MATH3013: Statistical Inference

MATH4075: Classical & Bayesian Inference

Autumn Semester, 2023–2024

Dr. Yordan P. Raykov

School of Mathematical Sciences
University of Nottingham

Contents

Chapter 1

Introduction & Key concepts

In this module, we will study in depth the Frequentist and Bayesian approaches to statistical inference.

The goal of this module is to give you the fundamental theoretical toolbox relevant to a professional statistician / data scientist / data analyst.

Outline

1. Introduction & Key concepts
2. Point estimation (Frequentist)
3. Hypothesis testing (Frequentist)
4. The Delta-method (Frequentist)
5. Introduction to Bayesian statistics (Bayesian)
6. The case of categorical data (Bayesian)
7. The case of normal data (Bayesian)
8. An introduction to sample-based Bayesian inference using Markov Chain Monte Carlo (MCMC)

Statistical inference: the science of how to **infer** the parameters of a statistical model given **observed data** (that we assume were generated) from this model.

Example Given the lifetimes of 100 identical lightbulbs, estimate the average/median lifetime and/or lifetime distribution of a lightbulb.

Fundamental information for inference:

Probability model:
A model describing the type of random behaviour the system considered has. In other words, the model tells us how the data may have been generated.

Sample/Observed data:
Realisations of the system being modelled. Often independent repetitions of the same experiment.

Prior knowledge (only in Bayesian setting):
Summary of knowledge about unknown parameters prior to experiment.

1.1 Probability Models

A probability model is a way of describing the random behaviour of the system considered.

Probability models are generally expressed in terms of probability distributions, themselves expressed via their probability mass function (pmf) for discrete random variables (rvs) or probability density function (pdf) for continuous rvs.

Notation

$X, X_1, X_2, \ldots, X_n, Y, Z$ - random variables
$x, x_1, x_2, \ldots, x_n, y, z$ - observed values
Ω - sample space (of the possible values)

In this course, we are parametric: models are functions of an unknown parameter $\boldsymbol{\theta}$.

$p(x|\boldsymbol{\theta})$ - pmf/pdf at point x & parameter $\boldsymbol{\theta}$
Θ - parameter space (finite-dimensional)

Examples

(i) **Number X of tweets** by Donald Trump during fixed time (eg. a week): **Number of COVID-19 cases** in Nottingham per day: Poisson model, *i.e.* discrete distribution with pmf

$$p_X(x|\theta) = \frac{e^{-\theta}\theta^x}{x!},\ x \in \mathbb{N} = \{0, 1, 2, \ldots\},$$
$$\Omega = \mathbb{N},\ \ \Theta = (0, \infty) = \{\theta : \theta > 0\}.$$

(ii) **Service time** T in a bank; the **time it takes** to be served in a bank. Exponential model, *i.e.* continuous distribution with pdf

$$p_T(t|\theta) = \theta e^{-\theta t},\ t > 0,$$
$$\Omega = (0, \infty),\ \Theta = (0, \infty).$$

(iii) **Yields of crop** Y in relation to quantity of fertilizer x, in n fields. Linear regression model:

$$(Y_1, x_1), \ldots, (Y_n, x_n),\ Y_i \sim N(\mu_i, \sigma^2),$$
$$(Y_1, x_1), \ldots, (Y_n, x_n),\ Y_i \sim N(\alpha + \beta x_i, \sigma^2),$$
$$\Omega = \mathbb{R}^n,\ \boldsymbol{\theta} = (\alpha, \beta, \sigma^2),\ \Theta = \mathbb{R} \times \mathbb{R} \times (0, \infty).$$

[an example of linear regression]

How to choose a model? For instance:

(i) using previous work on the situation studied;

(ii) from a theoretical analysis of the situation;

(iii) for model simplicity, with meaningful parameters;

(iv) by trial and error.

In this module we will mostly assume the model has been chosen, and we shall develop tools to infer the parameters of this model from sample data.

Advertisement: Modules addressing how to choose the model in particular situations include Applied Statistical Modelling & Data Analysis and Modelling.

1.2 Likelihood, Score statistic and Fisher information

Definition Suppose that sample data

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$$

is described by a probability model whose joint pmf/pdf is $p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})$. Then the likelihood of the parameter vector $\boldsymbol{\theta}$ given the data $\boldsymbol{x}$ is

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}),$$

i.e. it is simply the (joint)pmf/pdf of the data, but seen as a function of the parameter vector.

Fundamental case If the data $\boldsymbol{x} = (x_1, \ldots, x_n)$ are independent and identically distributed (IID) observations from the same pmf/pdf $p_X(\cdot|\boldsymbol{\theta})$ then, due to independence of the observations:

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{i=1}^{n} p_X(x_i|\boldsymbol{\theta}).$$

Examples (continued)

(i) Number of COVID-19 cases: Given $n = 1$ observation x,

$$L(\theta|x) = \begin{cases} \dfrac{e^{-\theta}\theta^x}{x!} & \text{if } \theta > 0, \\ 0 & \text{if } \theta \leq 0. \end{cases}$$

This is (of course) the pmf of a Poisson distribution with mean θ. It is equivalent to the likelihood since we have only observed one data point.

(ii) Lifetimes of n lightbulbs $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)$:

$$L(\theta|\boldsymbol{t}) = \prod_{i=1}^{n} \theta e^{-\theta t_i} = \theta^n \exp\left(-\theta \sum_{i=1}^{n} t_i\right),$$

for $\theta > 0$, and 0 otherwise. Here $\boldsymbol{t}$ denotes the observed data and we have assumed that we have IID.

(iii) Crop yield, simple linear regression: the data, also called response variable, is $\boldsymbol{y} = (y_1, \ldots, y_n)$ and the explanatory variable is $\boldsymbol{x} = (x_1, \ldots, x_n)$. Here:

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{\frac{-\sum(y_i - \alpha - \beta x_i)^2}{2\sigma^2}\right\}, \end{aligned}$$

for any $\boldsymbol{\theta} = (\alpha, \beta, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times (0, \infty)$.

As in the previous example, the term in the product is the pdf of a Normal distribution with mean $\alpha + \beta x_i$ and variance σ^2.

Log-likelihood Because the likelihood says how likely a value of the parameter is given the data, one way to infer $\boldsymbol{\theta}$ from the data is to maximise the likelihood, that is, find the most likely value of the parameter given the data.

However, this involves differentiating the likelihood, which does not mix well with its product structure in the IID case. It is much easier to maximise a sum, and this is why we always work with the log-likelihood:

$$l(\boldsymbol{\theta}|\boldsymbol{x}) = \ln L(\boldsymbol{\theta}|\boldsymbol{x}) = \log L(\boldsymbol{\theta}|\boldsymbol{x}).$$

Here and in what follows "log" denotes the natural logarithm. The log function is increasing, so:

$$\begin{aligned} & L(\boldsymbol{\theta} = \boldsymbol{\theta}_1|\boldsymbol{x}) > L(\boldsymbol{\theta} = \boldsymbol{\theta}_2|\boldsymbol{x}) \\ \Leftrightarrow \quad & \log L(\boldsymbol{\theta} = \boldsymbol{\theta}_1|\boldsymbol{x}) > \log L(\boldsymbol{\theta} = \boldsymbol{\theta}_2|\boldsymbol{x}) \\ \Leftrightarrow \quad & l(\boldsymbol{\theta} = \boldsymbol{\theta}_1|\boldsymbol{x}) > l(\boldsymbol{\theta} = \boldsymbol{\theta}_2|\boldsymbol{x}). \end{aligned}$$

So maximising the log-likelihood is exactly equivalent to maximising the likelihood, and it is much easier.

Example In the crop yield example:

$$\begin{aligned} l(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) &= \log \left((2\pi\sigma^2)^{-n/2} \right. \\ &\quad \left. \times \exp\left\{ \frac{-\sum(y_i - \alpha - \beta x_i)^2}{2\sigma^2} \right\} \right) \\ l(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum(y_i - \alpha - \beta x_i)^2}{2\sigma^2}. \end{aligned}$$

Because we want to maximise the log-likelihood, it is natural to study its derivative, called the score.

Definition For a model with a single parameter θ and differentiable pdf/pmf, the score statistic is given by

$$U(\boldsymbol{X}) = \frac{\partial}{\partial\theta} l(\theta|\boldsymbol{X}) = \frac{\partial}{\partial\theta} \log p_{\boldsymbol{X}}(\boldsymbol{X}|\theta).$$

where the symbol ∂ denotes the partial derivative.

Theorem If the first two derivatives of L w.r.t. θ exist and certain regularity conditions are satisfied (including that the sample space does not depend on the parameter) then

$$\mathbb{E}[U(\boldsymbol{X})] = \mathbb{E}\left[\frac{\partial}{\partial\theta} l(\theta|\boldsymbol{X})\right] = 0$$

$$\begin{aligned} \text{and} \quad \text{Var}[U(\boldsymbol{X})] &= \text{Var}\left[\frac{\partial}{\partial\theta} l(\theta|\boldsymbol{X})\right] \\ &= -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} l(\theta|\boldsymbol{X})\right]. \end{aligned}$$

Recap: If a random variable X follows a distribution with probability density function $f_X(x)$ then

$$E[g(x)] = \int g(x) f_X(x) \mathrm{d}x.$$

Proof We have

$$\begin{aligned} \mathbb{E}[U(\boldsymbol{X})] &= \int_\Omega U(\boldsymbol{X})\, p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)\, d\boldsymbol{x} \\ &= \int_\Omega \left(\frac{\partial}{\partial\theta} \log p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)\right) p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)\, d\boldsymbol{x} \\ &= \int_\Omega \frac{1}{p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)} \frac{\partial p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}{\partial\theta} p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)\, d\boldsymbol{x} \\ &= \int_\Omega \frac{\partial p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}{\partial\theta} d\boldsymbol{x}. \end{aligned}$$

Assuming that differentiation and integration can be interchanged this yields

$$\mathbb{E}[U(\boldsymbol{X})] = \frac{\partial}{\partial\theta}\int_{\Omega} p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)\, d\boldsymbol{x} = \frac{\partial}{\partial\theta}(1) = 0$$

as required because $p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)$ is a proper density function (ie it integrates to one).

To show that the variance of the *score* statistic is equal to $-\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} l(\theta|\boldsymbol{X})\right]$ will require a bit more work.

First of all

$$\begin{aligned}\text{Var}[U(\boldsymbol{X})] &= \mathbb{E}\left[(U(\boldsymbol{X}))^2\right] - \left(\mathbb{E}\left[U(\boldsymbol{X})\right]\right)^2 \\ &= \mathbb{E}\left[(U(\boldsymbol{X}))^2\right]\end{aligned}$$

since we showed above that the expectation of the score is zero.

Rather trying to calculate $\mathbb{E}[(U(\boldsymbol{X}))^2]$ directly (as we did above for $\mathbb{E}[U(\boldsymbol{X})]$) we start by calculating the 2nd derivative of the log-likelihood:

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2} l(\theta|\boldsymbol{X}) &= \frac{\partial}{\partial\theta}\left(\frac{\partial}{\partial\theta} l(\theta|\boldsymbol{X})\right) \\
&= \frac{\partial}{\partial\theta}\left(\frac{1}{p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}\,\frac{\partial p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}{\partial\theta}\right)
\end{aligned}$$

$$\begin{aligned}
&= -\frac{1}{(p_{\boldsymbol{X}}(\boldsymbol{x}|\theta))^2}\left(\frac{\partial p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}{\partial\theta}\right)^2 \\
&+ \frac{1}{p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}\,\frac{\partial^2 p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}{\partial\theta^2} \\
&= \frac{1}{p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}\,\frac{\partial^2 p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}{\partial\theta^2} - (U(\boldsymbol{X}))^2
\end{aligned}$$

since

$$U(\boldsymbol{X}) = \frac{\partial \log p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}{\partial\theta} = \frac{1}{p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}\frac{\partial p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}{\partial\theta}.$$

[Recap] So far we have got that:

$$\frac{\partial^2}{\partial\theta^2} l(\theta|\boldsymbol{X}) = \frac{1}{p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)} \frac{\partial^2 p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}{\partial\theta^2} - \left(U(\boldsymbol{X})\right)^2 .$$

Consequently, take expectations and re-arrange:

[for notational convenience (i.e. laziness!) write $p_{\boldsymbol{X}}$ for $p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)$ and l for $l(\theta|\boldsymbol{X})$]

$$\begin{aligned}
\mathbb{E}[(U(\boldsymbol{X}))^2] &= \mathbb{E}\left[\frac{1}{p_{\boldsymbol{X}}} \frac{\partial^2 p_{\boldsymbol{X}}}{\partial\theta^2}\right] - \mathbb{E}\left[\frac{\partial^2 l}{\partial\theta^2}\right] \\
&= \int_\Omega \frac{1}{p_{\boldsymbol{X}}} \frac{\partial^2 p_{\boldsymbol{X}}}{\partial\theta^2}\, p_{\boldsymbol{X}}\, d\boldsymbol{x} - \mathbb{E}\left[\frac{\partial^2 l}{\partial\theta^2}\right] \\
&= \frac{\partial^2}{\partial\theta^2} \int_\Omega p_{\boldsymbol{X}}\, d\boldsymbol{x} - \mathbb{E}\left[\frac{\partial^2 l}{\partial\theta^2}\right] \\
&= 0 - \mathbb{E}\left[\frac{\partial^2 l}{\partial\theta^2}\right] \\
&= -\mathbb{E}\left[\frac{\partial^2 l}{\partial\theta^2}\right],
\end{aligned}$$

and thus

$$\mathrm{Var}[U(\boldsymbol{X})] = -\mathbb{E}\left[\frac{\partial^2 l}{\partial \theta^2}\right].$$

This concludes the proof. ∎

Definition The quantity

$$\mathrm{Var}[U(\boldsymbol{X})] = -\mathbb{E}\left[\frac{\partial^2 l(\theta|\boldsymbol{X})}{\partial \theta^2}\right]$$

is called the Fisher information about θ, and is denoted by $I(\theta)$.

Interpretation The Fisher information is a measure of how sharply the (log-)likelihood varies on average:

$$I(\theta) = \mathbb{E}[(U(\boldsymbol{X}))^2] = \mathbb{E}\left[\left(\frac{1}{p_{\boldsymbol{X}}(\boldsymbol{X}|\theta)}\frac{\partial p_{\boldsymbol{X}}(\boldsymbol{X}|\theta)}{\partial \theta}\right)^2\right].$$

Since the expected value of the score is zero (as shown previously), the Fisher's information represents the variance of the score at a particular value of the parameter θ.

We shall see in Chapter 2 that a standard measure of how well θ can be estimated is given by $1/I(\theta)$.

Additivity of the Fisher information

If $X_1, \ldots, X_n$ are IID with pmf/pdf $p_X(\cdot|\theta)$ then

$$\begin{aligned} l(\theta|\boldsymbol{X}) &= \sum_{i=1}^{n} \log p_X(X_i|\theta) \\ \text{and } \frac{\partial^2}{\partial\theta^2} l(\theta|\boldsymbol{X}) &= \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log p_X(x_i|\theta), \\ \text{so } I(\theta) &= \mathbb{E}\left[-\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log p_X(X_i|\theta) \right] \\ &= \sum_{i=1}^{n} \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2} \log p_X(X_i|\theta) \right] \\ &= n\mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2} \log p_X(X|\theta) \right]. \end{aligned}$$

As a consequence, if we define $i(\theta)$ as the Fisher information contained in one observation, namely

$$i(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log p_X(X|\theta) \right],$$

then $I(\theta) = n\, i(\theta)$.

The more observations, the more information!

1.3 The Exponential Family

The log-likelihood is a very important tool in statistical inference. It is then natural to think of the distributions whose log-likelihood is easy to compute.

Definition The one-parameter exponential family of distributions consists of the distributions whose pdf/pmf can be written

$$p_X(x|\theta) = \exp\{A(\theta)B(x) + C(\theta) + D(x)\},$$

when $p_X(x|\theta) > 0$; here A, B, C and D are arbitrary real-valued functions of their arguments.

Example If $X \sim \text{Poisson}(\theta)$, then for $x \in \mathbb{N}$ and $\theta > 0$,

$$\begin{aligned} p_X(x|\theta) &= \frac{\theta^x e^{-\theta}}{x!} \\ &= \exp\{x \log\theta - \theta - \log x!\}, \end{aligned}$$

i.e. the Poisson family is part of the exponential family with $A(\theta) = \log\theta$, $B(x) = x$, $C(\theta) = -\theta$, and $D(x) = -\log x!$.

Other examples: Binomial(m, θ) (m known), normal $N(\theta, 1)$, Gamma$(\theta, 1)$. See exercise sheet.

1.4 Sufficiency

It often happens that all the information about an unknown parameter θ provided by a set $\boldsymbol{x}$ of observations is actually summarised by a specific subset of events/estimates of θ.

Example Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be the outcomes of n IID Bernoulli trials:

$$x_i = \mathbb{I}_{\{\text{trial } i \text{ successful}\}}$$

with success probability θ.

Claim: The total number of successes, $\sum_{i=1}^{n} x_i$, provides all the information about θ contained in the sample.

In other words, if the total number of successes is known, then no further information about θ can be extracted from the data. We are going to show that this is the case.

To make this precise, *partition the sample space* $\Omega = \{(x_1, \ldots, x_n) : x_i = 0 \text{ or } 1\}$ as

$$\Omega = \bigcup_{j=0}^{n} A_j, \quad A_j = \left\{ (x_1, \ldots, x_n) : \sum_{i=1}^{n} x_i = j \right\}.$$

Now for $\boldsymbol{x} = (x_1, \ldots, x_n)$ such that $\sum_{i=1}^{n} x_i = j$,

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{X} = \boldsymbol{x} | \boldsymbol{X} \in A_j, \theta) \\
&\quad = \frac{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{X} \in A_j | \theta)}{\mathbb{P}(\boldsymbol{X} \in A_j | \theta)} \\
&\quad = \frac{\theta^j (1-\theta)^{n-j}}{\binom{n}{j} \theta^j (1-\theta)^{n-j}} \\
&\quad = 1 / \binom{n}{j}; \quad \text{here } \binom{n}{j} = \frac{n!}{j!(n-j)!}.
\end{aligned}$$

The set of outcomes $(x_1, \ldots, x_n)$, given $\sum_{i=1}^{n} x_i$, is then independent of θ.

This exactly means that the knowledge of $\sum_{i=1}^{n} x_i$ is "sufficient" to describe θ.

Definition A partition $\mathcal{A}$ of the sample space Ω is called sufficient for θ, if for all $A_j \in \mathcal{A}$, $p_{\boldsymbol{X}}(\boldsymbol{x}|\theta, \boldsymbol{x} \in A_j)$ is independent of θ.

There is always at least one sufficient partition: the partition whose elements are individual points of Ω is clearly sufficient since if all the data is given, then of course no further information can be extracted.

An interesting question is then to find a minimal sufficient partition, or informally, the most concise way given by the data to describe θ.

Definition If a sufficient partition $\mathcal{A}$ is such that given any other sufficient partition $\mathcal{B}$, any set element of $\mathcal{B}$ is contained in a set element of $\mathcal{A}$, then $\mathcal{A}$ is said to be minimal sufficient.

We can already see that the definition of (minimal) sufficiency is interesting but difficult to work with. It is much easier to use a function of the data instead.

Definition Let J be an arbitrary set. A function $T : \Omega \to J$ independent of θ is called a statistic.

Definition A statistic $T : \Omega \to J$ is called sufficient (resp. minimal sufficient) for θ if the partition

$$\{\boldsymbol{x} \in \Omega \ : \ T(\boldsymbol{x}) = a\}, \ a \in J$$

is sufficient (resp. minimal sufficient) for θ.

It is important to note that T is sufficient for θ if

(i) for all $\boldsymbol{x}$ and a such that $T(\boldsymbol{x}) = a$,

$$p_{\boldsymbol{X}}(\boldsymbol{x}|\theta, T(\boldsymbol{x}) = a) \text{ is independent of } \theta,$$

or equivalently,

(ii) for all $\boldsymbol{x}$ and $\boldsymbol{x}'$,

$$T(\boldsymbol{x}) = T(\boldsymbol{x}') \Rightarrow \frac{p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}{p_{\boldsymbol{X}}(\boldsymbol{x}'|\theta)} \text{ is independent of } \theta.$$

(Note that either if (i) or (ii) is regarded a definition of sufficiency).

T is minimal sufficient for θ if there is equivalence ($\Leftrightarrow$) in (ii), instead of a simple implication:

$$T(\boldsymbol{x}) = T(\boldsymbol{x}') \Leftrightarrow \frac{p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}{p_{\boldsymbol{X}}(\boldsymbol{x}'|\theta)} \text{ is independent of } \theta.$$

Example In the Bernoulli trials example we saw earlier,

$$\frac{p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}{p_{\boldsymbol{X}}(\boldsymbol{x}'|\theta)} = \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\theta^{\sum x_i'}(1-\theta)^{n-\sum x_i'}} = \left(\frac{\theta}{1-\theta}\right)^{\sum x_i - \sum x_i'}$$

which is independent of θ if and only if $\sum x_i = \sum x_i'$. Thus the statistic $T = \sum X_i$ is minimal

sufficient. Equivalently, the partition

$$\Omega = \bigcup_{j=0}^{n} A_j, \quad A_j = \left\{ (x_1, \ldots, x_n) : \sum_{i=1}^{n} x_i = j \right\}$$

is minimal sufficient.

The next result is a fundamental characterisation of sufficiency.

Theorem (Neyman's Factorisation Theorem) A statistic T is sufficient for θ if and only if there exist functions g and h such that

$$p_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = g(\theta, T(\boldsymbol{x}))h(\boldsymbol{x})$$

i.e. the pmf/pdf can be factored into the product of one term depending only on $\boldsymbol{x}$ through $T(\boldsymbol{x})$ and another term depending on $\boldsymbol{x}$ only.

Proof (Discrete case: $X_1, \ldots, X_n$ are discrete r.v) If T is sufficient for θ, and $T(\boldsymbol{x}) = a$, say, then

$$\begin{aligned} p_{\boldsymbol{X}}(\boldsymbol{x}|\theta) &= \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|\theta) \\ &= \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}, T(\boldsymbol{X}) = a|\theta) \\ &= \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|\theta, T(\boldsymbol{X}) = a)\, \mathbb{P}(T(\boldsymbol{X}) = a|\theta) \\ &= h(\boldsymbol{x})g(\theta, a), \end{aligned}$$

since T is sufficient for θ.

Conversely, suppose that

$$p_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = g(\theta, T(\boldsymbol{x}))h(\boldsymbol{x}).$$

Then for any $\boldsymbol{x}$ such that $T(\boldsymbol{x}) = a$,

$$\begin{aligned}
\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|\theta, T(\boldsymbol{X}) = a) &= \frac{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}, T(\boldsymbol{X}) = a|\theta)}{\mathbb{P}(T(\boldsymbol{X}) = a|\theta)} \\
&= \frac{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}|\theta)}{\sum_{\boldsymbol{y}:T(\boldsymbol{y})=a} \mathbb{P}(\boldsymbol{X} = \boldsymbol{y}|\theta)} \\
&= \frac{g(\theta, a)h(\boldsymbol{x})}{\sum_{\boldsymbol{y}:T(\boldsymbol{y})=a} g(\theta, a)h(\boldsymbol{y})} \\
&= \frac{h(\boldsymbol{x})}{\sum_{\boldsymbol{y}:T(\boldsymbol{y})=a} h(\boldsymbol{y})},
\end{aligned}$$

which is independent of θ. ∎

Example For n independent Bernoulli trials with success probability θ we saw earlier that the likelihood can be written as:

$$p_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

and if we re-arrange

$$p_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = \underbrace{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}_{g(\theta, \sum x_i)} \times \underbrace{1}_{h(\boldsymbol{x})}$$

and we found earlier that $\sum_{i=1}^{n} X_i$ is sufficient.

Theorem If $X_1, \ldots, X_n$ are IID from a distribution of the exponential family

$$p_X(x|\theta) = \exp\{A(\theta)B(x) + C(\theta) + D(x)\}$$

then $T = \sum B(X_i)$ is sufficient for θ.

Proof Write

$$p_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \exp\{A(\theta)B(x_i) + C(\theta) + D(x_i)\}$$

$$= \underbrace{\exp\left\{A(\theta)\sum_{i=1}^{n} B(x_i) + nC(\theta)\right\}}_{g(\theta, \sum B(x_i))} \underbrace{\exp\left(\sum_{i=1}^{n} D(x_i)\right)}_{h(\boldsymbol{x})}.$$

By Neyman's factorisation theorem, $T = \sum B(X_i)$ is sufficient for θ. ∎

1.5 Completeness

[Disclaimer: this is a rather difficult concept to grasp!]

Definition A family $\{p_{\boldsymbol{X}}(\boldsymbol{x}|\theta) : \theta \in \Theta\}$ of distributions on Ω is called complete if

$$\begin{aligned} \mathbb{E}[h(\boldsymbol{X})] &= 0 \quad \text{for all} \quad \theta \in \Theta \\ \Rightarrow \mathbb{P}[h(\boldsymbol{X}) = 0 \mid \theta] &= 1 \quad \text{for all} \quad \theta \in \Theta, \end{aligned}$$

for any statistic $h(\boldsymbol{X})$ such that the above expectation makes sense.

If T is a statistic, let $\{p_T(t|\theta) : \theta \in \Theta\}$ be the family distributions of T.

Definition A statistic T is said to be complete if its family of distributions $\{p_T(t|\theta) : \theta \in \Theta\}$ is complete.

Why bother?

One interesting consequence of the completeness of a family of distributions of $\boldsymbol{X}$ is that if there is a statistic $h(\boldsymbol{X})$ such that

$$\mathbb{E}[h(\boldsymbol{X})] = \theta \text{ for all } \theta \in \Theta$$

then $h(\boldsymbol{X})$ is the only such statistic.

This rigid structure will make it very easy to find and study unbiased estimators of θ. More details later.

Example Suppose that we have n independent Bernoulli random variables with probability of success θ.

We consider T to be the number of successes we which know it follows a Binomial distribution with probability mass function:

$$p_T(t|\theta) = \binom{n}{t}\theta^t(1-\theta)^{n-t}, \quad t = 0, \ldots n, \quad \theta \in (0,1).$$

$$\mathbb{E}[h(T)] = \sum_{t=0}^{n} h(t)\binom{n}{t}\theta^t(1-\theta)^{n-t},$$

which we write

$$\mathbb{E}[h(T)] = (1-\theta)^n \sum_{t=0}^{n} h(t)\binom{n}{t}\left(\frac{\theta}{1-\theta}\right)^t$$

Since the factor $(1-\theta)^n \neq 0$, for this expectation to be 0 for all $\theta \in (0,1)$ then we must have

$$0 = \sum_{x=0}^{n} h(t)\binom{n}{t}\phi^t, \quad \forall \phi > 0, \quad \phi = \frac{\theta}{1-\theta}.$$

This is a polynomial in ϕ of degree n.

For this polynomial to be zero for all $\phi > 0$, it must be true that the coefficient of ϕ^t, which is $h(t)\binom{n}{x}$, is 0 for every t.

This shows that $h(t) = 0$ for $t = 0, \ldots, n$ and hence $\mathbb{P}[h(T) = 0 \mid \theta] = 1$ for all $\theta \in \Theta$. In other words T is a complete sufficient statistic.

For a one-parameter exponential family model in its natural parametrisation the following powerful theorem may be used to check for completeness.

Theorem Suppose $\boldsymbol{X} = (X_1, \ldots, X_n)$ is an IID sample from the probability model

$$p_X(x|\theta) = \exp\{\theta B(x) + C(\theta) + D(x)\}, \ \theta \in \Theta,$$

and let $T = \sum B(X_i)$ denote the corresponding sufficient statistic.

If Θ contains an open interval, then T is complete.

Example Consider the family of Poisson distributions with mean μ, where $\mu \in (0, \infty)$. For $x = 0, 1, 2, \ldots$ the pmf is given by

$$p_X(x|\mu) = \frac{\mu^x e^{-\mu}}{x!} = \exp(x \log \mu - \mu - \log x!).$$

Writing $\theta = \log \mu$ and putting $B(x) = x$,

$$C(\theta) = -\exp(\theta), \quad D(x) = -\log x!,$$

we see that the Poisson pmf is of the required exponential form. Since $\mu \in (0, \infty)$, it follows that

$$\theta = \log \mu \in \Theta = \mathbb{R},$$

which is an open interval. Thus the family of Poisson distributions is complete.

Note that we actually showed that the statistic X is complete. This is because completeness of X (as

a statistic, for $n = 1$) is exactly equivalent to completeness of the family of distributions of X.

We conclude by an illustration of why completeness matters in practice.

Bahadur Theorem If T is sufficient for θ and complete, then T is minimal sufficient.

Beware: the converse is not true in general.

Bahadur's Theorem is most easily applied in practice by showing that the model belongs to the exponential family (and reparametrising if needed).

Chapter 2

Point estimation

2.1 The point estimation problem

Let $\boldsymbol{\theta} \in \Theta$ be an unknown parameter determining the exact distribution of the observations $\boldsymbol{X} = (X_1, \ldots, X_n)$ in an experiment. We wish to estimate $g(\boldsymbol{\theta})$, where g is some known real-valued function of the unknown parameter vector $\boldsymbol{\theta}$.

Example Given the lifetimes of $n = 100$ lightbulbs with distribution $\text{Exp}(\theta)$, estimate the unknown average lifetime, namely $1/\theta$, of a lightbulb.

Let $\hat{g}(\boldsymbol{x})$ be our estimate of $g(\boldsymbol{\theta})$ when we observe $\boldsymbol{X} = \boldsymbol{x}$. In addition, we wish to find an estimator $\hat{g}$ which is in some sense optimal.

Terminology: An estimator is a function of the random variables $X_1, \ldots, X_n$. The corresponding estimate is obtained when the random variables are replaced by their observed values $x_1, \ldots, x_n$.

In this chapter, we shall first study two general techniques to construct estimators: the method of moments, and maximum likelihood estimation. We will then see how to discuss whether an estimator is optimal and, if not, how to find the optimal one.

2.2 Two methods of estimation

(i) The method of moments

Suppose here that $g(\boldsymbol{\theta}) = h(m_1, \ldots, m_k)$, where h is known and m_r is the r-th moment of the unknown distribution, *i.e.* $m_r = \mathbb{E}[X^r]$. In other words the first moment, m_1 is the expected value of X. The method of moments uses the law of large numbers to estimate m_r by the r-th moment of the sample, *i.e.*

$$m_r = \mathbb{E}[X^r] \text{ estimated by } \hat{m}_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r.$$

Thus $g(\boldsymbol{\theta})$ is estimated by the estimator

$$\hat{g}(\boldsymbol{X}) = h\left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} X_i^2, \ldots, \frac{1}{n}\sum_{i=1}^{n} X_i^k\right).$$

In short: replace the population average $\mathbb{E}(\cdot)$ by the sample average $n^{-1}\sum_{i=1}^{n}(\cdot)$.

Sometimes (for instance when estimating a variance) it is easier to work with centralised moments:

$$\forall r \geq 2, \quad m_r^C = \mathbb{E}[(X - \mathbb{E}[X])^r]$$

$$\text{estimated by } \hat{m}_r^C = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^r, \ \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Example 1 Let $X_1, \ldots, X_n$ be IID Exp(λ). Then $\mathbb{E}(X_i) = 1/\lambda$ and the method of moments gives the estimate $\hat{\lambda} = 1/\overline{x}$ for λ.

Example 2 Let $X_1, \ldots, X_n$ be IID Gamma(α, β), *i.e.* with common pdf

$$p_X(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \ x > 0.$$

Then (see handout on Gamma distribution!)

$$\mathbb{E}(X_i) = \frac{\alpha}{\beta}, \ \text{Var}(X_i) = \mathbb{E}([X_i - \mathbb{E}(X_i)]^2) = \frac{\alpha}{\beta^2}.$$

The method of moments to find estimates gives

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\hat{\alpha}}{\hat{\beta}}, \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{\hat{\alpha}}{\hat{\beta}^2}$$

$$\Rightarrow \hat{\alpha} = \frac{\bar{x}^2}{S^2}, \ \hat{\beta} = \frac{\bar{x}}{S^2}.$$

(ii) Maximum likelihood estimation

A maximum likelihood estimate (MLE), $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ say, of a parameter $\boldsymbol{\theta}$, is a value of $\boldsymbol{\theta}$ which maximises the likelihood $L(\boldsymbol{\theta}|\boldsymbol{x})$ w.r.t. $\boldsymbol{\theta}$ for fixed $\boldsymbol{x}$:

$$L(\hat{\boldsymbol{\theta}}(\boldsymbol{x})|\boldsymbol{x}) = \sup_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}|\boldsymbol{x}).$$

As mentioned earlier we shall instead equivalently maximise the log-likelihood $l(\boldsymbol{\theta}|\boldsymbol{x}) = \log L(\boldsymbol{\theta}|\boldsymbol{x})$:

$$l(\hat{\boldsymbol{\theta}}(\boldsymbol{x})|\boldsymbol{x}) = \sup_{\boldsymbol{\theta}\in\Theta} l(\boldsymbol{\theta}|\boldsymbol{x}) = \sup_{\boldsymbol{\theta}\in\Theta} \log L(\boldsymbol{\theta}|\boldsymbol{x}).$$

In most cases we find this estimator by cancelling the score function:

$$\frac{\partial}{\partial\theta}(l\hat{\boldsymbol{\theta}}(\boldsymbol{x})|\boldsymbol{x}) = 0.$$

For instance, this works when the log-likelihood is continuously differentiable and concave.

When this is not the case the log-likelihood has to be studied directly. See the Exercise handouts for such examples.

Example Let $X_1, \ldots, X_n$ be IID Poisson(λ). The likelihood is, for $\lambda > 0$:

$$L(\lambda|\boldsymbol{x}) = p_{\boldsymbol{X}}(\boldsymbol{x}|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\prod x_i!}.$$

Taking logs gives the log-likelihood:

$$l(\lambda|\boldsymbol{x}) = -n\lambda + \left(\sum x_i\right) \log \lambda - \log\left(\prod x_i!\right).$$

At an MLE, the first derivative must vanish (set the derivative equal to zero and solve for θ):

$$\frac{\partial l}{\partial \lambda}(\lambda|\boldsymbol{x}) = -n + \frac{\sum x_i}{\lambda} = 0 \Leftrightarrow \lambda = \bar{x} = \frac{\sum x_i}{n}.$$

To check that this point is a maximum, we compute the second derivative:

$$\frac{\partial^2}{\partial \lambda^2} l(\lambda|\boldsymbol{x}) = \frac{-\sum x_i}{\lambda^2} < 0 \quad \forall \lambda.$$

Since the second derivative is negative for any λ then that means that it will also be negative when evaluated at the MLE ($\hat{\lambda}$) which what is necessary for a

stationary point to be a maximum. Hence the MLE is then indeed $\hat{\lambda} = \bar{x}$.

2.3 Desirable properties of estimators

(i) The estimator should not take values outside the parameter space.

For instance, if $X \sim \text{Binomial}(n, p)$, then we always want $\hat{p} \in [0, 1]$.

(ii) Unbiasedness Suppose we write the expectation of an estimator $\hat{g}$ of $g(\boldsymbol{\theta})$ as

$$\mathbb{E}[\,\hat{g}(\boldsymbol{X})] = g(\boldsymbol{\theta}) + b_{\hat{g}}(\boldsymbol{\theta}),$$

where $b_{\hat{g}}(\boldsymbol{\theta}) = \mathbb{E}[\,\hat{g}(\boldsymbol{X}) - g(\boldsymbol{\theta})]$ is the bias function.

Definition An estimator $\hat{g}$ is called unbiased for $g(\boldsymbol{\theta})$ if $b_{\hat{g}}(\boldsymbol{\theta}) = 0$, *i.e.*

$$\mathbb{E}[\,\hat{g}(\boldsymbol{X})] = g(\boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \Theta.$$

An unbiased estimator is an estimator which recovers the value of the parameter on average.

(iii) Small volatility The estimator should have as small a Mean Squared Error (MSE)

$$\begin{aligned} \text{MSE} &= \mathbb{E}\left\{[\hat{g}(\boldsymbol{X}) - g(\boldsymbol{\theta})]^2\right\} \\ &= \text{Var}[\hat{g}(\boldsymbol{X})] + [b_{\hat{g}}(\boldsymbol{\theta})]^2 \end{aligned}$$

as possible.

The MSE of an estimator measures how far the estimator is from the target quantity on average. Suppose we have two estimators. One of them is biased and other one unbiased. But the biased estimator has smaller variance than the unbiased. One can decide which of the two estimators to use based on their MSEs.

Note that if $\hat{g}(\boldsymbol{X})$ is unbiased then

$$\text{MSE} = \text{Var}[\hat{g}(\boldsymbol{X})].$$

In that case, loosely speaking, a small MSE means a small variance, *i.e.* there is a high probability that the estimator is close to the true value.

(iv) Consistency An estimator $\hat{g}(\boldsymbol{X})$ is consistent if $\hat{g}(\boldsymbol{X}) \to g(\boldsymbol{\theta})$ as $n \to \infty$, *i.e.* for every $\varepsilon > 0$

$$\mathbb{P}(|\hat{g}(\boldsymbol{X}) - g(\boldsymbol{\theta})| > \varepsilon \,|\, \theta) \to 0 \;\text{ as } n \to \infty.$$

This is called convergence in probability: the estimator is then close to the target quantity with probability arbitrarily close to 1 when $n \to \infty$.

It can be shown that if

$$\mathbb{E}[\hat{g}(\boldsymbol{X})] \to g(\boldsymbol{\theta}) \text{ and } \mathrm{Var}(\hat{g}(\boldsymbol{X})) \to 0 \text{ as } n \to \infty$$

then $\hat{g}(\boldsymbol{X})$ is consistent for $g(\boldsymbol{\theta})$.

It is generally fairly easy to work on these four aspects for moment estimators because they are defined in terms of sample moments.

We therefore focus instead on the general theory of MLE in the next section.

2.4 Properties of MLEs

Here, we restrict our attention to MLEs that are unique and to models whose sample space is not parameter-dependent (no uniform distribution on an interval of unknown length, for instance).

If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, we define the Fisher information matrix to be the $p \times p$ matrix $\mathbf{I}(\boldsymbol{\theta})$ with components

$$\mathbf{I}(\boldsymbol{\theta})_{ij} = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} l(\boldsymbol{\theta}|\boldsymbol{X})\right), \quad i, j = 1, \ldots, p.$$

where $\mathbf{I}(\boldsymbol{\theta})_{ij}$ denotes the ij$_{th}$ element of $\mathbf{I}(\boldsymbol{\theta})$ and $l(\boldsymbol{\theta}|\boldsymbol{X})$ denotes the log-likelihood.

Also, we write $N_p(\boldsymbol{\mu}, \mathbf{V})$ for the p-dimensional multivariate normal distribution with vector mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$. We shall see study this distribution in more detail later; simply note that when $\mathbf{V}$ has full rank p, so that $\mathbf{V}^{-1}$ exists, then $N_p(\boldsymbol{\mu}, \mathbf{V})$ has multivariate pdf

$$(2\pi)^{-p/2}|\mathbf{V}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\},$$

where $|\mathbf{V}|$ is the determinant of $\mathbf{V}$.

(i) Invariance t If $\hat{\boldsymbol{\theta}}$ is the MLE for $\boldsymbol{\theta}$ and g is a one-to-one function (for instance, continuous and increasing) then $g(\hat{\boldsymbol{\theta}})$ is the MLE for $g(\boldsymbol{\theta})$.

(ii) Consistency and asymptotic normality

If $X_1, \ldots, X_n$ are IID with pmf/pdf $p_X(\cdot|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a p-dimensional parameter vector, then:

Theorem Under regularity conditions, the MLE $\hat{\boldsymbol{\theta}}$ is consistent and, for large n, satisfies

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \approx N_p(\mathbf{0}_p, \mathbf{i}(\boldsymbol{\theta})^{-1}),$$

where $\mathbf{0}_p$ is a size p null vector and $\mathbf{i}(\boldsymbol{\theta}) = n^{-1}\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information matrix for a single observation.

This is a **very** important result. A lot of theory and statistical analysis (including recent advances!) is based on this.

(iii) MLE and sufficiency

The next result will be useful when we work on optimal estimators later.

Lemma If the model $p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)$ admits a sufficient statistic $T(\boldsymbol{X})$, then the MLE $\hat{\theta}$ is a function of the sufficient statistic T.

Proof Since T is sufficient, by Neyman's Factorisation Theorem we have

$$\begin{aligned} L(\theta|\boldsymbol{x}) = p_{\boldsymbol{X}}(\boldsymbol{x}|\theta) &= g(\theta, T(\boldsymbol{x}))h(\boldsymbol{x}) \\ \Rightarrow \sup_{\theta\in\Theta} L(\theta|\boldsymbol{x}) &= h(\boldsymbol{x}) \sup_{\theta\in\Theta} g(\theta, T(\boldsymbol{x})). \end{aligned}$$

Therefore maximisation of L w.r.t. θ is equivalent to maximisation of $g(\theta, T(\boldsymbol{x}))$, which depends on $\boldsymbol{x}$ through T alone. ∎

2.5 Cramér-Rao Lower Bound

Our next step is to examine if any of the estimators we developed is "best" or "can be improved".

Theorem Assume θ is a scalar parameter. Let $\hat{g}(\boldsymbol{X})$ be an unbiased estimator of $g(\theta)$. Then, subject to mild regularity conditions,

$$\text{Var}[\hat{g}(\boldsymbol{X})] \geq \frac{[g'(\theta)]^2}{I(\theta)},$$

where $g'(\theta) = \frac{\partial g(\theta)}{\partial \theta)}$ and $I(\theta)$ is the Fisher information.

The right-hand side is known as the Cramér-Rao lower bound (CRLB) for $g(\theta)$.

Proof Recall the score statistic

$$U(\boldsymbol{X}) = \frac{\partial}{\partial \theta} l(\theta|\boldsymbol{X}) = \frac{\partial}{\partial \theta} \log p_{\boldsymbol{X}}(\boldsymbol{X}|\boldsymbol{\theta}).$$

The basic idea of the proof is to connect $\text{Var}[\hat{g}(\boldsymbol{X})]$ to $I(\theta)$. Because $I(\theta) = \text{Var}[U(\boldsymbol{X})]$, we can use the inequality $|\text{corr}(\widehat{g}(\boldsymbol{X}), U(\boldsymbol{X}))| \leq 1$, rewritten as

$$\left[\frac{\text{Cov}(\widehat{g}(\boldsymbol{X}), U(\boldsymbol{X}))}{\sqrt{\text{Var}(\widehat{g}(\boldsymbol{X}))\text{Var}(U(\boldsymbol{X}))}}\right]^2 \leq 1$$

$$\Leftrightarrow \text{Var}(\widehat{g}(\boldsymbol{X})) \geq \frac{[\text{Cov}(\widehat{g}(\boldsymbol{X}), U(\boldsymbol{X}))]^2}{I(\theta)}.$$

Now compute $\operatorname{cov}(\hat{g}(\boldsymbol{X}), U(\boldsymbol{X}))$

$$
\begin{aligned}
&= \mathbb{E}[\hat{g}(\boldsymbol{X})U(\boldsymbol{X})] - \mathbb{E}[\hat{g}(\boldsymbol{X})]\mathbb{E}[U(\boldsymbol{X})] \\
&= \int_{\Omega} \hat{g}(\boldsymbol{x})U(\boldsymbol{x})p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)\, d\boldsymbol{x} - 0 \\
&= \int_{\Omega} \hat{g}(\boldsymbol{x}) \frac{\partial \log p_{\boldsymbol{X}}}{\partial \theta}\, p_{\boldsymbol{X}}\, d\boldsymbol{x} \\
&= \int_{\Omega} \hat{g}(\boldsymbol{x}) \frac{1}{p_{\boldsymbol{X}}} \frac{\partial p_{\boldsymbol{X}}}{\partial \theta}\, p_{\boldsymbol{X}}\, d\boldsymbol{x} \\
&= \int_{\Omega} \hat{g}(\boldsymbol{x}) \frac{\partial p_{\boldsymbol{X}}}{\partial \theta}\, d\boldsymbol{x} \\
&= \frac{\partial}{\partial \theta} \int_{\Omega} \hat{g}(\boldsymbol{x}) p_{\boldsymbol{X}}\, d\boldsymbol{x} \\
&= \frac{\partial}{\partial \theta} \mathbb{E}[\hat{g}(\boldsymbol{X})] \\
&= \frac{\partial}{\partial \theta} (g(\theta)) = g'(\theta)
\end{aligned}
$$

and simply use the previous inequality. ∎

The regularity conditions actually assumed here allow the interchanging of differentiation and integration. They can fail if $p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)$ is sufficiently irregular, or if the sample space of $\boldsymbol{X}$ depends on θ, for example when $X \sim U(0, \theta)$.

Corollary There is an unbiased estimator $\widehat{g}(\boldsymbol{X})$ of $g(\theta)$ which attains the CRLB, *i.e.*

$$\text{Var}(\widehat{g}(\boldsymbol{X})) = \frac{[g'(\theta)]^2}{I(\theta)}$$

if and only if $U(\boldsymbol{X})$ can be written in the form

$$U(\boldsymbol{X}) = \alpha(\theta)[\widehat{g}(\boldsymbol{X}) - g(\theta)].$$

In this case:

$$\text{Var}(\widehat{g}(\boldsymbol{X})) = \frac{g'(\theta)}{\alpha(\theta)}.$$

Proof From the proof of the Cramér-Rao lower bound, $\text{Var}(\widehat{g}(\boldsymbol{X}))$ equals the CRLB if and only if

$$|\text{corr}(\widehat{g}(\boldsymbol{X}), U(\boldsymbol{X}))| = 1.$$

But the correlation of A and B is ± 1 if and only if there is a (nontrivial) purely linear relationship between A and B; here, this translates into

$$U(\boldsymbol{X}) = \alpha(\theta)\widehat{g}(\boldsymbol{X}) + \beta(\theta)$$

with the slope $\alpha(\theta) \neq 0$. Taking expectations of both sides, we obtain

$$0 = \alpha(\theta)g(\theta) + \beta(\theta) \Leftrightarrow \beta(\theta) = -\alpha(\theta)g(\theta)$$

i.e. $U(\boldsymbol{X}) = \alpha(\theta)\{\widehat{g}(\boldsymbol{X}) - g(\theta)\}$.

Furthermore, in the proof of the CRLB we showed that

$$\text{Cov}(\widehat{g}(\boldsymbol{X}), U(\boldsymbol{X})) = g'(\theta).$$

Using $U(\boldsymbol{X}) = \alpha(\theta)\{\widehat{g}(\boldsymbol{X}) - g(\theta)\}$, we get:

$$\text{Cov}(\widehat{g}(\boldsymbol{X}), \alpha(\theta)[\widehat{g}(\boldsymbol{X}) - g(\theta)]) = g'(\theta).$$

Using $\text{Cov}(Y, aY - c) = a\text{Cov}(Y, Y) = a\text{Var}(Y)$:

$$\alpha(\theta)\text{Var}[\widehat{g}(\boldsymbol{X})] = g'(\theta) \Leftrightarrow \text{Var}[\widehat{g}(\boldsymbol{X})] = \frac{g'(\theta)}{\alpha(\theta)}.$$

This completes the proof. ∎

Some consequences:

(i) The CRLB for $g(\theta)$ is the lowest possible variance of an unbiased estimator of $g(\theta)$.

(ii) The more information about θ provided by the sample $\boldsymbol{X}$, the larger the Fisher information and the smaller the CRLB. Especially, for a large sample size, we should expect the variance of a sensible unbiased estimator to be small. This is intuitive,

but how small the variance (ideally) should be can be made precise using the knowledge of the CRLB.

Example Suppose $X_1, \ldots, X_n$ are IID Exp(λ):

$$\begin{aligned} L(\lambda|\boldsymbol{x}) &= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i} \\ \Rightarrow l(\lambda|\boldsymbol{x}) &= n \log \lambda - \lambda \sum x_i. \end{aligned}$$

The score statistic is

$$U(\boldsymbol{X}) = \frac{\partial}{\partial \lambda} l(\lambda|\boldsymbol{x}) = \frac{n}{\lambda} - \sum x_i = \underbrace{-n}_{\alpha(\lambda)} \Big(\underbrace{\frac{\sum x_i}{n}}_{\widehat{g}(\boldsymbol{X})} - \underbrace{\frac{1}{\lambda}}_{g(\lambda)} \Big).$$

Thus $\bar{X} = n^{-1} \sum X_i$ is an unbiased estimator of $g(\lambda) = 1/\lambda$ which attains the CRLB for $1/\lambda$, and

$$\mathrm{Var}(\bar{X}) = \underbrace{(-1/\lambda^2)}_{g'(\lambda)} \Big/ \underbrace{(-n)}_{\alpha(\lambda)} = (n\lambda^2)^{-1}.$$

Theorem If $\tilde{\theta}$ is an unbiased estimator of a parameter θ that attains the CRLB, then $\tilde{\theta}$ is the unique maximum likelihood estimator of θ.

Proof By the previous corollary,

$$\frac{\partial}{\partial\theta}\log p_{\boldsymbol{X}}(\boldsymbol{X}\,|\,\theta) = \alpha(\theta)[\tilde{\theta} - \theta],$$

where the slope $\alpha(\theta) \neq 0$, so that the equation

$$\frac{\partial}{\partial\theta}\log p_{\boldsymbol{X}}(\boldsymbol{X}\,|\,\theta) = 0$$

has the unique solution $\widehat{\theta} = \tilde{\theta}$, where $\hat{\theta}$ denotes the MLE. ∎

We note that the converse is not true! An MLE need not be unbiased; besides, unbiasedness + uniqueness of MLE does not imply attainability of the CRLB.

2.6 Rao-Blackwell theory

So what can we do if the CRLB is not attainable?

Definition A minimum variance unbiased estimator (MVUE) of $g(\theta)$ is an unbiased estimator with smallest possible variance.

How do we find one (or the) MVUE?

Rao-Blackwell Theorem Let $\widehat{g}(\boldsymbol{X})$ be an unbiased estimator of $g(\theta)$. Then, if $T = T(\boldsymbol{X})$ is a

sufficient statistic for θ, the estimator $\tilde{g}(T)$ defined by

$$\tilde{g}(t) = \mathbb{E}[\widehat{g}(\boldsymbol{X}) \,|\, T(\boldsymbol{X}) = t]$$

also defines an unbiased estimator of $g(\theta)$ and

$$\mathrm{Var}(\tilde{g}(T)) \leq \mathrm{Var}(\widehat{g}(\boldsymbol{X})).$$

Proof First, note that since T is sufficient for θ, the distribution of $\boldsymbol{X} \,|\, T(\boldsymbol{X}) = t$ is independent of θ and hence so is $\tilde{g}(T)$ (independent on θ), making it an estimator indeed.

Now, since $\widehat{g}(\boldsymbol{X})$ is unbiased, the law of iterated expectations $\mathbb{E}[A] = \mathbb{E}[\mathbb{E}[A|B]]$ yields

$$g(\theta) = \mathbb{E}_X[\widehat{g}(\boldsymbol{X})] = \mathbb{E}_T\left[\mathbb{E}_X[\widehat{g}(\boldsymbol{X})|T]\right] = \mathbb{E}_T[\tilde{g}(T)],$$

so that $\tilde{g}(T)$ is unbiased.

Finally, recall Jensen's inequality: if h is a convex function then $\mathbb{E}[h(Z)] \geq h(\mathbb{E}[Z])$.

$$\begin{aligned}
\mathrm{Var}(\widehat{g}(\boldsymbol{X})) &= \mathbb{E}_X\left[(\widehat{g}(\boldsymbol{X}) - g(\theta))^2\right] \\
&= \mathbb{E}_T\left[\mathbb{E}_X\left[(\widehat{g}(\boldsymbol{X}) - g(\theta))^2 | T\right]\right] \\
&\geq \mathbb{E}_T\left[\left\{\mathbb{E}_X[\widehat{g}(\boldsymbol{X})|T] - g(\theta)\right\}^2\right] \\
&= \mathbb{E}_T\left[\left\{\tilde{g}(T) - g(\theta)\right\}^2\right]
\end{aligned}$$

by Jensen's inequality on the line before last with $h(x) = x^2$. The right-hand side is exactly $\mathrm{Var}(\tilde{g}(T))$ so that

$$\mathrm{Var}(\widehat{g}(\boldsymbol{X})) \geq \mathrm{Var}(\tilde{g}(T))$$

as announced. ∎

If completeness of T is added, the Rao-Blackwell construction actually gives the unique MVUE.

Lehmann-Scheffé Theorem Suppose there exists an unbiased estimator $\widehat{g}(\boldsymbol{X})$ of $g(\theta)$ and a statistic T which is sufficient and complete for θ. Then

$$\tilde{g}(t) = \mathbb{E}[\widehat{g}(\boldsymbol{X}) \,|\, T(\boldsymbol{X}) = t]$$

is the unique MVUE of $g(\theta)$.

Proof (not examinable)

Let $\widehat{g}_1(\boldsymbol{X})$ and $\widehat{g}_2(\boldsymbol{X})$ be any two unbiased estimators of $g(\theta)$. For $i = 1, 2$ define

$$\tilde{g}_i(t) = \mathbb{E}[\widehat{g}_i(\boldsymbol{X}) \,|\, T(\boldsymbol{X}) = t]$$

so that $\mathbb{E}[\tilde{g}_i(T)] = g(\theta)$ and thus

$$\mathbb{E}[\tilde{g}_1(T) - \tilde{g}_2(T)] = 0 \quad \text{for all } \theta \in \Theta.$$

Since T is complete, applying the definition of completeness to the function $\tilde{g}_1 - \tilde{g}_2$ entails

$$\mathbb{P}[\tilde{g}_1(T) - \tilde{g}_2(T) = 0 \,|\, \theta] = 1.$$

Thus whatever unbiased estimator $\widehat{g}(\boldsymbol{X})$ we use, $\tilde{g}(T)$ is always the same (recall also slide 24!)

By the Rao-Blackwell Theorem,

$$\mathrm{Var}(\tilde{g}(T)) \leq \mathrm{Var}(\widehat{g}(\boldsymbol{X}))$$

whatever the unbiased estimator $\widehat{g}(\boldsymbol{X})$ is. Thus $\tilde{g}(T)$ is unbiased and has a variance not larger than any other unbiased estimator, *i.e.* $\tilde{g}(T)$ is the unique MVUE of $g(\theta)$. ∎

Corollary If there is a statistic T which is sufficient and complete, then any unbiased function of T is the MVUE.

Summary To find the MVUE of $g(\theta)$:

(0) Check if there exists an estimator of $g(\theta)$ which attains the CRLB. If not, then:

(1) Find a sufficient statistic T which is also complete.

Then either:

(2a) Find a simple (often very crude!) unbiased estimator of $g(\theta)$, $\widehat{g}(\boldsymbol{X})$ say; compute

$$\tilde{g}(t) = \mathbb{E}[\widehat{g}(\boldsymbol{X}) \,|\, T(\boldsymbol{X}) = t],$$

and then $\tilde{g}(T)$ is the MVUE;

(2b) Or find an estimator function of T, $\tilde{g}(T)$ say (the MLE may be a good starting point,) which is unbiased for $g(\theta)$; then $\tilde{g}(T)$ is the MVUE.

Example If $X_1, \ldots, X_n$ are iid Exp(λ) with mean $1/\lambda$, then

$$p_{\boldsymbol{X}}(\boldsymbol{X}|\,\lambda) = \prod_{i=1}^{n} \lambda \mathrm{e}^{-\lambda x_i} = \lambda^n \mathrm{e}^{-\lambda \sum x_i}.$$

(a) Estimating λ^{-1}.

Note that $T = \sum X_i$ is sufficient for λ^{-1}, and we have already shown that this T is complete.

Now X_1 is a simple unbiased estimator for λ^{-1}. Consider

$$\tilde{g}(t) = \mathbb{E}[\, X_1 \,|T = t\,].$$

To calculate $\tilde{g}$ we need the conditional pdf

$$p_{X_1|T}(x|t) = \frac{p_{X_1,T}(x,t)}{p_{_T}(t)}.$$

First note that

$$\{X_1 = x, T = t\} = \left\{ X_1 = x, \sum_{i=2}^{n} X_i = t - x \right\}.$$

Recalling that the sum of m iid Exp(λ) random variables has a Gamma(m, λ) distribution now gives

$$\begin{aligned} \frac{p_{X_1,T}(x,t)}{p_{_T}(t)} &= \\ &= \frac{[\lambda \mathrm{e}^{-\lambda x}][\lambda^{n-1}\,(t-x)^{n-2}\,\mathrm{e}^{-\lambda(t-x)}/\Gamma(n-1)]}{\lambda^n\, t^{n-1}\,\mathrm{e}^{-\lambda t}/\Gamma(n)} \\ &= t^{-1}(n-1)[1-(x/t)]^{n-2}, \end{aligned}$$

using that $\Gamma(n) = (n-1)\Gamma(n-1)$.

Thus

$$\begin{aligned} E[X_1|T=t] = \tilde{g}(t) &= \int_0^t x p_{X_1|T}(x|t)\,dx \\ &= \int_0^t x t^{-1}(n-1)(1-x/t)^{n-2}\,dx \\ &= \frac{t}{n} \quad \text{(after some calculation)} \end{aligned}$$

and thus $\mathbb{E}[X_1|T] = T/n$.

Thus the MVUE of λ^{-1} is given by

$$\tilde{g}(T) = \frac{\sum X_i}{n} = \bar{X} \quad \text{i.e. sample mean.}$$

Note that we showed earlier that this estimator attains the CRLB.

(b) Estimating λ.

The common pdf of the X_i is

$$p(x|\lambda) = \lambda e^{-\lambda x} = \exp(-\lambda x + \log\lambda), \quad x > 0.$$

Note that, by the theorems for sufficiency and completeness in the exponential family, $T = \sum X_i$ is sufficient for λ, and complete.

Since $\mathbb{E}[T] = \mathbb{E}[\sum X_i] = \sum \mathbb{E}[X_i] = n\frac{1}{\lambda}$, (ie a function of $\frac{1}{\lambda}$) it makes sense to consider T^{-1}.

Since $T \sim \text{Gamma}(n, \lambda)$,

$$\begin{aligned}\mathbb{E}[T^{-1}] &= \int_0^\infty t^{-1} \frac{1}{\Gamma(n)} \lambda^n t^{n-1} e^{-\lambda t}\, dt \\ &= \lambda \frac{\Gamma(n-1)}{\Gamma(n)} \underbrace{\int_0^\infty \frac{1}{\Gamma(n-1)} \lambda^{n-1} t^{n-2} e^{-\lambda t}\, dt}_{=1 \quad \text{pdf of a } Ga(n-1,\lambda)} \\ &= \frac{\lambda}{n-1} \neq \lambda.\end{aligned}$$

The estimator $\tilde{g}(T) = (n-1)T^{-1}$ is then unbiased and is a function of the sufficient and complete statistic T.

Consequently

$$\tilde{g}(T) = \frac{n-1}{T} = \frac{n-1}{\sum X_i}$$

is the MVUE for λ.

It can be shown similarly (exercise!) that

$$\text{Var}(\tilde{g}(T)) = \frac{\lambda^2}{n-2}$$

and that the CRLB is λ^2/n. Thus

$$\text{Var}(\tilde{g}(T)) = \frac{\lambda^2}{n-2} > \frac{\lambda^2}{n}$$

meaning that $\tilde{g}(T)$ does not attain the CRLB, but it does "asymptotically attain" the CRLB since

$$\frac{\text{Var}(\tilde{g}(T))}{\text{CRLB}} = \frac{n}{n-2} \to 1 \text{ as } n \to \infty.$$

Note The concepts of CRLB and MVUE only apply to unbiased estimators!

Proof of the asymptotic normality of the MLE

The proof is not examinable. First of all, we assume consistency. We look at the case of a univariate θ. The idea is to write a Taylor expansion of the score statistic:

$$\frac{\partial l}{\partial \theta}(\widehat{\theta}|\boldsymbol{X}) = \frac{\partial l}{\partial \theta}(\theta|\boldsymbol{X}) + (\widehat{\theta} - \theta)\frac{\partial^2 l}{\partial \theta^2}(\theta|\boldsymbol{X}) + (\widehat{\theta} - \theta)R(\widehat{\theta} - \theta)$$

where $R = R(x)$ is a function converging to 0 as $x \to 0$, *i.e.* a small remainder term. Now, the left-hand side of the above equality is 0 because $\widehat{\theta}$ is the MLE, and the last term on the right is negligible. Rearranging yields, for large n,

$$\sqrt{n}(\widehat{\theta} - \theta) \times \left[-\frac{1}{n} \frac{\partial^2 l}{\partial \theta^2}(\theta | \boldsymbol{X}) \right] \approx \frac{1}{\sqrt{n}} \frac{\partial l}{\partial \theta}(\theta | \boldsymbol{X}).$$

Now

$$\begin{aligned} -\frac{1}{n} \frac{\partial^2 l}{\partial \theta^2}(\theta | \boldsymbol{X}) &= \frac{1}{n} \sum_{i=1}^{n} -\frac{\partial^2}{\partial \theta^2} \log p_X(X_i | \theta) \\ &\to \mathbb{E}\left[-\frac{\partial^2}{\partial \theta^2} \log p_X(X | \theta) \right] \end{aligned}$$

as $n \to \infty$ by the Law of Large Numbers.

In other words,

$$-\frac{1}{n} \frac{\partial^2 l}{\partial \theta^2}(\theta | \boldsymbol{X}) \to i(\theta) \text{ as } n \to \infty.$$

Besides

$$\frac{1}{\sqrt{n}} \frac{\partial l}{\partial \theta}(\theta | \boldsymbol{X}) = \sqrt{n} \left[\frac{1}{n} \sum_{i=1}^{n} \underbrace{\frac{\partial}{\partial \theta} \log p_X(X_i | \theta)}_{Y_i} \right].$$

The Y_i, being single-observation versions of the score statistic, are IID random variables such that

$$\mathbb{E}(Y_i) = 0 \text{ and } \operatorname{Var}(Y_i) = i(\theta),$$

see the Theorem on slide 11. By the Central Limit Theorem,

$$\frac{1}{\sqrt{n}}\frac{\partial l}{\partial \theta}(\theta|\boldsymbol{X}) \approx N(0, i(\theta)) \text{ as } n \to \infty.$$

Combining our two convergences, we obtain the required result:

$$\sqrt{n}(\widehat{\theta} - \theta) \approx [i(\theta)]^{-1} N(0, i(\theta)) = N(0, i(\theta)^{-1})$$

(use the relationship $a \times N(0, v) = N(0, a^2 v)$). ∎

Chapter 3

Hypothesis testing

3.1 Framework

Suppose the parameter space Θ is partitioned into two subsets:

$$\Theta = \Theta_0 \cup \Theta_1, \qquad \Theta_0 \cap \Theta_1 = \emptyset.$$

We wish to determine whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$ based on data. The former is called the null hypothesis, H_0, while the latter is the alternative hypothesis, H_1.

Example If $x_1, \ldots, x_n$ are n IID observations from a Bernoulli distribution with success probability θ, an interesting test is

$$H_0 : \ \theta = 1/2 \quad \text{versus} \quad H_1 : \ \theta \neq 1/2$$

i.e. fair coin against biased coin.

For a given test there are two actions: *Reject* H_0 or *Do not reject* H_0 Therefore, there are two types of error, namely:

Type I error: Reject H_0 when it is true.

Type II error: Do not reject H_0 when it is false.

The significance level or size of a test is the probability of a Type I error:

$$\alpha = \mathbb{P}(\text{reject } H_0 \,|\, H_0 \text{ true}).$$

For $\theta \in \Theta_1$, the power $1 - \beta(\theta)$ of a test is the probability of rejecting H_0 when it is false and the parameter is equal to θ. In other words, for $\theta \in \Theta_1$, $\beta(\theta)$ is the probability of Type II error:

$$\beta(\theta) = \mathbb{P}(\text{do not reject } H_0 \,|\, \theta)$$

and therefore the power is:

$$1 - \beta(\theta) = \mathbb{P}(\text{reject } H_0 \,|\, \theta), \quad \theta \in \Theta_1.$$

It is desirable for size to be small and power large; but they are inversely related so instead we fix a size $\alpha \in (0, 1)$ and look for the Most Powerful Test (MP) of size α, *i.e.* the test of size α having highest power.

A hypothesis H is called simple if H only contains one value; otherwise, it is called composite.

A convenient mathematical way to address testing is via a test function ϕ:

$$\phi(\boldsymbol{x}) = \mathbb{P}(\text{reject } H_0 | \boldsymbol{X} = \boldsymbol{x}).$$

In other words, if the sample of observed data is $\boldsymbol{x}$ then H_0 is rejected with probability $\phi(\boldsymbol{x})$ (which is a number between 0 and 1).

That can seem weird: intuitively, it would make more sense, given a value of $\boldsymbol{x}$, to always reject ($\phi(\boldsymbol{x}) = 1$) or always not reject H_0 ($\phi(\boldsymbol{x}) = 0$).

We will see very shortly that the theoretically optimal test function ϕ is in fact "almost always" equal to 0 or 1, and is moreover everywhere 0 or 1 when the statistical model is continuous.

In the particular case when $\phi(\boldsymbol{x}) = 0$ or 1 for all $\boldsymbol{x} \in \Omega$, then the test is said to be non-randomised.

3.2 Comparing two simple hypotheses

Neyman–Pearson Lemma

(i) If $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$ and $\alpha \in (0, 1)$ are fixed then there are constants $k > 0$ and $\gamma \in [0, 1]$ such that the test function below induces a test of size α:

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & \text{when } p_{\boldsymbol{X}}(\boldsymbol{x}|\theta_0) < k p_{\boldsymbol{X}}(\boldsymbol{x}|\theta_1), \\ \gamma & \text{when } p_{\boldsymbol{X}}(\boldsymbol{x}|\theta_0) = k p_{\boldsymbol{X}}(\boldsymbol{x}|\theta_1), \\ 0 & \text{when } p_{\boldsymbol{X}}(\boldsymbol{x}|\theta_0) > k p_{\boldsymbol{X}}(\boldsymbol{x}|\theta_1) \end{cases}$$

i.e. we reject H_0 with probability 1 when

$$\lambda(\boldsymbol{x}) = \text{likelihood ratio} = \frac{p_{\boldsymbol{X}}(\boldsymbol{x}|H_0)}{p_{\boldsymbol{X}}(\boldsymbol{x}|H_1)} < k,$$

and we reject H_0 with probability γ if $\lambda(\boldsymbol{x}) = k$.

(ii) For fixed k, the test in (i) is a most powerful test of size

$$\begin{aligned} \alpha &= \mathbb{E}[\phi(\boldsymbol{X})|H_0] \\ &= \mathbb{P}(\lambda(\boldsymbol{X}) < k \,|\, H_0) + \gamma \mathbb{P}(\lambda(\boldsymbol{X}) = k \,|\, H_0) \end{aligned}$$

for testing H_0 versus H_1.

Comments

(a) The optimal test consists in rejecting H_0 in favour of H_1 when the likelihood ratio in favour of H_0 against H_1 is small enough. How small the rejection level k should be is determined using the equation

$$\alpha = \mathbb{P}(\lambda(\boldsymbol{X}) < k \,|\, H_0) + \gamma \mathbb{P}(\lambda(\boldsymbol{X}) = k \,|\, H_0).$$

(b) If $\boldsymbol{X}$ has a probability density function then the probability $\mathbb{P}(\lambda(\boldsymbol{X}) = k \,|\, H_0)$ is always zero, no matter what k is. Consequently, we may always take $\gamma = 0$ or 1 in this case, and the MP test is non-randomised. We then say that the set

$$C = \{\boldsymbol{x} : \ \phi(\boldsymbol{x}) = 1\}$$

is the critical region (or rejection region) of the test. Its complement is (sometimes) called the acceptance region.

(c) It is, therefore, only when $\boldsymbol{X}$ is discrete that we may need to consider $\gamma \in (0, 1)$. This is because in the discrete case, non-randomised tests miss certain values of α. A test with $\gamma \in (0, 1)$ is an example of a randomised test.

(d) The test function can be rewritten in terms of a sufficient statistic (by the Neyman Factorisation Theorem).

Example 1 A sample of n independent observations $X_1, \ldots, X_n$ is available from a Normal distribution with mean μ and variance 100:

$$X_1, \ldots, X_n \sim N(\mu, 100).$$

Suppose $n = 25$, find the most powerful 5%-level test (i.e. assume $\alpha = 0.05$) of the null hypothesis

$$H_0 : \mu = 100$$

against the alternative

$$H_1 : \mu = 105$$

and compute the power of this test.

Solution To determine the required test we apply the Neyman–Pearson (NP) lemma. The likelihood ratio is

$$\begin{aligned} \lambda(\boldsymbol{x}) &= \frac{p_{\boldsymbol{X}}(\boldsymbol{x}|H_0)}{p_{\boldsymbol{X}}(\boldsymbol{x}|H_1)} \\ &= \frac{(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_0)^2\right\}}{(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_1)^2\right\}}. \end{aligned}$$

X is continuous which means that we do not need to worry about γ (in the NP Lemma) and hence we reject H_0 if $\lambda(\boldsymbol{x}) < k$.

Thus, with $\mu_0 = 100 < \mu_1 = 105$, $\lambda(\boldsymbol{x}) < k$ if and only if (by taking logs and simplifying):

$$2(\mu_0 - \mu_1)\sum x_i + n(\mu_1^2 - \mu_0^2) < 2\sigma^2 \log k = k_1$$

$$\Leftrightarrow \sum x_i > \frac{1}{2(\mu_0 - \mu_1)}\left(k_1 - n(\mu_1^2 - \mu_0^2)\right) = k_2$$

(for the last inequality, note that $\mu_0 - \mu_1 < 0$). Note that k_1 and k_2 are constants that do no depend on the data $\boldsymbol{x}$.

Thus an MP test of size α rejects H_0 if

$$\sum x_i > k_2,$$

where k_2 satisfies

$$P\left(\sum x_i > k_2 \,|\, H_0\right) = \alpha = 0.05.$$

To find k_2 note that, with $\mu_0 = 100$, $\mu_1 = 105$, $\sigma^2 = 100$ then under H_0 (why?):

$$\bar{X} = n^{-1} \sum X_i \sim N(100, 100/n).$$

Let $k_3 = n^{-1} k_2$. Then k_3 is such that

$$\mathbb{P}\left(\frac{\bar{X} - 100}{\sqrt{100/n}} > \frac{k_3 - 100}{\sqrt{100/n}} \right) = 0.05$$

that is

$$\mathbb{P}(Z > (k_3 - 100)/\sqrt{100/n}) = 0.05,$$

where Z is a Normal(0,1) random variable.

From tables (or using `R`) we find that $\Phi(1.645) = 0.95$ (here Φ is the distribution function (cdf) of the standard Normal) and hence

$$(k_3 - 100)/\sqrt{100/n} = 1.645,$$

from which $k_3 = 103.29$ with $n = 25$.

Thus the test is to reject H_0 if and only if $\bar{x} > 103.29$.

Finally, the power of the test is

$$\begin{aligned}
1-\beta &= P(\text{reject } H_0 \,|\, \text{assuming } H_1 \text{ is true}) \\
&= \mathbb{P}(\bar{X} > 103.29 \,|\, \mu = 105) \\
&= \mathbb{P}\left(\frac{\bar{X}-105}{\sqrt{100/n}} > \frac{103.29-105}{\sqrt{100/n}} \,|\, \mu = 105\right) \\
&= \mathbb{P}(Z > -0.855) \qquad \text{(can find from Tables)} \\
&= \Phi(0.855) = 0.8037. \qquad \text{(power of the test)}
\end{aligned}$$

Example 2 Let us look at an example where X is discrete – let $X_1, \ldots, X_8$ be eight IID Bernoulli trials with success probability θ.

Find a most powerful test of

$$H_0 : \theta = 1/2$$

versus

$$H_1 : \theta = 1/4,$$

having size (significance level α) 1/16.

Solution Apply the Neyman-Pearson lemma: with $\theta_0 = 1/2$, $\theta_1 = 1/4$ and $n = 8$, the likelihood ratio is

$$\lambda(\boldsymbol{x}) = \frac{p_{\boldsymbol{X}}(\boldsymbol{x}|\theta_0)}{p_{\boldsymbol{X}}(\boldsymbol{x}|\theta_1)} = \left(\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)}\right)^{\sum x_i} \left(\frac{1-\theta_0}{1-\theta_1}\right)^n.$$

Here $\theta_0(1-\theta_1) - \theta_1(1-\theta_0) = \theta_0 - \theta_1 > 0$ so that

$$\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)} > 1,$$

and thus λ is increasing in $\sum x_i$.

Therefore:

$$\lambda(\boldsymbol{x}) < k \;\Leftrightarrow\; \sum x_i < k_1$$

for some constant k_1 where $\sum x_i$ is the number of successes .

By the Neyman-Pearson lemma, a test of H_0 versus H_1 of the form

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & \text{when } \sum x_i < k_1, \\ \gamma & \text{when } \sum x_i = k_1, \\ 0 & \text{when } \sum x_i > k_1, \end{cases}$$

is the MP test of size α where

$$\begin{aligned} \alpha &= \mathbb{E}[\phi(\boldsymbol{X}) \,|\, H_0] \\ &= \mathbb{P}(\sum X_i < k_1 \,|\, H_0) + \gamma \mathbb{P}(\sum X_i = k_1 \,|\, H_0). \end{aligned}$$

We now determine k_1 and γ.

Under H_0, $\sum X_i \sim \text{Binomial}(8, 1/2)$.

Hence, under H_0,

$$\mathbb{P}\left(\sum_{i=1}^{8} X_i < 1\right) = \mathbb{P}\left(\sum_{i=1}^{8} X_i = 0\right) = \left(\frac{1}{2}\right)^8 = \frac{1}{256}$$

$$\mathbb{P}\left(\sum_{i=1}^{8} X_i < 2\right) = \mathbb{P}\left(\sum_{i=1}^{8} X_i \leq 1\right) = \left(\frac{1}{2}\right)^8 + 8\left(\frac{1}{2}\right)^8$$

$$= \frac{9}{256} < \frac{1}{16} = \alpha,$$

$$\mathbb{P}\left(\sum_{i=1}^{8} X_i < 3\right) = \mathbb{P}\left(\sum_{i=1}^{8} X_i \leq 2\right) = \frac{37}{256} > \frac{1}{16} = \alpha.$$

Thus an MP test of size $1/16$ for H_0 versus H_1 is

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & \text{when } \sum x_i < 2, \\ \gamma & \text{when } \sum x_i = 2, \\ 0 & \text{when } \sum x_i > 2, \end{cases}$$

where γ is determined by ensuring that

$$\alpha = \frac{1}{16} = \mathbb{P}\left(\sum X_i < 2 \,|\, H_0\right) + \gamma \mathbb{P}\left(\sum X_i = 2 \,|\, H_0\right),$$

which gives $\gamma = 1/4$.

Hence the MP test:

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & \text{when } \sum x_i < 2, \\ 1/4 & \text{when } \sum x_i = 2, \\ 0 & \text{when } \sum x_i > 2. \end{cases}$$

Notes: (a) This is an example of a randomised test.

(b) To use this test, one would always reject H_0 if $\sum x_i < 2$, never reject H_0 if $\sum x_i > 2$, and reject it with probability $\gamma = 1/4$ if $\sum x_i = 2$ (by flipping a biased coin).

(c) While randomised tests are a nice theoretical solution, they are very rarely used in practice. This is because two practitioners can have exactly the same data and arrive at two contrary conclusions!

3.3 Testing a simple vs a composite hypothesis

The Neyman-Pearson lemma provides a nice and optimal solution when H_1 is simple. But what about a problem of the form $H_0 : \theta = \theta_0$ versus $H_1 : \theta \in \Theta_1$, for instance

$$H_0 : \ \theta = 1/2 \quad \text{versus} \quad H_1 : \ \theta > 1/2$$

for IID Bernoulli trials?

Definition A test of size α is uniformly most powerful (UMP) if, for all $\theta \in \Theta_1$, its power is greater than or equal to the power of any other test of size α.

Lemma There exists a UMP test of size α for H_0 versus H_1 if and only if there is a test ϕ that is the MP test of H_0 versus $H_1' : \theta = \theta_1$ for all $\theta_1 \in \Theta_1$. In this case, the UMP test is exactly the test ϕ.

In practice, this Lemma reduces the search of a UMP test to the computation of the Neyman-Pearson MP test, although it has to be independent of $\theta_1 \in \Theta_1$.

Example Let $X_1, \ldots, X_n$ be IID $N(\theta, 1)$. Is there a UMP test of size α for $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$?

Solution In Example 1 we saw that the MP test of size α of $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1 > \theta_0$ rejects H_0 when $\sum X_i > k_2$, hence when $\bar{X}n > k_2$ or, equivalently when $\bar{X} > k_3$ where $k_3 = k_2/n$.

Under H_0:

$$\bar{X} \sim N\left(\theta_0, \frac{1}{n}\right)$$

and therefore (by standardising):

$$\mathbb{P}\left(\frac{\bar{X} - \theta_0}{\sqrt{1/n}} > \frac{k_3 - \theta_0}{\sqrt{1/n}} \,|\, \theta = \theta_0\right) = \alpha.$$

Namely, if Z is $N(0,1)$, k_3 is such that

$$P[Z > (k_3 - \theta_0)/(1/\sqrt{n})] = \alpha.$$

Thus, defining Φ^{-1} as the inverse of the $N(0,1)$ cdf (*i.e.* its quantile function),

$$\frac{k_3 - \theta_0}{1/\sqrt{n}} = \Phi^{-1}(1-\alpha),$$

so that

$$k_3 = \theta_0 + (1/\sqrt{n})\Phi^{-1}(1-\alpha).$$

Hence the MP test rejects H_0 when $\bar{X} > k_3$, *i.e.*

$$\sum x_i > n\theta_0 + \sqrt{n}\Phi^{-1}(1-\alpha).$$

But this test is independent of θ_1, and so holds for any value of θ_1.

By the previous Lemma, this test is then the UMP test for $H_0 : \theta = \theta_0$ (simple) versus $H_1 : \theta > \theta_0$ (composite).

Example An IID sample of n observations is obtained from a Weibull distribution having pdf

$$f_X(x|\mu) = \mu^{-2} x e^{-x^2/2\mu^2}, \quad x \geq 0,\ \mu > 0.$$

Suppose we wish to find a UMP test for the problem

$$H_0 : \mu = \mu_0 \ \text{ versus } \ H_1 : \mu > \mu_0.$$

Consider then first the simple alternative

$$H_1' : \mu = \mu_1, \text{ for some } \mu_1 > \mu_0,$$

for which we are going to find an MP test by the Neyman-Pearson lemma. The likelihood ratio is

$$\begin{aligned}
\lambda(\boldsymbol{x}) &= \frac{L(\boldsymbol{x}|H_0)}{L(\boldsymbol{x}|H_1')} \\
&= \frac{\mu_0^{-2n}(\prod x_i)\exp\left(-\frac{1}{2\mu_0^2}\sum x_i^2\right)}{\mu_1^{-2n}(\prod x_i)\exp\left(-\frac{1}{2\mu_1^2}\sum x_i^2\right)} \\
&= \left(\frac{\mu_1}{\mu_0}\right)^{2n} \exp\left[-\left(\frac{1}{2\mu_0^2} - \frac{1}{2\mu_1^2}\right)\sum x_i^2\right].
\end{aligned}$$

Taking logs,

$$\log \lambda(\boldsymbol{x}) = 2n \log(\mu_1/\mu_0) - \left(\frac{1}{2\mu_0^2} - \frac{1}{2\mu_1^2}\right) \sum x_i^2$$

and thus the inequality $\lambda(\boldsymbol{x}) < k$ is equivalent to

$$-\left(\frac{\mu_1^2 - \mu_0^2}{2\mu_0^2\mu_1^2}\right) \sum x_i^2 < k_2 = \log k - 2n \log(\mu_1/\mu_0).$$

Since $\mu_1 > \mu_0$ this is equivalent to

$$\sum x_i^2 > k_3 = -\frac{2\mu_0^2\mu_1^2}{\mu_1^2 - \mu_0^2} k_2.$$

To determine k_3 we need to find the distribution of X^2 (not that easy/obvious).

It is actually more convenient to consider $Y = X^2/\mu^2$. Then $X = \mu\sqrt{Y}$ and

$$\begin{aligned} p_Y(y) &= p_X(\mu\sqrt{y}) \times \left|\frac{dY}{dX}\right|_{X=\mu\sqrt{y}}^{-1} \quad \text{[change of variable]} \\ &= \mu^{-2}(\mu\sqrt{y})e^{-y/2}\left(\frac{\mu}{2\sqrt{y}}\right), \text{ when } y > 0 \\ &= \frac{1}{2}e^{-y/2}, \text{ when } y > 0 \end{aligned}$$

and thus Y has a Exp(1/2) distribution.

[Recall that $f_X(x|\mu) = \mu^{-2}xe^{-x^2/2\mu^2}$]

Having considered $Y = X^2/2$ we can now rewrite $\sum x_i^2 > k_3$ as $\sum y_i > c_0 = k_3/\mu_0^2$.

To determine c_0 we need to know the distribution of the sum of n independent Exponential random variables (Y_i).

We can show that the sum of n independent Exponential random variables with the same rate follows a Gamma distribution: $\sum Y_i$ has a Gamma$(n, 1/2)$ distribution.

Having known what distribution does $\sum Y_i$ follows we should choose k_3 as $c_0\mu_0^2$ where

$$\mathbb{P}(\text{Gamma}(n, 1/2) > c_0) = \alpha.$$

There are no Tables available for all Gamma distributions, but for a given n and α we can use R to find c_0.

The test is then to reject H_0 if $\sum x_i^2 > c_0\mu_0^2$. Because this procedure is independent of μ_1, it is also the UMP test required.

As an example, suppose $n = 10$, $\mu_0 = 1$ and $\alpha = 0.05$. From tables we find that

$$\mathbb{P}(\text{Gamma}(n, 1/2) > 31.41) = 0.05$$

and thus by typing in R

```
qgamma(0.05, shape = 10, rate = 1/2, lower=FALSE)
```

we find that $c_0 = 31.41$ So we reject H_0 if

$$\sum x_i^2 > 31.41 \times 1^2 = 31.41.$$

3.4 Likelihood Ratio Tests

It can also happen that there is no UMP test.

Example If $X_1, \ldots, X_n$ are IID $N(\theta, 1)$, is there a UMP test of size α for

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0 \, ?$$

By the previous results there is a UMP test if and only if there is a common MP test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ for any $\theta_1 \neq \theta_0$.

But the MP test rejects H_0 if

$$\sum x_i > n\theta_0 + \sqrt{n}\Phi^{-1}(1-\alpha) \quad \text{if } \theta_1 > \theta_0,$$

$$\sum x_i < n\theta_0 - \sqrt{n}\Phi^{-1}(1-\alpha) \quad \text{if } \theta_1 < \theta_0,$$

(exercise!) *i.e.* tests are different depending on θ_1. It follows that there is no UMP test for this problem.

So what if no UMP test is available?

One solution is to use a Likelihood Ratio Test.

Suppose we wish to test

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \in \Theta_1,$$

where H_0 and H_1 may be simple or composite, and $\boldsymbol{\theta}$ can be a vector. If both are simple then we are back to the Neyman-Pearson Lemma.

Definition The likelihood ratio test statistic is given by

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta_1} p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})}.$$

or in other words

$$\lambda(\boldsymbol{x}) = \frac{\text{value of likelihood when maximised under } H_0}{\text{value of likelihood when maximised under} H_1}.$$

Definition The Likelihood Ratio Test (LRT) of size α of $H_0 : \boldsymbol{\theta} \in \Theta_0$ versus $H_1 : \boldsymbol{\theta} \in \Theta_1$ is given by

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & \text{when } \lambda(\boldsymbol{x}) < k, \\ \gamma & \text{when } \lambda(\boldsymbol{x}) = k, \\ 0 & \text{when } \lambda(\boldsymbol{x}) > k, \end{cases}$$

where k and γ are chosen to satisfy

$$\sup_{\boldsymbol{\theta} \in \Theta_0} \mathbb{P}(\text{reject } H_0 \,|\, \boldsymbol{\theta}) = \sup_{\boldsymbol{\theta} \in \Theta_0} \mathbb{E}[\phi(\boldsymbol{X}) \,|\, \boldsymbol{\theta}] = \alpha.$$

This test is a generalisation of the MP test for simple hypotheses in the Neyman-Pearson lemma. Note that when H_0 and H_1 are both simple, the LRT is exactly the same as the MP test.

Example Suppose $X_1, \ldots, X_n$ are IID $N(\theta, 1)$.

Find the LRT of size α for

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0.$$

The likelihood is

$$p_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = (2\pi)^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (x_i - \theta)^2 \right\}$$

and the likelihood ratio is

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\theta=\theta_0} p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}{\sup_{\theta\neq\theta_0} p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)} = \frac{p_{\boldsymbol{X}}(\boldsymbol{x}|\theta_0)}{\sup_{\theta\neq\theta_0} p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)}.$$

Now the denominator is the maximum possible value of $p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)$, that is:

$$\sup_{\theta\neq\theta_0} p_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = p_{\boldsymbol{X}}(\boldsymbol{x}|\widehat{\theta}),$$

where $\widehat{\theta}$ is the MLE of θ. Here $\widehat{\theta} = \bar{x}$ (exercise!), so

$$p_{\boldsymbol{X}}(\boldsymbol{x}|\widehat{\theta}) = (2\pi)^{-n/2} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 \right\}.$$

The likelihood ratio is then

$$\lambda(\boldsymbol{x}) = \exp\left\{ -\frac{1}{2}\left[\sum_{i=1}^{n}(x_i - \theta_0)^2 - \sum_{i=1}^{n}(x_i - \bar{x})^2 \right] \right\}.$$

Expanding and simplifying, this is

$$\lambda(\boldsymbol{x}) = \exp\left\{ -\frac{n}{2}(\bar{x} - \theta_0)^2 \right\}.$$

We reject H_0 if and only if $\lambda(\boldsymbol{x}) < k$:

$$\exp\left\{-\frac{n}{2}(\bar{x}-\theta_0)^2\right\} < k \Leftrightarrow |\bar{x}-\theta_0| > K,$$

where K satisfies

$$\sup_{\theta\in\Theta_0} \mathbb{P}(\text{Reject } H_0 \,|\, H_0 \text{ true}) = \alpha,$$

$$\sup_{\theta\in\Theta_0} \mathbb{P}(|\bar{X}-\theta_0| > K \,|\, \theta) = \alpha,$$

$$\textit{i.e.} \quad \mathbb{P}(|\bar{X}-\theta_0| > K \,|\, \theta=\theta_0) = \alpha.$$

Under H_0, $\bar{X}-\theta_0 \sim N(0, n^{-1})$. Let again Φ be the cdf of the $N(0,1)$ distribution:

$$\begin{aligned}
& \mathbb{P}(|\bar{X}-\theta_0| > K \,|\, \theta=\theta_0) = \alpha \\
\Leftrightarrow\ & \mathbb{P}(\{\bar{X}-\theta_0 > K\} \text{ or } \{\bar{X}-\theta_0 < -K\} \,|\, \theta=\theta_0) = \alpha \\
\Leftrightarrow\ & 2\mathbb{P}(\bar{X}-\theta_0 < -K \,|\, \theta=\theta_0) = \alpha \qquad \text{(by symmetry)} \\
\Leftrightarrow\ & 2\Phi(-K/(1/\sqrt{n}\,)) = \alpha \\
\Leftrightarrow\ & K = -n^{-1/2}\Phi^{-1}(\alpha/2).
\end{aligned}$$

The LRT of size α is then

$$\phi(\boldsymbol{x}) = \begin{cases} 1 \text{ if } |\bar{x} - \theta_0| > -n^{-1/2}\Phi^{-1}(\alpha/2), \\ 0 \text{ if } |\bar{x} - \theta_0| \leq -n^{-1/2}\Phi^{-1}(\alpha/2), \end{cases}$$

i.e. reject if and only if $|\bar{x} - \theta_0| > -n^{-1/2}\Phi^{-1}(\alpha/2)$.

3.5 A particular problem: nested hypotheses

Suppose here that $\boldsymbol{\theta}$ is a vector of parameters with $q > 0$ components, and consider the sets

$$\begin{aligned} \Theta_0 &= \{\boldsymbol{\theta} : \boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\mu}),\ \boldsymbol{\theta}_0 \in \mathbb{R}^p\}, \\ \Theta_1 &= \left\{\boldsymbol{\theta} : \boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1),\ \boldsymbol{\theta}_0 \in \mathbb{R}^p,\ \boldsymbol{\theta}_1 \in \mathbb{R}^{q-p} \setminus \{\boldsymbol{\mu}\}\right\}. \end{aligned}$$

Here, $\boldsymbol{\mu}$ is a fixed vector having $q - p$ components. The integer $p \geq 0$ is thus the number of free parameters in the restricted statistical model offered by the space of parameters Θ_0, which is nested in the space Θ_1 of parameters having q free components.

Consider then the testing problem

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad \text{versus} \quad H_1 : \boldsymbol{\theta} \in \Theta_1.$$

Example Suppose $X_1, \ldots, X_n$ are IID $N(\theta, 1)$. The testing problem

$$H_0 : \theta = 0 \text{ versus } H_1 : \theta \neq 0$$

is one such problem, with $p = 0$, $q = 1$, $\boldsymbol{\mu} = 0$.

Wilks' Theorem Under general conditions, when n is large, the likelihood ratio

$$\lambda(\boldsymbol{X}) = \frac{\sup_{\boldsymbol{\theta}\in\Theta_0} p_{\boldsymbol{X}}(\boldsymbol{X}|\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\Theta_1} p_{\boldsymbol{X}}(\boldsymbol{X}|\boldsymbol{\theta})}$$

is such that $-2\log\lambda(\boldsymbol{X}) \approx \chi^2_{q-p}$ under H_0.

The case $\boldsymbol{\mu} = \mathbf{0}_{q-p}$ makes Wilks' Theorem interesting in model selection: not rejecting H_0 means that a simpler model (*i.e.* with fewer free components) might be favoured.

Example In a simple linear regression,

$Y_i = \alpha+\beta x_i+\varepsilon_i$ with $\varepsilon_i \sim N(0,\sigma^2)$, $i = 1, 2, \ldots, n$

we could test $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ by approximating $-2\log\lambda(\boldsymbol{X})$ by a χ^2_1 distribution.

It should be noted that Wilks' Theorem involves an approximation. As such, it is only suitable for large n and there is an uncertainty on its actual performance.

Example Suppose $X_1, \ldots, X_n$ are IID $N(\theta, 1)$. In the testing problem

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0$$

we saw that $p = 0$ (because θ_0 is fixed and known) and $q = 1$. Besides

$$\lambda(\boldsymbol{X}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} p_{\boldsymbol{X}}(\boldsymbol{X}|\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta_1} p_{\boldsymbol{X}}(\boldsymbol{X}|\boldsymbol{\theta})} = \frac{p_{\boldsymbol{X}}(\boldsymbol{X}|\theta = \theta_0)}{p_{\boldsymbol{X}}(\boldsymbol{X}|\theta = \widehat{\theta})}$$

where $\widehat{\theta}$ is the MLE of θ. Because

$$p_{\boldsymbol{X}}(\boldsymbol{X}|\boldsymbol{\theta}) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (X_i - \theta)^2 \right)$$

and $\widehat{\theta} = \bar{X} = n^{-1} \sum_{i=1}^{n} X_i$, we showed (see earlier)

$$\lambda(\boldsymbol{X}) = \exp\left(-\frac{n}{2} (\bar{X} - \theta_0)^2 \right).$$

Wilks' theorem predicts $-2 \log \lambda(\boldsymbol{X}) \approx \chi^2_1$ for n large.

In fact,

$$-2\log\lambda(\boldsymbol{X}) = n(\bar{X} - \theta_0)^2$$

and $\bar{X}$ has a $N(\theta_0, 1/n)$ distribution under H_0, so that

$$n(\bar{X} - \theta_0)^2 = \left[\sqrt{n}\left\{\bar{X} - \theta_0\right\}\right]^2 \stackrel{d}{=} [N(0,1)]^2 = \chi^2_1.$$

In this case, Wilks' theorem is actually exact.

3.6 An equivalent view: Confidence sets

Definition A $100(1-\alpha)\%$ confidence set $C(\boldsymbol{X})$ for the parameter θ is a random set such that

$$\mathbb{P}(\theta_0 \in C(\boldsymbol{X}) \,|\, \theta = \theta_0) = 1 - \alpha.$$

A confidence set provides a range/set of plausible values for an unknown parameter.

Example Let $X_1, \ldots, X_n$ be IID $N(\mu, \sigma^2)$, μ and σ^2 unknown; construct a $100(1-\alpha)\%$ confidence set for μ.

1. Find a pivotal quantity, *i.e.* a function of the observations and parameters whose distribution does not depend on any unknown parameters.

In this case,

$$Y = \frac{\bar{X} - \mu}{\sqrt{S_c^2/n}} \stackrel{d}{=} t_{n-1} \text{ (Student with } n-1 \text{ d.f.)}$$

where S_c^2 is the usual unbiased variance estimator:

$$S_c^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Since the parameters of t_{n-1} do not depend on μ or σ, $\frac{\bar{X}-\mu}{\sqrt{S_c^2/n}}$ is a pivotal quantity.

Recall that the Student distribution with $n-1$ d.f. has pdf

$$p(y) = \frac{\Gamma(n/2)}{\sqrt{(n-1)\pi}\,\Gamma((n-1)/2)} \left(1 + \frac{y^2}{n-1}\right)^{-n/2}.$$

Let $t_{n-1,\beta}$ be the quantile of the Student distribution with $n-1$ d.f. at level β:

$$\mathbb{P}(Y \leq t_{n-1,\beta}) = \beta.$$

The Student distribution is symmetric, so $t_{n-1,\beta} = -t_{n-1,1-\beta}$.

Thus

$$\mathbb{P}\left[-t_{n-1,1-\alpha/2} \leq Y \leq t_{n-1,1-\alpha/2}\right] = 1-\alpha.$$

2. Rearrange this statement around the parameter of interest:

$$\mathbb{P}\left[\bar{X} - \sqrt{\frac{S_c^2}{n}} t_{n-1,1-\alpha/2} \leq \mu \leq \bar{X} + \sqrt{\frac{S_c^2}{n}} t_{n-1,1-\alpha/2}\right]$$

$=1-\alpha$.

3. Extract the confidence set from the relationship obtained:

$$C(\boldsymbol{X}) = \left[\bar{X} - \sqrt{\frac{S_c^2}{n}} t_{n-1,1-\alpha/2},\ \bar{X} + \sqrt{\frac{S_c^2}{n}} t_{n-1,1-\alpha/2}\right]$$

is a $100(1-\alpha)\%$ confidence interval for μ.

When a pivotal quantity is difficult to obtain, one can try to find approximate confidence sets using the asymptotic behaviour of an estimator (for instance, the MLE).

Definition A $100(1-\alpha)\%$ approximate (or asymptotic) confidence set $C_n(\boldsymbol{X})$ for the parameter θ is a random set such that

$$\mathbb{P}(\theta_0 \in C_n(\boldsymbol{X}) \,|\, \theta = \theta_0) \to 1 - \alpha \quad \text{as } n \to \infty.$$

A key element for computing such a set in practice is the following result.

Slutsky's lemma Assume that $\widehat{\theta}$ is a consistent estimator of θ, and Z_n is a sequence of random variables such that $Z_n \approx Z$ when n is large. Then we have

$$Z_n + \widehat{\theta} \approx Z + \theta, \quad \widehat{\theta} Z_n \approx \theta Z, \quad \frac{Z_n}{\widehat{\theta}} \approx \frac{Z}{\theta}$$

when n is large.

Example If $X_1, \ldots, X_n$ are IID Poisson distributed with unknown parameter θ, find an asymptotic $100(1-\alpha)\%$ confidence set for θ.

In this case, the log-likelihood is

$$l(\theta|\boldsymbol{X}) \;=\; -n\theta + \log(\theta)\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\log(X_i!)$$

$$\Rightarrow \frac{\partial l}{\partial \theta}(\theta|\boldsymbol{X}) \;=\; -n + \frac{1}{\theta}\sum_{i=1}^{n} X_i.$$

The MLE is then $\widehat{\theta} = \bar{X} = n^{-1}\sum_{i=1}^{n} X_i$.

The Fisher information matrix for a *single* observation is

$$i(\theta) = \frac{1}{n}\mathbb{E}\left[-\frac{\partial^2 l}{\partial \theta^2}(\theta|\boldsymbol{X})\right] = \frac{\mathbb{E}(X)}{\theta^2} = \frac{1}{\theta}.$$

By standard ML theory, $\widehat{\theta}$ is consistent (mean square error tends to zero as $n \to \infty$ and satisfies

$$\sqrt{n}(\widehat{\theta} - \theta) \approx N(0, \theta)$$

as $n \to \infty$.

Equivalently, for large n, we have:

$$\sqrt{n}\frac{\widehat{\theta} - \theta}{\sqrt{\theta}} \approx N(0, 1).$$

Because (of consistency) $\widehat{\theta} \to \theta$, we have $\sqrt{\widehat{\theta}} \to \sqrt{\theta}$; then, by Slutsky's lemma,

$$\sqrt{n}\frac{\widehat{\theta} - \theta}{\sqrt{\widehat{\theta}}} \approx N(0, 1)$$

for large n.

Let then $q_{1-\alpha/2}$ be the quantile of the normal distribution at level $1 - \alpha/2$. The approximate $100(1 - \alpha)\%$ confidence interval for θ is:

$$C(\boldsymbol{X}) = \left[\bar{X} - q_{1-\alpha/2}\sqrt{\frac{\widehat{\theta}}{n}},\ \bar{X} + q_{1-\alpha/2}\sqrt{\frac{\widehat{\theta}}{n}}\right].$$

This is only suitable for large n!

Note that in large samples ($n \to \infty$): $\theta \approx N(\theta, I^{-1}(\theta)$ and hence an approximate confidence set can be obtained using the asymptotic behaviour of the MLE:

$$\left[\hat{\theta} - q_{1-\alpha/2}\frac{1}{\sqrt{I(\hat{\theta})}},\quad \hat{\theta} + q_{1-\alpha/2}\frac{1}{\sqrt{I(\hat{\theta})}}\right]$$

If $\hat{\theta}$ is based on an IID sample then $I(\hat{\theta}) = ni(\hat{\theta})$

Duality confidence sets–hypothesis testing

Suppose that there is a $100(1-\alpha)\%$ confidence set $C(\boldsymbol{X})$ for θ, *i.e.*

$$\mathbb{P}(\theta_0 \in C(\boldsymbol{X}) \,|\, \theta = \theta_0) = 1 - \alpha.$$

Then a test of size α for the null hypothesis $H_0 : \theta = \theta_0$ is defined by the test function

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \theta_0 \notin C(\boldsymbol{x}), \\ 0 & \text{if } \theta_0 \in C(\boldsymbol{x}). \end{cases}$$

Conversely, suppose that for every value θ_0 in the parameter space Θ, there is a non-randomised test at level α of $H_0 : \theta = \theta_0$. Denote the critical region of this test by $A(\theta_0)$. Then the set

$$C(\boldsymbol{X}) = \{\theta \in \Theta : \boldsymbol{X} \notin A(\theta)\}$$

is a confidence set for θ, with confidence level $100(1-\alpha)\%$.

Note that the alternative hypothesis is not specified here. Of course, the actual power of the test depends on the alternative!

Chapter 4

The Delta-method

Context Let $\widehat{\boldsymbol{\theta}}$ denote an estimator of an unknown parameter $\boldsymbol{\theta}$ based on a sample $X_1, \ldots, X_n$. Suppose the asymptotic properties of $\widehat{\boldsymbol{\theta}}$, as $n \to \infty$, are known.

Question: If $h\colon \Theta \to \mathbb{R}$ is a known function, what can we say about $h(\widehat{\boldsymbol{\theta}})$?

Example Let $X_1, \ldots, X_n$ be IID Exp(θ). If $\widehat{\theta}$ denotes the MLE then we have by ML theory

$$\sqrt{n}(\widehat{\theta} - \theta) \approx N(0, [i(\theta)]^{-1}) = N(0, \theta^2).$$

What about the estimator of the mean, $1/\widehat{\theta}$?

4.1 The univariate case

Let $\widehat{\theta}$ denote an estimator (MLE, or anything else) of an unknown parameter $\theta \in \Theta \subseteq \mathbb{R}$ based on a sample $X_1, \ldots, X_n$.

Suppose it is known that as $n \to \infty$,

$$\sqrt{n}(\widehat{\theta} - \theta) \approx N(0, V)$$

or

$$\hat{\theta} \approx N\left(\theta, \frac{V}{n}\right)$$

Theorem (Delta-method) Suppose that h has a nonzero derivative at θ. Then:

$$\sqrt{n}\big(h(\widehat{\theta}) - h(\theta)\big) \approx N\big(0, \{h'(\theta)\}^2 V\big) \text{ as } n \to \infty.$$

And hence an approximate 95% confidence interval is given by:

$$h(\hat{\theta}) \mp 1.96|h'(\hat{\theta})|V^{1/2}/\sqrt{n}$$

Proof (not examinable)

Recall that h has a first-order Taylor expansion at the point θ:

$$h(\widehat{\theta}) = h(\theta) + (\widehat{\theta} - \theta)h'(\theta) + (\widehat{\theta} - \theta)R(\widehat{\theta} - \theta)$$

where $R = R(x)$ is a function converging to 0 as $x \to 0$. Because

$$\sqrt{n}(\widehat{\theta} - \theta) \approx N(0, V) \text{ as } n \to \infty,$$

it follows that $\widehat{\theta}$ is consistent for θ. Hence, by Slut-

sky's lemma, the convergence

$$\underbrace{\sqrt{n}(\widehat{\theta}-\theta)}_{\to N(0,V)}\underbrace{R(\widehat{\theta}-\theta)}_{\to 0} \to 0.$$

Using the equality $a \times N(0,V) = N(0,a^2V)$, we obtain

$$\begin{aligned}\sqrt{n}\big(h(\widehat{\theta})-h(\theta)\big) &\approx \big[\sqrt{n}(\widehat{\theta}-\theta)\big]h'(\theta)\\ &\approx h'(\theta)\times N(0,V)\\ &= N\big(0,\{h'(\theta)\}^2V\big)\end{aligned}$$

as $n \to \infty$, which concludes the proof. ∎

Example Suppose $X_1,\ldots,X_n$ are IID $N(\theta,1)$. The MLE of θ is $\widehat{\theta}=\bar{X}=n^{-1}\sum_{i=1}^n X_i$.

What is the asymptotic distribution of $\widehat{\theta}^2$?
i.e. $h(\theta)=\theta^2$

Here:

$$\sqrt{n}(\widehat{\theta}-\theta) \approx N(0,1) \text{ as } n\to\infty,$$

and

$$\widehat{\theta}^2-\theta^2 = h(\widehat{\theta})-h(\theta)$$

with $h(x)=x^2$. Then $h'(x)=2x$ and so, by the delta-method with $V=1$:

$$\begin{aligned}\sqrt{n}(\widehat{\theta}^2 - \theta^2) &= \sqrt{n}\big(h(\widehat{\theta}) - h(\theta)\big) \\ &\approx N\big(0, \{h'(\theta)\}^2 V\big) \\ &= N(0, 4\theta^2) \text{ as } n \to \infty.\end{aligned}$$

Consequently, $\widehat{\theta}^2 \approx N(\theta^2, 4\theta^2/n)$.

4.2 The multivariate case

The case when the parameter is multivariate is even more interesting, because it allows for a wider range of applications. First of all, we give a formal definition of the multivariate normal distribution:

Definition A random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$ having p components is said to have a multivariate normal distribution if any linear combination of its components has a normal distribution:

$$\forall \mathbf{c} = (c_1, \ldots, c_p),\ \mathbf{c}^\intercal \boldsymbol{Y} \text{ has a normal distribution.}$$

If moreover the covariance matrix $\boldsymbol{V}$ of $\boldsymbol{Y}$, that is

$$\boldsymbol{V}_{ij} = \mathrm{Cov}(Y_i, Y_j),$$

is nonsingular (*i.e.* of full rank p), then $\boldsymbol{Y}$ has pdf

$$\begin{aligned} &\phi_p(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{V}) \\ &= (2\pi)^{-p/2}|\boldsymbol{V}|^{-1/2} \exp\{-\tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^\intercal \boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\}. \end{aligned}$$

Here $\boldsymbol{\mu}$ $(p \times 1)$ is the mean vector of $\boldsymbol{Y}$. We write $Y \sim N_p(\boldsymbol{\mu}, \boldsymbol{V})$.

Suppose now that $\widehat{\boldsymbol{\theta}}$ is an estimator of $\boldsymbol{\theta}$ that satisfies

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \approx N_p(\mathbf{0}_p, \boldsymbol{V}) \text{ as } n \to \infty$$

where $\mathbf{0}_p$ is the zero vector in $\mathbb{R}^p$ and $\boldsymbol{V}$ is the $p \times p$ covariance matrix of the limiting normal distribution.

Theorem (Delta-method) Suppose that h is a differentiable function having nonzero gradient $\nabla h(\boldsymbol{\theta})$ at $\boldsymbol{\theta}$. Then

$$\sqrt{n}\big(h(\widehat{\boldsymbol{\theta}}) - h(\boldsymbol{\theta})\big) \approx N\left(0, \{\nabla h(\boldsymbol{\theta})\}^\intercal \boldsymbol{V} \{\nabla h(\boldsymbol{\theta})\}\right)$$

as $n \to \infty$.

Recall that the gradient of h is the vector of its partial derivatives:

$$\nabla h(\boldsymbol{\theta}) = \begin{bmatrix} \dfrac{\partial h}{\partial \theta_1}(\boldsymbol{\theta}) \\ \vdots \\ \dfrac{\partial h}{\partial \theta_p}(\boldsymbol{\theta}) \end{bmatrix} = \left\{ \frac{\partial h}{\partial \theta_r}(\boldsymbol{\theta}) \right\}_{1 \le r \le p}.$$

Proof (non examinable) The idea is to mimic the proof of the univariate case. Let $\boldsymbol{Y} = \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$. Then $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta} + n^{-1/2}\boldsymbol{Y}$ and thus, by the multivariate Taylor formula:

$$\begin{aligned} h(\widehat{\boldsymbol{\theta}}) &= h(\boldsymbol{\theta}) + n^{-1/2}\{\nabla h(\boldsymbol{\theta})\}^\intercal \boldsymbol{Y} \\ &+ \text{ terms of order smaller than } n^{-1/2}. \end{aligned}$$

Rearranging yields

$$\sqrt{n}\big(h(\widehat{\boldsymbol{\theta}}) - h(\boldsymbol{\theta})\big) = \mathbf{c}^\intercal \boldsymbol{Y} + \text{terms converging to } 0$$

where $\mathbf{c} = \nabla h(\boldsymbol{\theta})$ is the gradient of h at $\boldsymbol{\theta}$. It follows that

$$\sqrt{n}\big(h(\widehat{\boldsymbol{\theta}}) - h(\boldsymbol{\theta})\big) \approx \mathbf{c}^\intercal \boldsymbol{Y} \stackrel{d}{=} N\left(0, \{\nabla h(\boldsymbol{\theta})\}^\intercal \boldsymbol{V} \{\nabla h(\boldsymbol{\theta})\}\right)$$

as $n \to \infty$, using the following property:

$$\mathbf{c}^\intercal \boldsymbol{Y} \stackrel{d}{=} N(\mathbf{c}^\intercal \boldsymbol{\mu}, \mathbf{c}^\intercal \boldsymbol{V} \mathbf{c}).$$

■

Example. Suppose $X_1, \ldots, X_n$ is an IID sample from $N(\mu, \sigma^2)$. We wish to estimate μ^2/σ^2, called the coefficient of variation or relative standard deviation.

Let $\boldsymbol{\theta} = (\theta_1, \theta_2)^\intercal = (\mu, \sigma^2)^\intercal$ and define $h\colon \mathbb{R}^2 \to \mathbb{R}$ by $h(\boldsymbol{\theta}) = \theta_1^2/\theta_2$, so that $\mu^2/\sigma^2 = h(\boldsymbol{\theta})$. We estimate $\boldsymbol{\theta}$ by the MLE $\widehat{\boldsymbol{\theta}}$. From standard ML theory,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \approx N_2(\mathbf{0}_2, \boldsymbol{i}(\boldsymbol{\theta})^{-1}) \text{ as } n \to \infty$$

where $\boldsymbol{i}(\boldsymbol{\theta})$ is the Fisher information for a single observation. To find $\boldsymbol{i}(\boldsymbol{\theta})$, we consider the log-likelihood for a single observation x:

$$l(\theta_1, \theta_2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\theta_2 - \frac{1}{2\theta_2}(x - \theta_1)^2.$$

Then $\dfrac{\partial l}{\partial \theta_1} = \dfrac{x - \theta_1}{\theta_2}$, $\dfrac{\partial l}{\partial \theta_2} = -\dfrac{1}{2\theta_2} + \dfrac{1}{2\theta_2^2}(x - \theta_1)^2$, so:

$$\frac{\partial^2 l}{\partial \theta_1^2} = -\frac{1}{\theta_2}, \quad \frac{\partial^2 l}{\partial \theta_2^2} = \frac{1}{2\theta_2^2} - \frac{1}{\theta_2^3}(x - \theta_1)^2$$

$$\text{and } \frac{\partial^2 l}{\partial \theta_1 \partial \theta_2} = -\frac{(x - \theta_1)}{\theta_2^2}.$$

Now $\mathbb{E}(X-\theta_1) = \mathbb{E}(X)-\mu = 0$ and $\mathbb{E}[(X-\theta_1)^2] = \mathbb{E}[(X-\mu)^2] = \sigma^2 = \theta_2$, so that

$$\boldsymbol{i}(\boldsymbol{\theta}) \;=\; -\begin{bmatrix} \mathbb{E}\left(\dfrac{\partial^2 l}{\partial\theta_1^2}\right) & \mathbb{E}\left(\dfrac{\partial^2 l}{\partial\theta_1\partial\theta_2}\right) \\ \mathbb{E}\left(\dfrac{\partial^2 l}{\partial\theta_1\partial\theta_2}\right) & \mathbb{E}\left(\dfrac{\partial^2 l}{\partial\theta_2^2}\right) \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{1}{\theta_2} & 0 \\ 0 & \dfrac{1}{2\theta_2^2} \end{bmatrix}.$$

As a consequence,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \approx N_2(\mathbf{0}_2, \boldsymbol{V})$$

where

$$\boldsymbol{V} = \boldsymbol{i}(\boldsymbol{\theta})^{-1} = \begin{bmatrix} \theta_2 & 0 \\ 0 & 2\theta_2^2 \end{bmatrix}.$$

To conclude, we compute the gradient of h:

$$\nabla h(\boldsymbol{\theta}) = \left(\frac{2\theta_1}{\theta_2}, -\frac{\theta_1^2}{\theta_2^2}\right)^\intercal$$

since ,

$$\frac{\partial h(\boldsymbol{\theta})}{\partial \theta_1} = \frac{1}{\theta_2} 2\theta_1 = \frac{2\theta_1}{\theta_2}$$

$$\frac{\partial h(\boldsymbol{\theta})}{\partial \theta_2} = -\frac{1}{\theta_2^2}\theta_1$$

and

$$\begin{aligned}
& \{\nabla h(\boldsymbol{\theta})\}^\intercal \boldsymbol{V} \{\nabla h(\boldsymbol{\theta})\} \\
= & \left(\frac{2\theta_1}{\theta_2}, -\frac{\theta_1^2}{\theta_2^2}\right) \begin{bmatrix} \theta_2 & 0 \\ 0 & 2\theta_2^2 \end{bmatrix} \begin{pmatrix} 2\theta_1/\theta_2 \\ -\theta_1^2/\theta_2^2 \end{pmatrix} \\
= & \frac{4\theta_1^2}{\theta_2} + \frac{2\theta_1^4}{\theta_2^2} \\
= & \frac{2\theta_1^2}{\theta_2^2}(2\theta_2 + \theta_1^2).
\end{aligned}$$

By the delta-method, we have as $n \to \infty$:

$$\sqrt{n}\big(h(\widehat{\boldsymbol{\theta}}) - h(\boldsymbol{\theta})\big) \approx N\left(0, \frac{2\theta_1^2(2\theta_2 + \theta_1^2)}{\theta_2^2}\right).$$

In other words,

$$\sqrt{n}\left(\frac{\widehat{\mu}^2}{\widehat{\sigma^2}} - \frac{\mu^2}{\sigma^2}\right) \approx N\left(0, \frac{2\mu^2(2\sigma^2 + \mu^2)}{\sigma^4}\right).$$

Chapter 5

Introduction to Bayesian statistics

5.1 The Bayesian paradigm

5.1.1 Basic concepts

In classical/frequentist statistics, the unknown parameter $\boldsymbol{\theta}$ is regarded as having some true fixed (non-random) value, and all probabilistic statements relate to observable random variables $X_1, \ldots, X_n$. Inference essentially depends on how likely the actual outcome is compared to other possible outcomes.

In Bayesian statistics, $\boldsymbol{\theta}$ is seen as a random variable, and probabilistic statements can be made about $\boldsymbol{\theta}$ and about any other unknown quantity. Inference depends on (i) initial beliefs about $\boldsymbol{\theta}$ and (ii) how knowledge of the data is used to update these beliefs. In other words, the observed data in the Bayesian paradigm are treated as fixed (non-random) quantities.

Bayesian inference

The following framework is adopted [same as before]:

(i) A probability model for the data

$$\boldsymbol{x} = (x_1, \ldots, x_n),$$

given the parameter vector $\boldsymbol{\theta}$, is denoted by $p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})$.

(ii) Initial belief about the parameter $\boldsymbol{\theta}$ is summarised by a distribution with pmf/pdf $\pi(\boldsymbol{\theta})$, called the prior distribution on $\boldsymbol{\theta}$.

With these elements, and using Bayes' Theorem:

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

we get the posterior pmf/pdf of $\boldsymbol{\theta}$ given the data $\boldsymbol{x}$:

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})\frac{\pi(\boldsymbol{\theta})}{m(\boldsymbol{x})}.$$

Here $m(\boldsymbol{x})$ is the marginal pdf/pmf of $\boldsymbol{x}$: if Θ is discrete then

$$m(\boldsymbol{x}) = \sum_{\boldsymbol{\theta} \in \Theta} p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}),$$

and if Θ is continuous then

$$m(\boldsymbol{x}) = \int_{\boldsymbol{\theta} \in \Theta} p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta}.$$

The posterior pmf/pdf of $\boldsymbol{\theta}$ given the data is an expression of our updated beliefs about the parameter $\boldsymbol{\theta}$ after the data has been observed.

Note: We shall often use the formula

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) \propto p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta,$$

where $\propto$ denotes the proportionality symbol.

It is indeed often unnecessary to compute the actual value of the normalising constant: in particular, any time we want to know which family of distributions the posterior belongs to, it shall be enough to merely compute

$$p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \quad (\text{for } \boldsymbol{\theta} \in \Theta)$$

and recognise the resulting pmf/pdf "up to a constant".

We illustrate this on the next example.

Example Let $X_1, \ldots, X_n$ be IID Bernoulli rvs with success probability θ, i.e. $P(X_i = 1) = \theta$ and $P(X_i = 0) = 1 - \theta$.

Suppose that the prior density on θ is Beta(g, h) (i.e. we express our prior beliefs about θ with a Beta distribution:

$$\forall \theta \in (0,1),\ \pi(\theta) = \frac{\theta^{g-1}(1-\theta)^{h-1}}{B(g,h)},$$

$$\text{with } B(g,h) = \int_0^1 \theta^{g-1}(1-\theta)^{h-1}\, d\theta,$$

Also, suppose that we observe some data, $\boldsymbol{x}$, where $\boldsymbol{x} = (x_1, \ldots, x_n)$. Then for $\theta \in (0, 1)$,

$$\pi(\theta|\boldsymbol{x}) = \frac{\left(\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}\right) \frac{\theta^{g-1}(1-\theta)^{h-1}}{B(g,h)}}{\int_0^1 \left(\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}\right) \frac{\theta^{g-1}(1-\theta)^{h-1}}{B(g,h)}\, d\theta}$$

$$= \frac{\theta^{g+\sum x_i-1}(1-\theta)^{h+n-\sum x_i-1}}{\int_0^1 \theta^{g+\sum x_i-1}(1-\theta)^{h+n-\sum x_i-1}\, d\theta}$$

$$= \frac{\theta^{g+\sum x_i-1}(1-\theta)^{h+n-\sum x_i-1}}{B(g+\sum x_i, h+n-\sum x_i)},$$

while if $\theta \notin (0, 1)$, $\pi(\theta|\boldsymbol{x}) = 0$. Thus

$$\pi(\theta|\boldsymbol{x}) \sim \text{Beta}(g + \sum x_i, h + n - \sum x_i).$$

Alternatively, using the $\propto$ symbol: for $\theta \in (0,1)$,

$$\begin{aligned}\pi(\theta|\boldsymbol{x}) &\propto \pi(\theta)p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)\\ &\propto \theta^{g-1}(1-\theta)^{h-1}\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}\\ &\propto \theta^{g+\sum x_i-1}(1-\theta)^{h+n-\sum x_i-1}.\end{aligned}$$

The posterior pdf is proportional to a Beta$(g+\sum x_i, h+n-\sum x_i)$ pdf, so the posterior distribution must be a Beta$(g+\sum x_i, h+n-\sum x_i)$ distribution. In other words, we recognise the last expression above looks like the pdf of a Beta distribution.

This is much quicker than the full computation!

Example Let $X_1, \ldots, X_n$ be IID Exp(θ) rvs. Suppose the prior density on θ is exponential Exp(g) and we observe $\boldsymbol{x} = (x_1, \ldots, x_n)$. Then for $\theta > 0$,

$$\begin{aligned}\pi(\theta|\boldsymbol{x}) &\propto p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)\pi(\theta)\\ &\propto e^{-g\theta}\theta^n e^{-\theta\sum x_i}\\ &\propto \theta^n e^{-\theta(g+\sum x_i)}\end{aligned}$$

and the posterior is then a Gamma$(n+1, g+\sum x_i)$ distribution.

5.1.2 Sufficiency

Theorem Suppose $T(\boldsymbol{X})$ is sufficient for θ. Then

$$\pi(\theta|\boldsymbol{x}) = \pi(\theta|T(\boldsymbol{X}) = t),$$

i.e. the posterior distribution of θ depends on the data through the sufficient statistic $T(\boldsymbol{X})$ only.

Proof If T is sufficient for θ then the distribution of X given T does not depend on θ. In other words,

$$\begin{aligned} p_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = p_{\boldsymbol{X}}(\boldsymbol{x}, t|\theta) &= p_{\boldsymbol{X}}(\boldsymbol{x}|t, \theta)p_{_T}(t|\theta) \\ &= p_{\boldsymbol{X}}(\boldsymbol{x}|t)p_{_T}(t|\theta). \end{aligned}$$

Therefore,

$$\begin{aligned} \pi(\theta|\boldsymbol{x}) &= \frac{\pi(\theta)p_{\boldsymbol{X}}(\boldsymbol{x}|t)p_{_T}(t|\theta)}{\int_\Theta \pi(\theta)p_{\boldsymbol{X}}(\boldsymbol{x}|t)p_{_T}(t|\theta)\, d\theta} \\ &= \frac{\pi(\theta)p_{_T}(t|\theta)}{\int_\Theta \pi(\theta)p_{_T}(t|\theta)\, d\theta} = \pi(\theta|t) \end{aligned}$$

which is what we wanted to prove. ∎

Example In the example of the n Bernoulli trials on slide 95, recall that

$$\pi(\theta|\boldsymbol{x}) \sim \text{Beta}(g + \sum x_i, h + n - \sum x_i),$$

i.e. the posterior depends on data only through $t = \sum x_i$. We have seen previously (slide 21) that $T = \sum X_i$ is sufficient for θ in this case.

5.2 Choice of prior distribution

Choosing a prior for θ consists of two steps:

(i) Choosing a suitable family of distributions, e.g. Beta(g, h);
(ii) Choosing suitable parameters for that distribution, i.e. values for g and h.

We will largely address (i) in what follows.

5.2.1 Non-informative priors

What if we have little or no prior knowledge about θ? In this case we want the prior to influence the posterior as little as possible, *i.e.* we want the posterior to be dominated by the likelihood.

We consider three ways of trying to do this: an ignorant prior, a vague prior and Jeffreys' prior.

Another possibility, beyond the scope of this module, is to use reference priors (*i.e.* maximise the expected difference between prior and posterior).

(i) Ignorant prior Represent ignorance by "all values of θ are equally likely."

If θ is discrete with a finite number m of possible values $\theta_1, \ldots, \theta_m$ then take the discrete uniform distribution over these values:

$$\pi(\theta_i) = \frac{1}{m}, \quad i = 1, \ldots, m.$$

In other words, before seeing the data all the values θ_i are equally likely.

If θ is continuous over a bounded interval (a, b) then take:

$$\theta \sim U(a, b), \quad i.e. \quad \pi(\theta) = (b - a)^{-1}, \quad \theta \in (a, b).$$

If θ is otherwise unbounded, take the improper constant prior equal to 1:

$$\pi(\theta) = 1 \ \forall \theta \in \Theta; \quad \text{here} \int_\Theta \pi(\theta)\, d\theta = \infty.$$

Note: improper priors (ie distributions whose pdfs do not integrate to 1) are OK so long as the corresponding posterior is proper.

Drawback If we are ignorant about θ, we should also be ignorant about θ^2, $e^{-\theta}$, etc. But if $\pi(\theta)$ is constant (ie improper), the prior on $\psi = \theta^2$, $\Pi(\psi) = \pi(\sqrt{\psi})/2\sqrt{\psi}$ is not a constant, etc. This is an awkward issue.

(ii) Vague prior knowledge Choose a prior with large or "infinite" variance, or a "very flat" prior.

Example $X_1, \ldots, X_n$ IID $N(\theta, \sigma^2)$ where σ^2 is known, and prior $\theta \sim N(b, d^2)$ in which d is large, *i.e.* $d \to \infty$. Then:

$$\pi(\theta|\boldsymbol{x}) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i-\theta)^2\right\}\exp\left\{-\frac{1}{2d^2}(\theta-b)^2\right\}$$

$$\propto \exp\left\{-\frac{1}{2\sigma^2}\left(n\theta^2 - 2\theta n\bar{x}\right) - \frac{1}{2d^2}\left(\theta^2 - 2\theta b\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{d^2}\right)\left[\theta^2 - 2\theta\frac{\left(\frac{n\bar{x}}{\sigma^2}+\frac{b}{d^2}\right)}{\left(\frac{n}{\sigma^2}+\frac{1}{d^2}\right)}\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{d^2}\right)\left[\theta - \frac{\left(\frac{n\bar{x}}{\sigma^2}+\frac{b}{d^2}\right)}{\left(\frac{n}{\sigma^2}+\frac{1}{d^2}\right)}\right]^2\right\},$$

and so $\pi(\theta|\boldsymbol{x}) \sim N(B, D^2)$ where

$$B = \frac{\left(\frac{n\bar{x}}{\sigma^2} + \frac{b}{d^2}\right)}{\left(\frac{n}{\sigma^2} + \frac{1}{d^2}\right)} \text{ and } D^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{d^2}}.$$

As $d \to \infty$, $B \to \bar{x}$, $D^2 \to \frac{\sigma^2}{n}$, so assuming vague prior knowledge for θ gives $\pi(\theta|\boldsymbol{x}) \sim N(\bar{x}, \sigma^2/n)$.

(iii) Jeffreys' Prior Jeffreys suggested the prior

$$\pi(\theta) \propto \sqrt{I(\theta)},$$

where $I(\theta)$ is the Fisher information in the statistical model $\boldsymbol{X}|\theta$.

Contrary to what happens for the ignorant prior, Jeffreys' prior does not depend on the parametrisation used.

Indeed, if the prior chosen when using the θ parametrisation is $\pi(\theta) \propto \sqrt{I(\theta)}$, and if $\psi = f(\theta)$ is a reparametrisation where f is a one-to-one monotonic and differentiable function, then the prior $\tilde{\pi}(\psi)$ in the ψ parametrisation will be

$$\tilde{\pi}(\psi) \propto \sqrt{\tilde{I}(\psi)}$$

where $\tilde{I}(\psi)$ is the Fisher information in the ψ parametrisation.

Example $X_1, \ldots, X_n$ IID Exp(λ). Then for $\lambda > 0$,

$$\begin{aligned} \text{Fisher Informatio: } I(\lambda) &= -\mathbb{E}\left[\frac{\partial^2}{\partial\lambda^2} \log p_{\boldsymbol{X}}(\boldsymbol{x}|\lambda)\right] \\ &= -\mathbb{E}\left[\frac{\partial^2}{\partial\lambda^2} \log(\lambda^n e^{-\lambda \sum x_i})\right] \\ &= \frac{n}{\lambda^2}, \end{aligned}$$

so Jeffreys' prior is $\pi(\lambda) \propto 1/\lambda$, for $\lambda > 0$. Note that this is an improper prior.

Consequently, for $\lambda > 0$,

$$\begin{aligned} \pi(\lambda|\boldsymbol{x}) &\propto p_{\boldsymbol{X}}(\boldsymbol{x}|\lambda)\pi(\lambda) \\ &\propto \frac{1}{\lambda} \lambda^n e^{-\lambda \sum x_i} \\ &\propto \lambda^{n-1} e^{-\lambda \sum x_i}, \end{aligned}$$

which means

$$\pi(\lambda|\boldsymbol{x}) \sim \text{Gamma}(n, \sum x_i).$$

5.2.2 Conjugate Priors

It is often convenient to choose a prior which gives rise to a posterior distribution of the same form as the prior. It helps in particular to assess the updating effect of the data on the posterior distribution. Such a prior is called a conjugate prior.

Definition A class $\mathcal{C}$ of distributions is said to form a conjugate family for $p_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta})$ if the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ is in $\mathcal{C}$ whenever the prior distribution $\pi(\boldsymbol{\theta})$ is.

Example $\mathcal{C} = \{\text{Beta}(\alpha, \beta)|\alpha, \beta > 0\}$ is a conjugate family for the Bernoulli distribution (see earlier example).

We now state and prove a general result on conjugate priors in the one-dimensional exponential family.

Theorem If $\boldsymbol{X} = (X_1, \ldots, X_n)$ is an IID sample from

$$p_X(x|\theta) = \exp\{A(\theta)B(x) + C(\theta) + D(x)\},$$

a conjugate family for $p_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = \prod_{i=1}^n p_X(x_i|\theta)$ is given by

$$\pi(\theta) = \frac{1}{k(u,v)} \exp\{uA(\theta) + vC(\theta)\}, \ \theta \in \Theta$$

where u, v are prior parameters and

$$k(u,v) = \int_\Theta \exp\{uA(\theta) + vC(\theta)\} \ d\theta$$

is a normalising constant.

Proof Here

$$p_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = \exp\left\{A(\theta)\sum B(x_i) + nC(\theta) + \sum D(x_i)\right\}$$

and thus, for $\theta \in \Theta$,

$$\begin{aligned}\pi(\theta|\boldsymbol{x}) &\propto \exp\left\{A(\theta)\sum B(x_i) + nC(\theta) + \sum D(x_i)\right\} \\ &\quad \times \exp\{uA(\theta) + vC(\theta)\} \\ &\propto \exp\left\{\left[u + \sum B(x_i)\right]A(\theta) + [n+v]C(\theta)\right\}.\end{aligned}$$

This is of the same form as $\pi(\theta)$, with u replaced by $u + \sum B(x_i)$ and v replaced by $v + n$. ∎

Example Suppose X has a Poisson distribution with parameter θ; then

$$p_X(x|\theta) = \frac{\theta^x e^{-\theta}}{x!} = \exp\{x \log\theta - \theta - \log x!\},$$

so that $X|\theta$ is part of the exponential family with $A(\theta) = \log\theta$ and $C(\theta) = -\theta$.

Thus, a conjugate prior is

$$\pi(\theta) \propto \exp\{u \log\theta - v\theta\} = \theta^u e^{-v\theta}, \ \theta > 0,$$

i.e. $\pi(\theta)$ has the form of a Gamma distribution.

To check this: suppose $\pi(\theta) \sim \text{Gamma}(g, h)$, so

$$\begin{aligned}
\pi(\theta|\boldsymbol{x}) &\propto p_{\boldsymbol{X}}(\boldsymbol{x}|\theta)\,\pi(\theta) \\
&\propto \theta^{\sum x_i}\, e^{-n\theta}\, \theta^{g-1}\, e^{-h\theta} \\
&\propto \theta^{g+\sum x_i - 1}\, e^{-(h+n)\theta}, \ \text{for } \theta > 0,
\end{aligned}$$

which means

$$\pi(\theta|\boldsymbol{x}) \sim \text{Gamma}(g + \sum x_i, h + n).$$

Note that the prior distribution was:

$$\pi(\theta) \sim \text{Gamma}(g, h).$$

Notes on conjugate priors

(i) A conjugate family may exist even if $p_X(x|\theta)$ is not in the exponential family.

(ii) Conjugate priors make the calculation of $\pi(\theta|\boldsymbol{x})$ easier: the form of the posterior distribution is known in advance, so it only remains to compute posterior parameters.

(iii) Conjugate priors make it easy to see the effect of the data on the posterior through the updating formulae.

(iv) Conjugate priors can sometimes be too restrictive, especially when θ is a vector. We shall see an example later (in Chapter 6) in which conjugate priors impose rigid conditions on the prior/posterior dependence structure.

5.3 Bayesian Point and interval estimation

Since we are interested in the entire posterior distribution of a parameter, point/interval estimates are (only) useful as summaries of the posterior.

5.3.1 Summaries of location

We focus on location summaries for a univariate parameter θ. Three popular possibilities:

(i) The posterior mean: $\mathbb{E}[\theta|\boldsymbol{x}]$.

(ii) A posterior median, *i.e.* θ_{med} such that

$$\mathbb{P}(\theta \leq \theta_{\text{med}}|\boldsymbol{x}) \geq 1/2 \text{ and } \mathbb{P}(\theta \geq \theta_{\text{med}}|\boldsymbol{x}) \geq 1/2.$$

(iii) The posterior mode, *i.e.* θ_M such that

$$\pi(\theta_M|\boldsymbol{x}) = \sup_{\theta \in \Theta} \pi(\theta|\boldsymbol{x}).$$

Note: if the prior is constant then the mode is the MLE based on the data.

The usefulness of these quantities depends on the shape of the posterior density: symmetric? how many peaks? (among others)

Connection with loss functions

Suppose we use a parameter value $\tilde{\theta}$ ("best guess", point estimate) when the true (random!) value is θ. The loss in doing so will be denoted by $L(\widetilde{\theta}, \theta)$. Common loss functions are:

(i) $L(\widetilde{\theta}, \theta) = (\widetilde{\theta} - \theta)^2$ called squared loss, or L^2-loss;

(ii) $L(\widetilde{\theta}, \theta) = |\widetilde{\theta} - \theta|$ called absolute loss, or L^1-loss;

(iii) $L(\widetilde{\theta}, \theta) = \begin{cases} 0 & \text{if } \widetilde{\theta} = \theta \\ 1 & \text{if } \widetilde{\theta} \neq \theta \end{cases}$, the all-or-nothing loss.

Posterior uncertainty about θ is described by $\pi(\theta|\boldsymbol{x})$ and so we may wish to use the value of $\widetilde{\theta}$ that minimises the expected posterior loss

$$\mathbb{E}[L(\widetilde{\theta}, \theta)|\boldsymbol{x}] = \int_{\theta \in \Theta} L(\widetilde{\theta}, \theta)\pi(\theta|\boldsymbol{x})d\theta.$$

Using the above loss functions results in the aforementioned choices of $\widetilde{\theta}$:

(i) $\widetilde{\theta} = \mathbb{E}[\theta|\boldsymbol{x}]$, the posterior mean;

(ii) $\widetilde{\theta} = \theta_{\text{med}}$, the posterior median;

(iii) [discrete θ only, no analogue in the continuous case] $\widetilde{\theta} = \theta_M$, the posterior mode.

Proof of (i) when $\mathbb{E}[\theta^2|\boldsymbol{x}] < \infty$:

The goal is to compute $\mathbb{E}[L(\widetilde{\theta}, \theta)$ and find the value of $\tilde{\theta}$ that minimises it!

$$\begin{aligned}
\mathbb{E}[L(\widetilde{\theta}, \theta)|\boldsymbol{x}] &= \int_{\theta\in\Theta} (\widetilde{\theta} - \theta)^2 \pi(\theta|\boldsymbol{x}) d\theta \\
&= A\widetilde{\theta}^2 - 2B\widetilde{\theta} + C \\
\text{with } A &= \int_{\theta\in\Theta} \pi(\theta|\boldsymbol{x}) d\theta = 1, \\
B &= \int_{\theta\in\Theta} \theta\pi(\theta|\boldsymbol{x}) d\theta = \mathbb{E}[\theta|\boldsymbol{x}], \\
\text{and } C &= \int_{\theta\in\Theta} \theta^2\pi(\theta|\boldsymbol{x}) d\theta = \mathbb{E}[\theta^2|\boldsymbol{x}].
\end{aligned}$$

This is a convex polynomial of degree 2, whose derivative w.r.t. $\widetilde{\theta}$ is zero if and only if

$$2A\widetilde{\theta} - 2B = 2(\widetilde{\theta} - \mathbb{E}[\theta|\boldsymbol{x}]) = 0 \Leftrightarrow \widetilde{\theta} = \mathbb{E}[\theta|\boldsymbol{x}].$$

The minimiser is then $\mathbb{E}[\theta|\boldsymbol{x}]$ as required. ∎

The proofs of (ii) and (iii) are omitted and are not examinable.

5.3.2 Interval Estimation

Definition The region C_α is a $100(1-\alpha)\%$ Bayesian Credible Region (BCR) for θ if

$$\mathbb{P}(\theta \in C_\alpha | \boldsymbol{x}) = 1 - \alpha.$$

Note that such regions are not unique, so we look for a "canonical" choice of credible region.

Definition The region C_α is a $100(1-\alpha)\%$ Highest Posterior Density Region (HPDR) for θ if

$$C_\alpha = \{\theta \, : \, \pi(\theta|\boldsymbol{x}) \geq \gamma\}$$

where γ is chosen so that $\mathbb{P}(\theta \in C_\alpha|\boldsymbol{x}) = 1 - \alpha$.

In short: the HPDR gives the BCR that contains the most plausible values for the parameter after looking at the data.

Notes

(i) The HPDR gives the smallest (*i.e.* shortest) $100(1-\alpha)\%$ BCR for θ.

(ii) Computing the HPDR may not be easy if the shape of $\pi(\theta|\boldsymbol{x})$ is complex (a bimodal nonsymmetric distribution, some non-monotonic pdfs...)

(iii) To compute the HPDR you should know the full form of the pdf/pmf. This implies that for this kind of calculation the proportionality symbol should not be used. Or, if it is used then we need to normalise the posterior density, i.e. find the marginal likelihood $m(\mathbf{x})$.

Example Suppose $\theta|\boldsymbol{x}$ is Beta$(2, 23)$ *i.e.*

$$\pi(\theta|\boldsymbol{x}) = 552\,\theta(1-\theta)^{22}, \quad 0 \leq \theta \leq 1.$$

This pdf increases towards its mode and then decreases.

To find the HPDR, we solve

$$\int_a^b 552\,\theta(1-\theta)^{22}\,d\theta = 1-\alpha,$$

with a and b such that

$$552\,a(1-a)^{22} = 552\,b(1-b)^{22}.$$

This should be done numerically!

Example Suppose $\theta|\boldsymbol{x}$ is Beta$(1, 24)$ *i.e.*

$$\pi(\theta|\boldsymbol{x}) = 24(1-\theta)^{23}, \quad 0 \le \theta \le 1.$$

This pdf is monotonically decreasing: the HPDR has the form $(0, b)$.

To find the HPDR explicitly, we solve

$$\int_0^b 24(1-\theta)^{23}\,d\theta = 1-\alpha,$$

that is, $b = 1-\alpha^{1/24}$. Thus

$$(0, 1-\alpha^{1/24})$$

is a $100(1-\alpha)\%$ HPDR for θ.

Example Recall that a Normal model with μ unknown, σ known and a vague Normal prior gave $\pi(\mu|\boldsymbol{x}) \sim N(\bar{x}, \sigma^2/n)$.

The posterior density is symmetric: we find (exercise!) that the $100(1-\alpha)\%$ HPDR is

$$\left(\bar{x} - q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

where $q_{1-\alpha/2}$ is the quantile of the standard normal distribution at level $1-\alpha/2$. This interval is identical to the frequentist confidence interval for μ.

Frequentist confidence intervals versus Bayesian credible intervals

These are interpreted differently.

1. If C_B is a $100(1-\alpha)\%$ BCR for θ, then

$$\mathbb{P}(\theta \in C_B|\boldsymbol{x}) = 1-\alpha$$

and θ should be thought of as a random variable.

2. If C_C is a $100(1-\alpha)\%$ classical confidence region, then θ (which is assume to be non-random) has a true value θ_0 and

$$\mathbb{P}(\theta_0 \in C_C \,|\, \theta = \theta_0) = 1-\alpha,$$

i.e. in the long run C_C contains the true value in $100(1-\alpha)\%$ of repetitions of the experiment.

5.4 Large sample properties of the posterior

Suppose $\boldsymbol{x} = (x_1, \ldots, x_n)$ is an IID sample from the distribution with pmf/pdf $p_X(\cdot|\theta)$. Write the likelihood of this sample:

$$L(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} p_X(x_i|\theta)$$

and $l(\theta|\boldsymbol{x}) = \log L(\theta|\boldsymbol{x})$.

Suppose that n is large. Then

$$\pi(\theta|\boldsymbol{x}) \approx N(\widehat{\theta}, \widehat{I}_{obs}^{-1}),$$

where $\widehat{\theta}$ is the maximum likelihood estimator of θ, *i.e.* the θ which maximises $l(\theta|\boldsymbol{x})$, and $\widehat{I}_{obs}$ is the observed Fisher information defined by

$$\widehat{I}_{obs} = -\frac{\partial^2 l}{\partial \theta^2}\Big|_{\theta=\widehat{\theta}}.$$

and may be derived either analytically or using the computer.

Ideas of proof which is not examinable When n is large:

(i) the effect of the prior distribution of θ on the posterior is negligible;
(ii) the log-likelihood is approximately quadratic in

a neighbourhood around $\widehat{\theta}$ (Taylor expansion) which contains most of the posterior probability.

5.5 Bayesian Model Comparison

Bayesians view hypothesis testing as model comparison. The issue is not whether or not a hypothesis is true but rather whether a model described under one hypothesis is preferable to one under another hypothesis.

Note that the model specification includes the prior distribution of the parameters, since the posterior distribution depends on the prior distribution.

Suppose we wish to compare the two models:

$$M_0 : p_{\boldsymbol{X}}^{(0)}(\boldsymbol{x}|\theta_0), \quad \pi_0(\theta_0)$$
$$M_1 : p_{\boldsymbol{X}}^{(1)}(\boldsymbol{x}|\theta_1), \quad \pi_1(\theta_1)$$

and we believe one of them is the "true" model.

Prior beliefs about the models can be expressed by probabilities $\mathbb{P}(M_0)$ and $\mathbb{P}(M_1) = 1 - \mathbb{P}(M_0)$.

If we observe data $\boldsymbol{X} = \boldsymbol{x}$ then, using Bayes' Theorem, the posterior probability of model M_i, $i = 0, 1$, is

$$\begin{aligned}\mathbb{P}(M_i|\boldsymbol{x}) &= \mathbb{P}(\boldsymbol{x}|M_i)\frac{\mathbb{P}(M_i)}{\mathbb{P}(\boldsymbol{x})} \text{ (abuse of notation)} \\ &= \frac{m_i(\boldsymbol{x})\mathbb{P}(M_i)}{m_0(\boldsymbol{x})\mathbb{P}(M_0) + m_1(\boldsymbol{x})\mathbb{P}(M_1)}\end{aligned}$$

where

$$m_i(\boldsymbol{x}) = \int_{\theta_i \in \Theta_i} p^{(i)}_{\boldsymbol{X}}(\boldsymbol{x}|\theta_i)\, \pi_i(\theta_i)\, d\theta_i$$

is the pmf/pdf of the data given that model M_i is true.

The posterior odds in favour of model M_1 are then

$$\begin{aligned}\frac{\mathbb{P}(M_1|\boldsymbol{x})}{\mathbb{P}(M_0|\boldsymbol{x})} &= \frac{m_1(\boldsymbol{x})}{m_0(\boldsymbol{x})}\frac{\mathbb{P}(M_1)}{\mathbb{P}(M_0)} \\ &= \frac{m_1(\boldsymbol{x})}{m_0(\boldsymbol{x})} \times \text{prior odds.}\end{aligned}$$

Therefore, the influence of the data on model selection is quantified by the ratio of the marginal pmfs/pdfs of $\boldsymbol{X}$, which is $m_1(\boldsymbol{x})/m_0(\boldsymbol{x})$.

Definition The quantity

$$\frac{m_1(\boldsymbol{x})}{m_0(\boldsymbol{x})} = \frac{\int p_{\boldsymbol{X}}^{(1)}(\boldsymbol{x}|\theta_1)\,\pi_1(\theta_1)\,d\theta_1}{\int p_{\boldsymbol{X}}^{(0)}(\boldsymbol{x}|\theta_0)\,\pi_0(\theta_0)\,d\theta_0}$$

is called the Bayes Factor in favour of model M_1 against model M_0.

The data indicate that M_1 is "more likely" than M_0 if and only if this Bayes Factor is greater than 1.

We do not think here about how large the Bayes Factor should be so as to ensure that model M_1 should be preferred to model M_0 "with given confidence". This is a context-dependent question.

Example Suppose $n = 1$ and

$$\begin{aligned} M_0 &: X \sim N(0,1) \\ M_1 &: X \sim N(\theta,1), \quad \theta \sim N(b,d^2). \end{aligned}$$

Then the Bayes Factor of model M_1 against M_0 is calculated as follows:

$$m_0(x) = (2\pi)^{-1/2} \exp(-x^2/2)$$

(no prior, so no integration to be done)

$$\begin{aligned} m_1(x) = & \int_{-\infty}^{\infty} (2\pi)^{-1/2} \exp[-(x-\theta)^2/2] \\ & \times (2\pi d^2)^{-1/2} \exp[-(\theta - b)^2/2d^2]\, d\theta \\ = & \frac{1}{2\pi d} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[x^2 + \frac{b^2}{d^2}\right]\right\} \\ & \times \exp\left\{-\frac{1}{2}\left[\left(1+\frac{1}{d^2}\right)\theta^2 - 2\theta\left(x+\frac{b}{d^2}\right)\right]\right\}\, d\theta \\ = & \frac{1}{2\pi d} \exp\left\{-\frac{1}{2}\left[x^2 + \frac{b^2}{d^2}\right]\right\} \times \\ & \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(1+\frac{1}{d^2}\right)\left[\theta^2 - \frac{2\theta\left(x+\frac{b}{d^2}\right)}{\left(1+\frac{1}{d^2}\right)}\right]\right\}\, d\theta. \end{aligned}$$

Now the integral (I, say), is

$$I = \int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{2}\left(1+\frac{1}{d^2}\right)\left[\theta - \frac{\left(x+\frac{b}{d^2}\right)}{\left(1+\frac{1}{d^2}\right)}\right]^2 \right\} d\theta \times \exp\left\{ \frac{1}{2}\frac{\left(x+\frac{b}{d^2}\right)^2}{\left(1+\frac{1}{d^2}\right)} \right\}.$$

Setting

$$\mu = \frac{\left(x+\frac{b}{d^2}\right)}{\left(1+\frac{1}{d^2}\right)} \quad \text{and} \quad \sigma^2 = \left(1+\frac{1}{d^2}\right)^{-1},$$

the integral J above is

$$J = \int_{-\infty}^{+\infty} \exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right) d\theta = \sqrt{2\pi\sigma^2} = \sqrt{\frac{2\pi}{1+1/d^2}}.$$

Thus

$$I = \sqrt{\frac{2\pi}{1+1/d^2}} \times \exp\left\{ \frac{1}{2}\frac{\left(x+\frac{b}{d^2}\right)^2}{\left(1+\frac{1}{d^2}\right)} \right\}.$$

Finally,

$$\begin{aligned} m_1(x) &= I \times \frac{1}{2\pi d} \exp\left\{ -\frac{1}{2}\left[x^2 + \frac{b^2}{d^2} \right] \right\} \\ &= \exp\left\{ -\frac{1}{2}\left[x^2 + \frac{b^2}{d^2} - \frac{\left(x + \frac{b}{d^2}\right)^2}{\left(1 + \frac{1}{d^2}\right)} \right] \right\} \\ &\quad \times \frac{1}{\sqrt{2\pi(d^2+1)}} \\ &= \exp\left\{ -\frac{1}{2}\left[\frac{x^2 + b^2 - 2bx}{d^2\left(1 + \frac{1}{d^2}\right)} \right] \right\} \\ &\quad \times \frac{1}{\sqrt{2\pi(d^2+1)}} \\ &= \frac{1}{\sqrt{2\pi(d^2+1)}} \exp\left\{ -\frac{(x-b)^2}{2(d^2+1)} \right\}. \end{aligned}$$

Hence the Bayes Factor:

$$\frac{m_1(x)}{m_0(x)} = \frac{1}{\sqrt{d^2+1}} \exp\left\{ -\frac{(x-b)^2}{2(d^2+1)} + \frac{x^2}{2} \right\}.$$

For instance, in the case $b = 0$, $d = 1$, the data indicates that model M_1 should be preferred to model

M_0 if and only if

$$\begin{aligned} \frac{m_1(x)}{m_0(x)} &> 1 \\ \Leftrightarrow 2^{-1/2} \exp\left\{-\frac{x^2}{4} + \frac{x^2}{2}\right\} &> 1 \\ \Leftrightarrow \frac{x^2}{4} &> \frac{1}{2}\log 2 \\ \Leftrightarrow |x| &> \sqrt{2\log 2}. \end{aligned}$$

Note This is a rather technical derivation, even though the prior chosen is a conjugate prior. It is fair to say that, in general, Bayesian model choice is harder to implement than frequentist model choice. Fortunately, there are many cases(in the Bayesian setting) in which choosing a conjugate prior makes computations very easy, and we study one such example below.

Example Suppose

$$\begin{aligned} M_0 &: X \sim \text{Binomial}(n, \theta), \quad \theta = 1/2 \\ M_1 &: X \sim \text{Binomial}(n, \theta), \quad \theta \sim \text{Beta}(a, b). \end{aligned}$$

Then for $x = 0, 1, \ldots, n$, $m_0(x) = \binom{n}{x}(1/2)^n$ and

$$\begin{aligned} m_1(x) &= \int_0^1 p_X(x|\theta)\,\pi(\theta)\,d\theta \\ &= \binom{n}{x}\frac{1}{B(a,b)}\int_0^1 \theta^{a+x-1}(1-\theta)^{b+n-x-1}d\theta \\ &= \binom{n}{x}\frac{B(a+x, b+n-x)}{B(a,b)}. \end{aligned}$$

The latter quantity is called the beta-binomial pmf.

It follows that the Bayes Factor in favour of model M_1 against model M_0 is:

$$\frac{m_1(x)}{m_0(x)} = 2^n \frac{B(a+x, b+n-x)}{B(a,b)}.$$

5.6 Predictive Inference

Much of statistical inference aims to make inference about a parameter θ, based on a sample $x_1, \ldots, x_n$ from a population. This can in turn lead to inference about the population itself.

In a Bayesian context, there is uncertainty about θ, as described by the posterior $\pi(\theta|\boldsymbol{x})$. Given θ, there is also uncertainty about a future observation Y in the observational model $p_Y(y|\theta)$.

The predictive distribution combines these uncertainties: by the law of total probability,

$$p(y|\boldsymbol{x}) = \int_\Theta p_Y(y|\theta)\pi(\theta|\boldsymbol{x})\, d\theta.$$

Example Let $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$, $\theta \sim N(b, d^2)$, and σ^2 is known. From the example earlier,

$$\theta|\boldsymbol{x} \sim N(B, D^2),$$

where

$$B = \frac{\left(\frac{n\bar{x}}{\sigma^2} + \frac{b}{d^2}\right)}{\left(\frac{n}{\sigma^2} + \frac{1}{d^2}\right)}, \quad D^2 = \left(\frac{n}{\sigma^2} + \frac{1}{d^2}\right)^{-1}.$$

Thus the predictive distribution for a future observation Y from the $N(\theta, \sigma^2)$ model is

$$\begin{aligned} p(y|\boldsymbol{x}) = & \int_{-\infty}^{\infty} (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(y-\theta)^2\right\} \\ & \times (2\pi D^2)^{-1/2} \exp\left\{-\frac{1}{2D^2}(\theta - B)^2\right\} d\theta, \end{aligned}$$

which is a similar integral to that of the example on

slide 121. Using similar techniques gives

$$p(y|\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\Sigma^2}} \exp\left\{-\frac{1}{2\Sigma^2}(y-B)^2\right\},$$

with $\Sigma^2 = D^2 + \sigma^2$, *i.e.*

$$Y|\boldsymbol{x} \sim N(B, D^2 + \sigma^2).$$

Example Suppose the data $X_1, \ldots, X_n$ are IID Poisson distributed with parameter θ, and the prior is $\theta \sim \text{Gamma}(g, h)$. Then (see slide 105)

$$\pi(\theta|\boldsymbol{x}) \sim \text{Gamma}(g + \sum x_i, h + n).$$

For a future observation $Y \sim \text{Poisson}(\theta)$, the predictive distribution is, for $y = 0, 1, 2, \ldots,$

$$\begin{aligned} p(y|\boldsymbol{x}) &= \\ &\int_0^\infty \frac{\theta^y e^{-\theta}}{y!} \frac{(h+n)^{g+\sum x_i}\, \theta^{g+\sum x_i - 1} e^{-(h+n)\theta}}{\Gamma(g+\sum x_i)}\, d\theta \\ &= \frac{(h+n)^{g+\sum x_i}}{y!\Gamma(g+\sum x_i)} \int_0^\infty \theta^{y+g+\sum x_i - 1} e^{-(h+n+1)\theta}\, d\theta \\ &= \frac{(h+n)^{g+\sum x_i}}{y!\Gamma(g+\sum x_i)} \frac{\Gamma(y+g+\sum x_i)}{(h+n+1)^{y+g+\sum x_i}} \\ &= \frac{(h+n)^{g+\sum x_i}}{(h+n+1)^{y+g+\sum x_i}} \frac{\Gamma(y+g+\sum x_i)}{\Gamma(y+1)\Gamma(g+\sum x_i)}, \end{aligned}$$

since $\Gamma(y+1) = y!$.

So for example, if $g = h \downarrow 0$, $n = 1$, $x_1 = 1$,

$$p(y|\boldsymbol{x}) = \frac{1}{2^{y+1}} \frac{\Gamma(y+1)}{\Gamma(y+1)\Gamma(1)} = \frac{1}{2^{y+1}},$$

for $y = 0, 1, 2, \ldots,$ *i.e.* essentially the geometric distribution for a fair coin toss.

Chapter 6

The case of categorical data

6.1 Binary data

An observation which can only take two possible values, usually 0 or 1 (*i.e.* a Bernoulli trial), is also called binary. This arises when coding whether an event occurs or not, or whether a treatment outcome is positive or negative, for instance.

The simplest model for binary data (e.g. 0, 1, 1, 0) is a sample of IID Bernoulli outcomes, $X_1, \ldots, X_n$, given θ with $P(X_i = 1|\theta) = \theta$. The likelihood is, for $\theta \in (0, 1)$:

$$\begin{aligned} p(\boldsymbol{x}|\theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^r(1-\theta)^{n-r} \end{aligned}$$

where $r = \sum_{i=1}^{n} x_i$ = number of successes.

We have seen in Chapter 1 that r, which has a Binomial(n, θ) distribution, is sufficient for θ.

A conjugate prior is $\theta \sim \text{Beta}(\alpha, \beta)$: recall that it is

$$\pi(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

for $\theta \in (0,1)$.

Since r is sufficient, we know (see slide 97) that the posterior distribution of θ given the data $\boldsymbol{x}$ is exactly the posterior distribution of θ given r. Thus, for $\theta \in (0,1)$,

$$\begin{aligned}\pi(\theta|r) = \pi(\theta|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\theta)\pi(\theta)\\ &\propto \theta^r(1-\theta)^{n-r}\theta^{\alpha-1}(1-\theta)^{\beta-1}\\ &= \theta^{r+\alpha-1}(1-\theta)^{n-r+\beta-1}.\end{aligned}$$

Consequently $\theta|r = \theta|\boldsymbol{x} \sim \text{Beta}(r+\alpha, n-r+\beta)$.

Recall that if $\theta \sim \text{Beta}(\alpha,\beta)$ then (see handout!)

$$\mathbb{E}(\theta) = \frac{\alpha}{\alpha+\beta} \text{ and } \text{Var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

As $\theta|r$ also has a Beta distribution,

$$\begin{aligned}\mathbb{E}(\theta|r) &= \frac{\alpha+r}{\alpha+\beta+n}\\ &= \frac{\alpha+\beta}{\alpha+\beta+n}\mathbb{E}(\theta) + \frac{n}{\alpha+\beta+n}\widehat{\theta}\end{aligned}$$

where $\widehat{\theta} = r/n$ is the classical MLE. So the posterior expectation of θ is a weighted average of the estimates from the prior and the likelihood.

6.2 Categorical data

Binary data are the simplest case of categorical data. An observation is categorical if we just observe which of k possible categories it falls into. Binary data correspond to $k = 2$.

Suppose that $X_1, \ldots, X_n$ are independent, identically distributed categorical observations given $\boldsymbol{\theta}$, where $\mathbb{P}(X_i = j) = \theta_j$, $j = 1, \ldots, k$.

The probabilities θ_j are such that

$$\theta_j \geq 0 \quad (j = 1, \ldots, k) \quad \text{and} \quad \sum_{j=1}^{k} \theta_j = 1.$$

If n_j of the x_i's are observed to fall in category j, and we write $\mathbf{n} = (n_1, \ldots, n_k)$, then

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{j=1}^{k} \theta_j^{n_j} \quad \text{and} \quad p(\mathbf{n}|\boldsymbol{\theta}) = \frac{n!}{n_1! \ldots n_k!} p(\boldsymbol{x}|\boldsymbol{\theta}).$$

Note that the multinomial coefficient

$$\binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1! \ldots n_k!}$$

is the number of ways to have n_1 observations in category 1, n_2 observations in category 2, ... among all $n = n_1 + n_2 + \ldots + n_k$ observations. It coincides with the usual binomial coefficient for $k = 2$.

6.3 Dirichlet distribution

The natural conjugate prior family for multinomial data is the Dirichlet distribution with parameters $\alpha_1, \alpha_2, \ldots, \alpha_k > 0$, which has pdf

$$\pi(\boldsymbol{\theta}) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \quad \text{with} \quad \theta_j \geq 0, \ \sum_{j=1}^{k} \theta_j = 1.$$

In short, $\boldsymbol{\theta} \sim \text{Dir}_k(\alpha_1, \alpha_2, \ldots, \alpha_k)$, and we will write $\mathcal{S}_k$ for the unit simplex

$$\mathcal{S}_k = \left\{ \theta_j \geq 0, \ \sum_{j=1}^{k} \theta_j = 1 \right\}.$$

Note that $\text{Dir}_2(\alpha_1, \alpha_2) = \text{Beta}(\alpha_1, \alpha_2)$, making the Dirichlet distribution a generalisation of the Beta. Observe that the posterior, given $\mathbf{n} = (n_1, \ldots, n_k)$, is

$$\begin{aligned} \pi(\boldsymbol{\theta}|\mathbf{n}) \propto p(\mathbf{n}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \ &\propto \prod_{j=1}^{k} \theta_j^{n_j} \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \\ &= \prod_{j=1}^{k} \theta_j^{n_j + \alpha_j - 1}, \ \boldsymbol{\theta} \in \mathcal{S}_k \end{aligned}$$

and so $\boldsymbol{\theta}|\mathbf{n} \sim \text{Dir}_k(n_1 + \alpha_1, n_2 + \alpha_2, \ldots, n_k + \alpha_k)$: the Dirichlet distribution is a conjugate of the multinomial distribution.

The Dirichlet pdf in full is given by

$$\pi(\boldsymbol{\theta}) = \frac{1}{D(\boldsymbol{\alpha})} \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}, \quad \boldsymbol{\theta} \in \mathcal{S}_k$$

where the Dirichlet function is

$$D(\boldsymbol{\alpha}) = D(\alpha_1, \ldots, \alpha_k) = \frac{\prod_{j=1}^{k} \Gamma(\alpha_j)}{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}.$$

This gives us the normalising constant when we need it. Note that the Dirichlet function is equal to the Beta function when $k = 2$.

The prior expectation of each θ_i is

$$\begin{aligned}
\mathbb{E}(\theta_i) &= \frac{1}{D(\alpha_1, \ldots, \alpha_k)} \int_{\mathcal{S}_k} \theta_i \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} d\boldsymbol{\theta} \\
&= \frac{1}{D(\alpha_1, \ldots, \alpha_k)} \int_{\mathcal{S}_k} \theta_i^{\alpha_i} \prod_{j \neq i} \theta_j^{\alpha_j - 1} d\boldsymbol{\theta} \\
&= \frac{D(\beta_1, \ldots, \beta_k)}{D(\alpha_1, \ldots, \alpha_k)}
\end{aligned}$$

where $\beta_i = \alpha_i + 1$ and $\beta_j = \alpha_j$ for $j \neq i$.

Writing $\alpha_0 = \sum_{j=1}^{k} \alpha_j$, it follows that

$$\begin{aligned}
\mathbb{E}(\theta_i) &= \frac{\Gamma(\alpha_i + 1) \prod_{j \neq i} \Gamma(\alpha_j)}{\Gamma(\alpha_0 + 1)} \frac{\Gamma(\alpha_0)}{\prod_{j=1}^{k} \Gamma(\alpha_j)} \\
&= \frac{\Gamma(\alpha_i + 1)}{\Gamma(\alpha_0 + 1)} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_i)} \\
&= \frac{\alpha_i}{\alpha_0} \text{ because } \Gamma(x+1) = x\Gamma(x) \\
&= \frac{\alpha_i}{\sum_{j=1}^{k} \alpha_j}.
\end{aligned}$$

It can similarly be shown that

$$\mathbb{E}(\theta_i^2) = \frac{\alpha_i(\alpha_i + 1)}{\alpha_0(\alpha_0 + 1)}.$$

Observe that these formulae are close to those for the Beta distribution in the previous section. In fact, this is because each θ_i marginally has a Beta distribution, and more precisely:

$$\theta_i \sim \text{Beta}\left(\alpha_i, \sum_{j \neq i} \alpha_j\right).$$

It is also possible to calculate covariances between the θ_j's. Write $\alpha_0 = \sum_{j=1}^k \alpha_j$ and note that

$$\begin{aligned}
\mathbb{E}(\theta_1\theta_2) &= \frac{1}{D(\alpha_1,\ldots,\alpha_k)}\int_{\mathcal{S}_k} \theta_1\theta_2 \prod_{j=1}^k \theta_j^{\alpha_j-1} d\boldsymbol{\theta} \\
&= \frac{1}{D(\alpha_1,\ldots,\alpha_k)}\int_{\mathcal{S}_k} \theta_1^{\alpha_1}\theta_2^{\alpha_2} \prod_{j=3}^k \theta_j^{\alpha_j-1} d\boldsymbol{\theta} \\
&= \frac{D(\alpha_1+1,\alpha_2+1,\alpha_3,\ldots,\alpha_k)}{D(\alpha_1,\ldots,\alpha_k)} \\
&= \frac{\Gamma(\alpha_1+1)\Gamma(\alpha_2+1)}{\Gamma(\alpha_0+2)}\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \\
&= \frac{\alpha_1\alpha_2}{\alpha_0(\alpha_0+1)}.
\end{aligned}$$

Hence

$$\begin{aligned}
\mathrm{Cov}(\theta_1,\theta_2) &= \mathbb{E}(\theta_1\theta_2) - \mathbb{E}(\theta_1)\mathbb{E}(\theta_2) \\
&= \frac{-\alpha_1\alpha_2}{\alpha_0^2(\alpha_0+1)}.
\end{aligned}$$

In general

$$\mathrm{Cov}(\theta_j,\theta_{j'}) = \frac{-\alpha_j\alpha_{j'}}{\left(\sum_{j=1}^k \alpha_j\right)^2\left(1+\sum_{j=1}^k \alpha_j\right)},\ j\neq j'.$$

The covariance formula makes intuitive sense. If one θ_j is large then we expect others to be correspondingly smaller, due to the constraints $\sum \theta_j = 1$ and $\theta_j \geq 0$, $j = 1,\ldots,k$.

Expectations and covariances may similarly be calculated for the posterior distribution as this is also a Dirichlet distribution.

Remember the Dirichlet distribution is $k-1$ dimensional and not k dimensional. There are really only $k-1$ free variables, since fixing any $k-1$ of them fixes the kth by $\sum \theta_j = 1$. The space of possible $\boldsymbol{\theta}$ values is the $(k-1)$-dimensional set $\mathcal{S}_k$.

6.4 Limitations of Dirichlet distribution

The Dirichlet has k parameters $\alpha_1, \ldots, \alpha_k$ to represent prior information about the $k-1$ free θ_j's.

So far, we assign some values to $E[\theta_1]$, $E[\theta_2]$ etc. Given these values, we can choose $\alpha_1, \alpha_2, \ldots, \alpha_k$.

Question: How feasible is this?

Suppose we fix the prior expectations

$$\mathbb{E}(\theta_1), \ldots, \mathbb{E}(\theta_{k-1})$$

by

$$\mathbb{E}(\theta_i) = \frac{\alpha_i}{\sum_{j=1}^{k} \alpha_j} = t_i, \;\; i = 1, \ldots, k-1.$$

Then $\mathbb{E}(\theta_k)$ is fixed as well; besides, $\alpha_i = t_i \alpha_0$, where as before $\alpha_0 = \sum_{j=1}^{k} \alpha_i$.

Having fixed the t_i's, only α_0 (or equivalently α_k) is

left for us to choose. But

$$\text{Var}(\theta_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)} = \frac{t_i(1 - t_i)}{\alpha_0 + 1}.$$

So fixing any one variance fixes α_0. Note that all other variances are fixed, and all covariances. So the Dirichlet family is not flexible enough to handle all possible configurations of prior beliefs about the θ_j's.

This can be compared to the multivariate normal prior $N_{k-1}(\boldsymbol{\mu}, \boldsymbol{V})$, which has $k-1$ free parameters in the mean vector $\boldsymbol{\mu}$ and $\frac{1}{2}(k-1)k$ free parameters in the covariance matrix $\boldsymbol{V}$. This multivariate prior, though, can yield negative values and therefore is not suitable to categorical data.

6.5 Marginal distribution/model comparison

Model comparison was previously considered in Chapter 5 in the single-parameter case but can be carried out on multinomial data as well.

Suppose a model M has

$$\mathbf{n}|\boldsymbol{\theta} \sim \text{Mult}_k(n, \boldsymbol{\theta}) \text{ and } \boldsymbol{\theta} \sim \text{Dir}_k(\alpha_1, \alpha_2, \ldots, \alpha_k)$$

where $\mathbf{n} = (n_1, n_2, \ldots, n_k)$ and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$.

Recall that the Bayes factor in favour of model M_1 against model M_0 given data $\boldsymbol{x}$ is $m_1(\boldsymbol{x})/m_0(\boldsymbol{x})$.

For the above model the marginal distribution of $\mathbf{n}$ is

$$
\begin{aligned}
& m(\mathbf{n}) \\
=& \int_{\mathcal{S}_k} p(\mathbf{n}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathbf{d}\boldsymbol{\theta} \\
=& \frac{n!}{n_1!n_2!\cdots n_k!}\frac{1}{D(\boldsymbol{\alpha})}\int_{\mathcal{S}_k}\prod_{j=1}^{k}\theta_j^{n_j}\prod_{j=1}^{k}\theta_j^{\alpha_j-1}d\boldsymbol{\theta} \\
=& \frac{n!}{n_1!n_2!\cdots n_k!}\frac{1}{D(\alpha_1,\alpha_2,\ldots,\alpha_k)}\int_{\mathcal{S}_k}\prod_{j=1}^{k}\theta_j^{n_j+\alpha_j-1}d\boldsymbol{\theta}.
\end{aligned}
$$

Now the above integral is the normalising constant of the $\text{Dir}_k(n_1+\alpha_1, n_2+\alpha_2, \ldots, n_k+\alpha_k)$ distribution, namely

$$D(n_1+\alpha_1, n_2+\alpha_2, \ldots, n_k+\alpha_k)$$

and so

$$m(\mathbf{n}) = \frac{n!}{n_1!n_2!\cdots n_k!}\frac{D(\mathbf{n}+\boldsymbol{\alpha})}{D(\boldsymbol{\alpha})}.$$

This identity can be used for model comparison.

Similar calculations can of course be carried out for the purpose of predictive inference (see Chapter 5; Section 5.6) of Multinomial-Dirichlet models.

6.6 Application: Contingency Tables

Contingency tables are a particular way of considering multinomial data in which the categories can be laid out in a rectangular grid.

For example, if we were interested in studying whether the probability of being left-handed depends on gender, we could use a 2×2 grid with categories LF, LM, RF, RM. The contingency table of the data would be given by

	Female	Male
Left-handed	n_{lf}	n_{lm}
Right-handed	n_{rf}	n_{rm}

The likelihood of this multinomial data with four categories is

$$\frac{n!}{n_{lf}!n_{lm}!n_{rf}!n_{rm}!}\theta_{lf}^{n_{lf}}\theta_{lm}^{n_{lm}}\theta_{rf}^{n_{rf}}\theta_{rm}^{n_{rm}}$$

where $\theta_{ij} \geq 0$ is the probability of using primarily hand i and being of gender j.

Of particular interest is whether handedness depends on gender, *i.e.*

$$\mathbb{P}(\text{Left}|\text{Female}) = \mathbb{P}(\text{Left}|\text{Male}).$$

Because

$$p = q \Leftrightarrow \frac{p}{1-p} = \frac{q}{1-q}$$

this is equivalent to the statement

$$\frac{\mathbb{P}(\text{Left}|\text{Female})}{\mathbb{P}(\text{Right}|\text{Female})} = \frac{\mathbb{P}(\text{Left}|\text{Male})}{\mathbb{P}(\text{Right}|\text{Male})}. \qquad (*)$$

Observe that $\mathbb{P}(\text{Left}|\text{Female}) = \theta_{lf}/(\theta_{lf} + \theta_{rf})$ with similar expressions for the others. Hence $(*)$ is equivalent to

$$\frac{\theta_{lf}}{\theta_{rf}} = \frac{\theta_{lm}}{\theta_{rm}}.$$

This is in turn equivalent to $\rho = 1$ where

$$\rho = \frac{\theta_{lf}\theta_{rm}}{\theta_{rf}\theta_{lm}}.$$

The quantity ρ is called the odds ratio.

Assuming a Dirichlet prior distribution, the posterior expectation and variance of the odds ratio may be calculated.

Chapter 7

The case of normal data

7.1 Likelihood

Suppose that a sample $x_1, \ldots, x_n$ has been obtained from the observation of an IID sample $X_1, \ldots, X_n$ having a normal distribution, *i.e.*

$$X_i|\mu, \sigma^2 \sim N(\mu, \sigma^2), \quad i = 1, 2, \ldots, n.$$

In other words, this is our model for the observed data.

Therefore

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

and the likelihood is

$$p(\boldsymbol{x}|\mu,\sigma^2) = \prod_{i=1}^{n} p(x_i|\mu,\sigma^2)$$
$$\propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}.$$

We now look for an alternative way of writing this, more suited to conjugate analysis and predictive inference. Note that:

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\mu)^2 &= \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2 \\ &= \sum_{i=1}^{n}(x_i-\bar{x})^2 + \sum_{i=1}^{n}(\bar{x}-\mu)^2 \\ &+ 2\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-\mu) \\ &= \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\end{aligned}$$

We then use the bias-variance decomposition:

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\mu)^2 &= n\left[(\mu-\bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] \\ &= n\left[(\mu-\bar{x})^2 + S^2\right]\end{aligned}$$

where S^2 is the sample variance:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

which does not depend on μ or σ^2.

Therefore

$$p(\boldsymbol{x}|\mu, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{n}{2\sigma^2}\left[(\mu - \bar{x})^2 + S^2\right]\right\}.$$

This way of writing the likelihood will prove very useful throughout this chapter.

7.2 Case σ^2 known, μ unknown

When σ^2 is known, the usual conjugate prior for μ is normal (exercise!). Suppose then $\mu \sim N(m, w)$, so

$$\pi(\mu) \propto \exp\left\{-\frac{1}{2w}(\mu - m)^2\right\}$$

and therefore $\pi(\mu|\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}Q\right)$

where Q is given by

$$\begin{aligned} Q &= \frac{1}{w}(\mu - m)^2 + \frac{n}{\sigma^2}(\mu - \bar{x})^2 \\ &= \frac{1}{w}(\mu^2 - 2m\mu + m^2) + \frac{n}{\sigma^2}(\mu^2 - 2\mu\bar{x} + \bar{x}^2) \\ &= \left[\frac{1}{w} + \frac{n}{\sigma^2}\right]\mu^2 - 2\mu\left[\frac{1}{w}m + \frac{n}{\sigma^2}\bar{x}\right] + \text{constant} \\ &= \frac{1}{w_1}(\mu - m_1)^2 + \text{constant} \end{aligned}$$

with

$$w_1 = \left(\frac{1}{w} + \frac{n}{\sigma^2}\right)^{-1} \text{ and } m_1 = w_1\left(\frac{1}{w}m + \frac{n}{\sigma^2}\bar{x}\right).$$

Therefore $\mu|\boldsymbol{x} \sim N(m_1, w_1)$.

Note that

$$\mathbb{E}(\mu|\boldsymbol{x}) = m_1 = \frac{w^{-1}}{w^{-1} + n\sigma^{-2}}m + \frac{n\sigma^{-2}}{w^{-1} + n\sigma^{-2}}\bar{x}$$

is a weighted average of prior mean m and sample mean $\bar{x}$, with weights reflecting the strength of each information source, *i.e.* prior and data. When $n \to \infty$ we have $\mathbb{E}(\mu|\boldsymbol{x}) \to \bar{x}$, *i.e.* the data takes over.

7.3 Case μ known, σ^2 unknown

When μ is known, the usual conjugate prior for σ^2 is an Inverse Gamma (IG) distribution. The pdf of the IG(a, d) distribution is

$$p(y) = \frac{(a/2)^{(d/2)}}{\Gamma(d/2)} y^{-(d+2)/2} \exp\left(-\frac{a}{2y}\right), \ y > 0.$$

If Y has an IG(a, d) distribution and $X = 1/Y$, observe that for $x > 0$,

$$\begin{aligned} p_X(x) &= p_Y(y = 1/x) \left|\frac{dy}{dx}\right| \\ &= \frac{(a/2)^{(d/2)}}{\Gamma(d/2)} \left(\frac{1}{x}\right)^{-(d+2)/2} \exp\left(-\frac{a}{2} \cdot \frac{1}{x^{-1}}\right) \frac{1}{x^2} \\ &= \frac{(a/2)^{(d/2)}}{\Gamma(d/2)} x^{(d/2)-1} \exp\left(-\frac{ax}{2}\right). \end{aligned}$$

Thus, $X = 1/Y$ has a Gamma$(d/2, a/2)$ distribution, hence the name Inverse Gamma for Y. Similarly, if $Y \sim$ IG(a, d) then $a/Y \sim \chi^2_d$, so this is sometimes called an inverse-chi squared distribution.

The expectation of an $\mathrm{IG}(a, d)$ is

$$\begin{aligned}\mathbb{E}(Y) &= \int_0^\infty y p(y) dy \\ &= \frac{(a/2)^{(d/2)}}{\Gamma(d/2)} \int_0^\infty y^{-d/2} \exp\left(-\frac{a}{2y}\right) dy.\end{aligned}$$

The integral just gives the normalising constant for an $\mathrm{IG}(a, d-2)$, and thus

$$\begin{aligned}\mathbb{E}(Y) &= \frac{(a/2)^{(d/2)}}{\Gamma(d/2)} \frac{\Gamma(d/2-1)}{(a/2)^{(d/2-1)}} \\ &= \frac{a/2}{d/2-1} \text{ because } \Gamma(x+1) = x\Gamma(x),\end{aligned}$$

that is, $\mathbb{E}(Y) = a/(d-2)$ if $d > 2$.

If now the prior distribution for the variance, $\sigma^2 \sim \mathrm{IG}(a, d)$, then for $\sigma^2 > 0$,

$$\pi(\sigma^2) \propto (\sigma^2)^{-(d+2)/2} \exp\left\{-\frac{a}{2\sigma^2}\right\}$$

and

$$\begin{aligned}
\pi(\sigma^2|\boldsymbol{x}) &\propto p(\boldsymbol{x}|\sigma^2)\pi(\sigma^2) \\
&\propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} S_\mu^2\right\} \\
&\quad \times (\sigma^2)^{-(d+2)/2} \exp\left\{-\frac{a}{2\sigma^2}\right\} \\
&= (\sigma^2)^{-(n+d+2)/2} \exp\left\{-\frac{1}{2\sigma^2}(S_\mu^2 + a)\right\}
\end{aligned}$$

where $S_\mu^2 = \sum_{i=1}^n (x_i - \mu)^2$. Hence

$$\sigma^2 \,|\, \boldsymbol{x} \sim \text{IG}\left(a + S_\mu^2,\, d + n\right).$$

We have just completed a conjugate analysis, with both prior and posterior given by an Inverse Gamma distribution.

7.4 General case μ and σ^2 unknown

When both parameters μ and σ^2 are unknown, a convenient class of conjugate prior distributions is

$$\pi(\mu, \sigma^2) \propto (\sigma^2)^{-(d+3)/2} \exp\left\{-\frac{1}{2\sigma^2}\left[\frac{1}{v}(\mu - m)^2 + a\right]\right\}$$

for $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

This is called the Normal Inverse Gamma (NIG) distribution, with parameters a, d, m, v. Symbolically,

$$(\mu, \sigma^2) \sim \text{NIG}(a, d, m, v).$$

The pdf of NIG(a, d, m, v), for $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, is

$$\begin{aligned}\pi(\mu, \sigma^2) &= \frac{(a/2)^{(d/2)}}{\Gamma(d/2)\sqrt{2\pi v}}(\sigma^2)^{-(d+3)/2} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma^2}\left[\frac{(\mu - m)^2}{v} + a\right]\right\}.\end{aligned}$$

This gives us the normalising constant when needed.

Here μ and σ^2 are not independent, but we can derive the conditional prior distribution $\mu|\sigma^2$ (ie we look at the distribution of μ conditional on some (fixed) value for σ^2) as

$$\begin{aligned}\pi(\mu|\sigma^2) = \frac{\pi(\mu,\sigma^2)}{\pi(\sigma^2)} &\propto \pi(\mu,\sigma^2) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\frac{(\mu-m)^2}{v}\right\}\end{aligned}$$

using the proportionality in μ and treating σ^2 as known. This clearly implies that $\mu|\sigma^2 \sim \mathrm{N}(m, v\sigma^2)$.

Similarly, treating μ as known, the prior conditional distribution $\sigma^2|\mu$ is, for $\sigma^2 > 0$,

$$\pi(\sigma^2|\mu) \propto \pi(\mu,\sigma^2) \propto (\sigma^2)^{-(d+3)/2} \times \exp\left\{-\frac{1}{2\sigma^2}\left[\frac{1}{v}(\mu-m)^2 + a\right]\right\}.$$

From the definition of an Inverse Gamma distribution (ie recognising that the above expression looks like the pdf of an IG distribution),

$$\sigma^2|\mu \sim \mathrm{IG}\left(a + \frac{(\mu-m)^2}{v}, d+1\right).$$

The prior marginal distribution of σ^2 is derived from

$$\begin{aligned}\pi(\sigma^2) &= \int_{-\infty}^{\infty} \pi(\mu, \sigma^2) d\mu \\ &= \frac{(a/2)^{(d/2)}}{\Gamma(d/2)\sqrt{2\pi v}} (\sigma^2)^{-(d+3)/2} \exp\left\{-\frac{a}{2\sigma^2}\right\} \\ &\quad \times \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}\frac{(\mu - m)^2}{v}\right\} d\mu \text{ if } \sigma^2 > 0.\end{aligned}$$

Now

$$\frac{1}{\sqrt{2\pi\sigma^2 v}} \exp\left\{-\frac{1}{2\sigma^2}\frac{(\mu - m)^2}{v}\right\}, \ \mu \in \mathbb{R}$$

is a $N(m, v\sigma^2)$ density and so integrates to one.

Therefore

$$\pi(\sigma^2) = \frac{(a/2)^{(d/2)}}{\Gamma(d/2)} (\sigma^2)^{-(d+2)/2} \exp\left\{-\frac{a}{2\sigma^2}\right\}.$$

Thus, $\sigma^2 \sim \text{IG}(a, d)$.

In other words, marginally, σ^2 follows an $IG(a, d)$ distribution.

The prior marginal distribution of μ can be evaluated similarly from the equation

$$\pi(\mu) = \int_0^\infty \pi(\mu, \sigma^2) d\sigma^2.$$

One can show that μ has a Generalised t distribution, with parameters d, m and av/d. The Generalised t Distribution $t_d(m, w)$ is a location-scale generalisation of the Student t distribution with d degrees of freedom. Its pdf is, for $x \in \mathbb{R}$,

$$p(x) = \frac{1}{\sqrt{\pi w d}} \frac{\Gamma\big((d+1)/2\big)}{\Gamma(d/2)} \left\{ 1 + \frac{(x-m)^2}{wd} \right\}^{-(d+1)/2}.$$

The Student distribution is obtained with an integer d (it does not have to be so in the generalised form), location $m = 0$ and scale $w = 1$.

Like the Student distribution, this distribution has symmetry properties: the $t_d(m, w)$ distribution is symmetric about m, the median (and unique mode) of the distribution.

The moments of the $t_d(m, w)$ distribution are

$$\mathbb{E}(X) = m \quad \text{if} \quad d > 1,$$

$$\mathrm{Var}(X) = \frac{wd}{d-2} \quad \text{if} \quad d > 2.$$

Thus, if $\mu \sim t_d\,(m, av/d)$:

$$\mathbb{E}(\mu) \;=\; m \;\text{ if }\; d > 1,$$

$$\mathrm{Var}(\mu) \;=\; \frac{av}{d-2} \;\text{ if }\; d > 2.$$

7.5 Conjugate analysis in the general case

Suppose (as will be the case from now on) we use the prior $(\mu, \sigma^2) \sim \mathrm{NIG}(a, d, m, v)$. Since we are told that the NIG pror is conjugate for μ and σ^2, that means that the posterior will also be of the same family, i.e. a NIG distribution.

The posterior based on an IID sample from the $N(\mu, \sigma^2)$ is proportional to

$$(\sigma^2)^{-(d+3)/2} \exp\left\{-\frac{1}{2\sigma^2}\left[\frac{1}{v}(\mu - m)^2 + a\right]\right\} \;\text{(prior)}$$

$$\times(\sigma^2)^{-n/2} \exp\left\{-\frac{n}{2\sigma^2}[(\mu - \bar{x})^2 + S^2]\right\} \;\text{(likelihood)}$$

$$=(\sigma^2)^{-(d+n+3)/2} \exp\left\{-\frac{1}{2\sigma^2}Q\right\}$$

where

$$\begin{aligned} Q &= \frac{1}{v}(\mu - m)^2 + a + n\left[(\mu - \bar{x})^2 + S^2\right] \\ &= \left(\frac{1}{v} + n\right)\mu^2 - 2\mu\left(\frac{1}{v}m + n\bar{x}\right) \\ &\quad + \frac{m^2}{v} + n\bar{x}^2 + a + nS^2 \\ &= \left(\frac{1}{v} + n\right)\left[\mu^2 - 2\mu m_1\right] + \frac{m^2}{v} + n\bar{x}^2 + a + nS^2 \end{aligned}$$

with

$$m_1 = \frac{\frac{1}{v}m + n\bar{x}}{\frac{1}{v} + n}.$$

Completing the square in μ and tidying up, we find

$$Q = \left(\frac{1}{v} + n\right)(\mu - m_1)^2 + a_1$$

where

$$a_1 = \frac{m^2}{v} + n\bar{x}^2 + a + nS^2 - \left(\frac{1}{v} + n\right)m_1^2.$$

(e.g. just a constant)

Since

$$
\begin{aligned}
& \frac{m^2}{v} + n\bar{x}^2 - \left(\frac{1}{v} + n\right) m_1^2 \\
= & \frac{(\frac{m^2}{v} + n\bar{x}^2)(\frac{1}{v} + n) - \frac{m^2}{v^2} - 2\frac{m}{v}n\bar{x} - n^2\bar{x}^2}{\frac{1}{v} + n} \\
= & \frac{n\frac{m^2}{v} - 2\frac{m}{v}n\bar{x} + n\bar{x}^2\frac{1}{v}}{\frac{1}{v} + n} \\
= & \frac{(\bar{x} - m)^2}{\frac{1}{n} + v}
\end{aligned}
$$

we get

$$a_1 = a + nS^2 + \left(v + \frac{1}{n}\right)^{-1} (\bar{x} - m)^2.$$

We then recognise the posterior as an NIG distribution

$$(\mu, \sigma^2)|\boldsymbol{x} \sim \text{NIG}(a_1, d_1, m_1, v_1)$$

with

$$d_1 = d + n, \;\; v_1 = \left(\frac{1}{v} + n\right)^{-1}.$$

As the posterior joint distribution of μ and σ^2 is of the same form as the prior, just with different parameters, the posterior marginal and conditional distributions of μ and σ^2 will have the same form also.

Using what we have done before:

$$\begin{aligned}
\mu \,|\, \boldsymbol{x} &\sim t_{d_1}\left(m_1, \frac{a_1 v_1}{d_1}\right) \\
\sigma^2 \,|\, \boldsymbol{x} &\sim \mathrm{IG}(a_1, d_1) \\
\mu \,|\, \sigma^2, \boldsymbol{x} &\sim \mathrm{N}(m_1, v_1\sigma^2) \\
\sigma^2 \,|\, \mu, \boldsymbol{x} &\sim \mathrm{IG}\left(a_1 + \frac{(\mu - m_1)^2}{v_1}, d_1 + 1\right).
\end{aligned}$$

7.6 Marginal inference about μ

What can we say about $\mu|\mathbf{x}$ i.e. by ignoring σ^2 (which is unknown)?

We will need to look at $\pi(\mu|\mathbf{x}) = \int_{\sigma^2} \pi(\mu, \sigma^2|\mathbf{x})\ \mathrm{d}\sigma^2$, integrating σ^2 of the joint posterior distribution.

The posterior mean of μ is

$$\mathbb{E}(\mu|\boldsymbol{x}) = m_1 = \frac{\frac{1}{v}m + n\bar{x}}{\frac{1}{v} + n} = \frac{\frac{1}{v}}{\frac{1}{v} + n}\mathbb{E}(\mu) + \frac{n}{\frac{1}{v} + n}\bar{x}$$

i.e. a weighted average of the prior mean $\mathbb{E}(\mu)$ and the sample mean $\bar{x}$, with weights depending on the balance between information in the prior and information in the data. When $n \to \infty$, $\mathbb{E}(\mu|\boldsymbol{x}) \approx \bar{x}$

and the influence of the prior disappears.

Note also that the posterior degrees of freedom $d_1 = d + n$ of $\mu|\boldsymbol{x}$ is the sum of the prior degrees of freedom, d, and the degrees of freedom in the data, n.

We saw earlier that

$$\text{Var}(\mu) = v\frac{a}{d-2} = v\mathbb{E}(\sigma^2).$$

Similarly, for the posterior:

$$\text{Var}(\mu|\boldsymbol{x}) = v_1\frac{a_1}{d_1-2} = v_1\mathbb{E}(\sigma^2|\boldsymbol{x}).$$

So the strength of our posterior knowledge about μ, and how that compares with the strength of our prior knowledge about μ, is affected by two components. One factor is

$$v_1 = \left(\frac{1}{v} + n\right)^{-1} < \left(\frac{1}{v}\right)^{-1} = v.$$

So if $\mathbb{E}(\sigma^2|\boldsymbol{x}) = \mathbb{E}(\sigma^2)$, our uncertainty about μ will definitely be reduced by the data and the larger n, the bigger that reduction. As $n \to \infty$, $v_1 \to 0$ and hence $\text{Var}(\mu|\boldsymbol{x}) \to 0$.

Ultimately, given enough data we will know μ.

But the other factor is what we learn about σ^2. If,

after seeing the data, our expectation of σ^2 exceeds a certain limit, it is possible that $\text{Var}(\mu|\boldsymbol{x}) > \text{Var}(\mu)$.

Thus, we could end up being less sure about μ. This happens if the data are much more variable than we expected (from prior information).

7.7 Marginal inference about σ^2

Recall that $\sigma^2|\boldsymbol{x} \sim \text{IG}(a_1, d_1)$ with posterior mean

$$\frac{a_1}{d_1 - 2} = \frac{a + nS^2 + (v + \frac{1}{n})^{-1}(\bar{x} - m)^2}{d + n - 2}.$$

We can see it as a weighted average of three things:

Prior mean of σ^2: $\dfrac{a}{d-2}$, with weight $\dfrac{d-2}{d+n-2}$.

Classical estimate: $\dfrac{nS^2}{n-1}$, with weight $\dfrac{n-1}{d+n-2}$.

Third component: $\dfrac{(\bar{x}-m)^2}{v+\frac{1}{n}}$, with weight $\dfrac{1}{d+n-2}$.

The third component arises because μ and σ^2 are not independent a priori. Hence $\bar{x}$ provides some information about σ^2.

More precisely:

$$\begin{aligned} & \bar{x}|\mu, \sigma^2 \sim N(\mu, \sigma^2/n) \text{ and } \mu|\sigma^2 \sim N(m, \sigma^2 v) \\ \Rightarrow\ & p(\bar{x}|\sigma^2) = \int_{\mathbb{R}} p(\bar{x}|\mu, \sigma^2)\pi(\mu|\sigma^2)d\mu \\ \Rightarrow\ & \bar{x}|\sigma^2 \sim N\left(m, \sigma^2\left(v + \frac{1}{n}\right)\right). \end{aligned}$$

7.8 Predictive inference

We saw in Chapter 5, Section 5.6 that the predictive distribution of a future independent observation y given that we have already observed $\boldsymbol{x}$ is

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

We have parameters $\boldsymbol{\theta} = (\mu, \sigma^2)$, an NIG posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ and $y|\boldsymbol{\theta} \sim N(\mu, \sigma^2)$.

Now

$$p(y|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta}|\boldsymbol{x})$$

$$\propto \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

$$\times(\sigma^2)^{-(d_1+3)/2}\exp\left\{-\frac{1}{2\sigma^2}\left[\frac{1}{v_1}(\mu-m_1)^2+a_1\right]\right\}$$

$$\propto(\sigma^2)^{-(d_1+4)/2}\exp\left\{-\frac{1}{2\sigma^2}Q\right\}$$

where

$$\begin{aligned} Q &= (y-\mu)^2+\frac{1}{v_1}(\mu-m_1)^2+a_1 \\ &= \frac{1}{v^*}(\mu-m^*)^2+a^* \end{aligned}$$

with

$$\begin{aligned} v^* &= \left(\frac{1}{v_1}+1\right)^{-1} = \left(\frac{1}{v}+n+1\right)^{-1} \\ m^* &= v^*\left(\frac{1}{v_1}m_1+y\right) = \frac{\frac{1}{v}m+n\bar{x}+y}{\frac{1}{v}+n+1} \\ a^* &= a_1+(v_1+1)^{-1}(y-m_1)^2 \end{aligned}$$

(after some algebra similar to that of Section 7.5).

Recall (from previous page) that $Q = \frac{1}{v^*}(\mu - m^*)^2 + a^*$

We can do the integration easily now:

$$p(y|\boldsymbol{x})$$

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

$$\propto \int_0^\infty \left[\int_{\mathbb{R}} (\sigma^2)^{-(d_1+4)/2} \exp\left\{-\frac{1}{2\sigma^2}Q\right\}\,d\mu\right] d\sigma^2$$

$$\propto \int_{\sigma^2=0}^\infty (\sigma^2)^{-(d_1+4)/2} \exp\left(-\frac{1}{2\sigma^2}a^*\right)$$

$$\left[\int_{\mu=-\infty}^\infty \exp\left\{-\frac{1}{2v^*\sigma^2}(\mu - m^*)^2\right\}\,d\mu\right] d\sigma^2$$

$$\propto \int_0^\infty (\sigma^2)^{-(d_1+3)/2}$$

$$\exp\left\{-\frac{1}{2\sigma^2}\left[(v_1+1)^{-1}(y-m_1)^2 + a_1\right]\right\}\,d\sigma^2$$

(note the change in power of σ^2 due to integration and that we have used $a^* = a_1 + (v_1+1)^{-1}(y-m_1)^2$).

This is the first marginal of an NIG$(a_1, d_1, m_1, v_1 + 1)$ distribution, and therefore

$$y|\boldsymbol{x} \sim t_{d_1}\left(m_1, \frac{a_1}{d_1}(v_1+1)\right),$$

see Section 7.4. This gives us the required conditional distribution.

7.9 Weak prior information?

Recall that Bayesian analysis aims to synthesise the data and prior information.

With strong prior information, the data has little impact, so posterior $\approx$ prior.

With weak prior information or large sample size, the data dictates the posterior and so posterior $\approx$ likelihood (which resembles the frequentist view).

How could we define a weak NIG prior?

Consider the posterior mean for μ:

$$\mathbb{E}(\mu|\boldsymbol{x}) = \frac{\frac{1}{v}m + n\bar{x}}{\frac{1}{v}+n} = \frac{\frac{1}{v}}{\frac{1}{v}+n}\mathbb{E}(\mu) + \frac{n}{\frac{1}{v}+n}\bar{x}.$$

So the prior mean m has weight $\frac{1}{v}$. For $v \to \infty$, then $\frac{1}{v} \to 0$ and $m_1 \to \bar{x}$.

Therefore $v \to \infty$ gives weak prior information for μ. In this case the posterior is $\text{NIG}(a_1, d_1, m_1, v_1)$ with

$$m_1 = \bar{x}, \;\; v_1 = \frac{1}{n}, \;\; a_1 = a + nS^2, \;\; d_1 = d + n.$$

Independently, and for σ^2,

$$\mathbb{E}(\sigma^2|\boldsymbol{x}) = \frac{a_1}{d_1 - 2} = \frac{a + nS^2 + (v + \frac{1}{n})^{-1}(\bar{x} - m)^2}{d + n - 2}.$$

Recovering something equivalent to the unbiased estimate of σ^2, which is $nS^2/(n-1)$, from this expression is done by setting $a = 0$ and $v \to \infty$.

This yields a posterior $\text{NIG}(a_1, d_1, m_1, v_1)$ with

$$m_1 = \bar{x}, \;\; v_1 = \frac{1}{n}, \;\; a_1 = nS^2, \;\; d_1 = d + n.$$

With this weak information on μ and σ^2,

$$\mu|\boldsymbol{x} \sim t_{d_1}\left(m_1, \frac{a_1 v_1}{d_1}\right) = t_{d+n}\left(\bar{x}, \frac{S^2}{d+n}\right).$$

In the classical setting, the MLE of μ is $\overline{X}$. If centred and divided by the empirical standard deviation, it

is Student distributed with $n-1$ d.f. This gives the idea to set $d=-1$. We then get

$$\mu|\boldsymbol{x} \sim t_{n-1}\left(\bar{x}, \frac{S^2}{n-1}\right),$$

so that

$$\frac{\mu-\bar{x}}{\sqrt{\frac{S^2}{n-1}}}|\boldsymbol{x} \sim t_{n-1} \text{ (standard Student distribution).}$$

Also

$$\sigma^2|\boldsymbol{x} \sim \text{IG}(nS^2, n-1), \text{ so } n\frac{S^2}{\sigma^2}|\boldsymbol{x} \sim \chi^2_{n-1}.$$

These closely resemble the corresponding frequentist results we saw in Section 3.6:

$$\frac{\bar{X}-\mu}{\sqrt{S_c^2/n}} \sim t_{n-1}$$

$$\text{and } \sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 = (n-1)\frac{S_c^2}{\sigma^2} \sim \chi^2_{n-1}$$

where S_c^2 is the corrected variance estimator:

$$S_c^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

Finally, for $a = 0$, $d = -1$ and $v \to \infty$, what does the "weak" prior distribution $\mathrm{NIG}(a, d, m, v)$ look like? We have in this case

$$\pi(\mu, \sigma^2) \propto (\sigma^2)^{-(d+3)/2} \exp\left\{ -\frac{1}{2\sigma^2} \left[\frac{1}{v}(\mu - m)^2 + a \right] \right\}$$

$$\propto 1/\sigma^2, \text{ for } \sigma^2 > 0.$$

This is of course an improper prior (*i.e.* it does not have a finite integral over $\mu \in \mathbb{R}$ and $\sigma^2 > 0$).

Don't forget, that it is fine to improper priors as long as the posterior distribution is proper!

Chapter 8

Introduction to Markov Chain Monte Carlo

8.1 Motivation

So far (see Chapters 5, 6 and 7) when performing Bayesian inference, we have managed to derive analytic expressions for the density of the posterior distribution $\pi(\theta|\boldsymbol{x})$ having avoided the calculation of the normalising constant $\pi(\boldsymbol{x}) = \int_\theta \pi(\theta)p(\boldsymbol{x}|\theta)\,\mathrm{d}\theta$. That was largely due to the choice of convenient (conjugate) prior distributions for θ.

In this chapter we will introduce a class of computational algorithms which will enable us to **sample** from the posterior distribution $\pi(\theta|x)$ **without knowing** (or having to compute) **the normalising constant** $\pi(\boldsymbol{x})$.

By *sampling from the posterior distribution* we mean the ability to generate samples from this distribution.

There are many situations that the normalising constant is difficult to compute e.g. the model has more than one parameters (i.e. θ is high-dimensional) and the intergral is difficult to compute.

In the next section we argue why being able to sample from the posterior distribution is almost as good as (if not better!) being able to write down the posterior density analytically.

8.2 Sample-based inference

If we can draw samples from the posterior distribution $\pi(\theta|\boldsymbol{x})$ then we can (pretty much) do everything we want/need:

- Estimate the posterior density (kernel density estimation, histogram);
- Estimate moments (eg means, variances), probabilities etc;
- Sample from the distribution of functions of the parameters $g(\theta)$ in a very straightforward manner;
- Visualise the relationship of two or more model parameters.

An Example

Suppose that the length of time spent in a queue at a supermarket, X, comes from a Gamma distribution with the following probability density function (pdf)

$$f_X(x|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \cdot \exp\{-\beta x\}, \quad \alpha,\beta > 0.$$

For any given α and β the expected value, i.e the mean of X is derived by calculating the integral

$$E[X] = \int_X x \cdot f_X(x)\, \mathrm{d}x = \ldots = \frac{\alpha}{\beta}$$

and we also know that the probability $P(X < 0.5)$ is given by $\int_0^{0.5} f_X(x)\, \mathrm{d}x$.

For some given values of α and β, e.g. $\alpha = 5$ and $\beta = 3$ we can compute these quantities:

- the mean is 5/3;
- $P(X < 0.5) = 0.01857$.

Let us assume that, for some reason, we couldn't do any of the integrals. Is there another way to estimate/approximate the mean and the probability that $X < 0.5$?

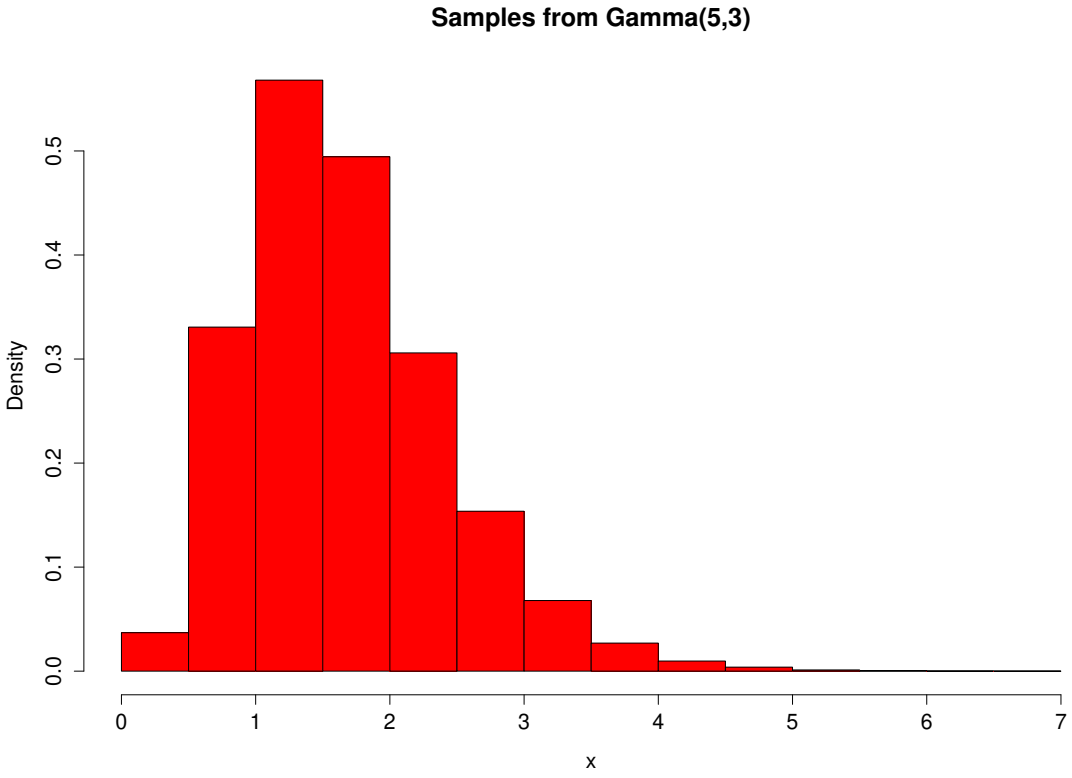
Sampling

Suppose that, somehow, we have a way (i.e. black-box) of simulating realizations from a Gamma distribution.

We can sample/draw N samples where N is a large number, e.g. $100,000$.

We plot the histogram of these N values by doing something like this in `R/Rstudio`

```
hist(rgamma(10^5, shape = 5,  rate = 3),
main="Samples from Gamma(5,3)",
xlab=expression(x),col=2, prob=TRUE)
```



Having got the N samples/realisations from the Gamma distribution we can do many things:

- "approximate" (or "estimate") the mean $E[X]$ by the sample mean, i.e.

$$\widehat{E[X]} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

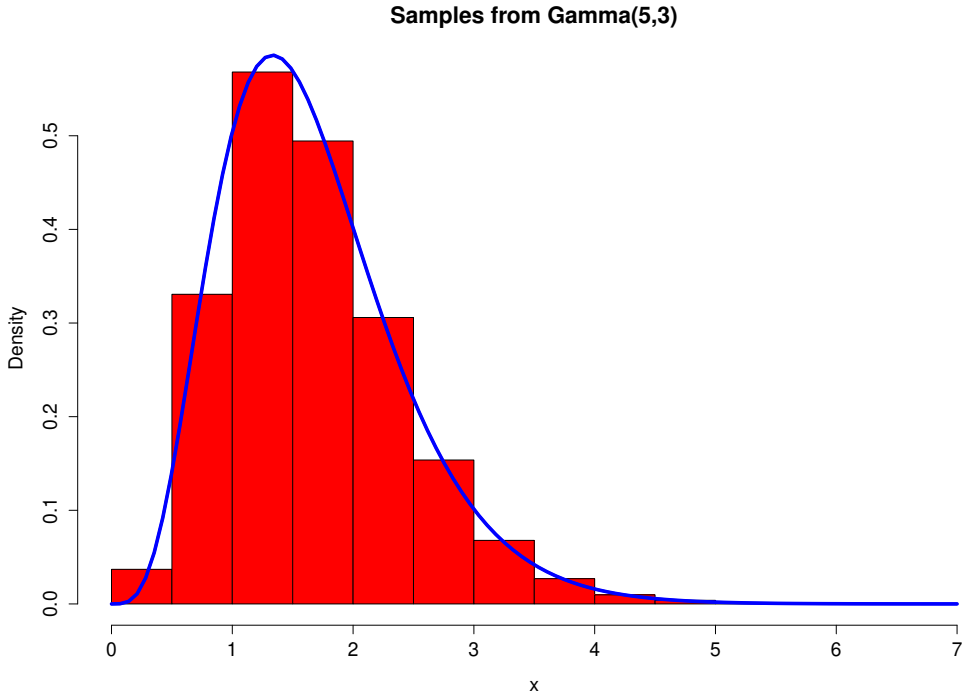
- and the probability $P(X < 0.5)$ by the proportion of the values in the sample which are less than 0.5, i.e.

$$\widehat{P(X < 0.5)} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(x_i < 0.5)$$

Of course, these "approximations" are getting better and better as $N \to> \infty$.

Exercise: run the `R` code/line above and see how close your estimates are to the true values. Consider different values of N, e.g. $N = 100$ or $N = 10^5$.

Recall that the histogram is an estimate of the (true) probability density function of the Gamma distribution:



We see that the histogram matches very well the true pdf (shown in blue).

Remark In the particular case above of the Gamma distribution we used the built-in command `rgamma` to sample from it.

The same applies for other well-known distributions, e.g. `rnorm` for the Normal distribution, `rexp` for the Exponential, `runif` for the Uniform etc.

In other words, someone, somewhere has written algorithms that allow us to draw independent samples from these distributions.

In the rest of this chapter we will introduce some general algorithms which, in principle, allow us to draw samples from any distribution.

8.3 Markov Chain Monte Carlo

We now start looking at a (class of) method(s) known as Markov Chain Monte Carlo, or more commonly as MCMC.

MCMC methods provide a way of simulating random variables from a distribution (which we often call the *target* distribution), whose probability density function need only be known up to a normalising constant.

The first MCMC algorithm was due to Metropolis (1953).

Although MCMC was unused by the statistics community the early 90s, since then, the methods have been used extensively and revolutionised Bayesian statistics.

Note that the target distribution can be can be (and in most real-life application is) high-dimensional.

Idea: The idea is to construct a Markov chain which has stationary distribution the target distribution. Then, we simulate this chain, and if run long enough, the chain forgets its starting point and the values in the chain are (approximately) from the target distribution.

The Markov chains in MCMC are discrete-time, but the state space can be discrete, continuous or both.

Fortunately, it is lot easier to implement than it sounds above!

8.3.1 Basic facts about Markov Chains

We first review some key facts about Markov chains which will be helpful/important when we talk about MCMC.

Definition: A sequence $X_1, X_2, \ldots$ of random variables taking values in a **state-space** S is a Markov chain if

$$\mathbb{P}(X_{t+1} \in A | X_1 = x_1, \ldots, X_t = x_t) = \mathbb{P}(X_{t+1} \in A | X_t = x_t)$$

for all $A \subseteq S$ for which $\mathbb{P}(A)$ is defined.

Definition: The transition kernel $Q(\cdot, \cdot)$ defines the dynamics of the chain:

$$Q(x, y) = f_{X_{t+1}|X_t}(y|x),$$

i.e. $Q(x, y)$ is the conditional pdf of a point y if the chain is currently at x (the likelihood of moving to y given that the current state is x). We assume time-homogeneity, so Q does not depend on t.

Q is the equivalent of a transition probability matrix for discrete-state-space Markov chains.

Defintion: A distribution π is a *stationary distribution* of a chain with transition kernel Q if

$$\pi = \pi Q$$

$$\text{which means } \pi(y) = \int \pi(x)Q(x,y)\mathrm{d}x \text{ for all } y.$$

For a discrete state space this is $\pi_i = \sum_j \pi_j q_{ji}$, where

$$q_{ji} = \mathbb{P}[X_{t+1} = i | X_t = j].$$

There are many other properties that Markov Chain may have, such as being aperiodic, irreducible etc. However, for the purposes of this chapter the key thing to keep in mind is the transition kernel of the Markov Chain. The latter determines how the chain evolves in time.

There is quite a lot of theory behind the validity and the properties of Markov Chain Monte Carlo and this is out of the scope of this chapter/module.

The Level 4 module "Computational Statistics" is concerned with the computational methods in statistics and MCMC algorithms are explored there in more detail.

8.3.2 Metropolis-Hastings

We will now present one of the most commonly used MCMC algorithm: Metropoils-Hastings.

Suppose we wish to simulate from a (possibly multivariate) distribution with probability density function $\pi(x)$.

To put it in some context, we can think of $\pi(x)$ to be the posterior density of interest in a Bayesian inference problem. In other words, when concerned with Bayesian inference we are interested in drawing samples from the posterior distribution.

We need to find a transition kernel $Q(x, y)$ such that

- Simulation of a Markov chain $\{X_1, X_2, \ldots\}$ with transition kernel Q is straightforward.
- The stationary distribution of the Markov chain is $\pi(x)$.

The Metropolis-Hastings (M-H) algorithm tells us how to build transition kernels such that the Markov chain converges to the distribution $\pi(x)$.

Metropolis-Hastings Algorithm

1. Choose a starting location $X_0 = x_0$.

2. Suppose at time t, we have $X_t = x$.

   Propose (sample) a candidate value y from proposal distribution $q(x, y) = q(y|x)$.

3. Calculate the 'Metropolis-Hastings acceptance probability' $\alpha(x, y)$ given by

$$\alpha(x, y) = \min\left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right)$$

4. Accept the move with probability $\alpha(x, y)$. If the move is rejected, stay at x.

   That is, we set

$$X_{t+1} = \begin{cases} y & \text{with probability } \alpha(x, y) \\ x & \text{with probability } 1 - \alpha(x, y) \end{cases}$$

Recall that Step 4 can be achieved by

1. Simulating $u \sim U(0, 1)$, and

2. accepting the move if $u < \alpha(x, y)$.

One can show show that the chain produced by the M-H algorithm does have stationary distribution $\pi(x)$ as claimed. Note that the proof is omitted and not examinable.

Remarks

- The algorithm works for any arbitrary choice for $q(x, y)$, so we can be inventive with our choices.

- It is only necessary to know $\pi(x)$ up to proportionality as any unknown normalising constant is cancelled in the numerator and denominator of the acceptance probability. We demonstrate this in the following example.

Sampling from a Gamma distribution

Before we start looking at how to sample from posterior distributions, let us assume that we want to drop samples from a Gamma(α, β) distribution with probability density function:

$$\pi(x) = f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{a-1} \exp\{-\beta x\}, \quad x > 0.$$

Note that the above expression says that $\pi(x) = 0$ for any $x < 0$.

Of course, we can sample *directly* from a Gamma distribution in `R` using the `rgamma` function (see beginning of this chapter)!

However, for the purposes of illustration we will pretend that we don't have access to `rgamma`.

To implement the M-H algorithm we need a proposal density. Consider a simple proposal distribution of the form

$$q(x, y) \sim N(x, \sigma^2)$$

i.e $y \sim N(x, \sigma^2)$.

In other words if the current value of the chain is x, we simulate y from a $N(x, \sigma^2)$ distribution.

Then

$$\alpha(x, y) = \min\left(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right)$$

$$= \min\left(1, \frac{\pi(y)}{\pi(x)}\right)$$

$$= \min\left(1, \frac{y^{a-1}\exp\{-\beta y\}}{x^{a-1}\exp\{-\beta x\}}\right)$$

Notes:

1. The normalising constants $\frac{\beta^\alpha}{\Gamma(\alpha)}$ have cancelled because the appear both in the numerator and denominator.

2. We have used the fact that $q(x, y)$ is symmetric in x and y, where $q(y, x) = q(x|y)$, the $N(y, \sigma^2)$ distribution evaluated at x:

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

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

3. It is important to appreciate that q was our choice, and we have to choose the value of σ^2.

4. In principle we can choose any value of σ we like, in the sense if we let the Markov Chain run long enough, then it will reach equilibrium and we will sample from the stationary distribution.

   In practice though, if σ^2 is too small then we will be making very small jumps (ie accepting too often). On other hand if we choose a large value for σ^2 then we will be making too big jumps (i.e. rejecting too often).

In the case where $q(y|x) \sim N(x, \sigma^2)$ the algorithm is called/known as **Gaussian Random Walk** and is on the most commonly used M-H algorithms due to its simplicity.

5. By q being any distribution we like, we a lot of freedom.

   However, we need to be careful that our choice of q is a good one – i.e. that the chain mixes well, meaning it quickly forgets its starting point and converges to the stationary distribution, and explores the support of the target distribution efficiently.

See the R Markdown document on Moodle for an implementation of the above algorithm `in R` where we discuss some details (e.g the effect of σ) and illustrate how the algorithm works in practice.

8.3.3 Gibbs Sampler

The Gibbs sampler was invented by Geman and Geman in 1984 and has revolutionised Bayesian statistics.

It is sometimes used to introduce the idea of MCMC, before moving on to look at Metropolis-Hastings updates.

However, the Gibbs sampler can be viewed naturally as a special case of the Metropolis-Hastings algorithm, which uses a proposal distribution which makes the Metropolis-Hastings acceptance probability equal to one — i.e., we always accept the proposed moves.

Basic idea

The basic idea is that to sample from $\pi(\boldsymbol{x})$, where $\boldsymbol{x}$ is multidimensional (of dimension d say), we sample from $\pi(x_i|\boldsymbol{x}_{-i})$ in turn for each i, where

$$\boldsymbol{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, x_d)$$

is $\boldsymbol{x}$ with the i^{th} element removed.

These conditional distributions are known as the *full-conditional distributions*.

How to find full conditionals?

Fortunately, it is often possible to calculate the full conditionals, even when the overall distribution is intractable.

The full conditional distribution of x_i describes the behaviour of x_i conditional on values of all the other variables. I.e, it describes the local characteristics of the full distribution in the direction of x_i, given the other variables.

Consider the conditional density of x_i given all the other elements of $\boldsymbol{x}$. We have

$$\begin{aligned}
\boldsymbol{x} &= (x_1, x_2, \ldots, x_d) \\
\boldsymbol{x}_{-i} &= (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) \\
\pi(x_i|\boldsymbol{x}_{-i}) &= \frac{\pi(x_i, \boldsymbol{x}_{-i})}{\pi(\boldsymbol{x}_{-i})} = \frac{\pi(\boldsymbol{x})}{\pi(\boldsymbol{x}_{-i})} = \frac{\pi(\boldsymbol{x})}{\int \pi(\boldsymbol{x})\, \mathrm{d}x_i} \\
&\propto \pi(\boldsymbol{x}).
\end{aligned}$$

That is, the conditional distribution is proportional to the full joint distribution, $\pi(\boldsymbol{x})$, viewed as a function of x_i alone.

Often, we can recognise the conditional distribution immediately in this way.

The distributions $\pi(x_i|\boldsymbol{x}_{(i)})$, $i = 1, \ldots, d$ are called the *full conditionals* of $\pi(\boldsymbol{x})$.

One can show that the Gibbs sampler is a special type of Metropolis-Hastings algorithm.

The Gibbs Sampler works by sampling from each of the full conditionals in turn (provided they are known and can be sampled from). The full algorithm is as follows:

The Gibbs Sampler

1. Start with an initial vector $\boldsymbol{x}^{(0)}$;

2. Given $\boldsymbol{x}^{(m)} = \left(x_1^{(m)}, \ldots, x_d^{(m)}\right)^T$,

   we generate $\boldsymbol{x}^{(m+1)}$ as follows:

   - Generate $x_1^{(m+1)}$ by sampling from
     $$\pi(x_1 \mid x_2^{(m)}, x_3^{(m)}, \ldots, x_d^{(m)});$$
   - Generate $x_2^{(m+1)}$ by sampling from
     $$\pi(x_2 \mid x_1^{(m+1)}, x_3^{(m)}, x_4^{(m)}, \ldots, x_d^{(m)});$$
   - Generate $x_3^{(m+1)}$ by sampling from
     $$\pi(x_3 | x_1^{(m+1)}, x_2^{(m+1)}, x_4^{(m)}, \ldots, x_d^{(m)});$$
     $$\vdots$$
   - Generate $x_d^{(m+1)}$ by sampling from
     $$\pi(x_d \mid x_1^{(m+1)}, x_2^{(m+1)}, \ldots, x_{d-1}^{(m+1)});$$

3. We now have
   $$\boldsymbol{x}^{(m+1)} = \left(x_1^{(m+1)}, x_2^{(m+1)}, \ldots, x_d^{(m+1)}\right)^T$$
   .

4. Set $m \leftarrow m + 1$ and return to step 2.

Under general conditions the distribution of $\boldsymbol{x}^{(m)}$ tends to $\pi(\boldsymbol{x})$.

Gibbs sampler: An Example

Consider generating a sample from the joint density

$$\pi(\theta,\phi) \propto \begin{cases} \phi^3 \exp\{-(1+5\phi)\theta^2 + 40\theta\phi - 81\phi\}, & \text{if } \theta \in \mathbb{R},\ \phi \in \mathbb{R}^+, \\ 0, & \text{otherwise.} \end{cases}$$

We need to sample in turn from the two full conditionals,

$$\pi(\theta|\phi)$$
$$\pi(\phi|\theta)$$

Conditional distribution of $\theta|\phi$

As a function of θ alone we have

$$\pi(\theta|\phi) \propto \exp\{-(1+5\phi)\theta^2 + 40\theta\phi\}.$$

We observe that it looks like we can manipulate the above expression by completing the square which gives:

$$\pi(\theta|\phi) \propto \exp\left\{-(1+5\phi)\left(\theta - \frac{20\phi}{1+5\phi}\right)^2\right\}.$$

Compare this with $\exp\left\{-\dfrac{(x-\mu)^2}{2\sigma^2}\right\}$, the form of the pdf of a $N(\mu,\sigma^2)$ distribution.

$$\mu = \frac{20\phi}{1+5\phi} \qquad \sigma^2 = \frac{1}{2(1+5\phi)}$$

$$\text{i.e.} \quad \theta|\phi \sim N\left(\frac{20\phi}{1+5\phi}, \frac{1}{2(1+5\phi)}\right).$$

So *given* ϕ, the full conditional distribution of θ is a Normal with mean

$$\frac{20\phi}{1+5\phi}$$

and variance

$$\frac{1}{2(1+5\phi)}$$

.

Conditional Distribution of $\phi|\theta$

Similarly, as a function of ϕ we have

$$\pi(\phi|\theta) \propto \phi^3 \exp\{-\phi(5\theta^2 - 40\theta + 81)\}, \ \phi > 0.$$

Compare this with the pdf of a Gamma (α, β), distribution which is proportional to

$$x^{\alpha-1} \exp\{-\beta x\}.$$

Thus,

$$\alpha = 4 \quad \beta = 5\theta^2 - 40\theta + 81$$

i.e. $\phi \mid \theta \sim$ Gamma $(4, 5\theta^2 - 40\theta + 81)$.

Since the two full conditionals are $\theta \mid \phi \sim$ Normal, $\phi \mid \theta \sim$ Gamma, they can both be easily simulated from.

The algorithm proceeds as follows:

1. Initialise ϕ at $\phi^{(0)}$

2. Generate $\theta^{(1)}$ by sampling from a

$$N\left(\frac{20\phi^{(0)}}{1+5\phi^{(0)}}, \frac{1}{2(1+5\phi^{(0)})}\right);$$

3. Generate $\phi^{(1)}$ by sampling from a

$$\text{Gamma}\left(4, 5\left(\theta^{(1)}\right)^2 - 40\theta^{(1)} + 81\right);$$

4. Generate $\theta^{(2)}$ by sampling from a

$$N\left(\frac{20\phi^{(0)}}{1+5\phi^{(0)}}, \frac{1}{2(1+5\phi^{(0)})}\right);$$

5. Generate $\phi^{(2)}$ by sampling from a

$$\text{Gamma}\left(4, 5\left(\theta^{(2)}\right)^2 - 40\theta^{(2)} + 81\right);$$

6. and so on.

Gibbs sampler: Another example

Let θ_1 and θ_2 have bivariate density

$$f(\theta_1, \theta_2) \propto \exp\{-\theta_1^2 - \theta_2^2 + 0.6\theta_1\theta_2 - 0.3\theta_1^2\theta_2^2\}$$

where $-\infty < \theta_1, \theta_2 < \infty$.

Then the full conditional distribution of θ_1 is

$$g(\theta_1|\theta_2) \propto \exp\{-\theta_1^2 + 0.6\theta_1\theta_2 - 0.3\theta_1^2\theta_2^2)$$

so that

$$g(\theta_1|\theta_2) \propto \exp\{-\theta_1^2(1 + 0.3\theta_2^2) + 2\theta_1(0.3\theta_2)\}.$$

Compare this with pdf of $X \sim N(\mu, \sigma^2)$ which is proportional to

$$\exp\left\{-\frac{x^2}{2\sigma^2} + 2x\left(\frac{\mu}{2\sigma^2}\right)\right\}.$$

Hence we deduce that $\theta_1|\theta_2$ has a normal distribution with mean and variance given by

$$\frac{0.3\theta_2}{1 + 0.3\theta_2^2}$$

and

$$\frac{1}{2(1+0.3\theta_2^2)}$$

respectively, so

$$\theta_1|\theta_2 \sim N\left(\frac{0.3\theta_2}{1+0.3\theta_2^2}, \frac{1}{2(1+0.3\theta_2^2)}\right).$$

Similarly the full conditional distribution of θ_2 is

$$h(\theta_2|\theta_1) \propto \exp\{-\theta_2^2 + 0.6\theta_1\theta_2 - 0.3\theta_1^2\theta_2^2\}$$

so that

$$h(\theta_2|\theta_1) \propto \exp\{-\theta_2^2(1+0.3\theta_1^2) + 2\theta_2(0.3\theta_1)\}.$$

Hence we deduce that

$$\theta_2|\theta_1 \sim N\left(\frac{0.3\theta_1}{1+0.3\theta_1^2}, \frac{1}{2(1+0.3\theta_1^2)}\right).$$

Exercise: Implement in R a Gibbs sampler to sample from the above distribution.

8.4 MCMC for Bayesian Inference

So far we implemented MCMC algorithms to sample from some (generic) distributions which were not related with a Bayesian inference problem.

In this section we will implement MCMC algorithms to draw samples from posterior distributions.

8.4.1 Normal data; Cauchy prior for the mean

Suppose that $X_1, \ldots, X_n$ are IID from a Normal distribution with mean θ and variance 1.

We saw in Chapter 7 that if we assume a Normal distribution as a (conjugate) prior for θ, then we end up with a Normal distribution for the posterior, from which he can sample directly (e.g. using `rnorm`).

Prior Suppose that we assign a Cauchy distribution as a prior distribution for θ, i.e.

$$p(\theta) = \frac{1}{\pi(1+\theta^2)}$$

Likelihood The likelihood function is given by

$$p_{\boldsymbol{X}}(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x_i - \theta)^2\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right\}$$

$$\propto \exp\left\{-\frac{n}{2}(\theta - \bar{x})^2\right\}$$

Posterior density The posterior density (up to proportionality) is also given as follows:

$$\pi(\theta|\boldsymbol{x}) \propto \exp\left\{-\frac{n}{2}(\theta - \bar{x})^2\right\} \times \frac{1}{1+\theta^2}$$

The above expression does not look like the pdf of any well-known distribution (e.g. Normal).

Hence, we will use the Metrpolis-Hastings algorithm to draw samples from the posterior distribution.

1. Choose an initial value for θ, say θ_0.

2. Suppose at time t, we have the current value of the chain $\theta_t = \theta_{\text{cur}}$.

Propose (sample) a candidate value θ_{can} from a Normal distribution with mean θ_{cur} and variance σ^2, that is

$$\theta_{\text{can}} \sim N(\theta_{\text{cur}}, \sigma^2)$$

3. Accept θ_{can} with probability acceptance probability' $\alpha(\theta_{\text{cur}}, \theta_{\text{can}})$ given by

$$\alpha(\theta_{\text{cur}}, \theta_{\text{can}}) \;=\; \min\left(1, \frac{\pi(\theta_{\text{can}}|\boldsymbol{x})}{\pi(\theta_{\text{cur}}|\boldsymbol{x})}\right)$$

If the move is rejected, stay at θ_{cur}.

That is, we set

$$= \theta_{t+1} = \begin{cases} \theta_{\text{can}} & \text{with probability } \alpha(\theta_{\text{cur}}, \theta_{\text{can}}) \\ \theta_{\text{cur}} & \text{with probability } 1 - \alpha(\theta_{\text{cur}}, \theta_{\text{can}}) \end{cases}$$

where $\pi(\theta|\boldsymbol{x}) \propto \exp\left\{-\frac{n}{2}(\theta - \bar{x})^2\right\} \times \frac{1}{1+\theta^2}$.

Have a look at the accompanied `R` markdown document where the above algorithm is implemented.

8.4.2 Gibbs Sampler for the Normal Model

In Chapter 7 we had to employ several algebraic tricks to derive analytically the marginal posterior distribution of the parameters μ and σ^2.

We now demonstrated how to sample from the joint posterior distribution of μ and σ^2 using a Gibbs sampler.

We saw in Section 7.6.

$$\mu \,|\, \sigma^2, \boldsymbol{x} \;\sim\; \mathrm{N}(m_1, v_1\sigma^2)$$

$$\sigma^2 \,|\, \mu, \boldsymbol{x} \;\sim\; \mathrm{IG}\left(a_1 + \frac{(\mu - m_1)^2}{v_1}, d_1 + 1\right).$$

where $\boldsymbol{x}$ is the observed data assuming that they come from a $N(\mu, \sigma^2)$.

Given that both full conditional distributions are well known distribution, i.e. we know how to sample them directly, It is then straightforward to implement a Gibbs sampler to sample from the joint posterior distribution.

1. Initialise σ^2 at $\sigma^{2^{(0)}}$

2. Generate $\mu^{(1)}$ by sampling from a

$$\mathrm{N}(m_1, v_1\sigma^{2^{(0)}})$$

3. Generate $\sigma^{2^{(1)}}$ by sampling from a

$$\mathrm{IG}\left(a_1 + \frac{(\mu^{(1)} - m_1)^2}{v_1}, d_1 + 1\right)$$

4. Generate $\mu^{(2)}$ by sampling from a

$$\mathrm{N}(m_1, v_1\sigma^{2^{(1)}})$$

5. Generate $\sigma^{2^{(2)}}$ by sampling from a

$$\text{IG}\left(a_1 + \frac{(\mu^{(2)} - m_1)^2}{v_1}, d_1 + 1\right)$$

6. and so on.

Recall that (from Chapter 7).

$$d_1 = d + n,$$
$$v_1 = \left(\frac{1}{v} + n\right)^{-1}$$
$$a_1 = \frac{m^2}{v} + n\bar{x}^2 + a + nS^2 - \left(\frac{1}{v} + n\right) m_1^2$$
$$m_1 = \frac{\frac{1}{v}m + n\bar{x}}{\frac{1}{v} + n.}$$

Have a look at the accompanied `R` markdown document where the above algorithm is implemented.

8.5 Remarks

MCMC methods made Bayesian inference possible in complex models and are widely used.

As mentioned in the beginning of this chapter, the motivation was to present a way of performing Bayesian inference when one cannot derive analytically the density of the posterior distribution .

We presented two widely used algorithms: the M-H and the Gibbs sampler. There are tons of variations of the two samplers that we discussed in this chapter and we have only scratched the surface.

Please note that there is a lot of theory behind MCMC (why it works etc) but that was not within the scope of this module and has been omitted.

it is important to be aware of, as well as there being several implementation issues, some of which are discussed in the `R` Markdown documents.

Finally, please remember that MCMC is not a black-box and it is important to be aware of several implementation issues, some of which are discussed in the `R` Markdown documents as well as those below:

Final remarks

1. We no longer have independent draws from the distribution of interest, as successive observations are correlated.

2. Before reporting summaries/histograms from the sampled distribution, we need to ensure that the chain has converged (i.e. we really are sampling from the stationary distribution of the Markov

chain).

For the purposes of this module this has been done visually (ie by looking at the trace plots). However, there are some tests that one can perform to check for convergence.

3. Poor proposal densities lead to poor mixing (exploration of the sample space).

— THE END —