MT4531: Bayesian Inference

MT5731: Advanced Bayesian Inference

Course Notes - part 1

Semester 1, 2021

Lecturer and Module co-ordinator: Dr Nicolò Margaritella

Notes compiled on: October 9, 2021

Rev. Thomas Bayes
(Probably)

University of St. Andrews

School of Mathematics and Statistics

CONTENTS

INTRODUCTION

0.1 Bayesian approach in statistics

The Bayesian approach is a complete theory in statistics, different in many ways from the standard (classical/frequentist) approach, that you have encountered in previous modules. We can use Bayesian statistics to design experiments, construct statistical models, estimate parameters, make predictions, test hypotheses, compare models, and in general measure the uncertainty that is associated with our inferences. In some situations, the actual outcomes, e.g. parameter estimates, of the Bayesian and the standard approach are not very different. However, the methodology that is used to derive such outcomes and especially their interpretation under the two approaches are typically very different. This is the reason why there is a long-lasting, and often passionate, philosophical debate about which is the most appropriate statistical approach.

The basic idea of Bayesian statistics is reasonably straightforward.

Example: Suppose that we wish to infer a quantity of interest, θ. For example, lets say that I wish to estimate the proportion of blemished apples that the apple tree in my garden produces in October. I first gather the existing information: the proportion of blemished apples produced in previous years, as far as I remember, and how certain I am about this estimate. I may also want to check any information available on the internet or in my library on blemished apples. I will use this information to construct a probability distribution, $p(\theta)$, for θ, which will express the uncertainty that I have for θ based on this information and before making any new observations. Then, from the 1st of October until Halloween, I will be gathering as data the number of blemished apples and the total number of apples produced. Once these data are gathered, probability theory (Bayes rule or Bayes theorem) allows us to update $p(\theta)$ based on the new data, to obtain the probability distribution $\pi(\theta|data)$.

We call $\pi(\theta|data)$ the *posterior* distribution of θ as opposed to the *prior* distribution, $p(\theta)$. These are the probability distributions before (*a priori*) and after (*a posteriori*) observing the data. Bayes theorem allows us to make the transition from prior to posterior probability distributions for an unknown quantity of interest. Bayesian statistics provides a way of formalising the gathering of information and uncertainty before new observations are made available and then updating this information based on the new observations. It is a natural way of scientific investigation.

The result on which Bayesian inference rests - Bayes Theorem - is a simple result in elementary probability theory, by the Presbyterian minister Reverend Thomas Bayes (1701-61), though this work was only published posthumously in 1763, by his friend Richard Price.

0.1.1 Putting Bayesian statistics in perspective

The statistician's tools of the trade are:

(a) **Probability theory.** Probability theory is used to measure uncertainty in a coherent manner (i.e., probabilities should not contradict one another). Probability theory derives properties for probabilities of related events, as well as the theory for random variables and their distributions.

(b) **Classical statistics.**

The standard approach that you have encountered in previous statistics modules is typically called "classical" or "frequentist". It is not a unified approach, but rather a collection of different approaches. Some of the key classical statistical methods are based solely on the likelihood function of observed data (e.g. maximum likelihood estimates, p-values, likelihood ratio function). Other methods are based on satisfying certain criteria, such as the least square estimation (LSE) that minimises squared differences between observations and fitted values. Methods typically focus on deriving procedures that satisfy optimal long-run properties (e.g. unbiasedness, consistency, fixed error rates) if the experiment is repeated infinite times. A key characteristic is that they consider parameters as having a *fixed* (typically unknown) value. The common criticism for the standard approach is:

1. the interpretation of outcomes (e.g. confidence intervals or p-values) based on their long-run properties, even in cases where the experiments, surveys, etc. are practically unrepeatable.
2. the lack of formal methods for incorporating background information and that they often use background information implicitly (e.g. experimental design, choice of likelihood, choice of error rates in testing).
3. the disobedience of some classical procedures (e.g. confidence intervals, p-values) to the likelihood principle.

(c) **Bayesian statistics.**

In Bayesian statistics all uncertain quantities, including parameters, are random and therefore they have probability distributions. These probability distributions are constructed using the rules of probability theory that ensure that they are well-defined. All probability distributions describe the "degree of belief" or uncertainty of one person about a quantity of interest. In Bayesian statistics, probability distributions are always conditional to the information used to construct them. Therefore the use of background information, or indeed any form of information, and the subjective method used to construct them is explicitly open to scrutiny.

The openness in the use of background information is one of the most common arguments in support of the Bayesian approach. It is particularly useful in situations where the data are either very limited or non-existent (e.g. experimental design, rare events). However, it is also one of its common criticism. The argument against it is that it is subjective, and represents personal judgements. Non-informative and empirical priors, as well as prior sensitivity analysis (comparing inferences that result from different priors), can be used to alleviate this criticism. Bayesian methods can also be more computationally intensive, although complex models are often easier to fit within the Bayesian framework.

Why use Bayesian statistics? More and more statisticians are accepting it in preference of classical statistics because:

- It is a natural and coherent way of thinking about science and learning based solely on standard probability theory. All unknowns are random variables and they can be studied using the available information that form conditional probability distributions. Their probability distributions are updated once more information becomes available. Different procedures (e.g. LSE, MLE) are not required.

- It is becoming more acceptable to use prior background information in many areas of application. Non-informative/vague priors that express huge uncertainty can be used, also to examine the degree to which informative priors have affected the analysis. The audience to which the analyses are directed can then assess the validity of the results and decide if the inferences are justified. Sometimes, when data are not easily obtained, using prior information is the only option!

- Posterior distributions and their summaries are easier to interpret compared to, say, p-values and confidence intervals.

- It is often computationally easier and also more informative to do the analysis with the Bayesian approach (this is particularly true for complex analyses).

0.2 Scope of the Course

This course focuses on the use of Bayesian methods. There has been a dramatic increase in the application of the Bayesian approach throughout all areas of statistics within recent years, and this has had a considerable impact in the analysis of complex data. The basic underlying principles will be discussed in detail within the course, which is broadly structured into two parts:

- The first part will describe the Bayesian paradigm, stating Bayes Theorem (the underlying principle that Bayesian inference hinges on) and discuss associated practical issues. Examples will be given and compared to the traditional (classical) estimates.

- The second part of the course will look at modern approaches to Bayesian inference and the associated computational techniques in more complex (i.e., realistic) inference problems. We shall aim at understanding the basics of how these techniques work, as well as learning software that enables them to be applied relatively easily.

There are two main aims to this course. The first is to gain an understanding and knowledge of the underlying concepts of Bayesian inference. The second is to understand the techniques and tools that are available (given sufficient computing power). These techniques are not merely a means to an end, but have intrinsic value in their own right as exercises in probability and statistics. The Bayesian statistical R-package NIMBLE will be used to illustrate the ideas and provide worked examples of the computational techniques that will be described. NIMBLE uses and extends the BUGS (Bayesian Inference Using Gibbs Sampling) language to develop statistical models in R. Instructions to install NIMBLE on your computer are in Appendix B. The emphasis of the course is to illustrate the various methods and their application on relatively simple examples.

0.3 Course Texts

There is a large, and ever increasing, set of books devoted to Bayesian statistics. Some focus on theoretical aspects, some are more practical, while others try to span both areas. A relatively concise and accessible book that we'd recommend if you're going to rely on one text is:

- Lee, P. M. (2012) *Bayesian Statistics: An Introduction. 4th Edition.* Wiley.

(Earlier editions are fine too!)

A suggested text book to learn the BUGS language is,

- Lunn, et al. (2012) *The Bugs Book. A practical introduction to Bayesian Analysis.* CRC Press

Another good textbook for BUGS (optional) is:

- Ntzoufras I. (2009) *Bayesian modelling using WinBugs.* Wiley.

Some more advanced/detailed books (optional), are listed below.

- Bernardo, J.M. and Smith, A.F.M. (1994) *Bayesian Theory.* Wiley.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D.B., Vehtari, A. and Rubin, D. B. (2013) *Bayesian Data Analysis. 3rd Edition.* Chapman and Hall/CRC.
- O'Hagan, A., and Forster, J. (2004) *Bayesian Inference. Second edition.* Chapman and Hall/CRC.
- O'Hagan, A. et al. (2006) *Uncertain Judgements: Eliciting Experts' Probabilities (Statistics in Practice).* Wiley
- Kadane J. B. (2011) *Principles of Uncertainty.* Chapman and Hall/CRC

0.4 Acknowledgments

Similar modules to this are now given at most universities and we would like to acknowledge the fact that some of this material is based upon the work of other people, most notably Steve Brooks at Cambridge, and Tony O'Hagan at Nottingham and Sheffield. Many St. Andrews staff, past and present, contributed to these notes, principally Ruth King, Ian Goudie, Len Thomas and Michail Papathomas.

Chapter 1

BAYESIAN STATISTICS

1.1 Introduction

The result on which Bayesian inference rests - Bayes Theorem - is uncontroversial. It is simply a result in elementary probability theory, by the Presbyterian minister Reverend Thomas Bayes (1701-61), though this work was only published posthumously in 1763, by his friend Richard Price. However, the way that Bayesian statisticians use this theorem to make inferences about unknown quantities is distinctly different from the way that classical statisticians operate.

Let's begin with a recap of the standard approach for statistical inference before outlining the Bayesian approach. Assume we have some population parameter θ on which we wish to make inference and a probability mechanism $f(x|\theta)$ which determines the probability of observing different data, x, under different parameter values, θ. In the standard approach, θ is considered to be some fixed, but unknown, constant. Inference is then based on the likelihood $f(\boldsymbol{x}|\theta)$, where $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ represents a sample of observations from the population. Thus, the standard approach looks at the distribution of the data given the parameter, to estimate the parameter θ. For example, we may calculate the maximum likelihood estimator (MLE) of θ.

Conversely in the Bayesian paradigm, we no longer assume that the parameters have a fixed value, but consider θ to be a random quantity. We then assume that θ has some unknown distribution, which we wish to estimate. This distribution is denoted by $\pi(\theta|\boldsymbol{x})$, and so we look at the distribution of the parameter, given the data. In many ways this is a more natural way to make inference, but we shall see that to achieve this we will have to specify a *prior probability distribution*, denoted by $p(\theta)$, which represents our initial beliefs about the distribution of θ *prior* to observing any data.

In most situations, when we are trying to estimate a parameter θ, we have some knowledge, or preconceptions, about the value of θ before we take into account the data that we observe.

Example 1. Suppose that you are working hard at your desk, and glance out of the window to see a large wooden looking object with branches covered in green things. You consider two alternatives: one that it is a tree, the other it is a postman. Obviously you choose that it is a tree, since the object does not look anything like a postman.

We can formulate the process here. Suppose that you denote the event that you see a wooden looking object with green things on by A. Then, let B_1 denote the event that it is a tree, and B_2 that it is a postman. Then, you reject event B_2 and accept event B_1, since,

$$\mathbb{P}(A|B_1) > \mathbb{P}(A|B_2).$$

Thus, we are essentially maximizing the likelihood.

However, suppose that we entertain a third possibility, event B_3, that the object is a plastic replica tree. In this case it might well be that $\mathbb{P}(A|B_1) = \mathbb{P}(A|B_3)$, and yet you would still reject this hypothesis in favour of B_1, i.e., it is a real tree. That is, even though the probability of seeing what you observe (a large wooden looking object with green bits on) is the same whether it is a real tree or a replica tree, your *prior* belief is that it is more likely to be a real tree, and you include this in your decision. However, this might change if, for example, you were working at a desk inside a replica tree factory. Then, your *prior* beliefs would reflect this additional information, and so you may conclude that what you see is a replica tree.

The essential point is that experiments are not abstract devices. Invariably we have some knowledge about the process being investigated before obtaining any data. It is sensible to include this into our inferential process, and Bayesian inference is the mechanism for drawing inference from this combined knowledge. It should be pointed out, however, that although the underlying probabilistic derivation of Bayes' Theorem is uncontroversial, the reliance on the prior beliefs is the main criticism of Bayesian statistics. Whilst advocates of the Bayesian approach see this reliance on a prior distribution as an advantage, opponents point out that using different prior beliefs will lead to different inferences. It is whether or not you see this as a good or bad thing that determines how acceptable you find the Bayesian approach.

In more mathematical terms; a standard approach in statistics is to obtain maximum likelihood estimates, by choosing the point in parameter space that maximizes the likelihood surface. In Bayesian statistics, we average across the likelihood surface, rather than maximizing. This averaging is weighted according to the prior distribution. However, in classical statistics, we often apply different weights to different pieces of information, thus the Bayesian approach is simply a method of incorporating that weighting procedure within a rigid mathematical framework.

1.2 Bayes' Theorem

There are several different ways that Bayes' Theorem can be written, dependent on whether the quantities of interest are discrete or continuous. For simplicity we shall consider the cases separately, beginning with the discrete case.

1.2.1 Discrete case

Let A and B denote possible events, such that $\mathbb{P}(B) > 0$. Then, Bayes' Theorem states that:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

Proof:

By definition of conditional probability,

$$\begin{aligned} \mathbb{P}(A|B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\ &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}. \end{aligned}$$

□

The denominator in the expression for Bayes' Theorem is most often expressed in an alternative way. For example, denoting by A^c the complement of A, by the law of total probability,

$$\begin{aligned}\mathbb{P}(A|B) &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B\cap A)+\mathbb{P}(B\cap A^c)}\\ &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A)+\mathbb{P}(B|A^c)\mathbb{P}(A^c)}.\end{aligned}$$

More generally, suppose that we let A_i, denote a set of mutually exclusive and exhaustive events, for $i = 1, \ldots$,n. Then, by the law of total probability we can express $\mathbb{P}(B)$ in the form,

$$\mathbb{P}(B) = \mathbb{P}(B\cap A_1) + \ldots + \mathbb{P}(B\cap A_n) = \sum_{i=1}^{n}\mathbb{P}(B|A_i)\mathbb{P}(A_i).$$

Thus, an alternative expression for Bayes theorem is given by,

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i=1}^{n}\mathbb{P}(B|A_i)\mathbb{P}(A_i)}.$$

Example

A test for Hepatitis C is given to a population of possible carriers, although it is believed that only 10% have Hepatitis C. The test itself is only 95% accurate for people who have Hepatitis C, and 80% accurate for those who do not have the disease. (The former is often called *sensitivity* and the latter *specificity.*) (i) Given that a person has a negative test result, what is the probability that they actually do have Hepatitis C (i.e., that the result is a false negative)? (ii) Given that a person has a positive test result, what is the probability that they do not have Hepatitis C (i.e., that the result is a false positive)? (iii) what is the probability that a person has Hepatitis C given a positive result? What is the same probability given two positive tests (assuming the tests are identical and independent given the disease status)?

Answers:

(i) Let the event that an individual has Hepatitis C be denoted by C, the event that their test result is positive be denoted by Po, and their complements be denoted C^c and Po^c respectively. Then, we have that,

$$\begin{aligned}\mathbb{P}(C) &= 0.1\\ \mathbb{P}(Po|C) &= 0.95\\ \mathbb{P}(Po^c|C^c) &= 0.8\\ \mathbb{P}(C^c) &= 0.9\\ \mathbb{P}(Po^c|C) &= 0.05\\ \mathbb{P}(Po|C^c) &= 0.2\end{aligned}$$

Then, for the probability of a false negative,

$$\begin{aligned}\mathbb{P}(\text{Hep C} \mid \text{negative test}) &= \mathbb{P}(C|Po^c)\\ &= \frac{\mathbb{P}(Po^c\cap C)}{\mathbb{P}(Po^c)}\\ &= \frac{\mathbb{P}(Po^c\cap C)}{\mathbb{P}(Po^c\cap C)+\mathbb{P}(Po^c\cap C^c)}\\ &= \frac{\mathbb{P}(Po^c|C)\mathbb{P}(C)}{\mathbb{P}(Po^c|C)\mathbb{P}(C)+\mathbb{P}(Po^c|C^c)\mathbb{P}(C^c)}\\ &= \frac{0.05\times 0.1}{(0.05\times 0.1)+(0.8\times 0.9)}\\ &= 0.0069\end{aligned}$$

(ii) Similarly, for a false positive result,

$$\begin{aligned}\mathbb{P}(\text{No Hep C} \mid \text{positive test}) &= \mathbb{P}(C^c|Po) \\ &= \frac{\mathbb{P}(Po|C^c)\mathbb{P}(C^c)}{\mathbb{P}(Po|C^c)\mathbb{P}(C^c)+\mathbb{P}(Po|C)\mathbb{P}(C)} \\ &= \frac{0.2\times 0.9}{(0.2\times 0.9)+(0.95\times 0.1)} \\ &= 0.6545455\end{aligned}$$

Note the high false positive probability! One way of looking at this is that 90% of people are healthy, which is a high percentage, in relation to the specificity ($\mathbb{P}(Po^c|C^c) = 0.8$) of the test. See how the false positive probability would drop for a higher specificity. For instance, for $\mathbb{P}(Po^c|C^c) = 0.99$, we obtain, $\mathbb{P}(\text{No Hep C} \mid \text{positive test}) = 0.048$.

(iii) The probability that a person has the disease given a positive test is $1-\mathbb{P}(C^c|Po) = 0.3454545$. Let $Po2$ be the event that the second test is positive. For some testing procedures, it may be justifiable to assume that the two tests are independent, given the disease status and the problem specifications. We will make this assumption for this problem. **Note**: It may not be sensible to assume the two tests are independent for unknown disease status, as learning about the outcome of the first test could affect our beliefs on the probabilities for the outcome of the second test. Now, after a second positive test,

$$\begin{aligned}\mathbb{P}(\text{Hep C} \mid \text{2 positive tests}) &= \mathbb{P}(C|Po,Po2) \\ &= \frac{\mathbb{P}(Po,Po2|C)\mathbb{P}(C)}{\mathbb{P}(Po,Po2|C)\mathbb{P}(C)+\mathbb{P}(Po,Po2|C^c)\mathbb{P}(C^c)} \\ &= \frac{\mathbb{P}(Po|C)\mathbb{P}(Po2|C)\mathbb{P}(C)}{\mathbb{P}(Po|C)\mathbb{P}(Po2|C)\mathbb{P}(C)+\mathbb{P}(Po|C^c)\mathbb{P}(Po2|C^c)\mathbb{P}(C^c)} \\ &= \frac{0.95\times 0.95\times 0.1}{(0.95\times 0.95\times 0.1)+(0.2\times 0.2\times 0.9)} \\ &= 0.7148515\end{aligned}$$

In Bayesian updating, one can use the previous posterior as the current prior, and update beliefs in a sequential manner. See Tutorial 1, for a different derivation of $\mathbb{P}(C|Po,Po2) = 0.7148515$.

Example

A new HIV test has 95% sensitivity and a 98% specificity. Assume that HIV prevalence is 1/1000. What is the probability that a patient that tested positive has HIV? What is the probability that a patient is healthy given that he tested negative? What is the probability that a patient is positive given a positive test? Why the counter intuitive result?

Answers:

Let the event that an individual is HIV positive be denoted by H, the event that their test result is positive be denoted by Po, and their complements be denoted H^c and Po^c respectively. Then, we have that,

$$\begin{aligned}\mathbb{P}(H) &= 0.001 \\ \mathbb{P}(Po|H) &= 0.95 \\ \mathbb{P}(Po^c|H^c) &= 0.98 \\ \mathbb{P}(H^c) &= 0.999 \\ \mathbb{P}(Po^c|H) &= 0.05 \\ \mathbb{P}(Po|H^c) &= 0.02\end{aligned}$$

For the probability that someone is healthy given a negative test,

$$\begin{aligned}\mathbb{P}(\text{Healthy} \mid \text{negative test}) &= \mathbb{P}(H^c|Po^c) \\ &= \frac{\mathbb{P}(Po^c|H^c)\mathbb{P}(H^c)}{\mathbb{P}(Po^c|H^c)\mathbb{P}(H^c)+\mathbb{P}(Po^c|H)\mathbb{P}(H)} \\ &= \frac{0.98\times 0.999}{(0.98\times 0.999)+(0.05\times 0.001)} \\ &= 0.999\end{aligned}$$

Looks like a great test! But, for the probability that someone is HIV positive given a positive test,

$$\begin{aligned}\mathbb{P}(\text{HIV pos.} \mid \text{positive test}) &= \mathbb{P}(H|Po) \\ &= \frac{\mathbb{P}(Po|H)\mathbb{P}(H)}{\mathbb{P}(Po|H)\mathbb{P}(H)+\mathbb{P}(Po|H^c)\mathbb{P}(H^c)} \\ &= \frac{0.95\times 0.001}{(0.95\times 0.001)+(0.02\times 0.999)} \\ &= 0.045\end{aligned}$$

So, over 95% of those tested positive will in fact NOT be HIV positive. Why the counter intuitive result for what appeared to be a good test? It is because of the very low prevalence of HIV, in relation to the sensitivity and specificity of the test. As in the previous example, check how the last calculation would change if the specificity was much higher.

So, when probabilities are 'inverted' the results may be counter intuitive. See some of the extra material uploaded on MMS on Bayesian inversion and the prosecutor's fallacy.

Note: Subjective, objective and long-run probability

Subjective (or personal) probability. Subjective probability measures the "degree of belief" or uncertainty of a given person, conveniently called You, about the event's occurrence. This subjective probability is conditional on Your personal assessment of the relevant available information. For example, if you are based in St Andrews, what is *your* probability that it will rain tomorrow (event E), given that you have observed the weather over the last month, and also that you know it is Autumn and we are in Fife (background information H)? The probability $P(E|H)$ is personal as, for non-trivial events, typically no 2 people have the same information H or make judgements in the same way. If $P(E|H) = 1$, You are certain that E will occur, and if $P(E|H) = 0$, You are certain that it will not occur. Probability is relevant wherever there is uncertainty, not just randomness. As your information changes, so do your probabilities.

Objective probability. Some probabilities arise from **objective** logical thinking, e.g. due to symmetry. For example, think of throwing a fair die. The item is symmetric and so we believe that the probability of observing, say, a 3 is 1/6. Module MT2504 gives many more examples of using Combinatorics to derive such probabilities. The Bayesian approach considers them as probabilities that everyone (or nearly everyone) would agree on, based on the current body of knowledge. These probabilities mostly concern carefully designed experiments, or events with a very specific underlying mechanism in place, such as the Mendelian Laws in biological reproduction.

Long-run probability. The long-run probability $P(E)$, is the proportion of times E will occur in an infinitely long series of repeated identical situations. For example, the probability of "tails" can be derived by repeatedly tossing a fair coin and counting the frequency of observing "tails". This is again a rather artificial framework, as in practice it is not easy to repeat the same experiment/situation/event a very large number of times.

1.2.2 Continuous (single parameter) case

So far we have only considered the discrete case. Bayes' Theorem is equally well defined for continuous parameters, and it is this specification of Bayes' Theorem that is most often used.

Suppose that we have a parameter $\theta \in \Theta$, on which we wish to make inference. We then observe data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from some known probability distribution $f(\boldsymbol{x}|\theta)$, which is a function of the parameter value θ. Then Bayes' Theorem states that,

$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)p(\theta)}{f(\boldsymbol{x})}$$

Here the term $p(\theta)$ is referred to as the **prior** distribution and $\pi(\theta|\boldsymbol{x})$ the **posterior** distribution. Essentially the prior represents the initial beliefs concerning the parameters prior to any data being observed, whereas the posterior distribution represents an update of these beliefs, following the data $\boldsymbol{x}$ being observed. The information contained in the data on the parameter $\boldsymbol{\theta}$ is represented by the term $f(\boldsymbol{x}|\boldsymbol{\theta})$, and is most usually called the **likelihood**. We can write the denominator of Bayes' Theorem as,

$$f(\boldsymbol{x}) = \int_{\Theta} f(\boldsymbol{x}|\theta)p(\theta)d\theta.$$

Thus, this term is independent of θ, the parameter of interest, and is simply equal to some constant. It 'shifts' the posterior distribution function up or down, so that it integrates to one. Then, $f(\boldsymbol{x})$ is usually referred to as the marginal likelihood, and $f(\boldsymbol{x})^{-1}$ as the normalization constant, and they are often found by inspection. Note that in many cases the integration for the normalization constant may be analytically intractable, or tedious to calculate. Then, one option is to calculate it using stochastic simulation (see second part of this course), although this calculation is often not necessary. More often Bayes' Theorem is quoted as,

$$\pi(\theta|\boldsymbol{x}) \propto f(\boldsymbol{x}|\theta)p(\theta).$$

This formula forms the essential core of Bayesian inference.

Example

Suppose that we observe data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, such that, given λ, each $X_i \overset{iid}{\sim} Exp(\lambda)$. We place the following prior on λ, namely that,

$$\lambda \sim \Gamma(\alpha, \beta).$$

Note: $E(\lambda) = \alpha/\beta$. Then, the corresponding posterior distribution for λ is given by,

$$\begin{aligned}
\pi(\lambda|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\lambda)p(\lambda) = f(x_1|\lambda) \times \ldots f(x_n|\lambda)p(\lambda) \\
&= \prod_{i=1}^{n} \lambda \exp(-x_i\lambda) \times \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp(-\lambda\beta) \\
&\propto \lambda^n \exp\left(-\lambda\sum_{i=1}^{n} x_i\right) \times \lambda^{\alpha-1}\exp(-\lambda\beta) \\
&= \lambda^{n+\alpha-1}\exp(-\lambda[n\bar{x}+\beta]) \\
&\propto \frac{(n\bar{x}+\beta)^{n+\alpha}}{\Gamma(n+\alpha)}\lambda^{n+\alpha-1}\exp(-\lambda[n\bar{x}+\beta]) \\
\Rightarrow \lambda|\boldsymbol{x} &\sim \Gamma(n+\alpha, n\bar{x}+\beta).
\end{aligned}$$

Given this, we can state that the constant of proportionality (the constant we multiply with to obtain a density that integrates to one) is equal to, $\frac{(n\bar{x}+\beta)^{n+\alpha}}{\Gamma(n+\alpha)}$. Or, by inspection, we can write that,

$$f(\boldsymbol{x}) = \frac{\Gamma(n+\alpha)\beta^{\alpha}}{(n\bar{x}+\beta)^{n+\alpha}\Gamma(a)}.$$

Application: Assume that $1/\lambda$ denotes the average lifetime of a laptop in years. Assume that prior beliefs are described by $\lambda \sim \Gamma(0.2, 0.6)$, so that $\mathrm{E}(\lambda) = 1/3$, and $\mathrm{Var}(\lambda) = 0.2/0.6^2 = 0.55$. Assume now that 20 laptops are tested, and their average lifetime turns out to be $\bar{x} = 5$. Then,

$$\lambda|\boldsymbol{x} \sim \Gamma(20 + 0.2, 20 \times 5 + 0.6) = \Gamma(20.2, 100.6),$$

so that $\mathrm{E}(\lambda|\boldsymbol{x}) = 20.2/100.6 = 0.2007$. Although the expectation of a ratio is not the ratio of expectations, this will roughly give a posterior expectation for the lifetime of a laptop equal to 5 years.

For a different prior $\lambda \sim \Gamma(10, 30)$, so that $\mathrm{E}(\lambda) = 1/3$, and $\mathrm{Var}(\lambda) = 10/900 = 0.011$, $\lambda|\boldsymbol{x} \sim \Gamma(20 + 10, 20 \times 5 + 30) = \Gamma(30, 130)$. Now, $\mathrm{E}(\lambda|\boldsymbol{x}) = 30/130 = 0.2307$. This will roughly give a posterior expectation for the lifetime of a laptop equal to 4.3 years.

Example

We observe data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, such that each $X_i \overset{iid}{\sim} Geom(\theta)$, where we wish to make inference on the parameter θ. Then, initially we must specify a prior on the parameter θ. Suppose that we set,

$$\theta \sim Beta(\alpha, \beta).$$

Find the posterior distribution.
Answer:

$$\begin{aligned}
\pi(\theta|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\theta)p(\theta) \\
&\propto \prod_{i=1}^{n} \theta(1-\theta)^{x_i-1} \times \theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \theta^{n+\alpha-1}(1-\theta)^{n\bar{x}-n+\beta-1} \\
&= \theta^{a-1}(1-\theta)^{b-1},
\end{aligned}$$

where $a = n + \alpha$ and $b = n\bar{x} - n + \beta$. Thus,

$$\theta|\boldsymbol{x} \sim Beta(a, b) = Beta(n + \alpha, n\bar{x} - n + \beta),$$

by inspection.

1.3 Conjugate Bayesian analysis

Often, a particular distributional family is chosen for the prior, such that the corresponding posterior distribution of the parameter belongs to the same family, irrespective of the sample size and any value of the observations. (This was the case in the previous two examples) Then, the prior distribution is known as a **conjugate** prior.

Definition 1.1: *A family of probability distributions, $\mathcal{F}$, is conjugate to a family of sampling distributions, $\mathcal{P}$, if whenever the prior belongs to the family, $\mathcal{F}$, then for any sample size and any value of observations, the posterior also belongs to the family, $\mathcal{F}$.*

Example

Suppose that a treatment (radiation) has a probability p of success in treating cancer. Success is denoted with $X = 1$, failure with $X = 0$. We monitor n randomly selected patients with (0 or 1) responses $x_1, x_2, ..., x_n$. We observe s positive responses in total, i.e.

$$\sum_{i=1}^{n} x_i = s.$$

Suppose that we are prepared to assume that $x_1, \dots, x_n$ are independently and identically distributed (iid), given p, with,

$$P(X_i = 1|p) = p, \quad i = 1, \dots, n.$$

Likelihood:

$$P(X_1 = x_1, ..., X_n = x_n|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^s(1-p)^{n-s}.$$

Alternatively, we can assume that the total number of successes S in n patients follows a Binomial distribution so that, $S|p \sim Bin(n, p)$. The likelihood will then be,

$$p(S = s|p) = \binom{n}{s} p^s(1-p)^{n-s}.$$

The posterior distribution **will be the same** under the two likelihoods, but the marginal distribution of the observations $f(\boldsymbol{x})$ will be different. See relevant question in Tutorial 1b.

Conjugate prior: We place a $Beta(a, b)$ prior on p, so that,

$$p(p) = \frac{1}{B(a,b)} p^{a-1}(1-p)^{b-1} \propto p^{a-1}(1-p)^{b-1},$$

with Beta function $B(a, b) = \int_0^1 z^{a-1}(1-z)^{b-1}dz = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.
Note that $E(p) = \frac{a}{a+b}$ and $Var(p) = \frac{ab}{(a+b)^2(a+b+1)}$.
(See lecture and material on Moodle for examples of different Beta densities using R.)

Posterior: By Bayes' Theorem,

$$\begin{aligned} \pi(p|s) &\propto f(s|p)p(p) \\ &\propto p^s(1-p)^{n-s} \times p^{a-1}(1-p)^{b-1} \\ &= p^{s+a-1}(1-p)^{n-s+b-1} \end{aligned}$$

Then, by inspection, we have that,

$$p|s \sim Beta(s + a, n - s + b).$$

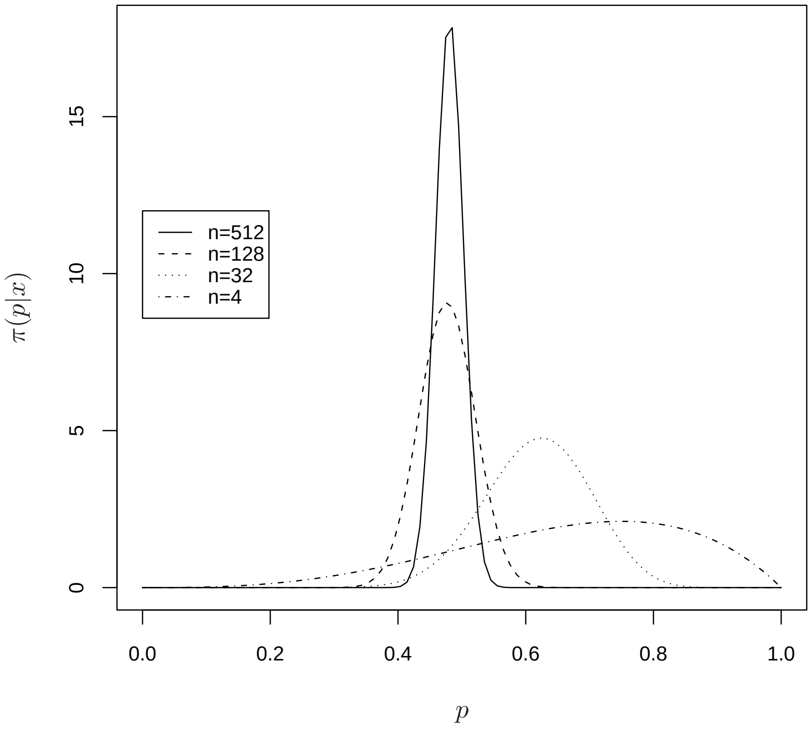


Fig. 1.1: The posterior distribution of p, the probability of successful treatment, for (i) $n = 4, s = 3$, (ii) $n = 32, s = 20$, (iii) $n = 128, s = 61$ and (iv) $n = 512, s = 248$, with a $U[0, 1]$ prior on p.

Thus, the Beta distribution is a conjugate prior to the Binomial distribution.

Consider the particular case where we specify our prior beliefs, such that $a = b = 1$, i.e. a $U[0, 1]$ prior on p. Then, Figure 1.1 shows the corresponding posterior distribution of p for one replicate of the experiment when $n = 4, 32, 128, 512$, and $x = 3, 20, 61, 248$, respectively.

We can see from Figure 1.1 that as we obtain more information about the parameter, through more trials, the posterior distribution becomes more and more peaked, and we become more certain about the unknown probability of success p.

We can obtain further insight into the way in which the posterior distribution combines information from the data with that from the prior, by considering the form of the posterior mean. We have that,

$$\mathbb{E}_\pi(p) = \frac{s+a}{a+s+b+n-s} = \frac{s+a}{n+a+b}.$$

We can rewrite this expectation in the form,

$$\mathbb{E}_\pi(p) = \frac{(a+b)\left(\frac{a}{a+b}\right) + n\left(\frac{s}{n}\right)}{n+a+b},$$

which can be reformulated as,

$$(1-w)\left(\frac{a}{a+b}\right) + w\left(\frac{s}{n}\right),$$

where $w = n/(n + a + b)$. In other words, the Bayes estimate is a *weighted average* of the two quantities,

$$\frac{a}{a+b} \qquad \text{and} \qquad \frac{s}{n}.$$

The first is the mean of the prior distribution and is the Bayes estimate we could use if we had no data. The latter is the “usual” classical estimate of p, derived via maximum likelihood or minimum variance unbiased estimation.

In an obvious sense, we see that our estimate is a combination of what the data tells us, s/n, and what we believed before observing any data, $\alpha/(\alpha + \beta)$. As the amount of data increases i.e. as n increases, more and more weight is placed on s/n; mathematically, in the limiting case, as $n \to \infty$, we have that $w \to 1$. Conversely, if we have no data, i.e. $n = 0$, then $w = 0$ and our only source of information on the parameter is contained within the prior.

Usually, we are not only interested in the Bayes estimate of the parameter, but the whole shape of the posterior distribution of the parameter. Often, we may also be interested in the “spread” of the distribution. Clearly, from Figure 1.1, as the number of coin tosses, n, increases, the precision of the posterior distribution for p increases, as we have more information on the parameter from the data. This can be seen formally, by considering the posterior variance for p,

$$Var_\pi(p) \quad = \quad \frac{(s+a)(n-s+b)}{(n+a+b)^2(n+a+b+1)}$$

So, that in the limiting case, as $n \to \infty$, we have that $Var_\pi(p) \to 0$. Thus, irrespective of our prior beliefs represented by the prior parameters a and b, as the amount of information increases, the posterior distribution becomes more and more dominated by the data, and our posterior beliefs become more and more concentrated on a value of p tending to a value of s/n.

In general, since the posterior distribution is formed by combining the likelihood with the prior, there is a trade-off between the information contained in the data and the strength of the prior beliefs. Posterior distributions are often said to be “data-driven” if the likelihood dominates the posterior; and “prior-driven” if the prior dominates the posterior.

Example

Consider that we toss n coins and obtain s number of heads. Then, suppose that we have two different priors on the probability p:

$$\begin{aligned} p &\sim Beta(2, 10) \\ p &\sim Beta(10, 2). \end{aligned}$$

We toss the coin 5 times and obtain 3 heads, i.e. $n = 5$, and that $s = 3$. The priors, likelihood function and corresponding posterior distributions are given in Figure 1.2. Clearly, here we can see that the prior dominates the posterior distribution, i.e. the posterior distribution looks more like the prior than the likelihood.

However, suppose that we continue to toss the coin, so that we toss the coin a total of 500 times, and obtain 242 heads. The corresponding priors, likelihood and posterior distributions are given in Figure 1.3. Clearly, here the posterior distribution is dominated by the likelihood term, which contains the information contained in the data. Thus, the posterior distribution is data-driven, with little influence from the prior distribution.

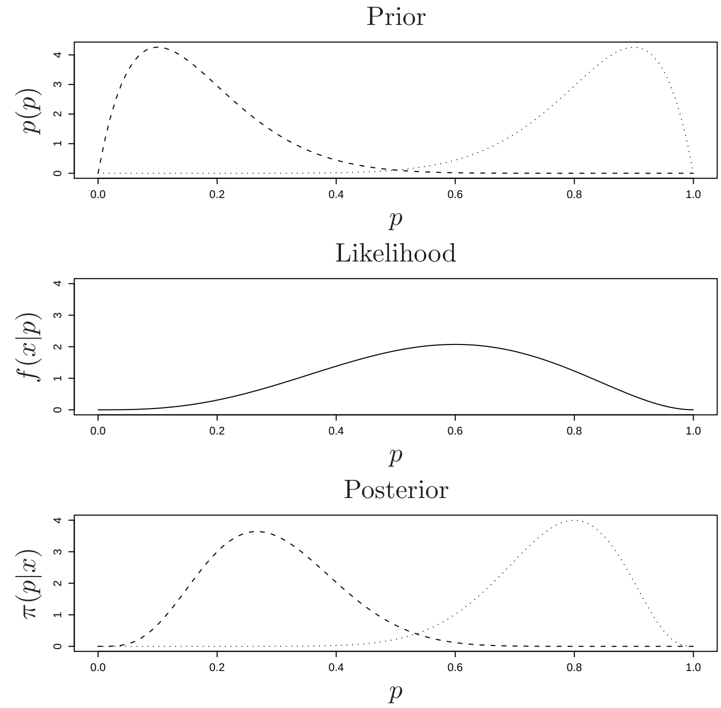


Fig. 1.2: The prior, likelihood and posterior distribution of p when $n = 5$, for prior 1: $p \sim Beta(2, 10)$ (dashed line) and prior 2: $p \sim Beta(10, 2)$ (dotted line).

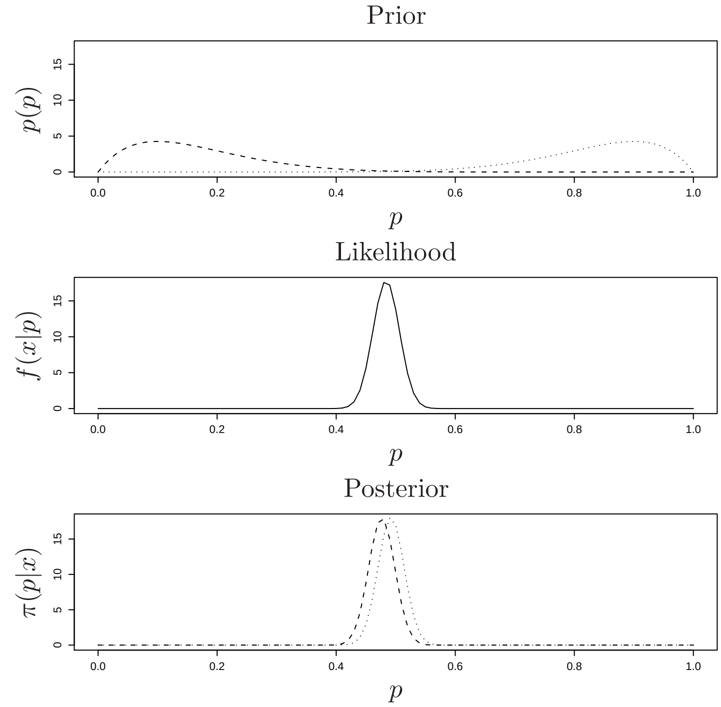


Fig. 1.3: The prior, likelihood and posterior distribution of p when $n = 500$, for prior 1: $p \sim Beta(2, 10)$ (dashed line) and prior 2: $p \sim Beta(10, 2)$ (dotted line).

Note that in typical Bayesian analyses, often a variety of different prior distributions will be used and the corresponding posterior distributions compared in order to see to what extent the priors influence the posterior distribution. This is often called a prior sensitivity analysis.

Aside: The Likelihood Principle:

Bayesian inference obeys what is commonly called the *likelihood principle*. According to this principle, for a given sample of data, $\boldsymbol{x}$, any probability models that lead to the same likelihood for the data should yield the same inference for θ. This means that the data only affect the posterior via the likelihood, $f(\boldsymbol{x}|\theta)$ and that the prior is independent of the data. Experiments may be different in various aspects, but those differences are irrelevant for inference about θ. To further probe into the likelihood principle, consider two experiments one yielding data x and the other y, so that $f(y|\theta) = cf(x|\theta)$ where c does not depend on θ. Then the two experiments contain identical information about θ and we can easily check that they lead to identical posterior distributions, as the constant c will cancel out in the posterior distribution calculations,

$$\pi(\theta|y) = \frac{f(y|\theta)p(\theta)}{\int f(y|\theta)p(\theta)d\theta} = \frac{cf(x|\theta)p(\theta)}{\int cf(x|\theta)p(\theta)d\theta} = \frac{f(x|\theta)p(\theta)}{\int f(x|\theta)p(\theta)d\theta} = \pi(\theta|x).$$

An example of this was shown above, where the likelihood for estimating a proportion can be assumed to be the Binomial distribution (where the order of obtaining successes and failures does not matter), or the product of the probabilities for the obtained successes and failures in the order they were observed. Both options result in the same posterior distribution.

Example: Normal prior/posterior, with unknown μ and known σ^2

We shall assume that we observe conditionally independent observations $\boldsymbol{x} = \{x_1, \dots, x_n\}$, drawn from the Normal distribution, i.e. given μ and σ, $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, \dots, n$. Assume that μ is unknown, whilst σ is known. We need to first place a prior on the unknown mean, μ. Suppose that we specify the prior,

$$\mu \sim N(\phi, \tau^2).$$

Find the corresponding posterior distribution:

$$\begin{aligned}
\pi(\mu|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\mu)p(\mu) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\mu-\phi)^2}{2\tau^2}\right) \\
&\propto \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right) \exp\left(-\frac{(\mu-\phi)^2}{2\tau^2}\right) \\
&\propto \exp\left(-\frac{(n\mu^2 - 2n\bar{x}\mu)}{2\sigma^2}\right) \exp\left(-\frac{(\mu^2 - 2\mu\phi)}{2\tau^2}\right) \\
&= \exp\left(-\frac{\mu^2(\tau^2 n + \sigma^2) - 2\mu(\tau^2 n\bar{x} + \sigma^2\phi)}{2\sigma^2\tau^2}\right) \\
&\propto \exp\left(-\frac{(\tau^2 n + \sigma^2)\left[\mu - \frac{(\tau^2 n\bar{x} + \sigma^2\phi)}{(\tau^2 n + \sigma^2)}\right]^2}{2\sigma^2\tau^2}\right).
\end{aligned}$$

By completing the square we can identify the posterior distribution for μ to be,

$$\mu|\boldsymbol{x} \sim N\left(\frac{\tau^2 n\bar{x} + \sigma^2\phi}{\tau^2 n + \sigma^2}, \frac{\sigma^2\tau^2}{\tau^2 n + \sigma^2}\right).$$

Thus, the Normal prior on μ is a conjugate prior. We can also note that $\bar{x}$ is sufficient in the Normal case. It is also clear from the form of the posterior distribution, that the posterior mean is a mixture of the prior mean (ϕ) and the classical MLE for the mean ($\bar{x}$), as we can write,

$$E(\mu|\boldsymbol{x}) = \frac{\tau^2 n\bar{x} + \sigma^2\phi}{\tau^2 n + \sigma^2} \quad = \quad w\bar{x} + (1-w)\phi,$$

where,

$$w = \frac{\tau^2 n}{\tau^2 n + \sigma^2}.$$

The value of the prior variance, τ^2, specifies the informativeness of the prior. A small variance, implies a "tight" prior distribution around ϕ for the parameter μ; whereas a large prior variance suggests that there is little information contained in the prior concerning the parameter value, with a relatively flat prior distribution. This can be clearly seen, if we consider the posterior distribution for μ.

Initially, consider the case for τ^2 small. In the limiting case, as $\tau^2 \to 0$, we have that the mean of the distribution tends to ϕ (i.e. the prior mean for μ), with corresponding variance,

$$\frac{\sigma^2\tau^2}{\tau^2 n + \sigma^2} \quad = \quad \frac{\tau^2}{n\tau^2/\sigma^2 + 1}.$$

As $\tau^2 \to 0$, clearly, the posterior variance tends to 0. Thus, the prior dominates the posterior distribution.

Conversely, consider τ^2 large, so that we have a vague prior on the parameter. Then, again in the limiting case $\tau^2 \to \infty$, the posterior mean for μ tends to $\bar{x}$. Additionally, the variance tends to σ^2/n. This can be compared to the result in classical statistics, where the sampling distribution of $\bar{X}$ is normal with mean μ and variance σ^2/n. However, these statements must not be confused, and although we may have the same answer, the question is different between the classical and Bayesian approaches. For a Bayesian, we apply the probability statement to the parameter μ; for a classicist, the probability statement is applied to the statistic $\bar{X}$.

Example

Data, $\boldsymbol{x}$, are collected on the length of time x_i it takes student i to answer a particular question within an examination. The data collected (in minutes) are:

36, 67, 44, 39, 56, 65, 43, 49.

So, the sample mean is 49.87. It is assumed that, given μ and σ, each $X_1, \ldots, X_{10} \overset{iid}{\sim} N(\mu, \sigma^2)$. The mean is unknown, and is to be estimated. The variance σ^2 can be taken to represent the variability in the ability of students: small σ^2 would represent a fairly homogeneous set of students, whereas a large σ^2 would represent a heterogeneous mix of students of varying abilities. Here, from previous examinations, we know that $\sigma^2 = 100$.

The corresponding prior for the mean is specified in the form,

$$\mu \sim N(\phi, \tau^2).$$

The mean, $\phi = 45$, is the length of time the examiner expects a student to take answering the question. The variance, τ^2 is to be specified by the expert (i.e. the examiner).

The posterior distribution of μ is,

$$\mu|\boldsymbol{x} \sim N\left(\frac{399\tau^2 + 4500}{8\tau^2 + 100}, \frac{100\tau^2}{8\tau^2 + 100}\right).$$

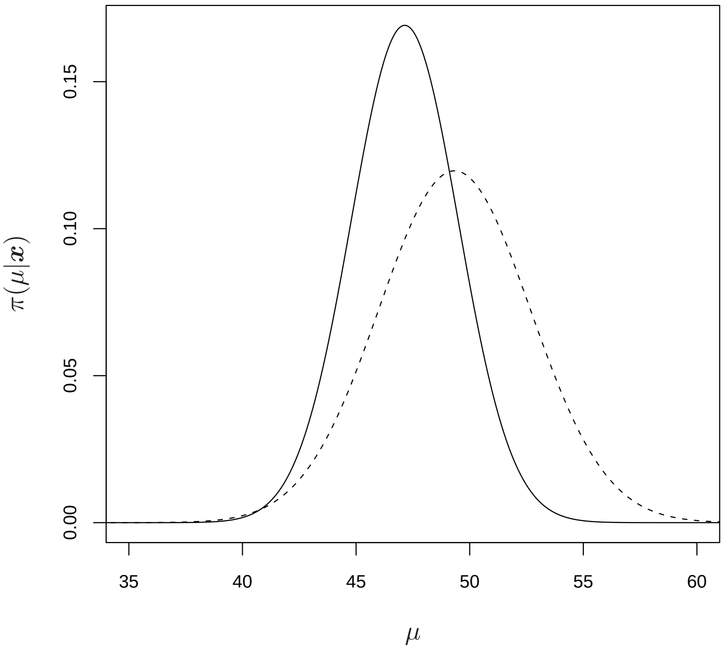


Fig. 1.4: The posterior distribution of μ, for (a) $\tau^2 = 100$ (dashed line) and (b) $\tau^2 = 1$ (solid line).

Figure 1.4 gives the corresponding posterior distribution for μ, under two different priors: (a) $\tau^2 = 100$ and (b) $\tau^2 = 1$.

The corresponding posterior mean for μ can be expressed as a weighted average of the prior mean and classical MLE, as shown above. So that in this case the mean can be expressed in the form,

$$\frac{8\tau^2}{8\tau^2 + 100}\bar{x} + \frac{100}{8\tau^2 + 100}\phi.$$

So that, when the prior variance on the parameter μ is equal to 100, we obtain mixture weights of $\frac{8}{9}$ and $\frac{1}{9}$, on the MLE and prior, respectively, and thus a posterior mean for μ of 49.33. Alternatively, when we consider the variance equal to 1, the corresponding weights are 0.074 and 0.926, giving a posterior mean of 45.36. Overall, for $\tau^2 = 100$, we obtain $\mu|\boldsymbol{x} \sim N(49.33, 100/9)$. For $\tau^2 = 1$, we obtain $\mu|\boldsymbol{x} \sim N(45.36, 100/108)$.

1.4 Prior distributions

The specification of a prior on the unknown parameters of interest, before observing any data is controversial. Bayesians argue that the Bayesian approach allows the introduction of any external information that may be available. Conversely, classicists/frequentists argue that the analysis of the data should be objective, and any results obtained should be purely on the evidence of the data, and not influenced by subjective priors that will generally differ between individuals. So, how should we assign the prior $p(\theta)$? Let us be clear from the beginning - there is no such thing as the *correct choice* of $p(\theta)$ for a given problem. The actual choice of prior lies entirely with the statistician and the information and experience s/he has at the time. Of course, the investigator should be able to persuade their audience that their choice of prior is sensible.

1.4.1 Non-informative/vague priors

An obvious question to ask is: what should we do if we do not have any prior information concerning the parameter of interest? Bayes himself suggested that when this is the case, the Uniform prior should be used, so that all parameter values are, *a priori*, equally likely. When this is the case, clearly we have that,

$$\pi(\theta|\boldsymbol{x}) \propto f(\boldsymbol{x}|\theta),$$

i.e. the posterior distribution is the same shape as the likelihood function. Note that in this case, the posterior mode of the distribution is equal to the MLE of the parameter.

However, non-linear transformations of the parameter θ, denoted by $\phi = g(\theta)$, say, will result in a non-Uniform prior on this transformed parameter ϕ. For example, suppose that we have no information about a parameter θ apart from that it lies in $[0,1]$. We place a Uniform prior, $p(\theta) = 1$, $\theta \in [0,1]$. Then the corresponding prior on $\phi = g(\theta) = \theta^2$ is non-Uniform. Note that according to the change of variable rule, $p(\phi) = p(g^{-1}(\phi)) \left|dg^{-1}(\phi)/d\phi\right|$.

We have that (using the transformation of variables rule),

$$p_\phi(\phi) = p\left(g^{-1}(\phi)\right)\left|\frac{dg^{-1}(\phi)}{d\phi}\right| = p\left(\sqrt{\phi}\right)\left|\frac{d\sqrt{\phi}}{d\phi}\right| = \left|\frac{d\sqrt{\phi}}{d\phi}\right| = \frac{1}{2\sqrt{\phi}}, \qquad \phi \in [0,1].$$

However, one may expect that ignorance about the value of θ would imply ignorance about $\phi = \theta^2$. More generally, it might be beneficial to be able to define a non-informative prior, $p(\theta)$, so that the prior for $\phi = g(\theta)$ is non-informative for ϕ **in the same way** that the prior for θ is not informative for θ.

Jeffreys' prior

Jeffreys suggested a prior based on an invariance rule for one-to-one (bijective) transformations. The idea is to derive a prior for θ so that for any $\phi = h(\theta)$ (h: bijective function) computing the prior for ϕ produces a prior that is uninformative for ϕ in exactly the same manner as the prior for θ is uninformative for θ.

Jeffreys' prior is given by,

$$p(\theta) \propto \sqrt{I(\theta|\boldsymbol{x})},$$

where $I(\theta|\boldsymbol{x})$ is the Fisher Information

$$I(\theta|\boldsymbol{x}) = \mathbb{E}_x\left(\frac{d\log f(\boldsymbol{x}|\theta)}{d\theta}\right)^2.$$

Essentially, Fisher's information is an indicator of the amount of information supplied by the model and observations about an unknown parameter θ. Looking at it as a function of θ, at the regions of the parameter space where it obtains high values, the amount of information brought by the data is high. If we use this function/curve as a prior for θ, we favour the values of θ for which $I(\theta|\boldsymbol{x})$ is large, i.e., we minimize the influence of the prior.

Under certain regularity conditions, Fisher's information can also be expressed in the form,

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

Fisher's information is generally more easily calculated using this latter expression.

Then, suppose that we wish to consider the parameter ϕ, which is a bijective transformation of the parameter θ, i.e. $\phi = h(\theta)$. If we specify,

$$p(\theta) \propto \sqrt{I(\theta|\boldsymbol{x})}$$

then,

$$p(\phi) \propto \sqrt{I(\phi|\boldsymbol{x})}.$$

To prove this result, first remember that for a transformed random variable X, so that $Y = g(X)$, where g is bijective,

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

Remember also that,

$$\frac{d}{dx}f(g(x)) = \frac{df(y)}{dy}\frac{dy}{dx}.$$

Notice now that Fisher's information is

$$\begin{aligned}
I(\theta|\boldsymbol{x}) &= \mathbb{E}\left(\frac{d\log f(\boldsymbol{x}|\theta)}{d\theta}\right)^2 \\
&= \mathbb{E}\left(\frac{d\log f(\boldsymbol{x}|\theta = h^{-1}(\phi))}{d\theta}\right)^2 \\
&= \mathbb{E}\left(\frac{d\log f(\boldsymbol{x}|\phi)}{d\phi}\times\frac{d\phi}{d\theta}\right)^2, \quad \text{as } h \text{ is bijective} \\
&= \left|\frac{d\phi}{d\theta}\right|^2\mathbb{E}\left(\frac{d\log f(\boldsymbol{x}|\phi)}{d\phi}\right)^2 \\
&= I(\phi|\boldsymbol{x})\left|\frac{d\phi}{d\theta}\right|^2. \qquad (1.1)
\end{aligned}$$

Assume that the prior on the parameter θ is specified in the form:

$$p(\theta) \propto \sqrt{I(\theta|\boldsymbol{x})}$$

Using the transformation of variable rule, we have that,

$$\begin{aligned}
p(\phi) &\propto \sqrt{I(\phi|\boldsymbol{x})\left|\frac{d\phi}{d\theta}\right|^2}\times\left|\frac{d\theta}{d\phi}\right| \\
&= \sqrt{I(\phi|\boldsymbol{x})},
\end{aligned}$$

as required.

□

Note: suppose that we have n independent observations $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ generated from the same distribution with pdf f. Then, it can be shown that Fisher's information is given by,

$$I(\theta|\boldsymbol{x}) \quad = \quad nI(\theta|x),$$

where $X \sim f$.

Jeffreys' prior can be extended to the case where there are several unknown parameters. Then, Fisher's information is defined as the matrix, with the element in row i and column j given by,

$$(I(\boldsymbol{\theta}|\boldsymbol{x}))_{i,j} = \mathbb{E}\left(\frac{d^2 \log f(\boldsymbol{x}|\boldsymbol{\theta})}{d\theta_i d\theta_j}\right).$$

Then, the prior is specified as,

$$p(\boldsymbol{\theta}) \propto \sqrt{\det I(\boldsymbol{\theta}|\boldsymbol{x})}.$$

Example

Let X denote the number of defective items in a batch of n fudge doughnuts, where each doughnut is defective with probability θ, independently of each other, given θ. Then $X \sim Bin(n, \theta)$ and, $f(x|\theta) \propto \theta^x(1-\theta)^{n-x}$, so that,

$$\log f(x|\theta) = x \log \theta + (n-x)\log(1-\theta) + C,$$

so that,

$$\frac{d^2 \log f(x|\theta)}{d\theta^2} = -\frac{x}{\theta^2} - \frac{(n-x)}{(1-\theta)^2}.$$

Then,

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

So that, Jeffreys' prior for the probability parameter θ is,

$$p(\theta) \propto \sqrt{I(\theta|x)} \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}.$$

In other words, $\theta \sim Beta\left(\frac{1}{2}, \frac{1}{2}\right)$.

The most important objection to Jeffrey's prior is that it does not satisfy the Likelihood Principle. (That is, probability models that lead to the same likelihood for the data should give the same inferences for θ.) Depending on the design of the experiment or the stopping rule, Jeffreys' prior may be different even if the likelihood for the observed data is the same, leading to different posterior distributions. For example, there are different Jeffrey's priors for Binomial and Negative Binomial experiments, and posterior inference using Jeffrey's prior in each case will violate the Likelihood Principle.

Alternative vague or non-informative prior distributions often have a reasonable mean for the distribution, but with a large variance parameter. Again, usually within a Bayesian analysis, several different priors may be considered, each of which may be described to be "vague" or "non-informative", and the sensitivity of the posterior on these priors investigated. Note that vague prior distributions do not have to be proper distributions, in the sense that they integrate to one. As long as the posterior distribution is proper, then it is acceptable to use an improper vague prior distribution.

Example

Assume $x \sim N(\mu, \phi)$, given μ, ϕ. Find the Jeffrey's prior for ϕ. Is it a proper prior? Check also that the posterior is a proper distribution.

Answer:

$$f(x|\phi) = \frac{1}{\sqrt{2\pi\phi}}\exp\{-\frac{1}{2\phi}(x-\mu)^2\}.$$

$$\log(f(x|\phi)) = -\frac{1}{2}\log(\phi) - \frac{1}{2\phi}(x-\mu)^2 + C.$$

$$\frac{d^2 \log f(x|\phi)}{d\phi^2} = \frac{1}{2}\phi^{-2} - \frac{(x-\mu)^2}{\phi^3}.$$

$$I(\theta|x) = -\mathbb{E}\left(\frac{d^2 \log f(x|\phi)}{d\phi^2}\right) = -\frac{1}{2}\phi^{-2} + \frac{\phi}{\phi^3}$$

So, $p(\phi) \propto \sqrt{I(\theta|x)} \propto \sqrt{\frac{1}{2}\phi^{-2}} \propto 1/\phi$, which is not a proper distribution (because the pdf does not integrate to 1). However, note that the pdf of the posterior distribution is

$$\pi(\phi|x,\mu) \quad \propto \quad \frac{1}{2\pi\phi}\exp\left(-\frac{1}{2\phi}(x-\mu)^2\right)\phi^{-1} \propto \phi^{-1+\frac{1}{2}}\exp\left(-\frac{1}{2\phi}(x-\mu)^2\right)$$

and therefore $(\phi|x) \sim \Gamma^{-1}(\frac{1}{2}, \frac{1}{2}(x-\mu)^2)$, which is a proper probability distribution.

1.4.2 Informative priors - Elicitation

Elicitation is the process of extracting prior knowledge (from someone - an expert, perhaps, or even yourself, or a group of experts) so as to express it as a prior distribution.

The goal is to identify and quantify the key aspects of prior knowledge, since the expert will find these easier to specify. It is, however, important to recognize that no matter how many summaries for a distribution we choose to elicit, they will not identify a complete distribution. For example, a continuous distribution for a parameter θ implies an infinite number of statements about this parameter. Once summaries have been elicited, there will be more than one distributions that fit those summaries. In practice, once we elicit some key aspects of prior knowledge, we fit them to a distribution and, if the expert is happy with our feedback, we consider it to be the prior.

Elicitation errors. There are two kinds of error when eliciting prior knowledge. First, fitting error, which is the error that takes place when selecting an arbitrary prior to fit the elicited summaries. Second, specification error, which is the error in the elicited summaries. After some prior summaries are elicited, asking for more may reduce fitting error, but will certainly increase the specification error.

Asking the right questions.

- Identify the right summaries/aspects of prior knowledge.
- Ask questions that are clear, in terms which the subject understands.

Scalar elicitation. Some useful summaries may be:

- Plot/sketch (more useful after other summaries have been elicited)
- Shape (Unimodal? Symmetric?)
- Location (mean? mode? median?)
- Dispersion

- Probabilities (quantiles?)

It is important to ask questions in understandable terms. For example, assume that you want to elicit a distribution that describes your flatmate's uncertainty when it comes to the price of a fudge doughnut at Fisher and Donaldson. 'Tell me the median of θ?' is not the correct approach. Better ask, "Tell me what price is, in your opinion, equally likely to be above or below the true price of a fudge doughnut?"

Other elicitation issues.

- Psychologists have found that experts are usually overconfident when it comes to assessing their own uncertainty, i.e. they provide distributions with unreasonably small variances.
- After eliciting some summaries, feed back to the subject their consequences (usually after fitting a convenient distribution). Give the subject the chance to rethink.
- Overfitting. Elicit more summaries than necessary to check for consistency. For example, for a Normal density, elicit the mean, variance, and also additional quantiles.
- Given enough data, elicitation error will usually not matter. However, it is always good advise to perform some sensitivity analysis, trying different prior distributions which will be within the bounds of what we judge elicitation error to be, and check if the corresponding posteriors differ considerably. If not, then the posterior inference is insensitive (robust) to elicitation error.

Example

70 cancer patients are followed for 3 months after treatment. Doctors expect a 30% survival rate. They are 90% certain that the survival rate will be between 20% and 40%. After 3 months, 34 patients are alive. To Answer the following questions you could use the R function *SimpleR_ElicitingaBetapdf.R* available on Moodle or just pen and paper!

a) Derive a prior distribution in accordance with the beliefs of the doctor.

b) What should the posterior beliefs of the doctor be?

b) What should the posterior beliefs of the doctor be for a vague prior? How do the posteriors compare?

A solution without using R:

a) Using a Beta prior whose support is $[0, 1]$ and which is also conjugate to the Binomial likelihood could be a good choice. The parameters can be obtained using:

$$\mathrm{E}_{\mathrm{Beta}} = \frac{\alpha}{\alpha+\beta} = 0.3,$$

$$\mathrm{Sd}_{\mathrm{Beta}} = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2+(\alpha+\beta+1)}} = 0.06098,$$

where $0.06098 = \sigma$, the standard deviation obtained considering a Normal approximation ($0.3 \pm 1.64 * \sigma = [0.2, 0.4]$). Solving for α and β should result in a prior distribution for the survival rate represented by a Beta(16.65,38.82) whose quantiles for a symmetric 90% interval are $[0.204, 0.405]$ [look how a similar result is obtained using the R function provided!]

b) A conjugate Beta prior- Binomial Likelihood returns a posterior that is also a Beta distribution,

$$\text{Beta}(s+a, \quad n-s+b),$$

where s is the number of successes observed, n the number of trials, and a, b the parameters of the prior. Therefore, in our example the posterior is a $\text{Beta}(34+16.65, \quad 36+38.82) = \text{Beta}(50.65, \quad 74.82)$. You could use the function *SimpleR_BetaPrior_BinomialLikelihood.R* available on Moodle (Lecture 3) to see what this posterior looks like.

(c) A Beta(1,1) or a Beta(0.5,0.5) (Jeffrey's prior) could be suitable choices for non-informative/vaguely informative priors. You could use the function *SimpleR_BetaPrior_BinomialLikelihood.R* available on Moodle (Lecture 3) to see how the posterior changes when different priors are selected. This is called *sensitivity analysis* and it is essential that you carry it out in all your Bayesian analyses!

1.5 Summarising posteriors

The posterior distribution incorporates all the available information concerning the parameter of interest, and so is the most informative description of the parameter. Often the distribution is displayed graphically, in order to illustrate the information in a readily interpretable way. However, although the complete specification of the posterior pdf $\pi(\theta|\boldsymbol{x})$ is, for Bayesian statisticians, the desired end-product, it may be somewhat of a sophisticated concept for a non-mathematical client, for example. Then, for convenience, a more easily understandable *summary* of the posterior distribution may be desirable. For example, the posterior mean and standard deviation may be given. We have already looked at credible intervals. We will now look theoretically at a variety of point estimates that are often used.

1.5.1 Point estimates

There are a variety of different point estimates that are often used to describe the posterior distribution. Some of these have a decision theoretic interpretation. We shall consider three in more detail, which give an indication of the location ("average" value) of the distribution. However, first we shall recall some decision theory.

Suppose that we wish to estimate the parameter $\theta \in \Theta$. Then, we define a loss function $L(\theta, \hat{\theta})$ to be the associated loss for the estimate $\hat{\theta}$, when the true value is θ. The corresponding Bayes estimator is then chosen to minimise the expectation of the loss function with respect to the posterior distribution, i.e. the posterior expected loss. The corresponding estimate, $\hat{\theta}$, chosen under this rule is called the Bayes estimate. Mathematically, the Bayes estimate, $\hat{\theta}_B$ is defined such that,

$$\begin{aligned}\hat{\theta}_B &= \arg\min_{\hat{\theta}\in\Theta} \mathbb{E}_\pi[L(\theta,\hat{\theta})] \\ &= \arg\min_{\hat{\theta}\in\Theta} \left[\int_{\theta\in\Theta} L(\theta,\hat{\theta})\pi(\theta|\boldsymbol{x})d\theta\right].\end{aligned}$$

We now consider possible summary estimates of a posterior distribution in the context of loss functions. We shall consider three results (the first two of which are derived from the continuous case, the third for the discrete case), relating to point estimates of the posterior distribution.

Theorem 1.1: *The mean of the posterior distribution is the Bayes estimate with respect to the quadratic loss function.*

Proof: The quadratic loss function is given by,

$$L(\theta, a) = (\theta - a)^2.$$

Then, the corresponding Bayes estimate $\hat{\theta}_B$ is defined to be the estimate of θ such that it minimises the posterior expected loss, given by,

$$\begin{aligned}\mathbb{E}_\pi[L(\theta, a)] = \mathbb{E}_\pi[(\theta - a)^2] = \\ = \int_{\theta\in\Theta} (\theta - a)^2 \pi(\theta|\boldsymbol{x}) d\theta \\ = a^2 \int \pi(\theta|\boldsymbol{x}) d\theta - 2a \int \theta\pi(\theta|\boldsymbol{x}) d\theta + \int \theta^2 \pi(\theta|\boldsymbol{x}) d\theta = g(\theta, a).\end{aligned}$$

We differentiate with respect to a and equate to zero,

$$2a \int \pi(\theta|\boldsymbol{x}) d\theta - 2 \int \theta\pi(\theta|\boldsymbol{x}) d\theta = 0.$$

Then, we obtain,

$$\begin{aligned} a &= \frac{\int_{\theta\in\Theta} \theta\pi(\theta|\boldsymbol{x}) d\theta}{\int_{\theta\in\Theta} \pi(\theta|\boldsymbol{x}) d\theta} \\ &= \int_{\theta\in\Theta} \theta\pi(\theta|\boldsymbol{x}) d\theta, \end{aligned}$$

since $\int_{\theta\in\Theta} \pi(\theta|\boldsymbol{x}) d\theta = 1$, as $\pi(\theta|\boldsymbol{x})$ is a probability density function. To show that this is a minimum, taking the second derivative, we obtain,

$$\begin{aligned} \frac{d^2 g(\theta, a)}{da^2} &= 2 \int_{\theta\in\Theta} \pi(\theta|\boldsymbol{x}) d\theta \\ &= 2 \\ &> 0, \end{aligned}$$

and hence a minimum.

Thus, we have that,

$$\hat{\theta} = \int_{\theta\in\Theta} \theta\pi(\theta|\boldsymbol{x}) d\theta = \mathbb{E}_\pi(\theta),$$

i.e. the Bayes estimate is the posterior mean. □

Theorem 1.2: *The median of the posterior distribution is the Bayes estimate with respect to the absolute error loss function.*

Proof: The absolute error loss function is given by,

$$L(\theta, a) = |\theta - a|.$$

We shall consider the case where $\theta \in \mathbb{R}$. Then, we choose a, such that it minimises the posterior expected loss,

$$\begin{aligned} \mathbb{E}_\pi[L(\theta, a)] &= \int_{-\infty}^{\infty} |\theta - a| \pi(\theta|\boldsymbol{x}) d\theta = \int_{-\infty}^{a} (a - \theta)\pi(\theta|\boldsymbol{x}) d\theta + \int_{a}^{\infty} (\theta - a)\pi(\theta|\boldsymbol{x}) d\theta \\ &= a \int_{-\infty}^{a} \pi(\theta|\boldsymbol{x}) d\theta - a \left(1 - \int_{-\infty}^{a} \pi(\theta|\boldsymbol{x}) d\theta\right) - \int_{-\infty}^{a} \theta\pi(\theta|\boldsymbol{x}) d\theta + \int_{a}^{\infty} \theta\pi(\theta|\boldsymbol{x}) d\theta \\ &= 2a \int_{-\infty}^{a} \pi(\theta|\boldsymbol{x}) d\theta - 2 \int_{-\infty}^{a} \theta\pi(\theta|\boldsymbol{x}) d\theta - a + \mathbb{E}_\pi[\theta] \end{aligned}$$

By differentiating wrt to a we have

$$\frac{d}{da}\mathbb{E}_{\pi}[L(\theta,a)] \;=\; 2\int_{-\infty}^{a}\pi(\theta|\boldsymbol{x})d\theta+2a\pi(a|\boldsymbol{x})-2a\pi(a|\boldsymbol{x})-1=2\int_{-\infty}^{a}\pi(\theta|\boldsymbol{x})d\theta-1$$

and

$$\frac{d^2}{da^2}\mathbb{E}_{\pi}[L(\theta,a)]=2\pi(a|\boldsymbol{x})>0.$$

Here we used that

$$\frac{d}{da}\left(\int_{-\infty}^{a}\pi(\theta|\boldsymbol{x})d\theta\right)=\pi(a|\boldsymbol{x})$$

and

$$\frac{d}{da}\left(\int_{-\infty}^{a}\theta\pi(\theta|\boldsymbol{x})d\theta\right)=a\pi(a|\boldsymbol{x})$$

Hence, by taking the first derivative equal to 0 we have that the Bayes estimate $a=\hat{\theta}_B$ satisfies the equation

$$\int_{-\infty}^{a}\pi(\theta|\boldsymbol{x})d\theta=1/2$$

and thus it is the median of the posterior distribution, $\pi(\theta|\boldsymbol{x})$.

Hence,

$$\int_{-\infty}^{a}\pi(\theta|\boldsymbol{x})d\theta=\int_{a}^{\infty}\pi(\theta|\boldsymbol{x})d\theta,$$

that is,

$$2\int_{-\infty}^{a}\pi(\theta|\boldsymbol{x})d\theta=\int_{-\infty}^{\infty}\pi(\theta|\boldsymbol{x})d\theta=1.$$

Thus,

$$\int_{-\infty}^{a}\pi(\theta|\boldsymbol{x})=\frac{1}{2},$$

and so $a=\hat{\theta}_B$ is the median of the posterior distribution. □

Theorem 1.3: *The mode of the posterior distribution is the Bayes estimate with respect to the zero-one loss function.*

Proof: We consider the discrete case only. The loss function is of the form,

$$L(\theta,a)=\begin{cases}1 & \text{if } a\neq\theta;\\ 0 & \text{if } a=\theta,\end{cases}$$

and we wish to minimise,

$$\sum_{\theta\in\Theta}L(\theta,a)\pi(\theta|\boldsymbol{x}).$$

if we choose $a=\theta^*$, this expression becomes,

$$\sum_{\theta\in\Theta\backslash\theta^*}\pi(\theta|\boldsymbol{x})=1-\pi(\theta^*|\boldsymbol{x}).$$

Thus to minimise this, we simply maximise $\pi(\theta^*|\boldsymbol{x})$. In other words we should take $\hat{\theta}$ to be the "most likely" value of θ in the posterior distribution. □

Within our results, we shall usually examine the posterior mean of the distribution, taking this point estimate to be a "reasonable" point estimate. Whatever form of point estimate we use, however, it provides a very poor description of the complete posterior distribution. A more adequate summary of the latter, and yet still readily understandable, is the idea of an interval estimate.

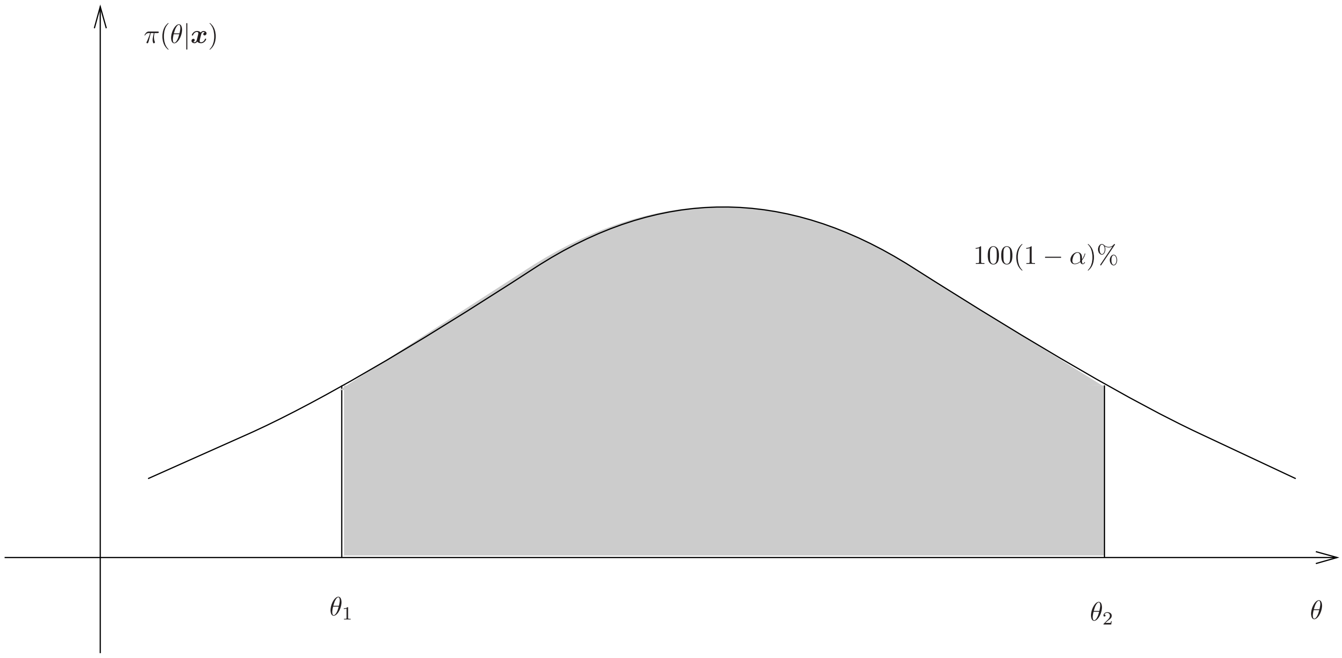


Fig. 1.5: Typical $100(1-\alpha)\%$ credible interval.

1.6 Interval estimates

Interval estimates give an estimate of the spread of the posterior distribution. These are analogous to confidence intervals within the classical case, and are called **credible intervals**. However, their interpretation is very different. A classical $100(1-\alpha)\%$ confidence interval is defined such that, if the data collection process is repeated again and again, then in the long run, $100(1-\alpha)\%$ of the confidence intervals formed are expected to contain the (fixed) unknown parameter value. Conversely, the interpretation of the Bayesian $100(1-\alpha)\%$ credible interval is that this interval contains $100(1-\alpha)\%$ of the posterior distribution of the parameter. More formally, suppose that we are interested in the parameter θ, which has posterior distribution $\pi(\theta|\boldsymbol{x})$.

Definition 1.2: *The interval (θ_1, θ_2) is defined as an $100(1-\alpha)\%$ credible interval if,*

$$\int_{\theta_1}^{\theta_2} \pi(\theta|\boldsymbol{x}) d\theta = 1 - \alpha, \qquad 0 \leq \alpha \leq 1.$$

Typical values of α are 0.1, 0.05, 0.01, and we speak of 90%, 95% and 99% credible intervals. The idea is illustrated in Figure 1.5.

Note that a $100(1-\alpha)\%$ credible interval is not unique, since, in general, there will be many choices of θ_1 and θ_2, such that, $\int_{\theta_1}^{\theta_2} \pi(\theta|\boldsymbol{x}) d\theta = 1 - \alpha$. For example, see Figure 1.6, where we have two $100(1-\alpha)\%$ credible intervals, $[\theta_1, theta_2]$ and $[\xi_1, \xi_2]$.

Often, a symmetric $100(1-\alpha)\%$ credible interval (θ_1, θ_2) is used. This credible interval is unique, and is defined such that,

$$\int_{-\infty}^{\theta_1} \pi(\theta|\boldsymbol{x}) d\theta = \frac{\alpha}{2} = \int_{\theta_2}^{\infty} \pi(\theta|\boldsymbol{x}) d\theta,$$

i.e. the credible interval such that θ_1 corresponds to the lower $\frac{\alpha}{2}$ quantile, and θ_2 to the upper $1 - \frac{\alpha}{2}$ quantile of the posterior distribution $\pi(\theta|\boldsymbol{x})$.

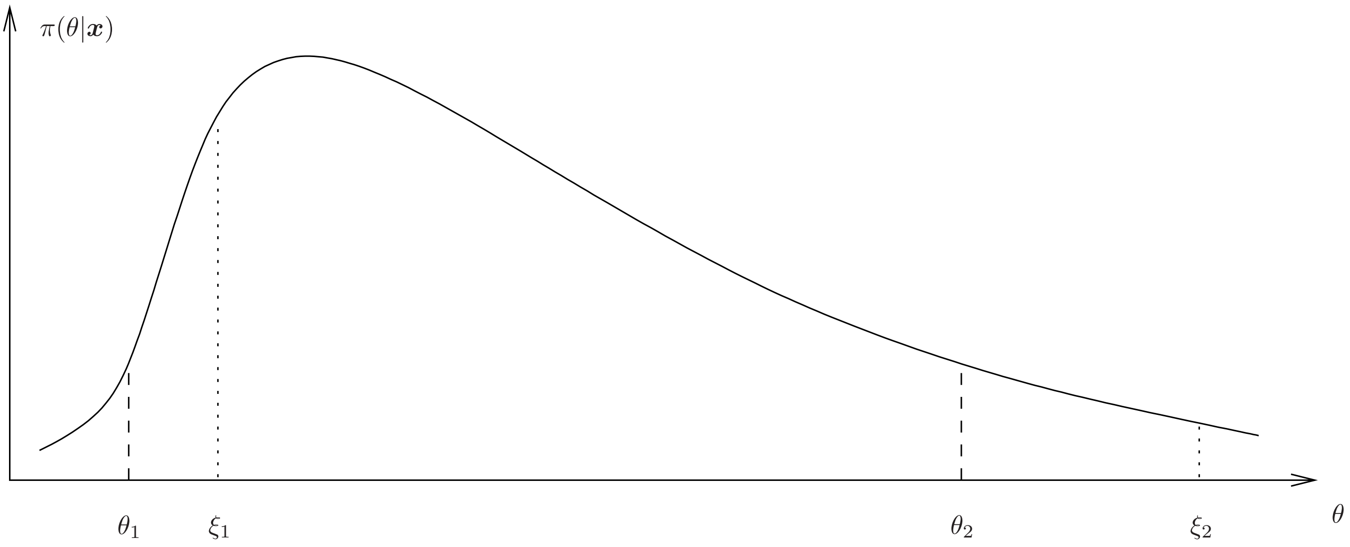


Fig. 1.6: Two alternative $100(1-\alpha)\%$ credible intervals.

Consider Figure 1.6, again. Both of the intervals contain $100(1-\alpha)\%$ of the distribution, so that, in betting terms, there is no objection to either of them. However, what about communicating information about θ? The interval $[\theta_1, \theta_2]$ is clearly more informative, since, for a given α, a shorter interval represents a "tighter" inference. This motivates the following refinement of a credible interval.

Definition 1.3: *The interval* $[\theta_1, \theta_2]$ *is a* $100(1-\alpha)\%$ ***highest posterior density interval*** *(HPDI) if:*

1. $[\theta_1, \theta_2]$ *is a* $100(1-\alpha)\%$ *credible interval; and*
2. *for all* $\theta' \in [\theta_1, \theta_2]$ *and* $\theta'' \notin [\theta_1, \theta_2]$, $\pi(\theta'|\boldsymbol{x}) \geq \pi(\theta''|\boldsymbol{x})$.

This is the required definition for the narrowest possible interval having a given credible level $1-\alpha$, and essentially centres the interval around the mode, in the uni-modal case. Clearly, if the distribution is symmetrical about the mean, such as the Normal distribution, the $100(1-\alpha)\%$ symmetric credible interval is identical to the $100(1-\alpha)\%$ HPDI. In the case where the posterior distribution is multi-modal, the corresponding HPDI may consist of several disjoint intervals.

Suppose that we have a symmetric credible interval $[\theta_1, \theta_2]$ for a given parameter θ. Then, if we consider a bijective (monotonic) transformation of the parameters, such as $g(\theta)$, the corresponding symmetric credible interval is given by $[g(\theta_1), g(\theta_2)]$, However, this is not always true for HPDI's. The corresponding HPDI on $g(\theta)$ is $[g(\theta_1), g(\theta_2)]$, if and only if, g is a linear transformation; else, the HPDI needs to be recalculated for $g(\theta)$.

In practice, for non-trivial problems, stochastic simulation is used for obtaining summary estimates of posterior distributions. If we are able to obtain a sample from the posterior distribution (using, for example, `R`), then we are able to use this sample to estimate corresponding summary statistics. For example, the mean of the posterior distribution can be estimated via the average of the corresponding sample from the posterior distribution (this is known as a Monte Carlo estimate). Any number of posterior summary statistics may be of interest. For example, suppose that we are interested in the posterior probability that the parameter of interest, θ has a value greater than 10. Then, we can estimate $\mathbb{P}_\pi(\theta > 10)$ by simply calculating the proportion of the sample from the posterior distribution for which the parameter value is greater than 10. Credible interval estimates are calculated in a similar fashion. These (and other) ideas will be discussed in greater detail in the latter half of the course. Here we are simply providing a sample of the types of summary estimates that we may be interested in, and giving a taster of how they may be calculated.

When describing a posterior distribution via summary statistics, a variety of point and interval estimates can be given to encapsulate the main properties of the distribution. In practice, the mean, median and 95% symmetric CI are the most common choices.

1.7 Prediction

Suppose that before we observe any data, we wish to make inference about a random vector $\boldsymbol{X}$, with pdf $f(\boldsymbol{x}|\theta)$, with unknown parameter $\theta \in \Theta$. If our uncertainty for θ is represented by $p(\theta)$, then we can derive the corresponding predictive pdf of $\boldsymbol{X}$ given by,

$$f(\boldsymbol{x}) = \int_{\theta\in\Theta} f(\boldsymbol{x},\theta)d\theta = \int_{\theta\in\Theta} f(\boldsymbol{x}|\theta)p(\theta)d\theta.$$

Thus, we take the expected pdf, where the expectation is with respect to the prior density $p(\theta)$. The term $f(\boldsymbol{x})$ is called the *prior predictive distribution.* (Recall that this is also the denominator in the expression for Bayes' Theorem). Essentially, we are weighting the pdf $f(\boldsymbol{x}|\theta)$ with the best description of our beliefs for θ, and since we have not observed any data, that is the prior distribution for θ.

Example

Suppose that X is a random variable, such that,

$$X \sim Exp(\lambda), \quad \lambda > 0,$$

and that our prior beliefs on λ are described by $\Gamma(\alpha,\beta)$ distribution ($\alpha,\beta > 0$). Then, the prior predictive distribution is given by,

$$\begin{aligned} f(x) &= \int_0^\infty \lambda\exp(-\lambda x) \times \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}\exp(-\beta\lambda)d\lambda \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+1)}{(\beta+x)^{\alpha+1}}\int_0^\infty \frac{(\beta+x)^{\alpha+1}}{\Gamma(\alpha+1)}\lambda^\alpha\exp(-\lambda(x+\beta))d\lambda \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+1)}{(\beta+x)^{\alpha+1}}, \end{aligned}$$

since the integral is over a $\Gamma(\alpha+1,\beta+x)$ pdf, and integrates to one.

Suppose that we now observe data $\boldsymbol{x}$, and wish to predict future observations $\boldsymbol{y}$, from the same process, assuming that conditional on the parameter θ in the process, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent. Then, the *posterior predictive distribution* for $\boldsymbol{Y}$ is given by,

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{x}) &= \int_\Theta f(\boldsymbol{y},\theta|\boldsymbol{x})d\theta \\ &= \int_\Theta f(\boldsymbol{y}|\boldsymbol{x},\theta)\pi(\theta|\boldsymbol{x})d\theta \\ &= \int_\Theta f(\boldsymbol{y}|\theta)\pi(\theta|\boldsymbol{x})d\theta, \end{aligned}$$

where, here we assume that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are conditionally independent, given θ. Thus, we are now weighting the corresponding pdf for $\boldsymbol{Y}$ with our current (posterior) beliefs for θ having already observed data $\boldsymbol{x}$.

Example

Suppose that the number of calls X to a telephone switchboard in z minutes has a $Poisson(\lambda z/10)$ distribution, where $\lambda > 0$ is unknown. (So, for a period of 10 minutes, $X \sim Poisson(\lambda)$) Being an enthusiastic Bayesian research graduate, the operator forms the following prior on λ, (from working in a similar telephone switchboard),

$$\lambda \sim Exp(10).$$

Being reasonably bored, they decide to calculate their prior predictive distribution for the number of calls that they receive in the first 10 minutes of work, denoted by X.

Then, the prior predictive distribution is given by,

$$\begin{aligned}
P(X = x) = f(x) &= \int_0^\infty f(x|\lambda)p(\lambda)d\lambda \\
&= \int_0^\infty \frac{\lambda^x}{x!}\exp(-\lambda) \times 10\exp(-10\lambda)d\lambda \\
&= \frac{10}{x!}\frac{\Gamma(x+1)}{11^{x+1}}\int_0^\infty \frac{11^{x+1}}{\Gamma(x+1)}\lambda^x\exp(-11\lambda)d\lambda \\
&= \frac{10}{x!}\frac{\Gamma(x+1)}{11^{x+1}} \\
&= \frac{10}{11^{x+1}},
\end{aligned}$$

since $\Gamma(x+1) = x!$, as x is a positive integer.

The operator then takes x calls within the first 10 minutes, and wants a 20 minute break. They decide to calculate the posterior predictive distribution for the number of calls that they will receive in this time, from which they can calculate the probability of no calls being missed in this time, i.e. receiving no calls in these twenty minutes.

Answer:

First they calculate the corresponding posterior distribution for λ, given that they have received a total of x calls in the first 10 minutes,

$$\begin{aligned}
\pi(\lambda|x) &\propto f(x|\lambda)p(\lambda) \\
&= \frac{\lambda^x}{x!}\exp(-\lambda) \times 10\exp(-10\lambda) \\
&\propto \lambda^x\exp(-11\lambda),
\end{aligned}$$

so that,

$$\lambda|x \sim \Gamma(x+1, 11).$$

Let Y denote the number of calls they receive in 20 minutes. Then,

$$Y|\lambda \sim Poisson(2\lambda).$$

The corresponding posterior predictive distribution is given by,

$$\begin{aligned}
P(Y = y|x) = f(y|x) &= \int_0^\infty f(y|\lambda)\pi(\lambda|x)d\lambda \\
&= \int_0^\infty \frac{(2\lambda)^y\exp(-2\lambda)}{y!} \times \frac{11^{x+1}}{\Gamma(x+1)}\lambda^x\exp(-11\lambda)d\lambda \\
&= \frac{11^{x+1}}{13^{x+y+1}}\frac{2^y}{y!}\frac{\Gamma(x+y+1)}{\Gamma(x+1)}\int_0^\infty \frac{13^{x+y+1}}{\Gamma(x+y+1)}\lambda^{x+y}\exp(-13\lambda)d\lambda \\
&= \frac{11^{x+1}}{13^{x+y+1}}\frac{2^y}{y!}\frac{\Gamma(x+y+1)}{\Gamma(x+1)}.
\end{aligned}$$

So that the posterior predictive probability of no calls within the next 20 minutes is,

$$P(Y=0|x)=f(0|x)=\left(\frac{11}{13}\right)^{x+1}.$$

Then, for example, if they have not had more than 3 calls in the last 10 minutes, their posterior predictive probability of not missing a call in their break exceeds 1/2.

1.8 Bayes' Theorem (multivariate)

Bayes' Theorem is easily generalized to the multi-parameter case. Suppose that we have a set of parameters $\boldsymbol{\theta}$ on which we wish to make inference, and that we observe data $\boldsymbol{x}$. Then the (joint) posterior distribution for $\boldsymbol{\theta}$ is given by,

$$\pi(\boldsymbol{\theta}|\boldsymbol{x})=\frac{f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{f(\boldsymbol{x})},$$

where,

$$f(\boldsymbol{x})=\int f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

Since this integration becomes increasingly more complex in higher dimensions, Bayes' Theorem is most often quoted in the form,

$$\pi(\boldsymbol{\theta}|\boldsymbol{x})\propto f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

The prior is then specified jointly over all parameters. Often the parameters $\boldsymbol{\theta}=\{\theta_1,\dots,\theta_n\}$ are assumed to be independent of each other, *a priori*, so that, $p(\boldsymbol{\theta})=\prod_{i=1}^n p(\theta_i)$. It is also common to specify the prior by introducing another hyper-parameter ϕ, so that,

$$p(\boldsymbol{\theta}|\phi)=\prod_{i=1}^n p(\theta_i|\phi), \qquad p(\phi)=\dots.$$

In multi-dimensions the posterior distribution significantly increases in complexity. However, often we may only be interested in the marginal distribution of a single parameter conditional on the data. For example, suppose that we are only interested in θ_1, then,

$$\pi(\theta_1|\boldsymbol{x})=\int \pi(\boldsymbol{\theta}|\boldsymbol{x})d\theta_2,\dots,d\theta_n.$$

This integration is often too complex to do in practice, however, the latter part of the course will show how we may be able to obtain summary statistics of a marginal posterior distribution, such as this, using an alternative method called Markov chain Monte Carlo (MCMC).

Note that the ideas presented for the single parameter case are all directly generalized to the multi-parameter case, for example, the concepts of sufficiency and conjugacy. Posterior credible intervals, however, now generalize to posterior credible regions with dimension equal to the number of parameters. For example, in the two parameter case, where $\boldsymbol{\theta}=\{\theta_1,\theta_2\}$, the $100(1-\alpha)\%$ posterior credible interval is now a two-dimensional region, R, such that,

$$\mathbb{P}((\theta_1,\theta_2)\in R|\boldsymbol{x})=1-\alpha.$$

Clearly, a computer is needed in order to plot these more complex regions. Often, in practice, the marginal posterior density intervals may be calculated for a single parameter, rather than the more complex higher-dimensional density regions for the full set of parameters.

1.9 Additional Examples: Further Bayesian Inference For Normal Distributions

Within this section, we shall assume that we observe data $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, such that, given μ and σ, $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, distribution.

1.9.1 Single parameter problems

Known μ, unknown σ^2

As usual, we need to place a prior on the unknown parameter σ^2. We specify,

$$\sigma^2 \sim \Gamma^{-1}(\alpha, \beta).$$

In other words $1/\sigma^2 \sim \Gamma(\alpha, \beta)$. Then, the corresponding posterior distribution is given by,

$$\begin{aligned}
\pi(\sigma^2|\boldsymbol{x}) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \times (\sigma^2)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right) \\
&\propto (\sigma^2)^{-(n/2+\alpha+1)} \exp\left(-\frac{\left(\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2+\beta\right)}{\sigma^2}\right) \\
\Rightarrow \sigma^2|\boldsymbol{x} &\sim \Gamma^{-1}\left(\frac{n}{2}+\alpha, \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2+\beta\right).
\end{aligned}$$

Consider again the posterior pdf for σ^2:

$$\pi(\sigma^2|\boldsymbol{x}) \propto (\sigma^2)^{-(n/2+\alpha+1)} \exp\left(-\frac{\left(\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2+\beta\right)}{\sigma^2}\right).$$

Writing $z^2 = \sum_{i=1}^{n}(x_i-\mu)^2$, we note that z^2 is simply a constant, so that,

$$\pi(\sigma^2|\boldsymbol{x}) \quad \propto \quad (\sigma^2)^{-(n/2+\alpha+1)} \exp\left(-\frac{z^2+2\beta}{2\sigma^2}\right)$$

Then, we can see that setting $a = \frac{n+2\alpha}{2}$ and $b = \frac{z^2+2\beta}{2}$ we have that,

$$\sigma^2|\boldsymbol{x} \sim \Gamma^{-1}(a, b) = \Gamma^{-1}\left(\frac{n+2\alpha}{2}, \frac{z^2+2\beta}{2}\right).$$

Example

Suppose that we observe data $\boldsymbol{x}$, for $n = 10$, given by,

$$2.1,\ 4.2,\ 6,\ 4.5,\ 3.5,\ 2.1,\ 4.4,\ 3.2,\ 3.7,\ 3.9$$

where $\mu = 4$. We place a $\Gamma^{-1}(0.5, 0.5)$ prior on σ^2.

Then, what is the corresponding distribution for σ^2? We have that,

$$z^2 = \sum_{i=1}^{n}(x_i-\mu)^2 = 12.66.$$

So that,

$$\sigma^2|\boldsymbol{x} \sim \Gamma^{-1}(5.5, 6.83).$$

Thus, the posterior mean for σ^2 is 1.518, with standard deviation 0.82.

The distribution is skewed, and can be see in Figure 1.7, which plots the corresponding posterior distribution for σ^2. One can also use R to derive credible intervals, working directly with the Inverse Gamma density. For example, the 95% symmetric CI is, (0.62,3.59).

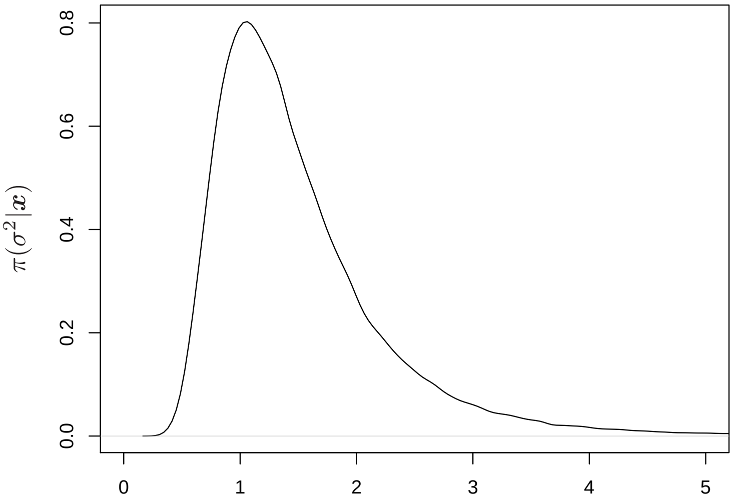


Fig. 1.7: Posterior distribution for σ^2.

Alternative parameterization

Suppose that the parameter of interest is the precision $\tau = 1/\sigma^2$ and that we place a $\Gamma(\alpha, \beta)$ prior on τ. This is the same as placing a $\Gamma^{-1}(\alpha, \beta)$ distribution on σ^2, i.e. the same prior as in the above example. This can be easily seen via a transformation of variables. Suppose that $\sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$, and $\tau = 1/\sigma^2$. Then, using a transformation of variables, the corresponding distribution on τ, is given by,

$$\begin{aligned} p_\tau(\tau) &= p_{\sigma^2}(\sigma^2(\tau)) \left| \frac{d\sigma^2}{d\tau} \right| \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\tau}\right)^{-(\alpha+1)} \exp\left(-\beta\tau\right) \frac{1}{\tau^2} \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} \exp\left(-\beta\tau\right), \end{aligned}$$

so that, $\tau \sim \Gamma(\alpha, \beta)$.

Then, since the posterior distribution for σ^2 is also Γ^{-1}, the corresponding posterior for τ is Γ, and in particular,

$$\tau|\boldsymbol{x} \sim \Gamma\left(\frac{n}{2} + \alpha, \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 + \beta\right)$$

This can be shown in the usual way, by combining the likelihood with the corresponding Gamma prior on τ, and is left as an exercise - see Tutorial 2.

For the data given above, we have that,

$$\tau^2|\boldsymbol{x} \sim \Gamma(5.5, 6.83),$$

so that the posterior mean of τ^2 is 0.805, with standard deviation 0.34. The 95% symmetric credible interval is (0.28,1.61). Note that the 95% symmetric credible interval for τ^2 is equal to the reciprocal of the 95% symmetric credible interval for $\sigma^2 = 1/\tau^2$. However, this is not true for all posterior summary statistics, for example, the mean, variance, 95% HPDI etc.

1.9.2 Multi-parameter problem

Multivariate Normal

Suppose that we observe random variables $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ from a Multivariate Normal distribution with known (symmetric) covariance matrix Σ and unknown mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$, which we wish to estimate. Then, we write $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$. (Note that if $\Sigma = \sigma^2 I$, then the random variables, $X_1, \ldots, X_p$ are independent Normal random variables, with mean $\mu_1, \ldots, \mu_p$, respectively).The pdf for $\boldsymbol{X}$, given parameters $\boldsymbol{\mu}$, is defined to be,

$$f(\boldsymbol{x}|\boldsymbol{\mu}) = \frac{1}{\sqrt{2\pi}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

We specify the prior,

$$\boldsymbol{\mu} \sim \mathcal{N}_p(\boldsymbol{\theta}, \Sigma_p).$$

Then, the corresponding posterior distribution for $\boldsymbol{\mu}$ is given by,

$$\begin{aligned}
\pi(\boldsymbol{\mu}|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\boldsymbol{\mu})p(\boldsymbol{\mu}) \\
&= \frac{1}{\sqrt{2\pi}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \times \frac{1}{\sqrt{2\pi}|\Sigma_p|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\theta})^T \Sigma_p^{-1}(\boldsymbol{\mu}-\boldsymbol{\theta})\right) \\
&\propto \exp\left(-\frac{1}{2}[\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{x} - \boldsymbol{x}^T\Sigma^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\Sigma_p^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^T\Sigma_p^{-1}\boldsymbol{\theta} - \boldsymbol{\theta}^T\Sigma_p^{-1}\boldsymbol{\mu}]\right) \\
&= \exp\left(-\frac{1}{2}[\boldsymbol{\mu}^T(\Sigma^{-1}+\Sigma_p^{-1})\boldsymbol{\mu} - \boldsymbol{\mu}^T(\Sigma^{-1}\boldsymbol{x}+\Sigma_p^{-1}\boldsymbol{\theta}) - (\boldsymbol{x}^T\Sigma^{-1}+\boldsymbol{\theta}^T\Sigma_p^{-1})\boldsymbol{\mu}]\right) \\
&\propto \exp\left(\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_\pi)^T\Sigma_\pi^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_\pi)\right),
\end{aligned}$$

where,

$$\boldsymbol{\mu}_\pi = \Sigma_\pi(\Sigma^{-1}\boldsymbol{x} + \Sigma_p^{-1}\boldsymbol{\theta}); \quad \text{and} \quad \Sigma_\pi^{-1} = \Sigma^{-1} + \Sigma_p^{-1}.$$

Thus,

$$\boldsymbol{\mu}|\boldsymbol{x} \sim N(\boldsymbol{\mu}_\pi, \boldsymbol{\Sigma}_\pi).$$

The Multivariate Normal prior on the mean vector $\boldsymbol{\mu}$ is a conjugate prior.

1.10 Motivating Examples

Throughout this section we have generally considered conjugate priors and particular vague priors. These have all resulted in standard posterior distributions for the parameters of interest. However, what happens if we wish to use a different prior for a given parameter(s), resulting in a more complex posterior distribution, or where the likelihood is complex, and there is no prior of standard form that results in a standard posterior distribution for the parameters. Consider a simple example. Suppose that we have a random variables, $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ which are independent and Normally distributed with mean μ and variance σ^2 (assumed to be known). Then, we may wish to specify a prior on μ of the form,

$$\mu \sim \log N(\phi, \tau^2).$$

(This means that $\log \mu \sim N(\phi, \tau^2)$).

The corresponding posterior distribution for μ is given by,

$$\pi(\mu|\boldsymbol{x}) \propto \exp\left(-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right) \frac{1}{\mu} \exp\left(-\frac{(\log \mu - \phi)^2}{2\tau^2}\right).$$

This posterior distribution is clearly non-standard, so that alternative (numerical) approaches need to be implemented in order to obtain any inference on the parameter μ. For example, in this case the distribution can be plotted within `R`. However, in general, as the number of dimensions increases the visual representation of the posterior distribution becomes increasingly difficult. Calculating a posterior marginal distribution of a particular parameter is also often difficult due to the complex integration needed (which may be analytically intractable), resulting in the need for approximations. As the number of dimensions increases, alternative approaches need to be considered.

Consider a relatively simple two-dimensional case, where we have data $x_1, \ldots, x_n$ such that each X_i are independent and Normally distributed with mean μ and (unknown) variance σ^2. We specify independent priors:,

$$\mu \sim N(\phi, \tau^2); \qquad \text{and} \qquad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta).$$

The corresponding joint posterior distribution is of the form,

$$\begin{aligned}\pi(\mu, \sigma^2|\boldsymbol{x}) &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \times \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(\mu - \phi)^2}{2\tau^2}\right) \times \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-(\alpha+1)} \exp\left(-\frac{\beta}{\sigma^2}\right) \\ &\propto (\sigma^2)^{-(n/2+\alpha+1)} \exp\left(-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right) \exp\left(-\frac{\beta}{\sigma^2}\right) \exp\left(-\frac{(\mu - \phi)^2}{2\tau^2}\right),\end{aligned}$$

after algebra. We note that the posterior conditional distributions for μ and σ^2 are of standard form, namely,

$$\begin{aligned}\mu|\boldsymbol{x}, \sigma^2 &\sim N\left(\frac{\tau^2 n\bar{x} + \sigma^2\phi}{\tau^2 n + \sigma^2}, \frac{\sigma^2\tau^2}{\tau^2 n + \sigma^2}\right); \\ \sigma^2|\boldsymbol{x}, \mu &\sim \Gamma^{-1}\left(\frac{n}{2} + \alpha, \frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2 + \beta\right),\end{aligned}$$

(see results on the Normal distribution derived earlier.) Thus, although the joint posterior distribution is of a complex form, the (uni-dimensional) posterior conditional distributions of the individual parameters are of standard form. So, what would happen if we repeatedly simulate values for μ and σ^2 from the posterior conditional distributions of the parameters? We will answer this in the following section of the course where we utilise this kind of property in obtaining inference on more complex posterior distributions.

1.11 Monte Carlo Integration

In many circumstances, we are faced with the problem of evaluating an integral that is too complex to calculate explicitly. With particular reference to Bayesian inference, posterior distributions are typically summarised using statistics such as the mean and/or variance. Such posterior summary statistics require integration of the posterior density (which is often analytically intractable, particularly for high-dimensional posterior distributions). For example, we may wish to estimate the posterior (marginal) expectation of a parameter θ, given observed data $\boldsymbol{x}$:

$$\mathbb{E}_{\pi}(\theta) = \int \theta \pi(\theta|\boldsymbol{x}) d\theta.$$

We can use the simulation technique of *Monte Carlo integration* to obtain an estimate of a given integral (and hence posterior expected value). The method is based upon drawing observations from the distribution of the variable of interest and simply calculating the empirical estimate of the expectation. For example, given a sample of observations, $\theta^1, \ldots, \theta^n \sim \pi(\theta|\boldsymbol{x})$, we can estimate the expectation by,

$$\frac{1}{n}\sum_{i=1}^{n} \theta^i.$$

For independent samples, the Law of Large Numbers ensures that

$$\frac{1}{n}\sum_{i=1}^{n} \theta^i \to \mathbb{E}_{\pi}(\theta) \quad \text{as } n \to \infty.$$

Independent sampling from $\pi(\theta|\boldsymbol{x})$ may be difficult, however this result still holds if we generate our samples, not independently, but via some other method (although this may be less effective than independently drawn samples in that larger sample sizes are needed to obtain the same level of accuracy).

In other words we estimate the posterior mean by the sample mean of observations taken from the posterior distribution. This is *Monte Carlo integration.* The idea extends directly to any function of θ, denoted by $f(\theta)$. For example, suppose that we wish to calculate the posterior mean of $f(\theta)$. Given a sample of observations, $\theta^1, \ldots, \theta^n \sim \pi(\theta|\boldsymbol{x})$, we can estimate the posterior mean of $f(\theta)$ by

$$\frac{1}{n}\sum_{i=1}^{n} f(\theta^i).$$

Similarly, we estimate the posterior variance, $\mathrm{Var}_{\pi}(\theta)$, by the sample variance of observations taken from the posterior distribution, i.e.

$$\frac{1}{n-1}\left[\sum_{i=1}^{n}(\theta^i)^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\theta^i\right)^2\right].$$

Finally we note that we can obtain (marginal) density plots of the parameters of interest by using standard software. For example, suppose that in `R` the sample values of the parameters are stored in the vector `theta`, we can obtain a density plot using the command `plot(density(theta),type="l")`.

Example

Suppose that $\theta \sim N(0,1)$. Directly we have that $\mathbb{E}(\theta) = 0$ and $\mathrm{Var}(\theta) = 1$. However, for the purposes of illustration we will use Monte Carlo integration to estimate these posterior summary statistics. We will use the function `rnorm` in `R` to independently simulate observations from this distribution and calculate the sample mean and standard deviation (SD) for different numbers of random deviates simulated (i.e. different values of n). We repeat this 3 times for each value of n. The corresponding sample means and variances I obtained were:

Repetition number	1		2		3	
n	Mean	SD	Mean	SD	Mean	SD
10	-0.39	1.26	-0.36	1.32	-0.09	0.87
100	0.36	0.88	-0.23	0.94	0.10	0.95
1000	-0.10	1.01	-0.01	1.01	0.02	0.98
10000	-0.01	1.00	0.00	0.99	0.00	1.00
100000	0.00	1.00	0.00	1.00	0.00	1.00

The above table illustrates the idea of *Monte Carlo error.* Monte Carlo error essentially measures the variation we would expect to see if multiple replications of the experiment are conducted. Monte Carlo error decreases with increasing sample size (i.e. n).Using the central limit theorem, we have the result that for θ values drawn independently from the distribution π,

$$\overline{\theta} \sim N\left(\mathbb{E}_\pi(\theta), \frac{\sigma^2}{n}\right).$$

The term σ^2/n is the Monte Carlo variance, or more commonly, $\sqrt{\sigma^2/n}$ is the Monte Carlo error. So, the "true" Monte Carlo errors given sample sizes of 10, 100, 1000, 10000 and 100000 are 0.31, 0.1, 0.031, 0.01 and 0.0031. However, we don't typically know the variance – we estimate it from our sample (as SD^2). For the above simulations, I obtained estimated Monte Carlo errors of 0.39, 0.088, 0.032, 0.01 and 0.0032 for the increasing values of n for repetition 1.

The above example concentrates on obtaining estimates of the posterior mean (and variance) of a distribution. However, any other posterior summary statistics may be of interest and they can all be approximated using the Monte Carlo samples (See Tutorial 4).

Thus, we can replace the integration problem by a sampling problem, but this creates two new problems. In general, $\pi(\theta|\boldsymbol{x})$ represents a high-dimensional and complex distribution from which samples would usually be difficult to obtain. In addition, large sample sizes are often required and so powerful computers are needed to generate these samples. So, how do we obtain a potentially large sample from the posterior distribution, when in general, this will be very complex and often high-dimensional?

1.12 Markov chain Monte Carlo

1.12.1 Basic Idea

A Markov chain is simply a stochastic sequence of numbers where each value in the sequence depends *only* upon the last. In other words suppose we have a sequence of numbers $\theta^0, \theta^1, \theta^2, \ldots, \theta^n$ then θ^1 is only a function of θ^0; θ^2 is only a function of θ^1; $\ldots$; θ^n is only a function of θ^{n-1}. We let θ^0 be chosen to be equal to some arbitrary value. Thus, we can simulate a Markov chain by generating a new state of the chain, say θ^{n+1}, from some distribution, dependent only on θ^n:

$$\theta^{n+1} \sim \mathcal{P}(\theta^n, \theta^{n+1}) \quad \left(\equiv \mathcal{P}(\theta^{n+1}|\theta^n)\right).$$

with corresponding density $\equiv \mathcal{P}(\theta^{n+1}|\theta^n)$. We call $\mathcal{P}$ the transition kernel for the chain. The transition kernel (or density) uniquely describes the dynamics of the chain.

Under certain conditions (that the chain is aperiodic and irreducible) the distribution over the states of the Markov chain will converge to a *stationary* distribution (in this course we shall always assume that these conditions are met). The stationary distribution is *independent* of the initial starting values specified for the chains. Our aim is to construct a Markov chain such that the stationary distribution is equal to the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$. If we can do this (which we obviously can or I would not be describing such a situation!), we can run the Markov chain until the stationary distribution is reached, so that realisations of the chain can be regarded as a *dependent* sample from the posterior distribution of interest (thus note that we need to discard the first portion of a Markov chain). We are then able to use this sample from the latter part of the chain, after it has converged, to obtain Monte Carlo estimates of the parameters of interest and/or plot their corresponding density function (see §1.11). Since we are combining a Markov chain with Monte Carlo integration this method is called Markov chain Monte Carlo (MCMC). The beauty of MCMC (as we shall see) is that the updating of the states in the Markov chain remains relatively simple, using standard techniques, irrespective of the complexity of the posterior distribution.

Clearly several questions immediately arise for such a technique:

1. How do we construct such a Markov chain?
2. Even if we can construct such a Markov chain with the correct stationary distribution, how long do we need to run the chain until the stationary distribution of interest has been reached?
3. How many samples do we need from the posterior distribution so that we accurately estimate the quantites of interest? (this should have already occurred to you for direct sampling methods!).

We will consider the latter two questions, before focussing on the first question (which clearly we need to know the answer to in order to use this method!).

1.12.2 Run Lengths

There are two issues to be considered when determining the number of iterations of the Markov chain (i.e. how many values to simulate):

(i) the time required for the Markov chain to reach the stationary distribution (i.e. for the chain to converge), and

(ii) the post-convergence sample size required for suitably small Monte Carlo errors.

Issue (i): Burn-in

We need to discard the realisations of the Markov chain before the chain has converged to the stationary distribution. The initial observations that we discard are referred to as the *burn-in*. The simplest method to determine the length of the burn-in period is to look at trace plots - these are simply the value of the parameter at each iteration of the Markov chain. It is often possible to see the individual parameters converging from their starting position to values based around a constant mean (i.e. the mean of the posterior distribution). For example, consider Figure 1.8, clearly the earliest values of the Markov chain do not look like the later values - these early values are dependent on the starting value, and hence would be discarded as burn-in. By eye we might suggest a suitable burn-in of around 200 iterations. Note it is always best to be conservative with regard to the burn-in and err on the side of overestimation to ensure that convergence has been achieved when obtaining the sample to be used to form Monte Carlo estimates of the parameters of interest.

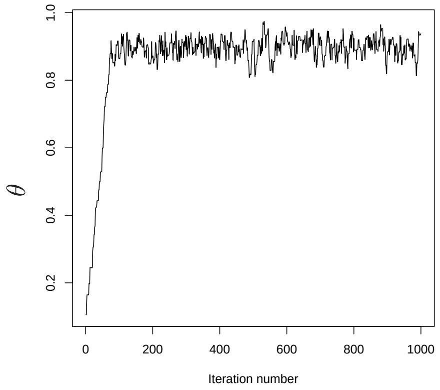


Fig. 1.8: A single MCMC trace plot.

This use of a trace plot is often a fairly efficient method, but it is not robust. For example, an ad hoc interpretation of the first trace plot in Figure 1.9 might suggest that the chain had converged after around 500 iterations. However, when the chain is run for longer, it is clear that the chain has not converged within these first 500 iterations.

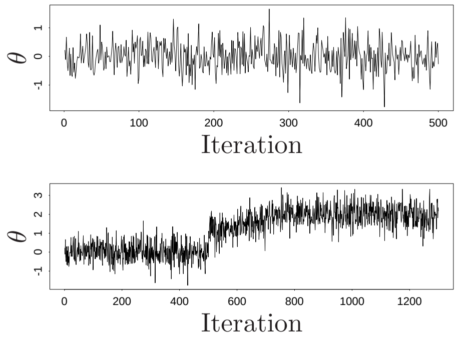


Fig. 1.9: MCMC sample paths.

Another early technique (sometimes referred to as the "thick-pen" technique) involves running two (or more chains) started at very different starting values and plotting the output on a single graph. A "thick pen" is then taken and run over either of the trace plots from one of the simulated Markov chains. When the pen touches both lines of the plot it could be concluded that the chains had converged. For example, consider Figure 1.10 where we run two chains and plot the values from each Markov chain (ie. trace plot) on the same axes. Using this technique we might once again suggest a burn-in of at least 200.

This idea motivated many of the more formal (and mathematical) techniques for assessing convergence to the stationary distribution via the assessment of multiple replications starting from over-dispersed starting points. Essentially, this means running the Markov chains several times (from different starting points) and checking that given a suitable burn-in period the posterior estimates of all of the chains are essentially the same, providing evidence that no major nodes have been missed. The most common approach is the Brooks-Gelman-Rubin (BGR) method. There are various im-

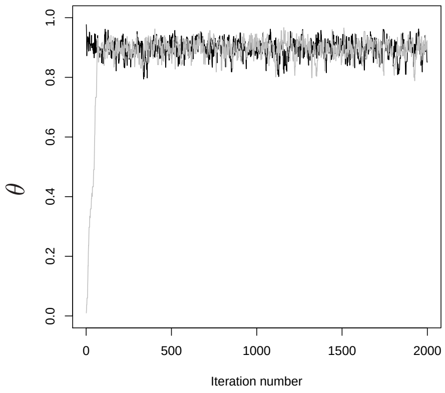


Fig. 1.10: Two MCMC trace plots from independent chains starting at different values

plementations of this diagnostic procedure, all based upon the idea of using an analysis of variance technique to check whether there are any differences in the posterior estimates obtained from the different replications. In order to implement this procedure at least two chains need to be simulated. The simplest implementation for a chain containing $2n$ iterations is to discard the first n iterations and take the ratio of the width of the empirical 80% credible interval obtained from all chains combined after the burn-in, with the corresponding mean within-chain 80% interval width, i.e., set

$$\hat{R} = \frac{\text{width of 80\% credible interval of pooled chains}}{\text{mean of width of 80\% credible interval of individual chains}}.$$

Convergence is assumed when these are roughly equal, implying that all chains have roughly equal variability, so that the $\hat{R} \approx 1$. The $\hat{R}$ value is plotted in Figure 1.11 for the two chains presented in Figure 1.10 for increasing value of n. Looking at this plot (and being conservative) we might suggest that convergence is achieved by around iteration 1000.[1]

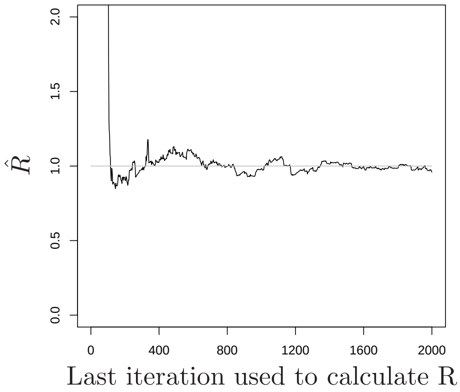


Fig. 1.11: BGR statistic for the trace plots provided in Figure 1.10.

[1] Another useful thing to check, not shown here is that the within chain and between chain variabilities are not changing as number of iterations increase.

Alternative techniques compare within-chain and between-chain variances rather than interval widths, but the principle remains the same. Note however that no convergence diagnostic can prove that the chain has converged, they can only identify when the chain has not converged. For example, consider the trace plot in Figure 1.12(a). This chain appears to converge very quickly, but the posterior distribution for this distribution is given in Figure 1.12(b). Starting several Markov chains in different starting values would quickly identify the bimodality in this example (but it can be more difficult to identify in high dimensional space).

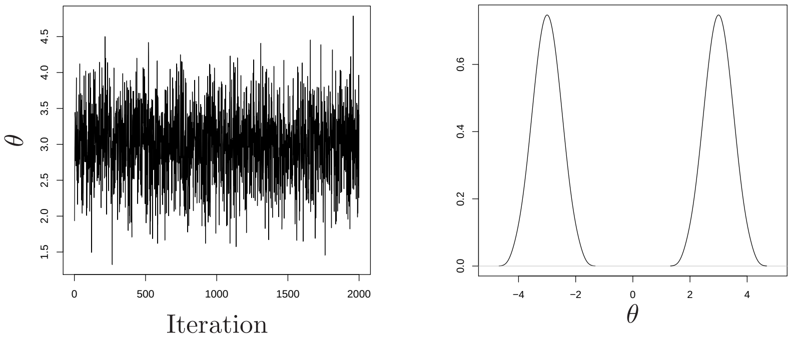


Fig. 1.12: (a) MCMC trace plot and (b) the underlying distribution.

Issue (ii): Monte Carlo error

To consider the issue of the number of iterations we need following the burn-in to obtain an accurate estimate of the summary statistics, we once more return to the idea of Monte Carlo error (see §1.11). Recall that the sample mean, $\overline{\theta} = \frac{1}{n}\sum_{i=1}^{n}\theta^i$ satisfies,

$$\overline{\theta} \sim N\left(\mathbb{E}_\pi(\theta), \frac{\sigma^2}{n}\right),$$

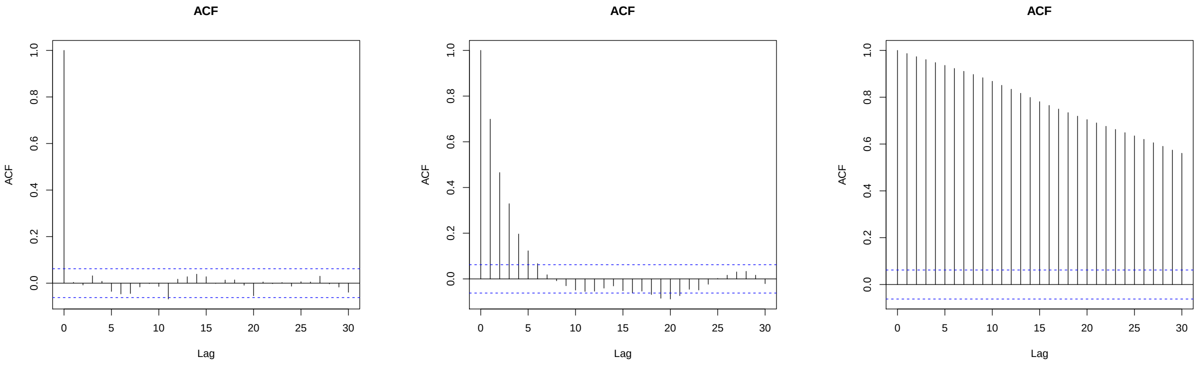
where $\frac{\sigma}{\sqrt{n}}$ is the Monte Carlo error. However, we do not know σ^2. This is most commonly estimated using *batching*. This involves dividing the chain into m distinct batches each of length T, so that $n = mT$ and where it is assumed that T is "large" leading to reasonably reliable sample mean estimates for each batch. Let $\overline{\theta}_1, \ldots, \overline{\theta}_m$ denote the sample means for each batch and $\overline{\theta}$ denote the mean over all n samples. The batch means estimate of σ^2 is given by,

$$\hat{\sigma}^2 = \frac{T}{m-1}\sum_{i=1}^{m}(\overline{\theta}_i - \overline{\theta})^2.$$

Thus an estimate of the Monte Carlo error is,

$$\sqrt{\frac{\hat{\sigma}^2}{n}}.$$

The performance of the Markov chain, in terms of exploring the parameter space and hence level of Monte Carlo error is often initially performed by eye from trace plots. Chains that quickly explore the full range of plausible parameter values will have lower Monte Carlo error than chains that only slowly move over the set of plausible values. This leads to assessing the performance of the Markov chain via the *autocorrelation function* (ACF). This is simply defined to be the correlation between the given parameter value in the Markov chain separated by j iterations. The term $j > 1$ is usually referred to as *lag*. Mathematically, suppose that we are interested in parameter θ, that takes value θ^t at iteration t of the Markov chain. The autocorrelation of the parameter at lag j is simply defined to be $cor(\theta^t, \theta^{t+j})$. This is typically calculated for values $j = 1, \ldots, j_{max}$ and plotted on a graph. Example ACF plots are provided in Figure 1.13.



(a) (b) (c)

Fig. 1.13: Sample ACF plots representing (a) ideal mixing; (b) typical good mixing; (c) poor mixing.

Note that the autocorrelation function is always equal to 1 for the value $j = 0$, since $cor(\theta^t, \theta^t) = 1$. Ideally, for efficient Markov chains (as in Figures 1.13(a) and (b)), there should be a fast decrease in the value of the autocorrelation function as the lag increases. In other words, in the ACF plot, this would be represented by a sharp gradient at low values of j. This would imply that there is little relationship between values of the Markov chain within a small number of iterations. Conversely, poorly mixing chains will typically have a very shallow gradient in the ACF plot, with high autocorrelation values for even relatively large values of j (say, $j \geq 20$, as in Figure 1.13(c)).

An alternative method for calculating the Monte Carlo error makes use of the autocorrelation function. If each successive sample were completely independent, then we would not need batching to estimate σ^2 – we could instead use the standard formula

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_i - \overline{\theta})^2$$

and estimate Monte Carlo error as $\sqrt{\hat{\sigma^2}/n}$ as before. However, the samples are dependent, so our *effective* sample size for calculating the standard error is smaller than n. It turns out that the effective sample size (which we will denote M is given by

$$M = \frac{n}{1 + 2\sum_{k=1}^{\infty}\rho_k}.$$

This can be thought of as the number of independent samples the Markov chain represents. (Note that the formula above involves a sum to infinity so isn't a practical way to estimate M; in practice another method is used that you don't need to know about – see any good text book for details.) Hence, we can estimate the Monte Carlo error as $\sqrt{\hat{\sigma^2}/\hat{M}}$.

Finally, we note that the requirement to store many (often highly) dependent samples on the computer can be reduced via the process of *thinning*. This involves simply taking every kth realisation (e.g. every 10th iteration) of the Markov chain and discarding the rest. This clearly reduces the autocorrelation of the MCMC sample being used to obtain posterior summary statistics of interest. The discarded values (although possibly very highly dependent on the previous value in the Markov chain) still provides information concerning the posterior distribution. Thus thinning should only be used if there are issues relating to the storage and/or memory allocation of the large number of sampled values.

In Chapter 2, and before we describe the (standard) algorithms for constructing such a Markov chain, we consider the `BUGS` programming language used to implement Bayesian analyses, using an MCMC algorithm as a “black-box” (i.e. essentially hidden to the user).

Appendix A

PROBABILITY DISTRIBUTIONS

This appendix gives the form of the pmf/pdf and summary statistics for common distributions, which are frequently used within these Bayesian Inference notes.

A.1 Discrete distributions

Distribution	Parameters	Mass function	Mean and variance
Binomial $\theta \sim Bin(n, p)$	sample size $n \in \mathbb{N}$ $p \in [0, 1]$	$f(\theta) = \left(\begin{array}{c} n \\ \theta \end{array} \right) p^{\theta}(1-p)^{n-\theta}$ $\theta = 0, 1, \ldots, n$	$\mathbb{E}(\theta) = np$ $Var(\theta) = np(1-p)$
Poisson $\theta \sim Poisson(\lambda)$	rate $\lambda > 0$	$f(\theta) = \lambda^{\theta} \exp(-\lambda)(\theta!)^{-1}$ $\theta = 0, 1, 2, \ldots$	$\mathbb{E}(\theta) = \lambda$ $Var(\theta) = \lambda$
Geometric $\theta \sim Geom(p)$	$p \in [0, 1]$	$f(\theta) = p(1-p)^{\theta - 1}$ $\theta = 0, 1, \ldots$	$\mathbb{E}(\theta) = 1/p$ $Var(\theta) = (1-p)/p^2$
Negative Binomial $\theta \sim Neg\text{-}Bin(\alpha, \beta)$	shape $\alpha > 0$ inverse scale $\beta > 0$	$f(\theta) = \left(\begin{array}{c} \theta + \alpha - 1 \\ \theta \end{array} \right) \left(\frac{\beta}{\beta+1}\right)^{\alpha} \left(\frac{1}{\beta+1}\right)^{\theta}$ $\theta = 0, 1, 2, \ldots$	$\mathbb{E}(\theta) = \alpha/\beta$ $Var(\theta) = \frac{\alpha}{\beta^2}(\beta + 1)$
Multinomial $\boldsymbol{\theta} \sim MN(n, \boldsymbol{p})$	sample size $n \in \mathbb{N}$ $p_i \in [0, 1]$; $\sum_{i=1}^{k} p_i = 1$	$f(\boldsymbol{\theta}) = \frac{n!}{\prod_{i=1}^{k} \theta_i!} \prod_{i=1}^{k} p_i^{\theta_i}$ $\theta_i = 0, 1, \ldots, n$; $\sum_{i=1}^{k} \theta_i = n$	$\mathbb{E}(\theta_j) = np_j$ $Var(\theta_i) = np_i(1-p_i)$

A.2 Continuous distributions

Distribution	Parameters	Density function	Mean and variance
Uniform $\theta \sim U[a,b]$	$b > a$	$f(\theta) = 1/(b-a)$ $\theta \in [a,b]$	$\mathbb{E}(\theta) = (a+b)/2$ $Var(\theta) = (b-a)^2/12$
Normal $\theta \sim N(\mu, \sigma^2)$	location μ scale $\sigma > 0$	$f(\theta) = \frac{\exp\left(-(\theta-\mu)^2/(2\sigma^2)\right)}{\sqrt{2\pi\sigma^2}}$ $\infty < \theta < \infty$	$\mathbb{E}(\theta) = \mu$ $Var(\theta) = \sigma^2$
log Normal $\theta \sim \log N(\mu, \sigma^2)$	μ $\sigma > 0$	$f(\theta) = \frac{\exp\left(-(\log\theta-\mu)^2/(2\sigma^2)\right)}{\sqrt{2\pi\sigma^2}\theta}$ $0 \leq \theta < \infty$	$\mathbb{E}(\theta) = \exp\left(\mu + \frac{\sigma^2}{2}\right)$ $Var(\theta) = \exp(2\mu+\sigma^2)(\exp(\sigma^2)-1)$
Beta $\theta \sim Beta(\alpha, \beta)$	$\alpha > 0$ $\beta > 0$	$f(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$ $\theta \in [0,1]$	$\mathbb{E}(\theta) = \frac{\alpha}{\alpha+\beta}$ $Var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Exponential $\theta \sim Exp(\lambda)$	$\lambda > 0$	$f(\theta) = \lambda\exp(-\lambda\theta)$ $\theta > 0$	$\mathbb{E}(\theta) = 1/\lambda$ $Var(\theta) = 1/\lambda^2$
Gamma $\theta \sim \Gamma(\alpha, \beta)$	shape $\alpha > 0$ rate $\beta > 0$	$f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta)$ $\theta > 0$	$\mathbb{E}(\theta) = \alpha/\beta$ $Var(\theta) = \alpha/\beta^2$
Inverse Gamma $\theta \sim \Gamma^{-1}(\alpha, \beta)$	shape $\alpha > 0$ rate $\beta > 0$	$f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-(\alpha+1)}\exp(-\beta/\theta)$ $\theta > 0$	$\mathbb{E}(\theta) = \beta/(\alpha-1)$, for $\alpha > 1$ $Var(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$, $\alpha > 2$
Chi-squared $\theta \sim \chi^2_\nu$	df $\nu > 0$ (deg. of freedom)	$f(\theta) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)}\theta^{\frac{\nu}{2}-1}\exp(-\theta/2)$ $\theta > 0$ (same as $\Gamma\left(\alpha = \frac{\nu}{2}, \beta = \frac{1}{2}\right)$)	$\mathbb{E}(\theta) = \nu$ $Var(\theta) = 2\nu$
Inverse Chi-squared $\theta \sim \chi^{-2}_\nu$	df $\nu > 0$ (deg. of freedom)	$f(\theta) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)}\theta^{-\left(\frac{\nu}{2}+1\right)}\exp(-1/2\theta)$ $\theta > 0$ (same as $\Gamma^{-1}\left(\alpha = \frac{\nu}{2}, \beta = \frac{1}{2}\right)$)	$\mathbb{E}(\theta) = \frac{1}{\nu-2}$ $Var(\theta) = \frac{2}{(\nu-2)^2(\nu-4)}$
Dirichlet $\theta \sim Dir(\alpha_1, \ldots, \alpha_k)$	$\alpha_i > 0$; $\alpha_0 \equiv \sum_{i=1}^k \alpha_i$	$f(\theta) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)}\prod_{i=1}^k \theta_i^{\alpha_i-1}$ $\theta_i > 0$; $\sum_{i=1}^k \theta_i = 1$	$\mathbb{E}(\theta_i) = \frac{\alpha_i}{\alpha_0}$ $Var(\theta_i) = \frac{\alpha_i(\alpha_0-\alpha_i)}{\alpha_0^2(\alpha_0+1)}$

Appendix B

NIMBLE

B.1 NIMBLE: Numerical Inference for Statistical Models for Bayesian and Likelihood Estimation

NIMBLE is a system for building and sharing analysis methods for statistical models, especially for hierarchical models and computationally-intensive methods. NIMBLE is built in R but compiles your models and algorithms using C++ for speed. It includes three components:

- *A system for using models written in BUGS model language as programmable objects in R.*
- *An initial library of algorithms for models written in BUGS, including basic MCMC, which can be used directly or can be customized from R before being compiled and run.*
- *A language embedded in R for programming algorithms for models, both of which are compiled through C++ code and loaded into R (from r-nimble.org).*

We will learn how to use `Nimble` in RStudio for Bayesian analysis practically, through examples presented and discussed in lectures and tutorials and relevant `R` code uploaded in Moodle. You can find further information and examples at https://r-nimble.org/. The Nimble manual and a Nimble cheatsheet are available also from Moodle. The very first step to start learning `Nimble` is to correctly install it on our computer! (see next section).

B.2 What is required for using the `R` package `Nimble`

Installing a C++ compiler

First, you need to install a C++ compiler on your computer for Nimble to work. Specifically, for the different operating systems, the Nimble Manual states:

MacOS

On MacOS, you should install Xcode. The command-line tools, which are available as a smaller installation, should be sufficient. This is freely available from the Apple developer site and the App Store. In the somewhat unlikely event you want to install from the source package rather than the CRAN binary package, the easiest approach is to use the source package provided at R-nimble.org. If you do want to install from the source package provided by CRAN, you'll need to install this gfortran package.

Linux

On Linux, you can install the GNU compiler suite (gcc/g++). You can use the package manager to install pre-built binaries. On Ubuntu, the following command will install or update make, gcc and libc: *sudo apt-get install build-essential*

Windows

On Windows, download and install Rtools.exe available at https://cran.r-project.org/bin/windows/Rtools/. Select the appropriate executable corresponding to your version of R (and update your version of R if you notice it is not the most recent).

Putting Rtools on the PATH

You need to put the location of the Rtools 'make' utilities (bash, make, etc.) on the PATH. This can be done by including the following command in your R code:

```
writeLines('PATH="${RTOOLS40_HOME}\\usr\\bin;${PATH}"', con = "~/.Renviron")
```

Installing required R packages

Finally, you need to install the following R packages:

```
install.packages("nimble")
install.packages("igraph")
install.packages("coda")
install.packages("R6")
```

It is easier to use the R studio drop down commands to install the packages, rather than the commands above. Also, remember to 'load' the packages:

```
library(nimble)
library(igraph)
library(coda)
library(R6)
```