Statistical Inference

Dr. Francisco Javier Rubio
email: f.j.rubio@ucl.ac.uk

August 27, 2021

Contents

These notes were developed by Dr. Francisco Javier Rubio. I would appreciate if you point out any typos you spot out to me: (f.j.rubio@ucl.ac.uk).

Disclaimer: These notes closely follow the material in the textbooks cited in the bibliography, with some additions by the author. These notes should *not* be distributed or used for commercial purposes.

Notation

- Probability density function (PDF or pdf).
- Probability mass function (PMF or pmf).
- Characteristic function (CF or cf).
- ∇f: gradient of f.
- ∇^2: Hessian matrix of f.
- $A^\top$: Transpose of the matrix A.

Chapter 1

Preliminary Material

This chapter contains preliminary material that will serve as a basis for the following chapters. For a deeper study of these topics, see the references and the lecture notes of the prerequisite modules.

The most basic definition is that of a random variable.

Definition 1. (Measurable function). Let f be a function from a measurable space $(\Omega, \mathcal{F})$ into the real numbers. We say that the function is measurable if for each Borel set $B \in \mathcal{B}$, the set $\{\omega \in \Omega : f(\omega) \in \mathcal{B}\} \in \mathcal{F}$.

Definition 2. (random variable). A random variable X is a measurable function from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into the real numbers $\mathbb{R}$ (or a subset).

1.1 Families of Distributions

The distributions presented here are *parametric distributions*. A parametric distribution is a distribution that has one or more *parameters* (also known as *statistical parameters*). Finally, a parameter (or statistical parameter) is a numerical characteristic that indexes a family of probability distributions.

Discrete distributions

A random variable X is said to be discrete if the range of X is countable. Some examples of discrete variables and their corresponding *probability mass functions* are presented below.

Example 1: The Bernoulli distribution. The simplest example of a discrete random variable corresponds to the case where the range of X is the set $\{0, 1\}$.

The distribution of X is:

$$P(X = x) = \begin{cases} p, & \text{for } x = 1 \\ 1 - p, & \text{for } x = 0. \end{cases}$$

This distribution is known as the **Bernoulli** distribution, and it is often denoted as $X \sim \text{Bernoulli}(p)$. Often, the event $\{x = 1\}$ is called a "success", and the event $\{x = 0\}$ is called a "failure". Thus, the parameter p is known as "the probability of success". It follows that:

$$\begin{aligned} E[X] &= 1 \cdot p + 0 \cdot (1 - p) = p, \\ Var[X] &= (1 - p)^2 \cdot p + (0 - p)^2(1 - p) = p(1 - p). \end{aligned}$$

A more particular example is the case where X is the outcome observed from tossing a fair coin once. In this case $P(X = \text{ heads }) = P(X = \text{ tails }) = \frac{1}{2}$.

Bernoulli random variables are used in many contexts, and they are often referred to as *Bernoulli trials*. A Bernoulli trial is an experiment with two, and only two, possible outcomes.

Parameters: $0 < p < 1$.

Example 2: The Binomial distribution. A Binomial random variable X is the total number of successes in n Bernoulli trials. Consequently, the range of X is the set $\{0, 1, 2, \ldots, n\}$. The probability of each outcome is given by:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x},$$

where $\binom{n}{x} = \dfrac{n!}{x!(n-x)!}$ is the Binomial coefficient (also known as *combination*). If a random variable X has Binomial distribution, it is denoted as $X \sim \text{Binomial}(n, p)$, where n is the number of trials and p is the probability of success.

The mean and variance of a Binomial random variable are $E[X] = np$ and $Var[X] = np(1-p)$.

A more particular example is the case where X is the number of heads in $n = 10$ fair coin tosses. Then, for $x = 0, 1, \ldots, 10$:

$$P(X = x) = \binom{10}{x} 0.5^x (0.5)^{n-x} = \binom{10}{x} (0.5)^n.$$

Parameters: $n \in \mathbb{Z}_+$ and $0 < p < 1$.

Example 3: The Poisson distribution. A random variable X has a Poisson distribution if it takes values in the non-negative integers and its distribution is given by:

$$P(X = x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, \ldots$$

It is possible to show that $E[X] = Var[X] = \lambda$.

Parameters: $\lambda > 0$.

Example 4: The Multinomial distribution. The Multinomial distribution is a generalisation of the binomial distribution. For n independent trials, each of which produces and outcome (success) in one of $k \geq 2$ categories, where each category has a given fixed success probability θ_i, $i = 1, \ldots, k$, the Multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories. Thus, the pmf is

$$p(x_1, \ldots, x_k; \theta_1, \ldots, \theta_k) = \frac{n!}{x_1! \cdots x_k!} \theta_1^{x_1} \cdots \theta_k^{x_k},$$

where $\theta_i > 0$ for $i = 1, \ldots, k$ and $\sum_{i=1}^k \theta_i = 1$.

Parameters: $\theta_i \geq 0$ for $i = 1, \ldots, k$ and $\sum_{i=1}^k \theta_i = 1$.

There are many discrete distributions of practical interest. To name but a few: the hypergeometric distribution, the discrete uniform distribution, among others (see [1, 6] for more examples).

Continuous distributions

A random variable X is said to be continuous if its range is uncountable and its distribution is continuous everywhere. Moreover, a random variable X is said to be *absolutely continuous* if there exists a non-negative function f such that for any open set B [16]:

$$P(X \in B) = \int_B f(x) dx$$

The function f is called the *probability density function* of X. This definition can be used to link the probability density function and the cumulative distribution F as follows:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

Definition 3. Suppose that $f : \mathcal{D} \to \mathbb{R}_+$ is a density function. Then, the *support* of f, $\text{supp}(f)$, is the set of points where f is positive:

$$\text{supp}(f) = \{x \in \mathcal{D} : f(x) > 0\}.$$

Definition 4. A random variable X, with cdf F has the characteristic function

$$\varphi_X(t) = E\left[e^{itX}\right] = \int_{-\infty}^{\infty} e^{itx}\, dF(x).$$

If the pdf f exists, then

$$\varphi_X(t) = E\left[e^{itX}\right] = \int_{-\infty}^{\infty} e^{itx} f(x)\, dx.$$

Another feature of continuous distributions is that $P(X = x) = 0$, for all x in the range of X. Some examples of continuous distributions are presented below.

Example 1: The Beta Distribution. The probability density function of a Beta random variable $X \in (0, 1)$ is:

$$f(x; a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)},$$

where $a, b > 0$ are shape parameters, $B(a, b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the Beta special function, and $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ is the Gamma special function. It is important to distinguish the Beta distribution from the Beta special function.

The mean of the Beta distribution is $E[X] = \dfrac{a}{a+b}$, and the mode (maximum) is $Mode(X) = \dfrac{a-1}{a+b-2}$ for $b > 1$. The variance is $Var[X] = \dfrac{ab}{(a+b)^2(a+b+1)}$.

Parameters: $a > 0$ and $b > 0$.

The uniform distribution is a special case of the Beta distribution for the case $a = b = 1$.

Example 2: Normal or Gaussian Distribution. The probability density function of a Gaussian random variable $X \in \mathbb{R}$ is:

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

where $-\infty < \mu < \infty$ and $\sigma > 0$ are parameters of this density function. In fact, $E[X] = \mu$ and $Var[X] = \sigma^2$. If a random variable X has normal distribution with mean μ and variance σ^2, we will denote it $X \sim N(\mu, \sigma^2)$. This is one of the most popular distributions in applications and it appears in a number of statistical and probability models.

Parameters: $-\infty < \mu < \infty$ and $\sigma > 0$.

The Normal distribution has many interesting properties. One of them is that it is closed under summation, meaning that the sum of normal random variables is normally distributed. This is, let $X_1, \ldots, X_n$ be *i.i.d.* random variables with distribution $N(\mu, \sigma^2)$, and $Y = \sum_{j=1}^n X_j$, $Z = \dfrac{1}{n}\sum_{j=1}^n X_j$. Then, $Y \sim N\left(n\mu, n\sigma^2\right)$, $Z \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$.

Example 3: The Logistic distribution. The probability density function of a logistic random variable $X \in \mathbb{R}$ is:

$$f(x;\mu,\sigma) = \frac{\exp\left\{-\dfrac{x-\mu}{\sigma}\right\}}{\sigma\left(1+\exp\left\{-\dfrac{x-\mu}{\sigma}\right\}\right)^2},$$

where $-\infty < \mu < \infty$ and $\sigma > 0$ are *location* and *scale* parameters, respectively. The mean and variance of X are given by $E[X] = \mu$ and $Var[X] = \dfrac{\pi^2\sigma^2}{3}$. The cdf of a logistic random variable is

$$F(x;\mu,\sigma) = \frac{\exp\left\{\dfrac{x-\mu}{\sigma}\right\}}{1+\exp\left\{\dfrac{x-\mu}{\sigma}\right\}}.$$

This distribution is very popular in practice as well. In particular, the so called "logistic regression model" is based on this distribution, as well as some *Machine Learning* algorithms (what is that?).

Parameters: $-\infty < \mu < \infty$ and $\sigma > 0$.

Example 4: The Exponential distribution. The probability density function of an Exponential random variable $X > 0$ is:

$$f(x;\lambda) = \lambda \exp\left\{-\lambda x\right\},$$

where $\lambda > 0$ is a rate parameter. The mean and variance are given by $E[X] = \frac{1}{\lambda}$ and $Var[X] = \dfrac{1}{\lambda^2}$.

This distribution is widely used in engineering and the analysis of survival times.

Parameters: $\lambda > 0$.

Example 5: The Gamma distribution. The probability density function of an Gamma random variable $X > 0$ is:

$$f(x;\kappa,\theta) = \frac{1}{\Gamma(\kappa)\theta^\kappa} x^{\kappa-1} \exp\left\{-\frac{x}{\theta}\right\},$$

where $\kappa > 0$ is a shape parameter, $\theta > 0$ is a scale parameter, and $\Gamma(z) = \int_0^\infty s^{z-1}e^{-s}ds$ is the Gamma function. The mean and variance of X are given by $E[X] = \kappa\theta$ and $Var[X] = \kappa\theta^2$.

This distribution is widely used in engineering and the analysis of survival times.

Parameters: $\kappa > 0$ and $\theta > 0$.

Example 6: Student's t-distribution. Student's t-distribution has pdf

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sigma\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}},$$

where $x \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma > 0$, and $\nu > 0$. The mean of the Student's t-distribution is μ for $\nu > 0$, and undefined for $\nu \leq 1$. The variance of this distribution is $\frac{\nu}{\nu-2}$ for $\nu > 2$, and undefined or infinite for $\nu \leq 2$. Moreover, the Student's t-distribution converges to the normal distribution (pointwise) as $\nu \to \infty$.

Parameters: $\mu \in \mathbb{R}$, $\sigma > 0$, and $\nu > 0$.

Example 7: The Chi-square distribution. The probability density function of the chi-square (χ^2_k) distribution with k degrees of freedom is

$$f(x;k) = \frac{x^{\frac{k}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}, \quad x > 0.$$

The mean of the chi-square distribution is k and the variance is $2k$.

Parameters: $k > 0$.

Relationship with other distributions:

- Let $X_1, \ldots, X_k$ be *i.i.d.* random variables with distribution $N(0,1)$. Let $Y = \sum_{j=1}^{k} X_j^2$, then $Y \sim \chi^2_k$.
- Let $X \sim N(0,1)$ and $W \sim \chi^2_k$. Then, $Y = \dfrac{X}{\sqrt{\frac{W}{k}}}$ has Student's t-distribution ($\mu = 0$, $\sigma = 1$) with k degrees of freedom.

There exist many continuous distributions of practical interest. To name but a few: the chi-square distribution, the generalised extreme value distribution, the Student-t distribution, the F distribution, among many others (see [6] for more examples).

Definition 5. The qth quantile of the distribution of a random variable X, is that value x such that $P(X < x) = q$. If $q = 0.5$, the value is called the median. The cases $q = 0.25$ and $q = 0.75$ correspond to the lower quartile and upper quartile, respectively.

Definition 6. The kernel of a probability density function (pdf) or probability mass function (pmf) is the factor of the pdf or pmf in which any factors that are not functions of any of the variables in the domain are omitted. For example, the kernel of the Beta distribution is:

$$K(x;a,b) = x^{a-1}(1-x)^{b-1}. \tag{1.1.1}$$

The kernel of the Gaussian (Normal) distribution is:

$$K(x;\mu,\sigma) = \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

Types of parameters

The parameters of a distribution are classified into three types: location parameters, scale parameters, and shape parameters.

- A parameter μ of a distribution function $F(x;\mu)$ is called a *location parameter* if $F(x;\mu) = F(x-\mu;0)$. An equivalent definition can be made on the probability density function. This is, a parameter μ of a density function $f(x;\mu)$ is called a *location parameter* if $f(x;\mu) = f(x-\mu;0)$.

  An example of a location parameter is the parameter μ in the Gaussian distribution.

- A parameter $\sigma > 0$ of a distribution function $F(x;\sigma)$ is called a *scale parameter* if $F(x;\sigma) = F\left(\frac{x}{\sigma};1\right)$. An equivalent definition can be made on the probability density function. This is, a parameter σ of a density function $f(x;\sigma)$ is called a *scale parameter* if $f(x;\sigma) = \dfrac{1}{\sigma} f\left(\frac{x}{\sigma};1\right)$.

  An example of a scale parameter is the parameter σ in the Gaussian distribution.

- A shape parameter is a parameter that is neither a location nor a scale parameter. This kind of parameters control the shape of a distribution (equivalently, density) function.

  An example of a shape parameter is the parameter κ in the Gamma distribution.

A distribution is said to belong to the **Location-Scale family** of distributions if it is parameterised in terms of a location and a scale parameter. Examples of members of the location-scale family are the Gaussian and Logistic distributions. The Gamma distribution *is not* a member of this family (why?).

A distribution $F(x;\theta)$ is said to be **Identifiable** if $F(x;\theta_1) = F(x;\theta_2)$, for all x, implies that $\theta_1 = \theta_2$ for all possible values of θ_1, θ_2.

1.2 Multivariate Normal Distribution

The multivariate normal distribution of a p-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_p)$ is said to be distributed as a multivariate Normal if and only if its probability density function is:

$$\phi_{\mathbf{X}}(x_1, \ldots, x_p; \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})\right),$$

where $\boldsymbol{\mu} = E[\mathbf{X}]$ is the location parameter and $\Sigma = E\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top\right]$ is the covariance matrix. We denote it as $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$.

1.3 Conditional Probability

For more details on this topic see [7].

Definition 7. A probability space is the triplet $(\Omega, \mathcal{F}, P)$, where Ω is the sample space, $\mathcal{F}$ is the collection of events (σ-algebra), and P is a probability measure defined over $\mathcal{F}$, with $P(\Omega) = 1$.

Definition 8. Two events A and B are independent if and only if the probability of their intersection equals the product of their individual probabilities, that is

$$P(A \cap B) = P(A)P(B).$$

Definition 9. Given two events A and B, with $P(B) > 0$, the conditional probability of A given B, denoted $P(A \mid B)$, is defined by the relation

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A, B)}{P(B)}.$$

In connection with these definitions, the following result holds. Let $\{C_j : j = 1, \ldots, n\}$ be a partition of Ω, this is, $\Omega = \cup_{j=1}^n C_j$ and $C_i \cap C_k = \emptyset$ for $i \neq k$. Let also A be an event. The *Law of Total Probability* states that

$$P(A) = \sum_{j=1}^{n} P(A \mid C_j) P(C_j).$$

Definition 10. Let X and Y be discrete, jointly distributed random variables. For $P(X = x) > 0$ the conditional probability function of Y given that $X = x$ is

$$p_{Y|X=x}(y) = P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)},$$

and the conditional cumulative distribution function of Y given that $X = x$ is

$$F_{Y|X=x}(y) = P(Y \leq y \mid X = x) = \sum_{z \leq y} p_{Y|X=x}(z).$$

Definition 11. Let X and Y have a joint continuous distribution. For $f_X(x) > 0$, the conditional density function of Y given that $X = x$ is

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)},$$

where $f_{X,Y}$ is the joint probability density function of X and Y, and f_X is the marginal probability density function of X. The conditional cumulative distribution function of Y given that $X = x$ is

$$F_{Y|X=x}(y) = \int_{-\infty}^{y} f_{Y|X=x}(z)dz.$$

Remark 1.3.1. *The law of total probability. Let X and Y have a joint continuous distribution. Suppose that $f_X(x) > 0$, and let $f_{Y|X=x}(y)$ be the conditional density function of Y given that $X = x$. The law of total probability states that*

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X=x}(y) f_X(x)dx.$$

1.4 Modes of Convergence

The following four modes of convergence will be used throughout the lectures. For more details on this topic see [7].

Definition 12. Let $X_1, X_2, \ldots$ be a sequence of independent random variables. X_n converges *almost surely* (a.s.) to the random variable X, as $n \to \infty$, if and only if

$$P\left(\{\omega \in \Omega : X_n(\omega) \to X(\omega) \quad as \quad n \to \infty\}\right) = 1.$$

Notation: $X_n \overset{a.s.}{\to} X$ as $n \to \infty$. Almost sure convergence is often referred to as *strong convergence*.

Definition 13. Let $X_1, X_2, \ldots$ be a sequence of independent random variables. X_n converges *in probability* to the random variable X, as $n \to \infty$, if and only if, for all $\varepsilon > 0$:

$$P\left(|X_n - X| > \varepsilon\right) \to 0 \quad as \quad n \to \infty.$$

Notation: $X_n \overset{P}{\to} X$ as $n \to \infty$.

Recall now that the expectation (or the mean) of a continuous random variable X with probability density function f is defined as:

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx,$$

the $n-$th moment of the random variable X is defined as:

$$E[X^n] = \int_{-\infty}^{\infty} x^n f(x)dx,$$

and the $n-$th absolute moment of the random variable X is defined as:

$$E|X|^n = \int_{-\infty}^{\infty} |x|^n f(x)dx.$$

Definition 14. Let $X_1, X_2, \ldots$ be a sequence of independent random variables. X_n converges in $r-$mean to the random variable X, as $n \to \infty$, if and only if, for all $\varepsilon > 0$:

$$E\left|X_n - X\right|^r \to 0 \quad as \quad n \to \infty.$$

Notation: $X_n \overset{r}{\to} X$ as $n \to \infty$.

Definition 15. Let $X_1, X_2, \ldots$ be a sequence of independent random variables. X_n converges in *distribution* to the random variable X, as $n \to \infty$, if and only if:

$$F_{X_n}(x) \to F_X(x) \quad as \quad n \to \infty \quad for\ all \quad x \in C(F_X),$$

where F_{X_n} and F_X are the cumulative distribution functions of X_n and X, respectively, and $C(F_X)$ is the continuity set of F_X (that is, the points where F_X is continuous). Notation: $X_n \xrightarrow{d} X$ as $n \to \infty$.

Theorem 1.4.1. *Slutsky's theorem. Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be sequences of random variables. Suppose that*

$$X_n \xrightarrow{d} X, \quad and \quad Y_n \xrightarrow{P} a, \quad as \quad n \to \infty,$$

where a is some constant. Then, as $n \to \infty$

$$\begin{aligned} X_n + Y_n &\xrightarrow{d} X + a, \\ X_n - Y_n &\xrightarrow{d} X - a, \\ X_n \cdot Y_n &\xrightarrow{d} X \cdot a, \\ \frac{X_n}{Y_n} &\xrightarrow{d} \frac{X}{a}, \quad for\ a \neq 0. \end{aligned}$$

Theorem 1.4.2. *Convergence of sums of sequences of random variables.*

(i) Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be sequences of random variables. Suppose that

$$X_n \xrightarrow{a.s.} X, \quad and \quad Y_n \xrightarrow{a.s.} Y, \quad as \quad n \to \infty.$$

Then,

$$X_n + Y_n \xrightarrow{a.s.} X + Y, \quad as \quad n \to \infty.$$

(ii) Let $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ be sequences of random variables. Suppose that

$$X_n \xrightarrow{P} X, \quad and \quad Y_n \xrightarrow{P} Y, \quad as \quad n \to \infty.$$

Then,

$$X_n + Y_n \xrightarrow{P} X + Y, \quad as \quad n \to \infty.$$

1.5 The Law of Large Numbers and the Central Limit Theorem

For more details on this topic see [7].

Definition 16. We say that two random variables X and Y are identically distributed if and only if $P(X \leq x) = P(Y \leq x)$, for all x.

If two variables are independent and identically distributed, we say that they are "*i.i.d.*".

Theorem 1.5.1. *The weak law of large numbers. Let $X_1, X_2, \ldots$ be a sequence of* i.i.d. *random variables with finite mean μ, and set $S_n = X_1 + X_2 + \cdots + X_n$, $n \geq 1$. Then*

$$\bar{X}_n = \frac{S_n}{n} \xrightarrow{P} \mu \quad as \quad n \to \infty.$$

Corollary 1.5.1. *Let h be a measurable function and $X_1, \ldots, X_n$ be a sequence of* i.i.d. *random variables with distribution F. Suppose that $E[h(X)] < \infty$, for $X \sim F$. Then, by the law of large numbers*

$$\frac{1}{n} \sum_{i=1}^{n} h(X_i) \xrightarrow{P} E[h(X)] \quad as \quad n \to \infty.$$

Theorem 1.5.2. *The strong law of large numbers. Let $X_1, X_2, \ldots$ be a sequence of* i.i.d. *random variables with finite mean μ and finite variance, and set $S_n = X_1 + X_2 + \cdots + X_n$, $n \geq 1$. Then*

$$\bar{X}_n = \frac{S_n}{n} \xrightarrow{a.s.} \mu \quad as \quad n \to \infty.$$

Theorem 1.5.3. *The central limit theorem (univariate case). Let $X_1, X_2, \ldots$ be a sequence of* i.i.d. *random variables with finite mean μ and finite variance σ^2, and set $S_n = X_1 + X_2 + \cdots + X_n$, $n \geq 1$. Then*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1) \quad as \quad n \to \infty.$$

Theorem 1.5.4. *The central limit theorem (multivariate case). Let $\mathbf{X}_1, \mathbf{X}_2, \ldots$ be a sequence of* i.i.d. *p-dimensional random vectors, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$. Suppose that*

$$E||\mathbf{X}_1||^2 = E\left(X_{11}^2 + \cdots + X_{1p}^2\right) < \infty,$$

the central limit theorem asserts that

$$\sqrt{n}\left(\mathbf{S}_n - E[\mathbf{X}_1]\right) \xrightarrow{d} N(0, Cov[\mathbf{X}_1]) \quad as \quad n \to \infty,$$

where $Cov[\mathbf{X}_1]$ is the $p \times p$ covariance matrix of the random vector $\mathbf{X}_1$.

Theorem 1.5.5. *(Continuous mapping theorem). Let $X_1, X_2, \ldots$ be a sequence of random variables on $\mathbb{R}$. Suppose that $g : \mathbb{R} \to \mathbb{R}$ is a continuous function (almost surely). Then,*

$$\begin{array}{lll} X_n \xrightarrow{d} X & \textit{implies} & g(X_n) \xrightarrow{d} g(X), \\ X_n \xrightarrow{P} X & \textit{implies} & g(X_n) \xrightarrow{P} g(X), \\ X_n \xrightarrow{a.s.} X & \textit{implies} & g(X_n) \xrightarrow{a.s.} g(X). \end{array}$$

1.6 Additional tools

- Markov's inequality

  $$P[|X| \geq \alpha] \leq \frac{1}{\alpha^k} E[|X|^k],$$

  where $\alpha > 0$.

- Chebyshev inequality. Let $m = E[X]$ and $\alpha > 0$

  $$P[|X - m| \geq \alpha] \leq \frac{1}{\alpha^2} Var[X].$$

- Jensen's inequality. Let φ be a convex function on an interval containing the range of X. Then,

  $$\varphi(E[X]) \leq E[\varphi(X)].$$

  The opposite is true for concave distributions.

- Hölder's inequality. Let $p > 1$, $q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$

  $$E[|XY|] \leq E[|X|^p]^{\frac{1}{p}} \cdot E[|Y|^q]^{\frac{1}{q}}.$$

  The case $p = q = 2$ is known as the Cauchy-Schwarz inequality.

- The rank of a matrix $\mathbf{M}$ is the dimension of the vector space generated by its columns. This corresponds to the maximal number of linearly independent columns of $\mathbf{M}$. The rank is commonly denoted rank$(\mathbf{M})$ or Rank$(\mathbf{M})$.

- The square root of a non-singular matrix $\mathbf{M}$, denoted $\mathbf{M}^{\frac{1}{2}}$, is the matrix that satisfies $\mathbf{M} = \mathbf{M}^{\frac{1}{2}}\mathbf{M}^{\frac{1}{2}}$.

- Reparameterisations.

  Definition 17. Let $f(\mathbf{x};\boldsymbol{\theta})$ be a pdf with parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top \in \Theta \subset \mathbb{R}^p$, $\mathbf{x} \in \mathcal{D} \subset \mathbb{R}^n$. A reparameterisation $\boldsymbol{\eta} = \varphi(\boldsymbol{\theta})$ is a change of variables $\theta_j \mapsto \eta_j$, $j = 1, \ldots, p$, via a one-to-one function φ such that, for each $\boldsymbol{\theta} \in \Theta$, there exists $\boldsymbol{\eta} \in \varphi(\Theta)$ such that $f(x;\boldsymbol{\theta}) = f(x;\varphi^{-1}(\boldsymbol{\eta}))$. Analogously, for each $\boldsymbol{\eta} \in \varphi(\Theta)$, there exists $\boldsymbol{\theta} \in \Theta$ such that $f(x;\boldsymbol{\eta}) = f(x;\varphi(\boldsymbol{\theta}))$.

  The use of reparameterisations is very common in statistics. For instance, the Exponential distribution is often parameterised in terms of the rate parameter λ or in terms of the mean $\beta = \dfrac{1}{\lambda}$. Another example is the Normal distribution, which is often parameterised in terms of the mean μ and the standard deviation σ; or in terms of the mean μ and the variance $\sigma_2 = \sigma^2$; or in terms of the mean μ and the precision $\tau = \dfrac{1}{\sigma^2}$. They are all equivalent as there exist a one-to-one function between the different parameterisations.

- The Beta function, is a special function defined by

$$\mathrm{B}(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt.$$

- The indicator function of the set A is a function

$$\mathbf{I}_A : \mathbb{X} \to \{0,1\},$$

  defined as:

$$\mathbf{I}_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

Chapter 2

Point Estimation Theory

2.1 Introduction

According to the Oxford Dictionary of Statistics, *Statistical Inference* is defined as:

"The process of drawing conclusions about the nature of some system on the basis of data subject to random variation. There are several distinguishable and apparently irreconcilable approaches to the process of inference; comfortingly, there are rarely any gross differences in the inferences that result. Approaches include Bayesian inference and fiducial inference; the approach first met by a student of Statistics is usually that based on the Neyman-Pearson lemma."

The word "inference" refers to drawing conclusions on the basis of some evidence. Thus, *Statistical Inference* refers to drawing conclusions on the basis of evidence obtained from the data. The main challenges to do so are:

(i) How to summarise the information in the data using formal mathematical tools?

(ii) How to use these summaries to answer questions about the phenomenon of interest?

The definition from the Oxford Dictionary of Statistics already points out that there is no unique way for answering these questions. In fact, there exist several schools of thought that perform statistical inference using different tools and starting from different philosophical views. These philosophical differences, as well as the different mathematical tools employed in these approaches, will be discussed in detail in this module. The two main schools of thought are the "Frequentist approach" and the "Bayesian approach". An appealing feature of this module is that both approaches are presented in parallel, giving the student a balanced perspective.

In many areas, researchers collect data as a means to obtain information about a phenomenon of interest or to collect information about a population. For example,

- The Office for National Statistics collects information about UK residents, such as the age of those persons.
- Cancer Registries in the UK monitor the survival times of cancer patients diagnosed in the UK in specific years (cohorts).
- Pharmaceutical companies conduct experiments (trials) to assess the effectiveness of new drugs on a group of people.
- The National Aeronautics and Space Administration (NASA) has a Data Portal (publicly available) with a number of data sets produced in their experiments and monitoring.
- Many companies monitor their financial performance by looking at the daily price of the saleable stocks of the company ("share price").

- Many others ...

We can understand the collected data as a sample of observations $x_1, \ldots, x_n$, where each x_j, $j = 1, \ldots, n$ can be either an scalar number or a vector. In statistical inference, this sample is interpreted as a realisation of the random variables (vectors) $X_1, \ldots, X_n$. We will use bold capital letters to denote the vector of random variables $\mathbf{X} = (X_1, \ldots, X_n)^\top$, and bold lowercase letters to denote the sample of observations $\mathbf{x} = (x_1, \ldots, x_n)^\top$.

2.2 Data Reduction

In order to analyse, understand, and communicate the information contained in the sample $\mathbf{x}$, we need tools to summarise it. The process of summarising the data is known as data reduction or data summary. There exist many quantities that are used as summaries. These quantities are functions of the sample, and they are called "statistics". In mathematical terms, a statistic is any function of the sample $T(\mathbf{x})$, with $T : \mathbb{R}^n \to \mathbb{R}^m$, $1 \leq m \leq n$. The statistic summarises the data in that, instead of reporting the entire sample, only the value of the statistic $T(\mathbf{x}) = t$ is reported. An example of a statistic (also known as "summary statistic") is the sample mean (or average) $T(\mathbf{x}) = \bar{\mathbf{x}} = \dfrac{x_1 + \cdots + x_n}{n}$. For instance, if the sample $\mathbf{x}$ consists of the age of individuals in the UK population, a way of summarising this sample is to report only the average age of the population, in which case $T(\mathbf{x})$ is the sample mean.

In many cases, a statistical data analysis consists only on summarising a data set, using a choice of different summary statistics. This kind of analysis is known as "Descriptive Statistics" or "Descriptive Analysis". In fact, a descriptive analysis is usually the first step in statistical data analysis as it helps the statistician gain understanding about the features of the data. Other summaries that are used in practice are: the median (0.5 quantile) as well as other quantiles, the minimum of the sample, the maximum of the sample, and etcetera. In fact, the set of summary statistics given by the minimum of the sample, first quartile (0.25 quantile), the median, the third quartile (0.75 quantile), and the maximum is known as the "Five Number Summary". Visual tools are also used in applied statistics to understand other features of the data. These include boxplots, violin plots, histograms, smooth density plots, scatter plots, and etcetera. These tools are not covered in this course, but you may want to have a look at them if you are planning to pursue a career involving data analysis.

2.3 Point Estimators

"We need good point estimators because we cannot inject a 95% confidence interval of insulin"

– Professor David A. Sprott.

In statistical inference, it is often assumed that the random variables $X_1, \ldots, X_n$ are *i.i.d.*and that they are characterised by a parametric distribution $F(\cdot; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top \in \mathbb{R}^p$ are the parameters of the distribution and are assumed to be unknown. This is denoted as:

$$X_j \overset{i.i.d.}{\sim} F(\cdot; \boldsymbol{\theta}), \quad j = 1, \ldots, n.$$

Equivalently, we will use the notation in terms of the probability density function (pdf) $X_j \overset{i.i.d.}{\sim} f(\cdot; \boldsymbol{\theta})$. This idea motivates the following definition.

Definition 18. Parametric Statistical Model. A parametric statistical model is a set $\mathcal{S}$ of parametric probability distributions on a sample space Ω. More specifically,

$$\mathbf{x} = (x_1, \ldots, x_n)^\top \sim f(\cdot; \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p.$$

The pdf $f(\cdot;\boldsymbol{\theta})$ characterises the stochastic nature of the random variables, but the parameter is unknown. There are no rules for selecting the pdf $f(\cdot;\boldsymbol{\theta})$, and this has to be careful thought in the first steps of the analysis, taking into account the nature of the observations. For instance, based on whether the observations are discrete or continuous, whether their range is bounded or unbounded, and whether the observations are positive or can take any real value. Once the pdf $f(\cdot;\boldsymbol{\theta})$ for your sample is chosen, what remains is to estimate the unknown parameter $\boldsymbol{\theta}$. The rationale behind point estimation consists of using methods for finding an estimator of the unknown parameter $\boldsymbol{\theta}$, using the information in the sample $\mathbf{x}$. The general definition of a point estimator is presented below [1].

Definition 19. A point estimator is any function $W(X_1,\ldots,X_n)$ of a sample; that is, any statistic is a point estimator.

From this definition, we can see that an estimator is a random variable (or random vector). Thus, an *estimator* is a function of the sample, while an *estimate* is a realisation of the estimator. Note also that this definition does not establish a correspondence between the value of $W(X_1,\ldots,X_n)$ and the parameter $\boldsymbol{\theta}$, as the idea is to keep it as general as possible. Later, we will study particular methods that establish this connection explicitly.

Example 2.3.1. Let $X_j \overset{i.i.d.}{\sim} N(\mu,\sigma^2)$, with $j=1,\ldots,n$ and σ known (say $\sigma^2=1$). Then, $\mu = E[X_j] = \int_{\mathbb{R}} x f(x;\mu,\sigma)dx$. An estimator of μ is

$$W(X_1,\ldots,X_n) = \bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j.$$

Note that this estimator is defined in terms of the random variables. In contrast, an estimate of μ requires a sample of observations $\mathbf{x} = (x_1,\ldots,x_n)^\top$. Thus, an estimate of μ is

$$W(x_1,\ldots,x_n) = \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j.$$

In particular, if the sample is $\mathbf{x} = (9,10,11,8,12)^\top$, then the estimate of μ is $W(x_1,\ldots,x_5) = 10$.

Remark 2.3.1. *Given that a statistic (in particular, an estimator) is a random variable (vector), it is possible to associate it to its distribution function.*

In the following examples, we will discuss some specific estimators of parameters. Later, we will study methods that produce these estimators.

Example 2.3.2. Let $X_1,\ldots,X_n$ be *i.i.d.*Bernoulli random variables with parameter $0<\theta<1$ (the probability of success). Then, $T(\mathbf{X}) = X_1 + \ldots X_n$ is distributed as a $Binomial(n,\theta)$. See [6].

Example 2.3.3. Let $X_j \overset{i.i.d.}{\sim} N(\mu,\sigma^2)$, with $j=1,\ldots,n$ and σ^2 known. Then, $\mu = E[X_j] = \int_{\mathbb{R}} x f(x;\mu,\sigma)dx$. Let

$$W = W(X_1,\ldots,X_n) = \bar{\mathbf{X}} = \frac{1}{n}\sum_{j=1}^{n} X_j,$$

be an estimator of μ. Then, the distribution of the random variable $W = \bar{\mathbf{X}}$ is normal with mean μ and variance $\dfrac{\sigma^2}{n}$ (see [6]). This is

$$W = \bar{\mathbf{X}} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Although, in this case, it was simple to obtain the distribution of the estimator (by using the properties of the normal distribution), this is not usually possible. In most cases, the distribution of an estimator is not easy to obtain, and may not correspond to any distribution in the catalogue of distributions used in the literature [6].

2.4 Sufficiency

In statistical inference, it is important to understand general properties of estimators (or statistics in general). The reason for this is that any statistic $T(\mathbf{X})$ divides the information of the sample into two parts: the information contained (marginally) in the statistic $T(\mathbf{x}) = t$ and the information contained (conditionally) on "$\mathbf{x}$ given t". In mathematical terms, and using the rules of conditional probability, this means [17]

$$f_{\mathbf{X}}(x;\boldsymbol{\theta}) = f_T(t;\boldsymbol{\theta}) f_{\mathbf{X}|T}(\mathbf{x};\boldsymbol{\theta} \mid t). \tag{2.4.1}$$

In this line, an important concept is the *Sufficiency Principle* presented below.

Definition 20. *Sufficiency Principle*. A statistic $T(\mathbf{X})$ is a *sufficient statistic* for $\boldsymbol{\theta}$ if the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{X})$ does not depend on $\boldsymbol{\theta}$.

Connecting this definition with equation (2.4.1): if T is a sufficient statistic, then it follows that $f(\mathbf{x};\boldsymbol{\theta} \mid t) = f(\mathbf{x} \mid t)$ (where we are omitting the subindex in order to simplify the notation). That is, the conditional distribution $f(\mathbf{x} \mid t)$ of the sample given t (given that we know the value of the statistic) does not depend on $\boldsymbol{\theta}$. Thus, if T is a sufficient statistic, equation (2.4.1) becomes

$$f_{\mathbf{X}}(x;\boldsymbol{\theta}) = f_T(t;\boldsymbol{\theta}) f_{\mathbf{X}|T}(\mathbf{x} \mid t) \quad \text{for } T \text{ sufficient.} \tag{2.4.2}$$

We will see that this definition is key in point estimation theory as it implies that we only need the value of a sufficient statistic if we want to estimate an unknown parameter. At a more intuitive level, a sufficient statistic for a parameter $\boldsymbol{\theta}$ is a statistic that captures all the information about the parameter $\boldsymbol{\theta}$ contained in the sample.

Equation (2.4.2) also indicates that the distribution (density) of the sample can be factorised into two terms, one depending on $\boldsymbol{\theta}$ and the sufficient statistics, and another one depending only on the sample (as t is a function of the sample). This idea is formalised in the following result, known as the *Factorisation Theorem*.

Theorem 2.4.1. *Factorisation Theorem. Let $f(\mathbf{x};\boldsymbol{\theta})$ denote the joint pdf or pmf of a sample $\mathbf{X}$. A statistic $T(\mathbf{X})$ is a sufficient statistic for $\boldsymbol{\theta}$ if and only if there exist non-negative functions $g(t;\boldsymbol{\theta})$ and $h(\mathbf{x})$ such that, for all sample points $\mathbf{x}$ and all parameter points $\boldsymbol{\theta}$,*

$$f(\mathbf{x};\boldsymbol{\theta}) = g\left(T(\mathbf{x});\boldsymbol{\theta}\right) h(\mathbf{x}). \tag{2.4.3}$$

Proof. The proof is not presented here as it is not essential for the course. For the discrete case, see Theorem 6.2.6 from [1]. For the continuous case, see Theorem 2.6.2 and Corollary 2.6.1 from [11].

An immediate result from the factorisation theorem is:

Corollary 2.4.1. *if $f(\mathbf{x};\boldsymbol{\theta})$ is the pdf of pmf of $\mathbf{X}$ and $q(t;\theta)$ is the pdf or pmf of $T(\mathbf{X})$, then $T(\mathbf{X})$ is a sufficient statistic for $\boldsymbol{\theta}$ if, for every $\mathbf{x}$ in the sample space, the ratio*

$$\frac{f(\mathbf{x};\boldsymbol{\theta})}{q\left(T(\mathbf{x});\boldsymbol{\theta}\right)},$$

is constant as a function of $\boldsymbol{\theta}$.

Although simple, this result is very useful to identify sufficient statistics, as shown in the following example.

Example 2.4.1. Consider again the case where $X_1, \ldots, X_n$ be *i.i.d.* Bernoulli random variables with parameter $0 < \theta < 1$, and consequently $T(\mathbf{X}) = X_1 + \ldots X_n$ is distributed as a $Binomial(n, \theta)$. Thus, using that the variables are *i.i.d.*, it follows that

$$\begin{aligned} f(\mathbf{x};\theta) &= \prod_{j=1}^{n} \theta^{x_j}(1-\theta)^{1-x_j} = \theta^{\sum_j x_j}(1-\theta)^{n-\sum_j x_j} = \theta^t(1-\theta)^{n-t}, \\ q(t;\theta) &= \binom{n}{t}\theta^t(1-\theta)^{n-t}, \end{aligned}$$

where $t = \sum_j x_j$. Then,

$$\frac{f(\mathbf{x};\theta)}{q(t;\theta)} = \frac{1}{\binom{n}{t}},$$

which does not depend upon θ, and we conclude that T is a sufficient statistic for θ.

Definition 21. A sufficient statistic $T(\mathbf{X})$ is called a *minimal sufficient statistic* if, for any other sufficient statistic $T'(\mathbf{X})$, $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$.

This definition has the following important implication.

Theorem 2.4.2. *Let $f(\mathbf{x};\boldsymbol{\theta})$ be the pdf or pmf of the sample $\mathbf{X}$. Suppose that there exists a function $T(\mathbf{x})$ such that, for every two sample points $\mathbf{x}$ and $\mathbf{y}$, the ratio $f(\mathbf{x};\boldsymbol{\theta})/f(\mathbf{y};\boldsymbol{\theta})$ is constant as a function of $\boldsymbol{\theta}$ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Then, $T(\mathbf{X})$ is a minimal sufficient statistic.*

Proof. For simplicity, suppose that $f(\mathbf{x};\boldsymbol{\theta}) > 0$ for all $\mathbf{x} \in \mathcal{X}$ and $\boldsymbol{\theta} \in \Theta$. We need to show that T is a sufficient statistic and that it is minimal.

Let $T : \mathcal{X} \to \mathcal{T}$. Define the sets $A_t : \{\mathbf{x} : T(\mathbf{x}) = t\}$. For any $\mathbf{x} \in \mathcal{X}$, let us denote $\mathbf{x}_{T(\mathbf{x})}$ an element of A_t. Since $\mathbf{x}$ and $\mathbf{x}_{T(\mathbf{x})}$ are in the same set, by definition, then $T(\mathbf{x}) = T\left(\mathbf{x}_{T(\mathbf{x})}\right)$. Hence, $f(\mathbf{x};\boldsymbol{\theta})/f(\mathbf{x}_{T(\mathbf{x})};\boldsymbol{\theta})$ is a constant function of $\boldsymbol{\theta}$. Define the function

$$h(\mathbf{x}) = \frac{f(\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{x}_{T(\mathbf{x})};\boldsymbol{\theta})},$$

which does not depend upon $\boldsymbol{\theta}$. Then, we can rewrite

$$f(\mathbf{x};\boldsymbol{\theta}) = \frac{f(\mathbf{x}_{T(\mathbf{x})};\boldsymbol{\theta})f(\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{x}_{T(\mathbf{x})};\boldsymbol{\theta})} = g(T(\mathbf{x});\boldsymbol{\theta})h(\mathbf{x}),$$

and, by the Factorisation Theorem, it follows that T is a sufficient statistic.

The next step is to show that it is minimal. In order to show this, let $T'(\mathbf{x})$ be any other sufficient statistic. By the Factorisation Theorem, there exist functions g' and h' such that $f(\mathbf{x};\boldsymbol{\theta}) = g'(T'(\mathbf{x});\boldsymbol{\theta})h'(\mathbf{x})$. Let $\mathbf{x}$ and $\mathbf{y}$ be any two sample points with $T'(\mathbf{x}) = T'(\mathbf{y})$. Then,

$$\frac{f(\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{y};\boldsymbol{\theta})} = \frac{g'(T'(\mathbf{x});\boldsymbol{\theta})h'(\mathbf{x})}{g'(T'(\mathbf{y});\boldsymbol{\theta})h'(\mathbf{y})} = \frac{h'(\mathbf{x})}{h'(\mathbf{y})}.$$

Since this ratio does not depend on $\boldsymbol{\theta}$, the assumptions of the theorem imply that $T(\mathbf{x}) = T(\mathbf{y})$. Thus, $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$ and $T(\mathbf{x})$ is minimal.

Example 2.4.2. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be *i.i.d.* $N(\mu, \sigma^2)$, with both μ and σ^2 unknown. Let $\mathbf{x}$ and $\mathbf{y}$ denote two sample points, and let $(\bar{\mathbf{x}}, s^2_{\mathbf{x}})$ and $(\bar{\mathbf{y}}, s^2_{\mathbf{y}})$ be the sample means and variances associated to $\mathbf{x}$ and $\mathbf{y}$, respectively. Then, noting that $s^2_{\mathbf{x}} = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$ is the sample variance, we obtain:

$$f(\mathbf{x};\mu,\sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\exp\left\{-\frac{n}{2\sigma^2}\left[s^2_{\mathbf{x}} + (\bar{\mathbf{x}} - \mu)^2\right]\right\}.$$

Analogously for $\mathbf{y}$. Consequently, after expanding the binomials:

$$\frac{f(\mathbf{x};\mu,\sigma)}{f(\mathbf{y};\mu,\sigma)} = \exp\left\{-\frac{1}{2\sigma^2}\left[n(\bar{\mathbf{x}}^2-\bar{\mathbf{y}}^2)-2n\mu(\bar{\mathbf{x}}-\bar{\mathbf{y}})+n(s_{\mathbf{x}}^2-s_{\mathbf{y}}^2)\right]\right\}.$$

This ratio will be constant as a function of μ and σ^2 if and only if $\bar{\mathbf{x}}=\bar{\mathbf{y}}$ and $s_{\mathbf{x}}^2=s_{\mathbf{y}}^2$. Thus, by Theorem 2.4.2, it follows that $(\bar{\mathbf{X}}, S^2)$ is a minimal sufficient statistic for (μ,σ^2).

2.5 Unbiased estimators

Another important class of unbiased estimators. An estimator $T(\mathbf{X})$, of a parameter $g(\boldsymbol{\theta})$, is a random variable (vector) itself. Consequently, for each observed sample, $T(\mathbf{x})$ has random variability that can be characterised with the same properties used to characterise other random variables (vectors). One important characteristic of a random vector is its mean (expected value), which tells us about the long-run average values the estimator takes in repetitions of the experiment (that is, for different samples). The concept of unbiasedness is connected with the expected value of an estimator as follows.

Definition 22. An estimator $T(\mathbf{X})$ of the parameter $g(\boldsymbol{\theta})$ is unbiased if

$$E[T(\mathbf{X})] = g(\boldsymbol{\theta}), \quad \text{for all } \boldsymbol{\theta}\in\Theta,$$

where g is a measurable function.

In particular, if g is the identity function, then we say that an estimator $T(\mathbf{X})$ of the parameter $\boldsymbol{\theta}$ is unbiased if $E[T(\mathbf{X})]=\boldsymbol{\theta}$.

Example 2.5.1. In Example 2.3.3, the estimator of $\boldsymbol{\mu}$ is $T(\mathbf{X})=\bar{\mathbf{X}}$ and its distribution was shown to be $N(\boldsymbol{\mu},\Sigma/n)$. Clearly, $E[T(\mathbf{X})]=\boldsymbol{\mu}$ for all n. Consequently, $T(\mathbf{X})$ is an unbiased estimator of $\boldsymbol{\mu}$.

Example 2.5.2. Let $\mathbf{X}_1,\ldots,\mathbf{X}_n$ *i.i.d.* $p-$dimensional random vectors with multivariate normal distribution with mean $\boldsymbol{\mu}$ and known covariance matrix Σ. Then, $T(\mathbf{X})=\bar{\mathbf{X}}=\frac{1}{n}\sum_{i=1}^n\mathbf{X}_i$ is distributed according to a multivariate normal $N_p(\boldsymbol{\mu},\Sigma/n)$. Clearly, $E[T(\mathbf{X})]=\boldsymbol{\mu}$ for all n. Consequently, $T(\mathbf{X})$ is an unbiased estimator of $\boldsymbol{\mu}$.

Unbiased estimators suffer from a number of drawbacks. The next example illustrates a case where there is an unbiased estimator, but this makes no sense as it gives implausible values.

Example 2.5.3. A nonsense unbiased estimator. Let X be a Poisson random variable with mean $\lambda>0$. Recall that the pmf of X is given by

$$p(x;\lambda)=P(X=x;\lambda)=\frac{\lambda^x e^{-\lambda}}{x!}.$$

Suppose that we are interested in estimating the parameter $\theta=e^{-3\lambda}$ based on a sample of size one. Let $T(X)=(-2)^X$. Then, the expectation is

$$\begin{aligned} E(T) &= \sum_{x=0}^{\infty}(-2)^x\frac{\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda}\sum_{x=0}^{\infty}\frac{(-2\lambda)^x}{x!} \\ &= e^{-\lambda}e^{-2\lambda}=e^{-3\lambda}. \end{aligned}$$

Therefore, T is unbiased for $e^{-3\lambda}$. However, T is unreasonable in the sense that it may be negative (for X odd), even though it is an estimator of a strictly positive quantity. This suggests that one should not automatically assume that a unique unbiased estimator is necessarily good.

Example 2.5.4. Unbiased estimators are not invariant (in general). Let $X_1, \ldots, X_n$ be *i.i.d.* random variables with mean $\lambda > 0$. Consider the estimator $T(\mathbf{X}) = \bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^n X_i$. Then, $E[T(\mathbf{X})] = E[X_1] = \lambda$, and we conclude that T is an unbiased estimators. Now, consider the estimator $T_2(\mathbf{X}) = T(\mathbf{X})^2$. Then, recalling that the second moment of the Poisson distribution is $\lambda^2 + \lambda$

$$\begin{aligned} E[T_2(\mathbf{X})] &= E[T(\mathbf{X})^2] = E\left[\frac{(X_1 + \cdots + X_n)^2}{n^2}\right] \\ &= \frac{1}{n^2} E\left[\sum_{i=1}^n E[X_i^2] + \sum_{i \neq j} E[X_i X_j]\right] \\ &= \frac{n(\lambda^2 + \lambda) + n(n-1)\lambda^2}{n^2} = \lambda^2 + \frac{\lambda}{n}. \end{aligned}$$

Thus, $E[T^2] \neq \lambda^2$.

In fact, using Jensen's inequality, we can also see that the bias can be higher or larger, depending on whether the function applied to the unbiased estimator is convex or concave.

2.6 Asymptotic Efficiency

The Fisher Information Matrix

Let $\mathbf{X}$ be a sample of independent observations from a distribution $F(\cdot; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. Suppose that the distribution $F(\cdot; \boldsymbol{\theta})$ possesses densities or mass functions $f(\cdot; \boldsymbol{\theta})$. The *Fisher information matrix* (FIM, also known as the Expected information matrix, or Expected Fisher information matrix) is defined as

$$I(\boldsymbol{\theta}) = \left[E_{\boldsymbol{\theta}} \left\{ \frac{\partial \log f(X; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i} \cdot \frac{\partial \log f(X; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} \right\} \right]_{p \times p}, \tag{2.6.1}$$

where $X \sim F(\cdot; \boldsymbol{\theta})$ and the expectation is taken with respect to this distribution. This matrix is defined and is positive definite under regularity conditions that mainly involve the finiteness of the integral in the expectation.

Calculating the FIM for specific distribution is often difficult and involves lengthy and tricky calculations. However, it is often more useful to study the FIM for general families of distributions in order to understand the functional dependence of this matrix on the parameters of the distribution. One example of this is the location scale family (see Preliminary Material)

Example 2.6.1. Consider the location-scale family with density $s(x; \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x - \mu}{\sigma}\right)$, where f is a symmetric density with mode (maximum) at $x = \mu$ and support on $\mathbb{R}$ (and does not depend on (μ, σ)). The FIM is a 2×2 matrix with entries (to be calculated in the lecture)

$$I_{11} = \frac{1}{\sigma^2} \int_{-\infty}^{\infty} \left[\frac{f'(y)}{f(y)}\right]^2 f(y) dy = \frac{1}{\sigma^2} \int_{-\infty}^{\infty} \frac{[f'(y)]^2}{f(y)} dy, \tag{2.6.2}$$

$$I_{22} = \frac{1}{\sigma^2} \int_{-\infty}^{\infty} \left[1 + y\frac{f'(y)}{f(y)}\right]^2 f(y) dy, \tag{2.6.3}$$

$$I_{12} = \frac{1}{\sigma^2} \int_{-\infty}^{\infty} y \left[\frac{f'(y)}{f(y)}\right]^2 f(y) dy = 0, \tag{2.6.4}$$

$$I_{21} = I_{12}. \tag{2.6.5}$$

Thus FIM only depends on σ, through the factor $\frac{1}{\sigma^2}$ for the entire location-scale family. Expressions for the FIM associated to particular choices of f can be obtained by calculating the derivatives and integrals in these formulae. For example, when f is the Normal or Logistic distribution.

The following result provides an alternative expression that is easier to calculate for some distributions.

Theorem 2.6.1. *Suppose that $p = 1$ (one parameter) and that the following conditions are satisfied:*

(i) $\frac{\partial}{\partial\theta}\int f(x;\theta)dx = \int \frac{\partial f(x;\theta)}{\partial\theta}dx = 0.$

(ii) $\frac{\partial}{\partial\theta}\int g(x)f(x;\theta)dx = \int g(x)\frac{\partial f(x;\theta)}{\partial\theta}dx$, *where g is any integrable function that does not depend on θ.*

(iii) $0 < E\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)^2 < \infty.$

(iv) $\log f(x;\boldsymbol{\theta})$ *is twice differentiable with respect to $\boldsymbol{\theta}$.*

Then,

$$I(\theta) = E\left\{\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)^2\right\} = -E\left\{\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right\}. \tag{2.6.6}$$

Proof. Note that

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\log f(X;\theta) &= \frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)} - \left(\frac{\frac{\partial}{\partial\theta}f(X;\theta)}{f(X;\theta)}\right)^2 \\ &= \frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)} - \left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2.\end{aligned}$$

Taking expectation with respect to $f(x;\theta)$ on both sides, we obtain

$$E\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right] = E\left[\frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)}\right] - E\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right].$$

From the assumptions, it follows that

$$\begin{aligned}E\left[\frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)}\right] &= \int \frac{\partial^2}{\partial\theta^2}f(x;\theta)dx \\ &= \frac{\partial^2}{\partial\theta^2}\int f(x;\theta)dx = \frac{\partial^2}{\partial\theta^2}1 = 0.\end{aligned}$$

This result can be extended to $p > 1$ by assuming that the assumptions (i)–(iv) are satisfied for each entry of the FIM. The corresponding expression is

$$I(\boldsymbol{\theta}) = -\left[E_{\boldsymbol{\theta}}\left\{\frac{\partial^2 \log f(X;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}\right\}\right]_{p\times p}.$$

Remark 2.6.1. *Let X and Y be independent with the Fisher information matrices $I_X(\boldsymbol{\theta})$ and $I_Y(\boldsymbol{\theta})$, respectively. Then, the Fisher information about $\boldsymbol{\theta}$ contained in (X, Y) is $I_X(\boldsymbol{\theta}) + I_Y(\boldsymbol{\theta})$. In particular, if $X_1, ..., X_n$ are* i.i.d. *and $I_1(\boldsymbol{\theta})$ is the Fisher information about $\boldsymbol{\theta}$ contained in a single X_i, then the Fisher information about $\boldsymbol{\theta}$ contained in $X_1, ..., X_n$ is $nI_1(\boldsymbol{\theta})$.*

The FIM appears in a number of theories and applications, ranging from asymptotic normality of estimators, to Bayesian theory, design of experiments, and physics. The next result represents an essential result in the study of unbiased estimators.

The Cramér-Rao lower bound

Theorem 2.6.2. *(***Cramér-Rao lower bound***) Let $X_1, \ldots, X_n$ be a sample with pdf $f(\mathbf{x};\theta)$, and let $W(\mathbf{X})$ be any estimator where $\varphi(\theta) = E[W(\mathbf{X})]$ is a differentiable function of θ. Suppose the joint pdf $f(\mathbf{x};\theta)$ satisfies*

$$\frac{\partial}{\partial\theta}\int_{\chi^n} h(\mathbf{x})f(\mathbf{x};\theta)dx_1\ldots dx_n = \int_{\chi^n} h(\mathbf{x})\frac{\partial f(\mathbf{x};\theta)}{\partial\theta}dx_1\ldots dx_n,$$

for any function $h(\mathbf{x})$ with $E|h(\mathbf{X})| < \infty$. We implicitly assume that we can exchange the integral and the derivative sign. Then,

$$Var[W(\mathbf{X})] \;\geq\; \frac{\left(\frac{d}{d\theta}E[W(\mathbf{X})]\right)^2}{E\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right)^2\right]}$$

Proof. The proof will follow by the Cauchy-Schwarz inequality (see Preliminary Material):

$$[Cov(\tilde{X},\tilde{Y})]^2 \leq (Var\tilde{X})(Var\tilde{Y}).$$

If we rearrange the previous equation, we can get a lower bound on the variance of $\tilde{X}$,

$$\frac{[Cov(\tilde{X},\tilde{Y})]^2}{Var\tilde{Y}} \leq Var\tilde{X}.$$

Under the assumptions in this theorem, we have that

$$\begin{aligned}
E\left[\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right] &= E\left[\frac{\frac{\partial f(\mathbf{X};\theta)}{\partial\theta}}{f(\mathbf{X};\theta)}\right] \\
&= \int_{\chi^n}\frac{\frac{\partial f(\mathbf{x};\theta)}{\partial\theta}}{f(\mathbf{x};\theta)}f(\mathbf{x};\theta)dx_1\ldots dx_n \\
&= \int_{\chi^n}\frac{\partial f(\mathbf{x};\theta)}{\partial\theta}dx_1\ldots dx_n \\
&= \frac{\partial}{\partial\theta}\int_{\chi^n} f(\mathbf{x};\theta)dx_1\ldots dx_n = \frac{\partial}{\partial\theta}1 = 0.
\end{aligned}$$

Then,

$$E\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right)^2\right] = Var\left[\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right].$$

Then, the proof follows by selecting $\tilde{X} = W(\mathbf{X})$ and $\tilde{Y} = \frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)$:

$$\begin{aligned}
Cov\left[W(\mathbf{X}),\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right] &= E\left[W(\mathbf{X})\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right] \\
&= E\left[W(\mathbf{X})\frac{\frac{\partial f(\mathbf{X};\theta)}{\partial\theta}}{f(\mathbf{X};\theta)}\right] \\
&= \int_{\chi^n} W(\mathbf{x})\frac{\frac{\partial f(\mathbf{x};\theta)}{\partial\theta}}{f(\mathbf{x};\theta)}f(\mathbf{x};\theta)dx_1\ldots dx_n
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\chi^n} W(\mathbf{x}) \frac{\partial f(\mathbf{x};\theta)}{\partial \theta} dx_1 \dots dx_n \\
&= \frac{\partial}{\partial \theta} \int_{\chi^n} W(\mathbf{x}) f(\mathbf{x};\theta) dx_1 \dots dx_n \\
&= \frac{d}{d\theta} E[W(\mathbf{X})].
\end{aligned}
$$

Combining the Cauchy-Schwarz inequality with the previous results:

$$
Var[W(\mathbf{X})] \geq \frac{\left(\frac{d}{d\theta} E[W(\mathbf{X})]\right)^2}{E\left[\left(\frac{\partial}{\partial \theta} \log f(\mathbf{X};\theta)\right)^2\right]}.
$$

Corollary 2.6.1. *If the sample $X_1, \dots, X_n$ be* i.i.d. *random variables. Then,*

$$
E\left[\left(\frac{\partial}{\partial \theta} \log f(\mathbf{X};\theta)\right)^2\right] = nE\left[\left(\frac{\partial}{\partial \theta} \log f(X;\theta)\right)^2\right] = nI(\theta).
$$

Proof.

$$
\begin{aligned}
E\left[\left(\frac{\partial}{\partial \theta} \log f(\mathbf{X};\theta)\right)^2\right] &= E\left[\left(\frac{\partial}{\partial \theta} \log \prod_{j=1}^{n} f(X_j;\theta)\right)^2\right] \\
&= E\left[\left(\sum_{j=1}^{n} \frac{\partial}{\partial \theta} \log f(X_j;\theta)\right)^2\right] \\
&= \sum_{j=1}^{n} E\left[\left(\frac{\partial}{\partial \theta} \log f(X_j;\theta)\right)^2\right] + \sum_{i \neq j} E\left[\frac{\partial}{\partial \theta} \log f(X_i;\theta) \frac{\partial}{\partial \theta} \log f(X_j;\theta)\right] \\
&= \sum_{j=1}^{n} E\left[\left(\frac{\partial}{\partial \theta} \log f(X_j;\theta)\right)^2\right] + \sum_{i \neq j} \cancelto{0}{E\left[\frac{\partial}{\partial \theta} \log f(X_i;\theta)\right]} \cancelto{0}{E\left[\frac{\partial}{\partial \theta} \log f(X_j;\theta)\right]} \\
&= nE\left[\left(\frac{\partial}{\partial \theta} \log f(X;\theta)\right)^2\right].
\end{aligned}
$$

There are some natural consequences of these results:

Remark 2.6.2.

(i) If $W(\mathbf{X})$ is an unbiased estimator of θ, then,

$$
Var[W(\mathbf{X})] \geq \frac{1}{nI(\theta)},
$$

where $I(\theta)$ is the Fisher information associated to a single random variable.

(ii) If $W(\mathbf{X})$ is an unbiased estimator of $g(\theta)$, then,

$$
Var[W(\mathbf{X})] \geq \frac{[g'(\theta)]^2}{nI(\theta)}.
$$

Definition 23. The *efficiency* of an unbiased estimator is defined as:

$$
e(W(\mathbf{X})) = \frac{I_n(\theta)^{-1}}{Var[W(\mathbf{X})]},
$$

where $I_n(\theta)$ is the Fisher information of the sample. Using the CRLB, it follows that $e(W(\mathbf{X})) \leq 1$.

Now, we present some illustrative examples of the Cramér-Rao lower bound (CRLB).

Example 2.6.2. Let $X_1, ..., X_n$ be *i.i.d.* from the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathbb{R}$ and a known σ^2. An unbiased estimator of μ is $T(\mathbf{X}) = \bar{X} = \frac{1}{n}\sum_{i=1}^n X_j$. By the properties of the normal distribution, $\bar{X} \sim N(\mu, \sigma^2/n)$. Thus, the $Var(T(\mathbf{X})) = \sigma^2/n$. Now, from the calculation in (2.6.2) and using that $\phi'(x) = x\phi(x)$ (try to prove this), where ϕ is the standard normal pdf, it follows that $I(\mu) = \frac{n}{\sigma^2}$. Consequently, the CRLB is attained with equality, as $Var(T(\mathbf{X})) = 1/I(\mu)$.

Example 2.6.3. Let X be a single sample from a $Bin(n, \theta)$, where n is known. The pmf is

$$f(x;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

Consequently,

$$\begin{aligned}
\log f(x;\theta) &= \log\binom{n}{x} + x\log(\theta) + (n-x)\log(1-\theta), \\
\frac{\partial}{\partial\theta}\log f(x;\theta) &= \frac{x}{\theta} - \frac{n-x}{1-\theta} = \frac{x-n\theta}{\theta(1-\theta)}. \\
I(\theta) &= E\left[\left(\frac{X-n\theta}{\theta(1-\theta)}\right)^2\right] = \frac{E\left[(X-n\theta)^2\right]}{\theta^2(1-\theta)^2} = \frac{Var(X)}{\theta^2(1-\theta)^2} = \frac{n\theta(1-\theta)}{\theta^2(1-\theta)^2} = \frac{n}{\theta(1-\theta)}.
\end{aligned}$$

Thus, the variance of any unbiased estimator of θ, $T(X)$, satisfies $Var(T(X)) \geq \frac{\theta(1-\theta)}{n}$.

In fact, an unbiased estimator θ is $T(X) = X/n$, which achieves the CRLB with equality again.

Example 2.6.4. The regularity conditions cannot be neglected.Let X be a random variable with uniform distribution on $(0, \theta)$, with $\theta > 0$. The pdf of X is

$$f(x;\theta) = \frac{1}{\theta}I_{(0,\theta)}(x),$$

where $I_{(0,\theta)}(x) = 1$ if $x \in (0, \theta)$ and $I_{(0,\theta)}(x) = 0$ otherwise, is the indicator function of the interval $(0, \theta)$. Let's omit the step of checking the regularity conditions, and proceed to calculate the Fisher information:

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right] = -E\left[\frac{1}{\theta^2}\right] = -\frac{1}{\theta^2}.$$

Now, by assumption (i), which we are not checking yet, we also know that

$$I(\theta) = Var\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right] = Var\left[-\frac{1}{\theta}\right] = 0.$$

On the other hand, if we use the original definition

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right] = E\left[\frac{1}{\theta^2}\right] = \frac{1}{\theta^2}.$$

Each of these would give a different value for the CLRB, the second one being ∞! The error here is to assume (i), which does not hold:

$$\begin{aligned}
-\frac{1}{\theta} &= E\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\right] = \int_0^\theta \frac{\partial}{\partial\theta}f(x;\theta)dx \\
&\neq \frac{\partial}{\partial\theta}\int_0^\theta f(x;\theta)dx = \frac{\partial}{\partial\theta}1 = 0.
\end{aligned}$$

Moreover, if we try to use the CRLB with the original definition of the Fisher information, we obtain that $Var(T) > \frac{1}{I(\theta)} = \theta^2$, for any unbiased estimator T. Now, consider the estimator $T(X) = 2X$, we can see that

$$E[T(X)] = 2\int_0^\theta \frac{x}{\theta} dx = \theta.$$

Thus, T is an unbiased estimator of θ. Now,

$$Var[T(X)] = \int_0^\theta \frac{(2x-\theta)^2}{\theta} dx = \frac{\theta^2}{3}.$$

Therefore, we obtain

$$Var[T(X)] = \frac{\theta^2}{3} < \frac{1}{I(\theta)} = \theta^2.$$

which contradicts the CRLB.

2.7 Maximum Likelihood Estimation

Historical Note 1. The method of maximum likelihood estimation was popularised by Sir Ronald A. Fisher in 1913, although earlier uses date back to 1778 by Daniel Bernoulli. Fisher was a British statistician and geneticist. His contributions in statistics include the development of the maximum likelihood method, the study of parametric distributions, and the design of experiments. The method of maximum likelihood estimation is often called "The Classical Approach".

The Likelihood Function

Now, we introduce the general definition of the *likelihood function*.

Definition 24. Let $f(\mathbf{x}; \boldsymbol{\theta})$ denote the joint pdf or pmf of the sample $\mathbf{X} = (X_1, \ldots, X_n)^\top$. Then, given that $\mathbf{X} = \mathbf{x}$ is observed, the function of $\boldsymbol{\theta}$ defined by

$$L(\boldsymbol{\theta} \mid \mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta}),$$

is called the *likelihood function*, where $\boldsymbol{\theta} \in \Theta$ and $\Theta \subset \mathbb{R}^p$ is the *parameter space*.

When, When $\mathbf{x} = (x_1, \ldots, x_n)$ is a collection of *i.i.d.* observations the likelihood function is

$$L(\boldsymbol{\theta} \mid \mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta}) = \prod_{j=1}^{n} f(x_j; \boldsymbol{\theta}).$$

Thus, the likelihood function is basically a role inversion of the argument of the joint pdf or pmf, in the sense that the likelihood function represents *a function of the parameters*, for a given sample, in contrast to the pdf (pmf) which represents a function of the sample, for a given parameter value.

In some cases, instead of using a semicolon, a vertical line is used to separate the role of the samples and the role of the parameters. These are equivalent notations, and their choice is essentially a matter of taste. In some textbooks you will find this definition in terms of the notation $L(\boldsymbol{\theta} \mid \mathbf{x}) = f(\mathbf{x} \mid \boldsymbol{\theta})$. Later, we will see that the vertical line notation is preferred in the context of "Bayesian statistics" as this has a connection with conditional probability.

Example 2.7.1. Let $x = 3$ be a sample from a Binomial distribution with $n = 10$ trials and probability of success $\theta \in (0, 1)$. Then, the likelihood function of θ is given by

$$L(\theta \mid x) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \propto \theta^3 (1-\theta)^7,$$

where the symbol "$\propto$" denotes "proportional to". In the previous formula, we are using that the binomial coefficient does not depend upon θ.

The likelihood function plays a key role is statistical inference, and defines a school of thought that utilises the likelihood function in order to estimate the parameter $\boldsymbol{\theta}$ of any statistical model. Later, we will discuss other approaches, such as the "Bayesian" approach. The justification for using the likelihood function as a tool for estimating parameters, and conducting statistical inference in general, is the **Likelihood Principle**. This principle asserts that all the information about a parameter of a model is contained in the observed likelihood function.

Definition 25. The Likelihood Principle. If $\mathbf{x}$ and $\mathbf{y}$ are two sample points such that $L(\boldsymbol{\theta} \mid \mathbf{x})$ is proportional to $L(\boldsymbol{\theta} \mid \mathbf{y})$, that is, there exists a constant $C(\mathbf{x}, \mathbf{y})$ such that

$$L(\boldsymbol{\theta} \mid \mathbf{x}) = C(\mathbf{x}, \mathbf{y}) L(\boldsymbol{\theta} \mid \mathbf{y}), \quad \text{for all} \quad \boldsymbol{\theta},$$

then the conclusions drawn from $\mathbf{x}$ and $\mathbf{y}$ should be identical.

This has important implications as it can be connected with the concept of sufficient statistic as well. Let $\mathbf{X}$ be a *i.i.d.* sample from a distribution $f(\cdot; \boldsymbol{\theta})$, and let $T(\mathbf{X})$ be a sufficient statistic. Denote $\mathbf{x}$ and t the realisations of $\mathbf{X}$ and $T(\mathbf{X})$, respectively. From the definition of sufficient statistic (2.4.2) and the factorisation theorem (2.4.3), it follows that:

$$L(\boldsymbol{\theta} \mid \mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta}) = g(T(\mathbf{x}); \boldsymbol{\theta}) h(\mathbf{x}) \propto g(t; \boldsymbol{\theta}).$$

This implies that the likelihood function can be defined, up to a proportionality constant, if we know the value of a sufficient statistic.

Remark 2.7.1. *In the case where $f(\cdot; \boldsymbol{\theta})$ is a pmf (that is, $\mathbf{X}$ is discrete), the likelihood function has a direct interpretation as the probability of observing the event $\mathbf{x}$ (the probability of observing that sample) given that the parameter takes the value $\boldsymbol{\theta}$. This is*

$$L(\boldsymbol{\theta} \mid \mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta}) = P(\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}). \tag{2.7.1}$$

Remark 2.7.2. *In the case where $f(\cdot; \boldsymbol{\theta})$ is a pdf (that is, $\mathbf{X}$ is continuous), the likelihood function cannot be directly interpreted as a probability, since a pdf assigns zero-probability to each observation (see preliminary material). Thus, the interpretation of the likelihood function in the continuous case is as follows.*

Let $\mathbf{X}$ be an i.i.d. *sample from $f(\cdot; \boldsymbol{\theta})$, with associated cdf $F(\cdot; \boldsymbol{\theta})$. When we observe a sample $\mathbf{x} = (x_1, \ldots, x_n)^\top$, each observation has an associated measurement error $\epsilon > 0$ (also known as "precision"). For instance, if you measure your height with a measuring tape that has centimetres as the smallest unit, then if your measured height is 175 centimetres, this value typically represents the rounding of the exact value, which is not observed due to the precision of the instrument. Thus, one can say that the height is actually $175 \pm \frac{1}{2}$ centimetres. Thus, for an observed sample x_i, which was measured with an instrument with precision ϵ, the probability of observing x_i is*

$$P\left(\textit{observing } x_i; \boldsymbol{\theta}\right) = P(X_i \in (x_i - \epsilon, x_i + \epsilon); \boldsymbol{\theta}).$$

Using the rules of probability, we get

$$P(X_i \in (x_i - \epsilon, x_i + \epsilon); \boldsymbol{\theta}) = F(x_i + \epsilon; \boldsymbol{\theta}) - F(x_i - \epsilon; \boldsymbol{\theta}).$$

Then,

$$\begin{aligned} P\left(\textit{observing } \mathbf{x}; \boldsymbol{\theta}\right) &= \prod_{i=1}^{n} P(X_i \in (x_i - \epsilon, x_i + \epsilon); \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \left[F(x_i + \epsilon; \boldsymbol{\theta}) - F(x_i - \epsilon; \boldsymbol{\theta})\right]. \end{aligned}$$

Now, if $\epsilon \approx 0$ (if the measurement error is small enough), then we can use the mid-value theorem to obtain

$$F(x_i + \epsilon; \boldsymbol{\theta}) - F(x_i - \epsilon; \boldsymbol{\theta}) = F'(c_i; \boldsymbol{\theta}) \cdot 2\epsilon = f(c_i; \boldsymbol{\theta}) \cdot 2\epsilon,$$

where $c_i \in (x_i - \epsilon, x_i + \epsilon)$. If f is continuous (and for small ϵ), then we can use the approximation $f(c_i; \boldsymbol{\theta}) \approx f(x_i; \boldsymbol{\theta})$. Consequently,

$$\begin{aligned} P\left(\textit{observing } \mathbf{x}; \boldsymbol{\theta}\right) &\approx 2^n \epsilon^n \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) \\ &\propto \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) = L(\boldsymbol{\theta} \mid \mathbf{x}). \end{aligned}$$

This link is often used to interpret the likelihood function as an approximation (up to a proportionality constant) to the probability of observing the sample $\mathbf{x}$, given a parameter value $\boldsymbol{\theta}$.

Example 2.7.2. Let $\mathbf{x} = (x_1, \ldots, x_n)^\top$ be an observed *i.i.d.* sample from a $N(\mu, \sigma^2)$, with $\mu \in \mathbb{R}$ and $\sigma > 0$ unknown. Then, the likelihood function of (μ, σ) is:

$$\begin{aligned} L(\mu, \sigma \mid \mathbf{x}) &= \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x_j - \mu)^2}{2\sigma^2} \right\} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n} (x_j - \mu)^2 \right\}. \end{aligned}$$

In fact, after some algebra, we can show that (exercise):

$$\sum_{j=1}^{n} (x_j - \mu)^2 = \sum_{j=1}^{n} (x_j - \bar{x})^2 + n(\bar{\mathbf{x}} - \mu)^2.$$

Then,

$$L(\mu, \sigma \mid \mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n} (x_j - \bar{\mathbf{x}})^2 - \frac{n}{2\sigma^2} (\bar{\mathbf{x}} - \mu)^2 \right\}.$$

Now, noting that $s^2 = \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x})^2$ is the sample variance, and removing the constant terms, we obtain:

$$\begin{aligned} L(\mu, \sigma \mid \mathbf{x}) &= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^n} \exp\left\{ -\frac{n}{2\sigma^2} \left[s^2 + (\bar{\mathbf{x}} - \mu)^2 \right] \right\} \\ &\propto \frac{1}{\sigma^n} \exp\left\{ -\frac{n}{2\sigma^2} \left[s^2 + (\bar{\mathbf{x}} - \mu)^2 \right] \right\}. \end{aligned}$$

Thus, we can observe that the likelihood is fully determined by the sample mean $\bar{\mathbf{x}}$ and the sample variance s^2. Consequently, the sufficient statistic for (μ, σ) is $T(\mathbf{x}) = (\bar{\mathbf{x}}, s^2)$.

Remark 2.7.3. *If, for two parameter values $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, we have that $L(\boldsymbol{\theta}_1 \mid \mathbf{x}) > L(\boldsymbol{\theta}_2 \mid \mathbf{x})$, we say that $\boldsymbol{\theta}_1$ is more "plausible" than $\boldsymbol{\theta}_2$.*

Note that we are using the word "plausible" rather than "probable". The reason for this is that the likelihood function *is not* a pdf or pmf, in general.

The Method of Maximum Likelihood Estimation

Suppose that a probability model has been formulated for an experiment, and that this model involves an unknown parameter $\boldsymbol{\theta} \in \Theta$. The outcome of the experiment is a sample of observations $\mathbf{x}$. The aim in statistical inference is to estimate the value of $\boldsymbol{\theta}$. In general terms, we want to determine the possible value (or values) of $\boldsymbol{\theta}$ that are plausible or likely in the light of the observations $\mathbf{x}$.

As previously discussed, the observed data can be interpreted as an event E in the sample space for the probability model. The probability of the event E can be determined from the probability model, and this probability can be seen as a function of the unknown parameter $P(E; \boldsymbol{\theta})$. The *Maximum Likelihood Estimate* (MLE) of $\boldsymbol{\theta}$ is the value of $\boldsymbol{\theta}$ which maximises $P(E; \boldsymbol{\theta})$. The MLE of $\boldsymbol{\theta}$ is typically denoted as $\widehat{\boldsymbol{\theta}}$. Thus, the MLE is interpreted as the value that best explains the data E in the sense that it maximises the probability of E under the model. As we saw in the previous section, the likelihood function can be directly interpreted as the probability of the event E in the discrete case. In the continuous case, the same definition applies if you account for the precision of the measuring instrument, and the use of the joint pdf to define the likelihood function represents an approximation.

Definition 26. The MLE is the value that maximises the likelihood function:

$$\widehat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta} \mid \mathbf{x}).$$

Thus, the process of estimation using the method of Maximum Likelihood Estimation can be translated as a maximisation (or optimisation) process. Since the properties and characteristics of the likelihood function depend on the probability model, and may contain more than one parameter, it becomes important to consider theoretical and numerical tools for maximising functions. In particular, the following tools are important.

- The value that maximises a function $f(x)$ (assuming it exists) is the same as the value that minimises the function $g(x) = -f(x)$.
- The value that maximises a function $f(x)$ (assuming it exists) is the same as the value that maximises the function $g(x) = cf(x)$, for any $c > 0$.
- The value that maximises a function $f(x)$ (assuming it exists) also maximises the function $g(x) = \log f(x)$.
- Let $f : \mathbb{R} \to \mathbb{R}_+$ be a continuously differentiable function (*i.e.* $f \in C^1(\mathbb{R}, \mathbb{R}_+)$), and let x_M be the value that maximises f, then, the first derivative of f satisfies $f'(x_M) = 0$.
- Let $f : \mathbb{R}^p \to \mathbb{R}_+$ be a continuously differentiable function (*i.e.* $f \in C^1(\mathbb{R}, \mathbb{R}_+)$), and let $\mathbf{x}_M$ be the value that maximises f, then, the gradient satisfies $\nabla f(\mathbf{x}_M) = \mathbf{0}$.

In addition, numerical optimisation such as the Newton's method and other numerical optimisation methods are useful to find the MLE numerically. In particular, the R command *optim()*, represents a powerful tool for maximising or minimising functions. In fact, some of these functions appear so often in maximum likelihood estimation that they even have specific names.

Remark 2.7.4. *The following notation, associated to a likelihood function $L(\boldsymbol{\theta} \mid \mathbf{x})$ with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top \in \Theta \subset \mathbb{R}^p$, will be used throughout.*

(a) In many cases, for the sake of simplicity of notation, the argument $\mathbf{x}$ is omitted from the likelihood function, as it is understood that the likelihood depends on the sample. This is, the notation $L(\boldsymbol{\theta})$ is often used instead of $L(\boldsymbol{\theta} \mid \mathbf{x})$.

(b) The logarithm of the likelihood function is called the log-likelihood function.

$$\ell(\boldsymbol{\theta} \mid \mathbf{x}) = \log L(\boldsymbol{\theta} \mid \mathbf{x}) = \sum_{i=1}^{n} \log f(x_i; \boldsymbol{\theta}).$$

(c) The gradient of the log-likelihood function, $S(\boldsymbol{\theta}) = \nabla \ell(\boldsymbol{\theta} \mid \mathbf{x})$ is known as the score function*. Here, we emphasise that the derivatives are taken with respect to (the entries of) $\boldsymbol{\theta}$. The MLE is the solution to $S(\boldsymbol{\theta}) = \nabla \ell(\boldsymbol{\theta} \mid \mathbf{x}) = \mathbf{0}$*

(d) For $p = 1$, the negative of the second derivative of the log-likelihood function is known as the observed information function. In one-parameter models

$$i(\theta) = -\frac{d^2}{d\theta^2} \ell(\theta \mid \mathbf{x}).$$

In models with two or more parameters ($p \geq 2$), the negative of the Hessian matrix is known as the observed information matrix:

$$i(\boldsymbol{\theta}) = -\nabla^2 \ell(\boldsymbol{\theta} \mid \mathbf{x}),$$

where ∇^2 represents the Hessian matrix, which is the matrix of second derivatives:

$$\nabla^2 \ell(\boldsymbol{\theta} \mid \mathbf{x}) = \begin{pmatrix} \frac{\partial^2}{\partial \theta_1^2} \ell(\boldsymbol{\theta} \mid \mathbf{x}) & \dots & \frac{\partial^2}{\partial \theta_1 \theta_p} \ell(\boldsymbol{\theta} \mid \mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial \theta_p \theta_1} \ell(\boldsymbol{\theta} \mid \mathbf{x}) & \dots & \frac{\partial^2}{\partial \theta_p^2} \ell(\boldsymbol{\theta} \mid \mathbf{x}) \end{pmatrix},$$

where, with a slightly abuse of notation, we are denoting

$$\frac{\partial^2}{\partial \theta_j^2} \ell(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{\partial^2}{\partial t^2} f(\theta_1, \dots, \theta_{j-1}, t, \theta_{j+1}, \dots, \theta_p) \Big|_{t=\theta_j}.$$

Next, we will illustrate how to find the MLE in several models. The examples are presented in increasing order of complexity.

Example 2.7.3. Binomial model.

Suppose that we want to estimate θ, the proportion of people with Diabetes mellitus (DM) in a large homogenous population. To do this, we randomly select n individuals for testing, and find that x of them have the disease. Since the population is large and homogeneous, we assume that n individuals tested are independent, and that each has probability θ of having DM. The probability of the observed event (data) is then

$$\begin{aligned} P(E; \theta) &= P(x \text{ out of } n \text{ have DM}) \\ &= L(\theta \mid x) \\ &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \\ &\propto \theta^x (1-\theta)^{n-x}, \end{aligned}$$

where $0 \leq \theta \leq 1$. The MLE $\widehat{\theta}$ is the value of θ that maximises the likelihood function $L(\theta; x, n)$. In order to obtain this maximum, we will use the log-likelihood function, as it is easier to maximise this function, and we will also omit the binomial coefficient, as this does not depend on θ. Thus,

$$\ell(\theta \mid x) = x \log(\theta) + (n-x) \log(1-\theta) \quad \text{for } 0 < \theta < 1.$$

For $0 < \theta < 1$, the score and observed information functions are:

$$\begin{aligned} S(\theta) &= \frac{d}{d\theta}\ell(\theta \mid x) = \frac{x}{\theta} - \frac{n-x}{1-\theta}, \\ i(\theta) &= -\frac{d^2}{d\theta^2}\ell(\theta \mid x) = \frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2}. \end{aligned}$$

Equating the score function to zero, $S(\theta) = 0$, and for the case $1 \leq x \leq n-1$, we obtain the unique solution $\frac{x}{n}$, and that $i(x/n) > 0$. Moreover, $L(\theta \mid x) = 0$ for $\theta = 0$ and $\theta = 1$. Thus, $\frac{x}{n}$ is the unique solution and consequently $\widehat{\theta} = \frac{x}{n}$. This value maximises the probability of the data, and we can say that the estimate of the portion of diseased person in the population is $\frac{x}{n}$, which corresponds to the portion of diseased persons in the sample.

Now, if $x = 0$, then $S(\theta) = 0$ has no solution, and the maximum is reached at the boundary of the parameter space $\Theta = [0, 1]$. More specifically,

$$L(\theta \mid x) \propto (1-\theta)^n,$$

which is a decreasing function of θ, and then $\widehat{\theta} = 0$.

Analogously, if $x = n$, then

$$L(\theta \mid x) \propto \theta^n,$$

which is an increasing function of θ, and then $\widehat{\theta} = 1$.

Consequently, the solution $\widehat{\theta} = \frac{x}{n}$ holds for all values of x.

Example 2.7.4. The tank problem.

"The enemy" has an unknown number N of tanks, which he has obligingly numbered $1, 2, \ldots, N$. Spies have reported sighting 8 tanks with numbers $\mathbf{x} = (137, 24, 86, 33, 92, 129, 17, 111)^\top$. Assume that sightings are independent, and that each of the N tanks has probability $1/N$ of being observed at each sighting. What is the MLE of N?

Let X_i be the serial number of tank i. Then, the pmf of these variables is:

$$P(x; N) = \begin{cases} \dfrac{1}{N}, & \text{for } x \leq N, \\ 0, & \text{for } x > N. \end{cases}$$

Given that each tank has the same probability of being observed, and that the largest sample value is $x_{(8)} = 137$, it follows that the likelihood function of N is

$$L(N \mid \mathbf{x}) = P(\text{Event}; N) = \begin{cases} \dfrac{1}{N^8}, & \text{for } N \geq 137, \\ 0, & \text{for } N < 137. \end{cases}$$

it is straightforward to see that the likelihood function is maximised at

$$\widehat{N} = \max_{i=1,\ldots,8} x_i = 137.$$

Q: Would you trust this estimate?

Example 2.7.5. Normal model.

Suppose that we want to know the distribution of the heights in the UK population. Let $\mathbf{x} = (x_1, \ldots, x_n)^\top$ be the sample of heights of a random sample of residents in the UK. We will assume that they are distributed according to a normal distribution with unknown mean μ and variance σ^2. This

assumption is, in fact, the usual assumption made by medical statisticians in practice. The likelihood function is

$$L(\mu, \sigma \mid \mathbf{x}) \quad \propto \quad \frac{1}{\sigma^n} \exp\left\{-\frac{n}{2\sigma^2}\left[s^2 + (\bar{\mathbf{x}} - \mu)^2\right]\right\}.$$

Then, the log-likelihood function is (omitting the proportionality constant):

$$\ell(\mu, \sigma \mid \mathbf{x}) \quad = \quad -n\log(\sigma) - \frac{n}{2\sigma^2}\left[s^2 + (\bar{\mathbf{x}} - \mu)^2\right].$$

The score function, which is now a vector, is:

$$S(\mu, \sigma) = \begin{pmatrix} \frac{\partial}{\partial \mu}\ell(\mu, \sigma \mid \mathbf{x}) \\ \frac{\partial}{\partial \sigma}\ell(\mu, \sigma \mid \mathbf{x}) \end{pmatrix} = \begin{pmatrix} \frac{n(\bar{\mathbf{x}} - \mu)}{\sigma^2} \\ -\frac{n}{\sigma} + \frac{n}{\sigma^3}\left[s^2 + (\bar{\mathbf{x}} - \mu)^2\right] \end{pmatrix}$$

Equating the score function to zero, we obtain a system of nonlinear equations in 2 variables. In order to solve this system, we start by solving the first entry of the score function:

$$\frac{n(\bar{\mathbf{x}} - \mu)}{\sigma^2} = 0,$$

which implies that the maximum with respect to μ does not depend upon the value of σ and $\widehat{\mu} = \bar{\mathbf{x}}$. Replacing this value into the second equation, we obtain

$$-\frac{n}{\sigma} + \frac{ns^2}{\sigma^3} = 0,$$

which has the solution $\widehat{\sigma}^2 = s^2$. That is, the MLE of the mean is the sample mean, while the MLE of the variance is the sample variance.

The observed information matrix I is a 2×2 matrix with entries:

$$\begin{aligned} i_{1,1} &= -\frac{\partial^2}{\partial \mu^2}\ell(\mu, \sigma \mid \mathbf{x}) = \frac{n}{\sigma^2}, \\ i_{2,2} &= -\frac{\partial^2}{\partial \sigma^2}\ell(\mu, \sigma \mid \mathbf{x}) = -\frac{n}{\sigma^2} + \frac{3n}{\sigma^4}\left[s^2 + (\bar{\mathbf{x}} - \mu)^2\right], \\ i_{1,2} &= -\frac{\partial^2}{\partial \sigma \partial \mu}\ell(\mu, \sigma \mid \mathbf{x}) = \frac{2n(\bar{\mathbf{x}} - \mu)}{\sigma^3}, \\ i_{2,1} &= i_{1,2}. \end{aligned}$$

Q1: Proof that this is a positive definite matrix (and also diagonal) when evaluated at $(\widehat{\mu}, \widehat{\sigma})$.

Q2: Other distributions have also been used to model heights such as the log-normal (which has positive support). Do you think that the assumption of normality is appropriate?

Example 2.7.6. Poisson Model: The MLE may not always exist Let $X_1, \ldots, X_n$ be *i.i.d.* $Poisson(\lambda)$, where $\lambda > 0$. Then, the likelihood function is:

$$L(\lambda \mid \mathbf{x}) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}.$$

Then, the log-likelihood function is:

$$\ell(\lambda \mid \mathbf{x}) \quad = \quad \log(\lambda)\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log(x_i!) - n\lambda,$$

$$= \quad n\bar{\mathbf{x}}\log(\lambda) - \sum_{i=1}^{n}\log(x_i!) - n\lambda$$

The score function is

$$S(\lambda) = \frac{n\bar{\mathbf{x}}}{\lambda} - n.$$

Equating this function to zero $S(\lambda) = 0$ we find the MLE $\widehat{\lambda} = \bar{\mathbf{x}}$.

Now, recall that the range of a Poisson random variable is the set $\{0, 1, \dots\}$. Note that the event $x_1 = \cdots = x_n = 0$ has positive probability, then if all of the observations are zero, the MLE $\widehat{\lambda} = 0$, which contradicts the assumption of the model $\lambda > 0$. This means that a sample of zeroes contains no information about the parameter λ.

Example 2.7.7. Multivariate Normal Distribution Let $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ be *i.i.d.* random vectors with Multivariate Normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and $p \times p$ covariance matrix Σ (assumed to be positive definite). The likelihood function is (omitting constants):

$$L(\boldsymbol{\mu}, \Sigma \mid \mathbf{x}_1, \dots, \mathbf{x}_n) \propto \frac{1}{\det(\Sigma)^{\frac{n}{2}}} \prod_{i=1}^{n} \exp\left[-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right].$$

Then, the log-likelihood is:

$$\ell(\boldsymbol{\mu}, \Sigma \mid \mathbf{x}_1, \dots, \mathbf{x}_n) = -\frac{n}{2}\log[\det(\Sigma)] - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}).$$

We will need the following results from vector and matrix calculus. Let $\mathbf{A}$ be a $p \times p$ symmetric matrix and $\mathbf{x}$ a $p \times 1$ vector. Then,

- $$\frac{\partial}{\partial \mathbf{x}}\mathbf{x}^\top \mathbf{A}\mathbf{x} = 2\mathbf{x}^\top \mathbf{A}.$$
- $$\mathbf{x}^\top \mathbf{A}\mathbf{x} = Tr(\mathbf{x}^\top \mathbf{A}\mathbf{x}) = Tr(\mathbf{x}\mathbf{x}^\top \mathbf{A}).$$

  where $Tr(\mathbf{A}) = \sum_i a_{ii}$ is the trace of the matrix.
- $$\frac{\partial}{\partial \mathbf{A}}\mathbf{x}^\top \mathbf{A}\mathbf{x} = \frac{\partial}{\partial \mathbf{A}}Tr(\mathbf{x}\mathbf{x}^\top \mathbf{A}) = \mathbf{x}\mathbf{x}^\top.$$
- $$\frac{\partial}{\partial \mathbf{A}}\log[det(\mathbf{A})] = (\mathbf{A}^{-\mathbf{1}})^\top.$$

Taking the derivative with respect to $\boldsymbol{\mu}$ we obtain (this requires vector-derivatives of a quadratic form):

$$\nabla_{\boldsymbol{\mu}}\ell(\boldsymbol{\mu}, \Sigma \mid \mathbf{x}_1, \dots, \mathbf{x}_n) = \sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}.$$

Equating this derivative to zero (and using that Σ is a positive definite matrix) we obtain $\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$. Now, we rewrite the log-likelihood as follows

$$\ell(\boldsymbol{\mu}, \Sigma \mid \mathbf{x}_1, \dots, \mathbf{x}_n) = \frac{n}{2}\log[\det(\Sigma^{-1})] - \frac{1}{2}\sum_{i=1}^{n} Tr\left((\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top \Sigma^{-1}\right).$$

Taking the derivative with respect to Σ^{-1}, we obtain:

$$\frac{\partial}{\partial \Sigma^{-1}} \ell(\boldsymbol{\mu}, \Sigma \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top.$$

Equating this derivative to zero we obtain

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})^\top.$$

Example 2.7.8. Normal Linear Regression Model

Suppose that you observe n responses $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ (also known as dependent variables), associated to n individuals. Suppose that, for each individual, you know p characteristics associated to them $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$ (also known as covariates or independent variables). The aim here is to explain the responses y_i, $i = 1, \ldots, n$, based on their corresponding individual characteristics $\mathbf{x}_i$. A possible way for doing so is to use *linear regression*, which is formulated as follows:

$$y_i = \alpha + \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i,$$

where $\mathbf{x}_i^\top$ denotes the transposed vector $\mathbf{x}_i$; $\alpha \in \mathbb{R}$ is an unknown parameter, and represents the intercept of the linear model; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is a vector of unknown parameters (also known as *regression coefficients*), and represent slopes of each covariate $\mathbf{x}_i$. Then, $\mathbf{x}_i^\top \boldsymbol{\beta}$ is the inner product

$$\mathbf{x}_i^\top \boldsymbol{\beta} = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p,$$

and ϵ_i are random residual errors which are assumed to be *i.i.d.* and to be distributed according to a normal distribution with mean zero and unknown variance σ^2. This is, $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. In matrix notation, this model can be written as:

$$\mathbf{y} = \mathcal{D}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where

$$\mathcal{D} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix},$$

which is known as the *design matrix*, and

$$\boldsymbol{\theta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},$$

and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$. We will assume that the Rank$(\mathcal{D}) = p + 1$. Recall that this corresponds to a generalised linear model [4] with mean

$$E[Y_i] = \mu_i = \mathcal{D}\boldsymbol{\theta}; \quad Y_i \sim N(\mu_i, \sigma^2).$$

By noting that $\epsilon_i = y_i - \alpha - \mathbf{x}_i^\top \boldsymbol{\beta}$ and recalling that these errors are normally distributed, the likelihood function is

$$L(\boldsymbol{\theta}, \sigma \mid \mathbf{x}, \mathbf{y}) \quad = \quad \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(y_i - \alpha - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2} \right\}$$

$$\propto \prod_{i=1}^{n} \frac{1}{\sigma} \exp\left\{ -\frac{(y_i - \alpha - \mathbf{x}_i^\top \boldsymbol{\beta})^2}{2\sigma^2} \right\}.$$

The log-likelihood function is (omitting constant values)

$$\begin{aligned} \ell(\boldsymbol{\theta}, \sigma \mid \mathbf{x}, \mathbf{y}) &= -n\log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \mathbf{x}_i^\top \boldsymbol{\beta})^2 \\ &= -n\log(\sigma) - \frac{1}{2\sigma^2} (\mathbf{y} - \mathcal{D}\boldsymbol{\theta})^\top (\mathbf{y} - \mathcal{D}\boldsymbol{\theta}). \end{aligned}$$

The entries of the score function are:

$$\begin{aligned} \frac{\partial}{\partial \sigma} \ell(\boldsymbol{\theta}, \sigma \mid \mathbf{x}, \mathbf{y}) &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} (\mathbf{y} - \mathcal{D}\boldsymbol{\theta})^\top (\mathbf{y} - \mathcal{D}\boldsymbol{\theta}), \\ \frac{\partial}{\partial \alpha} \ell(\boldsymbol{\theta}, \sigma \mid \mathbf{x}, \mathbf{y}) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \mathbf{x}_i^\top \boldsymbol{\beta}), \\ \frac{\partial}{\partial \boldsymbol{\beta}_k} \ell(\boldsymbol{\theta}, \sigma \mid \mathbf{x}, \mathbf{y}) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \mathbf{x}_i^\top \boldsymbol{\beta}) x_{ik}. \end{aligned}$$

By equating the entries of the score function to zero, we obtain a system of $p + 2$ non-linear equations in terms of the variables $\begin{pmatrix} \alpha \\ \boldsymbol{\beta} \\ \sigma \end{pmatrix}$. The first equation implies that

$$\widehat{\sigma}^2(\boldsymbol{\theta}) = \frac{1}{n} (\mathbf{y} - \mathcal{D}\boldsymbol{\theta})^\top (\mathbf{y} - \mathcal{D}\boldsymbol{\theta}).$$

Thus, the MLE of σ depends on the MLEs of $\boldsymbol{\theta}$ (*i.e.* α and $\boldsymbol{\beta}$). On the other hand, from the second and third system of equations, we can see that the solutions for α and $\boldsymbol{\beta}$ do not depend on the value of σ (check this by clearing/reflecting σ from these equations). In order to solve these equations, it is easier to write them in matrix form:

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}, \sigma \mid \mathbf{x}, \mathbf{y}) = \frac{1}{\sigma^2} \mathcal{D}^\top (\mathbf{y} - \mathcal{D}\boldsymbol{\theta}) = \mathbf{0}.$$

This expression can also be obtained if you recall the rules of vector derivatives. Solving this equation using basic linear algebra results lead to:

$$\widehat{\boldsymbol{\theta}} = (\mathcal{D}^\top \mathcal{D})^{-1} \mathcal{D}^\top \mathbf{y}.$$

From the matrix form of the score function, we can easily obtain the observed Information matrix:

$$i(\boldsymbol{\theta}) = \begin{pmatrix} i_{\boldsymbol{\theta},\boldsymbol{\theta}} & i_{\sigma,\boldsymbol{\theta}} \\ i_{\boldsymbol{\theta},\sigma} & i_{\sigma,\sigma} \end{pmatrix}.$$

After some algebra, we can show that

$$\begin{aligned} i_{\boldsymbol{\theta},\boldsymbol{\theta}} &= -\nabla^2_{\boldsymbol{\theta},\boldsymbol{\theta}} \, \ell(\boldsymbol{\theta}, \sigma \mid \mathbf{x}, \mathbf{y}) = \frac{\mathcal{D}^\top \mathcal{D}}{\sigma^2}, \\ i_{\sigma,\sigma} &= -\frac{\partial^2}{\partial \sigma^2} \, \ell(\boldsymbol{\theta}, \sigma \mid \mathbf{x}, \mathbf{y}) = -\frac{n}{\sigma^2} + \frac{3}{\sigma^4} (\mathbf{y} - \mathcal{D}\boldsymbol{\theta})^\top (\mathbf{y} - \mathcal{D}\boldsymbol{\theta}), \\ i_{\sigma,\boldsymbol{\theta}} &= -\nabla^2_{\sigma,\boldsymbol{\theta}} \, \ell(\boldsymbol{\theta}, \sigma \mid \mathbf{x}, \mathbf{y}) = \frac{2}{\sigma^3} \mathcal{D}^\top (\mathbf{y} - \mathcal{D}\boldsymbol{\theta}), \\ i_{\boldsymbol{\theta},\sigma} &= i_{\sigma,\boldsymbol{\theta}}^\top. \end{aligned}$$

Q: Proof that this matrix is block-diagonal and positive definite when evaluated at $(\widehat{\boldsymbol{\theta}}, \widehat{\sigma})$.

Connection with Least Squares

As we mentioned, the MLE of $\widehat{\boldsymbol{\theta}}$ does not depend on the value of the variance σ^2. Thus, if we restrict our attention on the optimisation problem with respect to $\boldsymbol{\theta}$:

$$\max_{\boldsymbol{\theta}} -(\mathbf{y} - \mathcal{D}\boldsymbol{\theta})^\top(\mathbf{y} - \mathcal{D}\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}}(\mathbf{y} - \mathcal{D}\boldsymbol{\theta})^\top(\mathbf{y} - \mathcal{D}\boldsymbol{\theta}).$$

This corresponds to finding the values of $\boldsymbol{\theta} = (\alpha, \boldsymbol{\theta})$ that minimises the sum of square errors

$$\text{argmin}_{\alpha,\beta} \sum_{i=1}^{n}(y_i - \alpha - \mathbf{x}_i^\top \boldsymbol{\beta})^2$$

This method is known a "least squares", and will be studied later. However, we can see that in the case of the linear regression model with normal errors, the maximum likelihood estimation method is equivalent to the least squares method.

So far, we have shown examples where the maximum likelihood estimation method leads to closed form (in terms of elementary functions) solutions. In general, MLEs cannot always be obtained in closed form, and the use of numerical methods is necessary to approximate the MLE. A powerful numerical method is Newton's method, which, interestingly, is closely linked with the score and information functions.

Remark 2.7.5. Newton's Method in Optimisation

Let $f : \mathbb{R}^p \to \mathbb{R}$ be a twice continuously differentiable function with positive definite Hessian matrix $\nabla^2 f$. Suppose that f has a unique maximum (minimum) at $f(\mathbf{x}^*)$. Then, for an initial point $\mathbf{x}_0$, the sequence:

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \left[\nabla^2 f(\mathbf{x}_n)\right]^{-1} \nabla f(\mathbf{x}_n), \quad n \geq 0,$$

converges to $\mathbf{x}^*$.

Interestingly, if f is the log-likelihood function, then the iterations in Newton's method involve the use of the score function (the gradient) and the observed Information matrix (the Hessian matrix), which emphasises the importance of these functions in maximum likelihood estimation. The following example illustrates a very useful model where the MLE cannot be found in closed form, and the use of numerical methods is necessary to calculate the MLE.

Definition 27. Let $\widehat{\boldsymbol{\theta}}$ be a point estimator of θ, the parameter of a distribution $f(\cdot; \boldsymbol{\theta})$. The *fitted distribution* or *fitted model* is the distribution obtained by replacing $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. This is, $f(\cdot; \widehat{\boldsymbol{\theta}})$.

The fitted model is often used to visualise the goodness of fit of a statistical model (how well the model captures the features of the data). It is important to notice that parametric models contain some assumptions. For instance, the normal distribution assumes that the data are symmetrically distributed around the mean. The linear regression model assumes that the relationship between the response variable and the covariates is linear. Thus, it is important to check how reasonable those assumptions are for the data at hand. There are formal tools to assess the fit of a model, and they are called *goodness of fit tests*. An alternative is to use a (informal) visual tool, and to judge, by eye, the fit of the model. For example, we can plot the fitted model and a histogram of the data in the same graph in order to assess the fit of the model.

The following example illustrates how to visualise the fit of a logistic regression model used in drug development (dose-response model).

Example 2.7.9. Logistic Regression Model

Let Y_i, $i = 1, \ldots, m$, be binomial random variables associated to n_i trials with π_i probability of success and $1 - \pi_i$ probability of failure. Then, their pmfs are:

$$P(Y_i = y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}, \quad y_i \in \{0, 1, \ldots, n_i\}.$$

Let $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^{p+1}$, $i = 1, \ldots, n$, be a vector of continuous (although discrete can be considered as well) measurements corresponding to covariates and dummy variables corresponding to factor levels and $\boldsymbol{\theta}$ be a parameter vector. This vector of covariates are assumed to include an intercept, by default. That is, $x_{i0} = 1$, for all i. Define the relationship

$$\operatorname{logit} \pi_i = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}_i^\top \boldsymbol{\theta}.$$

Equivalently

$$\pi_i = \frac{e^{\mathbf{x}_i^\top \boldsymbol{\theta}}}{1 + e^{\mathbf{x}_i^\top \boldsymbol{\theta}}} = F\left(\mathbf{x}_i^\top \boldsymbol{\theta}\right),$$

where F is the logistic cdf (see Preliminary Material). The inverse of F, F^{-1}, is called the link function, which in this case corresponds to the *logit* function. From the previous expression, you can observe that the idea is to link the probability of success of the binomial variables Y_i with the logistic cdf F of the value $\mathbf{x}_i^\top \boldsymbol{\theta}$. This equation also defines the popular *logistic regression model*, which is a model that it used to explain binary or binomial outcomes with respect to some available characteristics for each trial.

The likelihood and log-likelihood functions of $\boldsymbol{\theta}$ are:

$$\begin{aligned} L(\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{y}) &= \prod_{i=1}^{n} \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}, \\ \ell(\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{y}) &= \sum_{i=1}^{n} \left[\log \binom{n_i}{y_i} + y_i \log(\pi_i) + (n_i - y_i) \log(1 - \pi_i) \right], \end{aligned}$$

with $\pi_i = \dfrac{e^{\mathbf{x}_i^\top \boldsymbol{\theta}}}{1 + e^{\mathbf{x}_i^\top \boldsymbol{\theta}}}$. Let us define $u_i = e^{\mathbf{x}_i^\top \boldsymbol{\theta}}$. The log-likelihood can be simplified as follows

$$\ell(\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \left(\log \binom{n_i}{y_i} + y_i \log(u_i) - n_i \log(1 + u_i) \right), \tag{2.7.2}$$

Note now that the gradient of u_i with respect to $\boldsymbol{\theta}$ is given by $\nabla_{\boldsymbol{\theta}} u_i = \mathbf{x}_i u_i$. Then, the entries of the score function are given by

$$\begin{aligned} S(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \, \ell(\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{y}) &= \sum_{i=1}^{n} (y_i \mathbf{x}_i - n_i \pi_i \mathbf{x}_i) \\ &= \sum_{i=1}^{n} (y_i - n_i \pi_i) \, \mathbf{x}_i \end{aligned}$$

Thus, the MLE of $\boldsymbol{\theta}$ is the solution to the system of non-linear equations:

$$\sum_{i=1}^{n} (y_i - n_i \pi_i) \, \mathbf{x}_i = \mathbf{0}.$$

Unfortunately, this system of equations does not admit a closed-form solution. However, it has been shown that the log-likelihood function of $\boldsymbol{\theta}$ has a unique maximum, under some mild conditions on the covariates $\mathbf{x}_i$ (which are not studied in this course). Then, as we know that the maximum exists, we

can use numerical methods to calculate it for a specific sample. In particular, one of the most popular methods used to calculate this MLE is Newton's method. Thus, in order to implement Newton's method, we need to calculate the Information matrix. Using the simplified version of the log-likelihood, we can quickly obtain:

$$i(\boldsymbol{\theta}) = -\nabla^2_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} n_i \pi_i (1-\pi_i) \mathbf{x}_i \mathbf{x}_i^\top. \tag{2.7.3}$$

In the previous example, instead of using the logistic distribution F, one could employ any other cdf, such as the normal distribution. If F is the normal distribution, often denoted $F = \Phi$, then the model is known as the Probit Regression Model.

The MLE has an appealing property referred to as *Invariance Property* [13].

Theorem 2.7.1. Invariance Property. *Consider the likelihood function $L(\boldsymbol{\theta} \mid \mathbf{x})$. Suppose that the MLE of $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ exists and is denoted as $\widehat{\boldsymbol{\theta}}$. Let $g : \mathbb{R}^p \to \mathbb{R}^q$, $q \leq p$, be a function. Then, the MLE of the parametric function $g(\boldsymbol{\theta})$ is given by $g(\widehat{\boldsymbol{\theta}})$.*

Example 2.7.10. An experiment was conducted to evaluate the toxicity of gaseous carbon disulphide on flour beetle. Table 2.1 shows the numbers of beetles n_i that were exposed to gaseous carbon disulphide at concentrations x_i (Dose, expressed in $\log_{10}$ mg/L), as well as the numbers of beetles y_i dead after five hours of exposure. The data comes from an article by Bliss (1935). Thus, the responses y_i are binomial

Dose x_i	n_i	y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Table 2.1: Beetle mortality data.

realisations with n_i trials at concentrations x_i. This kind of experiment can be modelled using a logistic regression model with probabilities

$$\pi_i = \frac{\exp(\theta_1 + \theta_2 x_i)}{1 + \exp(\theta_1 + \theta_2 x_i)},$$

equivalently (please, check),

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \theta_1 + \theta_2 x_i.$$

This is a very useful model known as a *Dose-Response model*, and it still enjoys some popularity in real applications. The corresponding likelihood function is

$$L(\theta_1, \theta_2 \mid \mathbf{x}, \mathbf{y}) = \prod_{i=1}^{8} \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i},$$

and the log-likelihood (see (2.7.2)):

$$\ell(\theta_1, \theta_2 \mid \mathbf{x}, \mathbf{y}) \;=\; \sum_{i=1}^{n} \left(\log \binom{n_i}{y_i} + y_i(\theta_1 + \theta_2 x_i) - n_i \log(1 + \exp(\theta_1 + \theta_2 x_i) \right).$$

The entries of the score function are:

$$\begin{aligned} S_1(\theta_1,\theta_2) &= \frac{\partial l}{\partial \theta_1} = \sum_i (y_i - n_i\pi_i), \\ S_2(\theta_1,\theta_2) &= \frac{\partial l}{\partial \theta_2} = \sum_i x_i(y_i - n_i\pi_i). \end{aligned}$$

The observed information matrix is (see (2.7.3))

$$i(\theta_1,\theta_2) = \begin{pmatrix} \sum_i n_i\pi_i(1-\pi_i) & \sum_i n_i x_i \pi_i(1-\pi_i) \\ \sum_i n_i x_i\pi_i(1-\pi_i) & \sum_i n_i x_i^2 \pi_i(1-\pi_i) \end{pmatrix}.$$

As discussed before, the MLEs of (θ_1,θ_2) cannot be obtained in closed-form. However, using the score function and the information matrix, we can calculate the MLEs numerically. For instance, we can select the initial value $(\theta_1^{(0)},\theta_2^{(0)}) = (0,0)$, and run N iterations (for N large, say $N = 100$) of the sequence:

$$(\theta_1^{(j)},\theta_2^{(j)}) = (\theta_1^{(j-1)},\theta_2^{(j-1)}) + i^{-1}(\theta_1^{(j-1)},\theta_2^{(j-1)}) \cdot (S_1(\theta_1^{(j-1)},\theta_2^{(j-1)}), S_2(\theta_1^{(j-1)},\theta_2^{(j-1)})),$$

for $j = 1,\ldots,N$. After running 100 iterations of this sequence we obtain $(\widehat{\theta}_1,\widehat{\theta}_2) = (-60.72, 34.27)$.

Exercise. Calculate the estimated probabilities $\widehat{\pi}_i$ by replacing the value of the MLEs of (θ_1,θ_2) in the corresponding expressions. In addition, calculate $n_i\widehat{\pi}_i$ and compare these values to the observed values y_i.

- **Odds ratio**. From previous calculations, we know that the odds

  $$\frac{\pi_i}{1-\pi_i} = \exp(\theta_1 + \theta_2 x_i).$$

  Then, if we wish to compare the odds associated to $x_i + 1$ and the odds of x_i (increasing x_i by 1 unit), we can calculate the odds ratio

  $$OR = \frac{\exp(\theta_1 + \theta_2(x_i+1))}{\exp(\theta_1 + \theta_2 x_i)} = \exp(\theta_2).$$

  This number is often reported in dose-response analyses. In order to estimate OR, we can appeal to the invariance property of the MLEs. In this case, the OR represents a function (reparameterisation) of the parameter θ_2. Thus, the MLE of OR is simply $OR = \exp(\widehat{\theta}_2) = \exp(34.27)$.

- **Median Lethal Dose (LD50)**. Another quantity of interest in dose-response is the LD50, which is the concentration required to kill half of the members of the tested population. The LD50 is the solution to:

  $$\pi_{0.5} = 0.5 = \frac{\exp(\theta_1 + \theta_2 LD_{50})}{1 + \exp(\theta_1 + \theta_2 LD_{50})}.$$

  Then,

  $$\log\left(\frac{0.5}{1-0.5}\right) = 0 = \theta_1 + \theta_2 LD_{50}.$$

  Finally, $LD_{50} = -\dfrac{\theta_1}{\theta_2}$. Again, this is a function of the parameters (θ_1,θ_2). Using again the invariance of the MLEs, the MLE of LD_{50} is $\widehat{LD}_{50} = -\dfrac{\widehat{\theta}_1}{\widehat{\theta}_2}$ =1.77.

This example is implemented in the R (programming language) at the following link:

`http://rpubs.com/FJRubio/beetle`

Asymptotic properties of MLEs

Given that an estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{X})$ (or $T(\mathbf{X})$) is a function of the sample $\mathbf{X} = (X_1, \ldots, X_n)^\top$, it follows that it depends on $n \geq 1$ random variables. The dependency of $\widehat{\boldsymbol{\theta}}$ (or T) on $\mathbf{X}$ is usually omitted, and it is simply denoted as $\widehat{\boldsymbol{\theta}}_n$ (or $T(\mathbf{X}) = T_n$). Thus, it is of interest to analyse the behaviour of the random sequence $\widehat{\boldsymbol{\theta}}_n$, as $n \to \infty$. The study of estimators (or statistics in general) as $n \to \infty$ is known as *asymptotic theory* or *large sample theory*. Similar to the preliminary material, the asymptotic behaviour of the sequence $\widehat{\boldsymbol{\theta}}_n$ can be characterised using the different modes of convergence, as $\widehat{\boldsymbol{\theta}}_n$ is a random variable or random vector itself.

Suppose that $X_j \overset{iid}{\sim} F(\cdot; \boldsymbol{\theta}_0)$ for some fixed value $\boldsymbol{\theta}_0 \in \Theta \subset \mathbb{R}^p$. Let $\widehat{\boldsymbol{\theta}}$ be an estimator of $\boldsymbol{\theta}_0$ (for instance, a MLE). Then,

(i) The sequence $\widehat{\boldsymbol{\theta}}_n$ is said to be *consistent* (or weakly consistent) if $\widehat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$.

(ii) The sequence $\widehat{\boldsymbol{\theta}}_n$ is said to be *strongly consistent* if $\widehat{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \boldsymbol{\theta}_0$.

(iii) An estimator $\widehat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}_0$ is said to be asymptotically normal if $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \Sigma_\theta)$, as $n \to \infty$, for a non-negative definite matrix Σ_θ. Later, we will see that the matrix Σ_θ, which we can interpret as the *asymptotic variance*, is related to the expected Fisher information matrix (FIM).

Next, we will establish asymptotic results for the MLE in the case $p = 1$ under the following assumptions.

A1 $f(x; \theta)$, $\theta \in \Theta \subset \mathbb{R}$, $x \in \mathcal{D} \subset \mathbb{R}$, is **Identifiable**. That is, if $f(x; \theta_1) = f(x; \theta_2)$, for all values of x, this implies that $\theta_1 = \theta_2$.

A2 The support of $f(x; \theta)$, $\text{supp}(f) = \{x \in \mathcal{D} : f(x) > 0\}$, is the same for all values of θ.

A3 The observations $\mathbf{X} = (X_1, \ldots, X_n)$ are *i.i.d.* with probability density function $f(x_i; \theta_0)$.

A4 θ_0 is an interior point of Θ. This means that there is an open subset $O \subset \Theta$ that contains the true value θ_0.

The next results shows that the likelihood function evaluated at the true value θ_0 is bigger than the likelihood function evaluated at any other value in the limit when the sample size grows to infinity. This is,

Theorem 2.7.2. *If A1–A3 hold, then, for any* $\theta \neq \theta_0$

$$\lim_{n \to \infty} P_{\theta_0} \left(L(\theta_0 | \mathbf{X}) > L(\theta | \mathbf{X}) \right) = 1. \tag{2.7.4}$$

Proof. Note first that $L(\theta_0|\mathbf{X}) > L(\theta|\mathbf{X})$ is equivalent to $L(\theta|\mathbf{X})/L(\theta_0|\mathbf{X}) < 1$. Taking logarithms on both sides of this inequality, we obtain

$$\ell(\theta|\mathbf{X}) - \ell(\theta_0|\mathbf{X}) < 0.$$

which is equivalent to (dividing by n)

$$K_n(\theta, \theta_0) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{f(X_i; \theta)}{f(X_i; \theta_0)} \right) < 0.$$

Now, by the law of large numbers,

$$K_n(\theta, \theta_0) = \frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{f(X_i; \theta)}{f(X_i; \theta_0)} \right) \xrightarrow{P} E_{\theta_0} \left[\log \left(\frac{f(X; \theta)}{f(X; \theta_0)} \right) \right] \coloneqq K(\theta, \theta_0) \quad as \quad n \to \infty.$$

By Jensen's inequality, and using that the logarithm function is strictly concave

$$E_{\theta_0}\left[\log\left(\frac{f(X;\theta)}{f(X;\theta_0)}\right)\right] \leq \log E_{\theta_0}\left[\frac{f(X;\theta)}{f(X;\theta_0)}\right] = 0,$$

with equality only if $\theta = \theta_0$. Thus, we have that $K_n(\theta, \theta_0) \xrightarrow{P} K(\theta, \theta_0) < 0$, for $\theta \neq \theta_0$, which is equivalent to proving (2.7.4).

This result provides an intuitive justification for using the MLE $\widehat{\theta}$ as a point estimator of θ_0, as this maximiser should be closer to the true value θ_0 for large enough n. The following Theorem formalises this idea.

Theorem 2.7.3. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sequence of random variables satisfying A1–A4, and suppose that for almost all x, $f(x;\theta)$ is differentiable with respect to $\theta \in O$, with derivative $f'(x;\theta)$. Then, with probability tending to 1 as $n \to \infty$, the likelihood equation*

$$\frac{\partial}{\partial\theta}\ell(\theta|\mathbf{x}) = 0,$$

or, equivalently, the equation

$$\ell'(\theta|\mathbf{x}) = \sum_{i=1}^{n} \frac{f'(x_i;\theta)}{f(x_i;\theta)} = 0,$$

has a root $\widehat{\theta}_n = \widehat{\theta}_n(x_1, \ldots, x_n)$ such that $\widehat{\theta}_n(X_1, \ldots, X_n)$ tends to the true value θ_0 in probability.

Proof. Let a be small enough so that $(\theta_0 - a, \theta_0 + a) \subset O$, and define

$$C_n = \{\mathbf{x} : \ell(\theta_0|\mathbf{x}) > \ell(\theta_0 - a|\mathbf{x}) \ \textit{and} \ \ell(\theta_0|\mathbf{x}) > \ell(\theta_0 + a|\mathbf{x})\}.$$

By Theorem 2.7.2, $P(C_n) \to 1$ as $n \to \infty$. Thus C_n is not an empty set for large enough n. Now, for $\mathbf{x} \in C_n$ and by Rolle's Theorem (Calculus), there exists a value $\theta_0 - a < \widehat{\theta}_n < \theta_0 + a$ at which $\ell(\theta)$ has a local maximum, so that $\ell'(\widehat{\theta}_n) = 0$. Hence, for any $a > 0$ sufficiently small, there exists a sequence $\widehat{\theta}_n = \widehat{\theta}(a)$ of roots such that, as $n \to \infty$,

$$P_{\theta_0}\left[|\widehat{\theta}_n - \theta_0| < a\right] \to 1.$$

Thus, we only need to prove that we can determine such sequence, which does not depend on a (as in the definition of convergence in probability).

Let $\widehat{\theta}_n^*$ be the root closest to θ_0, then

$$P_{\theta_0}\left[|\widehat{\theta}_n - \theta_0| < a\right] \leq P_{\theta_0}\left[|\widehat{\theta}_n^* - \theta_0| < a\right].$$

Corollary 2.7.1. *Under the assumptions of Theorem 2.7.3, if the likelihood function has a unique root $\widehat{\theta}_n$ for each n and all $\mathbf{x}$, then $\{\widehat{\theta}_n\}$ is a consistent sequence of estimators of θ_0. If, in addition, the parameter space is an open interval (not necessarily finite), then with probability tending to 1, $\{\widehat{\theta}_n\}$ maximises the likelihood, that is, $\{\widehat{\theta}_n\}$ is the MLE, which is therefore consistent.*

Theorem 2.7.3 is a fundamental result in Statistical Inference, and thus it is important to understand what the theorem says, as well as what the theorem does not say.

- This theorem does not guarantee the existence or uniqueness of a root $\widehat{\theta}_n$ for a given sample $\mathbf{x}$. This has to be checked directly.
- This theorem establishes the existence of a sequence of local maxima of the likelihood function which is consistent.

- This theorem does not establish the existence of a consistent estimator since, as the true value θ_0 is unknown, the data do not tell us which root to choose in order to obtain a consistent sequence.
- If the root is unique, as in our previous examples, the previous theorem implies that the MLE is consistent.

In the examples we have studied so far, the MLE is unique and it is possible to show that the corresponding distributions satisfy the "regularity conditions" in the previous theorem, then we can appeal to this result directly in order to show consistency of the MLE.

The following result characterises the asymptotic distribution of the maximum likelihood estimator. It represents a fascinating general result that continues to be studied in modern statistical models. In order to establish this result, we need the following additional assumptions.

A5 For every x, the density $f(x;\theta)$ is three times differentiable with respect to θ, and the third derivative is continuous in θ.

A6 The integral $\int f(x;\theta)dx$ can be differentiated three times under the integral sign.

A7 The Fisher information $I(\theta)$ defined satisfies $0 < I(\theta) < \infty$.

A8 For any given $\theta_0 \in \Theta$, there exists a positive number c and a function $M(x)$ (both of which may depend on θ_0) such that

$$\left|\frac{\partial^3}{\partial\theta^3}\log f(x;\theta)\right| \leq M(x),$$

for all $x \in \mathcal{D}$, $\theta_0 - c < \theta < \theta_0 + c$, with

$$E_{\theta_0}[M(X)] = \int M(x)f(x;\theta_0)dx < \infty.$$

Theorem 2.7.4. *Under the assumptions A1-A8, any consistent sequence $\widehat{\theta}_n = \widehat{\theta}_n(X_1,\ldots,X_n)$ of roots of the likelihood function satisfies*

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right).$$

Proof. For any fixed $\mathbf{x}$, expand $\ell'(\widehat{\theta}_n)$ about θ_0 using a Taylor's series expansion of order one, and expressing the remaining orders using the Taylor's Remainder Theorem:

$$\ell'(\widehat{\theta}_n) = \ell'(\theta_0) + (\widehat{\theta}_n - \theta_0)\ell''(\theta_0) + \frac{1}{2}(\widehat{\theta}_n - \theta_0)^2\ell'''(\theta_n^*),$$

where θ_n^* lies between θ_0 and $\widehat{\theta}_n$. By assumption, the left side is zero, so that

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = \frac{(1/\sqrt{n})\ell'(\theta_0)}{-(1/n)\ell''(\theta_0) - (1/2n)(\widehat{\theta}_n - \theta_0)\ell'''(\theta_n^*)},$$

where it should be remembered that $\ell(\theta)$, $\ell'(\theta)$, and so on are functions of $\mathbf{X}$ as well as θ. We shall show that

(i)
$$\frac{1}{\sqrt{n}}\ell'(\theta_0) \xrightarrow{d} N(0, I(\theta_0)), \tag{2.7.5}$$

(ii) that
$$-\frac{1}{n}\ell''(\theta_0) \xrightarrow{P} I(\theta_0), \tag{2.7.6}$$

(iii) and that

$$\frac{1}{n}\ell'''(\theta_n^*), \tag{2.7.7}$$

is bounded in probability.

The result will then follow by Slutsky's theorem (see preliminary material).

(i) This follows from the fact that

$$\begin{aligned}\frac{1}{\sqrt{n}}\ell'(\theta_0) &= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{f'(X_i;\theta_0)}{f(X_i;\theta_0)} - E_{\theta_0}\left(\frac{f'(X_i;\theta_0)}{f(X_i;\theta_0)}\right)\right] \\ &= \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{f'(X_i;\theta_0)}{f(X_i;\theta_0)} - E_{\theta_0}\left(\frac{f'(X_i;\theta_0)}{f(X_i;\theta_0)}\right)\right].\end{aligned}$$

This relationship follows by noting that the second term is zero, by assumption. Define $S_n = \sum_{i=1}^{n}\frac{f'(X_i;\theta_0)}{f(X_i;\theta_0)}$, and note that $\mu = E(S_n) = 0$, and that $Var\left(\frac{f'(X_i;\theta_0)}{f(X_i;\theta_0)}\right) = I(\theta_0)$ (the Fisher information). Then, by the Central Limit Theorem (see Preliminary Material), we obtain

$$\frac{1}{\sqrt{n}}\ell'(\theta_0) = \sqrt{n}\,(S_n - \mu) \xrightarrow{d} N(0, I(\theta_0)).$$

(ii) Now,

$$\begin{aligned}-\frac{1}{n}\ell''(\theta_0) &= \frac{1}{n}\sum_{i=1}^{n}\frac{f'^2(X_i;\theta_0) - f(X_i;\theta_0)f''(X_i;\theta_0)}{f^2(X_i;\theta_0)} \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{f'^2(X_i;\theta_0)}{f^2(X_i;\theta_0)} - \frac{1}{n}\sum_{i=1}^{n}\frac{f''(X_i;\theta_0)}{f(X_i;\theta_0)}.\end{aligned}$$

By the law of large numbers, this converges in probability to (its expectation)

$$I(\theta_0) - E_{\theta_0}\left[\frac{f''(X;\theta_0)}{f(X;\theta_0)}\right] = I(\theta_0).$$

The second term is zero by assumption.

(iii) Finally, note that

$$\frac{1}{n}\ell'''(\theta_n^*) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3}{\partial\theta^3}\log f(X_i;\theta).$$

Then, by assumption, and by applying the triangle inequality sequentially,

$$\begin{aligned}\left|\frac{1}{n}\ell'''(\theta_n^*)\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^3}{\partial\theta^3}\log f(X_i;\theta)\right| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial^3}{\partial\theta^3}\log f(X_i;\theta)\right| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}M(X_i).\end{aligned}$$

Finally, by the law of large numbers $\frac{1}{n}\sum_{i=1}^{n}M(X_i) \xrightarrow{P} E[M(X)] < \infty$. Then,

$$-(1/2n)(\widehat{\theta}_n - \theta_0)\ell'''(\theta_n^*) \xrightarrow{P} 0.$$

Combining the above results and using Slutsky's theorem, it follows that

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right).$$

Thus, in order to prove consistency and asymptotic normality of the MLE in any model f, we need to check that the assumptions A1–A8 are satisfied. Checking that these conditions hold can be still challenging and typically requires lengthy algebraic calculations. Nonetheless, these results offer very general conditions that guarantee good asymptotic properties of the MLE. In fact, if a MLE is consistent and asymptotically normal, it is called an *efficient estimator*.

Example 2.7.11. A simple example of these results is the case where $X_1, \ldots, X_n$ are *i.i.d.* $N(\mu_0, 1)$. One way of showing consistency and asymptotic normality consists of checking that the normal distribution satisfies the assumptions in the previous theorems. A simpler way is to use the properties of the normal distribution as follows. The MLE of μ is the sample mean $\widehat{\mu} = \bar{X} \sim N(\mu_0, \sigma^2/n)$. Thus, as $n \to \infty$, $\widehat{\mu} \xrightarrow{P} \mu_0$ by the law of large numbers (consistency), and $\sqrt{n}(\widehat{\mu} - \mu_0) \sim N(0, \sigma)$ by the properties of the normal distribution (asymptotic normality, as well as normality for finite n).

Example 2.7.12. We will now see an example with all the properties studied previously. Let $X_1, \ldots, X_n$ be an *i.i.d.* sample from an Exponential distribution with mean $\beta_0 > 0$. The corresponding density function is:

$$f(x; \beta) = \frac{1}{\beta} e^{-\frac{x}{\beta}}, \quad x > 0.$$

Then, the likelihood and log-likelihood functions are:

$$\begin{aligned} L(\beta|\mathbf{x}) &= \frac{1}{\beta^n} e^{-\frac{n\bar{x}}{\beta}}, \\ \ell(\beta|\mathbf{x}) &= -n\log(\beta) - \frac{n\bar{x}}{\beta}. \end{aligned}$$

It is easy to show that the MLE is $\widehat{\beta} = \bar{x}$. First, we can show that $E[\widehat{\beta}] = \beta$, then the MLE is unbiased. We can also check that all the conditions A1-A8 are satisfied. Thus, the MLE is consistent and asymptotically normal. In order to obtain the asymptotic variance, we need the Fisher information (Exercise):

$$I(\beta) = \int_0^\infty \left(\frac{\partial}{\partial\beta} \log f(x; \beta)\right)^2 f(x; \beta) = \frac{1}{\beta^2}.$$

Then, $\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{d} N(0, \beta^2)$, as $n \to \infty$. Moreover, $Var(\widehat{\beta}) = \dfrac{\beta^2}{n}$, which coincides with the inverse of the Fisher information matrix. Thus, the CRLB is attained with equality. This shows that the Exponential distribution has good asymptotic properties, which is the reason why it is popular in applications.

Fake data example. Suppose that we use $n = 30$ AAA batteries of a certain brand until the end of their life (that is, until they can no longer power a device requiring these batteries). We record the exact time at which the batteries fail, which are $\mathbf{x} =$ (1.2, 1.31, 0, 1.3, 0.94, 0.55, 2.88, 1.15, 0.27, 5.49, 0.72, 0.91, 0.4, 1.91, 0.64, 0.78, 2.11, 0.13, 0.37, 0.24, 1.77, 0.76, 1.07, 1.3, 0.25, 2.44, 0.06, 0.89, 1, 0.14)-years. Suppose that these numbers are distributed according to an Exponential distribution with unknown mean β. Then, the MLE of β is $\widehat{\beta} = \bar{\mathbf{x}} = 1.099$. Thus, we can conclude that the mean battery life is roughly 1.1 years. Given that we know that this is a good estimation, we can also try to understand other properties of the battery life. For example, the $5\%, 50\%, 95\%$ quantiles of the exponential distribution with mean 1.1 are $(0.056, 0.762, 3.295)$ years, respectively. These represents the probabilities of a battery failing before the aforementioned times. These probabilities can even be presented in a graph by reporting the CDF $F(x; \widehat{\beta})$:

Consistency and asymptotic normality also applies for the case where the dimension of $\boldsymbol{\theta}$ is greater than one, but the proofs of these results are more complicated.

We now know that the MLE is invariant under reparameterisations. If the MLE $\widehat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$, by the continuous mapping theorem, it follows that $\varphi(\boldsymbol{\theta})$ is a consistent estimator of $\varphi(\boldsymbol{\theta}_0)$. Now, if the MLE $\widehat{\boldsymbol{\theta}}$ is asymptotically normal, a natural question is whether or not $\varphi(\widehat{\boldsymbol{\theta}})$ is also asymptotically normal. The answer is yes, as shown in the following theorem for the case where the dimension of $\boldsymbol{\theta}$ is one. The multivariate case is not presented here as the proof is more complicated.

Theorem 2.7.5. *(The Delta Method). Let $\widehat{\theta}_n$ be a sequence of estimators (random variables) indexed by n that satisfies*

(i) $\widehat{\theta}_n \xrightarrow{P} \theta_0$,

(ii) $\sqrt{n}(\widehat{\theta}_n - \theta_0)^2 \xrightarrow{P} 0$,

(iii) $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \sigma^2)$.

For a given function φ, suppose that $\varphi'(\theta_0)$ exists and is nonzero. Suppose also that the second derivative is bounded $\varphi''(\theta) < M$, $\theta \in (\theta_0 - \epsilon, \theta_0 + \epsilon)$, for some $M > 0$ and $\epsilon > 0$. Then,

$$\sqrt{n}(\varphi(\widehat{\theta}_n) - \varphi(\theta_0)) \xrightarrow{d} N(0, \sigma^2[\varphi'(\theta_0)]^2).$$

Proof. The Taylor expansion of $\varphi(\widehat{\theta}_n)$ around θ_0 and the Taylor's remainder theorem imply

$$\varphi(\widehat{\theta}_n) = \varphi(\theta_0) + \varphi'(\theta_0)(\widehat{\theta}_n - \theta_0) + \frac{\varphi''(\theta^*)}{2}(\widehat{\theta}_n - \theta_0)^2,$$

where $\theta^* = \rho\widehat{\theta}_n + (1-\rho)\theta_0$, for some $\rho \in (0, 1)$. Then,

$$\sqrt{n}(\varphi(\widehat{\theta}_n) - \varphi(\theta_0)) = \sqrt{n}\varphi'(\theta_0)(\widehat{\theta}_n - \theta_0) + \sqrt{n}\frac{\varphi''(\theta^*)}{2}(\widehat{\theta}_n - \theta_0)^2,$$

Since $\widehat{\theta}_n \xrightarrow{P} \theta_0$, the second derivative is bounded around θ_0, and assumption (3), it follows that the remainder (second term) converges to 0 in probability. By assumption,

$$\sqrt{n}(\varphi(\widehat{\theta}_n) - \varphi(\theta_0)) = \sqrt{n}\varphi'(\theta_0)(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \sigma^2[\varphi'(\theta_0)]^2).$$

The result follows by combining these two results and applying Slutsky's theorem.

Assumption (i) is typically a necessary condition to get the asymptotic normality in assumption (i). Assumption (ii) establishes a condition on the rate of convergence of $(\widehat{\theta}_n - \theta_0) \xrightarrow{P} 0$ since the factor $\sqrt{n} \to \infty$ as $n \to \infty$. The Delta Method is also considered a generalisation of the Central Limit Theorem.

Example 2.7.13. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be *i.i.d.* random variables with mean $E[X_i] = \mu_0 \neq 0$, and variance $Var[X_i] = \sigma^2 < \infty$. Let $T(\mathbf{X}) = \bar{\mathbf{X}}$. By the central limit theorem:

$$\sqrt{n}(\bar{\mathbf{X}} - \mu_0) \xrightarrow{d} N(0, \sigma^2).$$

Let $g(z) = \dfrac{1}{z}$. We can check that this function satisfies the conditions in Theorem 2.7.5, and $g'(z) = -\dfrac{1}{z^2}$. Then,

$$\sqrt{n}\left(\frac{1}{\bar{\mathbf{X}}} - \frac{1}{\mu_0}\right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{\mu_0^4}\right).$$

Multiparameter case

Under more complex assumptions that we will not cover in this course (See Theorem 18 from [5] if you want to read the details), it is possible to show that the MLE is consistent and asymptotically normal in the multiparameter case. is. This is,

(i) $\widehat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$.

(ii) $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_0)\right)$, where $I(\boldsymbol{\theta}_0)$ is the Fisher information matrix.

2.8 Least Squares Estimation

Historical Note 2. The method of least squares was introduced (in a formal way) by Legendre in 1805. In 1809 Carl Friedrich Gauss connected the method of least squares with the principles of probability and to the normal distribution.

Linear Least Squares

Consider the formulation in Example 2.7.8. Suppose that you observe n responses $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ (also known as dependent variables), associated to n individuals. Suppose that, for each individual, you know p characteristics associated to them $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$ (also known as covariates or independent variables). The aim here is to explain the responses y_i, $i = 1, \ldots, n$, based on their corresponding individual characteristics $\mathbf{x}_i$. A possible way for doing so is to use the *linear model*

$$\mathbf{y} = \mathcal{D}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where

$$\mathcal{D} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix},$$

which is known as the *design matrix*, $\boldsymbol{\theta} \in \mathbb{R}^{p+1}$, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ are *i.i.d.* random variables known as the residual errors. Let us denote $\mathbf{d}_i^\top$ the rows of $\mathcal{D}$.

A general procedure for the estimation of $\boldsymbol{\theta}$ is to minimise

$$\sum_{i=1}^{n} M(y_i - \mathbf{d}_i^\top \boldsymbol{\theta}),$$

for a suitable choice of the function $M : \mathbb{R} \to \mathbb{R}$. In this section, we will study the case when $M(z) = z^2$, leading to the estimation procedure

$$\text{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - \mathbf{d}_i^\top \boldsymbol{\theta})^2.$$

One immediate observation is that this minimisation process is equivalent to maximising the log-likelihood function in Example 2.7.8 with respect to $\boldsymbol{\theta}$ only. However, in this case we are not assuming a distribution on the residual errors and, in fact, we can see that the variance σ^2 in Example 2.7.8 does not appear in this approach. This approach is known as Least Squares Estimation (LSE) or Ordinary Least Squares Estimation (OLSE).

In order to study inferential properties of the LSE approach, we will make the following specific assumptions

(i) $E[\boldsymbol{\epsilon}] = \mathbf{0}$.

(ii) $E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \sigma^2 \mathbf{I}_n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

(iii) Rank$(\mathcal{D}) = p + 1$.

(iv) $\mathcal{D}$ is a non-stochastic matrix.

(v) $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

(vi) $\frac{\mathcal{D}^\top \mathcal{D}}{n} \overset{a.s.}{\to} \Sigma_{\mathcal{D}}$, as $n \to \infty$, where $\Sigma_{\mathcal{D}}$ is a non-stochastic and non-singular matrix, with finite entries.

(vii) $\frac{\mathcal{D}^\top \boldsymbol{\epsilon}}{n} = \frac{1}{n} \sum_{i=1}^n \mathbf{d}_i \epsilon_i \overset{P}{\to} E[\mathbf{d}_1 \epsilon_1] = \mathbf{0}$.

These assumptions can be relaxed without affect the asymptotic properties of the LSE $\widehat{\boldsymbol{\theta}}_{\text{LSE}}$, but this is beyond the aims of this module.

As mentioned before, the estimation procedure

$$\text{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^n (y_i - \mathbf{d}_i^\top \boldsymbol{\theta})^2 = \text{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^n (\mathbf{y} - \mathcal{D}\boldsymbol{\theta})^\top (\mathbf{y} - \mathcal{D}\boldsymbol{\theta}).$$

The minimisation process can be done by differentiating with respect to $\boldsymbol{\theta}$ and finding the roots of this system of equations (see Example 2.7.8). Thus, the LSE of $\boldsymbol{\theta}$ is $\widehat{\boldsymbol{\theta}}_{\text{LSE}} = (\mathcal{D}^\top \mathcal{D})^{-1} \mathcal{D}^\top \mathbf{y}$.

Theorem 2.8.1. *Under assumptions (i)–(vii), we obtain the following properties.*

(a) $E[\widehat{\boldsymbol{\theta}}_{LSE}] = \boldsymbol{\theta}$.

(b) $Var[\widehat{\boldsymbol{\theta}}_{LSE}] = \sigma^2 (\mathcal{D}^\top \mathcal{D})^{-1}$.

(c) $\widehat{\boldsymbol{\theta}}_{LSE} \sim N_{p+1}\left(\boldsymbol{\theta}, \sigma^2 (\mathcal{D}^\top \mathcal{D})^{-1}\right)$.

Proof. (a) Let $\mathbf{M} = (\mathcal{D}^\top \mathcal{D})^{-1} \mathcal{D}^\top$. Then, $\widehat{\boldsymbol{\theta}}_{\text{LSE}} = \mathbf{M}\mathbf{y}$, and

$$\begin{aligned} \mathbf{y} &= \mathcal{D}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \\ \mathbf{M}\mathbf{y} &= \boldsymbol{M}\mathcal{D}\boldsymbol{\theta} + \mathbf{M}\boldsymbol{\epsilon}, \\ \widehat{\boldsymbol{\theta}}_{\text{LSE}} &= \boldsymbol{\theta} + \mathbf{M}\boldsymbol{\epsilon}, \\ \widehat{\boldsymbol{\theta}}_{\text{LSE}} - \boldsymbol{\theta} &= \mathbf{M}\boldsymbol{\epsilon}. \end{aligned}$$

Then,

$$E[\widehat{\boldsymbol{\theta}}_{\text{LSE}} - \boldsymbol{\theta}] = E[\mathbf{M}\boldsymbol{\epsilon}] = \boldsymbol{M} E[\boldsymbol{\epsilon}] = 0.$$

(b) Since $\boldsymbol{\theta}$ is non-random and by using the assumptions

$$\begin{aligned} Var[\widehat{\boldsymbol{\theta}}_{\text{LSE}}] &= Var[\widehat{\boldsymbol{\theta}}_{\text{LSE}} - \boldsymbol{\theta}] \\ &= Var[\mathbf{M}\boldsymbol{\epsilon}] \\ &= \mathbf{M} Var[\boldsymbol{\epsilon}] \mathbf{M}^\top \\ &= \mathbf{M} E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] \mathbf{M}^\top \\ &= \mathbf{M} \sigma^2 \mathbf{I}_{p+1} \mathbf{M}^\top \\ &= \sigma^2 \mathbf{M}\mathbf{M}^\top \\ &= \sigma^2 (\mathcal{D}^\top \mathcal{D})^{-1}, \end{aligned}$$

since $\mathbf{M}\mathbf{M}^\top = (\mathcal{D}^\top \mathcal{D})^{-1} \mathcal{D}^\top \mathcal{D} (\mathcal{D}^\top \mathcal{D})^{-1} = (\mathcal{D}^\top \mathcal{D})^{-1}$.

(c) Using that $\widehat{\boldsymbol{\theta}}_{\text{LSE}} - \boldsymbol{\theta} = \mathbf{M}\boldsymbol{\epsilon}$, and that $\mathbf{M}\boldsymbol{\epsilon} \sim N_{p+1}\left(\mathbf{0}, \sigma^2 \mathbf{M}\mathbf{M}^\top\right)$, we obtain the result.

The LSE has good asymptotic properties (consistency and asymptotic normality), as shown in the following theorem.

Theorem 2.8.2. *Under assumptions (i)–(vi).*

(a) $\widehat{\boldsymbol{\theta}}_{LSE} \xrightarrow{P} \boldsymbol{\theta}$, *as* $n \to \infty$.

(b) $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{LSE} - \boldsymbol{\theta}) \xrightarrow{d} N_{p+1}(\mathbf{0}, \sigma^2 \Sigma_{\mathcal{D}}^{-1})$, *as* $n \to \infty$.

Proof. (a) First, we know that

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{\text{LSE}} - \boldsymbol{\theta} &= \mathbf{M}\boldsymbol{\epsilon} = (\mathcal{D}^\top \mathcal{D})^{-1} \mathcal{D}^\top \boldsymbol{\epsilon} \\ &= \left(\frac{\mathcal{D}^\top \mathcal{D}}{n}\right)^{-1} \frac{\mathcal{D}^\top \boldsymbol{\epsilon}}{n}. \end{aligned}$$

By assumption $\left(\frac{\mathcal{D}^\top \mathcal{D}}{n}\right)^{-1} \xrightarrow{a.s.} \Sigma_{\mathcal{D}}^{-1}$, which has finite entries, by assumption. By the law of large numbers $\frac{\mathcal{D}^\top \boldsymbol{\epsilon}}{n} = \frac{1}{n}\sum_{i=1}^n \mathbf{d}_i \epsilon_i \xrightarrow{P} E[\mathbf{d}_1 \epsilon_1] = \mathbf{0}$, as $n \to \infty$. The result follows by combining these two asymptotic results together with Slutsky's theorem.

(b) Note that $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\text{LSE}} - \boldsymbol{\theta}) = \sqrt{n}\mathbf{M}\boldsymbol{\epsilon}$. We know that $\mathbf{M}\boldsymbol{\epsilon} \sim N_{p+1}\left(\mathbf{0}, \sigma^2(\mathcal{D}^\top\mathcal{D})^{-1}\right)$, then $\sqrt{n}\mathbf{M}\boldsymbol{\epsilon} \sim N_{p+1}\left(\mathbf{0}, \sigma^2\left(\frac{1}{n}\mathcal{D}^\top\mathcal{D}\right)^{-1}\right)$.

Consequently, $\frac{1}{\sigma}\left(\frac{1}{n}\mathcal{D}^\top\mathcal{D}\right)^{\frac{1}{2}} \sqrt{n}\mathbf{M}\boldsymbol{\epsilon} \sim N_{p+1}\left(\mathbf{0}, \mathbf{I}_{p+1}\right)$ (where the square root is taken in a matrix sense, see preliminary material). Then, we can define the random vectors $\mathbf{Z}_n = \frac{1}{\sigma}\left(\frac{1}{n}\mathcal{D}^\top\mathcal{D}\right)^{\frac{1}{2}} \sqrt{n}\mathbf{M}\boldsymbol{\epsilon}$, and conclude that $\mathbf{Z}_n \xrightarrow{d} N_{p+1}\left(\mathbf{0}, \mathbf{I}_{p+1}\right)$. Now, we now that $\frac{1}{\sigma}\left(\frac{1}{n}\mathcal{D}^\top\mathcal{D}\right)^{\frac{1}{2}} \xrightarrow{P} \frac{1}{\sigma}\Sigma_{\mathcal{D}}^{\frac{1}{2}}$. Then, by Slutsky's theorem, $\sigma\left(\frac{1}{n}\mathcal{D}^\top\mathcal{D}\right)^{-\frac{1}{2}} \mathbf{Z}_n = \sqrt{n}\mathbf{M}\boldsymbol{\epsilon} \xrightarrow{d} N\left(0, \sigma^2\Sigma_{\mathcal{D}}^{-1}\right)$.

Remark 2.8.1. Extension. *Linear Least Squares Estimation can be generalised to non-linear least squares estimation. The idea is to formulate the more general relationship*

$$\mathbf{y} = g(\mathbf{x}_i, \boldsymbol{\theta}) + \boldsymbol{\epsilon},$$

where g *is a parametric function. Typically,* g *is chosen to be a polynomial (or a spline basis) on each of the covariates* $\mathbf{x}_i$*. The estimation procedure consists of minimising*

$$\text{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^n (y_i - g(\mathbf{x}_i, \boldsymbol{\theta}))^2.$$

In most cases, the solution to this equation is not available in closed-form, and the use of numerical methods is necessary. Now, you are familiar with Newton's algorithm, which can be used to find in this case. However, the solution to this equation can be difficult to find, even with the Newton's algorithm as some functions are not concave (a condition for convergence of this algorithm). In those cases, more sophisticated numerical methods need to be considered, such as gradient descent and coordinate descent algorithms.

2.9 The Method of Moments

[14].

Recall that the kth moment of a random variable X with pdf $f(x;\boldsymbol{\theta})$ and support $\text{supp}(f) = \mathcal{D} \subseteq \mathbb{R}$ is defined as:

$$\mu_k = E[X^k] = \int_{\mathcal{D}} x^k f(x;\boldsymbol{\theta}) dx.$$

If $X_1, \ldots, X_n$ are *i.i.d.* random variables from that distribution, the kth *sample moment* is defined as

$$\tilde{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

We can view $\tilde{\mu}_k$ as an estimator of μ_k. In fact, we can establish consistency of $\tilde{\mu}_k$, under some conditions (that allows us to use the law of large numbers), by appealing to the law of large numbers.

Consider the particular case where $\boldsymbol{\theta} = (\theta_1, \theta_2)$. Thus, we know that, in general, the moments of $f(x;\boldsymbol{\theta})$ are a function of the parameters. In particular

$$\begin{aligned} \mu_1 &= g_1(\theta_1, \theta_2), \\ \mu_2 &= g_2(\theta_1, \theta_2). \end{aligned}$$

Let us assume that we can invert this relationship and, consequently, we can find functions h_1 and h_2 such that

$$\begin{aligned} \theta_1 &= h_1(\mu_1, \mu_2), \\ \theta_2 &= h_2(\mu_1, \mu_2). \end{aligned}$$

Then, the method of moments is defined as

$$\begin{aligned} \tilde{\theta}_1 &= h_1(\tilde{\mu}_1, \tilde{\mu}_2), \\ \tilde{\theta}_2 &= h_2(\tilde{\mu}_1, \tilde{\mu}_2). \end{aligned}$$

The extension to the case where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ follows analogously. The idea of the method of moments (MM) consists of the following three steps

1. Calculate low order moments, finding expressions for the moments in terms of the parameters. Typically, the number of low order moments needed will be the same as the number of parameters.
2. Invert the expressions found in the preceding step, finding new expressions for the parameters in terms of the moments.
3. Insert the sample moments into the expressions obtained in the second step, thus obtaining estimates of the parameters in terms of the sample moments.

Next, we will illustrate this idea through several examples.

Example 2.9.1. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be *i.i.d.* Poisson random variables with parameter $\lambda > 0$. The first moment of the Poisson distribution is $\mu_1 = E[X] = \lambda$. The first sample moment is

$$\tilde{\mu}_1 = \bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{2.9.1}$$

Therefore, the method of moments estimates λ as $\tilde{\lambda} = \bar{\mathbf{X}}$.

Note that the method of moments estimator (MME) coincides with the MLE obtained previously. Consequently, it also fails to exist for samples where all the observations are zero (why?).

Example 2.9.2. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be *i.i.d.* Normal random variables with mean μ_0 and variance σ_0^2. The first and second moments of the Normal distribution are

$$\begin{aligned} \mu_1 &= E[X] = \mu, \\ \mu_2 &= E[X^2] = \mu^2 + \sigma^2. \end{aligned}$$

Therefore,

$$\begin{aligned} \mu &= \mu_1, \\ \sigma^2 &= \mu_2 - \mu^2. \end{aligned}$$

The corresponding MMEs of μ and σ^2 are obtained by inserting the sample moments

$$\begin{aligned} \tilde{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{\mathbf{X}}, \\ \tilde{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{\mathbf{X}}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{\mathbf{X}})^2. \end{aligned}$$

Thus, the MME coincide with the MLE again. Thus, the asymptotic properties are also the same. In particular, the MME $\tilde{\mu}$ and $\tilde{\sigma}^2$ are consistent estimators of μ_0 and σ_0^2.

Example 2.9.3. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be *i.i.d.* Gamma random variables with shape parameter $\kappa > 0$, and scale parameter $\theta > 0$. The first and second moments of the Gamma distribution are

$$\begin{aligned} \mu_1 &= E[X] = \kappa\theta, \\ \mu_2 &= E[X^2] = Var(X) + E[X]^2 = \kappa\theta^2 + \kappa^2\theta^2. \end{aligned}$$

Therefore,

$$\begin{aligned} \kappa\theta &= \bar{\mathbf{X}}, \\ \kappa\theta^2 + \kappa^2\theta^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2. \end{aligned}$$

Consequently, $\theta = \frac{\bar{\mathbf{X}}}{\kappa}$, and replacing this value in the second equation we obtain the MME

$$\begin{aligned} \tilde{\theta} &= \frac{S_n}{\bar{\mathbf{X}}}, \\ \tilde{\kappa} &= \frac{\bar{\mathbf{X}}^2}{S_n} \end{aligned}$$

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

In this case, and in contrast to the MLE, the MME has a closed form that is easy to evaluate.

In many cases, the expressions for the moments cannot be found and close form. The method of moments is still applicable in such cases, but the solution can only be found numerically by solving a system of non-linear equations.

Example 2.9.4. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be *i.i.d.* random variables with Kumaraswamy distribution with parameters $a > 0$ and $b > 0$. The probability density function of the Kumaraswamy distribution is

$$f(x; a, b) = abx^{a-1}(1 - x^a)^{b-1}, \text{ where } x \in [0, 1].$$

The Kumaraswamy distribution is an alternative to the Beta distribution, which offers similar flexibility in terms of the shape of the density. An appealing feature of the Kumaraswamy distribution is that is

only involves algebraic terms, in contrast to the Beta distribution, which requires the evaluation of the Beta function. The mean and variance of this distribution are

$$\begin{aligned} \mu_1 &= E[X] = \frac{b\Gamma(1+\frac{1}{a})\Gamma(b)}{\Gamma(1+\frac{1}{a}+b)}, \\ Var[X] &= E[X^2] - E[X]^2 = \mu_2 - \mu_1^2 = \frac{b\Gamma(1+2/a)\Gamma(b)}{\Gamma(1+b+2/a)} - \mu_1^2. \end{aligned}$$

The MME of a and b, $\tilde{a}$ and $\tilde{b}$, are the solution to the system of non-linear equations,

$$\begin{aligned} \bar{\mathbf{X}} &= \frac{b\Gamma(1+\frac{1}{a})\Gamma(b)}{\Gamma(1+\frac{1}{a}+b)}, \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{\mathbf{X}}^2 &= \frac{b\Gamma(1+2/a)\Gamma(b)}{\Gamma(1+b+2/a)} - \mu_1^2. \end{aligned}$$

which cannot be found in closed form. Thus, we can only obtain the MME using numerical methods.

The numerical solution can be found in the following R code:

`http://rpubs.com/FJRubio/KumaMM`

In some cases, it is not possible to obtain expressions for the moments in terms of elemental functions. In such cases, the implementation of the MME requires numerical integration as well.

Asymptotic properties of the MME

In order to establish consistency and asymptotic normality of the MMEs, let us first formalise some ideas. We will focus on the one-parameter case since the multi-parameter scenario is more challenging as it involves additional results on the convergence of random vectors.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be *i.i.d.* random variables with pdf $f(x;\theta_0)$, $x \in \mathcal{D} \subset \mathbb{R}$ and $\theta_0 \in \Theta \subseteq \mathbb{R}$. Suppose that we can find a function $g : \mathcal{D} \to \mathbb{R}$ such that the function

$$m(\theta) = E_\theta[g(X)],$$

has a continuous inverse m^{-1}. The use of the function g allows one to use higher order moments instead of restricting to the first moment. For instance, $g(x) = x^k$, for any choice of $k > 0$. Here E_θ denotes the expectation with respect to $f(\cdot;\theta)$. Define

$$\tilde{\theta} = m^{-1}(\bar{g}) = m^{-1}\left(\frac{g(X_1) + \cdots + g(X_n)}{n}\right),$$

as the estimate of θ_0. The following result shows the consistency of the MME under the same conditions of the weak law of large numbers.

Theorem 2.9.1. *Consistency of the MME. Let* $X_1, \ldots, X_n$ i.i.d. *random variables with pdf* $f(\cdot;\theta)$, $\theta \in \Theta \subset \mathbb{R}$. *Suppose that* $E[g(X_i)] < \infty$. *as* $n \to \infty$

$$\tilde{\theta} \xrightarrow{P} \theta_0.$$

Proof. The intuition is that the weak law of large numbers (WLLN) implies that the sample moments converge in probability to the population moments. If the functions relating the estimates to the sample moments are continuous, the estimates will converge to the parameters as the sample moments converge to the population moments. This is, by the WLLN, as $n \to \infty$

$$\bar{g} \xrightarrow{P} E_{\theta_0}[g(X_1)] = m(\theta_0).$$

Since the inverse m^{-1} is continuous, the continuous mapping theorem (see preliminary material) implies that

$$\tilde{\theta} = m^{-1}(\bar{g}) \xrightarrow{P} m^{-1}(m(\theta_0)) = \theta_0.$$

The MMEs are also asymptotically normal as shown in the following result.

Theorem 2.9.2. *The estimate $\tilde{\theta} = m^{-1}(\bar{g})$ obtained by the method of moments is asymptotically normal with asymptotic variance*

$$\sigma^2_{\theta_0} = \frac{Var_{\theta_0}(g(X_1))}{[m'(\theta_0)]^2}.$$

This is,

$$\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{d} N\left(0, \sigma^2_{\theta_0}\right).$$

Proof. By applying the Taylor's series expansion and the Taylor's Remainder Theorem of the function m^{-1} at a point $m(\theta_0)$, we have

$$m^{-1}(\bar{g}) = m^{-1}(m(\theta_0)) + (m^{-1})'(m(\theta_0))(\bar{g} - m(\theta_0)) + \frac{(m^{-1})''(c)}{2}(\bar{g} - m(\theta_0))^2,$$

where $c \in [m(\theta_0), \bar{g}]$, that is $c = \lambda m(\theta_0) + (1-\lambda)\bar{g}$ for some $\lambda \in [0,1]$. Since $m^{-1}(m(\theta_0)) = \theta_0$, we get

$$m^{-1}(\bar{g}) - \theta_0 = (m^{-1})'(m(\theta_0))(\bar{g} - m(\theta_0)) + \frac{(m^{-1})''(c)}{2}(\bar{g} - m(\theta_0))^2,$$

We need to show that the left hand side, multiplied by $\sqrt{n}$ is asymptotically normal. Now, note that the derivative of the inverse is given by

$$(m^{-1})'(m(\theta_0)) = \frac{1}{m'(m^{-1}(m(\theta_0)))} = \frac{1}{m'(\theta_0)}.$$

Then, for the first term we obtain, using the central limit theorem and that $E_{\theta_0}(g(X_1)) = m(\theta_0)$,

$$\frac{\sqrt{n}}{m'(\theta_0)}(\bar{g} - m(\theta_0)) \xrightarrow{d} N\left(0, \frac{Var_{\theta_0}(g(X_1))}{(m'(\theta_0))^2}\right).$$

Now, for the second term we have

$$\sqrt{n}\frac{(m^{-1})''(c)}{2}(\bar{g} - m(\theta_0))^2 = \frac{(m^{-1})''(c)}{2}\frac{1}{\sqrt{n}}(\sqrt{n}(\bar{g} - m(\theta_0)))^2.$$

Now, as $n \to \infty$, $c \xrightarrow{P} m(\theta_0)$, since $\bar{g} \xrightarrow{P} m(\theta_0)$. By the continuous mapping theorem $(m^{-1})''(c) \xrightarrow{P} (m^{-1})''(m(\theta_0))$. Also, $\frac{1}{\sqrt{n}} \to 0$. Finally, by the previous result $\sqrt{n}(\bar{g} - m(\theta_0)) \xrightarrow{d} N(0, Var_{\theta_0}(g(X_1)))$. Combining these results and using Slutsky's theorem it follows that

$$\sqrt{n}\frac{(m^{-1})''(c)}{2}(\bar{g} - m(\theta_0))^2 \xrightarrow{d} 0.$$

Since convergence in distribution to a constant implies convergence in probability

$$\sqrt{n}\frac{(m^{-1})''(c)}{2}(\bar{g} - m(\theta_0))^2 \xrightarrow{P} 0.$$

Finally, combining the two convergence results and applying Slutsky's theorem again, we obtain the desired result.

Example 2.9.5. (Non-existence of the MME).

(i) Let $X_1, \ldots, X_n$ *i.i.d.* random variables with Cauchy distribution. The pdf of the Cauchy distribution with location parameter $\theta \in \mathbb{R}$ and unit scale parameter is

$$f(x;\theta) = \frac{1}{\pi[1+(x-\theta)^2]}, \quad x \in \mathbb{R}. \tag{2.9.2}$$

It can be shown that

$$E[X] = \int_{-\infty}^{0} x f(x;\theta)dx + \int_{0}^{\infty} x f(x;\theta)dx,$$

is undefined for all values of θ, as both terms are infinite. Thus, the method of moments cannot be applied for this simple distribution.

(ii) Let $X_1, \ldots, X_n$ *i.i.d.* random variables with Uniform distribution on $(-\theta, \theta)$. Then, $E[X] = 0$, and the first moment cannot be used to construct a MME.

On the other hand, the MLE exists in both cases.

Example 2.9.6. Multivariate Normal Distribution Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p$ be *i.i.d.* random vectors with Multivariate Normal distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and $p \times p$ covariance matrix Σ (assumed to be positive definite).

The first moment of $\mathbf{X}_i$ is $E[\mathbf{X}_i] = \boldsymbol{\mu}$. Equating this to the first sample moment,

$$\tilde{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i.$$

The second moment of $\mathbf{X}_i$ is defined as $E[\mathbf{X}_i\mathbf{X}_i^\top] = \Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^\top$. Equating this to the second sample moment and replacing $\boldsymbol{\mu}$ by $\tilde{\boldsymbol{\mu}}$ we obtain (exercise):

$$\tilde{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top - \bar{\mathbf{x}}\bar{\mathbf{x}}^\top = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top.$$

Consequently, the MME and the MLE coincide for the multivariate normal distribution.

2.10 The Generalised Method of Moments

Historical Note 3. The Generalised Method of Moments was developed by Professor Lars Peter Hansen as a generalisation of the method of moments. Professor Hansen shared the 2013 Nobel Prize in Economics in part for this contribution.

The Generalised Method of Moments (GMM), as its name indicates, is a generalisation of the method of moments. In order to present the intuition behind GMM, consider the following example. Let $X_1, \ldots, X_n$ be *i.i.d.* random variables with distribution $N(\mu_0, 1)$. Thus, $E[X_i] = \mu_0$. Consider the quantity $\bar{\mathbf{X}}$. We know that $\bar{\mathbf{X}} \xrightarrow{P} \mu_0$. Thus, consider

$$E[X_i - \mu_0] = 0.$$

In the light of these results, it follows that a way to estimate μ_0 is to look at the solution of

$$\frac{1}{n}\sum_{i=1}^{n} x_i - \mu = 0,$$

and we can take the estimator

$$\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The main ingredient of a **GMM** estimation is a multivariate function $h(\boldsymbol{\theta}, x_i)$ with $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ parameters to be estimated and x_i data points. In our previous example concerning MME, this function was $\bar{\mathbf{X}} - \mu$. The idea in GMM is to consider more general functions. This function has to satisfy the "orthogonality condition" about its expectation:

$$E_{\boldsymbol{\theta}_0}[h(\boldsymbol{\theta}_0, X_i)] = 0,$$

under the true value of the parameters. Define the multivariate function

$$g(\boldsymbol{\theta}, \mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{\theta}, X_i).$$

If the number of parameters to estimate equals the number of orthogonality conditions, we can find the estimator $\tilde{\theta}$ directly as the solution to

$$g(\tilde{\boldsymbol{\theta}}, \mathbf{X}) = 0.$$

Otherwise, if the number of parameters to estimate is less than the number of orthogonality conditions, we can find $\tilde{\theta}$ as

$$\tilde{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}) = \operatorname{argmin}_{\boldsymbol{\theta}} g(\boldsymbol{\theta}, \mathbf{X})^\top W g(\boldsymbol{\theta}, \mathbf{X})$$

where W is a positive definite weighting matrix. There are different ways to specify this weighting matrix. Under some "regularity conditions", the GMM estimators are consistent and asymptotically normal. The proofs of these results are beyond the aims of this course. We will illustrate the generality and usefulness of the GMM formulation with some examples.

Example 2.10.1. Consider the simple linear regression model

$$y_i = x_i\beta + \epsilon_i, \quad i = 1, \ldots, n,$$

where $x_i, y_i, \beta \in \mathbb{R}$, and impose the following conditions

$$\begin{aligned} E[\epsilon_i] &= 0, \\ E[x_i\epsilon_i] &= 0, \\ Var[\epsilon_i] &= \sigma^2. \end{aligned}$$

Note that these conditions are similar to those that we imposed in the LSE. The moment condition $E[x_i\epsilon_i] = E[x_i(y_i - x_i\beta)] = 0$, which is the expectation of the function h. The equivalent sample condition is

$$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - x_i\beta) = 0,$$

which defines the function g. By solving this equation, we obtain the GMM estimator

$$\tilde{\beta}_{\text{GMM}} = \left(\sum_{i=1}^{n} x_i x_i\right)^{-1} \left(\sum_{i=1}^{n} x_i y_i\right).$$

In vector notation,

$$\tilde{\beta}_{\text{GMM}} = \left(\mathbf{x}^\top\mathbf{x}\right)^{-1}\mathbf{x}^\top\mathbf{y}.$$

Thus, the GMM estimator coincides with the LSE.

Example 2.10.2. Let $\mathbf{X} = (X_1, \ldots, X_n)$ *i.i.d.* random variables distributed according to $f(\cdot; \boldsymbol{\theta})$. Consider the system of equations:

$$E\left[\nabla_{\boldsymbol{\theta}} \log f(X_i, \boldsymbol{\theta})\right],$$

and the corresponding system of sample equations

$$\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log f(X_i, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \left(\sum_{i=1}^{n} \log f(X_i, \boldsymbol{\theta}) \right) = 0.$$

The GMM estimator is the solution to this system of equations. Moreover, we can identify this system of equations as the Score function of $f(\cdot; \boldsymbol{\theta})$. Thus, the GMM coincides with MLE under this formulation.

2.11 Robust Statistics and M-Estimation

Historical Note 4. M-Estimators were proposed by Peter J. Huber in the paper "Robust Estimation of a Location Parameter", in 1964.

In the maximum likelihood estimation method, the idea is to either minimise the negative log-likelihood (equivalently, the likelihood) function

$$\widehat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \; -\ell(\boldsymbol{\theta} \mid \mathbf{x}) = \text{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{n} -\log f(x_i; \boldsymbol{\theta}).$$

When the log-likelihood is differentiable, this is equivalent to finding the roots of the score function:

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta} \mid \mathbf{x}) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log f(x_i; \boldsymbol{\theta}) = \mathbf{0}.$$

M-Estimators are a broader class of estimators that generalise the idea of MLE. A M-Estimate is the value that minimises the function:

$$\widehat{\boldsymbol{\theta}}_M = \text{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \rho(x_i; \boldsymbol{\theta}), \tag{2.11.1}$$

or, as the roots of the implicit equation:

$$\sum_{i=1}^{n} \psi(x_i; \boldsymbol{\theta}) = \mathbf{0}. \tag{2.11.2}$$

where ρ is an arbitrary function, and $\psi(x_i; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \rho(x_i; \boldsymbol{\theta})$. Thus, we can see that MLE is a particular case of M-Estimation for $\rho(x_i; \boldsymbol{\theta}) = -\log f(x_i; \boldsymbol{\theta})$. In some literature, the estimators defined by (2.11.1) are called M-Estimators, while the estimators defined by (2.11.2) are referred to as Z-Estimators (are they are based on finding the zeroes of a function). The reason for this distinction is that the second definition requires differentiability of ρ with respect to $\boldsymbol{\theta}$, while the first one does not. Since the function ρ, in general, is not necessarily related to the model (pdf) f, it follows that this estimation method is not a probabilistic method. However, this does *not* mean that the M-Estimators do not have asymptotic properties, as $\widehat{\boldsymbol{\theta}}_M$ is still a random vector. Moreover, M-Estimates do not require assuming that the true distribution of the observations x_i is some pdf $f(; \boldsymbol{\theta})$.

Example 2.11.1. linear and non-linear least square estimation are particular cases of M-Estimation for $\rho(y_i, \mathbf{x}_i; \boldsymbol{\theta}) = (y_i - g(\mathbf{x}_i; \boldsymbol{\theta}))^2$.

M-Estimation theory is a general theory, but here we will only focus on the estimation of (pure, without nuisance parameters) location and scale parameters using this approach.

2.11.1 M-Estimates of Location

Let $\mu \in \mathbb{R}$ be a location parameter of a pdf $f(\cdot;\mu)$. We are interested in the class of M-Estimates of the type:

$$\begin{aligned} \widehat{\mu}_n &= \text{argmin}_\mu \sum_{i=1}^n \rho(x_i - \mu), \\ &\text{or} \\ \widehat{\mu}_n &: \sum_{i=1}^n \psi(x_i - \mu) = 0. \end{aligned}$$

Let us focus on the second equation, which, for the sake of simplicity of notation will be written as

$$\sum_{i=1}^n \psi(x_i - \widehat{\mu}_n) = 0.$$

Define the weights (sample-dependent) $w_i = \dfrac{\psi(x_i - \widehat{\mu}_n)}{x_i - \widehat{\mu}_n}$. Then, the equation that defines the M-estimate of μ can be written as

$$\sum_{i=1}^n w_i(x_i - \widehat{\mu}_n) = 0.$$

Solving this equation, we obtain

$$\widehat{\mu}_n = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}.$$

Although this is not the final solution, as we have not specified ψ, this equation indicates that the M-Estimates in this case are weighted means.

Example 2.11.2. Define $\rho(x;\mu) = \dfrac{(x-\mu)^2}{2}$. Then, $\psi(x;\mu) = \mu - x$. Consequently, the M-Estimate of μ is $\widehat{\mu}_n = \bar{\mathbf{x}}$.

As we have discussed before, this estimator is consistent and asymptotically normal. We will not study these asymptotic properties here as they require a bit more of theory.

Example 2.11.3. Define $\rho(x;\mu) = |x - \mu|$. Then, for a sample $\mathbf{x}$ of size n, we want to minimise

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

This is known as least absolute deviation. As you know from calculus (recall the example about the derivative of $f(x) = |x|$), this function is differentiable everywhere except at the values $\theta = x_1, \ldots, x_n$. For $\theta \neq x_1, \ldots, x_n$,

$$M'(\mu) = -\sum_{i=1}^n \frac{x_i - \mu}{|x_i - \mu|}.$$

The function $\text{sgn}(x-\mu) = \dfrac{x-\mu}{|x-\mu|}$ is known as the *sign* function. Let us now define $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$, the order statistics (this is, we are just ordering the sample and giving the ordered samples a subindex that indicates their order). Then, we have the following results:

1. If $\mu < x_{(1)}$, then $M'(\mu) = -n$.

2. If $x_{(1)} < \mu < x_{(2)}$, then $M'(\mu) = 2 - n$ (as the first term is 1, and the remaining $n-1$ terms are -1).

3. If $x_{(2)} < \mu < x_{(3)}$, then $M'(\mu) = 4 - n$.

4. If $x_{(k)} < \mu < x_{(k+1)}$, then $M'(\mu) = 2k - n$.

We can also see that $M(\mu)$ is a polygonal curve (piecewise linear). Consequently, if n is odd, then, using the previous results, it follows that $M(\mu)$ is strictly decreasing on the interval $\left(-\infty, x_{\left(\frac{n+1}{2}\right)}\right]$, and strictly increasing on the interval $\left[x_{\left(\frac{n+1}{2}\right)}, \infty\right)$, so that the minimum of $M(\mu)$ is achieved at: $\widehat{\mu}_n = x_{\left(\frac{n+1}{2}\right)}$. Now, if n is even, the lowest points (plural) of $M(\mu)$ lie on the segment

$$\left\{(\mu, M(\mu)) : x_{\left(\frac{n}{2}\right)} \leq \mu \leq x_{\left(\frac{n}{2}\right)+1}\right\}.$$

Thus, we can define the M-Estimate of μ, for any n, as the sample median:

$$\widehat{\mu}_n = \text{Median}(\mathbf{x}) = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{if } n \text{ is odd} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}\right)+1}}{2}, & \text{if } n \text{ is even.} \end{cases}$$

Note that the previous M-Estimate applies to any location parameter, such as those in the Normal and Logistic distributions. In R, you can calculate the median of a sample by using the command `median(.)`. In Example 2.11.2, we know that $\widehat{\mu}_n \xrightarrow{P} E[X_i]$, as $n \to \infty$, the expectation of the distribution that generated the sample. In Example 2.11.3, we know that $\widehat{\mu}_n \xrightarrow{P} \text{Median}[X_i]$, as $n \to \infty$, the median (quantile 0.5) of the distribution that generated the sample. If the distribution that generated the sample is symmetric, these estimators coincide. However, if such distribution is not symmetric (if it is asymmetric), then the estimates (and its limit) will be different. For this reason, this kind of M-Estimates are only used as estimates of a location parameter when the true generating distribution is assumed to be symmetric about some point μ_0.

Now, we will address the concept of robustness in a more formal way.

Definition 28. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be *i.i.d.* random variables with distribution $F(\cdot; \theta)$, $\theta \in \mathbb{R}$. Let $T(\mathbf{X})$ an estimator of θ. The **Sensitivity Curve** (or empirical influence function) is of the estimator T at a point x is defined as

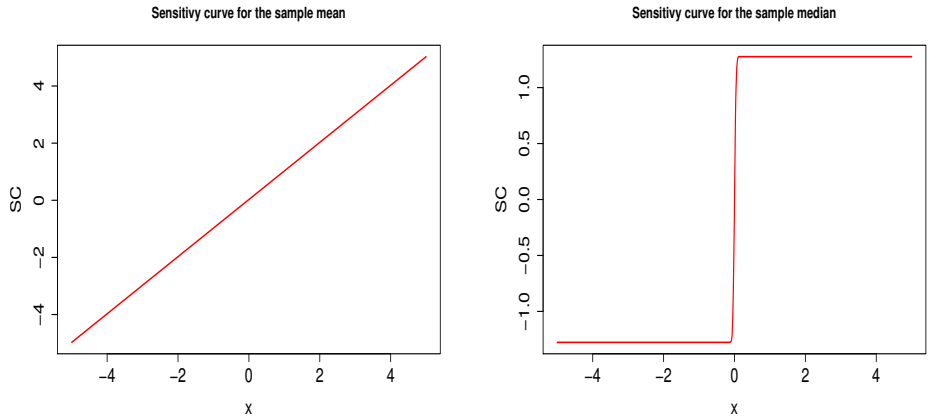
$$SC(x, T) = (n+1)\left[T(X_1, \ldots, X_n, x) - T(X_1, \ldots, X_n)\right].$$

It is possible to relax the assumption of independence, but that is beyond the aim of this course. This function measures the effect of adding one data point on a given sample, as a function of such additional point. For example, if T is the sample mean, we get that

$$SC(x, T) = (n+1)\left[\frac{(n\bar{\mathbf{X}} + x)}{n+1} - \bar{\mathbf{X}}\right] = x - \bar{\mathbf{X}}$$

Definition 29. Outlier (The Cambridge Dictionary of Statistics). "An observation that appears to deviate markedly from the other members of the sample in which it occurs. In the set of systolic blood pressures, $\{125; 128; 130; 131; 198\}$, for example, 198 might be considered an outlier. More formally the term refers to an observation which appears to be inconsistent with the rest of the data, relative to an assumed model. Such extreme observations may be reflecting some abnormality in the measured characteristic of a subject, or they may result from an error in the measurement or recording."

Thus, when the estimator is the sample mean, we obtain the effect of adding one new observation is linear and unbounded. The larger x, the larger the effect. This suggests that if our sample contains extreme observations or "outliers" (observations that are distant from most of the observations in the sample), these values will have a non-negligible effect on the estimator. On the other hand, if the estimator T is the sample median, then the sensitivity curve is bounded. Figure 2.11.1a shows an example of the sensitivity curve for a sample of size $n = 1,000$ coming from a $N(0,1)$. Figure 2.11.1b shows an approximation of the sensitivity curve for samples of size $n = 1,000$ coming from a $N(0,1)$. The effect of a new observation has only a bounded effect on the median.



(a) (b)

Figure 2.11.1: Sensitivity curves for the sample mean and the sample median.

Definition 30. The **Influence Function** is defined as the limit (in probability) of the Sensitivity Curve as $n \to \infty$. This is,

$$IF(x) = \lim_{n\to\infty} SC(x,T).$$

If the influence function of a statistic (estimator) is bounded, we say the estimator is *Robust*.

Example 2.11.4. Let T be the sample mean, and suppose that $X_i \sim F$, for an arbitrary distribution F with finite mean. Then, the influence function is

$$\begin{aligned} IF(x) &= \lim_{n\to\infty} SC(x,T) \\ &= \lim_{n\to\infty} [x - \bar{\mathbf{X}}] \\ &= x - E_F[X_1]. \end{aligned}$$

This function is unbounded and therefore the sample mean is not a robust estimator. Notice also that if the expectation does not exist or is infinite, then the influence function is undefined or infinite.

Example 2.11.5. (Extra). Let T be the sample mean, and suppose that $X_i \sim F$, for an arbitrary distribution F with density f. After some calculations, and the use of more advanced mathematical tools (such as Functional Analysis), it is possible to show that, if T is the sample median, then

$$\begin{aligned} IF(x) &= \lim_{n\to\infty} SC(x,T) \\ &= \lim_{n\to\infty} (n+1)[\text{Median}(\mathbf{X}, x) - \text{Median}(\mathbf{X})] \\ &= \frac{1}{2f\left(F^{-1}(0.5)\right)} \text{sign}\left(x - F^{-1}(0.5)\right), \end{aligned}$$

where the limit is taken in a probabilistic sense. Then, if the true generating distribution F is such that its 0.5 quantile (median) $F^{-1}(0.5)$ is in the support of the pdf f (so that the denominator is positive), then the influence function is bounded. Consequently, the median is a robust estimator.

Note that in this case we are not assuming that the expectation of X_i is finite, in contrast to the example about the sample mean estimator.

Before we proceed with the next example, let us introduce a useful statistical tool: Simulation (or random number generation).

Definition 31. Simulation (The Cambridge Dictionary of Statistics). "The artificial generation of random processes (usually by means of pseudo-random numbers and/or computers) to imitate the behaviour of particular statistical models."

Example 2.11.6. Now, we consider several scenarios where we are interested in estimating (purely) the scale parameter. We simulate $M = 1000$ samples (fake data, emulating sampling from those distributions) of size $n = 100$ from:

1. $N(20, 1)$.
2. $N(20, 3)$.
3. 95% from a $N(20, 1)$ and 5% from a Student's t-distribution with 2 degrees of freedom.
4. 90% from a $N(20, 1)$ and 10% from a Student's t-distribution with 2 degrees of freedom.

For each of these simulated samples, we calculate the mean and the median. Scenarios 1 and 2 correspond to the case where the true model is normal and, consequently, the SD is the MLE, which has good asymptotic and finite sample properties. In Scenarios 3 and 4, there is a contamination of the normal samples with different proportions of samples coming from another population (distribution). In fact, in real applications is very common to come across samples that apparently come from the same population, with similar characteristics, but where some individuals are actually different. An example of this is measuring the Body Mass Index (weight in kilograms divided by the square of the height in metres) of healthy (without chronic diseases) young adults (of the same sex). If the sample contains athletes, which is not uncommon, the BMI of those individuals tends to be higher.

The R code and results can be found at:

`http://www.rpubs.com/FJRubio/MM`

2.11.2 Robust Estimates of Scale

In this section, by *Scale Estimate* we will refer to a positive statistic that is equivariant under scale transformations. This is,

$$T(ax_1, \ldots, ax_n) = aT(x_1, \ldots, x_n), \quad a > 0.$$

One of the most popular scale estimates is the median absolute deviation (MAD):

$$\text{MAD}_n = \text{Median}\left\{|x_i - M_n|\right\},$$

where $M_n = \text{Median}(\mathbf{x})$. Note that this is an analogous definition to that of the sample variance $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2$, but using medians instead of means. In applications, it is preferred to use the normalised MAD, which is

$$\text{NMAD}_n = b\text{Median}\left\{|x_i - M_n|\right\},$$

for some $b > 0$. The choice of the factor $b > 0$ is based on making the estimator consistent for the parameter of interest. In the particular case where the data come from a normal distribution, the choice $b = \dfrac{1}{\Phi^{-1}(0.75)} \approx 1.4826$ makes the NMAD a consistent estimator for σ.

Example 2.11.7. Now, we consider several scenarios where we are interested in estimating (purely) the scale parameter. We simulate $M = 1000$ samples (fake data, emulating sampling from those distributions) of size $n = 100$ from:

1. $N(0, 1)$.
2. $N(0, 3)$.
3. 95% from a $N(0, 1)$ and 5% from a Student's t-distribution with 2.5 degrees of freedom.
4. 90% from a $N(0, 1)$ and 10% from a Student's t-distribution with 2.5 degrees of freedom.

For each of these simulated samples, we calculate the NMAD and the SD. Then, we compare the resulting estimators.

The R code and results can be found at:

`http://www.rpubs.com/FJRubio/NMADSD`

Example 2.11.8. Outlier detection. Given that the median is a robust estimator of the location, and the NMAD is a robust estimator of the scale, they are often combined to detect outliers. One empirical one way for doing so, which implicitly assumes normality of the data, is to estimate the location of the data using median, and to define an outlier as those observations that depart a times the NMAD_n from that location. The value of a has to be fixed by the user, and the usual rules are: $a = 3$ (very conservative), $a = 2.5$ (moderately conservative) or even $a = 2$ (poorly conservative). This is, our population is defined as the values:

$$\text{Median}(\mathbf{x}) - a \cdot \text{NMAD}_n < x_i < \text{Median}(\mathbf{x}) + a \cdot \text{NMAD}_n,$$

and outliers are identified as the values:

$$\left| \frac{x_i - \text{Median}(\mathbf{x})}{\text{NMAD}_n} \right| \geq a.$$

Consider the following data set:

$$\{3, 5, 5, 7, 9, 11, 11, 1000\}.$$

Visually, which values do you think are outliers?

Q. What would happen if we use the mean and the standard deviation instead?

R code and answers (and additional references) to these questions can be found at:

`http://rpubs.com/FJRubio/Outlier`

Chapter 3

Interval Estimation

3.1 Confidence intervals.

In the previous chapter, we focused on point estimation of a parameter θ, based on a sample $\mathbf{X} = (X_1, \ldots, X_n)$. In this chapter, we will focus on interval estimation of this parameter based on confidence intervals. Confidence intervals of $100\gamma\%$ level, $\gamma \in (0,1)$, for a parameter $\theta \in \mathbb{R}$ are *random intervals* of the type: $(\underline{\theta}(\mathbf{X}), \overline{\theta}(\mathbf{X}))$, for which the probability of containing the unknown parameter is at least γ:

$$P[\underline{\theta}(\mathbf{X}) \leq \theta \leq \overline{\theta}(\mathbf{X})] \geq \gamma.$$

However, we will only focus on those intervals that produce equality in the previous expression. This is,

Definition 32. A $100\gamma\%$ level, $\gamma \in (0,1)$, confidence interval (CI) for θ is a *random interval* $(\underline{\theta}(\mathbf{X}), \overline{\theta}(\mathbf{X}))$

$$P[\underline{\theta}(\mathbf{X}) \leq \theta \leq \overline{\theta}(\mathbf{X})] = \gamma.$$

It is important to notice that in this definition, θ is fixed, while the end points of the CI are random. The frequentist interpretation of CI is that, if we calculate the confidence intervals associated to a sequence of samples $\mathbf{x}_1, \ldots, \mathbf{x}_m$, $(\underline{\theta}(\mathbf{x}_1), \overline{\theta}(\mathbf{x}_1)), \ldots, (\underline{\theta}(\mathbf{x}_m), \overline{\theta}(\mathbf{x}_m))$, then, the proportion of intervals that contain θ converges to γ as $m \to \infty$.

Remark 3.1.1. *In the case where the random variables X_i, $i = 1, \ldots, n$, and the left- and right-end points of the interval, are continuous, then,*

$$P[\underline{\theta}(\mathbf{X}) \leq \theta \leq \overline{\theta}(\mathbf{X})] = P[\underline{\theta}(\mathbf{X}) < \theta < \overline{\theta}(\mathbf{X})] = \gamma.$$

Thus, one can report a closed or an open interval. Typically, an open interval is reported, by convention.

However, if the random variables X_i, $i = 1, \ldots, n$, and the left- and right-end points of the interval, are discrete, then, in general

$$P[\underline{\theta}(\mathbf{X}) \leq \theta \leq \overline{\theta}(\mathbf{X})] \neq P[\underline{\theta}(\mathbf{X}) < \theta < \overline{\theta}(\mathbf{X})].$$

Thus, one has to report the kind or intervals (open or closed) that provide the correct probability ($= \gamma$ or $\geq \gamma$, depending on the definition).

Warning 1. A common confusion is to interpret CIs as intervals with probability γ of containing the true value θ, for a particular sample $\mathbf{x}$. This interpretation is *incorrect*. The interpretation of CIs has to be made in terms of repeated sampling as discussed in the previous paragraph.

Another type of CIs that are often of interest, specially in complex models, are CIs with **asymptotic confidence**:

$$P[\underline{\theta}(\mathbf{X}) \leq \theta \leq \overline{\theta}(\mathbf{X})] \to \gamma, \quad n \to \infty.$$

Below, we study some classical examples of CIs for different types of parameters.

3.2 Confidence Intervals for Means

Example 3.2.1. (Known variance). Let $\mathbf{X} = (X_1, \ldots, X_n)$ be *i.i.d.* samples from a $N(\mu, 1)$. The aim is to construct a 95% confidence interval for the unknown parameter μ. As we have discussed previously, $\bar{\mathbf{X}} \sim N\left(\mu, \frac{1}{n}\right)$, and $\sqrt{n}(\bar{\mathbf{X}} - \mu) \sim N(0, 1)$. Thus, if $z_1 < z_2 \in \mathbb{R}$ satisfy

$$\Phi(z_2) - \Phi(z_1) = 0.95,$$

where Φ is the standard normal cdf, then

$$P\left[z_1 < \sqrt{n}(\bar{\mathbf{X}} - \mu) < z_2\right] = 0.95.$$

Rearranging terms we obtain (with some care on the signs)

$$P\left[\bar{\mathbf{X}} - \frac{z_2}{\sqrt{n}} < \mu < \bar{\mathbf{X}} - \frac{z_1}{\sqrt{n}}\right] = 0.95.$$

Therefore, $\left(\bar{\mathbf{X}} - \frac{z_2}{\sqrt{n}}, \bar{\mathbf{X}} - \frac{z_1}{\sqrt{n}}\right)$ is a 95% CI for μ. One remaining issue is that the values z_1, z_2 is not unique, as there are infinitely many pairs that cumulate 95% of the mass of the standard normal cdf within (z_1, z_2). In practice, one usually is interested in the shortest CI. Given that the standard normal distribution is symmetric about 0, then, the shortest interval is the one that satisfies (the proof follows by using Lagrange multipliers) $\phi(z_1) = \phi(z_2)$, where ϕ is the standard normal pdf. This in turn implies that $z = z_2 = -z_1$. Thus, z_1 is the 2.5% quantile of the standard normal distribution which, either using software or probability tables (old technique), we obtain $z = 1.96$, and the 95% confidence interval for μ is $\left(\bar{\mathbf{X}} - \frac{1.96}{\sqrt{n}}, \bar{\mathbf{X}} + \frac{1.96}{\sqrt{n}}\right)$.

This sort of intervals are often referred to as "Z-intervals".

Remark 3.2.1. *We can use Z-intervals to construct confidence intervals for μ for samples that are not necessarily normally distributed as long as the central limit theorem applies. The reason for this is that the quantity used to construct the Z-intervals is the same that appears in the CLT.*

Example 3.2.2. (Unknown variance). Let $\mathbf{X} = (X_1, \ldots, X_n)$ be *i.i.d.* samples from a $N(\mu, \sigma^2)$. The aim is to construct a 95% confidence interval for the unknown parameter μ.

We will use the classical results:

$$\begin{aligned} \frac{\sum_{i=1}^n (X_i - \bar{\mathbf{X}})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} &\sim \chi^2_{n-1}, \\ \bar{\mathbf{X}} &\sim N\left(\mu, \frac{\sigma^2}{n}\right), \end{aligned}$$

where $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{\mathbf{X}})^2$. Then,

$$\frac{\frac{\bar{\mathbf{X}} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma^2}}{n-1}}} = \frac{\bar{\mathbf{X}} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1},$$

where t_{n-1} is a Student's t-distribution with $n-1$ degrees of freedom. Thus, if $z_1 < z_2 \in \mathbb{R}$ satisfy

$$F_{n-1}(z_2) - F_{n-1}(z_1) = 0.95,$$

where F_{n-1} is the Student's t-distribution with $n-1$ degrees of freedom cdf, then

$$P\left[z_1 < \frac{\bar{\mathbf{X}} - \mu}{\frac{S}{\sqrt{n}}} < z_2\right] = 0.95.$$

Rearranging terms, and using that the Student's t-distribution is symmetric, a 95% CI for μ is

$$\left(\bar{\mathbf{x}} - Q^{n-1}_{0.975}\frac{S}{\sqrt{n}}, \bar{\mathbf{x}} + Q^{n-1}_{0.975}\frac{S}{\sqrt{n}}\right),$$

where $Q^{n-1}_{0.975}$ is the 0.975 quantile of the Student's t-distribution with $n-1$ degrees of freedom.

Recall that the Student's t-distribution converges to the normal distribution as $n \to \infty$.

Example 3.2.3. (matched pair t interval). Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ be two independent measurements on the same individual. This is, the pair (x_i, y_i) is obtained from the same individual. This kind or observations are often called *Paired Observations*. Suppose that the first set of measurements has mean μ_1 and the second set has mean μ_2. Suppose that we are interested in obtaining a confidence interval for $\mu = \mu_1 - \mu_2$. A strategy to do so consists of analysing the differences and apply the procedure from the previous example (assuming normality of the differences). This is, define $\mathbf{d} = \mathbf{x} - \mathbf{y} = (x_1 - y_1, \ldots, x_n - y_n)$. Then, we obtain the confidence interval for μ as :

$$\left(\bar{\mathbf{d}} - Q^{n-1}_{0.975}\frac{S_d}{\sqrt{n}}, \bar{\mathbf{d}} + Q^{n-1}_{0.975}\frac{S_d}{\sqrt{n}}\right),$$

where $S_d^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{\mathbf{d}})^2$.

A key assumption in this procedure is that the observations are paired. The pairing is defined by the design used to obtain such sample. Note that if the observations are not paired, then we cannot calculate the differences in a meaningful way.

In many cases, the data are not paired. The following examples illustrate how to construct confidence intervals for the difference of means of the two populations. An example with real data can be found at:

`http://rpubs.com/FJRubio/DarwinPaired`

Example 3.2.4. (two-sample Z-interval). Let $\mathbf{X} = (X_1, \ldots, X_m)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be two *independent* samples from a normal distribution, where the X_i's have mean μ_1 and variance σ_1^2, and the Y_i's have mean μ_2 and variance σ_2^2. Then, the difference of sample means $\bar{\mathbf{D}} = \bar{\mathbf{X}} - \bar{\mathbf{Y}}$ is normally distributed with mean $\mu = \mu_1 - \mu_2$ and variance

$$\sigma^2 = \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}.$$

Therefore, the random variable

$$Z = \frac{\bar{\mathbf{D}} - \mu}{\sigma} = \frac{(\bar{\mathbf{X}} - \bar{\mathbf{Y}}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{m} + \dfrac{\sigma_2^2}{n}}},$$

is a standard normal random variable. If σ_1^2 and σ_2^2 are known, then we obtain the $100(1-\gamma)\%$ confidence intervals for μ (compare them to the Z-intervals):

$$(\bar{\mathbf{X}} - \bar{\mathbf{Y}}) \pm Z_{\frac{1-\gamma}{2}}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}.$$

Note that this construction relies on knowing the variances and independence of the two samples. If the variances are unknown, and their values are replaced by point estimators, then the distribution of the random variable Z is not necessarily normal anymore.

Example 3.2.5. (two-sample t-interval: equal variance). Consider the scenario in the previous example. If we know that $\sigma_1^2 = \sigma_2^2$, then we can pool the data to compute the standard deviation. Let S_1^2 and S_2^2 be the sample variances from the two samples. Then, the **pooled sample variance** S_p^2 is the weighted average of the sample variances with weights equal to their respective degrees of freedom. This is,

$$S_p^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}.$$

This gives the statistic

$$T = \frac{\bar{\mathbf{X}} - \bar{\mathbf{Y}} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}},$$

which has a t distribution with $n_1 + n_2 - 2$ degrees of freedom. Thus, we have a γ-level confidence interval for μ (compare this to the t-intervals):

$$\bar{\mathbf{X}} - \bar{\mathbf{Y}} \pm t_{m+n-2,\frac{1-\gamma}{2}} S_p\sqrt{\frac{1}{m} + \frac{1}{n}}.$$

Note that the key assumption here is that the variances are the same.

Example 3.2.6. (two-sample t-interval: unequal variance). Consider the scenario in the previous example. If we do not know that $\sigma_1^2 = \sigma_2^2$, then the distribution of the statistic T is *no* longer a t distribution with $m + n - 2$ degrees of freedom. However, Welch and Satterthwaite proposed an alternative statistic (known as the *Studentised* statistic) $\tilde{T} = \dfrac{(\bar{\mathbf{X}} - \bar{\mathbf{Y}}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}}$, and showed that the distribution of this Studentised statistic is close to a t distribution with a different number of degrees of freedom. More specifically, they showed that the distribution of T can be approximated with a t distribution with *effective degrees of freedom* given by

$$\nu = \frac{\left(\frac{s_1^2}{m} + \frac{s_2^2}{n}\right)^2}{\frac{s_1^4}{m^2(m-1)} + \frac{s_2^4}{n^2(n-1)}}.$$

Giving the γ-level confidence interval

$$\bar{\mathbf{x}} - \bar{\mathbf{y}} \pm t_{\nu,\frac{1-\gamma}{2}}\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}.$$

As this is an approximation, the number of observations per group may need to be at least 40 to obtain a good approximation.

An example with real data can be found at:
`http://rpubs.com/FJRubio/Mosquitoes`

3.3 Pivotal quantities

The previous examples hint at a general method for constructing confidence intervals, based on a quantity that involves the sample and the parameter, but its distribution does not depend on them.

Definition 33. A **pivotal quantity** in general is a function $u(\mathbf{X}; \theta)$ of the observations $\mathbf{X}$ and the parameter θ that has a completely specified distribution $F(\cdot)$, that is, a distribution that does not involve any unknown parameters.

If we can construct a pivotal quantity, then we can find values $L < U$ such that, for $\gamma \in (0,1)$

$$P[L < u(\mathbf{X};\theta) < U] = \gamma,$$

which are quantiles of the distribution F, and invert this relationship to find a confidence interval for θ, as in the previous example. In many cases, it is not possible to obtain a closed-form expression for the CI, but it is still possible to find a numerical solution.

Example 3.3.1. Let $X_1, \ldots, X_n$ be *i.i.d.* random variables with distribution $N(0, \sigma^2)$. The aim is now to construct a 95% CI for the unknown parameter σ. Then, $\frac{X_i}{\sigma} \sim N(0,1)$, and (see preliminary material)

$$u(\mathbf{X};\sigma) = \sum_{i=1}^{n} \frac{X_i^2}{\sigma^2} \sim \chi_n^2.$$

Thus, $u(\mathbf{X};\sigma)$ is a pivotal quantity. In order to construct a confidence interval, we need to find the values $z_1, z_2 > 0$ such that

$$P[z_1 < u(\mathbf{X};\sigma) < z_2] = 0.95.$$

We will consider an interval with equal mass cumulated on the left of z_1 and on the right of z_2. This sort of intervals are often called *Equal Tails Confidence Interval* (ETCI). This implies that z_1 is the 2.5% quantile of the χ_n^2 distribution, and z_2 is the quantile 97.5%. The confidence interval is constructed as:

$$P\left[\sqrt{\frac{\sum_{i=1}^n X_i^2}{z_2}} < \sigma < \sqrt{\frac{\sum_{i=1}^n X_i^2}{z_1}}\right] = 0.95.$$

In order to illustrate this method, let us assume that $n = 30$. Then, $z_1 = 16.79$, and $z_2 = 46.98$. So, the 95% confidence interval is $\left(\sqrt{\frac{\sum_{i=1}^n X_i^2}{46.98}}, \sqrt{\frac{\sum_{i=1}^n X_i^2}{16.79}}\right)$

Q: How would you construct a CI for σ^2?

Now, we will study asymptotic confidence intervals. This kind of CI are perhaps the most commonly used in statistical applications. We will start with a particular example, and then move to a more general method. We will require the following result, which is a consequence of Slustsky's theorem.

Theorem 3.3.1. Plug-in principle *(one parameter). Let $\widehat{\theta}_n$ be an estimator of θ that satisfies:*

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2), \quad \textit{as } n \to \infty.$$

Let $\widehat{\sigma}$ be a consistent estimator of σ. Then,

$$\frac{\sqrt{n}(\widehat{\theta}_n - \theta)}{\widehat{\sigma}} \xrightarrow{d} N(0,1), \quad \textit{as } n \to \infty.$$

This result is used to argue that, for large enough samples,

$$\frac{\sqrt{n}(\widehat{\theta}_n - \theta)}{\widehat{\sigma}} \dot{\sim} N(0,1).$$

where $\dot{\sim}$ denotes "is approximately distributed as". Thus, a $100\gamma\%$ confidence interval for θ can be obtained as

$$\widehat{\theta}_n - Z_{\frac{1+\gamma}{2}} \frac{\widehat{\sigma}}{\sqrt{n}} < \theta < \widehat{\theta}_n + Z_{\frac{1+\gamma}{2}} \frac{\widehat{\sigma}}{\sqrt{n}},$$

where Z is the $\dfrac{1+\gamma}{2}$ quantile of the standard normal distribution.

The following example is an application of this result.

Example 3.3.2. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be *i.i.d.* random variables with distribution $Bernoulli(\theta)$, $\theta \in (0, 1)$. The aim now is to construct an asymptotic (approximate) 95% confidence interval.

We know that the MLE is $\widehat{\theta} = \bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and $Var[\bar{\mathbf{X}}] = \frac{\theta(1-\theta)}{n}$. The weak law of large numbers implies that $\widehat{\theta}$ is a consistent estimator of θ. We can apply the Central Limit Theorem to obtain

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N\left(0, \theta(1-\theta)\right).$$

Using the plug-in principle, for large enough samples we have,

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

Let $Z_{0.975} = 1.96$ be the 97.5% quantile of the standard normal distribution. Then,

$$P\left[-Z_{0.975} < \frac{\sqrt{n}(\widehat{\theta} - \theta)}{\sqrt{\widehat{\theta}(1-\widehat{\theta})}} < Z_{0.975}\right] \approx 0.95,$$

Rearranging terms we obtain:

$$P\left[\widehat{\theta} - Z_{0.975}\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}} < \theta < \widehat{\theta} + Z_{0.975}\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}\right] \approx 0.95,$$

Thus, for n large, an approximate (asymptotic) 95% normal confidence interval for θ is

$$\left(\widehat{\theta} - Z_{0.975}\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}, \widehat{\theta} + Z_{0.975}\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}\right)$$

Note that we have been using $\gamma = 0.95$, for illustrative purposes, but any other values of $\gamma \in (0, 1)$ can be used instead.

A similar idea can be applied to transformations of an estimator using the Delta Method (previously studied). Recall that the Delta method states that for a consistent and asymptotically normal estimator $\widehat{\theta}_n$ of θ,

$$\sqrt{n}\left[\varphi(\widehat{\theta}_n) - \varphi(\theta)\right] \xrightarrow{d} N\left(0, [\varphi'(\theta)]^2\sigma^2\right),$$

where $\sigma^2 = \sigma^2(\theta)$, depends on θ. In other words, for large n

$$\varphi(\widehat{\theta}_n) \dot{\sim} N\left(\varphi(\theta), \frac{[\varphi'(\theta)]^2\sigma^2}{n}\right).$$

Using a similar argument as in the plug-in principle theorem, we obtain

$$\frac{\sqrt{n}\left[\varphi(\widehat{\theta}_n) - \varphi(\theta)\right]}{|\varphi'(\widehat{\theta}_n)|\sigma(\widehat{\theta}_n)} \dot{\sim} N\left(0, 1\right),$$

where the idea is that we replace θ by $\widehat{\theta}$ in all instances. Using this result, a $100\gamma\%$ confidence interval for $\varphi(\theta)$ can be obtained as

$$\varphi(\widehat{\theta}_n) - Z_{\frac{1+\gamma}{2}}\frac{|\varphi'(\widehat{\theta}_n)|\sigma(\widehat{\theta}_n)}{\sqrt{n}} < \varphi(\theta) < \varphi(\widehat{\theta}_n) + Z_{\frac{1+\gamma}{2}}\frac{|\varphi'(\widehat{\theta}_n)|\sigma(\widehat{\theta}_n)}{\sqrt{n}},$$

where Z is the $\dfrac{1+\gamma}{2}$ quantile of the standard normal distribution.

Example 3.3.3. Let X be a random variable with binomial distribution with n trials and $\theta \in (0,1)$ probability of success. The MLE of θ, $\widehat{\theta} = \dfrac{X}{n}$, $E[\widehat{\theta}] = \theta$, $Var[\widehat{\theta}] = \frac{\theta(1-\theta)}{n}$. We know that the MLE is consistent and asymptotically normal.

A parameter of interest is the log-odds, which is the logarithm of the odds:

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

This function is also known as the *logit* function. The aim now is to construct a $100\gamma\%$ CI for $\varphi(\theta)$, using the consistent estimator $\varphi(\widehat{\theta})$ and the plug-in principle. It is easy to check that $\varphi'(\widehat{\theta}) = \dfrac{1}{\widehat{\theta}(1-\widehat{\theta})}$. Then, a $100\gamma\%$ confidence interval for $\varphi(\theta)$ can be obtained as

$$\varphi(\widehat{\theta}_n) - \frac{Z_{\frac{1+\gamma}{2}}}{\sqrt{n\widehat{\theta}(1-\widehat{\theta})}} < \varphi(\theta) < \varphi(\widehat{\theta}_n) + \frac{Z_{\frac{1+\gamma}{2}}}{\sqrt{n\widehat{\theta}(1-\widehat{\theta})}}.$$

The variance of the asymptotic normal approximation is the variance of the normal distribution that approximates the true distribution of the estimator. It is important to understand that this is not the variance of the distribution of the estimator. In fact, we can see that $\widehat{\theta}$ can be 0 or 1 with positive probability (why?). Then, the logit can be equal to $-\infty$ or ∞ with positive probability. However, this probability goes to zero as $n \to \infty$, which provides an interpretation of why this approximation works for large enough samples.

Historical Note 5. Interest on the odds dates back to the times of William Shakespeare:

"Knew that we ventured on such dangerous seas. That if we wrought out life 'twas ten to one"

–William Shakespeare,

Remark 3.3.1. *A natural question is: what does the phrase "for large samples" mean?* $n = 10, 30, 100, 10^6$*?*

There is no absolute answer to this question as it depends on the model. We have seen some models where the estimator is normal for any finite sample size. There exist models where the distribution of the estimator of a parameter (or a function of the parameter) converges quickly to a normal distribution, in the sense that $n \geq 10$ *might suffice to get a good normal approximation of the CIs. However, there also exist models where this convergence is very slow (in the sense that very large samples are necessary to obtain an accurate normal approximation).*

In practice, what statisticians do in order to empirically analyse the sample size required for a good normal approximation is a **Simulation Study**. *The idea is to simulate* N *(*e.g. 1000 *or more) samples of size* n *from a model with parameter* θ_0*, and calculate the corresponding CIs of interest for each sample. Then, check the proportions of those CIs that contain* θ_0*. The law of large numbers implies that this proportion is the probability that those CIs contain the true value of the parameter, thus it is an approximation of the confidence of those intervals. If that proportion is close to the "nominal value" (*e.g. *95%), then we can say that the normal approximation is good. In contrast, if it is far from the nominal value, it means that the normal approximation is not good for that sample size and that parameter value* θ*. This kind of simulations are also known as Monte Carlo studies.*

Simulations studies are typically easier to implement than proving that the normal approximation is good for finite samples with formal mathematical tools. They do not replace a formal mathematical proof, but they represent a very useful tool to check the performance of asymptotic confidence intervals (or any other kind of CIs).

A simulation study (implemented in R) to check the performance of the normal CIs in Example 3.3.3 can be found at:

`http://www.rpubs.com/FJRubio/CILO`

Multiparameter case

Consider now that we want to construct confidence intervals for each of the entries of a vector parameter $\boldsymbol{\theta}_0$ associated to a statistical model $f(\cdot;\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$, based on a sample $\mathbf{X} = (X_1, \ldots, X_n)$.

First, we have discussed in the previous chapter that the MLE, $\widehat{\boldsymbol{\theta}}_n$, is consistent and asymptotically normal under some regularity conditions. This is, $\widehat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$, and

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N\left(0, I^{-1}(\boldsymbol{\theta}_0)\right).$$

where $I^{-1}(\boldsymbol{\theta}_0)$ is the inverse of the Fisher information matrix. This result cannot be used directly to construct confidence intervals for $\boldsymbol{\theta}_0$ since we do not know the true parameter value $\boldsymbol{\theta}_0$. However, $\widehat{\boldsymbol{\theta}}_n \approx \boldsymbol{\theta}_0$. The FIM can be approximated as $I(\boldsymbol{\theta}) \approx I(\widehat{\boldsymbol{\theta}})$, but in many cases calculating the Fisher information matrix can be challenging as it involves calculating many multivariate integrals (which may not be available in closed form). However, the following double approximation can be used. First, note that, by the law of large numbers, the average of the Hessian of the log-likelihood converges to the Fisher information matrix. This is,

$$\begin{aligned}
\frac{1}{n}\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta} \mid \mathbf{X}) &= \frac{1}{n}\nabla^2_{\boldsymbol{\theta}}\sum_{i=1}^{n}\log f(X_i;\boldsymbol{\theta}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\nabla^2_{\boldsymbol{\theta}}\log f(X_i;\boldsymbol{\theta}) \\
&\xrightarrow{P} E[\nabla^2_{\boldsymbol{\theta}}\log f(X_1;\boldsymbol{\theta})] = -I(\boldsymbol{\theta}).
\end{aligned}$$

Thus, for large n, $-\frac{1}{n}\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta} \mid \mathbf{X}) \approx I(\boldsymbol{\theta})$. Let us denote $i(\boldsymbol{\theta}) = -\nabla^2_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta} \mid \mathbf{X})$. Then, we have that, for large n,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \dot{\sim} N\left(0, ni^{-1}(\widehat{\boldsymbol{\theta}})\right).$$

Equivalently,

$$\widehat{\boldsymbol{\theta}}_n \dot{\sim} N\left(\boldsymbol{\theta}_0, i^{-1}(\widehat{\boldsymbol{\theta}})\right).$$

Thus, for each entry of $\boldsymbol{\theta}_0$, we can construct an asymptotic normal confidence interval of level $100\gamma\%$ as we did previously, using

$$[\widehat{\boldsymbol{\theta}}_n]_j - Z_{\frac{1+\gamma}{2}}[i^{-1}(\widehat{\boldsymbol{\theta}})]^{\frac{1}{2}}_{jj} < [\boldsymbol{\theta}_0]_j < [\widehat{\boldsymbol{\theta}}_n]_j + Z_{\frac{1+\gamma}{2}}[i^{-1}(\widehat{\boldsymbol{\theta}})]^{\frac{1}{2}}_{jj},$$

where $j = 1, \ldots, p$. In order to obtain 95% CIs, $Z_{\frac{1+\gamma}{2}} = 1.96$. This process can be automatised using any numerical software, such as R. In fact, the Hessian matrix is often calculated using numerical derivatives (which are already implemented in some R packages), and the inverse is also calculated using numerical methods. This avoids the need for calculating, potentially cumbersome, second derivatives and inverse matrices.

An example about the calculation of 95% CIs for the parameters in a logistic regression model can be found at:

`http://rpubs.com/FJRubio/Challenger`

Remark 3.3.2. *In many cases, the entries of parameter $\boldsymbol{\theta}$ are a combination of positive $\theta_j > 0$ and real parameters $\theta_k \in \mathbb{R}$. For example, the normal distribution with parameters $\boldsymbol{\theta} = (\mu, \sigma)$. In such cases, the asymptotic normal CIs may not be good (the coverage is not close to the nominal value). In practice,*

statisticians usually transform the positive parameters using the logarithm. This is, the model is reparameterised in order to map all the parameters to $\mathbb{R}^p$. For instance we can use the reparameterisation $\eta = \log(\sigma)$, and calculate asymptotic normal CIs. Then, we can obtain the corresponding CIs for the original parameterisation by transforming back the end points of the CIs. The justification for doing so is

$$\begin{aligned} P(L < \theta < U) &= P\left(\log(L) < \log(\theta) < \log(U)\right), \\ P(\tilde{L} < \log(\theta) < \tilde{U}) &= P\left(\exp(\tilde{L}) < \theta < \exp(\tilde{U})\right). \end{aligned}$$

Therefore, if we have a $100\gamma\%$ confidence interval for $\log(\theta)$ $(\tilde{L}, \tilde{U})$, the corresponding $100\gamma\%$ confidence interval for θ is $(\exp(\tilde{L}), \exp(\tilde{U}))$.

Warning 2. Suppose that a parameter $\theta_j > 0$, and that you calculate an asymptotic normal CI for this parameter without log-transforming it. Then, particularly for small samples, there is no guarantee that the CI will only contain positive values, which makes no sense from a probabilistic perspective. However, even in this case, the probability that an asymptotic normal CI associated to a positive parameter contains negative values converges to zero as the sample size increases.

3.4 Likelihood-Confidence intervals: The Profile Likelihood

In many cases, a statistical model may contain a large number of parameters $\boldsymbol{\theta} \in \mathbb{R}^p$. However, only one or two may be of interest. The remaining parameters are often called "nuisance parameters". The fact that the nuisance parameters are not of (immediate) interest does not imply that they are not relevant to draw inference about the parameters of interest, though. For instance, in location scale models $\boldsymbol{\theta} = (\mu, \sigma)$, the location parameter μ is often of more interest than the parameter σ. In such case, the parameter μ is the parameter of interest, and the parameter σ is the nuisance parameter. In other cases, the parameter σ might be of interest, and μ a nuisance parameter. Another situation is where we are interested in analysing one parameter at a time, while the remaining parameters are unknown. This idea of studying specific parameters of interest is known as "elimination of parameters". An essential point to understand is that *likelihood functions are not densities*. Likelihood functions are functions of the parameters of a density function, so there is no guarantee that this inversion of roles will produce a density function. Following this argument, we cannot eliminate the parameters by integrating them out, as such integral might be infinite. Thus, it is necessary to come up with alternative approaches that overcome this issue. A popular alternative is the so called "Profile Likelihood" method.

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a sample from a distribution with density $f(\cdot; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. Suppose that $\boldsymbol{\theta} = (\boldsymbol{\delta}, \boldsymbol{\xi})$, where $\boldsymbol{\delta} \in \mathbb{R}^{p_1}$ is a parameter of interest, and $\boldsymbol{\xi} \in \mathbb{R}^{p_2}$ is a nuisance parameter, where $p = p_1 + p_2$. Let:

$$L(\boldsymbol{\theta} \mid \mathbf{x}) = L(\boldsymbol{\delta}, \boldsymbol{\xi} \mid \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}),$$

be the likelihood function of $\boldsymbol{\theta}$. Let $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\xi}})$ be the MLE of $\boldsymbol{\theta}$. Then, the *profile likelihood* (or profile likelihood ratio) of $\boldsymbol{\delta}$ is:

$$R(\boldsymbol{\delta} \mid \mathbf{x}) = \frac{L(\boldsymbol{\delta}, \widehat{\boldsymbol{\xi}}(\boldsymbol{\delta}) \mid \mathbf{x})}{L(\widehat{\boldsymbol{\theta}} \mid \mathbf{x})} = \frac{\max_{\boldsymbol{\xi}} L(\boldsymbol{\delta}, \boldsymbol{\xi} \mid \mathbf{x})}{\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta} \mid \mathbf{x})},$$

where $\max_{\xi} L(\boldsymbol{\delta}, \boldsymbol{\xi} \mid \mathbf{x})$ means that we are maximising the likelihood with respect to $\boldsymbol{\xi}$, for a fixed value of $\boldsymbol{\delta}$. Then, profile likelihood is bounded $0 < R(\boldsymbol{\delta} \mid \mathbf{x}) \leq 1$.

A natural question now is: what are the properties of the profile likelihood? If I calculate an interval based on this likelihood, does it have a confidence level associated to it? The following result can be used to answer these questions.

Theorem 3.4.1. *(Wilk's theorem). Suppose that $(\boldsymbol{\delta}_0, \boldsymbol{\xi}_0)$ are the true values of the parameters. Under similar conditions to those that guarantee consistency and asymptotic normality of the MLE:*

$$\Lambda(\boldsymbol{\delta}_0) = -2\left[\log L(\boldsymbol{\delta}_0, \widehat{\boldsymbol{\xi}}(\boldsymbol{\delta}_0) \mid \mathbf{x}) - \log L(\widehat{\boldsymbol{\theta}} \mid \mathbf{x})\right] \xrightarrow{d} \chi^2_{p_1}.$$

Now, suppose that the dimension of the parameter of interest is one ($p_1 = 1$). Then,

$$\Lambda(\delta_0) = -2\log R(\delta_0 \mid \mathbf{x}) \dot{\sim} \chi^2_1.$$

Thus, $\Lambda(\delta_0)$ can be seen as an approximate pivotal quantity which may be used to construct confidence intervals for δ_0. Let $Q_p(1-\alpha)$ be the $1-\alpha$ quantile of a χ^2_p distribution (with p degrees of freedom), $\alpha \in (0,1)$. Then,

$$P[\Lambda(\delta_0) \leq Q_1(1-\alpha)] \approx 1-\alpha.$$

Then, we say that values of δ for which $\Lambda(\delta) \leq Q_1(1-\alpha)$ may be regarded as plausible at the $(1-\alpha)$ level. This is equivalent to saying:

$$R(\delta) \geq \exp\left(-\frac{1}{2}Q_1(1-\alpha)\right).$$

If we want to obtain an approximate 95% interval for δ ($\alpha = 0.05$), then $Q_1(0.95) = 3.84$, and the approximate CI corresponds to the values that satisfy $R(\delta) \geq 0.147$. This is an interesting and powerful result as it allows for conducting point and interval estimation of a parameter using the likelihood function for both purposes. Moreover, opposite to the construction of confidence interval based on asymptotic normality, approximate CIs based on the likelihood function lie within the parameter space.

The points that satisfy $R(\delta) \geq 0.147$ are typically within an interval (since the profile likelihood has, in most cases, a unique maximum). The calculation of the end points of such intervals typically requires numerical methods. This is, we need to find the two end points that satisfy $R(\delta) = 0.147$, or equivalently, to find the roots of the function $f(\delta) = R(\delta) - 0.147$. General method for finding the roots of a univariate function are: bisection, Newton-Raphson, Brent's, the secant methods, among others. In R, the command 'uniroot()' implements Brent's method.

Since this method applies to any likelihood function, under general regularity conditions, it means that it can be applied to a broad kind of statistical models.

Example 3.4.1. Consider the a normal sample $\mathbf{x}$ of size n. The likelihood function is (up to a proportionality constant):

$$L(\mu, \sigma) \propto \frac{1}{\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

Then, the MLEs are $(\widehat{\mu}, \widehat{\sigma}) = (\bar{\mathbf{x}}, s)$, with $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$.

- Suppose that the parameter of interest is the standard deviation σ, while the mean of the population μ is the nuisance parameter. As we discussed before, the MLE of μ is $\widehat{\mu} = \bar{\mathbf{x}}$ for any value of σ (that is, the MLE of μ is independent of σ). Then, replacing these values we obtain:

$$R(\sigma) = \frac{\max_\mu L(\mu, \sigma)}{L(\widehat{\mu}, \widehat{\sigma})} = \left(\frac{\widehat{\sigma}}{\sigma}\right)^n \exp\left[\frac{n}{2}\left(1 - \left(\frac{\widehat{\sigma}}{\sigma}\right)^2\right)\right].$$

- Suppose that the parameter of interest is the mean of the population μ and the nuisance parameter is σ. As we discussed before, the MLE of σ, for a fixed value of μ, is $\widehat{\sigma}(\mu) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}$ (that is, the MLE of σ depends on μ). Then, replacing these values we obtain:

$$R(\mu) = \frac{\max_\sigma L(\mu, \sigma)}{L(\widehat{\mu}, \widehat{\sigma})} = \left(\frac{\widehat{\sigma}}{\widehat{\sigma}(\mu)}\right)^n = \left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\mu)^2}\right)^{\frac{n}{2}}.$$

An illustration of this example in R can be found at:

`http://www.rpubs.com/FJRubio/PLCIN`

Example 3.4.2. Challenger data.

The NASA space shuttle "Challenger" exploded shortly after its launch on 28 January 1986, with a loss of seven lives. The US Presidential Commission concluded that the accident was caused by leakage of gas from one of the fuel-tanks. Rubber insulating rings, so-called "O-rings", were not pliable enough after the overnight low temperature of 31°F, and did not plug the joint between the fuel in the tanks and the intense heat outside. Table 3.1 presents the data concerning previous flights which has been slightly modified, for illustration purposes, by only reporting the presence of failure of the O-rings. The last row corresponds to the conditions at which the Challenger was launched.

	Failure	Temperature	Pressure (psi)
1	0	66	50
2	1	70	50
3	0	69	50
4	0	68	50
5	0	67	50
6	0	72	50
7	0	73	100
8	0	70	100
9	1	57	200
10	1	63	200
11	1	70	200
12	0	78	200
13	0	67	200
14	1	53	200
15	0	67	200
16	0	75	200
17	0	70	200
18	0	81	200
19	0	76	200
20	0	79	200
21	1	75	200
22	0	76	200
23	1	58	200
C	–	31	200

Table 3.1: Challenger data. Failures out of 6 rings.

Let's now "fit" a logistic regression using maximum likelihood estimation. This model contains three parameters $\boldsymbol{\beta} = (\beta_0, \beta_T, \beta_P)$, which correspond to the intercept, the coefficient associated to the variable "temperature", and the coefficient associated to the variable "pressure". Let $\mathbf{x}_i$ denote the vector of covariates $\mathbf{x}_i = (1, T_i, P_i)^\top$. The model is

$$P[\text{Failure}_i = 1] = \frac{\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}{1 + \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)} \quad i = 1, \ldots, 23.$$

The MLE of $\boldsymbol{\beta}$ is, $\widehat{\boldsymbol{\beta}} = (2.520, -0.098, 0.008)^\top$. The CIs (based on asymptotic normality and Profile Likelihood) for each of these parameters are presented in the following tables, and the code can be found in the RPUBS link at the end of this example.

	2.5 %	97.5 %
Intercept	-4.3229	9.7726
temperature	-0.1941	-0.0136
pressure	-0.0043	0.0289

Table 3.2: CIs based on the profile likelihood.

	2.5 %	97.5 %
Intercept	-4.3139	9.3543
temperature	-0.1863	-0.0103
pressure	-0.0066	0.0235

Table 3.3: CIs based on asymptotic normality.

The probability of failure of the O-rings under the conditions of that day $\mathbf{x}_C = (1, 31, 200)^\top$ are:

$$P[\text{Failure}_C = 1] = \frac{\exp\left(\mathbf{x}_C^\top \widehat{\boldsymbol{\beta}}\right)}{1 + \exp\left(\mathbf{x}_C^\top \widehat{\boldsymbol{\beta}}\right)} = 0.76.$$

Note that NASA only had experience launching flights at higher temperatures, the minimum being $53°$F. At this temperature, the probability of failure of the O-rings is 0.27, which is still high.

Although this approach is often questionable as we need to extrapolate the probability of failure to temperatures far from what they had observed before, and there are more accurate methods nowadays (for instance, that consider dependence between the failure of the O-rings). However, a quick analysis like this (or using the probit link) would have warned them about the risk of launching the Challenger at that temperature.

The R code used to produce these calculations can be found at:

`http://rpubs.com/FJRubio/Challenger`

An article about the Challenger disaster can be found at:

`https://www.history.com/topics/1980s/challenger-disaster`

3.5 Second order approximation: Wald confidence intervals

Consider a second order approximation of Λ around the MLE $\widehat{\delta}$:

$$\begin{aligned}\Lambda(\delta) = -2\log R(\delta \mid \mathbf{x}) &\approx \Lambda(\widehat{\delta}) - 2\frac{R'(\widehat{\delta} \mid \mathbf{x})}{R(\widehat{\delta} \mid \mathbf{x})}\left(\delta - \widehat{\delta}\right) + \frac{1}{2}\Lambda''(\widehat{\delta})\left(\delta - \widehat{\delta}\right)^2 \\ &= \frac{1}{2}\Lambda''(\widehat{\delta})\left(\delta - \widehat{\delta}\right)^2,\end{aligned}$$

where we are using that the profile likelihood is maximised at $\widehat{\delta}$, then $R(\widehat{\delta} \mid \mathbf{x})$ and $R'(\widehat{\delta} \mid \mathbf{x}) = 0$. Compare this quantity with the quantity that we have studied in the context of asymptotic normality of the MLE, and notice that this is the square of such quantity. Using a formal argument, it is possible to prove that this quantity converges in distribution to a χ^2_1. Thus, for large enough n,

$$\frac{1}{2}\Lambda''(\widehat{\delta})\left(\delta - \widehat{\delta}\right)^2 \dot{\sim} \chi^2_1.$$

This provides an alternative way of approximating a confidence interval of level $100(1-\alpha)\%$, by calculating the values of δ that satisfy $\left(\delta - \widehat{\delta}\right)^2 \leq \dfrac{2Q_1(1-\alpha)}{\Lambda''(\widehat{\delta})}$. This requires the calculation of the

second derivative of the profile likelihood. In the case of the normal distribution, these derivatives can be found explicitly. However, for most models the second derivative cannot be obtained in closed-form, and the use of numerical methods is necessary in order to calculate this term.

An illustration of this method in R can be found at:

`http://www.rpubs.com/FJRubio/WCIN`

Compare the confidence intervals obtained with this method with those obtained with the Profile likelihood method. The data sets used in these example were generated using the same "seed", and thus they are the same data. Fixing the seed in simulations is useful in order to allow for reproducibility of the results. Which CIs are better and why?

Chapter 4

Hypothesis tests

Historical Note 6. The modern terminology of hypothesis testing was largely created by Ronald A. Fisher and the team of Jerzy Neyman and Egon S. Pearson in the 1920s and -30s. These parties disagreed about the theory of testing but their terminologies are blended in modern Statistics.

4.1 Introduction

Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be random *i.i.d.* random variables with pdf (pmf) $f(\cdot;\theta)$, and $\theta \in \Theta$. Suppose that the value $\theta = \theta_0$ is of particular interest.

Example 4.1.1. For instance, θ may represent the probability of success in tossing a coin n times (a sequence of Bernoulli trials), and we may be interested in checking whether or not the coin is fair, meaning $\theta = \frac{1}{2}$. The alternative to the hypothesis of a fair coin is $\theta \neq \frac{1}{2}$.

Definition 34. A hypothesis is a statement about a model parameter.

This idea can be formalised into a hypothesis, which we refer to as the *Null Hypothesis*:

$$H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta \neq \theta_0.$$

The alternative values H_1 is referred to as the *Alternative Hypothesis*. This idea can be generalised as,

$$H_0 : \theta \in \Theta_0 \quad vs. \quad H_1 : \theta \in \Theta_1.$$

where $\Theta = \Theta_0 \cup \Theta_1$ ($\cup =$ union of sets) and $\Theta_0 \cap \Theta_1 = \emptyset$ (where $\emptyset$ is the empty set, and $\cap =$ intersection of sets). In some textbooks, Θ_1 is denoted as Θ_0^c, which represents the complement of Θ_0.

Remark 4.1.1. *A hypothesis (alternative) is* simple *if Θ_0 (Θ_1) is a singleton set (for instance $\Theta_0 = \{\theta_0\}$). Otherwise, it is called* composite.

Remark 4.1.2. *A hypothesis test is called a* one-sided *test when:*

- $\Theta_0 = \{\theta : \theta \leq \theta_0\}$ *and* $\Theta_1 = \{\theta : \theta > \theta_0\}$, *or*
- $\Theta_0 = \{\theta : \theta \geq \theta_0\}$ *and* $\Theta_1 = \{\theta : \theta < \theta_0\}$,

for some choice of the parameter value θ_0.

A hypothesis test is called a two-sided *test when:*

$$\Theta_0 = \{\theta_0\} \quad \textit{and} \quad \Theta_1 = \{\theta \neq \theta_0\},$$

for some choice of the parameter value θ_0.

Definition 35. The two complementary hypotheses in a hypothesis testing are called the *null hypothesis* and the *alternative hypothesis*. They are denoted by H_0 and H_1, respectively.

The aim is to assess, in the light of the data, whether there is sufficient evidence to reject the null hypothesis H_0. This is,

Definition 36. A hypothesis testing procedure or hypothesis test is a rule that specifies

(i) For which sample values the decision is made to *not reject* H_0 as true.

(ii) For which sample values H_0 is rejected and H_1 is *not rejected* as true.

The subset of the sample space for which H_0 will be rejected is called the *rejection region* or *critical region*.

It is important to notice that we can only *reject* the null hypothesis. If there is not enough evidence to reject it, we say that we *do not reject* it. Strictly speaking, it is incorrect to say that "we accept the null" when we do not reject it, as we can only gather evidence against the null, although in some textbooks you will find phrases like "we accept the null hypothesis" for pedagogical purposes. Thus, our task can be interpreted as a binary decision problem, where the possible outcomes are presented in Table 4.1.

	Decision	
	Do not Reject H_0	Reject H_0
H_0 True	✓	Type I error (false positive)
H_1 True	Type II error (false negative)	✓

Table 4.1: Possible outcomes from a hypothesis test

Let us now illustrate this idea with a non-mathematical example.

Example 4.1.2. Consider a trial at the Supreme Court, where a person is charged with a crime. We do not know whether or not the accused person actually committed the crime, so we can only gather evidence about it. Based on this evidence, we can send the accused person to prison, if the person is found guilty, or set the accused person free, if there is not enough evidence of the crime (found not guilty).

There is "presumption of innocence" (presumed innocent) unless proven guilty.

"The burden of proof is on the one who declares, not on the one who denies."

So, the trial is a hypothesis test, where the null hypothesis is that the accused person is innocent, and the alternative hypothesis is that the accused person is guilty.

	Person is Innocent	Person is Guilty
Decision: Person Innocent	✓	Type II error
Decision: Person Guilty	Type I error	✓

Table 4.2: Possible outcomes from a trial

(i) Reducing Type I error means that we want to reduce the probability of sending innocent persons to prison. This implies setting free guilty persons with higher probability, as we need to be more lax about the evidence.

(ii) Reducing Type II errors means that we want to reduce the probability of setting free guilty persons. This implies that we need to be more strict with the evidence, implying that we may send innocent people to prison with a higher probability.

Q. What option do you prefer?

Next, we present the formal definition of a Hypothesis test.

Definition 37. Hypothesis testing consists of the following steps:

1. Select a **test statistic** $T_n = T(X_1, \ldots, X_n)$.
2. Select a **rejection region** R.
3. If $T_n \in R$, we reject H_0, otherwise we (retain) do not reject H_0.

Thus, in order to test a hypothesis, the ingredients are: the null and the alternative hypothesis, a test statistic, and a rejection region.

Later, we will explore general methods for testing hypothesis. As discussed in Table 4.1.2, when testing $H_0 : \theta \in \Theta_0$ *vs.* $H_1 : \theta \in \Theta_0^c$, there are two types of errors one can make: Type I Error and Type II Error. If $\theta \in \Theta_0$, but the hypothesis test incorrectly rejects H_0, then the test has made a Type I Error. On the other hand, if $\theta \in \Theta_0^c$ but the test does not reject H_0, we say that a Type II error has been made. We can also see that $P_\theta(T_n \in R) = 1 - P_\theta(T_n \in R^c)$. Here $P_\theta(A)$ denotes the probability of the event A, using the parameter value θ in the corresponding distribution. Thus,

$$P_\theta(T_n \in R) = \begin{cases} \text{probability of Type I Error} & \text{if } \theta \in \Theta_0 \\ \text{one minus the probability of Type II Error} & \text{if } \theta \in \Theta_0^c. \end{cases}$$

The probabilities of Type I and Type II errors must be understood in a Frequentist sense. For instance, if the probability Type I error is 0.05, this means that if we sample many times, we will reject H_0 5% of the times, when H_0 is true, as the number of samples (not the sample size) goes to infinity.

Example 4.1.3. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be *i.i.d.* random variables with Bernoulli distribution with parameter $\theta \in (0, 1)$. Suppose that we are interested in testing

$$H_0 : \theta = \frac{1}{2} \quad vs. \quad H_1 : \theta \neq \frac{1}{2}.$$

Let $T_n = \bar{\mathbf{X}}$ and $R = \left\{\mathbf{x} : \left|T_n - \frac{1}{2}\right| \geq \delta\right\}$, for some $\delta \in (0, 1/2)$. Thus, we reject the null hypothesis H_0 if $\left|T_n - \frac{1}{2}\right| \geq \delta$.

Another key concept in the context of hypothesis testing is the power function.

Definition 38. The *power function* of a hypothesis test with rejection region R is the function of θ defined by $\beta(\theta) = P_\theta(T_n \in R)$.

Ideally, we would like $\beta(\theta)$ to be as small as possible when $\theta \in \Theta_0$, and as close to one as possible when $\theta \in \Theta_0^c$. Qualitatively, a good test has power near one for most $\theta \in \Theta_0^c$, and near zero for most $\theta \in \Theta_0$. Concepts related to this idea are the size and the level of a test.

Definition 39. A test is size $\alpha \in (0, 1)$ if

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha.$$

A test is level α if

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha.$$

This definition is useful in cases where one cannot construct a test of size exactly equal to α.

At first sight, choosing a very low value of α sounds like a good idea, as we may want to reject the null hypothesis, when this is true, with low probability. However, given the relationship between the power function and *both* types of Errors, by selecting a low test size, we will also affect (in fact, increase) the Type II error.

Let us illustrate these concepts with an example in a more mathematical scenario.

Example 4.1.4. Let $\mathbf{X} = X_1, \ldots, X_n$, where $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, with σ^2 known. Suppose that we want to test

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu > \mu_0.$$

This sort of hypotheses are called *one-sided alternative*. Here $\Theta = [0, \infty]$ and $\Theta_0 = \{\theta_0\}$, so this formulation is still consistent with the original formulation, where the alternative hypothesis represents the complement set. Let

$$T_n = \frac{\bar{\mathbf{X}} - \mu_0}{\frac{\sigma}{\sqrt{n}}},$$

be the test statistic, and suppose that we reject H_0 if $T_n \geq c$, for some $c \in \mathbb{R}$. We know that $T_n \sim N(0, 1)$ from the properties of the normal distribution. Then, by definition

$$\beta(\mu) \quad = \quad P_\mu \left(\frac{\bar{\mathbf{X}} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq c \right)$$

After some algebra (and adding the same term on both sides of the inequality)

$$\begin{aligned} \beta(\mu) &= P\mu \left(\frac{\bar{\mathbf{X}} - \mu}{\frac{\sigma}{\sqrt{n}}} \geq c + \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} \right) \\ &= 1 - \Phi \left(c + \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}} \right), \end{aligned}$$

where Φ is the standard normal cdf. The last term reaches its maximum with respect to $\mu \in \Theta_0$ when the term $\Phi(\ldots)$ reaches its minimum, but Θ_0 is a singleton $\{\mu_0\}$. Therefore,

$$\sup_{\Theta_0} \beta(\mu) = \beta(\mu_0) = 1 - \Phi(c).$$

In order to obtain a test of size α, we need $\alpha = 1 - \Phi(c)$, and $c = \Phi^{-1}(1 - \alpha)$. This is, we need to choose c to be the $(1 - \alpha)$ quantile of the standard normal distribution, denoted as $Q_{1-\alpha}$. The test becomes: Reject H_0 when $T_n \geq Q_{1-\alpha}$.

Example 4.1.5. Let $\mathbf{X} = X_1, \ldots, X_n$, where $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$, with σ^2 known. Suppose that now we want to test

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \neq \mu_0.$$

This sort of hypotheses are called *two-sided alternative*. Here $\Theta = \mathbb{R}$ and $\Theta_0 = \{\theta_0\}$. Let

$$T_n = \frac{\bar{\mathbf{X}} - \mu_0}{\frac{\sigma}{\sqrt{n}}},$$

be the test statistic, and suppose that we reject H_0 if $|T_n| > c$, for some $c \in \mathbb{R}$. Then,

$$\begin{aligned} \beta(\mu) &= P_\mu (|T_n| \geq c) \\ &= P_\mu (T_n \geq c) + P_\mu (T_n \leq -c) \end{aligned}$$

After some algebra (using the same algebra tricks in the previous example and recalling that $\Phi(-t) = 1 - \Phi(t)$ since this cdf is symmetric)

$$\begin{aligned}
\beta(\mu) &= P\mu\left(\frac{\bar{\mathbf{X}} - \mu}{\frac{\sigma}{\sqrt{n}}} \geq c + \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) + P\mu\left(\frac{\bar{\mathbf{X}} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq -c + \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) \\
&= 1 - \Phi\left(c + \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) + \Phi\left(-c + \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) \\
&= \Phi\left(-c - \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}}\right) + \Phi\left(-c + \frac{\mu_0 - \mu}{\frac{\sigma}{\sqrt{n}}}\right),
\end{aligned}$$

Then, the size of the test is

$$\sup_{\Theta_0} \beta(\mu) = \beta(\mu_0) = 2\Phi(-c).$$

In order to obtain a test of size α, we need $\alpha = 2\Phi(-c)$, and $c = -\Phi^{-1}(\alpha/2) = \Phi^{-1}(1 - \alpha/2)$ (as ϕ, the standard normal pdf, is symmetric). This is, we need to choose c to be the $\left(1 - \frac{\alpha}{2}\right)$ quantile of the standard normal distribution, denoted as $Q_{1-\frac{\alpha}{2}}$. The test becomes: Reject H_0 when $|T_n| \geq Q_{1-\frac{\alpha}{2}}$.

4.2 Most Powerful Tests

In the previous section, we described some hypothesis tests that control the probability of Type I Error, for example, level α tests have Type I Error probabilities at most α for all $\theta \in \Theta_0$. A good test in such class would also have a small Type II Error probability, that is, a large power function for $\theta \in \Theta_0^c$. If one test had a smaller Type II Error probability than all other tests in the class, it would certainly be appealing and might be considered the best test in that class. This idea is formalised in the next definition.

Definition 40. Let $\mathcal{C}$ be a class of tests for testing

$$H_0 : \theta \in \Theta_0 \quad vs. \quad H_1 : \theta \in \Theta_0^c.$$

A test in class $\mathcal{C}$, with power function $\beta(\theta)$, is a *uniformly most powerful* (UMP) class $\mathcal{C}$ test if $\beta(\theta) \geq \beta'(\theta)$ for every $\theta \in \Theta_0^c$ and every $\beta'(\theta)$ that is a power function of a test in class $\mathcal{C}$.

We will consider $\mathcal{C}$ to be the class of all level α tests. A minimisation of the Type II Error probability without some control of the Type I Error probability is not very interesting. For instance, if we construct a test that rejects H_0 with probability one, then this test will never make a Type II Error. The requirements in the previous definition are so strict that UMP tests do not exist in many realistic problems. However, for those cases where a UMP test does exist, then that test might be considered to be the best test in that class (note that this is a very specific criterion of what is the "best" test). Then, it is desirable to identify UMP tests in those cases where they exist. The following theorem describes which tests are UMP level α tests in the situation where the null and the alternative hypothesis both consists of only one probability distribution for the sample (this is, when both $H_0$and H_1 are simple hypotheses).

Definition 41. A *test function*, $\varphi(\mathbf{x})$, for a hypothesis testing procedure is a function on the sample space whose value is one if $\mathbf{x}$ is in the rejection region, and zero if $\mathbf{x}$ is in the acceptance region. That is, $\varphi(\mathbf{x})$ is the indicator function of the rejection region.

Theorem 4.2.1. *(***Neyman-Pearson Lemma***). Consider testing*

$$H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta = \theta_1,$$

where the pdf or pmf corresponding to θ_i is $f(\mathbf{x}; \theta_i)$, $i = 0, 1$, using a test with rejection region R that satisfies

$$x \in R \text{ if } f(\mathbf{x}; \theta_1) > kf(\mathbf{x}; \theta_0),$$

and (4.2.1)

$$x \in R^c \text{ if } f(\mathbf{x};\theta_1) < kf(\mathbf{x};\theta_0),$$

for some $k \geq 0$, *and*

$$\alpha = P_{\theta_0}(\mathbf{X} \in R). \tag{4.2.2}$$

Then,

(i) (Sufficiency). Any test that satisfies (4.2.1) *and* (4.2.2) *is a UMP level* α *test.*

(ii) (Necessity). If there exists a test satisfying (4.2.1) *and* (4.2.2) *with* $k > 0$, *then every UMP level* α *test is a size* α *test (it satisfies* (4.2.2)*) and every UMP level* α *test satisfies* (4.2.1) *except perhaps on a set* A *satisfying* $P_{\theta_0}(\mathbf{X} \in A) = P_{\theta_1}(\mathbf{X} \in A) = 0$.

Proof. Note first that $f(\mathbf{x};\theta) = L(\theta \mid \mathbf{x})$ is the likelihood function. We will prove the theorem for the case that $f(\mathbf{x};\theta_0)$ and $f(\mathbf{x};\theta_1)$ are pdfs of continuous random variables. The proof for discrete random variables can be accomplished by replacing integrals with sums.

Note that any test satisfying (4.2.2) is a size α test and, hence, a level α test because

$$\sup_{\theta \in \Theta_0} P_\theta(\mathbf{X} \in R) = \alpha,$$

since Θ_0 has only one point.

Let $\varphi(\mathbf{x})$ be the test function of a test satisfying (4.2.1) and (4.2.2). Let $\varphi'(\mathbf{x})$ be the test function of any other level α test, and let $\beta(\theta)$ and $\beta'(\theta)$ be the power functions corresponding to the tests φ and φ', respectively. Because $0 \leq \varphi'(\mathbf{x}) \leq 1$, (4.2.1) implies that

$$[\varphi(\mathbf{x}) - \varphi'(\mathbf{x})][f(\mathbf{x};\theta_1) - kf(\mathbf{x};\theta_0)] \geq 0,$$

for every $\mathbf{x}$, since $\varphi = 1$ if $f(\mathbf{x};\theta_1) > kf(\mathbf{x};\theta_0)$ and $\varphi = 0$ if $f(\mathbf{x};\theta_1) < kf(\mathbf{x};\theta_0)$. Thus,

$$0 \leq \int [\varphi(\mathbf{x}) - \varphi'(\mathbf{x})][f(\mathbf{x};\theta_1) - kf(\mathbf{x};\theta_0)]d\mathbf{x} = \beta(\theta_1) - \beta'(\theta_1) - k[\beta(\theta_0) - \beta'(\theta_0)]. \tag{4.2.3}$$

The first point (i) is proved by noting that, since φ' is a level α test and φ is a size α test,

$$\beta(\theta_0) - \beta'(\theta_0) = \alpha - \beta'(\theta_0) \geq 0.$$

Thus, (4.2.3) and $k \geq 0$ imply that

$$0 < \beta(\theta_1) - \beta'(\theta_1) - k[\beta(\theta_0) - \beta'(\theta_0)] \leq \beta(\theta_1) - \beta'(\theta_1),$$

showing that $\beta(\theta_1) \geq \beta'(\theta_1)$, and hence φ has greater power than φ'. Since φ' was an arbitrary level α test and θ_1 is the only point in Θ_0^c, φ is a UMP level α test.

In order to prove point (ii), let φ' now be the test function for any UMP level α test. By part (i), φ, the test satisfying (4.2.1) and (4.2.2), is also a UMP level α test, thus $\beta(\theta_1) = \beta'(\theta_1)$. This fact, (4.2.3), and $k > 0$ imply

$$\alpha - \beta'(\theta_0) = \beta(\theta_0) - \beta'(\theta_0) \leq 0.$$

Now, since φ' is a level α test, $\beta'(\theta_0) \leq \alpha$. Thus $\beta'(\theta_0) = \alpha$, that is, φ' is a size α test, and this also implies that (4.2.3) is an equality in this case. But the nonnegative integrand $[\varphi(\mathbf{x}) - \varphi'(\mathbf{x})][f(\mathbf{x};\theta_1) - kf(\mathbf{x};\theta_0)]$ will have a zero integral only if φ' satisfies (4.2.1), except perhaps on a set A with $\int_A f(\mathbf{x};\theta_i) = 0$. This implies that (ii) is holds.

The Neyman-Pearson lemma has the following implication for tests based on a sufficient statistic.

Corollary 4.2.1. *Consider the hypothesis*

$$H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta = \theta_1.$$

Suppose that $T(\mathbf{X})$ is a sufficient statistic for θ and $g(t;\theta_i)$ is the pdf or pmf of T corresponding to θ_i, $i = 0, 1$. Then, any test based on T with rejection region S (a subset of the sample space of T) is a UMP level α test if it satisfies

$$\begin{aligned} & t \in S \ \text{ if } \ g(t;\theta_1) > kg(t;\theta_0) \\ \text{and} \qquad & \\ & t \in S^c \ \text{ if } \ g(t;\theta_1) < kg(t;\theta_0) \end{aligned} \tag{4.2.4}$$

for some $k \geq 0$, where $\alpha = P_{\theta_0}(T \in S)$.

Proof. In terms of the original sample $\mathbf{X}$, the test based on T has the rejection region $R = \{\mathbf{x} : T(\mathbf{x}) \in S\}$. By the Factorisation Theorem, the pdf or pmf of $\mathbf{X}$ can be written as

$$f(\mathbf{x};\theta_i) = g(T(\mathbf{x});\theta_i)h(\mathbf{x}), \ \ i = 0, 1,$$

for some nonnegative function $h(\mathbf{x})$. Multiplying the inequalities in (4.2.4) by this nonnegative function, we see that R satisfies

$$\mathbf{x} \in R \ \text{ if } \ f(\mathbf{x};\theta_1) = g(T(\mathbf{x});\theta_1)h(\mathbf{x}) > kg(T(\mathbf{x});\theta_0)h(\mathbf{x}) = kf(\mathbf{x};\theta_0),$$

and

$$\mathbf{x} \in R^c \ \text{ if } \ f(\mathbf{x};\theta_1) = g(T(\mathbf{x});\theta_1)h(\mathbf{x}) < kg(T(\mathbf{x});\theta_0)h(\mathbf{x}) = kf(\mathbf{x};\theta_0).$$

Also,

$$P_{\theta_0}(\mathbf{X} \in R) = P_{\theta_0}(T(\mathbf{X}) \in R) = \alpha.$$

Then, by the sufficiency part of the Neyman-Pearson Lemma (and cancelling h accordingly), the test based on T is a UMP level α test.

The following corollary indicates how to use the Neyman-Pearson Lemma for composite hypotheses.

Corollary 4.2.2. *Consider testing*

$$H_0 : \theta \in \Theta_0 \quad vs. \quad H_1 : \theta \in \Theta_0^c.$$

Suppose that a test based on a sufficient statistic T with rejection region S satisfies the following three conditions:

(i) The test is a level α test.

(ii) There exists a $\theta_0 \in \Theta_0$ such that $P_{\theta_0}(T \in S) = \alpha$.

(iii) Let $g(t;\theta)$ denote the pdf or pmf of T. For the same θ_0 as in (ii), and for each $\theta' \in \Theta_0^c$, there exists a $k' \geq 0$ such that

$$t \in S \ \text{ if } \ g(t;\theta') > k'g(t;\theta_0) \quad \text{and} \quad t \in S^c \ \text{ if } \ g(t;\theta') < k'g(t;\theta_0).$$

Then, this test is a UMP level α test of H_0 versus H_1.

Proof. Let $\beta(\theta)$ be the power function of the test with rejection region S. Fix $\theta' \in \Theta_0^c$. Consider testing

$$H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta = \theta'.$$

Corollary 4.2.1 and assumptions (i)–(iii) imply that $\beta(\theta') \geq \beta^*(\theta')$, where $\beta^*(\theta)$ is the power function of any other level α est of H_0', that is, any test satisfying $\beta(\theta_0) \leq \alpha$. However, any level α test of H_0 satisfies

$$\beta^*(\theta_0) \leq \sup_{\theta \in \Theta_0} \beta^*(\theta) \leq \alpha.$$

Thus, $\beta(\theta') \geq \beta^*(\theta')$, for any level α test of H_0. Since θ' was arbitrary, the result follows.

This corollary extends the Neyman-Pearson Lemma to composite hypotheses. However, conditions (i) and (ii) are very strong, as they imply such value θ_0 is an interior point of Θ_0. We now present some illustrative examples.

Example 4.2.1. Let $X \sim Binomial(2, \theta)$. We want to test

$$H_0 : \theta = \frac{1}{2} \quad vs. \quad H_1 : \theta = \frac{3}{4}.$$

Calculating the ratios of the pmfs gives

$$\begin{aligned} \frac{p(0;\theta = 3/4)}{p(0;\theta = 1/2)} &= \frac{1}{4} \\ \frac{p(1;\theta = 3/4)}{p(1;\theta = 1/2)} &= \frac{3}{4} \\ \frac{p(2;\theta = 3/4)}{p(2;\theta = 1/2)} &= \frac{9}{4}. \end{aligned}$$

Then,

- By choosing $\frac{3}{4} < k < \frac{9}{4}$, the Neyman-Pearson Lemma says that the test that rejects H_0 if $X = 2$ is the UMP level $\alpha = P(X = 2; \theta = 1/2) = \frac{1}{4}$ test.
- By choosing $\frac{1}{4} < k < \frac{3}{4}$, the Neyman-Pearson Lemma says that the test that rejects H_0 if $X = 1 \ \ or \ \ 2$ is the UMP level $\alpha = P(X = 1 \ \ or \ \ 2; \theta = 1/2) = \frac{3}{4}$ test.
- Choosing $k < \frac{1}{4}$ or $k > \frac{9}{4}$ yields the UMP level $\alpha = 1$ or level $\alpha = 0$ test.
- For the case that $k = \frac{3}{4}$, the Neyman-Pearson lemma says we must reject H_0 for the sample $x = 2$, and accept for H_0 for $x = 0$ but leaves our action for $x = 1$ undetermined. However, if we accept H_0 for $x = 1$, we get the UMP level $\alpha = \frac{1}{4}$ test, as discussed above. On the other hand, if we reject H_0 for $x = 1$, we get the UMP level $\alpha = \frac{3}{4}$ test as above.

This shows that, although the Neyman-Pearson Lemma can be used to identify UMP tests, the level of such tests still depends on the rejection region.

Example 4.2.2. Let $\mathbf{X} = X_1, \ldots, X_n$ be an *i.i.d.* random sample from a $N(\mu, \sigma^2)$, where σ^2 is known. The sample mean is a sufficient statistic for μ. Consider testing

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu = \mu_1,$$

where $\mu_0 > \mu_1$. The inequality $g(\bar{\mathbf{x}}; \mu_1) > kg(\bar{\mathbf{x}}; \mu_0)$ is equivalent to (after some algebraic calculations)

$$\bar{\mathbf{x}} < \frac{(2\sigma^2 \log(k))/n - \mu_0^2 + \mu_1^2}{2(\mu_1 - \mu_0)},$$

since $\bar{\mathbf{x}}$ is also normally distributed (as discussed previously) and the fact that $\mu_1 - \mu_0 < 0$. The right-hand side of this inequality increases from $-\infty$ to ∞ as k increases from 0 to ∞. Thus, by Corollary 4.2.1, the test with rejection region $\bar{\mathbf{x}} < c$ is the UMP level α test where $\alpha = P_{\mu_0}(\bar{\mathbf{X}} < c)$. If a particular α is specified, then the UMP test rejects H_0 if $\bar{\mathbf{X}} < c = \sigma Z_\alpha/\sqrt{n} + \mu_0$, where $Z_\alpha = \Phi^{-1}(\alpha)$. This choice of c guarantees the desired level α.

Example 4.2.3. Under the assumptions of the previous example, now consider testing

$$H_0 : \mu \geq \mu_0 \quad vs. \quad H_1 : \mu < \mu_0.$$

The test that rejects H_0 if

$$\bar{\mathbf{X}} < \frac{\sigma Z_\alpha}{\sqrt{n}} + \mu_0,$$

is a UMP level α test in this problem. Condition (ii) of Corollary 4.2.2 is satisfied (see previous example). Condition (iii) is true because, in the above argument, only the fact hat $\mu_1 < \mu_0$, not the exact value of μ_1, was used in determining the UMP level α test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$. Condition (i) is true because the power function of this test,

$$\beta(\mu) = P_\mu\left(\bar{\mathbf{X}} < \frac{\sigma Z_\alpha}{\sqrt{n}} + \mu_0\right),$$

is a decreasing function of μ, since μ is a location parameter in the distribution of $\bar{\mathbf{X}}$, which is normal (try to check this yourself by subtracting μ on both sides and dividing by the standard deviation of $\bar{\mathbf{X}}$). Thus, $\sup_{\mu \geq \mu_0} \beta(\mu) = \beta(\mu_0) = \alpha$, and the test is a level α test.

A large class of problems that admit UMP level α tests involve one-sided hypotheses and pdfs or pmfs with the monotone likelihood ratio property:

Definition 42. A family of pdfs or pmfs $\{f(t;\theta) : \theta \in \Theta\}$ for a univariate random variable T with real-valued parameter θ has a *monotone likelihood ratio* (MLR) is, for every $\theta_2 > \theta_1$, $\frac{f(t;\theta_2)}{f(t;\theta_1)}$ is a non-decreasing function of t on $\{t : f(t;\theta_1) > 0 \text{ or } f(t;\theta_2) > 0\}$. Note that $c/0$ is defined as ∞ if $c > 0$.

Many families of distributions have an MLR. For example, the Normal distribution with known variance and unknown mean, the Poisson distribution, and the Binomial distribution. In fact, the **Exponential Family** (not to be confused with the Exponential distribution), which has density function of the type:

$$f(t;\theta) = h(t)c(\theta)e^{w(\theta)t},$$

for some "regular" functions h, c, w, has an MLR if $w(\theta)$ is a non-decreasing function. This family of distributions contains many common distributions such as the normal distribution, the exponential distribution, the gamma distribution, the Poisson distribution, the Bernoulli distribution, among others.

Theorem 4.2.2. *(Karlin-Rubin). Consider testing*

$$H_0 : \theta \leq \theta_0 \quad vs. \quad H_1 : \theta > \theta_0.$$

Suppose that T is a sufficient statistic for θ and the family of pdfs or pmfs $\{g(t;\theta) : \theta \in \Theta\}$ of T has an MLR. Then, for any t_0, the test that rejects H_0 if and only if $T > t_0$ is a UMP level α test, where $\alpha = P_{\theta_0}(T > t_0)$.

Proof. Since the family of pdfs or pmfs of T has an MLR, the power function $\beta(\theta) = P_\theta(T > t_0)$ is nondecreasing, as we will show next. Note first that $\beta(\theta) = P_\theta(T > t_0) = 1 - G(t_0;\theta)$, where G is the cdf of T. We want to prove that, for $\theta_1 < \theta_2$, $\beta(\theta_1) \leq \beta(\theta_2)$, which is equivalent to $1 - G(t_0;\theta_1) \leq 1 - G(t_0;\theta_2)$, or $G(t_0;\theta_1) \geq G(t_0;\theta_2)$.

Let $g(t;\theta)$ be the pdf (pmf) of T, $\theta_1 < \theta_2$, and define

$$r(t) = \frac{g(t;\theta_2)}{g(t;\theta_1)}.$$

r is non-decreasing as T has a MLR. Define $\tilde{t} = \sup\{t : r(t) \leq 1\}$. Then, by definition, for all $t \leq \tilde{t}$ $G(t;\theta_1) \geq G(t;\theta_2)$, since on the interval $(-\infty, \tilde{t})$ $g(t;\theta_2) \leq g(t;\theta_1)$, and $G(t;\theta) = \int_{-\infty}^{t} g(x;\theta)dx$.

Now, for $t > \tilde{t}$, $g(t;\theta_2) \geq g(t;\theta_1)$ (as r is non-decreasing). Recall that $G(t;\theta) = 1 - \int_t^{\infty} g(x;\theta)dx$, then

$$\begin{aligned} G(t;\theta_1) - G(t;\theta_2) &= 1 - \int_t^{\infty} g(x;\theta_1)dx - 1 + \int_t^{\infty} g(x;\theta_2)dx \\ &= \int_t^{\infty} [g(x;\theta_2) - g(x;\theta_1)]dx \\ &\geq 0. \end{aligned}$$

This implies that the power function is non-decreasing. So, $\sup_{\theta \leq \theta_0} \beta(\theta) = \beta(\theta_0) = \alpha$, and this is a level α test. Condition (ii) of Corollary 4.2.2 is true by assumption. Condition (iii) can be verified by using

$$k' = \inf_{t \in \mathcal{T}} \frac{g(t;\theta')}{g(t;\theta_0)},$$

where $\mathcal{T} = \{t : t > t_0 \text{ and either } g(t;\theta') > 0 \text{ or } g(t;\theta_0) > 0\}$. More specifically, this can be verified by looking at the ratio $\frac{g(t;\theta')}{g(t;\theta_0)} > k'$. Thus, by Corollary 4.2.2 the test is a UMP level α test.

By an analogous argument, it can be shown that under the same conditions of the previous theorem, the test that rejects $H_0 : \theta \geq \theta_0$ in favour of $H_1 : \theta < \theta_0$ if and only if $T < t_0$ is a UMP level $\alpha = P_{\theta_0}(T < t_0)$ test. In fact, the test in the previous example is a test of this form.

4.3 Likelihood Ratio Test

The likelihood ratio (LR) method of hypothesis testing is related to the maximum likelihood estimation and profile likelihood methods discussed earlier. The LR test is as widely applicable as the maximum likelihood estimation method. Recall that the likelihood function is

$$L(\theta \mid \mathbf{x}) = \prod_{i=1}^{n} f(x_i \mid \theta),$$

where f is the pdf or pmf of the sample $\mathbf{x}$, $\theta \in \Theta$, and Θ denotes the parameter space. LR tests are defined as follows

Definition 43. The likelihood ratio test (LRT) statistic for testing

$$H_0 : \theta \in \Theta_0 \quad vs. \quad H_1 : \theta \in \Theta_0^c,$$

is

$$\lambda(\mathbf{x}) = \frac{\sup_{\Theta_0} L(\theta \mid \mathbf{x})}{\sup_{\Theta} L(\theta \mid \mathbf{x})},$$

where sup denotes the supremum (equivalently maximum, if the likelihood has a maximum). A likelihood ratio test (LRT) is any test that has a rejection region of the form $R = \{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}$, where c is any number satisfying $0 \leq c \leq 1$.

In the discrete case (f is a pmf), the numerator of the statistic $\lambda(\mathbf{x})$ can be interpreted as the maximum probability of the sample being computed over the values of the parameter in the null hypothesis. The denominator is the maximum probability of the sample over all the values of the parameter. Then, the ratio of these probabilities is expected to be small if there are values of the parameter in the alternative hypothesis for which the observed sample is more likely than for any parameter in the null hypothesis. Thus, we would like to reject the null hypothesis when the LRT is small. We will discuss methods to obtain the threshold value to reject the null. In some references, you will find the following shorter alternative notation

$$\lambda(\mathbf{x}) = \frac{\sup_{\Theta_0} L(\theta \mid \mathbf{x})}{\sup_{\Theta} L(\theta \mid \mathbf{x})} = \frac{L(\widehat{\theta}_0)}{L(\widehat{\theta})}.$$

The following example illustrates the use of the LRT in the one-parameter scenario for simple hypotheses.

Simple Null Hypothesis

Example 4.3.1. Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, 1)$. Consider testing

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \neq \mu_0.$$

We know that the MLE is $\widehat{\mu} = \bar{\mathbf{X}}$, then

$$\lambda(\mathbf{x}) = \frac{(2\pi)^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu_0)^2\right]}{(2\pi)^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2\right]}.$$

After some algebra (using results in previous sections),

$$\lambda(\mathbf{x}) = \exp\left[-\frac{n}{2}(\bar{\mathbf{x}} - \mu_0)^2\right].$$

Consequently, a LRT is a test that rejects H_0 for small values of $\lambda(\mathbf{x})$. Then, if $\lambda(\mathbf{x}) \leq c$, $c \in (0, 1)$, this implies that

$$\left\{\mathbf{x} : |\bar{\mathbf{x}} - \mu_0| \geq \sqrt{-2\log(c)/n}\right\}.$$

The following example illustrates the use of the LRT in the case where there are nuisance parameters, and the profile likelihood is used instead of the likelihood function in order to eliminate the nuisance parameters.

Example 4.3.2. Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$, with σ^2 unknown. Consider testing

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \neq \mu_0.$$

In this case, the LRT statistic is

$$\lambda(\mathbf{x}) = \frac{L(\mu_0, \widehat{\sigma}(\mu_0))}{L(\widehat{\mu}, \widehat{\sigma})},$$

where $\widehat{\sigma}(\mu_0) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2}$, $\widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2}$, and $\widehat{\mu} = \bar{\mathbf{x}}$. Then,

$$\lambda(\mathbf{x}) = \left(\frac{\widehat{\sigma}}{\widehat{\sigma}(\mu_0)}\right)^n = \left(\frac{\sum_{i=1}^{n}(x_i - \bar{\mathbf{x}})^2}{\sum_{i=1}^{n}(x_i - \mu_0)^2}\right)^{\frac{n}{2}},$$

Now, we also have that (exercise)

$$\sum_{i=1}^{n}(x_i-\mu_0)^2=\sum_{i=1}^{n}(x_i-\bar{\mathbf{x}})^2+n(\bar{\mathbf{x}}-\mu_0)^2.$$

Then,

$$\begin{aligned}\lambda(\mathbf{x}) &= \left(\frac{\widehat{\sigma}}{\widehat{\sigma}(\mu_0)}\right)^n=\left(\frac{\sum_{i=1}^{n}(x_i-\bar{\mathbf{x}})^2}{\sum_{i=1}^{n}(x_i-\bar{\mathbf{x}})^2+n(\bar{\mathbf{x}}-\mu_0)^2}\right)^{\frac{n}{2}}\\ &= \left(\frac{1}{1+\dfrac{n(\bar{\mathbf{x}}-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\bar{\mathbf{x}})^2}}\right)^{\frac{n}{2}},\end{aligned}$$

Solving the inequality $\lambda(\mathbf{x})\leq c$, we obtain

$$\frac{n(\bar{\mathbf{x}}-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\bar{\mathbf{x}})^2}\geq c^{-\frac{2}{n}}-1=c^*.$$

Equivalently,

$$\frac{(\bar{\mathbf{x}}-\mu_0)^2}{\frac{1}{n}S_u^2}\geq(n-1)c^*=c^{**},$$

where $S_u^2=\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{\mathbf{x}})^2$. Taking square roots on both sides:

$$\frac{|\bar{\mathbf{x}}-\mu_0|}{\frac{S_u}{\sqrt{n}}}\geq\sqrt{c^{**}}=k.$$

As discussed before, and under H_0 (when the null hypothesis is true), $T_n=\dfrac{\bar{\mathbf{x}}-\mu_0}{S_u/\sqrt{n}}$ has Student's t-distribution with $n-1$ degrees of freedom. Thus, if we want to obtain a test of size α based on the test statistic T_n (which was obtained from the LRT), we need to select k as the $\left(1-\frac{\alpha}{2}\right)$ quantile of the Student's t-distribution with $n-1$ degrees of freedom (so, c depends on other variables).

Historical Note 7. The test discussed in Example 4.3.2 is known as the (one-sample) Student's t-test. It was published by William Sealy Gosset, while he was working at the Guinness Brewery. Due to confidentiality reasons, he was not allowed to use his real name, and he published the paper under the pseudonym "Student".

Clearly, T_n is a pivotal quantity. Thus, we can construct a confidence interval of level $100(1-\alpha)\%$ for μ_0 by calculating the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles of the Student's t-distribution with $n-1$ degrees of freedom, $Q_{\frac{\alpha}{2}}, Q_{1-\frac{\alpha}{2}}$. Since the Student's t-distribution is symmetric, it follows that $Q_{\frac{\alpha}{2}}=-Q_{1-\frac{\alpha}{2}}$. Then,

$$P\left(-Q_{1-\frac{\alpha}{2}}\leq T_n\leq Q_{1-\frac{\alpha}{2}}\right)=1-\alpha,$$

implies that the $100(1-\alpha)\%$ confidence interval for μ_0 is

$$\bar{\mathbf{x}}-Q_{1-\frac{\alpha}{2}}\frac{S_u}{\sqrt{n}}\leq\mu_0\leq\bar{\mathbf{x}}+Q_{1-\frac{\alpha}{2}}\frac{S_u}{\sqrt{n}}.$$

Note that this king of confidence intervals can be used when the variance is unknown.

Thus, the Student's t-test can be obtained as a LRT.

In previous chapters, we studied the intuitive notion that all the information about θ in the sample $\mathbf{x}$ is contained in $T(\mathbf{x})$, a sufficient statistic. Intuitively, the test based on T should be as good as the test based on the complete sample $\mathbf{X}$. This is, if $T(\mathbf{X})$ is sufficient with pdf or pmf $g(t;\theta)$, then we might consider constructing an LRT based on T and its likelihood function $L^*(\theta \mid t) = g(t;\theta)$ rather than on the sample $\mathbf{x}$ and its likelihood function $L(\theta \mid \mathbf{x})$. Let $\lambda^*(t)$ denote the likelihood ratio test statistic based on T. Then, we have the following result.

Theorem 4.3.1. *If $T(\mathbf{X})$ is a sufficient statistic for θ and $\lambda^*(t)$ and $\lambda(\mathbf{x})$ are the LRT statistics based on T and $\mathbf{X}$, respectively, then $\lambda^*(T(\mathbf{x})) = \lambda(\mathbf{x})$ for every $\mathbf{x}$ in the sample space.*

Proof. From the factorisation theorem, the pdf or pmf of $\mathbf{X}$ can be written as $f(\mathbf{x};\theta) = g(T(\mathbf{x});\theta)h(\mathbf{x})$, where $g(t;\theta)$ is the pdf or pmf of T and $h(\mathbf{x})$ does not depend on θ. Then,

$$\begin{aligned}
\lambda(\mathbf{x}) &= \frac{\sup_{\Theta_0} L(\theta \mid \mathbf{x})}{\sup_{\Theta} L(\theta \mid \mathbf{x})} \\
\text{(by definition)} &= \frac{\sup_{\Theta_0} f(\mathbf{x};\theta)}{\sup_{\Theta} f(\mathbf{x};\theta)} \\
\text{(Fact. Th.)} &= \frac{\sup_{\Theta_0} g(T(\mathbf{x});\theta)h(\mathbf{x})}{\sup_{\Theta} g(T(\mathbf{x});\theta)h(\mathbf{x})} \\
(h \text{ cancels out}) &= \frac{\sup_{\Theta_0} g(T(\mathbf{x});\theta)}{\sup_{\Theta} g(T(\mathbf{x});\theta)} \\
\text{(by definition)} &= \frac{\sup_{\Theta_0} L^*(\theta \mid T(\mathbf{x}))}{\sup_{\Theta} L^*(\theta \mid T(\mathbf{x}))} \\
&= \lambda^*(T(\mathbf{x})).
\end{aligned}$$

Unfortunately, the distribution of the test statistic used in the LRT cannot always be obtained in closed-form, as was the case of our previous examples. However, an asymptotic approximation can be used that greatly simplifies the use of the LRT and allows its use in a wide range of cases.

Theorem 4.3.2. *Consider testing*

$$H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta \neq \theta_0,$$

where $\theta \in \mathbb{R}$. Under H_0,

$$-2\log\lambda(\mathbf{X}) \xrightarrow{d} \chi^2_1.$$

Then, if we use the test statistic $T_n = -2\log\lambda(\mathbf{X})$, and denote the $1-\alpha$ quantile of a χ^2_1 distribution as $\chi^2_{1,1-\alpha}$, we have that

$$P_{\theta_0}(T_n \geq \chi^2_{1,1-\alpha}) \to \alpha \quad as\ n \to \infty.$$

Proof. Let us present first and intuitive proof, and then we will discuss the steps for the formal proof.

- Intuitive proof. Using a second order Taylor expansion of the log-likelihood around the MLE:

$$\begin{aligned}
\ell(\theta) &\approx \ell(\widehat{\theta}) + \ell'(\widehat{\theta})(\theta - \widehat{\theta}) + \frac{\ell''(\widehat{\theta})}{2}(\theta - \widehat{\theta})^2 \\
&= \ell(\widehat{\theta}) + \frac{\ell''(\widehat{\theta})}{2}(\theta - \widehat{\theta})^2.
\end{aligned}$$

Then, using this approximation at $\theta = \theta_0$, we obtain the following approximation for the LRT

$$\begin{aligned}
-2\log\lambda(\mathbf{x}) &= 2\ell(\widehat{\theta}) - 2\ell(\theta_0) \\
&\approx 2\ell(\widehat{\theta}) - 2\ell(\widehat{\theta}) - \ell''(\widehat{\theta})(\theta - \widehat{\theta})^2
\end{aligned}$$

$$
\begin{aligned}
&= -\ell''(\widehat{\theta})(\theta - \widehat{\theta})^2 \\
&= -\frac{\ell''(\widehat{\theta})}{nI(\theta_0)} \times I(\theta_0) \left(\sqrt{n}(\theta - \widehat{\theta})\right)^2 = A_n \times B_n.
\end{aligned}
$$

As discussed before, the negative of the second derivative of the log-likelihood divided by n converges to the Fisher information in probability. Thus, $A_n \xrightarrow{P} 1$, as $n \to \infty$ by the weak law of large numbers. The second term B_n can be immediately connected with the CLT, which implies that $\sqrt{B_n} \xrightarrow{d} N(0,1)$. Thus, by Slutsky's theorem, $B_n \xrightarrow{d} \chi^2_1$. A second application of Slutsky's theorem implies that $-2\log\lambda(\mathbf{x}) \xrightarrow{d} \chi^2_1$.

- Sketch of the formal proof. What are we missing in the previous proof?

  The answer is: assumptions. In order to formally proof this result, we need some conditions on the likelihood function (or, equivalently, on the pdf/pmf). The approximation based on the Taylor expansion still contains a remainder term (residual term), which we need to show that converges to 0. In order to do this, we can use the remainder Taylor expansion of third order to obtain

$$
\ell(\theta) = \ell(\widehat{\theta}) + \frac{\ell''(\widehat{\theta})}{2}(\theta - \widehat{\theta})^2 + \frac{\ell''(\theta^*)}{6}(\theta - \widehat{\theta})^3,
$$

  where $\theta^* = \epsilon\theta + (1-\epsilon)\widehat{\theta}$, for some $\epsilon \in (0,1)$. Here, we are assuming that the log-likelihood is three times continuously differentiable. Now, in order to show that the remainder term vanishes as $n \to \infty$, we need to assume that the third derivative of the likelihood function is bounded, $|\ell''(\theta^*)| < M$, for some $M > 0$, in a neighbourhood of θ_0. This often requires assuming that the parameter space Θ is compact (closed and bounded).

  We were also assuming that the Fisher information exists, which requires some additional conditions, as we discussed previously.

Example 4.3.3. Let $X_1, \ldots, X_n$ be *i.i.d.* with Poisson distribution with mean $\theta > 0$. Suppose we are interested on testing,

$$
H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta \neq \theta_0,
$$

Recall that the pmf is $p(x) = \dfrac{\theta^x}{x!}e^{-\theta}$, and that the MLE of θ, $\widehat{\theta} = \bar{\mathbf{x}}$. Then,

$$
-2\log\lambda(\mathbf{x}) = 2n\left[(\theta_0 - \bar{\mathbf{x}}) - \bar{\mathbf{x}}\log\left(\frac{\theta_0}{\bar{\mathbf{x}}}\right)\right].
$$

Thus, based on the previous theorem, we reject if $-2\log\lambda(\mathbf{x}) \geq \chi^2_{1,1-\alpha}$.

The LRT is also asymptotically distributed as a ξ^2 as shown in the following result that we present without proof. The proof is similar to that of the previous theorem, but it requires additional multivariate tools.

Theorem 4.3.3. *(Wilks' Theorem for simple hypotheses). Let $\boldsymbol{\theta} = (\boldsymbol{\delta}, \boldsymbol{\xi})$, where $\boldsymbol{\delta} \in \mathbb{R}^p$ and $\boldsymbol{\xi} \in \mathbb{R}^q$, and suppose that $\boldsymbol{\theta}_0 = (\boldsymbol{\delta}_0, \boldsymbol{\xi}_0)$ are the true value of the parameters. Then, Under some regularity conditions (basically involving differentiability and existence of the Fisher information matrix),*

$$
-2\log\lambda(\mathbf{X}) = -2\left[\ell(\boldsymbol{\delta}_0, \widehat{\boldsymbol{\xi}}(\boldsymbol{\delta}_0) \mid \mathbf{X}) - \ell(\widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\xi}} \mid \mathbf{X})\right] \xrightarrow{d} \chi^2_p.
$$

This theorem can be used to test:

$$
H_0 : \boldsymbol{\delta} = \boldsymbol{\delta}_0 \quad vs. \quad H_1 : \boldsymbol{\delta} \neq \boldsymbol{\delta}_0.
$$

A test of approximate level α is obtained by rejecting H_0 when $T_n = -2\log\lambda(\mathbf{X}) \geq \chi^2_{p,1-\alpha}$.

The next example illustrates how to use this result in the case of testing a normal mean. Note that this is an approximate test, in contrast to the exact test we derived in a previous example. However, the importance of this method relies in its generality, as it can be used on other tests where the distribution of the test statistic is not available in closed form.

Example 4.3.4. Let $X_1, \ldots, X_n$ be an *i.i.d.* random sample from a $N(\mu, \sigma^2)$, with σ^2 unknown. Consider testing

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \neq \mu_0.$$

Thus, $p = 1$ in this case. The test statistic based on the LRT is

$$T_n = -2 \log \lambda(\mathbf{x}) = n \log \left(1 + \frac{n(\bar{\mathbf{x}} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{\mathbf{x}})^2} \right).$$

We know that $T_n \xrightarrow{d} \chi^2_1$, under H_0. Thus, a test of approximate level α rejects H_0 when $T_n \geq \chi^2_{1,1-\alpha}$. Note that this is equivalent to $\lambda(\mathbf{x}) \leq \exp\left\{ -\frac{\chi^2_{1,1-\alpha}}{2} \right\}$, which is consistent with the formulation of the LRT.

Theorem 4.3.4. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from a pdf or pmf $f(\cdot; \boldsymbol{\theta})$. Under some regularity conditions on the model $f(\cdot; \boldsymbol{\theta})$, if $\theta \in \Theta_0$ then the distribution of the statistic $-2 \log \lambda(\mathbf{X})$ converges to a χ^2_p distribution as the sample size $n \to \infty$. The degrees of freedom p of the limiting distribution is the difference between the number of free parameters specified by $\theta \in \Theta_0$ and the number of free parameters specified by $\boldsymbol{\theta} \in \Theta$.*

The regularity conditions, again, correspond to differentiability of the likelihood function conditions and the existence of the Fisher information matrix. These conditions are satisfied for many reasonable distributions.

Rejection of $H_0 : \boldsymbol{\theta} \in \Theta_0$ for small values of $\lambda(\mathbf{x})$ is equivalent for large values of $-2 \log \lambda(\mathbf{x})$. Then, H_0 is rejected if and only if $-2 \log \lambda(\mathbf{x}) \geq \chi^2_{p,1-\alpha}$. The Type I Error probability is approximately α if $\boldsymbol{\theta} \in \Theta_0$ (*i.e.* under the null), and the sample size is large. This is,

$$\lim_{n \to \infty} P_\theta(\text{reject } H_0) = \alpha,$$

for each $\theta \in \Theta_0$. However, this does not imply that $\sup_{\theta \in \Theta_0} P_\theta(\text{reject } H_0) \to \alpha$. That is, the convergence is *pointwise* on θ.

Example 4.3.5. Comparing nested models. Let $X_1, \ldots, X_n$ be an *i.i.d.* random sample from a distribution with pdf $f(\cdot; \boldsymbol{\theta})$. Let $\boldsymbol{\theta} = (\boldsymbol{\delta}, \boldsymbol{\xi})$, where $\boldsymbol{\delta} \in \mathbb{R}^p$ and $\boldsymbol{\xi} \in \mathbb{R}^q$. Suppose that the "nested model" $\boldsymbol{\xi} = \boldsymbol{\xi}_0$ is of interest, so we are interested on testing

$$H_0 : \boldsymbol{\xi} = \boldsymbol{\xi}_0 \quad vs. \quad H_1 : \boldsymbol{\xi} \neq \boldsymbol{\xi}_0.$$

Then, the likelihood ratio test statistic

$$\lambda(\mathbf{x}) = \frac{\sup_{\delta \in \mathbb{R}^p} L(\boldsymbol{\delta}, \boldsymbol{\xi}_0 \mid \mathbf{x})}{\sup_{\Theta} L(\boldsymbol{\theta} \mid \mathbf{x})},$$

satisfies $T_n = -2\lambda(\mathbf{x}) \xrightarrow{d} \chi^2_p$, under H_0.

This kind of hypothesis appear in many contexts where the statistician is interested in testing whether or not some parameters are necessary, or if a more parsimonious model (with fewer parameters) is preferred.

- For instance, if $X_i \sim Gamma(1, 1/\lambda)$ (shape -scale parameterisation), then X_i has an Exponential distribution with rate parameter λ. We say that the Exponential distribution is "nested" in the Gamma distribution. If $X_i \sim Gamma(a, b)$, we may be interested on testing:

$$H_0 : a = 1 \quad vs. \quad H_1 : a \neq 1,$$

in order to decide whether the data are exponentially distributed (with one parameter) or we need a more flexible (with two parameters) distribution (Gamma).

- Suppose that we have a linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

and that we are interested on testing whether some specific variables are spurious (this is, they are not important to explain the response variable, $\beta_j = 0$) or they are active (they are important to explain the response variable, $\beta_j \neq 0$). Let $\boldsymbol{\beta} \in \mathbb{R}^d$, and γ be the set of indices of the variables that are suspected to be spurios (*e.g.* $\{1,4,7,\dots\}$). So, we are interested on testing:

$$H_0 : \boldsymbol{\beta}_\gamma = \mathbf{0} \quad vs. \quad H_1 : \boldsymbol{\beta}_\gamma \neq \mathbf{0}.$$

Note that we can only test all of the γ variables simultaneously. It is important to emphasise that there exist better, and more formal, methods to select active variables.

One of the limitations of the likelihood ratio test is that it can only be used when the null model is nested in the alternative model.

Composite Null Hypothesis

Theorem 4.3.5. *(Wilks' Theorem for composite hypotheses). Let $\boldsymbol{\theta} = (\boldsymbol{\delta}, \boldsymbol{\xi}) \in \mathbb{R}^{p+q}$, where $\boldsymbol{\delta} \in \mathbb{R}^p$ and $\boldsymbol{\xi} \in \mathbb{R}^q$. Consider testing:*

$$H_0 : \boldsymbol{\delta} \in \Delta_0 \quad vs. \quad H_1 : \boldsymbol{\delta} \in \Delta_0^c.$$

Define the generalised profile likelihood

$$\lambda(\mathbf{x}) = \frac{\sup_{\boldsymbol{\delta}\in\Delta_0, \boldsymbol{\xi}\in\mathbb{R}^q} L(\boldsymbol{\theta} \mid \mathbf{x})}{\sup_{\boldsymbol{\theta}\in\mathbb{R}^{p+q}} L(\boldsymbol{\theta} \mid \mathbf{x})}.$$

Then, Under some regularity conditions,

$$-2\log\lambda(\mathbf{x}) \xrightarrow{d} \chi^2_p.$$

A test of approximate level α is obtained by rejecting H_0 when $T_n = -2\log\lambda(\mathbf{x}) \geq \chi^2_{p,1-\alpha}$.

Example 4.3.6. Let $X_1, \dots, X_n$ be a random sample from a $N(\mu, \sigma^2)$, with both parameters unknown. Consider testing

$$H_0 : \mu \leq \mu_0 \quad vs. \quad H_1 : \mu > \mu_0.$$

In this case, the LRT statistic is

$$\lambda(\mathbf{x}) = \frac{\max_{\mu\leq\mu_0, \sigma>0} L(\mu, \sigma \mid \mathbf{x})}{\max_{\mu\in\mathbb{R}, \sigma>0} L(\mu, \sigma \mid \mathbf{x})}.$$

The MLEs are, as we know, $\widehat{\mu} = \bar{\mathbf{x}}$ and $\widehat{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^n (x_i - \bar{\mathbf{x}})^2$. On the other hand, the restricted estimator of σ in the numerator is:

$$\widehat{\sigma}(\mu_0) = \begin{cases} \frac{1}{n}\sum_{i=1}^n (x_i - \bar{\mathbf{x}})^2, & \text{if } \bar{\mathbf{x}} \leq \mu_0 \\ \frac{1}{n}\sum_{i=1}^n (x_i - \mu_0)^2, & \text{if } \bar{\mathbf{x}} > \mu_0, \end{cases}$$

Therefore,

$$\lambda(\mathbf{x}) = \begin{cases} 1, & \text{if } \bar{\mathbf{x}} \leq \mu_0 \\ \dfrac{L(\mu_0, \widehat{\sigma}(\mu_0) \mid \mathbf{x})}{L(\widehat{\mu}, \widehat{\sigma} \mid \mathbf{x})}, & \text{if } \bar{\mathbf{x}} > \mu_0. \end{cases}$$

A test of approximate level α is obtained by rejecting H_0 when $T_n = -2 \log \lambda(\mathbf{x}) \geq \chi^2_{1,1-\alpha}$.

Alternatively, and similar to the construction of the one-sample t-test (exercise), we can show that the test that rejects the null hypothesis when

$$\bar{\mathbf{x}} \geq \mu_0 + Q_{1-\alpha} \frac{S_u}{\sqrt{n}},$$

is of size α.

4.4 Tests of Significance

The purpose of a test of significance is to measure the strength of the evidence provided by the experimental data $\mathbf{x} = x_1, \ldots, x_n$, against a hypothesis H_0.

A test of significance requires two ingredients:

1. a discrepancy measure, or test criterion, $D(\mathbf{x}) \geq 0$,
2. a probability distribution of D under H_0.

The discrepancy measure $D(\cdot)$ ranks the samples $\mathbf{x}$ according to their strength of evidence against H_0. So, a sample $\mathbf{x}'$ contains stronger evidence against H_0 than $\mathbf{x}''$ if and only if $D(\mathbf{x}') > D(\mathbf{x}'')$. Thus, evidence against H_0 is indicated by observing sufficiently large values of $D = d$. The observed discrepancy d is regarded as large if the probability of getting a larger value is small, so that D is out in the tail of its distribution, that is, the observed d is an outlier.

Definition 44. The observed significance level, or p-value, of the data in relation to H_0 is defined as

$$P = P[D(\mathbf{X}) \geq d \mid H_0].$$

This is the probability under H_0 of observing a discrepancy at least as great as the observed discrepancy d.

Then, the p-value measures the strength of the evidence against H_0 on the probability scale. The smaller the p-value, the stronger the evidence against H_0 provided by the observed $\mathbf{x}$.

An ordinary p-value (not extreme) means only that H_0 can be assumed without necessitating an unusually improbable or singular experiment. But the lack of evidence against H_0 cannot be interpreted as evidence in favour of H_0. p-values cannot be used to provide evidence in favour of a hypothesis. In fact, there is a popular saying related to this idea:

"Absence of evidence is not evidence of absence."

In some cases, an exceptionally large p-value also represents a red flag, as $P^* = 1 - P$ is often interpreted as a p-value, in the sense that observing very small discrepancies is considered to be rare, due to the stochastic nature of the sample. This can only be "taken with a pinch of salt", as it is possible that, if H_0 is true, the discrepancy associated to a sample under H_0 is low, but these cases often deserve further attention as they may be "too good to be true".

Historical Note 8. There is a connection between hypothesis testing and the logic of falsification/falsifiability (Karl Popper). The idea that we can only provide evidence against a hypothesis has been long studied in Philosophy.

Significance tests are criticised since the tail probabilities $D \geq d$ involve data that were not observed. Only d was observed; but values greater than d were not observed. This brings in the "Frequentist" definition of probability, which assumes repeatability of the experiment. We will discuss this definition in the section about Bayesian inference. Briefly, this criticism ignores the fact that experiments must be repeatable at least in a statistical sense.

Thus no single experiment can provide conclusive evidence against H_0. The statistical interpretation is that $P = P(D \geq d \mid H_0)$ is a cumulative distribution function (in the case that the discrepancy measure is continuous). Therefore, for continuous variates P is a random variable with, under H_0, the uniform distribution between 0 and 1, $U(0,1)$ (the probability integral transform). Let us now proof this statement.

Theorem 4.4.1. *(Probability Integral Transform). Let Z and T be two continuous random variables such that $Z = F(T)$, where F is the cdf of T. Then, $Z \sim U(0,1)$.*

Proof.

$$\begin{aligned} F(z) &= P(Z \leq z) \\ &= P(F(T) \leq z) \\ &= P(T \leq F^{-1}(z)) \\ &= F(F^{-1}(z)) \\ &= z. \end{aligned}$$

Since $P(Z \leq z) = z$, this implies that $Z \sim U(0,1)$.

Theorem 4.4.2. *(The p-value is uniformly distributed). Let $\mathbf{X} = X_1, \ldots, X_n$ be* i.i.d.*continuous random variables. Let $D(\mathbf{X})$ be a discrepancy measure with cumulative distribution function $F_0(\cdot)$, which is assumed to be continuous, under H_0. Then, under H_0,*

$$P = P[D(\mathbf{X}) \geq d \mid H_0] \sim U(0,1).$$

Proof. By assumption,

$$\begin{aligned} P &= P[D(\mathbf{X}) \geq d \mid H_0] \\ &= 1 - F_0(d). \end{aligned}$$

Now, recall that, for a continuous random variable X and a continuous increasing function g, $P(X \leq x) = P(g(X) < g(x))$. Then,

$$\begin{aligned} P &= P[D(\mathbf{X}) \geq d \mid H_0] \\ &= 1 - P[D(\mathbf{X}) \leq d \mid H_0] \\ &= 1 - P[F_0(D(\mathbf{X})) \leq F_0(d) \mid H_0] \\ &= 1 - F_0(d). \end{aligned}$$

Then, $P[F_0(D(\mathbf{X})) \leq F_0(d) \mid H_0] = F_0(d)$, and we can conclude that (recall the cdf of a uniform distribution) $F_0(D(\mathbf{X})) \sim U(0,1)$ (see the connection with the probability integral transform).

A common error in the interpretation of the p-value is to interpret it in terms of probabilistic statements about the null hypothesis.

Warning 3. The P-value is *not* the probability that H_0 is true given the observed data.

$$P \neq P(H_0 \mid \mathbf{x}).$$

A natural question now is: what discrepancy measure can we use?

A general answer consist of using the likelihood ratio test statistic as a discrepancy measure:

$$D(\mathbf{x}) = -2 \log \lambda(\mathbf{x}).$$

The interpretation is that if D is large, then λ is small. Examples of this choice were studied in the previous section. Next, we will present some alternative choices.

Example 4.4.1. (matched pair). Let $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be two measurements on paired individuals (or on the same individual), which are assumed to be normally distributed with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, respectively. Suppose that we are interested in testing that

$$H_0 : \mu = \mu_1 - \mu_2 = 0 \quad vs. \quad H_1 : \mu \neq 0.$$

Define $\mathbf{D} = \mathbf{X} - \mathbf{Y} = (X_1 - Y_1, \ldots, X_n - Y_n)$. We know that:

$$T = \frac{\bar{\mathbf{D}} - \mu}{S_d/\sqrt{n}},$$

has $t-$distribution with $n-1$ degrees of freedom. Then, we can use $D(\mathbf{d}) = |T|$ as the discrepancy measure. Thus, the P-value becomes, for a realisation t_0 of T,

$$P = P(|T| \geq t_0 \mid H_0) = P(T \geq t_0 \mid H_0) + P(T \leq -t_0 \mid H_0).$$

This is basically quantifying how close to zero the sample mean of differences (scaled by their variance) is.

An example using Darwin's data can be found at:

`http://rpubs.com/FJRubio/DarwinPaired`

Example 4.4.2. (unmatched samples). Let $\mathbf{X} = (X_1, \ldots, X_m)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be two independent samples, which are assumed to be normally distributed with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, respectively. Recall that the Welch-Satterthwaite statistic (known as the *Studentised* statistic)

$$\tilde{T} = \frac{(\bar{\mathbf{X}} - \bar{\mathbf{Y}}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{m} + \dfrac{S_2^2}{n}}},$$

has approximately a t distribution with *effective degrees of freedom* given by

$$\nu = \frac{\left(\dfrac{s_1^2}{m} + \dfrac{s_2^2}{n}\right)^2}{\dfrac{s_1^4}{m^2(m-1)} + \dfrac{s_2^4}{n^2(n-1)}}.$$

Suppose that we are interested on testing,

$$H_0 : \mu_1 = \mu_2 \quad vs. \quad H_1 : \mu_1 \neq \mu_2$$

Note that $H_0 : \mu_1 = \mu_2$ is equivalent to $H_0 : \mu_1 - \mu_2 = \mu = 0$. Then, in this case the p-value is

$$P = P(|\tilde{T}| \geq t_0 \mid H_0) = P(\tilde{T} \geq t_0 \mid H_0) + P(\tilde{T} \leq -t_0 \mid H_0).$$

This is basically quantifying how close to zero the difference of sample means (scaled by their variance) is.

An example using the Mosquitoes data can be found at:

`http://rpubs.com/FJRubio/Mosquitoes`

Remark 4.4.1. *If we were only interested in testing the one-sided hypothesis:*

$$H_0 : \mu_1 \leq \mu_2 \quad vs. \quad H_1 : \mu_1 > \mu_2,$$

then, the P-value is instead calculated as follows:

$$P = P(\tilde{T} \geq t_0 \mid H_0),$$

where $\tilde{T}$ is as described in the previous example. This is known as the one-sided (or right-tailed) two-sample t-test, and it mirrors the results that we obtained in the previous section for one-side alternative (one-sample) tests. This test, however, is obtained in a different way.

4.5 Interval Estimation by Inverting Statistics

In this section, we will study a strong connection between hypothesis testing and interval estimation. In fact, we can say in general that every confidence set corresponds to a test and vice versa. For instance, in previous sections, we found that the t-test statistic is used to test a hypothesis, but it was also used to construct confidence intervals about the mean. This idea is formalised in the following theorem.

Theorem 4.5.1. *For each $\theta_0 \in \Theta$, let $A(\theta_0) = R^c$ be the complement of the rejection region (often called the* acceptance region*) of a level α test of $H_0 : \theta = \theta_0$. For each $\mathbf{x}$, define the set $C(\mathbf{x})$ in the parameter space by*

$$C(\mathbf{x}) = \{\theta_0 : \mathbf{x} \in A(\theta_0)\}.$$

Then, the random set $C(\mathbf{X})$ is a $1 - \alpha$ confidence set. Conversely, let $C(\mathbf{X})$ be a $1 - \alpha$ confidence set. For any $\theta_0 \in \Theta$, define

$$A(\theta_0) = \{\mathbf{x} : \theta_0 \in C(\mathbf{x})\}.$$

Then, $A(\theta_0)$ is the acceptance region of a level α test of $H_0 : \theta = \theta_0$.

Proof. For the first part, since $A(\theta_0)$ is the acceptance region of a level α test,

$$P_{\theta_0}(\mathbf{X} \notin A(\theta_0)) \leq \alpha,$$

hence,

$$P_{\theta_0}(\mathbf{X} \in A(\theta_0)) \geq 1 - \alpha.$$

Since θ_0 is arbitrary, write θ instead of θ_0. The above inequality, together with the definition of $C(\mathbf{X})$, implies that the coverage probability of the set $C(\mathbf{X})$ is given by

$$P_{\theta_0}(\theta \in C(\mathbf{X})) = P_{\theta_0}(\mathbf{X} \notin A(\theta_0)) =\geq 1 - \alpha,$$

showing that $C(\mathbf{X})$ IS A $1 - \alpha$ confidence set.

For the second part, the Type I Error probability for the test of $H_0 : \theta = \theta_0$ with acceptance region $A(\theta_0)$ is

$$P_{\theta_0}(\mathbf{X} \notin A(\theta_0)) = P_{\theta_0}(\theta_0 \notin C(\mathbf{X})) \leq \alpha.$$

Thus, this is a level α test.

Â§

Chapter 5

Bayesian Inference

Historical Note 9. The idea behind Bayesian estimation was introduced by the Reverend Thomas Bayes in the paper "An Essay towards solving a Problem in the Doctrine of Chances". This paper was published by Bayes' friend Richard Price two years after Thomas Bayes died.

From a mathematical perspective, the concept of probability is formalised using Measure Theory. Thus, from a mathematical perspective, the concept of probability is relatively free from disagreement (although there are new mathematical theories that formalise the concept of probability in different ways, such as Quantum Probability). On the other hand, in the philosophy of science, there are two school of thoughts (there are others, but we will focus on these two) that define the concept of "Probability" in different ways.

- Frequentist probability: "The probability of an event is the limit of its relative frequency in a sequence of trials".
- Bayesian probability: "Is a measure of the degree of belief about an event".

Thus, we can identify a crucial different between these two definitions: repeatability. The frequentist definition of probability requires the possibility of repeating many (infinite) times an experiment. In contrast, the Bayesian definition does not require repeatability, but it is still valid on repeatable experiments. The definition and interpretation of Probability has been a matter of debate for many years, and there is no general consensus. Adopting either of these two definitions has great implications in Statistics, as one needs to develop logic axioms in order to create a mathematical theory which is also consistent with that definition. In fact, adopting the frequentist probability precludes one to define probabilities for events that are not repeatable such as

- What is the probability there is life in outer space?
- Will it rain tomorrow in London?
- Will you pass the Statistical Inference exam?

So far, we have seen only estimation procedures which are consistent with the first definition, and they are often referred to as Frequentist estimators. In this section, we will study a different approach, arising from the second definition.

5.1 The Bayes' Theorem for discrete events.

Theorem 5.1.1. *Let A and B be two events (in a $\sigma-$ algebra associated to the probability P), such that $P(B) > 0$. The Bayes' theorem for discrete events is:*

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}. \tag{5.1.1}$$

Proof. If $P(A) = 0$, the result is immediate by the laws of probability. If $P(A) > 0$, the proof follows by using the definition of conditional probability:

$$\begin{aligned} P(A \mid B) &= \frac{P(A, B)}{P(B)} \\ &= \frac{P(A, B)}{P(B)} \frac{P(A)}{P(A)} \\ &= \frac{P(A, B)}{P(A)} \frac{P(A)}{P(B)} \\ &= \frac{P(B \mid A) P(A)}{P(B)}. \end{aligned}$$

In some cases, the marginal probability $P(B)$ is not available but it is still possible to calculate it using the Law of Total Probability $P(B) = \sum_i P(B \mid C_i) P(C_i)$, where the sets C_i represent a partition of the sample space.

Example 5.1.1. (HIV test). An HIV test gives a positive result with probability 98% when the person is indeed infected by HIV, while it gives a negative result with 99% probability when the person is not affected by HIV. If a person is drawn at random from a population in which 0.1% of individuals are affected by HIV (this is known as "prevalence") and he/she is found positive, what is the probability that he/she is indeed infected by HIV?

The question can be translated as $P(\text{HIV} \mid \text{+ Result})$. We will use the Bayes' theorem to answer this question,

$$P(\text{HIV} \mid \text{+ Result}) = \frac{P(\text{+ Result} \mid \text{has HIV})\, P(\text{HIV})}{P(\text{+ Result})}.$$

Thus, we need to identify each term in the Bayes' theorem first.

- $P(\text{+ Result} \mid \text{HIV}) = 0.98$.
- $P(\text{+ Result} \mid \text{no HIV}) = 1 - P(\text{- Result} \mid \text{no HIV}) = 1 - 0.99 = 0.01$.
- $P(\text{HIV}) = 0.001$.
- $P(\text{no HIV}) = 1 - P(\text{HIV}) = 0.999$.

Then, we need an additional step to calculate $P(\text{+ Result})$ using the law of total probability:

$$\begin{aligned} P(\text{+ Result}) &= P(\text{+ Result} \mid \text{HIV}) P(\text{HIV}) + P(\text{+ Result} \mid \text{no HIV}) P(\text{no HIV}) \\ &= (0.98)(0.001) + (0.01)(0.999) \\ &= 0.01097. \end{aligned}$$

Therefore, using the Bayes' theorem

$$P(\text{HIV} \mid \text{+ Result}) = \frac{(0.98)(0.001)}{0.01097} = 0.08933.$$

Therefore, even though the test is conditionally very accurate, the unconditional probability of being affected by HIV when found positive is less than 10%!

In real life, tests are only conducted when there is an event that led the doctors to suspect infection, rather than at random. This is known as "belonging to a risk group", which implies that the prevalence of such subgroup is much larger. For instance if the prevalence in a risk group is 1 in 5 ($P(\text{HIV}) = 0.2$), then,

$$P(\text{HIV} \mid + \text{Result}) = \frac{(0.98)(0.2)}{(0.98)(0.2) + (0.01)(0.8)} = 0.96.$$

Moreover, tests are often repeated in order to avoid human and technical errors.

5.2 Prior Distributions

Let us now illustrate the idea behind choosing a prior distribution with an example-exercise.

Example 5.2.1. Consider a trial where 100 subjects are applied different treatments for a certain bacterial infection. This setting corresponds to a binomial sampling model with $n = 100$ and $\theta \in (0, 1)$ represents the probability of cure from the bacterial infection. We are interested in using Bayesian inference to analyse this data set, then we need a prior distribution for θ. Since $\theta \in (0, 1)$, we need to employ a distribution with support on $(0, 1)$. We will see an example of this later, but, for the moment, let's focus on the shape of the distribution according to what you believe about this parameter. Consider the following more specific scenarios.

(i) θ represents the probability of cure from a bacterial infection using a Penicillin treatment.

(ii) θ represents the probability of cure from a bacterial infection using a new experimental antibiotic treatment.

(iii) θ represents the probability of cure from a bacterial infection using a Homeopathic treatment. The idea of homeopathic treatments consists of diluting a part of an active ingredient (typically obtained from plants or animals) in 10^{30} parts of water. This produces a solution where the concentration of the active ingredient is lower than what you would get if you mix one drop of water in the ocean. The NHS stopped funding homeopathic treatments in the UK in August 2018.

5.3 Posterior Distribution

Bayesian inference is based on conditional probabilities. The conditional probability used in Bayesian statistics for quantifying the uncertainty about the parameters of a model is the "Posterior Distribution". The posterior distribution is the distribution of a parameter given a sample. Next, we will discuss this idea in detail.

A Bayesian parametric statistical model is a set of parametric probability distributions on a sample space Ω, and a prior distribution on the parameters. This is,

$$\begin{aligned} \mathbf{x} = (x_1, \ldots, x_n) &\sim f(\mathbf{x} \mid \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d, \quad \textbf{(Likelihood)}, \\ \boldsymbol{\theta} &\sim \pi_\Theta(\boldsymbol{\theta}), \quad \textbf{(Prior distribution)}. \end{aligned}$$

An important difference in the notation is that now we are using "|" to denote the conditional probability density function (probability mass function). That is, $f(\mathbf{x} \mid \boldsymbol{\theta})$ now denotes the conditional pdf (pmf) of the sample $\mathbf{x}$ given the value of the parameter $\boldsymbol{\theta}$. We are also assuming that $\boldsymbol{\theta}$ is a random variable or a random vector. At first, this may look counterintuitive in comparison with the estimation theories that we studied previously, as the parameter $\boldsymbol{\theta}$ is now treated as a random variable (random vector).

However, it is important to understand that we are not assuming that the true value of the parameter is random. Instead, the prior distribution reflects the *prior* uncertainty or the prior knowledge about the parameters before collecting the data. We can now construct the main ingredient of Bayesian statistics, for quantifying uncertainty about parameters, by combining the likelihood function and the prior distribution. The **Posterior Distribution** or posterior pdf (pmf) is obtained with the Bayes' Theorem for parametric models,

Theorem 5.3.1. Bayes' Theorem.

$$\pi(\boldsymbol{\theta} \mid \mathbf{x}) \;=\; \frac{f(\mathbf{x} \mid \boldsymbol{\theta})\pi_\Theta(\boldsymbol{\theta})}{\int_\Theta f(\mathbf{x} \mid \boldsymbol{\theta})\pi_\Theta(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{f(\mathbf{x} \mid \boldsymbol{\theta})\pi_\Theta(\boldsymbol{\theta})}{\pi_M(\mathbf{x})} \propto f(\mathbf{x} \mid \boldsymbol{\theta})\pi_\Theta(\boldsymbol{\theta}).$$

Proof. The proof follows by using first the definition of conditional pdf for continuous variables (analogously for discrete variables):

$$\begin{aligned}
\pi(\boldsymbol{\theta} \mid \mathbf{x}) &= \frac{f_{\boldsymbol{\theta},\mathbf{X}}(\boldsymbol{\theta}, \mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \\
&= \frac{f_{\boldsymbol{\theta},\mathbf{X}}(\boldsymbol{\theta}, \mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})} \cdot \frac{\pi_\Theta(\boldsymbol{\theta})}{\pi_\Theta(\boldsymbol{\theta})} \\
&= \frac{f(\mathbf{x} \mid \boldsymbol{\theta})\pi_\Theta(\boldsymbol{\theta})}{f_{\mathbf{X}}(\mathbf{x})}.
\end{aligned}$$

The result follows by applying the law of total probability (see preliminary material) to $f_{\mathbf{X}}(\mathbf{x})$:

$$f_{\mathbf{X}}(\mathbf{x}) = \pi_M(\mathbf{x}) = \int_\Theta f(\mathbf{x} \mid \boldsymbol{\theta})\pi_\Theta(\boldsymbol{\theta})d\boldsymbol{\theta}.$$

$\pi(\boldsymbol{\theta} \mid \mathbf{x})$ is known as the posterior distribution of $\boldsymbol{\theta}$ given the data $\mathbf{x}$. Let us discuss the interpretation of the elements in the posterior distribution.

- The prior distribution, $\pi_\Theta(\boldsymbol{\theta})$, represents the uncertainty about the (true value of) parameter $\boldsymbol{\theta}$ before the data are collected (note that it does not depend on the data). It is typically chosen by translating the prior "expert" knowledge about a phenomenon into a probability distribution. The parameters of the prior distribution are called **hyperparameters**.
- The likelihood function, $f(\mathbf{x} \mid \boldsymbol{\theta})$, is the same likelihood function that we studied before. This function links the data $\mathbf{x}$ and the parameters $\boldsymbol{\theta}$, and contains all the information in the data about the parameters (Likelihood Principle).
- The marginalisation constant (also known as marginal likelihood), $\pi_M(\mathbf{x})$, plays an important role in Bayesian hypothesis testing.
- The posterior distribution, $\pi(\boldsymbol{\theta} \mid \mathbf{x})$, represents the uncertainty about the parameters $\boldsymbol{\theta}$ given the data $\mathbf{x}$. It represents an update of the information we had *a priori* (before observing the data), once we have observed the data.

The idea of updating our prior knowledge after observing the data is known as the "Actualisation Principle":

Prior information $\rightarrow$ Collect data $\rightarrow$ Posterior distribution

Example 5.3.1. Beta-Binomial model. Let $X \sim Bin(n, \theta)$, for a fixed value of $n \in \mathbb{N}$, and $\theta \in (0, 1)$. Thus, the likelihood function is:

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

Suppose now that the prior distribution for θ is

$$\pi_\Theta(\theta) = I_{(0,1)}(\theta),$$

where $I_{(0,1)}(\theta)$ is the indicator function of the interval $(0, 1)$. That is, we are assuming that $\theta \sim U(0, 1)$. This assumption is often interpreted as the noninformative prior, as this prior gives equal probability to any two subintervals $I_1, I_2 \subset (0, 1)$ of equal length. Then, the posterior distribution is, up to a proportionality constant,

$$\pi(\theta \mid x) \propto f(x|\theta)\pi_\Theta(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \propto \theta^x(1-\theta)^{n-x}.$$

In order to obtain the normalising constant, we can go the long route, which consists of calculating the integral over $(0, 1)$ with respect to θ, or we can identify (by eye) the structure of this function (the kernel, see preliminary material), which corresponds to a Beta pdf (see preliminary material) with shape parameters $a = x + 1$ and $b = n - x + 1$. Thus, the posterior distribution is

$$\pi(\theta \mid x) = \frac{\theta^x(1-\theta)^{n-x}}{B(x+1, n-x+1)},$$

where $B(\cdot, \cdot)$ is the Beta function (see preliminary material). Recall that the Uniform distribution is a particular case of the Beta distribution for the case when both parameters are equal to 1. Thus, in this case, the posterior distribution and the prior distribution are Beta distributions (with different parameters). This example is essentially what Thomas Bayes published in his paper.

We can generalise this result to the case where the prior distribution is a Beta distribution with are arbitrary hyperparameters $a, b > 0$. This is,

$$\pi_\Theta(\theta \mid a_0, b_0) = \frac{\theta^{a_0-1}(1-\theta)^{b_0-1}}{B(a_0, b_0)},$$

Thus, the posterior distribution is, up to a proportionality constant,

$$\pi(\theta \mid x, a_0, b_0) \propto f(x|\theta)\pi_\Theta(\theta \mid a_0, b_0) = \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\theta^{a_0-1}(1-\theta)^{b_0-1}}{B(a_0, b_0)} \propto \theta^{x+a_0-1}(1-\theta)^{n-x+b_0-1}.$$

We can identify this function as the kernel of the Beta distribution with parameters $a = x + a_0$ and $b = n - x + b_0$. Thus, the prior and the posterior distribution are Beta distributions.

Additional details, examples, and R code can be found at:

`http://rpubs.com/FJRubio/BetaBinomial`

In the previous example, we illustrated a case where the prior and the distribution belong to the same family of distributions, and they only differ on the values of the corresponding parameters. The prior distribution only depends on the hyperparameters (a_0, b_0), while the posterior distribution is a Beta distribution with parameters $(x + a_0, n - x + b_0)$. So, we can see how the shape parameters are updated once we observe the data. The fact that the prior and the posterior distributions belong to the same family is an appealing property as we know the properties of this distribution. However, this phenomenon is not always the case, as we will see in a later exercise. The priors that produce a posterior distribution in the same family of distributions are known as *Conjugate Priors*, and they were appealing in the past as they greatly simplified any calculations. Formally,

Definition 45. A family $\mathcal{F}$ of probability distributions on Θ is said to be conjugate (or closed under sampling) for a likelihood function $f(\mathbf{x} \mid \boldsymbol{\theta})$ if, for every prior $\pi_\Theta \in \mathcal{F}$, the posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ also belongs to $\mathcal{F}$.

Of course if $\mathcal{F}$ is the space of all distributions, then any prior is conjugate! The idea is to restrict $\mathcal{F}$ to a known family of distributions. In the previous example, the family $\mathcal{F}$ corresponds to the family of Beta distributions with parameters $a, b > 0$. We say that the Beta distribution is a conjugate prior for the Binomial sampling model.

Example 5.3.2. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be an independent sample from the Poisson distribution with mean $\lambda > 0$. As calculated previously, the likelihood function is, up to a proportionality constant,

$$f(\mathbf{x} \mid \theta) \propto \lambda^{n\bar{\mathbf{x}}} e^{-n\lambda}. \tag{5.3.1}$$

Consider the Gamma prior distribution for the parameter λ,

$$\begin{aligned} \pi_\Theta(\lambda \mid \kappa_0, \theta_0) &= \frac{1}{\Gamma(\kappa_0)\theta_0^{\kappa_0}} \lambda^{\kappa_0 - 1} \exp\left\{-\frac{\lambda}{\theta_0}\right\} \\ &\propto \lambda^{\kappa_0 - 1} \exp\left\{-\frac{\lambda}{\theta_0}\right\}, \end{aligned}$$

where $\kappa_0 > 0$ and $\theta_0 > 0$. Then, the posterior distribution is proportional to (omitting the dependence on κ_0 and θ_0 in order to simplify notation)

$$\pi(\lambda \mid \mathbf{x}) \propto \lambda^{n\bar{\mathbf{x}} + \kappa_0 - 1} \exp\left\{-\lambda\left(\frac{1 + n\theta_0}{\theta_0}\right)\right\}.$$

Thus, we can identify this function as the kernel of a Gamma distribution with shape parameter $\kappa = n\bar{\mathbf{x}} + \kappa_0$, and scale parameter $\theta = \frac{\theta_0}{1 + n\theta_0}$. Consequently, the Gamma distribution is a conjugate prior distribution for Poisson sampling model.

Example 5.3.3. The normal distribution with known variance. We will analyse two cases: the first one illustrates the conjugate prior for μ, and the second one illustrates the elicitation of a prior in the case when there is expert knowledge. In order to illustrate the use of reparameterisations, we will use the parameterisation of the normal distribution in terms of the mean μ and the precision $\tau = \dfrac{1}{\sigma^2}$.

- Suppose the data point has distribution $y \sim N(\mu, \sigma^2 = \tau^{-1})$, that τ is known, and that the prior for μ is normal with mean ν and precision γ. Then, remembering that we are deriving the posterior for μ, so that all other terms are constants, the posterior of μ is $\pi(\mu \mid y) =$

$$\begin{aligned} \left(\frac{\tau}{2\pi}\right)^{1/2} \exp\left\{-\frac{\tau}{2}(y - \mu)^2\right\} \quad &\times \quad \left(\frac{\gamma}{2\pi}\right)^{1/2} \exp\left\{-\frac{\gamma}{2}(\mu - \nu)^2\right\} \\ &\propto \exp\left\{-\frac{\tau}{2}(y - \mu)^2 - \frac{\gamma}{2}(\mu - \nu)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[(\tau + \gamma)\mu^2 - 2(\tau y + \nu\gamma)\mu\right]\right\} \\ &\propto \exp\left\{-\frac{(\tau + \gamma)}{2}\left(\mu - \frac{\tau y + \nu\gamma}{\tau + \gamma}\right)^2\right\} \\ &\propto N\left\{\frac{\tau y + \gamma\nu}{\tau + \gamma}, \text{precision } = (\tau + \gamma)\right\}. \end{aligned}$$

  Thus, the prior and the posterior are normal and we can conclude that the Normal prior is a conjugate prior for μ when the variance (precision) is known.

  Additional details and examples in R can be found at:

  `http://www.rpubs.com/FJRubio/NormalNormalKV`

- Suppose that you want to model the heights of females in the age group $21-25$ years in Mexico. There is an expert from INEGI (National Institute of Statistics and Geography) Mexico who tells you that she expects that the average height should be $160cm$ based on previous studies, and with high probability around $150cm-170cm$, and that based on previous studies, the variance σ_0^2 is known. Suppose that you decide to use the previous normal model with a conjugate prior. How do you select the hyperparameters?

  One possibility is to use the expert information "the average height should be $155cm$" and translate it into $\nu = 155$. The second piece of information is that the heights should be "high probability around $150cm-170cm$". Thus, if we select $\gamma = \dfrac{1}{Var[\mu]} = 0.0385$ (equivalently, a variance of 5.1), we obtain that this prior cumulates approximately 95% of the mass on this range:

  $$\int_{150}^{170} \pi(\mu)d\mu \approx 0.95.$$

  Thus, selecting a normal prior with mean $\nu = 160$ and precision $\gamma = 0.0385$ reflects the prior knowledge provided by the expert.

  The work of Bayesian statisticians often relies on being able to obtain the prior information about the parameters of a model from the experts. This requires understanding the jargon used by the experts, as they may not necessarily be statisticians. In many cases, the experts have a background on medicine, biology, banking, politics, or economy, and may not be familiar with Bayesian statistics.

Improper priors

Definition 46. An improper prior is a function of the parameters that has infinite integral. This is,

$$\int_{\Theta} \pi_{\Theta}(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty.$$

The use of improper priors is highly controversial in practice as they are not probability distributions. However, in some cases, the use of such priors produce a posterior distribution. This is,

$$\int_{\Theta} f(\mathbf{x} \mid \boldsymbol{\theta})\pi_{\Theta}(\boldsymbol{\theta})d\theta < \infty.$$

and therefore, it is possible to normalise the product of the likelihood and the (improper prior).

Example 5.3.4. Consider the setting in Example 5.3.1, which is Binomial sampling model with a Beta prior. The parameters or the Beta prior a_0, b_0 are restricted to be positive in order to obtain a proper distribution. Omit this condition and consider the improper prior

$$\pi_{\Theta}(\boldsymbol{\theta}) = \frac{1}{\theta(1-\theta)}.$$

This prior is known as Haldane's prior. It is easy to see that this is an improper prior by checking that (for instance, integrate over $(0, 0.5)$)

$$\int_0^1 \frac{1}{\theta(1-\theta)}d\theta = \int_0^1 \left(\frac{1}{\theta} + \frac{1}{1-\theta}\right)d\theta = \infty.$$

Then,

$$\pi(\theta \mid x, a_0 = 0, b_0 = 0) \propto f(x|\theta)\pi_{\Theta}(\theta \mid a_0, b_0) \propto \theta^{x-1}(1-\theta)^{n-x-1}.$$

This is the kernel of a Beta distribution with parameters $a = x$, $b = n - x$, which is a distribution provided that $n > x > 0$. Therefore, the posterior distribution is a well-defined distribution despite the prior is improper under some additional conditions (in this case, that not all observations are zero).

Later, we will illustrate some scenarios where the use of improper priors might be of interest.

Warning 4. There exist cases where improper priors produce improper posteriors. This is,

$$\begin{aligned}\int_{\Theta} \pi_{\Theta}(\boldsymbol{\theta}) d\theta &= \infty, \\ \int_{\Theta} f(\mathbf{x} \mid \boldsymbol{\theta}) \pi_{\Theta}(\boldsymbol{\theta}) d\theta &= \infty.\end{aligned}$$

The use of improper posteriors is not justified from a formal probabilistic perspective. Thus, if at some point you employ an improper prior, always check that the corresponding posterior is proper.

5.4 The Jeffreys Prior

Sir Harold Jeffreys, a British Statistician, developed a theory for constructing prior distributions that are invariant under reparameterisations. He came up with the rule for constructing a prior using

$$\pi_J(\boldsymbol{\theta}) \propto (\det[I(\boldsymbol{\theta})])^{\frac{1}{2}} .$$

where $I(\cdot)$ represents the Fisher information matrix, and det denotes the determinant of this matrix. This prior has been widely used in many models. One problem with this prior is that it produces an improper prior for some models, making necessary to conduct additional studies about the propriety (properness) of the posterior distribution, as previously discussed. Let us illustrate the use of the Jeffreys prior with some examples.

Example 5.4.1. Let $X \sim Bernoulli(\theta)$ be a random variable. Recall that $X \in \{0, 1\}$. Then, the pdf of X is

$$f(x \mid \theta) = \theta^x (1-\theta)^{1-x}.$$

The Fisher information associated to this distribution can be calculated as follows.

$$\begin{aligned}\log f(x \mid \theta) &= x \log(\theta) + (1-x) \log(1-\theta), \\ \frac{\partial}{\partial \theta} \log f(x \mid \theta) &= \frac{x}{\theta} - \frac{1-x}{1-\theta}, \\ \left(\frac{\partial}{\partial \theta} \log f(x \mid \theta) \right)^2 &= \frac{(x-\theta)^2}{\theta^2 (1-\theta)^2}, \\ E\left(\frac{\partial}{\partial \theta} \log f(X \mid \theta) \right)^2 &= \frac{(1-\theta)^2}{\theta^2 (1-\theta)^2} \theta + \frac{(0-\theta)^2}{\theta^2 (1-\theta)^2} (1-\theta).\end{aligned}$$

Thus,

$$I(\theta) = \frac{1}{\theta(1-\theta)},$$

and the Jeffreys prior is

$$\pi_J(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}}.$$

We can identify this as the kernel of Beta distribution with parameters $a = b = 1/2$. Consequently, the Jeffreys prior is proper in this case, since it is a well-defined distribution.

Calculating the Bayes' estimates associated to this prior is simple, using the previous examples.

The following examples shows that, in many cases, the Jeffreys prior is improper.

Example 5.4.2. Let X be a random variable such that its distribution is a member of the location-scale family discussed in Example 2.6.1, with parameters $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$. In Example 2.6.1, we showed that the FIM is given by

$$I(\mu, \sigma) = \frac{1}{\sigma^2}\mathbf{M},$$

where $\mathbf{M}$ is a 2×2 matrix that does not depend on the parameters (μ, σ). Then, it follows that

$$\begin{aligned} \pi_J(\mu, \sigma) &\propto \det[I(\mu, \sigma)]^{\frac{1}{2}} \\ &= \frac{1}{\sigma^2}\det(\mathbf{M})^{\frac{1}{2}} \\ &\propto \frac{1}{\sigma^2}. \end{aligned}$$

So, the Jeffreys prior has a simple form for the entire location-scale family. However, it is easy to show that

$$\int_0^\infty \int_{-\infty}^\infty \pi_J(\mu, \sigma) = \infty.$$

Thus, the Jeffreys prior is improper.

Theorem 5.4.1. *Consider a pdf $f(x \mid \theta)$, $\theta \in \mathbb{R}$, and the Jeffreys prior*

$$\pi_J(\theta) \propto \sqrt{I(\theta)}.$$

The Jeffreys prior is invariant under reparameterisations.

Proof. Let $\theta \in \mathbb{R}$ and $\varphi : \mathbb{R} \to \mathbb{R}$. Let $\eta = \varphi(\theta)$, we want to prove that the Jeffreys prior of η is $\pi_J(\eta) \propto \sqrt{I(\eta)}$. In order to show this, we will use the change of variables theorem from calculus and the chain rule for derivatives. By the change of variables theorem,

$$\begin{aligned} \pi_J(\eta) &= \pi_J(\theta)\left|\frac{\partial\theta}{\partial\eta}\right| \\ &\propto \sqrt{I(\theta)\left(\frac{\partial\theta}{\partial\eta}\right)^2} \end{aligned}$$

By definition,

$$\begin{aligned} &= \sqrt{E\left[\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2\right]\left(\frac{\partial\theta}{\partial\eta}\right)^2} \\ &= \sqrt{E\left[\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\frac{\partial\theta}{\partial\eta}\right)^2\right]} \end{aligned}$$

By the chain rule,

$$\begin{aligned} &= \sqrt{E\left[\left(\frac{\partial}{\partial\eta}\log f(x;\eta)\right)^2\right]} \\ &= \sqrt{I(\eta)}. \end{aligned}$$

The multiparameter case follows similarly, although the notation is more complicated. It is common to come across textbooks or scientific papers that appeal to this invariance property of the Jeffreys prior in order to justify it as a "noninformative prior", since the prior has the same genesis under any parameterisation.

5.5 Bayesian Point Estimation

Previously, we analysed different frequentist estimation methods. These methods were originated by a system of estimating equations (which can be translated as an optimisation problem in some cases). The Bayesian estimation theory, in contrast, relates to a general mathematical theory called Decision Theory, in which estimating a parameter can be seen as "making a decision" in such a way that a certain "loss function" is minimised.

Suppose that we are interested in estimating a function of the parameter $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$, say $h(\boldsymbol{\theta})$, $h : \mathbb{R}^p \to \mathbb{R}^q$, $q \leq p$. In many cases, this function h is the identity function $h(\boldsymbol{\theta}) = \boldsymbol{\theta}$. The decision-theoretic approach to conduct statistical inference requires the specification of a *loss function* $L(\boldsymbol{\theta}, \boldsymbol{\delta})$, which represents the loss incurred by estimating $h(\boldsymbol{\theta})$ using the value $\boldsymbol{\delta}$. The aim is then to choose the estimator that minimises, with respect to $\boldsymbol{\delta}$,

$$\mathbb{E}_{\pi(\boldsymbol{\theta}|\mathbf{x})}\left[L(\boldsymbol{\theta}, \boldsymbol{\delta})\right] = \int_{\Theta} L(\boldsymbol{\theta}, \boldsymbol{\delta})\pi(\boldsymbol{\theta} \mid \mathbf{x})d\boldsymbol{\theta}.$$

This function is known as the Conditional Risk. More specifically,

Definition 47. A Bayes' estimator is the value that minimises the posterior expected value of a loss function.

$$\tilde{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\delta}\in\Theta}\mathbb{E}_{\pi(\boldsymbol{\theta}|\mathbf{x})}\left[L(\boldsymbol{\theta}, \boldsymbol{\delta})\right].$$

In general, it may be difficult to find the Bayes estimators for different choices of the loss function and the posterior distribution. However, there exist some loss functions that produce an explicit solution.

- **Quadratic loss**. Suppose that

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}) = ||h(\boldsymbol{\theta}) - \boldsymbol{\delta}||^2 = \left(h(\boldsymbol{\theta}) - \boldsymbol{\delta}\right)^\top \left(h(\boldsymbol{\theta}) - \boldsymbol{\delta}\right) = \sum_{i=1}^{q}(h(\boldsymbol{\theta})_i - \boldsymbol{\delta}_i)^2.$$

  This loss function is known as the quadratic loss (and has some resemblance to LSE). Then, we want to minimise the conditional risk

$$\int_{\Theta} \left(h(\boldsymbol{\theta}) - \boldsymbol{\delta}\right)^\top \left(h(\boldsymbol{\theta}) - \boldsymbol{\delta}\right) \pi(\boldsymbol{\theta} \mid \mathbf{x})d\boldsymbol{\theta}$$

  Using tools from multivariate calculus, the gradient of the conditional risk is (assuming we can interchange the integral and derivative)

$$\begin{aligned}\nabla_{\boldsymbol{\delta}} \int_{\Theta} \left(h(\boldsymbol{\theta}) - \boldsymbol{\delta}\right)^\top \left(h(\boldsymbol{\theta}) - \boldsymbol{\delta}\right) \pi(\boldsymbol{\theta} \mid \mathbf{x})d\boldsymbol{\theta} &= \int_{\Theta} \nabla_{\boldsymbol{\delta}} \left(h(\boldsymbol{\theta}) - \boldsymbol{\delta}\right)^\top \left(h(\boldsymbol{\theta}) - \boldsymbol{\delta}\right) \pi(\boldsymbol{\theta} \mid \mathbf{x})d\boldsymbol{\theta} \\ &= -2\int_{\Theta} \left(h(\boldsymbol{\theta}) - \boldsymbol{\delta}\right) \pi(\boldsymbol{\theta} \mid \mathbf{x})d\boldsymbol{\theta}\end{aligned}$$

  The Hessian matrix is twice the identity matrix

$$-2\int_{\Theta} \nabla_{\boldsymbol{\delta}}^\top \left(h(\boldsymbol{\theta}) - \boldsymbol{\delta}\right) \pi(\boldsymbol{\theta} \mid \mathbf{x})d\boldsymbol{\theta} = 2\mathbf{I}_q,$$

  which is positive definite. Thus, we have that the solution to

$$-2\int_{\Theta} \left(h(\boldsymbol{\theta}) - \boldsymbol{\delta}\right) \pi(\boldsymbol{\theta} \mid \mathbf{x})d\boldsymbol{\theta} = \mathbf{0},$$

  is a minimum. Therefore, the Bayes' estimator is

$$\int_{\Theta} h(\boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \mathbf{x})d\boldsymbol{\theta} = \int_{\Theta} \tilde{\boldsymbol{\theta}}\pi(\boldsymbol{\theta} \mid \mathbf{x})d\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}.$$

  Recall that the posterior mean of $h(\boldsymbol{\theta})$ is the mean of $h(\boldsymbol{\theta})$ with respect to the posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{x})$. We conclude that the Bayes' estimator is the posterior mean of $h(\boldsymbol{\theta})$.

  Note also that we are not assuming much about the Bayesian model (likelihood + prior).

- **0-1 loss**. Assume now that the loss function is

$$L_0(\boldsymbol{\theta}, \boldsymbol{\delta}) = \begin{cases} 1 & \text{if } \boldsymbol{\delta} \neq \boldsymbol{\theta}, \\ 0 & \text{if } \boldsymbol{\delta} = \boldsymbol{\theta}. \end{cases}$$

This loss function can be seen as the limit of the Uniform loss function:

$$L_\varepsilon(\boldsymbol{\theta}, \boldsymbol{\delta}) = \begin{cases} 1 & \text{if } ||\boldsymbol{\delta} - \boldsymbol{\theta}|| > \varepsilon, \\ 0 & \text{if } ||\boldsymbol{\delta} - \boldsymbol{\theta}|| \leq \varepsilon, \end{cases}$$

for $\varepsilon > 0$. Under continuity arguments that we will not cover here (as this is still a research topic):

$$\tilde{\boldsymbol{\theta}} = \lim_{\varepsilon \to 0} \tilde{\boldsymbol{\theta}}_\varepsilon = \text{argmax}_{\boldsymbol{\theta}}\, \pi(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta}.$$

Thus, the Bayes estimator is the maximum (mode) of the posterior distribution. This is known as the Maximum a Posteriori (MAP) estimator, and it is often interpreted as the analogous to the MLE.

- **Absolute loss**. For the case when $p = 1$ (one-parameter models) and $h(\delta) = \delta$, we can also obtain the Bayes estimator for the absolute loss function:

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

The conditional risk is

$$\begin{aligned} R(\delta) = E_{\pi(\theta|\mathbf{x})}\left[L(\theta, \delta)\right] &= \int_{-\infty}^{\infty} |\theta - \delta| \pi(\theta \mid \mathbf{x}) d\theta \\ &= \int_{-\infty}^{\delta} (\delta - \theta) \pi(\theta \mid \mathbf{x}) d\theta + \int_{\delta}^{\infty} (\theta - \delta) \pi(\theta \mid \mathbf{x}) d\theta \\ &= \delta \int_{-\infty}^{\delta} \pi(\theta \mid \mathbf{x}) d\theta - \int_{-\infty}^{\delta} \theta \pi(\theta \mid \mathbf{x}) d\theta \\ &+ \int_{\delta}^{\infty} \theta \pi(\theta \mid \mathbf{x}) d\theta - \delta \int_{\delta}^{\infty} \pi(\theta \mid \mathbf{x}) d\theta. \end{aligned}$$

Using that $\int_{-\infty}^{\delta} \pi(\theta \mid \mathbf{x}) d\theta = 1 - \int_{\delta}^{\infty} \pi(\theta \mid \mathbf{x}) d\theta$ and $\int_{\delta}^{\infty} \theta \pi(\theta \mid \mathbf{x}) d\theta = \int_{-\infty}^{\infty} \theta \pi(\theta \mid \mathbf{x}) d\theta - \int_{-\infty}^{\delta} \theta \pi(\theta \mid \mathbf{x}) d\theta$, we obtain

$$R(\delta) = 2\delta \int_{-\infty}^{\delta} \pi(\theta \mid \mathbf{x}) d\theta - 2 \int_{-\infty}^{\delta} \theta \pi(\theta \mid \mathbf{x}) d\theta + \int_{-\infty}^{\infty} \theta \pi(\theta \mid \mathbf{x}) d\theta - \delta.$$

Differentiating $R(\delta)$ with respect to δ and equating to zero we obtain the condition:

$$2 \int_{-\infty}^{\delta} \pi(\theta \mid \mathbf{x}) d\theta = 1,$$

and

$$\int_{-\infty}^{\delta} \pi(\theta \mid \mathbf{x}) d\theta = \frac{1}{2}.$$

Now, the second derivative is $2\pi(\delta \mid \mathbf{x}) > 0$, under the assumption that the posterior has support on the entire real line, then this is a minimum. An analogous argument can be applied to other supports by integrating on the corresponding sets. Finally, $\tilde{\theta}$ is the median of the posterior distribution $\pi(\theta \mid \mathbf{x})$, by definition of median.

Example 5.5.1. Consider the formulation presented in Example 5.3.1 concerning a Binomial sampling model and a Beta prior. Then,

- The Bayes estimator under the quadratic loss is (posterior mean):

$$\frac{x+a_0}{n+a_0+b_0}.$$

- The Bayes estimator under the 0-1 loss (MAP):

$$\frac{x+a_0-1}{n+a_0+b_0-2},$$

  for $n+a_0+b_0-2>0$.

- The Bayes estimator under the absolute loss cannot be written in closed-form as the median of the Beta distribution is not available in closed form. However, it can be written in terms of the inverse of the regularised incomplete beta function, which is implemented in most numerical softwares (for example R).

An interesting feature of the Bayes estimates is that, for large values of n,

$$\frac{x+a_0}{n+a_0+b_0}=\frac{\frac{x}{n}+\frac{a_0}{n}}{1+\frac{a_0+b_0}{n}}\approx\frac{x}{n},\quad \text{(post. mean)},$$

and

$$\frac{x+a_0-1}{n+a_0+b_0-2}=\frac{\frac{x}{n}+\frac{a_0-1}{n}}{1+\frac{a_0+b_0-2}{n}}\approx\frac{x}{n},\quad \text{(MAP)},$$

Recalling that the MLE $\widehat{\theta}=\dfrac{x}{n}$, we can conclude that the Bayes estimators and the MLE are close for large n and, in fact, they are asymptotically equivalent.

Another important characteristic of these estimators is that, when $a_0=b_0=1$ (uniform prior), then the MAP coincides with the MLE. This fact is used to justify the uniform prior as a "non-informative prior". However, we can also see that the posterior mean for this choice of hyperparameters is

$$\frac{x+1}{n+2},$$

which differs from the MLE.

Now, for $a_0=b_0=0$, the posterior is proper as long as $n>x>0$, and the posterior mean coincides with the MLE. However, the MAP (assuming $n>2$) does not coincide with the MLE

$$\frac{x-1}{n-2}.$$

Finally, the posterior variance (the variance of the posterior distribution) is

$$Var_{\theta|x}[\theta]=\frac{(x+a_0)(n-x+b_0)}{(n+a_0+b_0)^2(n+a_0+b_0+1)},$$

so, we can see that $Var_{\theta|x}[\theta]\to 0$ as $n\to\infty$, for any values of a_0 and b_0. Thus, the posterior distribution concentrate around a point (the true value of the parameter) as the sample grows.

Additional details and R code about this example can be found at:

`http://rpubs.com/FJRubio/TwoPriorsBinomial`

Remark 5.5.1. *Although studying the asymptotic behaviour of the Bayes estimates is beyond the aim of this course, from the previous example we can expect that some Bayes estimates are consistent and asymptotically normal as they coincide, or are very similar, to the MLE. In fact, under some general "regularity conditions", the Bayes estimate obtained with the square loss, $\tilde{\theta}$, is asymptotically normal. This is,*

$$\sqrt{n}\left(\tilde{\theta}-\theta_0\right) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right), \quad as \ n \to \infty,$$

where θ_0 is the true value of the parameter, and $I(\cdot)$ is the Fisher information. A surprising result is that this asymptotic behaviour is valid for any prior distribution that satisfies the regularity conditions. A general result that characterises the asymptotic properties of posterior distributions and their independence from the choice of the prior is the Bernstein-von Mises Theorem*, which is a research topic in modern statistics.*

In the previous examples of Bayes estimates, we focused on the use of symmetric losses. That is, the loss functions are symmetric in the sense that $L(\boldsymbol{\theta}, \boldsymbol{\delta}) = L(\boldsymbol{\theta}, -\boldsymbol{\delta})$. Thus, this implicitly assumes that the loss incurred by underestimating is the same as the loss incurred by overestimating the parameter $\boldsymbol{\theta}$. A natural question is: Are symmetric loss functions always the best choice? In order to answer this question, consider the following example.

Example 5.5.2. Consider the following *hypothetical* scenario in the context of optical power. **Suppose that the ideal vision score of a patient is** 0 and that this score can depart in both directions (positive and negative). For instance, positive values of the score indicate that the patient has far-sightedness (hyperopia), in which close objects appear to be blurry, while far objects appear normal. On the other hand, negative values indicate that the patient has near-sightedness (myopia) in which distant objects appear to be blurry while close objects appear normal. The magnitude of the score represents the severity of the sight problem, and it is obtained with an accurate medical test.

Now, suppose also that your vision score is $+5$. After a laser surgery, what would you prefer:

1. a score of $+0.25$,
2. a score of -0.25,
3. you don't mind as long as the magnitude is 0.25?

Suppose that the parameter $\theta \in \mathbb{R}$ of a Bayesian model controls the score obtained for a given intensity of the laser used for the surgery, and the score and θ are directly proportional. Thus, we need to calibrate the intensity order to obtain, with higher probability, the preferred outcome. For example, if you chose option 1, it is better to slightly underestimate it than slightly overestimate it. Consequently, the loss function is asymmetric. Only if you choose option 3, the corresponding loss function is symmetric.

5.6 The Predictive Distribution

A powerful tool in the Bayesian framework is that it can be formally used to make predictions. Predictions can be made based solely on prior knowledge or based on a sample.

Definition 48. Let $X \in \mathbb{R}$ be a random variable with pdf (pmf) $f(\cdot|\boldsymbol{\theta})$ and let $\pi_\Theta(\boldsymbol{\theta})$ be the prior distribution on $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. The **prior predictive probability density (mass) function** is defined as the marginal pdf of an observation, with respect to the prior distribution. This is,

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

The intuition behind the prior predictive pdf (pmf) is as follows. If we knew the true value of the parameter $\boldsymbol{\theta}_0$, we could predict a new observation with the distribution $f(\cdot|\boldsymbol{\theta}_0)$. For instance, suppose that

$$P(X \in (L, U)) = \int_L^U f(x|\boldsymbol{\theta}_0)dx = 0.95.$$

Then, we could say that, with 95% probability, the next observation would be within the interval (L, U). However, in real life we do not know the true value $\boldsymbol{\theta}_0$. Nonetheless, before collecting any data we have a prior distribution $\pi_\Theta(\boldsymbol{\theta})$ that quantifies the uncertainty about this parameter. Thus, we can include this uncertainty in our prediction by integrating $\boldsymbol{\theta}$ out with respect to this prior measure of uncertainty, using the rules of probability (law of total probability).

The prior predictive distribution is used to assess or design clinical trials, as we can predict the outcome of a trial if we can come up with a reasonable prior. This is particularly important in some cases where regulations require a trial to have a minimum number of successes.

Example 5.6.1. Let X be a binomial random variable for n trials with probability of success θ. Consider a Beta prior for θ with hyperparameters $a_0, b_0 > 0$. This is the Beta-Binomial model studied before. The prior predictive pmf for the outcome of n trials,

$$\begin{aligned} p(x^\star) &= \int_0^1 \binom{n}{x^\star} \theta^{x^\star}(1-\theta)^{n-x^\star} \frac{\theta^{a_0-1}(1-\theta)^{b_0-1}}{B(a_0, b_0)} d\theta \\ &= \frac{\binom{n}{x^\star}}{B(a_0, b_0)} \int_0^1 \theta^{x^\star+a_0-1}(1-\theta)^{n-x^\star+b_0-1} d\theta \\ &\quad \text{By definition of the Beta function,} \\ &= \binom{n}{x^\star} \frac{B(x^\star + a_0, n - x^\star + b_0)}{B(a_0, b_0)}. \end{aligned}$$

Thus, once we choose the values of the hyperparameters, we can calculate the probabilities of success for $x^\star = 0, 1, \ldots, n$.

Once we have collected some data, it is also possible to predict the following observations. This is also used in clinical trials, where usually a small group of individuals are studied, in order to obtain some information and combine it with the prior information. This information is then used in the following phases of the trial. This idea is formalised in the posterior predictive distribution, as follows.

Definition 49. Let $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be *i.i.d.* random variables with pdf (pmf) $f(\cdot|\boldsymbol{\theta})$, let $\pi_\Theta(\boldsymbol{\theta})$ be the prior distribution on $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$, and let $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ be the corresponding posterior distribution. The **posterior predictive probability density (mass) function** is defined as the marginal pdf of an observation, with respect to the posterior distribution. This is,

$$f(x^\star \mid \mathbf{x}) = \int_\Theta f(x^\star|\boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \mathbf{x})d\boldsymbol{\theta}.$$

The intuition behind the posterior predictive pdf (pmf) is as follows. Again, if we knew the true value of the parameter $\boldsymbol{\theta}_0$, we could predict a new observation with the distribution $f(\cdot|\boldsymbol{\theta}_0)$. However, we do not know the true value $\boldsymbol{\theta}_0$. After collecting a sample $\mathbf{x}$, we have a posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ that quantifies the uncertainty about this parameter in the light of the data. Thus, we can include this uncertainty in our prediction by integrating $\boldsymbol{\theta}$ out with respect to this posterior measure of uncertainty, using the rules of probability.

Example 5.6.2. Let $r \in \{0, 1, \ldots, n\}$ be the outcome of n trials with probability of success θ. Consider a Beta prior for θ with hyperparameters $a_0, b_0 > 0$. This is the Beta-Binomial model studied before, and the posterior is a Beta distribution with parameters $a = a_0 + r$ and $b = n - r + b_0$. The posterior predictive pmf for the outcome of $n^\star$ new trials is

$$p(x^\star \mid r) = \int_0^1 \binom{n^\star}{x^\star} \theta^{x^\star}(1-\theta)^{n^\star-x^\star} \frac{\theta^{a_0+r-1}(1-\theta)^{n-r+b_0-1}}{B(a_0+r, n-r+b_0)} d\theta$$

$$= \frac{\binom{n^\star}{x^\star}}{B(a_0+r, n-r+b_0)} \int_0^1 \theta^{x^\star+a_0+r-1}(1-\theta)^{n^\star-x^\star+n-r+b_0-1} d\theta$$

By definition of the Beta function,

$$= \binom{n^\star}{x^\star} \frac{B(x^\star+a_0+r, n^\star-x^\star+n-r+b_0)}{B(a_0+r, n-r+b_0)}.$$

Additional details and examples in R can be found at:

`http://www.rpubs.com/FJRubio/BetaBinomialPred`

Example 5.6.3. Suppose that we are interested in estimating the proportion of responders to a new treatment for a serious disease. The company that developed the treatment is allowed to run a trial for the first time on $n = 10$ (human) patients. You are the Statistician responsible for analysing the results of the trial. For illustrative purposes, consider the following scenarios:

(i) You decide to use a MLE approach to estimate θ, the probability of success.

(ii) You decide to use the Jeffreys prior, since you think that it is non-informative (in what sense?). This is, you use a prior distribution $\theta \sim Beta(1/2, 1/2)$.

(iii) You decide to use a flat priors, since this prior is non-informative (in what sense?). This is, you use a prior distribution $\theta \sim Beta(1, 1)$.

(iv) The team of Pharmaceutical scientists that developed the drug tell you that they are very optimistic about this drug and that, based on previous experiments on mice, they expect the drug to be successful in average on 75% of the cases, and they are certain that the probability of success should be higher than 60%. They also acknowledge that due to genetic variability, the drug may not work in 10% of the patients, so the probability of success is lower than 90%. If you use this information, you can construct a prior by trying different values of the Beta distribution to match this information. First, the mean is $\frac{a_0}{a_0+b_0} = 0.75$, then $a_0 = 3b_0$. Then, you play with the values of the parameters until you obtain that 90% of the mass of the prior distribution is cumulated between $(0.6, 0.90)$. Other values could be used as well instead of 90%, such as 95%. This is,

$$\int_{0.6}^{0.9} \pi_\Theta(\theta \mid a_0, b_0) d\theta \approx 0.90.$$

After some tests, you obtain $a_0 = 15$ and $b_0 = 5$. This is an *informative* prior (in the sense that it incorporates the information from the experts).

Suppose that the company runs the experiment and they obtain that $x = 3$ patients out of $n = 10$ are responders (successes), and $n - x = 7$ are non-responders. What are the posteriors distributions for each prior choice?

(i) No posterior.

(ii) $Beta(3.5, 7.5)$.

(iii) $Beta(4, 8)$

(iv) $Beta(18, 12)$.

Now, suppose that you decide to use the posterior mean as a point estimator. What are the point estimators and the posterior variances for each prior choice? Include the MLE and its variance. Table 5.6.3 shows the results.

Suppose that these results are used as an initial trial in order to predict the results in a new trial now involving $n^\star = 100$ patients. What is the predictive distribution associated to each prior? The predictive

	Frequentist	Informative	Jeffreys	Uniform
Estimators	0.3000	0.6000	0.3182	0.3333
Variances	0.0210	0.0077	0.0181	0.0171

distribution is Beta-Binomial distribution distribution with parameters given by the hyperparameters of each prior. Note that, in the frequentist scenario, prediction can only be done by using the "fitted model", which is a Binomial distribution for $n^\star = 100$ trials and probability of success $\widehat{\theta}$. This approach ignores the uncertainty about the parameter.

The R code to produce these numerical calculations can be found at:

`http://www.rpubs.com/FJRubio/BayesBinomTrial`

See also:

`http://rpubs.com/FJRubio/BetaBinomial`

`http://rpubs.com/FJRubio/BetaBinomialPred`

Prior-Data Conflict

In the Bayesian framework, the statistician is confronted with the task of coming up with a prior distribution, which represents a translation of the prior information coming from the experts. This information is then combined with the information in the data through the posterior distribution. There is a chance that the prior information and the data are in disagreement, as illustrated in Figure 5.6.1.

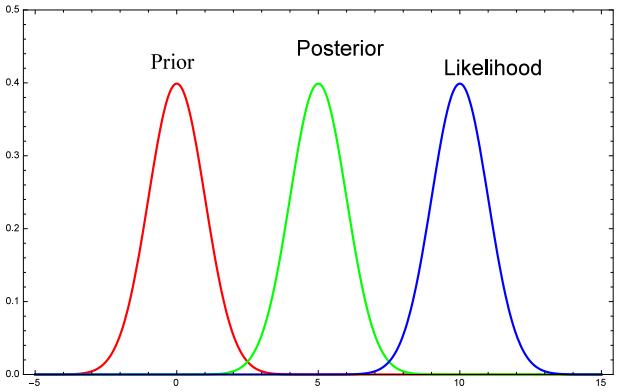


Figure 5.6.1: Prior-Data Conflict

The reasons for this possible disagreement are of very different nature, and it is not always a bad thing. For example:

(i) Bad prior information + Good Data. It is, of course, possible that the experts do not have good prior information as the phenomenon of interest may be challenging or new. In this case, it is also important to understand how accurate the prior information is, as some experts might be overconfident in reporting their knowledge. For this reason, sometimes more than one expert are consulted in order to contrast their prior information.

(ii) Good prior information + Bad Data. In this case, the prior knowledge is good, but there may be a problem with the data collection, leading to bad quality data. Either a bad design, an error in capturing and collecting the data, measurement errors, and etcetera.

(iii) Good prior information + Good Data but extreme observations - outliers. In this case, both the prior information and the data are of good quality. However, it is possible, particularly with small samples, to obtain extreme observations. For instance, consider a binomial trial with $n = 10$

individuals and $\theta_0 = 0.9$. Then, $P(X = 0) > 0$ and $P(X = 1) > 0$. Thus, one might observe a low number of successes due to chance, which may lead to an inaccurate estimation of θ_0. This problem is less prevalent with large or moderate samples (recall the consistency and asymptotic normality results).

(iv) Bad prior information + Bad Data ☹.

In general, prior-data conflict reflects a mismatch of the information in the prior and the data, and it is important to reflect about possible causes. Thus, it requires an extra step of reflection and understanding of the problem at hand, but if done carefully, it may lead to the detection of problems either in the prior understanding of the phenomenon of interest or in the data.

5.7 Bayesian Interval Estimation

One of the appealing properties of Bayesian estimation is that uncertainty about a parameter of interest is quantified through a probability distribution: the posterior distribution. This allows one to make probabilistic statements about the parameter, in contrast frequentist approaches. However, when reporting this distribution, it is often better to summarise it in terms of quantities of interest such as moments and quantiles. Related to reporting quantiles, it is also of interest to report probability intervals that cumulates a certain amount of mass. This motivates the following definition.

Definition 50. A **Bayesian Credible Interval** of size $1 - \alpha$, $\alpha \in (0, 1)$, is an interval $I = (L, U)$ such that:

$$P(L \leq \theta \leq U \mid \mathbf{x}) = \int_L^U \pi(\theta \mid \mathbf{x}) d\theta = 1 - \alpha.$$

Clearly, this definition does not lead to a unique credible interval. In fact, there are infinite credible intervals of size $1 - \alpha$. Some examples are shown in Figure 5.7.1

Two popular types of credible intervals are:

1. Highest Posterior Density (HPD) intervals. If the posterior distribution is unimodal, we can apply an argument similar to that used for confidence intervals to show that a HPD interval corresponds to an interval such that $\pi(L \mid \mathbf{x}) = \pi(U \mid \mathbf{x})$, and
$$\int_L^U \pi(\theta \mid \mathbf{x}) d\theta = 1 - \alpha.$$
2. Quantile intervals. This sort of intervals correspond to the case where $L = Q_{\frac{\alpha}{2}}$ and $U = Q_{1-\frac{\alpha}{2}}$. That is, the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of the posterior distribution. Although this sort of intervals might be difficult to construct when the prior is not conjugate, there exists numerical methods that can be used to approximate these quantiles.

Example 5.7.1. Calculate the 95% quantile posterior credible intervals for θ in Example 5.6.3. Note that these intervals can be constructed using the quantiles of a Beta distribution, which is already implemented in R. Report also the 95% confidence interval for θ. Recall that the interpretation of confidence intervals is different from that of credible intervals.

	Frequentist	Informative	Jeffreys	Uniform
L	0.02	0.42	0.09	0.11
U	0.58	0.76	0.61	0.61

Table 5.1: Credible intervals and confidence interval for θ.

The R code used to produce these numerical results can be found at:

`http://www.rpubs.com/FJRubio/BayesBinomTrial`

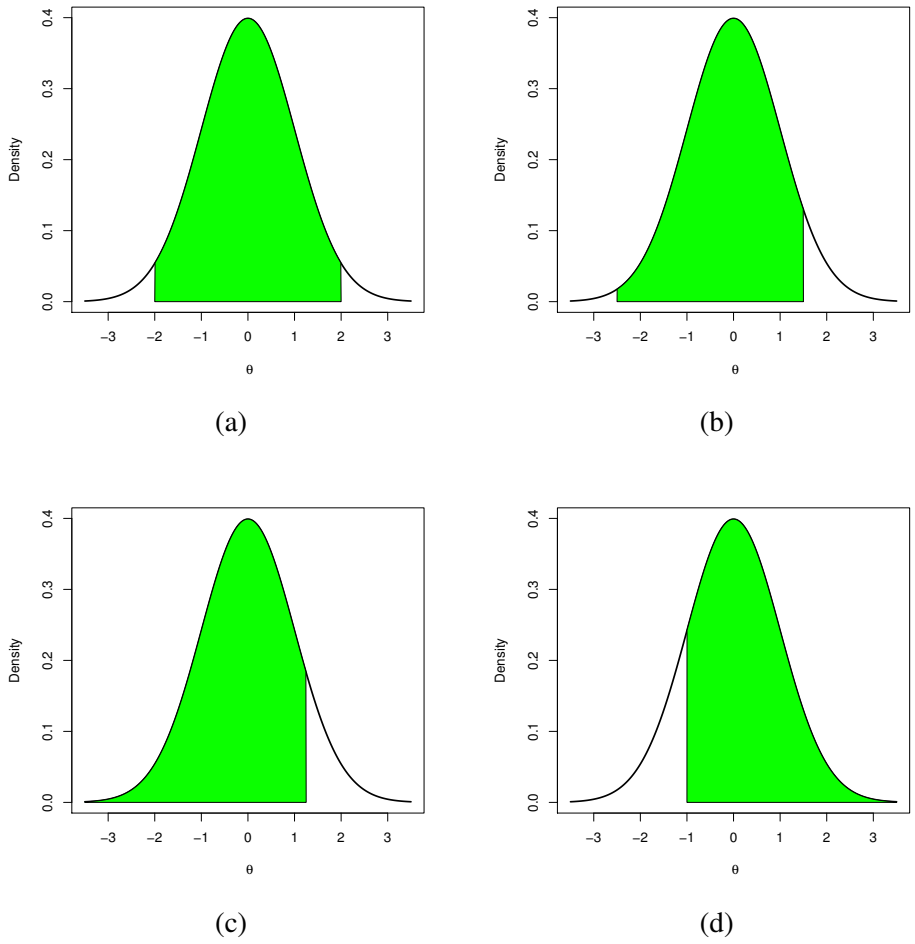


Figure 5.7.1: Different types of credible intervals.

Multiparameter case: Elimination of Parameters

In the case when the parameter $\boldsymbol{\theta} \in \mathbb{R}^p$, with $p \geq 2$, we may have again that some parameters are of interest and the remaining parameters are nuisance parameters. In the Bayesian framework, the elimination of parameters can be done using a probabilistic argument, by integrating them out of the posterior distribution. This is, suppose that $\boldsymbol{\theta} = (\boldsymbol{\delta}, \boldsymbol{\xi}) \in \mathbb{R}^p$, $\boldsymbol{\delta} \in \mathbb{R}^{p_1}$ is a parameter of interest, and $\boldsymbol{\xi} \in \mathbb{R}^{p_2}$ is a nuisance parameter, $p = p_1 + p_2$. Then, the *marginal posterior distribution* of $\boldsymbol{\delta}$ is

$$\pi(\boldsymbol{\delta} \mid \mathbf{x}) = \int_{\mathbb{R}^{p_2}} \pi(\boldsymbol{\delta}, \boldsymbol{\xi} \mid \mathbf{x}) d\boldsymbol{\xi}.$$

In particular, if $p_1 = 1$, then, we can obtain the marginal posterior distribution of δ, $\pi(\delta \mid \mathbf{x})$, by integrating out the remaining parameters from the joint posterior distribution $\pi(\delta, \boldsymbol{\xi} \mid \mathbf{x})$. This marginal distribution. This marginal posterior distribution can be used to calculate credible intervals or point estimators are discussed in previous sections.

Remark 5.7.1. Monte Carlo methods.

If we want to make inference about the model parameters, we do not need to calculate the expression of the posterior distribution, only some summary statistics: posterior mean, posterior median, posterior moments, credible intervals, posterior percentiles, standard deviation, visual tools (plots). In order to calculate these quantities, we can use a sample from the posterior distribution as follows (Monte Carlo Integration):

- *Suppose that we can get a sample from the posterior $\pi(\boldsymbol{\theta} \mid \mathbf{x})$: $(\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^N)$. Then,*

$$\textit{Posterior mean} = \int \boldsymbol{\theta}\pi(\boldsymbol{\theta} \mid \mathbf{x})d\theta \approx \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{\theta}^j.$$

$$E[g(\boldsymbol{\theta}) \mid \mathbf{x}] = \int g(\boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \mathbf{x})d\boldsymbol{\theta} \approx \frac{1}{N}\sum_{j=1}^{N} g(\boldsymbol{\theta}_j).$$

$$\textit{Posterior Prob. } \theta \in [0,1] = \int_0^1 \pi(\theta \mid \mathbf{x})d\theta \approx \frac{1}{N}\sum_{\theta_j \geq 0}^{\theta_j \leq 1} 1,$$

  where θ_j are the corresponding entries of the parameter of interest θ.

- *Other quantities, such as quantiles, can be approximated in R with the commands: $median(\cdot)$, $quantile(\cdot)$, $sd(\cdot)$, among many others.*

The justification for using these approximations is basically the law of large numbers and Slutsky's theorem. Thus, the challenge is to sample from the posterior distribution, which motivates the following remark.

Remark 5.7.2. Markov Chain Monte Carlo (MCMC).

MCMC methods are general methods to sample from a multivariate distribution. In particular, these methods can be used to sample from the posterior distribution of a parameter $\boldsymbol{\theta}$. Most of these methods do not require the normalising constant, which makes them appealing in practice.

Some examples of MCMC methods are: Metropolis-Hastings, Gibbs sampler, Hamiltonian Monte Carlo, among many others. These methods will not be covered in this course.

5.8 Bayesian Hypothesis Testing

Let $\mathbf{x} = x_1, \ldots, x_n$ be *i.i.d.* random samples from a distribution with density $f(\cdot \mid \theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$ (note that we are using conditional notation as we are not set in the Bayesian context where parameters are treated as random variables). Suppose that we are interested on testing the hypothesis:

$$H_0 : \theta \in \Theta_0 \quad vs. \quad H_1 : \theta \in \Theta_0^c.$$

Thus, for the case when, *a posteriori*, Θ_0 is a non-zero probability set, the idea is to calculate

$$P\left(H_0 \mid \mathbf{x}\right) = P\left(\theta \in \Theta_0 \mid \mathbf{x}\right) = \int_{\Theta_0} \pi(\theta \mid \mathbf{x})d\theta.$$

For instance, when the null hypothesis is $H_0 : \theta \geq 0$, then

$$P\left(H_0 \mid \mathbf{x}\right) = \int_0^{\infty} \pi(\theta \mid \mathbf{x})d\theta.$$

This quantity tells us the posterior probability of the null hypothesis, allowing you to make probabilistic statements about the null hypothesis directly, in contrast to frequentist testing. The rejection of the null hypothesis is based on this probability (for instance, if this is less than 0.5 or another value). However, the rejection is no longer based in terms of frequentist arguments.

Of course, if the posterior distribution of θ is continuous, any point null hypothesis will be assigned zero probability since

$$P\left(\theta = \theta_0 \mid \mathbf{x}\right) = 0,$$

for continuous distributions. In some cases, point hypothesis are replaced by a neighbourhood, such as $H_0 : \theta \in (-\epsilon, \epsilon)$, for some $\epsilon > 0$, instead of $H_0 : \theta = 0$. This is in line with the quote by H. Jeffreys:

"that the mere fact that it has been suggested that θ is zero corresponds to some presumption that it is fairly small"

Then, we can now calculate

$$0 < P\left(H_0 \mid \mathbf{x}\right) = \int_{-\epsilon}^{\epsilon} \pi(\theta \mid \mathbf{x}) d\theta.$$

However, there exist other alternative approaches to test point null hypothesis, as we will discuss later.

Remark 5.8.1. *In the above discussion, we are using a posterior distribution. This, of course, requires a prior distribution. The notation does not reflect the inclusion of a prior distribution, but keep in mind that a prior is implicitly there.*

Decision-Theoretic testing

Testing a hypothesis can be interpreted as making inference about an indicator function

$$I_{\Theta_0}(\theta) = \begin{cases} 1 & \text{if } \theta \in \Theta_0, \\ 0 & \text{if } \theta \notin \Theta_0. \end{cases}$$

In the Bayesian contest, the alternative is usually defined as the complement of the null, this is $H_1 : \theta \in \Theta_1 = \Theta_0^c$. However, in some cases, we may be interested in alternatives of the type $\Theta_1 \neq \Theta_0^c$, as there may be information about the support of the alternative available. For instance, $H_1 : \theta \geq 0$ instead of $H_1 : \theta \neq 0$. Such cases need to be clearly stated.

Under this formulation, every test procedure φ represents an estimator of $I_{\Theta_0}(\theta)$. Going back to the theory of Bayesian point estimation, we need a loss function $L(\theta, \varphi)$ to derive the Bayes estimator. For instance, the 0-1 loss function

$$L(\theta, \varphi) = \begin{cases} 1 & \text{if } \varphi \neq I_{\Theta_0}(\theta), \\ 0 & \text{otherwise.} \end{cases}$$

For this loss, the Bayesian solution is

$$\varphi^{\pi} = \begin{cases} 1 & \text{if } P(\theta \in \Theta_0 \mid \mathbf{x}) > P(\theta \in \Theta_0^c \mid \mathbf{x}), \\ 0 & \text{otherwise.} \end{cases}$$

This estimator is easily justified on an intuitive basis since it chooses the hypothesis with the largest posterior probability. A generalisation of this loss function is presented in the following theorem, in which a loss function that penalises differently errors when the null hypothesis is true or false.

Theorem 5.8.1. *Consider the loss function*

$$L(\theta, \varphi) = \begin{cases} 0 & \textit{if } \varphi = I_{\Theta_0}(\theta), \\ a_0 & \textit{if } \theta \in \Theta_0 \textit{ and } \varphi = 0, \\ a_1 & \textit{if } \theta \notin \Theta_0 \textit{ and } \varphi = 1. \end{cases}$$

Typically, $a_0 + a_1 = 1$, but other values can also be used.. The Bayes estimator associated with a prior π_Θ is

$$\varphi^{\pi}(\mathbf{x}) = \begin{cases} 1 & \textit{if } P(\theta \in \Theta_0 \mid \mathbf{x}) > \frac{a_1}{a_0 + a_1}, \\ 0 & \textit{otherwise.} \end{cases}$$

Proof. Since the posterior loss is

$$\begin{aligned} L(\varphi \mid \mathbf{x}) &= \int_{\Theta} L(\theta, \varphi)\pi(\theta \mid \mathbf{x})d\theta \\ &= a_0 P(\theta \in \Theta_0 \mid \mathbf{x})I_{\{0\}}(\varphi) + a_1 P(\theta \notin \Theta_0 \mid \mathbf{x})I_{\{1\}}(\varphi). \end{aligned}$$

In order to minimise this function, we observe that

- when $\varphi = 0$, $L(\varphi = 0 \mid \mathbf{x}) = a_0 P(\theta \in \Theta_0 \mid \mathbf{x})$.
- when $\varphi = 1$, $L(\varphi = 1 \mid \mathbf{x}) = a_1 P(\theta \notin \Theta_0 \mid \mathbf{x}) = a_1 - a_1 P(\theta \in \Theta_0 \mid \mathbf{x})$.

So, in order to minimise these values, we need to select $\varphi = 1$, when $a_1 - a_1 P(\theta \in \Theta_0 \mid \mathbf{x}) < a_0 P(\theta \in \Theta_0 \mid \mathbf{x})$, which implies selecting $\varphi = 1$, when

$$P(\theta \in \Theta_0 \mid \mathbf{x}) > \frac{a_1}{a_0 + a_1},$$

and $\varphi = 0$ otherwise, as desired.

This is, this kind of loss functions implies rejecting the null hypothesis H_0 when its posterior probability is too small, and the threshold is $\frac{a_1}{a_0+a_1}$. For instance, if we do not have a preference in terms of the meaning of a_0 and a_1, we can set $a_0 = a_1 = 1/2$, which leads to not rejecting the null hypothesis when its posterior probability is larger than $1/2$. In contrast, if we are very strict and we want to avoid accepting the null (not rejecting) when the null hypothesis is false, then we need a large value of a_1. For instance, if we set $a_1 = 9/10$ and $a_0 = 1/10$, this leads to accepting the null when its posterior probability is larger than 0.9, a very restrictive condition.

Example 5.8.1. Consider $x \sim Bin(n, \theta)$, and $H_0 : \theta \in \Theta_0$, where $\Theta_0 = [0, 0.5]$. Consider a Beta prior on θ, $\theta \sim Beta(a, b)$. Thus, we know that the posterior distribution is again Beta, since this is a conjugate prior. For simplicity, suppose that $a = b = 1$ (uniform prior). Then,

$$P(H_0 \mid x) = P(\theta \leq 0.5 \mid x) = \int_0^{0.5} \frac{1}{B(x+1, n-x+1)} \int_0^{0.5} \theta^x (1-\theta)^{n-x}.$$

This value can be calculated numerically for each value of n and x.

Example 5.8.2. Consider $x \sim N(\theta, \sigma^2)$ and $\theta \sim N(m, s^2)$ (conjugate prior), with σ^2 known. Then, the posterior $\pi(\theta \mid x)$ is normal with posterior mean $\mu(x) = \dfrac{\sigma^2 m + s^2 x}{\sigma^2 + s^2}$ and posterior variance $\gamma^2(x) = \dfrac{\sigma^2 s^2}{\sigma^2 + s^2}$. Consider the test

$$H_0 : \theta < 0.$$

We need to calculate

$$\begin{aligned} P(\theta < 0 \mid x) &= P\left(\frac{\theta - \mu(x)}{\gamma(x)} < -\frac{\mu(x)}{\gamma(x)}\right) \\ &= \Phi\left(-\frac{\mu(x)}{\gamma(x)}\right). \end{aligned}$$

Let Z_{a_0,a_1} be the $\frac{a_1}{a_0+a_1}$-quantile of the standard normal distribution. Then, we accept H_0 when

$$-\mu(x) > Z_{a_0,a_1}\gamma(x).$$

This is equivalent to:

$$x < -\frac{\sigma^2 + s^2}{s^2} Z_{a_0,a_1}\gamma(x) - \frac{\sigma^2}{s^2} m.$$

Exercise. It is possible to extend this result to the case with more than one observation. This implies a little more algebra, but a similar result can be obtained involving the sample mean.

A difficulty of this approach is the choice of the weights a_0 and a_1, since they are usually selected by the user rather than determined from utility considerations (mathematical rules).

The Bayes Factor

An important concept in Bayesian model comparison and hypothesis testing is the *Bayes Factor*. This quantity is based on the posterior probability of each model or test as follows.

Definition 51. The Bayes factor is the ratio of the posterior probabilities of the null and the alternative hypotheses over the ratio of the prior probabilities of the null and the alternative hypotheses, this is,

$$B_{0,1} = \frac{P(\theta \in \Theta_0 \mid \mathbf{x})}{P(\theta \in \Theta_1 \mid \mathbf{x})} \Big/ \frac{P_\pi(\theta \in \Theta_0)}{P_\pi(\theta \in \Theta_1)},$$

where P_π is the probability with respect to the prior distribution, and $P(\cdot \mid \mathbf{x})$ is the posterior probability.

Thus, the Bayes factors compares the prior and posterior probabilities of the hypotheses (null and alternative). In fact, what the Bayes factor compares are the *odds* of Θ_0 against Θ_1.

$$B_{0,1} = \frac{\text{Posterior Odds}}{\text{Prior Odds}}.$$

In the case of point hypothesis, where $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, the Bayes factors reduces to the likelihood ratio (the deduction of this result is not straightforward and consists of calculating the probabilities of the models associated to the two parameter values)

$$B_{0,1} = \frac{f(\mathbf{x} \mid \theta_0)}{f(\mathbf{x} \mid \theta_1)}.$$

Now, let us study the logic behind the Bayes factor.

Let $\mathbf{X} = X_1, \ldots, X_n$ be *i.i.d.* random variables with pdf $f(\cdot|\theta)$. Suppose that we want to test

$$H_0 : \theta \in \Theta_0 \quad vs. \quad H_1 : \theta \in \Theta_1.$$

In a Bayesian framework, since the null and the alternative hypothesis are statements about the parameters and we can make probabilistic statements about the parameters, we can come up with prior probabilities about the hypothesis: $P(H_0)$ and $P(H_1)$. We need prior densities on Θ_0 and Θ_1, denoted as $\pi_0(\theta)$ and $\pi_1(\theta)$. In the case where Θ_i is a point (point hypothesis), then $\pi_i(\theta)$ is a point mass (a distribution that assign probability one to that point). We now need to introduce the definition of marginal likelihoods under the hypotheses:

Definition 52. The marginal likelihoods of H_0 and H_1 are

$$m(\mathbf{x} \mid H_i) = \int_{\Theta_i} f(\mathbf{x} \mid \theta)\pi_i(\theta)d\theta, \;\; i = 0, 1.$$

We can now obtain the posterior probability of the hypotheses using the Bayes theorem (note that his conditional probability combines discrete probabilities and potentially continuous ones corresponding to the sample):

$$\begin{aligned} P(H_0 \mid \mathbf{x}) &= \frac{m(\mathbf{x} \mid H_0)P(H_0)}{m(\mathbf{x})} \\ &= \frac{m(\mathbf{x} \mid H_0)P(H_0)}{m(\mathbf{x} \mid H_0)P(H_0) + m(\mathbf{x} \mid H_1)P(H_1)} \\ &= 1 - P(H_1 \mid \mathbf{x}). \end{aligned}$$

Then, the posterior odds can be written as:

$$\frac{P(H_0 \mid \mathbf{x})}{P(H_1 \mid \mathbf{x})} = \frac{m(\mathbf{x} \mid H_0)P(H_0)}{m(\mathbf{x} \mid H_1)P(H_1)}$$

Consequently,

$$B_{0,1} = \frac{m(\mathbf{x} \mid H_0)}{m(\mathbf{x} \mid H_1)} = \frac{P(H_0 \mid \mathbf{x})}{P(H_1 \mid \mathbf{x})} \Big/ \frac{P(H_0)}{P(H_1)}.$$

The likelihood ratio can thus be calculated as the ratio of marginal likelihoods. Unfortunately, the marginal likelihood is not always available in closed form.

The natural question is now, how large does the Bayes factor need to be to provide evidence in favour of the null hypothesis? There is no unique answer to this question, but two popular scales were proposed by Harold Jeffreys and Kass and Raftery. These are presented in the tables below.

Bayes Factor	Evidence
$1 - \sqrt{10}$	Barely worth mentioning
$\sqrt{10} - 10$	Substantial
$10 - \sqrt{1000}$	Strong
$\sqrt{1000} - 100$	Very strong
$100-$	Decisive

Table 5.2: Jeffreys' scale

Bayes Factor	Evidence
$1 - 3$	Not worth more than a bare mention
$3 - 20$	Positive
$20 - 150$	Strong
$150-$	Very strong

Table 5.3: Kass and Raftery's scale

Point Null Hypothesis and the Laplace's approximation

Suppose that we are interested on testing

$$H_0 : \theta_k = 0 \quad vs. \quad H_1 : \theta_k \neq 0.$$

$\theta_k \in \mathbb{R}$, and $\theta_k \in \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$. This is a common problem where we are interested in comparing two models: a null model (M_0) that does not contain θ_k, and an alternative model (M_1) that contains θ_k. In particular θ_k may represent a variable in a linear or logistic regression regression model, and the idea is to test whether or not to include that variable in the model. This is known a "variable selection" or more generally as "model selection". Thus, we can use the Bayes factor to compare these two models. Suppose that we do not want to favour any of these model *a priori*. This implies that $P(M_0) = P(M_1) = 0.5$, and that

$$\begin{aligned} B_{0,1} &= \frac{m(\mathbf{x} \mid M_0)}{m(\mathbf{x} \mid M_1)} \\ &= \frac{P(M_0 \mid \mathbf{x})}{P(M_1 \mid \mathbf{x})}, \end{aligned}$$

where

$$m(\mathbf{x} \mid M_0) = \int_{\mathbb{R}^{p-1}} f(\mathbf{x} \mid \eta)\pi_0(\eta)d\eta,$$

and $\eta = \boldsymbol{\theta} \setminus \{\theta_k\}$, and

$$m(\mathbf{x} \mid M_1) = \int_{\mathbb{R}^p} f(\mathbf{x} \mid \boldsymbol{\theta})\pi_1(\boldsymbol{\theta})d\theta.$$

Thus, we have translated the problem of selecting between the two models into an integration problem. Numerical integration is quite challenging in dimensions higher than 2, and closed-form expressions are rarely available. However, there exist general methods that can be used under some conditions. One of these is the Laplace's approximation, which is explained below.

Suppose that we want to calculate the integral:

$$\begin{aligned} I &= \int_{\mathbb{R}^p} f(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\theta \\ &= \int_{\mathbb{R}^p} \exp\left\{\log f(\mathbf{x} \mid \boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta})\right\} d\theta. \end{aligned}$$

Define $g(\boldsymbol{\theta}) = -\log f(\mathbf{x} \mid \boldsymbol{\theta}) - \log \pi(\boldsymbol{\theta})$, a second order Taylor approximation to this function around its maximum $\tilde{\boldsymbol{\theta}}$ is

$$g(\boldsymbol{\theta}) \approx g(\tilde{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^\top H(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}),$$

where $H(\theta) = \nabla^2_{\boldsymbol{\theta}} g(\boldsymbol{\theta})$ (the hessian matrix). Replacing this approximation we obtain,

$$I \approx \exp\left\{-g(\tilde{\boldsymbol{\theta}})\right\} \int_{\mathbb{R}^p} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^\top H(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\right\} d\theta.$$

We can recognise the integrand and an unnormalised multivariate normal density function with mean $\tilde{\boldsymbol{\theta}}$ and variance-covariance matrix $H(\tilde{\boldsymbol{\theta}})^{-1}$. Then, we know that this integral is equal to the normalising constant of the normal density function:

$$I \approx f(\mathbf{x} \mid \tilde{\boldsymbol{\theta}})\pi(\tilde{\boldsymbol{\theta}})(2\pi)^{\frac{p}{2}} \det\left[H(\tilde{\boldsymbol{\theta}})\right]^{-\frac{1}{2}}.$$

Then, again, we have reduced the problem to maximising the posterior distribution, which can be done using numerical methods, such as the Newton's method previously studied.

Since now we have all the ingredients, we can compare the two models using this approximation to the Bayes factor. This method can be used to test whether or not a certain variable is relevant in logistic regression.

Bibliography

[1] G. Casella and R.L. Berger. ***Statistical Inference***, volume 2. Duxbury Pacific Grove, CA, 2002.

[2] D.R. Cox. *Principles of Statistical Inference*. Cambridge University Press, 2006.

[3] A.C. Davison. ***Statistical Models***, volume 11. Cambridge University Press, 2003.

[4] A.J. Dobson and A.G. Barnett. *An introduction to Generalized Linear Models*. Chapman and Hall/CRC, 2008.

[5] T.S. Ferguson. *A Course in Large Sample Theory*. Routledge, 2017.

[6] C. Forbes, M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions*. John Wiley & Sons, 2011.

[7] A. Gut. *Probability: a Graduate Course*, volume 75. Springer Science & Business Media, 2013.

[8] J.G. Kalbfleisch. ***Probability and Statistical Inference, Volume 2: Statistical Inference***. Springer Science & Business Media, 2012.

[9] D. Kunin. Seeing theory. `https://seeing-theory.brown.edu/`, 2018.

[10] E.L. Lehmann and G. Casella. ***Theory of Point Estimation***. Springer Science & Business Media, 2006.

[11] E.L. Lehmann and J.P. Romano. *Testing Statistical Hypotheses*. Springer Science & Business Media, 2006.

[12] J. Miller. Earliest known uses of some of the words of mathematics. `http://jeff560.tripod.com/mathword.html`, 2017.

[13] N. Mukhopadhyay. *Probability and Statistical Inference*. CRC Press, 2000.

[14] J.A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, 2007.

[15] C. Robert. ***The Bayesian Choice: from Decision-Theoretic Foundations to Computational Implementation***. Springer Science & Business Media, 2007.

[16] S. Ross. *A First Course in Probability 8th Edition*. Pearson, 2009.

[17] D.A. Sprott. *Statistical Inference in Science*. Springer Science & Business Media, 2008.

[18] L. Wasserman. ***All of Statistics: a Concise Course in Statistical Inference***. Springer Science & Business Media, 2013.

[19] G.A. Young and R.L. Smith. *Essentials of Statistical Inference*, volume 16. Cambridge University Press, 2005.