Decide whether it is sufficient for the purpose at hand:

   (a) If yes, draw appropriate conclusions, make appropriate decisions, and take appropriate actions.

   (b) If no, go to the next step.

   Example: literature review in another location indicates that planting vegetation x at density y may lower water temperatures during hottest summer month by 2^{o} C, and increase survival of fish species z by 5%. But that vegetation, stream environment, and fish species are very different from our situation: need more information.

3. Determine what additional information is needed and design a sample, a study, or an experiment to get it.

   Example: may use an “artificial” setting with aquariums and controlled water temperatures, and different fish species, or conduct a field experiment with vegetation planting along randomly selected stream banks.

4. Collect new information: select the sample or carry out the study or experiment from step 3.

5. Use the data obtained in step 4 to update what was previously known. Return to step 2.

Statistics in general are central to steps 2, 3, and 5. For example, with step 2 one might have data from previous samples or studies or experiments that could be analysed. With step 3, statistics are used to draw probability samples from one or more existing populations, or use methods for setting up a designed experiment for comparing two or more procedures, for example. Step 5 involves the analysis of sample or experimental data using a variety of methods, e.g., estimating parameters.

Bayesian statistics is particularly well suited for steps 2 and 5 as it providing a quantitative framework for assessing current knowledge (step 2) and then updating or revising that knowledge by incorporating the new information gained (step 5).

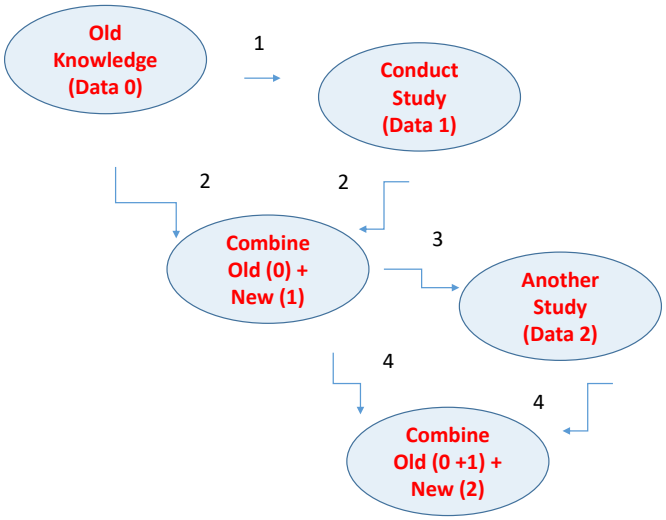


Figure 1.4: A general schematic for Bayesian statistics

Neither frequentistic nor Bayesian methods provide an answer to how to (in step 5) update the knowledge about the fundamental model structure, but they can provide toolkits for helping with such tasks. However, Bayesian methodology provides a coherent framework for updating probabilistic information related to the model structure, and general probability theory can be used to design model assessment procedures.

1.3 Prerequisites

1.3.1 Probability Basics

You are expected to know the basics of probability including:

The general mathematical structure of probability: a Probability Space or Probability Triple, (Ω, $\mathcal{F}$, Pr).

1. A sample space, Ω, which is the set of all possible outcomes.
2. A set of events (a σ-algebra) $\mathcal{F}$, where each event is a subset from Ω that contains zero or more outcomes.
3. A function Pr that maps events in $\mathcal{F}$ to numbers between 0 and 1 (inclusive), i.e., probabilities.

For any event E in $\mathcal{F}$, $0 \leqslant \Pr(\mathrm{E}) \leqslant 1$, and $\Pr(\Omega)$=1.

Given an event A in $\mathcal{F}$, the complement of A, denoted A^c is that part of Ω that does not include A: thus $\mathrm{A} \cap \mathrm{A}^c = \emptyset$ and $\mathrm{A} \cup \mathrm{A}^c = \Omega$.

Conditional probability. Given two events A and B in $\mathcal{F}$, where $\Pr(\mathrm{B}) > 0$, the conditional probability of A given B is written $\Pr(\mathrm{A}|\mathrm{B})$ and is defined as $\Pr(\mathrm{A} \cap \mathrm{B}) / \Pr(\mathrm{B})$. An alternative definition: $\Pr(\mathrm{A} \cap \mathrm{B}) = \Pr(\mathrm{B}) \Pr(\mathrm{A}|\mathrm{B})$.)

General probability results and definitions. Given events A and B in $\mathcal{F}$:

- Probability *at least one occurs*

$$\Pr(\mathrm{A} \cup \mathrm{B}) = \Pr(\mathrm{A}) + \Pr(\mathrm{B}) - \Pr(\mathrm{A} \cap \mathrm{B})$$

  If events are *Mutually Exclusive*, then $\Pr(\mathrm{A} \cap \mathrm{B})$ = 0. And the *Addition Rule* applies:
  $\Pr(\mathrm{A} \cup \mathrm{B}) = \Pr(\mathrm{A}) + \Pr(\mathrm{B})$.

- Probability *that both occur*:

$$\Pr(\mathrm{A} \cap \mathrm{B}) = \Pr(\mathrm{A}|\mathrm{B}) \Pr(\mathrm{B}) = \Pr(\mathrm{B}|\mathrm{A}) \Pr(\mathrm{A})$$

  If the events are *Independent*, then $\Pr(\mathrm{A}|\mathrm{B}) = \Pr(\mathrm{A})$. And the *Multiplication Rule* applies:
  $\Pr(\mathrm{A} \cap \mathrm{B}) = \Pr(\mathrm{A}) \Pr(\mathrm{B})$.

  Note: the term *joint probability* is a synonym for the probability that two, or more, events both occur.

  Note: We will typically write $\Pr(\mathrm{AB})$ for $\Pr(\mathrm{A} \cap \mathrm{B})$; $\Pr(\mathrm{A}, \mathrm{B})$ is another expression.

- Law of Total Probability: if events A_1, A_2, ..., A_K are mutually exclusive and exhaustive[1]

$$\Pr(B) = \sum_{i=1}^{K} \Pr(A_i \cap B) = \sum_{i=1}^{K} \Pr(B|A_i) \Pr(A_i)$$

1.3.2 Random variables (rv's)

Random variables, e.g., X, are, usually, real-valued functions defined on sample spaces, i.e., mappings from the outcomes of a random process, the events, to the real number line:

$$X : \Omega \to \mathbb{R}$$

More loosely stated, they are numbers associated with random events.

- The mapping induces a probability distribution for the random variables; e.g., $\Pr(X = x) \equiv \sum_{i:X(\omega_i)=x} \Pr(\omega_i)$, where ω_i are events in Ω such that $X(\omega_i)$=x. More generally $\Pr(X \in A)$ is the probability assigned to the set $A \in \mathcal{F}$, which tells us the probability that $X(\omega)$ falls in the set A.
- Random vectors of length p of real numbers, $\mathbf{X} = (X_1, \ldots, X_p)^T$, are mappings from Ω to $\mathbb{R}^p$.

When the random variables have discrete values, e.g., counts, the induced probability distribution has a *probability mass function*, pmf. When the random variables have continuous values, e.g., lengths, the induced probability distribution has a *probability density function*, pdf. Distributions that are neither discrete nor continuous are either *singular* or *mixtures* of discrete, continuous, and singular distributions. For examples, if values below some detection limit are assigned the value 0, and values at or above the detection limit retain their measured value, it may be modelled as a mixture between a probability mass at zero, and a continuous distribution above the detection limit.

We assume familiarity with the following discrete random variables, their pmfs, expected values, and variances.

- $X \sim \text{Bernoulli}(\theta)$: E[X]=$\theta$, Var[X]=$\theta(1-\theta)$
- $X \sim \text{Binomial}(n, \theta)$: E[X]=$n\theta$, Var[X]=$n\theta(1-\theta)$
- $X \sim \text{Poisson}(\mu)$: E[X]=$\mu$, Var[X]=$\mu$

And the discrete random vector:

- $\mathbf{X} = (X_1, X_2, \ldots, X_k) \sim \text{Multinomial}(n,\theta_1,\theta_2,\ldots,\theta_k)$: E[$X_i$]=$n\theta_i$, Var[$X_i$]=$n\theta_i(1-\theta_i)$, $\text{Cov}(X_i, X_j) = -n\theta_i\theta_j$, for $i \neq j$.

And the following continuous random variables and their pdfs.

[1] Mutually exclusive means $A_i \cap A_j = \emptyset$ and exhaustive means $\cup_{i=1}^{K} A_i = \Omega$.

- $X \sim$ Uniform(α, β): $\mathrm{E}[X]=\frac{\alpha+\beta}{2}$, $\mathrm{Var}[X] = \frac{(\beta-\alpha)^2}{12}$
- $X \sim$ Beta(α, β): $\mathrm{E}[X]=\frac{\alpha}{\alpha+\beta}$, $\mathrm{Var}[X] = \frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\frac{1}{\alpha+\beta+1} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- $X \sim$ Normal(μ, σ^2): $\mathrm{E}[X]=\mu$, $\mathrm{Var}[X]=\sigma^2$
- $X \sim$ Gamma(α, β): $\mathrm{E}[X] = \frac{\alpha}{\beta}$, $\mathrm{Var}[X]=\frac{\alpha}{\beta^2}$; *(the shape-rate parameterisation)*.
- $X \sim$ Exponential(λ): $\mathrm{E}[X]=\frac{1}{\lambda}$, $\mathrm{Var}[X]=\frac{1}{\lambda^2}$. Recall: Exponential($\lambda$) $\equiv$ Gamma(1,λ).

And the continuous random vector:

- $\mathbf{X} \sim$ Multivariate Normal $(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}$ is p by 1 column vector and Σ is a p by p symmetric (positive semi-definite) matrix.

Notes.

- See Appendix A of the notes by King and Ross for other probability distributions that will be used in the course.
- Given a *joint* probability distribution for two or more random variables, say $\mathrm{p}(X, Y)$, we will refer to the distribution for one variable alone as a *marginal* distribution, e.g., $\mathrm{p}(X)$ is the marginal distribution for X.
- We will sometimes refer to the *Coefficient of Variation* or CV of a random variable (when that makes sense), where $\mathrm{CV}(X) = \sqrt{\mathrm{Var}[X]}/\mathrm{E}[X]$ and is often expressed in percent; e.g., $\mathrm{CV}(X)$=15%.

1.3.3 Some Notation

- Population parameters will usually be denoted by Greek letters, with θ a generic representation. Scalar, single parameters will often be represented by lower case letter, e.g.,

  $$\mu,\ \sigma^2,\ \alpha,\ \beta,\ \rho, \gamma$$

  And upper case letter sometimes denoting vectors of parameters, e.g., $\Theta = (\theta_1, \ldots, \theta_\mathrm{p})^\mathsf{T}$.
- Random variables (that are not parameters) will usually be denoted by Roman letters with upper case letters often denoted unrealized, yet-to-be-observed, random variables and lower case letters denoted realised, observed values. For example, $X \sim$ Binomial(10,0.3) and x=6.

A notational distinction will not always be made between a probability density function (pdf) and a probability mass function (pmf) for a random variable; e.g., $\mathrm{p}(Y)$ is the probability density or mass function for a random variable Y.

A generic representation of the pdf/pmf:s of probability distributions (from Gelfand and Smith, 1990) is based on squared brackets, as in $[X]$ represents the distribution for X, $[X|Y]$ represents the conditional probability distribution for X given Y, and $[X, Y]$ represents the joint distribution for X and Y.

To indicate that a rv follows a specific distribution, the notation $\sim$ followed by a distribution name or shorthand will be used:

$$y \sim \text{Normal}\left(\mu, \sigma^2\right) \equiv N(\mu, \sigma^2)$$

Reminders.

- A pdf evaluated at a single value, $p(y)$, where $p(y) > 0$, does not yield the probability of the event y as the probability that a specific value occurs is 0. However the probability that Y falls in some interval $[a, b]$, $Pr[a \leqslant Y \leqslant b]$ is found by integrating the pdf over the interval, $\int_a^b p(y)dy$.
- Another perspective on pdfs: the limit of

  $$\frac{1}{2\epsilon} Pr(y - \epsilon < Y \leqslant y + \epsilon)$$

  as ϵ goes to 0, is the pdf $p(y)$, and is equal to the derivative of the cumulative distribution function (cdf) $Pr(Y \leqslant y)$ at y.
- Let (x, y) be a pair of random variables with with jointly continuous distribution with density $p(x, y)$. Then $p(x|y) = \frac{p(x,y)}{p(y)}$ is the conditional density for x given y, when $p(y) > 0$.

1.4 Real World Meaning of Probability

While the mathematical notion of probability seems clear, there are a variety of real world interpretations or definitions of probability. These interpretations have relevance to Bayesian statistics. Here we just consider two: long-run frequency and subjective.

Long-run frequency interpretation. The long-run frequency interpretation of probability is that the probability of an event A is the fraction of times the event A would occur if the same random process is repeatedly carried out many many times ("infinitely" often). For example, if one rolls a die many, many times, $Pr(2)$ is the fraction of times that the side with 2 pips lands face up. Another example is that if a simple random sample of two people is drawn without replacement from a list of six people, the probability that a given person will be in the sample is 2/6—where such draws could be done repeatedly.

This definition runs into difficulties when a process, for which there is uncertainty about its outcome, cannot be repeated many times. For example, will there be an earthquake above 3.0 Richter here tomorrow? There either will or there won't, but we are uncertain about the outcome, and while we can't repeat "tomorrow", we would like to have quantitative assessment or measure of our belief about whether an earthquake will occur or not.

Subjective probability. The subjective probability of an event is defined to be a number between 0 and 1 that quantifies an individual's belief that the event will occur (or did occur but the outcome is unknown to that individual). Thus subjective probabilities are *personal*[2] and will differ between people. My subjective probability for an earthquake above 3.0 occurring tomorrow is 1/1,000,000. A geoscientist would have a different subjective probability, presumably informed by their professional expertise.

A key component of Bayesian statistical inference is a statement of the probability, or probabilities, about some state of nature, what is called the prior probability. While not the only way to define prior probabilities, the subjective definition or interpretation is generally more applicable than a long-run frequency definition. How one defines prior probabilities is a *hugely important* issue in Bayesian statistics.

1.5 Bayes Theorem and Examples

Bayes Theorem can be viewed as a particular formulation of a conditional pmf, for discrete random outcomes, or a conditional pdf, for continuous random outcomes.

1.5.1 Discrete case

Let X and Y be two discrete valued random variables with possible values denoted $x_1,\ldots,x_n$ and $y_1,\ldots,y_m$, respectively. The conditional probability that $X = x_i$ given $Y = y_j$ is defined by:

$$\Pr(X = x_i|Y = y_j) = \frac{\Pr(X = x_i, Y = y_j)}{\Pr(Y = y_j)}$$

which can be rewritten as

$$\Pr(X = x_i|Y = y_j) = \frac{\Pr(Y = y_j|X = x_i)\Pr(X = x_i)}{\Pr(Y = y_j)}$$

The law of total probability can be used to rewrite the denominator $\Pr(Y = y_j)$:

$$\begin{aligned}\Pr(X = x_i|Y = y_j) &= \frac{\Pr(Y = y_j|X = x_i)\Pr(X = x_i))}{\sum_{k=1}^{n}\Pr(X = x_k, Y = y_j)} \\ &= \frac{\Pr(Y = y_j|X = x_i)\Pr(X = x_i))}{\sum_{k=1}^{n}\Pr(Y = y_j|X = x_k)\Pr(X = x_k)}\end{aligned} \tag{1.2}$$

Equation 1.2 is often the way that Bayes Theorem is defined, for discrete valued random variables. Sometimes it is written without denoting the exact value of X and Y, namely:

$$\Pr(X|Y) = \frac{\Pr(Y|X)\Pr(X))}{\sum_x \Pr(X = x, Y)} = \frac{\Pr(Y|X)\Pr(X))}{\sum_x \Pr(Y|X = x)\Pr(X = x)} \tag{1.3}$$

[2]Assuming that a parameter is a fixed but unknown quantity and then viewing it as "random" is an example of epistemic uncertainty. More on this later.

Bayes Theorem also holds for outcomes that are random events but not random variables, in other words, numerical values are not defined on the events. For example, X could refer to the presence of a disease and X^c is the absence of the disease, and similarly Y could be the result of a positive test result for the disease and Y^c is a negative test result.

Example 1. Karen has two housemates, Jill and Mary, who do all the cooking. Mary cooks about 2/3s of the time while Jill cooks 1/3 of the time. Mary tends to burn the dinner less often than Jill, about 1/6th of the time, while Jill burns the dinner 1/4th of the time. Karen comes home, steps through the door, smells burned food, and yells: "Jill, you must be cooking!". What is the probability that Karen guessed correctly? We use Bayes Theorem to find out the probability that it was Jill. Letting M and J denote Mary and Jill and B=burnt dinner

$$\Pr(\mathrm{J}|\mathrm{B}) = \frac{\Pr(\mathrm{J}\cap\mathrm{B})}{\Pr(\mathrm{B})} = \frac{\Pr(\mathrm{B}|\mathrm{J})\Pr(\mathrm{J})}{\Pr(\mathrm{B}|\mathrm{J})\Pr(\mathrm{J}) + \Pr(\mathrm{B}|\mathrm{M})\Pr(\mathrm{M})} = \frac{\frac{1}{4}\frac{1}{3}}{\frac{1}{4}\frac{1}{3} + \frac{1}{6}\frac{2}{3}} = 0.429$$

Thus, she has more likely guessed wrong.

Note that *prior* to acquiring additional information the probability that Jill is cooking is 1/3. But given additional information, namely the dinner has been burnt, changes our probability for Jill doing the cooking, it is now 0.429.

Example 2. (From Efron 2005, Jr of the Am. Stat. Assoc.) A woman was pregnant and a sonogram showed that she was going to have two twin boys. Her doctor told her that in the general population of all twins, fraternal and identical, the proportion of identical twins is 1/3. Given that she knew she had two boys, she wanted to know what the probability was that *her* boys would be identical twins. Assume that if twins are fraternal the probability of either sex for each twin is 1/2, the probability of two boys is 1/4, two girls is 1/4, and a boy and a girl is 1/2. Note that given one is having identical twins, the probability of them being girls is the same as them being boys, namely 1/2.

$$\begin{aligned}
\Pr(\text{Identical}|\text{Twin Boys}) &= \frac{\Pr(\text{Twin Boys}\cap\text{Identical})}{\Pr(\text{Twin Boys})} \\
&= \frac{\Pr(\text{Twin Boys}|\text{Identical})\Pr(\text{Identical})}{\Pr(\text{Twin Boys}|\text{Identical})\Pr(\text{Identical}) + \Pr(\text{Twin Boys}|\text{Fraternal})\Pr(\text{Fratern} \\
&= \frac{\frac{1}{2}\frac{1}{3}}{\frac{1}{2}\frac{1}{3} + \frac{1}{4}\frac{2}{3}} \\
&= \frac{\frac{1}{6}}{\frac{1}{6} + \frac{1}{6}} = \frac{1}{2}
\end{aligned}$$

Thus a 50% probability that they would be identical.

Note that *prior* to acquiring additional information the probability that the woman will have identical twins is 1/3. But given additional information, namely that she has twin *boys*, changes our probability for identical twins, it is now 1/2.

Comment. Note that in these two examples, the applications of Bayes Theorem were simply exercises in probability, not in statistics, as there are no unknown parameters to estimate nor hypotheses to test. The focus in this course will, of course, be on statistical inference, particularly parameter estimation and hypothesis testing.

1.5.2 Continuous case

Here we consider random variables, thus numerical outcomes, not categorical. Let X and Y be two continuous random variables with corresponding pdf's $f(x)$ and $g(y)$. The continuous version of Bayes Theorem is then expressed as a conditional probability density function:

$$f(X|Y) = \frac{f(X,Y)}{g(Y)} = \frac{g(Y|X)f(X)}{\int g(Y|X)f(X)dX} \tag{1.4}$$

Bayes Theorem can be applied to hybrid cases, where X is discrete and Y is continuous, and one of the examples below is such a case.

1.6 Bayesian Statistical Inference

The most basic setup for Bayesian statistical inference has two components:

1. An *observation model* for observations y, with pdf/pmf $f(y|\theta)$ (or $p(y|\theta)$). This is often represented through a *likelihood* for the parameter, θ, $L(\theta|y)$, which is equal or proportional to the pdf/pmf of the observation model, viewed as a function of θ, for fixed values of y.
2. A *prior probability distribution* for parameters θ, with pdf/pmf $\pi(\theta)$ (or simply $p(\theta)$.

The model building blocks results in a joint distribution for the observations and parameter, with pmf/pdf $p(\theta, y) = p(\theta)p(y|\theta)$, from the definition of conditional probability/density.

Bayesian inference for θ produces a third component, the conditional probability of θ given the data, what is called the *posterior probability distribution* for θ is determined by the pdf/pmf:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} \tag{1.5}$$

where $p(y)$ can be referred to as the *marginal distribution* for y, i.e. $p(y) = \int p(\theta, y)\, d\theta$. Since we are conditioning on a specific y, this is a constant, and $p(y)$ is often referred to as a *normalizing constant*.

Eq'n 1.5 can be rewritten in several ways, using ordinary probability theory:

$$p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int p(\theta, y)d\theta} = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta} \propto f(y|\theta)\pi(\theta) \tag{1.6}$$

where the second ratio is the classical form of "Bayes' theorem" or "Bayes' formula", written for continuous parameter distributions. In general, one might have a combination of continuous and discrete parameter distributions, turning the integrals into combinations of sums and integrals.

Comments.

1. Bayesian inference can be viewed as taking the knowledge that one has about θ before seeing sample data and *updating* or *modifying* that knowledge after seeing data.
2. Given that we are considering probability distributions for θ, prior and posterior, implies that θ is being viewed as a random variable. This does not mean that θ is necessarily a random quantity in nature. It is a random variable in the sense that one is uncertain about the value of the parameter.
3. Bayesian inference may look deceptively simple. Just calculate the conditional probability distribution. But it's not—which is why there are many books and papers written on it. There are two complications that we will focus on that make up a large part of this course.
   (a) Complication 1: Calculation of $p(y)$, an integration over θ, can be extremely difficult, in particular when θ is a multi-dimensional parameter.
   (b) Complication 2: Specification of the prior distribution can be both difficult, particularly for high dimensional θ, and controversial.
4. Sequential updating of one's knowledge about θ given new information, e.g., a new sample, is carried out by applying Bayes formula again, using the posterior from the last iteration as the prior for the current iteration. Letting y_{first} denote the first sample and y_{second} denote the second sample:

$$p(\theta|y_{\text{first}}, y_{\text{second}}) = \frac{f(y_{\text{second}}|\theta, y_{\text{first}})p(\theta, y_{\text{first}})}{\int p(\theta, y_{\text{first}}, y_{\text{second}})d\theta} = \frac{f(y_{\text{second}}|\theta, y_{\text{first}})p(\theta|y_{\text{first}})}{\int f(y_{\text{second}}|\theta, y_{\text{first}})p(\theta|y_{\text{first}})d\theta} \tag{1.7}$$

5. In a few simple settings the posterior distribution can be deduced by just looking at the numerator of Eq'n 1.6 and examining terms that include the parameter θ, the so-called "kernel". The kernel for a well-known pdf or pmf might be evident. Example 3 will demonstrate this.

Example 3. A mobile phone company is releasing a new phone. Based on experience with the previous model, only 85% of the phones still work after two years. The company advertises that 95% of the *new* models will still be working after two years. A consumer advocacy group is skeptical that the new model will be much better than the old model. It plans to randomly sample n=200 purchasers of the new model and after 2 years and find out how many still have working phones. To give the phone company some allowance, the group chooses a prior for the probability of working, θ,

that has a mean value of 0.90 and a CV of 20%. In particular the prior for $\theta \sim$ Beta(1.6, 0.178). After waiting two years, the sample is taken and 172 phones, or 86%, are still working. How reasonable is the phone company's claim?

Assuming independence between the phones and an in-common "survival" probability, the number of phones, y, still working after two years has a Binomial(n=200,θ) distribution. This is the sampling distribution for the data and determines the likelihood for θ.

The posterior distribution for θ is then:

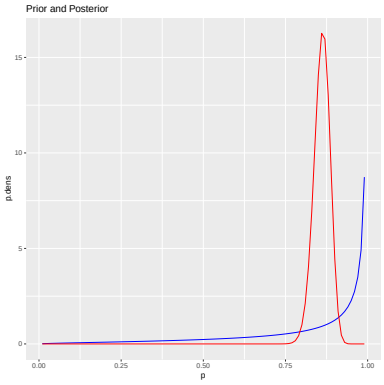
$$p(\theta|y) \propto \pi(\theta) f(y|n,\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\binom{n}{y}\theta^{y}(1-\theta)^{n-y}$$
$$\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{y}(1-\theta)^{n-y} = \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}$$

which is recognized as the kernel of a Beta(α+y, β+n−y), in this case Beta(1.6+172, 0.178+200-172) = Beta(173.6,28.178).

Notes:

- Figure 1.5 shows the prior (in blue) and posterior (in red) densities for the probability a mobile phone is still working after 2 years. Note that the inclusion of the data has shifted the prior distribution noticeably to the left. This is an example of the informal expression that data "dominated" the prior.
- Given the posterior probability distribution it is easy to calculate various summaries such as the mean of the posterior distribution, in this case 173.6/(173.6+28.178)= 0.86.
- Further, one can examine the claim made by the phone company and calculate that the probability that the survival probability θ is 90% or higher is 0.042. Referring to the phone company claim "more accurately", the probability that $\theta \geqslant 0.95$ is 8.450503e-07.

Figure 1.5: Prior (in blue) and posterior (in red) distribution for the probability that mobile phone is still working after 2 years.



Note: the parameters in the prior distribution for θ are called *hyperparameters*.

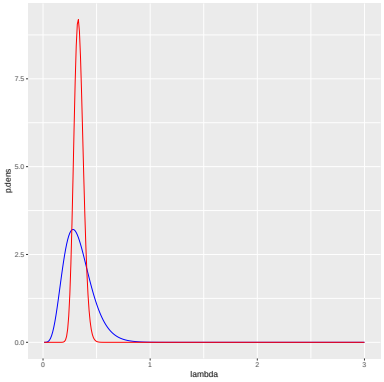
Example 4. After entering a queue at a bank on Fridays between 11:30AM and 1:00PM, the waiting time for seeing a teller follows a exponential distribution with rate parameter λ. Thus the average waiting time is $1/\lambda$. You are considering switching to a new bank and will visit that bank next Friday. In your experience at your current bank you currently wait from 2 to 6 minutes. You select a prior distribution for λ where the expected value for λ is 1/3 minutes with a CV of 40%, in particular the prior for λ is Gamma(6.25, 18.75). You then visited the bank and observed 52 customers enter the queue during the 11:30-1:00PM period. The total amount of waiting time was 156 minutes. The posterior distribution for λ:

$$p(\lambda|y) \propto \pi(\lambda) f(y_1, \ldots, y_n|\lambda) \propto \lambda^{\alpha-1} \exp(-\beta\lambda) \prod_{i=1}^{n} \lambda \exp(-\lambda y_i)$$

$$= \lambda^{\alpha+n-1} \exp(-\lambda(\beta + \sum_{i=1}^{n} y_i))$$

which is the kernel for a Gamma($\alpha + n$, $\beta + \sum_{i=1}^{n} y_i$) distribution. Thus the posterior distribution for λ is Gamma(6.25+52, 18.75+156) = Gamma(58.25, 174.75). See Figure 1.6.

Figure 1.6: Prior (in blue) and posterior (in red) distribution for the exponential waiting time parameter λ.



1.7 Prior and Posterior Predictive Distributions

The *predictive distribution* is another name for the marginal distribution for the data, i.e., the distribution for the data after the parameter has been integrated out.

There are two predictive distributions: the prior predictive and the posterior predictive. Letting y^{new} denote a new, not-yet-observed data point, the *Prior Predictive* distribution is

$$p(y^{new}) = \int_{\theta} f(y^{new}|\theta)\pi(\theta)d\theta \tag{1.8}$$

And, letting y^{old} denote already observed data, e.g., an available sample, and the *Posterior Predictive* distribution is

$$p(y^{new}|y^{old}) = \int_\theta f(y^{new}|\theta, y^{old})p(\theta|y^{old})d\theta \tag{1.9}$$

where $p(\theta|y^{old})$ is the *posterior* probability distribution for θ.

Demonstration. Consider the binomial sampling model with a Beta prior on θ and the prior predictive distribution. Let n be the binomial sample size.

$$\begin{aligned}
p(y^{new}) &= \int_\theta f(y^{new}|\theta)\pi(\theta)d\theta = \int \binom{n}{y^{new}} \theta^{y^{new}}(1-\theta)^{n-y^{new}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\
&= \binom{n}{y^{new}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int \theta^{\alpha+y^{new}-1}(1-\theta)^{\beta+n-y^{new}-1}d\theta \\
&= \binom{n}{y^{new}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y^{new})\Gamma(\beta+n-y^{new})}{\Gamma(\alpha+\beta+n)} \int \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+y^{new})\Gamma(\beta+n-y^{new})}\theta^{\alpha+y^{new}-1}(1-\theta \\
&= \binom{n}{y^{new}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y^{new})\Gamma(\beta+n-y^{new})}{\Gamma(\alpha+\beta+n)}
\end{aligned}$$

This is a Beta-Binomial distribution, which we denote Beta-Binomial(n, α, β).

For the posterior predictive distribution, given that the posterior distribution for θ is still Beta, one can substitute $\alpha^{post} = \alpha + y^{old}$ and $\beta^{post} = \beta + n - y^{old}$ in the above results. Letting n^{new} be the new sample size, the posterior predictive distribution is then:

$$p(y^{new}|y^{old}) = \binom{n^{new}}{y^{new}} \frac{\Gamma(\alpha^{post}+\beta^{post})}{\Gamma(\alpha^{post})\Gamma(\beta^{post})} \frac{\Gamma(\alpha^{post}+y^{new})\Gamma(\beta^{post}+n^{new}-y^{new})}{\Gamma(\alpha^{post}+\beta^{post}+n^{new})}$$

namely, a Beta-Binomial$(n^{new}, \alpha^{post}, \beta^{post})$.

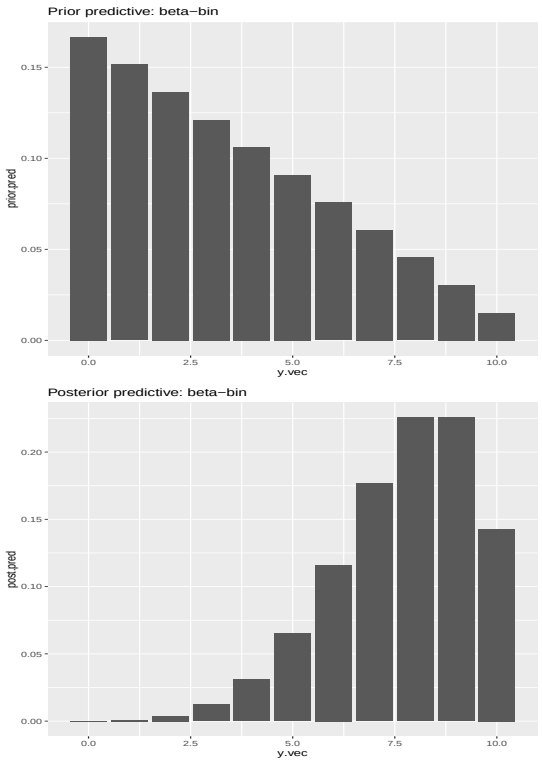
Example 5. Assume a binomial sampling model, Binomial(n, θ), with n=10 and the prior for θ is Beta(1,2). Suppose that y=9 is observed, then the posterior for θ is Beta(10,3). Letting n^{new}=10, Figure 1.7 shows the prior and the posterior predictive distributions for y^{new}. Before observing the data, the predicted probability of y^{new}=9 was 0.0303. After observing the data, the predicted probability of y^{new}=9 is now 0.2256.

Comparing prior and posterior predictive distributions is one way of evaluating the influence of the prior.

1.8 Likelihood principle and Bayesian vs Frequentist inference

Of relevance to comparisons between Bayesian and Frequentist inference is the Likelihood Principle. The Likelihood Principle says that given a sample of data, y, any two sampling models for y, say $f_1(y|\theta)$ and $f_2(y|\theta)$, that have likelihoods that have the

Figure 1.7: (Example 5.) Prior and posterior predictive distributions for y^{new} where $y|\theta \sim$ Binomial(n=10,θ). Thus the prior and posterior predictive distributions are both Beta-Binomial distributions. The prior for θ was Beta(1,2) and y=9 was observed (the posterior for θ was then Beta(10,3)).



same kernel (identical up to a constant of proportionality) yield the same inference for θ. Thus f_1=$c_1 g(\theta)$ and f_2 = $c_2 g(\theta)$, where c_1 and c_2 are constants. The main point is that inference for θ depends on the observed y alone not on unobserved values of y.

One aspect of Frequentist inference that does not obey the Likelihood Principle is the use of p-values in hypothesis tests. Suppose that the sampling model for the data is Poisson(θ) and that there are two hypotheses about θ: a null hypothesis, $H_0 : \theta = 1$, and an alternative hypothesis, $H_1 : \theta = 2$. Suppose that a single observation is made yielding the value y=2. The standard frequentist approach is to calculate the p-value: the probability of the observed value and any values in a direction away from H_0 in the direction of H_1. In this case the p-value is $Pr(Y \geqslant 2|H_0)$ = 1- $Pr(Y = 0|H_0)$ - $Pr(Y = 1|H_0)$ = 1-exp(-1) - exp$(-1) * 1$ = 0.264. Thus, one would not reject H_0.

This procedure is violating the Likelihood Principle, however, in that inference is being based on more than the likelihood of the data. The probability of events that *did not occur* are included in the inference.

A more straightforward approach to testing these two hypotheses would be to compare the likelihoods of θ under the two hypotheses. Under H_0, θ=1 and $Pr(Y = 2|\theta = 1)$ = $L(\theta = 1|Y = 2)$ = exp$(-1) * 1/2!$ = 0.184. Under H_1, θ=2 and $Pr(Y = 2|\theta = 2)$ = $L(\theta = 2|Y = 2)$ = exp$(-2) * 2^2/2!$ = 0.271. Thus, given a value of y=2, H_1 seems

more consistent with the data than does H_0, as can be seen by the likelihood ratio $\mathrm{L}(\theta = 1|Y = 2)/\mathrm{L}(\theta = 2|Y = 2)$ = 0.6796.

Part of the problem with p-values in hypothesis testing is the "inherent" special status of one hypothesis, the Null Hypothesis, over another hypothesis, the Alternative Hypothesis. The p-value is calculated *conditional* on H_0 *being true*.

We will discuss hypothesis testing in a Bayesian framework in detail later, but here we will discuss a Bayesian approach for this particular setting. To do so we will bring the notion of Odds. The odds for an event is the ratio of the probability of the event to the probability of its complement. For example, if a fair die is thrown, the odds of the sides 2, 3, 4, 5, or 6 landing face up is $\Pr(2 \cup 3 \cup 4 \cup 5 \cup 6)/\Pr(1) = (5/6)/(1/6) = 5$.

A Bayesian approach to testing the above hypotheses is to specify prior probabilities for both hypotheses. In this case, assuming no expert knowledge, prior probabilities of 0.5 for H_0 and H_1 seem reasonable and the prior *odds* is the ratio 0.5/0.5 = 1. The posterior probability for H_0 is proportional to the product of the prior and the likelihood, namely 0.5*0.184 = 0.092, and the posterior probability for H_1 is proportional to 0.5*0.271 = 0.1355. Thus the posterior odds of H_0 to H_1 is 0.092/0.1335 = 0.6796, in this case simply the likelihood ratio because the prior probabilities were equal. Thus the Bayesian conclusion would be that there is more evidence for H_1 than H_0.

Bayesian inference is consistent with the Likelihood Principle in that the posterior distribution is proportional to the product of the prior distribution and the likelihood alone, i.e., the effect of the data on the posterior is only through the likelihood.

1.9 Some History

Quoting from Wikipedia: "Bayes' theorem is named after Reverend Thomas Bayes (1701—1761), who first provided an equation that allows new evidence to update beliefs in his *An Essay towards solving a Problem in the Doctrine of Chances* (1763). It was further developed by Pierre-Simon Laplace, who first published the modern formulation in his 1812 "Theorie analytique des probabilites". Sir Harold Jeffreys put Bayes' algorithm and Laplace's formulation on an axiomatic basis. Jeffreys wrote that Bayes' theorem 'is to the theory of probability what the Pythagorean theorem is to geometry'."

Thus Bayes Theorem has been around for over 250 years. However, widespread use of Bayesian statistics is a relatively recent development–over the last 30 years.

While statistics were certainly gathered and used in the 1800s, statistics started becoming a more established and mature discipline in the early 1900s due to contributions by Karl Pearson, R.A. Fisher, Jerzy Neyman, Egan Pearson and others. Contributions from Fisher, which included the development of maximum likelihood theory, and the hypothesis testing framework of Neyman and E. Pearson became the roots of Frequentist Statistical methods.

Bayesian methods, however, did not thrive in the 1900s. One reason was concern

over the influence of priors on inference. There were heated arguments over the merits of frequentist and Bayesian inference in the 60's, 70's, and 80's—but in a practical sense these arguments had little effect because people generally could not carry out Bayesian inference except for a relatively restricted range of problems. The hang-up was calculating the posterior distribution. Specifically, the integration involved in calculating the normalizing constant or the marginal pdf/pmf for the data, $p(\mathbf{y}) = \int_\theta f(\mathbf{y}|\theta)\pi(\theta)d\theta$, was simply too difficult.

All this began to change in the early 1990s with the application of, and further development of Monte Carlo simulation methods, which provided ways of generating samples from the posterior distribution. Note this is in contrast to being able to writing down a closed-form expression for the posterior density, or being able to exactly evaluate the posterior pdf for a given value of θ (or a vector of θ's). Given a large enough sample from the posterior distribution, however, gives one the ability to make all sorts of approximate inferences about the posterior, e.g., the mean from a large enough sample from the posterior distribution can be a "good enough" approximation of the true posterior mean.

Since then many problems that have proven quite hard to solve in a Frequentist framework have become more readily solved in a Bayesian framework.

Controversy over the influence of priors remains and considerable thought and research into formulating and selecting priors continues as well, but the ability to solve problems previously unsolveable has led to a revolution of usage of Bayesian methods.

Among the things that we will discuss in this class are

1. Key "classical" Bayesian statistical principles and methods that have been around for decades;
2. Various perspectives on prior distributions;
3. Some of the Monte Carlo procedures used for making Bayesian inference possible.

Optional, but interesting reading

- Lavine, M. "What is Bayesian statistics and why everything else is wrong." `http://people.math.umass.edu/~lavine/whatisbayes.pdf`
- Efron, B. 2005. "Bayesians, frequentists, and scientists." `https://courses.physics.ucsd.edu/2015/Fall/physics210b/REFERENCES/Efron_Bayesians_Frequentists.pdf`

R code to produce figures

Example 3: Mobile phone failure.

```
#-- Mobile phone failure example
library(ggplot2)
xseq <- seq(0.01,0.99,by=0.01)
prior.a <- 1.6; prior.b <- 0.178
n <- 200; y <- 172
post.a  <- prior.a+y; post.b <- prior.b + n - y
prior.pdf <- dbeta(xseq,prior.a,prior.b)
posterior.pdf <- dbeta(xseq,post.a,post.b)
df <- data.frame(x=xseq,prior=prior.pdf,post=posterior.pdf)
g1 <- ggplot(df,mapping=aes(x,prior)) + geom_line(col="blue") +
  geom_line(mapping=aes(y=post),col="red") + xlab(expression(theta))
g1

cat("Pr(survival)>0.9=",1-pbeta(0.9,post.a,post.b),"\n")
# Pr(survival)>0.9= 0.04176059
```

Example 4: waiting time at bank.

```
 #-- Example 4: Waiting time at bank
xseq <- seq(0.01,3,by=0.01)
prior.a <- 6.25; prior.b <- 18.75
n <- 52; ttl.wait <- 156
post.a <- prior.a+n; post.b <- prior.b + ttl.wait
prior.pdf <- dgamma(xseq,shape=prior.a,rate=prior.b)
posterior.pdf  <- dgamma(xseq,shape=post.a,rate=post.b)
df <- data.frame(x=xseq,prior=prior.pdf,post=posterior.pdf)
g1 <- ggplot(df,mapping=aes(x,prior)) + geom_line(col="blue") +
  geom_line(mapping=aes(y=post),col="red") + xlab(expression(lambda)) +
  ggtitle("Prior and Posterior: Exponential dist parameter")
g1
```

Example: prior and posterior predictive with Beta-Binomial distributions.

```
# Code to evaluate Beta-Binomial pmf
dbetabin <- function(y,size,alpha,beta) {
  bin.coef <- choose(n=size,k=y)
  num <- beta(alpha+y,beta+n-y)
  den <- beta(alpha,beta)
  out <- bin.coef*num/den
  return(out)
}
```

```
library(ggpubr)
n <- 10; y.obs <- 9

#prior and posterior
a.prior <- 1; b.prior  <- 2
a.post <- a.prior+y.obs; b.post <- b.prior+n-y.obs

y.vec <- 0:n
#prior predictive
prior.pred <- dbetabin(y=y.vec, size=n, alpha=a.prior, beta=b.prior)
#posterior predictive
post.pred <- dbetabin(y=y.vec,size=n,alpha=a.post,beta=b.post)

df <- data.frame(y.vec=y.vec,prior.pred=prior.pred,post.pred=post.pred)

g.prior <- ggplot(df,aes(x=y.vec,y=prior.pred)) + geom_col() +
             labs(title="Prior predictive: beta-bin")
g.post <- ggplot(df,aes(x=y.vec,y=post.pred)) + geom_col() +
             labs(title="Posterior predictive: beta-bin")
ggarrange(g.prior,g.post)
```

Supplement to L1: Bayesian Theory

Posterior Predictive Dist'n for Exponential

Here the sampling distribution is Exponential(λ) and a Gamma(α, β) prior is used for λ. As shown in example 4 in L1, given n iid random Exponential(λ) variables, the posterior for λ is Gamma($\alpha + n$, $\beta + \sum_{i=1}^{n} y_i$).

The general equation for the posterior predictive density for y is

$$p(y^{new}|y^{old}) = \int f(y^{new}|\lambda)p(\lambda|y^{old})d\lambda$$

To reduce notation substitute y for y^{new}, drop the conditioning notation of y^{old}, and let $a = \alpha + n$ and $b = \beta + \sum_{i=1}^{n} y_i$. The posterior predictive density in this case:

$$p(y) = \int_0^\infty \lambda e^{-\lambda y} \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} d\lambda = \frac{b^a}{\Gamma(a)} \int_0^\infty \lambda^{a+1-1} e^{-\lambda(b+y)} \tag{1.10}$$

$$= \frac{b^a}{\Gamma(a)} \frac{\Gamma(a+1)}{(b+y)^{a+1}} = \frac{b^a a}{(b+y)^{a+1}} \tag{1.11}$$

To show that this is a pdf, i.e., this function integrates to 1:

$$p(y) = \int_0^\infty \frac{b^a a}{(b+y)^{a+1}} dy = b^a a \times \left(-\frac{1}{a(b+y)^a} \Big|_0^\infty \right) = b^a a \times \left(0 + \frac{1}{ab^a} \right) = 1$$

Referring to Example 4, the posterior distribution for λ was Gamma(58.25, 174.75). The probability of waiting between 1 and 2 minutes is then:

$$\Pr(1 \leqslant Y \leqslant 2) = \int_1^2 \frac{b^a a}{(b+y)^{a+1}} dy = b^a a \times \left(-\frac{1}{a(b+y)^a} \Big|_1^2 \right)$$

$$= 174.75^{58.25} 58.25 \times \left(-\frac{1}{58.25(174.75+2)^5 8.25} + \frac{1}{58.25(174.75+1)^5 8.25} \right) = 0.2018$$

Aside: ignoring uncertainty in λ, the expected value for λ, a/b = 0.3333333, can be used directly in an exponential distribution:

$$\Pr(1 \leqslant Y \leqslant 2) \approx \int_1^2 0.333 e^{-0.333*y} dy = F(2) - F(1)$$

$$= (1 - e^{0.333*2}) - (1 - e^{0.333*1}) = e^{0.333*1} - e^{0.333*2} = 0.2031142$$

R code

```
#-- posterior predictive for exponential
post.pred.exp.gamma <- function(a,b,LB,UB) {
 out <- (b^a)*a * ( -1/(a*(b+UB)^a) + 1/(a*(b+LB)^a))
  return(out)
```

```
}
a <- 58.25
b <- 174.75
# Pr(1 <= Y <= 2):

post.pred.exp.gamma(a=a,b=b,LB=1,UB=2)
# [1] 0.2018478
```

2 Bayesian Theory—Priors, Part 1

Other Reading

1. (Optional) Read Chapter 3 of *Applied Bayesian Statistics With R and OpenBUGS Examples* by Cowles. Section 3.6 focuses on R, which is not strictly required at this point—but will be used for demonstration frequently. It is a good exercise to simply type the code from this section into R to reproduce the results—you can learn a lot that way.
2. (Optional) Read Chapter 2 of *Bayesian Statistical Methods* by Reich and Ghosh.

Optional Homework. Please see Section 2.5.

Notation for prior, likelihood, and posterior

We will sometimes use the following notation (following Reich and Ghosh 2019):

- *Prior distribution pdf/pmf* for θ: $\pi(\theta)$
- *Conditional observation distribution pdf/pmf for* $\mathbf{Y}$ *given* θ: $\mathrm{f}(\mathbf{Y}|\theta)$
- *Marginal distribution pdf/pmf* for observations $\mathbf{Y}$: $\mathrm{m}(\mathbf{Y})$
  Note: The term *likelihood* is strictly reserved for $\mathrm{f}(Y|\theta)$ seen as a function of θ, and will sometimes be denoted $\mathrm{L}(\theta|Y)$.
- *Posterior distribution pdf/pmf* for θ: $\mathrm{p}(\theta|\mathbf{Y})$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)$ denotes a data vector.

Then Bayes Theorem can be written as follows:

$$\mathrm{p}(\theta|\mathbf{y}) = \frac{\pi(\theta)\mathrm{f}(\mathbf{Y}|\theta)}{\mathrm{m}(\mathbf{Y})} \propto \pi(\theta)\mathrm{f}(\mathbf{Y}|\theta) \tag{2.1}$$

Often, we will use $\mathrm{p}(\cdot)$ for each and every pdf/pmf, emphasizing the general probability framework that applies to *all* random variables. The quantities Y and θ are merely special cases, so using separate notations can obscure the generality. Also, in more complex models the distinction between observations, parameters, and other variables isn't always clear.

Comments.

- In some literature, the term likelihood is used interchangeably with the data observation distribution (data sampling distribution), but that should be avoided. The likelihood exclusively refers to the *effect* of the data model on the posterior distribution.
- The likelihood function is sometimes written as $\mathrm{L}(\theta|\mathbf{Y})$.
- Alternative terminology for the Marginal distribution is the *Normalising Constant*, as it is this factor that ensures that the Posterior distribution integrates to 1.
- Crucially, the likelihood is *not* a density with respect to θ, and is usually also allowed to be unnormalised with respect to Y, so that the general relationship only holds up to some constant proportionality factor: $\mathrm{L}(\theta|\mathbf{Y}) \propto \mathrm{p}(\mathbf{Y}|\theta)$

2.1 Overview

Prior distributions or simply priors, $\pi(\theta)$, along with the sampling distribution for the data, $\mathrm{f}(\mathbf{y}|\theta)$, are the two fundamental components of Bayesian inference that yield the posterior distribution, $\mathrm{p}(\theta|\mathbf{y})$.

Priors, however, have been the focus of considerable discussion and debate as they are a choice made by the data analyst. As Gelman et al. (2017) say: "A key sticking point of Bayesian analysis is the choice of prior distribution ...". In short, that choice can have a considerable effect on the resulting posterior.

Thus a central question: "How does one choose priors?"

2.1.1 What does the prior mean?

The prior is a probability distribution for an unobserved quantity, in this case a parameter θ. It reflects our knowledge about θ prior to observing data.

- Remember that in the setting we are considering, the parameter itself is not random, it is a fixed quantity, but it is unknown, i.e., we are uncertain as to its value.

  Another way to say this: there is *epistemic* uncertainty about the parameter, θ, which is in contrast to "truly" random uncertainty, also called *aleatoric* uncertainty[1].
- The prior distribution is meant to be a measure of one's uncertainty about the parameter value.
- In a pure Bayesian sense, the prior distribution for a parameter reflects one's personal beliefs, they are *subjective* probabilities about the possible value of the parameter (see Lecture Note 1). Thus different people typically have different priors[2].
- Often, however, Bayesian inference is carried out because it is potentially easier than classical or frequentist inference, or lends itself to more easily understood explanations, or less convoluted explanations, e.g., confidence intervals. In this case one simply wants a prior that allows Bayesian inference without "unduly" influencing the posterior, namely, one such that the data dominate the posterior[3].

  Reich & Ghosh (p 59) go so far as to say "In the absence of prior information, selecting the prior can be viewed as a nuisance to be avoided if possible".
- So at one extreme a prior represents very specific beliefs and at the other extreme it is simply a necessary component of Bayesian inference.

2.1.2 Subjective vs Objective priors

One categorisation of priors is that of subjective priors versus objective priors. The line between the two is not necessarily sharply defined.

Subjective priors are the personal priors that have been selected by an individual. Two different heart surgeons, for example, may have individual and different prior probability distributions for the average number of hours required for a particular type of heart by-pass surgery. For example, one surgeon thinks the average is around 4 hours (with a range of 3.5 to 4.5 hours), while another surgeon thinks the average is around 4.5 hours (with a range of 4.3 to 4.7 hours, thus a narrower range: more certainty).

Objective priors are priors chosen according to some "rule" or procedure whereby two individuals carrying out the same rule will end up with the same prior distribution. Some examples of objective priors are Jeffreys prior, reference priors, maximum entropy priors, empirical Bayes priors, and penalised complexity priors (see Section 2.3 of Reich and Ghosh). We will later discuss Jeffreys priors.

[1] From *Aleatoric and Epistemic Uncertainty in Machine Learning*: "Aleatoric (aka statistical) uncertainty refers to the notion of randomness, that is, the variability in the outcome of an experiment which is due to inherently random effects. Epistemic (aka systematic) uncertainty refers to uncertainty caused by a lack of knowledge, i.e., to the epistemic state of the agent." https://www.gdsd.statistik.uni-muenchen.de

[2] See page 2 of Gelman, et al 2017 where they refer to one extreme of choosing priors as a "maximalist" position where priors are chosen completely independently of the sample data, one knows nothing about the likelihood.

[3] Gelman, et al 2017 refer to this as a "minimalist" perspective on priors.

2.1.3 "Non-informative" priors

Another categorisation of priors is so-called "non-informative" priors. Quotation marks are used here because the label is extremely imprecise and nearly all priors are informative in the sense that the prior has *some* influence on the posterior distribution. In particular, the interpretation of "non-informativeness" depends strongly on the model parameterisation; what is deemed to be non-informative of a parameter θ can be strongly informative about e.g. $\exp(\theta)$.

With these caveats in mind: in the absence of expert knowledge, or in situations where one does not want to have the results depend too heavily upon the prior, one would like to use what is sometimes labelled a *non-informative* prior. What is loosely meant is that the posterior will not be *too* influenced by the prior and is *dominated* by the likelihood.

Here we consider two examples.

- The set of possible parameter values is finite: $\theta_1, \ldots, \theta_K$. A uniform prior distribution gives equal prior probability to each possible value, namely $\pi(\theta_i)$=1/K, $i = 1, \ldots, K$.
- The parameter θ is continuous but with bounded support, e.g., (0,20), then a Uniform(0,20) prior would be considered non-informative.

A problem with treating all uniform prior distributions as non-informative will be apparent in Section 2.1.6 if one were to apply a "uniform" density for e.g. a standard deviation parameter σ, $\sigma \sim$ Uniform$(0, c)$ for some very large c. This might appear non-informative, as it has "the same effect" for all σ values below c, but the interpretation in the limit as $c \to \infty$ is that $\Pr(\sigma > a) = 1$ for every value a, which from a philosophical point of view is *highly informative*, in that it strongly favours large values of σ.

A synonym for non-informative is "uninformative". More precise terms for priors that attempt to be "permissive" in what parameter values are deemed likely include "vague" or "diffuse" priors. These may also be parameterisation dependent, but avoid some of the problems with the term non-informative. [4].

2.1.4 Terminology: hyperparameters

The (fixed) parameters of the prior distribution that are specified by the individual are called *hyperparameters*.[5]

For example, if the sampling distribution is $y \sim$ Binomial(n, ϕ) and the prior distribution is $\phi \sim$ Beta$(2, 7)$, then 2 and 7 are the hyperparameters.

Another example, if the sampling distribution is $y \sim$ Poisson(λ) and the prior distribution is $\lambda \sim$ Gamma(20,5), then 20 and 5 are hyperparameters.

2.1.5 Hierarchical models

For more complex models, e.g., random effects models, the hyperparameters can appear at different locations in the overall model structure.

[4] Some statisticians dislike all of these labels due to the lack of precise meaning.

[5] In some software, the parameters of the parameter prior distributions may themselves be random, with hyper-hyperparameters, but in classic terminology, the term hyperparameter is reserved for the *fixed* "leaf node" parameters of a hierarchical model, and all other variables are either observations, parameters, or *latent variables*, which is a term used for unobserved variables. Sometimes these are also called parameters, bringing the terminology back to the basic "observation, parameter, hyperparameter" structure.

For example, each spring, young salmon are released from a hatchery and the number of fish released in year t that survive one year is Binomial(n_t, ϕ_t) where n_t is the number released in year t and ϕ_t is the probability of surviving in year t. The survival probability ϕ_t varies between years is itself a random variable. A Beta distribution is to model the between year ("environmental") variation. Thus:

$$y_t|\phi_t \sim \text{Binomial}(n_t, \phi_t)$$
$$\phi_t|\alpha, \beta \sim \text{Beta}(\alpha, \beta)$$

The unknown parameters are now α and β and prior distributions must be specified for them.

Suppose that Gamma distributions are used as the priors for both:

$$\alpha \sim \text{Gamma}(a_1, b_1)$$
$$\beta \sim \text{Gamma}(a_2, b_2)$$

In this case the hyperparameters are a_1, b_1, a_2, and b_2.

This is a setting where there is both *aleatoric* uncertainty and *epistemic* uncertainty:

- Aleatoric uncertainty about both the number of surviving salmon, y, and the year-specific survival probabilities, ϕ_t.
- Epistemic uncertainty about the fixed but unknown parameters α and β.

2.1.6 Terminology: improper priors

When the support of a parameter is finite, and one would like to use a non-informative prior, then the notion of using a flat or uniform prior over the support is attractive. This is in effect saying that $\pi(\theta) \propto c$, where c is some constant. For example, if θ can take on values from -10 to 10, then $\pi(\theta) = 1/20$.

In the case of infinite support, one cannot define a constant pdf (or pmf) as $\int_{-\infty}^{\infty} c \, d\theta$ is not integrable (the integral is not finite). Of course, what one can do in practice is choose lower and upper bounds that are "extreme" enough that one is fairly certain that θ will not be outside these bounds and then use a Uniform prior; e.g., bounds = $[-c, c]$, then $\theta \sim \text{Uniform}(-c, c)$ and $\pi(\theta) = 1/2c$.

However, sometimes the specified priors are not proper probability distributions in that the prior pdf/pmf does not integrate to 1. For example, suppose the sampling model for data y is Normal(μ, 4^2). A "non-informative" prior for μ is to say that all values are equally probable, in other words the prior for μ is uniform; the pdf is "flat" over the real number line, say $\pi(\mu) \propto 1$. However, such a "prior" is not a probability distribution, as it does not integrate to one. It is what is called an *improper prior*. This improper prior can be viewed the result of taking a Uniform($-c, c$) distribution and letting c go to ∞.

Improper priors can be acceptable as long as the resulting posterior is *not* improper.

Referring to the above normal distribution example, n=1 and y_1=3, then the posterior for μ:

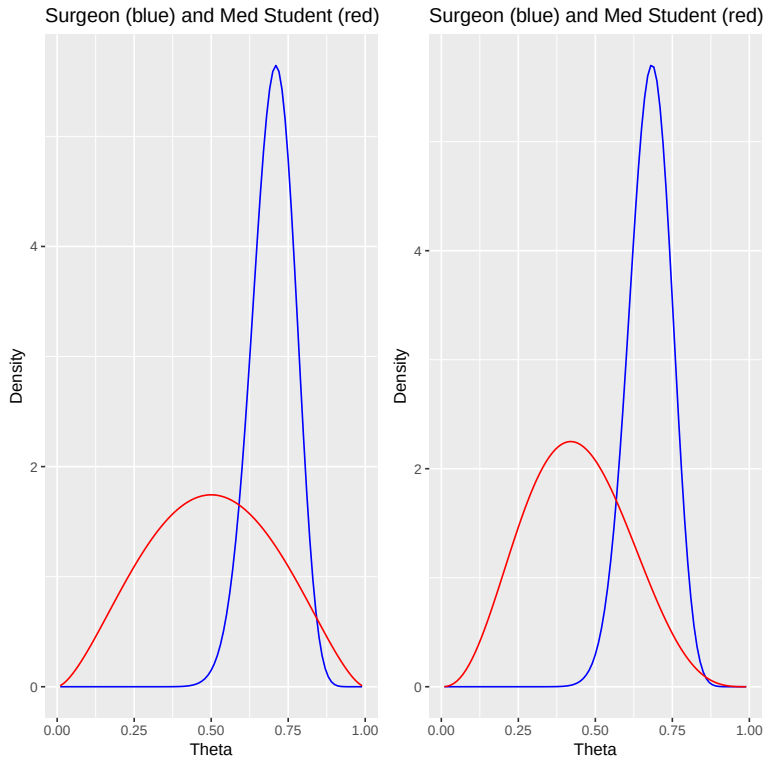
$$p(\mu|y = 3) = \frac{\frac{1}{\sqrt{2\pi 4^2}} e^{-\frac{(3-\mu)^2}{2*4^2}} \times 1}{\int \frac{1}{\sqrt{2\pi 4^2}} e^{-\frac{(3-\mu)^2}{2*4^2}} \times 1 \, d\mu} = \frac{1}{\sqrt{2\pi 4^2}} e^{-\frac{(\mu-3)^2}{2*4^2}}$$

or $\mu|y \sim$ Normal(3, 4^2) which integrates to 1 and is thus a *proper posterior*. Apart from a lack of clear interpretability, there is nothing inherently wrong with using improper priors but one must check that the posteriors are proper.

2.1.7 Competing priors and their effect

Different individuals with different degrees of knowledge and experience of the underlying random process (the data distribution) are likely to have different priors. For example, an experienced heart surgeon who is assessing the probability of survival for a month following coronary by-pass surgery for 60 year old females, θ, may have a prior that is more concentrated over a range of values, e.g., most of the probability is between 60% and 80%, than a first year medical student, e.g., a prior with most of the probability spread relatively evenly between 10% and 90% (see left panel of Figure 2.1). The fundamental, and sometime controversial, issue is how much effect the choice of the prior has on the posterior. For example, after observing the survival or not of n=3 60 year old females who had by-pass surgery, where 1 survived, the posterior distribution for the experienced heart surgeon and the first year medical student might be much different. For example, suppose the surgeon's prior for θ was Beta(29.3, 12.6) and the medical student's was Beta(2.63, 2.63). The differences between their priors are shown in the left panel of Figure 2.1. After observing n=3 with y=1 surviving, thus 33% survival, the posterior distributions are Beta(30.3, 14.6) for the surgeon and Beta(3.63, 4.63) for the medical student. The right panel of the figure shows that the medical student's posterior was more influenced by the data than the surgeon. Whose posterior distribution are you more likely to believe? And why?

Figure 2.1: (Hypothetical) Prior (left panel) and posterior (right panel) distributions for the probability of by-pass surgery survival for experienced heart surgeon and first year medical student. The posterior is based on a sample of n=3 with y=1 survivors, thus the observed survival fraction is 1/3 = 0.33.



2.1.8 Mathematical perspectives

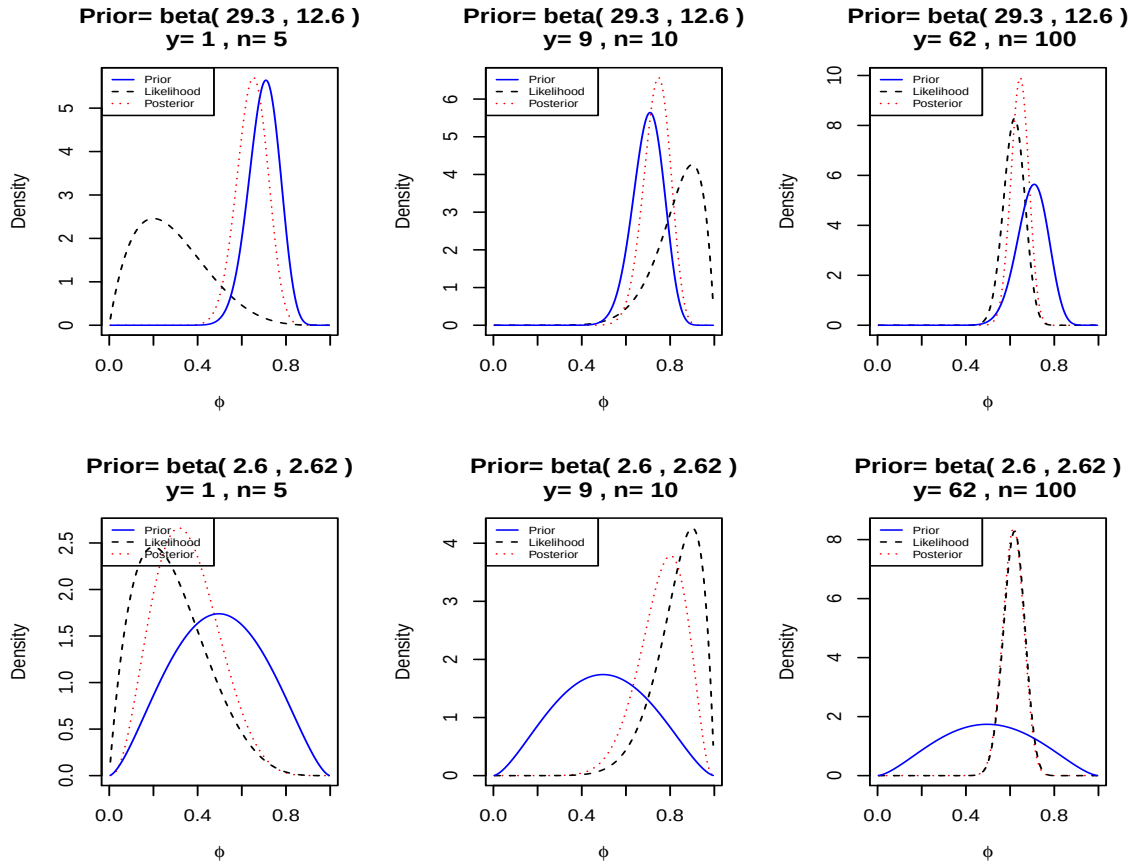
Mathematically it can be seen that the effect of the prior on the posterior depends upon the prior itself and the data distribution (likelihood):

$$p(\theta|y) \propto \pi(\theta) f(y|\theta) = \pi(\theta) f(y|\theta)$$

The influence of the data distribution is, in many cases, a function of sample size—typically, as n increases, the influence of the likelihood on the posterior increases.

- *Graphical display.* In situations where there is just one parameter or relatively few parameters, visual examination of the prior and posterior distributions is a practical, and feasible, procedure for seeing the relative influence. Figure 2.2 shows the effects of different priors. As sample size n increases, the likelihood begins to *dominate* the prior in the sense that the posterior is very similar to the likelihood.

Figure 2.2: (Hypothetical) Changes in Prior (blue solid line) and Posterior (red dotted line) distributions for the probability of by-pass surgery survival for an experienced heart surgeon and first year medical student as sample size increases.



- *Implied sample size of a prior.* In some settings one can quantify the relative contribution of the prior and the likelihood to specific features of the posterior like the posterior mean.

  For a binomial parameter θ with Beta(α, β) prior and a sample size n with y successes, the posterior mean for θ can be written:

$$E[\theta|y] = \frac{\alpha + y}{\alpha + \beta + n} = \frac{\alpha + \beta}{\alpha + \beta + n} E[\theta] + \frac{n}{\alpha + \beta + n} \overline{y}$$

  Thus the prior mean is contributing a relative weight of $\alpha + \beta$ to the posterior mean. The sum $\alpha + \beta$ can be viewed as a sample size. Thus if $\alpha + \beta$ is large relative to n, the prior can dominate the posterior. For example, suppose n=10. If $\alpha + \beta$=2, the prior is like a sample of size 2, while

if $\alpha + \beta$=200, then the prior has a much larger contribution to the posterior mean. Or another view, if $\alpha + \beta$=2, the relative contribution of the prior to the posterior mean is 2/12 = 0.17, while if $\alpha + \beta$=200, it's 200/210=0.95.

2.1.9 Priors for multiple parameters

Given how difficult it can be to specify a prior for one parameter, one can imagine the difficulty with multiple parameters, as one needs to specify a joint prior distribution. The simplest setting is to specify independent marginal priors for each parameter, and then define the joint prior as the product of the pmf:s/pdf:s. For example with two parameters:

$$\pi(\theta_1, \theta_2) = \pi(\theta_1)\pi(\theta_2)$$

Another approach is to construct the joint prior is to begin with a marginal prior for one parameter, then define a conditional prior for a second parameter given the first, then another conditional prior for the third parameter given the first and second priors, and so on. For example,

$$\pi(\theta_1, \theta_2, \theta_3) = \pi(\theta_1)\pi(\theta_2|\theta_1)\pi(\theta_3|\theta_1, \theta_2)$$

With many parameters, this can be difficult and even with a few parameters, say 3 or 4, determining which parameter to start with, and the subsequent conditional priors can be challenging. Thus, the use of independent priors is common.

Note: independent priors do *not* imply independent posteriors.

2.2 Conjugate Priors

As we saw in Lecture 1 (and the above heart by-pass surgery example), there are situations where the posterior distribution is the same as the prior distribution (other than updated hyperparameters in the posterior). Such priors are called *Conjugate Priors*.

Before computer intensive algorithms for computing posterior distributions, like Markov chain Monte Carlo, became available, the use of Conjugate Priors was more a technical convenience than a desired reflection of parameter uncertainty. While much more complicated posteriors can now be calculated, conjugate priors are worth knowing about and can be useful.

Note: conjugate distributions typically are subjective distributions in the sense that an individual chooses the hyperparameters for the prior distribution.

2.2.1 Examining the likelihood

One way of determining whether or not there is a probabiity distribution that could serve as a conjugate prior is to examine the kernel of the likelihood, to see if one can identify a known probability distribution for θ. Some examples:

$$\text{Binomial pmf for } y \propto \theta^y(1-\theta)^{n-y} \sim \theta^{\alpha-1}(1-\theta)^{\beta-1}: \text{ Beta pdf for } \theta$$
$$\text{Poisson pmf for } y \propto \exp(-\theta)\theta^y \sim \exp(-\beta\theta)\theta^{\alpha-1}: \text{ Gamma pdf for } \theta$$
$$\text{Exponential pdf for } y \propto \theta\exp(-\theta y) \sim \exp(-\beta\theta)\theta^{\alpha-1}: \text{ Gamma pdf for } \theta$$

Reich and Ghosh (2019, Section 2.1.5) discuss the following general procedure for finding what they call "natural" conjugate priors.

- For a given data distribution $f(Y|\theta)$, imagine m distinct and fixed values denoted $y_1^0, y_2^0, \ldots, y_m^0$. They call these *pseudo-observations*.
- Postulate a prior distribution for θ with these m pseudo-observations as hyperparameters and that the prior will be proportional to the data distribution (likelihood):

$$\pi(\theta|y_1^0, y_2^0, \ldots, y_m^0, m) \propto \prod_{i=1}^{m} f(y_i^0|\theta) \tag{2.2}$$

- Given a sample of n independent and real (observed) values from $f(\mathbf{Y}|\theta)$, the posterior is then:

$$\begin{aligned} p(\theta|\mathbf{Y}) &\propto \pi(\theta|y_1^0, y_2^0, \ldots, y_m^0, m) \prod_{i=1}^{n} f(y_i|\theta) \\ &= \prod_{j=1}^{m+n} f(y_j^*|\theta) \end{aligned}$$

  where for j=1,...,m, y_j^*=y_i^0, $i = 1, \ldots, m$ and for j=$m+1$,...,$m+n$, y_j^*=y_i, $i = 1, \ldots, n$.
- The posterior is the same form as the prior and thus the prior is conjugate

As an example consider the Poisson(θ) data distribution. The prior is constructed as follows.

$$\pi(\theta|y_1^0, \ldots, y_m^0, m) \propto \prod_{i=1}^{m} e^{-\theta}\theta^{y_i^0} = e^{-m\theta}\theta^{\sum_{i=1}^{m} y_i^0}$$

which is seen to be the kernel of a Gamma(s_0+1,m), where $s_0 = \sum_{i=1}^{m} y_i^0$.

Using the "natural conjugate" construction, the following conjugate priors for data distributions with a single unknown parameter result.

Sampling Dist'n $f(y\|\theta)$	Parameter	Prior Dist'n $\pi(\theta)$
Bernoulli(θ)	θ	Beta(α, β)
Poisson(θ)	θ	Gamma(α, β)
Exponental(θ)	θ	Gamma(α, β)
Normal(μ,σ_o^2)	μ (σ_o^2 known)	Normal(μ_h, σ_h^2)
Normal(μ_o,σ^2)	σ^2 (μ_o known)	Inverse Gamma(α, β)
Normal(μ_o,$\frac{1}{\tau}$)	τ (μ_o known)	Gamma(α, β)

In the last example, $\tau = 1/\sigma^2$ is the *precision* of a Normal distribution.

Note: key to this approach is that the kernel in Eq'n 2.2 is either something recognizable from a "known" distribution or, if not, it is something integrable that can easily be normalized to produce a proper probability distribution, i.e., it sums or integrates to one[6].

[6] It might be interesting to note that any non-negative integrable function $h(\theta)$ can be made a pdf. Given $h(\theta)$ where $h(\theta) \geqslant 0$ for all θ in "some" domain, where

$$\int h(\theta)d\theta = K < \infty$$

then

$$p(\theta) = \frac{1}{K}h(\theta)$$

is a pdf.

2.2.2 Technical Aside: Exponential family distributions

Definition A probability distribution which has a single parameter is in the exponential family if the pdf/pmf can be written as follows:

$$f(y \mid \theta) = h(y)g(\theta)\exp\left(\eta(\theta) \cdot T(y)\right) \tag{2.3}$$

This includes a large number of distributions, e.g., binomial, Poisson, exponential, Normal with known variance, Normal with known mean. For example, the binomial distribution can be written:

$$f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = \binom{n}{y}(1-\theta)^n \exp\left(\ln\left(\frac{\theta}{1-\theta}\right)y\right)$$

where

$$h(y) = \binom{n}{y}; \quad g(\theta) = (1-\theta)^n; \quad \eta(\theta) = \ln\left(\frac{\theta}{1-\theta}\right); \quad T(y) = y$$

Note that $\eta(\theta)$ is referred to as the *canonical parameterisation*.

For multiple parameters:

$$f(y \mid \boldsymbol{\theta}) = h(y)g(\boldsymbol{\theta})\exp\left(\boldsymbol{\eta}(\boldsymbol{\theta}) \cdot \mathbf{T}(y)\right) \tag{2.4}$$

One feature of exponential family distributions is that their support cannot depend upon the parameter θ. This rules out the Uniform(0,θ) distribution for example. One class of modelling techniques, generalized linear models (GLMs), is based on exponential family distributions. Another feature is that when expressed in the above form, the sufficient statistic, $T(y)$, is readily identified by the Factorization Theorem. There are many other important features. Here we bring these up primarily for one reason:

Conjugate priors. Exponential family distributions have conjugate priors.

$$\pi(\boldsymbol{\eta} \mid \boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu)g(\boldsymbol{\eta})^\nu \exp\left(\boldsymbol{\eta}^\top \boldsymbol{\chi}\right), \qquad \boldsymbol{\chi} \in \mathbb{R}^s \tag{2.5}$$

where s is the dimension of $\boldsymbol{\eta}$, and $\nu > 0$ and $\boldsymbol{\chi}$ are hyperparameters.

2.3 Mixture Priors

For the case of a single parameter θ, one can imagine situations where no single conjugate prior adequately describes one's prior knowledge. Mixture priors are one way of "hedging one's bets" so to speak. The simplest mixture prior is a mixture of two probability distributions (note: these need not be conjugate):

$$\pi(\theta) = q\pi_1(\theta) + (1-q)\pi_2(\theta)$$

where $0 \leqslant q \leqslant 1$.

More generally, a mixture of K distributions:

$$\pi(\theta) = \sum_{k=1}^{K} q_k \pi_k(\theta)$$

where $0 \leqslant q_k \leqslant 1$ and $\sum_{k=1}^{K} q_k = 1$.

The posterior will also be a mixture of what would be the individual posteriors, but the weightings will change, from q_k to Q_k.

$$p(\theta|\mathbf{Y}) = \sum_{k=1}^{K} Q_k p_k(\theta|\mathbf{Y})$$

The weightings, Q_k, will involve the marginal distributions from the individual priors and the marginal distribution across all priors.

For example, consider the case of a mixture of two distributions. Let

$$m_i(\mathbf{Y}) = \int \pi_i(\theta) f(\mathbf{Y}|\theta) d\theta, \ i = 1, 2$$

and note that the marginal for the mixture is as follows:

$$m(\mathbf{Y}) = \int \pi(\theta) f(\mathbf{Y}|\theta) d\theta = \int \left[q_1\pi_1(\theta) + q_2\pi_2(\theta)\right] f(\mathbf{Y}|\theta) d\theta$$

The posterior distribution:

$$\begin{aligned} p(\theta|\mathbf{Y}) &= \frac{p(\theta, \mathbf{Y})}{m(\mathbf{Y})} = \frac{\pi(\theta) f(\mathbf{Y}|\theta)}{m(\mathbf{Y})} = \frac{\left[q_1\pi_1(\theta) + q_2\pi_2(\theta)\right] f(\mathbf{Y}|\theta)}{m(\mathbf{Y})} \\ &= \frac{q_1\pi_1(\theta) f(\mathbf{Y}|\theta)}{m(\mathbf{Y})} + \frac{q_2\pi_2(\theta) f(\mathbf{Y}|\theta)}{m(\mathbf{Y})} \\ &= \frac{q_1 m_1(\mathbf{Y})}{m(\mathbf{Y})} \frac{\pi_1(\theta) f(\mathbf{Y}|\theta)}{m_1(\mathbf{Y})} + \frac{q_2 m_2(\mathbf{Y})}{m(\mathbf{Y})} \frac{\pi_2(\theta) f(\mathbf{Y}|\theta)}{m_2(\mathbf{Y})} \\ &= Q_1 p_1(\theta|\mathbf{Y}) + Q_2 p_2(\theta|\mathbf{Y}) \end{aligned}$$

where $Q_i = \frac{q_i m_i(\mathbf{Y})}{m(\mathbf{Y})}$.

Example. Returning to the heart surgeon and the medical student example, suppose that you thought that both of them had knowledge and opinions that were credible. Perhaps the heart surgeon was overly optimistic and perhaps the medical student had had training in relevant fields that the surgeon knew nothing about. You decide to give 60% weight (q_1=0.6) to the surgeon, and 40% (q_2=0.4) to the student. Then the resulting mixture prior is

$$\pi(\theta) = 0.6 * \text{Beta}(29.3, 12.557) + 0.4 * \text{Beta}(2.625, 2.625)$$

The denominator of the posterior:

$$\begin{aligned} \int \left[q_1\pi_1(\theta) + q_2\pi_2(\theta)\right] f(y|\theta) d\theta &= \binom{n}{y} \left[q_1 \int \frac{1}{B(a_1, b_1)} \theta^{a_1+y-1}(1-\theta)^{b_1+n-y-1} d\theta + \right. \\ &\quad \left. q_2 \int \frac{1}{B(a_2, b_2)} \theta^{a_2+y-1}(1-\theta)^{b2+n-y-1} d\theta \right] \\ &= \binom{n}{y} \left[q_1 \frac{B(a_1+y, b_1+n-y)}{B(a_1, b_1)} + q_2 \frac{B(a_2+y, b_2+n-y)}{B(a_2, b_2)} \right] \end{aligned}$$

where

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Then the posterior:

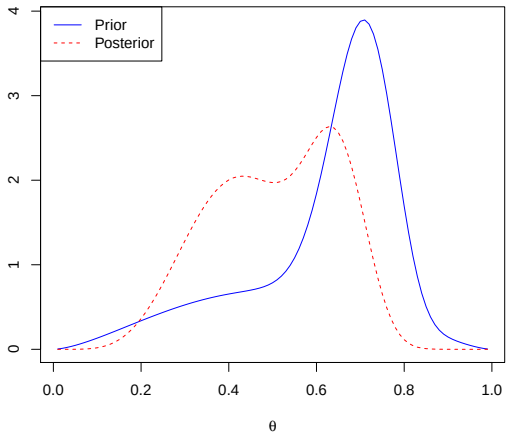
$$p(\theta|y) = Q_1 \text{ Beta}(a_1 + y, b_1 + n - y) + Q_2 \text{ Beta}(a_2 + y, b_2 + n - y)$$

where

$$Q_1 = \frac{q_1 \frac{B(a_1+y, b_1+n-y)}{B(a_1, b_1)}}{\left[q_1 \frac{B(a_1+y, b_1+n-y)}{B(a_1, b_1)} + q_2 \frac{B(a_2+y, b_2+n-y)}{B(a_2, b_2)} \right]}$$

And after calculation, Q_1 = 0.3442245 and Q_2 = 0.6557755. Figure 2.3 shows the prior and the posterior when n=10 and y=4.

Figure 2.3: Mixture prior and posterior distributions for the probability of by-pass surgery survival for experienced heart surgeon and first year medical, where n=10,y=4.



2.4 References

- Hidden dangers of specifying noninformative priors. Seaman, Seaman, and Stamey. 2012. The American Statistician.
- The prior can generally only be understood in the context of the likelihood. Gelman, Simpson, and Betancourt. arXiv- 28 Aug 2017.
- Penalising model component complexity: A principled, practical approach to constructing priors. Simpson, Rue, Martins, Riebler, and Sorbye. 2015. Statistical Science. Describes a method for constructing penalised complexity priors that may become a commonly used approach.

2.5 Week 2 Homework

Exercises: not to turn in

1. The data Y are Binomial(n, θ). The prior for θ is Beta(α, β).
   (a) Show that the posterior distribution for θ is Beta($\alpha + y$, $\beta + n - y$).
   (b) Show that the posterior mean for θ, $E[\theta|y]$ denoted $\overline{\theta}_{post}$, can be written as a weighted combination of the maximum likelihood estimate (mle), namely $\hat{\theta} = y/n$, and the prior mean, namely $\overline{\theta}_{prior} = \frac{\alpha}{\alpha+\beta}$:

   $$\overline{\theta}_{post} = w\hat{\theta} + (1-w)\overline{\theta}_{prior}$$

   What is w?
2. The data $\mathbf{Y} = y_1, y_2, \ldots, y_n$ are independent and identically distributed (iid) Poisson(θ) random variables. The prior for θ is Gamma(α, β).
   (a) Show that the posterior distribution for θ is Gamma($\alpha + \sum_{i=1}^{n} y_i$, $\beta + n$).
   (b) Show that the posterior mean for θ can be written as a weighted combination of the mle, namely $\hat{\theta} = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, and the prior mean, $\overline{\theta}_{prior} = \alpha/\beta$:

   $$\overline{\theta}_{post} = w\hat{\theta} + (1-w)\overline{\theta}_{prior}$$

   What is w?
3. Joint probability distribution exercise (from Reich & Ghosh). X_1 and X_2 have a Bivariate Normal Distribution pdf:

   $$X_1, X_2 \sim \text{BVN}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right)$$

   where the joint pdf can be written as follows:

   $$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1-\rho^2)}\right]$$

   Note: ρ is the correlation between X_1 and X_2.

   Let $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$.
   (a) Derive the marginal distribution of X_1.
   (b) Derive the conditional distribution of $X_1|X_2$.

3 Bayesian Theory—Priors, Part 2

3.1 Specifying hyperparameters

Hyperparameters are sometimes calculated on the basis of subjective specifications of summary measures for the prior distribution. One common procedure is to equate *a priori* guesses of expected values, variances, or coefficients of variances to the algebraic formulas for those values in the prior distribution. This is sometimes referred as moment matching (see **Bayesian Models: A Statistical Primer for Ecologists** (2015), by Hobbs and Hooten). Another is to specify the probability that θ lies in some range $[\theta_L, \theta_U]$; e.g., $\Pr(3 \leqslant \theta \leqslant 9)$ = 0.8.

Beta prior for Binomial. For example, a Beta(α, β) prior will be used for the probability parameter θ in a binomial data distribution, Binomial(n, θ), and the desired expected value of θ is 0.8 with a coefficient of variation of 0.25 ($\sqrt{V[\theta]}/E[\theta]$). This implies a desired variance of $(0.25 * 0.8)^2$ = 0.04. One can calculate α and β by solving the following system of two equations with two unknowns[1].

$$E[\theta] = \frac{\alpha}{\alpha + \beta}$$
$$\mathrm{Var}[\theta] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Gamma prior for Poisson. A similar exercise can be carried out when specifying a Gamma(α, β) prior for a Poisson data distribution, Poisson(θ). Desired characteristics of the prior distribution are that the expected value of θ is 10 with a coefficient of variation of 0.30 ($\sqrt{\mathrm{Var}[\theta]}/E[\theta]$). One can calculate α and β using the following[2]:

$$E[\theta] = \frac{\alpha}{\beta}$$
$$\mathrm{Var}[\theta] = \frac{\alpha}{\beta^2}$$

Normal prior for Normal μ. Instead of specifying moment-related measures of parameters, one might specify desired quantiles. For example, the data distribution is Normal(θ, σ^2), where σ^2 is known, and a normal distribution will be used as the prior for θ, i.e., $\theta \sim$ Normal(μ_o, σ_o^2). Specifying a 95% credible interval, the 2.5th and 97.5th percentiles for the prior should be 15 and 30. One can calculate μ_0 and σ_0^2 using the following equations[3]:

$$15 = \mu_o - 1.960 * \sigma_o$$
$$30 = \mu_o + 1.960 * \sigma_o$$

[1] Exercise: Given $E[\theta]$=0.8 and CV=0.25, show that α=2.4 and β=0.6.

[2] Exercise: Given $E[\theta]$=12 and CV=0.3, show that α=11.11111 and β=0.9259259.

[3] Exercise: Given 2.5th and 97.5th percentiles are equal to 15 and 30, show that μ_o=22.5; σ_o= 3.826531.

In general, letting z_p denote the standard normal, Normal(0,1), quantile for probability p, and y_p denote the normal quantile for probability p for Normal(μ_o, σ_o^2):

$$y_{p1} = \mu_o + z_{p1} * \sigma_o$$
$$y_{p2} = \mu_o + z_{p2} * \sigma_o$$

where $p1 < p2$ (and $y_{p1} < y_{p2}$).

3.2 Normal Distribution Priors

Given the ubiquity and utility of the normal distribution, it is useful to thoroughly explore the posterior for the parameters of that distribution for various priors. We begin with conjugate priors for just one parameter of a univariate normal distribution, here denoted Normal(μ, σ^2), either μ or σ^2. Later we'll examine the case of specifying joint prior distributions for μ and σ^2.

To derive the conjugate posteriors, there is a fair amount of algebraic manipulation involved. It is worth understanding and being able to reproduce the following for yourself as such techniques are useful in other circumstances.

In all cases considered, the data distribution is Normal(μ, σ^2) from which there are n independent and identically distributed (iid) observations, $\mathbf{y}$ = (y_1, y_2, ..., y_n). The joint probability density function is

$$f(\mathbf{y}|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2}\right) \tag{3.1}$$

Remarks.

- *Re-expressing the Normal pdf.* A useful re-expression of the normal pdf is the following.

$$f(\mathbf{y}|\mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{2\sigma^2}\right) \exp\left(-\frac{n(\bar{y} - \mu)^2}{2\sigma^2}\right) \tag{3.2}$$

  Check the validity of this re-expression. (Hint: rewrite $(y_i - \mu)^2$ as $([y_i - \bar{y}] + [\bar{y} - \mu])^2$.)

- *Sufficient statistics.* Suppose that σ^2 is known. Note the only term in eq'n 3.2 containing μ includes $\bar{y}$, thus $\bar{y}$ is the relevant data for inference about μ. We say that $\bar{y}$ is a *sufficient statistic* for μ, as this statistic contains all the information in the data that is useful for estimating the unknown parameter, namely μ[4] .

  Now suppose that σ^2 is also unknown, letting s^2=$\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$, then

$$f(\mathbf{y}|\mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{(n-1)}{2\sigma^2}s^2\right) \exp\left(-\frac{n(\bar{y} - \mu)^2}{2\sigma^2}\right) \tag{3.3}$$

  and $(\bar{y}, s^2)$ is the *joint sufficient statistic* for (μ, σ^2).

- *Precision.* Instead of writing the normal pdf in terms of μ and σ^2, it is sometimes written in terms of μ and τ, where τ is called the Precision, and is the inverse of the variance, $\tau = 1/\sigma^2$. Symbolically, $y \sim$ Normal$\left(\mu, \frac{1}{\tau}\right)$, and mathematically

$$f(y|\mu, \tau) = \frac{\sqrt{\tau}}{\sqrt{2\pi}} \exp\left(-\frac{\tau(y - \mu)^2}{2}\right) \tag{3.4}$$

[4]Consider an *iid* sample $\mathbf{Y} = Y_1, Y_2, \ldots, Y_n$ from a probability distribution $f(Y|\theta)$. Define a statistic $T(\mathbf{Y})$, a function of $\mathbf{Y}$. That statistic $T(\mathbf{Y})$ is a *sufficient statistic for* θ if $Pr[\mathbf{Y}|T(\mathbf{Y})] \neq g(\theta)$, i.e., the conditional probability distribution is **not** a function of θ. Note sufficient statistics are not necessarily unique; e.g., $\bar{Y}$ and $\sum_{i=1}^{n} Y_i$ are both sufficient for μ with a Normal(μ, σ^2) distribution.

Some software packages for Bayesian inference, e.g. JAGS and WinBUGS, specify the normal distribution in terms of the precision τ.

The term precision has an intuitive interpretation, if a random variable is more precise, it is less variable, i.e., as τ increases, σ^2 decreases.

3.2.1 Normal distribution: unknown μ and known σ^2

While assuming that σ^2 is known is usually not realistic, the methods used for inference for μ are helpful building blocks for more realistic situations.

To begin we write the likelihood for μ, given that σ^2 is known, and examine the portion of the pdf in Eq'n (3.2) that involves μ alone:

$$f(\mathbf{y}|\mu, \sigma^2) \equiv L(\mu|\mathbf{y}, \sigma^2) \propto \exp\left(-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}\right) = \exp\left(-\frac{(\mu-\bar{y})^2}{2\sigma^2/n}\right) \tag{3.5}$$

Note that the last expression in (3.5) can be seen as the kernel of a Normal distribution. Thus the *conjugate distribution* for μ (when σ^2, or τ, is known) is the Normal distribution:

$$\text{Prior for } \mu: \quad \mu \sim \text{Normal}\left(\mu_0, \sigma_0^2\right)$$

where μ_0 and σ_0^2 are the hyperparameters of the prior. An alternative formulation for the prior (see Reich and Ghosh, p 47), which leads to a tidier posterior distribution, is to write $\sigma_0^2 = \sigma^2/m$, where m is some positive number. Of course if one specifies σ_0^2 first, then m is σ^2/σ_0^2.

$$\text{Prior for } \mu: \quad \mu \sim \text{Normal}\left(\mu_0, \frac{\sigma^2}{m}\right) \tag{3.6}$$

The **Posterior Distribution** for μ given a normal prior is then:

$$\begin{aligned}
p(\mu|\mathbf{y}, \sigma^2) &\propto \pi(\mu) f(\mathbf{y}|\mu) \\
&= (2\pi\sigma^2/m)^{-\frac{1}{2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma^2/m}\right) * \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}\right) \\
&\propto \exp\left(-\frac{m(\mu-\mu_0)^2}{2\sigma^2}\right) \exp\left(-\frac{n(\bar{y}-\mu)^2}{2\sigma^2}\right) \\
&\propto \exp\left[-\frac{m\mu^2 - 2m\mu_0\mu + n\mu^2 - 2n\bar{y}\mu}{2\sigma^2}\right] \\
&= \exp\left[-\frac{(m+n)\mu^2 - 2(m\mu_0 + n\bar{y})\mu}{2\sigma^2}\right] \\
&\propto \exp\left[-\frac{(\mu - \frac{m\mu_0+n\bar{y}}{m+n})^2}{2\sigma^2/(m+n)}\right]
\end{aligned} \tag{3.7}$$

where eq'n 3.7 is the kernel for a normal distribution, in other words:

$$\text{Posterior for } \mu|\mathbf{y}, \sigma^2: \quad \text{Normal}\left(\frac{m\mu_0 + n\bar{y}}{m+n}, \frac{\sigma^2}{m+n}\right) \tag{3.8}$$

The posterior mean for μ is thus a weighted combination of the prior mean, μ_0, and the sample mean, $\bar{y}$:

$$E[\mu|\mathbf{y}, \sigma^2] = \frac{m\mu_0 + n\bar{y}}{m+n} = \frac{m}{m+n}\mu_0 + \frac{n}{m+n}\bar{y} = (1-w)\mu_0 + w\bar{y}$$

where w=$\frac{n}{m+n}$.

Comments.

- As n increases, the posterior mean is dominated by the sample mean:

$$\lim_{n\to\infty} E[\mu|\mathbf{y}, \sigma^2] = \lim_{n\to\infty}\left[\frac{m}{m+n}\mu_0 + \frac{n}{m+n}\bar{y}\right] = \bar{y}$$

  And the posterior becomes concentrated at $\bar{y}$

$$\lim_{n\to\infty} \mathrm{Var}[\mu|\mathbf{y}, \sigma^2] = \lim_{n\to\infty}\frac{\sigma^2}{m+n} = 0$$

- Conversely, as m increases, σ^2/m decreases, and the posterior is dominated by the prior.

$$\lim_{m\to\infty} E[\mu|\mathbf{y}, \sigma^2] = \lim_{m\to\infty}\frac{m\mu_0 + n\bar{y}}{m+n} = \mu_0$$
$$\lim_{m\to\infty} \mathrm{Var}[\mu|\mathbf{y}, \sigma^2] = \lim_{m\to\infty}\frac{\sigma^2}{m+n} = 0$$

  Thus a point mass at μ_0 as σ_0^2 goes to 0.

- Conversely, as m decreases, σ^2/m increases (a "vaguer" prior), the influence of the prior decreases:

$$\lim_{m\to 0} E[\mu|\mathbf{y}, \sigma^2] = \lim_{m\to 0}\frac{m\mu_0 + n\bar{y}}{m+n} = \bar{y}$$
$$\lim_{m\to 0} \mathrm{Var}[\mu|\mathbf{y}, \sigma^2] = \lim_{m\to 0}\frac{\sigma^2}{m+n} = \frac{\sigma^2}{n}$$

Unknown μ and known τ

If express the sampling distribution for $\mathbf{y}$ in terms of precision, $\tau = 1/\sigma^2$, then the prior for μ (given known τ) is:

$$\mu \sim \text{Normal}\left(\mu_0, \frac{1}{\tau m}\right) \tag{3.9}$$

And the posterior for μ:

$$\text{Posterior for } \mu|\mathbf{y}, \sigma^2\text{: Normal}\left(\frac{m\mu_0 + n\bar{y}}{m+n}, \frac{1}{\tau(m+n)}\right) \tag{3.10}$$

Posterior predictive distribution

As discussed in Lecture 1, the posterior predictive distribution is the marginal probability distribution for a new scalar value, y^{new}, given the past data, $\mathbf{y}^{old}$:

$$p(y^{new}|\mathbf{y}^{old}) = \int p(y^{new}|\theta, \mathbf{y}^{old})p(\theta|\mathbf{y}^{old})d\theta = \int p(y^{new}|\theta)p(\theta|\mathbf{y}^{old})d\theta \tag{3.11}$$

The second equality results from y^{new} being conditionally independent on $\mathbf{y}^{old}$ given θ.

In this normal distribution case with known σ^2, letting μ_1 and σ_1^2 denote the posterior mean and variance for μ:

$$\begin{aligned} p(y^{new}|\mathbf{y}^{old}) &= \int p(y^{new}|\mu)p(\mu|\mathbf{y}^{old})d\mu \\ &= \int \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y^{new}-\mu)^2}{2\sigma^2}\right)\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp\left(-\frac{(\mu-\mu_1)^2}{2\sigma_1^2}\right)d\mu \qquad (3.12) \\ &= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma^2)}}\exp\left(-\frac{(y^{new}-\mu_1)^2}{2(\sigma_1^2+\sigma^2)}\right) \qquad (3.13) \end{aligned}$$

Thus, $y^{new}|y^{old} \sim$ Normal$(\mu_1, \sigma_1^2 + \sigma^2)$[5]. Note that the variance of y^{new} is the sum of the variance for the distribution of y if μ were known, namely, σ^2, and the variance due to the uncertainty in μ[6].

Example: inference for μ

Assume that the amount of money spent during September on food by individual students, y, is Normally distributed with an unknown mean μ and known standard deviation, σ, of £50, or a precision, τ, of $1/50^2$ = 0.0004. You would like to estimate μ and will take a simple random sample of n=20 students.

Before doing so, you specify a Normal(μ_0, $\frac{50^2}{m}$) prior for μ. You think that the average is around μ_0 = £200 and set μ_0 = 200. Further, you guess that the 25th and 75th percentiles are 150 to 250. Given that the corresponding standard normal percentiles are -0.6744898 and 0.6744898, you can solve for m using the following equation:

$$-0.6744898 = \frac{150 - 200}{50/\sqrt{m}}$$

Thus m = 0.6744898^2 = 0.4549365. Then the prior for μ:

$$\mu \sim \text{Normal}\left(200, \frac{50^2}{0.4549365}\right)$$

The simple random sample of n=20 was then taken and the average was £165.

The mean and variance for the posterior distribution for μ are based on (3.8):

$$\mathrm{E}[\mu|\bar{y}] = \frac{0.4549365 * 200 + 20 * 165}{0.4549365 + 20} = 165.7784$$
$$\mathrm{Var}[\mu|\bar{y}] = \frac{50^2}{0.4549365 + 20} = 122.2199 = 11.05531^2$$

and the posterior for μ is

$$\mu|\bar{y} \sim \text{Normal}\left(165.778, 11.05531^2\right)$$

A weight of 0.4549365/20.4549365, about 2%, was given to the prior mean (and 98% to the sample mean) and the posterior standard deviation for μ went from 74 in the prior ($50/\sqrt{0.4549365}$)) to 11.05531 in the posterior.

Figure 3.1 shows the prior and posterior distributions for μ.

Posterior predictive distributions. If a student was randomly sampled, the prior and posterior predictive distributions for that student's food expenditures are:

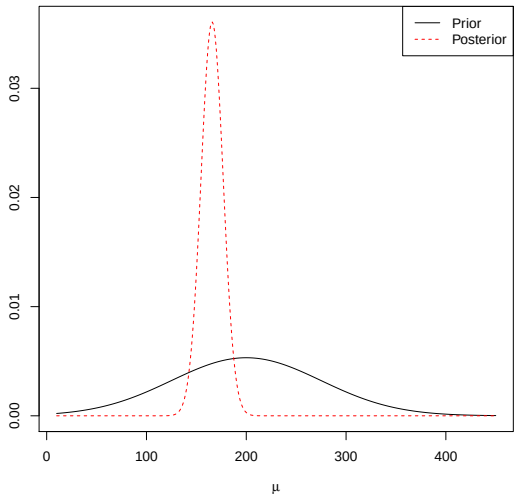
$$\text{Prior } y^{new} \sim \text{Normal}\left(200, 50^2 + \frac{50^2}{0.4549365} = 89.41629^2\right)$$
$$\text{Posterior } y^{new}|\bar{y}^{old} = 165 \sim \text{Normal}\left(165.7784, 50^2 + 11.05531^2 = 51.20762^2\right)$$

Comparing the posterior to the prior, the mean has decreased by around £34. Also, the prediction variance of 51.2^2 is slightly larger than 50^2, with the additional variance due to the uncertainty in the value of μ.

[5] For details on the derivation see (handwritten) Note on Calculating Posterior Predictive Dist'n for a Normal in Week 3 material on *Learn*.

[6] If you have taken a frequentist statistics course including prediction intervals in linear regression, you will have seen this combination of 2 sources of uncertainty manifested by *Prediction Intervals* for $y|x$, say, being wider than *Confidence Intervals* for $\mathrm{E}[Y|x]$.

Figure 3.1: Prior and posterior distribution for μ, average amount spent on food during September, assuming Normal(μ, 50^2) distribution.



3.2.2 Normal distribution: unknown τ and known μ

Again this is usually not a realistic situation, but the results are useful for more complex and realistic models. In this section inference is for the precision, $\tau = 1/\sigma^2$, while the next section has inference for the variance σ^2.

The conjugate prior for τ

We examine the likelihood for τ, keeping only terms that involve τ (see Eq'n 3.1).

$$f(\mathbf{y}|\mu,\tau) \equiv L(\tau|\mathbf{y},\mu) \propto (\tau)^{\frac{n}{2}} \exp\left(-\frac{\tau\sum_{i=1}^{n}(y_i-\mu)^2}{2}\right) = (\tau)^{\frac{n}{2}} \exp\left(-\frac{\tau z^2}{2}\right) \tag{3.14}$$

where to reduce notation

$$z^2 = \sum_{i=1}^{n}(y_i-\mu)^2$$

The likelihood (3.14) is the kernel of a Gamma distribution. Thus the conjugate prior for τ when μ is known is

$$\tau \sim \text{Gamma}\,(\alpha,\beta) \tag{3.15}$$

and the posterior for τ:

$$\begin{aligned}\text{Posterior:}\quad \tau|\mathbf{y},\mu &\propto (\tau)^{\alpha-1}\exp(-\beta\tau)\,(\tau)^{\frac{n}{2}}\exp\left(-\frac{z^2\tau}{2}\right)\\ &= (\tau)^{\left(\alpha+\frac{n}{2}-1\right)}\exp\left(-(\beta+z^2/2)\tau\right) &(3.16)\\ &\Rightarrow &(3.17)\\ \tau|\mathbf{y},\mu &\sim \text{Gamma}\left(\alpha+\frac{n}{2},\beta+\frac{z^2}{2}\right) &(3.18)\end{aligned}$$

Comments.

- Recall that if $\theta \sim$ Gamma(α, β), then $E[\theta]=\alpha/\beta$ and $V[\theta] = \alpha/\beta^2$.
- Examining the posterior mean:

$$E[\tau|\mathbf{y}] = \frac{\alpha + n/2}{\beta + z^2/2} = \frac{\alpha}{\beta + z^2/2} + \frac{n/2}{\beta + z^2/2} = \frac{2\alpha}{2\beta + z^2} + \frac{n}{2\beta + z^2} \tag{3.19}$$

  Referring to $\frac{2\alpha}{2\beta+z^2}$, as n increases z^2 increases, thus the term goes to zero. Referring to $\frac{n}{2\beta+z^2}$ goes to $n/\sum_{i=1}^{n}(y_i - \mu)^2 = \frac{1}{\hat{\sigma}^2}$, where $\hat{\sigma}_2$ is the maximum likelihood estimate (mle) for σ^2. Thus as n increases $E[\tau|\mathbf{y}]$ approaches $1/\hat{\sigma}^2$ or $\hat{\tau}$, the mle for τ.
- One approach for selecting the hyperparameters for the Gamma dist'n prior is to specify approximate values for $E[\tau]$ and $\mathrm{Var}[\tau]$ and then solve for α and β. (Admittedly, specifying a value for $V[\tau]$ might be a little involved.)
- *An Aside: Re-expression as a χ^2 distribution.* The χ^2 distribution with ν degrees of freedom has the following pdf (see Appendix A of King and Ross's notes).

$$p(\theta) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)}\theta^{\frac{\nu}{2}-1}\exp\left(-\frac{\theta}{2}\right)$$

  Note that this is the same pdf as for Gamma$\left(\frac{\nu}{2}, \frac{1}{2}\right)$. Focusing on the kernel of the posterior for τ in (3.16):

$$\begin{aligned} p(\tau|\mathbf{y}, \mu) &\propto (\tau)^{\left(\alpha+\frac{n}{2}-1\right)}\exp\left(-(\beta + z^2/2)\tau\right) \\ &\propto (2\beta + z^2)^{\frac{2\alpha+n}{2}-1}\tau^{\frac{2\alpha+n}{2}-1}\exp\left(-\frac{\tau(2\beta + z^2)}{2}\right) \qquad (3.20) \\ &= (\tau(2\beta + z^2))^{\frac{2\alpha+n}{2}-1}\exp\left(-\frac{\tau(2\beta + z^2)}{2}\right) \qquad (3.21) \end{aligned}$$

  where the multiplier $(2\beta + z^2)^{\frac{2\alpha+n}{2}-1}$ in (3.20) is a constant that does not affect the kernel. Then it can be seen that (3.21) is the kernel of a χ^2 distribution for $\tau(2\beta + z^2)$ with $2\alpha + n$ degrees of freedom:

$$\tau(2\beta + z^2) \sim \chi^2_{2\alpha+n} \tag{3.22}$$

  This can also be written as what is called a *scaled* χ^2 distribution[7]:

$$\tau \sim \frac{1}{2\beta + z^2}\chi^2_{2\alpha+n} \tag{3.23}$$

3.2.3 Normal distribution: unknown σ^2 and known μ

The conjugate prior for σ^2

Examine the likelihood for σ^2 and only keep terms that involve σ^2.

$$f(\mathbf{y}|\mu, \sigma^2) \equiv L(\sigma^2|\mathbf{y}, \mu) \propto \left(\sigma^2\right)^{-\frac{n}{2}}\exp\left(-\frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2}\right) = \left(\sigma^2\right)^{-\frac{n}{2}}\exp\left(-\frac{z^2}{2\sigma^2}\right) \tag{3.24}$$

[7] If a random variable θ multiplied by a constant c follows a distribution D, $c\theta \sim D$, then θ follows a scaled distribution $(1/c)D$.

The term on the right-hand side of Eq'n (3.24) is the kernel of an Inverse Gamma distribution. The pdf for an Inverse Gamma with parameters α and β (see Appendix A of King and Ross's notes):

$$p(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}(\theta)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\theta}\right) \tag{3.25}$$

Thus Inverse Gamma(α, β) is the conjugate prior for σ^2 when μ is known.

Then the posterior for σ^2:

$$\begin{aligned} \text{Posterior:} \quad \sigma^2|\mathbf{y}, \mu &\propto (\sigma^2)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right) (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{z^2}{2\sigma^2}\right) \\ &= (\sigma^2)^{-\left(\alpha+\frac{n}{2}+1\right)} \exp\left(-\frac{\beta + z^2/2}{\sigma^2}\right) \end{aligned} \tag{3.26}$$

where eq'n 3.26 is the kernel for the Inverse Gamma:

$$\text{Posterior for } \sigma^2|\mu: \quad \text{Inverse Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{z^2}{2}\right) \tag{3.27}$$

Comments.

- If $\theta \sim$ Inverse Gamma(α, β), $E[\theta] = \beta/(\alpha-1)$ if $\alpha > 1$. and $\mathrm{Var}[\theta]$ is $\beta^2/((\alpha-1)^2(\alpha-2))$ if $\alpha > 2$. Thus, values for the hyperparameters, α and β, can be deduced given prior notions about the mean and the variance of σ^2.
- Relatively small values for α and β, e.g., 0.01 or 0.001, are often used in practice. Such choices are not universally considered a good idea. However, examining the sensitivity of the posterior to choices of α and β is good practice.
- As for τ, the posterior for σ^2 can be written as a scaled negative χ^2 distribution (See Appendix A of King and Ross).
- If $x \sim$ Gamma(α, β), then $y = 1/x \sim$ Inverse Gamma(α, β)[8].

Example: inference for σ^2

The construction timber called 2x4 has cross-piece dimensions of 1.5 inches by 3.5 inches on average. Let Y be the longer dimension and assume that $Y \sim$ Normal(μ, σ^2). The higher the quality control the closer the value of Y should be to 3.5 inches, in other words, σ^2 should be relatively small. A building company is considering purchasing timber from a new supplier but before doing so would like to see just how precisely the 2 by 4's are cut. They will take a random sample of n=10 2x4s and measure the length of the longer side. They are comfortable assuming that the average length μ is 3.5, thus the data distribution is $Y \sim$ Normal(3.5, σ^2).

Before taking the sample, they decide to use a Inverse Gamma(α, β) prior distribution for σ^2, and need to specify the hyperparameters. They would like to be cautious and will assume *a priori* that the average value of σ^2 is 0.05 with a variance of 0.02. This results in Inverse Gamma$(2.125, 0.05625)$ (check this). A simple random sample of n=10 2x4s yielded the following measurements:

```
3.527 3.387 3.466 3.382 3.612 3.680 3.471 3.475 3.603 3.680
```

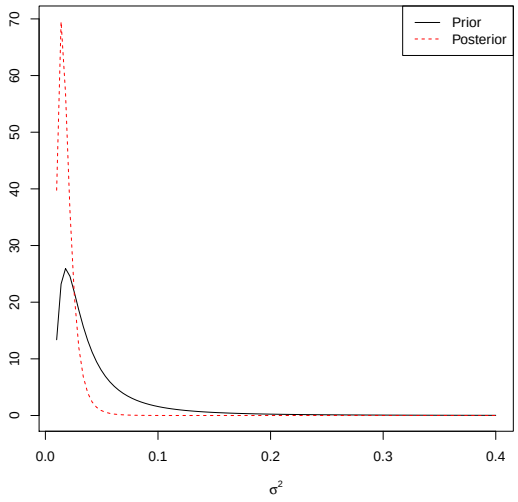
where $\sum_{i=1}^{10}(y_i - 3.5)^2$ = 0.117997. The posterior distribution for $\sigma^2|\mathbf{y}$:

$$\begin{aligned} \sigma^2|\mathbf{y} &\sim \text{Inverse Gamma}\left(2.125 + \frac{10}{2}, 0.05625 + \frac{0.117997}{2}\right) \\ &= \text{Inverse Gamma}\,(7.125, 0.1152385) \end{aligned}$$

[8]This can be shown using the so-called change of variable theorem.

Thus the posterior mean for σ^2 is $\frac{0.1152385}{7.125-1}$ = 0.0188 and posterior variance is $\frac{0.1152385^2}{(7.125-1)^2*(7.125-2)}$ = 6.908e-05. Figure 3.2 plots the prior and posterior distributions for σ^2.

Figure 3.2: Prior and posterior distribution for σ^2, variance of the longer side of 2x4s, assuming sides are Normal(3.5, σ^2).



R does not have "built-in" Inverse Gamma distribution functions which could be used to calculate the quantiles of the posterior distribution for σ^2. However, we can use the quantile function for the Gamma distribution in R, namely `qgamma` which will yield quantiles for $\tau = 1/\sigma^2$, and invert the results. The posterior distribution for τ is $\Gamma(7.125, 0.1152385)$ and the 2.5 and 97.5 percentiles can be found with `qgamma(c(0.025,0.975),shape=7.125, rate=0.1152385)` = (25.10391, 114.81646). Thus

$$0.95 = \Pr(25.104 \leqslant \tau \leqslant 114.8) = \Pr(25.104 \leqslant \frac{1}{\sigma^2} \leqslant 114.8)$$
$$= \Pr\left(\frac{1}{114.8} \leqslant \sigma^2 \leqslant \frac{1}{25.104}\right) = \Pr\left(.00871 \leqslant \sigma^2 \leqslant 0.03983\right)$$

Thus a 95% credible interval for σ^2 is (0.00871, 0.03983).

3.3 Bayes Theorem for multiple parameters

Often there will be $q > 1$ parameters:

$$\Theta = \{\theta_1, \theta_2, \ldots, \theta_q\}$$

Given data $\mathbf{y}$, Bayes theorem has the same form

$$\Pr(\Theta|\mathbf{y}) = \frac{\Pr(\mathbf{y}|\Theta)\Pr(\Theta)}{\Pr(\mathbf{y})} \propto \Pr(\mathbf{y}|\Theta)\Pr(\Theta)$$

3.3.1 Comments

- The posterior distribution for multiple parameters can be high dimensional, e.g., q parameters = q dimensions, and summarising a high dimensional space can be complicated.

- Often one dimensional summaries, namely marginal posterior distributions, are examined instead:

$$p(\theta_i|\mathbf{y}) = \int p(\Theta|\mathbf{y}) d\theta_1 d\theta_2 \dots d\theta_{i-1} d\theta_{i+1} \dots d\theta_q$$

  This is the posterior distribution for θ_i found by "averaging" over all the other parameters, and thus "projecting" (collapsing) the q-dimensional posterior distribution on a single dimension. One can then examine a single posterior density plot, for example, calculate posterior mean, variance, and credible interval for θ_i.
- However, one dimensional summaries can fail to capture important features of the joint posterior.
- Two-dimensional graphical summaries are useful: contour plots or perspective plots, or in the case of samples from the posterior distribution, pairwise scatterplots.
- Two-dimensional numerical summaries include correlations or covariances.
- With more than two-dimensions, however, detecting patterns, if they exist, can be more difficult and complex.

3.3.2 Normal Dist'n: Unknown μ and σ^2

Without deriving any results, we discuss two approaches to the situation where both μ and σ^2 are unknown.

Joint prior constructed with independent marginal priors

One approach to arriving at a joint prior for both μ and σ^2 is to specify independent informative marginal priors for μ and σ^2 and multiply the two to yield a joint prior.

The joint posterior distribution:

$$p(\mu, \sigma^2|\mathbf{y}) = \frac{\pi(\mu)\pi(\sigma^2)\left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (y_i-\mu)^2}{2\sigma^2}\right)}{\int\int \pi(\mu)\pi(\sigma^2)\left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (y_i-\mu)^2}{2\sigma^2}\right) d\mu\, d\sigma^2} \tag{3.28}$$

In general (3.28) will not be something that can be calculated analytically depending on the choices of $\pi(\mu)$ and $\pi(\sigma^2)$, because of the integral in the denominator. Numerical or simulation-based integration methods, which we will discuss later, can be used to yield approximate results.

Consider the 2x4 timber example, but now assume that both μ and σ^2 are unknown. Suppose one specified that the prior for μ is Lognormal(μ_0, σ_0^2) and the prior for σ^2 is Gamma(α, β). The denominator of (3.28) in this case:

$$\int_0^\infty \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma_0^2}} \frac{1}{\mu} \exp\left(-\frac{(\ln(\mu)-\mu_0)^2}{2\sigma_0^2}\right)$$
$$* \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{\alpha-1} \exp(-\beta\sigma^2) \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (y_i-\mu)^2}{2\sigma^2}\right) d\mu\, d\sigma^2$$

which is "probably" not analytically tractable (no closed form solution, I'm guessing as I've not tried to solve it).

Comment. While the joint prior distribution for μ and σ^2 was constructed by multiplying two independent marginal pdfs, the joint posterior distribution is *not* the product of two independent marginal pdfs. This is common: joint priors constructed as products of independent distributions do not usually yield a joint posterior that can be written as products of independent distributions.

Conjugate prior for μ and σ^2

There is a joint conjugate prior density for μ and σ^2. It is defined in terms of a marginal prior density for σ^2, which is an Inverse Gamma, and then a conditional prior density for μ *given* σ^2, which is a Normal where the variance hyperparameter is a function of σ^2. Namely,

$$\begin{aligned} \mu, \sigma^2 &\sim \text{Inverse Gamma}(\alpha, \beta) \\ &\times \text{N}\left(\mu_0, \frac{\sigma^2}{\kappa}\right) \end{aligned} \tag{3.29}$$

There are 4 hyperparameters, α, β, μ_0, and κ. As can be seen, conditional on the value of σ^2, prior uncertainty about the variance of μ around the expected value μ_0 increases as σ^2 increases and as κ decreases. Thus decreasing the values for α and β and the values for κ leads to a more dispersed prior for μ.

The resulting joint posterior density is then the product of an inverse gamma density and a normal density (conditional on σ^2):

$$\begin{aligned} \mu, \sigma^2|\mathbf{y} &\sim \text{Inverse Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{(n-1)s^2}{2} + \frac{\kappa n(\bar{y} - \mu_0)^2}{2(\kappa + n)}\right) \\ &\times \text{Normal}\left(\frac{\kappa\mu_0 + n\bar{y}}{\kappa + n}, \frac{\sigma^2}{\kappa + n}\right) \end{aligned} \tag{3.30}$$

The marginal density for σ^2 is Inverse Gamma, while the marginal density for μ is a students' t distribution (see "Applied Bayesian Statistics", 2013, Cowles, M.K.).

3.3.3 Example: Multiple Parameter Inference

As a demonstration of multiple parameter inference, we fit a Bayesian linear regression of the lengths of dugongs[9] (sea cows, a type of marine mammal; see Figure 3.3) against their age in years. There were n=27 dugongs measured and the sampling model was the following.

$$\text{Length}_i \overset{iid}{\sim} \text{Normal}\left(\beta_0 + \beta_1 \ln(\text{age}), \sigma^2\right)$$

The relationship is shown in Figure 3.4.

Figure 3.3: Dugong (image from Wikipedia)

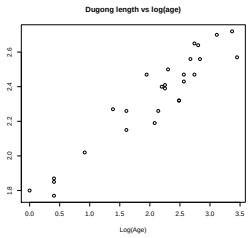


Figure 3.4: Dugong lengths plotted against log(age).

There are three parameters, $\Theta = \{\beta_0, \beta_1, \sigma^2\}$. The following independent marginal prior distributions were chosen to construct a joint prior distribution.

$$\beta_0, \beta_1 \sim \text{Uniform}\,(-50, 50)\,, \quad \sigma^2 \sim \text{Uniform}\,(0.01, 20)$$

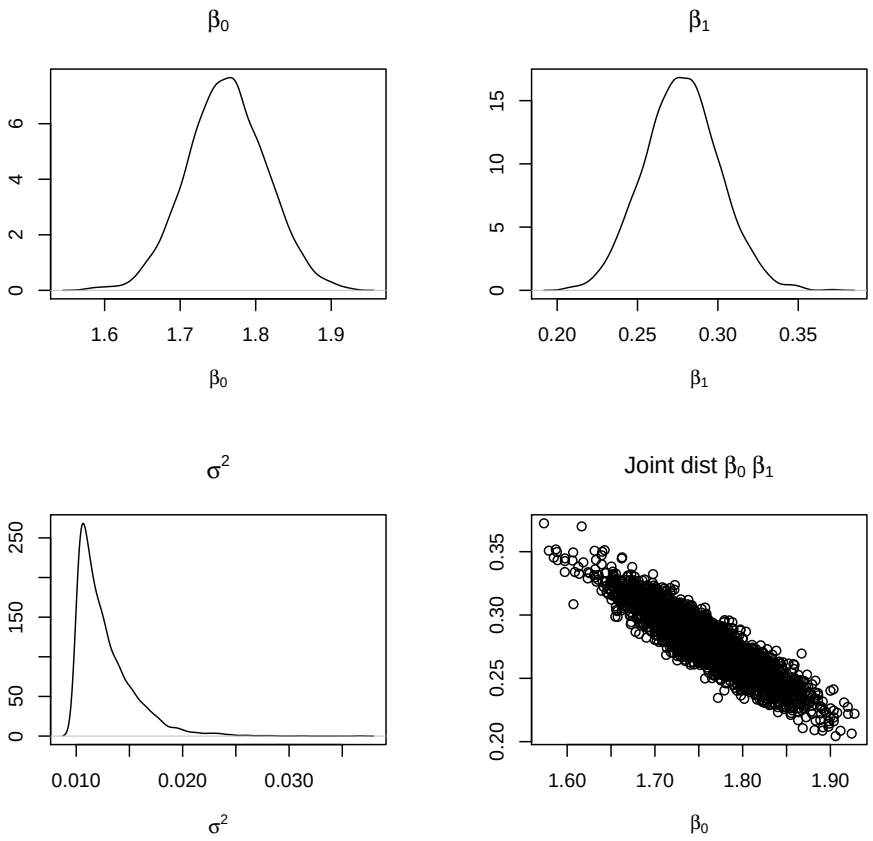
[9] Example motivated from Bayesian Methods for Data Analysis, 2009. Carlin and Louis.

The resulting (estimated) posterior distribution was found using JAGS (code in the Appendix). The posterior means and standard deviation are shown below, along with the maximum likelihood estimates (except $\hat{\sigma}^2$ is bias-correction of the mle).

	Bayesian		Frequentist	
	Mean	SD	mle	std error
β_0	1.76098	0.053089	1.762	0.0424
β_1	0.27757	0.023531	0.277	0.0188
σ^2	0.01277	0.002719	0.0081	

Figure 3.5 shows the marginal posterior distributions for the three parameters as well as a scatterplot of the samples of β_0 and β_1. The scatterplot shows that there is a negative relationship between β_0 and β_1.

Figure 3.5: Marginal posterior distributions for β_0, β_1, σ^2, and the joint distribution of β_0 and β_1



4 Jeffreys' Prior, Eliciting and Analysing Priors

4.1 Jeffreys' prior

4.1.1 Example of the problem of "uninformative" prior

Priors considered "uninformative" for a parameter, θ, may not be considered uninformative for transformations of the parameter, ψ=$g(\theta)$. For example, consider a Uniform(0,20) prior for a parameter θ. The problem with such uniform priors is that the *induced prior* for simple transformations of the parameter, e.g., ϕ=$\sqrt{\theta}$, will not be uniform. Given $\theta \sim$ Uniform(0,20), the distribution for ϕ is found by the change of variable theorem[1]:

$$\pi(\phi) = \frac{1}{20}\left|\frac{d\phi^2}{d\phi}\right| \mathbb{I}(0 < \phi < \sqrt{20}) = \frac{1}{20} 2\phi \mathbb{I}(0 < \phi < \sqrt{20}) = \frac{\phi}{10} \mathbb{I}(0 < \phi < \sqrt{20})$$

where $\mathbb{I}(0 < \phi < \sqrt{20})$ is an indicator function that equals 1 when $0 < \phi < \sqrt{20}$ and 0 otherwise. This *induced prior* clearly is informative, as it linearly increases from 0 to $\sqrt{20}$.

4.1.2 Definition and calculation of Jeffreys' prior

Jeffreys' prior is an example of an *objective prior* which can be seen as a remedy to the just discussed problem of the induced priors resulting from transformations. It is a prior that is invariant to strictly monotonic (1-1 or bijective) transformations of the parameter, say ϕ =$g(\theta)$, where g is strictly monotonic.

Jeffreys' prior is proportional to the square root of the Fisher Information, $I(\theta)$:

$$\text{Jeffreys' prior } \pi_{JP}(\theta) \propto \sqrt{I(\theta)} \tag{4.1}$$

where

$$I(\theta) = \mathbb{E}_{y|\theta}\left[\left(\frac{d \log f(y|\theta)}{d\theta}\right)^2\right] \tag{4.2}$$

[1]The change of variable theorem is a procedure for determining the pdf of a (continuous) random variable Y that is a strictly monotonic (1:1) transformation of another (continuous) random variable X, i.e., Y=$g(X)$. The pdf for Y:

$$p_Y(y) = p_X(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right|$$

See Section 4.5 for more details.

Note that under certain regularity conditions (e.g. that the differentiation operation can be moved inside the integral), Fisher Information can be calculated from the second derivative of the log likelihood:

$$I(\theta) = -\mathbb{E}_{y|\theta}\left[\frac{d^2 \log f(y|\theta)}{d\theta^2}\right] \tag{4.3}$$

which is often much easier to calculate than Eq'n 4.2. For more discussion of Fisher Information see Section 4.6.

4.1.3 Jeffreys' prior is invariant to 1:1 transformations

Remark 4.1. Let $f(y|\theta)$ denote a probability density or mass function for a random variable y where θ is a scalar. Let $\phi = g(\theta)$ where g is a strictly monotonic (1:1 or bijective) transformation. If we specify a Jeffreys' prior for θ, namely, $\pi_{JP}(\theta) \propto \sqrt{I(\theta)}$, then the induced prior on ϕ, $\pi(\phi)$, is proportional to $\sqrt{I(\phi)}$. In other words, a 1:1 transformation of a parameter that has a Jeffreys' prior yields a Jeffreys' prior for the transformed parameter.

Proof. This proof uses the chain rule, $\frac{dy}{dx} = \frac{dy}{dz}\frac{dz}{dx}$, and the change of variable theorem.

Write the Fisher information for θ as follows:

$$\begin{aligned} I(\theta) &= \mathbb{E}_{y|\theta}\left[\left(\frac{d \log f(y|\theta)}{d\theta}\right)^2\right] = \mathbb{E}_{y|\theta}\left[\left(\frac{d \log f(y|\phi))}{d\phi}\frac{d\phi}{d\theta}\right)^2\right] \\ &= \mathbb{E}_{y|\phi}\left[\left(\frac{d \log f(y|\phi))}{d\phi}\right)^2\right]\left|\frac{d\phi}{d\theta}\right|^2 = I(\phi)\left|\frac{d\phi}{d\theta}\right|^2 \end{aligned}$$

Thus the Jeffreys' prior for θ can be written:

$$\pi_{JP}(\theta) \propto \sqrt{I(\theta)} = \sqrt{I(\phi)}\left|\frac{d\phi}{d\theta}\right|$$

Then the induced distribution for ϕ given this prior:

$$\pi(\phi) = \pi_{JP}(\theta)\left|\frac{d\theta}{d\phi}\right| \propto \sqrt{I(\phi)}\left|\frac{d\theta}{d\phi}\right|\left|\frac{d\phi}{d\theta}\right| = \sqrt{I(\phi)}$$

□

4.1.4 Example A. Binomial distribution

Suppose that the prevalence of Potato Virus Y in a population of aphids is an unknown parameter θ. A random sample of n aphids is taken (using a trap) and the number of aphids with the virus is x. Assuming independence between the aphids and that they all have the same probability of having the virus, $x \sim$ Binomial(n, θ). The Fisher information:

$$\begin{aligned} I(\theta) &= -\mathbb{E}_{x|\theta}\left[\frac{d^2\left(\log\binom{n}{x} + x\log(\theta) + (n-x)\log(1-\theta)\right)}{d\theta^2}\right] = -\mathbb{E}_{x|\theta}\left[\frac{-x}{\theta^2} - \frac{n-x}{(1-\theta)^2}\right] \\ &= \frac{\mathbb{E}(x)}{\theta^2} + \frac{n-\mathbb{E}(x)}{(1-\theta)^2} = \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)} \end{aligned}$$

Thus the Jeffreys' prior is

$$\pi_{JP}(\theta) \propto \sqrt{\frac{1}{\theta(1-\theta)}} = \theta^{-1/2}(1-\theta)^{-1/2}$$

which is the kernel for a Beta(1/2,1/2) distribution.

Aside. Note that the mle for θ is $\hat{\theta} = x/n$ and the variance of $\hat{\theta}$ is $\frac{\theta(1-\theta)}{n}$, which equals $I(\theta)^{-1}$.

4.1.5 Example B. Exponential distribution

Let $x_1, \ldots, x_n$ be an iid sample from an exponential distribution with rate parameter λ, namely, $x_i \overset{iid}{\sim}$ Exponential(λ), $i = 1, \ldots, n$. For example, suppose that x_i is amount of time individual i waits in a queue at a bank during lunch hour until seeing a teller.

The Fisher information for a single random variable:

$$I(\lambda) = \mathbb{E}\left[\left(\frac{d \log f(x|\lambda)}{d\lambda}\right)^2\right] = \mathbb{E}\left[\left(\frac{d \log \lambda - \lambda x}{d\lambda}\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{\lambda} - x\right)^2\right]$$
$$= \mathbb{E}\left[\frac{1}{\lambda^2} - 2\frac{x}{\lambda} + x^2\right] = \frac{1}{\lambda^2} - \frac{2}{\lambda^2} + \frac{2}{\lambda^2} = \frac{1}{\lambda^2}$$

where we used the facts that $E(x)=1/\lambda$ and $E(x^2)=2/\lambda^2$ (use the moment generating function $\lambda/(\lambda - t)$). Alternatively, we could use the other expression for $I(\lambda)$:

$$I(\lambda) = -\mathbb{E}\left[\frac{d^2 \log f(x|\lambda)}{d\lambda^2}\right] = -\mathbb{E}\left[\frac{-1}{\lambda^2}\right] = \frac{1}{\lambda^2}$$

Thus the Fisher information for the entire sample is then $I_n(\lambda) = \frac{n}{\lambda^2}$.

Then the Jeffreys' prior:

$$\pi_{JP}(\lambda) \propto \sqrt{I(\lambda)} = \frac{1}{\lambda}$$

Note that this is an *improper* prior as it does not integrate over the domain of λ, namely $(0, \infty)$. However this is a situation where the posterior for λ is proper:

$$p(\lambda|x_1, \ldots, x_n) \propto \pi(\lambda) f(x_1, \ldots, x_n|\lambda) = \frac{1}{\lambda}\lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) = \lambda^{n-1} \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)$$

which is the kernel for a Gamma$(n, \sum_{i=1}^{n} x_i)$ density function.

To demonstrate the invariance under a 1:1 transformation, reparameterize the exponential with θ=$g(\lambda)$ = $\sqrt{\lambda}$, then

$$f(x|\theta) = \theta^2 \exp(-\theta^2 x)$$

and $g^{-1}(\theta) = \theta^2 = \lambda$. Given $\pi(\lambda) \propto 1/\lambda$, the induced prior for θ:

$$\pi_\theta(\theta) = \left|\frac{d g^{-1}(\theta)}{d\theta}\right| \pi_\lambda(g^{-1}(\theta)) = \left|\frac{d\theta^2}{d\theta}\right| \frac{1}{\theta^2} = \frac{2}{\theta}$$

Checking that this is indeed the Jeffreys' prior for θ:

$$I(\theta) = -E\left[\frac{d^2 2\log(\theta) - \theta^2 x}{d\theta^2}\right] = \frac{4}{\theta^2}$$

Thus the Jeffreys' prior is $\pi_{JP}(\theta) \propto \sqrt{I(\theta)} = 2/\theta$.

4.1.6 Example C. Normal dist'n with known variance

The sampling distribution for $y_1, \ldots, y_n$ is Normal(μ, σ^2) where σ^2 is known. It can be shown that the Jeffreys' prior for μ when σ^2 is known is

$$\pi_{JP}(\mu) \propto 1, \ -\infty < \mu < \infty$$

This is an improper prior because $\int_{-\infty}^{\infty} 1 d\mu$ is not finite. However the posterior distribution is proper:

$$\pi(\mu|\mathbf{y}) \propto \exp\left(-\frac{(\mu-\bar{y})^2}{2\sigma^2/n}\right) * 1$$

namely the kernel for a Normal $\left(\bar{y}, \frac{\sigma^2}{n}\right)$.

4.1.7 Example D. Normal dist'n with known mean

The sampling distribution for $y_1, \ldots, y_n$ is Normal(μ, σ^2) where μ is known. The Jeffreys' prior for σ^2 (given known μ) can be shown to be the following:

$$\pi_{JP}(\sigma^2) \propto \frac{1}{\sigma^2}, \ \ 0 < \sigma^2 < \infty \tag{4.4}$$

which is an improper prior because $\int_0^{\infty} \frac{1}{\sigma^2} d\sigma^2$ is not finite. The posterior distribution for σ^2,

$$p(\sigma^2|\mathbf{y}) \propto \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{z^2}{2\sigma^2}\right) \frac{1}{\sigma^2} = \left(\sigma^2\right)^{-(\frac{n}{2}+1)} \exp\left(-\frac{z^2}{2\sigma^2}\right) \tag{4.5}$$

where $z^2 = \sum_{i=1}^{n}(y_i - \mu)^2$. This is the kernel for $\Gamma^{-1}\left(\frac{n}{2}, \frac{z^2}{2}\right)$ so long as $n/2 > 0$ (obviously so), and $z^2 > 0$, which simply means that not all values in $\mathbf{y}$ are identical.

4.1.8 Jeffreys' prior for a multivariate parameter vector.

To calculate Jeffreys' prior for a multivariate parameter vector, θ, with p parameters, one calculates the score function, which is now the gradient of the log likelihood:

$$S(\theta) = \begin{bmatrix} \frac{d}{d\theta_1} \ln(f(x)) \\ \vdots \\ \frac{d}{d\theta_p} \ln(f(x)) \end{bmatrix} \tag{4.6}$$

and then calculates the Hessian of the log likelihood, namely, the matrix of the partial derivatives of the score function

$$H(\theta) = \begin{bmatrix} \frac{d}{d\theta_1}S(\theta)[1] & \frac{d}{d\theta_1}S(\theta)[2] & \ldots & \frac{d}{d\theta_1}S(\theta)[p] \\ \frac{d}{d\theta_2}S(\theta)[1] & \frac{d}{d\theta_2}S(\theta)[2] & \ldots & \frac{d}{d\theta_2}S(\theta)[p] \\ \vdots & \vdots & \vdots & \vdots \\ \frac{d}{d\theta_p}S(\theta)[1] & \frac{d}{d\theta_p}S(\theta)[2] & \ldots & \frac{d}{d\theta_p}S(\theta)[p] \end{bmatrix} = \begin{bmatrix} \frac{d^2}{d\theta_1^2} \ln(f(x)) & \frac{d^2}{d\theta_1 d\theta_2} \ln(f(x)) & \ldots & \frac{d^2}{d\theta_1 d\theta_p} \ln(f(x)) \\ \frac{d^2}{d\theta_2 d\theta_1} \ln(f(x)) & \frac{d^2}{d\theta_2^2 d\theta_2} \ln(f(x)) & \ldots & \frac{d^2}{d\theta_2 d\theta_p} \ln(f(x)) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{d^2}{d\theta_p d\theta_1} \ln(f(x)) & \frac{d^2}{d\theta_2 d\theta_p} \ln(f(x)) & \ldots & \frac{d^2}{d\theta_p^2} \ln(f(x)) \end{bmatrix} \tag{4.7}$$

Fisher information is

$$I(\theta|x) = -\mathbb{E}[H(\theta)] \tag{4.8}$$

Finally, the Jeffreys' prior for the vector of parameters is proportional to the square root of the determinant of the Fisher information matrix:

$$\pi_{JP}(\theta) \propto \sqrt{\det(I(\theta|x))} \tag{4.9}$$

Example E. Normal(μ,σ^2) Jeffreys' prior

To make the differentiation a little less awkward, the normal distribution is parameterized with θ=σ^2. Given $y_1, \ldots, y_n$ (iid) Normal(μ, θ) pdf can be written:

$$f(\mathbf{y}; \mu, \theta) = (2\pi\theta)^{-n/2} e^{-\frac{1}{2\theta}\sum_{i=1}^{n}(y_i-\mu)^2} \tag{4.10}$$

Then the log likelihood:

$$\log L(\mu, \theta) = l \propto -n \log(\theta) - \theta^{-1} \sum_{i=1}^{n} (y_i - \mu)^2 \tag{4.11}$$

The score vector:

$$\frac{dl}{d\mu} = 2\frac{\sum_{i=1}^{n}(y_i - \mu)}{\theta} = 2\frac{n(\bar{y} - \mu)}{\theta} \tag{4.12}$$

$$\frac{dl}{d\theta} = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{\theta^2} \tag{4.13}$$

The Hessian matrix (H) components:

$$\frac{d^2l}{d\mu^2} = -\frac{2n}{\theta} \tag{4.14}$$

$$\frac{d^2l}{d\mu d\theta} = \frac{d^2l}{d\theta d\mu} = -\frac{2n(\bar{y} - \mu)}{\theta^2} \tag{4.15}$$

$$\frac{d^2l}{d\theta^2} = \frac{n}{\theta^2} - \frac{2\sum_{i=1}^{n}(y_i - \mu)^2}{\theta^3} \tag{4.16}$$

Then the Fisher Information matrix

$$I(\mu, \theta) = -E[H] = \begin{bmatrix} \frac{2n}{\theta} & 0 \\ 0 & \frac{n}{\theta^2} \end{bmatrix} = \begin{bmatrix} \frac{2n}{\sigma^2} & 0 \\ 0 & \frac{n}{\sigma^4} \end{bmatrix} \tag{4.17}$$

Then the Jeffreys' prior for (μ, σ^2):

$$\pi_{JP}(\mu, \sigma^2) \propto \sqrt{|I(\mu, \sigma^2)|} = \sqrt{\frac{2n}{\sigma^2}\frac{n}{\sigma^4}} = \sqrt{\frac{2n^2}{\sigma^6}} \propto \frac{1}{(\sigma^2)^{\frac{3}{2}}} \tag{4.18}$$

The posterior can be shown to be the product of an inverse Chi-square (Gamma) and a normal (conditional on σ^2).

Example F. Normal with two independent Jeffreys' priors.

Note: this example is Not using the Determinant as above.

If the Jeffreys' prior for μ (given σ^2 is known) and the Jeffreys' prior for σ^2 (given μ is known) are treated as independent priors, the Jeffreys' prior is

$$\pi(\mu, \sigma^2) \propto 1 * \frac{1}{\sigma^2},\ -\infty < \mu < \infty,\ 0 < \sigma^2 < \infty$$

then it turns out that the posterior marginal distribution for σ^2 is

$$\sigma^2|\mathbf{y} \sim \Gamma^{-1}\left(\frac{n-1}{2}, \frac{(n-1)s^2}{2}\right)$$

where s^2 = $(1/(n-1))\sum_{i=1}^{n}(y_i-\bar{y})^2$, the usual sample variance. And the posterior marginal distribution for μ is a student's t distribution with mean $\bar{y}$, scale parameter s^2/n, and $n-1$ degrees of freedom.

Referring to the 2x4 timber example from Lecture 3 notes, $\bar{y}$=3.5283, s^2= 0.0122209, with 9 degrees of freedom. Thus the posterior expected value of μ is 3.5283. A 95% credible interval for μ can be found by calculating the 2.5 and 97.5 percentiles of the standard $t_{9\ df}$ distribution, multiplying them by the square root of the scale parameter and then adding the mean:

```
qt(c(0.025,0.975),df=9)*sqrt(0.0122209/10) + 3.5283
 3.449219  3.607381
```

Thus 95% credible interval for μ of (3.45, 3.61).

For σ^2, the 95% credible interval is calculated for the Gamma(9/2, 9*0.0122209/2) distribution and then inverting the results:

```
gam.bds <- qgamma(c(0.025,0.975),shape=9/2,rate=9*0.0122209/2)
inv.gam.bds <- sort(1/gam.bds)
inv.gam.bds
0.005781919 0.040730458
```

Thus a 95% credible interval for σ^2 of (0.0058, 0.0407).

4.2 Reference Priors

As mentioned in LN 2, the Jeffreys' prior is one of several objective priors, namely procedures for selecting priors which will yield the same prior for anyone who uses the procedure. Reich and Ghosh (2019) discuss four other objective priors that are at least worth knowing about and I recommend reading the approximately two pages of discussion. Admittedly understanding the general concepts behind the procedures and being able to implement the procedures can have quite different degrees of difficulty, with the latter generally more difficult than the former. Here we will just examine one other type of objective prior, the Reference Prior, denoted $\pi_{\mathrm{RP}}(\theta)$. An example of a middle ground between fully subjective and fully "objective" priors are Penalised Complexity Priors, discussed in Section 4.7.

4.2.1 KL divergence

Before introducing Reference Priors, the notion of Kullback-Leibler (KL) divergence is introduced. KL divergence is a measure of the difference between two pmfs or two pdfs. A KL divergence value of 0 means that the two distributions are identical.

For the continuous case let f and g denote two pdfs. The KL divergence is defined "conditional" on one of the two distributions, here denoted $\mathrm{KL}(f, g)$ or $\mathrm{KL}(g, f)$ where $\mathrm{KL}(f, g) \neq \mathrm{KL}(g, f)$, except when f and g are identical (almost everywhere). More exactly, $\mathrm{KL}(f, g)$ is the expected value of $\log(f(x)/g(x)$ assuming that f is "true"—more accurately stated that the expectation is with respect to $f(x)$, and $\mathrm{KL}(g, f)$ is the reverse:

$$\begin{aligned}\mathrm{KL}(f, g) &= \int \log\left[\frac{f(x)}{g(x)}\right] f(x)dx = \int \log(f(x))f(x)dx - \int \log(g(x))f(x)dx \\ &= \mathrm{E}_f\left[\log(f(X)\right] - \mathrm{E}_f\left[\log(g(X)\right] \end{aligned} \tag{4.19}$$

and

$$\mathrm{KL}(g, f) = \int \log\left[\frac{g(x)}{f(x)}\right] g(x)dx = \mathrm{E}_g\left[\log(g(x))\right] - \mathrm{E}_g\left[\log(f(x)\right] \tag{4.20}$$

Notes: (1) if $g(x) = f(x)$, then $\log\left[\frac{f(x)}{g(x)}\right] = \log(1) = 0$; and (2) $\mathrm{KL}(f, g) \geqslant 0$.

KL divergence examples. As a simple example consider a discrete valued random variable with values 0, 1, or 2. Let $f(x)$ be the Binomial(2,p=0.2) pmf with probabilities 0.64, 0.32, and 0.04 for X=0, 1, and 2, respectively. Let $g(x)$ be the discrete uniform where $g(0) = g(1) = g(2) = 1/3$. Then

$$\mathrm{KL}(f, g) = \sum_{x=0}^{2} f(x) \log\left[\frac{f(x)}{g(x)}\right] = 0.64 \log[0.64/(1/3)] + 0.32 \log[0.32/(1/3)] + 0.04 \log[0.04/(1/3)] = 0$$

$$\mathrm{KL}(g, f) = \sum_{x=0}^{2} g(x) \log\left[\frac{g(x)}{f(x)}\right] = (1/3) \log[(1/3)/0.64] + (1/3) \log[(1/3)/0.32] + (1/3) \log[(1/3)/0.04]$$

For another example let $g(x)$ be a Binomial(2, p=0.25). Then $\mathrm{KL}(f, g)$= 0.01400421 and $\mathrm{KL}(g, f)$ = 0.01476399, thus the KL divergence measures are both close to 0 (and close to each other).

4.2.2 Reference prior

Given data y, the KL divergence between the prior and posterior (with respect to the posterior) is

$$\mathrm{KL}(p(\theta|y), \pi(\theta)) = \int p(\theta|y) \log\left[\frac{p(\theta|y)}{\pi(\theta)}\right] d\theta \tag{4.21}$$

The key idea of a RP is that the KL divergence between the posterior and the prior should be as large as possible, thus implying that the data are dominating the prior.

However, this measure is conditional on the data, y, which does not help for determining a prior. Thus to remove the conditioning on the data, the data are integrated out, and $\pi_{RP}(\theta)$ is the probability distribution (pmf or pdf) that maximizes the following:

$$E_y\left[\mathrm{KL}(p(\theta|y), \pi(\theta))\right] = \int \left[\mathrm{KL}(p(\theta|y), \pi(\theta))\right] m(y) dy \tag{4.22}$$

where $m(y)$ is the marginal distribution for the data.

While this approach is conceptually attractive, it can be technically challenging as one is trying to find an entire probability distribution $\pi(\theta)$ that maximizes eq'n (4.22), noting that determining $m(y)$ involves integration as well.

4.3 Eliciting Informative Priors

Often the scientist or subject-matter specialist will have a definite opinion as to what the range of parameter values should be. For example, an experienced heart surgeon who has done 1000s of coronary by-pass surgeries on a variety of patients will have a definite opinion on post-surgery survival probability.

How does one translate that prior knowledge into a prior probability distribution?

- First, think about whether it's even feasible or are there too many parameters, or is the underlying sampling model fairly complex, e.g., a hierarchical model. For example, suppose the sampling model is a multiple regression with 4 covariates, thus these five parameters, β_0, β_1, β_2, β_3, and β_4, and the variance parameter. How might the expert's knowledge be translated into prior distributions for these 5 parameters? The expert might have an opinion about the signs, positive or negative, of each coefficient, and perhaps the relative importance of each covariate; e.g., dealing with standardized covariates, then the effects of x_1 and x_2 are thought to be positive but the effect of x_1 may be twice as large as x_2.

- Single parameter case. This is generally the most feasible situation. If the expert is not familiar with probability distributions, the statistician may need to work with the expert to arrive at a prior and can help by asking questions about the parameter without using statistics jargon.

  For example, instead of asking for the median, ask "For what value of the parameter do you think that it's equally likely that values are either below or above it?"

  "What do you think the range of values might be?"

  " What do you think the relative variation around an average value be, for example, if your best guess for θ is 15, is your uncertainty within $\pm$ 10% of that value ($\pm$ 1.5), or 20% ($\pm$ 3.0)?" Thus potentially getting a measure of the coefficient of variation, $\mathrm{CV}=\sigma/\mu$.

  Given a mean value, and a range, standard deviation, or CV, and assuming a particular standard probability distribution might suffice, rough estimates of hyperparameters for the prior might be calculated. This is the "moment matching" idea that we've examined previously.

  For example, with the above example of the surgeon, the surgeon was thinking that θ would be 0.7 on average. Further questioning about the surgeon's uncertainty led to a determination that a CV of 0.1 would be appropriate. Using a Beta distribution for the prior, the mean is $\alpha/(\alpha+\beta)$ and the variance is $(\alpha\beta)/[(\alpha+\beta)^2(\alpha+\beta+1)]$, some algebra yields Beta(29.3, 12.6).

- Discrete histogram priors. Another simple way to elicit priors is to partition parameter values into non-overlapping bins, and have the expert present relative weights for each bin. For example, θ is grouped into three bins, [0,10), [10,25), [25,30], and the expert gives relative weights of 0.2, 0.5, and 0.3. A proper histogram, pdf, is constructed using the result that bin area = height $\times$ width, where bin area corresponds to probability or weight. For the bin [0,10), area=0.2, width=10, thus height equals area/width = 0.2/10 = 0.02. For [10,25) height is 0.5/15 = 0.033, and for [25,30] height is 0.3/5 = 0.06.

References

- "The elicitation of prior distributions", Chaloner, 1996, in *Bayesian Biostatistics*, eds., Berry and Stangl.
- *Uncertain Judgements: Eliciting Experts' Probabilities*, O'Hagan, et al. 2006. This book is available online as a pdf via the University Library.

4.4 Sensitivity analysis of priors

Using the term *sensitivity analysis* loosely here, we mean an examination of the effects of different priors on the posterior. For example, the comparison of the posterior distribution for the probability of survival after by-pass surgery for the surgeon's prior and the medical student's prior is a sensitivity analysis.

Another loosely put phrase, if the posterior distributions for different priors look much "the same", e.g., have similar means and variances, then then one might say that the results are *robust* to the priors.

With large enough samples, unless the prior is particularly concentrated over a narrow range of possible values, sometimes called a pig-headed prior, the posterior will look much the same for a wide range of priors as the data (the likelihood) are *dominating* the prior.

Problematic issues

- Practical issue: if 100s of parameters, then tedious at least, to carry out a sensitivity analysis for all the priors.

- Generalized linear models where data come from an exponential family distribution with parameter θ and covariate(s) x, say $\mathcal{F}$, and, $g(\theta, x)$, the "link" function, is a linear model:

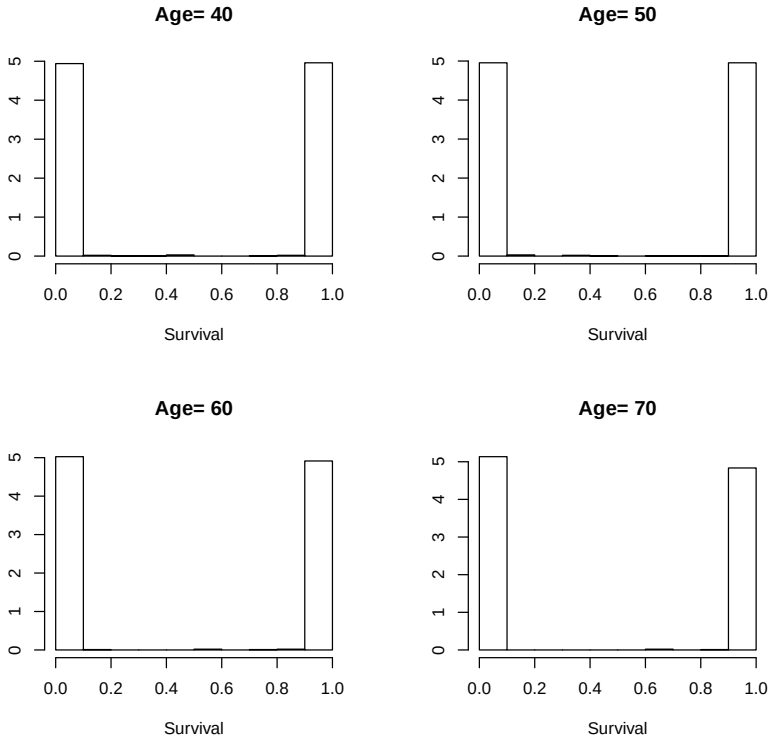
$$y|\theta, x \sim \mathcal{F}(\theta, x)$$
$$g(\theta, x) = \beta_0 + \beta_1 x$$

Apparently uninformative priors in the link function may *induce* quite informative priors at a lower level. For example, a logistic regression for the number of patients surviving heart bypass surgery where the probabilities differ with age:

$$y_i|Age_i \sim \text{Bernoulli}(\theta(Age_i))$$
$$\text{where } \theta(Age) = \frac{\exp(\beta_0 + \beta_1 Age)}{1 + \exp(\beta_0 + \beta_1 Age)}$$
$$\text{equivalently } \ln\left(\frac{\theta(Age)}{1 - \theta(Age)}\right) = \beta_0 + \beta_1 Age$$

A seemingly innocuous prior for both β_0 and β_1 is Normal(μ=0,σ^2=5^2). Suppose that the ages range from 40 to 70. Figure 4.1 shows the results of simulating from the priors for β_0 and β_1 on the induced priors for survival for four different ages. Note how the probabilities are massed near 0 and 1. Such simulation exercises can be quite valuable for detecting such effects.

Figure 4.1: Induced prior for survival probability (θ) as a function of age given Normal(0,5) priors for logit transformation, $\ln(\theta/(1-\theta)) = \beta_0 + \beta_1 Age$.



4.5 Supplement A: Change of Variable Theorem

The Problem

A continuous random variable X has a pdf $f_X(X)$.

A new random variable Y is "constructed" from X by a strictly monotonic function g (a 1:1 function) $Y = g(X)$.

Thus X can be "recovered" from Y by an inverse function, g^{-1}, where $X = g^{-1}(Y)$.

The problem: what is the pdf for Y, namely, $f_Y(Y)$?

The Solution

The pdf for Y is found as follows:

$$f_Y(Y) = f_X(g^{-1}(Y)) \left| \frac{dg^{-1}(Y)}{dY} \right| \tag{4.23}$$

Example 1. X is Uniform(0,20) and $Y=g(X) = 3X$. Thus $g^{-1}(Y) = Y/3$.

Note that the support of Y will be (0,60) as $Y=3X$. Then

$$f_Y(Y) = \frac{1}{20} \left| \frac{dY/3}{dY} \right| = \frac{1}{20}\frac{1}{3} = \frac{1}{60}, \quad I_Y(0 < Y < 60)$$

- $I_Y(\text{condition})$ is an *Indicator* function which takes on one of two values: 1 when the condition is met, is True, and 0 when it is not met, is False.
- This is an "intuitive" result, $Y \sim$ Uniform(0,60)

Example 2. $X \sim$ Uniform(16,64), thus pdf of X is $\frac{1}{64-16} = \frac{1}{48}\, I_X(16 < X < 64)$. Define $Y= g(X) = \sqrt{X}$, noting that the support for Y is (4,8). Then $g^{-1}(Y) = Y^2$, and

$$f_Y(Y) = \frac{1}{48}\frac{dY^2}{dY} = \frac{1}{48}2Y = \frac{1}{24}Y\, I_Y(4 < Y < 8)$$

Skeleton of Proof

The essence of the change of variables theorem is that for $x=g^{-1}(y)$, $Pr(y \leqslant Y \leqslant y + dy)$ should equal $Pr(g^{-1}(y) < X < g^{-1}(y) + dg^{-1}(y) \equiv Pr(x \leqslant X \leqslant x + dx)$. This will happen if the following relationship between the areas under the two pdfs holds:

$$|f_Y(y)dy| = |f_X(g^{-1}(y))dg^{-1}(y)|$$

Note that $f_Y(y)dy$ is approximately $Pr(y < Y < y + dy)$. Think of the area of a rectangle, Area=Height $\times$ Width, where Area is probability, Height is the pdf evaluated at y, and Width is dy. Then

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|$$

Biological Allometry Example

This example is based on http://www.biology.arizona.edu/biomath/tutorials/applications/allometry.html:

> Male fiddler crabs (Uca pugnax) possess an enlarged major claw for fighting or threatening other males. In addition, males with larger claws attract more female mates.
>
> The sex appeal (claw size) of a particular species of fiddler crab 4.2 is determined by the following allometric equation:
>
> $$M_c = 0.036 M_b^{1.356}$$
>
> where M_c is the mass of the major claw and M_b is the body mass of the crab minus the mass of the claw.

Suppose that M_b is on average 2000 mg with a CV of 0.20. Assuming a Gamma distribution for M_b, then $M_b \sim$ Gamma(25, 0.0125).

Figure 4.2: Fiddler Crab. Image from Southeastern Regional Taxonomic Center (SERTC), South Carolina Department of Natural Resources.

What is the pdf for M_c? To reduce notation momentarily, let X=M_b, Y=M_c, a=0.036, b=1.856, α=25, and β=0.0125. Then

$$Y = g(X) = aX^b \quad \text{and} \quad X = g^{-1}(Y) = \left(\frac{Y}{a}\right)^{\frac{1}{b}}$$

and

$$\frac{dg^{-1}(y)}{dy} = a^{-1/b}\frac{1}{b}y^{\frac{1-b}{b}}$$

The pdf for Y (M_c):

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right| = \frac{\beta^\alpha}{\Gamma(\alpha)}\left[\left(\frac{Y}{a}\right)^{\frac{1}{b}}\right]^{\alpha-1} e^{-\beta*\left(\frac{Y}{a}\right)^{\frac{1}{b}}} \times a^{-1/b}\frac{1}{b}y^{\frac{1-b}{b}}$$

Substituting the original values:

$$f_{M_c}(M_c) = \frac{0.0125^{25}}{\Gamma(25)}\left[\left(\frac{Y}{0.036}\right)^{\frac{1}{1.856}}\right]^{25-1} e^{-0.125*\left(\frac{Y}{0.036}\right)^{\frac{1}{1.856}}} \frac{1}{0.036^{1/1.856}}\frac{1}{1.856}y^{\frac{1-1.856}{1.856}} \tag{4.24}$$

The accuracy of the derivation was examined by simulating body mass from a Gamma(25,0.0125) and then transforming using the allometric equation (the R code is shown below). The empirical and theoretical pdfs are plotted in Figure 4.3, and the two are quite similar.

```
 #---- Change of variable with Fiddler Crabs ----
body.alpha <- 25
body.beta <- 0.0125
```

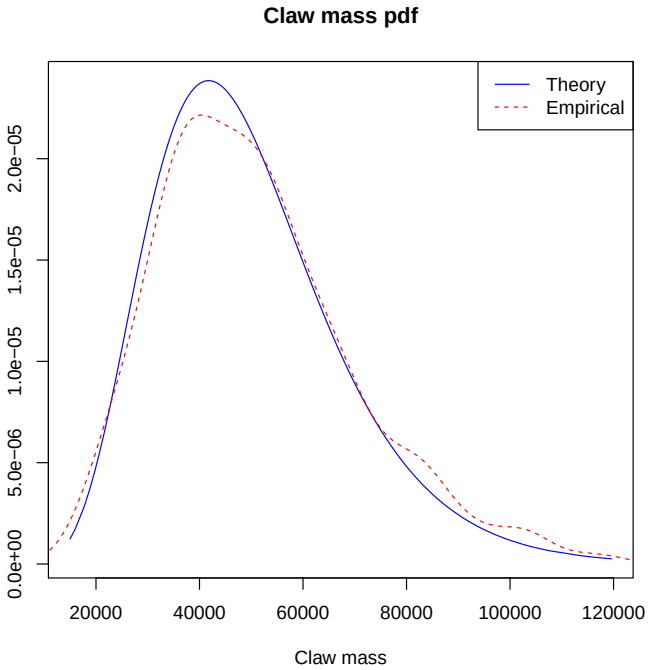


Figure 4.3: Theoretical and empirical pdf for crab claw mass.

```
n <- 500
set.seed(931)
sim.mass <- sort(rgamma(n=n,shape=body.alpha,rate=body.beta))

claw.a <- 0.036
claw.b <- 1.856
sim.claw <- claw.a*sim.mass^claw.b
plot(density(sim.claw))

Mc.density <- function(y,alpha,beta,a,b) {
  x <- (y/a)^(1/b)
  p1 <- dgamma(x,alpha,beta)
  p2 <- a^(-1/b)*(1/b)*y^((1-b)/b)
  out <- p1*p2
}

theory.density <- Mc.density(y=sim.claw,alpha=body.alpha,beta=body.beta,
                             a=claw.a,b=claw.b)

plot(sim.claw,theory.density,xlab="Claw mass",ylab="",main="Claw mass pdf",
     type="l",col="blue")
lines(density(sim.claw),col="red",lty=2)
legend("topright",legend=c("Theory","Empirical"),col=c("blue","red"),lty=1:2)
```

4.6 Supplement B: Fisher Information

In the following we begin with the case of a single (scalar) parameter θ.

Definition of Fisher Information:

$$I(\theta|x) = \mathbb{E}\left[\left(\frac{d \log f(x|\theta)}{d\theta}\right)^2\right] \tag{4.25}$$

Note that under certain regularity conditions[2], Fisher Information can be calculated from the second derivative of the log likelihood:

$$I(\theta|x) = -\mathbb{E}\left[\frac{d^2 \log f(x|\theta)}{d\theta^2}\right] \tag{4.26}$$

which is often much easier to calculate than Eq'n 4.25.

Remarks.

1. Given n iid random variables $x_1, \ldots, x_n$ from the same distribution with parameter θ, the Fisher information for θ = $nI_1(\theta|x)$, where $I_1(\theta|x)$ denotes the information for a single observation:

$$\begin{aligned} I(\theta|x_1, \ldots, x_n) &= -\mathbb{E}\left(\frac{d^2 \log f(x_1, \ldots, x_n|\theta)}{d\theta^2}\right) = -\mathbb{E}\left(\frac{d^2 \sum_{i=1}^n \log f(x_i|\theta)}{d\theta^2}\right) \\ &= \sum_{i=1}^n \left[-\mathbb{E}\left(\frac{d^2 \log f(x_i|\theta)}{d\theta^2}\right)\right] = nI_1(\theta|x) \end{aligned} \tag{4.27}$$

2. Inverse of $I(\theta)$ as lower bound on variance of $\hat{\theta}$. Under the previously mentioned regularity conditions, the inverse of Fisher information is the lower bound on the variance of an unbiased estimator of a parameter. In other words, given a probability distribution with parameter θ which satisfies certain regularity conditions, if $\hat{\theta}$ is unbiased for θ, then

$$V(\hat{\theta}) \geqslant I(\theta)^{-1}$$

   The right hand term is called the Cramer-Rao bound.

   Thus the variance of an unbiased estimate can never be less than the Cramer-Rao bound.

3. Maximum likelihood estimators. In the particular case of maximum likelihood estimates (mles), the inverse of $I(\theta|x)$ evaluated at the mle, $\hat{\theta}$, is often used as an estimate of the variance of $\hat{\theta}$:

$$\widehat{Var}(\hat{\theta}) = I(\theta)^{-1}$$

[2]There are three conditions in this case. (1) For all x such that $f(x|\theta) > 0$, $\frac{d \log(f(x|\theta)}{dx}$ exists and is finite. (2) The order of operations of integration with respect to x and differentiation with respect to θ for the expectation of a function of $T(x)$ can be interchanged, i.e.,

$$\frac{d}{d\theta}\left[\int T(x) f(x|\theta) dx\right] = \int T(x) \frac{df(x|\theta)}{d\theta} dx$$

. (3) The order of operations of integration and differentiation can also be reversed for the second derivative of $f(x|\theta)$ with respect to θ, i.e.,

$$\frac{d^2}{d\theta^2}\left[\int T(x) f(x|\theta) dx\right] = \int T(x) \frac{d^2 f(x|\theta)}{d\theta^2} dx$$

.

Example. if $y \sim \text{Binomial}(n, p)$, then the mle for p is $\hat{p} = \frac{y}{n}$.
The variance of $\hat{p}$ is

$$Var[\hat{p}] = Var\left[\frac{y}{n}\right] = \frac{1}{n^2}Var[y] = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

It can be shown that Fisher information for p is

$$I(p) = \frac{n}{p(1-p)}$$

Observe that $I^{-1}(p) = \frac{p(1-p)}{n}$, which is the variance of $\hat{p}$.

4. Observed Fisher Information is the Fisher information without the integration, i.e., without taking the expectation of the second derivative of the log likelihood:

$$\mathcal{J}(\theta) = \frac{d^2 \log(f(x|\theta))}{d\theta^2} \tag{4.28}$$

and an estimate of θ, namely $\hat{\theta}$, is substituted for θ.

5. Multivariate Θ. Extension of Fisher Information to the case of multiple parameters, $\Theta = (\theta_1, \ldots, \theta_q)$, is similar to much of the above. The differences are that instead of having a single first derivative of $\log(f(x|\theta))$, there is a vector of first derivatives, namely the gradient:

$$\nabla \log(f(x|\Theta)) = \begin{bmatrix} \frac{d \log(f(x)|\theta_1))}{d\theta_1} \\ \frac{d \log(f(x)|\theta_2))}{d\theta_2} \\ \vdots \\ \frac{d \log(f(x)|\theta_2))}{d\theta_q} \end{bmatrix} \tag{4.29}$$

And instead of a single second derivative of $\log(f(x)|\theta)$, there is a matrix of second derivatives, namely the Hessian:

$$\nabla\nabla^T \log(f(x|\Theta)) = \begin{bmatrix} \frac{d^2 \log(f(x)|\Theta))}{d\theta_1^2} & \frac{d^2 \log(f(x)|\Theta))}{d\theta_1 d\theta_2} & \cdots & \frac{d^2 \log(f(x)|\Theta))}{d\theta_1 d\theta_q} \\ \frac{d^2 \log(f(x)|\Theta))}{d\theta_2 d\theta_1} & \frac{d^2 \log(f(x)|\Theta))}{d\theta_1^2} & & \\ \vdots & & & \\ \frac{d^2 \log(f(x)|\Theta))}{d\theta_q d\theta_1} & \frac{d^2 \log(f(x)|\Theta))}{d\theta_q d\theta_2} & \cdots & \frac{d^2 \log(f(x)|\Theta))}{d\theta_q^2} \end{bmatrix} \tag{4.30}$$

4.7 Supplement C: Penalised Complexity Priors

4.7.1 Motivation and basic idea

One of the key features if Jeffreys priors is that they are *parameterisation invariant*, i.e. the observation, or effect, model becomes the same regardless of which internal parameter choice is used when constructing the prior. The downside of the Jeffreys priors is that they typically involve improper densities. The idea behind the Penalised Complexity Prior method (Simpson et al, 2017[3]) is to instead consider how the observation or effect model deviates from some *base model*, and use a measure of that deviation to define a prior distribution for the parameters, that will be parameterisation invariant, by construction. This approach can be helpful in hierarchical models, where analysing the whole model can be challenging, but each model building block is simple and interpretable. An example of this is basic linear regression models with random effect components, where the base model might be that the random effects have variance zero, i.e. aren't needed at all, and the more complex version of the model has some strictly positive variance.

4.7.2 Penalised Complexity Priors

Let θ be the parameter vector of interest, and let y be an observation or effect model, with pdf or pmf $f(y|\theta)$. For simplicitly, we will assume that the base model can be written as a special case of the general model for a specific parameter values $\theta = \theta_0$, so that the base model has effect pdf/pmf $f(y|\theta_0)$. Let $\mathrm{dist}[f(y|\theta_0), f(y|\theta)]$ be a measure of the complexity of the model for y. A Penalised Complexity Prior for θ is then defined implicitly as the distribution that induces an Exponential distribution on the complexicy measure, $\mathrm{dist}[f(y|\theta_0), f(y|\theta)] \sim \mathrm{Exp}(\lambda)$, for some scale parameter λ. The scale parameter λ can be chosen to reflect the prior belief in the complexity, which is typically practical after converting the prior to a distribution statement for an appropriate parameterisation. In the original paper, the Kullback-Leibler divergence was used as the complexity measure, with

$$\begin{aligned}
\mathrm{dist}[f(y|\theta_0), f(y|\theta)] &= \sqrt{2 D_{KL}[f(y|\theta] \| f(y|\theta_0)]}, \\
D_{KL}[f(y|\theta] \| f(y|\theta_0)] &= \int f(y|\theta) \log\left[\frac{f(y|\theta)}{f(y|\theta_0)}\right] dy,
\end{aligned}$$

but other measures can be used as well. One benefit of the Kullback-Leibler divergence is that it is invariant also under transformations of y, so that the prior is invariant not only under reparameterisations of θ, but also under transformations of y, so that e.g. units of measurements of y do not affect the prior for θ directly.

4.7.3 Example: Random Normal random effects with unknown mean or variance

Consider the effect model $y_i = \mu + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$. We wish to construct penalised complexity priors for μ and σ. The Kullback-Leibler divergence between the a base model with given $\mu = \mu_0$ and $\sigma = \sigma_0$ and the general model is

$$D_{KL}[f(y|\theta] \| f(y|\theta_0)] = \log\left(\frac{\sigma_0}{\sigma}\right) + \frac{\sigma^2 + (\mu - \mu_0)^2}{2\sigma_0^2} - \frac{1}{2},$$

[3]Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G. Martins, Sigrunn H. Sørbye, "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors." Statistical Science 32(1): 1-28 (February 2017). DOI: 10.1214/16-STS576, https://projecteuclid.org/journals/statistical-science/volume-32/issue-1/Penalising-Model-Component-Complexity--A-Principled-Practical-Approach-to/10.1214/16-STS576.full

where $(y|\theta) \sim \mathrm{N}(y|\mu, \sigma^2)$ and $(y|\theta_0) \sim \mathrm{N}(y|\mu_0, \sigma_0^2)$.

First, consider the a base model where $\mu = \mu_0$ for some $\mu_0 \in \mathbb{R}$ and known $\sigma = \sigma_0$. The complexity measure is then

$$z_\mu(\mu) := \mathrm{dist}[f(y|\theta_0), f(y|\theta)] = \frac{|\mu - \mu_0|}{\sigma_0},$$

on which we impose a $\mathrm{Exp}(\lambda_\mu)$ distribution, for some λ_μ value. Then, for every $\nu > 0$,

$$\begin{aligned}
\Pr\left(\frac{|\mu - \mu_0|}{\sigma_0} \leqslant \nu\right) &= 1 - \exp(-\lambda_\mu \nu), \\
\Pr\left(-\sigma_0\nu \leqslant \mu - \mu_0 \leqslant \sigma_0\nu\right) &= 1 - \exp(-\lambda_\mu \nu), \\
\Pr\left(\mu_0 - u \leqslant \mu \leqslant \mu_0 + u\right) &= 1 - \exp(-\lambda_\mu u/\sigma_0),
\end{aligned}$$

where $u = \sigma_0\nu$. Due to symmetry around μ_0, and noting that the expression is the difference between two CDF values, we obtain the density function

$$p(\mu) = \frac{\lambda_\mu}{2\sigma_0} \exp\left(-\lambda_\mu |\mu - \mu_0|/\sigma_0\right), \qquad \mu \in \mathbb{R},$$

which is the density of a *Laplace distribution* (also known as a *double exponential* distribution) for μ with location μ_0 and scale σ_0/λ_μ, or $\mu \sim \mathrm{Laplace}(\mu_0, \sigma_0/\lambda_\mu)$.

Note: If our prior knowledge about μ is not relative to the σ_0 value, we may introduce a new parameter $\tilde{\lambda}_\mu = \lambda_\mu/\sigma_0$, and use that when eliciting the prior, with $\mu \sim \mathrm{Laplace}(\mu_0, 1/\tilde{\lambda}_\mu)$.

Next, consider a sequence of base models where $\sigma_0 \to 0$, i.e. the random effect is not needed. The complexity measure for models with fixed $\mu = \mu_0$ and $\sigma > \sigma_0$ is then

$$\begin{aligned}
z_\sigma(\sigma) &= \sqrt{2\log\left(\frac{\sigma_0}{\sigma}\right) + \frac{\sigma^2}{\sigma_0^2} - 1} \\
&= \frac{\sigma}{\sigma_0}\sqrt{\frac{\sigma_0^2}{\sigma^2}\log\left(\frac{\sigma_0^2}{\sigma^2}\right) + 1 - \frac{\sigma_0^2}{\sigma^2}},
\end{aligned}$$

which is strictly increasing for $\sigma > \sigma_0$. Scaling by σ_0 and taking the limit $\sigma_0 \to 0$, we obtain

$$\lim_{\sigma_0 \to 0} \sigma_0 z_\sigma(\sigma) = \sigma,$$

on which we impose a $\mathrm{Exp}(\lambda_\sigma)$ distribution, for some λ_σ value.

Note: This kind of limit renormalisation is typically required when the desired base model and the more complex models are mutually *singular*, in that the base model cannot generate all possible outcomes that are possible in the more complex models.

In practice, one would typically choose a Normal prior with a large variance for μ and use the penalised complexity prior only for the standard deviation σ, with λ_σ chosen to reflect the prior belief in the complexity of the model. For example, to impose a prior that σ is unlikely to be larger than some σ_{large}, we could do

$$\begin{aligned}
\Pr(\sigma > \sigma_{\text{large}}) &= 0.01, \\
\exp(-\lambda_\sigma \sigma_{\text{large}}) &= 0.01, \\
\lambda_\sigma &= -\log(0.01)/\sigma_{\text{large}}.
\end{aligned}$$

Week 4 Homework.

Reading

1. (Optional) Read Section 5.3 of *Applied Bayesian Statistics With R and OpenBUGS Examples* by Cowles, particularly focusing on 5.3.3 on Jeffreys prior.
2. (Optional) Read Section 2.3 on Objectives Priors in Chapter 2 of *Bayesian Statistical Methods* by Reich and Ghosh.

Exercises: not to turn in

1. (From Cowles, exercise 5.3.) The prior for a binomial parameter, θ, is defined as

$$\theta \sim \text{Uniform}(0, 1)$$

   A new parameter ϕ is defined as a logit transformation of θ:

$$\phi = \log\left(\frac{\theta}{1-\theta}\right)$$

   Verify that the induced density for ϕ is

$$\pi(\phi) = \frac{\exp(\phi)}{(1+\exp(\phi))^2}, \quad -\infty < \phi < \infty$$

2. The sampling distribution is Poisson(θ).
   (a) Verify that the Jeffreys' prior for θ is $\pi_{\text{JP}}(\theta) \propto 1/\sqrt{\theta}$.
   (b) Reparameterise the Poisson with $\phi = \sqrt{\theta}$. Use the change of variable theorem to show that the induced prior for ϕ is proportional to a constant, in particular proportional to 2.
   (c) Calculate the Fisher information for ϕ from the following Poisson pmf:

$$\frac{e^{-\phi^2}(\phi^2)^y}{y!}$$

   Then show that Jeffreys' prior, $\pi_{\text{JP}}(\phi)$ is proportional to a constant, again 2.

5 Summarising Posteriors & Hypothesis Testing

5.1 Summarising Posteriors

Overview

A complete summary of a posterior is the posterior distribution. In the case of a single parameter, θ, simply drawing a picture of the pmf (a histogram) or the pdf (a curve) "shows everything" about the parameter. In the case of two parameters, Θ=(θ_1, θ_2), a three dimensional (perspective) plot or contour plot can be drawn. Given that the θs often have some degree of dependency in the joint posterior distribution, e.g., $p(\theta_1, \theta_2|y) \neq p(\theta_1|y)p(\theta_2|y)$, one would want to look at the joint distribution, by looking at contour plots, for example. However, in the case of dozens or even 100s of parameters, sorting through dozens or 100s of histograms or density plots may be burdensome.

Thus while graphical summaries of the posterior distributions are indeed good statistical practice, there is utility in simpler numerical summaries, such as simple point estimates, like posterior mean, measures of spread like posterior standard deviations or interval estimates.

5.1.1 Point estimates

Point estimates are single number summaries, such as the mean, or median, or mode. As it turns out, each of these summaries, in the context of *Decision Theory*, can be viewed as optimal estimators for particular types of *Risk*, where risk is the expected value of a *Loss Function*.

Components of Decision Theory with emphasis on parameter estimation

Decision theory includes parameter estimation, hypothesis testing, and prediction: here we emphasize parameter estimation. We note below that the notation for the loss function is somewhat atypical.

1. *State Space, Θ:* There is a true state of nature, θ, which is a member of a set of states, what is called the *State Space* and will be denoted Θ.

   For example, you sit down next to a person on a bus, and suppose that that person is either infected with Covid 19 or not infected. Then the State Space is Θ=(Infected, Not Infected).

   For our discussion here, the State Space will be the possible values for an unknown parameter. For example, a Bernoulli parameter θ is a point on the [0,1] line segment. Here Θ = [0,1], and $\theta \in \Theta$. Furthermore, we will be imagining a probability distribution for the state space. For example, Pr(Infected)=0.05 and Pr(Not Infected)=0.95.

   Note: The *State Space* can also be a set of hypotheses; e.g., $\Theta = (H_1, H_2, H_3)$.

2. *Action Space, $\mathcal{A}$:* The decision maker will choose a possible action a from a set of possible actions, the *Action Space*, which is denoted $\mathcal{A}$; thus $a \in \mathcal{A}$. The action might be viewed as a "decision".

   Referring to the above example of sitting down next to someone on a bus, the action space could be $\mathcal{A}$ = (Put on a Mask, Don't Put on a Mask).

   For our discussion here the action space, $\mathcal{A}$, is the same as the state space, Θ. The action is to choose a value from Θ as a parameter estimate; e.g., $a \equiv \hat{\theta}$.

   Similarly, for hypothesis testing, the action space $\mathcal{A}$ could be selecting one of the hypotheses (H_1, H_2, H_3).

3. *Loss function, $\mathcal{L}(a|\theta)$:* Given a specific state, θ, and an action, a, there is a loss or "cost" for any action or decision. This loss is denoted $\mathcal{L}(a|\theta)$[1] and typically this loss is a numerical value. Here we assume that one wants to minimise loss.

   Continuing the bus, Covid, mask-wearing example, a hypothetical loss matrix is the following.

	State Space (Θ)	
Action Space ($\mathcal{A}$)	Infected	Not Infected
Wear Mask	4	1
Don't Wear Mask	10	2

   The loss values in this example are completely subjective, and will vary between individuals.

   Again the context here will be a loss associated with a true parameter value (the state) and a parameter estimate (the action): $\mathcal{L}(\hat{\theta}|\theta)$.

4. *Risk, $R(a)$:* The Risk of action a, $R(a)$, is the expected loss where the expectation is taken over the state space, Θ.

$$R(a) = E_\Theta[L(a|\theta)] = \int L(a|\theta)\pi(\theta)d\theta \tag{5.1}$$

   or if Θ is a finite space with K values:

$$R(a) = E_\Theta[L(a|\theta)] = \sum_{k=1}^{K} L(a|\theta_k)\pi(\theta_k) \tag{5.2}$$

   Continuing with the bus example, using the previously mentioned probabilities and loss function, letting WM be "wear mask", NoM be not wear mask, I be infected, and NI be not infected.

$$R(WM) = L(WM|I)\Pr(I) + L(WM|NI)\Pr(NI) = 4 * 0.05 + 1 * 0.95 = 1.15$$
$$R(NoM) = L(NoM|I)\Pr(I) + l(NoM|NI)\Pr(NI) = 10 * 0.05 + 2 * 0.95 = 2.40$$

   If one was to choose the action with minimal risk, then one would choose to wear a mask. Note how this choice of action very much depends on both the probabilities for the states and the loss function values. A different set of probabilities and different loss values could lead to not wearing a mask as the choice.

5. *Adjustments based on data, i.e., learning.* Estimates of risk will (usually) change after data, y, have been collected that provide information about the states. A "parametric" sampling distribution for y makes explicit the connection between data and the state space, $f(y|\theta)$. The probability distribution for the state space is essentially a prior distribution, and after collecting data, the posterior distribution is used for calculating risk:

$$R(a|y) = E_\Theta L(a|\Theta, y) = \int L(a|\theta, y)p(\theta|y)d\theta \tag{5.3}$$
$$= \int L(a|\theta)\frac{\pi(\theta)f(y|\theta)}{m(y)}d\theta \tag{5.4}$$

[1] This is not typical notation, $l(\theta, a)$ is more common but that does not make clear that θ is fixed and a is the variable.

where the last line follows from assuming that loss for a given action and state is independent of the data.

when Θ is a parameter space and the action a is to estimate a parameter, the posterior risk is

$$R(\hat{\theta}|\mathbf{y}) = \int L(\hat{\theta}|\theta)p(\theta|\mathbf{y})d\theta \tag{5.5}$$

6. *Bayes estimator of a parameter:* The *Bayes Estimator* of θ is *defined* as the value of θ that minimizes *Posterior Risk*:

$$\text{Bayes Estimator: } \hat{\theta}_{BE} = \texttt{argmin}_{\hat{\theta}\in\Theta} R_{\Theta}(\hat{\theta}(\mathbf{y})|\mathbf{y}) \tag{5.6}$$

For clarity the `argmin` operation yields the value of $\hat{\theta}$ with the smallest risk. Note that the subscript BE is *not* standard notation. The dependence of $\hat{\theta}$ on $\mathbf{y}$ will be implicit in the following.

Comments.

- Determining what loss function to use is not necessarily simple, and can be arrived at via "elicitation", much like priors.
- Finding the Bayes Estimator is equivalent to finding the action, $a \equiv \hat{\theta}$, that minimises Risk. In the case of continuous Θ, the estimate is sometimes found by calculus, setting the derivative of risk with respect $\hat{\theta}$ equal to zero and solving for $\hat{\theta}$.
- See Reich & Ghosh (2019, pp 218-220) for a description of Decision Theory, focusing on applications to point estimation, hypothesis testing, and prediction.

Example distinguishing loss and risk. Suppose that θ is restricted to a finite interval, $L \leqslant \theta \leqslant U$, e.g., [2,8], and suppose that the true value of θ is 5. Figure 5.1 shows plots of $\mathcal{L}(\hat{\theta}|\theta = 5)$ for four different loss functions (to be discussed shortly) against varying values of $\hat{\theta}$. Note that the loss is minimized when $\hat{\theta}$=5, namely θ, for all four loss functions.

In practice one will not know the true value of θ. Uncertainty about the value of θ is reflected by its posterior distribution $p(\theta|\mathbf{y})$. This leads to the risk function calculation. Suppose that the posterior distribution for θ is right-triangular on [2,8] (see Figure 5.2):

$$p(\theta|\mathbf{y}) = -\frac{1}{9} + \frac{1}{18}\theta, \;\; 2 \leqslant \theta \leqslant 8$$

The risk function for the different lost functions and the triangle pdf is calculated using eq'n 5.3. Plots of the risk versus $\hat{\theta}$ for four loss functions are shown in Figure 5.3 and the corresponding Bayes Estimators are indicated on the plot. For R Code for Loss Function Plots and Bayes Estimators See L5_R_Code_Loss_Functions.html on Learn.

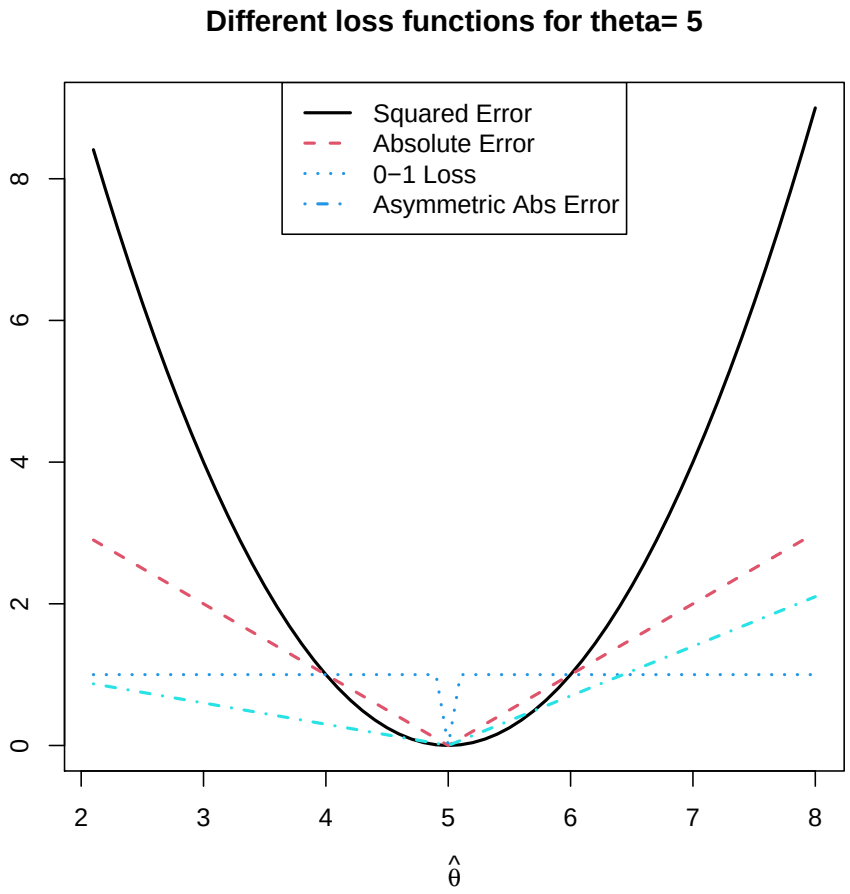


Figure 5.1: Different loss function values for $\hat{\theta}$ when Θ = [0,8] and true value of θ is 5.

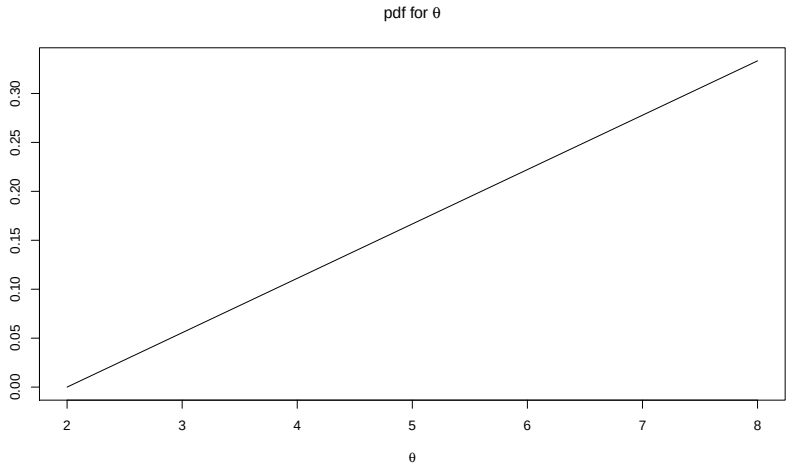


Figure 5.2: Triangle pdf for θ: $-\frac{1}{9} + \frac{1}{18}\theta$,

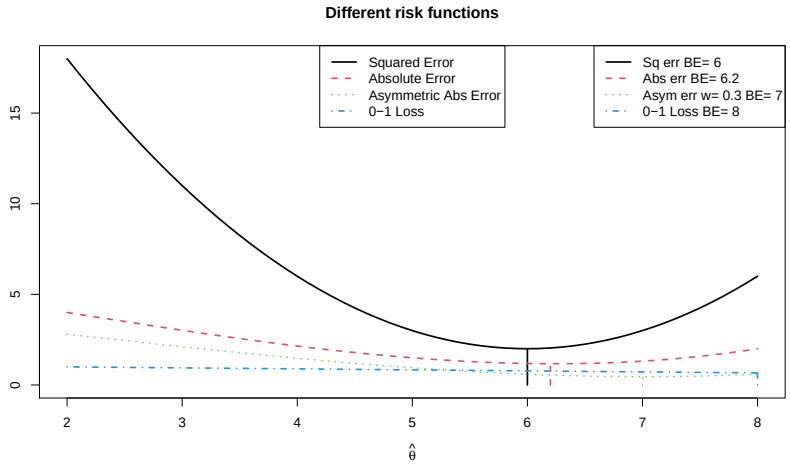


Figure 5.3: Risk function for triangle pdf (on [2,8]) for four loss functions with corresponding θ_{BE} indicated.

Common loss functions for estimators

Squared error loss

Squared error loss = $\mathcal{L}(\hat{\theta}|\theta)=(\theta-\hat{\theta})^2$. (The notation for $\mathbf{y}$ is omitted.) Here we consider the continuous case. The Bayes Estimator, $\hat{\theta}$, for minimizing the Risk with squared error loss is the *posterior mean*, $E[\theta|\mathbf{y}]$.

Proof:

$$\begin{aligned}\frac{d}{d\hat{\theta}}R(\hat{\theta}) = \frac{d}{d\hat{\theta}}E_{\Theta}\left[(\theta-\hat{\theta})^2|\mathbf{y}\right] &= \frac{d}{d\hat{\theta}}\int(\theta-\hat{\theta})^2 p(\theta|\mathbf{y})d\theta \\ &= \int\frac{d}{d\hat{\theta}}(\theta-\hat{\theta})^2 p(\theta|\mathbf{y})d\theta = \int -2(\theta-\hat{\theta})p(\theta|\mathbf{y})d\theta\end{aligned}$$

Setting the above equal to 0 and solving for $\hat{\theta}$:

$$\begin{aligned}\int(\theta-\hat{\theta})p(\theta|\mathbf{y})d\theta = 0 \quad &\Rightarrow \quad \int\theta p(\theta|\mathbf{y})d\theta = \hat{\theta}\int p(\theta|\mathbf{y})d\theta \\ &\Rightarrow \hat{\theta} = E(\theta|\mathbf{y}) \qquad (5.7)\end{aligned}$$

Thus $E(\theta|\mathbf{y})$ is a critical point. To check that it is a minimum, take the 2nd derivative:

$$\frac{d}{d\hat{\theta}}\int -2(\theta-\hat{\theta})p(\theta|\mathbf{y})d\theta = 2\int p(\theta|\mathbf{y})d\theta = 2 > 0$$

Thus $E(\theta|\mathbf{y})$ is a minimum.

Absolute error loss

Absolute error loss= $\mathcal{L}(\hat{\theta}|\theta)=|\theta-\hat{\theta}|$. Here again we consider the continuous case. The Bayes Estimator, $\hat{\theta}$, for minimizing expected absolute error loss is the *posterior median*, denoted $\theta_{0.5}$.

Proof:

First re-express the Risk as follows:

$$\begin{aligned}R(\hat{\theta}) = E_{\Theta}\left[|\theta-\hat{\theta}||\mathbf{y}\right] &= \int_{-\infty}^{\infty}|\theta-\hat{\theta}|p(\theta|\mathbf{y})d\theta \\ &= \int_{-\infty}^{\hat{\theta}}(\hat{\theta}-\theta)p(\theta|\mathbf{y})d\theta + \int_{\hat{\theta}}^{\infty}(\theta-\hat{\theta})p(\theta|\mathbf{y})d\theta \\ &= \hat{\theta}\int_{-\infty}^{\hat{\theta}}p(\theta|\mathbf{y})d\theta - \int_{-\infty}^{\hat{\theta}}\theta p(\theta|\mathbf{y})d\theta + \int_{\hat{\theta}}^{\infty}\theta p(\theta|\mathbf{y})d\theta - \hat{\theta}\int_{\hat{\theta}}^{\infty}p(\theta|\mathbf{y})d\theta \\ &= \hat{\theta}F_{\theta|\mathbf{y}}(\hat{\theta}) - \int_{-\infty}^{\hat{\theta}}\theta p(\theta|\mathbf{y})d\theta + \int_{\hat{\theta}}^{\infty}\theta p(\theta|\mathbf{y})d\theta - \hat{\theta}(1-F_{\theta|\mathbf{y}}(\hat{\theta})) \\ &= \hat{\theta}2F_{\theta|\mathbf{y}}(\hat{\theta}) - \hat{\theta} - \int_{-\infty}^{\hat{\theta}}\theta p(\theta|\mathbf{y})d\theta + \int_{\hat{\theta}}^{\infty}\theta p(\theta|\mathbf{y})d\theta\end{aligned}$$

where $F_{\theta|\mathbf{y}}(\theta))$ is the cumulative posterior distribution function for $p(\theta|\mathbf{y})$.

Now differentiate with respect to $\hat{\theta}$[2]:

$$\frac{d}{d\hat{\theta}}E_{\Theta}\left[|\theta-\hat{\theta}||\mathbf{y}\right] = 2F_{\theta|\mathbf{y}}(\hat{\theta}) + 2\hat{\theta}p(\hat{\theta}) - 1 - \hat{\theta}p(\hat{\theta}) - \hat{\theta}p(\hat{\theta}) = 2F_{\theta|\mathbf{y}}(\hat{\theta}) - 1$$

[2]Recall the Fundamental theorem of calculus: $\frac{d}{dx}\int_c^x f(t)dt$ = $f(x)$, and similarly $\frac{d}{dx}\int_x^c f(t)dt$ = -$f(x)$.

Setting the above equal to 0,

$$2F_{\theta|\mathbf{y}}(\hat{\theta}) = 1 \quad \Rightarrow \quad F_{\theta|\mathbf{y}}(\hat{\theta}) = 1/2 \tag{5.8}$$

Thus $\hat{\theta}$=50th percentile, $\theta_{0.5}$, is a critical value. To check that it is a minimum, take the 2nd derivative:

$$\frac{d}{d\hat{\theta}}\left[2F_{\theta|\mathbf{y}}(\hat{\theta}) - 1\right] = 2p(\hat{\theta})$$

where $p(\hat{\theta})$ is the probability density function which is greater than or equal to 0. If greater than 0, then a=median is a minimum. If equal to 0 then need to check further (we will not worry about this).

Exercise. Show that the Bayes estimator for the following loss function:

$$\mathcal{L}(\hat{\theta}|\theta) = \begin{cases} (1-\tau)(\hat{\theta} - \theta) & \text{if } \hat{\theta} > \theta \\ \tau(\theta - \hat{\theta}) & \text{if } \hat{\theta} < \theta \end{cases}$$

where $0 < \tau < 1$, is θ_τ, the τ^{th} quantile.

0-1 Loss

Fo 0-1 loss we treat the case of discrete parameters separate from continuous parameters. For discrete valued parameters, we define 0-1 loss as

$$\mathcal{L}(\hat{\theta}|\theta) = \begin{cases} 1, & \text{if } \hat{\theta} \neq \theta. \\ 0, & \text{if } \hat{\theta} = \theta. \end{cases}$$

i.e. $\mathcal{L}(\hat{\theta}|\theta) = \mathbb{I}(\theta \neq \hat{\theta})$. We want to minimize

$$R(\hat{\theta}) = E_\Theta\left[\mathcal{L}(\hat{\theta}|\theta)\right] = \sum_{\theta \in \Theta} I(\theta \neq \hat{\theta}) p(\theta|\mathbf{y}). \tag{5.9}$$

Before giving the general result, suppose there are 3 values of θ: θ_1, θ_2, and θ_3. Thus $\hat{\theta}$ can be chosen to be one of these 3 values. The expected loss (risk) for each choice is then:

$$\begin{aligned}
E_\Theta\left[\mathcal{L}(\hat{\theta} = \theta_1|\theta))\right] &= 0 * p(\theta_1|\mathbf{y}) + 1 * p(\theta_2|\mathbf{y}) + 1 * p(\theta_3|\mathbf{y}) = p(\theta_2|\mathbf{y}) + p(\theta_3|\mathbf{y}) = 1 - p(\theta_1|\mathbf{y}) \\
E_\Theta\left[\mathcal{L}(\hat{\theta} = \theta_2|\theta))\right] &= 1 * p(\theta_1|\mathbf{y}) + 0 * p(\theta_2|\mathbf{y}) + 1 * p(\theta_3|\mathbf{y}) = p(\theta_1|\mathbf{y}) + p(\theta_3|\mathbf{y}) = 1 - p(\theta_2|\mathbf{y}) \\
E_\Theta\left[\mathcal{L}(\hat{\theta} = \theta_3|\theta))\right] &= 1 * p(\theta_1|\mathbf{y}) + 1 * p(\theta_2|\mathbf{y}) + 0 * p(\theta_3|\mathbf{y}) = p(\theta_1|\mathbf{y}) + p(\theta_2|\mathbf{y}) = 1 - p(\theta_3|\mathbf{y})
\end{aligned}$$

The value of $\hat{\theta}$ minimizing expected loss is the θ_i that has the largest probability, the most likely value or the *posterior mode*[3].

The more general solution given q possible values for θ. Let Θ denote a finite set of possible parameters. The risk is

$$R(\hat{\theta}) = E_\Theta\left[\mathcal{L}(\hat{\theta}|\theta)\right] = \sum_{\theta \in \Theta} I(\theta \neq \hat{\theta}) p(\theta|\mathbf{y}) = 1 - p(\hat{\theta}|\mathbf{y}) \tag{5.10}$$

[3]Note the Bayes estimate would not be unique if there were two or more modes that were the largest and of equal value.

Thus the risk is minimized by $\hat{\theta}$ equal to θ with the largest probability, namely the mode.

For continuous parameters, one possible definition of 0-1 loss is $\mathcal{L}(\hat{\theta}|\theta) = -\delta_\theta(\hat{\theta})$, where $\delta_\theta(\cdot)$ is a Dirac Delta function. This can be interpreted as an infinite negative loss (i.e. a gain) at the true value, and zero loss elsewhere. We then obtain

$$R(\hat{\theta}) = E_\Theta\left[\mathcal{L}(\hat{\theta}|\theta)\right] = -\int_\Theta \delta_\theta(\hat{\theta})p(\theta|\mathbf{y})d\theta = -p_{\theta|\mathbf{y}}(\hat{\theta}|\mathbf{y}), \tag{5.11}$$

which is minimised for the mode[4] of the posterior density, $\hat{\theta} = \text{argmax}_{\theta\in\Theta}\, p(\theta|\mathbf{y})$.

Examples

Normal μ with Normal prior. Let θ be the average amount students in Edinburgh will spend on entertainment this month. In this case the sample space is a range of positive numbers; e.g., Θ = [£1, £1000]. A random sample of n students will be taken next month and the data are the amounts spent by each sampled student, $\mathbf{y} = y_1, \ldots, y_n$. The "action" or decision is to estimate θ based on the data, a=$\hat{\theta}(\mathbf{y})$. The sampling distribution of the y_i is assumed Normal(θ,50^2). The prior for θ is Normal(30, $\frac{50^2}{5}$)[5]. A random sample of n=100 students was selected and the average spent was £42. The posterior mean[6] is the Bayes Estimator for θ, namely, 41.43.

Because the posterior distribution is normal, the median equals the mean, thus the Bayes estimator of absolute loss is again £41.43. Similarly, the posterior distribution is unimodal with the mode equal to the posterior mean, and the Bayes estimator for 0-1 loss is the same.

Binomial θ with a Beta prior. Given a Beta(α, β) prior for a Binomial(n, θ) sampling distribution, the posterior distribution is Beta($\alpha + y$, $\beta + n - y$). Suppose $\alpha = 2$, $\beta = 3$, n=10, and $y = 4$. Then the posterior is Beta(6,9). The Bayes estimators for the loss functions given above:

$$\begin{aligned}
\text{Quadratic loss (posterior mean): } \hat{\theta} &= \frac{\alpha + y}{\alpha + \beta + n} = \frac{2+4}{2+3+10} = 0.4 \\
\text{Absolute error loss (posterior median): } \hat{\theta} &= \texttt{qbeta}(0.5, \texttt{shape1} = \alpha + y, \texttt{shape2} = \beta + n - y) \\
&= \texttt{qbeta}(0.5, 6, 9) = 0.395 \\
\text{0-1 loss (posterior mode): } \hat{\theta} &= \frac{\alpha + y - 1}{\alpha + \beta + n - 2} = \frac{2+4-1}{2+3+10-2} = 0.385
\end{aligned}$$

The three values are shown in Figure 5.4.

Poisson θ with a Gamma prior. As shown previously, given a Gamma(α, β) prior for a Poisson(θ) sampling distribution, the posterior distribution (given n=1) is Gamma($\alpha + y$, $\beta + 1$). Suppose $\alpha = 5$, $\beta = 3$, and $y = 3$. The posterior for θ is Gamma(8,4). The Bayes estimators for the loss functions given above:

$$\begin{aligned}
\text{Quadratic loss: } \hat{\theta} &= \frac{\alpha + y}{\beta + 1} = \frac{5+3}{3+1} = 2 \\
\text{Absolute error loss: } \hat{\theta} &= \text{qgamma(0.5,shape1=}\alpha + y\text{, shape2=}\beta + 1) = \texttt{qgamma}(0.5, 8, 4) = 1.917 \\
\text{0-1 loss: } \hat{\theta} &= \frac{\alpha + y - 1}{\beta + 1} = \frac{5+3-1}{3+1} = 1.75
\end{aligned}$$

Values are shown in Figure 5.5.

[4] Just as in the discrete case, this isn't necessarily unique.

[5] This is not entirely realistic as θ must be non-negative.

[6] The posterior distribution is Normal$\left(\frac{5*30+100*42}{5+100}, \frac{50^2}{5+100}\right)$ or Normal(41.43, 7.29^2).

Figure 5.4: Posterior distribution for Binomial θ, Beta(6,9), with posterior mean, median, and mode.

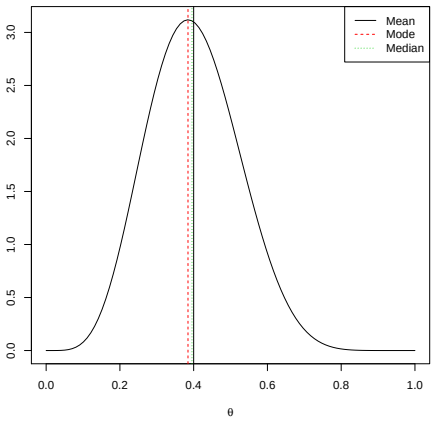
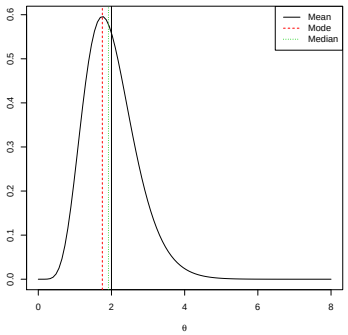


Figure 5.5: Posterior distribution for Poisson θ, Gamma(8,4), with posterior mean, median, and mode.



5.1.2 Interval estimates

Analogous to frequentist *Confidence Intervals* are Bayesian *Credible Intervals* but the latter are more easily interpreted. To begin we just consider the situation of a single parameter, θ, and then discuss intervals, "regions", for two or more parameters.

For a given probability P where P is often expressed as $1-\alpha$, where $0 < \alpha < 1$, a $\mathrm{P}\%$ Bayesian Credible interval is *defined* as an interval $[\mathrm{LB}, \mathrm{UB}]$ where

$$\int_{\mathrm{LB}}^{\mathrm{UB}} \mathrm{p}(\theta|\mathbf{y})\mathrm{d}\theta = \mathrm{P} \tag{5.12}$$

Suppose α equals 0.05, then P=1-0.05=0.95. Then a 95% credible interval are those values LB and UB such that

$$\int_{\mathrm{LB}}^{\mathrm{UB}} \mathrm{p}(\theta|\mathbf{y})\mathrm{d}\theta = 0.95$$

Similar to one sided-confidence bounds, one can define lower credible bounds, $[\mathrm{LB}, \infty]$, and upper credible bounds, $[-\infty, \mathrm{UB}]$.

Example: Binomial θ with Beta prior. Suppose the prior for θ is Beta(2,3) and for n=10 Bernoulli trials the observed number of successes is y=4. Then the posterior for θ is Beta(6,9).

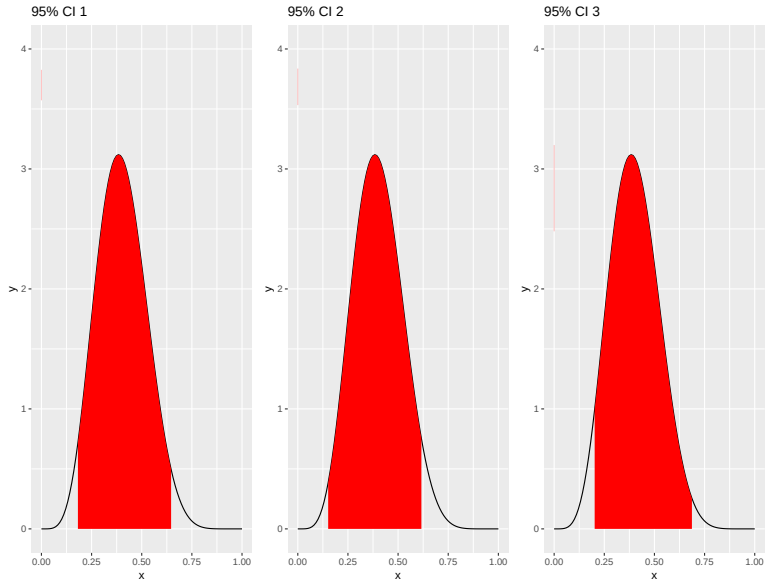
A 95% credible interval will be any combination of LB and UB such that:

$$\int_{\mathrm{LB}}^{\mathrm{UB}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = \int_{\mathrm{LB}}^{\mathrm{UB}} \frac{\Gamma(6+9)}{\Gamma(6)\Gamma(9)}\theta^{6-1}(1-\theta)^{9-1}d\theta = 0.95$$

Note that these bounds are not unique. For example, the following three intervals are all 95% credible intervals (See Figure 5.6).

$$[0.177, 0.649] \quad [0.146, 0.623] \quad [0.196, 0.692]$$

Figure 5.6: Three 95% credible intervals for Binomial θ, with posterior Beta(6,9).



Bimodal posteriors. If the posterior distribution is bimodal, then a credible interval composed of two line segments may be sensible. (Draw a picture.)

Symmetric credible intervals

A less arbitrary approach to constructing credible intervals is to use symmetric intervals where the probability 1-P, or α, is divided evenly in the tails:

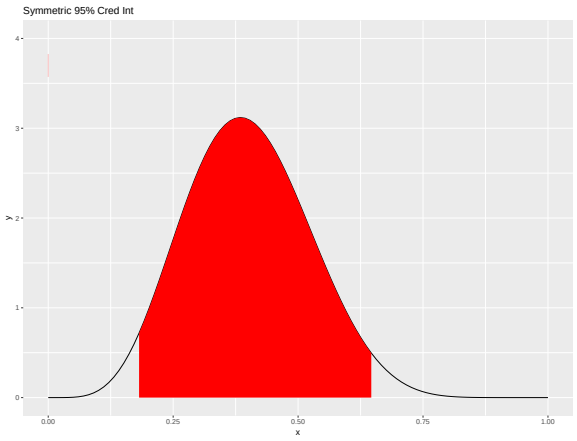
$$\int_{\mathrm{LB}_{sym}}^{\mathrm{UB}_{sym}} p(\theta|\mathbf{y})d\theta = P \tag{5.13}$$

where $\Pr(\theta \leqslant \mathrm{LB}_{sym}) = \Pr(\theta \geqslant \mathrm{UB}_{sym}) = \alpha/2$.

- One advantage of the symmetric approach is that for monotonic transformations of θ, $g(\theta)$, the 1-α symmetric interval for $g(\theta)$ is $[g(LB), g(UB)]$ where $[LB, UB]$ are the symmetric bounds for θ.

Referring to the previous Beta example: a 95% symmetric credible interval is [0.177, 0.649], and is shown in Figure 5.7.

Figure 5.7: Symmetric 95% credible interval for Binomial θ, with posterior Beta(6,9).



Highest Posterior Density Intervals

Another non-arbitrary approach is Highest Posterior Density Intervals (HPDIs). These are 1-α intervals where the densities (or probabilities) for values in the interval are higher than for values not in the interval. Another way to say this: "it is the interval with the shortest interval width, U-L, while maintaining appropriate coverage" (Reich and Ghosh, p 26)[7]. Formally: The interval $[LB, UB]$ is the 1-α HPDI if

1. $[LB, UB]$ is a 1-α credible interval.
2. For all $\theta' \in [LB, UB]$ and all $\theta'' \notin [LB, UB]$, $p(\theta'|y) \geqslant p(\theta''|y)$.

- This will yield the shortest 1-α credible interval.
- If the posterior is unimodal and symmetric, the symmetric and HPDI are the same.
- If the posterior is multimodal, the HPDI may consist of two or more line segments.
- In contrast to symmetric intervals, HPDI's are not invariant to monotonic transformations.

Software for HPDI calculations. The `R` package `HDInterval` has a function `hdi` that can be used to calculate HPDIs for a variety of situations. Below are two examples. R Code used to produce plots is on Learn in R Code Credible Intervals.

Demonstration with Binomial(n, θ). A Beta(0.5,0.5) prior (Jeffreys') was selected for the binomial parameter θ. For n=5 trials there were y=4 successes. Thus the posterior is Beta(4.5, 1.5). The 99% symmetric and HPDI CIs are shown below and drawn in Figure 5.8.

```
Symmetric  0.256037 0.9924612 length= 0.736424
HPD        0.300140 0.9998463 length= 0.699706
```

The symmetric CIs are indeed longer than the HPDIs are nearly identical.

[7] Although bimodal distributions may have two non-overlapping intervals that are shorter in total length than a single interval.

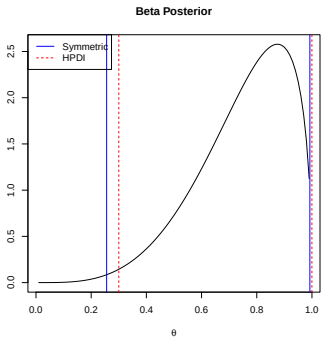


Figure 5.8: Example of 99% CIs for binomial parameter using symmetric and HPDI approaches.

Demonstration with Poisson. A Gamma(25, 2.5) prior was chosen for the Poisson parameter (EV=10, CV=0.2). A random sample of n=5 was generated from a Poisson(θ=1.5) yielding y= 1, 0, 1, 0, 0, 1, 2. The posterior was then Gamma(30, 9.5). 95% Symmetric and HPDIs are shown below and plotted in Figure 5.9.

```
Symmetric   2.130618 4.38409 length= 2.25347
HPD         2.069656 4.30633 length= 2.23667
```

In this case the symmetric and HPDI interval endpoints differ slightly but the lengths are nearly identical.

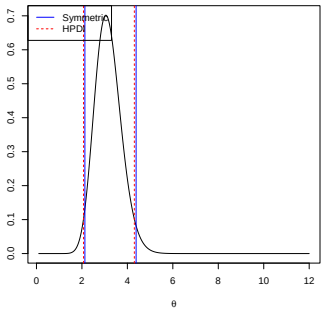


Figure 5.9: Example of 95% CIs for Poisson parameter using symmetric and HPDI approaches.

Credible regions

Given two or more parameters, 1-α, credible regions can be defined as well. There are a variety of ways to construct such regions and the construction can be quite complex. As a simple example suppose that there are two parameters, thus the joint posterior is a surface over the (θ_1, θ_2) plane. A rectangular region can be defined with the four corners being $[(\mathrm{LB}_1, \mathrm{LB}_2), (\mathrm{LB}_1, \mathrm{UB}_2), (\mathrm{UB1}, \mathrm{LB}_2), (\mathrm{UB}_1, \mathrm{UB}_2)]$ such that

$$\int_{\mathrm{LB}_2}^{\mathrm{UB}_2} \int_{\mathrm{LB}_1}^{\mathrm{UB}_1} p(\theta_1, \theta_2|\mathbf{y}) d\theta_1 d\theta_2 = 1 - \alpha$$

The region need not be rectangular, however. An HPD region can be calculated as well, generally by numerical methods, and it will usually not be rectangular. (Draw a picture.)

Contrast with frequentist intervals

Often, particularly if the data dominates the prior, 1-α frequentist confidence intervals and Bayesian credible intervals can be quite similar. Their interpretation is quite different however.

For example, the sampling model is Normal(θ,1) and the prior for θ is Normal(0,100). One then takes a random sample of n=10 observations and the sample average $\bar{y}$ is 1.54. The posterior distribution is Normal(1.538462,0.0999001) (we will show why this is in a later lecture).

- Frequentist 95% confidence interval:

$$[\bar{y} - 1.96\sqrt{1/n},\ \bar{y} + 1.96\sqrt{1/n}] = [0.920,\ 2.162]$$

  All randomness is in terms of the data as θ is treated as a fixed but unknown quantity.

  The interpretation of a 95% confidence interval is that if such intervals are constructed repeatedly based on repeated random samples of the data, the interval will include the unknown parameter θ 95% of the time.

  For this given sample the interval either contains θ or it doesn't. Once the sample has been observed there is no more randomness, and one cannot make a probability statement about an observed event *after the fact*. That's like rolling a die and seeing a 5 and then trying to say there's a 90% probability that it is a 3.

- Bayesian 95% credible interval (symmetric and HPDI are the same here), letting $\theta_q|\mathbf{y}$ denote the qth quantile from the Normal(1.538462, 0.0999001) posterior:

$$[\theta_{0.025}|\mathbf{y},\ \theta_{0.975}|\mathbf{y}] = [0.9189762,\ 2.157947]$$

  Thus the credible interval is quite similar to the confidence interval.

  However, the interpretation the credible interval is that there is a 95% probability that this particular interval contains the unknown parameter θ.

5.1.3 Other summaries

Standard numerical output. Numerical summaries of posterior distributions are routinely produced by software (e.g., JAGS). These include:

Parameter	Min	1st q	Median	Mean	3rd q	Max	StdDev
θ_1	3	12	21	19	31	45	8
θ_2	83	101	122	126	154	198	27

Posterior probabilities for specific events. Given a posterior density one can calculate many quantitative summaries some of which could be quite complex. For example if one has a joint posterior density for two parameters, θ_1 and θ_2, one can calculate:

$$\Pr(\theta_1 + \theta_2 > 7)$$
$$\Pr[(2 \leqslant \theta_1 \leqslant 4) \cap (1 \leqslant \theta_2)]$$
$$\Pr(\exp(\theta_1) < 10)$$

Posterior distributions for functions of parameters. The Schaeffer surplus production model is used to model the biomass of a harvested fish stock. Suppose that the sampling model is

$$B_{t+1} \sim \text{Lognormal}\left(\ln\left[B_t + r_{max}B_t\left(1 - \frac{B_t}{K}\right) - C_t\right], \sigma^2\right)$$

where B_t is fish biomass in year t, C_t is the catch and r_{max} and K are unknown parameters. Suppose one has n years of catch and biomass data, and carries out a Bayesian analysis of the unknown parameters, and arrives at a joint posterior distribution for r_{max} and K. The maximum sustainable harvest (the maximum harvest that can be taken in "perpetuity") is calculated by

$$Harvest_{MSY} = \frac{r_{max}K}{4}$$

Figure 5.10 shows the biomass projections with and without harvest. Given the joint posterior distribution for r_{max} and K (or a sample from it), one can calculate a joint posterior distribution for $Harvest_{MSY}$.

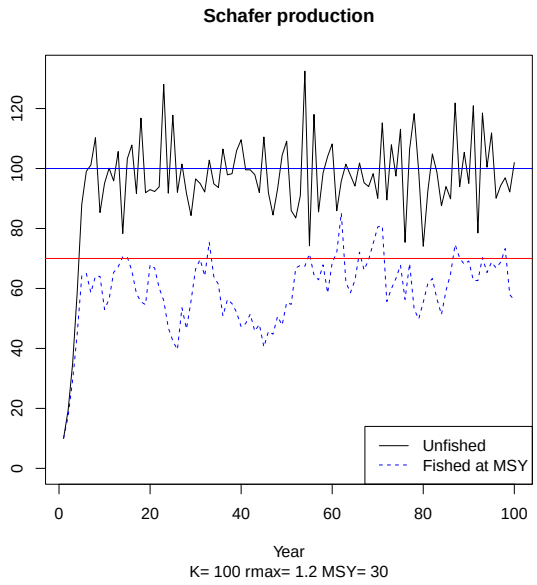


Figure 5.10: Schaefer Production model simulated without harvest and with harvest (at MSY).

5.2 Hypothesis Testing

Overview

Hypothesis testing is a formal procedure for comparing competing theories about natural phenomena. It can be viewed as a key component of the scientific method and, in general, a means of advancing knowledge and understanding.

The simplest scenario has two competing hypotheses, one labelled the Null Hypothesis and denoted H_0 and the other labelled the Alternative Hypothesis and denoted H_1. In our statistical framework these hypotheses are typically statements about the possible values of a parameter (or parameters):

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

The sets defined by the hypotheses are mutually exclusive, $\Theta_0 \cap \Theta_1 = \varnothing$, and (usually) exhaustive, i.e., their union includes the entire parameter space, $\Theta_0 \cup \Theta_1 = \Theta$.

Example: Event Probability. The fraction of a specific variety of potatoes infected by a virus is approximately 0.15. A virus resistant variety of potatoes will be planted this year and the hope is that the fraction infected will be less than 0.15. Letting θ denote the probability that plant is infected, the two competing hypotheses are:

$$H_0 : \theta \geqslant 0.15$$
$$H_1 : \theta < 0.15$$

Example: Regression covariate. In multiple regression one is interested in knowing if one or more covariates have a linear relationship with a response variable. For example, is this model, $E[Y]=\beta_0 + \beta_1 x_1 + \beta_2 x_2$, correct? In particular does x_1 have a relationship with Y? Two hypotheses might be:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Comments

- A Null Hypothesis that includes only a single value for θ it is called a Point Null Hypothesis (or Simple Hypothesis). There are often practical problems with such hypotheses; e.g., referrring to the above regression example, β_1 = 0.001 would be contrary to $H_0 : \beta_1 = 0$.
- Competing hypotheses can (sometimes) be viewed as competing models about phenomena. The above hypotheses could be written as:

  $$H_0 \equiv M_0 \text{ is the correct model : } E[Y] = \beta_0 + \beta_2 x_2$$
  $$H_1 \equiv M_1 \text{ is the correct model : } E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

  And one can imagine a larger set of hypotheses or alternative models:

  $$M_1 : E[Y] = \beta_0$$
  $$M_2 : E[Y] = \beta_0 + \beta_1 x_1$$
  $$M_3 : E[Y] = \beta_0 + \beta_2 x_2$$
  $$M_4 : E[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

5.2.1 Classical Hypothesis Testing

The classical or frequentist approach to testing two hypotheses is:

- Assume that one hypothesis is true, H_0.
- Calculate a test statistic based on the observed sample data, $T(\mathbf{y}_{obs})$ (that should be informative about H_0 and H_1).
- *Conditional* on H_0 being true, calculate the probability of observing sample data that would yield test statistics as extreme or more extreme than $T(\mathbf{y}_{obs})$ *in the direction of* H_1.

  That probability is called the *p-value* and is formally defined:

  $$p - value = \Pr(T(\mathbf{y}) \text{ more extreme than } T(\mathbf{y}_{obs}) | \theta, H_0) \tag{5.14}$$

  where "extremeness" is in the direction of the alternative hypothesis, H_1.
- If that probability is
  - "sufficiently small", "Reject H_0" and "Accept H_1".
  - "relatively large", "Do not reject H_0." (*But Do Not Say* "Accept H_0.")

Example: Normal(μ,σ^2). The sampling model is Normal(μ, σ^2), where μ is unknown but σ^2 is known and equals 2 (admittedly seldom realistic). The null hypothesis is that μ is less than or equal to 3 while the alternative hypothesis is that μ is greater than 3:

$$H_0 : \mu \leqslant 3$$
$$H_1 : \mu > 3$$

A random sample of n=10 is taken and the sample average is $\bar{y}$ = 4. The test statistic:

$$T(\mathbf{y}) = \frac{\bar{y} - \mu_0}{\sqrt{\sigma^2/n}}$$

where μ_0 is a value in the set $\mu \leqslant 3$. Note that conditional on H_0, $T(\mathbf{y})$ is Normal(0,1)[8]. Given that there are infinite number of values in the set Θ_0, the convention is to select the value of $\theta \in \Theta_0$ that would yield the largest p-value, in this case μ_0=3, and then the p-value is

$$\Pr\left(T(\mathbf{y}) \geqslant T(observed)\right) = \Pr\left(T(\mathbf{y}) \geqslant \frac{4-3}{\sqrt{2/10}} = 2.236\right) = 1 - \Phi(2.236) = 0.013$$

where $\Phi(z)$ is the cumulative distribution function for a standard normal random variable. Note that extremeness here is in the direction of H_1, namely, towards values of $\mu > 3$. Such a p-value of 0.013 would be considered by many to be "sufficiently small", or *statistically significant*, and H_0 would be rejected.

Example: Nested linear models. Two linear models for an expected outcome are proposed, where one model is nested inside the other model:

$$M1 : E[Y] = \beta_0 + \beta_1 x$$
$$M2 : E[Y] = \beta_0 + \beta_1 x + \beta_2 x^2$$

Equivalently,

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

Assuming normality of Y, the common test statistic is the t-statistic, $t = \frac{\hat{\beta}_2 - 0}{\text{std.error}(\hat{\beta}_2)}$. And extremeness in this case would be values of t that are relatively far from 0, $t << 0$ or $t >> 0$.

Problems with classical hypothesis testing.

1. H_0 and H_a must be structured such that "extremeness" in the direction of H_a is definable in order to calculate the p-value. If one is comparing models that are not nested, "extremeness" is not readily definable. For example, exponential "growth" versus linear "growth" models:

   $$M1 : E[Y] = \beta_0 \exp(\beta_1 t)$$
   $$M2 : E[Y] = \beta_0 + \beta_1 t$$

   If H_0 is that $M1$ is true, and H_1 is that $M2$ is true, then assuming that H_0 is true, what is a measure of extremeness in the direction of H_1?

2. The evidence is only *against* H_0 as the p-value is calculated *assuming* that H_0 is true.
   - A small p-value indicates that the data are not what would be expected if H_0 is true.

[8]The test statistic $T(\mathbf{y})$ for this setting is sometimes written z and is called the z-statistic.

- A large p-value, however, does not mean that H_0 is true, that the model implied by H_0 is true, as the calculation is made *assuming* that H_0 is true—-so there is no weight of evidence *for* H_0.
- This is the reason that the frequentist conclusion given a large p-value is to say "fail to reject" H_0, and *Not* to say "accept" H_0. You can't accept something that you assumed was true in the first place.

3. The p-value itself, e.g., 0.01, does not provide "weight of evidence" for the H_0. The p-value is a long-run relative frequency measure: if H_0 was true, only 1% of the time would the observed results or *more extreme* results. The p-value is not the probability that H_0 is true.
4. Calculation of p-values involves including values that were not even observed. This violates the Likelihood Principle[9].

   Example: Poisson with two possibilities. (This example was discussed previously in Lecture Notes 1.) The sampling model for the data is Poisson(θ) and there are two hypotheses about θ:

   $$H_0 : \theta = 1 \qquad H_1 : \theta = 2$$

   A sample size n=1 is drawn and yields the value y=2. The standard frequentist approach is to calculate the p-value: the probability of the observed value and any values in a direction away from H_0 in the direction of H_1. In this case the p-value is $Pr(Y \geqslant 2|H_0)$ = 1- $Pr(Y = 0 \cup Y = 1|\theta = 1)$ = 0.264[10]. Thus, one would not reject H_0.

   This procedure is violating the Likelihood Principle, however, in that inference is being based on more than the likelihood of the data: the probability of events that *did not occur*, such as Y=3 or Y=4, is being used as the basis for inference.

5.2.2 Bayesian Hypothesis Testing

Suppose that there are two hypotheses about a parameter θ:

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

where $\Theta_0 \cup \Theta_1 = \Theta$, the entire parameter space, and $\Theta_0 \cap \Theta_1 = \emptyset$.

The Bayesian approach is to specify prior probabilities for each hypothesis:

$$Pr(H_0 \text{ is true}) = Pr(H_0) = Pr(\theta \in \Theta_0)$$
$$Pr(H_1 \text{ is true}) = Pr(H_1) = Pr(\theta \in \Theta_1),$$

where the shortened expressions $Pr(H_0$ and $Pr(H_1$ will be used from here on, and by definition of the parameter subspaces, $Pr(H_0)$ + $Pr(H_1)$ = 1.

Data, $\mathbf{y}$, are collected and the posterior probabilities for each hypothesis are calculated:

$$Pr(H_0|\mathbf{y}) = Pr(\theta \in \Theta_0|\mathbf{y})$$

And $Pr(H_1|\mathbf{y})$ = 1- $Pr(H_0|\mathbf{y})$.

The complexity of the calculation of the posterior probability is affected by the nature of the hypotheses, i.e., simple or composite.

[9] Reminder from Lecture 1 Notes: The Likelihood Principle says that given a sample of data, $\mathbf{y}$, any two sampling models for $\mathbf{y}$, say $f_1(\mathbf{y}|\theta)$ and $f_2(\mathbf{y}|\theta)$, that have proportional likelihoods yield the same inference for θ. The main point is that, according to the Likelihood Principle, inference for θ depends on the observed $\mathbf{y}$ alone, not on unobserved values of $\mathbf{y}$.

[10] In R: `1-ppois(q=1,lambda=1)`=0.2642411.

Simple

Simple hypotheses have single parameter values:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1$$

where $\Theta = \{\theta_0, \theta_1\}$. Then

$$\Pr(H_0|\mathbf{y}) = \Pr(\theta = \theta_0|\mathbf{y}) = \frac{\Pr(H_0)f(\mathbf{y}|\theta_0)}{m(\mathbf{y})} = \frac{\Pr(H_0)f(\mathbf{y}|\theta_0)}{\Pr(H_0)f(\mathbf{y}|\theta_0) + \Pr(H_1)f(\mathbf{y}|\theta_1)}$$
$$\Pr(H_1|\mathbf{y}) = \Pr(\theta = \theta_1|\mathbf{y}) = \frac{\Pr(H_1)f(\mathbf{y}|\theta_1)}{m(\mathbf{y})} = \frac{\Pr(H_1)f(\mathbf{y}|\theta_1)}{\Pr(H_1)f(\mathbf{y}|\theta_1) + \Pr(H_0)f(\mathbf{y}|\theta_0)}$$

$\Pr(H_1|\mathbf{y})$ is simply $1 - \Pr(H_0|\mathbf{y})$.

Note that to calculate posterior *odds*, the ratio of the probability of an event to the probability of its complement, the normalizing constant $m(\mathbf{y})$ need not be calculated:

$$\frac{\Pr(H_0|\mathbf{y})}{\Pr(H_1|\mathbf{y})} = \frac{\Pr(H_0)f(\mathbf{y}|\theta_0)}{\Pr(H_1)f(\mathbf{y}|\theta_1)}$$

Composite

Composite hypotheses include sets of parameter values:

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

The probabilities for each hypothesis can be specified in two ways. One way is to simply assert values, e.g., $\Pr(H_0)$=0.7 and $\Pr(H_1)$=0.3.

The other way, which is more common and *what we will assume to be the case*, is to calculate *induced* priors for the hypotheses on the basis of a prior on the parameter space, Θ. Using our usual notation, $\pi(\theta)$ is the prior probability over the entire parameter space, the prior probability for Hypothesis i is

$$\Pr(H_i) = \int_{\theta \in \Theta_i} \pi(\theta) d\theta$$

Thus the prior distribution for the parameter θ, $\pi(\theta)$, is *inducing* the prior for the hypothesis, $\Pr(H_i)$. Now

$$\begin{aligned}
\Pr(H_i|\mathbf{y}) &= \frac{p(\mathbf{y}, H_i)}{m(\mathbf{y})} = \frac{\Pr(H_i)\ p(\mathbf{y}|H_i)}{m(\mathbf{y})} \\
&= \frac{\Pr(H_i) \int p(\mathbf{y}, \theta|H_i)\ d\theta}{m(\mathbf{y})} = \frac{\Pr(H_i) \int f(\mathbf{y}|\theta)\pi(\theta|H_i)\ d\theta}{m(\mathbf{y})} \\
&= \frac{\Pr(H_i) \int_{\theta \in \Theta_i} f(\mathbf{y}|\theta) \frac{\pi(\theta)}{\Pr(H_i)}\ d\theta}{m(\mathbf{y})} = \frac{\int_{\theta \in \Theta_i} f(\mathbf{y}|\theta)\pi(\theta) d\theta}{m(\mathbf{y})} \\
&= \int_{\theta \in \Theta_i} p(\theta|\mathbf{y})\ d\theta = \Pr(\theta \in \Theta_i|\mathbf{y})
\end{aligned}$$

The key step in the above is the equality of integrating $\pi(\theta|H_i)$ over the entire parameter space Θ and integrating $\frac{\pi(\theta)}{\Pr(H_i)}$ over the reduced parameter space Θ_i. This result may seem "after-the-fact" obvious: the posterior probability of H_i is simply the integral of the posterior for θ over Θ_i.

Note: if $\pi(\theta)$ is *improper*, then the posterior $p(\theta|\mathbf{y})$ needs to be *proper* to calculate $\Pr(H_i|\mathbf{y})$

Remarks

- The posterior odds of H_0 against H_1 can be written:

$$\frac{\Pr(H_0|\mathbf{y})}{\Pr(H_1|\mathbf{y})} = \frac{\int_{\theta\in\Theta_0} f(\mathbf{y}|\theta)\pi(\theta)d\theta}{\int_{\theta\in\Theta_1} f(\mathbf{y}|\theta)\pi(\theta)d\theta} = \frac{\Pr(\theta\in\Theta_0|\mathbf{y})}{\Pr(\theta\in\Theta_1|\mathbf{y})} = \frac{\Pr(\theta\in\Theta_0|\mathbf{y})}{1-\Pr(\theta\in\Theta_0|\mathbf{y})}$$

  Note that this does Not require calculation of the marginal distribution, $m(\mathbf{y})$,

- **Multiple Hypotheses.** Multiple hypotheses can be handled similarly. The different hypotheses could correspond to different sets of models: $M_1, \ldots, M_K$:

$$H_i : \text{The correct model is model } M_i$$

  One assigns priors to each hypothesis, $\Pr(H_i)$, where $\sum_{i=1}^{K} \Pr(H_i) = 1$. Then the posterior probability for model i:

$$\Pr(H_i|\mathbf{y}) = \frac{\Pr(H_i,\mathbf{y})}{\Pr(\mathbf{y})} = \frac{\Pr(H_i)\Pr(\mathbf{y}|H_i)}{\sum_{k=1}^{K}\Pr(H_k)\Pr(\mathbf{y}|H_k)}$$

  where the form of $\Pr(H_i,\mathbf{y})$ depends upon whether H_i was simple or composite.

- **Computational difficulties.** For composite hypotheses, the integration needed to calculate $\Pr(\theta\in\Theta_i|\mathbf{y})$ may not be analytically tractable.

5.2.3 Bayes Factors

An alternative to calculating posterior probabilities for the hypotheses is Bayes factors. A Bayes factor is the ratio of the posterior odds (for one hypothesis compared to another) to the prior odds. The *prior odds* for H_0 against H_1 is the ratio $\frac{\Pr(H_0)}{\Pr(H_1)}$. E.g., if $\Pr(H_0)$=0.6 and $\Pr(H_1)$=0.4, then 0.6/0.4 = 1.5 are the prior odds. The *posterior odds* for H_0 against H_1 is the ratio $\Pr(H_0|\mathbf{y})/\Pr(H_1|\mathbf{y})$. The Bayes Factor for H_0 against H_1, which is written BF_{01}, is

$$BF_{01} = \frac{\Pr(H_0|\mathbf{y})/\Pr(H_1|\mathbf{y})}{\Pr(H_0)/\Pr(H_1)} \tag{5.15}$$

Rules of thumb for interpreting Bayes Factors are given by Kass and Raftery (Journal of the American Statistical Association Volume 90, 1995 - Issue 430):

BF_{01}	Interpretation
< 3	No evidence for H_0 over H_1
> 3	Positive evidence for H_0
> 20	Strong evidence for H_0
> 150	Very strong evidence for H_0

Note: $BF_{10} = 1/BF_{01}$. And

- $BF_{01} < \frac{1}{3} \Rightarrow BF_{10} > 3 \Rightarrow$ positive evidence for H_1
- $BF_{01} < \frac{1}{20} \Rightarrow BF_{10} > 20 \Rightarrow$ strong evidence for H_1

Simple vs Simple

$H_0 : \theta = \theta_0$ **vs** $H_1 : \theta = \theta_1$**.**

$$BF_{01} = \frac{\Pr(H_0|\mathbf{y})/\Pr(H_1|\mathbf{y})}{\Pr(H_0)/\Pr(H_1)} = \frac{f(\mathbf{y}|\theta_0)\Pr(H_0)/f(\mathbf{y}|\theta_1)\Pr(H_1)}{\Pr(H_0)/\Pr(H_1)} = \frac{f(\mathbf{y}|\theta_0)}{f(\mathbf{y}|\theta_1)} \tag{5.16}$$

Thus the Bayes Factor is simply the ratio of the likelihoods, and the priors for the hypotheses are irrelevant.

Poisson Example. The sampling distribution for the data is Poisson(θ) and $H_0 : \theta = 1$ and $H_1 : \theta = 2$. The prior for H_0 is $\Pr(H_0)$=0.8, thus $\Pr(H_1)$=0.2. A single observation, $n = 1$, is observed with $y = 2$. Then $BF_{01} = e^{-1}1^2/e^{-2}2^2$ = 0.6796, and BF_{10} = 1.4715. Thus there is no evidence for H_0 over H_1, or for H_1 over H_0.

Composite vs Composite

$H_0 : \theta \in \Theta_0$ **vs** $H_1 : \theta \in \Theta_1$**;** $\Theta_0 \cup \Theta_1 = \Theta$. Below we are assuming that the priors for the hypotheses are induced by the prior for θ; e.g., $\Pr(H_0) = \int_{\theta \in H_0} \pi(\theta) d\theta$.

$$BF_{01} = \frac{\Pr(H_0|\mathbf{y})/\Pr(H_1|\mathbf{y})}{\Pr(H_0)/\Pr(H_1)} = \frac{\left[\int_{\theta\in\Theta_0} f(\mathbf{y}|\theta)\pi(\theta)d\theta\right] \Big/ \left[\int_{\theta\in\Theta_1} f(\mathbf{y}|\theta)\pi(\theta)d\theta\right]}{\Pr(H_0)/\Pr(H_1)} \tag{5.17}$$

$$= \frac{\Pr(\theta \in \Theta_0|\mathbf{y})/\Pr(\theta \in \Theta_1|\mathbf{y})}{\Pr(H_0)/\Pr(H_1)} \tag{5.18}$$

Simple vs Composite

$H_0 : \theta = \theta_0$ **vs** $H_1 : \theta \neq \theta_0$**.**

As before, we require that the prior probabilities for the hypotheses are induced by the prior for θ; e.g., $\Pr(H_0) = \int_{\theta \in H_0} \pi(\theta) d\theta$. For a simple hypothesis, this requires θ to have a probability point mass for H_0, which can be achieved by assigning a Dirac Delta at $\theta = \theta_0$, and expressing the prior density as a mixture, $\pi(\theta) = \alpha_0 \delta_{\theta_0}(\theta) + \alpha_1 \pi(\theta|H_1)$, where $\pi(\theta|H_1)$ is the conditional prior density under H_1, and α_0, α_1 are the prior probabilities for H_0 and H_1, respectively.

$$\begin{aligned} BF_{01} &= \frac{\Pr(H_0|\mathbf{y}) / \Pr(H_1|\mathbf{y})}{\Pr(H_0) / \Pr(H_1)} = \frac{[f(\mathbf{y}|\theta_0) \Pr(H_0)] / \left[\Pr(H_1) \int_{-\infty}^{\infty} f(\mathbf{y}|\theta) \pi(\theta|H_1) d\theta\right]}{\Pr(H_0) / \Pr(H_1)} \\ &= \frac{f(\mathbf{y}|\theta_0)}{\int_{-\infty}^{\infty} f(\mathbf{y}|\theta) \pi(\theta|H_1) d\theta} \end{aligned} \tag{5.19}$$

- As for the simple versus simple case, the prior probabilities for the hypotheses are cancelling out.
- Note that (5.19) is *not* the same as $\frac{f(\mathbf{y}|\theta_0)}{m(\mathbf{y})}$, as the normalisation of that is different, with $m(\mathbf{y}) = \alpha_0 f(\mathbf{y}|\theta_0) + \alpha_1 \int_{-\infty}^{\infty} f(\mathbf{y}|\theta) \pi(\theta|H_1) d\theta$.
- The hypothesis test here may be viewed as a *model selection* test; where $H_0 : \theta = \theta_0$ and $H_1 : \theta \sim \pi(\theta|H_1)$, i.e. the hypotheses are statements about the distribution of θ. More generally, we can have hypotheses of the type

  $$\begin{aligned} H_0 &: \theta \sim \pi(\theta|H_0), \\ H_1 &: \theta \sim \pi(\theta|H_1), \end{aligned}$$

  with some strictly positive prior hypothesis probabilities $\Pr(H_0)$ and $\Pr(H_1)$. The Bayes factor then becomes

  $$BF_{01}(\mathbf{y}) = \frac{E_{\theta|H_0}[f(\mathbf{y}|\theta)]}{E_{\theta|H_1}[f(\mathbf{y}|\theta)]},$$

  which is the ratio of the prior expectations with respect to θ of the likelihood function for fixed observations $\mathbf{y}$. This is similar to a likelihood ratio, but we take expectations over θ instead of the θ modes under H_0 and H_1.

5.2.4 Example: Simple Null and Simple Alternative

The number of hairs per square inch of mohair fabric used by a teddy bear manufacturer is assumed to have a Poisson(θ) distribution (King and Ross, 2017). The manufacturer wants to test the hypotheses:

$$H_0 : \theta = 100; \quad H_1 : \theta = 110$$

To test these hypotheses an independent random sample of n pieces of fabric is drawn and the number of hairs per square inch, $\mathbf{y} = y_1, \ldots, y_n$, is recorded.

1. The Bayes Factor, BF_{01}:

$$\begin{aligned} BF_{01} &= \frac{\Pr(H_0|\mathbf{y}) / \Pr(H_1|\mathbf{y})}{\Pr(H_0) / \Pr(H_1)} = \frac{\Pr(\mathbf{y}|H_0) \Pr(H_0) / \Pr(\mathbf{y}|H_1) \Pr(H_1)}{\Pr(H_0) / \Pr(H_1)} \\ &= \frac{\Pr(\mathbf{y}|H_0)}{\Pr(\mathbf{y}|H_1)} = \frac{\exp(-100 * n) 100^{n\bar{y}}}{\exp(-110 * n) 110^{n\bar{y}}} = \exp(10n) \left(\frac{100}{110}\right)^{n\bar{y}} \end{aligned}$$

2. Given n=10 and $\bar{y}$ = 102.7:

$$\text{BF}_{01} = \exp(10 * 10)\left(\frac{100}{110}\right)^{10*102.7} = 8.301576$$

which is between 3 and 20, thus "positive evidence" for H_0.

3. Assume the priors for H_0 and H_1 were $Pr(H_0)$=$Pr(H_1)$=0.5. The posterior probabilities for each hypothesis can be calculated directly from the Bayes Factor as follows:

$$Pr(H_0|\mathbf{y}) = \frac{\text{BF}_{01}}{1 + \text{BF}_{01}} = \frac{8.301576}{1 + 8.301576} = 0.8924913$$

Exercise: show why the above formula works.

Exercise: Given y_i, i=1,...,10 are independent Normal(μ,1) random variables, where observe data:

`3.4, 2.9, 3.0, 3.5, 3.3, 3.7, 2.7, 3.9, 2.7, 2.9`

Test the simple hypothesis: $H_0 : \mu = 3$ vs $H_1 : \mu = 3.5$. Show that the Bayes Factor BF_{01} = 1.28.

5.2.5 Example: Composite Null and Composite Alternative

[*R Code for this example can be found on Learn in Composite_vs_Composite_Example.*]

A food manufacturer is considering releasing a new flavour of hoummus, but before doing so wants to carry out an experiment with volunteers to see whether this new flavour is liked better than a competitor's version (based on example from Carlin and Louis, 2009). They would like to be "pretty sure" that the new flavour is preferred by at least 60% of hoummus consumers. Letting θ be the probability that the new flavour is preferred, there are two hypotheses:

$$H_0 : \theta \geqslant 0.6 \quad vs \quad H_1 : \theta < 0.6$$

If there is strong evidence for H_0, they will release the new flavour. The manufacturer would prefer to be cautious and selects a Beta prior for θ that has an expected value of 0.5 and a coefficient of variation of 0.3 (thus a standard deviation of 0.5*0.3=0.15). That translates into Beta(5.056, 5.056).

The *induced* prior probability for H_0 is then[11]:

$$\begin{aligned} Pr(H_0) &= \int_{0.6}^{1} \frac{1}{Be(5.056, 5.056)} \theta^{4.06}(1-\theta)^{4.06} d\theta \\ &= \int_{0.6}^{1} \frac{\Gamma(10.13)}{\Gamma(5.065)\Gamma(5.065)} \theta^{4.06}(1-\theta)^{4.06} d\theta = 0.265 \end{aligned}$$

Thus $Pr(H_1)$ = 0.735.

To test these hypotheses, a taste preference study is carried with n=16 volunteers. How would you recommend that such a study be carried out?

Assume that the probability of preferring the new flavour is the same for all volunteers and the responses are independent. Then, letting y be the number preferring the new flavour, $y \sim$ Binomial(16, θ). After the study was completed, 13 of the 16 volunteers preferred the new flavour. What are the posterior probabilities for H_0 and H_1? And what is BF_{01}?

To begin, note that $Pr(H_0|y)$ is the same as $Pr(\theta \geqslant 0.6|y)$. We know that the Beta distribution is conjugate for the Binomial distribution and the posterior is Beta($\alpha + y$, $\beta + n - y$), or in this case,

[11] In R: 1-pbeta(0.6,5.056,5.056)=0.265.

Beta(5.056+13, 5.056+16-13) = Beta(18.056, 8.056). Therefore:

$$\Pr(H_0|y = 13) = \int_{0.6}^{1} \frac{1}{Be(18.056, 8.056)} \theta^{18.056-1}(1-\theta)^{8.056-1} d\theta = 0.8448$$
$$\Pr(H_1|y = 13) = 1 - \Pr(H_0|13) = 0.1552$$

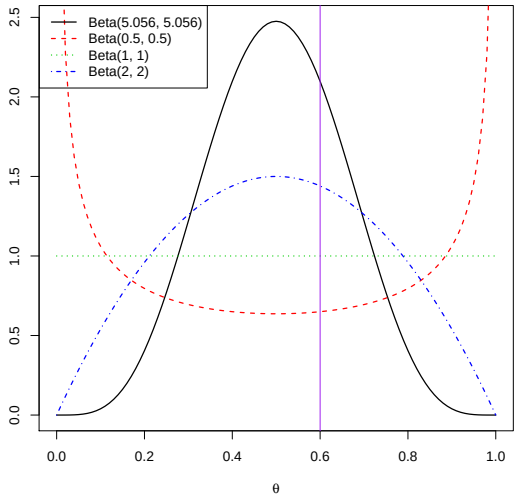
Note: the R code for $\Pr(\theta \geqslant 0.6)$ = `1-pbeta(0.6,18.056,8.056)` = 0.8447625. And the Bayes Factor:

$$BF_{01} = \frac{\Pr(H_0|y = 13)/\Pr(H_1|y = 13)}{\Pr(H_0)/\Pr(H_1)} = \frac{0.8448/0.1552}{0.265/0.735} = 15.1$$

which is between 3 and 20, thus "positive evidence" for H_0.

To evaluate the sensitivity of the resulting posterior probabilities and the Bayes Factor, three other priors for θ were considered: Beta(0.5,0.5) or Jeffreys' prior, Beta(1,1) or a Uniform(0,1), and Beta(2,2). The four prior densities are shown in Figure 5.11.

Figure 5.11: Four prior distributions for θ in the hoummus taste preference study. The vertical line at 0.6 marks the division between H_0 and H_1.

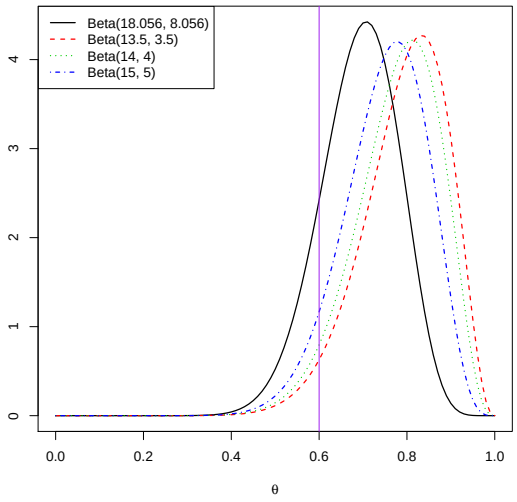


The posterior quantiles and means for θ, given y=13 for n=16, as well as $\Pr(H_0)$ and BF_{01} are shown in Table 5.1. The resulting posterior distributions are shown in Figure 5.12. As can be seen, the initial prior is the most skeptical regarding the preference for the new hoummus.

Table 5.1: Numerical summaries of posterior quantities for taste preference study.

$\pi(\theta)$	$\Pr(H_0)$	0.25	0.50	0.75	Mean	Pr(Ho|y)	BF(0,1)
Beta(5.056, 5.056)	0.265	0.63	0.70	0.76	0.69	0.84	15.06
Beta(0.5, 0.5)	0.436	0.74	0.81	0.87	0.79	0.96	34.43
Beta(1, 1)	0.400	0.72	0.79	0.85	0.78	0.95	30.81
Beta(2, 2)	0.352	0.69	0.76	0.82	0.75	0.93	24.60

Figure 5.12: Four posterior distributions for θ in the hoummus taste preference study given y=13 in n=16 trials. The vertical line at 0.6 marks the division between H_0 and H_1.



5.2.6 Example: Simple Null and Composite Alternative

The sampling distribution is Poisson(θ). The null hypothesis is $H_0 : \theta = 5$ and the alternative is $H_1 : \theta \neq 5$, where $Pr(H_0)$=0.7. Under H_1, a Gamma prior distribution is chosen for θ such that $E[\theta|H_1]$=5 with a CV of 0.1, thus a Gamma(100,20).[12]

A random sample of n=8 is drawn yielding the following values

$$3, 3, 3, 3, 5, 7, 7, 4$$

Note: $\bar{y}$=4.375, and $(\theta|\mathbf{y}, H_1) \sim$ Gamma(100+$\sum_{i=1}^{8} y_i$, 20+n) = Gamma(135,28).

To find the posterior probabilities:

$$Pr(H_0|\mathbf{y}) = \frac{Pr(H_0, \mathbf{y})}{m(\mathbf{y})} \propto Pr(H_0) \prod_{i=1}^{8} \frac{e^{-5} 5^{y_i}}{y_i!} = 0.7 * \frac{e^{-40} 5^{35}}{\prod_{i=1}^{8} y_i!} = 9.12873 \cdot 10^{-8}$$

$$Pr(H_1|\mathbf{y}) = \frac{Pr(H_1, \mathbf{y})}{m(\mathbf{y})} \propto Pr(H_1) \int_0^{\infty} \frac{e^{-\theta*8} \theta^{35}}{\prod_{i=1}^{8} y_i!} \frac{20^{100}}{\Gamma(100)} \theta^{100-1} e^{-20*\theta} d\theta$$

$$= 0.3 * \frac{1}{\prod_{i=1}^{8} y_i!} \frac{20^{100}}{\Gamma(100)} \frac{\Gamma(135)}{(28)^{135}} = 3.684845 \cdot 10^{-8}$$

Then

$$Pr(H_0|\mathbf{y}) = \frac{9.12873 \cdot 10^{-8}}{9.12873 \cdot 10^{-8} + 3.684845 \cdot 10^{-8}} = 0.7124$$

$$Pr(H_1|\mathbf{y}) = \frac{3.684845 \cdot 10^{-8}}{9.12873 \cdot 10^{-8} + 3.684845 \cdot 10^{-8}} = 0.2876$$

[12] Equivalently, the prior distribution for θ can be expressed as a mixture of a probability point mass 0.7 at $\theta = 5$, and a Gamma(100,20) density scaled by 0.3.

And the Bayes Factor[13] for H_0 against H_1:

$$BF_{01} = \frac{0.7124/0.2376}{0.7/0.3} = 1.0617$$

which implies no evidence of H_0 over H_1 or vice versa.

5.2.7 Multiple Hypotheses

As said previously, multiple models can be viewed as multiple hypotheses. From an example by Lavine[14], a primary (elementary) school in Fresno, California had two high-voltage transmission lines nearby and the cancer rate amongst staff was a concern as 8 of the 145 staff had developed invasive cancers. Assume independence between staff and identical probabilities for cancer. Let y denote the number developing cancer and θ the probability of cancer. Then $y \sim$ Binomial(n=145, θ) is the sampling model.

Based on data collected at a national level (for approximately the same age of the staff, mostly women, and number of years of working), the expected number of cancers for 145 staff was estimated to be 4.2. Translating that into a probability, one hypothesis was that θ=4.2/145 $\approx$ 0.03. However, different individuals thought the rate was higher and three alternative hypotheses were postulated:

$$H_1 : \theta = 0.03, \quad H_2 : \theta = 0.04, \quad H_3 : \theta = 0.05, \quad H_4 : \theta = 0.06$$

These four hypotheses can be viewed as 4 models. Lavine proposed that *a priori*, H_1 was as likely to be right as it was to be wrong, thus the prior for H_1 was $Pr(H_1) = 1/2$. Then he assumed that any of the remaining hypotheses was equally likely, thus $Pr(H_2) = Pr(H_3) = Pr(H_4)$=1/6. The posterior probabilities for the four hypotheses can be viewed as the relative weight of evidence for the competing theories:

$$\begin{aligned} Pr(H_1|y=8) &= \frac{Pr(y=8|H_1)\,Pr(H_1)}{\sum_{i=1}^{4} Pr(y=8|H_i)\,Pr(H_i)} \\ &= \frac{0.03^8 \cdot 0.97^{137} \cdot \frac{1}{2}}{0.03^8 \cdot 0.97^{137} \cdot \frac{1}{2} + 0.04^8 \cdot 0.96^{137} \cdot \frac{1}{6} + 0.05^8 \cdot 0.95^{137} \cdot \frac{1}{6} + 0.06^8 \cdot 0.94^{137} \cdot \frac{1}{6}} \\ &= 0.23 \end{aligned}$$

Repeating for H_2, H_3, and H_4:

$$Pr(H_1|y=8) = 0.23, \quad Pr(H_2|y=8) = 0.21, \quad Pr(H_3|y=8) = 0.28, \quad Pr(H_4|y=8) = 0.28$$

Thus, one could conclude that given the data and the priors, each of the four hypotheses are about equally likely. Or that the weight of evidence for each model is about the same. The posterior odds that the cancer rate is higher than the national average, or the posterior odds of H_2 or H_3 or H_4 against H_1 is (0.21+0.28+0.28)/0.23 =3.3. Given that the prior odds of H_2 or H_3 or H_4 and H_1 are 1, this is also the Bayes Factor and by the Kass and Raftery criteria this is just above the "positive evidence" lower bound of 3.

[13]Note: a simpler calculation based on the right-most term in Eq'n 5.19 is $f(y|H_0)/m(y)$, where $f(y|H_0)$=$\frac{e^{-40}5^{35}}{\prod_{i=1}^{8} y_i!} = 1.304104 \cdot 10^{-7}$ and $m(y)$ = $\frac{1}{\prod_{i=1}^{8} y_i!}\frac{20^{100}}{\Gamma(100)}\frac{\Gamma(135)}{(28)^{135}} = 1.228282 \cdot 10^{-7}$, and $BF_{01} = 1.0617$.

[14]"What is Bayesian statistics and why everything else is wrong".

Contrast with Frequentist Approach. Lavine also carried out the frequentist analysis $\mathrm{H}_0 : \theta = 0.3$ against the alternative $\mathrm{H}_1 : \theta > 0.3$. The P-value is the probability of observing an outcome equal to what was observed, 8 occurrences of cancer in 145 staff, and anything more extreme in the direction of H_1[15]:

$$\begin{aligned}\mathrm{Pr}(Y \geqslant 8|\theta = 0.3) &= \mathrm{Pr}(Y = 8|\theta = 0.3) + \mathrm{Pr}(Y = 9|\theta = 0.3) + \ldots + \mathrm{Pr}(Y = 145|\theta = 0.3) \\ &= 1 - \mathrm{Pr}(Y < 8|\theta = 0.3, n = 145) = 0.0717\end{aligned}$$

This would be considered "significant" evidence against H_0 if the cut-off was 0.10. However, as Lavine points out this P-value does not account for how well the other hypotheses explain the data, information about things that did not happen (e.g., there were Not 9, nor 10, nor 11, and so on incidences of cancer), and the Likelihood Principle is not obeyed.

[15] This can be calculated in R by `1-pbinom(7,size=145,prob=0.03)`.

5.3 Supplement A: Hypothesis Testing in a Decision Theory Framework

Consider just two hypotheses, H_0 and H_1.

1. The State Space has two states: $\Theta = (H_0, H_1)$.
2. The Action Space has two actions: $\mathcal{A} = (a_1, a_2)$, where a_1 is choose H_0 and a_2 is choose H_1. (By choose is meant to then act as if one of these hypotheses is in fact true.)
3. The Loss Function is defined such that there is no loss if correctly choose the true state (the correct hypothesis) but there is a loss when choose wrongly.

	State Space	
Action	H_0	H_1
a_1 (choose H_0)	0	λ_2
a_2 (choose H_1)	λ_1	0

4. After data are collected, $\mathbf{y}$, the *Bayes* Risks for each action are defined as follows:

$$R(a_1|\mathbf{y}) = 0 * Pr(H_0|\mathbf{y}) + \lambda_2 Pr(H_1|\mathbf{y}) = \lambda_2 Pr(H_1|\mathbf{y})$$
$$R(a_2|\mathbf{y}) = \lambda_1 Pr(H_0|\mathbf{y}) + 0 * Pr(H_1|\mathbf{y}) = \lambda_1 Pr(H_0|\mathbf{y})$$

   Note that $Pr(H_0|\mathbf{y})$=1-$Pr(H_1|\mathbf{y})$. Thus the above Risks could be re-written:

$$R(a_1|\mathbf{y}) = \lambda_2 Pr(H_1|\mathbf{y})$$
$$R(a_2|\mathbf{y}) = \lambda_1(1 - Pr(H_1|\mathbf{y}))$$

5. The Bayes Decision Rule is to choose the action with the smaller risk; e.g.
   Choose a_2, select H_1, if $R(a_2|\mathbf{y}) < R(a_1|\mathbf{y})$, which is equivalent to:

$$\frac{R(a_2|\mathbf{y})}{R(a_1|\mathbf{y})} < 1 \quad \equiv \quad \frac{\lambda_1(1 - Pr(H_1|\mathbf{y}))}{\lambda_2 Pr(H_1|\mathbf{y})} < 1 \quad \equiv \quad Pr(H_1|\mathbf{y}) > \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

Comments.

- In classical hypothesis testing wrongly choosing H_1 (conditional on H_0 being true) is called a *Type I error* and in the above setting the loss λ_1 equates to Type I error.
- Similarly wrongly choosing H_0 (conditional on H_1 being true) is a *Type II error*, thus λ_2 equates to a Type II error.
- It is common to choose H_0 to represent "status quo" and to attach primary importance to the value of a Type I error. Experimenters often want to keep the probability of a Type I error, conditional on H_0 being true, the "α"-level, "small", and values like 0.01, 0.05, or 0.10 are common.
- The probability of a Type II error, denoted β, is a function of the magnitude of the difference between H_1 and H_0 (the "effect" size) and the sample size. The experimenter, if possible, may choose a particular sample size such that the magnitude of β is "small", or 1-β is "large", where 1-β is called the Power of the test.

The above notions of α and β can be related to the losses λ_1 and λ_2. Making α, the "risk" of a Type I error, "small", equates, more or less, to making λ_1 "large". Suppose that λ_1 is chosen to be 10 times larger than λ_2, λ_1=10λ_2. Then the Bayes Decision to choose H_1 (action a_2) occurs when

$$Pr(H_1|\mathbf{y}) > \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{10\lambda_2}{10\lambda_2 + \lambda_2} = 0.91$$

Thus the probability of choosing H_1 must be greater than 90%

Example 1. Suppose that data are coming from a Poisson(μ) distribution and the following simple null and simple alternative hypotheses are assumed.

$$H_0 : \mu = 1 \qquad H_1 : \mu = 2$$

Let the prior probabilities for these hypotheses be $Pr(H_0)$=0.8 and $Pr(H_1)$=0.2. Define losses such that λ_1=10λ_2; then $\lambda_1/(\lambda_1 + \lambda_2)$=0.91.

A single Poisson observation is made with a value of $y = 3$. The posterior probability for H_1 is then:

$$\begin{aligned} Pr(H_1|y = 3) &= \frac{Pr(H_1)f(y = 3|\mu = 2)}{Pr(H_1)f(y = 3|\mu_2) + Pr(H_0)f(y = 3|\mu = 1)} \\ &= \frac{0.2\exp(-2)2^3}{0.2\exp(-2)2^3 + 0.8\exp(-1)1^3} = \frac{0.2165365}{0.2165365 + 0.2943036} = 0.4238831 \end{aligned}$$

Thus $Pr(H_1|y = 3) <$ 0.91, and the Bayes Decision is to choose H_0.

Question: How large much y be before the Bayes Decision would have been to choose H_1?

Example 2. Suppose again that data are Poisson(μ) but the two hypotheses are composite.

$$H_0 : \mu \leqslant 1 \qquad H_1 : \mu > 1$$

The same losses as in Example 1 are chosen: λ_1=10λ_2. Suppose that a Gamma(4, 4) was chosen for μ; this is based on $E[\mu]$=1 with a CV of 0.50. Then the (induced) prior probability for H_0 is $Pr(\mu < 1)$ = 0.57.

A random sample of n=10 observations was taken where the sample mean, $\bar{y}$, was 1.5. The posterior distribution for μ is then Gamma(4+10*1.5, 4+10) = Gamma(19, 14). Then $Pr(H_1|\mathbf{y})$ = $Pr(\mu > 1|\Gamma(19, 14))$ = 0.88, which is less than $\lambda_1/(\lambda_1 + \lambda_2)$ = 0.91, so choose H_0.

5.4 Supplement B: Composite Hypothesis and Conflicting Priors

In this note we compare the consequences of specifying prior probabilities for two composite hypotheses that are *not induced* from priors for the parameter(s).

There is one parameter, θ, and two composite hypotheses are posed.

$$H_0 : \theta \leqslant \theta_0$$
$$H_1 : \theta > \theta_0$$

Priors for H_0 and H_1 are simply specified, independent of $\pi(\theta)$ and not equal to what would be induced priors:

$$Pr(H_0) \neq p_0^* = \int_{\theta \leqslant \theta_0} \pi(\theta) d\theta$$
$$Pr(H_1) \neq p_1^* = \int_{\theta > \theta_0} \pi(\theta) d\theta$$

If we ignore the incompatibility, we might derive an expression of the following form:

$$\begin{aligned} Pr(H_0|\mathbf{y}) &= \frac{Pr(H_0, \mathbf{y})}{m(\mathbf{y})} = \frac{Pr(H_0) f(\mathbf{y}|H_0)}{m(\mathbf{y})} \\ &= \frac{Pr(H_0) \int f(\mathbf{y}|\theta, H_0) \pi(\theta|H_0) d\theta}{m(\mathbf{y})} \\ &= \frac{Pr(H_0)}{p_0^*} \frac{\int_{\theta \in \Theta_0} f(\mathbf{y}|\theta) \pi(\theta) d\theta}{m(\mathbf{y})} \\ &= \frac{Pr(H_0)}{p_0^*} Pr(\theta \leqslant \theta_0|\mathbf{y}) \end{aligned}$$

Comments:

- Since the probability specification for the model is self-inconsistent, the resulting expression is not well-defined.
- There is a potential for strange situations where a specified prior for H_0 is extremely different than what the induced prior for H_0 would have been. For example, let θ be a Poisson parameter with two hypotheses: H_0: $\theta \leqslant 3$ and H_1: $\theta > 3$. Specify probabilities for the hypotheses of $Pr(H_0)$=0.1 and $Pr(H_1)$=0.9, but let the prior for θ be Gamma(7,3.5). Thus the induced probabilities are $Pr(H_0)$=0.9 and $Pr(H_1)$=0.1.
- Using the ill-defined expression from above for $Pr(H_0|\mathbf{y})$ in the definition of the Bayes Factor, we get the expression

  $$BF_{01} = \frac{Pr(H_0|\mathbf{y}) / Pr(H_1|\mathbf{y})}{Pr(H_1) / Pr(H_1)} = \frac{Pr(\theta \leqslant \theta_0|\mathbf{y}) / Pr(\theta > \theta_0|\mathbf{y})}{p_0^* / p_1^*}.$$

  This happens to be the same as the Bayes Factor under the consistent prior model where $Pr(H_0) = Pr(\theta \leqslant \theta_0)$, but that is by *luck*, and cannot be relied on in general.
- Inconsistent (self contradicting) probability models should not be used, as it is impossible to derive general well-defined results from them.

5.5 Supplement C: Change of variables in densities and integrals

When there is a deterministic relationship between two parameters ϕ and θ, we can convert the density for θ into a density for ϕ, and vice versa.

Let $\phi = g(\theta)$ for some monotonic (increasing or decreasing) and differentiable function $g(\cdot)$. Then $\Pr_\theta$ and $\Pr_\phi$ denote the probability measures for each parameter and let $\mathcal{F}_\theta$ and $\mathcal{F}_\phi$ denote the associated sigma-algebras (the collections of events that the probability measures can measure, i.e. give well-defined probabilities to). For any set $A_\theta \subseteq \mathcal{F}_\theta$ there is then a corresponding set $A_\phi \subseteq \mathcal{F}_\phi$ with $A_\phi = \{\phi; g^{-1}(\phi); \theta \in A_\theta\}$, and $A_\theta = \{\theta; g(\theta) \in A_\phi\}$. Furthermore, since ϕ and θ are deterministically linked, the probability measures agree, so that $\Pr_\phi(A_\phi) = \Pr_\theta(A_\theta)$.

For the case of continuous random variables, we can expand the probabilities into integrals over the respective densities:

$$\Pr_\phi(A_\phi) = \int_{A_\phi} p_\phi(\phi)\, d\phi = \int_{A_\theta} p_\theta(\theta)\, d\theta = \Pr_\theta(A_\theta). \tag{5.20}$$

We can then use a change of variables in the integrals, by using our relation $\phi = g(\theta)$:

$$\int_{A_\phi} p_\phi(\phi)\, d\phi = \int_{A_\theta} p_\phi[g(\theta)] \left|\frac{d\phi}{d\theta}\right| d\theta = \int_{A_\theta} p_\phi[g(\theta)]\, |g'(\theta)|\, d\theta, \tag{5.21}$$

$$\int_{A_\theta} p_\theta(\theta)\, d\theta = \int_{A_\phi} p_\theta[g^{-1}(\phi)] \left|\frac{d\theta}{d\phi}\right| d\phi = \int_{A_\phi} p_\theta[g^{-1}(\phi)] \frac{1}{|g'[g^{-1}(\phi)]|}\, d\phi, \tag{5.22}$$

where the second line used the relation between the derivative of a function ($\theta = g^{-1}(\phi)$) and the derivative of its inverse ($g'(\theta)$).

Since the integrals are equal for all pairs A_θ, A_ϕ of the type defined above, we can equate the integrands with matching $d\phi$ or $d\theta$, so that

$$p_\phi(\phi) = p_\theta[g^{-1}(\phi)] \frac{1}{|g'[g^{-1}(\phi)]|}, \tag{5.23}$$

$$p_\theta(\theta) = p_\phi[g(\theta)]\, |g'(\theta)|\,. \tag{5.24}$$

In particular, we see that the final expression can be remembered via the informal construction

$$p_\theta(\theta)\, d\theta = p_\phi(\phi)\, d\phi \quad \Longrightarrow \quad p_\theta(\theta) = p_\phi(\phi) \left|\frac{d\phi}{d\theta}\right| = p_\phi[g(\theta)]\, |g'(\theta)|\,.$$

5.6 Supplement D: Decision theory

In the Bayesian decision theory context, real world problems rarely come in the neatly formalised setting of only the model parameter uncertainty being relevant to the action taken and the resulting loss. More commonly, the loss depends on future realisations of random phenomena. We can structure such a scenario as follows:

- Model parameters $\theta \sim p(\theta)$, $\theta \in \Theta$
- Past observations of relevant quantities $\mathbf{y} \sim p(\mathbf{y}|\theta)$, $\mathbf{y} \in \mathcal{Y}$
- Action/decision $a(\mathbf{y})$ taken after having observed $\mathbf{y}$, $a \in \mathcal{A}$
- Future outcomes $\mathbf{z} \sim p(\mathbf{z}|\theta, \mathbf{y}, a)$, $\mathbf{z} \in \mathcal{Z}$ (Note: $\mathbf{z}$ may include additional past unknown variables)
- Loss function $\mathcal{L}(a|\theta, \mathbf{y}, \mathbf{z}) \in \mathbb{R}$, quantifying the effect of the action, usually based on the $\mathbf{z}$ values.
- Expected loss, after having observed $\mathbf{y}$ (but not $\mathbf{z}$ or θ):

$$\begin{aligned} R(a) &= \mathsf{E}_{\mathbf{z},\theta|\mathbf{y}}\left[\mathcal{L}(a|\theta, \mathbf{y}, \mathbf{z})|\mathbf{y}\right] \\ &= \int_{\Theta} p(\theta|\mathbf{y}) \left[\int_{\mathcal{Z}} \mathcal{L}(a|\theta, \mathbf{y}, \mathbf{z})\, p(\mathbf{z}|\theta, \mathbf{y}, a)\, d\mathbf{z}\right] d\theta \\ &= \frac{1}{\int_{\Theta} p(\theta)p(\mathbf{y}|\theta)\, d\theta} \int_{\Theta} p(\theta)p(\mathbf{y}|\theta) \left[\int_{\mathcal{Z}} \mathcal{L}(a|\theta, \mathbf{y}, \mathbf{z})\, p(\mathbf{z}|\theta, \mathbf{y}, a)\, d\mathbf{z}\right] d\theta \end{aligned}$$

  Under this general framework, minimising $R(a)$ may in general be quite complicated. The benefit is that is allows systematic reasoning about the effect of modelling assumptions and choices of loss functions (e.g. it may be chosen for computational convenience in some settings, whereas in other settings it is dictated by real world considerations, such as health outcomes or societal impact).

6 Bayesian Computation: Numerical Methods

6.1 General Problem: Integration

Integration, in the continuous parameter case, or summation, in the discrete parameter case, is central to Bayesian inference, and arises in several areas.

1. Calculating the normalising constant, $m(y)$.

   The posterior distribution, $p(\theta|y)$:

   $$p(\theta|y) = \frac{p(\theta, y)}{m(y)} = \frac{\pi(\theta)f(y|\theta)}{\int \pi(\theta)f(y|\theta)d\theta} \tag{6.1}$$

   where y is the observed sample data, which could be a scalar, a vector, a matrix, etc, and θ could be a scalar, a vector, a matrix, etc.

   For a given value of θ, calculation of the numerator is often feasible as that involves evaluating the prior distribution at θ, $\pi(\theta)$, and the likelihood at θ, $f(y|\theta)$ ($\equiv L(\theta|y)$).

   Calculation of the denominator, however, can be a difficult integration problem, particularly when θ is multidimensional, e.g., $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$.

2. Calculating marginal posterior distributions. Given $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$, the posterior distribution for a single component, θ_i, requires integrating over the other q-1 components.

   $$p(\theta_i|y) = \int p(\theta_1, \theta_2, \ldots, \theta_q|y d\theta_1 d\theta_2 \ldots d\theta_{i-1} d\theta_{i+1} \ldots d\theta_q \tag{6.2}$$

3. Numerical summaries of the posterior distribution. For example, letting θ be a scalar, to calculate the posterior mean, $E[\theta|y]$:

   $$E[\theta|y] = \int \theta p(\theta|y) d\theta \tag{6.3}$$

   Calculation of the probabilities such as $Pr(\theta > \theta^*)$, where θ is a scalar, also requires integration

   $$Pr(\theta > \theta^*) = \int_{\theta^*}^{\infty} p(\theta|y) d\theta \tag{6.4}$$

   Or with a continuous random variable, for arbitrarily small $\epsilon > 0$, "$Pr(\theta = \theta^*)$":

   $$Pr(\theta^* - \epsilon \leqslant \theta \leqslant \theta^* + \epsilon) = \int_{\theta^* - \epsilon}^{\theta^* + \epsilon} p(\theta|y) d\theta \tag{6.5}$$

4. The posterior predictive distribution for future or unobserved sample data. Letting y^{new} denote a future or unobserved value, the distribution is found by integration:

$$p(y^{new}|\mathbf{y}^{old}) = \int p(y^{new}|\theta, \mathbf{y}^{old})p(\theta|\mathbf{y}^{old})d\theta \quad (6.6)$$

   Reminder: In many cases, y^{new} is conditionally independent of $\mathbf{y}$: $p(y^{new}|\theta, \mathbf{y}^{old}) = f(y^{new}|\theta)$.

In most of the examples given so far, the integration can be carried out *analytically*, i.e., there were exact analytic solutions for the posterior distribution. This has been the case with the conjugate prior distributions:

- Beta prior for θ when $y \sim$ Binomial(n,θ),
- Dirichlet prior for $\theta_1, \theta_2, \ldots, \theta_k$ when $y_1, y_2, \ldots, y_k \sim$ Multinomial($n, \theta_1, \theta_2, \ldots, \theta_k$).
- Gamma prior for θ when $y \sim$ Poisson(θ).
- Gamma prior for θ when $y \sim$ Exponential(θ).
- Pareto prior for θ when $y \sim$ Uniform(0,θ).
- Normal prior for μ when $y \sim$ Normal(μ, σ^2), when σ^2 is known.
- Inverse Gamma prior for σ^2 when $y \sim$ Normal(μ, σ^2), when μ is known.
- Inverse Gamma-Conditional Normal prior for σ^2 and μ when $y \sim$ Normal(μ, σ^2).

However, for many of the models being used, exact solutions are the exception not the rule, and approximate solutions are used.

6.2 Overview of integration methods

We will denote the integral to be evaluated generically as $I(\cdot)$, where $\cdot$ is some value; e.g., $I(\mathbf{y}) = \int \pi(\theta)p(\mathbf{y}|\theta)d\theta$.

In most of the methods to be discussed below, one will arrive at an estimate of $I(\cdot)$, which will be denoted $\hat{I}(\cdot)$, thus $\hat{I}(\cdot) \approx I(\cdot)$.

A broad categorization of methods for calculating $I(\cdot)$ is *Deterministic* and *Stochastic*. With the deterministic methods, the result is a single constant value. If you carry out the method and get $\hat{I}(\mathbf{y})$ =3, and if someone else carries out the same deterministic method, they will get the same $\hat{I}(\mathbf{y})$ = 3. With stochastic methods, however, you might get $\hat{I}(\mathbf{y})$ = 3.02, and another person will get $\hat{I}(\mathbf{y})$=2.93, and if you were to repeat the method yourself a second time, you might get $\hat{I}(\mathbf{y})$ = 3.04.

Our focus is on integration for Bayesian inference where the integrals involve probability distributions. Methods used for Bayesian inference (but not just Bayesian inference), that lie within each of these broad categories, include the following.

- Deterministic
  - Numerical integration (quadrature)
  - Asymptotic approximations: Bayesian Central Limit Theorem, Laplace Approximation
- Stochastic
  - Monte Carlo integration with Independent samples, both sampling from the "right" distribution (e.g., Direct sampling) and sampling from the "wrong" distribution (e.g., Rejection sampling, Importance sampling, Sampling Importance Resampling (SIR), Sequential Importance Sampling)
  - Monte Carlo integration with Dependent samples from "wrong" distribution, especially Markov Chain Monte Carlo (MCMC) (e.g., Metropolis-Hastings, Gibbs sampling)

6.3 Example: Modelling time between hurricanes

(Example from Gordon Ross.) An island community regularly experiences large and damaging hurricanes. The government wishes to build a statistical model that can quantify the probability of such hurricanes occurring in the future. Over the last 50 years, there have been 21 recorded large hurricanes. The number of years which occur between each hurricane (known as the inter-event times, of which there are n=20) are:

```
hurricane.gaps <- c( 0.30, 4.61, 5.75, 0.24, 0.09, 0.18, 7.38, 1.20, 2.40, 0.18,
  0.02, 10.07, 0.23, 0.44, 3.34, 0.06, 0.01, 0.71, 0.06, 0.42)
```

The goal is to estimate the distribution of the time between hurricanes.

Let y_i denote the time between hurricane i and hurricane $i+1$. A simple sampling model for the y_i's, which assumes independence between the times till the event occurs, is the Exponential distribution, a commonly used model for times between events[1]. Letting $\mathbf{y}$=$(y_1, \ldots, y_n)$, the sampling distribution is

$$f(\mathbf{y}|\lambda) = \prod_{i=1}^{n} \lambda \exp(-y_i\lambda) = \lambda^n \exp(-\lambda \sum_{i=1}^{n} y_i) = \lambda^n \exp(-\lambda n \bar{y}) \tag{6.7}$$

which is the kernel for a Gamma distribution. Thus the conjugate prior for the exponential distribution is Gamma(α,β). The posterior distribution is then Gamma($\alpha + n$, $\beta + n\bar{y}$). For a "relatively uninformative" prior, let $\alpha = 0.01$ and $\beta = 0.01$, and with n=20 and $\bar{y}$=1.8845, the posterior is Gamma(20.01,37.70).

A diagnostic for goodness of fit is to compare the posterior predictive distribution to the observed distribution, e.g., the empirical density. The posterior predictive density for $\tilde{y}$, letting α'=$\alpha + n$ and β'=$\beta + n\bar{y}$:

$$p(y^{new}|\mathbf{y}^{old}) = \int \lambda \exp(-\lambda y^{new}) \frac{\beta'^{\alpha'}}{\Gamma(\alpha')} \lambda^{\alpha'-1} \exp(-\lambda\beta') d\lambda = \frac{\beta'^{\alpha'}\alpha'}{(\beta' + y^{new})^{\alpha'+1}}$$

Figure 6.1 compares the empirical density and the posterior predictive densities. Also plotted is the density using the mle for λ ($1/\bar{y}$= 0.5306), which is nearly identical to the posterior predictive density. The most important thing to note is the mismatch between the empirical and the modelled results, indicating poor goodness of fit.

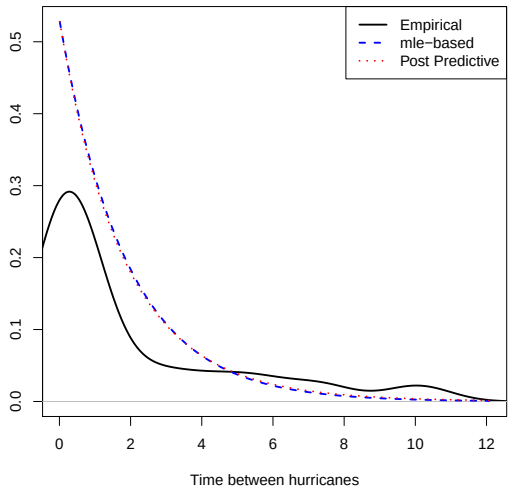
An alternative distribution that is more flexible and may fit the data better is the *Weibull*(α, β) distribution:

$$f(y|\alpha, \beta) = \frac{\alpha}{\beta}\left(\frac{y}{\beta}\right)^{\alpha-1} \exp\left(-(y/\beta)^{\alpha}\right) \tag{6.8}$$

The Weibull is often used to model between-event times. The exponential distribution is a special case with α=1 and β=$1/\lambda$. The parameter α in the Weibull thus provides more flexibility than the exponential.

[1] Recall the example from L1 notes for waiting time to see a teller in a bank.

Figure 6.1: Empirical, (mle) fitted, and post predictive densities for between hurricane times. (The mle fitted and posterior predictive are nearly exactly aligned.)



For a Bayesian analysis, the most general approach is to specify a joint prior for (α, β). If α is fixed, Γ^{-1} is conjugate for β. There is a conjugate prior for α, but it is awkward to work with. For simplicity (and to demonstrate some points about integration), we assume that β=1 and specify a non-conjugate prior for α, Gamma(κ, λ). The likelihood for α:

$$f(\mathbf{y}|\alpha, \beta = 1) \equiv L(\alpha|\mathbf{y}, \beta = 1) = \prod_{i=1}^{n} \alpha \left(y_i^{\alpha-1}\right) \exp\left(-y_i^{\alpha}\right) = \alpha^n \exp\left(-\sum_{i=1}^{n} y_i^{\alpha}\right) \prod_{i=1}^{n} y_i^{\alpha-1}$$

and the prior for α:

$$\pi(\alpha|\kappa, \lambda) = \frac{\lambda^{\kappa}}{\Gamma(\kappa)} \alpha^{\kappa-1} \exp(-\lambda\alpha)$$

Then the posterior:

$$\begin{aligned} p(\alpha|\mathbf{y}) &= \frac{p(\alpha, \mathbf{y})}{m(\mathbf{y})} = \frac{\pi(\alpha) f(\mathbf{y}|\alpha)}{\int \pi(\alpha) f(\mathbf{y}|\alpha) d\alpha} \\ &= \frac{\frac{\lambda^{\kappa}}{\Gamma(\kappa)} \alpha^{\kappa-1} \exp(-\lambda\alpha) \times \alpha^n \exp\left(-\sum_{i=1}^{n} y_i^{\alpha}\right) \prod_{i=1}^{n} y_i^{\alpha-1}}{\int \frac{\lambda^{\kappa}}{\Gamma(\kappa)} \alpha^{\kappa-1} \exp(-\lambda\alpha) \times \alpha^n \exp\left(-\sum_{i=1}^{n} y_i^{\alpha}\right) \prod_{i=1}^{n} y_i^{\alpha-1} d\alpha} \end{aligned} \tag{6.9}$$

where the marginal distribution for $\mathbf{y}$, the denominator of (6.9) is a complicated integral. Such complex integrals often occur for nonconjugate priors.

6.4 Deterministic Numerical Integration

While stochastic methods for estimating integrals are sometimes called “numerical” methods, here we will use the term “numerical” for non-stochastic, or deterministic

methods. These methods are also called quadrature. Such methods are a focus of a numerical methods course and we only briefly discuss basic and elementary forms, refer to some R software, and discuss some limitations of such methods.

6.4.1 Grid-based methods for calculating posterior expectations

Grid-based integration methods are a numerical method for calculating expectations. Consider a one-dimensional parameter, θ, where the objective is to calculate the expected value of a function of the parameter, $g(\theta)$.

$$E[g(\theta)|\mathbf{y}] = \int g(\theta)p(\theta|\mathbf{y})d\theta$$

For simplicity, assume that the support (domain) of θ is a finite interval, $[\theta_L, \theta_U]$. Partition the interval into m equal length intervals, let θ_i^* be the midpoint of the ith interval. The expected value can be estimated as follows.

$$\widehat{E}[g(\theta)|\mathbf{y}] = \sum_{i=1}^{m} g(\theta_i^*)W_i$$

where

$$W_i = \frac{\pi(\theta_i^*)f(\mathbf{y}|\theta_i^*)}{\sum_{j=1}^{m}\pi(\theta_j^*)f(\mathbf{y}|\theta_j^*)}$$

6.4.2 Simple numerical integration example

Consider a one-dimensional integral of the form:

$$\int_a^b f(x)dx$$

which may or may not have an analytic solution. The gist of many of the numerical solutions is to first partition $[a, b]$ into n subintervals:

$$[x_0 = a, x_1], [x_2, x_3], \ldots, [x_{n-1}, x_n = b]$$

Thus breaking the integration problem into n components:

$$\int_a^b f(x)dx = \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} f(x)dx$$

Then the integral over each subinterval is approximated using some relatively simple formula or rule. For example, the integral within a subinterval is estimated by the area of a rectangle with width equal to the width of the subinterval and height equal to the average of the function evaluated at the endpoints of the interval:

$$\int_{x_i}^{x_{i+1}} f(x)dx \approx \text{ width * } \overline{\text{height}} = (x_{i+1} - x_i) \times \frac{f(x_i) + f(x_{i+1})}{2}$$

This particular approach is called the Trapezoid rule.

For example consider the integral $\int_0^3 \frac{1}{2}e^{-x/2}dx$, which does have an analytic solution (the integrand is the pdf for an Exponential(1/2) random variable):

$$\int_0^3 \frac{1}{2}e^{-x/2}dx = -e^{-x/2}|_0^3 = 1 - \exp(-1.5) = 0.7768698$$

Breaking [0,3] into four intervals of equal length (3/4=0.75) and applying the trapezoid rule:

$$\begin{aligned}\int_0^3 \frac{1}{2}e^{-x/2}dx &= \int_0^{0.75} \frac{1}{2}e^{-x/2}dx + \int_{0.75}^{1.50} \frac{1}{2}e^{-x/2}dx + \int_{1.5}^{2.25} \frac{1}{2}e^{-x/2}dx + \int_{2.25}^{3} \frac{1}{2}e^{-x/2}dx \\ &\approx \frac{3}{4}\frac{0.5(e^{-0.75/2} + e^{-0/2})}{2} + \frac{3}{4}\frac{0.5(e^{-1.5/2} + e^{-0.75/2})}{2} + \frac{3}{4}\frac{0.5(e^{-2.25/2} + e^{-1.5/2})}{2} + \frac{3}{4}\frac{0.5(e \\ &= 0.3163667 + 0.2174355 + 0.1494411 + 0.1027092 = 0.7859525\end{aligned}$$

Increasing the number of subintervals improves the approximation, e.g., with 20 subintervals the estimate is 0.777234.

The general solution.

- Select $m + 1$ points within each subinterval; for interval i, $[x_i, x_{i+1}]$: the m points are x_{ij}, $j = 0, 1, 2, \ldots, m$:

$$[x_i = x_{i0} \leqslant x_{i1} < x_{i2} < \ldots < x_{i,m-1} \leqslant x_{i,m} = x_{i+1}]$$

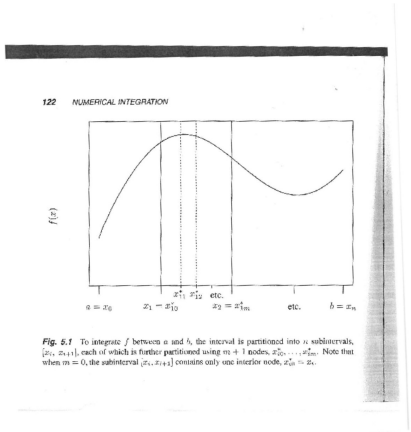
  Figure 6.2 shows the subintervals and points within an interval.

- Evaluate the function at each of these points, $f(x_{ij})$.
- Estimate the integral with a weighted combination of the function evaluations

$$\int_{x_i}^{x_{i+1}} f(x)dx \approx \sum_{j=1}^{m} w_{ij} f(x_{ij})$$

With the Trapezoid rule, m=1, and the $m + 1$=2 selected points are the endpoints of the subinterval. Two other simple rules are the Riemann rule, m=0, where one of the endpoints of the subinterval is selected, and Simpson's rule, m=2, where the endpoints x_i and x_{i+1} are used as well as the midpoint, $(x_i + x_{i+1})/2$. Figure 6.3 shows how the three rules differ.

Figure 6.2: Partitions of an integral with points within an interval. From Givens and Hoeting, Computational Statistics, 2013, p130.



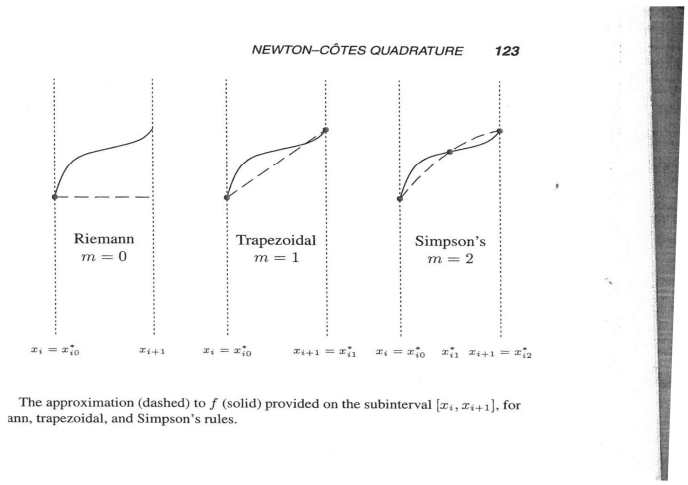
122 NUMERICAL INTEGRATION

$f(x)$

$a = x_0$ $x_1 = x_{10}^*$ x_{11}^* x_{12}^* etc. $x_2 = x_{1m}^*$ etc. $b = x_n$

Fig. 5.1 To integrate f between a and b, the interval is partitioned into n subintervals, $[x_i, x_{i+1}]$, each of which is further partitioned using $m+1$ nodes, $x_{i0}^*, \ldots, x_{im}^*$. Note that when $m = 0$, the subinterval $[x_i, x_{i+1}]$ contains only one interior node, $x_{i0}^* = x_i$.

Figure 6.3: Three rules: Reimann, Trapezoid, and Simpson, for evaluation of the integral over a subinterval. From Givens and Hoeting, Computational Statistics, 2013, p131.



NEWTON–CÔTES QUADRATURE 123

Riemann
$m = 0$

Trapezoidal
$m = 1$

Simpson's
$m = 2$

$x_i = x_{i0}^*$ x_{i+1} $x_i = x_{i0}^*$ $x_{i+1} = x_{i1}^*$ $x_i = x_{i0}^*$ x_{i1}^* $x_{i+1} = x_{i2}^*$

The approximation (dashed) to f (solid) provided on the subinterval $[x_i, x_{i+1}]$, for ann, trapezoidal, and Simpson's rules.

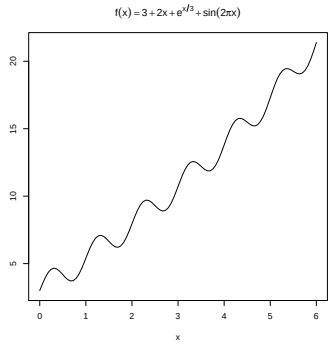
Demonstration with R. The above methods, while simple to understand, are relatively slow, and inefficient compared to other more sophisticated methods, where, for example, subinterval widths need not be equal[2]. R has a built-in function called `integrate` that can be applied to one dimensional integrals, which uses a more sophisticated procedure. The R package `pracma` has a similar function called `integral`.

To demonstrate the usage of the R functions, consider the integral:

$$\int_0^5 2 + 2x + \exp(x/3) + \sin(2\pi x)\mathrm{d}x$$

which can be evaluated analytically and equals 47.88347. The function is plotted in Figure 6.4.

Figure 6.4: Function, $f(x) = 2 + 2x + \exp(x/3) + \sin(2\pi x)$, to demonstrate R's integrate and integral functions.



The R code that describes the function, plots it over (0,6), and estimates the integral using `integral` and `integrate` functions is shown below.

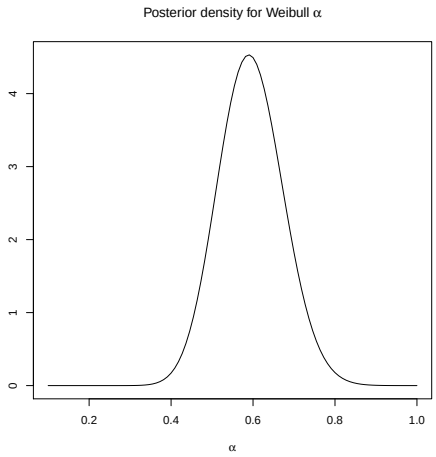
```
# example function to be integrated over (0,5)
f <- function(x) {
   2+2*x+exp(x/3) + sin(2*pi*(x))
}
x.seq <- seq(0,6,length=100)
plot(x.seq,f(x.seq),xlab="x",ylab="",
 main= expression(f(x)==paste(3+2*x+e^{x/3}+sin(2*pi*x))),
     type="l")

antideriv <- function(x) 2*x+x^2+3*exp(x/3)-(1/2*pi)*cos(2*pi*x)
true.value <- antideriv(5)-antideriv(0)

cat("true value=",true.value,
    "pracma:",integral(f=f,0,5),
```

[2]A highly readable chapter on numerical integration can be found in Computational Statistics, 2nd Edition by Givens and Hoeting (2013).

```
    "base:",integrate(f=f,0,5)$value,"\n")
#  true value= 47.88347 pracma: 47.88347 base: 47.88347
```

Demonstration with Weibull example. The normalising constant was estimated to be 2.008407e-14. The posterior distribution for α is shown in Figure 6.5.

Figure 6.5: Posterior distribution for the α parameter, $p(\alpha|\mathbf{y})$, in the Weibull example of times between hurricanes based on R's integral function.



6.4.3 Multiple integrals

To numerically integrate a double integral where the region of integration is a rectangle, $(a \leqslant x \leqslant b)$ and $(c \leqslant y \leqslant d)$, i.e.,

$$\int_a^b \int_c^d f(x, y) dy dx,$$

the double integral can be rewritten as:

$$\int_a^b g(x) dx$$

where

$$g(x) = \int_c^d f(x, y) dy$$

Quadrature methods can be applied in an iterative manner: estimating $g(x)$ by quadrature over $[c, d]$, then given a set of $g(x)$ values (over the sub-intervals of $[a, b]$) carry out quadrature over $[a, b]$. If there are n subintervals for x and y, then there are n^2 evaluations. Two practical problems:

- With higher dimension integrals the number of evaluations can become infeasible. Imagine a joint posterior $(\theta_1, \theta_2, \ldots, \theta_9)$ with n=5 subintervals: 5^9 = 1,953,125 evaluations are needed.

- The integration region will often not be rectangular (or a hyper-rectangle, $[a, b], [c, d], [e, f]$), and the problem becomes even more difficult[3].

Thus while understanding the principles of numerical integration are useful, and are sometimes used for small dimension θ, they "do not scale well with the number of parameters k" (Reich and Ghosh, p 74) and are not as commonly used as Monte Carlo methods.

6.5 Normal approximation to posterior

Suppose that the data $y_i \overset{iid}{\sim} f(y|\theta)$. Letting $\mathbf{y} = (y_1, \ldots, y_n)$, the joint probability is $f(\mathbf{y}|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$. If n is relatively large, the likelihood will be relatively peaked and the small changes in the prior will have little effect on the posterior. Assume that a mode for the posterior distribution, denoted $\hat{\theta}^p$, exists (where in the multivariate case θ is a vector). Further assume that $\pi(\theta)f(\mathbf{y}|\theta)$ is positive and twice differentiable at the mode. Then under suitable regularity conditions the posterior distribution will be approximately normal with mean $\hat{\theta}^p$ and covariance matrix equal to the negative of the inverse of the second derivative matrix of the log posterior distribution evaluated at $\hat{\theta}^p$:

$$p(\theta|\mathbf{y}) \approx \text{Multivariate Normal}\left(\hat{\theta}^p, (I^p(\mathbf{y}))^{-1}\right) \tag{6.10}$$

where $I^p(\mathbf{y})$ is the "generalised" observed Fisher information matrix:

$$I^p(\mathbf{y}) = -\left[\frac{d^2}{d\theta_i d\theta_j} \ln\left(\pi(\theta)f(\mathbf{y}|\theta)\right)\right]_{\theta=\hat{\theta}^p} \tag{6.11}$$

The result in (6.10) is sometimes referred to as the Bayesian Central Limit Theorem.

To provide some insight into this result (but not a rigorous proof[4]), consider the univariate θ case and let $l(\theta)$ denote $\ln(\pi(\theta)f(\mathbf{y}|\theta))$. The posterior distribution, $p(\theta|\mathbf{y})$, is

[3] See http://www.aip.de/groups/soe/local/numres/bookfpdf/f4-6.pdf for more discussion of these difficulties.

[4] Note in particular that the demonstration is working with a single parameter θ, not a vector of parameters.

approximated by a 2nd order Taylor series expansion of $l(\theta)$ around $\hat{\theta}^p$:

$$
\begin{aligned}
p(\theta|\mathbf{y}) \propto \pi(\theta) f(\mathbf{y}|\theta) &= \exp\left[\ln(\pi(\theta) f(\mathbf{y}|\theta))\right] = \exp(l(\theta)) \\
&\approx \exp\left\{ l(\hat{\theta}^p) + \frac{d}{d\theta} l(\theta)|_{\theta=\hat{\theta}^p} (\theta - \hat{\theta}^p) + \frac{1}{2}\frac{d^2}{d\theta^2} l_{\theta=\hat{\theta}^p} \left(\theta - \hat{\theta}^p\right)^2 \right\} \\
&= \exp\left\{ l(\hat{\theta}^p) + \frac{1}{2}\frac{d^2}{d\theta^2} l_{\theta=\hat{\theta}^p} \left(\theta - \hat{\theta}^p\right)^2 \right\} \quad \text{[because the 2nd term is 0 at the mode]} \\
&= \exp\left\{ l(\hat{\theta}^p) - \frac{1}{2} I^p(\mathbf{y}) \left(\theta - \hat{\theta}^p\right)^2 \right\} \quad \text{[because the likelihood dominates over the prior]} \\
&= \exp\left\{ l(\hat{\theta}^p) - \frac{1}{2}\frac{\left(\theta - \hat{\theta}^p\right)^2}{(I^p(\mathbf{y}))^{-1}} \right\} \\
&\propto \exp\left\{ -\frac{1}{2}\frac{\left(\theta - \hat{\theta}^p\right)^2}{(I^p(\mathbf{y}))^{-1}} \right\}
\end{aligned}
$$

The last term is the kernel of Normal $\left(\hat{\theta}^p, (I^p(\mathbf{y}))^{-1}\right)$.

Remarks.

- The posterior mode, $\hat{\theta}^p$, is called the *Maximum a posteriori* or *MAP* estimator. Note that calculating its value typically requires differentiation. Thus the problem of integration has been replaced by a problem of differentiation, in particular an *optimization* problem.
- The process of solving an integral problem, $\int f(x)dx$, where $f(x)$ is postive valued, by first rewriting the integrand as $\exp(\log(f(x))$, and then approximating $\log(f(x))$ by a second-order Taylor expansion at the mode of $f(x)$ is known as a *Laplace approximation*. The Laplace approximation method applies to higher dimensional integrals and is the basis of the popular model fitting packages INLA (Integrated Nested Laplace Approximations) and Template Model Builder (TMB).

Beta-Binomial example. The data y are Binomial(n,θ) and the prior for θ is Beta(α, β). Then

$$
\ln(p(\theta|y)) \equiv l(\theta) \propto \ln\left(\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}\right) = (y+\alpha-1)\ln(\theta) + (n-y+\beta-1)\ln(1-\theta)
$$

The mode of the posterior distribution is found by differentiating $l(\theta)$ with respect to θ:

$$
\frac{dl(\theta)}{d\theta} = \frac{y+\alpha-1}{\theta} - \frac{n-y+\beta-1}{1-\theta}
$$

setting the result equal to 0 and solving for θ, yielding the MAP estimate

$$
\hat{\theta}^p = \frac{y+\alpha-1}{n+\alpha+\beta-2}
$$

Then $I^p(y)$ is:

$$I^p(y) = -\frac{d^2}{d\theta^2} l(\theta)|_{\hat{\theta}} = \frac{\alpha + y - 1}{\theta^2} + \frac{\beta + n - y - 1}{(1-\theta)^2}|_{\hat{\theta}}$$
$$= \frac{(\alpha + \beta + n - 2)^3}{(\alpha + y - 1)(\beta + n - y - 1)}$$

Then

$$p(\theta|y) \approx \text{Normal}\left(\frac{y + \alpha - 1}{n + \alpha + \beta - 2}, \frac{(\alpha + y - 1)(\beta + n - y - 1)}{(\alpha + \beta + n - 2)^3}\right)$$

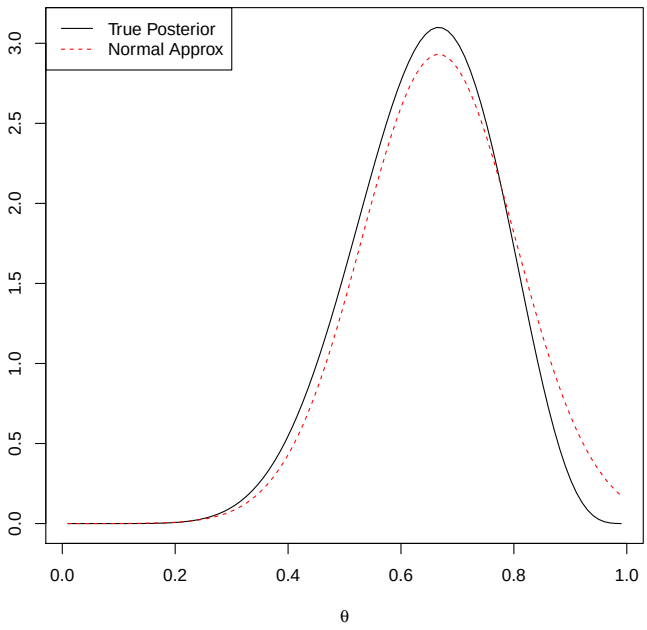
Note that if α=β=1,

$$p(\theta|y) \approx \text{Normal}\left(\frac{y}{n}, \frac{y(n-y)}{n^3}\right) \equiv \text{Normal}\left(\hat{p}, \frac{\hat{p}(1-\hat{p})}{n}\right)$$

where $\hat{p} = y/n$, the familiar sample proportion.

To examine the quality of the approximation, the prior for θ is Beta(1,3) (thus an expected value of 0.25) and n=10 with y=8 successes. The posterior for θ is Beta(9,5) (thus an expected value of 0.643). The mode, $\hat{\theta}$, is 0.667 and the approximate variance, $I^p(y)^{-1}$ is 0.0185. The true posterior density and the normal approximation are plotted in Figure 6.6. The modes are quite similar but the normal approximation attaches too much probability to $\theta > 0.8$. The differences are apparent in the lower and upper quantiles: e.g., the true 0.025 quantile is 0.386 compared to the normal approximation value of 0.400, and the true 0.975 quantile is 0.861 compared to 0.933 for the normal approximation.

Figure 6.6: Posterior distribution for θ (Beta(9,5)) and Normal(0.667,0.0185) approximation.



Homework Week 6: Bayesian Theory.

Optional Reading

1. Numerical Integration. Chapter 5, Section 1 of Computational Statistics, 2nd Ed (2013), Givens and Hoeting, pp 129-139.

Exercise: not to turn in

1. Given a Poisson(θ) sampling distribution for the data, $\mathbf{y}$=($y_1, \ldots, y_n$) and a Gamma(α, β) prior for θ, the posterior distribution is known and tractable. Use the Normal distribution approximation for the posterior and compare it to the exact distribution.

7 Bayesian Computation: Independent Monte Carlo Methods

7.1 Initial Look at Monte Carlo integration methods

While we will look at a variety of integration methods, both deterministic and stochastic, the most commonly used procedures for complex Bayesian models are Monte Carlo methods. As a preview of the Monte Carlo methods, consider the following problem. Suppose one wants to calculate the expected value of a random variable, θ:

$$E[\theta] = \int \theta p(\theta) d\theta \tag{7.1}$$

Suppose that calculation of the integral is difficult but one can relatively easily generate an arbitrarily large, and for now, independent, sample of size N from the probability distribution $p(\theta)$:

$$\theta^1, \theta^2, \ldots, \theta^N$$

Then a Monte Carlo estimate of the integral (7.1):

$$\hat{E}[\theta] = \frac{1}{N} \sum_{i=1}^{N} \theta^i \tag{7.2}$$

This is simply the sample average from the random (Monte Carlo) sample. If $E[\theta]$ exists (is finite), then by the strong Law of Large Numbers, $\hat{E}[\theta]$ converges with probability 1, which means:

$$Pr(\lim_{N \to \infty} \hat{E}[\theta] = E[\theta]) = 1 \tag{7.3}$$

Many Bayesian integrals can be viewed as expectations. Two examples.

Probabilities: $Pr(\theta > \theta^*)$. Calculating a probability over some interval(s) can be viewed as the expected value of the indicator random variable, $I(\theta > \theta^*)$:

$$E[I(\theta > \theta^*)] = \int p(\theta|\mathbf{y}) I(\theta > \theta^*) d\theta = \int_{\theta^*}^{\infty} p(\theta|\mathbf{y}) d\theta$$

Thus if one can generate a sample from $p(\theta|\mathbf{y})$, then

$$\hat{E}[I(\theta > \theta^*)] = \frac{1}{N}\sum_{i=1}^{N} I(\theta^i > \theta^*)$$

Normalising constants. The normalising constant for the posterior distribution in the denominator, namely $m(\mathbf{y})$, can be viewed as an expected value:

$$m(\mathbf{y}) = E[f(\mathbf{y}|\theta)] = \int f(\mathbf{y}|\theta)\pi(\theta)d\theta$$

If one draws a large independent sample of θ from the prior distribution $\pi(\theta)$, the normalising constant can be estimated as follows.

$$\hat{E}[f(\mathbf{y}|\theta)] = \frac{1}{N}\sum_{i=1}^{N} f(\mathbf{y}|\theta^i)$$

Demonstration with Weibull example. Here we demonstrate Monte Carlo integration for the posterior distribution of α in the Weibull model for times between hurricanes (see Lecture Notes 6), in particular to estimate the normalising constant, $m(\mathbf{y})$. The integral to estimate is then

$$\int \pi(\alpha)f(\mathbf{y}|\alpha)d\alpha = \int \left[\frac{\lambda^\kappa}{\Gamma(\kappa)}\alpha^{\kappa-1}\exp(-\lambda\alpha)\right] \times \alpha^n \exp\left(-\sum_{i=1}^{n} y_i^\alpha\right)\prod_{i=1}^{n} y_i^{\alpha-1}d\alpha$$

Given a random sample of α^i, i=1,. . . , N,from a Gamma(κ, λ) distribution, the integral can be estimated as:

$$\int \pi(\alpha)f(\mathbf{y}|\alpha)d\alpha \approx \frac{1}{N}\sum_{j=1}^{N} (\alpha^j)^n \exp\left(-\sum_{i=1}^{n} y_i^{\alpha^j}\right)\prod_{i=1}^{n} y_i^{\alpha^j-1} \tag{7.4}$$

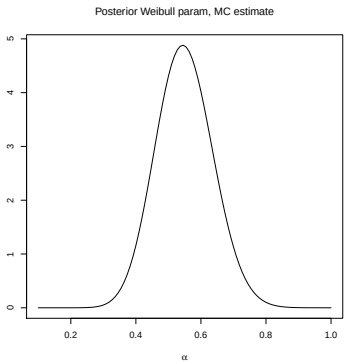
Let the hyperparameters for prior Gamma distribution be κ=λ=0.1. The R code for estimating the normalising constant:

```
hurricane.gaps <- c(0.30, 4.61, 5.75, 0.24, 0.09, 0.18, 7.38, 1.20, 2.40, 0.18,
                    0.02, 10.07, 0.23, 0.44, 3.34, 0.06, 0.01, 0.71, 0.06, 0.42)
n <- length(hurricane.gaps)

#-- Monte Carlo estimate of the normalizing constant with the Weibull --
# simulating from the prior, Gamma(kappa,lambda)
set.seed(7301)
N          <- 1000000
kappa      <- lambda <- 0.01
alpha.star <- rgamma(n=N,shape=kappa,rate=lambda)
```

```
#--- Calculating the sum by looping over the N samples
 integ.est <- 0
 for(i in 1:N) {
  #integ.est <- integ.est + (1/N)*(alpha.star[i]^n *
  #                                exp(-sum(hurricane.gaps^(alpha.star[i]))) *
  #                                exp(sum((alpha.star[i]-1)*log(hurricane.gaps))))
  integ.est <- integ.est + (1/N)*(exp(n*log(alpha.star[i]) -
                                      sum(hurricane.gaps^(alpha.star[i])) +
                                      sum((alpha.star[i]-1)*log(hurricane.gaps))))
 }
 cat("Estimate of normalising constant=",integ.est,"\n")
 # Estimate of normalising constant= 1.992099e-14
 #   Compare to LN 6 numerical result with pracma: norm constant= 2.008407e-14
```

Additional R code is in Appendix 7.7.1 that also generated an estimate of the posterior distribution and the posterior mean for α, which was 0.5522097. A plot of the estimated posterior distribution is shown in Figure 7.1.

Figure 7.1: Posterior distribution for the α parameter in the Weibull example of times between hurricanes based on Monte Carlo integration with samples taken from the prior distribution for α



Important implementation issue: Note how small the normalising constant is. This is not unusual for even moderate size samples such as this, n=20. Computational overflows and underflows are a common problem, e.g., $\prod_{i=1}^{n} y_i^{\alpha^i - 1}$ may be a number that is too large or too small for the computer. The commands that have been commented out, those preceded by #, carried out the calculation in a manner paralleling equation 7.4, but led to computational errors:

```
Posterior distribution for the Weibull example of times between hurricanes based on
Estimate of normalising constant= NaN
```

A general solution is to work with logarithms for intermediate calculations and then to exponentiate as late as possible in the calculations. A quote from Gelman et al (BDA):

> To avoid computational overflows and underflows, one should compute with the logarithms of posterior densities whenever possible. Exponenti-

ation should be performed only when necessary and as late as possible; for example, in the Metropolis algorithm, the required ratio of two densities should be computed as the exponential of the difference of the log densities.

7.2 Direct Sampling

At the heart of the Monte Carlo procedures is the generation of random variables that will be used to either

- Estimate an integral or
- Yield a sample from a desired distribution.

We will call the desired distribution the *target distribution*. For Bayesian inference, the target distribution is often the posterior distribution, $p(\theta|\mathbf{y})$, and the integrals of interest can be viewed as expected values:

$$E[h(\theta|\mathbf{y})] = \int h(\theta)p(\theta|\mathbf{y})d\theta$$

Common choices for $h(\theta)$, some of which were discussed previously, include:

$h(\theta) = \theta$, the posterior mean
$h(\theta) = (\theta - E[\theta])^2$, the posterior variance
$h(\theta) = I(a \leqslant \theta \leqslant b)$, the probability that θ is between a and b

As seen earlier, if we can generate an independent sample of size N from $p(\theta|\mathbf{y})$, θ^1, θ^2, ..., θ^N, we can estimate the expected values:

$$E[h(\theta|\mathbf{y})] \approx \hat{E}[h(\theta|\mathbf{y})] = \frac{1}{N}\sum_{i=1}^{N} h(\theta^i)$$

And such estimates are consistent, i.e., they converge to $E[h(\theta|\mathbf{y})]$ by the strong law of large numbers.

We will call $\hat{E}[h(\theta|\mathbf{y})] = \frac{1}{N}\sum_{i=1}^{N} h(\theta^i)$ a Monte Carlo estimate of $E[h(\theta|\mathbf{y})]$.

Such estimates will not exactly equal $E[h(\theta|\mathbf{y})]$, i.e., there will be "Monte Carlo" error:

$$error = \frac{1}{N}\sum_{i=1}^{N} h(\theta^i) - E[h(\theta|\mathbf{y})]$$

Comments.

- Given iid data, the Central Limit Theorem can be applied to approximate the sampling distribution of $\widehat{E}[h(\theta|\mathbf{y})]$:

  $$\widehat{E}[h(\theta|\mathbf{y})] \dot{\sim} \text{Normal}\left(E[h(\theta|\mathbf{y})], \frac{\sigma^2}{N}\right)$$

  where σ^2/N is the Monte Carlo variance, and $\sigma/\sqrt{N}$ is called the Monte Carlo error.

- Assuming that the Monte Carlo estimate is unbiased, σ^2 can be estimated via the sample variance[1].

$$\begin{aligned}\widehat{\sigma^2} = \widehat{V}\left(h(\theta|\mathbf{y})\right) &= \frac{1}{N-1}\sum_{i=1}^{N}\left(h(\theta^i) - \widehat{E}[h(\theta|\mathbf{y})]\right)^2 \\ &= \frac{1}{N-1}\left[\sum_{i=1}^{N} h(\theta^i)^2 - N\left(\frac{1}{N}\sum_{i=1}^{N} h(\theta^i)\right)^2\right] \\ &\approx \frac{1}{N}\sum_{i=1}^{N} h(\theta^i)^2 - \left(\frac{1}{N}\sum_{i=1}^{N} h(\theta^i)\right)^2\end{aligned}$$

The second line is simply an approach to estimating the variance in a single pass through the data, and in the last line we note that Monte Carlo methods usually involve large N, so dividing by $N-1$ is practically indistinguishable from dividing by N. The resulting Monte Carlo error is $\widehat{\sigma}/\sqrt{N}$.

Note that the single-pass version of the computation can be numerically unstable, but that problem can be reduced by shifting the data. Let h_0 be a value in the range of the simulated values, such as the first sample, $h(\theta^1)$. Then

$$\begin{aligned}\sum_{i=1}^{N}\left(h(\theta^i) - \widehat{E}[h(\theta|\mathbf{y})]\right)^2 &= \sum_{i=1}^{N}\left(h(\theta^i) - h_0 + h_0 - \widehat{E}[h(\theta|\mathbf{y})]\right)^2 \\ &= \sum_{i=1}^{N}[h(\theta^i) - h_0]^2 - N\left(\frac{1}{N}\sum_{i=1}^{N}[h(\theta^i) - h_0]\right)^2\end{aligned}$$

Another single-pass approach is to use a recursive estimator formulation. Let $\widehat{m}_n$ and $\widehat{\sigma^2}_n$ be the expectation and variance estimate based on n samples. Then

$$\widehat{m}_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} h(\theta^i) = \frac{1}{n+1}\left[n\widehat{m}_n + h(\theta^{n+1})\right],$$

$$\begin{aligned}\widehat{\sigma^2}_{n+1} &= \frac{1}{n}\sum_{i=1}^{n+1}\left(h(\theta^i) - \widehat{m}_{n+1}\right)^2 = \frac{1}{n}\left(\sum_{i=1}^{n}\left(h(\theta^i) - \widehat{m}_{n+1}\right)^2 + \left[h(\theta^{n+1}) - \widehat{m}_{n+1}\right]^2\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}\left(h(\theta^i) - \widehat{m}_n + \widehat{m}_n - \widehat{m}_{n+1}\right)^2 + \left[h(\theta^{n+1}) - \widehat{m}_{n+1}\right]^2\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n}\left(h(\theta^i) - \widehat{m}_n\right)^2 + 2\left(\widehat{m}_n - \widehat{m}_{n+1}\right)\sum_{i=1}^{n}\left(h(\theta^i) - \widehat{m}_n\right) + n\left(\widehat{m}_n - \widehat{m}_{n+1}\right)^2 + \left[h(\theta^{n+1}) - \widehat{m}_{n+1}\right]^2\right) \\ &= \frac{1}{n}\left((n-1)\widehat{\sigma^2}_n + n\left(\widehat{m}_n - \widehat{m}_{n+1}\right)^2 + \left[h(\theta^{n+1}) - \widehat{m}_{n+1}\right]^2\right) \\ &= [\ldots] = \frac{n-1}{n}\widehat{\sigma^2}_n + \frac{n+1}{n^2}\left[h(\theta^{n+1}) - \widehat{m}_{n+1}\right]^2\end{aligned}$$

[1]Note: estimates of a population variance, σ^2, with s^2 typically involve division by $N-1$: $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$. With Monte Carlo methods, N is often large enough that division by N instead of $N-1$ is deemed adequate.

This recursive method guarantees a numerically non-negative estimate of the variance.

- The Monte Carlo variance can (in theory) be reduced to some arbitrary value by increasing N. (In practice, the size of N to achieve a desired precision might be impractical, e.g., take too long to achieve.)

Example A. $\theta \sim$ from a Gamma(3,0.2), thus $\mathrm{E}[\theta]$= 3/0.2=15 (and $\mathrm{Var}[\theta] = 3/0.04 = 75$). A sample of size N is simulated from this distribution and the sample average is used to estimate $\mathrm{E}[\theta]$:

$$\hat{\mathrm{E}}[\theta] = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \theta^{i}$$

By the Central Limit Theorem

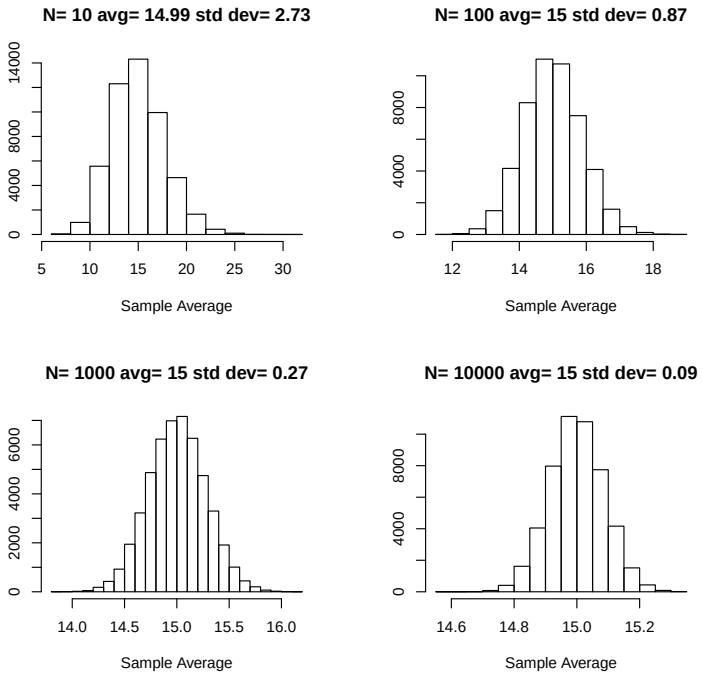
$$\hat{\mathrm{E}}[\theta] \dot{\sim} \text{Normal}\left(\mathrm{E}[\theta] = 15, \frac{\sigma^2}{\mathrm{N}} = 75/\mathrm{N}\right)$$

The table below shows the average of the $\hat{\mathrm{E}}[\theta]$, the empirical standard deviation of $\hat{\mathrm{E}}[\theta]$, and the theoretical standard deviation for sample sizes N=10, 100, 1000, and 10000 based upon 50,000 simulations of each sample size.

N	Average $\hat{\mathrm{E}}[\theta]$	Std Dev of $\hat{\mathrm{E}}[\theta]$	Empirical Std Dev
10	14.992	2.739	2.732
100	14.998	0.866	0.866
1000	15.002	0.274	0.274
10000	15.000	0.086	0.087

Figure 7.2 shows the distribution of $\hat{\mathrm{E}}[\theta]$.

Figure 7.2: Sampling distribution of $\hat{\mathrm{E}}[\theta]$ from a Gamma(3,0.2) for sample sizes N=10, 100, 1000, and 10000 based upon 50,000 simulations of each sample size. True value, $\mathrm{E}[\theta]$=15.



Example B. Now we estimate $\Pr(\theta < 5)$=$\mathrm{E}[\mathrm{I}(\theta < 5)]$:

$$\hat{\mathrm{E}}[\mathrm{I}(\theta < 5)] = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{I}(\theta^{\mathrm{i}} < 5)$$

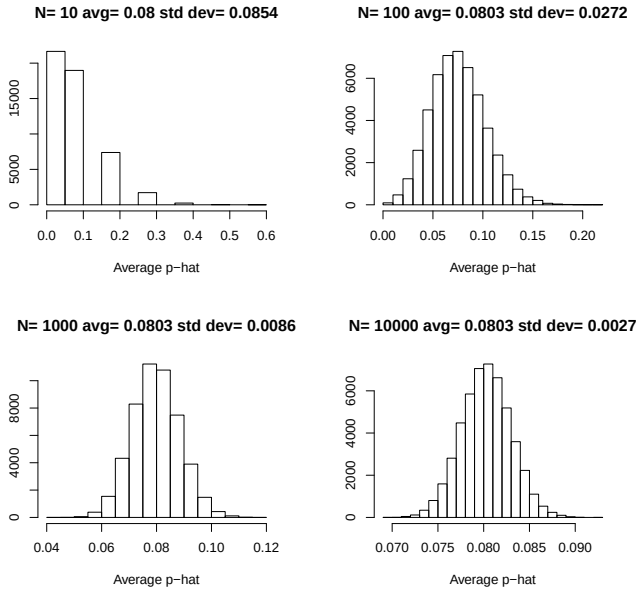
where the true value is $\mathrm{pgamma}(5, \mathrm{shape} = 3, \mathrm{rate} = 0.2)$=0.08030, and the theoretical standard error of the estimate is $\sqrt{\mathrm{p} * (1 - \mathrm{p})/\mathrm{N}}$, where p=0.08030. The table below shows the average of the $\hat{\mathrm{E}}[\theta]$, the empirical standard deviation of $\hat{\mathrm{E}}[\theta]$, and the theoretical standard deviation for sample sizes N=10, 100, 1000, and 10000 based upon 50,000 simulations of each sample size. R code for both examples is in Appendices 7.7.2 and 7.7.3.

N	Average $\hat{\mathrm{E}}[\mathrm{I}(\theta < 5)]$	Std Dev of $\hat{\mathrm{E}}[\mathrm{I}(\theta < 5)]$	Empirical Std Dev
10	0.080	0.085	0.086
100	0.080	0.027	0.027
1000	0.080	0.009	0.009
10000	0.080	0.003	0.003

Figure 7.3 shows the distribution of $\hat{\mathrm{E}}[\mathrm{I}(\theta < 5)]$.

Dependent samples. When the samples generated from $\mathrm{p}(\theta|\mathbf{y})$ are dependent samples, then:

Figure 7.3: Sampling distribution of $\hat{\Pr}(\theta < 5) = \hat{\mathrm{E}}[\mathrm{I}(\theta < 5)]$ from a Gamma(3,0.2) for sample sizes N=10, 100, 1000, and 10000 based upon 50,000 simulations of each sample size. True value, $\Pr(\theta < 5)$=0.08030.



- For the methods that we will be considering, the Monte Carlo estimates will still be consistent, i.e., they will converge in probability to the expected value—but it's not the standard law of large numbers argument used for independent samples.
- Calculation of the Monte Carlo error is more complicated; e.g., time series methods are sometimes used.

7.3 Inverse Probability Integral Transform method

For the moment we ignore the particular problem of calculating an integral and consider a general method that is occasionally feasible for simulating random variables from specific probability distributions, the inverse probability integral transform (inverse PIT)[2].

7.3.1 Continuous random variable

We will focus on univariate random variables and begin with continuous random variables. First recall that given a random variable Y and a function f, we can define new

[2]R, of course, has built-in functions for many distributions, runif, rnorm, rbeta, rgamma, $\mathrm{rweibull}$, and the aim here is to introduce an approach that can be useful at times.

random variable X equal to $f(Y)$. For example, we've demonstrated that with $\theta \sim$ Uniform(0,20) and defining $\phi = g(\theta) = \sqrt{\theta}$.

- Let Y be a continuous random variable with cumulative distribution function, $F_Y(y)$:

$$F_Y(y) = \int_{-\infty}^{y} f(y)dy$$

  where $f(y)$ is the probability density function for Y. Note that for whatever value of y passed to $F_Y(y)$ the value is on [0,1].

- Define a new random variable $U = F_Y(Y)$.

- Claim: $U \sim$ Uniform(0,1) distribution.
  Proof: We prove this by deriving the cumulative distribution function for U:

$$\begin{aligned} F_U(u) = Pr(U \leqslant u) &= Pr(F_Y(Y) \leqslant u) \\ &= Pr(Y \leqslant F_Y^{-1}(u)) \\ &= F_Y(F_Y^{-1}(u)) = u \end{aligned}$$

  which is the CDF for a Uniform(0,1) random variable.

The Inverse PIT method. This result has a corollary that can be used to "go the other way": transform a Uniform random variable using the inverse of the CDF of Y, $g(U) = F_Y^{-1}(U)$, and then the resulting transformed random variable is from the distribution for Y.

Proof: Let $Z = g(U)$, the CDF for Z evaluated at y:

$$Pr(Z \leqslant y) = Pr(F_Y^{-1}(U) \leqslant y) = Pr(U \leqslant F_Y(y)) = F_Y(y)$$

namely the CDF for the random variable Y.

Thus if one can evaluate the inverse cdf for a random variable Y, F^{-1}, one can generate a sample from that distribution by simulating a Uniform(0,1) random variable, u, and setting y=$F^{-1}(u)$.

A simple example is generating a random variable from the Exponential(λ) distribution. The cdf for the exponential:

$$F_Y(y) = \int_0^y \lambda \exp(-\lambda x)dx = 1 - \exp(-\lambda y)$$

Letting $u = F_Y(y)$, the inverse, $F^{-1}(u)$, is found as follows:

$$\begin{aligned} u = F_Y(y) = 1 - \exp(-\lambda y) &\Rightarrow \\ \exp(-\lambda y) = 1 - u &\Rightarrow \\ -\lambda y = \ln(1-u) &\Rightarrow \\ y = -\frac{\ln(1-u)}{\lambda} &= F^{-1}(u) \end{aligned}$$

Thus one can generate a random sample from Exponential(λ) by generating a random sample from Uniform(0,1), $u_1^*, u_2^*, \ldots, u_n^*$, and setting $y_i^* = -\frac{\ln(1-u_i^*)}{\lambda}$. Because $1-U$ and U have the same distribution, $y_i^* = -\frac{\ln(u_i^*)}{\lambda}$ can be used.

R code:

```
lambda <- 2
n      <- 100
u      <- runif(n=n,min=0,max=1)
y      <- -log(u)/lambda
```

7.3.2 Discrete random variable

The inverse PIT method works for discrete random variables as well. The key distinction is how the inverse CDF is defined:

$$F^{-1}(u) = \min_y \{y : F_Y(y) \geqslant u\}$$

Binomial(2,θ) example. Let $Y \sim$ Binomial(n=2,θ) random variable; thus Y has three possible values, 0, 1, and 2. The CDF, $F_Y(y)$, is a discontinuous function with three jumps:

$$\begin{aligned} F_Y(y) &= 0, y < 0 \\ F_Y(y) &= (1-\theta)^2, 0 \leqslant y < 1 \\ F_Y(y) &= 1-\theta^2, 1 \leqslant y < 2 \\ F_Y(y) &= 1, 2 \leqslant y \end{aligned}$$

To make things concrete, let θ=0.7. Then

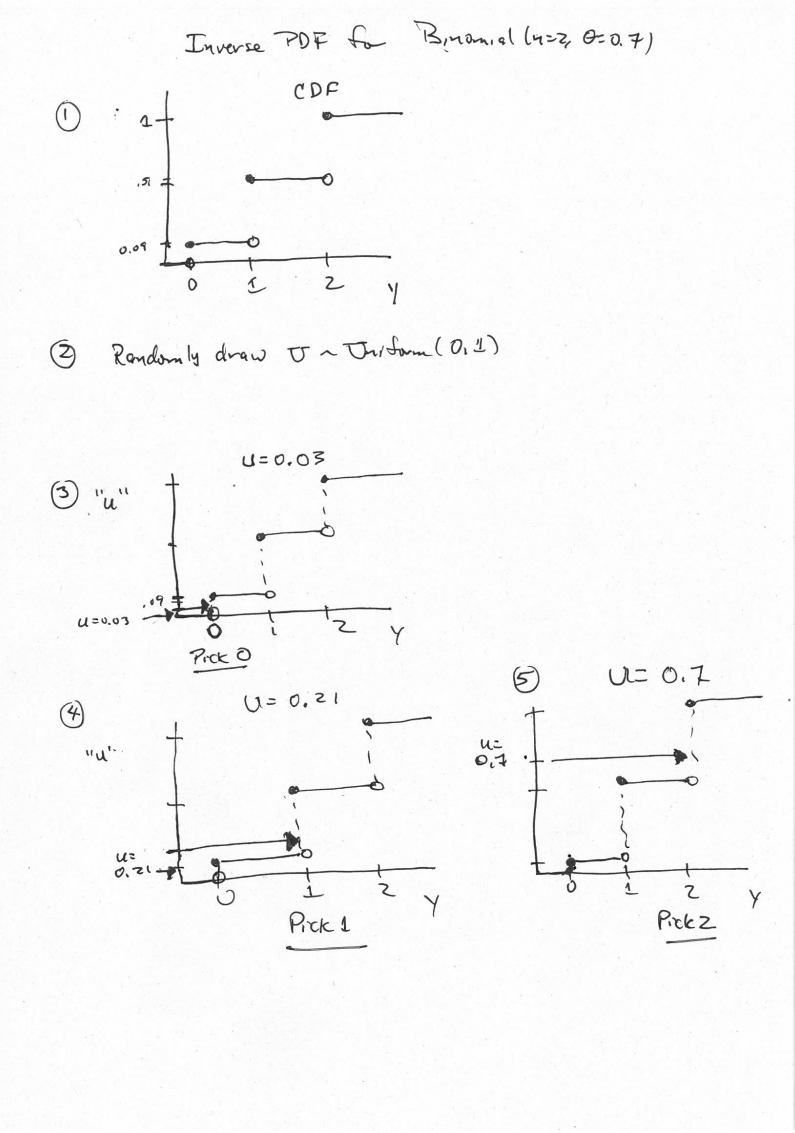
y	$F_Y(y)$
$y < 0$	0
$0 \leqslant y < 1$	0.09
$1 \leqslant y < 2$	0.51
$2 \leqslant y$	1.00

Consider 3 uniform random variable values: 0.03, 0.21, 0.70:

$$\begin{aligned} F^{-1}(u = 0.03) &= \min_y \{y : F_Y(y) \geqslant 0.03\} = 0 \\ F^{-1}(u = 0.21) &= \min_y \{y : F_Y(y) \geqslant 0.21\} = 1 \\ F^{-1}(u = 0.70) &= \min_y \{y : F_Y(y) \geqslant 0.70\} = 2 \end{aligned}$$

Note, for example, that the probability of generating Y=1 is the probability of $0.09 < U \leqslant 0.51$ = 0.42. Figure 7.4 demonstrates the results for these 3 uniform draws.

Figure 7.4: Demonstration of Inverse PIT method for a discrete random variable, Binomial(n=2,θ=0.7).



7.4 Rejection sampling

Rejection sampling is another Monte Carlo method that generates *independent* samples from the target distribution, e.g., $\mathrm{p}(\theta|\mathbf{y})$.

- It does so by generating independent samples from another distribution, which we will call the envelope distribution, and denote $\mathrm{g}(\theta)$.
- Only some of the generated values are kept or used, while others are discarded: hence the name "rejection" sampling. Thus to produce a sample of size n, one might generate $\mathrm{n} + \mathrm{K}$ samples from $\mathrm{g}(\theta)$ before ending up with n samples.
- It is more often used for generating a univariate random variable than a multivariate random vector.

To reduce notation, the target distribution will be denoted $\mathrm{p}(\theta)$ even when it is a posterior distribution.

7.4.1 Simple demonstration where $\mathrm{p}(\theta)$ has bounded support

As a simple demonstration of the main idea, consider a target distribution, $\mathrm{p}(\theta)$, where $0 \leqslant \theta \leqslant \mathrm{a}$, where a is a constant, and the envelope is $\mathrm{Uniform}(0, \mathrm{a})$. See Figure 7.5 for a graphical explanation. Note that two random variables are being generated in a given iteration: one from the envelope and another from a Uniform distribution (different from the envelope in this case).

The proof that the resulting kept value is a sample from the target distribution is shown in Figure 7.6.

Figure 7.5: Demonstration of rejection sampling when $\mathrm{p}(\theta)$ has support on $(0, \mathrm{a})$ and envelope is Uniform(0,a).

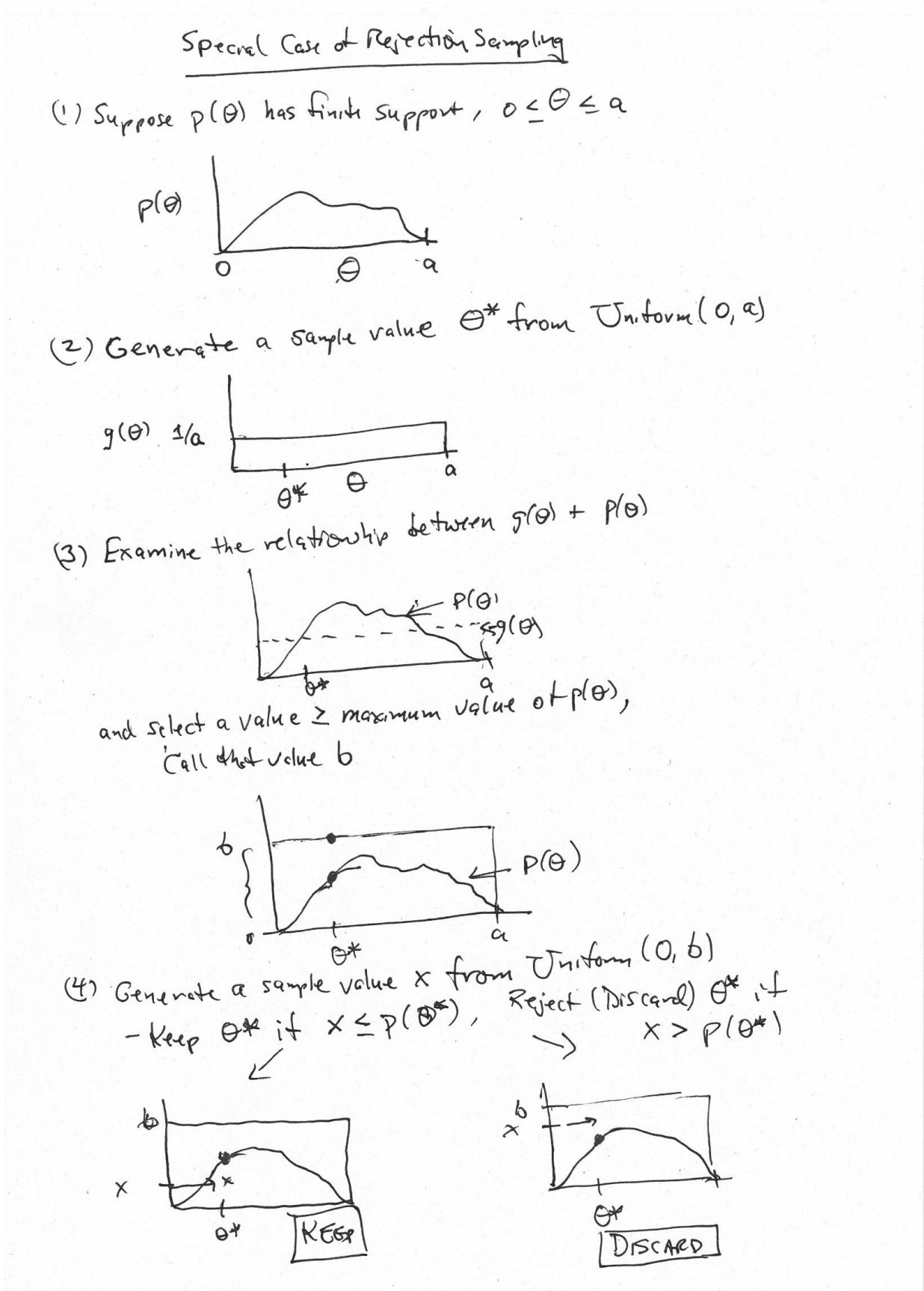
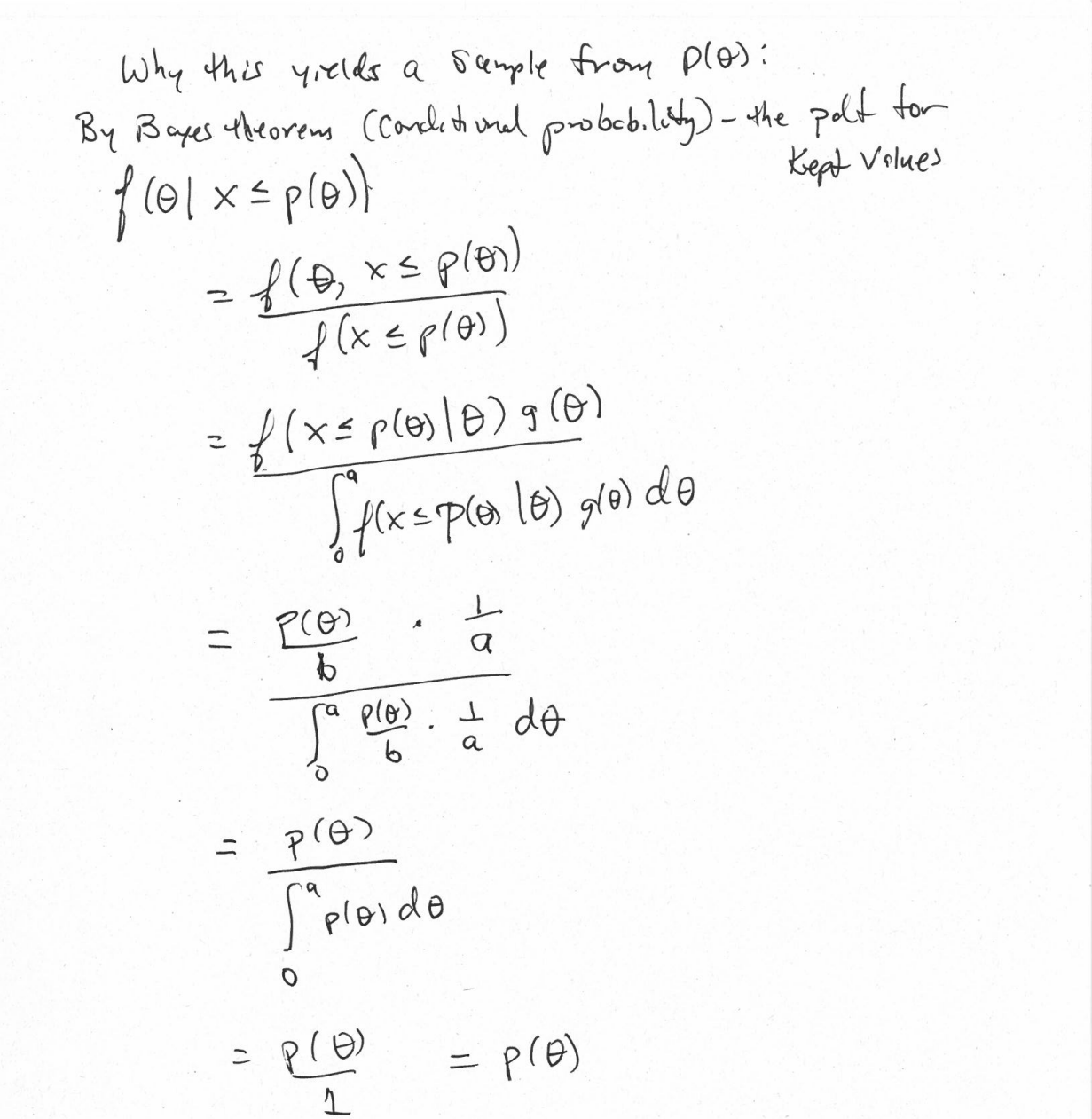


Figure 7.6: Proof for Figure 7.5 that the resulting kept value is from $\mathrm{p}(\theta)$ (with support on $(0, \mathrm{a})$) given the envelope is Uniform(0,a).



7.4.2 General rejection sampling algorithm

In general, given any target distribution, $p(\theta)$, which may or may not have finite support, the only constraint on $g(\theta)$ is that its support includes the support of $p(\theta)$, i.e. if $p(\theta) > 0$, then $g(\theta) > 0$. One then finds a bound M on the ratio of $p(\theta)/g(\theta)$:

$$M \geqslant \frac{p(\theta)}{g(\theta)}, \text{ for all } \theta$$

Equivalently:

$$Mg(\theta) \geqslant p(\theta)$$

The algorithm to generate a single value from $p(\theta)$.

1. Generate θ^* from $g(\theta)$.
2. Generate u from Uniform$(0, Mg(\theta^*))$.
3. If $u \leqslant p(\theta^*)$, keep θ^*, else go back to 1.

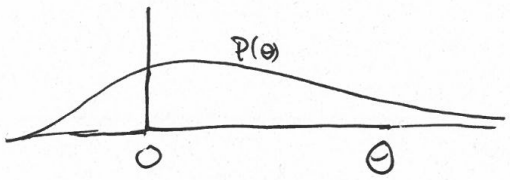
Variation. Instead of sampling from Uniform$(0, Mg(\theta^*))$ and keeping if $u \leqslant p(\theta^*)$, one can sample u from Uniform(0,1) and keep if $u \leqslant p(\theta^*)/Mg(\theta^*)$.

Figure 7.7 is a graphical representation of the procedure in this general setting.

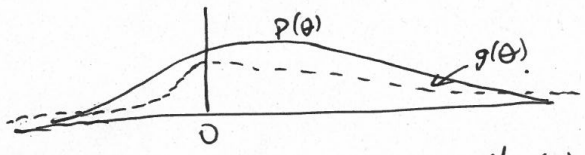
Figure 7.7: Demonstration of rejection sampling in the case of arbitrary $\mathrm{p}(\theta)$.

General Rejection Sampling Algorithm

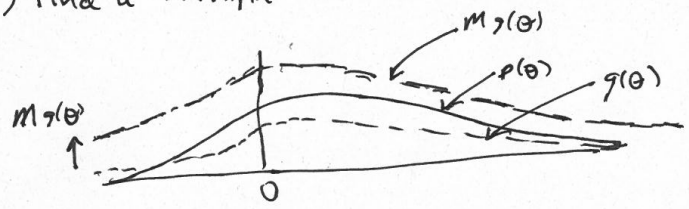
(1) Want sample from target dist^{n} $p(\theta)$ – need not have bounded support



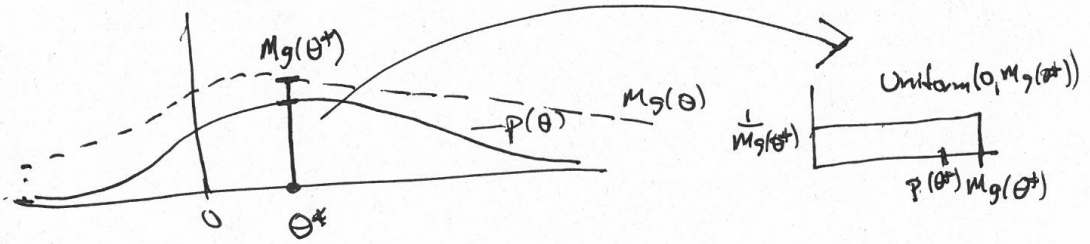
(2) Choose an envelope dist^{n} $g(\theta)$ which contains/includes support of $p(\theta)$: $p(\theta) > 0 \Rightarrow g(\theta) > 0$



(3) Find a "multiplier" M such that $M g(\theta) \geq p(\theta)$ for all θ



(4) – Randomly generate θ^* from $g(\theta)$
– Randomly generate u from $U(0, Mg(\theta^*))$
– If $u \leq p(\theta^*)$, keep θ^*
else Reject θ^*



Proof that this yields a sample from the target distribution

Conditional on being kept, the pdf for θ is the following.

$$f(\theta|u \leqslant p(\theta)) = \frac{f(\theta, u \leqslant p(\theta))}{f(u \leqslant p(\theta))} = \frac{Pr(u \leqslant p(\theta)|\theta)g(\theta)}{\int Pr(u \leqslant p(\theta)|\theta)g(\theta)d\theta}$$
$$= \frac{\frac{p(\theta)}{Mg(\theta)}g(\theta)}{\int \frac{p(\theta)}{Mg(\theta)}g(\theta)d\theta} = \frac{\frac{p(\theta)}{M}}{\int \frac{1}{M}p(\theta)d\theta} = \frac{p(\theta)}{\int p(\theta)d\theta} = p(\theta)$$

The acceptance rate is $1/M$

The probability that θ is kept is $1/M$:

$$Pr(\theta \text{ is kept}) = Pr(u \leqslant p(\theta)) = \int Pr(u \leqslant p(\theta)|\theta)g(\theta)d\theta$$
$$= \int \frac{p(\theta)}{Mg(\theta)}g(\theta)d\theta = \frac{1}{M}\int p(\theta)d\theta = \frac{1}{M}$$

Thus for a high acceptance rate, want $1/M$ close to 1, thus M to be relatively small, e.g., a little larger than 1. (Can't have $M < 1$.)

Thus ideally, select:

$$M_{opt} = \sup\left\{\frac{p(\theta)}{g(\theta)}\right\}$$

Namely, find the Least Upper Bound of the ratio of the target to the envelope.

In some cases for some target distribution and some envelope distribution, can find M_{opt}, but not always.

Example C

The target distribution, $p(\theta)$, Beta$(3, 2)$, and the envelope distribution is $Uniform(0, 1)$. To find a "suitable" M (actually an optimal M), examine the ratio $p(\theta)/g(\theta)$:

$$\frac{p(\theta)}{g(\theta)} = \frac{\Gamma(5)}{\Gamma(3)\Gamma(2)}\theta^{3-1}(1-\theta)^{2-1} = 12\theta^2(1-\theta)$$

Find a critical point by differentiating $\ln(12\theta^2(1-\theta)) \propto 2\ln(\theta)+\ln(1-\theta)$, setting equal to 0 and solving for θ. The result is θ=2/3 is a critical point (and 2nd derivative at θ=2/3 is 4.5 $>$ 0, therefore the ratio is minimized). And

$$\sup_\theta \frac{p(\theta)}{g(\theta)} = 12 * (2/3)^2 * (1/3) = 16/9$$

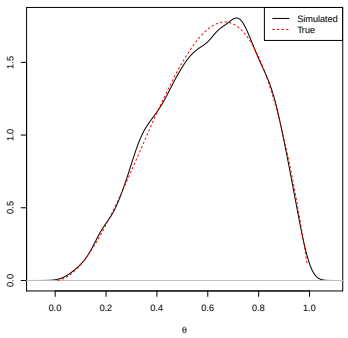
Then set M=16/9. Note the acceptance rate will be $1/M$ = 9/16 = 0.5625.

Demonstration with R:

```
#- Rejection sampling with target Beta(3,2) and envelope U(0,1)
a          <- 3; b          <- 2
n          <- 10000
M.opt      <- 16/9
theta.sim  <- numeric(n)
total.sims <- 0
tally      <- 0
while(tally < n) {
  total.sims <- total.sims+1
  theta.star <- runif(n=1,min=0,max=1)
  u.star     <- runif(n=1,min=0,max=M.opt*1)
  if(u.star <= dbeta(theta.star,a,b)) {
     tally <- tally+1
     theta.sim[tally] <- theta.star=
  }
}
accept.rate <- n/total.sims
cat("acceptance rate=",accept.rate,"Theory=",1/M.opt,"\n")
# acceptance rate= 0.5617978 Theory= 0.5625
```

Note that the empirical acceptance rate, 0.5618, was very close to the expected rate of 9/16=0.5625. Figure 7.8 compares the true Beta(3,2) density to the empirical density based on the rejection sample.

Figure 7.8: Demonstration of rejection sampling target Beta(3,2) and envelope U(0,1).



7.4.3 Target density need only be known up to a proportionality constant

The target density $p(\theta)$ need only be evaluated up to a proportionality constant, i.e., just need to be able to evaluate the terms in $p(\theta)$ that include θ. For example, if θ is Beta(α, β), need only evaluate $\theta^{\alpha-1}(1-\theta)^{\beta-1}$, not $\frac{\Gamma(\alpha+\beta}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$.

The algorithm is essentially the same as before. Write $p(\theta)$ as $p_1(\theta)c$, where c is the

proportionality constant. Find M_1 such that

$$M_1 \geqslant \frac{p_1(\theta)}{g(\theta)}$$

Then

- Generate θ^* from $g(\theta)$.
- Generate u from Uniform(0,$M_1 g(\theta^*)$).
- Keep θ^* if $u < p_1(\theta^*)$

The proof is just like for the case with $p(\theta)$. Conditional on being kept, the pdf for θ is the following.

$$\begin{aligned} f(\theta|u \leqslant p_1(\theta)) &= \frac{f(\theta, u \leqslant p_1(\theta))}{f(u \leqslant p(\theta))} = \frac{Pr(u \leqslant p_1(\theta)|\theta)g(\theta)}{\int Pr(u \leqslant p_1(\theta)|\theta)g(\theta)d\theta} \\ &= \frac{\frac{p_1(\theta)}{M_1 g(\theta)}g(\theta)}{\int \frac{p_1(\theta)}{M_1 g(\theta)}g(\theta)d\theta} = \frac{\frac{p_1(\theta)c}{M_1 g(\theta)}g(\theta)}{\int \frac{p_1(\theta)c}{M_1 g(\theta)}g(\theta)d\theta} \\ &= \frac{\frac{p(\theta)}{M_1}}{\int \frac{1}{M_1}p(\theta)d\theta} = \frac{p(\theta)}{\int p(\theta)d\theta} = p(\theta) \end{aligned}$$

What has happened here is that $M_1 \equiv M/c$. Note also that the acceptance probability is now $1/(M_1 c)$.

7.4.4 Example D

Not needing to evaluate the normalising constant, $m(y)$, is what makes rejection sampling very attractive for Bayesian inference.

For example (from Givens and Hoeting, 2013), a sample of size n=10 is generated from a Poisson(θ) distribution. The observed sample is:

$$8, 3, 4, 3, 1, 7, 2, 6, 2, 7$$

and $\bar{y}$ = 4.3. A lognormal prior distribution is assumed for θ: $\theta \sim$ Lognormal($\log(5)$, 0.5^2).

The objective is to generate a sample from the posterior distribution which is proportional to the Poisson likelihood and the lognormal prior:

$$\pi(\theta)f(y|\theta) = \left[\frac{1}{\sqrt{2\pi 0.5^2}}\frac{1}{\theta}e^{-\frac{1}{2*0.5^2}(\ln(\theta)-\log(5))^2}\right] \times \left[\frac{\exp(-n\theta)\theta^{n\bar{y}}}{\prod_{i=1}^{n} y_i!}\right]$$

which will have a complicated normalising constant.

For a given envelope distribution, need to find M_1 such that:

$$M_1 \geqslant \sup_{\theta} \frac{\pi(\theta)f(y|\theta)}{g(\theta)}$$

A convenient envelope distribution, $g(\theta)$, is the prior (not necessarily the most efficient however) as that makes calculation of M_1 easier:

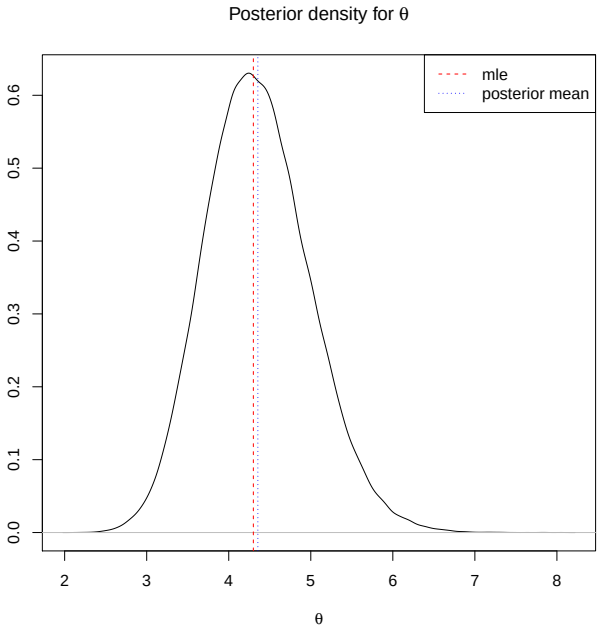
$$M_1 \geqslant \sup_{\theta} \frac{\pi(\theta) f(\mathbf{y}|\theta)}{\pi(\theta)} = \max_{\theta} f(\mathbf{y}|\theta)$$

The largest value of the likelihood function, $f(\mathbf{y}|\theta)$, is at the mle ($\hat{\theta}$), in this case the sample mean, 4.3. Thus

$$M_1 = \frac{\exp(-n\hat{\theta})\hat{\theta}^{n\bar{y}}}{\prod_{i=1}^{n} y_i!} = 1.439438e-10$$

The R code is shown below and Figure 7.9 shows the posterior density for θ. The estimate of the posterior mean is 4.355 and the estimate of the posterior variance is 0.398. The acceptance was about 28%.

Figure 7.9: Rejection Sampling: Posterior distribution for Poisson parameter θ given a Lognormal prior.




```
set.seed(214)
y <- c(8, 3, 4, 3, 1, 7, 2, 6, 2, 7)
n <- length(y)
mle <- mean(y)
M1 <- prod(dpois(y,lambda=mle))

log.mu <- log(5)
log.sd <- 0.5
N         <- 100000
M.opt      <- M1
```

```
theta.sim  <- numeric(n)
total.sims <- 0
tally      <- 0
while(tally < N) {
  total.sims <- total.sims+1
  theta.star <- rlnorm(n=1,meanlog = log.mu, sdlog=log.sd)
  #due to cancellations, g.theta need not be calculated,
  # but this is just to make the ratio calculations explicit
  g.theta    <- dlnorm(theta.star, meanlog=log.mu, sdlog=log.sd)
  u.star     <- runif(n=1,min=0,max=M.opt*g.theta)
  if(u.star <= prod(dpois(y,lambda=theta.star))*g.theta) {
    tally <- tally+1
    theta.sim[tally] <- theta.star
  }
}
cat("Estimated posterior mean=",mean(theta.sim),"posterior variance=",var(theta.sim),"\n")
#Estimated posterior mean= 4.356888 posterior variance= 0.395814
```

7.4.5 Advantages and Disadvantages of Rejection Sampling

Advantages:

1. It's a very general method that can be used for any distribution (so long as the target distribution, $\mathrm{p}(\theta)$, can be calculated).
2. Relatively easy to implement.
3. The distribution only need to be evaluated to a constant of proportionality. This is most advantageous for Bayesian inference for $\mathrm{p}(\theta|\mathbf{y})$—don't need $\mathrm{m}(\mathbf{y})$.
4. If can find an envelope distribution, $\mathrm{g}(\theta)$, where M is small, can be very efficient.

Disadvantages (or difficulties):

1. If for the chosen envelope, $\mathrm{g}(\theta)$, M is large, the rejection rate can be quite high, thus inefficient.
2. Can be difficult to find a good envelope, $\mathrm{g}(\theta)$, even for just three or four dimensions, e.g., $\mathrm{p}(\theta_1, \theta_2, \theta_3|\mathbf{y})$.

Main point: if $\mathrm{g}(\theta)$ is roughly proportional to the target distribution $\mathrm{p}(\theta)$, has similar shape, then rejection sampling can be quite efficient.

7.4.6 Two refinements on rejection sampling

Reference: Givens and Hoeting, Computational Statistics, 2nd Ed., 2013.

One refinement on rejection sampling is *Squeezing*: the essence of the idea is "enclose" the target distribution, $\mathrm{p}(\theta)$, above by the envelope function $\mathrm{Mg}(\theta)$ and below by one or more simple functions, e.g., a constant function. See top graph in Figure 7.10.

Another refinement, similar to squeezing is *Adaptive Rejection Sampling*. The essence of ARC is to define envelope and squeezing functions that are tangents and chords, respectively, to the target distribution. The tangents and chords get updated when generated values fall outside of the envelope and the squeezing functions. See bottom graph in Figure 7.10.

Figure 7.10: (From Givens and Hoeting, Computational Statistics, 2013.) Demonstration of “squeezing” (top figure) and Adaptive Rejection Sampling (ARS) (bottom figure).

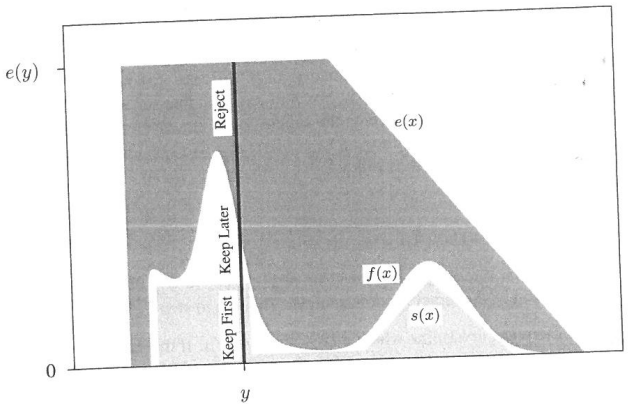


Fig. 6.3 Illustration of squeezed rejection sampling for a target distribution, f, using envelope e and squeezing function s. ‘Keep first’ and ‘Keep later’ correspond to steps 3 and 4 of the algorithm, respectively.

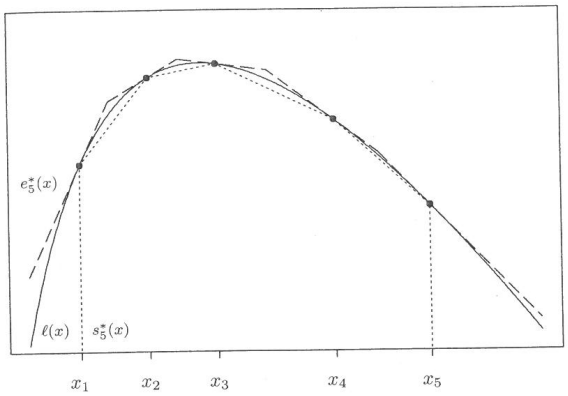


Fig. 6.4 Piecewise linear outer and inner hulls for $\ell(x) = \log f(x)$ used in adaptive rejection sampling when $k = 5$.

7.5 Importance Sampling

7.5.1 Basic idea

Similar to Rejection Sampling, Importance Sampling is a procedure

- where an *independent* sample of θ is generated from the *wrong* distribution, an importance sampling or envelope distribution, $g(\theta)$, but then the values are "adjusted" to correspond to the "right" distribution, the target distribution, $p(\theta)$ or $p(\theta|\mathbf{y})$ for Bayesian applications.

Importance Sampling differs from Rejection Sampling, however, in a crucial way: All the generated θ are kept—there is no rejection.

While importance sampling can be used to generate a sample directly from a target distribution, its utility for estimating integrals is perhaps more apparent. Suppose the objective is to estimate the following integral

$$E_p(h(\theta|\mathbf{y})) = \int h(\theta)p(\theta|\mathbf{y})d\theta$$

where $p(\theta|\mathbf{y})$ is a posterior distribution and the subscript p has been added to E_p to emphasise the probability distribution $p(\theta|\mathbf{y})$.

Let $g(\theta)$ be another distribution that has the same support as $p(\theta|\mathbf{y})$[3]. Then the integral can be rewritten:

$$\begin{aligned} E_p(h(\theta|\mathbf{y})) &= \int h(\theta)p(\theta|\mathbf{y})\frac{g(\theta)}{g(\theta)}d\theta = \int h(\theta)\frac{p(\theta|\mathbf{y})}{g(\theta)}g(\theta)d\theta \\ &= \int h(\theta)w(\theta)g(\theta)d\theta = E_g\left(h(\theta)w(\theta)\right) \end{aligned}$$

where $w(\theta) = p(\theta|\mathbf{y})/g(\theta)$ is the importance ratio. The subscript g in E_g is emphasising the probability distribution $g(\theta)$.

If one can generate a (iid) sample of size N from $g(\theta)$, then a Monte Carlo estimate of $E_g(w(\theta))$ can be calculated:

$$\hat{E}_g(h(\theta)w(\theta)) \equiv \hat{E}_p(h(\theta|\mathbf{y})) = \frac{1}{N}\sum_{i=1}^{N} h(\theta^i)w(\theta^i)$$

Again by the Strong Law of Large Numbers, the Monte Carlo estimate will converge with probability 1 to $E_g(h(\theta)w(\theta))$, which as showed above, is also $E_p(h(\theta|\mathbf{y}))$.

Comments.

1. The Monte Carlo error in the estimated integral is a function of the distribution of weights (or importance ratios). The expected value of importance ratios is 1:

$$E[p(\theta|\mathbf{y})/g(\theta)] = \int \frac{p(\theta|\mathbf{y})}{g(\theta)}g(\theta)d\theta = \int p(\theta|\mathbf{y})d\theta = 1$$

[3]Technically $g(\theta)$ just needs to have the same support as the product, $h(\theta)p(\theta|\mathbf{y})$, not the posterior distribution alone. For example, if $h(\theta) = I(\theta < 5)$, then the support of $g(\theta)$ just needs to include $\theta < 5$.

But the variation in these ratios is what causes Monte Carlo error.

$$V[p(\theta|\mathbf{y})/g(\theta)] = \int \left(\frac{p(\theta|\mathbf{y})}{g(\theta)} - 1\right)^2 g(\theta)d\theta$$

Very large weights, e.g., $p(\theta^i|\mathbf{y})/g(\theta^i) >> 1$, can result when g is "light" in the tails of the distribution of $p(\theta|\mathbf{y})$. A common preference for $g(\theta)$ is a distribution with relatively "heavy" tails and t and multivariate t distributions are popular choices.

2. Similar to Rejection Sampling, the Monte Carlo error decreases as the distributional shape of importance sampler becomes more similar to that of the integrand, $h(\theta)p(\theta|\mathbf{y})$.

3. Also similar to Rejection Sampling, the target distribution, or the integrand, need only be known up to a constant of proportionality. To begin the expected value is written as a ratio of two integrals:

$$E_p(h(\theta|\mathbf{y})) = \int h(\theta)p(\theta|\mathbf{y})d\theta = \int h(\theta)\frac{\pi(\theta)f(\mathbf{y}|\theta)}{m(\mathbf{y})}d\theta = \frac{\int h(\theta)\pi(\theta)f(\mathbf{y}|\theta)d\theta}{\int \pi(\theta)f(\mathbf{y}|\theta)d\theta}$$

Use importance sampling to estimate the integrals in the numerator and denominator:

$$\frac{1}{N}\sum_{i=1}^{N} h(\theta^i)\frac{\pi(\theta^i)f(\mathbf{y}|\theta^i)}{g(\theta^i)} \approx \int h(\theta)\pi(\theta)f(\mathbf{y}|\theta)d\theta \tag{7.5}$$

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\pi(\theta^i)f(\mathbf{y}|\theta^i)}{g(\theta^i)} \approx \int \pi(\theta)f(\mathbf{y}|\theta)d\theta = m(\mathbf{y}) \tag{7.6}$$

Then

$$\hat{E}_p(h(\theta|\mathbf{y})) = \frac{\sum_{i=1}^{N} h(\theta^i)\frac{\pi(\theta^i)f(\mathbf{y}|\theta^i)}{g(\theta^i)}}{\sum_{i=1}^{N} \frac{\pi(\theta^i)f(\mathbf{y}|\theta^i)}{g(\theta^i)}} = \sum_{i=1}^{N} h(\theta^i)w^{*i}$$

where

$$w^{*i} = \frac{\frac{\pi(\theta^i)f(\mathbf{y}|\theta^i)}{g(\theta^i)}}{\sum_{j=1}^{N} \frac{\pi(\theta^j)f(\mathbf{y}|\theta^j)}{g(\theta^j)}}$$

The estimator has some bias, but is consistent (asymptotically unbiased).

7.5.2 Example E: Poisson-Lognormal case again

Importance sampling was applied to the previous Poisson data with lognormal distribution. The prior was used as the envelope distribution (also called importance sampler), thus the importance ratio:

$$w^i = \frac{\pi(\theta^i)f(\mathbf{y}|\theta^i)}{\pi(\theta^i)} = f(\mathbf{y}|\theta^i) \propto e^{-n*\theta^i}(\theta^i)^{n\bar{y}}$$

The weights were then scaled to cancel the normalising constant:

$$w^{*i} = \frac{w^i}{\sum_{j=1}^{N} w^j}$$

And the posterior mean and variance were estimated as:

$$\hat{E}(\theta) = \sum_{i=1}^{N} w^{*i}\theta^i$$

$$\hat{V}(\theta) = \sum_{i=1}^{N} w^{*i}(\theta^i - \hat{E}(\theta))^2$$

The R code is shown below. The estimate of the posterior mean is 4.359 and variance is 0.400, quite similar to the rejection sampling estimates of 4.355 and 0.398, respectively.

```
#--- Importance sampling to estimate the mean and variance of the Poisson-Lognormal posteric
y <- c(8, 3, 4, 3, 1, 7, 2, 6, 2, 7)
n <- length(y)
ybar <- mean(y)
set.seed(381)
N <- 1000000
theta.star <- rlnorm(n=N,meanlog = log.mu, sdlog=log.sd)
importance.ratio <- exp(-n*theta.star)*theta.star^(n*ybar)
weight.star  <- importance.ratio/sum(importance.ratio)
hat.E.theta <- sum(theta.star*weight.star)
hat.V.theta <- sum((theta.star-hat.E.theta)^2*weight.star)
alt.V       <- sum(theta.star^2*weight.star)-hat.E.theta^2

cat("Estimated posterior mean=",hat.E.theta,"\n")
# Estimated posterior mean= 4.358722
cat("Estimated posterior variance=",hat.V.theta,"\n")
# Estimated posterior variance= 0.3996452
cat("alt variance=",alt.V,"\n")
#alt variance= 0.3996452
```

7.6 Sampling Importance Re-Sampling

Importance sampling can be seen as a means to estimate integrals, but the generated θ and the importance ratios, w^i, can be used to produce *an approximate sample from the target distribution*, e.g., $p(\theta|y)$. To generate a sample of size n:

1. Choosing $N > n$, generate N samples of θ^i, i=1,...,N, from the envelope distribution (importance sampler), $g(\theta)$.
2. Calculate the importance ratios, w^i, and scale them by the sum of the weights:

$$w^{*i} = \frac{w^i}{\sum_{j=1}^{N} w^j}$$

   Thus $0 < w^{*i} < 1$ and $\sum_{j=1}^{N} w^{*j}$ = 1.

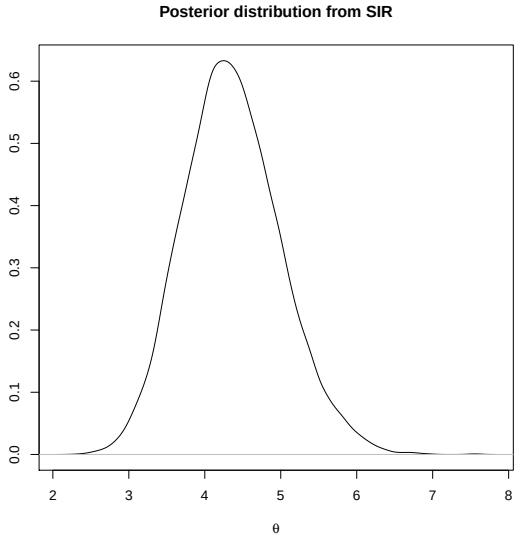
3. Randomly sample n θ^i, with replacement, with probabilities w^{*1}, w^{*2}, ..., w^{*N}.

Note that for Bayesian inference when the target distribution is the posterior, due to the scaling of the importance ratios, the normalising constant is cancelled out and does not need to be known.

Example F: SIR with the Poisson-Lognormal example. The scaled importance weights from the Importance Sampling example above were used for the SIR sampling, and the estimated posterior density for θ is plotted in Figure 7.11, which looks much like the result from the Rejection Sampling (Figure 7.9).

```
#--- SIR sample of the posterior ----
set.seed(9371)
n <- 10000
theta.post.sample <- sample(theta.star,size=n,prob=weight.star,replace=TRUE)
plot(density(theta.post.sample),type="l",xlab=expression(theta),ylab="",
     main="Posterior distribution from SIR")
```

Figure 7.11: Estimated posterior density for θ from the Poisson-Lognormal example based a SIR sampling.



7.7 R Code

7.7.1 Monte Carlo inference for the Weibull example.

```
 hurricane.gaps <- c(0.30, 4.61, 5.75, 0.24, 0.09, 0.18, 7.38, 1.20, 2.40, 0.18,
                     0.02, 10.07, 0.23, 0.44, 3.34, 0.06, 0.01, 0.71, 0.06, 0.42)
 n <- length(hurricane.gaps)

#-- Monte Carlo estimate of the normalizing constant with the Weibull --
# simulating from the prior, Gamma(kappa,lambda)
set.seed(730)
N          <- 100000
kappa      <- lambda <- 0.01
alpha.star <- rgamma(n=N,shape=kappa,rate=lambda)

#--- Calculating the sum by looping over the N samples
integ.est <- 0
natural.approach <- FALSE
for(i in 1:N) {
  if(natural.approach) {
    integ.est <- integ.est + (1/N)*(alpha.star[i]^n *
                                    exp(-sum(hurricane.gaps^(alpha.star[i]))) *
                                    exp(sum((alpha.star[i]-1)*log(hurricane.gaps))))
  } else {
    integ.est <- integ.est + (1/N)*(exp(n*log(alpha.star[i]) -
                                      sum(hurricane.gaps^(alpha.star[i])) +
                                      sum((alpha.star[i]-1)*log(hurricane.gaps))))
  }
}
cat("Estimate of normalising constant=",integ.est,"\n")
# Estimate of normalising constant= 1.864341e-14

#--- sample from posterior distribution....
alpha.seq <- seq(0.1,1,length=100)
mc.density <- numeric(length(alpha.seq))
for(i in 1:length(alpha.seq)) {
  mc.density[i] <- dgamma(alpha.seq[i],shape=kappa,rate=lambda)*
    prod(dweibull(x=hurricane.gaps,shape= alpha.seq[i],scale=1))/integ.est
}
plot(alpha.seq,mc.density,type="l",xlab=expression(alpha),ylab="",
    main=expression("Posterior Weibull param, MC estimate"))

#-- estimating expected value for posterior, can use the previously generate alpha.star
alpha.star.weights <- numeric(length(alpha.star))
for(i in 1:length(alpha.star)) {
  alpha.star.weights[i] <- prod(dweibull(x=hurricane.gaps,shape=alpha.star[i],scale=1))/integ.est
}

#--   problem with NAs when alpha.star=0 or is too large, >300,..."cheating"
ok <- !is.na(alpha.star.weights)
ok.N <- sum(ok)
# post.alpha.mean <- (1/N)* sum(alpha.star*alpha.star.weights,na.rm=TRUE)

post.alpha.mean <- (1/ok.N)* sum(alpha.star[ok]*alpha.star.weights[ok])
cat("Est of posterior mean=",post.alpha.mean,"\n")
# Est of posterior mean= 0.5522097
```

7.7.2 Demonstration of Monte Carlo error in estimation of $\mathrm{E}[\theta]$ where $\theta \sim$ Gamma(3,0.2).

```
#--- Monte Carlo error demonstration---- simulating from a Gamma and estimating expected value
set.seed(530)
number.simulations <- 50000
```

```
N.vec              <- c(10,100,1000,10000)
avg.averages       <- numeric(4)
sd.averages        <- numeric(4)
a      <- 3;  b    <- 0.2
true.ExpVal <- a/b
cat("Expected value=",true.ExpVal,"\n")
# Expected value= 15
true.stderr <- sqrt(a/b^2)/sqrt(N.vec)
cat("Theoretical Std Errors",round(true.stderr,3),"\n")

par(mfrow=c(2,2),oma=c(0,0,3,0))
for(i in 1:4) {
  out <- matrix(rgamma(N.vec[i]*number.simulations,shape=a,rate=b),
                nrow=number.simulations,ncol=N.vec[i])
  per.sample.average <- apply(out,1,mean)
  avg.averages[i] <- mean(per.sample.average)
  sd.averages[i]  <- sd(per.sample.average)
  hist(per.sample.average,xlab="Sample Average",ylab="",
       main=paste("N=",N.vec[i],"avg=",round(avg.averages[i],2),
             "std dev=",round(sd.averages[i],2)))
}
par(mfrow=c(1,1))
```

7.7.3 Demonstration of Monte Carlo error in estimation of $\Pr(\theta < 5)$ where $\theta \sim$ Gamma(3,0.2).

```
# simulating from a Gamma and estimating Pr(theta < 5)
set.seed(530)
number.simulations <- 50000
N.vec              <- c(10,100,1000,10000)
avg.averages       <- numeric(4)
sd.averages        <- numeric(4)
a      <- 3;  b    <- 0.2
true.prob <- pgamma(5,shape=a,rate=b)
cat("Pr(theta<5)=",true.prob,"\n")
# Pr(theta<5)= 0.0803014
true.stderr <- sqrt(true.prob*(1-true.prob))/sqrt(N.vec)
cat("Theoretical Std Errors",round(true.stderr,3),"\n")
#Theoretical Std Errors 0.086 0.027 0.009 0.003

indicator.fun <- function(x,y) x<y

par(mfrow=c(2,2),oma=c(0,0,3,0))
for(i in 1:4) {
  out <- matrix(rgamma(N.vec[i]*number.simulations,shape=a,rate=b),
                nrow=number.simulations,ncol=N.vec[i])
  out <- indicator.fun(out,5)
  per.sample.est <- apply(out,1,mean)
  avg.averages[i] <- mean(per.sample.est)
  sd.averages[i]  <- sd(per.sample.est)
  hist(per.sample.est,xlab="Average p-hat",ylab="",
       main=paste("N=",N.vec[i],"avg=",round(avg.averages[i],4),
                  "std dev=",round(sd.averages[i],4)))
}
par(mfrow=c(1,1))
```

Homework Week 7: Bayesian Theory.

Optional Reading

1. Monte Carlo Integration.
   (a) Chapter 6, Sections 6.1 to 6.31, of Computational Statistics, 2nd Ed (2013), Givens and Hoeting, pp 151-168.
   (b) Lecture Notes of King and Ross, Sections 2.1-2.2, pp 33-42.

Exercises: not to turn in

1. Explain how to use the Inverse Probability Integral Transform to generate a sample of size n from the Kumaraswamy distribution (an alternative to the Beta distribution):

$$f(y|\alpha, \beta) = \alpha\beta y^{\alpha-1}(1-y^{\alpha})^{\beta-1}, \;\; 0 < y < 1$$

   If you want, write a short piece of R code to generate such a sample using α=2 and β=5.

2. Describe how to use rejection sampling to generate a sample from a Beta(4,6) distribution using a Uniform(0,1) envelope distribution. The description needs to include the optimal value of M.

3. Describe the importance sampling algorithm to calculate $E[Y^2]$, where Y follows a Kumaraswamy(α, β) distribution, using a Uniform(0,1) envelope distribution.

4. Describe a SIR algorithm for generating a sample of size n from a Kumaraswamy(α, β) distribution using a Uniform(0,1) envelope distribution.

8 Dependent Monte Carlo Sampling, Overview

8.1 Overview of MCMC

1. **Dependent samples:** In contrast to direct sampling using the inverse probability integral transform, rejection sampling, and importance sampling, which all yield independent samples from a target distribution $p(\theta)$, or for Bayesian inference, $p(\theta|y)$, Markov Chain Monte Carlo (MCMC) methods yield dependent samples from the target distribution.

   In particular the *dependent* samples are generated from a Markov chain which has a *limiting stationary distribution* that equals the target distribution $p(\theta)$: the chain values have *converged* to be a sample from $p(\theta)$.

2. **Iterative conditional generation:** Given an initial value for θ, denoted θ^0, the next value θ^1 is generated from a conditional probability distribution given θ^0, say $g(\theta^1|\theta^0)$. This is subsequently repeated in an iterative manner: at the i^{th} iteration, θ^i is generated from $g(\theta^i|\theta^{i-1})$. Let N denote the total number of sample values generated.

3. **Burn-in:** In most cases the initial values, say θ^1, θ^2, ..., θ^B, are *not* distributed according to the target distribution, $p(\theta)$. These initial sample values are then discarded and are not used for making inferences about $p(\theta)$. This initial sample generation period is called the *Burn-in period*.

4. **Sample for inference:** The additional $N - B$ sample values that are generated, θ^{B+1}, θ^{B+2}, ..., θ^N, are the ones that used for inference. For example, suppose that θ is a scalar, the expected value of a function of θ, $h(\theta)$, is estimated by:

$$\hat{E}[h(\theta)] = \frac{1}{N-B}\sum_{i=B+1}^{N} h(\theta^i) \tag{8.1}$$

   Other sorts of inferences described previously can be carried out using this sample. For example:

$$\widehat{Pr}(a \leqslant \theta \leqslant b) = \frac{1}{N-B}\sum_{i=B+1}^{N} I(a \leqslant \theta^i \leqslant b)$$

   and

$$\widehat{Var}(\theta) = \frac{1}{N-B}\sum_{i=B+1}^{N}\left(\theta^i - \hat{\theta}\right)^2$$

   or

$$\widehat{Var}(\theta) = \left(\frac{1}{N-B}\sum_{i=B+1}^{N}(\theta^i)^2\right) - \left(\frac{1}{N-B}\sum_{i=B+1}^{N}\theta^i\right)^2$$

   Remarks on Variance.

   - The 2nd estimate for $\mathrm{Var}(\theta)$ is based on $V(X) = E[X^2] - (E[X])^2$.
   - For $\widehat{Var}(\theta)$, the divisor $N - B - 1$ would be technically better, but with large $N - B$ the effect is negligible.
   - Alternative notation is to let N=number of iterations used for inference, thus the total number of iterations is $B + N$.

Comments

- The art of MCMC is choosing or constructing the Markov chain that will yield (eventually) samples from the target distribution, $p(\theta)$. In other words, Markov chains that have the right limiting stationary distribution.
- The Metropolis-Hastings and the Gibbs Sampling algorithms are two popular methods that achieve this.
- Determining the length of the burn-in period is not an exact science. The commonly used methods for determining B are indicative that the chain has not converged, rather than it has converged.
- Multiple chains: one common sense approach to determining burn-in is to run multiple chains, e.g., 3, with different initial values, θ_j^0, for chains j=1,2,3. Then examine the point in the iterations at which the simulated values have similar probability distributions; e.g., after discarding B=1000 iterations, do the histograms for the remain $N - B$ values look similar?
- Burn-in time is very much a function of the Markov chain that is used—some converge faster than others.
- The MCMC estimate, e.g., eq'n (8.1), is again an estimate of an integral. The properties of the estimate, e.g., consistency, are not based on the standard Strong Law of Large Numbers (SLLN) or the Central Limit Theorem (CLT) because the sample values are dependent, not independent. Instead the properties depend upon various *ergodic* theorems that are Markov chain versions of the SLLN and CLT.
- Related to the previous point, the Monte Carlo error in the MCMC estimates cannot be calculated as simply as for the independent samples. Instead, various time series type approaches are used to estimate Monte Carlo error.
- Some argue that burn-in is unnecessary and a waste of samples. See Geyer, 1992, Statistical Science; see also "Burn-in is unnecessary" available at
`http://users.stat.umn.edu/~geyer/mcmc/burn.html`

8.2 Brief Introduction to Markov Chains

Here we give a brief introduction to discrete time and discrete state-space Markov chains[1].

Let $\{\theta^t\}$, $t = 0, 1, \ldots$ be a sequence of random variables where each θ^t can take one of a finite (or countably infinite) number of values. These possible values are called *states*.

The notation $\theta^t = j$ means that the "process" at time t is in state j.

The *state space*, denoted S, is the set of all possible states.

For example, if θ^t was a Poisson random variable, then possible states are 0, 1, 2, 3, ... (a countably infinite number) and S = the set of non-negative integers.

8.2.1 Definition of a Markov chain

- The joint distribution for $\theta^0, \theta^1, \ldots, \theta^T$ can be written as the product of conditional distributions:

$$\begin{aligned}\Pr(\theta^0 = x_0, \theta^1 = x_1, \ldots, \theta^T = x_T) = \Pr(\theta^T = x_T | \theta^0 = x_0, \theta^1 = x_1, \ldots, \theta^{T-1} = x_{T-1}) \\ \times \Pr(\theta^{T-1} = x_{T-1} | \theta^0 = x_0, \theta^1 = x_1, \ldots, \theta^{T-2} = x_{T-2})) \times \ldots \\ \times \Pr(\theta^0 = x_0)\end{aligned}$$

- If θ^t given θ^{t-1} is (conditionally) independent of other values, the conditional distribution for θ^t simplifies:

$$\Pr(\theta^t = x_t | \theta^{t-1} = x_{t-1}, \ldots, \theta^0 = x_0) = \Pr(\theta^t = x_t | \theta^{t-1} = x_{t-1})$$

[1] The explanation is based on Section 1.7 Markov chains from Computational Statistics (2013, Givens and Hoeting)

This is called the Markov property. And the joint distribution simplifies considerably:

$$\Pr(\theta^0 = x_0, \theta^1 = x_1, \ldots, \theta^T = x_T) = \prod_{t=1}^{T} \Pr(\theta^t = x_t | \theta^{t-1} = x_{t-1}) \times \Pr(\theta^0 = x_0)$$

- A sequence of random variables, θ^0, θ^1, θ^2, ..., with the Markov property is called a *Markov chain*.

8.2.2 Terminology and Notation of a Markov chain

- Shorthand notation for $\Pr(\theta^t = j | \theta^{t-1} = i)$ is p_{ij}^t.
- p_{ij}^t is called the *one-step transition probability*.
- If the p_{ij}^t are the same for times t, i.e., $p_{ij}^t = p_{ij}$, then the chain is called *time homogeneous*.
- If the p_{ij}^t vary over time, the chain is called *time inhomogeneous*.
- Assume there are K states and denote the values by 1, 2, ..., K. The "movement" probabilities of a Markov chain can be neatly represented by a K by K *transition probability matrix, tpm*, $P_t =$

	t			
$t-1$	1	2	...	K
1	p_{11}^t	p_{12}^t	$\cdots$	p_{1K}^t
2	p_{21}^t	p_{22}^t	$\cdots$	p_{2K}^t
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
K	p_{K1}^t	p_{K2}^t	$\cdots$	p_{KK}^t

Note that the probabilities in each row sum to 1. If the state at time t is i, then it will go be in *some* state at time t. It row can be viewed as a probability mass function.

In the case of a time homogeneous chain, let P denote the transition probability matrix.

8.2.3 Example

This example is taken from Givens and Hoeting. Precipitation "states", wet (more than 0.01 inches of precipitation) and dry, for San Francisco were recorded for 1814 pairs of consecutive days during the months November through March, starting from November 1990 and ending March 2002. The table below summarizes the transitions in states from day $t-1$ to day t:

	t	
$t-1$	Wet	Dry
Wet	418	256
Dry	256	884

So the state space S has 2 states: wet and dry. Let θ^t be the state on day t. Assuming time-homogeneity, the estimated tpm for θ^t is:

$$\widehat{P} = \begin{bmatrix} 0.620 & 0.380 \\ 0.224 & 0.775 \end{bmatrix}$$

For example, $\Pr(\theta^t = \text{Wet} | \theta^{t-1} = \text{Wet})$ = 0.620.

8.2.4 Limiting behaviour of Markov chains

First some definitions of certain types of states and chains:

Recurrent state: a state to which a Markov chain returns with probability 1 is a *recurrent state*.

Nonnull state: a state for which the expected time (number of steps) until recurrence is finite is called a *nonnull state*.

Note: recurrent states in Markov chains with finite state spaces are nonnull.

Irreducible Markov chain: a chain where a state j can be reached in a finite number of steps from *any state* i is an *irreducible Markov chain*.

In other words, there is a number $m > 0$ such that $\Pr(\theta^{m+n} = j | \theta^n = i) > 0$.

Periodic Markov chain: a chain where the number of steps required to return some portions of the state space is a multiple of some integer, say d, is a *periodic Markov chain*. In other words, $\Pr(\theta^{m+n} = j | \theta^n = j) > 0$ only if m/d is an integer.

For example, the chain with the following (homogeneous) tpm, P, has period d=3:

	t		
$t-1$	a	b	c
a	0.0	1.0	0.0
b	0.0	0.0	1.0
c	1.0	0.0	0.0

A chain that is *not periodic* is an *aperiodic Markov chain*.

Ergodic Markov chain: a chain that is *irreducible*, *aperiodic*, and all its states are *nonnull* and *recurrent* is an *ergodic Markov chain*.

Marginal and Stationary distributions.

- The marginal probability distribution for the state at time t is denoted π_t. It is a vector of probabilities that sum to 1, where the ith element $\pi_t(i)$ is the *marginal probability* that $\theta^t = i$, $\pi_t(i) = \Pr(\theta^t = i)$. For example, letting superscript T mean transpose, π_t^T=(0.35, 0.15, 0.3, 0.15, 0.05), and $\pi_t(2)$ =$\Pr(\theta^t = 2)$=0.15. In other words, π_t is the marginal distribution for θ^t.
- If the tpm at time $t+1$ is P, then the marginal distribution for θ^{t+1}, π_{t+1}, is $\pi_{t+1}^T = \pi_t^T P$.
- Any discrete probability distribution π is a *stationary distribution* for P (or a Markov chain with tpm P) **if**

$$\pi^T P = \pi^T$$

(Note the dropping of the subscript t.) This means that the probability of being in state j at time t equals the probability of being in state j at time $t+1$.

Referring to the San Francisco wet and dry days P, the stationary distribution is π^T=(0.3715, 0.6284):

$$[0.3715\ 0.6284] \begin{bmatrix} 0.6202 & 0.3798 \\ 0.2245 & 0.7754 \end{bmatrix} = [0.3715\ 0.6284]$$

Three simulations from the chain with T=100,000 (see Appendix 8.4.1 for the R code) was generated from this Markov chain. The results are shown in Table 8.1.

Table 8.1: Fraction of dry days at intermediate time points in the San Francisco Wet-Dry day Markov chain. The stationary distribution probability of dry days is 0.6284.

	Intermediate time points				
Simulation	10	100	1000	10000	100,000
1	0.400	0.680	0.669	0.636	0.629
2	0.500	0.630	0.662	0.633	0.626
3	0.800	0.620	0.639	0.630	0.626

Comments

- What this means: if a Markov chain at time t has a marginal probability distribution that is also a stationary distribution, the marginal probability distribution for all future random variables in the chain is the same marginal probability distribution. The marginal distributions for θ^t, θ^{t+1}, θ^{t+2}, . . ., are identical.
- A given P can have more than one stationary distribution. For example,

	t						
t − 1	0	1	2	3	4	5	6
0	0	1/2	0	0	0	0	1/2
1	0	1/3	1/3	1/3	0	0	0
2	0	1/3	1/3	1/3	0	0	0
3	0	1/3	1/3	1/3	0	0	0
4	0	0	0	0	1/3	1/3	1/3
5	0	0	0	0	1/3	1/3	1/3
6	0	0	0	0	1/3	1/3	1/3

Main idea of MCMC. **In the context of Bayesian inference, the central idea of Markov chain Monte Carlo methods, is to generate samples from a Markov chain such that "eventually" a stationary distribution will be reached and that distribution is the posterior distribution, $p(\theta|y)$.**

Of course MCMC methods can be used to generate samples from an arbitrary distribution.

The "art" of MCMC methods is to construct chains that have a tpm which will yield a (limiting) stationary distribution equal to the target distribution.

Reversible chains and the Detailed Balance Condition. One can roughly view what are reversible chains as a means of 'creating' a Markov chain that has the 'right' P for producing a (limiting) stationary distribution equal to the target distribution.

Reversible chain: a time homogeneous Markov chain with tpm P such that

$$\pi(i)\Pr(\theta^t = j|\theta^{t-1} = i) = \pi(j)\Pr(\theta^t = i|\theta^{t-1} = j) \tag{8.2}$$

then π is *a* stationary distribution for the chain and such a chain is called a *reversible Markov chain*.

- Equation 8.2 is called the *detailed balance equation*.
- Equation 8.2 is sometimes written

$$\pi(i)p_{ij} = \pi(j)p_{ji}$$

- Such a chain is called reversible because the joint distribution for two consecutive realisations is the same whether the chain is run forwards or backwards.
- The San Francisco wet and dry days chain is reversible (showing more digits to reduce rounding errors):

$$\pi(Wet)\Pr(\theta^t = Dry|\theta^{t-1} = Wet) = \pi(Dry)\Pr(\theta^t = Wet|\theta^{t-1} = Dry), namely$$
$$0.3715546 * 0.3798220 = 0.6284454 * 0.2245614 = 0.1411246$$

Key Theoretical Results. If a Markov chain with tpm P is irreducible and aperiodic and it has a stationary distribution π, then

- **π is *unique*, i.e., there is only one stationary distribution.**
- **The limiting distribution is the stationary distribution**:

$$\lim_{n\to\infty} \Pr(\theta^{t+n} = j|\theta^t = i) = \pi(j) \tag{8.3}$$

- Another way to state eq'n (8.3): "if θ^1, θ^2, . . ., are realisations from an irreducible and aperiodic Markov chain with stationary distribution π, then θ^n converges in distribution to the distribution π" (Givens and Hoeting, p 16).
- Given the tpm P, the components of π are solutions to the following equations:

$$\sum_{i \in S} \pi(i) = 1$$
$$\pi(j) = \sum_{i \in S} \pi(i) \Pr(\theta^t = j | \theta^{t-1} = i)$$

  subject to the constraint that $\pi(j) \geqslant 0$, for all j.

Another theoretical consequence is the following *ergodic theorem*, which is a generalisation of the strong law of large numbers: For any function h such that $E_\pi[h(\theta)]$ exists,

$$\Pr\left(\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} h(\theta^t) = E_\pi[h(\theta)] \right) = 1 \tag{8.4}$$

(almost sure convergence.)

Thus, a sample from a Markov chain can be used to estimate integrals.

8.3 Metropolis-Hastings Algorithm: Part 1

The objective is to generate a sample from a target distribution, $p(\Theta)$, where Θ is usually a vector, $\Theta = (\theta_1, \theta_2, \ldots, \theta_q)$. The cases we are primarily interested in are where $p(\Theta)$ is a posterior distribution: $p(\Theta|\mathbf{y})$. However, we will momentarily omit explicit conditioning on data.

The Metropolis-Hastings (MH) algorithm is an iterative sampling procedure that generates a sequence of random variables, $\Theta^0, \Theta^1, \ldots, \Theta^N$. In contrast to previous generation methods, like direct sampling, or rejection sampling, the value generated at iteration t depends on the value from iteration $t-1$, thus the resulting sample of values are not individually independent. MH has similarities with rejection sampling in that (at iteration t) there is a distribution that generates Θ^t and with a certain probability the generated value will be kept, i.e., added to the sample, otherwise it will be discarded. However, in contrast to rejection sampling, if the generated value is rejected, the previous value, Θ^{t-1}, is used as the t value: $\Theta^t = \Theta^{t-1}$.

In rejection sampling, we called the distribution used to generate Θ the envelope distribution, and we denoted it $g(\theta)$. In MH sampling that distribution is sometimes called the *proposal distribution*, which will denote it q and to indicate it dependence on previous values, it will sometimes be written $q(\Theta^t|\Theta^{t-1})$. We will call the generated value the *candidate value* and denote it Θ^c.

The steps in the MH algorithm are the following.

At iteration t, given a sample value at iteration $t-1$, Θ^{t-1}:

1. Generate a candidate value, Θ^c, from the proposal distribution, $q(\Theta^c|\Theta^{t-1})$.
2. Calculate the Metropolis Hastings ratio, MHR:

$$MHR(\Theta^{t-1}, \Theta^c) = \frac{p(\Theta^c)\ q(\Theta^{t-1}|\Theta^c)}{p(\Theta^{t-1})\ q(\Theta^c|\Theta^{t-1})} \tag{8.5}$$

   Note that this ratio can be larger than 1.
3. Generate a Uniform(0,1) random variable u.
4. If $u \leqslant \min\left(1, MHR(\Theta^{t-1}, \Theta^c)\right)$, then set $\Theta^t = \Theta^c$, thus "keep" Θ^c.
   Else set $\Theta^t = \Theta^{t-1}$.
5. Set $t = t+1$ and go back to Step 1

Note: this a generating samples from a Markov chain because the probability of selecting Θ^t is dependent on the value of Θ^{t-1}.

Key point from Bayesian perspective.

- The normalising constant of the posterior distribution does not need to be known due to the cancellation in MHR:

$$\begin{aligned} MHR(\Theta^{t-1}, \Theta^c) &= \frac{p(\Theta^c|\mathbf{y})\ q(\Theta^{t-1}|\Theta^c)}{p(\Theta^{t-1}|\mathbf{y})\ q(\Theta^c|\Theta^{t-1})} \\ &= \frac{\frac{\pi(\Theta^c)f(\mathbf{y}|\Theta^c)}{m(\mathbf{y})}\ q(\Theta^{t-1}|\Theta^c)}{\frac{\pi(\Theta^{t-1})f(\mathbf{y}|\Theta^{t-1})}{m(\mathbf{y})}\ q(\Theta^c|\Theta^{t-1})} \\ &= \frac{\pi(\Theta^c)f(\mathbf{y}|\Theta^c)\ q(\Theta^{t-1}|\Theta^c)}{\pi(\Theta^{t-1})f(\mathbf{y}|\Theta^{t-1})\ q(\Theta^c|\Theta^{t-1})} \end{aligned}$$

8.4 R Code

8.4.1 Simulation of the San Francisco Wet and Dry days Markov chain

```
#-- Demonstrating time homogeneous Markov chain, using the San Francisco wet and dry days
wet.dry.data <- matrix(c(418,256,256,884),nrow=2,ncol=2,byrow=TRUE,
                       dimnames=list(c("Wet.t-1","Dry.t-1"),c("Wet.t","Dry.t")))

P <- wet.dry.data/apply(wet.dry.data,1,sum)
pi.station <- c(P[2,1]/(P[1,2]+P[2,1]),P[1,2]/(P[1,2]+P[2,1]))

cat("stationary dist=",pi.station,"\n")
# 0.3715546 0.6284454

cat("p*P =\n")
rbind(pi.station) %*% P
#                Wet.t     Dry.t
#  pi.station 0.3715546 0.6284454

day.type <- c("Wet","Dry")

 set.seed(573)
  T <- 100000
  num.sims <- 3
  out <- matrix(data=NA,nrow=num.sims,ncol=T)

  for(i in 1:num.sims) {
   # 0=Wet, and 1= Dry (success)
   out[i,1]   <- rbinom(n=1,size=1,prob=0.5)
   for(j in 2:T) {
     #if wet, probability dry = 0.38
     #if dry, probability dry = 0.78
     p     <- P[out[i,j-1]+1,2]
     temp  <- rbinom(n=1,size=1,prob=p)
     out[i,j] <- temp
   }
  cat("Fraction dry=",sum(out[i,]==1)/T,"\n")
  }

check.points <- c(10,100,1000,10000,10000)
frac.dry.mat <- matrix(data=NA,nrow=num.sims,ncol=length(check.points),
    dimnames=list(paste("sim",1:num.sims),check.points))

for(i in 1:num.sims) {
  for(j in 1:length(check.points)) {
    frac.dry.mat[i,j] <- sum(out[i,1:check.points[j]]==1)/check.points[j]
  }
}

my.ylim <- c(-0.1,1.1)
time.block <- 500:530
for(i in 1:num.sims) {
  y <- out[i,time.block]
```

```
if(i==1) {
plot(time.block,y,type="l",xlab="Time",ylab="",
     main="Part of Markov chain",ylim=my.ylim)
} else {
  lines(time.block,y,col=i)
}
ok <- y==1
text(time.block[ok],jitter(y[ok],factor=2),"D",col=i)
text(time.block[!ok],jitter(y[!ok],factor=2),"W",col=i)
```

Homework Week 8: Bayesian Theory.

Optional Reading

1. Markov Chains. Chapter 1, Section 1.7 of Computational Statistics, 2nd Ed (2013), Givens and Hoeting, pp 14-17.
2. Metropolis-Hastings. Chapter 7, Section 7.1 of Givens and Hoeting, pp 201-209.

9 Dependent MC Sampling, Metropolis-Hastings & MCMC Diagnostics

9.1 Metropolis-Hastings Algorithm: Example with Hurricane Event Time Data

To demonstrate the Metropolis-Hastings algorithm in a simple case, we return to the time between hurricane data in Lecture 6 where the Weibull distribution was the likelihood for shape parameter α, and a Gamma distribution was used as the prior for α. Then the posterior:

$$p(\alpha|\mathbf{y}) \propto \pi(\alpha) f(\mathbf{y}|\alpha) = \frac{\lambda^\kappa}{\Gamma(\kappa)} \alpha^{\kappa-1} \exp(-\lambda\alpha) \times \alpha^n \exp\left(-\sum_{i=1}^n y_i^\alpha\right) \prod_{i=1}^n y_i^{\alpha-1} \tag{9.1}$$

With the Metropolis-Hastings algorithm, the normalising constant can be ignored. In addition all terms in the joint distribution not involving α can be ignored.

Pseudo-code for the t^{th} iteration of the algorithm is the following:

1. Generate candidate value, α^c, from the proposal, $q(\alpha|\alpha^{t-1})$.
2. Evaluate the MHR:

$$MHR(\alpha^{t-1}, \alpha^c) = \frac{(\alpha^c)^{\kappa-1} \exp(-\lambda\alpha^c) \times (\alpha^c)^n \exp\left(-\sum_{i=1}^n y_i^{\alpha^c}\right) \prod_{i=1}^n y_i^{\alpha^c-1}}{(\alpha^{t-1})^{\kappa-1} \exp(-\lambda\alpha^{t-1}) \times (\alpha^{t-1})^n \exp\left(-\sum_{i=1}^n y_i^{\alpha^{t-1}}\right) \prod_{i=1}^n y_i^{\alpha^{t-1}-1}} \times \frac{q(\alpha^{t-1}}{q(\alpha^c|\alpha}$$

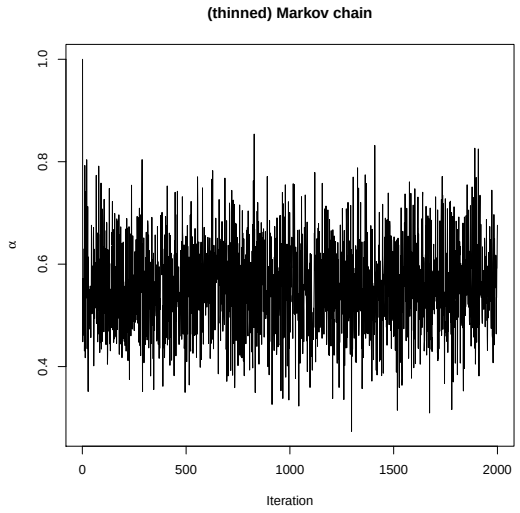
3. Generate a Uniform(0,1) random variable, u^*.
   If $u^* \leqslant \min(1, MHR)$, set α^t=α^c, else set α^t=α^{t-1}.

R code to generate a sample from the posterior using a $q(\alpha|\alpha^{t-1})$ = Gamma(2,3) proposal distribution is in Appendix 9.7.1. A portion of the code is shown below. The MHR has been calculated on the log scale first to lessen the chance of under- or over-flows.

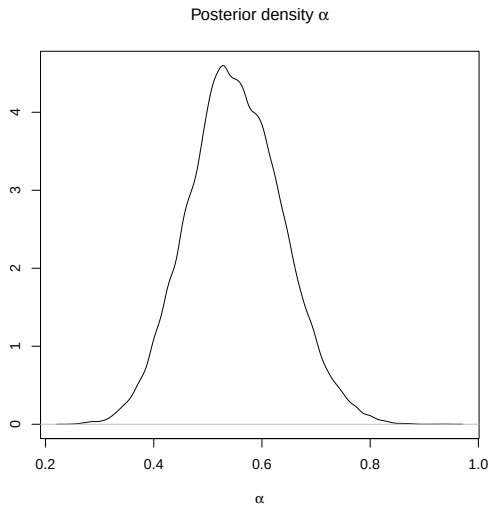
```
tally <- tally+1
candidate <- rgamma(n=1,shape=a.q,rate=b.q)
previous  <- alpha.star[tally-1]

log.joint.cand <- (n+kappa-1)*log(candidate) -  sum(hurricane.gaps^candidate) -
                  lambda*candidate +   (candidate-1)*sum(log(hurricane.gaps))
log.joint.prev <- (n+kappa-1)*log(previous) -  sum(hurricane.gaps^previous) -
```

```
                  lambda*previous +   (previous-1)*sum(log(hurricane.gaps))
  log.q.cand     <- dgamma(candidate,a.q,b.q,log=TRUE)
  log.q.prev     <- dgamma(previous,a.q,b.q,log=TRUE)
  MHR            <- exp(log.joint.cand + log.q.prev - log.joint.prev - log.q.cand)

  u.star         <- runif(1)
  if(u.star <= min(1,MHR)) {
    # keep the candidate
    alpha.star[tally] <- candidate
    number.accept <- number.accept+1
  } else {
    # reject the candidate
    alpha.star[tally] <- previous
  }
```

A Markov chain of length N=100,000 was generated with an acceptance rate of 0.26. The resulting α values (thinned to show every 50th value) are plotted in Figure 9.1a and the posterior distribution is plotted in Figure 9.1b.



(a) Trace Plot of α for the hurricane event time data based on a Weibull(α,$\beta = 1$) distribution with a Gamma(0.1,0.1) prior and Gamma(2,3) proposal distribution. Every 50th value is shown.



(b) Posterior distribution for α for the hurricane event time data based on a Weibull(α,$\beta = 1$) distribution with a Gamma(0.1,0.1) prior and Gamma(2,3) proposal distribution. B=1000.

Using a Burn-in B=1000, the following summary statistics were calculated:

```
summary(alpha.star)
#     Min. 1st Qu.  Median    Mean   3rd Qu.    Max.
#  0.2443  0.4919   0.5489   0.5512  0.6097   0.9477
```

Arbitrary probabilities can be calculated as well, such as $\Pr(0.40 \leqslant \alpha \leqslant 0.71)$:

$$\widehat{\Pr}(0.40 \leqslant \alpha \leqslant 0.71) = \frac{1}{N-B} \sum_{t=B+1}^{N} I(0.40 \leqslant \alpha^t \leqslant 0.71)$$

Using R:

```
sum(alpha.star>=0.40 & alpha.star<=0.71)/(N-Burnin) # 0.9257273
```

9.2 Metropolis-Hastings Algorithm: Why it works

We will apply the results from Lecture Notes 8 on discrete time, discrete state space Markov chains to include discrete time, continuous state space Markov chains, but we do not explain or justify the theory that makes this acceptable, however.

We will consider the case where the generated random variable, θ, is a scalar.

Recall from Lecture Notes 8 that:

1. Reversible Markov chains, namely,
$$\pi(i)\Pr(\theta^t = j|\theta^{t-1} = i) = \pi(j)\Pr(\theta^t = i|\theta^{t-1} = j) \tag{9.2}$$
have stationary distributions.
2. If an irreducible, aperiodic Markov chain has a stationary distribution, π, then that stationary distribution is unique and is the limiting distribution:
$$\lim_{n\to\infty} \Pr(\theta^{t+n} = j|\theta^t = i) = \pi(j) \tag{9.3}$$

Thus if a Markov chain is irreducible, aperiodic, and reversible, then it has a unique stationary and limiting distribution.

Claim. The Metropolis-Hastings Algorithm satisfies the above two conditions and is constructed such that the stationary and limiting distribution is the target distribution.

Proof. (This is based on *Introduction to Probability Models, 8th Ed.*, S. Ross.) This proof assumes a discrete, countable state space. Without loss of generality, let the state space be the set of positive integers, S=$\{1, 2, \ldots, \}$. Let $p(\theta)$ be a probability mass function with $\theta \in S$.

- Let θ^t be a stochastic process (a time indexed random variable) and,
- Let Q be any specified irreducible Markov tpm on S where $q(j|i)$ is the jth column of row i.
- Given θ^{t-1}=o, where $o \in S$ (and stands for "old"), generate a random variable Y such that Y=c with probability $q(c|o)$, where $c \in S$ (and stands for "candidate").
- If Y=c, set $\theta^t = c$ with probability $\alpha(o, c)$, otherwise set $\theta^t = o$. The process θ^t is thus a Markov chain:
$$\Pr(\theta^t = c|\theta^{t-1} = o) = q(c|o)\alpha(o, c), c \neq o$$
$$\Pr(\theta^t = o|\theta^{t-1} = o) = q(o|o) + \sum_{k\neq o} q(k|o)(1 - \alpha(o, k))$$
- θ^t will be a reversible Markov chain with stationary probability mass function $p(\theta)$ if
$$p(o)\Pr(\theta^t = c|\theta^{t-1} = o) = p(c)\Pr(\theta^t = o|\theta^{t-1} = c), \quad \forall c \neq o \tag{9.4}$$
- Define $\alpha(o, c)$ as follows[1]:
$$\alpha(o, c) = \min\left(1, \frac{p(c)q(o|c)}{p(o)q(c|o)}\right) \tag{9.5}$$
- There are two cases to consider which affect the value of $\alpha(o, c)$:

[1] This was the ingenious part.

– Case 1: $p(c)q(o|c) < p(o)q(c|o)$. Then

$$\alpha(o, c) = \frac{p(c)q(o|c)}{p(o)q(c|o)} \text{ and } \alpha(c, o) = 1$$

and

$$\begin{aligned} p(o)\Pr(\theta^t = c|\theta^{t-1} = o) &= p(o)q(c|o)\alpha(o, c) = p(o)q(c|o)\frac{p(c)q(o|c)}{p(o)q(c|o)} \\ &= p(c)q(o|c) = p(c)q(o|c)\alpha(c, o) = p(c)\Pr(\theta^t = o|\theta^{t-1} = c) \end{aligned}$$

– Case 2: $p(c)q(o|c) \geqslant p(o)q(c|o)$. Then

$$\alpha(o, c) = 1 \text{ and } \alpha(c, o) = \frac{p(o)q(c|o)}{p(c)q(o|c)}$$

and

$$\begin{aligned} p(c)\Pr(\theta^t = o|\theta^{t-1} = c) &= p(c)q(o|c)\alpha(c, o) = p(c)q(o|c)\frac{p(o)q(c|o)}{p(c)q(o|c)} \\ &= p(o)q(c|o) = p(o)q(c|o)\alpha(o, c) = p(o)\Pr(\theta^t = c|\theta^{t-1} = o) \end{aligned}$$

Thus θ^t is a reversible Markov chain with stationary probability $p(\theta)$.

9.3 Metropolis-Hastings Algorithm: Special case proposals

9.3.1 Random walk proposals

A random walk proposal has the following structure.

$$\theta^c = \theta^o + \epsilon$$

where θ^o is the old value and θ^c is the candidate value and ϵ is a mean zero random variable with probability distribution, f. Some special cases:

$$\begin{aligned} &\epsilon \sim \text{Normal}(0, \sigma^2) \\ &\epsilon \sim t_{2\ df} \\ &\epsilon \sim \text{Uniform}(-b, b) \end{aligned}$$

These three cases are examples of *symmetric* proposal distributions:

$$q(\theta^c|\theta^o) = q(\theta^o|\theta^c) \tag{9.6}$$

For example, if $\epsilon \sim$ Normal(0,σ^2),

$$q(\theta^c|\theta^o) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(\theta^c - \theta^o)^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(\theta^o - \theta^c)^2\right) = q(\theta^o|\theta^c)$$

When the proposal is symmetric,

$$\frac{q(\theta^o|\theta^c)}{q(\theta^c|\theta^o)} = 1$$

and the Metropolis Hastings Ratio simplifies to

$$MHR(\theta^o, \theta^c) = \frac{p(\theta^c)q(\theta^o|\theta^c)}{p(\theta^o)q(\theta^c|\theta^o)} = \frac{p(\theta^c)}{p(\theta^o)} \tag{9.7}$$

Note symmetric distributions often readily extend to multivariate settings, e.g, use a Bivariate normal as a proposal for Θ=(θ_1, θ_2).

Metropolis sampler. Warning: this paragraph is incorrect, as it ignores the asymmetric supports of the two pdfs, which makes the MHR infinite (or undefined) due to division by zero for $\epsilon > 1$; the proposal kernel doesn't permit going back far enough. Random walk proposals are a special case of a Metropolis sampler. A Metropolis sampler is one where $q(\theta^c|\theta^o)$ = $q(\theta^o|\theta^c)$ and thus $q(\theta^o|\theta^c)/q(\theta^c|\theta^o)$ = 1, and thus the proposal need not be evaluated. An example of a Metropolis sampler that is *not* a random walk is θ^c = θ^o + ϵ, $\epsilon \sim$ Uniform(-1, 2), and the pdf for $\theta^c|\theta^o$ = pdf for $\theta^o|\theta^c$ = 1/3, so long at the support for θ is unbounded.

Transforming parameters to facilitate random walk proposals. When the support of a parameter θ is restricted, e.g., $\theta > 0$, then a random walk proposal will be wasteful when a proposed value is outside of the support of θ. For example, suppose $\theta > 0$, and the proposal is Uniform($\theta^o - 1$, $\theta^o + 1$). Suppose θ^o=0.2, then the proposal is Uniform(-0.8,1.2) and there is a 0.8/2 = 0.4 probability of getting $\theta^c \leqslant 0$. If $\theta^c < 0$, $p(\theta^c)$=0, so the value will be rejected.

A less wasteful procedure is to transform a range restricted parameter with an invertible mapping to the real number line, say ϕ=$g(\theta)$, where g is a 1:1 function, and then applying a random walk proposal to the transformed parameter. The prior distribution for the transformed parameter is then

$$\pi_\phi(\phi) = \pi_\theta(g^{-1}(\phi)) \left| \frac{dg^{-1}}{d\phi} \right|$$

For example, the sampling distribution for the data, y, is Exponential(θ) distribution, θ must be greater than 0. Suppose the prior distribution for θ is Gamma(α, β). Define ϕ = $g(\theta)$ = $\ln(\theta)$, then $g^{-1}(\phi)$ = $\exp(\phi)$ and $dg^{-1}(\phi)/d\phi$ = $\exp(\phi)$. The prior distribution for ϕ:

$$\begin{aligned}\pi_\phi(\phi) &= \frac{\beta^\alpha}{\Gamma(\alpha)}(\exp(\phi))^{\alpha-1}\exp(-\beta * \exp(\phi)) \times \exp(\phi) \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}(\exp(\phi))^{\alpha}\exp(-\beta * \exp(\phi))\end{aligned}$$

The data distribution (the likelihood) will need to be rewritten in terms of ϕ, i.e., substitute $g^{-1}(\phi)$ for θ. Then use random walk proposal for ϕ such as

$$\phi^c = \phi^o + \epsilon_t$$

where $\epsilon_t \sim$ Normal(0,0.5), and one is left with a Metropolis sampler where the proposal distributions are not evaluated for the MHR.

9.3.2 Independence proposals

An independence proposal is one where the proposal does not depend on the previous value, θ^o. For example, $\theta^c \sim$ Normal(0,10). The example of the Weibull given in Section 9.1 of the Gamma(2,3) proposal was an independence proposal.

The MHR is then:

$$MHR(\theta^o, \theta^c) = \frac{p(\theta^c)q(\theta^o)}{p(\theta^o)q(\theta^c)}$$

Note that the generated values are still coming from a Markov chain because the transition density or kernel is the product of the proposal density and the Metropolis-Hastings ratio.

9.4 Multidimensional Θ: One-at-a-time or Single updates

In the case of multidimensional $\Theta=(\theta_1, \theta_2, \ldots, \theta_k)$, the entire vector can be updated at once, or subsets ("blocks") can be updated, or each individual parameter, θ_i, i=1,2,...,q can be updated one-at-a time. The one-at-a-time update leads to some simplification in the calculation of the Metropolis Hasting ratio.

For example, let $\Theta= (\theta_1, \theta_2, \theta_3)$. Set the initial values, $(\theta_1^0, \theta_2^0, \theta_3^0)$. The values will be updated in the order 1, 2, and 3. First generate a candidate value for θ_1 from $q(\theta_1^c|\theta_1^0, \theta_2^0, \theta_3^0)$, then evaluate the MHR:

$$\begin{aligned} MHR(\theta_1^0, \theta_1^c) &= \frac{p(\theta_1^c, \theta_2^0, \theta_3^0)q(\theta_1^0|\theta_1^c, \theta_2^0, \theta_3^0)}{p(\theta_1^0, \theta_2^0, \theta_3^0)q(\theta_1^c|\theta_1^0, \theta_2^0, \theta_3^0)} \\ &= \frac{p(\theta_1^c|\theta_2^0, \theta_3^0)p(\theta_2^0, \theta_3^0)q(\theta_1^0|\theta_1^c, \theta_2^0, \theta_3^0)}{p(\theta_1^0|\theta_2^0, \theta_3^0)p(\theta_2^0, \theta_3^0)q(\theta_1^c|\theta_1^0, \theta_2^0, \theta_3^0)} \\ &= \frac{p(\theta_1^c|\theta_2^0, \theta_3^0)q(\theta_1^0|\theta_1^c, \theta_2^0, \theta_3^0)}{p(\theta_1^0|\theta_2^0, \theta_3^0)q(\theta_1^c|\theta_1^0, \theta_2^0, \theta_3^0)} \end{aligned}$$

Note the cancellation of the joint distribution of θ_2^0, θ_3^0. Thus only terms involving θ_1 in $p(\theta_1, \theta_2, \theta_3)$ need to evaluated when calculating the MHR. Generate U^* from Uniform(0,1) and set $\theta_1^1 = \theta_1^c$ if $U^* \leqslant \min(1, MHR(\theta_1^0, \theta_1^c))$, else set θ_1^1=θ_1^0.

Next generate a candidate value for θ_2 from $q(\theta_2^c|\theta_1^1, \theta_2^0, \theta_3^0)$. Then evaluate the MHR:

$$\begin{aligned} MHR(\theta_2^0, \theta_2^c) &= \frac{p(\theta_1^1, \theta_2^c, \theta_3^0)q(\theta_2^0|\theta_1^1, \theta_2^c, \theta_3^0)}{p(\theta_1^1, \theta_2^0, \theta_3^0)q(\theta_2^c|\theta_1^1, \theta_2^0, \theta_3^0)} \\ &= \frac{p(\theta_2^c|\theta_1^1, \theta_3^0)q(\theta_2^0|\theta_1^1, \theta_2^c, \theta_3^0)}{p(\theta_2^0|\theta_1^1, \theta_3^0)q(\theta_2^c|\theta_1^1, \theta_2^0, \theta_3^0)} \end{aligned}$$

Again generate U^* from Uniform(0,1) and set $\theta_2^1 = \theta_2^c$ if $U^* \leqslant \min(1, MHR(\theta_2^0, \theta_2^c))$, else set θ_2^1=θ_2^0.

To complete the update at iteration 1: generate a candidate value for θ_3 from $q(\theta_3^c|\theta_1^1, \theta_2^1, \theta_3^0)$. Then evaluate the MHR:

$$\begin{aligned} MHR(\theta_3^0, \theta_3^c) &= \frac{p(\theta_1^1, \theta_2^1, \theta_3^c)q(\theta_3^0|\theta_1^1, \theta_2^1, \theta_3^c)}{p(\theta_1^1, \theta_2^1, \theta_3^0)q(\theta_3^c|\theta_1^1, \theta_2^1, \theta_3^0)} \\ &= \frac{p(\theta_3^c|\theta_1^1, \theta_2^1)q(\theta_3^0|\theta_1^1, \theta_2^1, \theta_3^c)}{p(\theta_3^0|\theta_1^1, \theta_2^1)q(\theta_3^c|\theta_1^1, \theta_2^1, \theta_3^0)} \end{aligned}$$

Again generate U^* from Uniform(0,1) and set $\theta_3^1 = \theta_3^c$ if $U^* \leqslant \min(1, MHR(\theta_3^0, \theta_3^c))$, else set θ_3^1=θ_3^0.

9.5 Metropolis-Hastings example with 2 parameters

To demonstrate Metropolis-Hastings in the multiparameter case, we will use the between event times hurricane data but now estimate both parameters of the Weibull distribution, α and β. The Weibull distribution pdf is:

$$f(y|\alpha, \beta) = \frac{\alpha}{\beta}\left(\frac{y}{\beta}\right)^{\alpha-1} \exp\left(-(y/\beta)^{\alpha}\right) \tag{9.8}$$

The joint distribution for a sample of size n:

$$f(\mathbf{y}|\alpha,\beta) = \left(\frac{\alpha}{\beta}\right)^n \prod_{i=1}^{n} \left(\frac{y_i}{\beta}\right)^{\alpha-1} \exp\left[-\sum_{i=1}^{n}(y_i/\beta)^\alpha\right] \tag{9.9}$$

For simplicity, we will use identical, and independent Gamma(κ,λ) priors for α and β. Thus the prior pdf:

$$\pi(\alpha,\beta) = \left(\frac{\lambda^\kappa}{\Gamma(\kappa)}\right)^2 \alpha^{\kappa-1}\beta^{\kappa-1}\exp(-\lambda\alpha)\exp(-\lambda\beta) \propto (\alpha\beta)^{\kappa-1}\exp(-\lambda(\alpha+\beta)) \tag{9.10}$$

The normalising constants can be ignored because they will cancel in the MHR.

To lessen the chance of underflows/overflows we will work with log transformed values in the Metropolis Hasting ratio and then exponentiate at the end. The log of the product of the priors and the likelihood is then:

$$\ln\left[\pi(\alpha,\beta)f(\mathbf{y}|\alpha,\beta)\right] = (\kappa-1)\ln(\alpha\beta) - \lambda(\alpha+\beta) + n\ln(\alpha/\beta) + \sum_{i=1}^{n}(\alpha-1)\ln(y_i/\beta) - \sum_{i=1}^{n}(y_i/\beta)^\alpha \tag{9.11}$$

For the proposals we will use a normal random walk and update both parameters simultaneously (as a pair):

$$\alpha^c \sim \text{Normal}\left(\alpha^{t-1}, \sigma_\alpha\right) \tag{9.12}$$
$$\beta^c \sim \text{Normal}\left(\beta^{t-1}, \sigma_\beta\right) \tag{9.13}$$

If either of the candidate values are less than or equal to zero, they will be rejected. The values of σ_α and σ_β will be experimented with to keep the probability of rejection relatively low.

With a symmetric proposal, we have a Metropolis algorithm and the MHR simplifies to the following:

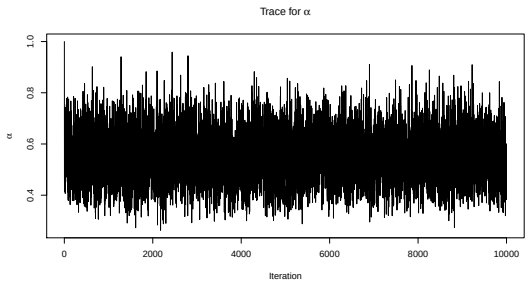
$$\begin{aligned}MHR([\alpha^{t-1},\beta^{t-1}],[\alpha^c,\beta^c]) = \exp\Bigg[&(\kappa-1)\ln(\alpha^c\beta^c) - \lambda(\alpha^c+\beta^c) + n\ln(\alpha^c/\beta^c) + \sum_{i=1}^{n}(\alpha^c-1)\ln(y_i/\beta^c) - \sum_{i=1}^{n}(y_i/\beta^c)^{\alpha^c} \\ &- (\kappa-1)\ln(\alpha^{t-1}\beta^{t-1}) + \lambda(\alpha^{t-1}+\beta^{t-1}) - n\ln(\alpha^{t-1}/\beta^{t-1}) - \sum_{i=1}^{n}(\alpha^{t-1}-1)\ln(y_i/\beta^{t-1}) + \sum_{i=1}^{n}(y_i/\beta^{t-1})\end{aligned} \tag{9.14}$$

Complete R code for the sampler is shown in Appendix 9.7.2. A portion of the code is shown below.

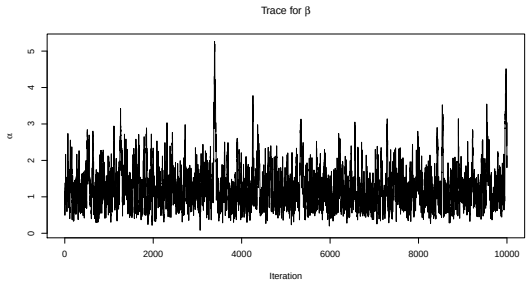
```
tally         <- 1
number.accept <- 0
while(tally < N) {
  tally <- tally+1
  # Proposals are Normal(alpha^{-1},sigma.alpha), Normal(beta^{-1},sigma.beta), thus
  # symmetric (random walk)
  alpha.c  <- rnorm(n=1,mean=alpha.star[tally-1],sd=sigma.alpha)
  beta.c   <- rnorm(n=1,mean=beta.star[tally-1],sd=sigma.beta)

  if(alpha.c < 0 || beta.c < 0) {
```

```
    alpha.star[tally] <- alpha.star[tally-1]
    beta.star[tally]  <- beta.star[tally-1]

  } else {

  alpha.p <- alpha.star[tally-1]
  beta.p  <- beta.star[tally-1]

  log.joint.cand <- (kappa-1)*log(alpha.c*beta.c)-lambda*(alpha.c+beta.c) +
     n*log(alpha.c/beta.c) + (alpha.c-1)*sum(log(y/beta.c)) -
     sum((y/beta.c)^alpha.c)

  log.joint.prev <- (kappa-1)*log(alpha.p*beta.p)-lambda*(alpha.p+beta.p) +
  n*log(alpha.p/beta.p) + (alpha.p-1)*sum(log(y/beta.p)) -
  sum((y/beta.p)^alpha.p)

  MHR            <- exp(log.joint.cand - log.joint.prev)
  u.star         <- runif(1)
  if(u.star <= min(1,MHR)) {
    # keep the candidate
    alpha.star[tally] <- alpha.c
    beta.star[tally]  <- beta.c
    number.accept <- number.accept+1
  } else {
    # reject the candidate
    alpha.star[tally] <- alpha.p
    beta.star[tally]  <- beta.p
   }
  }
}
```

The trace plots for α and β are shown in Figures 9.2a and 9.2b.



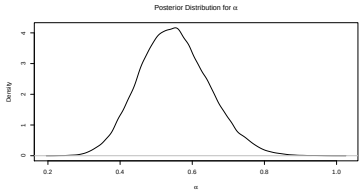
(a) Markov chain of α for the hurricane event time data based on a Weibull(α,β) distribution with a Gamma(0.1,0.1) prior and normal random walk proposal.



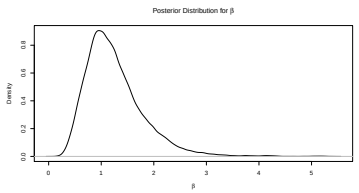
(b) Markov chain of β for the hurricane event time data based on a Weibull(α,β) distribution with a Gamma(0.1,0.1) priors and normal random walk proposal.

The posterior densities (without burn-in) are shown in Figures 9.3a and 9.3b.

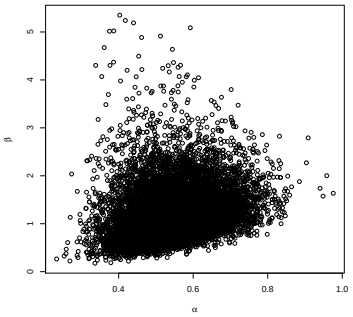
Summary statistics are given below.



(a) Posterior distribution for α for the hurricane event time data based on a Weibull(α,β) distribution with a Gamma(0.1,0.1) prior and and normal random walk proposal.



(b) Posterior distribution for β for the hurricane event time data based on a Weibull(α,β) distribution with a Gamma(0.1,0.1) prior and and normal random walk proposal.



(c) Scatterplot of β versus α for the hurricane event time data.

```
  summary(alpha.star)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2214  0.4830  0.5463  0.5501  0.6130  0.9767
  summary(beta.star)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.08241 0.86944 1.15283 1.25528 1.52438 5.42604
```

Note that the posterior mean for β is similar to the previously fixed value of β=1, and the posterior distribution for α is similar to that shown in Section 9.1. The correlation between α and β was estimated to be 0.22 and Figure 9.3c is a scatterplot of β versus α.

9.6 MCMC diagnostics

Inference using MCMC output is based on the sequence of samples $\mathrm{B}+1$, $\mathrm{B}+2$, . . ., $\mathrm{B}+\mathrm{N}$. Two reasonable questions then are

- What should B (burn-in length) be?
- How large should N, or really $\mathrm{N}+\mathrm{B}$ be?

As mentioned previously (Lecture Notes 8), some argue that discarding initial samples, the burn-in B, is wasteful and that a long enough chain (N) compensates for initial lack of convergence to the target distribution. We will not enter into that debate here, however, and simply note that that is a minority opinion—most users of MCMC output do discard initial chain values.

9.6.1 Burn-in

The ideal burn-in length B is the point at which the chain has run long enough that the resulting sample distribution equals the limiting distribution (the target distribution), i.e., the chain has converged.

Here we will discuss a couple of commonly used methods for specifying burn-in length. We note that none of these methods, however, *prove* that the chain has reached its limiting dis-

tribution. Instead the methods, more or less, help determine *if* the chain has *not* reached its limiting distribution.

Trace plots/Sample paths

The simplest method for specifying a burn-in length is to create a "time series" plot of the generated values, what is called a trace plot, as in Figures 9.1a, 9.2a, and 9.2b. These particular trace plots have been thinned (to reduce file size) but there is no evidence for any kind of pattern of being "stuck" in a certain region of the state space.

"Rapid" *apparent* movement about the state space is called *good mixing*.

Multiple Chains

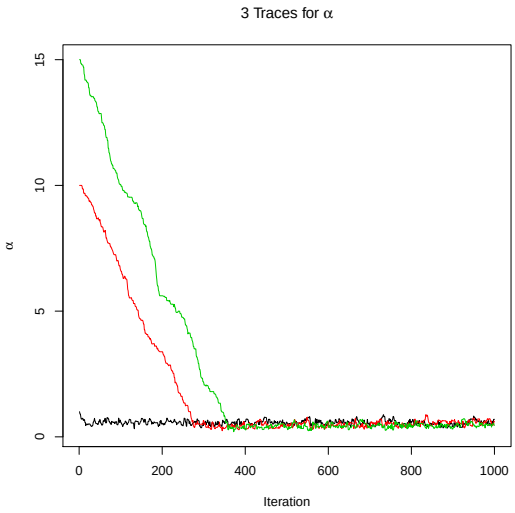
However, looking at a single chain's trace plots has limited value as it is possible that there are multiple modes in the posterior and only one mode is being sampled. A better approach is to run multiple chains with widely differing initial values and examine if and when the chains start to overlap. Figures 9.4a and 9.4b show the results for three chains for the hurricane data with initial values for α and β equal to 1, 10, and 15. The "convergence" of the α chains is considerably sooner than for the β parameter.

Again such plots do not prove convergence but they can indicate a lack of convergence. A burn-in pf 500 might be adequate for α while the burn-in for β might be at least 1,000. Sample values beyond these respective burn-in values do suggest "good mixing".

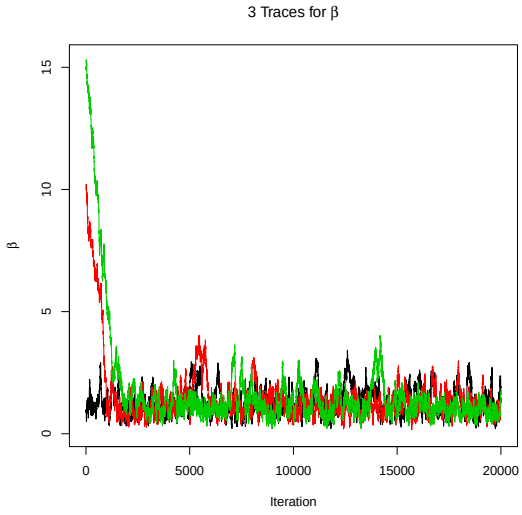
A more quantitative assessment of mixing (or lack of) is the Brooks-Gelman-Rubin (BGR) statistic. The idea is to compare the variability in the simulated values within each chain to the variability between chains. When chains have converged the variation within and between chains should be quite similar. The BGR statistic, denoted R, is defined as follows:

$$\mathrm{R} = \frac{\text{Width of 80\% credible interval of all chains combined}}{\text{Average width of 80\% credible intervals for the individual chains}}$$

If the chain has converged, then R should be around 1. If the chains have not converged, then the width of the interval based on all chains combined will be greater than 1, as the between chain variation will be "large". The underlying principle is similar to that of an F-statistic in an analysis of variance (ANOVA) of K treatments, where K corresponds to the number of chains.



(a) Trace plots for α based on three different initial values, α^0 = 1, 10, and 15.



(b) Trace plots for β based on three different initial values, β^0 = 1, 10, and 15.

9.7 R Code

9.7.1 Metropolis-Hastings algorithm applied to hurricane event times data (α)

```
#-- Metropolis Hastings applied to the time between hurricanes data set
# with Weibull sampling distribution and Gamma prior for alpha
hurricane.gaps <- c(0.30, 4.61, 5.75, 0.24, 0.09, 0.18, 7.38, 1.20, 2.40, 0.18,
                    0.02, 10.07, 0.23, 0.44, 3.34, 0.06, 0.01, 0.71, 0.06, 0.42)
n <- length(hurricane.gaps)

set.seed(730)
N          <- 100000
kappa      <- lambda <- 0.1 #parameters for prior
a.q        <- 2             #parameters for proposal
b.q        <- 3
alpha.star <- numeric(N)
alpha.star[1] <- 1          # initial value

#--- To lessen change of underflows/overflows with MHR
# will log and then exponentiate

tally         <- 1
number.accept <- 0
while(tally < N) {
  tally <- tally+1
  # Proposal is a Gamma(a.q,b.q)
  candidate <- rgamma(n=1,shape=a.q,rate=b.q)
  previous  <- alpha.star[tally-1]
  log.joint.cand <- (n+kappa-1)*log(candidate) -  sum(hurricane.gaps^candidate) -
                    lambda*candidate +   (candidate-1)*sum(log(hurricane.gaps))
  log.joint.prev <- (n+kappa-1)*log(previous) -  sum(hurricane.gaps^previous) -
                    lambda*previous +   (previous-1)*sum(log(hurricane.gaps))
  log.q.cand     <- dgamma(candidate,a.q,b.q,log=TRUE)
  log.q.prev     <- dgamma(previous,a.q,b.q,log=TRUE)
  MHR            <- exp(log.joint.cand + log.q.prev - log.joint.prev - log.q.cand)
  u.star         <- runif(1)
  if(u.star <= min(1,MHR)) {
    # keep the candidate
    alpha.star[tally] <- candidate
    number.accept <- number.accept+1
  } else {
    # reject the candidate
    alpha.star[tally] <- previous
  }
}

cat("Acceptance rate=",round(number.accept/N,3),"\n")
# Acceptance rate= 0.262
#thin for plotting to reduce file size
plot(alpha.star[seq(1,N,length=2000)],type="l",
     xlab="Iteration",ylab=expression(alpha),main="(thinned) Markov chain")
```

```
plot(density(alpha.star),type="l",xlab=expression(alpha),
     main=expression(paste("Posterior Distribution for ",alpha)))

Burnin <- 1000
alpha.star <- alpha.star[-c(1:Burnin)]
summary(alpha.star)
#  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0.2443  0.4919  0.5489  0.5512  0.6097  0.9477

cat("Probability alpha between 0.40 and 0.71=",
    sum(alpha.star>=0.40 & alpha.star<=0.71)/(N-Burnin),"\n")
# Probability alpha between 0.40 and 0.71= 0.9257273
```

9.7.2 MH joint update of α and β for hurricane data

```
#-- Metropolis-Hastings updating 2 parameters: first will do joint updating.
hurricane.gaps <- c(0.30, 4.61, 5.75, 0.24, 0.09, 0.18, 7.38, 1.20, 2.40, 0.18,
                    0.02, 10.07, 0.23, 0.44, 3.34, 0.06, 0.01, 0.71, 0.06, 0.42)
n <- length(hurricane.gaps)

set.seed(730)
N          <- 100000
kappa      <- lambda <- 0.1 #parameters for priors
sigma.alpha <- 0.1          #parameters for proposal
sigma.beta  <- 0.1
alpha.star <- beta.star <- numeric(N)
alpha.star[1] <- 1          # initial values
beta.star[1]  <- 1

#--- To lessen change of underflows/overflows with MHR
# will log and then exponentiate

y <- hurricane.gaps

tally         <- 1
number.accept <- 0
while(tally < N) {
  tally <- tally+1
  # Proposals are Normal(alpha^{-1},sigma.alpha), Normal(beta^{-1},sigma.beta), thus
  # symmetric (random walk)
  alpha.c  <- rnorm(n=1,mean=alpha.star[tally-1],sd=sigma.alpha)
  beta.c   <- rnorm(n=1,mean=beta.star[tally-1],sd=sigma.beta)

  if(alpha.c < 0 || beta.c < 0) {
     alpha.star[tally] <- alpha.star[tally-1]
     beta.star[tally]  <- beta.star[tally-1]

  } else {

  alpha.p <- alpha.star[tally-1]
  beta.p  <- beta.star[tally-1]
```

```
  log.joint.cand <- (kappa-1)*log(alpha.c*beta.c)-lambda*(alpha.c+beta.c) +
     n*log(alpha.c/beta.c) + (alpha.c-1)*sum(log(y/beta.c)) -
     sum((y/beta.c)^alpha.c)

  log.joint.prev <- (kappa-1)*log(alpha.p*beta.p)-lambda*(alpha.p+beta.p) +
  n*log(alpha.p/beta.p) + (alpha.p-1)*sum(log(y/beta.p)) -
  sum((y/beta.p)^alpha.p)

  MHR            <- exp(log.joint.cand - log.joint.prev)
  u.star         <- runif(1)
  if(u.star <= min(1,MHR)) {
    # keep the candidate
    alpha.star[tally] <- alpha.c
    beta.star[tally]  <- beta.c
    number.accept <- number.accept+1
  } else {
    # reject the candidate
    alpha.star[tally] <- alpha.p
    beta.star[tally]  <- beta.p
   }
  }
}

cat("Acceptance rate=",round(number.accept/N,3),"\n")
# Acceptance rate= 0.658

#Trace plots and densities
thin.N <- 10000
plot(alpha.star[seq(1,N,length=thin.N)],type="l",
     xlab="Iteration",ylab=expression(alpha),main=expression(paste("Trace for ",alpha)))

plot(density(alpha.star),type="l",xlab=expression(alpha),
     main=expression(paste("Posterior Distribution for ",alpha)))

plot(beta.star[seq(1,N,length=thin.N)],type="l",
     xlab="Iteration",ylab=expression(alpha),main=expression(paste("Trace for ",beta)))

plot(density(beta.star),type="l",xlab=expression(beta),
     main=expression(paste("Posterior Distribution for ",beta)))

# distributional summaries
Burnin <- 1000
alpha.star <- alpha.star[-c(1:Burnin)]
beta.star  <- beta.star[-c(1:Burnin)]
summary(alpha.star)
#   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#  0.2214  0.4830  0.5463  0.5501  0.6130  0.9767
summary(beta.star)
#    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#  0.08241 0.86944 1.15283 1.25528 1.52438 5.42604
```

9.8 L9A: Brief comments on Model Selection or Evaluation

Posterior model probabilities (which can be viewed as probabilities about different hypotheses) are just one aspect of the general procedure of evaluating or selecting models.

Paraphrasing (and quoting) Carlin and Louis (2009), there are three general aspects to evaluating and selecting model.

1. Model Robustness: How sensitive are my results to the assumptions made?
2. Model Assessment: How well does my model fit the data? (and somewhat related—how well might my model predict new observations?)
3. Model Selection: "Which model (or models) should I ultimately choose for the final presentation of my results?"

We won't go into detail here on answers to any of these questions and simply want to draw attention to them early on.

Remarks

- re: Model Robustness. One thing to do is to examine the effect of the priors on the posterior distribution. For example, with a normal linear regression model for the observations:

  $$y \sim \text{Normal}\left(\beta_0 + \beta_1 x, \sigma^2\right)$$

  Suppose the initial prior for σ^2 was Inverse Gamma(7,20) (with a mean of 3.3 and standard deviation of 1.49). How much would the posterior mean for σ^2 change with a Uniform(0,15) prior?

- re: Model Assessment. There are several measures. We just mention one here that can be used to check for outliers: standardized residuals given a training set and a test set. Here the n data values have been divided (randomly) into two sets, n_{train} and n_{test} where $n = n_{train} + n_{test}$[2]. Denote the values in the training set by $z = (z_1, z_2, \ldots, z_{n_{train}})$ and those in the test set by y_i, $i = 1, \ldots, n_{test}$. The model is then fit using the training set, and Bayesian residuals are defined as:

  $$r_i = y_i - \mathrm{E}(Y_i|z)$$

  where $\mathrm{E}(Y_i|z)$ is the expected value of the posterior predictive distribution for Y_i. The standardized residual, which removes the effects of scale, is

  $$d_i = \frac{r_i}{\sqrt{\mathrm{Var}(Y_i|z)}} = \frac{y_i - \mathrm{E}(Y_i|z)}{\sqrt{\mathrm{Var}(Y_i|z)}}$$

- re: Model Selection. Posterior model probabilities and Bayes factors can be used for model selection. However, they are measures of the *relative* "evidence" for one model versus another. For example, if there are three models, M1, M2, and M3, with posterior probabilities of 0.15, 0.05, and 0.80, then model M3 seems the "best" choice amongst the three.

  However, that does NOT mean that M3 fits the data well, or will predict future values well, or is robust to model assumptions. For example, suppose that the true model is $y \sim \text{Exponential}(\exp(\gamma_0 + \gamma_1 x_1 + \gamma_2 x_2))$, and models M_1, M_2, and M_3 assume that $y \sim \text{Normal}(\beta_0, \sigma^2)$, $\text{Normal}(\beta_0 + \beta_1 z_1, \sigma^2)$, and $\text{Normal}(\beta_0 + \beta_1 z_2, \sigma^2)$, respectively. Then M_3 might have the highest posterior probability but that does not mean that it fit the exponentially distributed data that well.

[2] For example, take a simple random sample of size n_{train} from n; e.g., in R, if n=100 and n_{train}=50, `n.train.set = sample(1:100,size=50)`

Without going into much detail here, alternatives for selecting models include BIC and DIC.

- Bayesian Information Criterion or BIC. For model M_i,

$$BIC_i = -2\ln(L_i(\hat{\theta}_i)) + p_i \ln(n)$$

where L_i is the likelihood for model M_i, $\hat{\theta}_i$ is the maximum likelihood estimate, n is the number of observations and p_i is the number of parameters. When comparing two models in terms of likelihood alone, the one with the higher likelihood would be seen as "better". Similarly the one with the lower negative likelihood would be better. E.g., suppose there were n=20 observations and $\ln(L_1(\hat{\theta_1}))$=10 and $\ln(L_2(\hat{\theta_2}))$=6, or $-\ln(L_1(\hat{\theta_1}))$=-10 and $-\ln(L_2(\hat{\theta_2}))$=-6, thus Model 1 is "better".

Suppose, however, that Model 1 had 5 parameters and Model 2 only had 2 parameters. BIC gives some "weight" to simpler models by penalising more complex models. The BIC values for these two models:

$$BIC_1 = -2 * 10 + 5 * \ln(20) = -5.02$$
$$BIC_2 = -2 * 6 + 2 * \ln(20) = -6.01$$

Thus Model 2 is slightly better according to BIC.

Schwarz[3] showed that if n is "relatively large", BIC values can be used as approximations to Bayes Factors.

$$\exp\left(-\frac{1}{2}(BIC_1 - BIC_2)\right) \approx BF_{1,2} \tag{9.15}$$

- The Deviance Information Criterion or DIC, is similar in concept (but not in calculation) to BIC. It too is a measure that can be *approximately* viewed as a combination of the "model misfit" to the data and a "penalty" for model complexity:

$$DIC(M) \approx misfit + model\ \ complexity$$

and lower DIC values are better. Thus when comparing 2 models, say M_1 and M_2, M_1 might fit the data better but have 30 parameters, while M_2 might not fit the data so well but only have 3 parameters. Suppose DIC(M_1)=10 and DIC(M_2)=2, then M_2 is "better" in a DIC sense.

Model Averaging: An alternative to selecting a single model as "best" for making inferences is to use multiple models for making inferences and the weighting the inferences based on some criterion, e.g., posterior model probability. For example suppose there are 3 models for the systolic blood pressure, SBP, of females as a function of age with posterior probabilities 0.15, 0.05, 0.80. Let $E[SBP|age = 20, M_i]$ denote the posterior mean for SBP for an 20 year old female, and suppose they are 115, 118, and 117. A model averaged estimate:

$$E[SBP|age = 20] = 0.15 * 115 + 0.05 * 118 + 0.80 * 117 = 116.75$$

[3]Schwarz, Gideon E. (1978), "Estimating the dimension of a model", Annals of Statistics, 6 (2): 461–464.

10 More MCMC Diagnostics & Gibbs Sampling

10.1 MCMC diagnostics: How Large a Sample?

Once the burn-in length, B, is determined, there is still the question of how much longer the chain should be run. Denote the number of iterations beyond B by N. Thus the total chain length is $B + N$.

Ultimately, the size of N depends upon the desired precision of the estimate, the "Monte Carlo error" of the estimate.

We'll consider the following estimator:

$$\hat{E}(\theta) = \frac{1}{N} \sum_{j=B+1}^{N+B} \theta^j$$

To reduce notation, let $n = N$ and let $i = j - B$, so the indexing i starts at 1:

$$\hat{E}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \theta^i \tag{10.1}$$

10.1.1 Variance of $\hat{\mathrm{E}}(\theta)$

Again, remember that we are measuring the Monte Carlo error of the estimate $\hat{\mathrm{E}}(\theta)$. This is **NOT** the variance of θ, nor the variance of $\mathrm{E}[\theta]$, a constant, which is then 0.

In the following we are assuming we have a stationary Markov chain, in particular that correlation between realizations at "times" i and j in the chain, corr$(\theta_{\mathrm{i}}, \theta_{\mathrm{j}})$ depends only on the absolute difference between i and j, $|\mathrm{i} - \mathrm{j}|$.

Notation:

- γ_{k}=Cov$(\theta^{\mathrm{i}}, \theta^{\mathrm{i}+\mathrm{k}})$, is the autocovariance of lag k ($\mathrm{k} \geqslant 0$) of the chain
- σ^2=γ_0, the variance of θ^{t}
- ρ_{k} = $\gamma_{\mathrm{k}}/\sigma^2$, the autocorrelation of lag k
- $\tau^2_{\mathrm{n}}/\mathrm{n}$, the variance of $\hat{\mathrm{E}}(\theta)$ (details below)

The variance of $\hat{E}(\theta)$:

$$\begin{aligned}
Var\left[\hat{E}(\theta)\right] &= Var\left[\frac{1}{n}\sum_{i=1}^{n}\theta^i\right] \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{n}\mathrm{Var}(\theta^i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathrm{Cov}(\theta^i,\theta^j)\right] \\
&= \frac{1}{n^2}\left[n\sigma^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\gamma_{j-i}\right] \\
&= \frac{1}{n^2}\left[n\sigma^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sigma^2\rho_{j-i}\right] \\
&= \frac{1}{n^2}\left[n\sigma^2 + 2\sigma^2\sum_{k=1}^{n}(n-k)\rho_k\right] = \frac{\sigma^2}{n}\left[1 + 2\sum_{k=1}^{n}\frac{n-k}{n}\rho_k\right] = \frac{\tau_n^2}{n}
\end{aligned}$$

where

$$\tau_n^2 = \sigma^2\left[1 + 2\sum_{k=1}^{n}\frac{n-k}{n}\rho_k\right]$$

The following is *not* a rigorous proof. We are assuming that ρ_k is decreasing as k increases, in particular $\sum_{k=1}^{\infty}\rho_k$ is finite. As n gets "large"

$$\tau_n^2 = \sigma^2\left[1 + 2\sum_{k=1}^{n}\frac{n-k}{n}\rho_k\right] \approx \sigma^2\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right) \tag{10.2}$$

Thus

$$V(\hat{E}(\theta)) \approx \frac{\sigma^2\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right)}{n} \tag{10.3}$$

And as usual, the variance, a measure of the Monte Carlo error, of $\hat{E}(\theta)$ can be reduced by increasing n, but that variance is a function of the autocorrelation, too.

10.1.2 Effective Sample Size, ESS

"The term $1 + 2\sum_{k=1}^{\infty} \rho_k$ is called the *inefficiency factor* or *integrated autocorrelation* because it measures how far the θ^i's are from being a random sample and how much $V(\hat{E}(\theta))$ increases because of that"[1]. From this the *effective sample size*, n_{eff}, can be calculated:

$$n_{eff} = \frac{n}{1 + 2\sum_{k=1}^{\infty} \rho_k} \tag{10.4}$$

Then

$$V(\hat{E}(\theta)) \approx \frac{\sigma^2}{n_{eff}} \tag{10.5}$$

A 2018 paper by Elvira, et al[2] is a critical examination of the notion of ESS for importance sampling, thus independent samples, and I just want to add a cautionary note that procedures for estimating ESS may change in the future.

[1] *Markov Chain Monte Carlo, 2nd edition*, 2006, Gamerman and Lopes, p26.

[2] "Rethinking the Effective Sample Size" which can be found at https://arxiv.org/abs/1809.04129.

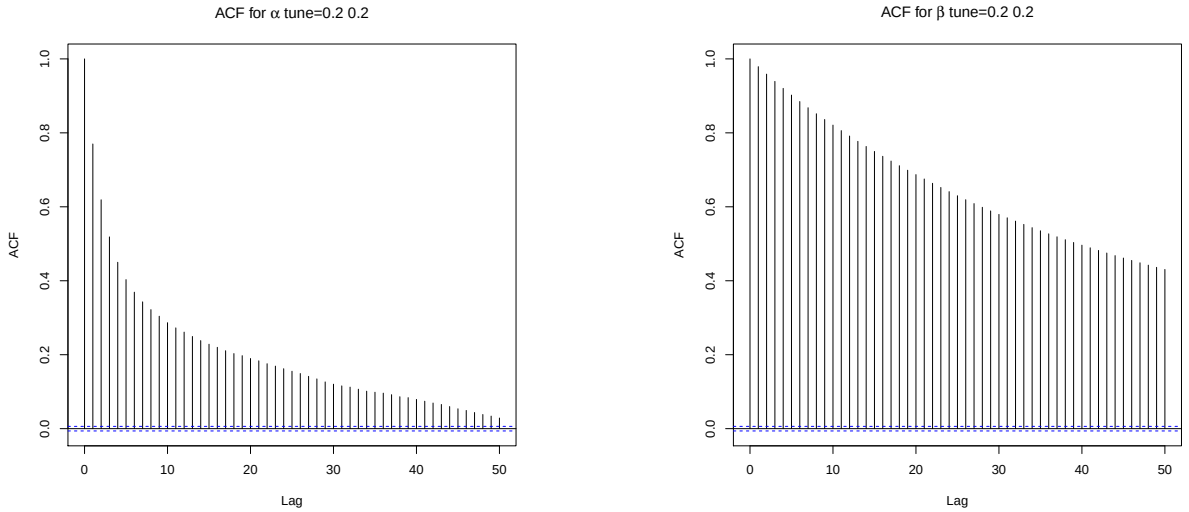
10.1.3 Autocorrelation plots

Thus if the autocorrelation is positive and large, then the variance of $V(\hat{E}(\theta))$ is large relative to the variance of $\hat{E}(\theta)$ based on *independent* samples of θ^i, i=1,...,n. A useful diagnostic for assessing the potential severity of autocorrelation is an autocorrelation plot, a plot of $\hat{\rho}_k$ against lag k where[3]

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\sigma^2}} = \frac{\sum_{i=1}^{n-k}(\theta^i - \bar{\theta})(\theta^{i+k} - \bar{\theta})}{\sum_{i=1}^{n}(\theta^i - \bar{\theta})^2} \tag{10.6}$$

where $\bar{\theta}$ is the sample average.

Figure 10.1 shows the autocorrelation plots for the Weibull parameters α and β when they are jointly updated using a Normal random walk, $\theta^c \sim$ Normal(θ^{t-1}, σ^2_θ), where the "tuning parameters" σ^2_θ were set equal to 0.2^2 for both parameters. The acceptance rate was 0.44 with these "tuning parameters". Reducing the acceptance rate by increasing the size of σ^2_α and σ^2_β has some effect on the degree of autocorrelation. For example increasing σ^2_α to 0.5^2 and σ^2_β to 1.2^2 lowered the acceptance rate to 0.12 but reduced the autocorrelation in both α and β (see Figure 10.2).

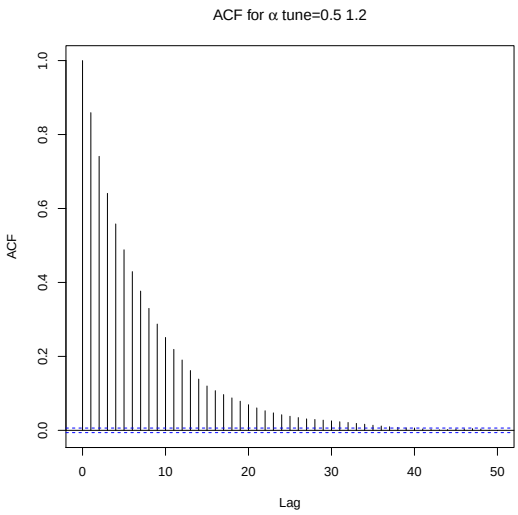


(a) Autocorrelation for α.

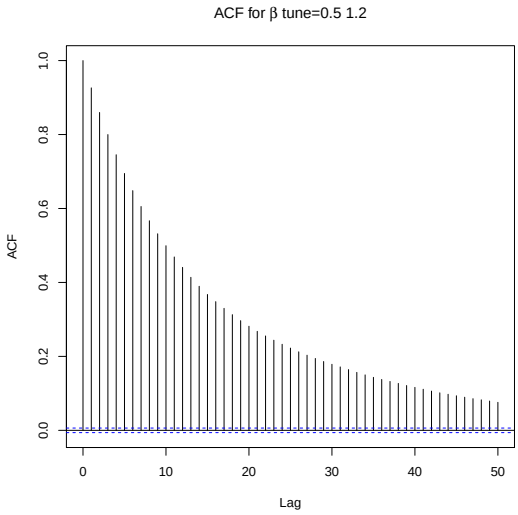
(b) Autocorrelation for β.

Figure 10.1: Autocorrelations for α and β with the Weibull model for hurricane event times. Normal random walk proposals used for α and β where σ^2_α=0.2^2 and σ^2_β=0.2^2.

[3]From Time Series Analysis and Its Applications, 4th Ed, Shumway and Stoffer, 2017, $\hat{\rho}_k$ = $\gamma(k)/\gamma(0)$, where $\gamma(k)$ = $\frac{1}{n}\sum_{i=1}^{n-k}(\theta^i - \bar{\theta})(\theta^{i+k} - \bar{\theta})$. Thus the divisor n cancels in numerator and denominator to yield eq'n 10.6.



(a) Autocorrelation for α.



(b) Autocorrelation for β.

Figure 10.2: Autocorrelation for α and β with the Weibull model for hurricane event times. Normal random walk proposals used for α and β where σ^2_α=0.4^2 and σ^2_β=1.2^2.

10.1.4 Central limit theorem for $\hat{E}(\theta)$

Under certain conditions on the Markov chain (that it be uniformly ergodic) and if $E[\theta^2]$ exists, then

$$\hat{E}(\theta) \sim \text{Asymptotically Normal}\left(E[\theta], \frac{\sigma^2}{n_{eff}}\right)$$

where by Asymptotically Normal we mean:

$$\frac{\hat{E}(\theta) - E(\theta)}{\sigma/\sqrt{n_{eff}}} \xrightarrow{\text{distribution}} \text{Normal}(0, 1)$$

This then allows us to calculate approximate interval estimates for $E[\theta]$ using the normal distribution. However, the main difficulty is estimating the variance, σ^2/n_{eff}.

10.1.5 Estimating $V(\hat{E}(\theta))$

Due to the autocorrelation, estimation of $V[\hat{E}(\theta)]$ is not a trivial matter. A popular method for calculating the variance is *batching*.

The algorithm for batching:

- Partition the chain of length n into m batches of T successive values.
- For each batch calculate the average value of θ within the batch. In batch i

$$\bar{\theta}_i = \frac{1}{T}\sum_{j=1}^{T} \theta_{i,j}, i = 1, \ldots, m$$

- The estimate of $\frac{\tau^2}{n}$ is then:

$$\frac{\widehat{\tau^2}}{n} = \frac{\frac{T}{m-1}\sum_{i=1}^{m}(\bar{\theta}_i - \bar{\theta})^2}{n}$$

"Values of T should be between 10 and 30" (Gamerman and Lopes, 2006, p 134).

The R package coda includes a function called batchSE that will calculate the standard errors with this method. If you run the code in Appendix A.2 of Lecture Notes 9, with the tuning parameters of σ_α=0.5 and σ_β=1.2:

```
library(coda)
#Burnin=1000, N=100,000
a.b.star <- mcmc(data=cbind(alpha.star,beta.star),start=Burnin,end=N)
MonteCarlo.se <- batchSE(x=a.b.star,batchSize=round(N/30))
print(MonteCarlo.se)
#  alpha.star   beta.star
# 0.001151362 0.012550150
```

The package coda will also calculate effective sample size, n_{eff}, with the function. For example,

```
effectiveSize(a.b.star)
# alpha.star  beta.star
# 6835.639   3130.651
```

The effective sample sizes are relatively low given N=100,000, roughly 6% and 3% for α and β due to the relatively high autocorrelation.

Note: Sometimes individuals will thin the chain, say keep every 100th value, to reduce autocorrelation. This does *not* reduce Monte Carlo error. However, it can be useful for reducing file sizes.

10.1.6 Improving performance

Monte Carlo error is exacerbated by slow mixing, relatively low acceptance rates, high autocorrelation, and generally limited or poor coverage of the state space. Potential remedies include:

1. Changing the proposal distribution. With random walk proposals one can change the variance of the proposal distribution. For example, the random walk sampler for the Weibull problem was:

$$\alpha^c \sim \text{Normal}(\alpha^t, \sigma^2_\alpha)$$
$$\beta^c \sim \text{Normal}(\beta^t, \sigma^2_\beta)$$

   Changing the values of σ^2_α and σ^2_β affected the autocorrelation (and the acceptance rates). Such manipulation of proposals is an example of *tuning the proposals*. This is a completely legitimate exercise, it is not "cheating" in any way as the prior distributions do not change. It's just a technique to increase effective sample size.

2. Reparameterising the model. For example, with Bayesian regression models, centering the covariates can reduce autocorrelation. For example, compare the following two parameterisations of a two covariate multiple regression:

$$\text{Formulation 1: } y \sim \text{Normal}\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2, \sigma^2\right)$$
$$\text{Formulation 2: } y \sim \text{Normal}\left(\beta_0 + \beta_1 (x_1 - \bar{x}_1) + \beta_2 (x_2 - \bar{x}_2), \sigma^2\right)$$

   Formulation 2 will generally lead to lower autocorrelation in the simulated values of β_0, β_1, and β_2.

3. Blocking: when parameters have some degree of positive correlation, it is often better to jointly update such parameters. For example if there are 4 parameters where θ_1 and θ_2 are highly correlated, then it may be better to generate (θ^c_1, θ^c_2) simultaneously to increase the acceptance rate, improve mixing, increase effective sample size.

10.2 Gibbs Sampler

The Gibbs Sampling is another very popular MCMC method for generating samples from multivariate distributions. While it can be shown to be a special case of a Metropolis-Hastings sampler, it has two unique features:

1. The proposal distributions are the conditional distributions for the θ's.
2. All candidate values are accepted.

10.2.1 The algorithm

The algorithm as described herein will closely parallel the single update Metropolis-Hastings procedure described previously.

Let $\Theta = (\theta_1, \theta_2, \ldots, \theta_q)$ with distribution $p(\Theta)$.

Denote the full conditional distribution for θ_i by

$$p(\theta_i|\theta_{(-i)}) = p(\theta_i|\theta_1, \theta_2, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_q), \ i = 1, 2, \ldots, q$$

Initialise the chain with $\Theta^0 = (\theta_1^0, \theta_2^0, \ldots, \theta_q^0)$.

At iteration $t + 1$ given the values Θ^t, generate Θ^{t+1} as follows:

- Generate θ_1^{t+1} from $p(\theta_1|\theta_2^t, \theta_3^t, \ldots, \theta_q^t)$
- Generate θ_2^{t+1} from $p(\theta_2|\theta_1^{t+1}, \theta_3^t, \ldots, \theta_q^t)$
- ...
- Generate θ_i^{t+1} from $p(\theta_i|\theta_1^{t+1}, \theta_2^{t+1}, \ldots, \theta_{i-1}^{t+1}, \theta_{i+1}^t, \ldots, \theta_q^t)$
- ...
- Generate θ_q^{t+1} from $p(\theta_q|\theta_1^{t+1}, \theta_2^{t+1}, \ldots, \theta_{q-1}^{t+1})$

The (joint) transition probability of going from Θ^t to Θ^{t+1} is then

$$\mathcal{K}_{Gibbs}(\Theta^t, \Theta^{t+1}) = \prod_{i=1}^{q} p(\theta_i^{t+1}|\theta_j^{t+1}, j < i, \text{ and } \theta_j^t, j > i)$$

and the stationary distribution is the target distribution $p(\Theta)$.

10.2.2 Example 1: Bivariate Normal with known correlation

The Gibbs Sampler is demonstrated for a two-dimensional parameter, $\Theta=(\theta_1, \theta_2)$, where the target distribution is a "standard" bivariate normal with means equal to 0, standard deviations equal to 1, and known correlation ρ:

$$\theta_1, \theta_2 \sim \text{BVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$$

The joint pdf is:

$$p(\theta_1, \theta_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{\theta_1^2 - 2\rho\theta_1\theta_2 + \theta_2^2}{2(1-\rho^2)}\right)$$

To find the conditional distribution for θ_1 given θ_2, focus on just those terms in the joint distribution involving θ_1:

$$p(\theta_1|\theta_2) \propto \exp\left(-\frac{\theta_1^2 - 2\rho\theta_1\theta_2}{2(1-\rho^2)}\right) \propto \exp\left(-\frac{(\theta_1 - \rho\theta_2)^2}{2(1-\rho^2)}\right)$$

which can be seen to be the kernel of a Normal($\rho\theta_2$, $1-\rho^2$) density function. Similarly, $\theta_2|\theta_1 \sim$ Normal$(\rho\theta_1, 1-\rho^2)$.

The Gibbs Sampler: Specify initial values (θ_1^0, θ_2^0). At iteration $t+1$,

1. Generate $\theta_1^{t+1}|\theta_2^t$ from Normal($\rho\theta_2^t$, $(1-\rho^2)$).
2. Then generate $\theta_2^{t+1}|\theta_1^{t+1}$ from Normal($\rho\theta_1^{t+1}$, $(1-\rho^2)$).

Example R code (additional code in Appendix 10.3.1):

```
rho <- 0.2
N <- 10000
theta1.star <- theta2.star <- numeric(N)
theta1.star[1] <- 1.3
theta2.star[1] <- -0.2
for(i in 2:N) {
  theta1.star[i] <- rnorm(n=1,mean=rho*theta2.star[i-1],sd=sqrt(1-rho^2))
  theta2.star[i] <- rnorm(n=1,mean=rho*theta1.star[i],sd=sqrt(1-rho^2))
}
```

The Monte Carlo errors and effective sample sizes, are shown below (Note: N=10,000).

	θ_1	θ_2
MC error	0.01204288	0.00953518
n_{eff}	8647	8041

The trace plots and density plots for the generated sample are shown in Figure 10.3. The joint distribution of θ_1 and θ_2 and the autocorrelation for θ_1 are plotted in Figure 10.4. The trace plots, low autocorrelation, Monte Carlo errors, and large effective sample sizes (over 80% of N) all indicate good mixing.

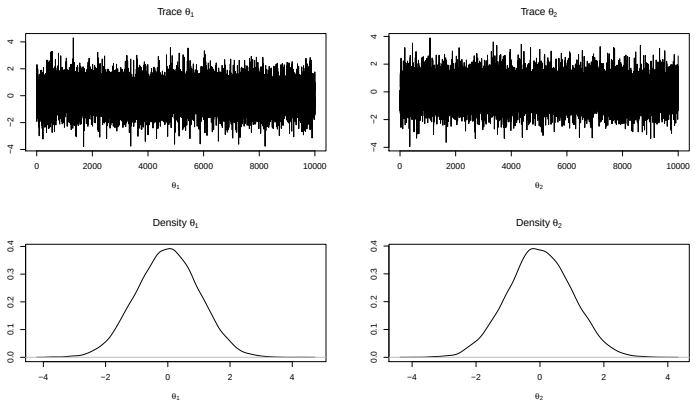
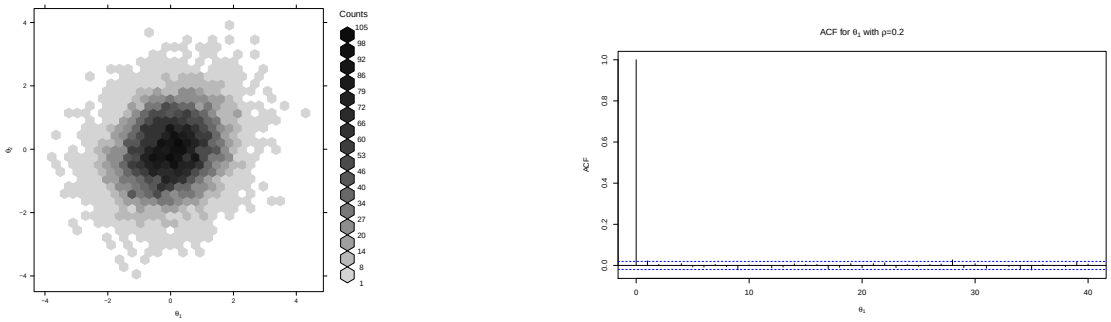


Figure 10.3: Trace and density plots for Gibbs Sampler values of θ_1 and θ_2 from a standard bivariate normal with ρ=0.2



(a) Joint distribution.

(b) ACF.

Figure 10.4: Joint distribution for Gibbs Sampler values of θ_1 and θ_2 from a standard bivariate normal with ρ=0.2.

By way of contrast the above exercise was repeated with a strong negative correlation, ρ= -0.95. The trace and density plots are shown in Figure 10.5 and a pattern in the trace plots can be seen. The joint density and the autocorrelation for θ_1 are shown in Figure 10.6 and make clear the effects of the high correlation. Monte Carlo errors are considerably larger and the effective sample sizes considerably lower:

	θ_1	θ_2
MC error	0.0526	0.0521
n_{eff}	447	467

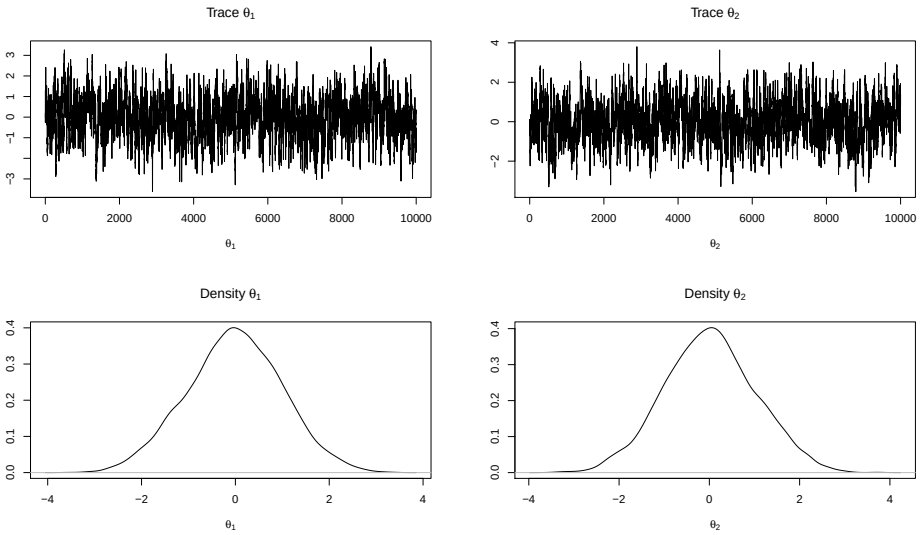


Figure 10.5: Trace and density plots for Gibbs Sampler values of θ_1 and θ_2 from a standard bivariate normal with ρ= -0.95.

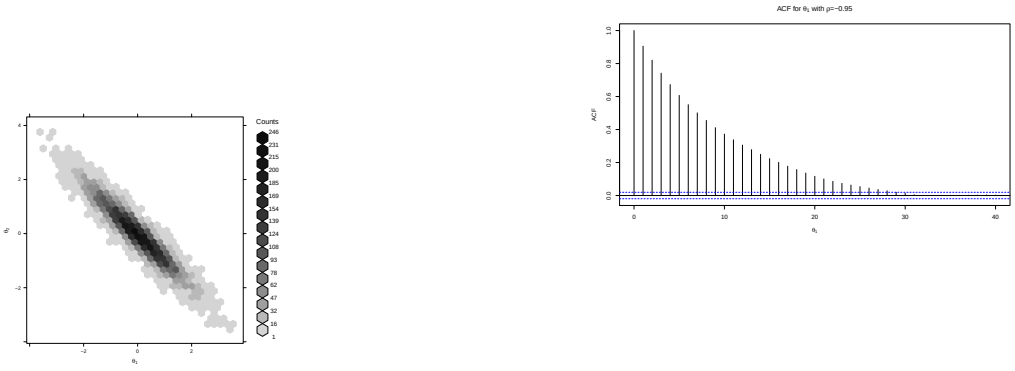


Figure 10.6: Joint distribution for Gibbs Sampler values of θ_1 and θ_2 from a standard bivariate normal with ρ= -0.95 and autocorrelation for θ_1.

10.2.3 A special case of Metropolis-Hastings

Consider the Metropolis-Hastings algorithm with single updates using the conditional distribution as the proposal. As a concrete example let q=3, thus $\Theta = (\theta_1, \theta_2, \theta_3)$, and let the updating be in the order 1, 2, and then 3. At the beginning of iteration $t+1$ we have $\Theta^t = (\theta_1^t, \theta_2^t, \theta_3^t)$. The "candidate" value for θ_1 is generated from $p(\theta_1|\theta_2^t, \theta_3^t)$, and the Metropolis Hastings ratio is then:

$$MHR(\theta_1^t, \theta_1^c) = \frac{p(\theta_1^c, \theta_2^t, \theta_3^t)}{p(\theta_1^t, \theta_2^t, \theta_3^t)} \frac{q(\theta_1^t|\theta_1^c, \theta_2^t, \theta_3^t)}{q(\theta_1^c|\theta_1^t, \theta_2^t, \theta_3^t)} \tag{10.7}$$

$$= \frac{p(\theta_1^c|\theta_2^t, \theta_3^t)p(\theta_2^t, \theta_3^t)}{p(\theta_1^t|\theta_2^t, \theta_3^t)p(\theta_2^t, \theta_3^t)} \frac{p(\theta_1^t|\theta_2^t, \theta_3^t)}{p(\theta_1^c|\theta_2^t, \theta_3^t)} \tag{10.8}$$

$$= \frac{p(\theta_2^t, \theta_3^t)}{p(\theta_2^t, \theta_3^t)} = 1 \tag{10.9}$$

Thus the candidate value is always accepted.

10.2.4 Gibbs Sampler vs (General) Metropolis-Hastings

- Gibbs Sampler Strengths
  - The proposal distribution is automatically defined (thus no "tuning" of a proposal)
  - Always "accept" the candidate value
  - Because the proposal is conditional on the other parameters, can view it as an "adaptive" algorithm as changes in one subset of parameters likely lead to changes in another subset of parameters
- Gibbs Sampler Weaknesses
  - The conditional distributions may not be tractable
  - Relatedly, sampling from the conditionals may be computationally intensive
  - 100% acceptance rate does not necessarily mean good mixing
- Metropolis-Hastings Strengths
  - Proposal distribution quite flexible, can be fast to sample from and fast to evaluate MHR (particularly with symmetric proposals)
  - Don't need to know the conditional distributions
  - Block updating can be relatively easy; e.g., multivariate t distribution as a proposal.[4].
- Metropolis-Hastings Weaknesses
  - May be hard to find a good proposal, one with good mixing
  - Can take time to "tune" a proposal, e.g., a good variance value for a normal random walk proposal

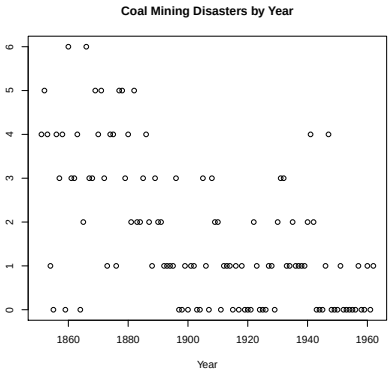
[4]Block updating can be done with the Gibbs sampler, too, so long as the conditional distribution for the "block" is tractable. For example, generate (θ_1, θ_2) simultaneously with $p(\theta_1, \theta_2|\theta_3, \ldots, \theta_q)$.

10.2.5 Combining Gibbs Sampling and Metropolis-Hastings

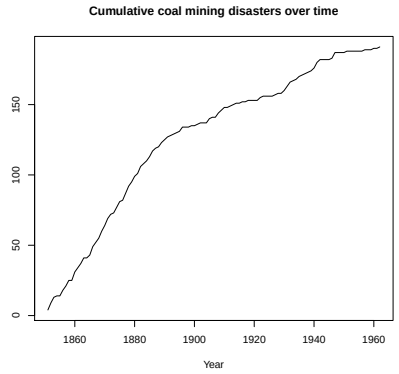
Given $\Theta = (\theta_1, \theta_2, \ldots, \theta_q)$, if full conditionals for all θ_i, i=1, ..., q are not tractable, then Metropolis-Hastings can be used in those cases. For example, suppose $\Theta = (\theta_1, \theta_2, \theta_3)$ and $p(\theta_1|\theta_2, \theta_3)$ and $p(\theta_2|\theta_1, \theta)$ are tractable, but $p(\theta_3|\theta_1, \theta_2)$ is not. Then the Gibbs Sampler can be used to generate θ_1 and θ_2 and Metropolis-Hastings, e.g. a random walk proposal, can be used to generate θ_3. Of course, θ_1 and θ_2 will have 100% acceptance rates, but θ_3 will not.

10.2.6 Example 2: Coal mining disasters with change point

This example[5] models the occurrence of coal mining disasters (with 10 or more fatalities) in Great Britain from the years 1851 through 1962 (n=112)[6]. Figure 10.7 shows the occurrences by year and the cumulative number of disasters over time.



(a) Number of occurrences vs year.



(b) Cumulative disasters over time.

Figure 10.7: Occurrences of coal mining disasters in Great Britain.

A Poisson distribution is a commonly used distribution for such count data. However as the plots in Figure 10.7 indicate there appears to be a reduction in the number of disasters some years before 1900, thus a single Poisson distribution does not seem adequate. A *change point* model is one means of accounting for a potential change in the incidence rates with one Poisson distribution before the change and a different Poisson distribution afterwards. Let y_i denote the number of disasters in year i and let τ indicate the last year that the first Poisson occurred, where τ could equal 1, ..., $n-1$. Assume that the distribution of y_i, i=1,..., τ is Poisson(μ_b) and y_i, i=$\tau+1$, ..., n is Poisson(μ_a), where b and a denote before and after.

Specify Gamma distribution priors for the Poisson parameters: $\mu_b \sim$ Gamma(α, β) and $\mu_a \sim$ Gamma(γ, δ). Assume that any of the change point years are equally likely, thus τ is uniformly distributed over 1,2,..., $n-1$, or $Pr(\tau = i)$ = 1/$n-1$, where n=112.

[5] Taken from Markov Chain Monte Carlo (2006) by Gamerman and Lopes, pp 143–146).

[6] These data can be accessed at https://conservancy.umn.edu/handle/11299/200478.

The posterior distribution for μ_b, μ_a, and τ, letting $\mathbf{y}$= $(y_1, \ldots, y_n)$:

$$\begin{aligned}
p(\mu_b, \mu_a, \tau|\mathbf{y}) &\propto \pi(\mu_b)\pi(\mu_a)\pi(\tau)f(\mathbf{y}|\mu_b, \mu_a, \tau) \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)}\mu_b^{\alpha-1}e^{-\mu_b\beta}\frac{\delta^\gamma}{\Gamma(\gamma)}\mu_a^{\gamma-1}e^{-\mu_a\delta}\frac{1}{n-1}\prod_{i=1}^{\tau}\frac{e^{-\mu_b}\mu_b{}^{y_i}}{y_i!} \times \prod_{i=\tau+1}^{n}\frac{e^{-\mu_a}\mu_a{}^{y_i}}{y_i!} \\
&\propto \mu_b^{\alpha-1}e^{-\mu_b\beta}\mu_a^{\gamma-1}e^{-\mu_a\delta}\prod_{i=1}^{\tau}e^{-\mu_b}\mu_b{}^{y_i} \times \prod_{i=\tau+1}^{n}e^{-\mu_a}\mu_a{}^{y_i} \\
&= \mu_b^{\alpha+s_\tau-1}e^{-(\beta+\tau)\mu_b}\mu_a^{\gamma+s_n-s_\tau-1}e^{-(\delta+n-\tau)\mu_a}
\end{aligned}$$

where $s_\tau = \sum_{i=1}^{\tau} y_i$ and $s_n = \sum_{i=1}^{n} y_i$.

The "trick" to finding the conditional for a parameter θ is to just consider terms including θ. Doing that for μ_b, μ_a, and τ:

$$p(\mu_b|\mu_a, \tau, \mathbf{y}) \propto \mu_b^{\alpha+s_\tau-1}e^{-(\beta+\tau)\mu_b} \tag{10.10}$$

$$p(\mu_a|\mu_b, \tau, \mathbf{y}) \propto \mu_a^{\gamma+s_n-s_\tau-1}e^{-(\delta+n-\tau)\mu_a} \tag{10.11}$$

$$p(\tau|\mu_a, \mu_b, \mathbf{y}) \propto \mu_b^{\alpha+s_\tau-1}e^{-(\beta+\tau)\mu_b}\mu_a^{\gamma+s_n-s_\tau-1}e^{-(\delta+n-\tau)\mu_a} \tag{10.12}$$

For the posterior conditional of μ_b, the righthand side of (10.10) is the kernel of a Gamma($\alpha+s_\tau$, $\beta+\tau$), while the righthand side of (10.11) is the kernel of a Gamma($\gamma+s_n-s_\tau$, $\delta+n-\tau$). For τ the conditional distribution is not a standard form but because there are a finite number of values for τ (1, 2, ..., $n-1$) and the sum of the conditionals over all possible values must sum to 1, the posterior for τ is then a discrete probability distribution:

$$p(\tau|\mu_a, \mu_b, \mathbf{y}) = \frac{\mu_b^{\alpha+s_\tau-1}e^{-(\beta+\tau)\mu_b}\mu_a^{\gamma+s_n-s_\tau-1}e^{-(\delta+n-\tau)\mu_a}}{\sum_{\tau=1}^{n-1}\mu_b^{\alpha+s_\tau-1}e^{-(\beta+\tau)\mu_b}\mu_a^{\gamma+s_n-s_\tau-1}e^{-(\delta+n-\tau)\mu_a}} \tag{10.13}$$

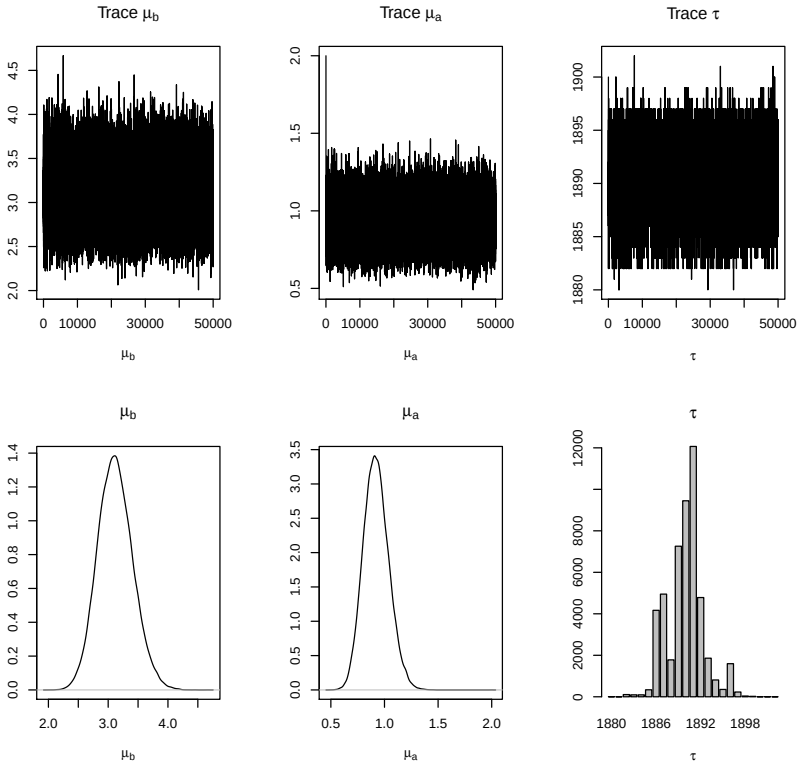
Setting the hyperparameters of the priors for μ_b, α and β, and for μ_a, γ and δ, all equal 0.001, the Gibbs sampler was implemented with the following R code.

```
N <- 50000
y <- coal.data$Disasters
s.n <- sum(y)
n <- length(y)
cum.y <- cumsum(y[-n])
tau.vec <- 1:(n-1)

alpha.b <- beta.b <- gamma.a <- delta.a <- 0.001

mub.star <- mua.star <- tau.star <- numeric(N)
mub.star[1] <- 3; mua.star[1] <- 2; tau.star[1] <- 50

for(i in 2:N) {
  s.tau <- sum(y[1:tau.star[i-1]])
  mub.star[i] <- rgamma(n=1,alpha.b+s.tau,beta.b+tau.star[i-1])

  mua.star[i] <- rgamma(n=1,gamma.a+s.n-s.tau,delta.a+n-tau.star[i-1])

  temp <-        (alpha.b+cum.y-1)*log(mub.star[i]) -(beta.b+tau.vec)*mub.star[i] +
             (gamma.a+s.n-cum.y-1)*log(mua.star[i]) -(delta.a+n-tau.vec)*mua.star[i]
  prob.vector <- exp(temp)
  prob.vector <- prob.vector/sum(prob.vector)
  tau.star[i] <- sample(x=1:(n-1),size=1,prob=prob.vector)
}
```

The trace plots and estimated posterior distributions are shown in Figure 10.8. Summary statistics are shown below. The year 1890 is the expected year of a change.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
μ_b	2.01	2.92	3.11	3.12	3.31	4.45
μ_a	0.49	0.84	0.92	0.92	1.00	1.46
τ	1880.00	1889.00	1890.00	1889.97	1891.00	1901.00

Figure 10.8: Trace plots and estimated posterior distributions for parameters of change point Poisson model for occurrences of coal mining disasters in Great Britain.



10.3 R Code

10.3.1 Gibbs Sampler for standard BVN with known ρ

```
#--- Gibbs Sampler ------------------------
# BVN with standard normal and known rho
set.seed(92)
rho <- 0.2
# rho <- -0.95   # the high negative correlation example
N <- 10000
theta1.star <- theta2.star <- numeric(N)
theta1.star[1] <- 1.3
theta2.star[1] <- -0.2
for(i in 2:N) {
  theta1.star[i] <- rnorm(n=1,mean=rho*theta2.star[i-1],sd=sqrt((1-rho^2)))
  theta2.star[i] <- rnorm(n=1,mean=rho*theta1.star[i],sd=sqrt((1-rho^2)))
}

# Trace and Density plots
par(mfrow=c(2,2),oma=c(0,0,3,0))
plot(1:N,theta1.star,xlab=expression(theta[1]),ylab="",main=expression(paste("Trace ",theta[1])),
     type="l")
plot(1:N,theta2.star,xlab=expression(theta[2]),ylab="",main=expression(paste("Trace ",theta[2])),
     type="l")
plot(density(theta1.star),xlab=expression(theta[1]),ylab="",
     main=expression(paste("Density ",theta[1])),type="l")
plot(density(theta2.star),xlab=expression(theta[2]),ylab="",
     main=expression(paste("Density ",theta[2])),type="l")
par(mfrow=c(1,1))

#-- Joint density
library(hexbin)
hexbinplot(theta2 ~ theta1,data=data.frame(theta1=theta1.star,theta2=theta2.star),
           xlab=expression(theta[1]),ylab=expression(theta[2]))

#-- examine autocorrelation
plot(acf(theta1.star),xlab=expression(theta[1]),
     main=bquote(paste("ACF for ",theta[1]," with ",rho,"=",.(rho))))

#-- examine Monte Carlo standard error and effective sample size
library(coda)
Burnin <- 1000
t1.t2.star <- mcmc(data=cbind(theta1.star,theta2.star),start=Burnin,end=N)
MonteCarlo.se <- batchSE(x=t1.t2.star,batchSize=round(N/30))
print(MonteCarlo.se)
#  theta1.star theta2.star
#   0.01204288  0.00953518

round(effectiveSize(t1.t2.star))
# theta1.star theta2.star
# 8647        8041
```

10.3.2 Gibbs sampler for coal mine disasters

```
set.seed(73)
N <- 50000
y <- coal.data$Disasters
s.n <- sum(y); n <- length(y)
cum.y <- cumsum(y[-n])
tau.vec <- 1:(n-1)

alpha.b <- beta.b <- gamma.a <- delta.a <- 0.001

mub.star <- mua.star <- tau.star <- numeric(N)
mub.star[1] <- 3; mua.star[1] <- 2; tau.star[1] <- 50

for(i in 2:N) {
  s.tau <- sum(y[1:tau.star[i-1]])
  mub.star[i] <- rgamma(n=1,alpha.b+s.tau,beta.b+tau.star[i-1])

  mua.star[i] <- rgamma(n=1,gamma.a+s.n-s.tau,delta.a+n-tau.star[i-1])

  temp <-        (alpha.b+cum.y-1)*log(mub.star[i]) -(beta.b+tau.vec)*mub.star[i] +
             (gamma.a+s.n-cum.y-1)*log(mua.star[i]) -(delta.a+n-tau.vec)*mua.star[i]
  prob.vector <- exp(temp)
  #prob.vector <- mua.star[i]^(alpha.b+cum.y-1)*exp(-(beta.b+tau.vec)*mub.star[i]) *
  #               mub.star[i]^(gamma.a+s.n-cum.y-1)*exp(-(delta.a+n-tau.vec)*mua.star[i])
  prob.vector <- prob.vector/sum(prob.vector)
  tau.star[i] <- sample(x=1:(n-1),size=1,prob=prob.vector)
}

tau.star <- tau.star + 1850
par(mfrow=c(2,3),oma=c(0,0,3,0))
plot(1:N,mub.star,xlab=expression(mu[b]),ylab="",main=expression(paste("Trace ",mu[b])),type="l")
plot(1:N,mua.star,xlab=expression(mu[a]),ylab="",main=expression(paste("Trace ",mu[a])),type="l")
plot(1:N,tau.star,xlab=expression(tau),ylab="",main=expression(paste("Trace ",tau)),type="l")
plot(density(mub.star),xlab=expression(mu[b]),ylab="",main=expression(mu[b]),type="l")
plot(density(mua.star),xlab=expression(mu[a]),ylab="",main=expression(mu[a]),type="l")
plot(density(tau.star),xlab=expression(tau),ylab="",main=expression(tau),type="l")
par(mfrow=c(1,1))

#-- summary statistics
burnin <- 10000
summary(mub.star[-c(1:burnin)])
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 2.008   2.917   3.110   3.118   3.307   4.449

summary(mua.star[-c(1:burnin)])
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 0.4906  0.8411  0.9177  0.9224  0.9984  1.4646

summary(tau.star[-c(1:burnin)])
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 1880    1889    1890    1890    1891    1901

theta.star <- mcmc(data=cbind(mub.star,mua.star,tau.star),start=Burnin,end=N)
```

```
MonteCarlo.se <- batchSE(x=theta.star,batchSize=round(N/30))
print(MonteCarlo.se)
#   mub.star     mua.star     tau.star
# 0.0017489203 0.0006727873 0.0126560505

round(effectiveSize(theta.star))
# mub.star mua.star tau.star
# 41322    42955    38232
```