
UNIVERSITÀ DEGLI STUDI DI TORINO
DEPARTMENT OF MATHEMATICS

M.Sc. Stochastics and Data Science

# Statistical Inference

Prof:
Stefano Favaro

Notes by:
Nicolas Destefanis

2018/2019

# Preface

These lecture notes are based on the semester-long course on Statistical Inference, given in 2018-2019 at the University of Torino.

Please, feel free to send me corrections and feedback.

*Nicolas Destefanis*
nicolas.destefanis@edu.unito.it

# Contents

# 1

# Week I

## 1.1 26/09/18

### 1.1.1 Remarkable Inequalities

Denote by $X$ a random variable with finite second moment, and let $\sigma = \sqrt{\mathrm{Var}(X)}$. The **Chebyshev's inequality** states that, for any $\varepsilon > 0$

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\mathrm{Var}(X)}{\varepsilon^2}.$$

It is a particular case of the **Markov's inequality**: if $Y$ is a non-negative random variable with finite expected value, then

$$\mathbb{P}(Y \geq t) \leq \frac{\mathbb{E}[Y]}{t}, \quad \forall t > 0. \tag{1.1}$$

Also, if $X$ and $Y$ are two different random variables with finite second moment, then by Hölder's inequality (taking $p = q = 2$), we get

$$(\mathbb{E}[XY])^2 \leq \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]. \tag{1.2}$$

**Theorem 1.1.** *If $X$ is a random variable taking values in some set $I$, with $\mathbb{E}[X] = \mu$, and $f$ is a convex function on $I$, then we have the following inequality:*

$$f(X) \geq f(\mu) + b(X - \mu),$$

*held with probability 1 for some choice of $b$. By integrating both sides of the inequality with respect to the distribution of $X$, we obtain*

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]),$$

*known as* ***Jensen's inequality****.*

### 1.1.2 Continuous Distributions

#### The Gaussian Distribution

**Definition 1.2.** *A continuous random variable $Y$ is said to have a* ***Gaussian distribution*** *with parameters $\mu$ and $\sigma^2$ if the density function at $t$ is given by*

$$f(t;\mu,\sigma^2) \coloneqq \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\} \mathbb{1}_{\mathbb{R}}(t).$$

*If $Y \sim \mathcal{N}(0,1)$, we say that $Y$ is the* ***standardised Gaussian random variable****, with density and cumulative distribution functions denoted by*

$$\phi(t) \coloneqq \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \mathbb{1}_{\mathbb{R}}(t) \quad \textit{and} \quad \Phi(t) \coloneqq \int_{-\infty}^{t} \phi(y)dy$$

*respectively.*

*Remark 1.3.* The Gaussian distribution has some interesting properties.

1. If $Y \sim \mathcal{N}(\mu,\sigma^2)$, its *characteristic function* and *moment generating function* are

$$\mathbb{E}\left[e^{itY}\right] = \exp\left\{it\mu - \frac{1}{2}\sigma^2t^2\right\} \quad \textit{and} \quad \mathbb{E}\left[e^{tY}\right] = \exp\left\{t\mu + \frac{1}{2}\sigma^2t^2\right\},$$

   respectively. By taking the derivatives, using the characteristic function we can obtain the *moments* of $Y$. Also, if we take

$$\mathbb{E}\left[(Y-\mu)^k\right] = \begin{cases} 0 & \text{for } k \text{ odd} \\ 1\cdot 3\ldots(k-1)\sigma^k & \text{for } k \text{ even} \end{cases}.$$

2. If $Y \sim \mathcal{N}(\mu,\sigma^2)$ and $a, b$ are constants, then

$$a + bY \sim \mathcal{N}(a+b\mu, b^2\sigma^2).$$

   This means that the entire family of distributions can be generated by linear transformations, starting from any member of the family: this property is known as **location scale**.
3. If $Y_1 \sim \mathcal{N}(\mu_1,\sigma_1^2)$ and $Y_2 \sim \mathcal{N}(\mu_2,\sigma_2^2)$ are independent, then

$$Y_1 + Y_2 \sim \mathcal{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2).$$

   This result extends to any linear combination of Gaussian random variables, meaning if $Y_1,\ldots,Y_N \sim \mathcal{N}(\mu_i,\sigma_i^2)$ and $Y_i \perp\!\!\!\perp Y_j$ for every $i \neq j$, then

$$\sum_{i=1}^{n}(a_iY_i+b_i) \sim \mathcal{N}\left(\sum_{i=1}^{n}(a_iY_i+b_i), \sum_{i=1}^{n}a_i^2b_i^2\right).$$

### The Uniform Distribution

**Definition 1.4.** *A continuous random variable $Y$ is said to have a* ***uniform distribution in*** $(a, b)$ *if the density function at $t$ is given by*

$$f(t; a, b) \coloneqq \frac{1}{b-a} \mathbb{1}_{(a,b)}(t).$$

*Remark 1.5.* An important result concerning the uniform distribution is the following: if $Z$ is a continuous random variable with distribution function given by $F(\cdot)$, then the random variable $W$ defined as $W = F(Z)$ is uniformly distributed on $(0, 1)$. If we wanted to find the probability of $W$, we would compute

$$\mathbb{P}(W \leq t) = \mathbb{P}(F(Z) \leq t) = \mathbb{P}\left(Z \leq F^{-1}(t)\right) = F(F^{-1}(t)) = t,$$

which is the distribution function of $\mathcal{U}(0, 1)$.

### The Gamma Distribution

**Definition 1.6.** *First of all, we define the* ***Gamma function*** *as*

$$\Gamma(x) \coloneqq \int_0^{+\infty} t^{x-1} e^{-t} dt.$$

*Note that:*

1. *Using the integration by parts, we get $\Gamma(x+1) = x\Gamma(x)$. In particular, when $n \in \mathbb{N}$, $\Gamma(n+1) = n!$, with $\Gamma(1) = 1$ and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.*
2. *$n! \sim \sqrt{\pi} n^{\frac{n+1}{2}} e^{-n}$ as $n \to \infty$.*

**Definition 1.7.** *A continuous random variable $Y$ is said to have a* ***Gamma distribution*** *with shape parameter $w$ and scale parameter $\lambda$ if the density function at $t$ is given by*

$$f(t; w, \lambda) \coloneqq \frac{\lambda^w}{\Gamma(w)} t^{w-1} e^{-\lambda t} \mathbb{1}_{(0,+\infty)}(t).$$

**Exercise 1.8.** Derive the moment formula for the Gamma distribution. Also, try to extract a Gamma distribution from a negative exponential.

**Exercise 1.9.** If $Y_1 \sim \Gamma(w_1, \lambda)$ and $Y_2 \sim \Gamma(w_2, \lambda)$, with $Y_1 \perp\!\!\!\perp Y_2$, then

$$Y_1 + Y_2 \sim \Gamma(w_1 + w_2, \lambda).$$

### The Beta Distribution

**Definition 1.10.** *First of all, we define the* ***Beta function*** *as*

$$B(p, q) \coloneqq \int_0^1 x^{p-1}(1-x)^{q-1} dx.$$

*Having said that, a continuous random variable $Y$ is said to have a* ***Beta distribution*** *with parameter $p$ and $q$ if the density function at $t$ is given by*

$$f(t; p, q) \coloneqq \frac{t^{p-1}(1-t)^{q-1}}{B(p, q)} \mathbb{1}_{(0,1)}(t).$$

### 1.1.3 Discrete Distributions

#### The Binomial Distribution

**Definition 1.11.** *A discrete random variable $Y$ is said to have a* ***Binomial distribution*** *with parameters $n$ and $p$ if the mass function at $t$ is given by*

$$\mathbb{P}(Y = t) \coloneqq \binom{n}{t} p^t (1-p)^{n-t} \mathbb{1}_{\{0,\dots,n\}}(t)$$

*for $p \in (0,1)$. A special case of the Binomial distribution is the Bernoulli distribution, where $n = 1$.*

#### The Hypergeometric Distribution

**Definition 1.12.** *A discrete random variable $Y$ is said to have a* ***Hypergeometric distribution*** *with parameters $N$ and $M$ if the mass function at $t$ is given by*

$$\mathbb{P}(Y = t) \coloneqq \frac{\binom{M}{t}\binom{N-M}{n-t}}{\binom{N}{n}}$$

*where $n, N, M$ are positive integers such that $N \geq M$ and $N \geq n$, and $t$ is an integer such that*

$$\max\{0, n - N + M\} \leq t \leq \min\{n, M\}.$$

*Remark 1.13.* This distribution provides the probability of throwing $t$ *success* balls when $n$ balls are drawn *without replacement* from an urn containing $M$ *success* balls and $N - M$ *failure* balls.

#### The Poisson Distribution

The **Poisson distribution** is a widely applied discrete distribution, and can serve as a model for a number of different types of experiments. For example, if we are modelling a phenomenon in which we are waiting for an occurrence (such as waiting for a bus, waiting for costumers to arrive in a bank), the number of occurrences in a given time interval can sometimes be modelled by the Poisson distribution. One of the basic assumptions on which the Poisson distribution is built is that, for small time intervals, the probability of an arrival is proportional to the length of waiting time. This makes it a reasonable model for situations like those indicated above. For example, it makes sense to assume that the longer we wait, the more likely it is that a customer will enter the bank.

Another area of application is in spatial distributions, where, for example, the Poisson may be used to model the distribution of bomb hits in an area or the distribution of fish in a lake.

**Definition 1.14.** *A random variable $Y$, taking values in the non-negative integers, has a* ***Poisson distribution****, with a single parameter $\lambda$ (sometimes called the intensity parameter), if*

$$\mathbb{P}(Y = t) \coloneqq \frac{e^{-\lambda}\lambda^t}{t!}, \quad t \in \mathbb{N}.$$

### The Negative Binomial Distribution

The Binomial distribution counts the number of successes in a fixed number of Bernoulli trials. Suppose that, instead, we count the number of Bernoulli trials required to get a fixed number of successes. This latter formulations leads to the Negative Binomial distribution.

**Definition 1.15.** *In a sequence of independent Bernoulli(p) trials, let the random variable $Y$ denote the trial at which the $r$-th success occurs, where $r$ is a fixed integer. Then*

$$\mathbb{P}(Y = t) \coloneqq \binom{t-1}{r-1} p^r (1-p)^{t-r}, \quad t = r, r+1, \dots \tag{1.3}$$

*and we say that $Y$ has a **Negative Binomial Distribution** with parameters $(r, p)$.*

*Remark 1.16.* The derivation of (1.3) follows quickly from the Binomial distribution. The event $\{Y = t\}$ can occur only if there are exactly $r-1$ successes in the first $t-1$ trials, and a success on the $t$-th trial. The probability of $r-1$ successes in $t-1$ trials is the binomial probability

$$\binom{t-1}{r-1} p^{r-1} (1-p)^{t-r}$$

and with probability $p$ there is a success on the $t$-th trial. Multiplying these probabilities, gives us (1.3).

The Negative Binomial distributions is sometimes defined in terms of the "number of failures before the $r$-th success". This formulation is statistically equivalent to the one given above in terms of "trial at which the $r$-th success occurs". Therefore, an alternative form of the negative binomial distribution is

$$\mathbb{P}(Y = t) \coloneqq \binom{r+t-1}{t} p^r (1-p)^t, \quad t \in \mathbb{N}.$$

Unless otherwise noted, when we refer to the negative binomial distribution with parameters $(r, p)$, we will use this probability mass function.

In the case that $r$ is an integer, the Negative Binomial is called a **Pascal distribution**, and it has the following interpretation: consider an infinite sequence of trials whose outcome is *success* with probability $p$ and *failure* with probability $1-p$, and assume that the trials are independent. Then, the above probability mass gives us the probability of expecting $t$ failures before achieving $r$ successes. If $r = 1$, we obtain the **Geometric distribution**.

### The Geometric Distribution

**Definition 1.17.** *The **Geometric distribution** is the simplest of the waiting time distributions, and is a special case of the negative binomial distribution. If we set $r = 1$ in (1.3), we obtain*

$$\mathbb{P}(Y = t) \coloneqq p(1-p)^{t-1}, \quad t \in \mathbb{N}_+,$$

*which defines the probability mass function of a Geometric random variable with success probability $p$. It can be interpreted as the trial at which the first success occurs, so we are "waiting for a success".*

## 1.2 27/09/18

### 1.2.1 Matrix Algebra

- Denote by $I_k$ the *identity matrix* of order $k$, by $\mathbf{1}_n$ the $n \times 1$ vector with all elements equal to 1 and by $\mathbf{0}_n$ the $n \times 1$ vector with all elements equal to 0.
- For two conformable matrices $A$ and $B$, we have

$$\det(AB) = \det(A)\det(B).$$

- If $\det(A) \neq 0$, then $A$ is non-singular, and there exist an *inverse matrix* $A^{-1}$ such that $AA^{-1} = A^{-1}A = I_n$. Moreover,

$$({}^{\mathrm{t}}A)^{-1} = {}^{\mathrm{t}}(A^{-1}) \quad \text{and} \quad (AB)^{-1} = B^{-1}A^{-1}.$$

- A matrix $A$ is *positive definite* if $A$ is symmetric and ${}^{\mathrm{t}}\boldsymbol{u}A\boldsymbol{u} > 0$ for all non-null vectors $\boldsymbol{u} \in \mathbb{R}^k$.
- For any two conformable matrices we have $\mathrm{tr}(AB) = \mathrm{tr}(A)\,\mathrm{tr}(B)$.
- If $A$ is idempotent, then $\mathrm{rk}(A) = \mathrm{tr}(A)$.
- Taking into account that the matrices $A, B, C, D$ are conformable, we have the following identities:

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

and

$$(A + b\,{}^{\mathrm{t}}d)^{-1} = A^{-1} - \frac{A^{-1}b\,{}^{\mathrm{t}}dA^{-1}}{1 + {}^{\mathrm{t}}dA^{-1}b}.$$

- **Spectral Decomposition**: if $A$ is a symmetric $k \times k$ matrix, there exist $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$ and and orthogonal matrix $Q$ such that

$$A = Q\Lambda\,{}^{\mathrm{t}}Q,$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$, $\lambda_1, \ldots, \lambda_k$ are the *eigenvalues* of $A$ and the column of order $j$ of $Q$ is called the *eigenvector corresponding* to $\lambda_j$. Also

$$\det(A) = \det(\Lambda) = \prod_{i=1}^{k} \lambda_i.$$

- Given a positive semidefinite matrix $A$, we define the *square root of a matrix* $A$ as the matrix $B$ such that $A = B\,{}^{\mathrm{t}}B$. In this case $A = \mathrm{diag}(a_1, \ldots, a_k)$, therefore it is natural to set $B = \mathrm{diag}(\sqrt{a_1}, \ldots, \sqrt{a_k})$. In the general case of a positive semidefinite matrix $A$, we put $B = Q\Lambda^{\frac{1}{2}} = Q\,\mathrm{diag}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_k}\right)$ such that

$$B\,{}^{\mathrm{t}}B = \left(Q\Lambda^{\frac{1}{2}}\right)\left(\Lambda^{\frac{1}{2}}\,{}^{\mathrm{t}}Q\right) = Q\Lambda^{\frac{1}{2}}\,{}^{\mathrm{t}}Q = A;$$

we then say that $B$ is the square root of $A$. Note that if $A > 0$

$$B^{-1} = (Q\Lambda^{\frac{1}{2}})^{-1} = \Lambda^{-\frac{1}{2}}\,{}^{\mathrm{t}}Q.$$

### 1.2.2 Random Vectors

**Definition 1.18.** *If $X_1, \dots, X_k$ are random variables defined on the same probability space, we define a* ***random vector*** *(or* ***multivariate random variable****) as*

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}.$$

*The* ***mean value*** *of $\boldsymbol{X}$ is obtained by forming the vector of mean values with components*

$$\mathbb{E}[\boldsymbol{X}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_k] \end{pmatrix}.$$

*Similarly, we define the* ***variance matrix*** *(or* ***dispersion matrix****) as*

$$\operatorname{Var}(\boldsymbol{X}) = \begin{pmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1, X_2) \dots & \operatorname{Cov}(X_1, X_k) \\ \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_k, X_1) & \operatorname{Cov}(X_k, X_2) \dots & \operatorname{Var}(X_k) \end{pmatrix}.$$

*We also define the* ***correlation matrix*** *whose generic element is*

$$\operatorname{Cor}(X_i, X_j) \coloneqq \frac{\operatorname{Cov}(X_i, X_j)}{\sqrt{\operatorname{Var}(X_i)\operatorname{Var}(X_j)}}.$$

**Lemma 1.19.** *If $A = (a_{ij})$ is a $n \times k$ matrix and $\boldsymbol{b} = {}^{\mathrm{t}}(b_1, \dots, b_n)$ an $n \times 1$ vector, define*

$$\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$$

*where $\boldsymbol{X} = {}^{\mathrm{t}}(X_1, \dots, X_k)$, $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}$ and $\operatorname{Var}(\boldsymbol{X}) = \mathbb{V}$. Then,*

$$\mathbb{E}[\boldsymbol{Y}] = A\boldsymbol{\mu} + \boldsymbol{b} \quad \text{and} \quad \operatorname{Var}(\boldsymbol{Y}) = A\mathbb{V}\,{}^{\mathrm{t}}A.$$

**Lemma 1.20.** *The variance matrix $\mathbb{V}$ of the random vector $\boldsymbol{X}$ is positive semidefinite, and is positive definite if there exist no vector such that ${}^{\mathrm{t}}\boldsymbol{b}\boldsymbol{X}$ is a degenerate random variable (meaning, a random variable which takes a single value with probability 1).*

**Lemma 1.21.** *If $\mathbb{V}(\boldsymbol{X}) > 0$, there exist a square matrix $C$ such that $\boldsymbol{Y} \coloneqq C\boldsymbol{X}$ has uncorrelated components with unit variance.*

**Lemma 1.22.** *Let $A = (a_{ij})$ be a square matrix of order $k$, and let $\boldsymbol{X}$ be a random vector such that $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}]$ and $\mathbb{V} = \operatorname{Var}(\boldsymbol{X})$. Then,*

$$\mathbb{E}[{}^{\mathrm{t}}\boldsymbol{X}A\boldsymbol{X}] = {}^{\mathrm{t}}\boldsymbol{\mu}A\boldsymbol{\mu} + \operatorname{tr}(A\mathbb{V}).$$

*Proof.* Now,

$$\mathbb{E}[{}^{\mathrm{t}}\boldsymbol{X}A\boldsymbol{X}] = \mathbb{E}\left[\sum_{i=1}^{k}\sum_{j=1}^{k} X_i a_{ij} X_j\right]$$

$$= \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}\mathbb{E}\left[X_i X_j\right]$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}(\mu_i\mu_j + \mathbb{V}_{ij})$$

where $\mu_i$ and $\mathbb{V}_{ij}$ are the generic elements of $\boldsymbol{\mu}$ and $\mathbb{V}$ respectively. Therefore,

$$\begin{aligned}\mathbb{E}\left[{}^{\mathrm{t}}\boldsymbol{X}A\boldsymbol{X}\right] &= \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}\mu_i\mu_j + \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}\mathbb{V}_{ij}\\ &= {}^{\mathrm{t}}\boldsymbol{\mu}A\boldsymbol{\mu} + \sum_{i=1}^{k}(A\mathbb{V})_{ii}\\ &= {}^{\mathrm{t}}\boldsymbol{\mu}A\boldsymbol{\mu} + \mathrm{tr}\,(A\mathbb{V}),\end{aligned}$$

where $(A\mathbb{V})_{ii}$ is the generic element of the main diagonal of $A\mathbb{V}$.

□

### 1.2.3 Multivariate Gaussian Distribution

Consider a vector $\boldsymbol{Z} = {}^{\mathrm{t}}(Z_1, \ldots, Z_k)$ where $Z_1, \ldots, Z_k$ are independent and identically distributed standard Gaussian random variables. Now, set

$$\boldsymbol{Y} = A\boldsymbol{Z} + \boldsymbol{\mu},$$

where $A$ is a non-singular $k \times k$ matrix, and $\boldsymbol{\mu}$ is a $k \times 1$ vector. It is natural to think about $\boldsymbol{Y}$ as a $k$-dimensional generalisation of the Gaussian distribution. We wish to find its distribution.

We start from the standard case

$$f_{\boldsymbol{Z}}(\boldsymbol{t}) = \frac{1}{(2\pi)^{\frac{k}{2}}}\exp\left\{-\frac{1}{2}\,{}^{\mathrm{t}}\boldsymbol{t}\boldsymbol{t}\right\}.$$

Then since $\boldsymbol{Z} = A^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$, we find the *Jacobian matrix* of the transformation, which is (considering all the partial derivatives)

$$\det\left|\frac{dz_i}{dy_j}\right| = \det\left(A^{-1}\right) = \left(\det\left(\mathbb{V}\right)\right)^{-\frac{1}{2}}.$$

Taking into account that

$$|\mathbb{V}| = |A\,{}^{\mathrm{t}}A| = |A|^2$$

and setting $\boldsymbol{y} = A\boldsymbol{t} + \boldsymbol{\mu}$ we obtain that

$${}^{\mathrm{t}}\boldsymbol{t}\boldsymbol{t} = {}^{\mathrm{t}}[A^{-1}(\boldsymbol{y} - \boldsymbol{\mu})][A^{-1}(\boldsymbol{y} - \boldsymbol{\mu})] = {}^{\mathrm{t}}(\boldsymbol{y} - \boldsymbol{\mu})\mathbb{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}).$$

Therefore, the density of $\boldsymbol{Y}$ is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) \coloneqq \frac{1}{(2\pi)^{\frac{k}{2}}(\det\mathbb{V})^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\,{}^{\mathrm{t}}(\boldsymbol{y} - \boldsymbol{\mu})\mathbb{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right\}. \tag{1.4}$$

**Definition 1.23.** *We say that the random variable* $\boldsymbol{Y} = {}^{\mathrm{t}}(Y_1, \dots, Y_k)$ *with density function* $f_{\boldsymbol{Y}}(\cdot)$ *is a* ***multivariate Gaussian random variable*** *with mean* $\boldsymbol{\mu}$ *and variance* $\mathbb{V}$*, or*

$$\boldsymbol{Y} \sim \mathcal{N}_k(\boldsymbol{\mu}, \mathbb{V}).$$

**Exercise 1.24.** Take $X_1, X_2$ independent random variables such that

$$X_1, X_2 \sim \mathcal{U}([0,1]).$$

Find the distribution of $Y = X_1 + X_2$, $Z = \frac{X_2}{X_1}$ and $W = X_1 X_2$.

**Exercise 1.25.** Take three independent negative exponential random variables, with parameter $\frac{1}{2}$. Find the distribution of $U = \frac{Y}{X}$ and $W = X + Y + Z$.

**Exercise 1.26.** Let $(X, Y)$ be a bivariate random variable such that the marginal distribution of $X \sim \mathcal{U}([-1,1])$ and the conditional distribution

$$Y \mid \{X = x\} \sim \mathcal{U}((x, x+1)).$$

Find the distribution of $Z = -\log(Y - X)$.

**Exercise 1.27.** Let $A$ and $B$ be two independent random variables such that

$$A, B \sim \mathcal{U}([0,h]).$$

Compute the probability that the equation $Z^2 - 2AZ + B = 0$ admits complex roots.

**Exercise 1.28.** Let $X \sim \Gamma(r,1)$ and $Y \sim \Gamma(s,1)$ be two independent random variables. Find the distribution of $\left(\frac{X}{X+Y}, X+Y\right)$, $Z = \frac{X}{X+Y}$ and $W = X + Y$.

# 2

# Week II

## 2.1 02/10/18

### Moment Generating Function of a Multivariate Gaussian

Last week, we derived the multivariate Gaussian distribution (1.4) starting from a vector of independent and identically distributed standard Gaussian random variables. We wish to compute the moment generating function of $\boldsymbol{Y}$.

**Lemma 2.1.** *Consider a positive definite $n \times n$ matrix $A$, a $n \times 1$ vector $\boldsymbol{b}$ and*

$$I := \int_{\mathbb{R}^n} \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}({}^t\boldsymbol{y}A\boldsymbol{y} - 2\,{}^t\boldsymbol{b}\boldsymbol{y})\right\} d\boldsymbol{y}. \tag{2.1}$$

*Then,*

$$I = \frac{\exp\left\{\frac{1}{2}\,{}^t\boldsymbol{b}A^{-1}\boldsymbol{b}\right\}}{\det(A)^{\frac{1}{2}}}. \tag{2.2}$$

*Proof.* Let $\boldsymbol{\mu} = A^{-1}\boldsymbol{b}$ and expand, within the integral, $\exp\{\cdot\}$ by adding and subtracting

$$\frac{1}{2}\,{}^t\boldsymbol{\mu}A\boldsymbol{\mu}:$$

$$\begin{aligned}
I &= \int_{\mathbb{R}^n} \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\,({}^t\boldsymbol{y}A\boldsymbol{y} - 2\,{}^t\boldsymbol{b}\boldsymbol{y})\right\} d\boldsymbol{y} \\
&= \int_{\mathbb{R}^n} \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\,{}^t\boldsymbol{y}A\boldsymbol{y} + {}^t\boldsymbol{\mu}A\boldsymbol{y} + \frac{1}{2}\,{}^t\boldsymbol{\mu}A\boldsymbol{\mu} - \frac{1}{2}\,{}^t\boldsymbol{\mu}A\boldsymbol{\mu}\right\} d\boldsymbol{y} \\
&= \exp\left\{\frac{1}{2}\,{}^t\boldsymbol{\mu}A^{-1}\boldsymbol{\mu}\right\} \int_{\mathbb{R}^n} \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2}\,({}^t\boldsymbol{y}A\boldsymbol{y} - 2\,{}^t\boldsymbol{\mu}A\boldsymbol{y} + {}^t\boldsymbol{\mu}A\boldsymbol{\mu})\right\} d\boldsymbol{y} \\
&= \det(A^{-1})^{\frac{1}{2}} \exp\left\{\frac{1}{2}\,{}^t\boldsymbol{\mu}A^{-1}\boldsymbol{\mu}\right\} \int_{\mathbb{R}^n} g(\boldsymbol{y})d\boldsymbol{y} \\
&= \det(A^{-1})^{\frac{1}{2}} \exp\left\{\frac{1}{2}\,{}^t\boldsymbol{b}A^{-1}AA^{-1}\boldsymbol{b}\right\} \\
&= \frac{\exp\left\{\frac{1}{2}\,{}^t\boldsymbol{b}A^{-1}\boldsymbol{b}\right\}}{\det(A)^{\frac{1}{2}}},
\end{aligned}$$

where $g(\cdot)$ is the density function of the multivariate gaussian (1.4) with $\mathbb{V}$ replaced by $A^{-1}$.

□

Now the computation of the moment generating function of $\boldsymbol{Y}$ is immediate:

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{{}^{\mathrm{t}}\boldsymbol{t}\boldsymbol{y}\right\}\right] &= \int_{\mathbb{R}^n} \exp\left\{{}^{\mathrm{t}}\boldsymbol{t}\boldsymbol{y}\right\} f_{\boldsymbol{Y}}(\boldsymbol{y})d\boldsymbol{y} \\
&= \exp\left\{-\frac{1}{2}\,{}^{\mathrm{t}}\boldsymbol{\mu}\mathbb{V}^{-1}\boldsymbol{\mu}\right\} \int_{\mathbb{R}^n} \frac{\exp\left\{-\frac{1}{2}\left({}^{\mathrm{t}}\boldsymbol{y}\mathbb{V}^{-1}\boldsymbol{y} - 2\,{}^{\mathrm{t}}(\mathbb{V}^{-1}\boldsymbol{\mu}+\boldsymbol{t})\boldsymbol{y}\right)\right\}}{(2\pi)^{\frac{n}{2}}\det(\mathbb{V})^{\frac{1}{2}}} d\boldsymbol{y} \\
&= \exp\left\{-\frac{1}{2}\,{}^{\mathrm{t}}\boldsymbol{\mu}\mathbb{V}^{-1}\boldsymbol{\mu}\right\} \exp\left\{\frac{1}{2}\,{}^{\mathrm{t}}(\mathbb{V}^{-1}\boldsymbol{\mu}+\boldsymbol{t})\mathbb{V}(\mathbb{V}^{-1}\boldsymbol{\mu}+\boldsymbol{t})\right\} \\
&= \exp\left\{-\frac{1}{2}\,{}^{\mathrm{t}}\boldsymbol{\mu}\mathbb{V}^{-1}\boldsymbol{\mu} + \frac{1}{2}({}^{\mathrm{t}}\boldsymbol{t}\boldsymbol{\mu} + {}^{\mathrm{t}}\boldsymbol{t}\mathbb{V}\boldsymbol{t} + {}^{\mathrm{t}}\boldsymbol{\mu}\mathbb{V}^{-1}\boldsymbol{\mu} + {}^{\mathrm{t}}\boldsymbol{\mu}\boldsymbol{t})\right\} \\
&= \exp\left\{\frac{1}{2}\,{}^{\mathrm{t}}\boldsymbol{t}\mathbb{V}\boldsymbol{t} + {}^{\mathrm{t}}\boldsymbol{t}\boldsymbol{\mu}\right\},
\end{aligned}$$

where we applied Lemma 2.1 with $A = \mathbb{V}^{-1}$ and $\boldsymbol{b} = \mathbb{V}^{-1}\boldsymbol{\mu} + \boldsymbol{t}$. Therefore, the **characteristic function** of $\boldsymbol{Y}$ is given by

$$\mathbb{E}\left[\exp\left\{i\,{}^{\mathrm{t}}\boldsymbol{t}\boldsymbol{y}\right\}\right] = \exp\left\{i\,{}^{\mathrm{t}}\boldsymbol{t}\boldsymbol{\mu} - \frac{1}{2}\,{}^{\mathrm{t}}\boldsymbol{t}\mathbb{V}\boldsymbol{t}\right\}.$$

*Remark 2.2.* If we were interested in the *marginal distributions*, we first have to partition our vector $Y$ as

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

where $Y_1 \in \mathbb{R}^r$ and $Y_2 \in \mathbb{R}^{n-r}$ (for $1 < r < n$). We also need to partition the rest of the elements, meaning

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad B = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \quad \text{and} \quad \mathbb{V} = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix},$$

with $V_{21} = V_{12}$. How can we extract the marginal distribution of $Y_1$, for example? Simply using the characteristic function formula imposing $t_2 = 0$: we therefore obtain

$$\exp\left\{i\,{}^{\mathrm{t}}t_1\mu_1 - \frac{1}{2}\,{}^{\mathrm{t}}t_1 V_{11} t_1,\right\}$$

which is the characteristic function of a multivariate Gaussian random variable with mean $\mu_1$ and variance $V_{11}$. Note that all marginal distributions of multivariate Gaussian are themselves multivariate Gaussians.

### The Chi-squared Distribution

If $\boldsymbol{Z} = {}^{\mathrm{t}}(Z_1, \dots, Z_k) \sim \mathcal{N}_k(\mathbf{0}_k, I_k)$, define $U_k \coloneqq {}^{\mathrm{t}}\boldsymbol{Z}\boldsymbol{Z}$: this is a **central Chi-squared random variable** with parameter $k$ (known as *degrees of freedom*), and we denote it by

$$U_k \sim \chi^2_k.$$

This random variable is closely related to the Gamma random variable, with a certain characterisation. To compute the density function of $U_k$, we can start from $k = 1$:

$$\begin{aligned}
\mathbb{P}(U_1 \leq t) &= \mathbb{P}\left(Z_1^2 \leq t\right) \\
&= \mathbb{P}\left(-\sqrt{t} \leq Z_1 \leq \sqrt{t}\right) \\
&= 2\mathbb{P}\left(0 \leq Z_1 \leq \sqrt{t}\right) \\
&= 2\left(\Phi(\sqrt{t}) - \frac{1}{2}\right),
\end{aligned}$$

therefore

$$f_1(t) = \frac{t^{-\frac{1}{2}} e^{-\frac{t}{2}}}{\sqrt{2\pi}} \mathbb{1}_{(0,+\infty)}(t),$$

which is the density function of a Gamma random variable with parameters $(\frac{1}{2}, \frac{1}{2})$. Then, by means of the additive property of the Gamma distribution, we have that $U_k \sim \Gamma(\frac{k}{2}, \frac{1}{2})$. Basically,

$$f_{U_k}(u) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} u^{\frac{k}{2}-1} e^{-\frac{u}{2}} \mathbb{1}_{(0,\infty)}(u).$$

If $\boldsymbol{Z} = {}^{\mathrm{t}}(Z_1, \dots, Z_k) \sim N_k(\boldsymbol{\mu}, I_k)$, we say that $U_k = {}^{\mathrm{t}}\boldsymbol{Z}\boldsymbol{Z}$ is a **non-central Chi-squared** random variable, with $k$ degrees of freedom and *non-centrality parameter* $\delta = {}^{\mathrm{t}}\boldsymbol{\mu}\boldsymbol{\mu}$. We can use the following shorthand notation: $U_k \sim \chi^2_k(\delta)$.

### The $t$ Distribution

If $Z \sim \mathcal{N}(0,1)$ and $U \sim \chi^2_k$ are independent, we say that

$$T \coloneqq \frac{Z}{\sqrt{U/k}}$$

is a *Student* $t$ random variable, with $k$ degrees of freedom.

### The $F$ Distribution

If $V \sim \chi^2_m$ and $U \sim \chi^2_k$ are independent, we say that

$$F \coloneqq \frac{V/m}{U/k}$$

is a *Fisher* $F$ random variable, with $(m, k)$ degrees of freedom.

### Multinomial Distribution

The multinomial distribution is a model for the following kind of experiment. It consists of $n$ independent trials. Each trial results in one of $k$ distinct possible outcomes. The probability of the $i$-th outcome is $p_i$ on every trial, and $X_i$ is the count of the number of times the $i$-th outcome occurred in the $n$ trials. For $k = 2$, this is just a binomial experiment, in which each trial has $k = 2$ possible outcomes and $X_1$ counts the number of "successes" and $X_2 = n - X_1$ counts the number of "failures" in $n$ trials. In a general multinomial experiment, there are $k$ possible outcome to count.

**Definition 2.3.** *The outcome of a trial may be one of $k$ mutually exclusive events $E_1, \ldots, E_k$, with associated probabilities given by*

$$p_j = \mathbb{P}(E_j), \quad j = 1, \ldots, k \qquad \sum_{j=1}^{k} p_j = 1.$$

*If $n$ independent replicates of the trial take place, denote by $Y_j$ the number of times that $E_j$ occurs. We say that the random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_k)$ is a **multinomial** random variable with parameters $(n, (p_1, \ldots, p_k))$. The probability mass is given by*

$$\mathbb{P}(\boldsymbol{Y} = \boldsymbol{t}) \coloneqq \mathbb{P}((Y_1 = t_1, \ldots, Y_k = t_k)) = \frac{n!}{t_1! \ldots t_k!} \prod_{i=1}^{n} p_i^{t_i},$$

*where the factor*

$$\frac{n!}{t_1! \ldots t_k!}$$

*is called **multinomial coefficient**. It is the number of ways the $n$ objects can be divided into $k$ groups, with $t_i$ in the $i$-th group.*

In order to find the marginal distributions, we could apply the following theorem.

**Theorem 2.4.** *Let $n, k \in \mathbb{N}_+$ and consider a set $A$ of vectors $\boldsymbol{t} = (t_1, \ldots, t_k)$ such that each $t_i$ is a non-negative integer, and $\sum_{i=1}^{k} x_i = n$. Then, for any real numbers $p_1, \ldots, p_k$,*

$$(p_1 + \ldots + p_k)^n = \sum_{\boldsymbol{t} \in A} \frac{n!}{t_1! \ldots t_k!} p_1^{t_1} \ldots p_k^{t_k}.$$

## 2.2 03/10/18

### 2.2.1 Properties of a Random Sample

Often, the data collected in an experiment consists of several observations on a variable of interest. Here, we present a model for data collection that is often used to describe this situation, a model referred to as *random sample*.

**Definition 2.5.** *Let $X \sim f_X(x;\vartheta)$. Consider $n$ independent and identically distributed (as $X$) random variables $X_1, \dots, X_n$. A **random sample** from $X$ is a collection of those random variables and it is denoted by*

$$\boldsymbol{X} = (X_1, \dots, X_n),$$

*with*

$$X_i \overset{d}{=} X \quad \text{and} \quad X_i \perp\!\!\!\perp X_j, \quad \forall i \neq j.$$

*We denote by $\boldsymbol{x} = (x_1, \dots, x_n)$ the **realisation** of the random sample.*

We know that the joint probability density function, or probability mass function, of $\boldsymbol{X}$ is given by

$$f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) = \prod_{i=1}^{n} f_{X_i}(x_i;\vartheta) = \prod_{i=1}^{n} f_X(x_i;\vartheta).$$

Since $X_1, \dots, X_n$ are identically distributed, all the marginal densities are the same function.

When a sample $X_1, \dots, X_n$ is drawn, some summary of the values is usually computed. Any well-defined summary may be expressed mathematically as a function $T : \mathbb{R}^n \to \mathbb{R}^m$ (with the assumption that $m < n$, usually $m = 1, 2$) whose domain includes the sample space of the random vector $\boldsymbol{X}$. The function $T$ may be real-valued or vector-valued; thus the summary is a random variable (or vector) $\boldsymbol{Y} = T(\boldsymbol{X})$. Since the random sample $\boldsymbol{X}$ has a simple probabilistic structure (because the $X_i$'s are independent and identically distributed), the distribution of $\boldsymbol{Y}$ is particularly tractable. Because this distribution is usually derived from the distribution of the variables in the random sample, it is called the *sampling distribution* of $\boldsymbol{Y}$. This distinguishes the probability distribution of $\boldsymbol{Y}$ from the distribution of the population, that is, the marginal distribution of each $X_i$.

**Definition 2.6.** *Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be a random sample of size $n$ from a population and let $T(\boldsymbol{X})$ be a real-valued (or vector-valued) function, whose domain includes the sample space of $\boldsymbol{X}$. Then, the random variable (or random vector) $\boldsymbol{Y} = T(\boldsymbol{X})$ is called a **statistic**. The probability distribution of a statistic $\boldsymbol{Y}$ is called the **sampling distribution** of $\boldsymbol{Y}$.*

Here are some examples of statistics.

**Definition 2.7.** *The **sample mean** is the arithmetic average of the values in a random sample, usually denoted by*

$$\overline{X}_n \coloneqq \frac{1}{n} \sum_{i=1}^{n} X_i.$$

**Definition 2.8.** *The* ***uncorrected sample variance*** *is the statistic defined by*

$$S_n^2 \coloneqq \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2,$$

*while the* ***corrected sample variance*** *is*

$$\hat{S}_n^2 \coloneqq \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2,$$

As is commonly done, we have suppressed the functional notation in the above definitions of these statistic. That is, we have written $S$ rather than $S(\boldsymbol{X})$. As before, we will denote observed values of statistics with lowercase letters: $\overline{x}$ and $s_n^2$ will denote observed values of $\overline{X}_n$ and $S_n$, respectively.

**Definition 2.9.** *The* ***sample moments of order*** $r$ *is the statistic defined by*

$$M_{r,n} \coloneqq \frac{1}{n}\sum_{i=1}^{n} X_i^r,$$

*while the* ***sample centred moments of order*** $r$ *is*

$$\overline{M}_{r,n} \coloneqq \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^r.$$

Sample values such as the *smallest*, *largest* or *middle* observation from a random sample can provide additional summary information. These are examples of *order statistics*.

**Definition 2.10.** *The* ***order statistics*** *of a random sample* $\boldsymbol{X}$ *are the sample values placed in ascending order, and they are denoted by* $X_{(1)}, \ldots, X_{(n)}$.

*Remark 2.11.* The order statistics are random variables that satisfy $X_{(1)} \leq \ldots \leq X_{(n)}$, in particular

$$\begin{aligned} X_{(1)} &= \min_{1\leq i\leq n} X_i \\ X_{(2)} &= \text{“second smallest } X_i\text{”} \\ &\vdots \\ X_{(n)} &= \max_{1\leq i\leq n} X_i \end{aligned}$$

**Definition 2.12.** *The* ***sample median*** $M$ *is a number such that approximately one-half of the observations are less than* $M$ *and one-half are greater. In terms of order statistics,* $M$ *is defined by*

$$M = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}\right) & \text{if } n \text{ is even} \end{cases}.$$

*Remark 2.13.* The median is a measure of location that might be considered an alternative to the sample mean. One advantage of the sample median over the sample mean is that it is less affected by extreme observations. Envision a particular set of sample values and then consider the effect of increasing the largest observation. The sample median is unchanged by this change of the sample values. The sample mean, on the other hand, increases without bound as the largest observation increases. This insensitivity to extreme observations is sometimes considerer an asset of the sample median.

**Theorem 2.14.** *Let* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *be a random sample from* $\mathcal{N}(\mu, \sigma^2)$*. Then,*

1. $\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$;
2. $\hat{S}_n^2 \sim \frac{\sigma^2}{n-1} R$, *where* $R \sim \chi^2_{(n-1)}$.

*Proof.* (1) Using the Moment Generating function,

$$\begin{aligned} M_{\overline{X}_n}(t) &= \mathbb{E}\left[e^{t\overline{X}_n}\right] \\ &= \mathbb{E}\left[\exp\left\{t\frac{X_1 + \ldots + X_n}{n}\right\}\right] \\ &= \mathbb{E}\left[\exp\left\{\frac{t}{n}(X_1 + \ldots + X_n)\right\}\right] \\ &= \mathbb{E}\left[\exp\left\{\frac{t}{n}X_1\right\} \ldots \exp\left\{\frac{t}{n}X_n\right\}\right] \\ &= \left(M_X\left(\frac{t}{n}\right)\right)^n, \end{aligned}$$

because, since $X_1, \ldots, X_n$ are identically distributed, $M_{X_i}(t)$ is the same function for each $i$. Now, the moment generating function for a Gaussian distribution is given by

$$M_X(t) = \exp\left\{\mu t + \frac{1}{2}\sigma^2 t^2\right\},$$

therefore our case simply becomes

$$\begin{aligned} M_{\overline{X}_n}(t) &= \left(M_X\left(\frac{t}{n}\right)\right)^n \\ &= \left(\exp\left\{\mu\frac{t}{n} + \frac{1}{2}\sigma^2\left(\frac{t}{n}\right)^2\right\}\right)^n \\ &= \exp\left\{n\left(\mu\frac{t}{n} + \frac{1}{2}\sigma^2\left(\frac{t}{n}\right)^2\right)\right\} \\ &= \exp\left\{\mu t + \frac{1}{2}\left(\frac{\sigma^2}{n}\right)t^2\right\}. \end{aligned}$$

We conclude that, since the moment generating function uniquely determines the distribution,

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

(2) Consider the following statistic

$$W \coloneqq \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2$$

and re-arrange the components,

$$\begin{aligned} W &= \sum_{i=1}^{n} \left(\frac{(X_i - \overline{X}_n) + (\overline{X}_n - \mu)}{\sigma}\right)^2 \\ &= \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}_n}{\sigma}\right)^2 + \sum_{i=1}^{n} \left(\frac{\overline{X}_n - \mu}{\sigma}\right)^2 + 2\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \sum_{i=1}^{n} (X_i - \overline{X}_n) \\ &= \sum_{i=1}^{n} \frac{(X_i - \overline{X}_n)^2}{\sigma^2} + \frac{n(\overline{X}_n - \mu)^2}{\sigma^2}, \end{aligned} \tag{2.3}$$

since $\sum_{i=1}^{n}(X_i - \overline{X}_n) = 0$. Now, consider the definition of the sample variance and multiply both sides by $n-1$:

$$(n-1)\hat{S}_n^2 = \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

and replace it in (2.3). So far, we have that

$$W = \underbrace{\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2}_{\| \; A} = \underbrace{\frac{(n-1)\hat{S}_n^2}{\sigma^2}}_{\| \; B} + \underbrace{\frac{n(\overline{X}_n - \mu)^2}{\sigma^2}}_{\| \; C},$$

where

- $A \sim \chi^2_{(n)}$. Indeed, observe that it is a sum of $n$ independent $\chi^2_{(1)}$ random variables: that is because we have assumed that $X_1, \ldots, X_n$ are observations of a random sample of size $n$ from $\mathcal{N}(\mu, \sigma^2)$, therefore

  $$\frac{X_i - \mu}{\sigma}$$

  follows a standard normal distribution with one degree of freedom.
- $C \sim \chi^2_{(1)}$. As we showed in (1), the sample mean is normally distributed with mean $\mu$ and variance $\frac{\sigma^2}{n}$, therefore

  $$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

  So, if we square $Z$, we get a $\chi^2_{(1)}$.

In order to conclude the proof, we use a very important result: given three random variables $Q \sim \chi^2_{(g)}$, $Q_1 \sim \chi^2_{(g_1)}$ and $Q_2$ such that we can write

$$Q = Q_1 + Q_2,$$

then

$$Q_2 \sim \chi^2_{(g_2)}, \qquad \text{with } g_2 = g - g_1.$$

This result implies that $B \sim \chi^2_{(n-1)}$, meaning

$$\frac{(n-1)\hat{S}_n^2}{\sigma^2} \sim \chi^2_{(n-1)} \iff \hat{S}_n^2 \sim \frac{\sigma^2}{n-1} R,$$

where $R \sim \chi^2_{(n-1)}$.

□

## 2.3 04/10/18

*Remark 2.15.* Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from $X \sim \mathcal{N}(\mu, \sigma^2)$, we know that the statistic

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \tag{2.4}$$

is distributed as a $\mathcal{N}(0,1)$ random variable. If we knew the value of $\sigma$ and we measured $\overline{X}_n$, then we could use (2.4) as a basis for inference about $\mu$, since it would be the only unknown quantity. Most of the time, however, $\sigma$ is unknown. In this case, we could consider the following statistic,

$$T_n := \frac{\overline{X}_n - \mu}{\hat{S}_n/\sqrt{n}} = \frac{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\hat{S}_n^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{R}{n-1}}}, \tag{2.5}$$

where $R \sim \chi^2_{(n-1)}$. Now, if we do not assume independence between $Z$ and $R$, this is how far we get. However, it can be proved that, given a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

are independent. In this case, (2.5) $\sim t_{(n-1)}$. Notice that also the vice versa holds.

### 2.3.1 Concentrations of Measures

Consider a non-negative random variable $X$ and let $t > 0$. Then, Markov's inequality (1.1) tells us that

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Now, consider a strictly monotonically increasing function $\phi$, with non-negative values: then, we can write

$$\mathbb{P}(X \geq t) = \mathbb{P}(\phi(X) \geq \phi(t)) \leq \frac{\mathbb{E}[\phi(X)]}{\phi(t)}.$$

For instance, if $\phi(x) = x^2$, we obtain Chebyshev's inequality. More in general, we can take $\phi(x) = x^q$ for $x \geq 0$ and $q > 0$, so that we have

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^q]}{t^q}.$$

Notice that this is a functional in $q$; also, it is possible to choose $q$ in such a way to optimise the upper bound.

*Remark 2.16.* A related idea is at the basis of *Chernoff's bounding method*. Consider a function $\phi(x) = e^{sx}$, where $s > 0$. For every random variable $X$ and any $t \in \mathbb{R}$, we have

$$\mathbb{P}(X \geq t) = \mathbb{P}\left(e^{sX} \geq e^{st}\right) \leq \frac{\mathbb{E}\left[e^{sX}\right]}{e^{st}}. \tag{2.6}$$

Our goal is to find the value of $s > 0$ for which that upper bound is minimised. Observe that it can be proven that the upper bound given by $q$ is better than the one obtained minimising $s$ in here.

**Theorem 2.17.** *Let $t \geq 0$, then*

$$\mathbb{P}\left(X - \mathbb{E}\left[X\right] \geq t\right) \leq \frac{\mathrm{Var}\left(X\right)}{\mathrm{Var}\left(X\right) + t^2}. \tag{2.7}$$

*Proof.* For simplicity's sake, assume $\mathbb{E}\left[X\right] = 0$: this is not restrictive, since otherwise it is enough to rescale the final result. Now, for every $t$ we could write

$$t = \mathbb{E}\left[t - X\right] \leq \mathbb{E}\left[(t - X)\mathbb{1}\{X < t\}\right].$$

Then, thanks to the Cauchy-Schwarz inequality (1.2),

$$\begin{aligned} t^2 &\leq \mathbb{E}\left[(t - X)^2\right] \mathbb{E}\left[(\mathbb{1}\{X < t\})^2\right] \\ &= \mathbb{E}\left[(t - X)^2\right] \mathbb{P}\left(X < t\right) \\ &= (\mathrm{Var}\left(X\right) + t^2)\mathbb{P}\left(X < t\right), \end{aligned}$$

meaning that

$$\mathbb{P}\left(X < t\right) \geq \frac{t^2}{\mathrm{Var}\left(X\right) + t^2}.$$

In order to conclude, we simply take the complementary:

$$\mathbb{P}\left(X \geq t\right) = 1 - \mathbb{P}\left(X < t\right).$$

□

**Theorem 2.18.** *Let $f$ and $g$ be non-decreasing real-valued functions defined over the real line. If $X$ is a real-valued random variable, then*

$$\mathbb{E}\left[f(X)g(X)\right] \geq \mathbb{E}\left[f(X)\right]\mathbb{E}\left[g(X)\right].$$

*If, however, $f$ is non-increasing and $g$ is non-decreasing, then*

$$\mathbb{E}\left[f(X)g(X)\right] \leq \mathbb{E}\left[f(X)\right]\mathbb{E}\left[g(X)\right].$$

**Theorem 2.19.** *Consider two non-decreasing functions $f, g : \mathbb{R}^n \to \mathbb{R}$. Let $X_1, \ldots, X_n$ be independent real-valued random variables and define the random vector*

$$\boldsymbol{X} = (X_1, \ldots, X_n)$$

*taking values in $\mathbb{R}^n$. Then,*

$$\mathbb{E}\left[f(\boldsymbol{X})g(\boldsymbol{X})\right] \geq \mathbb{E}\left[f(\boldsymbol{X})\right]\mathbb{E}\left[g(\boldsymbol{X})\right].$$

*If $f$ is non-increasing and $g$ is non-decreasing, then*

$$\mathbb{E}\left[f(\boldsymbol{X})g(\boldsymbol{X})\right] \leq \mathbb{E}\left[f(\boldsymbol{X})\right]\mathbb{E}\left[g(\boldsymbol{X})\right].$$

### 2.3.2 Concentrations for Sums of Independent Random Variables

We would like to bound

$$\mathbb{P}\left(S_n - \mathbb{E}\left[S_n\right] \geq t\right),$$

where $S_n = \sum_{i=1}^n X_i$ and $X_1, \ldots, X_n$ are independent real-valued random variables.

An application of Chebyshev's inequality is

$$\mathbb{P}\left(\left|S_n - \mathbb{E}\left[S_n\right]\right| \geq t\right) \leq \frac{\operatorname{Var}\left(S_n\right)}{t^2} = \sum_{i=1}^n \frac{\operatorname{Var}\left(X_i\right)}{t^2}.$$

If we set $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \operatorname{Var}\left(X_i\right)$ and take $t = n\varepsilon$, then

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}\left[X_i\right])\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}. \tag{2.8}$$

However, (2.8) is unsatisfactory as $n \to \infty$ (even if it works for small values of $n$). To understand why, apply the Central Limit Theorem:

$$\mathbb{P}\left(\sqrt{\frac{n}{\sigma^2}}\left|\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}\left[X_i\right])\right| \geq y\right) \xrightarrow{n \to \infty} 1 - \phi(y) \leq \frac{1}{\sqrt{2\pi}}\frac{e^{-\frac{y^2}{2}}}{y}.$$

Indeed,

$$\begin{aligned}
1 - \phi(y) &= \int_y^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \qquad x = z + y, \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{(z+y)^2}{2}} dz \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2} - zy - \frac{y^2}{2}} dz \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \int_0^\infty e^{-\frac{z^2}{2} - zy} dz \\
&\leq \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \left(-\frac{1}{y}\right) \int_0^\infty y e^{-zy} dz \\
&= \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \left[-\frac{e^{-zy}}{y}\right]_{z=0}^{z=\infty} \\
&= \frac{e^{-\frac{y^2}{2}}}{y\sqrt{2\pi}}.
\end{aligned}$$

Therefore, since it does not depend on $n$ anymore,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq \varepsilon\right) \approx \exp\left\{-\frac{n\varepsilon^2}{2\sigma^2}\right\}, \quad n \to \infty$$

with $\varepsilon := \sqrt{n/\sigma^2}y$. It is useful to compare the quantities

$$\frac{\sigma^2}{n\varepsilon^2} \quad \text{and} \quad \exp\left\{-\frac{n\varepsilon^2}{2\sigma^2}\right\}:$$

the Chebyshev's inequality applied to the sum of random variables does not get the tails' behaviour when $n \to \infty$. We would like to obtain a proper bound for the probability of the form provided by the Central Limit Theorem.

# 3

# Week III

## 3.1 09/10/18

### 3.1.1 Concentration Inequalities

#### Hoeffding's Inequalities

The Chernoff's bounding is convenient for bounding the tail probabilities of sums of independent random variables: recalling (2.6),

$$\begin{aligned}\mathbb{P}\left(S_n - \mathbb{E}\left[S_n\right] \geq t\right) &\leq e^{-st}\mathbb{E}\left[\exp\left\{s\sum_{i=1}^{n}(X_i - \mathbb{E}\left[X_i\right])\right\}\right] \\ &= e^{-st}\prod_{i=1}^{n}\mathbb{E}\left[e^{s(X_i - \mathbb{E}[X_i])}\right]. \end{aligned} \tag{3.1}$$

Therefore, the problem of finding bounds for the tail probabilities reduces to the problem of finding upper bounds for moment generating function (3.1) of $X_i - \mathbb{E}\left[X_i\right]$.

**Lemma 3.1.** *Let $X$ be a random variable with $\mathbb{E}\left[X\right] = 0$ and such that $0 \leq X \leq b$ almost everywhere. Then, for $s > 0$,*

$$\mathbb{E}\left[e^{sX}\right] \leq \exp\left\{\frac{s^2(b-a)^2}{8}\right\}.$$

*Proof.* By the convexity of the exponential function,

$$e^{sX} \leq \frac{X-a}{b-a}e^{sb} + \frac{b-X}{b-a}e^{sa}, \quad a \leq X \leq b.$$

Using the null-mean hypothesis and defining $p \coloneqq -\frac{a}{b-a}$,

$$\begin{aligned}\mathbb{E}\left[e^{sX}\right] &\leq \frac{b - \mathbb{E}\left[X\right]}{b-a}e^{sa} + \frac{-a + \mathbb{E}\left[X\right]}{b-a}e^{sb} \\ &= (1-p)e^{sa} + pe^{sb} \\ &= e^{sa}(1 - p + pe^{s(b-a)})\end{aligned}$$

$$= (1 - p + pe^{s(b-a)})e^{s(-p)(b-a)}$$
$$=: e^{\phi(u)}.$$

Let $u = s(b-a)$ and define $\phi : \mathbb{R} \to \mathbb{R}$ by taking the log of the above inequality, meaning

$$\phi(u) = -pu + \log(1 - p + pe^u).$$

Now,

$$\phi'(u) = -p + \frac{pe^u}{-p + 1 + pe^u},$$

therefore $\phi(0) = \phi'(0) = 0$. Moreover,

$$\begin{aligned}
\phi''(u) &= \frac{pe^u(1-p+pe^u) - pe^u pe^u}{(1-p+pe^u)^2} \\
&= \frac{pe^u(1-p)}{(1-p+pe^u)^2} \\
&= \frac{pe^u}{1-p+pe^u}\left(1 - \frac{pe^u}{1-p+pe^u}\right) \\
&= t(1-t) \\
&\leq \frac{1}{4}.
\end{aligned}$$

By Taylor's Theorem,

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{1}{2}u^2\phi''(\vartheta) \leq \frac{s^2(b-a)^2}{8}, \quad \vartheta \in [0, u],$$

which is a bound for the moment generating function of $X$.

□

We could use Lemma 3.1 in (3.1):

$$\begin{aligned}
\mathbb{P}\left(S_n - \mathbb{E}\left[S_n\right] \geq t\right) &\leq e^{-st} \prod_{i=1}^{n} \exp\left\{\frac{s^2(b_i - a_i)^2}{8}\right\} \\
&= e^{-st} \exp\left\{\frac{s^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2\right\} \\
&= \exp\left\{-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right\},
\end{aligned}$$

where

$$s = \frac{4t}{\sum_{i=1}^{n}(b_i - a_i)^2}.$$

**Theorem 3.2.** *Let $(X_1, \ldots, X_n)$ be independent random variables such that*

$$X_i \in [a_i, b_i].$$

*Then, for every $t > 0$,*

$$\mathbb{P}\left(S_n - \mathbb{E}\left[S_n\right] \geq t\right) \leq \exp\left\{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}$$

*and*

$$\mathbb{P}\left(S_n - \mathbb{E}\left[S_n\right] \leq -t\right) \leq \exp\left\{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}.$$

*These are also known as* ***Hoeffding inequalities****. Notice that we lost the dependence from the variance.*

These inequalities have the same form as the one based on the Central Limit Theorem, except that the average variance $\sigma^2$ is replaced by the upper bound

$$\frac{1}{4}\sum_{i=1}^n (b_i - a_i)^2.$$

**Bernstein's Inequality**

Assume $\mathbb{E}\left[X_i\right] = 0$: we will try to bound

$$\mathbb{E}\left[e^{sX_i}\right].$$

Now, set $\sigma_i^2 = \mathbb{E}\left[X_i^2\right]$ and define

$$F_i := \sum_{r=2}^{\infty} \frac{s^{r-2}\mathbb{E}\left[X_i^r\right]}{r!\sigma_i^2}.$$

Since

$$e^{sx} = 1 + sx + \sum_{r=2}^{\infty} \frac{s^r x^r}{r!},$$

then

$$\mathbb{E}\left[e^{sX_i}\right] = 1 + s\mathbb{E}\left[X_i\right] + \sum_{r=2}^{\infty} \frac{s^r \mathbb{E}\left[X_i^r\right]}{r!} = 1 + s^2\sigma_i^2 F_i \leq e^{s^2\sigma_i^2 F_i}.$$

Assuming that the $X_i$'s are bounded, meaning $|X_i| \leq c$, for every index $r$

$$\mathbb{E}\left[X_i^r\right] \leq c^{r-2}\sigma_i^2.$$

Indeed,

$$\mathbb{E}\left[X_i^{r-2}X_i^2\right] \leq \mathbb{E}\left[X_i^{r-2}\right]\mathbb{E}\left[X_i^2\right].$$

Thus,

$$F_i \leq \sum_{r=2}^{\infty} \frac{s^{r-2}c^{r-2}\sigma_i^2}{r!\sigma_i^2} = \frac{1}{(sc)^2}\sum_{r=2}^{\infty} \frac{(sc)^r}{r!} = \frac{e^{sc} - 1 - sc}{(sc)^2}$$

and we conclude

$$\mathbb{E}\left[e^{sX_i}\right] \leq \exp\left\{s^2\sigma_i^2 \frac{e^{sc} - 1 - sc}{s^2c^2}\right\}.$$

**Theorem 3.3.** *Let $(X_1, \dots, X_n)$ be independent real-valued random variables, with $\mathbb{E}[X_i] = 0$ and $|X_i| \leq c$. Let*

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \operatorname{Var}(X_i).$$

*For $t > 0$,*

$$\mathbb{P}\left( \sum_{i=1}^{n} X_i \geq t \right) \leq \exp\left\{ -\frac{n\sigma^2 h\left(\frac{ct}{n\sigma^2}\right)}{c^2} \right\},$$

*where $h(u) = (1+u)\log(1+u) - u$, for $u \geq 0$.*

**Corollary 3.4.** *Since*

$$h(u) \geq \frac{u^2}{2 + \frac{2u}{3}},$$

*we have that*

$$\mathbb{P}\left( \frac{1}{n} \sum_{i=1}^{n} X_i > \varepsilon \right) \leq \exp\left\{ -\frac{n\varepsilon^2}{2\sigma^2 + \frac{2c\varepsilon}{3}} \right\}.$$

*Here we recover the dependence from the variance.*

### Efron-Stein's Inequality

Let $A$ be some set and let $g : A^n \to \mathbb{R}$ be a measurable function. We consider the problem of deriving inequalities, for the variance of $g(X_1, \dots, X_n) = Z$, where $X_1, \dots, X_n$ are independent random variables in $A$.

**Lemma 3.5.** *Let $\mathbb{E}_i[Z]$ be the expected value of $Z$ with respect to $X_i$, that is*

$$\mathbb{E}_i[Z] \coloneqq \mathbb{E}[Z \mid X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n].$$

*Then,*

$$\operatorname{Var}(Z) \leq \sum_{i=1}^{n} \mathbb{E}\left[ (Z - \mathbb{E}_i[Z])^2 \right].$$

*Proof.* Recall that if $X$ and $Y$ are arbitrary bounded random variables, then

$$\mathbb{E}[\mathbb{E}[XY \mid Y]] = \mathbb{E}[Y\mathbb{E}[X \mid Y]].$$

For simplicity's sake, we will call $Z - \mathbb{E}[Z] = V$. Define

$$V_i = \mathbb{E}[Z \mid X_1, \dots, X_i] - \mathbb{E}[Z \mid X_1, \dots, X_{i-1}], \quad i = 1, \dots, n.$$

Now, by definition, $V = \sum_{i=1}^{n} V_i$, therefore

$$\operatorname{Var}(Z) = \mathbb{E}\left[ \left( \sum_{i=1}^{n} V_i \right)^2 \right],$$

since

$$\mathrm{Var}(Z) = \mathbb{E}\left[(Z - \mathbb{E}[Z])^2\right] = \mathbb{E}\left[V^2\right].$$

Therefore,

$$\mathrm{Var}(Z) = \mathbb{E}\left[\sum_{i=1}^{n} V_i^2\right] + 2\mathbb{E}\left[\sum_{i>j} V_i V_j\right] = \mathbb{E}\left[\sum_{i=1}^{n} V_i^2\right]$$

because, for every $i > j$,

$$\mathbb{E}[V_i V_j] = \mathbb{E}[\mathbb{E}[V_i V_j \mid X_1, \ldots, X_j]] = \mathbb{E}[V_j \mathbb{E}[V_i \mid X_1, \ldots, X_j]] = 0,$$

since $\mathbb{E}[V_i \mid X_1, \ldots, X_j] = 0$; indeed,

$$\begin{aligned}
\mathbb{E}[V_i \mid X_1, \ldots, X_j] &= \mathbb{E}[\mathbb{E}[Z \mid X_1, \ldots, X_i] - \mathbb{E}[Z \mid X_1, \ldots, X_{i-1}] \mid X_1, \ldots, X_j] \\
&= \mathbb{E}[\mathbb{E}[Z \mid X_1, \ldots, X_i] \mid X_1, \ldots, X_j] - \mathbb{E}[\mathbb{E}[Z \mid X_1, \ldots, X_{i-1}] \mid X_1, \ldots, X_j] \\
&= \mathbb{E}[\mathbb{E}[Z \mid X_1, \ldots, X_j]] - \mathbb{E}[\mathbb{E}[Z \mid X_1, \ldots, X_j]] \\
&= 0,
\end{aligned}$$

since $i > j$ and $i - 1 \geq j$.

Now, to bound $\mathbb{E}\left[V_i^2\right]$ we apply Jensen's inequality:

$$\begin{aligned}
V_i^2 &= (\mathbb{E}[Z \mid X_1, \ldots, X_i] - \mathbb{E}[Z \mid X_1, \ldots, X_{i-1}])^2 \\
&= (\mathbb{E}[\mathbb{E}[Z \mid X_1, \ldots, X_n] - \mathbb{E}[Z \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n] \mid X_1, \ldots, X_i])^2 \\
&\leq \mathbb{E}\left[(\mathbb{E}[Z \mid X_1, \ldots, X_n] - \mathbb{E}[Z \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n])^2 \,\middle|\, X_1, \ldots, X_i\right] \\
&= \mathbb{E}\left[(Z - \mathbb{E}_i[Z])^2 \,\middle|\, X_1, \ldots, X_i\right].
\end{aligned}$$

Finally, taking the expected values on both sides, we conclude.

□

**Theorem 3.6 (Efron-Stein).** *Let $X_1', \ldots, X_n'$ (identically distributed to $X_1, \ldots, X_n$) form an independent copy of $X_1, \ldots, X_n$ and define*

$$Z_i' = g(X_1, \ldots, X_{i-1} X_i' X_{i+1}, \ldots, X_n).$$

*Then, we have*

$$\mathrm{Var}(Z) \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left[(Z - Z_i')^2\right].$$

*Whenever*

$$g(X_1, \ldots, X_n) = X_1 + \ldots + X_n,$$

*the Efron-Stein's inequality becomes an equality.*

*Proof.* Remember that, given two independent and identically random variables $X$ and $Y$,

$$\mathrm{Var}(X) = \frac{1}{2}\mathbb{E}\left[(X - Y)^2\right].$$

Therefore, given

$$\mathrm{Var}(Z) = \mathbb{E}_i\left[(Z - \mathbb{E}_i[Z])^2\right]$$

and considering $Z$ and $Z_i'$ (which are independent and identically distributed),

$$\mathbb{E}_i\left[(Z - \mathbb{E}_i[Z])^2\right] = \frac{1}{2}\mathbb{E}_i\left[(Z - Z_i')^2\right],$$

then apply Lemma 3.5.

□

*Remark 3.7.* When $Z = g(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ is a sum of independent random variable, then the inequality is a proper equality and the bound is not improvable.

**Corollary 3.8.** *We have that*

$$\mathrm{Var}(Z) \leq \sum_{i=1}^n \mathbb{E}\left[(Z - Z_i)^2\right].$$

*Proof.* Recall that, for every random variable $X$,

$$\mathrm{Var}(X) \leq \mathbb{E}\left[(X - a)^2\right], \quad \forall a \in \mathbb{R}.$$

Therefore, for every $i = 1, \ldots, n$ we have

$$\mathbb{E}_i\left[(Z - \mathbb{E}_i[Z])^2\right] \leq \mathbb{E}_i\left[(Z - Z_i)^2\right],$$

where $Z_i = g_i(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ for an arbitrary $g_i : A^{n-1} \to \mathbb{R}$. Taking the expectations and applying Lemma 3.5 we get the result.

□

## 3.2 10/10/18

### 3.2.1 Likelihood Function

The Likelihood function contains all the statistical information required to make inference.

**Definition 3.9.** *Consider a random sample* $\boldsymbol{X} = (X_1, \dots, X_n)$ *from* $X \sim f_X(x; \vartheta)$. *Then, given an observation* $\boldsymbol{X} = \boldsymbol{x}$, *the function of* $\vartheta$ *defined by*

$$\mathcal{L}(\vartheta, \boldsymbol{x}) \coloneqq f_{\boldsymbol{X}}(\boldsymbol{x}; \vartheta) = \prod_{i=1}^{n} f_{X_i}(x_i; \vartheta)$$

*is called* ***Likelihood function****.*

This Definition almost seems to be defining the Likelihood function to be the same as the probability density function, or the probability mass function. The only distinction between these two is which variable is considered fixed and which is varying. When we consider the probability density function, or the probability mass function, $f(\boldsymbol{x}; \vartheta)$, we are considering $\vartheta$ as fixed and $\boldsymbol{x}$ as the variable; when we consider the Likelihood function $\mathcal{L}(\vartheta; \boldsymbol{x})$, we are considering $\boldsymbol{x}$ to be the observed sample point and $\vartheta$ to be varying over all possible values. Notice that, in Probability, we are interested in computing something like $\mathbb{P}(\boldsymbol{X} \in \{\cdot\})$, and $\vartheta$ must be known to achieve the result. In Statistics, on the other hand, observations are given and used to infer information about the parameter $\vartheta$.

For mathematical convenience, we shall use the **Log-Likelihood function**,

$$\log \mathcal{L}(\vartheta; \boldsymbol{x}) \coloneqq \sum_{i=1}^{n} \log f_{X_i}(x_i; \vartheta) :$$

indeed, we are generally interested in knowing where the Likelihood function reaches its maximum value. Since the logarithm is a strictly increasing function, so the logarithm of a function achieves its maximum value at the same points as the function itself.

If both $\boldsymbol{x}$ and $\vartheta$ are fixed, then $\mathcal{L}(\vartheta; \boldsymbol{x})$ is related to the probability that the particular value that we fixed for $\vartheta$ has generated $\boldsymbol{x}$.

**Definition 3.10.** *Consider the Likelihood function* $\mathcal{L}(\vartheta; \boldsymbol{x})$ *and* $\vartheta_1, \vartheta_2 \in \Theta$. *Suppose we have fixed* $\vartheta_1, \vartheta_2 \in \Theta$ *and we compute* $\mathcal{L}(\vartheta_i; \boldsymbol{x})$ *for* $i = 1, 2$. *We say that* $\vartheta_1$ *is* ***more likely*** *than* $\vartheta_2$ *if*

$$\mathcal{L}(\vartheta_1, \boldsymbol{x}) > \mathcal{L}(\vartheta_2; \boldsymbol{x}).$$

*Remark 3.11.* This means that $\boldsymbol{x}$ has been more likely generated under $\vartheta_1$, since under $\vartheta_1$ the probability of generating $\boldsymbol{x}$ is higher than under $\vartheta_2$. In general, we would like to optimise the Likelihood function with respect to $\vartheta$: this will not always be possible through differentiation.

The statistical inference based on Likelihood functions is a consequence of two principles.

**Weak Likelihood Principle**

For a fixed parametric model $X \sim F_X(x;\vartheta)$, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are two observed sample points such that

$$\mathcal{L}(\vartheta;\boldsymbol{x}) \propto \mathcal{L}(\vartheta;\boldsymbol{y}),$$

then the two samples provide the same inference on $\vartheta$.

**Strong Likelihood Principle**

Let $\boldsymbol{X} = \boldsymbol{x}$ be an observed sample point under the model $X \sim F_X(x;\vartheta)$ with Likelihood function $\mathcal{L}_F(\boldsymbol{x};\vartheta)$ and let $\boldsymbol{Y} = \boldsymbol{y}$ be an observed sample point under the model $Y \sim G_Y(y;\vartheta)$ with Likelihood function $\mathcal{L}_G(\boldsymbol{y};\vartheta)$. If $\mathcal{L}_F(\boldsymbol{x};\vartheta) \propto \mathcal{L}_G(\boldsymbol{y};\vartheta)$, then the two samples provide the same inference on $\vartheta$.

*Example 3.12.* Suppose we have a coin and we want to determine the "success" probability. To do so, we flip the coin three times and we try to infer some information about its probability, according to the outcomes. This means that we are working with a random sample $(X_1, X_2, X_3)$. Compare the following cases:

$$X \sim Ber(\vartheta_1), \quad \vartheta_1 = \frac{1}{2};$$

$$X \sim Ber(\vartheta_2), \quad \vartheta_2 = \frac{1}{3};$$

$$X \sim Ber(\vartheta_3), \quad \vartheta_3 = \frac{1}{4}.$$

The results are the following:

| $(x_1, x_2, x_3)$ | $\vartheta_1$ | $\vartheta_2$ | $\vartheta_3$ |
|---|---|---|---|
| (0, 0, 0) | 1/8 | 8/27 | 27/64* |
| (0, 0, 1) | 1/8 | 4/27* | 9/64 |
| (0, 1, 0) | 1/8 | 4/27* | 9/64 |
| (1, 0, 0) | 1/8 | 4/27* | 9/64 |
| (0, 1, 1) | 1/8* | 2/27 | 3/64 |
| (1, 0, 1) | 1/8* | 2/27 | 3/67 |
| (1, 1, 0) | 1/8* | 2/27 | 3/67 |
| (1, 1, 1) | 1/8* | 1/27 | 1/64 |

According to the Likelihood Principle, we know that our coin's probability is the one with the highest values.

## 3.3 11/10/18

### 3.3.1 Likelihood Function Method: Assumptions

**Definition 3.13.** *Consider $X \sim f_X(x;\vartheta)$ and fix $\vartheta \in \Theta$, where $\Theta$ is an open set of parameters. The following conditions are called* ***conditions of regularity*** *for $\{f_X(\cdot,\vartheta) \mid \vartheta \in \Theta\}$:*

1. *for every $\vartheta \in \Theta$ there exist derivatives of the Likelihood function with respect to $\vartheta$, at least up to the third order;*
2. *for every $\vartheta_0 \in \Theta$ there exist three functions, namely $g(\cdot)$, $h(\cdot)$ and $H(\cdot)$, that are integrable in a neighbourhood of $\vartheta_0$ and such that*

$$\left|\frac{d}{d\vartheta} f_X(x;\vartheta)\right| \leq g(x),$$
$$\left|\frac{d^2}{d\vartheta^2} f_X(x;\vartheta)\right| \leq h(x),$$
$$\left|\frac{d^3}{d\vartheta^3} \log(f_X(x;\vartheta))\right| \leq H(x);$$

3. *for every $\vartheta \in \Theta$,*

$$0 < \mathbb{E}\left[\left(\frac{d}{d\vartheta}\log \mathcal{L}(\vartheta;\boldsymbol{x})\right)^2\right] < \infty.$$

Note that $\log \mathcal{L}(\vartheta;\boldsymbol{x})$ is a well-defined random variable.

**Definition 3.14.** *A statistical model is* ***identifiable*** *if, for every $\vartheta_1, \vartheta_2 \in \Theta$, there exist at least an event $E$ such that*

$$\mathbb{P}(X \in E;\vartheta_1) \neq \mathbb{P}(X \in E;\vartheta_2).$$

**Definition 3.15.** *The random variable*

$$V_n(\vartheta) \coloneqq \log \mathcal{L}(\vartheta,\boldsymbol{x}) = \sum_{i=1}^{n} \log f_{X_i}(x_i;\vartheta)$$

*is called* ***Log-Likelihood random variable****. Moreover, we call*

$$V_n'(\vartheta) \coloneqq \frac{d}{d\vartheta} V_n(\vartheta) = \frac{\mathcal{L}'(\vartheta;\boldsymbol{x})}{\mathcal{L}(\vartheta;\boldsymbol{x})}$$

*the* ***score function****.*

**Proposition 3.16.** *The following properties are satisfied:*

1. $\mathbb{E}[V_n'(\vartheta)] = 0$;
2. $\mathrm{Var}(V_n'(\vartheta)) = \mathbb{E}\left[(V_n'(\vartheta))^2\right] = -\mathbb{E}[V_n''(\vartheta)]$.

*Proof.* (1) First,

$$\begin{aligned}\mathbb{E}\left[V_n'(\vartheta)\right] &= \int_{\mathbb{R}^n} V_n'(\vartheta) f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) d\boldsymbol{x} \\ &= \int_{\mathbb{R}^n} \frac{f_{\boldsymbol{X}}'(\boldsymbol{x};\vartheta)}{f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)} f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) d\boldsymbol{x} \\ &= \int_{\mathbb{R}^n} \frac{d}{d\vartheta} f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) d\boldsymbol{x} \\ &= \frac{d}{d\vartheta} \int_{\mathbb{R}^n} f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) d\boldsymbol{x} \\ &= 0.\end{aligned}$$

(2) Second,

$$\begin{aligned}V_n''(\vartheta) &= \frac{d^2}{d\vartheta^2} \log \mathcal{L}(\vartheta;\boldsymbol{x}) \\ &= \frac{d}{d\vartheta} \frac{f_{\boldsymbol{X}}'(\boldsymbol{x};\vartheta)}{f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)} \\ &= \frac{f_{\boldsymbol{X}}''(\boldsymbol{x};\vartheta) f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) - f_{\boldsymbol{X}}'(\boldsymbol{x};\vartheta) f_{\boldsymbol{X}}'(\boldsymbol{x};\vartheta)}{(f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta))^2} \\ &= \frac{f_{\boldsymbol{X}}''(\boldsymbol{x};\vartheta)}{f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)} - \left(\frac{f_{\boldsymbol{X}}'(\boldsymbol{x};\vartheta)}{f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)}\right)^2 \\ &= \frac{f_{\boldsymbol{X}}''(\boldsymbol{x};\vartheta)}{f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)} - (V_n'(\vartheta))^2,\end{aligned}$$

therefore

$$\begin{aligned}\mathbb{E}\left[V_n''(\vartheta)\right] &= \int_{\mathbb{R}^n} \frac{f_{\boldsymbol{X}}''(\boldsymbol{x};\vartheta)}{f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)} f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) d\boldsymbol{x} - \mathbb{E}\left[(V_n'(\vartheta))^2\right] \\ &= \int_{\mathbb{R}^n} \frac{d^2}{d\vartheta^2} f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) d\boldsymbol{x} - \mathbb{E}\left[(V_n'(\vartheta))^2\right] \\ &= \frac{d^2}{d\vartheta^2} \int_{\mathbb{R}^n} f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) d\boldsymbol{x} - \mathbb{E}\left[(V_n'(\vartheta))^2\right] \\ &= -\mathbb{E}\left[(V_n'(\vartheta))^2\right].\end{aligned}$$

□

**Definition 3.17.** $I_n(\vartheta) := -\mathbb{E}\left[V_n''(\vartheta)\right]$ *is called* ***Fisher information****.*

*Remark 3.18.* If $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a random sample from $X \sim f_X(x, \vartheta)$ regular, then

$$I_n(\vartheta) = nI_1(\vartheta).$$

*Example 3.19.* Consider a random sample $(X_1, \ldots, X_n)$ from $X \sim Ber(\vartheta)$. Then,

$$\begin{aligned}\mathcal{L}(\vartheta, \boldsymbol{x}) &= \prod_{i=1}^n f_{X_i}(x_i;\vartheta) \\ &= \vartheta^{\sum_{i=1}^n X_i} (1-\vartheta)^{n-\sum_{i=1}^n X_i};\end{aligned}$$

$$
\begin{aligned}
V_n(\vartheta) &= \log\vartheta \sum_{i=1}^{n} X_i + \log(1-\vartheta)\left(n - \sum_{i=1}^{n} X_i\right);\\
V_n'(\vartheta) &= \frac{\sum_{i=1}^{n} X_i}{\vartheta} - \frac{n - \sum_{i=1}^{n} X_i}{1-\vartheta};\\
V_n''(\vartheta) &= \frac{-\sum_{i=1}^{n} X_i}{\vartheta^2} - \frac{n - \sum_{i=1}^{n} X_i}{(1-\vartheta)^2};\\
\mathbb{E}\left[V_n'(\vartheta)\right] &= \frac{1}{\vartheta}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] - \frac{1}{1-\vartheta}\left(n - \sum_{i=1}^{n}\mathbb{E}\left[X_i\right]\right)\\
&= \frac{n\vartheta}{\vartheta} - \frac{n - n\vartheta}{1-\vartheta}\\
&= 0;\\
\mathbb{E}\left[V_n''(\vartheta)\right] &= \frac{1}{\vartheta^2}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] - \frac{1}{(1-\vartheta)^2}\left(n - \sum_{i=1}^{n}\mathbb{E}\left[X_i\right]\right)\\
&= \frac{n\vartheta}{\vartheta^2} - \frac{n - n\vartheta}{(1-\vartheta)^2}\\
&= -I_n(\vartheta).
\end{aligned}
$$

*Remark 3.20.* If $\Theta$ is $m$-dimensional, then $\operatorname{Im}\boldsymbol{\vartheta}$ is an $m \times m$ matrix.

### 3.3.2 The Exponential Family

**Definition 3.21.** *A single-parameter **exponential family** (EF) is a set of probability distributions whose probability density function (or probability mass function) can be expressed in the following form: if $X \sim EF(\vartheta)$*

$$f_X(x;\vartheta) \coloneqq \exp\left\{Q(\vartheta)A(x) + C(x) - K(\vartheta)\right\}. \tag{3.2}$$

*On the other hand, if we consider a sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from $X \sim f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)$,*

$$f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) \coloneqq \exp\left\{Q(\vartheta)\sum_{i=1}^{n} A(x_i) + \sum_{i=1}^{n} C(x_i) - K(\vartheta)\right\}.$$

*Example 3.22.* Consider $X \sim Ber(\vartheta)$, $X \in \{0,1\}$ and $\Theta = (0,1)$. We know that

$$p_X(x) = \vartheta^x(1-\vartheta)^{1-x}\mathbb{1}_{\{0,1\}}(x).$$

Now,

$$
\begin{aligned}
\exp\left\{\log\left(\vartheta^x(1-\vartheta)^{1-x}\right)\right\} &= \exp\left\{x\log\vartheta + (1-x)\log(1-\vartheta)\right\}\\
&= \exp\left\{x\log\vartheta + \log(1-\vartheta) - x\log(1-\vartheta)\right\}\\
&= \exp\left\{x\log\left(\frac{\vartheta}{1-\vartheta}\right) + \log(1-\vartheta)\right\},
\end{aligned}
$$

where

$$Q(\vartheta) = \log\left(\frac{\vartheta}{1-\vartheta}\right), \quad C(x) = 0, \quad A(x) = x, \quad K(\vartheta) = -\log(1-\vartheta).$$

**Properties of the Exponential Family**

**Proposition 3.23.** *Let $X \sim EF(\vartheta)$. Then,*

1. $\mathbb{E}[A(x)] = \frac{K'(\vartheta)}{Q'(\vartheta)}$;
2. $\mathrm{Var}(A(x)) = \frac{K''(\vartheta)}{[Q'(\vartheta)]^2} - \mathbb{E}[A(x)] \frac{Q''(\vartheta)}{[Q'(\vartheta)]^2}$.

*Proof.* 1. Differentiating (3.2) we obtain

$$\frac{d}{d\vartheta} f_X(x;\vartheta) = [A(x)Q'(\vartheta) - K'(\vartheta)] f_X(x;\vartheta),$$

therefore, by taking the integral on both sides of the equality,

$$\begin{aligned}\int_{\mathbb{R}} \frac{d}{d\vartheta} f_X(x;\vartheta)dx &= \int_{\mathbb{R}} [A(x)Q'(\vartheta) - K'(\vartheta)] f_X(x;\vartheta)dx \\ &= Q'(\vartheta)\int_{\mathbb{R}} A(x) f_X(x;\vartheta)dx - K'(\vartheta)\int_{\mathbb{R}} f_X(x;\vartheta)dx\end{aligned}$$

we get

$$0 = Q'(\vartheta)\mathbb{E}[A(x)] - K'(\vartheta).$$

We conclude that

$$\mathbb{E}[A(x)] = \frac{K'(\vartheta)}{Q'(\vartheta)}.$$

2. For the variance, we differentiate (3.2) twice:

$$\frac{d^2}{d\vartheta^2} f_X(x;\vartheta) = [A(x)Q''(\vartheta) - K''(\vartheta)] f_X(x;\vartheta) + [A(x)Q'(\vartheta) - K'(\vartheta)]^2 f_X(x;\vartheta),$$

therefore, by taking the integral on both sides of the equality,

$$\begin{aligned}\int_{\mathbb{R}} \frac{d^2}{d\vartheta^2} f_X(x;\vartheta)dx = &\int_{\mathbb{R}} [A(x)Q''(\vartheta) - K''(\vartheta)] f_X(x;\vartheta)dx + \\ &+ \int_{\mathbb{R}} [A(x)Q'(\vartheta) - K'(\vartheta)]^2 f_X(x;\vartheta)dx\end{aligned}$$

we get

$$\begin{aligned}0 &= Q''(\vartheta)\mathbb{E}[A(x)] - K''(\vartheta) + [Q'(\vartheta)]^2 \int_{\mathbb{R}} \left[A(x) - \frac{K'(\vartheta)}{Q'(\vartheta)}\right]^2 f_X(x;\vartheta)dx \\ &= Q''(\vartheta)\frac{K'(\vartheta)}{Q'(\vartheta)} - K''(\vartheta) + [Q'(\vartheta)]^2 \mathrm{Var}(A(x)).\end{aligned}$$

We conclude that

$$\begin{aligned}\mathrm{Var}(A(x)) &= \frac{K''(\vartheta) - Q''(\vartheta)\frac{K'(\vartheta)}{Q'(\vartheta)}}{[Q'(\vartheta)]^2} \\ &= \frac{K''(\vartheta)}{[Q'(\vartheta)]^2} - \mathbb{E}[A(x)]\frac{Q''(\vartheta)}{[Q'(\vartheta)]^2}.\end{aligned}$$

□

*Remark 3.24.* If $Q(\vartheta) = \vartheta$, then

$$\mathbb{E}[A(x)] = K'(\vartheta) \quad \text{and} \quad \mathrm{Var}(A(x)) = K''(\vartheta).$$

### 3.3.3 The *Natural* Exponential Family

A *natural exponential family* (NEF) is a class of probability distributions that is a special case of an exponential family. Every distribution possessing a moment-generating function is a member of a natural exponential family, and the use of such distributions simplifies the theory and computation of generalized linear models.

**Definition 3.25.** *The* ***natural exponential family*** *is a subset of the exponential family. It is an exponential family in which* $A(x) = x$ *and* $Q(\vartheta) = \eta$*:*

$$f_X(x;\eta) = \exp\left\{\eta x + C(x) - K(\eta)\right\}.$$

# 4

# Week IV

## 4.1 16/10/18

An experimenter uses the information in a sample $X_1, \ldots, X_n$ to make inferences about an unknown parameter $\vartheta$. If the sample size $n$ is large, then the observed sample $x_1, \ldots, x_n$ is a long list of numbers that may be hard to interpret. An experimenter might wish to summarise the information in a sample by determining a few key features of the sample values. This is usually done by computing statistics, functions of the sample. For example, the *sample mean*, the *sample variance*, the *largest observation* and the *smallest observation* are four statistics that might be used to summarise some key features of the sample.

Any statistic $T(\boldsymbol{X})$ defines a form of data reduction or data summary. An experimenter who uses only the observed value of the statistic, $T(\boldsymbol{X}) = T(\boldsymbol{x})$, rather than the entire observed sample, $\boldsymbol{X} = \boldsymbol{x}$, will treat as equal two samples, $\boldsymbol{X} = \boldsymbol{x}$ and $\boldsymbol{Y} = \boldsymbol{y}$, that satisfy $T(\boldsymbol{x}) = T(\boldsymbol{y})$ even though the actual sample values may be different in some ways.

Data reduction in terms of a particular statistic can be thought of as a partition of the sample space $\mathcal{X}$. Let $\mathcal{T} = \{t \,|\, t = T(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $T(\boldsymbol{x})$. Then, $T(\boldsymbol{x})$ partitions the sample space into sets $A_t = \{\boldsymbol{x} \,|\, T(\boldsymbol{x}) = t\}$. The statistic summarises the data in that, rather than reporting the entire sample $\boldsymbol{x}$, it reports only $T(\boldsymbol{x}) = t$ or, equivalently, $\boldsymbol{x} \in A_t$.

We study two principles of data reduction. We are interested in methods of data reduction that do not discard important information about the unknown parameter $\vartheta$ and methods that successfully discard information that is irrelevant as far as gaining knowledge about $\vartheta$ is concerned. The **Sufficiency Principle** promotes a method of data reduction that does not discard information about $\vartheta$ while achieving some summarisation of the data. The **Likelihood Principle** describes a function of the parameter, determined by the observed sample, that contains all the information about $\vartheta$ that is available from the sample.

### 4.1.1 Sufficiency Principle

**Definition 4.1.** *A statistic $T_n = T(\boldsymbol{X})$ is a **sufficient statistic** for $\vartheta$ if the conditional distribution of the sample $\boldsymbol{X}$, given the values of $T(\boldsymbol{X})$ does not depend on $\vartheta$.*

Basically, a sufficient statistic for a parameter $\vartheta$ is a statistic that, in a certain sense, captures all the information about $\vartheta$ contained in the sample. Any additional information in the sample, besides the values of the sufficient statistic, does not contain any more information about $\vartheta$.

To use this Definition to verify that a statistic $T_n = T(\boldsymbol{X})$ is a sufficient statistic for $\vartheta$, we must verify that for any fixed values of $\boldsymbol{x}$ and $t$, the conditional probability

$$\mathbb{P}\left(\boldsymbol{X} = \boldsymbol{x} \,|\, T(\boldsymbol{X}) = t\right)$$

is the same for all values of $\vartheta$. Now, this probability is zero for every value of $\vartheta$ if $T(\boldsymbol{x}) \neq t$. So, we must verify only that

$$\mathbb{P}\left(\boldsymbol{X} = \boldsymbol{x} \,|\, T(\boldsymbol{X}) = T(\boldsymbol{x})\right)$$

does not depend on $\vartheta$. But since $\{\boldsymbol{X} = \boldsymbol{x}\}$ is a subset of $\{T(\boldsymbol{X}) = T(\boldsymbol{x})\}$,

$$\begin{aligned}
\mathbb{P}\left(\boldsymbol{X} = \boldsymbol{x} \,|\, T(\boldsymbol{X}) = T(\boldsymbol{x})\right) &= \frac{\mathbb{P}\left(\boldsymbol{X} = \boldsymbol{x}, T(\boldsymbol{X}) = T(\boldsymbol{x})\right)}{\mathbb{P}\left(T(\boldsymbol{X}) = T(\boldsymbol{x})\right)} \\
&= \frac{\mathbb{P}\left(\boldsymbol{X} = \boldsymbol{x}\right)}{\mathbb{P}\left(T(\boldsymbol{X}) = T(\boldsymbol{x})\right)} \\
&= \frac{h_{(\boldsymbol{X},T_n)}(\boldsymbol{x}, t; \vartheta)}{g_{T_n}(t; \vartheta)} \\
&=: \phi_{\boldsymbol{X} \,|\, T_n = t}(\boldsymbol{x}; t, \vartheta), \qquad (4.1)
\end{aligned}$$

where

- $h_{(\boldsymbol{X},T_n)}(\boldsymbol{x}, t; \vartheta) := \mathbb{P}\left(\boldsymbol{X} = \boldsymbol{x}\right)$ is the joint distribution of $\boldsymbol{X}$ and $T_n$;
- $g_{T_n}(t; \vartheta) := \mathbb{P}\left(T(\boldsymbol{X}) = T(\boldsymbol{x})\right)$ is the marginal distribution of $T_n$;
- $\phi_{\boldsymbol{X} \,|\, T_n = t}(\boldsymbol{x}; t, \vartheta)$ is the conditional distribution of $\boldsymbol{X}$ given $T_n$.

Thus, $T_n = T(\boldsymbol{X})$ is a sufficient statistic for $\vartheta$ if and only if, for every $\boldsymbol{x}$, (4.1) is a constant as a function of $\vartheta$. Please note that if $\boldsymbol{X}$ and $T(\boldsymbol{X})$ have continuous distributions, then the above conditional probabilities cannot be interpreted as we did. However, it is still appropriate to use the above criterion to determine if $T(\boldsymbol{X})$ is a sufficient statistic for $\vartheta$, simply considering the distributions.

*Example 4.2.* Consider a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from $X \sim Ber(\vartheta)$. Then, $T(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for $\vartheta$.

*Proof.* Note that $T_n = T(\boldsymbol{X})$ counts the number of $X_i$'s that equal 1, so $T_n = T(\boldsymbol{X})$ is a Binomial with parameters $(n, \vartheta)$. The ratio of the probability mass functions is thus

$$\begin{aligned}
\frac{h_{(\boldsymbol{X},T_n)}(\boldsymbol{x}, t; \vartheta)}{g_{T_n}(t; \vartheta)} &= \frac{\prod_{i=1}^{n} \vartheta^{x_i}(1-\vartheta)^{1-x_i}}{\binom{n}{t}\vartheta^t(1-\vartheta)^{n-t}} \qquad t = \sum_{i=1}^{n} x_i \\
&= \frac{\vartheta^{\sum_{i=1}^{n} x_i}(1-\vartheta)^{\sum_{i=1}^{n}(1-x_i)}}{\binom{n}{t}\vartheta^t(1-\vartheta)^{n-t}} \\
&= \frac{\vartheta^t(1-\vartheta)^{n-t}}{\binom{n}{t}\vartheta^t(1-\vartheta)^{n-t}}
\end{aligned}$$

$$= \frac{1}{\binom{n}{t}}.$$

Therefore, $T_n = T(\boldsymbol{X})$ is sufficient for $\vartheta$.

Similarly, we can approach this problem from another direction: let us call

$$I_0 = \left\{ (x_1, \dots, x_n) \,\middle|\, x_i \in \{0,1\} \wedge \sum_{i=1}^{n} x_i = 0 \right\}$$
$$I_1 = \left\{ (x_1, \dots, x_n) \,\middle|\, x_i \in \{0,1\} \wedge \sum_{i=1}^{n} x_i = 1 \right\}$$
$$\vdots$$
$$I_n = \left\{ (x_1, \dots, x_n) \,\middle|\, x_i \in \{0,1\} \wedge \sum_{i=1}^{n} x_i = n \right\}$$

such that $I_i \cap I_j = \emptyset$ and $\bigcup_{j=1}^{n} I_j = \{0,1\}^n$. This means that the probability is simply

$$\mathbb{P}\left(\boldsymbol{X} = \boldsymbol{x} \,|\, T_n = t\right) = \frac{1}{\binom{n}{t}}.$$

□

It may be unwieldy to use the definition of a sufficient statistic to find a sufficient statistic for a particular model. To use the definition, we must guess a statistic $T_n = T(\boldsymbol{X})$ to be sufficient, find the probability mass function (or probability density function) of $T(\boldsymbol{X})$, and check that the ratio of the probability mass functions (or probability density functions) does not depend on $\vartheta$. The first step requires a good deal of intuition and the second sometimes requires some tedious analysis.

**Theorem 4.3 (Fisher Factorisation).** *If $h_{(\boldsymbol{X},T_n)}(\boldsymbol{x}, t; \vartheta)$ is the joint probability mass function (or probability density function) of $\boldsymbol{X}$ and $T_n$, and $g_{T_n}(t;\vartheta)$ is the probability density function (or probability mass function) of $T_n$, then $T_n = T(\boldsymbol{X})$ is a sufficient statistic for $\vartheta$ if and only if, for every $\boldsymbol{x} \in \mathcal{X}$, $\phi_{\boldsymbol{X} \,|\, T_n = t}(\boldsymbol{x}; t, \vartheta)$ is constant as a function of $\vartheta$.*

## 4.2 17/10/18

### 4.2.1 Savage's Factorisation Theorem

Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from $f_X(x;\vartheta)$. The idea is to determine that a statistic $T_n$ is sufficient for $\vartheta$ if and only if there exist two non-negative functions $g(\cdot)$ and $h(\cdot)$ such that

$$\mathcal{L}(\vartheta;\boldsymbol{x}) = g(T(\boldsymbol{X});\vartheta)h(\boldsymbol{x}),$$

where $g(\cdot)$ is a function of the observable sample via $T(\cdot)$, while $h(\cdot)$ is a function of the observed sample which does not depend on $\vartheta$.

**Theorem 4.4 (Savage's Factorisation).** *Consider a random sample $\boldsymbol{X}$ from $X \sim f_X(x;\vartheta)$. Let $f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)$ be the joint probability density function (or probability mass function) of $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is sufficient for $\vartheta$ if and only if there exist two functions, namely $g(\cdot)$ and $h(\cdot)$, such that, for every $\boldsymbol{x} \in \mathcal{X}$ and every $\vartheta \in \Theta$,*

$$f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) = g(T(\boldsymbol{x});\vartheta)h(\boldsymbol{x}). \tag{4.2}$$

*Proof.* We give the proof only for discrete distributions.

Suppose $T(\boldsymbol{X})$ is a sufficient statistic. Choose

$$g(t;\vartheta) = \mathbb{P}\left(T(\boldsymbol{x}) = t\right) \quad \text{and} \quad h(\boldsymbol{x}) = \mathbb{P}\left(\boldsymbol{X} = \boldsymbol{x} \,|\, T(\boldsymbol{X}) = T(\boldsymbol{x})\right).$$

Because $T(\boldsymbol{X})$ is sufficient, the conditional probability defining $h(\boldsymbol{x})$ does not depend on $\vartheta$. Thus, this choice of $h(\boldsymbol{x})$ and $g(t;\vartheta)$ is legitimate, and for this choice we have

$$\begin{aligned} f_X(\boldsymbol{x};\vartheta) &= \mathbb{P}\left(\boldsymbol{X} = \boldsymbol{x}\right) \\ &= \mathbb{P}\left(\boldsymbol{X} = \boldsymbol{x}, T(\boldsymbol{X}) = T(\boldsymbol{x})\right) \\ &= \mathbb{P}\left(T(\boldsymbol{X}) = T(\boldsymbol{x})\right)\mathbb{P}\left(\boldsymbol{X} = \boldsymbol{x} \,|\, T(\boldsymbol{X}) = T(\boldsymbol{x})\right) \\ &= g(T(\boldsymbol{x});\vartheta)h(\boldsymbol{x}). \end{aligned}$$

Then, the factorisation (4.2) has been exhibited. Notice that

$$\mathbb{P}\left(T(\boldsymbol{X}) = T(\boldsymbol{x})\right) = g(T(\boldsymbol{x});\vartheta),$$

meaning that $g(T(\boldsymbol{x});\vartheta)$ is the probability mass function of $T(\boldsymbol{X})$.

Now, assume that the factorisation (4.2) exists. Let $q(t;\vartheta)$ be the probability mass function of $T(\boldsymbol{X})$. To show that $T(\boldsymbol{X})$ is sufficient, we examine the ratio

$$\frac{f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)}{q(T(\boldsymbol{x});\vartheta)}.$$

Define

$$A_{T(\boldsymbol{x})} \coloneqq \{\boldsymbol{y} \,|\, T(\boldsymbol{y}) = T(\boldsymbol{x})\};$$

then,

$$\begin{aligned} \frac{f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)}{q(T(\boldsymbol{x});\vartheta)} &= \frac{g(T(\boldsymbol{x});\vartheta)h(\boldsymbol{x})}{q(T(\boldsymbol{x});\vartheta)} \\ &= \frac{g(T(\boldsymbol{x});\vartheta)h(\boldsymbol{x})}{\sum_{\boldsymbol{y}\in A_{T(\boldsymbol{x})}} g(\boldsymbol{y};\vartheta)h(\boldsymbol{y})} \end{aligned}$$

$$= \frac{g(T(\boldsymbol{x});\vartheta)h(\boldsymbol{x})}{\sum_{\boldsymbol{y}\in A_{T(\boldsymbol{x})}} g(T(\boldsymbol{x});\vartheta)h(\boldsymbol{y})}$$
$$= \frac{g(T(\boldsymbol{x});\vartheta)h(\boldsymbol{x})}{g(T(\boldsymbol{x});\vartheta)\sum_{\boldsymbol{y}\in A_{T(\boldsymbol{x})}} h(\boldsymbol{y})}$$
$$= \frac{h(\boldsymbol{x})}{\sum_{\boldsymbol{y}\in A_{T(\boldsymbol{x})}} h(\boldsymbol{y})}$$

since $T$ is a constant on $A_{T(\boldsymbol{x})}$. Therefore, because the ratio does not depend on $\vartheta$, we conclude thanks to Theorem 4.3 that $T(\boldsymbol{X})$ is a sufficient statistic for $\vartheta$.

□

*Remark 4.5.* To use Theorem 4.4 to find a sufficient statistic, we factor the probability density function (or probability mass function) into two parts, with one part not depending on $\vartheta$. The part that does not depend on $\vartheta$ constitutes $h(\boldsymbol{x})$. The other one, the one that does depend on $\vartheta$, usually depends on the sample $\boldsymbol{x}$ only through some function $T(\boldsymbol{x})$, and this function is a sufficient statistic for $\vartheta$.

*Example 4.6.* Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from $X \sim Ber(\vartheta)$. Then,

$$\mathcal{L}(\vartheta;\boldsymbol{x}) = \prod_{i=1}^{n} \vartheta^{x_i}(1-\vartheta)^{1-x_i} = \vartheta^{\sum_{i=1}^n x_i}(1-\vartheta)^{n-\sum_{i=1}^n x_i}.$$

Here

$$h(T(\boldsymbol{x})) = 1 \quad \text{and} \quad g(\boldsymbol{x}) = g\left(\sum_{i=1}^{n} x_i\right) = \vartheta^{\sum_{i=1}^n x_i}(1-\vartheta)^{n-\sum_{i=1}^n x_i}.$$

*Example 4.7.* Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from $X \sim \mathcal{N}(0,1)$. Then,

$$\begin{aligned}
\mathcal{L}(\vartheta;\boldsymbol{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x_i-\vartheta)^2\right\} \\
&= (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\vartheta)^2\right\} \\
&= (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} x_i^2 - \frac{n\vartheta^2}{2} + \vartheta\sum_{i=1}^{n} x_i\right\} \\
&= (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right\} \exp\left\{-\frac{n\vartheta^2}{2} + \vartheta\sum_{i=1}^{n} x_i\right\} \\
&= h(\boldsymbol{x})g(T(\boldsymbol{x});\vartheta).
\end{aligned}$$

Here

$$h(\boldsymbol{x}) = (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right\}$$
$$g(T(\boldsymbol{x});\vartheta) = \exp\left\{-\frac{n\vartheta^2}{2} - \vartheta\sum_{i=1}^{n} x_i\right\}.$$

Consider now a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from $X \sim \mathcal{N}(\mu, \sigma^2)$, with $\mu$ unknown. Then,

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}; \mu, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x}_n + \overline{x}_n - \mu)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 + n(\overline{x}_n - \mu)^2\right]\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2\right\} \exp\left\{-\frac{n(\overline{x}_n - \mu)^2}{2\sigma^2}\right\}. \end{aligned}$$

We can define

$$h(\boldsymbol{x}) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2\right\}$$

which does not depend on the unknown parameter $\mu$. The factor that contains $\mu$ depends on the sample $\boldsymbol{x}$ only through the sample mean $T(\boldsymbol{x}) = \overline{x}_n$. So we have

$$g(t; \mu, \sigma^2) = \exp\left\{-\frac{n(\overline{x}_n - \mu)^2}{2\sigma^2}\right\},$$

and note that

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \mu, \sigma^2) = h(\boldsymbol{x}) g(T(\boldsymbol{x}); \mu).$$

Thus, by Theorem 4.4, $T(\boldsymbol{X}) = \overline{X}_n$ is a sufficient statistic for $(\mu, \sigma^2)$.

### 4.2.2 Minimal Sufficient Statistic

In the preceding section, we found one sufficient statistic for each model considered. In any problem there are, however, many sufficient statistics.

It is always true that the complete sample $\boldsymbol{X}$ is a sufficient statistic. We can factor the probability density function (or probability mass function) of $\boldsymbol{X}$ as

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \vartheta) = g(T(\boldsymbol{x}); \vartheta) h(\boldsymbol{x}),$$

where

$$g(T(\boldsymbol{x}); \vartheta) = \mathcal{L}(\vartheta; \boldsymbol{x}) \quad \text{and} \quad h(\boldsymbol{x}) = 1, \quad \forall \boldsymbol{x} \in \mathcal{X}.$$

By Theorem 4.4, $T(\boldsymbol{X}) = \boldsymbol{X}$ is a sufficient statistic.

Also, it follows that any one-to-one function of a sufficient statistic is a sufficient statistic. Suppose $T(\boldsymbol{X})$ is a sufficient statistic and define

$$T^*(\boldsymbol{x}) = r(T(\boldsymbol{x})), \quad \forall \boldsymbol{x} \in \mathcal{X},$$

where $r(\cdot)$ is a one-to-one function with inverse $r^{-1}$. Then, by Theorem 4.4, there exist $g$ and $h$ such that

$$\mathcal{L}(\vartheta; \boldsymbol{x}) = g(T(\boldsymbol{x}); \vartheta) h(\boldsymbol{x}) = g(r^{-1}(T^*(\boldsymbol{x}); \vartheta)) h(\boldsymbol{x}).$$

Defining $g^*(t;\vartheta) = g(r^{-1}(t);\vartheta)$, we see that

$$\mathcal{L}(\vartheta;\boldsymbol{x}) = g^*(T^*(\boldsymbol{x});\vartheta)h(\boldsymbol{x}).$$

So, by Theorem 4.4, $T^*(\boldsymbol{X})$ is a sufficient statistic for $\vartheta$.

Because of the numerous sufficient statistics in a problem, we might ask whether one sufficient statistic is any better than another. Recall that the purpose of a sufficient statistic is to achieve data reduction without loss of information about the parameter $\vartheta$; thus, a statistic that achieves the most data reduction while still retaining all the information about $\vartheta$ might be considered preferable.

**Definition 4.8.** *A sufficient statistic $T(\boldsymbol{X})$ is called a* ***minimal sufficient statistic*** *if, for any other sufficient statistic $T'(\boldsymbol{X})$, $T(\boldsymbol{x})$ is a function of $T'(\boldsymbol{x})$.*

To say that $T(\boldsymbol{x})$ *is a function* of $T'(\boldsymbol{x})$ simply means that if $T'(\boldsymbol{x}) = T'(\boldsymbol{y})$, then $T(\boldsymbol{x}) = T(\boldsymbol{y})$. In terms of the partition sets, if $\{B_{t'} \,|\, t' \in \mathcal{T}'\}$ are partition sets for $T'(\boldsymbol{x})$ and $\{A_t \,|\, t \in \mathcal{T}\}$ are the partition sets for $T(\boldsymbol{x})$, then Definition 4.8 states that every $B_{t'}$ is a subset of some $A_t$. Thus, the partition associated with a minimal sufficient statistic is the "largest" possible partition for a sufficient statistic and a minimal sufficient statistic achieves the greatest possible data reduction for a sufficient statistic.

## 4.3 18/10/18

Using Definition 4.8 to find a minimal sufficient statistic is impractical, as was using Definition 4.1 to find sufficient statistics. We would need to guess that $T(\boldsymbol{X})$ was a minimal sufficient statistic and then verify the condition in the definition. Fortunately, the following result gives an easier way to find a minimal sufficient statistic.

**Theorem 4.9 (Lehmann-Scheffé).** *Let $f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)$ be the probability density function (or probability mass function) of a random sample $\boldsymbol{X}$. Suppose there exist a function $T(\boldsymbol{x})$ such that, for two sample points $\boldsymbol{x}$ and $\boldsymbol{y}$, the ratio*

$$\frac{f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)}{f_{\boldsymbol{X}}(\boldsymbol{y};\vartheta)}$$

*is constant as a function of $\vartheta$ if and only if $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Then, $T(\boldsymbol{X})$ is a minimal sufficient statistic for $\vartheta$.*

*Proof.* In order to simplify the proof, we assume $f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) > 0$ for every $\boldsymbol{x} \in \mathcal{X}$.

First, we show that $T(\boldsymbol{X})$ is a sufficient statistic. Let

$$\mathcal{T} \coloneqq \{t \,|\, t = T(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}\}$$

be the image of $\mathcal{X}$ under $T(\cdot)$. Define the *partition sets induced by* $T(\cdot)$ as

$$\mathcal{A}_t = \{\boldsymbol{x} \,|\, T(\boldsymbol{x}) = t\}.$$

For each $\mathcal{A}_t$, choose and fix one element $\boldsymbol{x} \in \mathcal{A}_t$. For any $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{x}_{T(\boldsymbol{x})}$ is the fixed element that is in the same set, $\mathcal{A}_t$, as $\boldsymbol{x}$. Since $\boldsymbol{x}$ and $\boldsymbol{x}_{T(\boldsymbol{x})}$ are in the same set $\mathcal{A}_t$,

$$T(\boldsymbol{x}) = T\left(\boldsymbol{x}_{T(\boldsymbol{x})}\right)$$

and hence

$$\frac{f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)}{f_{\boldsymbol{X}}(\boldsymbol{x}_{T(\boldsymbol{x})};\vartheta)}$$

is a constant as a function of $\vartheta$ by assumption. Thus, we can define a function on $\mathcal{X}$ by

$$h(\boldsymbol{x}) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)}{f_{\boldsymbol{X}}(\boldsymbol{x}_{T(\boldsymbol{x})};\vartheta)}$$

and observe $h$ does not depend on $\vartheta$. Now, define a function on $\mathcal{T}$ by

$$g(t;\vartheta) = f_{\boldsymbol{X}}(\boldsymbol{x}_t;\vartheta).$$

Then, it can be seen that

$$f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) = \frac{f_{\boldsymbol{X}}(\boldsymbol{x}_{T(\boldsymbol{x})};\vartheta) f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)}{f_{\boldsymbol{X}}(\boldsymbol{x}_{T(x)};\vartheta)} = g(T(\boldsymbol{x});\vartheta)h(\boldsymbol{x})$$

and, by Theorem 4.4, $T(\boldsymbol{X})$ is a sufficient statistic for $\vartheta$.

To show that $T(\boldsymbol{X})$ is minimal, let $T'(\boldsymbol{X})$ be any other sufficient statistic. By Theorem 4.4, there exist functions $g'$ and $h'$ such that

$$f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) = g'(T'(\boldsymbol{x});\vartheta)h'(\boldsymbol{x}).$$

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be any two sample points with $T'(\boldsymbol{x}) = T'(\boldsymbol{y})$. Then,

$$\frac{f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)}{f_{\boldsymbol{X}}(\boldsymbol{y};\vartheta)} = \frac{g'(T'(\boldsymbol{x});\vartheta)h'(\boldsymbol{x})}{g'(T'(\boldsymbol{y});\vartheta)h'(\boldsymbol{y})} = \frac{h'(\boldsymbol{x})}{h'(\boldsymbol{y})}.$$

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

*Remark 4.10.* The notion of *minimal sufficiency* is good, because it allows us to choose only statistics with the less number of block in the partition. However, this is still not enough, since there are many sufficient statistics and we need further properties to distinguish them.

### 4.3.1 Methods of Evaluating Estimators

The methods discussed in the previous section have outlined reasonable techniques for finding point estimators of parameters. A difficulty that arises, however, is that since we can usually apply more than one of these methods in a particular situation, we are often faced with the task of choosing between estimators. Of course, it is possible that different methods of finding estimators will yield the same answer, which makes evaluation a bit easier, but, in many cases, different methods will lead to different estimators.

**Definition 4.11.** *The **bias** of a point estimator $T_n$, of a parameter $\vartheta$, is the difference between the expected value of $T_n$ and $\vartheta$, that is*

$$b(T_n) \coloneqq \mathbb{E}[T_n] - \vartheta.$$

*An estimator $T_n$ whose bias is identically equal to zero is called **unbiased** for $\vartheta$ and satisfies $\mathbb{E}[T_n] = \vartheta$.*

*Remark 4.12.* This is a good property, because it is preferable for a statistic who is estimating $\vartheta$ to be centred in $\vartheta$. However, for completeness, we are looking for a measure of dispersion around the expected value, which is the variance of $T_n$. Such value measures the goodness of the estimator only in the case it is unbiased, since otherwise the variance would measure the dispersion around something that is not of interest. Indeed,

$$\operatorname{Var}(T_n) = \mathbb{E}\left[(T_n - \mathbb{E}[T_n])^2\right].$$

For a more general approach, consider Chebyshev's Inequality (taking the complementary)

$$\mathbb{P}(|T_n - \vartheta| < k) \geq 1 - \frac{\mathbb{E}\left[(T_n - \vartheta)^2\right]}{k^2},$$

which suggests to control the following quantity.

**Definition 4.13.** *The **Mean Squared Error** (*MSE*) of an estimator $T_n$ of a parameter $\vartheta$ is the function of $\vartheta$ described by*

$$\operatorname{MSE}(T_n) \coloneqq \mathbb{E}\left[(T_n - \vartheta)^2\right].$$

Notice that the Mean Squared Error measures the average squared difference between the estimator $T_n$ and the parameter $\vartheta$, a somewhat reasonable measure of performance for a point estimator. In general, any increasing function of the absolute distance $|T_n - \vartheta|$ would serve to measure the goodness of an estimator (*mean absolute error*, $\mathbb{E}\left[|T_n - \vartheta|\right]$, is a reasonable alternative), but the Mean Squared Error has at least two advantages over the distance measures. First, it is quite tractable analytically and, second, it has the interpretation

$$\begin{aligned}
\mathrm{MSE}(T_n) &= \mathbb{E}\left[(T_n - \vartheta)^2\right] \\
&= \mathbb{E}\left[(T_n - \mathbb{E}\left[T_n\right] + \mathbb{E}\left[T_n\right] - \vartheta)^2\right] \\
&= \mathbb{E}\left[(T_n - \mathbb{E}\left[T_n\right])^2\right] + \mathbb{E}\left[(T_n - \vartheta)^2\right] + 2\underbrace{\mathbb{E}\left[(T_n - \mathbb{E}\left[T_n\right])(\mathbb{E}\left[T_n\right] - \vartheta)\right]}_{\| \\ 0} \\
&= \mathrm{Var}\left(T_n\right) + \left(\mathbb{E}\left[T_n\right] - \vartheta\right)^2 \\
&= \mathrm{Var}\left(T_n\right) + \left(b(T_n)\right)^2
\end{aligned}$$

*Remark 4.14.* Thus, the Mean Squared Error incorporates two components, one measuring the variability of the estimator (*precision*), and the other measuring its bias (*accuracy*). An estimator that has good Mean Squared Error properties has small combined variance and bias. To find one, we need to find estimator that control both variance and bias. Clearly, unbiased estimators do a good job of controlling bias. For an unbiased estimator we have

$$\mathbb{E}\left[(T_n - \vartheta)^2\right] = \mathrm{Var}\left(T_n\right),$$

and so, if an estimator is unbiased, its Mean Squared Error is equal to its variance.

Although many unbiased estimators are also reasonable from the standpoint of Mean Squared Error, be aware that controlling bias does not guarantee that Mean Squared Error is controlled. In particular, it is sometimes the case that a trade-off occurs between variance and bias in such a way that a small increase in bias can be traded for a larger degrease in variance, resulting in an improvement in Mean Squared Error.

**Definition 4.15.** *Consider two estimators $T_n'$ and $T_n''$ for $\vartheta$. We say that $T_n'$ is* ***more efficient*** *than $T_n''$ if* $\mathrm{MSE}(T_n') < \mathrm{MSE}(T_n'')$.

# 5

# Week V

## 5.1 23/10/18

### 5.1.1 The Cramér-Rao Lower Bound

**Theorem 5.1 (Cramér- Rao).** *Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from a regular parametric model $X \sim f_X(x; \vartheta)$. Then, for any estimator $T_n$ of $\vartheta$, we have that*

$$\operatorname{Var}(T_n) \geq \frac{(1 + b'(T_n))^2}{I_n(\vartheta)}, \tag{5.1}$$

*where $b(T_n)$ is the bias of $T_n$ and $I_n$ is the Fisher information.*

*Proof.* The proof of this theorem is elegantly simple, and is a clever application of the Cauchy-Schwarz inequality, or stated statistically, the fact that any two random variables $X$ and $Y$,

$$\begin{aligned}(\operatorname{Cov}(X, Y))^2 &= (\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])])^2 \\ &\leq \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \mathbb{E}\left[(Y - \mathbb{E}[Y])^2\right] \\ &= \operatorname{Var}(X) \operatorname{Var}(Y).\end{aligned}$$

If we rearrange this, we can get a lower bound on the variance of $X$,

$$\operatorname{Var}(X) \geq \frac{(\operatorname{Cov}(X, Y))^2}{\operatorname{Var}(Y)}.$$

Observe that

$$\mathbb{E}[T_n] = \vartheta + b(T_n) \quad \text{and} \quad \frac{d}{d\vartheta}\mathbb{E}[T_n] = 1 + b'(T_n).$$

We know that, from Proposition 3.16,

$$\mathbb{E}[V_n'(\vartheta)] = 0,$$

therefore

$$\operatorname{Cov}(T_n, V_n'(\vartheta)) = \mathbb{E}[T_n V_n'(\vartheta)] - \mathbb{E}[T_n]\,\mathbb{E}[V_n'(\vartheta)] = \mathbb{E}[T_n V_n'(\vartheta)].$$

Continuing our evaluation, we obtain

$$\begin{aligned}
\mathbb{E}\left[T_n V_n'(\vartheta)\right] &= \int_{\mathbb{R}^n} T_n V_n'(\vartheta) f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) d\boldsymbol{x} \\
&= \int_{\mathbb{R}^n} T_n \frac{f_{\boldsymbol{X}}'(\boldsymbol{x};\vartheta)}{f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)} f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) d\boldsymbol{x} \\
&= \int_{\mathbb{R}^n} T_n \frac{d}{d\vartheta} f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) d\boldsymbol{x} \\
&= \frac{d}{d\vartheta} \int_{\mathbb{R}^n} T_n f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta) d\boldsymbol{x} \\
&= \frac{d}{d\vartheta} \mathbb{E}\left[T_n\right] \\
&= 1 + b'(T_n).
\end{aligned}$$

Since $\mathbb{E}\left[V_n'(\vartheta)\right] = 0$, it follows that

$$\mathrm{Var}\left(V_n'(\vartheta)\right) = \mathbb{E}\left[\left(V_n'(\vartheta)\right)^2\right] = -\mathbb{E}\left[V_n''(\vartheta)\right] = I_n(\vartheta).$$

Then, by Cauchy-Schwarz inequality, we conclude that

$$\mathrm{Var}\left(T_n\right) \geq \frac{\left(\mathrm{Cov}\left(T_n, V_n'(\vartheta)\right)\right)^2}{\mathrm{Var}\left(V_n'(\vartheta)\right)} = \frac{\left(1 + b'(T_n)\right)^2}{I_n(\vartheta)}.$$

□

*Remark 5.2.* Remember that, since

$$\mathrm{MSE}(T_n) = \mathrm{Var}\left(T_n\right) + b^2(T_n),$$

we obtain

$$\mathrm{MSE}(T_n) \geq \frac{\left(1 + b'(\vartheta)\right)^2}{I_n(\vartheta)} + b^2(T_n).$$

Also, the variance of an unbiased estimator $T_n(\vartheta)$ cannot be smaller than the Cramér-Rao lower bound (5.1). In general, however, we do not know if there exist an estimator whose variance is equal to the Cramér-Rao lower bound. The proper lower bound involves the Mean Squared Error

$$\mathrm{MSE}(T_n) = \frac{\left(1 + b'(\vartheta)\right)^2}{I_n(\vartheta)} + b^2(\vartheta).$$

If we add the assumption of independent samples, then the calculation of the lower bound is simplified. The expectation in the denominator becomes a univariate calculation, as the following corollary shows.

**Corollary 5.3.** *Consider a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from a regular parametric model $X \sim f_X(x;\vartheta)$. If there exist an unbiased estimator $T_n$ for $\vartheta$ whose variance is equal to the Cramér-Rao lower bound (5.1), then the estimator is unique.*

*Proof.* Consider two unbiased estimators $T_{1n}$ and $T_{2n}$ for $\vartheta$ such that

$$\mathrm{Var}\left(T_{1n}\right) = \mathrm{Var}\left(T_{2n}\right) = \frac{1}{I_n(\vartheta)} =: v.$$

Define another estimator $T_n$ as follows:

$$T_n \coloneqq \frac{T_{1n} + T_{2n}}{2}.$$

This is unbiased, since

$$\mathbb{E}[T_n] = \mathbb{E}\left[\frac{T_{1n} + T_{2n}}{2}\right] = \frac{2\vartheta}{2} = \vartheta.$$

Now,

$$\begin{aligned}
\operatorname{Var}(T_n) &= \operatorname{Var}\left(\frac{T_{1n} + T_{2n}}{2}\right) \\
&= \frac{1}{4}\left[\operatorname{Var}(T_{1n}) + \operatorname{Var}(T_{2n}) + 2\operatorname{Cov}(T_{1n}, T_{2n})\right] \\
&= \frac{(\operatorname{Var}(T_{1n})\operatorname{Var}(T_{2n}))^{1/2}}{(\operatorname{Var}(T_{1n})\operatorname{Var}(T_{2n}))^{1/2}}\frac{1}{4}\left[\operatorname{Var}(T_{1n}) + \operatorname{Var}(T_{2n}) + 2\operatorname{Cov}(T_{1n}, T_{2n})\right] \\
&= \frac{1}{2}\left(v + \frac{\operatorname{Cov}(T_{1n}, T_{2n})}{(\operatorname{Var}(T_{1n})\operatorname{Var}(T_{2n}))^{1/2}}v\right) \\
&= \frac{1}{2}(1+\rho)v,
\end{aligned}$$

where $\rho$ is the correlation coefficient. Now, if $\rho < 1$, we get that

$$\operatorname{Var}(T_n) = \frac{1}{2}(1+\rho)v$$

is less than the Cramér-Rao lower bound (5.1), which is impossible. Therefore, $\rho = 1$, meaning

$$\operatorname{Cor}(T_{1n}, T_{2n}) = 1 \quad T_{2n} = a + bT_{1n},$$

from which we obtain

$$\mathbb{E}[T_{2n}] = \mathbb{E}[a + bT_{1n}] \implies \vartheta = a + b\vartheta \Leftrightarrow a = 0 \wedge b = 1.$$

This implies that $T_{1n} = T_{2n}$. Therefore, it exist a unique unbiased estimator whose variance is equal to the Cramér-Rao lower bound (5.1).

□

**Definition 5.4.** *Consider a regular model $X \sim f_X(x;\vartheta)$. We say that an estimator is **efficient** if its variance is equal to the Cramér-Rao lower bound (5.1). Moreover, we say that an unbiased efficient estimator $T_n$ for $\vartheta$ is **absolute efficient**, i.e.*

$$\operatorname{Var}(T_n) = \frac{1}{I_n(\vartheta)}.$$

*Finally, the **efficiency** of $T_n$ is defined as*

$$\operatorname{eff}(T_n) \coloneqq \frac{1}{\operatorname{Var}(T_n)\, I_n(\vartheta)} \in [0,1].$$

*Remark 5.5.* We introduced the absolute efficiency at the cost of assuming regularity for the parametric model.

**Proposition 5.6.** *Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from a regular parametric model $X \sim f_X(x;\vartheta)$. Let $T_n$ be an unbiased estimator for $\vartheta$. Then, $T_n$ is efficient for $\vartheta$ if and only if*

$$V_n'(\vartheta) = I_n(\vartheta)(T_n - \vartheta). \tag{5.2}$$

*Proof.* Cramér-Rao inequality (5.1) can be expressed as

$$(\operatorname{Cov}(T_n, V_n'(\vartheta)))^2 \leq \operatorname{Var}(T_n)\, I_n(\vartheta).$$

Suppose we have efficiency, then

$$\operatorname{Var}(T_n) = \frac{1}{I_n(\vartheta)}$$

if and only if $\operatorname{Cor}(T_n, V_n'(\vartheta)) = 1$, if and only if $T_n$ is proportional to $V_n'(\vartheta)$, meaning

$$V_n'(\vartheta) = aT_n + b. \tag{5.3}$$

Now,

$$\mathbb{E}[V_n'(\vartheta)] = a\mathbb{E}[T_n] + b \quad \Longrightarrow \quad 0 = a\vartheta + b$$

hence $b = -a\vartheta$. This means that

$$I_n(\vartheta) = \operatorname{Var}(V_n'(\vartheta)) = a^2 \operatorname{Var}(T_n) = \frac{a^2}{I_n(\vartheta)} \quad \Longrightarrow \quad a = I_n(\vartheta),$$

($a$ cannot be equal to $-I_n(\vartheta)$, because in this case we would have $\operatorname{Cor}(T_n, V_n'(\vartheta)) = -1$) therefore (5.3) becomes

$$V_n'(\vartheta) = I_n(\vartheta)\,(T_n - \vartheta)\,.$$

Conversely, imagine that the equality (5.2) holds: now,

$$V_n'(\vartheta) = I_n\,(T_n - \vartheta)$$
$$T_n = \frac{V_n'(\vartheta)}{I_n} + \vartheta$$

and we simply evaluate its variance, meaning

$$\operatorname{Var}(T_n) = \operatorname{Var}\left(\frac{V_n'(\vartheta)}{I_n} - \vartheta\right) = \frac{1}{I_n^2}\operatorname{Var}(V_n'(\vartheta)) = \frac{1}{I_n}.$$

Therefore, $T_n$ is (absolutely) efficient by definition.

□

**Theorem 5.7 (Rao-Blackwell).** *Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from a parametric model $X \sim f_X(x;\vartheta)$. Let $T_{1n}$ be a sufficient estimator for $\vartheta$ and $T_{2n}$ be an unbiased estimator for $\vartheta$. Define*

$$T_n \coloneqq \mathbb{E}[T_{2n} \mid T_{1n}]\,.$$

*Then,*

1. *$T_n$ is a function of $T_{1n}$;*
2. *$\mathbb{E}[T_n] = \vartheta$;*
3. *$\operatorname{Var}(T_n) < \operatorname{Var}(T_{2n})$.*

### 5.1.2 Convergence

This section treats the somewhat fanciful idea of allowing the sample size to approach infinity and investigates the behaviour of certain sample quantities as this happens. Although the notion of an infinite sample size is a theoretical artefact, it can often provide us with some useful approximations for the finite-sample case, since it usually happens that expressions become simplified in the limit.

We are mainly concerned with three types of convergence, and treat them in varying amount of detail. But first, let us recall some useful definitions.

**Definition 5.8.** *A sequence of random variables $X_1, X_2, \dots$ is said to* ***converge in distribution*** *to a random variable $X$ if*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

*at all points $x$ where $F_X(x)$ is continuous.*

**Definition 5.9.** *A sequence of random variables $X_1, X_2, \dots$ is said to* ***converge in probability*** *to a random variable $X$ if, for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| \geq \varepsilon\right) = 0,$$

*or equivalently*

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| < \varepsilon\right) = 1$$

**Definition 5.10.** *Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from $X \sim f_X(x;\vartheta)$. We say that $T_n$ is* ***asymptotically unbiased*** *for $\vartheta$ if*

$$\lim_{n\to\infty} \mathbb{E}\left[T_n\right] = \vartheta,$$

*or equivalently*

$$\lim_{n\to\infty} b(T_n) = 0.$$

*We say that an estimator $T_n$ is* ***consistent in mean-square*** *for $\vartheta$ if*

$$\lim_{n\to\infty} \mathrm{MSE}(T_n) = 0. \tag{5.4}$$

*Since $\mathrm{MSE}(T_n) = \mathrm{Var}\left(T_n\right) + b^2(T_n)$, (5.4) is equivalent to say that*

$$\lim_{n\to\infty} \mathrm{Var}\left(T_n\right) = 0 \quad \textit{and} \quad \lim_{n\to\infty} b^2(T_n) = 0.$$

*A consistent estimator (in mean-square) is also asymptotically unbiased.*

*Finally, we say that $T_n$ is* ***consistent in probability*** *for $\vartheta$ if, for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(|T_n - \vartheta| < \varepsilon\right) = 1, \quad \textit{or} \quad \lim_{n\to\infty} \mathbb{P}\left(|T_n - \vartheta| \geq \varepsilon\right) = 0.$$

*Remark 5.11.* The consistency in mean-square implies the consistency in probability:

$$\mathbb{P}\left(|T_n - \vartheta| < \varepsilon\right) \geq 1 - \frac{\mathrm{MSE}(T_n)}{\varepsilon^2}$$

$$\lim_{n\to\infty} \mathbb{P}\left(|T_n - \vartheta| < \varepsilon\right) \geq 1 - \lim_{n\to\infty} \frac{\mathrm{MSE}(T_n)}{\varepsilon^2}.$$

**Exercise 5.12.** Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from $X \sim \mathcal{N}(\mu, \sigma^2)$. For

1. the corrected sample variance

$$\hat{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2,$$

2. the uncorrelated sample variance

$$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2,$$

show that (1) is unbiased for $\operatorname{Var}(X) = \sigma^2$ and (2) in biased for $\operatorname{Var}(X) = \sigma^2$.

*Proof.* (1) First,

$$\begin{aligned}
\mathbb{E}\left[\hat{S}_n^2\right] &= \mathbb{E}\left[\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n-1} \sum_{i=1}^{n} X_i^2 - \frac{2\overline{X}_n}{n-1} \sum_{i=1}^{n} X_i + \frac{n}{n-1} \mathbb{E}\left[\overline{X}_n^2\right]\right] \\
&= \frac{1}{n-1} \sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] - \frac{2n}{n-1} \mathbb{E}\left[\overline{X}_n^2\right] + \frac{n}{n-1} \mathbb{E}\left[\overline{X}_n^2\right] \\
&= \frac{1}{n-1} \sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] - \frac{n}{n-1} \mathbb{E}\left[\overline{X}_n^2\right]. \qquad (5.5)
\end{aligned}$$

Now,

$$\begin{aligned}
\mathbb{E}\left[X_i^2\right] &= \operatorname{Var}(X_i) + (\mathbb{E}[X_i])^2 = \sigma^2 + \mu^2, \\
\mathbb{E}\left[\overline{X}_n^2\right] &= \operatorname{Var}\left(\overline{X}_n\right) + \left(\mathbb{E}\left[\overline{X}_n\right]\right)^2 \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \operatorname{Var}(X_i) + \mu^2 \\
&= \frac{\sigma^2}{n} + \mu^2,
\end{aligned}$$

therefore

$$\begin{aligned}
(5.5) &= \frac{1}{n-1} \sum_{i=1}^{n} (\sigma^2 + \mu^2) - \frac{n}{n-1} \left(\frac{\sigma^2}{n} + \mu^2\right) \\
&= \frac{n}{n-1} (\sigma^2 + \mu^2) - \frac{n}{n-1} \left(\frac{\sigma^2}{n} + \mu^2\right) \\
&= \frac{n}{n-1} \sigma^2 - \frac{\sigma^2}{n-1} \\
&= \sigma^2.
\end{aligned}$$

This shows that the corrected sample variance is an unbiased estimator for the variance of a model.

□

**Exercise 5.13.** Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from $X \sim f_X(x; \vartheta)$ such that $\mathbb{E}[X] = \vartheta$ and $\mathrm{Var}(X) = 1$. Check if the following estimators are unbiased:

1. $T_{n1} = \frac{1}{n}\sum_{i=1}^{n} X_i$;
2. $T_{n2} = \frac{2}{n(n+1)}\sum_{i=1}^{n} iX_i$;
3. $T_{n3} = \sum_{i=1}^{n}(-1)^{i+1}X_i$;
4. $T_{n4} = \frac{1}{2^n-1}\sum_{i=1}^{n}\binom{n}{i}X_i$.

*Proof.* (1) *Unbiased*:

$$\mathbb{E}[T_{1n}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{n\vartheta}{n} = \vartheta.$$

(2) *Unbiased*:

$$\mathbb{E}[T_{n2}] = \frac{2}{n(n+1)}\sum_{i=1}^{n} i\mathbb{E}[X_i] = \frac{2\vartheta}{n(n+1)}\sum_{i=1}^{n} i = \frac{2\vartheta}{n(n+1)}\frac{n(n+1)}{2} = \vartheta.$$

(3) *Unbiased, when $n$ is odd*:

$$\mathbb{E}[T_{n3}] = \sum_{i=1}^{n}\mathbb{E}\left[(-1)^{i+1}X_i\right] = \vartheta\sum_{i=1}^{n}(-1)^{i+1} = \vartheta$$

(4) Observe that

$$\mathbb{P}(X = i) = \binom{n}{i}\vartheta^i(1-\vartheta)^{n-i}$$

and

$$\sum_{i=0}^{n}\binom{n}{i}\vartheta^i(1-\vartheta)^{n-i} = 1.$$

If we take $\vartheta = 1/2$,

$$\left(\frac{1}{2}\right)^n\sum_{i=0}^{n}\binom{n}{i} = 1 \Longrightarrow \sum_{i=1}^{n}\binom{n}{i} = 2^n - 1.$$

Therefore, $T_{n4}$ is unbiased.

□

## 5.2 24/10/18

### 5.2.1 Maximum Likelihood Estimator

Let $X_1, \dots, X_n$ be an independent and identically distributed random sample from a population $X \sim f_X(x; \vartheta)$. As we have seen, the Likelihood function is given by

$$\mathcal{L}(\vartheta; \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i; \vartheta_1, \dots, \vartheta_n).$$

**Definition 5.14.** *For each sample point $\boldsymbol{X} = \boldsymbol{x}$, let $\hat{\vartheta}(\boldsymbol{x})$ be a parameter value at which $\mathcal{L}(\vartheta, \boldsymbol{x})$ attains its maximum as a function of $\vartheta$, with $\boldsymbol{x}$ fixed. A **maximum Likelihood estimator** of the parameter $\vartheta$, based on a sample $\boldsymbol{X}$ is*

$$\arg\max_{\vartheta \in \Theta} \log \mathcal{L}(\vartheta, \boldsymbol{x}). \tag{5.6}$$

This can be done in the following situations:

1. when the model is *regular*: this allows us to solve the Maximum Likelihood equation
$$V_n'(\vartheta) = 0; \tag{5.7}$$
2. when the parameter $\vartheta$ is a (positive) integer, take the ratio
$$\frac{f_{n+1}(x)}{f_n(x)}.$$

**Exercise 5.15.** A computer takes an exponential time to solve a task. Assume that each time is modelled as a negative exponential random variable $T$, meaning

$$X \sim f_T(t; \vartheta) = \vartheta e^{-\vartheta t} \mathbb{1}_{\mathbb{R}_+}(t).$$

Suppose we have $n$ independent tasks are assigned to the computer, and let $T_i \overset{d}{=} T$ be the times, for $i = 1, \dots, n$.

1. Show that
$$\overline{T}_n = \frac{1}{n} \sum_{i=1}^{n} T_i$$
is an unbiased estimator for $1/\vartheta$, where $\vartheta$ is the parameter of the negative exponential distribution.
2. Find the distribution of $\overline{T}_n$: can you say that $\tilde{T}_n = (\overline{T}_n)^{-1}$ is an unbiased estimator for $\vartheta$?
3. Find an unbiased estimator for $\vartheta$ and compute the corresponding Mean Squared Error.
4. Are these estimators consistent for $\vartheta$ and $1/\vartheta$?

*Proof.* (1) First of all, $\overline{T}_n$ is clearly unbiased, since

$$\mathbb{E}\left[\overline{T}_n\right] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} T_i\right] = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\vartheta} = \frac{1}{\vartheta}.$$

Also,

$$\text{Var}\left(\overline{T}_n\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left(T_i\right) = \frac{1}{n^2}\frac{n}{\vartheta^2} = \frac{1}{n\vartheta^2}.$$

(2) We know that $\sum_{i=1}^{n} T_i \sim \Gamma(n,\vartheta)$. Define $X \coloneqq \sum_{i=1}^{n} T_i$: therefore, our goal is to find the distribution of $\frac{X}{n}$:

$$F_{X/n}(t) = \mathbb{P}\left(\frac{X}{n} \leq t\right) = \mathbb{P}\left(X \leq nt\right) = F_X(nt),$$

hence,

$$f_{X/n}(t) = nf_X(nt) \quad \Longrightarrow \quad \frac{T_i}{n} \sim \mathcal{E}(n\vartheta) = \Gamma(1,n\vartheta).$$

Therefore,

$$\frac{1}{n}\sum_{i=1}^{n} T_i = \sum_{i=1}^{n}\frac{T_i}{n} \sim \Gamma(n,n\vartheta).$$

Now,

$$\begin{aligned}
\mathbb{E}\left[\tilde{T}_n\right] &= \mathbb{E}\left[\frac{n}{\sum_{i=1}^{n} T_i}\right] \\
&= \int_0^\infty \frac{n}{t}\frac{\vartheta^n}{\Gamma(n)}t^{n-1}e^{-\vartheta t}dt \\
&= \frac{\vartheta^n n}{\Gamma(n)}\int_0^\infty t^{n-2}e^{-\vartheta t}dt \\
&= \frac{\vartheta^n n}{\Gamma(n)}\frac{\Gamma(n-1)}{\vartheta^{n-1}}\underbrace{\int_0^\infty \frac{\vartheta^{n-1}}{\Gamma(n-1)}t^{n-2}e^{-\vartheta t}dt}_{\| \\ 1} \\
&= \frac{\vartheta^n n\Gamma(n-1)}{\Gamma(n)\vartheta^{n-1}} \\
&= \frac{\vartheta n}{n-1},
\end{aligned}$$

therefore we conclude that it is indeed biased, but asymptotically unbiased.
(3) We want to find an unbiased estimator for $\vartheta$; $S_n$ must be unbiased for $\vartheta$. Indeed,

$$\mathbb{E}\left[S_n\right] = \mathbb{E}\left[\frac{n-1}{n}\hat{S}_n\right] = \frac{n-1}{n}\mathbb{E}\left[\tilde{T}_n\right] = \frac{n-1}{n}\frac{n\vartheta}{n-1} = \vartheta,$$

therefore it is unbiased. The corresponding Mean Squared Error is

$$\begin{aligned}
\text{MSE}(S_n) &= \text{Var}\left(S_n\right) + b^2(S_n) \\
&= \text{Var}\left(S_n\right) \\
&= \mathbb{E}\left[S_n^2\right] - \vartheta^2 \\
&= \mathbb{E}\left[\left(\frac{n-1}{\sum_{i=1}^{n} T_i}\right)^2\right] - \vartheta^2
\end{aligned}$$

$$
\begin{aligned}
&= (n-1)^2 \int_0^{+\infty} \frac{1}{t^2} \frac{\vartheta^n}{\Gamma(n)} t^{n-3} e^{-\vartheta t} dt - \vartheta^2 \\
&= \frac{(n-1)^2 \vartheta^n}{\Gamma(n)} \frac{\Gamma(n-2)}{\vartheta^{n-2}} - \vartheta^2 \\
&= \vartheta^2 (n-1)^2 \frac{\Gamma(n-2)}{\Gamma(n)} - \vartheta^2 \\
&= \vartheta^2 (n-1)^2 \frac{(n-3)!}{(n-1)(n-2)(n-3)!} - \vartheta^2 \\
&= \frac{\vartheta^2}{n-2}.
\end{aligned}
$$

(4) Is

$$\overline{T}_n = \frac{1}{n} \sum_{i=1}^{n} T_i$$

consistent for $1/\vartheta$? Yes, by the Weak Law of Large Numbers, since

$$\overline{T}_n \longrightarrow \frac{1}{\vartheta}.$$

Is $S_n$ consistent for $\vartheta$? Also, yes. Indeed

$$S_n = \frac{n-1}{\sum_{i=1}^n T_i} = \frac{(n-1)/n}{\left(\sum_{i=1}^n T_i\right)/n} \xrightarrow{n\to\infty} \frac{1}{\vartheta^{-1}} = \vartheta.$$

□

**Exercise 5.16.** Consider a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from $X \sim f_X(x;\vartheta)$ such that

$$f_X(x;\vartheta) = \vartheta x^{\vartheta-1} \mathbb{1}_{(0,1)}(x).$$

1. Find a sufficient statistic and a Maximum Likelihood estimator for $\vartheta$.
2. Is the Maximum Likelihood estimator unbiased for $\vartheta$? Compute the Mean Squared Error.
3. Is the Maximum Likelihood estimator consistent for $\vartheta$?
4. Propose another estimator for $\vartheta$.

*Proof.* (1) We have that

$$
\begin{aligned}
\mathcal{L}(\vartheta; x) &= \prod_{i=1}^{n} \vartheta x_i^{\vartheta-1} \\
&= \vartheta^n \left( \prod_{i=1}^{n} x_i \right)^{\vartheta-1} \\
&= \underbrace{\vartheta^n \left( \prod_{i=1}^{n} x_i \right)^{\vartheta}}_{\| \; g(T(\boldsymbol{x}),\vartheta)} \cdot \underbrace{\left( \prod_{i=1}^{n} x_i \right)^{-1}}_{\| \; h(\boldsymbol{x})},
\end{aligned}
$$

and

$$\begin{aligned}\log \mathcal{L}(\vartheta; x) &= n \log \vartheta + (\vartheta - 1) \log \left( \prod_{i=1}^{n} x_i \right) \\ &= n \log \vartheta + (\vartheta - 1) \left( \sum_{i=1}^{n} \log x_i \right). \end{aligned} \tag{5.8}$$

Therefore, thanks to the Factorisation Theorem, $\prod_{i=1}^{n} X_i$ is a sufficient statistic for $\vartheta$, whilst the Maximum Likelihood estimator is obtained deriving (5.8):

$$0 = \frac{n}{\vartheta} + \sum_{i=1}^{n} \log x_i \quad \Longrightarrow \quad \hat{\vartheta}_n = -\frac{n}{\sum_{i=1}^{n} \log x_i}.$$

(2) First of all, we compute the distribution of $Y \coloneqq -\log X$:

$$\begin{aligned}\mathbb{P}(Y \leq y) &= \mathbb{P}(-\log X \leq y) \\ &= \mathbb{P}\left(X \geq e^{-y}\right) \\ &= 1 - \mathbb{P}\left(X < e^{-y}\right) \\ &= 1 - e^{-y\vartheta},\end{aligned}$$

therefore $Y \sim \mathcal{E}_n(\vartheta)$, and consequently $-\sum_{i=1}^{n} \log x_i \sim \Gamma(n, \vartheta)$. This implies that

$$\mathbb{E}\left[\hat{\vartheta}_n\right] = \int_0^\infty \frac{n}{t} \frac{\vartheta^n}{\Gamma(n)} t^{n-1} e^{-\vartheta t} dt = \frac{n\vartheta}{n-1}$$

is biased for $\vartheta$, but asymptotically unbiased.
(3) Since

$$\hat{\vartheta}_n = -\frac{n}{\sum_{i=1}^{n} \log x_i},$$

by the Weak Law of Large Numbers

$$-\frac{1}{n} \sum_{i=1}^{n} \log x_i \xrightarrow{n \to \infty} \frac{1}{\hat{\vartheta}_n}$$

and

$$\frac{1}{\hat{\vartheta}_n} \xrightarrow{n \to \infty} \vartheta.$$

This implies that the Maximum Likelihood estimator is (weakly) consistent.
(4) We could use the Method of Moments. We know that the theoretical moment of order 1 of a parametric model is

$$\mathbb{E}[X] = \int_0^1 x \vartheta x^{\vartheta - 1} dx = \frac{\vartheta}{\vartheta + 1}.$$

We compare the theoretical moment of order 1 with the empirical model of the same order, which is the sample mean:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\vartheta}{\vartheta+1} \quad \Longrightarrow \quad \overline{\vartheta}_n = \frac{\overline{X}_n}{1-\overline{X}_n}.$$

Therefore,

$$\overline{\vartheta}_n = \frac{\overline{X}_n}{1-\overline{X}_n} \longrightarrow \frac{\vartheta/(\vartheta+1)}{1-\vartheta/(\vartheta+1)} = \vartheta.$$

□

## 5.3 25/10/18

When sampling from a population described by a probability density function (or probability mass function) $f_X(x;\vartheta)$, knowledge of $\vartheta$ yields knowledge of the entire population. Hence, it is natural to seek a method of finding a good estimator of the point $\vartheta$, that is, a good *point estimator*. It is also the case that the parameter $\vartheta$ has a meaningful physical interpretation (as in the case of a population mean) so there is the direct interest in obtaining a good point estimate of $\vartheta$. It may also be the case that some function of $\vartheta$ is of interest.

**Definition 5.17.** *A **point estimator** is any function of a sample. That is, any statistic is a point estimator.*

Notice that the definition makes no mention of any correspondence between the estimator and the parameter it is to estimate. While it might be argued that such statement should be included in the definition, such a statement would restrict the available set of estimators. Also, there is no mention in the definition of the range of the statistic.

There is one distinction that must be made clear, the difference between an estimate and an estimator. An *estimator* is a function of the sample, while an *estimate* is the realised value of an estimator (that is, a number) that is obtained when a sample is actually taken. Notationally, when a sample is taken, an estimator is a function of the random variables $X_1, \ldots, X_n$, while an estimate is a function of the realised values $x_1, \ldots, x_n$.

In many cases, there will be an obvious or natural candidate for a point estimator of a particular parameter. For example, the sample mean is a natural candidate for a point estimator of the population mean. However, when we leave a simple case like this, intuition may not only desert us, it may also lead us astray. Therefore, it is useful to have some techniques that will at least give us a some reasonable candidates for consideration.

### 5.3.1 Method of Moments

The Method of Moments is, perhaps, the oldest method of finding point estimators. It has the virtue of being quite simple to use and almost always yields some sort of estimate. In many cases, unfortunately, this method yields estimators that may be improved upon. However, it is a good place to start when other methods prove intractable.

**Definition 5.18.** *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a parametric model $X \sim f_X(x;\vartheta)$. Assume that $\vartheta \in \Theta$ such that $\dim \Theta = n$. Method of Moments estimators are found by equating the first $k$ sample moments to the corresponding $k$ population moments, and solving the resulting system of simultaneous equations. More precisely, define*

$$m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j, \quad \mu_j = \mathbb{E}\left[X^j\right], \quad j = 1, \ldots, k$$

*The population moment $\mu_j$ will typically be a function of $\vartheta_1, \dots, \vartheta_k$. The* ***Method of Moments*** *estimator $(\tilde{\vartheta}_1, \dots, \tilde{\vartheta}_k)$ of $(\vartheta_1, \dots, \vartheta_k)$ is obtained by solving the following system of equations of $(\vartheta_1, \dots, \vartheta_k)$ in terms of $(m_1, \dots, m_k)$:*

$$\begin{cases} m_1 = \mu_1(\vartheta_1, \dots, \vartheta_k) \\ \vdots \\ m_k = \mu_k(\vartheta_1, \dots, \vartheta_k) \end{cases}. \tag{5.9}$$

*Example 5.19.* Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from $X \sim \mathcal{N}(\mu, \sigma^2)$. In the preceding notation, $\vartheta_1 = \mu$ and $\vartheta_2 = \sigma^2$. We have that

$$m_1 = \overline{X}_n, \quad m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad \mu_1 = \mu, \quad \mu_2 = \mu^2 + \sigma^2$$

and hence we must solve

$$\overline{X}_n = \mu, \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mu^2 + \sigma^2.$$

Solving for $\mu$ and $\sigma^2$ yields the Method of Moments estimators

$$\tilde{\mu} = \overline{X}_n, \quad \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

In this simple example, the Method of Moments solutions coincides with our intuition, and perhaps give some credence to both. The method is somewhat more helpful, however, when no obvious estimator suggests itself.

**Exercise 5.20.** We want to estimate the proportion $\vartheta$ of individuals in a population, for which a certain feature $X$ takes value in a set $\mathcal{A}$. We take a sample of the population of size $n$ and we measure the feature $X$.

Let $Z_n$ be the number of individuals with feature $X$ in $\mathcal{A}$.

1. Is $\frac{Z_n}{n}$ unbiased for $\vartheta$? Is it consistent? Is it asymptotically Gaussian?
2. Find a Maximum Likelihood estimator for $\vartheta$.

*Proof.* (1) Denote

$$Y_i = \begin{cases} 1 & \text{if } X \text{ takes values in } \mathcal{A} \text{ for an individual } i \\ 0 & \text{otherwise} \end{cases}.$$

Obviously, $Y_i \sim Ber(\vartheta)$. Therefore

$$Z_n = \sum_{i=1}^{n} Y_i, \qquad \frac{Z_n}{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

and

$$\mathbb{E}\left[\frac{Z_n}{n}\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[Y_i\right] = \vartheta,$$

meaning that it is unbiased.

Also, it is consistent, since

$$\frac{Z_n}{n} \longrightarrow \vartheta, \quad n \to \infty,$$

and asymptotically Gaussian: indeed

$$\sqrt{n}\left(\frac{Z_n}{n} - \vartheta\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \vartheta\right) = \sqrt{n}(\overline{Y}_n - \vartheta) = \frac{\overline{Y}_n - \vartheta}{\frac{\sigma}{\sqrt{n}}}\sigma \longrightarrow \mathcal{N}(0, \vartheta(1-\vartheta)).$$

(2) Take $Y_1, \dots, Y_n$ from $Y \sim Ber(\vartheta)$ and consider

$$\begin{aligned}\log \mathcal{L}(\vartheta; \boldsymbol{x}) &= \log\left(\prod_{i=1}^{n} \vartheta^{x_i}(1-\vartheta)^{n-x_i}\right) \\ &= \log\left(\vartheta^{\sum_{i=1}^n x_i}(1-\vartheta)^{n-\sum_{i=1}^n x_i}\right) \\ &= \sum_{i=1}^{n} x_i \log\vartheta + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-\vartheta),\end{aligned}$$

so that

$$\frac{d}{d\vartheta}\log \mathcal{L}(\vartheta; \boldsymbol{x}) = \frac{\sum_{i=1}^n x_i}{\vartheta} - \frac{n - \sum_{i=1}^n x_i}{1-\vartheta} = 0,$$

and therefore

$$\hat{\vartheta} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

□

**Exercise 5.21.** Consider a population such that only 40% of individuals survive after one week. After one week we check the population and we find $k$ individuals. We want to estimate the size of the initial population.

*Proof.* The direct way to compute the number of survivals after one week knowing the parameter $\vartheta$ is represented by a binomial distribution

$$\mathbb{P}(K = k) = \binom{N}{k}\vartheta^k(1-\vartheta)^{N-k} = \mathcal{L}(N; k, \vartheta).$$

To find an estimate for $N$ consider

$$\begin{aligned}\frac{\mathcal{L}(N+1; k, \vartheta)}{\mathcal{L}(N; k, \vartheta)} &= \binom{N+1}{k}\binom{N}{k}^{-1}\frac{\vartheta^k(1-\vartheta)^{N-k+1}}{\vartheta^k(1-\vartheta)^{N-k}} \\ &= \frac{(N+1)!}{(N+1-k)!k!}\frac{k!(N-k)!}{N!}(1-\vartheta) \\ &= \frac{N+1}{N+1-k}(1-\vartheta) \qquad (5.10)\end{aligned}$$

The ratio (5.10) is greater or equal than one for $k = N, N+1$ and until $N \leq \frac{k}{\vartheta} - 1$. It is less or equal than one for any $N \geq \frac{k}{\vartheta} - 1$. Then we can take as an estimator of $N$ smallest integer larger than $\frac{k}{\vartheta} - 1$.

□

**Exercise 5.22.** Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from $X \sim f_X(x;\vartheta)$, where

$$f_X(x;\vartheta) = 2\vartheta x e^{-\vartheta x^2} \mathbb{1}_{\mathbb{R}_+}(x), \quad \vartheta > 0.$$

1. Find the Maximum Likelihood estimator for $\vartheta$.
2. Find a sufficient statistic for $\vartheta$.
3. Find the Moment estimator for $\vartheta$.
4. Compare the two estimators.

*Proof.* (1) The Likelihood Function is

$$\mathcal{L}(\vartheta;\boldsymbol{x}) = \prod_{i=1}^{n} 2\vartheta x_i e^{-\vartheta x_i^2} = (2\vartheta)^n \prod_{i=1}^{n} x_i \exp\left\{-\vartheta \sum_{i=1}^{n} x_i^2\right\},$$

and if we take the log

$$\begin{aligned}\mathcal{L}(\vartheta;\boldsymbol{x}) &= n\log 2\vartheta + \sum_{i=1}^{n} \log x_i - \vartheta \sum_{i=1}^{n} x_i^2 \\ &\propto n\log\vartheta - \vartheta \sum_{i=1}^{n} x_i^2.\end{aligned}$$

Now,

$$\frac{d}{d\vartheta}\log\mathcal{L}(\vartheta;\boldsymbol{x}) = n\frac{1}{\vartheta} - \sum_{i=1}^{n} x_i^2 = 0 \quad \Longleftrightarrow \quad \hat{\vartheta}_n = \frac{n}{\sum_{i=1}^{n} x_i^2}.$$

(2) We can apply the Fisher Factorisation Theorem:

$$\mathcal{L}(\vartheta;\boldsymbol{x}) = \underbrace{\prod_{i=1}^{n} x_i}_{\substack{\| \\ h(\boldsymbol{x})}} \cdot \underbrace{(2\vartheta)^n \exp\left\{-\vartheta \sum_{i=1}^{n} x_i^2\right\}}_{\substack{\| \\ g(T(\boldsymbol{x}),\vartheta)}}$$

and conclude that $\sum_{i=1}^{n} X_i^2$ is sufficient for $\vartheta$.
(3) For the Moment based estimator, remember that

$$m_1 = \mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Also,

$$\begin{aligned}\mu_1 &= \mathbb{E}[X] \\ &= \int_{\mathbb{R}} x f_X(x) dx \\ &= \int_0^{+\infty} 2\vartheta x^2 \exp\left\{-\vartheta x^2\right\} dx \\ &= \int_0^{+\infty} y^{\frac{1}{2}} \vartheta^{-\frac{1}{2}} e^{-y} dy, \qquad y = \vartheta x^2\end{aligned}$$

$$
\begin{aligned}
&= \vartheta^{-\frac{1}{2}} \int_0^{+\infty} y^{\frac{3}{2}-1} e^{-y} dy, \qquad \Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx \\
&= \vartheta^{-\frac{1}{2}} \Gamma\left(\frac{3}{2}\right) \\
&= \vartheta^{-\frac{1}{2}} \Gamma\left(1+\frac{1}{2}\right), \qquad \Gamma(\alpha+1) = \alpha\Gamma(\alpha) \\
&= \frac{\vartheta^{-\frac{1}{2}}}{2} \Gamma\left(\frac{1}{2}\right) \\
&= \frac{\sqrt{\pi}}{2} \vartheta^{-\frac{1}{2}},
\end{aligned}
$$

therefore

$$
\frac{\sqrt{\pi}}{2} \vartheta^{-\frac{1}{2}} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \Longrightarrow \quad \hat{\vartheta}_n = \frac{n^2 \pi}{4 \left(\sum_{i=1}^{n} X_i\right)^2}
$$

and it is not a function of the sufficient statistic.

□

# 6

# Week VI

## 6.1 30/10/18

### 6.1.1 Exercises

**Exercise 6.1.** The connection of a user to a server is accepted with probability $\vartheta \in (0,1)$. Assume that the trials for connecting are independent and denote by $X_1, \dots, X_n$ the number of trials that are necessary to get the connection in $n$ different days. Set $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

1. Is $\overline{X}_n$ an unbiased estimator for $\vartheta^{-1}$?
2. Study the convergence of

$$\sqrt{n}\left(\overline{X}_n - \frac{1}{\vartheta}\right), \quad X_i \in \{1, 2, \dots\}.$$

*Proof.* First of all $(X_i) \sim Geom(\vartheta)$ if and only if

$$\mathbb{P}(X_i = k) = \vartheta(1-\vartheta)^{k-1}, \quad k = 1, 2, \dots.$$

Now,

$$\mathbb{E}[X_i] = \frac{1}{\vartheta},$$

and

$$\mathbb{E}\left[\overline{X}_n\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{\vartheta},$$

therefore the estimator is unbiased for $\vartheta^{-1}$.
(2) Note that

$$\sqrt{n}\left(\overline{X}_n - \frac{1}{\vartheta}\right) = \frac{\sum_{i=1}^{n} X_i - n/\vartheta}{\sqrt{\frac{1-\vartheta}{\vartheta^2}}\sqrt{n}} \cdot \sqrt{\frac{1-\vartheta}{\vartheta^2}} \xrightarrow{n\to\infty} N\left(0, \frac{1-\vartheta}{\vartheta^2}\right),$$

therefore the estimator is asymptotically Gaussian.

□

**Exercise 6.2.** Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from $X \sim f_X(x;\vartheta)$, where

$$f_X(x;\vartheta) = \frac{4x^2}{\vartheta^3\sqrt{\pi}} \exp\left\{-\frac{x^2}{\vartheta^2}\right\} \mathbb{1}_{\mathbb{R}_+}(x), \quad \vartheta > 0.$$

1. Find the Maximum Likelihood estimator for $\vartheta$.
2. Find a sufficient statistic for $\vartheta$.
3. Can you find an unbiased estimator for $\vartheta$?

*Proof.* (1) The Likelihood function is

$$\begin{aligned}\mathcal{L}(\vartheta;\boldsymbol{x}) &= \prod_{i=1}^{n} \frac{4x_i^2}{\vartheta^3\sqrt{\pi}} \exp\left\{-\frac{x_i^2}{\vartheta^2}\right\} \\ &= \left(\frac{4}{\vartheta^3\sqrt{\pi}}\right)^n \prod_{i=1}^{n} x_i^2 \exp\left\{-\sum_{i=1}^{n}\frac{x_i^2}{\vartheta^2}\right\},\end{aligned}$$

and

$$\log \mathcal{L}(\vartheta,\boldsymbol{x}) = n\log\left(\frac{4}{\vartheta^3\sqrt{\pi}}\right) + \sum_{i=1}^{n}\log x_i^2 - \frac{1}{\vartheta^2}\sum_{i=1}^{n} x_i^2.$$

Now,

$$\frac{d}{d\vartheta}\log \mathcal{L}(\vartheta;\boldsymbol{x}) = n\frac{\vartheta^3\sqrt{\pi}}{4}\frac{4}{\sqrt{\pi}}(-3)\frac{1}{\vartheta^4} + \frac{2}{\vartheta^3}\sum_{i=1}^{n} x_i^2.$$

Therefore,

$$\frac{3n}{\vartheta^2} = \frac{1}{\vartheta^3}\sum_{i=1}^{n} x_i^2 \quad \Longrightarrow \quad \hat{\vartheta}_n = \frac{1}{3n}\sum_{i=1}^{n} x_i^2$$

is a Maximum Likelihood estimator for $\vartheta$.
(2) $\sum_{i=1}^{n} X_i^2$ is a sufficient statistic for $\vartheta$. Indeed,

$$\mathcal{L}(\vartheta;\boldsymbol{x}) = \left(\frac{4}{\vartheta^3\sqrt{\pi}}\right)^n \exp\left\{\frac{1}{\vartheta}\sum_{i=1}^{n} x_i^2\right\} \prod_{i=1}^{n} x_i^2.$$

(3) To check unbiasedness, first compute

$$m_1 = \mu_1 \quad \Longrightarrow \quad \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}[X].$$

$$\mathbb{E}\left[\hat{\vartheta}_n\right] = \mathbb{E}\left[\frac{1}{3n}\sum_{i=1}^{n} X_i^2\right] = \frac{1}{3n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] = \frac{1}{3}\mathbb{E}\left[X^2\right].$$

The second moment of $X$ is

$$\begin{aligned}\mathbb{E}\left[X^2\right] &= \int_{\mathbb{R}} x^2 f_X(x)dx \\ &= \int_0^{+\infty} x^2 \frac{4x^2}{\vartheta^3\sqrt{\pi}} \exp\left\{-\frac{x^2}{\vartheta^2}\right\} dx\end{aligned}$$

$$
\begin{aligned}
&= \int_0^{+\infty} \frac{4x^4}{\sqrt{\pi}\vartheta^2} \exp\left\{-\frac{x^2}{\vartheta^2}\right\} dx \\
&= \int_0^{+\infty} \frac{2x}{\vartheta^2} \cdot \frac{2x^3}{\sqrt{\pi}\vartheta} \exp\left\{-\frac{x^2}{\vartheta^2}\right\} dx, \qquad t = \frac{x^2}{\vartheta^2} \\
&= \int_0^{+\infty} \frac{1}{\sqrt{\pi}} t^{\frac{3}{2}} \frac{\vartheta^3}{\vartheta} e^{-t} dt \\
&= \frac{2}{\sqrt{\pi}}\vartheta^2 \int_0^{+\infty} t^{\frac{5}{2}-1} e^{-t} dt \\
&= \frac{2}{\sqrt{\pi}}\vartheta^2 \Gamma\left(\frac{5}{2}\right) \\
&= \frac{2}{\sqrt{\pi}}\vartheta^2 \Gamma\left(1+\frac{3}{2}\right) \\
&= \frac{2}{\sqrt{\pi}}\vartheta^2 \Gamma\left(\frac{3}{2}\right)\frac{3}{2} \\
&= \frac{2}{\sqrt{\pi}}\vartheta^2 \frac{\sqrt{\pi}}{2}\frac{3}{2} \\
&= \frac{3}{2}\vartheta^2,
\end{aligned}
$$

therefore

$$\mathbb{E}\left[\hat{\vartheta}_n\right] = \frac{\vartheta^2}{2}.$$

□

**Exercise 6.3.** A server gets requests of jobs from users. Every job takes an exponential time with parameter $\lambda$, which is known. Assuming that the server has been active for a time $X = x$, find the Maximum Likelihood estimator of the number $n$ of requests.

*Proof.* Since we are dealing with sums of exponentials, we have that $X \sim \Gamma(n, \lambda)$. Easily, we can compute

$$\mathbb{E}[X] = \frac{n}{\lambda} \iff \hat{n} = \lfloor x\lambda \rfloor,$$

which could be an estimator. As we did last time, consider

$$f_n(x) = \frac{\lambda^n}{\Gamma(n)} x^{n-1} e^{-\lambda x}$$

and compute

$$\frac{f_{n+1}(x)}{f_n(x)} = \frac{\lambda x}{n}.$$

Observe that the mapping $n \mapsto \frac{f_{n+1}(x)}{f_n(x)}$ is increasing until $\lambda x \geq n$ and then decreasing. Therefore, we have indeed that

$$\hat{n} = \lceil x\lambda \rceil.$$

□

**Exercise 6.4.** For estimating the number of individuals in a population we can use the following method: we take a first sample of $k$ individuals, all distinct, and we mark them. After that, we insert them again in the population. We take another sample of size $k$ (all distinct). Let $N$ be the number of the population.

1. What is the probability that in the second group there are $r$ individuals from the first group?
2. Find an estimator for $N$.

*Proof.* (1) The distribution in question is clearly an Hypergeometric:

$$\mathbb{P}(X = r) = \frac{\binom{K}{r}\binom{N-k}{k-r}}{\binom{N}{k}}.$$

(2) For an estimator of $N$, consider

$$\begin{aligned}\mathbb{P}(X = r) &= g(N) \\ &= c \cdot \frac{(N-k)!}{\frac{N!}{(N-k)!}(N-2k+r)!} \\ &= c \cdot \frac{[(N-k)!]^2}{N!(N-2k+r)!}\end{aligned}$$

where $c$ is a constant that does not depend on $N$. Finally, we consider the the usual ratio

$$\frac{g(N+1)}{g(N)} = \ldots = \frac{k^2 - r}{r}.$$

□

**Exercise 6.5.** Consider a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from $X \sim U([a-b],\ [b+a])$.

1. Find the Moment based estimator of $a$ and $b$.
2. Find the Maximum Likelihood estimator of $a$ and $b$.
3. Compare the estimators.

*Proof.* (1) Compute

$$\begin{aligned}\mathbb{E}[X] &= \int_{\mathbb{R}} \frac{x}{a+b-(a-b)} dx \\ &= \frac{1}{2b}\frac{x^2}{2}\Big|_{a-b}^{a+b} \\ &= \frac{1}{4b}(a^2 + b^2 + 2ab - a^2 - b^2 + 2ab) \\ &= a, \\ \mathbb{E}\left[X^2\right] &= \int_{a-b}^{a+b} \frac{x^2}{2b} dx \\ &= \frac{x^3}{6b}\Big|_{a-b}^{a+b}\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{6b}(a^3 + b^3 + 3a^2b + 3ab^2 - (a^3 - b^3 - 3a^2b + 3ab^2)) \\
&= \frac{1}{6b}(2b^3 + 6ba^2b) \\
&= a^2 + \frac{b^2}{3}
\end{aligned}
$$

so that

$$
\mathbb{E}[X] = a = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \mathbb{E}\left[X^2\right] = a^2 + \frac{b^2}{3} = \frac{1}{n}\sum_{i=1}^{n} X_i^2.
$$

Therefore

$$
\overline{T}_n(a) = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \overline{T}_n(b) = \left[\frac{3}{n}\sum_{i=1}^{n} X_i^2 - 3\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right]^{\frac{1}{2}}.
$$

(2) For the Maximum Likelihood estimator, first compute

$$
\mathcal{L}(a, b; \boldsymbol{x}) = \left(\frac{1}{2b}\right)^n \prod_{i=1}^{n} \mathbb{1}_{[a-b,b+a]}(x_i), \quad x_i.
$$

We have that

$$
x_i \in [a-b, b+a] \iff a - b \leq \min_i x_i \wedge \max_i x_i \leq a + b. \tag{6.1}
$$

We want $b$ to be as small al possible, because it is at the denominator: indeed, we are looking for the smallest $b$ such that (6.1) is satisfied for some $a$:

$$
b \geq \frac{\max_i x_i - \min_i x_i}{2} \implies b = \frac{\max_i x_i - \min_i x_i}{2},
$$

where (6.1) becomes an equality if

$$
a = \frac{\max_i x_i + \min_i x_i}{2}.
$$

Therefore the Maximum Likelihood estimators for $a$ and $b$ are

$$
\begin{aligned}
\hat{T}_n(a) &= \frac{\max_i x_i + \min_i x_i}{2} \\
\hat{T}_n(b) &= \frac{\max_i x_i - \min_i x_i}{2}
\end{aligned}
$$

and

$$
\mathcal{L}(a, b; \boldsymbol{x}) = \left(\frac{1}{2b}\right)^n \mathbb{1}_A(x_i),
$$

where $A \coloneqq a - b \leq \min_i x_i \wedge \max_i x_i \leq a + b$.

□

## 6.2 31/10/18

### 6.2.1 Exercises

**Exercise 6.6.** Suppose you want to achieve knowledge about a private information. Instead of asking directly, a commonly used technique is to ask the sample to flip a coin. Each person would then check either 0 or 1. Then the interviewer may ask wether the answer they should have given to a direct question corresponds to the result of the coin flipped.

Consider the experiment of flipping a the coin:

$$\textit{YES} \text{ with probability } p, \quad \textit{NO} \text{ with probability } 1-p,$$

with $p$ known. We would like to estimate the proportion $\vartheta$ of population that have a feature $A$ without asking directly, but through this randomisation process.

*Proof.* For example, consider an individual $i$: he/she either has or does not have the feature $A$. Assume he/she has the feature $A$ and flips the coin. With probability $p$ he obtains heads. Therefore, the answer to the direct and to the randomised question coincide.

Now, define an indicator random variable for the two answers being the same:

$$X_i = \begin{cases} 1 & \text{if } \textit{YES} \\ 0 & \text{if } \textit{NO} \end{cases}.$$

Therefore,

$$\begin{aligned} \mathbb{P}(X_i = 1) &= \mathbb{P}(X_i = 1 \mid A)\,\mathbb{P}(A) + \mathbb{P}(X_i = 1 \mid \neg A)\,\mathbb{P}(\neg A) \\ &= p\vartheta + (1-p)(1-\vartheta) \\ &= p\vartheta + 1 - p - \vartheta + p\vartheta \\ &= (1-p) - \vartheta(1-2p). \end{aligned}$$

Notice that $\boldsymbol{X} = (X_1, \dots, X_n)$ is a random sample form $X \sim Ber((1-p)-\vartheta(1-2p))$. If we would like to estimate $\vartheta$ using the Likelihood function, then set

$$\hat{T} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

and therefore

$$\hat{\vartheta} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - (1-p)}{2p-1}, \quad p \neq \frac{1}{2}.$$

Observe that it is unbiased, since

$$\mathbb{E}\left[\hat{\vartheta}\right] = \frac{\frac{1}{n}n(\vartheta p + (1-p)(1-\vartheta)) - (1-p)}{2p-1} = \vartheta$$

it is consistent. Also, thanks to the Weak Law of Large Numbers,

$$\operatorname{Var}\left(\hat{\vartheta}\right) = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - (1-p)}{2p-1} \xrightarrow{n\to\infty} = \vartheta.$$

□

**Exercise 6.7.** Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$, where $X_i$ is the survival time of a lamp, meaning $X_i \sim \mathcal{E}_n(\vartheta)$. We already know that one possible unbiased estimator for $\vartheta$ is

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

Fix a time $T$ and denote by $Z$ the number of lamps survived at the time $T$.

Find the Maximum Likelihood estimator for $\vartheta$ based on the information about $Z$.

*Proof.* Being $Z \sim \mathcal{B}(n,p)$, the probability that $z$ lamps are still working at time $T$ is

$$\mathbb{P}(Z=z) = \binom{n}{z} p^z (1-p)^{n-z}, \quad p = \mathbb{P}(X_i \geq T) = e^{-\vartheta T},$$

where instead $p$ is the probability that a lamp is still working after time $T$, being $X_i \sim \mathcal{E}_n(p)$. Now, combining our knowledge of a Maximum Likelihood estimator for the binomial coefficient

$$\hat{p} = \frac{z}{n}$$

and

$$e^{-\vartheta T} = \frac{z}{n},$$

we have that

$$\hat{\vartheta} = -\frac{\log\left(\frac{z}{n}\right)}{T}.$$

We have faced problems of maximising the Log-Likelihood function with respect to $\vartheta$. This is particularly difficult when

- the parameter of interest is in $\mathbb{N}$;
- the parameter of interest defines the support of the model, for example the Maximum Likelihood estimator function of $\max\{a,b\}$ or $\min\{a,b\}$;
- Laplace model, where
$$f_X(x;\vartheta) \propto \exp\{-|x-\vartheta|\};$$
- when the model is not regular or the Likelihood equation is unsolvable. Suppose for example to have a regular model $\boldsymbol{X} = (X_1, \dots, X_n)$ from $X \sim f_X(x;a,b)$, meaning that we can differentiate with respect to both $a$ and $b$. Also, suppose that we cannot solve the Likelihood equation explicitly. For that matter, consider a $\Gamma(\alpha,\beta)$:

$$\mathcal{L}(\alpha,\beta;\boldsymbol{x}) = \prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{\beta x_i} = \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^n \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} \exp\left\{-\beta \sum_{i=1}^{n} x_i\right\}$$

and

$$\log \mathcal{L}(\alpha,\beta;\boldsymbol{x}) = n\alpha \log\beta - n\log\Gamma(\alpha) + (\alpha-1)\sum_{i=1}^{n} \log x_i - \beta \sum_{i=1}^{n} x_i.$$

Taking the derivatives with respect to $\alpha$,

$$n\log\beta - \frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n} \log x_i = 0$$

and $\beta$

$$n\alpha\frac{1}{\beta} - \sum_{i=1}^{n} x_i = 0.$$

These are numerically approachable thanks to the Newton-Rapson method.

# 7

# Week VII

## 7.1 06/11/18

### 7.1.1 Properties of the Maximum Likelihood Approach

As we have stated before, concerning the Maximum Likelihood estimator, we would like to solve the equation (5.6). If the aforementioned cases are not relevant for the treated situation, we must implement different approaches. For example,

1. when the parameter defines the support of the model;
2. when the model is *regular*, but we cannot solve explicitly the Maximum Likelihood equation
$$V_n'(\vartheta) = 0.$$
If we were in this case, we could follow two different approaches.
   a) The *substitution approach*, if we had some idea of the value we are looking for.
   b) ***Newton-Rapson Method***.
   Let $\mathcal{L}(\vartheta; \boldsymbol{x})$ be the Likelihood function, with $\vartheta \in \Theta$. First of all, consider the *gradient*
   $$g(\vartheta) = V_n'(\vartheta)$$
   and the *hessian*
   $$H(\vartheta) = V_n''(\vartheta).$$
   Then, fix a value $\vartheta_0 \in \Theta$ and consider the Taylor expansion for $g(\vartheta)$ around $\vartheta_0$,
   $$g(\vartheta) = g(\vartheta_0) + (\vartheta - \vartheta_0)H(\vartheta_0) + \dots.$$
   Next, replace $\vartheta$ with the Maximum Likelihood estimator $T_n$ for $\vartheta$:
   $$g(T_n) = g(\vartheta_0) + H(\theta_0)(T_n - \vartheta_0) + \dots.$$
   This must be null, since the Maximum Likelihood estimator maximises the Likelihood function. Now,
   $$T_n - \vartheta_0 = -\frac{g(\vartheta_0)}{H(\vartheta_0)} + \dots.$$

Therefore, we can write an approximate solution for the Maximum Likelihood estimator of $\vartheta$, meaning

$$T_n = \vartheta_0 - \frac{g(\vartheta_0)}{H(\vartheta_0)^{-1}} \quad \Longrightarrow \quad T_n = \vartheta_0 - \frac{V_n'(\vartheta)}{V_n''(\vartheta_0)^{-1}}.$$

Finally, write down the iteration

$$t^{(i)} = t^{(i-1)} - \frac{V_n'(t^{(i-1)})}{V_n''(t^{(i-1)})}, \quad \forall i \in \mathbb{N}_+.$$

Note that the iterative procedure naturally stops when $t^{(i)} = t_{(i+1)}$, that is when $V_n'(\vartheta) = 0$. However, there are to problems to solve. First, $t^{(0)}$ must be specified, and second $[V_n''(\cdot)]^{-1}$ must exist. How do we solve the existence of that inverse function? We can use a variation of the Newton-Rapson method, which is called **Score algorithm**. This consists in replacing $V_n''(t^{(i)})$ with $\mathbb{E}\left[V_n''(t^{(i)})\right]$, which is always positive definite.

*Example 7.1.* Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from $X$, where $X$ is distributed as a zero-truncated Poisson distribution with parameter $\vartheta$. Now, for the generic Poisson,

$$\mathbb{P}(Y = x) = \frac{e^{-\vartheta}\vartheta^x}{x!}\mathbb{1}_{\mathbb{N}}(x),$$

where $e^{-\vartheta}$ is the constant of normalisation. For the truncated Poisson, we would like to change the constant in order to "truncate" the Poisson at 0:

$$\mathbb{P}(X = x) = \frac{\vartheta^x e^{-\vartheta}}{x!(1 - e^{-\vartheta})}\mathbb{1}_{\mathbb{N}_+}(x),$$

since

$$\sum_{i=0}^{\infty} \frac{e^{-\vartheta}\vartheta^i}{i!} = 1 \quad \Longrightarrow \quad \sum_{i=1}^{\infty} \frac{e^{-\vartheta}\vartheta^i}{i!} = 1 - e^{-\vartheta}.$$

Now,

$$\mathcal{L}(\vartheta, x_i) = \prod_{i=1}^{n} \frac{\vartheta^{x_i} e^{-\vartheta}}{x_i!(1 - e^{-\vartheta})}\mathbb{1}_{\mathbb{N}_+}(x_i) = \frac{e^{-n\vartheta}\vartheta^{\sum_{i=1}^n x_i}}{(\prod_{i=1}^n x_i!)\,(1 - e^{-\vartheta})^n}$$

and

$$\log \mathcal{L}(\vartheta, x_i) \propto \log \vartheta \sum_{i=1}^{n} x_i - n\vartheta - n \log\left(1 - e^{-\vartheta}\right);$$

therefore

$$\frac{d}{d\vartheta} \log \mathcal{L}(\vartheta, x_i) = -n + \frac{\sum_{i=1}^n x_i}{\vartheta} - \frac{ne^{-\vartheta}}{1 - e^{-\vartheta}} = \frac{n\overline{X}_n}{\vartheta} - \frac{n}{1 - e^{\vartheta}}$$

and we have

$$\frac{n\overline{X}_n}{\vartheta} - \frac{n}{1 - e^{-\vartheta}} = 0 \Longrightarrow T_n = \overline{X}_n\left(1 - e^{-T_n}\right) \Longrightarrow t^{(i)} = \overline{X}_n\left(1 - e^{-t^{(i-1)}}\right),$$

for $i = 1, 2, \ldots$. If we wanted to apply the *Newton-Rapson* method, first we compute

$$V_n'(\vartheta) = \frac{n\overline{X}_n}{\vartheta} - \frac{n}{1-e^{-\vartheta}}, \quad V_n''(\vartheta) = -\frac{n\overline{X}_n}{\vartheta^2} + \frac{ne^{-\vartheta}}{(1-e^{-\vartheta})^2},$$

then simply calculate

$$t^{(i)} = t^{(i-1)} - \underbrace{\left[\frac{n\overline{X}_n}{t^{(i-1)}} - \frac{n}{1-e^{-t^{(i-1)}}}\right]}_{\| \; V_n'(t^{(i-1)})} \cdot \underbrace{\left[-\frac{n\overline{X}_n}{\left(t^{(i-1)}\right)^2} + \frac{ne^{-t^{(i-1)}}}{\left(1-e^{-t^{(i-1)}}\right)^2}\right]^{-1}}_{\| \; (V_n''(t^{(i-1)}))}$$

### 7.1.2 Delta Method (a generalisation of the Central Limit Theorem for functionals)

Consider $X_1, \ldots, X_n$ independent Bernoulli random variables, with parameter $p$. The typical parameter of interest is $p$, but sometimes we also want to look out for

$$\frac{p}{1-p}. \tag{7.1}$$

Since a valid estimator for $p$ is $\hat{p} = \frac{1}{n}\sum_{i=1}^n x_i$, one may consider

$$\frac{\hat{p}}{1-\hat{p}}$$

as an estimator for (7.1). What are the properties of this estimator? Can we say something about his variance? And what about its distribution? In order to answer to this questions, we could apply the *Delta method*.

Let $T_1, \ldots, T_n$ be random variables, with means $\vartheta_1, \ldots, \vartheta_n$ respectively, and consider

$$\boldsymbol{T} = (T_1, \ldots, T_n), \quad \boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_n).$$

Suppose there is a differentiable function $g(\boldsymbol{T})$ for which we want to approximate *mean* and *variance*. Define

$$g_i'(\boldsymbol{\vartheta}) \coloneqq \frac{d}{dt_i} g(\boldsymbol{t})\Big|_{t_1=\vartheta_1,\ldots,t_n=\vartheta_n}$$

and consider the first order Taylor expansion

$$g(\boldsymbol{t}) = g(\boldsymbol{\vartheta}) + \sum_{i=1}^n g_i'(\boldsymbol{\vartheta})(t_i - \vartheta_i) + R.$$

Therefore,

$$g(\boldsymbol{t}) \approx g(\boldsymbol{\vartheta}) + \sum_{i=1}^n g_i'(\boldsymbol{\vartheta})(t_i - \vartheta_i)$$

$$g(\boldsymbol{T}) \approx g(\boldsymbol{\vartheta}) + \sum_{i=1}^n g_i'(\boldsymbol{\vartheta})(T_i - \vartheta_i)$$

$$\mathbb{E}\left[g(\boldsymbol{T})\right] \approx g(\boldsymbol{\vartheta}) + \sum_{i=1}^{n} g_i'(\boldsymbol{\vartheta})\mathbb{E}\left[T_i - \vartheta_i\right].$$

By similar calculation, we can obtain an approximation of the variance of $g(\boldsymbol{T})$:

$$\operatorname{Var}\left(g(\boldsymbol{T})\right) \approx \sum_{i=1}^{n} (g_i'(\boldsymbol{\vartheta}))^2 \operatorname{Var}\left(T_i\right) + 2\sum_{i>j} g_i'(\boldsymbol{\vartheta}) g_j'(\boldsymbol{\vartheta}) \operatorname{Cov}\left(T_i, T_j\right).$$

Then, by using the Taylor approximation of the mean and variance of $g(\boldsymbol{T})$, we can state the following generalisation of the Central Limit Theorem, which is known as *Delta method*.

**Theorem 7.2 (Delta Method).** *Let $(Y_n)_{n\geq 1}$ be a sequence of random variables such that*

$$\sqrt{n}(Y_n - \vartheta) \xrightarrow{n\to\infty} \mathcal{N}(0, \sigma^2).$$

*For a given function $g(\cdot)$ and a specific value of $\vartheta$, assume that $g'(\vartheta)$ exists and it is not zero. Then,*

$$\sqrt{n}(g(Y_n) - g(\vartheta)) \xrightarrow{n\to\infty} \mathcal{N}(0, \sigma^2(g'(\vartheta))^2).$$

*Proof.* Consider the Taylor expansion of $g(Y_n)$ around $\vartheta$:

$$g(Y_n) = g(\vartheta) + g'(\vartheta)(Y_n - \vartheta) + o(Y_n - \vartheta).$$

Now,

$$\sqrt{n}\left(g(Y_n) - g(\vartheta)\right) = \sqrt{n}g'(\vartheta)(Y_n - \vartheta) + \sqrt{n}o(Y_n - \vartheta) \xrightarrow{n\to\infty} \mathcal{N}(0, \sigma^2(g'(\vartheta))^2),$$

since

$$\sqrt{n}o\left(Y_n - \vartheta\right) = \sqrt{n}\left(Y_n - \vartheta\right) \cdot \frac{o\left(Y_n - \vartheta\right)}{\left(Y_n - \vartheta\right)} \xrightarrow{n\to\infty} 0.$$

□

## 7.2 07/11/18

### 7.2.1 Maximum Likelihood Approach

A useful property of the Maximum Likelihood estimator is the *invariance property*. Suppose we have a parametric family indexed by a parameter $\vartheta$, but we are interested in estimating $\tau(\vartheta)$. The invariance property of Maximum Likelihood estimators says that, if $\hat{\vartheta}$ is the Maximum Likelihood estimator for $\vartheta$, then $\tau(\hat{\vartheta})$ is the Maximum Likelihood estimator for $\tau(\vartheta)$.

In general, the invariance property of the Maximum Likelihood estimator is true for any functional of $\vartheta$. To show this for an arbitrary function $\tau(\cdot)$, we define

$$\mathcal{L}^*(\eta, \boldsymbol{x}) := \sup_{\{\vartheta \in \Theta \mid \tau(\vartheta) = \eta\}} \mathcal{L}(\vartheta, \boldsymbol{x}). \tag{7.2}$$

The value $\hat{\eta}$ that maximises (7.2) is called the **Maximum Likelihood estimator** of $\eta = \tau(\vartheta)$, and is defined as

$$\mathcal{L}^*(\hat{\eta}; \boldsymbol{x}) = \sup_{\eta} \mathcal{L}^*(\eta; \boldsymbol{x}).$$

**Theorem 7.3 (Invariance Property of Maximum Likelihood estimators).** *If $\hat{\vartheta}$ is the Maximum Likelihood estimator for $\vartheta$, then, for every other function $\tau(\vartheta)$, the Maximum Likelihood estimator of $\tau(\vartheta)$ is $\tau(\hat{\vartheta})$.*

*Proof.* Let $\hat{\eta}$ be the value that maximises the function $\mathcal{L}^*(\eta; \boldsymbol{x})$. We must show that

$$\mathcal{L}^*\left(\tau(\hat{\vartheta}); \boldsymbol{x}\right) = \sup_{\eta} \mathcal{L}^*(\eta; \boldsymbol{x}).$$

Therefore, we have that

$$\begin{aligned}
\mathcal{L}^*(\eta; \boldsymbol{x}) &= \sup_{\eta} \left( \sup_{\{\vartheta \in \Theta \mid \tau(\vartheta) = \eta\}} \mathcal{L}(\vartheta; \boldsymbol{x}) \right) \\
&= \sup_{\vartheta} \mathcal{L}(\vartheta; \boldsymbol{x}) \\
&= \mathcal{L}(\hat{\vartheta}; \boldsymbol{x}) \\
&= \sup_{\{\vartheta \in \Theta \mid \tau(\vartheta) = \tau(\hat{\vartheta})\}} \mathcal{L}(\vartheta; \boldsymbol{x}) \\
&= \mathcal{L}(\hat{\vartheta}; \boldsymbol{x}).
\end{aligned}$$

Therefore, $\vartheta(\hat{\vartheta})$ is the Maximum Likelihood estimator of $\tau(\vartheta)$, where $\hat{\vartheta}$ is the Maximum Likelihood estimator of $\vartheta$.

□

*Remark 7.4.* Theorem 7.3 holds also in the context of *multidimensional* parametric spaces.

*Example 7.5.* Consider a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from $X \sim \mathcal{N}(\mu, \sigma^2)$. Find the Maximum Likelihood estimator for $(\mu, \sigma^2)$.

*Proof.* We know that

$$\begin{aligned}\mathcal{L}(\mu,\sigma^2;\boldsymbol{x}) &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}\\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\},\end{aligned}$$

therefore

$$\begin{aligned}\log\mathcal{L}(\mu,\sigma^2;\boldsymbol{x}) &= -\frac{n}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\\ &\propto -\frac{n}{2}\log\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.\end{aligned}$$

Next, we impose

$$\begin{aligned}\frac{d}{d\mu}\log\mathcal{L}(\mu,\sigma^2;\boldsymbol{x}) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0\\ \frac{d}{d\sigma^2}\log\mathcal{L}(\mu,\sigma^2;\boldsymbol{x}) &= -\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2=0\end{aligned}$$

to obtain

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2.$$

We want to verify that $(\hat{\mu},\hat{\sigma}^2)$ is a maximum. For every $\mu\neq\overline{x}_n$, we have

$$\sum_{i=1}^{n}(x_i-\hat{\mu})^2>\sum_{i=1}^{n}(x_i-\overline{x}_n)^2:$$

indeed,

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 &= \sum_{i=1}^{n}[(x_i-\overline{x}_n)+(\overline{x}_n-\hat{\mu})]^2\\ &= \sum_{i=1}^{n}(x_i-\overline{x}_n)^2+\sum_{i=1}^{n}(\overline{x}_n-\hat{\mu})^2+\underbrace{2\sum_{i=1}^{n}(x_i-\overline{x}_n)(\overline{x}_n-\hat{\mu})}_{\|\ 0}\\ &= \sum_{i=1}^{n}(x_i-\overline{x}_n)^2+\sum_{i=1}^{n}(\overline{x}_n-\hat{\mu})^2\\ &> \sum_{i=1}^{n}(x_i-\overline{x}_n)^2.\end{aligned}$$

By using this inequality, we have

$$\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2\right\} \geq \left(\frac{1}{2\pi\hat{\sigma}^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(x_i-\hat{\mu})^2\right\},$$

therefore, the probability of verifying that $(\hat{\mu}, \hat{\sigma}^2)$ is a maximum reduces to a one-dimensional problem: verifying that

$$(\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2\right\}$$

achieves its global maximum at

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2.$$

Indeed,

$$\begin{aligned}
0 &= \frac{d}{d\sigma^2}\left((\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2\right\}\right) \\
&= -\frac{n}{2}(\sigma^2)^{-\frac{n}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2\right\} + \\
&\quad + (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2\right\}\left(-\frac{(-1)}{2\sigma^4}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2\right) \\
&= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2\right\}\left(-\frac{n}{2}(\sigma^2)^{-\frac{n+2}{2}} + (\sigma^2)^{-\frac{n}{2}}\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2\right)
\end{aligned}$$

if and only if

$$\begin{aligned}
\frac{n}{2}\sigma^{-n-2} &= \frac{1}{2}\sigma^{-n-4}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2 \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2.
\end{aligned}$$

□

From a computational level, the invariance property is a good one, but it is also the reason why the Maximum Likelihood estimator can be biased for a finite sample size.

**Proposition 7.6.** *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a parametric model $X \sim f_X(x;\vartheta)$. If the Maximum Likelihood estimator $T_n$ for $\vartheta$ exists and is unique, then $T_n$ is a* ***function of the sufficient statistic*** *for $\vartheta$.*

Indeed, if a sufficient statistic for $\vartheta$ exists, then the Likelihood function $\mathcal{L}(\vartheta, \boldsymbol{x})$ can be factorised as follows:

$$\mathcal{L}(\vartheta, \boldsymbol{x}) = g(T(\boldsymbol{x}), \vartheta)h(\boldsymbol{x}),$$

and therefore

$$\log \mathcal{L}(\vartheta, \boldsymbol{x}) = \log(g(T(\boldsymbol{x}), \vartheta)) + \log h(\boldsymbol{x}) \propto \log(g(T(\boldsymbol{x})), \vartheta).$$

### 7.2.2 Asymptotic Properties of the Maximum Likelihood Estimator

**Theorem 7.7.** *Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be a random sample from a parametric model $X \sim f_X(x;\vartheta)$ and consider the Likelihood function $\mathcal{L}(\vartheta;\boldsymbol{x})$. Let $\hat{\vartheta}_n$ be the Maximum Likelihood estimator for $\vartheta$ and let $\tau(\cdot)$ be a continuous function of $\vartheta$. Under the usual assumptions of regularity, if, for every $\varepsilon > 0$ and every $\vartheta \in \Theta$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\tau(\hat{\vartheta}_n) - \tau(\vartheta)\right| \geq \varepsilon\right) = 0,$$

*then the Maximum Likelihood estimators are consistent in probability.*

**Theorem 7.8.** *Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be a random sample from a parametric model $X \sim f_X(x;\vartheta)$ and consider the Likelihood function $\mathcal{L}(\vartheta;\boldsymbol{x})$. Let $\hat{\vartheta}_n$ be the Maximum Likelihood estimator for $\vartheta$ and let $\tau(\cdot)$ be a continuous function of $\vartheta$. Then,*

$$\sqrt{n}[\tau(\hat{\vartheta}_n) - \tau(\vartheta)] \xrightarrow{n\to\infty} \mathcal{N}\left(0, \frac{1}{I_1(\vartheta)}\right),$$

*where $I_1(\vartheta)$ is the Fisher information of the parametric model, then the Maximum Likelihood estimators are*

- *consistent (weakly);*
- *asymptotically unbiased;*
- *asymptotically Gaussian;*
- *asymptotically efficient.*

*Proof.* We will prove the result for $\vartheta$. The proof for $\tau(\vartheta)$ is simply an application of the Delta method.

We know that

$$\mathcal{L}(\vartheta;\boldsymbol{x}) = \sum_{i=1}^{n} \log f_{X_i}(x_i;\vartheta).$$

Denote by $\mathcal{L}'$ and $\mathcal{L}''$ the first and second derivatives of $\mathcal{L}$ with respect to $\vartheta$, respectively. Expand the first derivative around $\vartheta_0$

$$\mathcal{L}'(\vartheta;\boldsymbol{x}) = \mathcal{L}'(\vartheta_0;\boldsymbol{x}) + (\vartheta - \vartheta_0)\mathcal{L}''(\vartheta_0,\boldsymbol{x}) + \dots$$

and then substitute $\vartheta$ with its Maximum Likelihood estimator:

$$0 = \mathcal{L}'(\vartheta_0;\boldsymbol{x}) + (\hat{\vartheta}_n - \vartheta_0)\mathcal{L}''(\vartheta_0;\boldsymbol{x}) + \dots.$$

Therefore, we would like to study the convergence of

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) = -\sqrt{n}\frac{\mathcal{L}'(\vartheta_0;\boldsymbol{x})}{\mathcal{L}''(\vartheta_0;\boldsymbol{x})} = -\frac{\frac{1}{\sqrt{n}}\mathcal{L}'(\vartheta_0;\boldsymbol{x})}{\frac{1}{n}\mathcal{L}''(\vartheta_0;\boldsymbol{x})}.$$

Now, we can apply the Central Limit Theorem to the numerator and the Weak Law of Large Numbers to the denominator. Indeed,

$$\frac{1}{\sqrt{n}}\mathcal{L}'(\vartheta_0;\boldsymbol{x}) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} w_i\right),$$

where

$$w_i = \frac{\frac{d}{d\vartheta} f_{X_i}(x_i;\vartheta)}{f_{X_i}(x_i;\vartheta)}$$

with $\mathbb{E}[w_i] = 0$ and $\mathrm{Var}(w_i) = I_n(\vartheta_0) = nI_1(\vartheta_0)$. By the Central Limit Theorem,

$$\frac{1}{\sqrt{n}}\mathcal{L}'(\vartheta_0;\boldsymbol{x}) \longrightarrow \mathcal{N}(0, I_1(\vartheta_0)).$$

For the denominator,

$$-\frac{1}{n}\mathcal{L}''(\vartheta_0;\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} w_i^2 - \frac{1}{n}\sum_{i=1}^{n} \frac{\frac{d^2}{d\vartheta^2} f_{X_i}(x_i;\vartheta)}{f_{X_i}(x_i;\vartheta)},$$

with

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} w_i^2\right] = I_1(\vartheta_0)$$

and

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{\frac{d^2}{d\vartheta^2} f_{X_i}(x_i;\vartheta)}{f_{X_i}(x_i;\vartheta)}\right] = 0.$$

Therefore, by the Weak Law of Large Numbers,

$$-\frac{1}{n}\mathcal{L}''(\vartheta_0;\boldsymbol{x}) \longrightarrow I_1(\vartheta_0).$$

In conclusion, we have proved that

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta) \longrightarrow \frac{X}{I_1(\vartheta_0)},$$

where $X \sim \mathcal{N}(0, I_1(\vartheta_0))$, meaning that

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta) \longrightarrow \mathcal{N}\left(0, \frac{1}{I_1(\vartheta_0)}\right).$$

□

## 7.3 08/11/18

*Example 7.9.* Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from $X \sim \mathcal{N}(\vartheta, \sigma^2)$. We already showed that, by the Factorisation Theorem,

$$\left( \sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2 \right)$$

is a sufficient statistic for $\vartheta$. Is it also a minimal statistic?

*Proof.* Compute

$$\begin{aligned}
\frac{\mathcal{L}(\vartheta, \boldsymbol{x})}{\mathcal{L}(\vartheta, \boldsymbol{y})} &= \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\} \\
&= \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 + \frac{2\mu}{2\sigma^2} \sum_{i=1}^{n} x_i - \frac{1}{2\sigma^2} n\mu^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} y_i^2 - \frac{2\mu}{2\sigma^2} \sum_{i=1}^{n} y_i + \frac{1}{2\sigma^2} n\mu^2 \right\} \\
&= \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2} n\overline{x}_n - \frac{1}{2\sigma^2} n\mu^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{n} y_i^2 - \frac{\mu}{\sigma^2} n\overline{y}_n + \frac{1}{2\sigma^2} n\mu^2 \right\} \\
&= \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[ (x_i - \overline{x}_n)^2 - (y_i - \overline{y}_n)^2 \right] \right\} \exp\left\{ n\left(\overline{x}_n - \overline{y}_n\right)(-2\mu) + \left(\overline{x}_n^2 - \overline{y}_n^2\right) \right\},
\end{aligned}$$

where

$$\begin{aligned}
\sum_{i=1}^{n} (x_i - \mu)^2 &= \sum_{i=1}^{n} (x_i - \overline{x}_n)^2 + n(\overline{x}_n - \mu)^2 \\
\sum_{i=1}^{n} (y_i - \mu)^2 &= \sum_{i=1}^{n} (y_i - \overline{y}_n)^2 + n(\overline{y}_n - \mu)^2.
\end{aligned}$$

If $\overline{x}_n = \overline{y}_n$, then

$$\frac{\mathcal{L}(\vartheta; \boldsymbol{x})}{\mathcal{L}(\vartheta; \boldsymbol{y})}$$

does not depend on $\mu$. Moreover,

$$\sum_{i=1}^{n} (x_i - \overline{x}_n)^2 = \sum_{i=1}^{n} (y_i - \overline{y}_n)^2,$$

therefore

$$\frac{\mathcal{L}(\vartheta; \boldsymbol{x})}{\mathcal{L}(\vartheta; \boldsymbol{y})}$$

does not depend on $\sigma^2$ either. This means that

$$\left( \overline{X}_n; \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \right)$$

is a minimal sufficient statistic for $\vartheta$, otherwise

$$\left(\sum_{i=1}^{n} X_i; \sum_{i=1}^{n} X_i^2\right)$$

is a minimal sufficient statistic for $\vartheta$, since it is a bijection of it. Being minimal sufficient, we have the best possible predictor.

□

*Example 7.10.* Consider a random sample $\boldsymbol{X} = (X_i, \dots, X_n)$ from $X \sim f_X(x; \vartheta)$, where

$$f_X(\boldsymbol{x}; \vartheta) = \frac{1}{\pi} \cdot \frac{1}{1 + (\boldsymbol{x} - \vartheta)^2},$$

a Cauchy distribution. Now,

$$\mathcal{L}(\vartheta; \boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\pi} \frac{1}{1 + (x_i - \vartheta)^2} = \underbrace{\left(\frac{1}{\pi}\right)^n}_{\| \; h(\boldsymbol{x})} \cdot \underbrace{\prod_{i=1}^{n} \frac{1}{1 + (x_i - \vartheta)^2}}_{\| \; g(T(\boldsymbol{x}), \vartheta)},$$

meaning that the random sample is the sufficient statistic, according to the Factorisation Theorem.

# 8

# Week VIII

## 8.1 13/11/18

### 8.1.1 Hypothesis Testing

**Definition 8.1.** *We define a* ***statistical hypothesis*** *as any sentence that specifies completely (or partially) the statistical model*

$$\{f_X(x;\vartheta) \,|\, \vartheta \in \Theta\}.$$

*If the specification is complete, then the hypothesis is called* ***simple****. If the specification is partial, then the statistical hypothesis is* ***composite****.*

*Example 8.2.* For example, $\vartheta = \vartheta^*$ is a simple hypothesis, $\vartheta > \vartheta^*$ is a composite *unidirectional* hypothesis.

The definition of a hypothesis is rather general, but the important point is that a hypothesis makes a statement about the population. The goal of a hypothesis test is to decide, based on a sample from the population, which of the two complementary hypotheses is true.

**Definition 8.3.** *The two complementary hypotheses in a hypothesis testing problem are called the* ***null hypothesis*** *and the* ***alternative hypothesis****, denoted by $H_0$ and $H_1$, respectively.*

We denote by $\Theta_0 \subset \Theta$ the set of values of $\Theta$ that are specified by the null hypothesis $H_0$. If $\vartheta$ denotes a population parameter, the general format of the null and alternative hypotheses is

$$H_0 : \vartheta \in \Theta_0 \quad \text{and} \quad H_1 : \vartheta \notin \Theta_0.$$

The null statistical hypothesis is something that exists before the experiment. It is a hypothesis that exists until proven otherwise. The alternative of $H_0$ (denoted by $H_1$, as we said) is the complement of $H_0$. In a hypothesis testing problem, after observing the sample, the experimenter must decide either to accept $H_0$ as true or to reject $H_0$ as false and decide $H_1$ is true.

The decision rule about rejecting (or accepting) $H_0$ has different interpretations. If we decide (based on the data) to reject $H_0$, this has the unique consequence that $H_0$ is false. Otherwise, if we decide (based on the data) to accept $H_0$, this does not imply the support of $H_0$.

The hypothesis test is a decision rule based on the sample space, and this decision rule is characterised by a function defined on the set of all the possible values of $\boldsymbol{X} = (X_1, \dots, X_n)$. There will be some points $\boldsymbol{x} \in \mathcal{R}_0 \subset \mathbb{R}^n$ (the *sample space*) such that the decision rule leads to rejecting $H_0$, and some other points $\boldsymbol{x} \notin \mathcal{R}_0$ such that the decision rule leads to *not* rejecting $H_0$.

**Definition 8.4.** *A **hypothesis testing procedure** of **hypothesis test** is a rule that specifies:*

1. *for which sample values the decision is made to accept $H_0$ as true;*
2. *for which sample values $H_0$ is rejected and $H_1$ is accepted as true.*

*The subset of the sample space for which $H_0$ will be rejected is called the **rejection region** (or **critical region**), defined as $\mathcal{R}_0$. The complement of the rejection region is called the **acceptance region**.*

*Remark 8.5.* On a "philosophical" level, some people worry about the distinction between *rejection* $H_0$ and *accepting* $H_1$. In the first case, there is nothing implied about what the experimenter *is* accepting, only that the state defined bu $H_0$ is being rejected. Similarly, a distinction can be made between *accepting* $H_0$ and *not rejecting* $H_0$. The first phrase implies that the experimenter is willing to assert the state of nature specified by $H_0$, while the second phrase implies that the experimenter really does not believe $H_0$, but does not have evidence to reject it. For the most part, we will not be concerned with these issues. We view a hypothesis testing problem as a problem in which one of two actions is going to be taken (the actions being the assertion of $H_0$ and $H_1$).

At this stage, the decision rule provides a partition of the sample space in two subsets. However, we can work with such a partition only for values like $n = 1, 2, 3$ for example, which of course is not realistic since $n$ is the sample's dimension. We must therefore reduce the dimensionality of the decision rule (or the dimensionality of the partition of the sample space). We can do that by using a *sufficient statistic*

$$T(X_1, \dots, X_n) = T(\boldsymbol{X}) = T_n,$$

a function $T : \mathbb{R}^n \to \mathbb{R}^m$ defined on the set of all possible values of $\boldsymbol{X}$. There will be some points $\boldsymbol{X} \in \mathcal{R}_0 \subseteq \mathbb{R}^n$ such that the decision rule leads to reject $H_0$, and some other points $\boldsymbol{X} \notin \mathcal{R}_0 \subseteq \mathbb{R}^n$ such that th decision rule leads to not reject $H_0$. The decision rule, therefore, is determined by the following criterion:

- if $\boldsymbol{X} \in \mathcal{R}_0$, then we reject $H_0$;
- if $\boldsymbol{X} \notin \mathcal{R}_0$, then we do not reject $H_0$.

This allows us to "translate" the original decision rule into the *new* decision rule, which is

- if $T_n \in C_0$, then we reject $H_0$;
- if $T_n \notin C_0$, then we do not reject $H_0$,

where $C_0 \subset T(\mathbb{R}_n)$ is the *rejection region* obtained by transforming $\mathcal{R}_0$ through the statistic $T_n$.

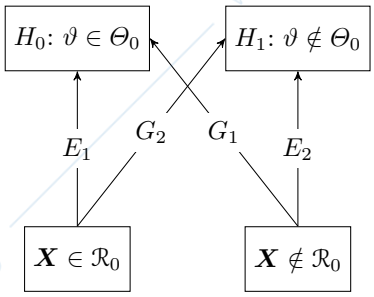
Therefore, briefly, what we obtained is the following: the **statistical hypothesis** defines a partition of the parametric space $\Theta$, meaning

$$\begin{cases} H_0 : & \vartheta_0 \in \Theta_0 \\ H_1 : & \vartheta \notin \Theta_0 \end{cases} . \tag{8.1}$$

The decision rule, on the other hand, defines a partition of the sample space, or a lower dimensional counterpart, meaning

$$\begin{cases} \text{if } T_n \in C_0, \text{ then we reject } H_0 \\ \text{if } T_n \notin C_0, \text{ then we do not reject } H_0 \end{cases} . \tag{8.2}$$

A test is a link between this two partitions. The possible links are:




Now, consider the statistical hypothesis (8.1) and the decision rule (8.2).

- $G_1$ is the decision of accepting $H_0$ (on the basis of the sample), where $H_0$ is true: *correct decision.*
- $G_2$ is the decision of rejecting $H_0$ (on the basis of the sample), where $H_0$ is false: *correct decision.*
- $E_1$ is the decision of rejecting $H_0$ (on the basis of the sample), despite $H_0$ is true: *wrong decision.*
- $E_2$ is the decision of accepting $H_0$ (on the basis of the sample), where $H_0$ is false: *wrong decision.*

If we reject $H_0$, then we can take a correct or a wrong decision (the same is valid if we do not reject $H_0$).

**Definition 8.6.** *Define*

- $\alpha := \mathbb{P}(E_1) = \mathbb{P}(\boldsymbol{X} \in \mathcal{R}_0, \vartheta \in \Theta_0)$;
- $\beta := \mathbb{P}(E_2) = \mathbb{P}(\boldsymbol{X} \notin \mathcal{R}_0, \vartheta \notin \Theta_0)$;
- $1 - \alpha := \mathbb{P}(G_1) = \mathbb{P}(\boldsymbol{X} \notin \mathcal{R}_0, \vartheta \in \Theta_0)$;
- $1 - \beta := \mathbb{P}(G_2) = \mathbb{P}(\boldsymbol{X} \in \mathcal{R}_0, \vartheta \notin \Theta_0)$,

*where $\alpha$ is the probability of the **first type error** and $\beta$ is the probability of the **second type error**. Also, $\gamma := 1 - \beta$ is known as the **power of the test**. Also, the **power function** will be defined as*

$$\gamma(\vartheta^*) := \mathbb{P}(\boldsymbol{X} \in \mathcal{R}_0, \vartheta = \vartheta^*) .$$

Our goal is to define a rejection region $\mathcal{R}_0$ such that $\alpha$ and $\beta$ are as small as they can get. Fix a rejection region $C_0$ and define

$$\alpha = \alpha(C_0) \quad \beta = \beta(C_0).$$

Suppose that $C_0'$ is a different rejection region such that

$$\alpha(C_0') < \alpha(C_0) \implies \beta(C_0') > \beta(C_0). \tag{8.3}$$

Indeed, (8.3) is possible if we consider $C_0' \subset C_0$ such that

$$\alpha(C_0') = \mathbb{P}(\boldsymbol{X} \in C_0' \mid H_0) < \mathbb{P}(\boldsymbol{X} \in C_0 \mid H_0) = \alpha(C_0),$$

therefore $(C_0')^c \subset (C_0)^c$ implies

$$\begin{aligned}
\beta(C_0') &= \mathbb{P}(\boldsymbol{X} \notin C_0' \mid H_1) \\
&= \mathbb{P}(\boldsymbol{X} \in (C_0')^c \mid H_1) \\
&> \mathbb{P}(\boldsymbol{X} \in (C_0)^c \mid H_1) \\
&= \mathbb{P}(\boldsymbol{X} \notin C_0 \mid H_1) \\
&= \beta(C_0),
\end{aligned}$$

meaning that $\alpha$ and $\beta$ are negatively correlated. Our way to proceed is the following: fix $\alpha$ and look for the rejection region with the minimum $\beta$.

## 8.2 14/11/18

*Example 8.7.* Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_{30})$ from $X \sim Ber(\vartheta)$. We know that

$$\hat{\vartheta} = \frac{1}{30} \sum_{i=1}^{30} X_i.$$

Consider the following test

$$\begin{cases} H_0 : & \vartheta = \frac{1}{2} \\ H_1 : & \vartheta = \frac{2}{3} \end{cases} . \tag{8.4}$$

Of course,

$$\sum_{i=1}^{30} X_i \sim \mathcal{B}(30, \vartheta).$$

Now, take

1. $C_0'$: we reject $H_0$ if the number of heads is greater than 25, meaning

$$\sum_{i=1}^{30} X_i > 25;$$

2. $C_0''$: we reject $H_0$ if the number of heads is greater than 17

$$\sum_{i=1}^{30} X_i > 17.$$

(1) and (2) are two possible rejection regions; if we use $C_0'$ as a decision rule, we would expect not to reject much, therefore it is likely to have a higher value of $\beta$. Otherwise, if we use $C_0''$, the value of $\alpha$ would be greater than the one we would obtain adopting $C_0'$. Now,

$$\alpha(C_0') = \mathbb{P}\left( \sum_{i=1}^{30} X_i > 25 \,\middle|\, \vartheta = \frac{1}{2} \right) = \sum_{r=26}^{30} \binom{30}{r} \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{30-r} = 0,000029$$

$$\beta(C_0') = \mathbb{P}\left( \sum_{i=1}^{30} X_i \leq 25 \,\middle|\, \vartheta = \frac{2}{3} \right) = \sum_{r=0}^{30} \binom{30}{r} \left(\frac{2}{3}\right)^r \left(\frac{1}{3}\right)^{30-r} = 0,9877$$

$$\alpha(C_0'') = \mathbb{P}\left( \sum_{i=1}^{30} X_i > 17 \,\middle|\, \vartheta = \frac{1}{2} \right) = \sum_{r=18}^{30} \binom{30}{r} \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{30-r} = 0,18$$

$$\beta(C_0'') = \mathbb{P}\left( \sum_{i=1}^{30} X_i \leq 17 \,\middle|\, \vartheta = \frac{2}{3} \right) = \sum_{r=0}^{30} \binom{30}{r} \left(\frac{2}{3}\right)^r \left(\frac{1}{3}\right)^{30-r} = 0,166.$$

### 8.2.1 Neyman-Pearson Lemma

For a fixed sample size, it is usually impossible to make both types of error probabilities arbitrarily small.

**Definition 8.8.** *For $0 \leq \alpha \leq 1$, a test with power function $\gamma(\vartheta)$ is a **size** $\alpha$ **test** if*

$$\sup_{\vartheta \in \Theta_0} \gamma(\vartheta) = \alpha.$$

**Definition 8.9.** *For $0 \leq \alpha \leq 1$, a test with power function $\gamma(\vartheta)$ is a **level** $\alpha$ **test** if*

$$\sup_{\vartheta \in \Theta_0} \gamma(\vartheta) \leq \alpha.$$

According to our definitions, the set of level $\alpha$ tests contains the set of size $\alpha$ tests. Moreover, the distinction becomes important in complicated models and complicated testing situations, where it is often computationally impossible to construct a size $\alpha$ test. In such situations, an experimenter must be satisfied with a level $\alpha$ test, realising that some compromises may be made.

**Definition 8.10.** *Let $\mathcal{C}$ be a class of tests for testing*

$$H_0 : \vartheta \in \Theta_0$$

*versus*

$$H_1 : \vartheta \in \Theta_0^c.$$

*A test in the class $\mathcal{C}$, with power function $\gamma(\vartheta)$, is a **uniformly most powerful (UMP) class** $\mathcal{C}$ **test** if*

$$\gamma(\vartheta) \geq \gamma'(\vartheta), \quad \forall \vartheta \in \Theta_0^c$$

*and every $\gamma'(\cdot)$ that is a power function of a test in class $\mathcal{C}$.*

In this section, the class $\mathcal{C}$ will be the class of all level $\alpha$ tests. Then, the test in Definition 8.10 is called a UMP level $\alpha$ test. However, the requirements in Definition 8.10 are so strong that UMP tests do not exist in many realistic problems. But in problems that have UMP tests, a UMP test might well be considered the best test in the class. Thus, we would like to be able to identify UMP tests if they exist.

The following famous Theorem clearly describes which tests are UMP level $\alpha$ tests in the the situation where the null and alternative hypotheses both consists of only one probability distribution for the sample (that is, when both $H_0$ and $H_1$ are *simple* hypotheses).

**Theorem 8.11 (Neyman-Pearson Lemma).** *Let $f_{\boldsymbol{X}}(\boldsymbol{x}; \vartheta_i)$ be the probability density function (or probability mass function) of $\boldsymbol{X} = (X_1, \ldots, X_n)$ corresponding to $\vartheta_i$. Consider testing*

$$H_0 : \vartheta = \vartheta_0$$

*versus*

$$H_1 : \vartheta = \vartheta_1$$

*using a test with rejection region $\mathcal{R}_0$ that satisfies*

$$\begin{cases} \boldsymbol{x} \in \mathcal{R}_0 & \text{if } f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta_1) > k f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta_0) \\ \boldsymbol{x} \in \mathcal{R}_0^c & \text{if } f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta_1) < k f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta_0), \end{cases} \tag{8.5}$$

*the* ***shape condition of the rejection region****, for some* $k \geq 0$ *and*

$$\alpha = \mathbb{P}\left(\boldsymbol{X} \in \mathcal{R}_0; \vartheta = \vartheta_0\right), \tag{8.6}$$

*the* ***size condition of the rejection region****. Then,*

1. ***sufficiency****: any test that satisfies (8.5) and (8.6) is a UMP level* $\alpha$ *test;*
2. ***necessity****: if there exist a test satisfying (8.5) and (8.6) with* $k > 0$*, then every UMP level* $\alpha$ *test is a size* $\alpha$ *test - satisfies (8.6) - and every UMP level* $\alpha$ *test satisfies (8.5) except perhaps on a set* $\mathcal{A}$ *satisfying* $\mathbb{P}\left(\boldsymbol{X} \in \mathcal{A}; \vartheta_0\right) = \mathbb{P}\left(\boldsymbol{X} \in \mathcal{A}; \vartheta_1\right) = 0$.

*Proof.* We will prove the Theorem for the case that $f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta_0)$ and $f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta_1)$ are probability density functions of continuous random variables. The proof for discrete random variables can be accomplished by replacing integrals with sums.

Note first that any test that satisfies (8.6) is a size $\alpha$ test and, hence, a level $\alpha$ test: indeed,

$$\sup_{\vartheta \in \Theta_0} \mathbb{P}\left(\boldsymbol{X} \in \mathcal{R}_0; \vartheta\right) = \mathbb{P}\left(\boldsymbol{X} \in \mathcal{R}_0; \vartheta_0\right) = \alpha,$$

since $\Theta_0$ has only one point. Define

$$\phi(\boldsymbol{x}) \coloneqq \begin{cases} 1 & \text{if } \boldsymbol{x} \in \mathcal{R}_0 \\ 0 & \text{if } \boldsymbol{x} \in \mathcal{R}_0^c \end{cases},$$

so that

- $\phi(\cdot)$ be a test function that satisfies (8.5) and (8.6);
- $\phi'(\cdot)$ be a test function of any other level $\alpha$ test;
- $\gamma(\vartheta)$ and $\gamma'(\vartheta)$ be the power functions corresponding to the test $\phi(\cdot)$ and $\phi'(\cdot)$, respectively.

Now, (8.5) implies that

$$\left[\phi(\boldsymbol{x}) - \phi'(\boldsymbol{x})\right]\left(f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta_1) - k f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta_0)\right) \geq 0, \quad \forall \boldsymbol{x}$$

since $\phi(\cdot) = 1$ if $f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta_1) > k f_{\boldsymbol{X}}(\boldsymbol{x},\vartheta_0)$ and $\phi(\cdot) = 0$ if $f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta_1) < k f_{\boldsymbol{X}}(\boldsymbol{x},\vartheta_0)$. Thus, for every $\vartheta \in \Theta$,

$$\begin{aligned} 0 &\leq \int \left[\phi(\boldsymbol{x}) - \phi'(\boldsymbol{x})\right]\left(f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta_1) - k f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta_0)\right) dx \\ &= \gamma(\vartheta_1) - \gamma'(\vartheta_1) - k\left[\gamma(\vartheta_0) - \gamma'(\vartheta_0)\right]. \end{aligned} \tag{8.7}$$

Statement (1) is proved by noting that, since $\phi'$ is a level $\alpha$ test and $\phi$ is a size $\alpha$ test,

$$\gamma(\vartheta_0) - \gamma'(\vartheta_0) = \alpha - \gamma'(\vartheta_0) \geq 0.$$

Thus (8.7) and $k \geq 0$ imply that

$$0 \leq \gamma(\vartheta_1) - \gamma'(\vartheta_1) - k[\gamma(\vartheta_0) - \gamma'(\vartheta_0)] \leq \gamma(\vartheta_1) - \gamma'(\vartheta_1),$$

showing that $\gamma(\vartheta_1) \geq \gamma'(\vartheta_1)$, and hence $\phi(\cdot)$ has greater power than $\phi'(\cdot)$. Since $\phi'(\cdot)$ was an arbitrary level $\alpha$ test and $\vartheta_1$ is the only point in $\Theta_0^c$, $\phi(\cdot)$ is a UMP level $\alpha$ test.

To prove statement (2), let $\phi'(\cdot)$ now be a test function for any UMP level $\alpha$ test. By the previous part, the test $\phi(\cdot)$ satisfying (8.5) and (8.6) is also a UMP level $\alpha$ test, which implies that

$$\gamma(\vartheta_1) = \gamma'(\vartheta_1).$$

This fact, (8.7) and $k > 0$ imply

$$\alpha - \gamma'(\vartheta_0) = \gamma(\vartheta_0) - \gamma'(\vartheta_0) \leq 0.$$

Now, since $\phi'$ is a level $\alpha$ test, we have that $\gamma'(\vartheta_0) \leq \alpha$, thus $\gamma'(\vartheta_0) = \alpha$. This implies that $\phi'(\cdot)$ is a size $\alpha$ test; this also implies that (8.7) is an equality in this case. But the non-negative integrand

$$[\phi(\boldsymbol{x}) - \phi'(\boldsymbol{x})][f_{\boldsymbol{X}}(\boldsymbol{x}; \vartheta_1) - k f_{\boldsymbol{X}}(\boldsymbol{x}; \vartheta_0)]$$

will have a zero integral only if $\phi'(\cdot)$ satisfies (8.5) except perhaps on a set $\mathcal{A}$ with

$$\int_A f_{\boldsymbol{X}}(\boldsymbol{x}; \vartheta_i) d\boldsymbol{x} = 0, \quad i = 0, 1.$$

This implies that the last assertion in statement (2) is true.

□

*Example 8.12.* Consider a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from $X \sim \mathcal{N}(\vartheta, \sigma^2)$, where $\sigma^2$ is supposed to be know. We would like to test

$$H_0 : \vartheta = \vartheta_0$$

versus

$$H_1 : \vartheta = \vartheta_1,$$

where $\vartheta_1 > \vartheta_0$. For a specified value of $\alpha$, a level $\alpha$ test in this problem is any test that satisfies

$$\mathbb{P}\left(\boldsymbol{X} \in \mathcal{R}_0; \vartheta_0\right).$$

We would like to compute

$$\frac{\mathcal{L}(\vartheta_1; \boldsymbol{x})}{\mathcal{L}(\vartheta_0; \boldsymbol{x})} \geq c,$$

so

$$\begin{aligned}
&\frac{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \vartheta_1)^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \vartheta_0)^2\right\}} \geq c \\
&\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n (x_i - \vartheta_1)^2 - \sum_{i=1}^n (x_i - \vartheta_0)^2\right]\right\} \geq c \\
&\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n x_i^2 + n\vartheta_1^2 - 2\vartheta_1 \sum_{i=1}^n x_i - \sum_{i=1}^n x_i^2 - n\vartheta_0^2 + 2\vartheta_0 \sum_{i=1}^n x_i\right]\right\} \geq c
\end{aligned}$$

$$\exp\left\{-\frac{1}{2\sigma^2}\left[n(\vartheta_1^2-\vartheta_0^2)-2(\vartheta_1-\vartheta_0)\sum_{i=1}^n x_i\right]\right\}\geq c$$

$$\exp\left\{-\frac{1}{2\sigma^2}(\vartheta_1-\vartheta_0)\left[n(\vartheta_1+\vartheta_0)-2\sum_{i=1}^n x_i\right]\right\}\geq c$$

$$-(\vartheta_1-\vartheta_0)\left[n(\vartheta_1-\vartheta_0)-2\sum_{i=1}^n x_i\right]\geq 2\sigma^2\log c$$

$$n(\vartheta_1-\vartheta_0)-2\sum_{i=1}^n x_i\leq -\frac{2\sigma^2\log c}{\vartheta_1-\vartheta_0}$$

$$-2\sum_{i=1}^n x_i\leq -\frac{2\sigma^2\log c}{\vartheta_1-\vartheta_0}-n(\vartheta_1-\vartheta_0)$$

$$\sum_{i=1}^n x_i\geq \frac{\sigma^2\log c}{\vartheta_1-\vartheta_0}+\frac{n}{2}(\vartheta_1-\vartheta_0)$$

is the shape of the rejection region.

*Example 8.13.* Consider a random sample $\boldsymbol{X}=(X_1,\dots,X_n)$ from $X\sim Ber(\vartheta)$. We would like to test

$$H_0:\vartheta=\vartheta_0$$

versus

$$H_1:\vartheta=\vartheta_1,$$

where $\vartheta_1>\vartheta_0$. Recall the alternative statement (**??**): we would like to compute

$$\frac{\mathcal{L}(\vartheta_0;\boldsymbol{x})}{\mathcal{L}(\vartheta_1;\boldsymbol{x})}\leq k,$$

so

$$\frac{\vartheta_0^{\sum_{i=1}^n x_i}(1-\vartheta_0)^{n-\sum_{i=1}^n x_i}}{\vartheta_1^{\sum_{i=1}^n x_i}(1-\vartheta_1)^{n-\sum_{i=1}^n x_i}}\leq k$$

$$\left(\frac{\vartheta_0}{\vartheta_1}\right)^{\sum_{i=1}^n x_i}\left(\frac{1-\vartheta_0}{1-\vartheta_1}\right)^{n-\sum_{i=1}^n x_i}\leq k$$

$$\sum_{i=1}^n x_i\log\left(\frac{\vartheta_0}{\vartheta_1}\right)+\left(n-\sum_{i=1}^n x_i\right)\log\left(\frac{1-\vartheta_0}{1-\vartheta_1}\right)\leq\log k$$

$$\sum_{i=1}^n x_i\left[\log\left(\frac{\vartheta_0}{\vartheta_1}\right)-\log\left(\frac{1-\vartheta_0}{1-\vartheta_1}\right)\right]\leq\log k-n\log\left(\frac{1-\vartheta_0}{1-\vartheta_1}\right)$$

$$\sum_{i=1}^n x_i\geq k'.$$

Here,

$$\alpha=\mathbb{P}\left(\sum_{i=1}^n x_i\geq k';\vartheta=\vartheta_0\right)=\sum_{r=k'}^n\binom{n}{r}\vartheta_0^r(1-\vartheta_0)^{n-r}.$$

We are looking for a value of $k'$ such that a very close $\alpha$ to the $\alpha$ chosen is provided, that is because $k \in \{0, \dots, n\}$, and in general it does not exist $k$ that determines the exact value of $\alpha$.

## 8.3 15/11/18

*Example 8.14.* Consider a random sample $\boldsymbol{X} = (X_1, \dots, X_n)$ from $X \sim EF(\cdot)$, the exponential family, where

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \vartheta) = \exp\{A(\boldsymbol{x})Q(\vartheta) + C(\boldsymbol{x}) - K(\vartheta)\},$$

and $Q$ is monotone increasing. We want to test

$$H_0 : \vartheta = \vartheta_0, \quad H_1 : \vartheta = \vartheta_1.$$

To do so, we can apply the Neyman-Pearson Lemma:

$$\frac{\mathcal{L}(\vartheta_0; \boldsymbol{x})}{\mathcal{L}(\vartheta_1; \boldsymbol{x})} \leq k$$

$$\frac{\prod_{i=1}^n \exp\{A(x_i)Q(\vartheta_0) + C(x_i) - K(\vartheta_0)\}}{\prod_{i=1}^n \exp\{A(x_i)Q(\vartheta_1) + C(x_i) - K(\vartheta_1)\}} \leq k$$

$$\exp\left\{Q(\vartheta_0)\sum_{i=1}^n A(x_i) + \sum_{i=1}^n C(x_i) - nK(\vartheta_0) - Q(\vartheta_1)\sum_{i=1}^n A(x_i) - \sum_{i=1}^n C(x_i) + nK(\vartheta_1)\right\} \leq k$$

$$\exp\left\{[Q(\vartheta_0) - Q(\vartheta_1)]\sum_{i=1}^n A(x_i) - n[K(\vartheta_0) - K(\vartheta_1)]\right\} \leq k$$

$$[Q(\vartheta_0) - Q(\vartheta_1)]\sum_{i=1}^n A(x_i) - n[K(\vartheta_0) - K(\vartheta_1)] \leq \log k$$

$$\sum_{i=1}^n A(x_i) \geq \frac{\log k + n[K(\vartheta_0) - K(\vartheta_1)]}{Q(\vartheta_0) - Q(\vartheta_1)}.$$

We conclude that

$$\sum_{i=1}^n A(x_i) \geq k'$$

is the shape of the rejection region for an arbitrary exponential family with $Q(\cdot)$ monotone increasing, meaning that $Q(\vartheta_0) < Q(\vartheta_1)$.

We will see an extension of the Neyman-Pearson Lemma for

1. *unidirectional hypothesis* (*Rubin Test*);
2. *arbitrary hypothesis* (*Likelihood Ratio Test*).

**Corollary 8.15.** *Consider the hypotheses of the Neyman-Pearson Lemma, and suppose that $T(\boldsymbol{X})$ is a sufficient statistic for the parameter $\vartheta$. Denote by $f_{\boldsymbol{X}}(\boldsymbol{x}; \vartheta_i)$, for $i = 0, 1$, the joint density function of the random sample, and by $g_T(t; \vartheta_i)$, for $i = 0, 1$, the density function of the sufficient statistic. Then, any test based on $T$ with rejection region $S$ (a subset of the sample space defined by $T$) is a UMP level $\alpha$ test if it satisfies*

$$\begin{aligned} g_T(t; \vartheta_1) &> k g_T(t; \vartheta_0), \quad t \in S \\ g_T(t; \vartheta_1) &< k g_T(t; \vartheta_0), \quad t \in S^c \end{aligned} \tag{8.8}$$

*for some $k \geq 0$, where*

$$\alpha = \mathbb{P}\left(T(\boldsymbol{X}) \in S; \vartheta_0\right). \tag{8.9}$$

*Proof.* In terms of the random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$, the test based on the sufficient statistic $T$ has the following rejection region:

$$\mathcal{R}_0 = \{\boldsymbol{x} \,|\, T(\boldsymbol{x}) \in S\}.$$

By the Factorisation Theorem, the probability density function (or probability mass function) of the random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ can be written as

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \vartheta_i) = g_T(T(\boldsymbol{x}); \vartheta_i)h(\boldsymbol{x}), \quad i = 0, 1$$

for some non-negative function $h(\cdot)$. Multiplying the inequalities in (8.8) by this non-negative function, we see that $\mathcal{R}_0$ satisfies

$$\boldsymbol{x} \in \mathcal{R}_0 \text{ if } f_{\boldsymbol{X}}(\boldsymbol{x}; \vartheta_1) = g_T(T(\boldsymbol{x}); \vartheta_1)h(\boldsymbol{x}) > kg_T(T(\boldsymbol{x}); \vartheta_0)h(\boldsymbol{x}) = kf_{\boldsymbol{X}}(\boldsymbol{x}; \vartheta_0)$$

and

$$\boldsymbol{x} \in \mathcal{R}_0^c \text{ if } f_{\boldsymbol{X}}(\boldsymbol{x}; \vartheta_1) = g_T(T(\boldsymbol{x}); \vartheta_1)h(\boldsymbol{x}) < kg_T(T(\boldsymbol{x}); \vartheta_0)h(\boldsymbol{x}) = kf_{\boldsymbol{X}}(\boldsymbol{x}; \vartheta_0).$$

Also, by condition (8.9), we have

$$\mathbb{P}(\boldsymbol{X} \in \mathcal{R}_0; \vartheta_0) = \mathbb{P}(T(\boldsymbol{X}) \in S; \vartheta_0) = \alpha.$$

So, by the sufficiency part of the Neyman-Pearson Lemma, the test based on $T$ is a UMP level $\alpha$ test.

□

*Example 8.16.* Consider

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \vartheta) = \exp\{A(\boldsymbol{x})Q(\vartheta) + C(\boldsymbol{x}) - K(\vartheta)\} :$$

a sufficient statistic could be $\sum_{i=1}^n A(x_i)$. If we wanted to test

$$H_0 : \vartheta = \vartheta_0, \quad H_1 : \vartheta = \vartheta_1 > \vartheta_0,$$

we could use the fact that the shape of the rejection region is

$$\sum_{i=1}^n A(x_i) \geq c.$$

In general, by the previous corollary, the shape of the rejection region defined by the Neyman-Pearson Lemma is a function of the sufficient statistic for the parameter of interest.

# 9

# Week IX

## 9.1 20/11/18

### 9.1.1 Rubin Test

As we have seen, the Neyman-Pearson Lemma comes in handy when dealing with *simple* hypotheses. However, in more realistic problems, the hypotheses of interest specify more than one possible distribution for the sample (*composite hypotheses*). A large class of problems that admit a UMP level $\alpha$ test involve one-side hypotheses and probability density functions, or probability mass functions, with the Monotone Likelihood Ratio property.

**Definition 9.1.** *A family of probability density functions, or probability mass functions, namely*

$$\{g_T(t;\vartheta) \,|\, \vartheta \in \Theta\}$$

*for a univariate random variable $T$ with real-valued parameter $\vartheta$ has a* ***Monotone Likelihood Ratio (MLR)*** *if, for every $\vartheta_2 > \vartheta_1$,*

$$\frac{g_T(t,\vartheta_2)}{g_T(t;\vartheta_1)}$$

*is a monotone (non-increasing or non-decreasing) function of $t$ on*

$$\{t \,|\, g_T(t;\vartheta_1) > 0 \vee g_T(t;\vartheta_2) > 0)\}.$$

*Example 9.2.* Many common families of distributions have the MLR property.

1. *Gaussian*, with respect to the mean for a fixed variance;
2. *Binomial*;
3. *Poisson*.

In general, any regular exponential family, with density function or probability mass function

$$g(t;\vartheta) = h(t)c(\vartheta)e^{w(\vartheta)t}$$

has the MLR property if $w(\vartheta)$ is a non-decreasing function.

There exist a generalisation of the Neyman-Pearson Lemma for statistical models with MLR property, which allows us to consider unidirectional (*composite*) hypotheses. This is called **Rubin test**.

**Theorem 9.3 (Rubin Test).** *Consider the problem of testing*

$$H_0 : \vartheta \leq \vartheta_0, \quad H_1 : \vartheta > \vartheta_0.$$

*Suppose that $T$ is a sufficient statistic for $\vartheta$, and that the family of probability density functions, or probability mass functions,*

$$\{g(t;\vartheta) \,|\, \vartheta \in \Theta\}$$

*of $T$ has the MLR property. Then, for any $t_0$, the statistical test that rejects $H_0$ if and only if $T > t_0$ is a UMP level $\alpha$ test, where $\alpha = \mathbb{P}(T > t_0; \vartheta_0)$.*

### 9.1.2 Likelihood Ratio Test

**Definition 9.4.** *Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be a random sample from a parametric model $X \sim f_{\boldsymbol{X}}(\boldsymbol{x};\vartheta)$. Assume we want to test something very general, like*

$$H_0 : \vartheta \in \Theta_0, \quad H_1 : \vartheta \notin \Theta_0.$$

*We define the **Likelihood Ratio** $\lambda(\boldsymbol{x})$ as*

$$\lambda(\boldsymbol{x}) \coloneqq \frac{\sup_{\vartheta \in \Theta_0} \mathcal{L}(\vartheta;\boldsymbol{x})}{\sup_{\vartheta \in \Theta} \mathcal{L}(\vartheta;\boldsymbol{x})}.$$

*We consider a rejection region of the form*

$$\lambda(\boldsymbol{x}) \leq c,$$

*such that $\mathbb{P}(\lambda(\boldsymbol{x}) \leq c \,|\, H_0) = \alpha$. Of course, $H_0$ is not specified.*

*Remark 9.5.* The numerator of $\lambda(\boldsymbol{x})$ is the maximum probability of the observed sample. Supposing the parameter lives in the null hypothesis, the denominator is the maximum probaility of the observed sample over all possible parameters. The ratio is small if there are parameter points in the alternative, for which the observed sample is much more likely than any other parameter in the null.

*Example 9.6.* Let $\boldsymbol{X} = (X_1, \dots, X_n)$ be a random sample from $X \sim \mathcal{N}(\vartheta, \sigma^2)$, where $\sigma^2$ is unknown. Consider the following test:

$$H_0 : \vartheta = \vartheta_0, \quad H_1 : \vartheta \neq \vartheta_0.$$

In this case,

$$\Theta = \mathbb{R} \times \mathbb{R}_+ = \{(\vartheta, \sigma^2) \in (\mathbb{R} \times \mathbb{R}_+)\}, \quad \Theta_0 = \{(\vartheta_0, \sigma^2) \in (\vartheta_0, \mathbb{R}_+)\}.$$

Without any restriction on the parametric space $\Theta$, the Maximum Likelihood estimators of $(\vartheta, \sigma^2)$ are

$$\hat{\vartheta} = \overline{X}_n, \quad \hat{\sigma}^2 = S_n^2.$$

Therefore,

$$\begin{aligned}\mathcal{L}(\overline{X}_n, S_n^2; \boldsymbol{x}) &= \left(\frac{1}{\sqrt{2\pi S_n^2}}\right)^n \exp\left\{-\frac{1}{2S_n^2}\sum_{i=1}^n (X_i - \overline{X}_n)^2\right\} \\ &\propto (S_n^2)^{-\frac{n}{2}} e^{-\frac{n}{2}} \\ &\propto (S_n^2)^{-\frac{n}{2}}\end{aligned}$$

Now, we maximise the Likelihood function with respect to $\Theta_0$: if $\hat{\vartheta} = \vartheta_0$,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \vartheta_0)^2 = S_0^2,$$

therefore

$$\mathcal{L}(\vartheta_0, S_0^2) = \left(\frac{1}{\sqrt{2\pi S_0^2}}\right)^n \exp\left\{-\frac{1}{2S_0^2}\sum_{i=1}^n (X_i - \vartheta_0)^2\right\} \propto (S_0^2)^{-\frac{n}{2}},$$

meaning that the Likelihood Ratio test is given by

$$\lambda(\boldsymbol{x}) = \left(\frac{S_n^2}{S_0^2}\right)^{n/2} \leq c.$$

Since

$$\begin{aligned}S_0^2 &= \frac{1}{n}\sum_{i=1}^n (X_i - \vartheta_0)^2 \\ &= \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n + \overline{X}_n - \vartheta_0)^2 \\ &= \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 + (\overline{X}_n - \vartheta_0)^2 + \underbrace{\frac{2}{n}\sum_{i=1}^n (X_i\overline{X}_n - X_i\vartheta_0 - \overline{X}_n^2 + \overline{X}_n\vartheta_0)}_{(*)} \\ &= S_n^2 + (\overline{X}_n - \vartheta_0)^2,\end{aligned}$$

where $(*) = 0$ because

$$\begin{aligned}&\frac{2}{n}\vartheta_0\sum_{i=1}^n (\overline{X}_n - X_i) + \frac{2}{n}\overline{X}_n\sum_{i=1}^n (X_i - \overline{X}_n) \\ &= 2\vartheta_0\left(\overline{X}_n - \frac{1}{n}\sum_{i=1}^n X_i\right) + 2\overline{X}_n\left(\frac{1}{n}\sum_{i=1}^n X_i - \overline{X}_n\right) \\ &= 0.\end{aligned}$$

Hence

$$\lambda(\boldsymbol{x}) = \left(\frac{S_n^2}{S_n + (\overline{X}_n - \vartheta_0)^2}\right)^{\frac{n}{2}} \leq c$$

$$\left(1+\frac{(\overline{X}_n-\vartheta_0)^2}{S_n^2}\right)^{-\frac{n}{2}} \leq c$$
$$\frac{(\overline{X}_n-\vartheta_0)^2}{S_n^2} \geq c'$$
$$\frac{(\overline{X}_n-\vartheta_0)^2}{\frac{n-1}{n}S_n^2} \geq c'$$
$$\frac{\sqrt{n}\,|X_n-\vartheta_0|}{S_n} \geq c'$$
$$|T_n| \geq c',$$

where

$$T_n = \frac{\sqrt{n}(\overline{X}_n-\vartheta_0)}{s_n} \sim t_{n-1}$$

under $H_0$. Therefore,

$$\mathbb{P}\left(-t_{\alpha/2} \leq T_n \leq t_{\alpha/2}\right) = 1-\alpha$$

and the *rejection region of level* $\alpha$ is

$$|T_n| > t_{\alpha/2}.$$

*Remark 9.7.* 1. The rejection region defined by the Likelihood Ratio Test is obtained by taking the ratio between the Likelihood function evaluated in the point of maximum, with respect to a parametric space **restricted** by a null hypothesis, and the Likelihood function evaluated in the point of maximum with respect to the **unrestricted** parametric space.
2. The rejection region defined by the Likelihood Ratio test is always a function of the sufficient statistic, for the parameter of interest.
3. If the hypotheses are "simple enough", then the Likelihood Ratio test becomes the Neyman-Pearson Lemma.

**Exercise 9.8.** The number of breaks of an object every week is a Poisson random variable with parameter $\lambda$. Consider the number of breaks in subsequent weeks to be independent. Let $X_1, \ldots, X_{10}$ be the number of breaks in ten weeks, and let

$$\overline{X}_{10} = \frac{1}{10}\sum_{i=1}^{10} X_i.$$

1. Find $t$ such that $\{\overline{X}_{10} > t\}$ is a rejection region of level $\alpha = 0,05$ for testing

$$H_0 : \lambda = 2, \quad H_1 : \lambda = 3.$$

   Compute the power of the test.
2. Is the test above a UMP level $\alpha$ test?

*Proof.* (1) Under $H_0$, the random variables $X_1, \ldots, X_{10}$ are Poisson random variables with parameter $\lambda = 2$. If we set $S := \sum_{i=1}^{10} X_i$, then

$$\mathbb{P}\left(\overline{X}_{10} > t\right) = \mathbb{P}\left(S > 10t\right).$$

We need to find the point $k$ such that $\mathbb{P}(S \leq k) \geq 0,95$; then, we set $t = k/10$ to solve

$$\mathbb{P}(S \leq k) \geq 0,95.$$

Recall that $S$, being the sum of Poisson random variables, is itself a Poisson random variable. Therefore,

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{10} X_i \leq 10t; \lambda = 2\right) &= \mathbb{P}(S \leq 10t; \lambda = 2) \\
&= \mathbb{P}\left(\frac{R - 10\lambda}{\sqrt{10\lambda}} \leq \frac{10t - 10\lambda}{\sqrt{10\lambda}}; \lambda = 2\right) \\
&\approx \Phi\left(\frac{10t - 10\lambda}{\sqrt{10\lambda}}; \lambda = 2\right) \\
&= \Phi\left(\frac{10t - 20}{\sqrt{20}}\right) \\
&\geq 0.95
\end{aligned}$$

if and only if

$$\sqrt{5}(t - 2) \geq Z_{0.05}.$$

Therefore,

$$t \geq \frac{1.64}{\sqrt{5}} + 2.$$

In order to find the power of the test with, we have to compute a similar probability as above:

$$\begin{aligned}
\gamma &:= 1 - \beta \\
&= 1 - \mathbb{P}\left(\sum_{i=1}^{10} X_i \leq 10t; \lambda = 3\right) \\
&= \mathbb{P}(S \leq 10t; \lambda = 3) \\
&= 1 - \mathbb{P}\left(\frac{R - 10\lambda}{\sqrt{10\lambda}} \leq \frac{10t - 10\lambda}{\sqrt{10\lambda}}; \lambda = 3\right) \\
&\approx 1 - \Phi\left(\frac{10t - 10\lambda}{\sqrt{10\lambda}}; \lambda = 3\right) \\
&= 1 - \Phi\left(\frac{10t - 30}{\sqrt{30}}\right)
\end{aligned}$$

(2) In order to find a test with the highest power, we can apply the Neyman-Pearson Lemma.

$$\frac{\mathcal{L}(\lambda_1; \boldsymbol{x})}{\mathcal{L}(\lambda_0; \boldsymbol{x})} = \frac{\prod_{i=1}^{10} e^{-\lambda_1} \frac{\lambda_1^{x_i}}{x_i!}}{\prod_{i=1}^{10} e^{-\lambda_0} \frac{\lambda_0^{x_i}}{x_i!}}.$$

Then, the rejection region of point (1) is a UMP level $\alpha$ rejection region.

□

## 9.2 21/11/18

So far, we have found the following techniques for the problem

$$\alpha = \mathbb{P}\left(T \geq k; H_0\right) :$$

1. Neyman-Pearson Lemma: UMP level $\alpha$ test for simple hypotheses and for an arbitrary statistical model;
2. Rubin Test: UMP level $\alpha$ test for composite unidirectional hypotheses, for statistical models with Monotone Likelihood Ratio;
3. Likelihood Ratio Test: level $\alpha$ test for any hypothesis and for arbitrary statistical models.

One of the most useful approach to obtain the shape of the rejection region is the *Likelihood Ratio Test*:

$$\lambda(\boldsymbol{x}) = \frac{\max_{\vartheta \in \Theta_0} \mathcal{L}(\vartheta; \boldsymbol{x})}{\max_{\vartheta \in \Theta} \mathcal{L}(\vartheta; \boldsymbol{x})},$$

with its form is given by

$$\{\boldsymbol{x} \,|\, \lambda(\boldsymbol{x}) \leq c\}.$$

After we gather the data $\boldsymbol{X} = \boldsymbol{x}$, the Likelihood function $\mathcal{L}(\vartheta; \boldsymbol{x})$ is a completely defined function of the variable $\vartheta$. Even if we cannot compute analytically the points that maximise the Likelihood function over the sets $\Theta_0$ and $\Theta$, these can be always evaluated *numerically*. Thus, the test statistic $\lambda(\boldsymbol{x})$ can be obtained for the observed data point even if no convenient formula defining $\lambda(\boldsymbol{x})$ is available.

To define a level $\alpha$ test, the constant $c$ must be chosen so that

$$\sup_{\vartheta \in \Theta_0} \mathbb{P}\left(\lambda(\boldsymbol{X}) \leq c; \vartheta\right) \leq \alpha. \tag{9.1}$$

If we cannot derive a simple formula for $\lambda(\boldsymbol{x})$, it might seem that it is hopeless to derive the sampling distribution of $\lambda(\boldsymbol{X})$ and thus know how to pick $c$ to ensure (9.1). This means that we must know the distribution of $\lambda(\boldsymbol{x})$. A solution to this problem consists in determining a general result for the asymptotic behaviour of $\lambda(\boldsymbol{x})$, as $n \to \infty$.

**Theorem 9.9 (Wilks).** *Consider a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from a probability density function (or probability mass function) $X \sim f_X(\boldsymbol{x}, \vartheta)$. Under some regularity conditions on the model $f_X(x; \vartheta)$, if $\Theta_0 = \{\vartheta_0\}$, then the distribution of the statistic*

$$-2 \log \lambda(\boldsymbol{x}) \xrightarrow{n \to \infty} \chi^2_1,$$

*where $\chi^2_1$ is a Chi-squared random variable with one degree of freedom.*

*Example 9.10.* Consider the problem of testing

$$H_0 : \lambda = \lambda_0, \quad H_1 : \lambda \neq \lambda_0.$$

Consider a random sample from $X \sim \mathcal{P}(\lambda)$. Therefore, applying Theorem 9.9,

$$-2 \log \lambda(\boldsymbol{x}) = -2 \log \left( \frac{e^{-n\lambda_0} \lambda_0^{\sum_{i=1}^n x_i}}{e^{-n\hat{\lambda}} \hat{\lambda}^{\sum x_i}} \right) = 2n \left[ (\lambda_0 - \hat{\lambda}) - \hat{\lambda} \log \left( \frac{\lambda_0}{\hat{\lambda}} \right) \right],$$

where

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

By a direct application of Wilks' Theorem, we reject $H_0$ at level $\alpha$ when

$$-2\log\lambda(\boldsymbol{x}) > \chi^2_{1,\alpha},$$

where the right had side of the inequality indicates the quantile of order $\alpha$ of a Chi-squared distribution.

*Remark 9.11.* As the Cramer Theorem, Wilks' Theorem is a "qualitative" result. That is, it does not provide a level of the sample size for which the theorem can be applied.

*Example 9.12.* Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from $X \sim f_X(x;\vartheta)$, where

$$f_X(x,\vartheta) = \exp\{-(x-\vartheta)\}\mathbb{1}_{(\vartheta,\infty)}(x).$$

Consider the problem of testing

$$H_0 : \vartheta = \vartheta_0, \quad H_1 : \vartheta > \vartheta_0.$$

If we apply the Likelihood Ratio Test,

$$\lambda(\boldsymbol{x}) = \frac{\prod_{i=1}^{n}\exp\{-(x_i-\vartheta_0)\}}{\prod_{i=1}^{n}\exp\{-(x_i-t_n)\}} = \exp\{-n(t_n-\vartheta_0)\},$$

where $t_n = x_{(1)}$. Therefore,

$$-2\log\lambda(\boldsymbol{x}) = 2n(T_n-\vartheta_0),$$

where $T_n = X_{(1)} = \min\{X_1, \ldots, X_n\}$. If we define $H = 2n(T_n-\vartheta_0)$, under $H_0$, the distribution of $H$ is given by

$$\mathbb{P}(H \le h) = \mathbb{P}\left(T_n - \vartheta_0 \le \frac{h}{2n}\right),$$

where $T_n$ is a negative exponential of parameter $n$ and $T_n - \vartheta_0$ is a negative exponential of parameter one.

Recall that, if $X_1, \ldots, X_n$ with $X_i \sim \mathcal{E}_n(1)$, then $\min_i X_i \sim \mathcal{E}_n(n)$. Therefore

$$\mathbb{P}(2n(T_n-\vartheta_0) \le h) = 1 - \exp\left\{-\frac{h}{2}\right\},$$

meaning that $2n(T_n-\vartheta_0) \sim \mathcal{E}_n\left(\frac{1}{2}\right)$: therefore, the Negative Exponential distribution with parameter $1/2$ is a Chi-squared distribution with two degrees of freedom.

Wilks' Theorem can be stated also for statistical (regular) models, whose parametric space has dimension greater than one.

**Theorem 9.13 (Kendal-Stuart).** *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample from a probability density function (or probability mass function) $f_X(x;\vartheta)$. Under the usual conditions of regularity, the statistic $-2\log\lambda(\boldsymbol{x})$ converges weakly, as $n \to \infty$, to a Chi-squared distribution. The degrees of freedom of the limiting distribution is the difference between the number of free parameters specified by the null hypothesis and the number of free parameters specified of the model.*

*Example 9.14 (Multinomial model).* Let $\vartheta = (p_1, \dots, p_5)$, where the $p_j$'s are non-negative and sum up to 1. Suppose that we consider a random sample $(X_1, \dots, X_5)$ such that

$$\mathbb{P}(X_i = j) = p_j, \quad j = 1, \dots, 5.$$

Then, the Likelihood function is proportional to

$$\prod_{i=1}^{5} p_i^{y_i},$$

where $y_i$ is the number of $x_1, \dots, x_n$ equal to $i$. Consider the problem of testing

$$H_0 : p_1 = p_2 = p_3, p_4 = p_5 \quad H_1 : H_0 \text{ is not true.}$$

The parametric space is $\Theta$ and it is a four-dimensional parametric space: indeed, since $p_5 = 1 - \sum_{i=1}^{4} p_i$, there are only four free parameters. The parameter set is defined as follows:

$$\left\{ (p_1, \dots, p_5) \,\middle|\, p_i \geq 0 \wedge \sum_{i=1}^{4} p_1 \leq 1 \right\},$$

a subset of $\mathbb{R}^4$ containing an open subset of $\mathbb{R}^4$. Thus, $q = 4$. How many free parameters are there under the null hypotheses? Only one, because once $p_1$ is fixed $(0 \leq p_1 \leq 1/3)$, then $p_2 = p_3$ must be equal to $p_1$ and $p_4 = p_5$ must be equal to

$$\frac{1 - 3p_1}{2}.$$

Thus $p_1 = 1$, and the degrees of freedom is $4 - 1 = 3$. According to Wilks' Theorem we have that

$$-2 \log \lambda(\boldsymbol{x}) \longrightarrow \chi^2_3.$$

Then, asymptotically, for $n$ large, the test rejects $H_0$ if

$$-2 \log \lambda(\boldsymbol{x}) \geq \chi^2_{3,\alpha}.$$

Wilks' Theorem is not the only approach for asymptotic test. A common method to construct large sample test statistic is based on estimators that are asymptotically Gaussian. In particular, the Maximum Likelihood estimators. Suppose that we want to test a hypothesis about a real valued parameter, and $W_n = W(X_1, \dots, X_n)$ is a point estimator for $\vartheta$, for instance the Maximum Likelihood estimator. An approximate test, based on the Gaussian approximation, can be justified as follows: if $\sigma^2$ denoted the variance of $W_n$ and if we use some form of the Central Limit Theorem (or the Delta method) in order to show that, as $n \to \infty$,

$$\frac{W_n - \vartheta}{\sigma_n} \longrightarrow \mathcal{N}(0, 1),$$

then $(W_n - \vartheta)/\sigma_n$ can be used as a term of comparison to the standard Gaussian. This implies that we have an approximate test.

# 10

# Week X

## 10.1 27/11/18

### 10.1.1 Linear Models

The technique or regression, in particular *linear* regression, probably wins as the most popular statistical tool. The major purpose of regression is to explore the dependence of one variable on another. What we did before was to consider a random sample $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ from a parametric model $Y \sim f_Y(y, \vartheta)$. Now, we say that

$$Y_i = f(\vartheta, x) \quad \text{with } x \text{ fixed,}$$

and we want to make inference on $\vartheta$.

The first restriction that we introduce is the presence of a **response variable** $y$ and one or more **explanatory variables** $x_1, \dots, x_p$. These variables, $y$ and $(x_1, \dots, x_p)$, are treated asymmetrically, since the response variable is considered to vary according to variations of the explanatory variables.

*Example 10.1.* In medical statistics, the variable $x$ could represent the dose of a drug which is given to individuals, and this is the role of the explanatory variable; the variable $y$ could be, for instance, a binary response which denotes the effect of the drug.

Now, consider

$$y = f(x_1, \dots, x_p) :$$

for the moment being, we assume $f$ to be linear (this is the reason why we call it *linear* model).

The second restriction is that the variable $y$ is the sum of two terms, namely

$$y = f(x_1, \dots, x_p) + \varepsilon. \tag{10.1}$$

The term $f(x_1, \dots, x_p)$ is called the **systematic component** and it is the contribution of $x_1, \dots, x_p$ to $y$. The term $\varepsilon$ is called the **erratic component**, and it is the random discrepancy between $y$ and $f(x_1, \dots, x_p)$. This is *free* of any connection with $x_1, \dots, x_p$ (meaning that they do not interact with each other).

**Definition 10.2.** *An equation of the type* (10.1) *is called a* ***regression model****.*

*Example 10.3.* Consider an electric circuit. Suppose we measure the voltage $V$ starting when the circuit is closed. It is known that there exist the following relationship

$$V = E\left(1 - \exp\left\{-\frac{t}{T}\right\}\right),$$

where

- $E$ denotes the *voltage of the power supply*;
- $t$ is the *time since the closure of the circuit*;
- $T = RC$ is a constant characteristic value of the circuit depending on $R$ (the *resistor*) and $C$ (the *capacitor*).

In this framework, $V$ is the response variable, $t$ is the explanatory variable, whilst $E$ and $T$ are parameters which must be evaluated. Under closed circuit, we take some measurements at different points in time: what we get is a collection of various pairs of the form $(t, V)$.

Next, the error enters into the game: if we assume that the instruments that we use to take measurements are error-free, then the measurements are completely accurate, and the pairs $(t_i, V_i)$, for $i = 1, \ldots, n$, are sufficient to determine the parameters. Therefore, we just have to solve a non-linear equation, with some constraints. However, in our case, we will consider the error's contribution, and we want to model its behaviour. In practice, the circuit and the instruments are not expected to be under ideal conditions. In this way we can model this uncertainty with

$$V = E\left(1 - \exp\left\{-\frac{t}{T}\right\}\right) + \varepsilon, \tag{10.2}$$

where $\varepsilon$ includes all the sources of departure from the ideal behaviour. Since we do not know anything about $\varepsilon$, we assume it to be random. Therefore, (10.2) can not be solved as a non-linear equation, since $V$ is random.

**Definition 10.4.** *We consider a special case of regression model (10.1), where $f(\cdot)$ is a linear function of the parameters,*

$$y = \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon, \tag{10.3}$$

*where $\beta_i$ are unknown parameters. This model is called* ***linear regression model****.*

This restriction on $f(\cdot)$ could appear as a limitation of the model, and indeed it is. However, it turns out to be useful in a large number of real examples. Why? First, (10.3) is extremely simple. Second, even if there are cases where $f(\cdot)$ is non-linear, they can be approximated by a linear relationship.

We consider $n$ observations from the model (10.3): the objective is to make inference on the values of the parameters $\beta_1, \ldots, \beta_p$. By inference, we mean *point estimation* and *testing*. Since the error component is considered to be generated at random, $y$ is also regarded as a random quantity. Therefore, the first step would be to specify the distributional assumption on $\varepsilon$.

### 10.1.2 Second Order Assumptions of the Linear Regression Model

**Definition 10.5.** *Suppose that $n$ observations $(y_1, \dots, y_n)$ are available, where each component is generated by (10.3). These observations are assumed to be a random sample from the random variables $(Y_1, \dots, Y_n)$ that satisfy*

$$Y_i = \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \dots, n,$$

*where $Y_i$ is the $i$-th component of the response variable, $x_i$ is the $i$-th value of the explanatory variable $X_i$. Using the matrix notation for convenience, we can write*

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

*where*

- $\boldsymbol{Y} = {}^{\mathrm{t}}(Y_1, \dots, Y_n)$ *is the random vector with the $n$ component of the response variable;*
- $X = (x_{ij})$ *is an $n \times p$ matrix (with $n \gg p$) called* ***regression matrix****, which contains the values of the $p$ explanatory variables;*
- $\boldsymbol{\varepsilon} = {}^{\mathrm{t}}(\varepsilon_1, \dots, \varepsilon_n)$ *is the vector containing the $n$ components of the error term;*
- $\boldsymbol{\beta} = {}^{\mathrm{t}}(\beta_1, \dots, \beta_p)$ *is the vector of the regression parameters.*

*Now, the* ***second order assumptions*** *are the following:*

- $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$;
- $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n$, *for some unknown $\sigma^2$, implying that*

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \textit{and} \quad \mathrm{Var}(\varepsilon_i) = \sigma^2, \qquad \forall i \neq j;$$

- $X$ *is a non-random matrix with full rank $p$.*

*We also refer to $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with the above assumptions, as a* ***linear regression model****.*

*Remark 10.6.* The linear regression model is particularly suited for the mathematical description of problems arising from *controlled experiments*, where the *experimenter* can control the values taken by relevant features to examine the corresponding value of the response variable. In this setting, $X$ contains the values of the experimental features, which are non-stochastic since they are chosen by the *experimenter*. The error term is determined by the measuring errors and, if the instruments are not biased, then it follows that

$$\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}.$$

Finally, if the various experiments are conducted in such a way as not to influence each other, then the stochastic independence assumption is satisfied, and it implies uncorrelated error.

We aim to make inference on the set of $(p+1)$ parameters $\beta_1, \dots, \beta_p, \sigma^2$. By the second order assumptions, we have that

- $\mathbb{E}[\boldsymbol{Y}] = X\boldsymbol{\beta}$;
- $\mathrm{Var}(\boldsymbol{Y}) = \sigma^2 I_n$.

However, we do not have enough ingredients to write down the probability distribution of $\boldsymbol{Y}$. Therefore, we do not have the Likelihood function of the model for the observed $\boldsymbol{y}$. This means that we need more assumptions for $\boldsymbol{\varepsilon}$ to introduce a probabilistic model for $\boldsymbol{Y}$. We must find a way to estimate $\boldsymbol{\beta}$ by using only the second order assumptions. A reasonable proposal is to choose $\boldsymbol{\beta}$ so as to minimise the *euclidean distance*

$$\|\boldsymbol{y} - \boldsymbol{\mu}\|$$

between the observed vector $\boldsymbol{y}$ and its mean value $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{Y}] = X\boldsymbol{\beta}$, that it to minimise

$$\|\boldsymbol{y} - X\boldsymbol{\beta}\|.$$

## 10.2 28/11/18

### 10.2.1 Least Square Method

We are looking for some estimate of the parameter $\boldsymbol{\beta}$ and $\sigma^2$. As we noted last time, the problem is to choose $\boldsymbol{\beta}$ such that the *euclidean distance* $\|\boldsymbol{y}-\boldsymbol{\mu}\|$ is minimum. Equivalently, we can minimise the square of the distance; basically

$$Q(\boldsymbol{\beta}) = \|\boldsymbol{y}-\boldsymbol{\mu}\|^2 = {}^{\mathrm{t}}(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu}) = {}^{\mathrm{t}}(\boldsymbol{y}-X\boldsymbol{\beta})(\boldsymbol{y}-X\boldsymbol{\beta}).$$

With this choice, we are establishing the **Least Square criterion**, whose name reminds that it operates by minimising the sum of squared differences between the component of $\boldsymbol{y}$ and the corresponding component of $\boldsymbol{\mu}$.

Consider the vectors $\boldsymbol{y}, \boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ containing the response and the explanatory variables as elements of the vector space $\mathbb{R}^n$. As the parameter $\boldsymbol{\beta}$ varies in $\mathbb{R}^p$, the expression

$$X\boldsymbol{\beta} = \beta_1\boldsymbol{x}_1 + \ldots + \beta_p\boldsymbol{x}_p$$

is a linear combination of the column vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_p$ with coefficients $\boldsymbol{\beta}$. In other terms, we have a parametric equation (in $\boldsymbol{\beta}$) of the linear subspace of $\mathbb{R}^n$ generated by the columns of the design matrix. We denote this linear subspace by $\mathcal{C}(X)$ (with dimension $p$). In the model $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ we stated that $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{Y}]$ lies in $\mathcal{C}(X)$ and the Least Squares criterion selects an element of $\mathcal{C}(X)$ which minimises the *euclidean distance* between $\boldsymbol{y}$ and the space $\mathcal{C}(X)$. We denote this element by $\hat{\boldsymbol{\mu}} \coloneqq X\hat{\boldsymbol{\beta}}$, which is uniquely identified by the coefficient $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$.

In order to identify $\hat{\boldsymbol{\beta}}$, we minimise the function $Q(\boldsymbol{\beta})$ by making use of differentiation.

*Remark 10.7.* Remember that, given two vectors

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \boldsymbol{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

such that $\boldsymbol{y} = {}^{\mathrm{t}}\boldsymbol{x}\boldsymbol{a}$,

$$\frac{d}{d\boldsymbol{x}}({}^{\mathrm{t}}\boldsymbol{x}\boldsymbol{a}) = {}^{\mathrm{t}}\left(\frac{d}{dx_1}({}^{\mathrm{t}}\boldsymbol{x}\boldsymbol{a}), \dots, \frac{d}{dx_n}({}^{\mathrm{t}}\boldsymbol{x}\boldsymbol{a})\right) = \boldsymbol{a}, \tag{10.4}$$

$$\frac{d}{d\,{}^{\mathrm{t}}\boldsymbol{x}}({}^{\mathrm{t}}\boldsymbol{x}\boldsymbol{a}) = \left(\frac{d}{dx_1}({}^{\mathrm{t}}\boldsymbol{x}\boldsymbol{a}), \dots, \frac{d}{dx_n}({}^{\mathrm{t}}\boldsymbol{x}\boldsymbol{a})\right) = \boldsymbol{a}, \tag{10.5}$$

$$\frac{d}{d\boldsymbol{x}}A\boldsymbol{x} = {}^{\mathrm{t}}A,$$

$$\frac{d}{d\boldsymbol{x}}\,{}^{\mathrm{t}}\boldsymbol{x}B\boldsymbol{x} = 2B\boldsymbol{x},$$

where $B$ has to be symmetric (note that in (10.4) we considered the transpose since $\boldsymbol{x}$ is a column vector, and in (10.5) we did not since ${}^{\mathrm{t}}\boldsymbol{x}$ is a row vector).

By using these two rules, we finally differentiate $Q(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$:

$$\begin{aligned}\frac{d}{d\boldsymbol{\beta}}Q(\boldsymbol{\beta}) &= \frac{d}{d\boldsymbol{\beta}}\,{}^{\mathrm{t}}(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu}) \\ &= \frac{d}{d\boldsymbol{\beta}}({}^{\mathrm{t}}\boldsymbol{y}\boldsymbol{y} - 2\,{}^{\mathrm{t}}\boldsymbol{y}X\boldsymbol{\beta} + {}^{\mathrm{t}}\boldsymbol{\beta}\,{}^{\mathrm{t}}XX\boldsymbol{\beta}) \\ &= \frac{d}{d\boldsymbol{\beta}}(-2\,{}^{\mathrm{t}}\boldsymbol{y}X\boldsymbol{\beta}) + \frac{d}{d\boldsymbol{\beta}}({}^{\mathrm{t}}\boldsymbol{\beta}\,{}^{\mathrm{t}}XX\boldsymbol{\beta}) \\ &= -2\,{}^{\mathrm{t}}X\boldsymbol{y} + 2\,{}^{\mathrm{t}}XX\boldsymbol{\beta} \\ &= 2({}^{\mathrm{t}}XX\boldsymbol{\beta} - {}^{\mathrm{t}}X\boldsymbol{y}).\end{aligned}$$

If we equate the previous expression to zero, then we deduce that the required minimum point satisfies

$${}^{\mathrm{t}}XX\boldsymbol{\beta} = {}^{\mathrm{t}}X\boldsymbol{y},$$

also known as **normal equation**. Observe that the inversion of the matrix ${}^{\mathrm{t}}XX$ is legittimate, since we are assuming the condition on the rank of $X$: recall that if $\operatorname{rk} X = p$, then $\operatorname{rk}{}^{\mathrm{t}}XX = p$. This implies that a point which minimises $Q(\boldsymbol{\beta})$ is

$$\hat{\boldsymbol{\beta}} = ({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X\boldsymbol{y}. \tag{10.6}$$

The estimates $\hat{\boldsymbol{\beta}}$ are called **regression coefficients**. To check that (10.6) gives a minimum point for $Q(\boldsymbol{\beta})$ we simply consider the matrix of the second derivatives

$$\frac{d}{d\boldsymbol{\beta}d\,{}^{\mathrm{t}}\boldsymbol{\beta}}Q(\boldsymbol{\beta}) = 2\,{}^{\mathrm{t}}XX,$$

which is a positive definite matrix (therefore $Q(\boldsymbol{\beta})$ is convex).

### 10.2.2 The Projection Matrix Operator

**Definition 10.8.** *Associated to $\hat{\boldsymbol{\beta}}$ there is the projection vector $\boldsymbol{y}$ onto $\mathcal{C}(X)$, which is called the **vector of fitted values**, that is*

$$\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}} = X({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X\boldsymbol{y} = P\boldsymbol{y},$$

*where $P \coloneqq X({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X$ is the **projection matrix operator** (or simply **projection matrix**) on $\mathcal{C}(X)$. The projection matrix defines another operator associated with the matrix $X$ whose role is to project any vector $\boldsymbol{y} \in \mathbb{R}^n$ and transform it into another vector, that we call $P\boldsymbol{y} \in \mathcal{C}(X)$: this minimises the distance from $\boldsymbol{y}$.*

*Remark 10.9.* $P$ has some interesting properties, that we will often use.

- $P$ is *symmetric*.
- $P$ is *idempotent*, that is $P^2\boldsymbol{y} = P\boldsymbol{y}$.
- $\operatorname{rk} P = \operatorname{tr} P = \operatorname{tr}(({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}XX) = p$.

We can decompose $\boldsymbol{y}$ into two terms:

1. the projection $\hat{\boldsymbol{\mu}}$ onto $\mathcal{C}(X)$;

2. the *residual component*

$$\boldsymbol{y} - \hat{\boldsymbol{\mu}} = \boldsymbol{y} - X({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X\boldsymbol{y} = (I_n - P)\boldsymbol{y}.$$

*Remark 10.10.* The equation ${}^{\mathrm{t}}XX\boldsymbol{\beta} = {}^{\mathrm{t}}X\boldsymbol{y}$ can be also obtained by a simple geometric argument. Indeed, the vector $\hat{\boldsymbol{\mu}}$ which minimises the distance from $\boldsymbol{y}$ is such that

$$(\hat{\boldsymbol{y}} - X\hat{\boldsymbol{\beta}}) \perp \mathcal{C}(X),$$

In particular, we have that $(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ is *orthogonal to all the elements in the subspace* $\mathcal{C}(X)$. Indeed, for every $X\boldsymbol{a} \in \mathcal{C}(X)$, we can write

$$\begin{aligned}
{}^{\mathrm{t}}(X\boldsymbol{a})(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) &= {}^{\mathrm{t}}(X\boldsymbol{a})(\boldsymbol{y} - P\boldsymbol{y}) \\
&= {}^{\mathrm{t}}\boldsymbol{a}\,{}^{\mathrm{t}}X(\boldsymbol{y} - X({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X\boldsymbol{y}) \\
&= {}^{\mathrm{t}}\boldsymbol{a}\,{}^{\mathrm{t}}X\boldsymbol{y} - {}^{\mathrm{t}}\boldsymbol{a}\,{}^{\mathrm{t}}XX({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X\boldsymbol{y} \\
&= {}^{\mathrm{t}}\boldsymbol{a}\,{}^{\mathrm{t}}X\boldsymbol{y} - {}^{\mathrm{t}}\boldsymbol{a}\,{}^{\mathrm{t}}X\boldsymbol{y} \\
&= \mathbf{0}.
\end{aligned}$$

In other terms, we need that

$${}^{\mathrm{t}}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})X = 0,$$

which is equivalent to the normal equation.

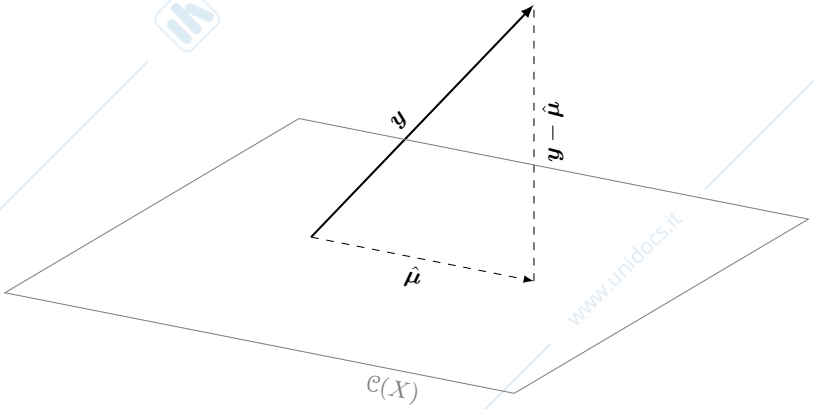



**Fig. 10.1.** Graphical representation of the projection onto $\mathcal{C}(X)$.

### 10.2.3 Estimation of $\hat{\boldsymbol{\beta}}$

Consider the estimates $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Can we say something about $\hat{\boldsymbol{\beta}}$? For instance, we have that

$$\begin{aligned}
\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] &= \mathbb{E}\left[({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X\boldsymbol{y}\right] \\
&= ({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X\mathbb{E}\left[\boldsymbol{y}\right] \\
&= ({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}XX\boldsymbol{\beta}
\end{aligned}$$

$$= \boldsymbol{\beta}.$$

The Least Squares estimator $\hat{\boldsymbol{\beta}}$ is **unbiased** for $\boldsymbol{\beta}$. Moreover, this implies that

$$\mathbb{E}\left[\hat{\boldsymbol{\mu}}\right] = \mathbb{E}\left[X\hat{\boldsymbol{\beta}}\right] = X\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = X\boldsymbol{\beta} = \boldsymbol{\mu}.$$

The dispersion matrix of the estimates is

$$\begin{aligned}
\operatorname{Var}\left(\hat{\boldsymbol{\beta}}\right) &= \operatorname{Var}\left(({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X\boldsymbol{y}\right) \\
&= ({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X\operatorname{Var}\left(\boldsymbol{y}\right){}^{\mathrm{t}}(({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X) \\
&= ({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X\sigma^2 I_n\,{}^{\mathrm{t}}(({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X) \\
&= \sigma^2 I_n({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}XX({}^{\mathrm{t}}XX)^{-1} \\
&= \sigma^2({}^{\mathrm{t}}XX)^{-1}
\end{aligned}$$

Also,

$$\begin{aligned}
\operatorname{Var}\left(\hat{\boldsymbol{\mu}}\right) &= X\operatorname{Var}\left(\boldsymbol{\beta}\right){}^{\mathrm{t}}X \\
&= \sigma^2 X({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X \\
&= \sigma^2 P
\end{aligned}$$

We can also check consistency. By the above variances, consistency holds if and only if the diagonal element of $V = ({}^{\mathrm{t}}XX)^{-1}$ converges to 0 as $n \to \infty$. This is determined by the design matrix $X$, whose number of rows must increase as $n \to \infty$. In general, it is impossible to assess consistency of $\hat{\boldsymbol{\beta}}$ in general terms without making some assumptions over the designed matrix.

*Remark 10.11.* So far, we have regarded the matrix $X$ as a non-stochastic matrix. However, it is very frequently the case that the response and explanatory variables are determined simultaneously. Therefore, we deal with stochastic explanatory variables. The most typical example occurs when two or more variables are observed on the same statistical unit, such as weight ($\boldsymbol{y}$) and height ($\boldsymbol{x}$).

The assumption of a non-stochastic matrix $X$ is supported by the following arguments. In most cases, the distribution of the explanatory variables does not contain any information on the relationship with the response variable. With a regression model, we are interested in making inference on these relationships, and not on the distribution of the explanatory variables. In some way, we examine the variables conditionally on the values taken by $X$.

## 10.3 29/11/18

### 10.3.1 Estimation of $\sigma^2$

So far, we considered the estimation of $\boldsymbol{\beta}$. However, we are also interested in estimating $\sigma^2$, but the Least Square criterion does not say how to estimate $\sigma^2$. Since the generic term $\varepsilon_i$ is such that $\mathrm{Var}(\varepsilon_i) = \sigma^2$, it is reasonable to estimate $\sigma^2$ by the arithmetic mean of $\hat{\varepsilon}_i$, where $\hat{\varepsilon}_i$ is the generic component of the residual vector

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{\mu}}.$$

Therefore, we can consider

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{n} \tag{10.7}$$

as an estimate of $\sigma^2$. Now,

$$\begin{aligned}\|\hat{\boldsymbol{\varepsilon}}\|^2 &= Q(\hat{\boldsymbol{\beta}}) \\ &= {}^{\mathrm{t}}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \\ &= {}^{\mathrm{t}}\boldsymbol{y}\,{}^{\mathrm{t}}(I_n - P)(I_n - P)\boldsymbol{y} \\ &= {}^{\mathrm{t}}\boldsymbol{y}(I_n - P)\boldsymbol{y},\end{aligned}$$

since $(I_n - P)$ is idempotent. We can compute the expectation of $\hat{\sigma}^2$ as follows, using (10.7):

$$\mathbb{E}\left[n\hat{\sigma}^2\right] = \mathbb{E}\left[{}^{\mathrm{t}}\boldsymbol{y}(I_n - P)\boldsymbol{y}\right].$$

In order to do the computation, recall Lemma 1.22; in our case,

$$\mathbb{E}\left[{}^{\mathrm{t}}\boldsymbol{y}(I_n - P)\boldsymbol{y}\right] = {}^{\mathrm{t}}\boldsymbol{\mu}(I_n - P)\boldsymbol{\mu} + \mathrm{tr}\left((I_n - P)\sigma^2 I_n\right):$$

now, the first addend in the right-hand side vanishes, since $(I_n - P)$ projects onto the orthogonal space $\mathcal{C}(X)$ which contains $\boldsymbol{\mu}$, and hence

$$(I_n - P)\boldsymbol{\mu} = (I_n - X({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X)X\boldsymbol{\beta} = 0.$$

This implies that

$$\mathbb{E}\left[{}^{\mathrm{t}}\boldsymbol{y}(I_n - P)\boldsymbol{y}\right] = \sigma^2(n - p):$$

we proved that $\mathbb{E}\left[n\hat{\sigma}^2\right] = \sigma^2(n-p)$, which implies that the estimator $\hat{\sigma}^2$ is *biased* for $\sigma^2$, but it is *asymptotically unbiased* for $\sigma^2$.

*Remark 10.12.* We could easily find an unbiased estimator for $\sigma^2$ by using the relation

$$\mathbb{E}\left[n\hat{\sigma}^2\right] = (n-p)\sigma^2.$$

This estimator is

$$\hat{\sigma}^2\frac{n}{n-p} = \frac{Q(\hat{\boldsymbol{\beta}})}{n-p}.$$

*Example 10.13.* The simplest case that we can consider has $p = 1$ and $\boldsymbol{X} = \mathbf{1}_n$. Then, the linear model is the following:

$$\boldsymbol{Y} = \mathbf{1}_n \beta + \boldsymbol{\varepsilon},$$

where $\beta$ is a scalar. This is just an unusual way of introducing a model where all components of $\boldsymbol{Y}$ have the same mean and the same variance, and are uncorrelated. In this scenario,

$$({}^{\mathrm{t}}XX)^{-1} = \frac{1}{n}, \quad {}^{\mathrm{t}}X\boldsymbol{y} = \sum_{i=1}^{n} \boldsymbol{y}, \quad \hat{\boldsymbol{\beta}} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

*Example 10.14.* We consider $p = 2$, with the first column of $\boldsymbol{X}$ equal to $\mathbf{1}_n$, and the second column of $\boldsymbol{X}$ equal to $\boldsymbol{x} = {}^{\mathrm{t}}(x_1, \dots, x_n)$, meaning

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

In this situation, the linear regression model

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \beta_1 + \beta_2 \boldsymbol{x} + \boldsymbol{\varepsilon}$$

is called **simple linear regression model**. Here,

$$\begin{aligned} {}^{\mathrm{t}}XX &= \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} \\ ({}^{\mathrm{t}}XX)^{-1} &= \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \\ {}^{\mathrm{t}}X\boldsymbol{y} &= \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix} \\ \hat{\boldsymbol{\beta}} &= \begin{pmatrix} \overline{y} - \frac{s_{xy}}{s_{xx}}\overline{x} \\ \frac{s_{xy}}{s_{xx}} \end{pmatrix} \end{aligned} \tag{10.8}$$

where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^n y_i, \quad s_{xx} = \sum_{i=1}^n (x_i - \overline{x})^2, \quad s_{xy} = \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}).$$

How did we obtain (10.8)? Indeed,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= ({}^{\mathrm{t}}XX)^{-1}({}^{\mathrm{t}}X\boldsymbol{y}) \\ &= \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \cdot \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{pmatrix} \\ &= \frac{1}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i \\ n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i \end{pmatrix}. \end{aligned}$$

In order to show that this is equal to

$$\begin{pmatrix} \overline{y} - \frac{s_{xy}}{s_{xx}}\overline{x} \\ \frac{s_{xy}}{s_{xx}} \end{pmatrix}$$

first notice that

$$\begin{aligned} s_{xx} &= \sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \sum_{i=1}^{n}\left(x_i^2 - 2x_i\overline{x} + \overline{x}^2\right) \\ &= \sum_{i=1}^{n} x_i^2 - 2\left(\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 \\ &= \sum_{i=1}^{n} x_i^2 - \frac{2}{n}\left(\sum_{i=1}^{n} x_i\right)^2 + \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \\ &= \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \\ &= \frac{1}{n}, \end{aligned}$$

while

$$\begin{aligned} s_{xy} &= \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \\ &= \sum_{i=1}^{n}(x_i y_i - x_i\overline{y} - \overline{x}y_i + \overline{xy}) \\ &= \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) \\ &= \frac{1}{n}. \end{aligned}$$

Finally,

$$\begin{aligned} \overline{y} - \frac{s_{xy}}{s_{xx}}\overline{x} &= \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\ &= \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} y_i \left(\sum_{i=1}^{n} x_i\right)^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i + \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ &= \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}. \end{aligned}$$

The expression for $\hat{\beta}$ simplifies if $\overline{x} = 0$. This condition can always be satisfied by changing the parametrisation from

$$\boldsymbol{y} = \mathbf{1}_n\beta_1 + x\beta_2 + \boldsymbol{\varepsilon}$$

to

$$\boldsymbol{y} = \mathbf{1}_n(\beta_1 + \beta_2\overline{x}) + (x - \mathbf{1}_n\overline{x})\beta_2 + \boldsymbol{\varepsilon}$$

which is an equation that has

- *intercept* $\alpha = \beta_1 + \beta_2\overline{x}$;
- *slope* $\beta_2$.

Under the new parametrisation, the new explanatory variables $\boldsymbol{z} = \boldsymbol{x} - \mathbf{1}_n\overline{x}$ has $\sum_{i=1}^{n} z_i = 0$.

### 10.3.2 Sample Correlation and Coefficient of Determination

The *sample correlation* is defined as

$$r \coloneqq \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}},$$

where

$$s_{yy} \coloneqq \sum_{i=1}^{n}(y_i - \overline{y})^2.$$

This quantity is typically used as a descriptive statistic to quantify the degree of closeness of the observed data points to the fitted line. In fact, $r \in [-1, 1]$; if $r = \pm 1$, this means that the fitted point lie exactly on the regression line. On the other hand, $r = 0$ denotes a poor fit of the regression line to the points.
A related quantity is the *coefficient of determination*, which is defined as a function of the sample correlation

$$r^2 \coloneqq \frac{s_{xy}^2}{s_{xx}s_{yy}},$$

which represents the fracture of variability of the $y_i$'s that can be imputed to the linear component of the relationship.

*Remark 10.15.* One of the second order hypotheses is that the error component $\boldsymbol{\varepsilon}$ must have $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$. In a large number of examples, this assumption does not work, but at the same time it does not represent an issue. In particular, suppose that one column of the designed matrix $X$, say the first, is $\mathbf{1}_n$. Denote by $\mu_{\varepsilon_i}$ the mean value of $\varepsilon_i$, and by $\overline{X}$ and $\overline{\boldsymbol{\beta}}$ the remaining column of $X$ and the remaining rows of $\boldsymbol{\beta}$ after deleting the first. Then, we can write our linear model

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

as

$$\boldsymbol{Y} = \mathbf{1}_n\beta_1 + \overline{X}\,\overline{\boldsymbol{\beta}} + \boldsymbol{\varepsilon} = \mathbf{1}_n(\beta_1 + \mu_{\varepsilon_i}) + \overline{X}\,\overline{\boldsymbol{\beta}} + (\boldsymbol{\varepsilon} - \mathbf{1}_n\mu_{\varepsilon_i}),$$

where the error component is $\boldsymbol{\varepsilon} - \mathbf{1}_n\mu_{\varepsilon_i}$ which has expected value equal to zero. Therefore, **the assumption on the mean value of $\boldsymbol{\varepsilon}$ is not a relevant assumption**.

**Theorem 10.16 (Gauss-Markov).** *Under the second order assumptions for the linear model $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the Least Squares estimator $\hat{\boldsymbol{\beta}}$ is such that*

$$\operatorname{Var}\left(\hat{\boldsymbol{\beta}}\right) \leq \operatorname{Var}(\boldsymbol{T}),$$

*where $T$ is any estimator of the form*

$$\boldsymbol{T} = {}^{\mathrm{t}}C\boldsymbol{Y}$$

*and $C$ is an $n \times p$ matrix such that $\mathbb{E}[\boldsymbol{T}] = \boldsymbol{\beta}$, for every $\boldsymbol{\beta}$.*

*Proof.* Since the equality

$$\mathbb{E}[{}^{\mathrm{t}}C\boldsymbol{Y}] = {}^{\mathrm{t}}CX\boldsymbol{\beta} = \boldsymbol{\beta}$$

for every $\boldsymbol{\beta}$, it follows that

$${}^{\mathrm{t}}CX = I_p = {}^{\mathrm{t}}XC.$$

Now, recall that the variance of $\hat{\boldsymbol{\beta}}$ is

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2({}^{\mathrm{t}}XX)^{-1}$$

and since

$$\mathrm{Var}(\boldsymbol{T}) = \mathrm{Var}({}^{\mathrm{t}}C\boldsymbol{Y}) = C\,\mathrm{Var}(\boldsymbol{Y})\,{}^{\mathrm{t}}C = \sigma^2\,{}^{\mathrm{t}}CC,$$

we have that the statement of the Theorem is equivalent to prove that

$${}^{\mathrm{t}}CC - ({}^{\mathrm{t}}XX)^{-1} \geq 0.$$

Since ${}^{\mathrm{t}}CX = I_p = {}^{\mathrm{t}}XC$, we must show

$$\begin{aligned} {}^{\mathrm{t}}CC - {}^{\mathrm{t}}CX({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}XC &= {}^{\mathrm{t}}C[C - X({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}XC] \\ &= {}^{\mathrm{t}}C(I_n - P)C \\ &\geq 0. \end{aligned}$$

Indeed, we have

$$\begin{aligned} {}^{\mathrm{t}}\boldsymbol{a}\,{}^{\mathrm{t}}C(I_n - P)C\boldsymbol{a} &= {}^{\mathrm{t}}\boldsymbol{a}\,{}^{\mathrm{t}}C\,{}^{\mathrm{t}}(I_n - P)(I_n - P)C\boldsymbol{a} \\ &= \|(I_n - P)C\boldsymbol{a}\|^2 \\ &\geq 0, \quad \forall \boldsymbol{a} \in \mathbb{R}^p. \end{aligned}$$

We proved that in the class of *linear unbiased estimators* for the parameters of a linear model $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the Least Squares estimator $({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X\boldsymbol{y}$ is the one with the minimum variance.

□

# 11

# Week XI

## 11.1 04/12/18

*Remark 11.1.* An implication of the Gauss-Markov Theorem is the following: consider a linear combination of the parameter $\vartheta$, usually known as *contrast*, of the form

$$\psi = {}^{\mathrm{t}}\boldsymbol{a}\boldsymbol{\beta},$$

where $\boldsymbol{a} \in \mathbb{R}^p$ is a vector of given constants. It follows that the estimator $\hat{\psi} = {}^{\mathrm{t}}\boldsymbol{a}\hat{\boldsymbol{\beta}}$ is unbiased for $\psi$ and its variance is the minimum achievable variance among linear estimators.

In particular, if we choose a vector $\boldsymbol{a}$ with null components except the component of order $j$-th, we have that $\hat{\psi} = \hat{\beta}_j$ is the best linear estimator for the component $j$-th.

*Remark 11.2.* The Gauss-Markov Theorem states that, in the class of unbiased linear estimators of $\boldsymbol{\beta}$, the Least Squares estimator has the minimum variance. Therefore,

$$\hat{\boldsymbol{\beta}} = ({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X\boldsymbol{y}$$

is the best linear estimator. This provides with the (unique) theoretical guarantee for the Least Squares approach.

### 11.1.1 Constrained Estimates

#### Sums-of-squares Decomposition

We discussed about the orthogonality of $P\boldsymbol{y}$ and of the residual vector. If we write the squared norm of $\hat{\boldsymbol{\mu}} + (\boldsymbol{y} - \hat{\boldsymbol{\mu}})$, then by orthogonality

$$\|\boldsymbol{y}\|^2 = \|\hat{\boldsymbol{\mu}}\|^2 + \|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2 .$$

Now, if we replace $\boldsymbol{y}$ with the corresponding random variable, we have that

$$\|\boldsymbol{Y}\|^2 = \|P\boldsymbol{Y}\|^2 + \|(I_n - P)\boldsymbol{Y}\|^2 ,$$

or alternatively

$$ {}^{\mathrm{t}}\boldsymbol{Y} I_n \boldsymbol{Y} = {}^{\mathrm{t}}\boldsymbol{Y} P \boldsymbol{Y} + {}^{\mathrm{t}}\boldsymbol{Y}(I_n - P)\boldsymbol{Y}. $$

These are quadratic forms with associated (idempotent) matrices

$$ I_n, \quad P, \quad (I_n - P), $$

and such that

$$ \operatorname{rk} I_n = n, \quad \operatorname{rk} P = p, \quad \operatorname{rk}(I_n - P) = n - p. $$

The above quadratic form (on the above square norms) are called

- $\|\boldsymbol{Y}\|^2$ *total sum of squares*;
- $\|P\boldsymbol{Y}\|^2$ *regression sum of squares*;
- $\|(I_n - P)\boldsymbol{Y}\|^2$ *residual sum of squares.*

It is important to study the distribution of these quantities. However, this study is possible only if we specify a distributional assumption on $\boldsymbol{\varepsilon}$. For the moment, with the second order assumptions, we can only compute the expectation of this quadratic form by applying Lemma 1.22:

$$ \begin{aligned} \mathbb{E}\left[\|P\boldsymbol{Y}\|^2\right] &= p\sigma^2 + {}^{\mathrm{t}}\boldsymbol{\mu} P \boldsymbol{\mu}, \\ \mathbb{E}\left[\|(I_n - P)\boldsymbol{Y}\|^2\right] &= (n-p)\sigma^2, \quad \text{since } {}^{\mathrm{t}}\boldsymbol{\mu}(I_n - P)\boldsymbol{\mu} = 0 \\ \mathbb{E}\left[\|\boldsymbol{Y}\|^2\right] &= n\sigma^2 + {}^{\mathrm{t}}\boldsymbol{\mu}\boldsymbol{\mu}. \end{aligned} $$

### 11.1.2 Constrained Optimisation: Lagrange Multipliers

Consider the following problem: we want to estimate $\boldsymbol{\beta}$ in the linear model when there exist some linear constraints among the components of $\boldsymbol{\beta}$. For instance, imagine that the vector $\boldsymbol{\beta}$ is such that

$$ H\boldsymbol{\beta} = \mathbf{0}, $$

where $H$ is a $q \times p$ ($q \ll p$) matrix of rank $q$, whose elements are given constants. The solution of this constrained estimation problem is particularly useful in the contest of hypotheses testing of the parameter $\boldsymbol{\beta}$. With respect to the Least Squares approach, we are dealing with the additional condition

$$ H\boldsymbol{\beta} = \mathbf{0}. $$

This condition constraints $\boldsymbol{\beta}$ to stay in a certain subset of the linear space $\mathcal{C}(X)$, say $\mathcal{C}_0(X)$. It is easy to show that $\mathcal{C}_0(X)$ is a linear subspace of $\mathcal{C}(X)$, with dimension $p - q$.

We can solve the problem of constrained optimisation by using *Lagrange Multipliers*. In other terms, we want to minimise the function $Q(\boldsymbol{\beta})$ under $H\boldsymbol{\beta} = 0$. Basically, we must minimise the function

$$ f(\boldsymbol{\alpha}, \boldsymbol{\beta}) \coloneqq {}^{\mathrm{t}}(\boldsymbol{y} - X\boldsymbol{\beta})(\boldsymbol{y} - X\boldsymbol{\beta}) + 2\,{}^{\mathrm{t}}(H\boldsymbol{\beta})\boldsymbol{\alpha}, $$

where $\boldsymbol{\alpha}$ is a vector of Lagrange multipliers. By differentiating the function $f$ with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and equating these expressions to zero, we obtain the following system of equations:

$$\begin{cases} \frac{\partial}{\partial \boldsymbol{\beta}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}) = {}^{\mathrm{t}}XX\boldsymbol{\beta} + {}^{\mathrm{t}}H\boldsymbol{\alpha} = {}^{\mathrm{t}}X\boldsymbol{y} \\ \frac{\partial}{\partial \boldsymbol{\alpha}} f(\boldsymbol{\alpha}, \boldsymbol{\beta}) = H\boldsymbol{\beta} = \mathbf{0} \end{cases}.$$

Consider the first equation and multiply it by $({}^{\mathrm{t}}XX)^{-1}$. Then, we can isolate $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} = ({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X\boldsymbol{y} - ({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}H\boldsymbol{\alpha} = \hat{\boldsymbol{\beta}} - ({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}H\boldsymbol{\alpha} \tag{11.1}$$

and (11.1) leads, multiplying by $H$, to

$$0 = H\boldsymbol{\beta} = H\hat{\boldsymbol{\beta}} - H({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}H\boldsymbol{\alpha},$$

thus isolating $\boldsymbol{\alpha}$ we obtain

$$\boldsymbol{\alpha} = [H({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}H]^{-1}H\hat{\boldsymbol{\beta}}.$$

Therefore, the minimum of the function $f$ is obtained for $\boldsymbol{\beta}$ equal to

$$\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}} - ({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}HKH\hat{\boldsymbol{\beta}},$$

where $K \coloneqq [H({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}H]^{-1}$. The vector $\hat{\boldsymbol{\beta}}_0$ is the estimator of $\boldsymbol{\beta}$ under $H\boldsymbol{\beta} = \mathbf{0}$. This implies that we have a fitted value with respect to $\hat{\boldsymbol{\beta}}_0$ which lives in $\mathcal{C}_0(X)$. The fitted value is obtained as follows:

$$\begin{aligned} \hat{\boldsymbol{\mu}}_0 &= X\hat{\boldsymbol{\beta}}_0 \\ &= X[\hat{\boldsymbol{\beta}} - ({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}HKH\hat{\boldsymbol{\beta}}] \\ &= \hat{\boldsymbol{\mu}} - X({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}HKH\hat{\boldsymbol{\beta}} \\ &= P\boldsymbol{y} - X({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}HKH({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X\boldsymbol{y} \\ &= (P - P_H)\boldsymbol{y} \\ &= P_0\boldsymbol{y}, \end{aligned}$$

where

$$P_0 \coloneqq P - P_H, \quad P_H \coloneqq X({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}HKH({}^{\mathrm{t}}XX)^{-1}\,{}^{\mathrm{t}}X.$$

We obtained a new projection matrix $P_0$ that applies to $\boldsymbol{y}$ in order to obtain the fitted value $\hat{\boldsymbol{\mu}}_0$ that minimises the distance with respect to $\boldsymbol{y}$ and lives in $\mathcal{C}_0(X)$.

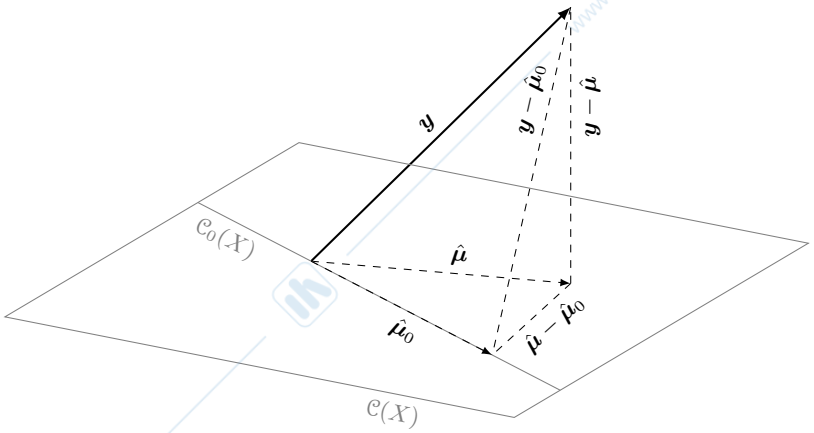



**Fig. 11.1.** Graphical representation of the projection onto $\mathcal{C}(X)$, taking $\mathcal{C}_0(X)$ into account.

**Properties**

We obtained a new projection matrix $P_0$ that projects any vector of $\mathbb{R}^n$ into the linear subspace $\mathcal{C}_0(X)$. The following properties of the projection hold.

1. The vector $\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0$ is orthogonal to all the elements of $\mathcal{C}_0(X)$. Indeed, consider $\boldsymbol{c}$ to be any element of $\mathcal{C}_0(X)$ such that $H\boldsymbol{c} = 0$; then, we have
$$^{\mathrm{t}}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0)X\boldsymbol{c} = 0,$$
and in particular
$$(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \perp \hat{\boldsymbol{\mu}}_0.$$
2. The projection of $\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0$ onto $\mathcal{C}(X)$ is
$$P(\boldsymbol{y} - \hat{\boldsymbol{\mu}}_0) = \hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_0,$$
such that
$$\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_0 \perp \hat{\boldsymbol{\mu}}_0.$$
We can have a decomposition of $\boldsymbol{y}$ which is similar to the one obtained without constraint on $\boldsymbol{\beta}$:
$$\boldsymbol{y} = \hat{\boldsymbol{\mu}}_0 + (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_0) + (\boldsymbol{y} - \hat{\boldsymbol{\mu}}),$$
where the three elements on the right-hand side are orthogonal and then we have an even more general version of Pitagora's Theorem:
$$\|\boldsymbol{y}\|^2 = \|\hat{\boldsymbol{\mu}}_0\|^2 \|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_0\|^2 + \|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2,$$
which extends the one used above.

### 11.1.3 Gaussian Theory

So far, we considered the linear model under the second order assumptions on $\boldsymbol{\varepsilon}$. However, it is natural to specify a distributional assumption on $\boldsymbol{\varepsilon}$, and then make use of the theory developed in the first part of this course. The most natural (and simplest) assumption is to consider $\boldsymbol{\varepsilon}$ to be Gaussian. In particular, we assume that
$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 I_n),$$
therefore
$$\boldsymbol{Y} \sim \mathcal{N}_n(X\boldsymbol{\beta}, \sigma^2 I_n).$$
We have a distributional assumption for $\boldsymbol{Y}$, that is, we have a parametric model for the observable values. Let $\boldsymbol{\vartheta} = {}^{\mathrm{t}}({}^{\mathrm{t}}\boldsymbol{\beta}, \sigma^2)$. Then, the Log-Likelihood function is
$$\begin{aligned}\log \mathcal{L}(\vartheta, \boldsymbol{y}) &= -\frac{1}{2} n \log \sigma^2 - \frac{1}{2}\sigma^{-2} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 \\ &= -\frac{1}{2} n \log \sigma^2 - \frac{1}{2}\sigma^{-2}({}^{\mathrm{t}}\boldsymbol{y}\boldsymbol{y} - 2\,{}^{\mathrm{t}}\boldsymbol{y}X\boldsymbol{\beta} + {}^{\mathrm{t}}\boldsymbol{\beta}\,{}^{\mathrm{t}}XX\boldsymbol{\beta}).\end{aligned}$$
This Log-Likelihood function has a clear exponential structure, such that we can identify the sufficient statistic:

$$^{t}(^{t}\boldsymbol{y}X, {}^{t}\boldsymbol{y}\boldsymbol{y}),$$

which is a vector of dimension $p+1$. In the Log-Likelihood function, we can identify the function $Q(\boldsymbol{\beta})$, which is the same quantity used by the Least Squares approach. This implies that, for $\sigma^2$ fixed, the maximum of the Log-Likelihood function coincides with the minimum of $Q(\boldsymbol{\beta})$. We conclude that **the Maximum Likelihood estimator of $\boldsymbol{\beta}$ coincides with the Least Squares estimator of $\boldsymbol{\beta}$**,

$$\hat{\boldsymbol{\beta}} = (^{t}XX)^{-1}\, {}^{t}X\boldsymbol{y}.$$

This happens because we assumed that $\boldsymbol{\varepsilon}$ is a multi-dimensional Gaussian, and then $\boldsymbol{Y}$ is a multi-dimensional Gaussian. We can also maximise the Log-Likelihood function with respect to $\sigma^2$: if we did that, we obtain that the Maximum Likelihood estimator for $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{n}.$$

Recall that $\hat{\sigma}^2$ is a biased estimator for $\sigma^2$. We can fix this issue and consider the unbiased estimator

$$\hat{\sigma}^2 \frac{n}{n-p}.$$

## 11.2 05/12/18

### 11.2.1 The Fisher-Cochran Theorem

**Theorem 11.3.** *Consider $\boldsymbol{Y} = (Y_1, \dots, Y_k) \sim \mathcal{N}_k(\boldsymbol{\mu}, I_k)$ and let $A_1, \dots, A_m$ be positive semidefinite matrices of ranks $r_1, \dots, r_m$, respectively. Assume that*

$$\sum_{i=1}^{m} A_i = I_k.$$

*Then, the following are equivalent:*

1. *the quadratic form*
$$Q_j = {}^{\mathrm{t}}\boldsymbol{Y} A_j \boldsymbol{Y}, \quad j = 1, \dots, m$$
*are distributed as non-centred Chi-squared distributions, with degrees of freedom $r_j$ and parameter of non-centrality ${}^{\mathrm{t}}\boldsymbol{\mu} A_j \boldsymbol{\mu}$. Moreover, $Q_i$ is independent of $Q_j$, for $j \neq i$;*
2. *it holds that*
$$r_1 + \dots + r_m = k.$$

*Proof.* (1) $\Rightarrow$ (2) We have that

$${}^{\mathrm{t}}\boldsymbol{Y}\boldsymbol{Y} = {}^{\mathrm{t}}\boldsymbol{Y}(A_1 + \dots + A_m)\boldsymbol{Y} = Q_1 + \dots + Q_m \sim \chi^2_r(\delta),$$

where

$$r = \sum_{j=1}^{m} r_j \quad \text{and} \quad \delta = {}^{\mathrm{t}}\boldsymbol{\mu} \left( \sum_{j=1}^{m} A_j \right) \boldsymbol{\mu} = {}^{\mathrm{t}}\boldsymbol{\mu}\boldsymbol{\mu},$$

$\boldsymbol{\delta}$ also known as the *non-centrality parameter*. Therefore, ${}^{\mathrm{t}}\boldsymbol{Y}\boldsymbol{Y} \sim \chi^2_k({}^{\mathrm{t}}\boldsymbol{\mu}\boldsymbol{\mu})$, meaning that $r = k$.
(2) $\Rightarrow$ (1) We can write $A_j = B_j \, {}^{\mathrm{t}}B_j$, where $B_j$ is a matrix of order $k \times r_j$. We can do that by taking a square root of $A_j$ and then removing the columns of the null eigenvalues. Define

$$B = (\boldsymbol{B}_1, \dots, \boldsymbol{B}_m),$$

which is by hypothesis a square matrix, so that

$$B \, {}^{\mathrm{t}}B = \boldsymbol{B}_1 \, {}^{\mathrm{t}}\boldsymbol{B}_1 + \dots + \boldsymbol{B}_m \, {}^{\mathrm{t}}\boldsymbol{B}_m = A_1 + \dots + A_m = I_k.$$

Therefore, the matrix $B$ is orthogonal. Now, define a random vector $\boldsymbol{Z}$ as

$$\boldsymbol{Z} = {}^{\mathrm{t}}B\boldsymbol{Y}.$$

This means that $\boldsymbol{Z} \sim N_k({}^{\mathrm{t}}B\boldsymbol{\mu}, I_k)$. But then we have

$$Q_j = {}^{\mathrm{t}}\boldsymbol{Y} A_j \boldsymbol{Y} = {}^{\mathrm{t}}\boldsymbol{Y} B_j \, {}^{\mathrm{t}}B_j \boldsymbol{Y} = {}^{\mathrm{t}}\boldsymbol{Z}_j \boldsymbol{Z}_j,$$

where $\boldsymbol{Z} = ({}^{\mathrm{t}}Z_1, \dots, {}^{\mathrm{t}}Z_m)$. The $\boldsymbol{Z}_j$'s are independent and each $\boldsymbol{Z}_j$ is distributed as $N_{r_j} \sim ({}^{\mathrm{t}}B_j\boldsymbol{\mu}, I_{r_j})$. This implies that

$$Q_j = {}^{\mathrm{t}}\boldsymbol{Z}_j \boldsymbol{Z}_j \sim \chi^2_{r_j}({}^{\mathrm{t}}\boldsymbol{\mu} A_j \boldsymbol{\mu})$$

and $Q_j$ independent of $Q_i$, for $i \neq j$.

□

**Corollary 11.4.** *If $\boldsymbol{Y} \sim N_k(\boldsymbol{\mu}, I_k)$ and $A$ is a symmetric idempotent matrix of order $k$, then*

$$ {}^{\mathrm{t}}\boldsymbol{Y} A \boldsymbol{Y} \sim \chi^2_r({}^{\mathrm{t}}\boldsymbol{\mu} A \boldsymbol{\mu}), $$

*where $r = \operatorname{rk} A = \operatorname{tr} A$.*

*Proof.* The proof is an application of Theorem 11.3, with $m = 2$ and

$$ A_1 = A, \quad A_2 = I_k - A. $$

□

We introduced the decomposition

$$ {}^{\mathrm{t}}\boldsymbol{Y} I_n \boldsymbol{Y} = {}^{\mathrm{t}}\boldsymbol{Y} P \boldsymbol{Y} + {}^{\mathrm{t}}\boldsymbol{Y}(I_n - P)\boldsymbol{Y}. $$

Assuming the second order hypotheses, we only found the expectation of the above three terms. Now, since we have distributional assumptions on $\boldsymbol{\varepsilon}$, we can determine the distributions of ${}^{\mathrm{t}}\boldsymbol{Y}\boldsymbol{Y}$, ${}^{\mathrm{t}}\boldsymbol{Y}P\boldsymbol{Y}$ and ${}^{\mathrm{t}}\boldsymbol{Y}(I_n - P)\boldsymbol{Y}$. These distributions follow from Theorem 11.3

- the *total residual*
$$ {}^{\mathrm{t}}\boldsymbol{Y}\boldsymbol{Y} \sim \sigma^2 \chi^2_n(\delta), $$
where $\delta = \frac{{}^{\mathrm{t}}\boldsymbol{\mu}\boldsymbol{\mu}}{\sigma^2}$;
- the *regression residual*
$$ {}^{\mathrm{t}}\boldsymbol{Y}P\boldsymbol{Y} \sim \sigma^2 \chi^2_p(\delta) $$
- the *residual*
$$ {}^{\mathrm{t}}\boldsymbol{Y}(I_n - P)\boldsymbol{Y} \sim \sigma^2 \chi^2_{n-p}(0), $$
since ${}^{\mathrm{t}}\boldsymbol{\mu}(I_n - P)\boldsymbol{\mu} = \mathbf{0}$.

Finally, we have that ${}^{\mathrm{t}}\boldsymbol{Y}P\boldsymbol{Y}$ is independent of ${}^{\mathrm{t}}\boldsymbol{Y}(I_n - P)\boldsymbol{Y}$.

*Remark 11.5.* From the perspective of point estimation, the use of the Maximum Likelihood approach does not make any difference, therefore we have the same estimator. The major improvement in introducing distributional hypotheses on $\boldsymbol{\varepsilon}$ is related to hypothesis testing on $\boldsymbol{\beta}$.

The Least Squares approach does not provide any insight on testing hypotheses on the parameter $\boldsymbol{\beta}$ and/or $\sigma^2$.

*Example 11.6.* Consider the problem of testing if the parameter $\boldsymbol{\beta}$ is equal or different from zero:

$$ H_0 : \boldsymbol{\beta} = \mathbf{0}, \quad H_1 : \boldsymbol{\beta} \neq \mathbf{0}. $$

To do that, we can apply the Likelihood Ratio test. Note that the parameter $\sigma^2$ is assumed to be unknown.

- Under $H_0$, the maximum of the Log-Likelihood function is achieved at the point
$$ \boldsymbol{\vartheta} = \begin{pmatrix} 0 \\ \hat{\sigma}^2_0 \end{pmatrix} \qquad \hat{\sigma}^2_0 = \frac{\|\boldsymbol{Y}\|^2}{n}, $$
as we have seen before (notice that $\|\boldsymbol{y}\|^2 = \|\boldsymbol{\varepsilon}\|^2$ when $\boldsymbol{\beta} = \mathbf{0}$).

· The unconstrained maximum value of the Log-Likelihood function is achieved at the point

$$\boldsymbol{\vartheta} = \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\sigma}^2 \end{pmatrix},$$

with $\hat{\boldsymbol{\beta}}$ obtained by the Least Squares approach.

The Log-Likelihood ratio criterion leads to the following test statistic

$$\lambda(\boldsymbol{y}) = \frac{(\hat{\sigma}_0^2)^{-n/2} \exp\left\{-\frac{1}{2}\hat{\sigma}_0^2 \left\|\boldsymbol{y}\right\|^2\right\}}{(\hat{\sigma}^2)^{-n/2} \exp\left\{-\frac{1}{2}\hat{\sigma}^2 \left\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\right\|^2\right\}} \propto \left(\frac{\left\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\right\|^2}{\left\|\boldsymbol{y}\right\|^2}\right)^{n/2}.$$

The Likelihood ratio $\lambda$ can be replaced by the monotone function

$$\lambda^*(\boldsymbol{y}) = \lambda^{-2/n}(\boldsymbol{y}) - 1 = \frac{\left\|\hat{\boldsymbol{\mu}}\right\|^2}{\boldsymbol{y} - \hat{\boldsymbol{\mu}}^2} = \frac{\frac{\left\|P\boldsymbol{Y}\right\|^2}{p}}{\frac{\left\|(I_n - P)\boldsymbol{Y}\right\|^P}{n-p}}.$$

According to the Likelihood ratio test, we reject the null hypothesis if $\lambda$ is sufficiently small or, in other terms, if $\lambda^*$ is sufficiently large. In order to do that, we have to determine the distribution of $\lambda^*$:

$$\lambda^*(\boldsymbol{y}) = \frac{\left\|\hat{\boldsymbol{\mu}}\right\|^2}{\left\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\right\|^2} = \frac{{}^{\mathrm{t}}\boldsymbol{Y} P \boldsymbol{Y}}{{}^{\mathrm{t}}\boldsymbol{Y} \left(I_n - P\right) \boldsymbol{Y}}.$$

The terms on the right-hand side have a Chi-squared distribution up to a multiplicative constant $\sigma^2$. Moreover, the terms $\left\|\hat{\boldsymbol{\mu}}\right\|^2$ and $\left\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\right\|^2$ are independent. This means that

$$F = \lambda^* \left(\frac{n-p}{p}\right)$$

is a statistic distributed as a Fisher distribution with $(p, n-p)$ degrees of freedom and non-centrality parameter

$$\delta = \frac{{}^{\mathrm{t}}\boldsymbol{\beta}\, {}^{\mathrm{t}}X X \boldsymbol{\beta}}{\sigma^2}.$$

Note that the non-centrality parameter vanishes under $H_0$.

# 12

# Week XII

## 12.1 11/12/18

### 12.1.1 Generalised Linear Models

What are the limitations of linear models? There are some kind of problems that cannot be reduced to the form required by the linear model. This might happen in different scenarios.

1. The relationship between the response and the explanatory variables is of the type
$$y = r(x_1, \dots, x_p) + \varepsilon,$$
where the function $r(\cdot)$ is non-linear in the parameters. In this case, linearisation does not work.
2. In addition, even the mathematical form of $r(\cdot)$ might not be known exactly. We often know enough about the nature of the problem. The most common example is related to the range of $y$. For instance, suppose that $y$ is the maximum fraction of skin which is affected by a certain disease and the explanatory variables are some factor for these particular diseases. In this case, $y \in [0, 1]$ but a fixed regression hyperplane cannot satisfy this constraint.
3. The variance of the error term, and then the response variable, should be constant while in practical situations the observed data does not satisfy this requirement.
4. In linear models, we assumed that the response variable is distributed according to a Gaussian distribution. There are distributions in which this is not the case: *binary response*, *counting response*, *positive response*.

We consider a new class of models to deal with the above situations. This class is called **Generalised Linear Models**. They rely on the exponential family.

For this purpose, we introduce a specific notation for the exponential family, different from the one we used before.

**Definition 12.1.** *A random variable $Y$ is in the **exponential family**, meaning*
$$Y \sim EF\left(b(\vartheta), \frac{\psi}{w}\right),$$

*if the probability density function (or the probability mass function) has the following form:*

$$f_Y(y) = \exp\left\{\frac{w}{\psi}[y\vartheta - b(\vartheta)] + c(y,\psi)\right\}, \tag{12.1}$$

*where $\vartheta$ and $\psi$ are scalar parameters, $w$ is a known constant, whilst $b(\cdot)$ and $c(\cdot,\cdot)$ are fixed functions that determine the specific parametric family of distributions.*

For any given choice of the **dispersion parameter** $\psi$, (12.1) forms an exponential family with parameter $\vartheta$. However, (12.1) is not an exponential family when $\vartheta$ and $\psi$ vary simultaneously. The variability of the random variable $Y$ distributed as (12.1) is not controlled by a single parameter, but by the ratio $\psi/w$ instead. The parameter $w$ is known as the **weight**.

**Proposition 12.2.** *Given a random variable $Y$ such that*

$$Y \sim EF\left(b(\vartheta), \frac{\psi}{w}\right),$$

*then*

$$\mathbb{E}[Y] = b'(\vartheta) \quad \mathrm{Var}(Y) = b''(\vartheta).$$

*Proof.* Take

$$\log(f_Y(y)) = \frac{w}{\psi}[y\vartheta - b(\vartheta)] + c(y,\psi)$$

and consider

$$\begin{aligned}\frac{\partial}{\partial\vartheta}\log(f_Y(y)) &= \frac{\partial}{\partial\vartheta}\left(\frac{(y\vartheta - b(\vartheta))w}{\psi} + c(y,\psi)\right)\\ &= \frac{y - b'(\vartheta)}{\psi}w.\end{aligned}$$

Now, by Bartlett's Identities, we have

$$\begin{aligned}0 &= \mathbb{E}\left[\frac{\partial}{\partial\vartheta}\log f_Y(y)\right]\\ &= \mathbb{E}\left[\frac{y - b'(\vartheta)}{\psi}w\right]\\ &= \frac{\mathbb{E}[Y] - b'(\vartheta)}{\psi}w,\end{aligned}$$

and

$$\begin{aligned}0 &= \mathrm{Var}\left(\frac{\partial}{\partial\vartheta}\left(\frac{y\vartheta - b(\vartheta)}{\psi}w + c(y,\psi)\right)\right) + \mathbb{E}\left[\frac{\partial^2}{\partial\vartheta^2}\left(\frac{y\vartheta - b(\vartheta)}{\psi}w + c(y,\psi)\right)\right]\\ &= \frac{w^2}{\psi^2}\mathrm{Var}(y - b'(\vartheta)) + \mathbb{E}\left[-\frac{b''(\vartheta)}{\psi}w\right].\end{aligned}$$

Therefore,

$$\mathbb{E}[Y] = b'(\vartheta) \quad \text{and} \quad \mathrm{Var}(Y) = b''(\vartheta)\frac{\psi}{w}.$$

□

Usually, for simplicity's sake, we write

$$\mu \coloneqq b'(\vartheta) \quad \text{and} \quad V(\mu) \coloneqq b''(\vartheta),$$

noting that $V$ is also known as the **variance function**.

*Example 12.3.* Consider $Y \sim \mathcal{P}(\mu)$. We can rewrite as usual, for $y = 0, 1, \ldots$,

$$\begin{aligned} f_Y(y) &= \frac{e^{-\mu}\mu^y}{y!} \\ &= \exp\{y \log \mu - \mu - \log y!\} \\ &= \exp\left\{y\vartheta - e^{\vartheta} - \log y!\right\} \end{aligned}$$

where $\vartheta = \log \mu$ and, consequently, $\mu = e^{\vartheta}$. Therefore, by (12.1), we have

$$\begin{aligned} &b(\vartheta) = b'(\vartheta) = b''(\vartheta) = e^{\vartheta}, \\ &\psi = w = 1, \\ &c(\psi, y) = \log y!, \end{aligned}$$

with variance function equal to $V(\mu) = \mu$. In practice, here $\psi$ is not present.

*Example 12.4.* Consider $Y \sim \mathcal{N}(\mu, \sigma^2)$, where

$$\begin{aligned} f_Y(y) &= \exp\left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y-\mu)^2\right\} \\ &= \exp\left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y^2 - 2y\mu + \mu^2) - \frac{1}{2}\right\} \\ &= \exp\left\{\frac{1}{\sigma^2}\left(y\mu - \frac{\mu^2}{2}\right) - \frac{1}{2}\left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right]\right\}. \end{aligned}$$

The density function is inside the exponential family, from where we conclude that

$$\begin{aligned} &\vartheta = \mu, \\ &b(\vartheta) = \frac{\mu^2}{2}, \quad b'(\vartheta) = \mu, \quad b''(\vartheta) = 1, \\ &\psi = \sigma^2, \\ &w = 1, \\ &\frac{\psi}{w} = \sigma^2. \end{aligned}$$

*Remark 12.5.* The class of Generalised Linear Models has the main advantage of providing (by means of the exponential family) a unified treatment for a set of relevant models. Consider the linear model. For each observation $Y_i$, we defined a linear predictor

$$\eta_i = {}^{\mathrm{t}}\boldsymbol{x}_i\boldsymbol{\beta},$$

where $\boldsymbol{x}_i$ is the row of order $i$ of the matrix $X$, for $i = 1, \ldots, n$. Suppose that each observed $y_i$ is sampled from $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, where the relation between the mean value $\mu_i$ and the linear predictor $\eta_i$ is the identity. In other terms, we assumed that

$$\eta_i = \mu_i, \quad \eta_i = {}^{\mathrm{t}}\boldsymbol{x}_i\boldsymbol{\beta}.$$

The Generalised Linear Models are obtained by expanding the above formulation of the linear model in two directions:

1. the distribution of $Y_i$ is not restricted to being Gaussian, but it can be any distribution of the form $EF(b(\vartheta_i), \psi/w_i)$, such that $b'(\vartheta_i) = \mu_i$ is provided;
2. other forms of relationship between the linear predictor $\eta_i$ and the mean value $\mu_i$ are possible. We consider
$$g(\mu_i) = \eta_i,$$
where $g(\cdot)$ is a differentiable monotonic function called **link function**.

In a schematic way, we can specify a Generalised Linear Model as follows:

1. $Y_i \sim EF(b(\vartheta_i), \psi/w_i)$ known as **error structure**;
2. $g(\mu_i) = \eta_i$ known as **link function**;
3. $\eta_i = {}^{\mathrm{t}}\boldsymbol{x}_i\boldsymbol{\beta}$ known as **linear predictor**.

**Definition 12.6.** *A statistical model is a Generalised Linear Model when it satisfies the following requirements:*

1. *the observations $y_1, \dots, y_n$ are realisations of independent random variables $Y_1, \dots, Y_n$;*
2. *each $Y_i \sim EF(b(\vartheta_i), \psi/w_i)$ with $\mathbb{E}[Y_i] = \mu_i = b'(\vartheta_i)$, for $i = 1, \dots, n$;*
3. *there exist a function $g(\cdot)$ such that*
$$g(\mu_i) = {}^{\mathrm{t}}\boldsymbol{x}_i\boldsymbol{\beta},$$
*where $\boldsymbol{x}_i$ is a vector of constants and $\boldsymbol{\beta}$ is a vector of parameters;*
4. *the function $b(\cdot)$ and the parameter $\psi$ are common to all the random variables $Y_i$, and the parameter $w_i$ can change;*
5. *the functions $b(\cdot)$, $c(\cdot, \cdot)$ and $g(\cdot)$ are known, as also the weight $w_i$.*

Note that, when the link function $g(\cdot)$ is the identity function and the error structure is Gaussian, then we go back to the linear model. The terminology "error structure" is preserved because of this connection with the linear model. However, in the generalised linear model, there is not an explicit error term $\varepsilon$.

In fact, in the Gaussian case, we can equivalently write

$$Y_i \sim \mathcal{N}(\mu, \sigma^2)$$

or

$$Y_i = \mu_i + \varepsilon_i,$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. In this case, there is a complete separation between the systematic component $\mu_i$ (which depends on the explanatory variable $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$) and the purely erratic term $\varepsilon_i$ (which does not). In the case of generalised linear models, this precise separation of the response variable in two components is no longer possible.

*Example 12.7 (Poisson Regression).* Consider $Y_1, \dots, Y_n$ independent Poisson random variables, with mean values $\mu_1, \dots, \mu_n$, respectively. Assume in addition that

$$\log \mu_i = {}^{\mathrm{t}}\boldsymbol{x}_i\boldsymbol{\beta} = \eta_i,$$

where the meanining of $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ is the usual. In this particular situation, $Y_i$ cannot be decomposed as

$$Y_i = \mu_i + \varepsilon_i,$$

where $\varepsilon_i$ represents a purely erratic component, because in order to satisfy an identity as the above, the random variable $\varepsilon_i$ should have a distribution depending on $\mu_i$, and this is not appropriate.

*Example 12.8 (Logistic Regression).* An important case of Generalised Linear Models is $Y_i \sim Ber(\mu_i)$. These are distributed as

$$\begin{aligned} f_{Y_i}(y_i) &= \exp\left\{\log\left(\mu_i^{y_i}(1-\mu_i)^{1-y_i}\right)\right\} \\ &= \exp\left\{y_i \log \mu_i + (1-y_i)\log(1-\mu_i)\right\} \\ &= \exp\left\{y_i \log\left(\frac{\mu_i}{1-\mu_i}\right) - \log(1-\mu_i)\right\} \\ &= \exp\left\{y_i\vartheta_i - \log\left(1 - \frac{e^{\vartheta_i}}{1+e^{\vartheta_i}}\right)\right\} \\ &= \exp\left\{y_i\vartheta_i - \log(1+e^{\vartheta_i})\right\}, \end{aligned}$$

where

$$\vartheta_i = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \text{LOGIT}(\mu_i) \quad \Longrightarrow \quad \mu_i = \frac{e^{\vartheta_i}}{1+e^{\vartheta_i}}$$

with the assumption that $\vartheta_i$ is a linear function of an explanatory variable. We can assume that $\vartheta_i$ is a linear function of an explanatory variable, therefore

$$\begin{aligned} b(\vartheta_i) &= \log(1+e^{\vartheta_i}) \\ b'(\vartheta_i) &= \frac{e^{\vartheta_i}}{1+e^{\vartheta_i}} = \mu_i \\ b''(\vartheta_i) &= \frac{e^{\vartheta_i}}{(1+e^{\vartheta_i})^2} = \mu(1-\mu_i) = V(\mu_i) \end{aligned}$$

This is the classical form of the logistic regression model.

## 12.2 12/12/18

We can extend the basic logistic regression in two directions. The first extension consists in allowing several explanatory variables, forming the vector $\boldsymbol{x}_i$, and assuming

$$\vartheta_i = {}^{\mathrm{t}}\boldsymbol{x}_i\boldsymbol{\beta}.$$

The second extension consists in the possibility of replicating the experiment for any given combination of experimental factors.

If we denote by $m_i$ the number of replicates at a given combination $\boldsymbol{x}_i$ of the explanatory variables, then the corresponding number of successes is

$$\overline{Y}_i \sim \mathcal{B}(m_i, \mu_i).$$

Since our goal is to investigate the relation between the explanatory variables and the probability of success, we do not use $\overline{Y}_i$ as a response variable, but rather the proportion of successes

$$Y_i = \frac{\overline{Y}_i}{m_i}.$$

In this way, $\mathbb{E}[Y_i] = \mu_i$, which is the probability under consideration. In this case, the probability function is

$$\begin{aligned}
f_Y(y_i) &= \binom{m_i}{m_i y_i} \mu_i^{m_i y_i} (1-\mu_i)^{m_i(1-y_i)} \\
&= \exp\left\{ \log\binom{m_i}{m_i y_i} + \log \mu_i^{m_i y_i} + \log(1-\mu_i)^{m_i(1-y_i)} \right\} \\
&= \exp\left\{ m_i y_i \log\mu_i + m_i(1-y_i)\log(1-\mu_i) + \log\binom{m_i}{m_i y_i} \right\} \\
&= \exp\left\{ m_i y_i \left[\log\mu_i - \log(1-\mu_i)\right] + m_i \log(1-\mu_i) + \log\binom{m_i}{m_i y_i} \right\} \\
&= \exp\left\{ m_i y_i \log\left(\frac{\mu_i}{1-\mu_i}\right) - m_i \log\left(\frac{1}{1-\mu_i}\right) + \log\binom{m_i}{m_i y_i} \right\} \\
&= \exp\left\{ m_i \left[ y_i \log\left(\frac{\mu_i}{1-\mu_i}\right) - \log\left(\frac{1}{1-\mu_i}\right)\right] + \log\binom{m_i}{m_i y_i} \right\} \\
&= \exp\left\{ m_i \left[ y_i\vartheta_i - \log(1+e^{\vartheta_i})\right] + \log\binom{m_i}{m_i y_i} \right\},
\end{aligned}$$

where

$$y_i \in \left\{ 0, \frac{1}{m_i}, \frac{2}{m_i}, \dots, 1 \right\}.$$

**How to make inference on the parameters $\boldsymbol{\beta}$ and $\psi$**

Denote by $p$ the size of $\boldsymbol{\beta}$ and let $X = (x_{ij})$ be the $n \times p$ matrix. Let ${}^{\mathrm{t}}\boldsymbol{x}_i$ be the $i$-th row of $X$. Considering the observations $Y_1, \dots, Y_n$ with mean values $\mu_1, \dots, \mu_n$, respectively, such that

$$g(\mu_i) = {}^{\mathrm{t}}\boldsymbol{x}_i\boldsymbol{\beta},$$

we wish to estimate the parameters $\boldsymbol{\beta}$ and $\psi$.

Let us first focus on the parameter $\boldsymbol{\beta}$, considering $\psi$ as a noise. Since the $Y_i$'s are independent and we have a precise distributional assumption on their nature, we can write the Log-Likelihood function (which will be a sum):

$$\begin{aligned}\log \mathcal{L}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left( \frac{w_i[y_i\vartheta_i - b(\vartheta_i)]}{\psi} + c_i(y_i, \psi) \right) \\ &= \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\beta}).\end{aligned}$$

To obtain the Likelihood equations, we can write

$$\frac{d\mathcal{L}_i}{d\beta_j} = \frac{d\mathcal{L}_i}{d\vartheta_i}\frac{d\vartheta_i}{d\mu_i}\frac{d\mu_i}{d\eta_i}\frac{d\eta_i}{d\beta_j},$$

whose components are

$$\begin{aligned}\frac{d\mathcal{L}_i}{d\vartheta_i} &= \frac{w_i[y_i - b'(\vartheta_i)]}{\psi} = \frac{w_i(y_i - \mu_i)}{\psi} \\ \frac{d\mu_i}{d\vartheta_i} &= b''(\vartheta_i) = \frac{w_i \operatorname{Var}(Y_i)}{\psi} \\ \frac{d\eta_i}{d\beta_j} &= x_{ij}.\end{aligned}$$

Now,

$$\frac{d\mathcal{L}_i}{d\beta_j} = \frac{w_i(y_i - \mu_i)}{\psi}\frac{\psi}{w_i \operatorname{Var}(Y_i)}\frac{d\mu_i}{d\eta_i}x_{ij}$$

and we obtain the Likelihood equation

$$\sum_{i=1}^{n} \frac{d\mathcal{L}_i}{d\beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\operatorname{Var}(Y_i)}\frac{d\mu_i}{d\eta_i} = 0, \quad j = 1, \dots, p.$$

We can also compute the Fisher information. Consider the second derivatives of $\mathcal{L}_i(\boldsymbol{\beta})$ and write

$$\begin{aligned}I_n(\boldsymbol{\beta}) &= -\mathbb{E}\left[\frac{d^2\mathcal{L}_i}{d\beta_j d\beta_k}\right] \\ &= -\mathbb{E}\left[\frac{d\mathcal{L}_i}{d\beta_j}\frac{d\mathcal{L}_i}{d\beta_k}\right] \\ &= -\mathbb{E}\left[\left(\frac{(Y_i - \mu_i)x_{ij}}{\operatorname{Var}(Y_i)}\frac{d\mu_i}{d\eta_i}\right)\left(\frac{(Y_i - \mu_i)x_{ik}}{\operatorname{Var}(Y_i)}\frac{d\mu_i}{d\eta_i}\right)\right] \\ &= -\frac{x_{ij}x_{ik}}{\operatorname{Var}(Y_i)}\left(\frac{d\mu_i}{d\eta_i}\right)^2, \quad i = 1, \dots, n \quad j, k = 1, \dots, p.\end{aligned}$$

This provides the entry of order $(j, k)$ of the Fisher information matrix. In general, we have

$$-\sum_{i=1}^{n}\mathbb{E}\left[\frac{d^2\mathcal{L}_i}{d\beta_j d\beta_k}\right]=-\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{\text{Var}\,(Y_i)}\left(\frac{d\mu_i}{d\eta_i}\right)^2,$$

which can be re-written as $I(\boldsymbol{\beta})={}^{\text{t}}X\tilde{W}X$, where

$$\tilde{W}=\begin{pmatrix}\tilde{w}_1 & 0 & \dots & 0\\ 0 & \tilde{w}_2 & \dots & 0\\ \vdots & & \ddots & 0\\ 0 & \dots & \dots & \tilde{w}_p\end{pmatrix}$$

and

$$\tilde{w}_i=\frac{1}{\text{Var}\,(Y_i)}\left(\frac{d\mu_i}{d\eta_i}\right)^2.$$

This is all we need to make inference.

*Example 12.9.* Consider a random sample $\boldsymbol{Y}=(Y_1,\dots,Y_n)$ from $Y\sim f_Y(y)$, where

$$f_Y(y)=\exp\{-\rho y+\log\rho\}=\exp\{\vartheta y+\log(-\vartheta)\},$$

where

$$\vartheta=-\rho=e^{-\eta}.$$

Here

$$\begin{aligned}b(\vartheta)&=-\log(-\vartheta),\\ b'(\vartheta)&=-\frac{1}{\vartheta}=\mu,\\ b''(\vartheta)&=\frac{1}{\vartheta^2}=\mu^2=V(\mu),\\ \frac{d\mu}{d\eta}&=e^{\eta}=-\frac{1}{\vartheta}.\end{aligned}$$

We can apply the Likelihood equations. We have

$$\sum_{i=1}^{n}\frac{(y_i+\vartheta_i^{-1})}{\vartheta_i^{-2}}x_{ij}\frac{1}{\vartheta_i}.$$

Suppose for instance that $p=2$. Set $x_{i1}=1$, $x_{i2}=x_i$. The Likelihood equations are

$$\begin{cases}\sum_{i=1}^{n}y_i\rho_i=0\\ \sum_{i=1}^{n}x_iy_i\rho_i-\sum_{i=1}^{n}x_i=0\end{cases}.$$

Try to compute the Fisher information

$$I(\boldsymbol{\beta})={}^{\text{t}}X\hat{W}X.$$

**Canonical Link and Sufficient Statistics**

The link function $g(\mu)$ does not need to satisfy strong conditions, and it can be specified in different ways. There exist a specific choice of the limit function, also known as ***canonical link***, which enjoys special properties: $g(\mu_i) = \vartheta_i$, such that $\eta_i = \vartheta_i$. The reason behind the name choice is evident, since the linear predictor $\eta_i = {}^{\mathrm{t}}\boldsymbol{x}_i\boldsymbol{\beta}$ coincides with the parameter $\vartheta_i$ of the exponential family. In the case of the canonical link, we have that

$$\begin{aligned}\log \mathcal{L}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{w_i}{\psi}[y_i\vartheta_i - b(\vartheta_i)] + \sum_{i=1}^{n} c_i(y_i, \eta_i) \\ &= \sum_{i=1}^{n} \frac{1}{\psi}\left[w_i y_i\, {}^{\mathrm{t}}\boldsymbol{x}_i\boldsymbol{\beta} - w_i b({}^{\mathrm{t}}\boldsymbol{x}_i\boldsymbol{\beta})\right] + \sum_{i=1}^{n} c_i(y_i, \psi) \\ &= \frac{1}{\psi}\left[{}^{\mathrm{t}}\left(\sum_{i=1}^{n} w_i y_i \boldsymbol{x}_i\right)\boldsymbol{\beta} - \sum_{i=1}^{n} w_i b({}^{\mathrm{t}}x_i\boldsymbol{\beta})\right] + \sum_{i=1}^{n} c_i(y_i, \psi).\end{aligned}$$

This implies that, by the factorisation result,

$$\sum_{i=1}^{n} w_i y_i \boldsymbol{x}_i$$

is a sufficient statistic for $\boldsymbol{\beta}$, when $\psi$ is a fixed quantity.

*Example 12.10.* Suppose that $Y_i \sim \Gamma\left(w_i, w_i/\mu_i\right)$, where the index $w$ is constant on $i$:

$$\begin{aligned}f_{Y_i}(y_i) &= \frac{\left(\frac{w}{\mu_i}\right)^{w}}{\Gamma(w)} y_i^{w-1} \exp\left\{-\frac{w y_i}{\mu_i}\right\} \\ &= \exp\left\{w\left(-\frac{y_i}{\mu_i} - \log\mu_i\right) + (w-1)\log y_i - \log\Gamma(w) + w_i \log w_i\right\} \\ &= \exp\left\{w(\vartheta y_i + \log(-\vartheta)) + c_i(y_i, w)\right\},\end{aligned}$$

where we set $\vartheta = -1/\mu_i$ and

$$c_i(y_i, m_i) = (w-1)\log y_i - \log\Gamma(w) + w\log w.$$

The canonical link can be obtained when

$$-\mu_i^{-1} = \vartheta_i = {}^{\mathrm{t}}\boldsymbol{x}_i\boldsymbol{\beta}.$$

Therefore,

$$\mathcal{L}(\boldsymbol{\beta}) = \exp\left\{w\left[{}^{\mathrm{t}}\left(\sum_{i=1}^{n}\boldsymbol{x}_i y_i\right)\boldsymbol{\beta} + \sum_{i=1}^{n}\log(-\,{}^{\mathrm{t}}\boldsymbol{x}_i\boldsymbol{\beta})\right] + \sum_{i=1}^{n} c(y_i, w)\right\},$$

confirming that

$$\sum_{i=1}^{n}\boldsymbol{x}_i y_i$$

is the sufficient statistic for $\boldsymbol{\beta}$ if $w$ is a fixed quantity.

The main advantage of using the canonical link is in terms of the Likelihood equations: indeed, they simplify quite a bit, since

$$\frac{d\mu_i}{d\eta_i} = \frac{d\mu_i}{d\vartheta_i} = \frac{d}{d\vartheta_i} b'(\vartheta_i) = b''(\vartheta_i).$$

This allows to write, using the simplification, more Likelihood equations:

$$\frac{d\mathcal{L}_i}{d\beta_j} = \left( \frac{y_i - \mu_i}{\text{Var}\,(Y_i)} \right) x_{ij} b''(\vartheta_i) = \frac{w_i (y_i - \mu_i) x_{ij}}{\psi} = 0, \quad j = 1, \dots, p,$$

which become independent from the specification of a particular distribution of $Y_i$. Moreover, with the canonical link, the Fisher information can also be simplified. Indeed, we have

$$\frac{d^2 \mathcal{L}_i}{d\beta_j d\beta_k} = -\frac{w_i x_{ij}}{\psi} \frac{d\mu_i}{d\beta_k},$$

which does not depend on the observations. Therefore,

$$\frac{d^2 \mathcal{L}}{d\beta_j d\beta_k} = -\mathbb{E}\left[ \frac{d^2 \mathcal{L}}{d\beta_j d\beta_k} \right],$$

so that the expected Fisher information and the observed information coincide.

## 12.3 13/12/18

### 12.3.1 Solving the Likelihood Equations

In most cases, the Likelihood equations

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\text{Var}\,(Y_i)} \frac{d\mu_i}{d\eta_i} = 0$$

cannot be solved explicitly. Therefore, we resort to numerical methods. We have all the ingredients to apply Newton-Rhapsody method.

One of the reasons of the success of Generalised Linear Models, is indeed the possibility of using a single algorithm with a small adjustment to be made for the choice of the link function and the distribution of $Y$. The algorithm operates by solving a sequence of Least Squares problems.

The method works by adjusting the $k$-th approximate solution $\beta^{(k)}$ to obtain the next approximation $\beta^{(k+1)}$. We will repeat this process until convergence.

Denote by $u^{(k)}$ the Fisher score computed at the solution $\beta^{(k)}$, and use the upscript notation for denoting iterations. In order to improve the approximate solution at the step $k$, we will use the Fisher scoring algorithm which can be expressed as follows:

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + \left[I\left(\boldsymbol{\beta}^{(k)}\right)\right]^{-1} \boldsymbol{u}^{(k)}, \tag{12.2}$$

which is similar to the Newton-Rapson method but with the negative Hessian matrix replaced by the expected Fisher information. Expression (12.2) can be re-written as follows

$$I\left(\boldsymbol{\beta}^{(k)}\right)\boldsymbol{\beta}^{(k+1)} = I\left(\boldsymbol{\beta}^{(k)}\right)\boldsymbol{\beta}^{(k)} + \boldsymbol{u}^{(k)}.$$

The right-hand side is a vector with the following $h$-th component:

$$\sum_{j=1}^{n}\left[\sum_{i=1}^{n} \frac{x_{ih}x_{ij}}{\text{Var}\,(Y_i)}\left(\frac{d\mu_i}{d\eta_i}\right)^2\right]\beta_j^{(k)} + \sum_{i=1}^{n} \frac{\left(y_i - \mu_i^{(k)}\right)x_{ih}}{\text{Var}\,(Y_i)}\frac{d\mu_i}{d\eta_i},$$

therefore we can write

$$I(\boldsymbol{\beta}^{(k)})\boldsymbol{\beta}^{(k)} + \boldsymbol{u}^{(k)} = {}^{\text{t}}X\tilde{W}^{(k)}\boldsymbol{z}^{(k)},$$

where

$$\begin{aligned} z_i^{(k)} &= \sum_{j=1}^{n} x_{ij}\beta_j^{(k)} + \left(y_i - \mu_i^{(k)}\right)\left(\frac{\beta_i^{(k)}}{d\mu_i^{(k)}}\right) \\ &= \eta_i^{(k)} + \left(y_i - \mu_i^{(k)}\right)\left(\frac{d\eta_i^{(k)}}{d\mu_i^{(k)}}\right) \end{aligned}$$

and the entries of $\tilde{W}$ have been defined yesterday. By combining

$$I\left(\boldsymbol{\beta}^{(k)}\right)\boldsymbol{\beta}^{(k+1)} = I\left(\boldsymbol{\beta}^{(k)}\right)\boldsymbol{\beta}^{(k)} + \boldsymbol{u}^{(k)}$$

and

$$I\left(\boldsymbol{\beta}\right) = {}^{\mathrm{t}}X\tilde{W}X$$

we get

$$\begin{aligned}\left({}^{\mathrm{t}}X\tilde{W}X\right)\boldsymbol{\beta}^{(k+1)} &= \left({}^{\mathrm{t}}X\tilde{W}X\right)\boldsymbol{\beta}^{(k)} + \boldsymbol{u}^{(k)} \\ &= {}^{\mathrm{t}}X\tilde{W}^{(k)}\boldsymbol{z}^{(k)}\end{aligned}$$

which is equivalent to

$$\boldsymbol{\beta}^{(k+1)} = \left({}^{\mathrm{t}}X\tilde{W}^{(k)}X\right)^{-1}{}^{\mathrm{t}}X\tilde{W}^{(k)}\boldsymbol{z}^{(k)}.$$

This is the recursion to set the approximate solution at level $(k+1)$ given the approximate solution at level $k$.

*Remark 12.11.* Using this approach, the dispersion parameter is not relevant for estimating $\boldsymbol{\beta}$.

**The Algorithm**

The algorithm proceeds in two main steps.

1. When $\boldsymbol{\beta}^{(k)}$ is given, the vector $\boldsymbol{z}^{(k)}$ and the matrix $\tilde{W}^{(k)}$ can be computed.
2. Using
$$\boldsymbol{\beta}^{(k+1)} = \left({}^{\mathrm{t}}X\tilde{W}^{(k)}X\right)^{-1}{}^{\mathrm{t}}X\tilde{W}^{(k)}\boldsymbol{z}^{(k)}$$
we can compute $\boldsymbol{\beta}^{(k+1)}$.

These two steps are iterated until the sequence $\left(\boldsymbol{\beta}^{(k)}\right)_{k\geq 1}$ converges. This algorithm is called **Iteratively Reweighed Least Squares**.

*Remark 12.12.* If we consider the following Taylor expansion of $g(y_i)$ around $\mu_i$,

$$\begin{aligned}g(y_i) &\approx g(\mu_i) + (y_i - \mu_i)g'(\mu_i) \\ &= \mu_i(y_i - \mu_i)\frac{d\eta_i}{d\mu_i} \\ &= z_i,\end{aligned}$$

which means that the quantity $z_i$ (component of the vector $\boldsymbol{z}$) is a local approximation of $g(y_i)$.

This Remark also provides us with a hint for selecting the starting point of the algorithm. In particular, we can set

$$z_i^{(0)} = g(y_i), \quad \tilde{W}^{(0)} = I.$$

Given these two, we obtain $\boldsymbol{\beta}^{(1)}$, and therefore a starting point for the iteration.

One application of this algorithm is the Logistic Regression (repeated experiment):

$$\begin{aligned}
\text{Var}\,(Y_i) &= \frac{1}{m_i}\mu_i(1-\mu_i);\\
g(\mu_i) &= \log\left(\frac{\mu}{1-\mu}\right) = \eta;\\
\mu &= \frac{e^\eta}{1+e^\eta};\\
\eta_i &= {}^{\text{t}}\boldsymbol{x}_i\boldsymbol{\beta} = \vartheta_i;\\
\frac{d\mu}{d\eta} &= \frac{e^\eta}{(1+e^\eta)^2} = \mu(1-\mu).
\end{aligned}$$

By combining these, we have

$$\begin{aligned}
z_i^{(k)} &= \eta_i^{(k)} + \frac{y_i - \mu_i^{(k)}}{\mu_i^{(k)}\left(1-\mu_i^{(k)}\right)}\\
\tilde{W}_i^{(k)} &= m_i\mu_i^{(k)}\left(1-\mu_i^{(k)}\right).
\end{aligned}$$

The equality $\eta_i = \vartheta_i$ means that the LOGIT function is the canonical link for the logistic regression. To start, we can set

$$z_i^{(0)} = g(y_i) = \log\left(\frac{y_i}{1-y_i}\right) = \log\left(\frac{\tilde{y}_i}{m_i-\tilde{y}_i}\right).$$

To avoid issues with $\tilde{y}_i = 0$ and $\tilde{y}_i = m_i$, typically we start with

$$z_i^{(0)} = \log\left(\frac{\tilde{y}_i + \frac{1}{2}}{m_i - \tilde{y}_i + \frac{1}{2}}\right).$$

*Initial Settings*: we need a value for $z_i$ we assign

$$z_i^{(0)} \longleftarrow \log\left(\frac{\tilde{y}_i + 2^{-1}}{m_i - \tilde{y}_i + 2^{-1}}\right)$$

and the identity matrix to $\tilde{W}$.
*Repeat these steps until convergence*:

$$\begin{aligned}
&R \longleftarrow ({}^{\text{t}}X\tilde{W}X)^{-1},\\
&\boldsymbol{\beta} \longleftarrow R\,{}^{\text{t}}X\tilde{W}X,\\
&\boldsymbol{\eta} \longleftarrow X\boldsymbol{\beta},\\
&\mu_i \longleftarrow \frac{e^{\eta_i}}{(1+e^{\eta_i})},\\
&\Delta_i \longleftarrow \mu_i(1-\mu_i),\\
&z_i \longleftarrow \frac{\mu_i + (\tilde{y}_i - \mu_i)}{\Delta_i,} \quad i = 1,\dots,n\\
&\tilde{W} \longleftarrow \text{diag}\,(\Delta_1 m_1,\dots,\Delta_n m_n)
\end{aligned}$$

*Final setting*:

$\hat{\boldsymbol{\beta}} \longleftarrow$ last value of $\boldsymbol{\beta}$,

$\text{Var}\left(\hat{\boldsymbol{\beta}}\right) \longleftarrow$ last value of R.