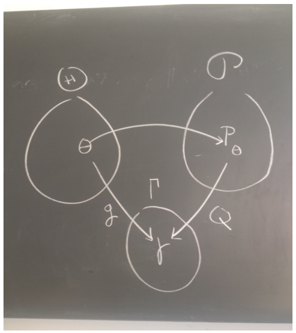
Mathematical Statistics



Sara van de Geer

September 2019

Contents

These notes in English closely follow *Mathematische Statistik*, by H.R. Künsch (2005). *Mathematische Statistik* can be used as supplementary reading material in German.

Throughout the notes measurability assumptions are not stated explicitly.

Mathematical rigour and clarity often bite each other. At some places, not all subtleties are fully presented. A snake will indicate this.

Chapters or (sub)sections with a $\star$ are not part of the exam.

Chapter 1

Introduction

Statistics is about the mathematical modeling of observable phenomena, using stochastic models, and about analyzing data: estimating parameters of the model, constructing confidence intervals and testing hypotheses. In these notes, we study various estimation and testing procedures. We consider their theoretical properties and we investigate various notions of optimality.

Some notation and model assumptions
The data consist of measurements (observations) $x_1, \ldots, x_n$, which are regarded as realizations of random variables $X_1, \ldots, X_n$. In most of the notes, the X_i are real-valued: $X_i \in \mathbb{R}$ (for $i = 1, \ldots, n$), although we will also consider some extensions to vector-valued observations.

1.1 Speed of light example

Fizeau and Foucault developed methods for estimating the speed of light (1849, 1850), which were later improved by Newcomb and Michelson. The main idea is to pass light from a rapidly rotating mirror to a fixed mirror and back to the rotating mirror. An estimate of the velocity of light is obtained, taking into account the speed of the rotating mirror, the distance travelled, and the displacement of the light as it returns to the rotating mirror.

Fig. 1

The data are Newcomb's measurements of the passage time it took light to travel from his lab, to a mirror on the Washington Monument, and back to his lab.

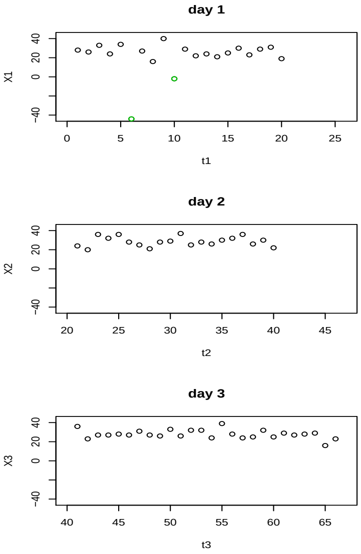
distance: 7.44373 km.

66 measurements on 3 consecutive days

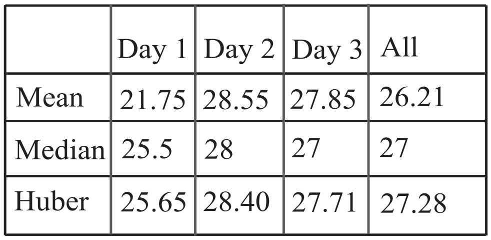
first measurement: 0.000024828 seconds= 24828 nanoseconds

The dataset has the deviations from 24800 nanoseconds.

The measurements on 3 different days:



One may estimate the speed of light using e.g. the mean, or the median, or Huber's estimate (see below). This gives the following results (for the 3 days separately, and for the three days combined):



	Day 1	Day 2	Day 3	All
Mean	21.75	28.55	27.85	26.21
Median	25.5	28	27	27
Huber	25.65	28.40	27.71	27.28

Table 1

The question which estimate is "the best one" is one of the topics of these notes.

1.2 Notation

The collection of observations will be denoted by $\mathbf{X} = \{X_1, \ldots, X_n\}$. The distribution of $\mathbf{X}$, denoted by $\mathbb{P}$, is generally unknown. A statistical model is a collection of assumptions about this unknown distribution.

We will usually assume that the observations $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.). Or, to formulate it differently, $X_1, \ldots, X_n$ are i.i.d. copies from some population random variable, which we denote by X. The common distribution, that is: the distribution of X, is denoted by P. For $X \in \mathbb{R}$, the distribution function of X is written as

$$F(\cdot) = P(X \leq \cdot).$$

Recall that the distribution function F determines the distribution P (and vise versa).

Further model assumptions then concern the modeling of P. We write such a model as $P \in \mathcal{P}$, where $\mathcal{P}$ is a given collection of probability measures, the so-called model class. Typically the distributions in $\mathcal{P}$ are indexed by a parameter, say θ, in some parameter space, say Θ. Then $\mathcal{P} = \{P_\theta : \ \theta \in \Theta\}$, and $P = P_\theta$ for some $\theta \in \Theta$. Then θ is often called the "true parameter".[1] The parameter space Θ may be high-dimensional, or even ∞-dimensional. Often, one is only interested in a certain aspect of the parameter. We write the parameter of interest as $\gamma := \ g(\theta)$ where $g : \ \Theta \rightarrow \Gamma$ a given function with values in some space Γ.

1.3 Example: the location model

The following example will serve to illustrate the concepts that are to follow. Let X be real-valued. The location model is

$$\mathcal{P} := \{P_\theta(X \leq \cdot) := F_0(\cdot - \mu), \ \theta := (\mu, F_0), \ \mu \in \mathbb{R}, \ F_0 \in \mathcal{F}_0\}, \tag{1.1}$$

where $\mathcal{F}_0$ is a given collection of distribution functions. Assuming the expectation exist, we center the distributions in $\mathcal{F}_0$ to have mean zero. Then P_{μ,F_0} has mean μ. We call μ a location parameter. Often, only μ is the parameter of interest, and F_0 is a so-called nuisance parameter. Then $g(\mu, F_0) = \mu$.

The class $\mathcal{F}_0$ is for example modeled as the class of all symmetric distributions, that is

$$\mathcal{F}_0 := \{F_0(x) = 1 - F_0(-x), \forall \ x\}. \tag{1.2}$$

This is an ∞-dimensional collection: it is not parametrized by a finite dimensional parameter. We then call F_0 an infinite-dimensional parameter.

A finite-dimensional model is for example

$$\mathcal{F}_0 := \{\Phi(\cdot/\sigma) : \ \sigma > 0\}, \tag{1.3}$$

where Φ is the standard normal distribution function.

[1] To be mathematically correct one should write $P_\theta \in \{P_\vartheta : \ \vartheta \in \Theta\}$ to make the distinction in notation between the true parameter θ and the parameter ϑ indexing the class $\mathcal{P}$. We actually need this distinction as the theory develops.

Thus, the location model is

$$X_i = \mu + \epsilon_i, \ i = 1, \ldots, n,$$

with $\epsilon_1, \ldots, \epsilon_n$ i.i.d. and, under model (1.2), symmetrically but otherwise unknown distributed and, under model (1.3), $\mathcal{N}(0, \sigma^2)$-distributed with unknown variance σ^2.

1.4 Some further examples of statistical models

Example 1.4.1 Poisson distribution
Consider a small insurance company. Let X be the number of claims on a particular day. Then a possible model is the Poisson model, which assumes that X has a Poisson distribution with parameter $\theta > 0$:

$$P_\theta(X = x) = \frac{\theta^x}{x!} \mathrm{e}^{-\theta}, \ x = 0, 1, 2, \ldots .$$

A parameter of interest could be for example the probability of at least 4 claims on a particular day. Then

$$\begin{aligned} \gamma &= P_\theta(X \geq 4) \\ &= 1 - P_\theta(X \leq 3) \\ &= 1 - \left(1 + \theta + \frac{\theta^2}{2} + \frac{\theta^3}{3!}\right) \mathrm{e}^{-\theta} \\ &:= g(\theta). \end{aligned}$$

Suppose we observed the number of claims $X_1, \ldots, X_n$ during $n = 200$ days. A possible estimator of θ is the sample average $\bar{X} := \sum_{i=1}^n X_i/n$. For γ we may use the "plug in" principle, that is, we plug in the estimator $\bar{X}$ for θ in the function g. This gives $\hat{\gamma} := g(\bar{X})$ as estimator of $g(\theta)$.

Here is a data example

x_i	# days
0	100
1	60
2	32
3	8
≥ 4	0

The observed average is $\bar{x} = .74$ and the estimate of $P_\theta(X \geq 4)$ is .00697.

Example 1.4.2 Pareto distribution
Suppose X has density (with respect to Lebesgue measure ν)

$$p_\theta(x) = \theta(1 + x)^{-(1+\theta)}, \ x > 0$$

where $\theta > 0$ is unknown. This is the Pareto density which is often used to model the distribution of income. The parameter θ is sometimes called the Pareto index. We write the model class for the distribution as

$$\mathcal{P} = \{P_\theta : \ dP_\theta / d\nu = p_\theta\}.$$

A parameter of interest may be the Gini index. It describes income inequality and is defined for $\theta \geq 1$ as $\gamma(\theta) = 1/(2\theta - 1)$. The value $G(1) = 1$ for $\theta = 1$ corresponds to complete income inequality.

Example 1.4.3 Classification
Let $X = (Y, Z)$ where $Z = \text{body mass index} \in \mathbb{R}$ and $Y \in \{0, 1\}$ indicates having diabetes or not. We assume the model

$$P_\theta(Y = 1 | Z = z) = \theta(z), \ z \in \mathbb{R},$$

with

$$\theta(\cdot) \in \Theta := \biggl\{ \text{all increasing functions } \theta : \ \mathbb{R} \to [0, 1] \biggr\}.$$

The parameter space is ∞-dimensional. A parameter of interest is for example

$$\gamma := \theta^{-1}(\tfrac{1}{2}),$$

that is, the value γ such that for $z \geq \gamma$ you are at risk:

$$P_\theta(Y = 1 | Z = z) \geq \tfrac{1}{2}.$$

Example 1.4.4 Social networks
Consider p individuals. which either do or do not communicate with each other. If they do, we say there is a connection between them, or we call them friends. Let X be a $p \times p$ matrix $X = (X_{j,k})$ coding the connections: for $j \neq k$

$$X_{j,k} := \begin{cases} 1 & \textit{if there is a connection between } j \textit{ and } k \\ 0 & \textit{else} \end{cases} .$$

The "stochastic block model" assumes that the $\{X_{j,k} : j < k\}$ are independent and

$$P_\theta(X_{j,k} = 1) = \begin{cases} \beta_m & \textit{if } j \textit{ and } k \textit{ are in the same community } m \\ \delta & \textit{if } j \textit{ and } k \textit{ are in different communities} \end{cases} .$$

If the number of communities is known, say M, but otherwise nothing further, then the parameter space is

$$\Theta = \biggl\{ (\beta_1, \ldots, \beta_M, \delta) \in [0, 1]^{M+1}, \ \mathcal{M} \biggr\}$$

where $\mathcal{M}$ is the collection of all community configurations. There are M^p possible community configurations, so $|\mathcal{M}| = M^p$. If p is large this is a huge number, i.e. the parameter space is very complex.

Remark: typically we observe only one realization of X, i.e. $n = 1$.

Example 1.4.5 Causal models
Suppose $X = (Z_1, \ldots, Z_p) \in \mathbb{R}^p$ *is a* p*-dimensional random variable, for example* $Z_1 = \text{rainfall}$, $Z_2 = \text{tea consumption}$, $Z_3 = \text{number of tall people}$, $Z_4 = \text{mountain height}$, $\cdots$ *within a particular canton of Switzerland. A causal model aims to find out which variables are causes and which are consequences. The structural relations model is*

$$Z_{\pi(j)} = f_j \biggl(Z_{\pi(1)}, \ldots, Z_{\pi(j-1)}, \epsilon_j \biggr), \ j = 2, \ldots, p.$$

Here π *is a permutation of* $\{1, \ldots, p\}$, ϵ_j *is unobservable noise and* f_j *is a partly unknown function. If we assume (for simplicity) the noise distribution to be known the parameter space is*

$$\Theta := \biggl\{ \text{all permutations } \pi \text{ and structural relations } (f_2, \ldots, f_p) \biggr\}.$$

A parameter of interest is for example the causal graph.

A sub-example is where the structural relations are modeled as being linear:

$$f_j(z_1, \ldots, z_{j-1}, \varepsilon_j) = \beta_{j,1} z_1 + \cdots + \beta_{j,j-1} z_{j-1} + \varepsilon_j, \ j = 2, \ldots, p$$

with $\{\beta_{j,k}\}$ *unknown coefficients.*

Chapter 2

Estimation

2.1 What is an estimator?

Recall that the data consist of observations $\mathbf{X} = (X_1, \ldots, X_n)$ with partly unknown distribution. A parameter is an aspect of the unknown distribution. We typically assume that $X_1, \ldots, X_n$ are i.i.d. copies of a random variable X where X has distribution $P_\theta \in \{P_\vartheta : \vartheta \in \Theta\}$.

An estimator is constructed to estimate some unknown parameter, γ say. Its formal definition is

Definition 2.1.1 *An estimator $T(\mathbf{X})$ is some given (measurable) function $T(\cdot)$ evaluated at the observations $\mathbf{X}$. The function $T(\cdot)$ is not allowed to depend on unknown parameters.*

An estimator is also called a *statistic* or a *decision.*

The reason why T is not allowed to depend on unknown parameters is that one should be able to calculate it in practice, using only the data. We will often use the same notation T for the estimator $T(\mathbf{X})$ (i.e., we write $T = T(\mathbf{X})$ omitting the argument $\mathbf{X}$) and the function $T = T(\cdot)$. It should be clear from the context which of the two is meant.

2.2 The empirical distribution function

Let $X_1, \ldots, X_n$ be real-valued observations. An example of a nonparametric estimator is the empirical distribution function

$$\hat{F}_n(\cdot) := \frac{1}{n} \#\{X_i \le \cdot,\ 1 \le i \le n\}.$$

This is an estimator of the theoretical distribution function

$$F(\cdot) := P(X \le \cdot).$$

Most estimators are constructed according the so-called a *plug-in principle* (*Einsetzprinzip*). That is, the parameter of interest γ is written as $\gamma = Q(F)$, with Q some given map. The empirical distribution $\hat{F}_n$ is then "plugged in", to obtain the estimator $T := Q(\hat{F}_n)$. (We note however that problems can arise, e.g. $Q(\hat{F}_n)$ may not be well-defined ....).

2.3 Some estimators for the location model

In the location model of Section 1.3, one may consider the following estimators $\hat{\mu}$ of μ (among others):

- The average

$$\hat{\mu}_1 := \frac{1}{n}\sum_{i=1}^n X_i.$$

Note that $\hat{\mu}_1$ minimizes over μ the squared loss[1]

$$\sum_{i=1}^n (X_i - \mu)^2,$$

that is

$$\hat{\mu} = \arg\min_{\mu\in\mathbb{R}} \sum_{i=1}^n (X_i - \mu)^2. \tag{2.1}$$

It can be shown that $\hat{\mu}_1$ is a "good" estimator if the model (1.3) holds, i.e., if $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$. When (1.3) is not true, in particular when there are *outliers* (large, "wrong", observations) (*Ausreisser*), then one has to apply a more *robust* estimator.

- The (sample) median is

$$\hat{\mu}_2 := \begin{cases} X_{((n+1)/2)} & \text{when } n \text{ odd} \\ \{X_{(n/2)} + X_{(n/2+1)}\}/2 & \text{when } n \text{ is even} \end{cases},$$

where $X_{(1)} \le \cdots \le X(n)$ are the order statistics. Note that $\hat{\mu}_2$ is a minimizer of the absolute loss

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

[1] To avoid misunderstanding, we note that e.g. in (2.1), μ is used as variable over which is minimized and is there not the unknown parameter μ. It is a general convention to abuse notation and employ the same symbol μ. When further developing the theory we shall often introduce a different symbol for the variable, to distinguish it from the "true parameter" μ, e.g., (2.1) is written as

$$\hat{\mu}_1 := \arg\min_{c\in\mathbb{R}} \sum_{i=1}^n (X_i - c)^2.$$

- The Huber estimator is

$$\hat{\mu}_3 := \arg\min_{\mu \in \mathbb{R}} \sum_{i=1}^n \rho(X_i - \mu),$$

where

$$\rho(x) = \begin{cases} x^2 & \text{if } |x| \le k \\ k(2|x| - k) & \text{if } |x| > k \end{cases},$$

with $k > 0$ some given threshold.

- We finally mention the α-trimmed mean, defined, for some $0 < \alpha < 1$, as

$$\hat{\mu}_4 := \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{(i)}.$$

The above estimators $\hat{\mu}_1, \ldots, \hat{\mu}_4$ are plug-in estimators of the location parameter μ. We define the maps

$$Q_1(F) := \int x dF(x)$$

(the mean, or point of gravity, of F), and

$$Q_2(F) := F^{-1}(1/2)$$

(the median of F), and

$$Q_3(F) := \arg\min_{\mu} \int \rho(\cdot - \mu) dF,$$

and finally

$$Q_4(F) := \frac{1}{1 - 2\alpha} \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x dF(x).$$

Then $\hat{\mu}_k$ corresponds to $Q_k(\hat{F}_n)$, $k = 1, \ldots, 4$. If the model (1.2) is correct (i.e. if F is symmetric around μ) $\hat{\mu}_1, \ldots, \hat{\mu}_4$ are all estimators of μ. If the model is incorrect, each $Q_k(\hat{F}_n)$ is still an estimator of $Q_k(F)$ (assuming the latter exists), but the $Q_k(F)$ may all be different aspects of F.

2.4 How to construct estimators

2.4.1 Plug-in estimators

For real-valued observations, one can define the distribution function

$$F(\cdot) = P(X \le \cdot).$$

An estimator of F is the empirical distribution function

$$\hat{F}_n(\cdot) = \frac{1}{n} \sum_{i=1}^n \mathrm{l}\{X_i \le \cdot\}.$$

Note that when knowing only $\hat{F}_n$, one can reconstruct the order statistics $X_{(1)} \le \ldots \le X_{(n)}$, but not the original data $X_1, \ldots, X_n$. Now, the order at which the data are given carries no information about the distribution P. In other words, a "reasonable"[2] estimator $T = T(X_1, \ldots, X_n)$ depends only on the sample $(X_1, \ldots, X_n)$ via the order statistics $(X_{(1)}, \ldots X_{(n)})$ (i.e., shuffling the data should have no influence on the value of T). Because these order statistics can be determined from the empirical distribution $\hat{F}_n$, we conclude that any "reasonable" estimator T can be written as a function of $\hat{F}_n$:

$$T = Q(\hat{F}_n),$$

for some map Q.

Similarly, the distribution function $F_\theta := P_\theta(X \le \cdot)$ completely characterizes the distribution P_θ. Hence, a parameter is a function of F_θ:

$$\gamma = g(\theta) = Q(F_\theta).$$

If the mapping Q is defined at all F_θ as well as at $\hat{F}_n$, we call $Q(\hat{F}_n)$ a plug-in estimator of $Q(F_\theta)$.

The idea is not restricted to the one-dimensional setting. For arbitrary observation space $\mathcal{X}$, we define the empirical measure

$$\hat{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i},$$

where δ_x is a point-mass at x. The empirical measure puts mass $1/n$ at each observation. This is indeed an extension of $\mathcal{X} = \mathbb{R}$ to general $\mathcal{X}$, as the empirical distribution function $\hat{F}_n$ jumps at each observation, with jump height equal to the number of times the value was observed (i.e. jump height $1/n$ if all X_i are distinct). As in the real-valued case, if the map Q is defined at all P_θ as well as at $\hat{P}_n$, we call $Q(\hat{P}_n)$ a plug-in estimator of $Q(P_\theta)$.

We stress that typically, the representation $\gamma = g(\theta)$ as function Q of P_θ is not unique, i.e., that there are various choices of Q. Each such choice generally leads to a different estimator. Moreover, the assumption that Q is defined at $\hat{P}_n$ is often violated. One can sometimes modify the map Q to a map Q_n that, in some sense, approximates Q for n large. The modified plug-in estimator then takes the form $Q_n(\hat{P}_n)$.

2.4.2 The method of moments

Let $X \in \mathbb{R}$ and suppose (say) that the parameter of interest is θ itself, and that $\Theta \subset \mathbb{R}^p$. Let $\mu_1(\theta), \ldots, \mu_p(\theta)$ denote the first p moments of X (assumed

[2] What is "reasonable" has to be considered with some care. There are in fact "reasonable" statistical procedures that do treat the $\{X_i\}$ in an asymmetric way. An example is splitting the sample into a training and test set (for model validation).

to exist), i.e.,

$$\mu_j(\theta) = E_\theta X^j = \int x^j dF_\theta(x),\ j = 1, \ldots, p.$$

Also assume that the map

$$m : \Theta \to \mathbb{R}^p,$$

defined by

$$m(\theta) = [\mu_1(\theta), \ldots, \mu_p(\theta)],$$

has an inverse

$$m^{-1}(\mu_1, \ldots, \mu_p),$$

for all $[\mu_1, \ldots, \mu_p] \in \mathcal{M}$ (say). We estimate the μ_j by their sample counterparts

$$\hat{\mu}_j := \frac{1}{n}\sum_{i=1}^n X_i^j = \int x^j d\hat{F}_n(x),\ j = 1, \ldots, p.$$

When $[\hat{\mu}_1, \ldots, \hat{\mu}_p] \in \mathcal{M}$ we can plug them in to obtain the estimator

$$\hat{\theta} := m^{-1}(\hat{\mu}_1, \ldots, \hat{\mu}_p).$$

Example 2.4.1 Method of moments for the negative binomial distribution
Let X have the negative binomial distribution with known parameter k and unknown success parameter $\theta \in (0,1)$:

$$P_\theta(X = x) = \binom{k+x-1}{x}\theta^k(1-\theta)^x, x \in \{0, 1, \ldots\}.$$

This is the distribution of the number of failures till the k^{th} success, where at each trial, the probability of success is θ, and where the trials are independent. It holds that

$$E_\theta(X) = k\frac{(1-\theta)}{\theta} := m(\theta).$$

Hence

$$m^{-1}(\mu) = \frac{k}{\mu + k},$$

and the method of moments estimator is

$$\hat{\theta} = \frac{k}{\bar{X} + k} = \frac{nk}{\sum_{i=1}^n X_i + nk} = \frac{\text{number of successes}}{\text{number of trails}}.$$

Example 2.4.2 Method of moments for the Pareto distribution
Suppose X has density

$$p_\theta(x) = \theta(1+x)^{-(1+\theta)},\ x > 0,$$

with respect to Lebesgue measure, and with $\theta \in \Theta \subset (0, \infty)$ (see Example 1.4.2). Then, for $\theta > 1$

$$E_\theta X = \frac{1}{\theta - 1} := m(\theta),$$

with inverse

$$m^{-1}(\mu) = \frac{1+\mu}{\mu}.$$

The method of moments estimator would thus be

$$\hat{\theta} = \frac{1+\bar{X}}{\bar{X}}.$$

However, the mean $E_\theta X$ does not exist for $\theta < 1$, so when Θ contains values $\theta < 1$, the method of moments is perhaps not a good idea. We will see that the maximum likelihood estimator does not suffer from this problem.

2.4.3 Likelihood methods

Suppose that $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ is dominated by a σ-finite measure ν. We write the densities as

$$p_\theta := \frac{dP_\theta}{d\nu},\ \theta \in \Theta.$$

Definition 2.4.1 *The* likelihood function *(with data $\mathbf{X} = (X_1, \ldots, X_n)$) is the function $L_{\mathbf{X}}: \ \Theta \to \mathbb{R}$ given by*

$$L_{\mathbf{X}}(\vartheta) := \prod_{i=1}^n p_\vartheta(X_i),\ \vartheta \in \Theta.$$

The MLE *(maximum likelihood estimator) is*

$$\hat{\theta} := \arg\max_{\vartheta\in\Theta} L_{\mathbf{X}}(\vartheta).$$

Note We use the symbol ϑ for the variable in the likelihood function, and the slightly different symbol θ for the parameter we want to estimate. It is however a common convention to use the same symbol for both (as already noted in the footnotes in Sections 1.2 and 2.3). As we will see, different symbols are needed for the development of the theory.

Note Alternatively, we may write the MLE as the maximizer of the *log*-likelihood

$$\hat{\theta} = \arg\max_{\vartheta\in\Theta} \log L_{\mathbf{X}}(\vartheta) = \arg\max_{\vartheta\in\Theta} \sum_{i=1}^n \log p_\vartheta(X_i).$$

The log-likelihood is generally mathematically more tractable. For example, if the densities are differentiable, one can typically obtain the maximum by setting the derivatives to zero, and it is easier to differentiate a sum than a product.

Note The likelihood function may have local maxima. Moreover, the MLE is not always unique, or may not exist (for example, the likelihood function may be unbounded).

We will now show that maximum likelihood is a plug-in method. First, as noted above, the MLE maximizes the log-likelihood. We may of course normalize the log-likelihood by $1/n$:

$$\hat{\theta} = \arg\max_{\vartheta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log p_{\vartheta}(X_i). \tag{2.2}$$

Replacing the average $\sum_{i=1}^{n} \log p_{\vartheta}(X_i)/n$ in (2.2) by its theoretical counterpart $E_{\theta} \log p_{\vartheta}(X)$ gives

$$\arg\max_{\vartheta \in \Theta} E_{\theta} \log p_{\vartheta}(X)$$

which is indeed equal to the parameter θ we are trying to estimate:

Lemma 2.4.1 *We have*

$$\theta = \arg\max_{\vartheta \in \Theta} E_{\theta} \log p_{\vartheta}(X).$$

Proof. By the inequality $\log x \le x - 1$, $x > 0$, for all $\vartheta \in \Theta$

$$E_{\theta} \log \frac{p_{\vartheta}(X)}{p_{\theta}(X)} \le E_{\theta}\left(\frac{p_{\vartheta}(X)}{p_{\theta}(X)} - 1 \right) = 0.$$

□

Example 2.4.3 MLE for the Pareto distribution
Suppose X has density

$$p_{\theta}(x) = \theta(1+x)^{-(1+\theta)}, \ x > 0,$$

with respect to Lebesgue measure, and with $\theta \in \Theta = (0, \infty)$. Then

$$\log p_{\vartheta}(x) = \log \vartheta - (1+\vartheta)\log(1+x),$$

$$\frac{d}{d\vartheta} \log p_{\vartheta}(x) = \frac{1}{\vartheta} - \log(1+x).$$

We put the derivative of the log-likelihood based on n observations to zero and solve:

$$\frac{n}{\hat{\theta}} - \sum_{i=1}^{n} \log(1+X_i) = 0$$

$$\Rightarrow \hat{\theta} = \frac{1}{\{\sum_{i=1}^{n} \log(1+X_i)\}/n}.$$

(One may check that this is indeed the maximum.)

Example 2.4.4 MLE for some location/scale models
Let $X \in \mathbb{R}$ and $\theta = (\mu, \sigma^2)$, with $\mu \in \mathbb{R}$ a location parameter, $\sigma > 0$ a scale parameter. We assume that the distribution function F_{θ} of X is

$$F_{\theta}(\cdot) = F_0\left(\frac{\cdot - \mu}{\sigma} \right),$$

where F_0 is a given distribution function, with density f_0 w.r.t. Lebesgue measure. The density of X is thus

$$p_\theta(\cdot) = \frac{1}{\sigma} f_0\biggl(\frac{\cdot - \mu}{\sigma}\biggr).$$

Case 1 *If $F_0 = \Phi$ (the standard normal distribution function), then*

$$f_0(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} \exp\biggl[-\frac{1}{2}x^2\biggr],\ x \in \mathbb{R},$$

so that

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\biggl[-\frac{1}{2\sigma^2}(x-\mu)^2\biggr],\ x \in \mathbb{R}.$$

The MLE of μ resp. σ^2 is

$$\hat{\mu} = \bar{X},\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2.$$

Case 2 *The (standardized) double exponential or Laplace distribution has density*

$$f_0(x) = \frac{1}{\sqrt{2}} \exp\biggl[-\sqrt{2}|x|\biggr],\ x \in \mathbb{R},$$

so

$$p_\theta(x) = \frac{1}{\sqrt{2\sigma^2}} \exp\biggl[-\frac{\sqrt{2}|x-\mu|}{\sigma}\biggr],\ x \in \mathbb{R}.$$

The MLE of μ resp. σ is now

$$\hat{\mu} = \text{sample median},\ \hat{\sigma} = \frac{\sqrt{2}}{n}\sum_{i=1}^n |X_i - \hat{\mu}_2|.$$

Example 2.4.5 An example where the MLE does not exist
Here is a famous example, from Kiefer and Wolfowitz (1956), where the likelihood is unbounded, and hence the MLE does not exist. It concerns the case of a mixture of two normals: each observation, is either $\mathcal{N}(\mu, 1)$-distributed or $\mathcal{N}(\mu, \sigma^2)$-distributed, each with probability $1/2$ (say). The unknown parameter is $\theta = (\mu, \sigma^2)$, and X has density

$$p_\theta(x) = \frac{1}{2}\phi(x-\mu) + \frac{1}{2\sigma}\phi((x-\mu)/\sigma),\ x \in \mathbb{R},$$

w.r.t. Lebesgue measure. Then

$$L_{\mathbf{X}}(\tilde{\mu}, \tilde{\sigma}^2) = \prod_{i=1}^n \biggl(\frac{1}{2}\phi(X_i - \tilde{\mu}) + \frac{1}{2\tilde{\sigma}}\phi((X_i - \tilde{\mu})/\tilde{\sigma})\biggr).$$

Taking $\tilde{\mu} = X_1$ yields

$$L_{\mathbf{X}}(X_1, \tilde{\sigma}^2) = \frac{1}{\sqrt{2\pi}}(\frac{1}{2} + \frac{1}{2\tilde{\sigma}}) \prod_{i=2}^n \biggl(\frac{1}{2}\phi(X_i - X_1) + \frac{1}{2\tilde{\sigma}}\phi((X_i - X_1)/\tilde{\sigma})\biggr).$$

Now, since for all $z \neq 0$

$$\lim_{\tilde\sigma \downarrow 0} \frac{1}{\tilde\sigma} \phi(z/\tilde\sigma) = 0,$$

we have

$$\lim_{\tilde\sigma \downarrow 0} \prod_{i=2}^n \biggl(\frac{1}{2}\phi(X_i - X_1) + \frac{1}{2\tilde\sigma}\phi((X_i - X_1)/\tilde\sigma) \biggr) = \prod_{i=2}^n \frac{1}{2}\phi(X_i - X_1) > 0.$$

It follows that

$$\lim_{\tilde\sigma \downarrow 0} L_{\mathbf{X}}(X_1, \tilde\sigma^2) = \infty.$$

2.5 Asymptotic tests and confidence intervals based on the likelihood

This section is a look ahead for what is to come in Chapter 14. Suppose that Θ is an open subset of $\mathbb{R}^p$. Define the log-likelihood ratio

$$Z(\mathbf{X}, \theta) := 2 \biggl\{ \log L_{\mathbf{X}}(\hat\theta) - \log L_{\mathbf{X}}(\theta) \biggr\}.$$

Note that $Z(\mathbf{X}, \theta) \geq 0$, as $\hat\theta$ maximizes the (log)-likelihood. We will see in Chapter 14 that, under some regularity conditions,

$$Z(\mathbf{X}, \theta) \xrightarrow{\mathcal{D}_\theta} \chi^2_p, \ \forall \ \theta.$$

Here, " $\xrightarrow{\mathcal{D}_\theta}$ " means convergence in distribution under $\mathbb{P}_\theta$, and χ^2_p denotes the Chi-squared distribution with p degrees of freedom.

We say that $Z(\mathbf{X}, \theta)$ is an asymptotic pivot: its asymptotic distribution does not depend on the unknown parameter θ. For the null-hypothesis
$H_0 : \ \theta = \theta_0$,
a test at asymptotic level α is: reject H_0 if $Z(\mathbf{X}, \theta_0) > \chi^2_p(1-\alpha)$, where $\chi^2_p(1-\alpha)$ is the $(1-\alpha)$-quantile of the χ^2_p-distribution. An asymptotic $(1-\alpha)$-confidence set for θ is

$$\{\theta : \ Z(\mathbf{X}, \theta) \leq \chi^2_p(1-\alpha)\}$$
$$= \{\theta : 2\log L_{\mathbf{X}}(\hat\theta) \leq 2 \log L_{\mathbf{X}}(\theta) + \chi^2_p(1-\alpha)\}.$$

Example 2.5.1 Likelihood ratio for the normal distribution
Here is a toy example. Let X have the $\mathcal{N}(\mu, 1)$-distribution, with $\mu \in \mathbb{R}$ unknown. The MLE of μ is the sample average $\hat\mu = \bar X$. It holds that

$$\log L_{\mathbf{X}}(\hat\mu) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^n (X_i - \bar X)^2,$$

and

$$2\biggl\{ \log L_{\mathbf{X}}(\hat\mu) - \log L_{\mathbf{X}}(\mu) \biggr\} = n(\bar X - \mu)^2.$$

The random variable $\sqrt{n}(\bar X - \mu)$ is $\mathcal{N}(0,1)$-distributed under $\mathbb{P}_\mu$. So its square, $n(\bar X - \mu)^2$, has a χ^2_1-distribution. Thus, in this case the above test (confidence interval) is exact.

Chapter 3

Intermezzo: distribution theory

3.1 Conditional distributions

Recall the definition of conditional probabilities: for two sets A and B, with $P(B) \neq 0$, the conditional probability of A given B is defined as

$$P(A|B) := \frac{P(A \cap B)}{P(B)}.$$

It follows that

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

and that, for a partition $\{B_j\}$[1]

$$P(A) = \sum_j P(A|B_j)P(B_j).$$

Consider now two random vectors $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$. Let $f_{X,Y}(\cdot,\cdot)$, be the density of (X, Y) with respect to Lebesgue measure (assumed to exist). The marginal density of X is

$$f_X(\cdot) = \int f_{X,Y}(\cdot, y)dy,$$

and the marginal density of Y is

$$f_Y(\cdot) = \int f_{X,Y}(x, \cdot)dx.$$

Definition 3.1.1 *The* conditional density *of X given $Y = y$ is*

$$f_X(x|y) := \frac{f_{X,Y}(x, y)}{f_Y(y)},\ x \in \mathbb{R}^n.$$

[1] $\{B_j\}$ is a partition if $B_j \cap B_k = \emptyset$ for all $j \neq k$ and $P(\cup_j B_j) = 1$.

Thus, we have

$$f_Y(y|x) = f_X(x|y)\frac{f_Y(y)}{f_X(x)},\ (x,y) \in \mathbb{R}^{n+m},$$

and

$$f_X(x) = \int f_X(x|y)f_Y(y)dy,\ x \in \mathbb{R}^n.$$

Definition 3.1.2 *Let* $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ *be some function. The* conditional expectation *of* $g(X,Y)$ *given* $Y = y$ *is*

$$E[g(X,Y)|Y=y] := \int f_X(x|y)g(x,y)dx.$$

Note thus that

$$E[g_1(X)g_2(Y)|Y=y] = g_2(y)E[g_1(X)|Y=y].$$

Notation We define the random variable $E[g(X,Y)|Y]$ as

$$E[g(X,Y)|Y] := h(Y),$$

where $h(y)$ is the function $h(y) := E[g(X,Y)|Y=y]$.

Lemma 3.1.1 *(Iterated expectations lemma) It holds that*

$$E\Big[E[g(X,Y)|Y]\Big] = Eg(X,Y).$$

Proof. Define

$$h(y) := E[g(X,Y)|Y=y].$$

Then

$$Eh(Y) = \int h(y)f_Y(y)dy = \int E[g(X,Y)|Y=y]f_Y(y)dy$$
$$= \int\int g(x,y)f_{X,Y}(x,y)dxdy = Eg(X,Y).$$

□

3.2 The multinomial distribution

In a survey, people were asked their opinion about some political issue. Let X be the number of `yes`-answers, Y be the number of `no` and Z be the number of `perhaps`. The total number of people in the survey is $n = X + Y + Z$. We consider the votes as a sample with replacement, with $p_1 = P(\texttt{yes})$, $p_2 = P(\texttt{no})$, and $p_3 = P(\texttt{perhaps})$, $p_1 + p_2 + p_3 = 1$. Then

$$P(X=x, Y=y, Z=z) = \binom{n}{x\ y\ z} p_1^x p_2^y p_3^z,\ (x,y,z) \in \{0,\ldots,n\},\ x+y+z=n.$$

Here
$$\binom{n}{x\ y\ z} := \frac{n!}{x!y!z!}.$$
It is called a *multinomial* coefficient.

Lemma 3.2.1 *The marginal distribution of X is the Binomial(n, p_1)-distribution.*

Proof. For $x \in \{0, \ldots, n\}$, we have

$$\begin{aligned} P(X = x) &= \sum_{y=0}^{n-x} P(X = x, Y = y, Z = n - x - y) \\ &= \sum_{y=0}^{n-x} \binom{n}{x\ y\ n-x-y} p_1^x p_2^y (1 - p_1 - p_2)^{n-x-y} \\ &= \binom{n}{x} p_1^x \sum_{y=0}^{n-x} \binom{n-x}{y} p_2^y (1 - p_1 - p_2)^{n-x-y} \\ &= \binom{n}{x} p_1^x (1 - p_1)^{n-x}. \end{aligned}$$

□

Definition 3.2.1 *We say that the random vector $(N_1, \ldots, N_k)$ has the* multinomial distribution *with parameters n and $p_1, \ldots, p_k$ (with $\sum_{j=1}^k p_j = 1$), if for all $(n_1, \ldots, n_k) \in \{0, \ldots, n\}^k$, with $n_1 + \cdots + n_k = n$, it holds that*

$$P(N_1 = n_1, \ldots, N_k = n_k) = \binom{n}{n_1\ \cdots\ n_k} p_1^{n_1} \cdots p_k^{n_k}.$$

Here
$$\binom{n}{n_1\ \cdots\ n_k} := \frac{n!}{n_1! \cdots n_k!}.$$

Example 3.2.1 Histograms
Let $X_1, \ldots, X_n$ be i.i.d. copies of a random variable $X \in \mathbb{R}$ with distribution F, and let $-\infty = a_0 < a_1 < \cdots < a_{k-1} < a_k = \infty$. Define, for $j = 1, \ldots, k$,

$$\begin{aligned} p_j &:= P(X \in (a_{j-1}, a_j]) &= F(a_j) - F(a_{j-1}), \\ \frac{N_j}{n} &:= \frac{\#\{X_i \in (a_{j-1}, a_j]\}}{n} &= \hat{F}_n(a_j) - \hat{F}_n(a_{j-1}). \end{aligned}$$

Then $(N_1, \ldots, N_k)$ has the Multinomial($n, p_1, \ldots, p_k$)-distribution.

3.3 The Poisson distribution

Definition 3.3.1 *A random variable $X \in \{0, 1, \ldots\}$ has the* Poisson *distribution with parameter $\lambda > 0$, if for all $x \in \{0, 1, \ldots\}$*

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

(see also Example 1.4.1).

Lemma 3.3.1 *Suppose X and Y are independent, and that X has the Poisson(λ)-distribution, and Y the Poisson(μ)-distribution. Then $Z := X + Y$ has the Poisson($\lambda + \mu$)-distribution.*

Proof. For all $z \in \{0, 1, \ldots\}$, we have

$$\begin{aligned} P(Z = z) &= \sum_{x=0}^{z} P(X = x, Y = z - x) \\ &= \sum_{x=0}^{z} P(X = x)P(Y = z - x) \\ &= \sum_{x=0}^{z} \mathrm{e}^{-\lambda} \frac{\lambda^x}{x!} \mathrm{e}^{-\mu} \frac{\mu^{z-x}}{(z-x)!} \\ &= \mathrm{e}^{-(\lambda+\mu)} \frac{1}{z!} \sum_{x=0}^{z} \binom{z}{x} \lambda^x \mu^{z-x} \\ &= \mathrm{e}^{-(\lambda+\mu)} \frac{(\lambda+\mu)^z}{z!}. \end{aligned}$$

□

Lemma 3.3.2 *Let $X_1, \ldots, X_n$ be independent, and (for $i = 1, \ldots, n$), let X_i have the Poisson(λ_i)-distribution. Define $Z := \sum_{i=1}^n X_i$. Let $z \in \{0, 1, \ldots\}$. Then the conditional distribution of $(X_1, \ldots, X_n)$ given $Z = z$ is the multinomial distribution with parameters z and $p_1, \ldots, p_n$, where*

$$p_j = \frac{\lambda_j}{\sum_{i=1}^n \lambda_i}, \ j = 1, \ldots, n.$$

Proof. First note that Z is Poisson(λ_+)-distributed, with $\lambda_+ := \sum_{i=1}^n \lambda_i$. Thus, for all $(x_1, \ldots, x_n) \in \{0, 1, \ldots, z\}^n$ satisfying $\sum_{i=1}^n x_i = z$, we have

$$\begin{aligned} P(X_1 = x_1, \ldots, X_n = x_n | Z = z) &= \frac{P(X_1 = x_1, \ldots, X_n = x_n)}{P(Z = z)} \\ &= \frac{\prod_{i=1}^n \left(\mathrm{e}^{-\lambda_i} \lambda_i^{x_i} / x_i! \right)}{\mathrm{e}^{-\lambda_+} \lambda_+^z / z!} \\ &= \binom{z}{x_1 \ \cdots \ x_n} \left(\frac{\lambda_1}{\lambda_+} \right)^{x_1} \cdots \left(\frac{\lambda_n}{\lambda_+} \right)^{x_n}. \end{aligned}$$

□

3.4 The distribution of the maximum of two random variables

Let X_1 and X_2 be independent and both have distribution F. Suppose that F has density f w.r.t. Lebesgue measure. Let

$$Z := \max\{X_1, X_2\}.$$

Lemma 3.4.1 *The distribution function of Z is F^2. Moreover, Z has density*

$$f_Z(z) = 2F(z)f(z),\ z \in \mathbb{R}.$$

Proof. We have for all z,

$$\begin{aligned} P(Z \le z) &= P(\max\{X_1, X_2\} \le z) \\ &= P(X_1 \le z, X_2 \le z) = F^2(z). \end{aligned}$$

If F has density f, then (Lebesgue)-almost everywhere,

$$f(z) = \frac{d}{dz}F(z).$$

So the derivative of F^2 exists almost everywhere, and

$$\frac{d}{dz}F^2(z) = 2F(z)f(z).$$

□

The conditional distribution function of X_1 given $Z = z$ is

$$F_{X_1}(x_1|z) = \begin{cases} \frac{F(x_1)}{2F(z)}, & x_1 < z \\ 1, & x_1 \ge z \end{cases}.$$

Note thus that this distribution has a jump of size $\frac{1}{2}$ at z.

Chapter 4

Sufficiency and exponential families

In this chapter, we denote the data by $X \in \mathcal{X}$. (In examples X is often replaced by $\mathbf{X} = (X_1, \ldots, X_n)$ with $X_1, \ldots, X_n$ i.i.d. copies of X.) We assume X has distribution $P \in \{P_\theta : \ \theta \in \Theta\}$.

4.1 Sufficiency

Let $S : \mathcal{X} \to \mathcal{Y}$ be some given map. We consider the statistic $S = S(X)$. Throughout, by the phrase *for all possible* s, we mean for all s for which conditional distributions given $S = s$ are defined (in other words: for all s in the support of the distribution of S, which may depend on θ).

Definition 4.1.1 *We call S* sufficient *for $\theta \in \Theta$ if for all θ, and all possible s, the conditional distribution*

$$P_\theta(X \in \cdot | S(X) = s)$$

does not depend on θ.

Example 4.1.1 Sufficiency and Bernoulli trials
Let $\mathbf{X} = (X_1, \ldots, X_n)$ with $X_1, \ldots, X_n$ i.i.d. with the Bernoulli distribution with probability $\theta \in (0, 1)$ of success: (for $i = 1, \ldots, n$)

$$P_\theta(X_i = 1) = 1 - P_\theta(X_i = 0) = \theta.$$

Take $S = \sum_{i=1}^n X_i$. Then S is sufficient for θ: for all possible s,

$$\mathbb{P}_\theta\biggl(X_1 = x_1, \ldots, X_n = x_n \biggm| S = s\biggr) = \frac{1}{\binom{n}{s}}, \ \sum_{i=1}^n x_i = s.$$

Example 4.1.2 Sufficiency and the Poisson distribution
Let $\mathbf{X} := (X_1, \ldots, X_n)$, with $X_1, \ldots, X_n$ i.i.d. and Poisson(θ)-distributed. Take

$S = \sum_{i=1}^n X_i$. Then S has the Poisson($n\theta$)-distribution. For all possible s, the conditional distribution of $\mathbf{X}$ given $S = s$ is the multinomial distribution with parameters s and $(p_1, \ldots, p_n) = (\frac{1}{n}, \ldots, \frac{1}{n})$:

$$\mathbb{P}_\theta\biggl(X_1 = x_1, \ldots, X_n = x_n \biggm| S = s\biggr) = \binom{s}{x_1 \ \cdots \ x_n}\left(\frac{1}{n}\right)^s, \ \sum_{i=1}^n x_i = s.$$

This distribution does not depend on θ, so S is sufficient for θ.

Example 4.1.3 Sufficiency and the exponential distribution
Let X_1 and X_2 be independent, and both have the exponential distribution with parameter $\theta > 0$. The density of e.g., X_1 is then

$$f_{X_1}(x; \theta) = \theta \mathrm{e}^{-\theta x}, \ x > 0.$$

Let $S = X_1 + X_2$. Verify that S has density

$$f_S(s; \theta) = s\theta^2 \mathrm{e}^{-\theta s}, \ s > 0.$$

(This is the Gamma(2, θ)-distribution.) For all possible s, the conditional density of (X_1, X_2) given $S = s$ is thus

$$f_{X_1,X_2}(x_1, x_2 | S = s) = \frac{1}{s}, \ x_1 + x_2 = s.$$

Hence, S is sufficient for θ.

Example 4.1.4 Sufficiency of the order statistics
Let $X_1, \ldots, X_n$ be an i.i.d. sample from a continuous distribution F. Then $S := (X_{(1)}, \ldots, X_{(n)})$ is sufficient for F: for all possible $s = (s_1, \ldots, s_n)$ $(s_1 < \ldots < s_n)$, and for $(x_{q_1}, \ldots, x_{q_n}) = s$,

$$\mathbb{P}_\theta\biggl((X_1, \ldots, X_n) = (x_1, \ldots, x_n) \biggm| (X_{(1)}, \ldots, X_{(n)}) = s\biggr) = \frac{1}{n!}.$$

Example 4.1.5 Sufficiency and the uniform distribution
Let X_1 and X_2 be independent, and both uniformly distributed on the interval $[0, \theta]$, with $\theta > 0$. Define $Z := X_1 + X_2$.

Lemma *The random variable Z has density*

$$f_Z(z; \theta) = \begin{cases} z/\theta^2 & \textit{if } 0 \le z \le \theta \\ (2\theta - z)/\theta^2 & \textit{if } \theta \le z \le 2\theta \end{cases}.$$

Proof. First, assume $\theta = 1$. Then the distribution function of Z is

$$F_Z(z) = \begin{cases} z^2/2 & 0 \le z \le 1 \\ 1 - (2 - z)^2/2 & 1 \le z \le 2 \end{cases}.$$

So the density is then

$$f_Z(z) = \begin{cases} z & 0 \le z \le 1 \\ 2 - z & 1 \le z \le 2 \end{cases}.$$

For general θ, the result follows from the uniform case by the transformation $Z \mapsto \theta Z$, which maps f_Z into $f_Z(\cdot/\theta)/\theta$. □

The conditional density of (X_1, X_2) given $Z = z \in (0, 2\theta)$ depends on θ, so Z is not sufficient for θ.

Consider now $S := \max\{X_1, X_2\}$. The conditional density of (X_1, X_2) given $S = s \in (0, \theta)$ is

$$f_{X_1,X_2}(x_1, x_2|S = s) = \frac{1}{2s},\ 0 \le x_1 < s,\ x_2 = s \text{ or } x_1 = s,\ 0 \le x_2 < s.$$

This does not depend on θ, so S is sufficient for θ.

4.2 Factorization Theorem of Neyman

Theorem 4.2.1 *(Factorization Theorem of Neyman) Suppose $\{P_\theta : \theta \in \Theta\}$ is dominated by a σ-finite measure ν. Let $p_\theta := dP_\theta/d\nu$ denote the densities. Then S is sufficient for θ if and only if one can write p_θ in the form*

$$p_\theta(x) = g_\theta(S(x))h(x),\ \forall\ x,\ \theta$$

for some functions $g_\theta(\cdot) \ge 0$ and $h(\cdot) \ge 0$.

Proof in the discrete case. Suppose X takes only the values $a_1, a_2, \ldots\ \forall\ \theta$ (so we may take ν to be the counting measure). Let Q_θ be the distribution of S:

$$Q_\theta(s) := \sum_{j:\ S(a_j)=s} P_\theta(X = a_j).$$

The conditional distribution of X given S is

$$P_\theta(X = x|S = s) = \frac{P_\theta(X = x)}{Q_\theta(s)},\ S(x) = s.$$

($\Rightarrow$) If S is sufficient for θ, the above does not depend on θ, but is only a function of x, say $h(x)$. So we may write for $S(x) = s$,

$$P_\theta(X = x) = P_\theta(X = x|S = s)Q_\theta(S = s) = h(x)g_\theta(s),$$

with $g_\theta(s) = Q_\theta(S = s)$.

($\Leftarrow$) Inserting $p_\theta(x) = g_\theta(S(x))h(x)$, we find

$$Q_\theta(s) = g_\theta(s) \sum_{j:\ S(a_j)=s} h(a_j),$$

This gives in the formula for $P_\theta(X = x|S = s)$,

$$P_\theta(X = x|S = s) = \frac{h(x)}{\sum_{j:\ S(a_j)=s} h(a_j)}$$

which does not depend on θ. □

Remark The proof for the general case is along the same lines, but does have some subtle elements!

Example 4.2.1 Sufficiency for the uniform distribution with unknown endpoint
Let $X_1, \ldots, X_n$ be i.i.d., and uniformly distributed on the interval $[0, \theta]$. Then the density of $\mathbf{X} = (X_1, \ldots, X_n)$ is

$$\begin{aligned}\mathbf{p}_\theta(x_1, \ldots, x_n) &= \frac{1}{\theta^n}\mathrm{l}\{0 \le \min\{x_1, \ldots, x_n\} \le \max\{x_1, \ldots, x_n\} \le \theta\} \\ &= g_\theta(S(x_1, \ldots, x_n))h(x_1, \ldots, x_n),\end{aligned}$$

with

$$g_\theta(s) := \frac{1}{\theta^n}\mathrm{l}\{s \le \theta\},$$

and

$$h(x_1, \ldots, x_n) := \mathrm{l}\{0 \le \min\{x_1, \ldots, x_n\}\}.$$

Thus, $S = \max\{X_1, \ldots, X_n\}$ is sufficient for θ.

Corollary 4.2.1 *The likelihood is $L_X(\theta) = p_\theta(X) = g_\theta(S)h(X)$. Hence, the maximum likelihood estimator $\hat{\theta} = \arg\max_\theta L_X(\theta) = \arg\max_\theta g_\theta(S)$ depends only on the sufficient statistic S.*

4.3 Exponential families

Definition 4.3.1 *A* k-dimensional exponential family *is a family of distributions $\{P_\theta : \theta \in \Theta\}$, dominated by some σ-finite measure ν, with densities $p_\theta = dP_\theta / d\nu$ of the form*

$$p_\theta(x) = \exp\Bigg[\sum_{j=1}^{k} c_j(\theta)T_j(x) - d(\theta)\Bigg]h(x).$$

Note In case of a k-dimensional exponential family, the k-dimensional statistic $S(X) = (T_1(X), \ldots, T_k(X))$ is sufficient for θ.

Note If $X_1, \ldots, X_n$ is an i.i.d. sample from a k-dimensional exponential family, then the distribution of $\mathbf{X} = (X_1, \ldots, X_n)$ is also in a k-dimensional exponential family. The density of $\mathbf{X}$ is then (for $\mathbf{x} := (x_1, \ldots, x_n)$),

$$\mathbf{p}_\theta(\mathbf{x}) = \prod_{i=1}^{n} p_\theta(x_i) = \exp[\sum_{j=1}^{k} nc_j(\theta)\bar{T}_j(\mathbf{x}) - nd(\theta)] \prod_{i=1}^{n} h(x_i),$$

where, for $j = 1, \ldots, k$,

$$\bar{T}_j(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} T_j(x_i).$$

Hence $S(\mathbf{X}) = (\bar{T}_1(\mathbf{X}), \ldots, \bar{T}_k(\mathbf{X}))$ is then sufficient for θ.

Note The functions $\{T_j\}$ and $\{c_j\}$ are not uniquely defined.

Example 4.3.1 Poisson distribution
If X is Poisson(θ)-distributed, we have

$$\begin{aligned} p_\theta(x) &= e^{-\theta}\frac{\theta^x}{x!} \\ &= \exp[x\log\theta - \theta]\frac{1}{x!}. \end{aligned}$$

Hence, we may take $T(x) = x$, $c(\theta) = \log\theta$, and $d(\theta) = \theta$.

Example 4.3.2 Binomial distribution
If X has the Binomial(n, θ)-distribution, we have

$$\begin{aligned} p_\theta(x) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= \binom{n}{x}\left(\frac{\theta}{1-\theta}\right)^x(1-\theta)^n \\ &= \binom{n}{x}\exp\left[x\log(\frac{\theta}{1-\theta}) + n\log(1-\theta)\right]. \end{aligned}$$

So we can take $T(x) = x$, $c(\theta) = \log(\theta/(1-\theta))$, and $d(\theta) = -n\log(1-\theta)$.

Example 4.3.3 Negative binomial distribution
If X has the Negative Binomial(m, θ)-distribution with m known we have

$$\begin{aligned} p_\theta(x) &= \frac{\Gamma(x+m)}{\Gamma(m)x!}\theta^m(1-\theta)^x \\ &= \frac{\Gamma(x+m)}{\Gamma(m)x!}\exp[x\log(1-\theta) + k\log(\theta)]. \end{aligned}$$

So we may take $T(x) = x$, $c(\theta) = \log(1-\theta)$ and $d(\theta) = -m\log(\theta)$.

Example 4.3.4 Gamma distribution with one parameter
Let X have the Gamma(m, θ)-distribution with m known. Then

$$\begin{aligned} p_\theta(x) &= \mathrm{e}^{-\theta x}x^{m-1}\frac{\theta^m}{\Gamma(m)} \\ &= \frac{x^{m-1}}{\Gamma(m)}\exp[-\theta x + m\log\theta]. \end{aligned}$$

So we can take $T(x) = x$, $c(\theta) = -\theta$, and $d(\theta) = -m\log\theta$.

Example 4.3.5 Gamma distribution with two parameters
Let X have the Gamma(m, λ)-distribution, and let $\theta = (m, \lambda)$. Then

$$\begin{aligned} p_\theta(x) &= \mathrm{e}^{-\lambda x} x^{m-1} \frac{\lambda^m}{\Gamma(m)} \\ &= \exp[-\lambda x + (m-1)\log x + m \log \lambda - \log \Gamma(m)]. \end{aligned}$$

So we can take $T_1(x) = x$, $T_2(x) = \log x$, $c(\theta) = -\lambda$, $c_2(\theta) = (m-1)$, and $d(\theta) = -m \log \lambda + \log \Gamma(m)$.

Example 4.3.6 Normal distribution
Let X be $\mathcal{N}(\mu, \sigma^2)$-distributed, and let $\theta = (\mu, \sigma)$. Then

$$\begin{aligned} p_\theta(x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{\sqrt{2\pi}} \exp\left[\frac{x\mu}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \log \sigma\right]. \end{aligned}$$

So we can take $T_1(x) = x$, $T_2(x) = x^2$, $c_1(\theta) = \mu/\sigma^2$, $c_2(\theta) = -1/(2\sigma^2)$, and $d(\theta) = \mu^2/(2\sigma^2) + \log(\sigma)$.

4.4 Intermezzo: the mean and covariance matrix of a random vector

Let $Z \in \mathbb{R}^k$ be a random vector. Then its mean (if it exists) is defined as the vector consisting of the means of each entry of Z:

$$EZ = \begin{pmatrix} EZ_1 \\ \vdots \\ EZ_k \end{pmatrix}.$$

The covariance matrix of X (if it exists) is defined as the symmetric $k \times k$ matrix Σ containing the (co)variances between each pair of entries in X:

$$\Sigma := EZZ' - EZEZ' = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,k} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,k} & \sigma_{2,k} & \cdots & \sigma_k^2 \end{pmatrix}.$$

Here, Z' denotes the transpose[1] of the vector Z, $\sigma_j^2 := \mathrm{var}(Z_j)$ $(j = 1, \ldots, k)$ and for all $j_1 \neq j_2$ $\sigma_{j_1,j_2} = \mathrm{cov}(Z_{j_1}, Z_{j_2})$. We will often write the covariance matrix Σ as $\mathrm{Cov}(Z)$.

[1] We alternatively write the transpose of a vector, say v, as v^T.

4.5 Canonical form of an exponential family

In this subsection, we assume regularity conditions, such as existence of derivatives, and inverses, and permission to interchange differentiation and integration.

Let $\Theta \subset \mathbb{R}^k$, and let $\{P_\theta : \theta \in \Theta\}$ be a family of probability measures dominated by a σ-finite measure ν. Define the densities

$$p_\theta := \frac{dP_\theta}{d\nu}.$$

Definition *We call* $\{P_\theta : \theta \in \Theta\}$ *an exponential family in* canonical form*, if*

$$p_\theta(x) = \exp\left[\sum_{j=1}^k \theta_j T_j(x) - d(\theta)\right] h(x).$$

Note that $d(\theta)$ is the normalizing constant

$$d(\theta) = \log\left(\int \exp\left[\sum_{j=1}^k \theta_j T_j(x)\right] h(x) d\nu(x)\right).$$

We let

$$\dot{d}(\theta) := \frac{\partial}{\partial\theta} d(\theta)$$

denote the vector of first derivatives. Let

$$\ddot{d}(\theta) := \frac{\partial^2}{\partial\theta\partial\theta'} d(\theta) = \left(\frac{\partial^2}{\partial\theta_{j_1}\partial\theta_{j_2}} d(\theta)\right)$$

denote the $k \times k$ matrix of second derivatives. Further, we write

$$T(X) := \begin{pmatrix} T_1(X) \\ \vdots \\ T_k(X) \end{pmatrix},\ E_\theta T(X) := \begin{pmatrix} E_\theta T_1(X) \\ \vdots \\ E_\theta T_k(X) \end{pmatrix},$$

and we write the $k \times k$ covariance matrix of $T(X)$ as

$$\mathrm{Cov}_\theta(T(X)) := E_\theta T(X) T'(X) - E_\theta T(X) E_\theta T'(X).$$

Lemma 4.5.1 *We have (under regularity)*

$$E_\theta T(X) = \dot{d}(\theta),\ \mathrm{Cov}_\theta(T(X)) = \ddot{d}(\theta).$$

Proof. By the definition of $d(\theta)$, we find

$$\begin{aligned}
\dot{d}(\theta) &= \frac{\partial}{\partial\theta}\log\left(\int \exp\left[\theta' T(x)\right]h(x)d\nu(x)\right) \\
&= \frac{\int \exp\left[\theta' T(x)\right]T(x)h(x)d\nu(x)}{\int \exp\left[\theta' T(x)\right]h(x)d\nu(x)} \\
&= \int \exp\left[\theta' T(x) - d(\theta)\right]T(x)h(x)d\nu(x) \\
&= \int p_\theta(x)T(x)d\nu(x) = E_\theta T(X),
\end{aligned}$$

and (omitting the integration x variable to shorten the expressions)

$$\begin{aligned}
\ddot{d}(\theta) &= \frac{\int \exp\left[\theta' T\right]TT'hd\nu}{\int \exp\left[\theta' T\right]hd\nu} \\
&- \frac{\left(\int \exp\left[\theta' T\right]Thd\nu\right)\left(\int \exp\left[\theta' T\right]Thd\nu\right)'}{\left(\int \exp\left[\theta' T\right]hd\nu\right)^2} \\
&= \int \exp\left[\theta' T - d(\theta)\right]TT'hd\nu \\
&- \left(\int \exp\left[\theta' T - d(\theta)\right]Thd\nu\right) \times \left(\int \exp\left[\theta' T - d(\theta)\right]T'hd\nu\right) \\
&= \int TT'p_\theta d\nu - \left(\int p_\theta T d\nu\right)\left(\int p_\theta T' d\nu\right) \\
&= E_\theta T(X)T'(X) - \left(E_\theta T(X)\right)\left(E_\theta T'(X)\right) \\
&= \mathrm{Cov}_\theta(T(X)).
\end{aligned}$$

□

4.6 Reparametrizing in the one-dimensional case

Let us now simplify to the one-dimensional case, that is $\Theta \subset \mathbb{R}$. Consider an exponential family, not necessarily in canonical form:

$$p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x).$$

We can put this in canonical form by reparametrizing

$$\theta \mapsto c(\theta) := \gamma \text{ (say)},$$

to get

$$\tilde{p}_\gamma(x) = \exp[\gamma T(x) - d_0(\gamma)]h(x),$$

where when c is one-to-one

$$d_0(\gamma) = d(c^{-1}(\gamma)).$$

It follows that

$$E_\theta T(X) = \dot{d}_0(\gamma) = \frac{\dot{d}(c^{-1}(\gamma))}{\dot{c}(c^{-1}(\gamma))} = \frac{\dot{d}(\theta)}{\dot{c}(\theta)}, \qquad (4.1)$$

and

$$\begin{aligned}
\text{var}_\theta(T(X)) &= \ddot{d}_0(\gamma) = \frac{\ddot{d}(c^{-1}(\gamma))}{[\dot{c}(c^{-1}(\gamma))]^2} - \frac{\dot{d}(c^{-1}(\gamma))\ddot{c}(c^{-1}(\gamma))}{[\dot{c}(c^{-1}(\gamma))]^3} \\
&= \frac{\ddot{d}(\theta)}{[\dot{c}(\theta)]^2} - \frac{\dot{d}(\theta)\ddot{c}(\theta)}{[\dot{c}(\theta)]^3} \\
&= \frac{1}{[\dot{c}(\theta)]^2}\left(\ddot{d}(\theta) - \frac{\dot{d}(\theta)}{\dot{c}(\theta)}\ddot{c}(\theta)\right).
\end{aligned}$$

4.7 Score function and Fisher information

Consider an arbitrary (but regular) family of densities $\{p_\theta : \theta \in \Theta\}$, with (again for simplicity) $\Theta \subset \mathbb{R}$.

Definition 4.7.1 *The* score function *is*

$$s_\theta(x) := \frac{d}{d\theta}\log p_\theta(x).$$

The Fisher information *for estimating θ is*

$$I(\theta) := \text{var}_\theta(s_\theta(X)).$$

More generally, the Fisher information for estimating a differentiable function $g(\theta)$ of the parameter θ, is equal to $I(\theta)/[\dot{g}(\theta)]^2$.

See also Chapters 5 and 13.

Lemma 4.7.1 *We have (under regularity)*

$$E_\theta s_\theta(X) = 0,$$

and

$$I(\theta) = -E_\theta \dot{s}_\theta(X),$$

where $\dot{s}_\theta(x) := \frac{d}{d\theta}s_\theta(x)$.

Proof. The results follow from the fact that densities integrate to one, assuming that we may interchange derivatives and integrals:

$$\begin{aligned}
E_\theta s_\theta(X) &= \int s_\theta(x) p_\theta(x) d\nu(x) \\
&= \int \frac{d \log p_\theta(x)}{d\theta} p_\theta(x) d\nu(x) = \int \frac{\dot{p}_\theta(x)}{p_\theta(x)} p_\theta(x) d\nu(x) \\
&= \int \dot{p}_\theta(x) d\nu(x) = \frac{d}{d\theta} \int p_\theta(x) d\nu(x) = \frac{d}{d\theta} 1 = 0,
\end{aligned}$$

and

$$\begin{aligned}
E_\theta \dot{s}_\theta(X) &= E_\theta \left[\frac{\ddot{p}_\theta(X)}{p_\theta(X)} - \left(\frac{\dot{p}_\theta(X)}{p_\theta(X)} \right)^2 \right] \\
&= E_\theta \left[\frac{\ddot{p}_\theta(X)}{p_\theta(X)} \right] - E_\theta s_\theta^2(X).
\end{aligned}$$

Now, $E_\theta s_\theta^2(X)$ equals $\text{var}_\theta s_\theta(X)$, since $E_\theta s_\theta(X) = 0$. Moreover,

$$\begin{aligned}
E_\theta \left[\frac{\ddot{p}_\theta(X)}{p_\theta(X)} \right] &= \int \frac{d^2}{d\theta^2} p_\theta(x) d\nu(x) \\
&= \frac{d^2}{d\theta^2} \int p_\theta(x) d\nu(x) \\
&= \frac{d^2}{d\theta^2} 1 = 0.
\end{aligned}$$

□

4.8 Score function for exponential families

In the special case that $\{P_\theta : \theta \in \Theta\}$ is a one-dimensional exponential family, the densities are of the form

$$p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x).$$

Hence

$$s_\theta(x) = \dot{c}(\theta)T(x) - \dot{d}(\theta).$$

The equality $E_\theta s_\theta(X) = 0$ implies that

$$E_\theta T(X) = \frac{\dot{d}(\theta)}{\dot{c}(\theta)},$$

which re-establishes (4.1). One moreover has

$$\dot{s}_\theta(x) = \ddot{c}(\theta)T(x) - \ddot{d}(\theta).$$

Hence, the inequality $\text{var}_\theta(s_\theta(X)) = -E_\theta \dot{s}_\theta(X)$ implies

$$\begin{aligned}
[\dot{c}(\theta)]^2 \text{var}_\theta(T(X)) &= -\ddot{c}(\theta) E_\theta T(X) + \ddot{d}(\theta) \\
&= \ddot{d}(\theta) - \frac{\dot{d}(\theta)}{\dot{c}(\theta)} \ddot{c}(\theta),
\end{aligned}$$

which re-establishes (4.2). In addition, it follows that

$$I(\theta) = \ddot{d}(\theta) - \frac{\dot{d}(\theta)}{\dot{c}(\theta)} \ddot{c}(\theta).$$

The Fisher information for estimating $\gamma = c(\theta)$ is

$$I_0(\gamma) = \ddot{d}_0(\gamma) = \frac{I(\theta)}{[\dot{c}(\theta)]^2}.$$

Example 4.8.1 Bernoulli-distribution in canonical form
Let $X \in \{0, 1\}$ have the Bernoulli-distribution with success parameter $\theta \in (0, 1)$:

$$p_\theta(x) = \theta^x (1-\theta)^{1-x} = \exp\left[x \log\left(\frac{\theta}{1-\theta} \right) + \log(1-\theta) \right], \ x \in \{0, 1\}.$$

We reparametrize:

$$\gamma := c(\theta) = \log\left(\frac{\theta}{1-\theta} \right),$$

which is called the log-odds ratio. Inverting gives

$$\theta = \frac{\mathrm{e}^\gamma}{1 + \mathrm{e}^\gamma},$$

and hence

$$d(\theta) = -\log(1-\theta) = \log\left(1 + \mathrm{e}^\gamma \right) := d_0(\gamma).$$

Thus

$$\dot{d}_0(\gamma) = \frac{\mathrm{e}^\gamma}{1 + \mathrm{e}^\gamma} = \theta = E_\theta X,$$

and

$$\ddot{d}_0(\gamma) = \frac{\mathrm{e}^\gamma}{1 + \mathrm{e}^\gamma} - \frac{\mathrm{e}^{2\gamma}}{(1 + \mathrm{e}^\gamma)^2} = \frac{\mathrm{e}^\gamma}{(1 + \mathrm{e}^\gamma)^2} = \theta(1-\theta) = \mathrm{var}_\theta(X).$$

The score function is

$$\begin{aligned} s_\theta(x) &= \frac{d}{d\theta}\left[x \log\left(\frac{\theta}{1-\theta} \right) + \log(1-\theta) \right] \\ &= \frac{x}{\theta(1-\theta)} - \frac{1}{1-\theta}. \end{aligned}$$

The Fisher information for estimating the success parameter θ is

$$E_\theta s_\theta^2(X) = \frac{\mathrm{var}_\theta(X)}{[\theta(1-\theta)]^2} = \frac{1}{\theta(1-\theta)},$$

whereas the Fisher information for estimating the log-odds ratio γ is

$$I_0(\gamma) = \theta(1-\theta).$$

4.9 Minimal sufficiency

Definition 4.9.1 *We say that two likelihoods $L_x(\theta)$ and $L_{\tilde{x}}(\theta)$ are proportional at $(x, \tilde{x})$, if*

$$L_x(\theta) = L_{\tilde{x}}(\theta) c(x, \tilde{x}), \forall \ \theta,$$

for some constant $c(x, \tilde{x})$.
We write this as

$$L_x(\theta) \propto L_{\tilde{x}}(\theta).$$

A sufficient statistic S is called minimal sufficient *if $S(x) = S(\tilde{x})$ for all x and $\tilde{x}$ for which the likelihoods are proportional.*

Example 4.9.1 Minimal sufficiency for the normal distribution
Let $X_1 \ldots, X_n$ be independent and $\mathcal{N}(\theta, 1)$-distributed. Then $S = \sum_{i=1}^n X_i$ is sufficient for θ. We moreover have

$$\log L_{\mathbf{x}}(\theta) = S(\mathbf{x})\theta - \frac{n\theta^2}{2} - \frac{\sum_{i=1}^n x_i^2}{2} - \log(2\pi)/2.$$

So

$$\log L_{\mathbf{x}}(\theta) - \log L_{\tilde{\mathbf{x}}}(\theta) = (S(\mathbf{x}) - S(\tilde{\mathbf{x}}))\theta - \frac{\sum_{i=1}^n x_i^2 - \sum_{i=1}^n (\tilde{x}_i)^2}{2},$$

which equals,

$$\log c(\mathbf{x}, \tilde{\mathbf{x}}), \ \forall \ \theta,$$

for some function c, if and only if $S(\mathbf{x}) = S(\tilde{\mathbf{x}})$. So S is minimal sufficient.

Example 4.9.2 Minimal sufficiency for the Laplace distribution
Let $X_1, \ldots, X_n$ be independent and Laplace-distributed with location parameter θ. Then

$$\log L_{\mathbf{x}}(\theta) = -(\log 2)/2 - \sqrt{2} \sum_{i=1}^n |x_i - \theta|,$$

so

$$\log L_{\mathbf{x}}(\theta) - \log L_{\tilde{\mathbf{x}}}(\theta) = -\sqrt{2} \sum_{i=1}^n (|x_i - \theta| - |\tilde{x}_i - \theta|)$$

which equals

$$\log c(\mathbf{x}, \tilde{\mathbf{x}}), \ \forall \ \theta,$$

for some function c, if and only if $(x_{(1)}, \ldots, x_{(n)}) = (\tilde{x}_{(1)}, \ldots, \tilde{x}_{(n)})$. So the order statistics $X_{(1)}, \ldots, X_{(n)}$ are minimal sufficient.

Chapter 5

Bias, variance and the Cramér Rao lower bound

5.1 What is an unbiased estimator?

Let $X \in \mathcal{X}$ denote the observations. The distribution P of X is assumed to be a member of a given class $\{P_\theta : \ \theta \in \Theta\}$ of distributions. The parameter of interest is $\gamma := g(\theta)$, with $g : \Theta \to \mathbb{R}$. Except for the last section in this chapter, the parameter γ is assumed to be one-dimensional.

Let $T : \mathcal{X} \to \mathbb{R}$ be an estimator of $g(\theta)$.

Definition 5.1.1 *The* bias *of $T = T(X)$ is*

$$\text{bias}_\theta(T) := E_\theta T - g(\theta).$$

The estimator T is called unbiased *if*

$$\text{bias}_\theta(T) = 0, \ \forall \ \theta.$$

Thus, unbiasedness means that there is no systematic error: $E_\theta T = g(\theta)$. We require this **for all** θ.

Example 5.1.1 Unbiased estimators in the Binomial case
Let $X \sim \text{Binomial}(n, \theta)$, $0 < \theta < 1$. We have

$$E_\theta T(X) = \sum_{k=0}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} T(k) =: q(\theta).$$

Note that $q(\theta)$ is a polynomial in θ of degree at most n. So only parameters $g(\theta)$ which are polynomials of degree at most n can be estimated unbiasedly. It means that there exists no unbiased estimator of, for example, $\sqrt{\theta}$ or $\theta/(1-\theta)$.

Example 5.1.2 Unbiased estimators in the Poisson case
Let $X \sim$ Poisson(θ). Then

$$E_\theta T(X) = \sum_{k=0}^{\infty} \mathrm{e}^{-\theta} \frac{\theta^k}{k!} T(k) =: e^{-\theta} \mathrm{p}(\theta).$$

Note that $\mathrm{p}(\theta)$ is a power series in θ. Thus only parameters $g(\theta)$ which are a power series in θ times $\mathrm{e}^{-\theta}$ can be estimated unbiasedly. An example is the probability of early failure

$$g(\theta) := \mathrm{e}^{-\theta} = P_\theta(X = 0).$$

An unbiased estimator of $\mathrm{e}^{-\theta}$ is for instance

$$T(X) = \mathrm{l}\{X = 0\}.$$

As another example, suppose the parameter of interest is

$$g(\theta) := \mathrm{e}^{-2\theta}.$$

An unbiased estimator is

$$T(X) = \begin{cases} +1 & \textit{if } X \textit{ is even} \\ -1 & \textit{if } X \textit{ is odd} \end{cases} .$$

This estimator does not make sense at all!

Example 5.1.3 Unbiased estimator of the variance
Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$, and let $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$. Then

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

is an unbiased estimator of σ^2. But S is not an unbiased estimator of σ. In fact, one can show that there does not exist any unbiased estimator of σ!

We conclude that requiring unbiasedness can have disadvantages: unbiased estimators do not always exist, and if they do, they can be nonsensical. Moreover, the property of unbiasedness is not preserved under taking nonlinear transformations.

5.2 UMVU estimators

Definition 5.2.1 *The* mean square error *of T is*

$$\mathrm{MSE}_\theta(T) := E_\theta \biggl(T - g(\theta) \biggr)^2 .$$

Lemma 5.2.1 *We have the following decomposition for the mean square error:*

$$\mathrm{MSE}_\theta(T) = \mathrm{bias}_\theta^2(T) + \mathrm{var}_\theta(T).$$

Proof. Write $E_\theta := q(\theta)$. Then

$$\begin{aligned} E_\theta\Big(T-g(\theta)\Big)^2 &= \underbrace{E_\theta\Big(T-q(\theta)\Big)^2}_{=\mathrm{var}_\theta(T)} + \underbrace{\Big(q(\theta)-g(\theta)\Big)^2}_{=\mathrm{bias}^2_\theta(T)} \\ &+ 2\Big(q(\theta)-g(\theta)\Big)\underbrace{E_\theta\Big(T-q(\theta)\Big)}_{=0}. \end{aligned}$$

□

In other words, the mean square error consists of two components, the (squared) bias and the variance. This is called the bias-variance decomposition. As we will see, it is often the case that an attempt to decrease the bias results in an increase of the variance (and vise versa).

Example 5.2.1 Estimators in case of the normal distribution
Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$-distributed. Both μ and σ^2 are unknown parameters: $\theta := (\mu, \sigma^2)$.

Case i *Suppose the mean μ is our parameter of interest. Consider the estimator $T := a\bar{X}$, where $0 \le a \le 1$. Then the bias is decreasing in a, but the variance is increasing in a:*

$$\mathrm{MSE}_\theta(T) = E_\theta(T-\mu)^2 = (1-a)^2\mu^2 + a^2\sigma^2/n.$$

The right hand side can be minimized as a function of α. The minimum is attained at

$$a_{\mathrm{opt}} := \frac{\mu^2}{\sigma^2/n + \mu^2}.$$

However, $\alpha_{\mathrm{opt}}\bar{X}$ is not an estimator as it depends on the unknown parameters.

Case ii *Suppose σ^2 is the parameter of interest. Let S^2 be the sample variance:*

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

It is known that S^2 is unbiased. But does it also have small mean square error? Let us compare it with the estimator

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

To compute the mean square errors of these two estimators, we first recall that

$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1},$$

a χ^2-distribution with $n-1$ degrees of freedom. The χ^2-distribution is a special case of the Gamma-distribution, namely

$$\chi^2_{n-1} = \Gamma\left(\frac{n-1}{2}, \frac{1}{2}\right).$$

Thus [1]

$$E_\theta\left(\frac{\sum_{i=1}^n (X_i-\bar{X})^2}{\sigma^2}\right) = n-1,\ \mathrm{var}\left(\frac{\sum_{i=1}^n (X_i-\bar{X})^2}{\sigma^2}\right) = 2(n-1).$$

It follows that

$$\mathrm{MSE}_\theta(S^2) = E_\theta\left(S^2-\sigma^2\right)^2 = \mathrm{var}(S^2) = \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{n-1}.$$

Moreover

$$E_\theta\hat{\sigma}^2 = \frac{n-1}{n}\sigma^2,\ \mathrm{bias}_\theta(\hat{\sigma}^2) = -\frac{1}{n}\sigma^2,$$

so that

$$\begin{aligned}\mathrm{MSE}_\theta(\hat{\sigma}^2) &= E_\theta\left(\hat{\sigma}^2-\sigma^2\right)^2\\ &= \mathrm{bias}_\theta^2(\hat{\sigma}^2)+\mathrm{var}_\theta(\hat{\sigma}^2)\\ &= \frac{\sigma^4}{n^2}+\frac{2(n-1)\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2}.\end{aligned}$$

Conclusion: the mean square error of $\hat{\sigma}^2$ is smaller than the mean square error of S^2!

Generally, it is not possible to construct an estimator that possesses the best among all of all desirable properties. We therefore fix one property: unbiasedness (despite its disadvantages), and look for good estimators among the unbiased ones.

Definition 5.2.2 *An unbiased estimator T^* is called* UMVU *(*U*niform* M*inimum* V*ariance* U*nbiased) if for any other unbiased estimator T,*

$$\mathrm{var}_\theta(T^*) \le \mathrm{var}_\theta(T),\ \forall\ \theta.$$

Suppose that T is unbiased, and that S is sufficient. Let

$$T^* := E(T|S).$$

The distribution of T given S does not depend on θ, so T^* is also an estimator. Moreover, it is unbiased: by the iterated expectations lemma

$$E_\theta T^* = E_\theta(E(T|S)) = E_\theta T = g(\theta).$$

By conditioning on S, "superfluous" variance in the sample is killed. Indeed, the following lemma (which is a general property of conditional distributions) shows that T^* cannot have larger variance than T:

$$\mathrm{var}_\theta(T^*) \le \mathrm{var}_\theta(T),\ \forall\ \theta.$$

[1] If Y has a $\Gamma(k,\lambda)$-distribution, then $EY = k/\lambda$ and $\mathrm{var}(Y) = k/\lambda^2$.

Lemma 5.2.2 *Let Y and Z be two random variables. Then*

$$\text{var}(Y) = \text{var}(E(Y|Z)) + E\text{var}(Y|Z).$$

Proof. It holds that

$$\begin{aligned} \text{var}(E(Y|Z)) &= E\Big[E(Y|Z)\Big]^2 - \Big[E(E(Y|Z))\Big]^2 \\ &= E[E(Y|Z)]^2 - [EY]^2, \end{aligned}$$

and

$$\begin{aligned} E\text{var}(Y|Z) = E\Big[E(Y^2|Z) - [E(Y|Z)]^2\Big] \\ = EY^2 - E[E(Y|Z)]^2. \end{aligned}$$

Hence, when adding up, the term $E[E(Y|Z)]^2$ cancels out, and what is left over is exactly the variance

$$\text{var}(Y) = EY^2 - [EY]^2.$$

□

5.3 The Lehmann-Scheffé Lemma

The question arises: can we construct an unbiased estimator with even smaller variance than $T^* = E(T|S)$? Note that T^* depends on X only via $S = S(X)$, i.e., it depends only on the sufficient statistic. In our search for UMVU estimators, we may restrict our attention to estimators depending only on S. Thus, if there is only one unbiased estimator depending only on S, it has to be UMVU.

Definition 5.3.1 *A statistic S is called* complete *if we have the following implication:*

$$E_\theta h(S) = 0 \ \forall \ \theta \Rightarrow h(S) = 0, \ P_\theta - a.s., \forall \ \theta.$$

Here, h is a function of S not depending on θ.

Lemma 5.3.1 *(Lehmann-Scheffé) Let T be an unbiased estimator of $g(\theta)$, with, for all θ, finite variance. Moreover, let S be sufficient and complete. Then $T^* := E(T|S)$ is UMVU.*

Proof. We already noted that $T^* = T^*(S)$ is unbiased and that $\text{var}_\theta(T^*) \leq \text{var}_\theta(T) \ \forall \ \theta$. If $T'(S)$ is another unbiased estimator of $g(\theta)$, we have

$$E_\theta(T(S) - T'(S)) = 0, \forall \ \theta.$$

Because S is complete, this implies

$$T^* = T', \ P_\theta - a.s.$$

□

To check whether a statistic is complete, one often needs somewhat sophisticated tools from analysis/integration theory. In the next two examples, we only sketch the proofs of completeness.

Example 5.3.1 UMVU estimator in the Poisson case
Let $X_1, \ldots, X_n$ be i.i.d. Poisson(θ)*-distributed. We want to estimate $g(\theta) := \mathrm{e}^{-\theta}$, the probability of early failure. An unbiased estimator is*

$$T(X_1, \ldots, X_n) := \mathrm{l}\{X_1 = 0\}.$$

A sufficient statistic is

$$S := \sum_{i=1}^n X_i.$$

We now check whether S is complete. Its distribution is the Poisson$(n\theta)$*-distribution. We therefore have for any function h,*

$$E_\theta h(S) = \sum_{k=0}^\infty \mathrm{e}^{-n\theta} \frac{(n\theta)^k}{k!} h(k).$$

The equation

$$E_\theta h(S) = 0 \ \forall \ \theta,$$

thus implies

$$\sum_{k=0}^\infty \frac{(n\theta)^k}{k!} h(k) = 0 \ \forall \ \theta.$$

Let f be a function with Taylor expansion at zero. Then

$$f(x) = \sum_{k=0}^\infty \frac{x^k}{k!} f^{(k)}(0).$$

The left hand side can only be zero for all x if $f \equiv 0$, in which case also $f^{(k)}(0) = 0$ for all k. Thus ($h(k)$ takes the role of $f^{(k)}(0)$ and $n\theta$ the role of x), we conclude that $h(k) = 0$ for all k, i.e., that S is complete.

So we know from the Lehmann-Scheffé Lemma that $T^* := E(T|S)$ is UMVU. Let us now calculate T^*. First,

$$\begin{aligned} P(T = 1|S = s) &= P(X_1 = 0|S = s) \\ &= \frac{\mathrm{e}^{-\theta} e^{-(n-1)\theta}[(n-1)\theta]^s/s!}{\mathrm{e}^{-n\theta}(n\theta)^s/s!} \\ &= \left(\frac{n-1}{n}\right)^s. \end{aligned}$$

Hence

$$T^* = \left(\frac{n-1}{n}\right)^S$$

is UMVU.

Example 5.3.2 UMVU estimation for uniform distribution
Let $X_1, \ldots, X_n$ be i.i.d. Uniform$[0, \theta]$*-distributed, and $g(\theta) := \theta$. We know*

that $S := \max\{X_1, \ldots, X_n\}$ *is sufficient (see Example 4.2.1). The distribution function of* S *is*

$$F_S(s) = P_\theta(\max\{X_1, \ldots, X_n\} \le s) = \left(\frac{s}{\theta}\right)^n, \ 0 \le s \le \theta.$$

Its density is thus

$$f_S(s) = \frac{ns^{n-1}}{\theta^n}, \ 0 \le s \le \theta.$$

Hence, for any (measurable) function h*,*

$$E_\theta h(S) = \int_0^\theta h(s) \frac{ns^{n-1}}{\theta^n} ds.$$

If

$$E_\theta h(S) = 0 \ \forall \ \theta,$$

it must hold that

$$\int_0^\theta h(s) s^{n-1} ds = 0 \ \forall \ \theta.$$

Differentiating w.r.t. θ *gives*

$$h(\theta)\theta^{n-1} = 0 \ \forall \ \theta,$$

which implies $h \equiv 0$*. So* S *is complete.*

It remains to find a statistic T^* *that depends only on* S *and that is unbiased. We have*

$$E_\theta S = \int_0^\theta s \frac{ns^{n-1}}{\theta^n} ds = \frac{n}{n+1}\theta.$$

So S *itself is not unbiased, it is too small. But this can be easily repaired: take*

$$T^* = \frac{n+1}{n} S.$$

Then, by the Lehmann-Scheffé Lemma, T^* *is UMVU.*

5.4 Completeness for exponential families

In the case of an exponential family, completeness holds for a sufficient statistic if the parameter space is "of the same dimension" as the sufficient statistic. This is stated more formally in the following lemma. We omit the proof.

Lemma 5.4.1 *Let for* $\theta \in \Theta$*,*

$$p_\theta(x) = \exp\left[\sum_{j=1}^k c_j(\theta) T_j(x) - d(\theta)\right] h(x).$$

Consider the set

$$\mathcal{C} := \{(c_1(\theta), \ldots, c_k(\theta)) : \ \theta \in \Theta\} \subset \mathbb{R}^k.$$

Suppose that $\mathcal{C}$ is truly k-dimensional (that is, not of dimension smaller than k), i.e., it contains an open ball in $\mathbb{R}^k$. (Or an open cube $\prod_{j=1}^k (a_j, b_j)$.) Then $S := (T_1, \ldots, T_k)$ is complete.

Example 5.4.1 Completeness for Gamma distribution
Let $X_1, \ldots, X_n$ be i.i.d. with $\Gamma(k, \lambda)$-distribution. Both k and λ are assumed to be unknown, so that $\theta := (k, \lambda)$. We moreover let $\Theta := \mathbb{R}_+^2$. The density f of the $\Gamma(k, \lambda)$-distribution is

$$f(z) = \frac{\lambda^k}{\Gamma(k)} \mathrm{e}^{-\lambda z} z^{k-1}, \ z > 0.$$

Hence,

$$p_\theta(x) = \exp\left[-\lambda \sum_{i=1}^n x_i + (k-1) \sum_{i=1}^n \log x_i - d(\theta) \right] h(x),$$

where

$$d(k, \lambda) = -nk \log \lambda + n \log \Gamma(k),$$

and

$$h(x) = \mathrm{l}\{x_i > 0, \ i = 1, \ldots, n\}.$$

It follows that

$$\left(\sum_{i=1}^n X_i, \sum_{i=1}^n \log X_i \right)$$

is sufficient and complete.

Example 5.4.2 Completeness for two normal samples
Consider two independent samples from normal distributions: $X_1, \ldots X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$-distributed and $Y_1, \ldots, Y_m$ be i.i.d. $\mathcal{N}(\nu, \tau^2)$-distributed.

Case i *If $\theta = (\mu, \nu, \sigma^2, \tau^2) \in \mathbb{R}^2 \times \mathbb{R}_+^2$, one can easily check that*

$$S := \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2, \sum_{j=1}^m Y_j, \sum_{j=1}^m Y_j^2 \right)$$

is sufficient and complete.

Case ii *If μ, σ^2 and τ^2 are unknown, and $\nu = \mu$, then S of course remains sufficient. One can however show that S is not complete. Difficult question: does a sufficient and complete statistic exist?*

5.5 The Cramér Rao lower bound

Let $\{P_\theta : \theta \in \Theta\}$ be a collection of distributions on $\mathcal{X}$, dominated by a σ-finite measure ν. We denote the densities by

$$p_\theta := \frac{dP_\theta}{d\nu}, \ \theta \in \Theta.$$

In this section, we assume that Θ is a one-dimensional open interval (the extension to a higher-dimensional parameter space will be handled in the next section).

We will impose the following two conditions:

Condition I *The set*

$$A := \{x : \ p_\theta(x) > 0\}$$

does not depend on θ.

Condition II *(Differentiability in L_2) For all θ and for a function $s_\theta : \mathcal{X} \to \mathbb{R}$ satisfying*

$$I(\theta) := E_\theta s_\theta^2(X) < \infty,$$

it holds that

$$\lim_{h \to 0} E_\theta \biggl(\frac{p_{\theta+h}(X) - p_\theta(X)}{hp_\theta(X)} - s_\theta(X) \biggr)^2 = 0.$$

Definition 5.5.1 *If I and II hold, we call s_θ the score function, and $I(\theta)$ the* Fisher information.

Comparing the above with Definition 4.7.1, we see that they coincide if $\theta \mapsto p_\theta$ is differentiable and "regularity conditions" hold. Recall also that in Lemma 4.7.1 we assumed unspecified "regularity conditions". We now present a rigorous proof of the first part of this lemma.

Lemma 5.5.1 *Assume Conditions I and II. Then*

$$E_\theta s_\theta(X) = 0, \forall \ \theta.$$

Proof. Under P_θ, we only need to consider values x with $p_\theta(x) > 0$, that is, we may freely divide by p_θ, without worrying about dividing by zero.

Observe that

$$E_\theta \biggl(\frac{p_{\theta+h}(X) - p_\theta(X)}{p_\theta(X)} \biggr) = \int_A (p_{\theta+h} - p_\theta) d\nu = 0,$$

since densities integrate to 1, and both $p_{\theta+h}$ and p_θ vanish outside A. Thus,

$$\begin{aligned} |E_\theta s_\theta(X)|^2 &= \biggl| E_\theta \biggl(\frac{p_{\theta+h}(X) - p_\theta(X)}{hp_\theta(X)} - s_\theta(X) \biggr) \biggr|^2 \\ &\le E_\theta \biggl(\frac{p_{\theta+h}(X) - p_\theta(X)}{hp_\theta(X)} - s_\theta(X) \biggr)^2 \to 0. \end{aligned}$$

□

Note Thus $I(\theta) = \text{var}_\theta(s_\theta(X))$.

Remark As already noted, if $p_\theta(x)$ is differentiable for all x, we can take (under regularity conditions)

$$s_\theta(x) := \frac{d}{d\theta} \log p_\theta(x) = \frac{\dot{p}_\theta(x)}{p_\theta(x)},$$

where

$$\dot{p}_\theta(x) := \frac{d}{d\theta} p_\theta(x).$$

Remark Suppose $X_1, \ldots, X_n$ are i.i.d. with density p_θ, and $s_\theta = \dot{p}_\theta / p_\theta$ exists. The joint density is

$$\mathbf{p}_\theta(\mathbf{x}) = \prod_{i=1}^n p_\theta(x_i),$$

so that (under conditions I and II) the score function for n observations is

$$\mathbf{s}_\theta(\mathbf{x}) = \sum_{i=1}^n s_\theta(x_i).$$

The Fisher information for n observations is thus

$$\mathbf{I}(\theta) = \text{var}_\theta(\mathbf{s}_\theta(\mathbf{X})) = \sum_{i=1}^n \text{var}_\theta(s_\theta(X_i)) = nI(\theta).$$

Theorem 5.5.1 *(The Cramér-Rao lower bound) Suppose Conditions I and II are met, and that T is an unbiased estimator of $g(\theta)$ with finite variance. Then $g(\theta)$ has a derivative, $\dot{g}(\theta) := dg(\theta)/d\theta$, equal to*

$$\dot{g}(\theta) = \text{cov}(T, s_\theta(X)).$$

Moreover,

$$\text{var}_\theta(T) \geq \frac{\dot{g}^2(\theta)}{I(\theta)}, \ \forall \ \theta.$$

Proof. We first show differentiability of g. As T is unbiased, we have

$$\begin{aligned}
\frac{g(\theta+h) - g(\theta)}{h} &= \frac{E_{\theta+h}T(X) - E_\theta T(X)}{h} \\
&= \frac{1}{h} \int T(p_{\theta+h} - p_\theta) d\nu = E_\theta T(X) \frac{p_{\theta+h}(X) - p_\theta(X)}{hp_\theta(X)} \\
&= E_\theta T(X) \Bigg(\frac{p_{\theta+h}(X) - p_\theta(X)}{hp_\theta(X)} - s_\theta(X) \Bigg) + E_\theta T(X) s_\theta(X) \\
&= E_\theta \Bigg(T(X) - g_\theta \Bigg) \Bigg(\frac{p_{\theta+h}(X) - p_\theta(X)}{hp_\theta(X)} - s_\theta(X) \Bigg) \\
&+ \ E_\theta T(X) s_\theta(X) \\
&\to \ E_\theta T(X) s_\theta(X),
\end{aligned}$$

as $h \to 0$. This is becasue by the Cauchy-Schwarz inequality

$$\left| E_\theta \biggl(T(X) - g_\theta \biggr) \biggl(\frac{p_{\theta+h}(X) - p_\theta(X)}{h p_\theta(X)} - s_\theta(X) \biggr) \right|^2$$

$$\begin{array}{rl} \leq & \mathrm{var}_\theta(T) E_\theta \biggl(\frac{p_{\theta+h}(X) - p_\theta(X)}{h p_\theta(X)} - s_\theta(X) \biggr)^2 \\ \to & 0. \end{array}$$

Thus,

$$\dot{g}(\theta) = E_\theta T(X) s_\theta(X) = \mathrm{cov}_\theta(T, s_\theta(X)).$$

The last inequality holds because $E_\theta s_\theta(X) = 0$. By Cauchy-Schwarz,

$$\begin{array}{rcl} \dot{g}^2(\theta) & = & \biggl(\mathrm{cov}_\theta(T, s_\theta(X)) \biggr)^2 \\ & \leq & \mathrm{var}_\theta(T) \mathrm{var}_\theta(s_\theta(X)) = \mathrm{var}_\theta(T) I(\theta). \end{array}$$

□

Definition 5.5.2 *We call $\dot{g}^2(\theta)/I(\theta)$, $\theta \in \Theta$, the* Cramer Rao lower bound (CRLB) *(for estimating $g(\theta)$).*

Example 5.5.1 CRLB for exponential case
Let $X_1, \ldots, X_n$ be i.i.d. Exponential(θ), *$\theta > 0$. The density of a single observation is then*

$$p_\theta(x) = \theta \mathrm{e}^{-\theta x}, \ x > 0.$$

Let $g(\theta) := 1/\theta$, and $T := \bar{X}$. Then T is unbiased, and $\mathrm{var}_\theta(T) = 1/(n\theta^2)$. *We now compute the CRLB. With $g(\theta) = 1/\theta$, one has $\dot{g}(\theta) = -1/\theta^2$. Moreover,*

$$\log p_\theta(x) = \log \theta - \theta x,$$

so

$$s_\theta(x) = 1/\theta - x,$$

and hence

$$I(\theta) = \mathrm{var}_\theta(X) = \frac{1}{\theta^2}.$$

The CRLB for n observations is thus

$$\frac{\dot{g}^2(\theta)}{nI(\theta)} = \frac{1}{n\theta^2}.$$

In other words, T reaches the CRLB.

Example 5.5.2 CRLB for Poisson case *Suppose $X_1, \ldots, X_n$ are i.i.d.* Poisson(θ), *$\theta > 0$. Then*

$$\log p_\theta(x) = -\theta + x \log \theta - \log x!,$$

so

$$s_\theta(x) = -1 + \frac{x}{\theta},$$

and hence

$$I(\theta) = \mathrm{var}_\theta\left(\frac{X}{\theta}\right) = \frac{\mathrm{var}_\theta(X)}{\theta^2} = \frac{1}{\theta}.$$

One easily checks that $\bar{X}$ reaches the CRLB for estimating θ.

Let now $g(\theta) := \mathrm{e}^{-\theta}$. The UMVU estimator of $g(\theta)$ is

$$T := \left(1 - \frac{1}{n}\right)^{\sum_{i=1}^n X_i}.$$

To compute its variance, we first compute

$$\begin{aligned} E_\theta T^2 &= \sum_{k=0}^\infty \left(1 - \frac{1}{n}\right)^{2k} \frac{(n\theta)^k}{k!} \mathrm{e}^{-n\theta} \\ &= \mathrm{e}^{-n\theta} \sum_{k=0}^\infty \frac{1}{k!} \left(\frac{(n-1)^2\theta}{n}\right)^k \\ &= \mathrm{e}^{-n\theta} \exp\left[\frac{(n-1)^2\theta}{n}\right] = \exp\left[\frac{(1-2n)\theta}{n}\right]. \end{aligned}$$

Thus,

$$\begin{aligned} \mathrm{var}_\theta(T) &= E_\theta T^2 - [E_\theta T]^2 = E_\theta T^2 - \mathrm{e}^{-2\theta} \\ &= \mathrm{e}^{-2\theta}\left(\mathrm{e}^{\theta/n} - 1\right) \\ & \begin{cases} > & \theta \mathrm{e}^{-2\theta}/n \\ \approx & \theta \mathrm{e}^{-2\theta}/n \text{ for } n \text{ large} \end{cases}. \end{aligned}$$

As $\dot{g}(\theta) = -\mathrm{e}^{-\theta}$, the CRLB is

$$\frac{\dot{g}^2(\theta)}{nI(\theta)} = \frac{\theta \mathrm{e}^{-2\theta}}{n}.$$

We conclude that T does not reach the CRLB, but the gap is small for n large.

5.6 CRLB and exponential families

For the next result, we:

Recall Let X and Y be two real-valued random variables. The correlation between X and Y is

$$\rho(X,Y) := \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}.$$

We have

$$|\rho(X,Y)| = 1 \Leftrightarrow \exists \text{ constants } a, b : \ Y = aX + b \text{ (a.s.)}.$$

The next lemma shows that the CRLB can only be reached within exponential families, thus is only tight in a rather limited context.

Lemma 5.6.1 *Assume Conditions I and II, with $s_\theta = \dot{p}_\theta / p_\theta$. Suppose T is unbiased for $g(\theta)$, and that T reaches the Craméř Rao lower bound. Then $\{P_\theta : \ \theta \in \Theta\}$ forms a one-dimensional exponential family: there exist functions $c(\theta)$, $d(\theta)$, and $h(x)$ such that for all θ,*

$$p_\theta(x) = \exp[c(\theta)T(X) - d(\theta)]h(x), \ x \in \mathcal{X}.$$

Moreover, $c(\theta)$ and $d(\theta)$ are differentiable, say with derivatives $\dot{c}(\theta)$ and $\dot{d}(\theta)$ respectively. We furthermore have the equality

$$g(\theta) = \dot{d}(\theta)/\dot{c}(\theta), \ \forall \ \theta.$$

Proof. By Theorem 5.5, when T reaches the CRLB, we must have

$$\text{var}_\theta(T) = \frac{|\text{cov}_\theta(T, s_\theta(X))|^2}{\text{var}_\theta(s_\theta(X))},$$

i.e., then the correlation between T and $s_\theta(X)$ is ± 1. Thus, there exist constants $a(\theta)$ and $b(\theta)$ (depending on θ), such that

$$s_\theta(X) = a(\theta)T(X) - b(\theta). \tag{5.1}$$

But, as $s_\theta = \dot{p}_\theta/p_\theta = d \log p_\theta / d\theta$, we can take primitives:

$$\log p_\theta(x) = c(\theta)T(x) - d(\theta) + \tilde{h}(x),$$

where $\dot{c}(\theta) = a(\theta)$, $\dot{d}(\theta) = b(\theta)$ and $\tilde{h}(x)$ is constant in θ. Hence,

$$p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x),$$

with $h(x) = \exp[\tilde{h}(x)]$.

Moreover, the equation (5.1) tells us that

$$E_\theta s_\theta(X) = a(\theta)E_\theta T - b(\theta) = a(\theta)g(\theta) - b(\theta).$$

Because $E_\theta s_\theta(X) = 0$, this implies that $g(\theta) = b(\theta)/a(\theta)$. □

5.7 Higher-dimensional extensions

Expectations and covariance matrices of random vectors

Let $Z \in \mathbb{R}^k$ be a k-dimensional random vector. Then EZ is a k-dimensional vector, and

$$\Sigma := \text{Cov}(Z) := EZZ' - (EZ)(EZ')$$

is a $k \times k$ matrix containing all variances (on the diagonal) and covariances (off-diagonal). Note that Σ is positive semi-definite: for any vector $a \in \mathbb{R}^k$, we have

$$\text{var}(a'Z) = a'\Sigma a \geq 0.$$

Some matrix algebra

Let V be a symmetric matrix. If V is positive (semi-)definite, we write this as $V > 0$ ($V \geq 0$). One then has that $V = W^2$, where W is also positive (semi-)definite.

Auxiliary lemma. *Suppose $V > 0$. Then*

$$\max_{a \in \mathbb{R}^p} \frac{|a'c|^2}{a'Va} = c'V^{-1}c.$$

Proof. Write $V = W^2$, and $b := Wa$, $d := W^{-1}c$. Then $a'Va = b'b = \|b\|^2$ and $a'c = b'd$. By Cauchy-Schwarz

$$\max_{b \in \mathbb{R}^p} \frac{|b'd|^2}{\|b\|^2} = \|d\|^2 = d'd = c'V^{-1}c.$$

□

We will now present the CRLB in higher dimensions. To simplify the exposition, we will not carefully formulate the regularity conditions, that is, we assume derivatives to exist and that we can interchange differentiation and integration at suitable places.

Consider a parameter space $\Theta \subset \mathbb{R}^k$. Let

$$g : \Theta \to \mathbb{R},$$

be a given function. Denote the vector of partial derivatives as

$$\dot{g}(\theta) := \begin{pmatrix} \partial g(\theta)/\partial\theta_1 \\ \vdots \\ \partial g(\theta)/\partial\theta_k \end{pmatrix}.$$

The score vector is defined as

$$s_\theta(\cdot) := \begin{pmatrix} \partial \log p_\theta/\partial\theta_1 \\ \vdots \\ \partial \log p_\theta/\partial\theta_k \end{pmatrix}.$$

The Fisher information matrix is

$$I(\theta) = E_\theta s_\theta(X) s'_\theta(X) = \text{Cov}_\theta(s_\theta(X)).$$

Theorem 5.7.1 *Let T be an unbiased estimator of $g(\theta)$. Then, under regularity conditions,*

$$\mathrm{var}_\theta(T) \ge \dot{g}(\theta)' I(\theta)^{-1} \dot{g}(\theta).$$

Proof. As in the one-dimensional case, one can show that, for $j = 1, \ldots, k$,

$$\dot{g}_j(\theta) = \mathrm{cov}_\theta(T, s_{\theta,j}(X)).$$

Hence, for all $a \in \mathbb{R}^k$,

$$\begin{aligned} |a'\dot{g}(\theta)|^2 &= |\mathrm{cov}_\theta(T, a's_\theta(X))|^2 \\ &\le \mathrm{var}_\theta(T)\mathrm{var}_\theta(a's_\theta(X)) \\ &= \mathrm{var}_\theta(T) a' I(\theta) a. \end{aligned}$$

Combining this with the auxiliary lemma gives

$$\mathrm{var}_\theta(T) \ge \max_{a \in \mathbb{R}^k} \frac{|a'\dot{g}(\theta)|^2}{a'I(\theta)a} = \dot{g}'(\theta) I(\theta)^{-1} \dot{g}(\theta).$$

□

Corollary 5.7.1 *As a consequence, one obtains a lower bound for unbiased estimators of higher-dimensional parameters of interest. As example, let $g(\theta) := \theta = (\theta_1, \ldots, \theta_k)'$, and suppose that $T \in \mathbb{R}^k$ is an unbiased estimator of θ: $E_\theta T = \theta \ \forall \ \theta$. Then, for all $a \in \mathbb{R}^k$, $a'T$ is an unbiased estimator of $a'\theta$. Since $a'\theta$ has derivative a, the CRLB gives*

$$\mathrm{var}_\theta(a'T) \ge a' I(\theta)^{-1} a.$$

But

$$\mathrm{var}_\theta(a'T) = a' \mathrm{Cov}_\theta(T) a.$$

So for all a,

$$a' \mathrm{Cov}_\theta(T) a \ge a' I(\theta)^{-1} a,$$

in other words, $\mathrm{Cov}_\theta(T) \ge I(\theta)^{-1}$, that is, $\mathrm{Cov}_\theta(T) - I(\theta)^{-1}$ is positive semidefinite.

Chapter 6

Tests and confidence intervals

6.1 Intermezzo: quantile functions

Let F be a distribution function on $\mathbb{R}$. Then F is *cadlag* (continue à droite, limite à gauche). Define the quantile functions

$$q_{\sup}^F(u) := \sup\{x : \ F(x) \le u\},$$

and

$$q_{\inf}^F(u) := \inf\{x : \ F(x) \ge u\} := F^{-1}(u).$$

It holds that

$$F(q_{\inf}^F(u)) \ge u$$

and, for all $h > 0$,

$$F(q_{\sup}^F(u) - h) \le u.$$

Hence

$$F(q_{\sup}^F(u)-) := \lim_{h \downarrow 0} F(q_{\sup}^F(u) - h) \le u.$$

6.2 How to construct tests

Consider a model class

$$\mathcal{P} := \{P_\theta : \ \theta \in \Theta\}.$$

Moreover, consider a space Γ, and a map

$$g : \Theta \to \Gamma, \ g(\theta) := \gamma.$$

We think of γ as the parameter of interest.

Definition 6.2.1 *Let $\gamma_0 \in \Gamma$ and $\alpha \in [0, 1]$ be given. A (non-randomized) test at level α for the hypothesis*
$H_0 : \ \gamma = \gamma_0$
is a statistic $\phi(X, \gamma_0) \in \ \{0, 1\}$ such that $P_\theta(\phi(X, \gamma_0) = 1) \le \alpha$ for all $\theta \in \{\vartheta : g(\vartheta) = \gamma_0\}$

We often omit the dependence of ϕ on γ_0, i.e., we write $\phi(X) := \phi(X, \gamma_0)$.

Note Typically a test ϕ is based on a test statistic T, i.e. it is of the form

$$\phi(X) = \begin{cases} 1 & \text{if } T(X) > c \\ 0 & \text{else} \end{cases} .$$

The constant c is called the *critical value.*

To test $\quad H_{\gamma_0}: \ \gamma = \gamma_0$,
we look for a *pivot* (*Tür-Angel*). This is a function $Z(\mathbf{X}, \gamma)$ depending on the data $\mathbf{X}$ and on the parameter γ, such that for all $\theta \in \Theta$, the distribution

$$\mathbb{P}_\theta(Z(\mathbf{X}, g(\theta)) \le \cdot) =: G(\cdot)$$

does not depend on θ. We note that to find a pivot is unfortunately not always possible. However, if we *do* have a pivot $Z(\mathbf{X}, \gamma)$ with distribution G, we can compute its quantile functions

$$q_L := q_{\sup}^G \left(\frac{\alpha}{2}\right), \ q_R := q_{\inf}^G \left(1 - \frac{\alpha}{2}\right).$$

and the test

$$\phi(\mathbf{X}, \gamma_0) := \begin{cases} 1 & \text{if } Z(\mathbf{X}, \gamma_0) \notin [q_L, q_R] \\ 0 & \text{else} \end{cases} .$$

Then the test has level α for testing H_{γ_0}, with $\gamma_0 = g(\theta_0)$:

$$\mathbb{P}_{\theta_0}(\phi(\mathbf{X}, g(\theta_0)) = 1) = P_{\theta_0}(Z(\mathbf{X}, g(\theta_0)) > q_R) + \mathbb{P}_{\theta_0}(Z(\mathbf{X}), g(\theta_0)) < q_L)$$

$$= 1 - G(q_R) + G(q_L-) \le 1 - \left(1 - \frac{\alpha}{2}\right) + \frac{\alpha}{2} = \alpha.$$

Asymptotic pivot Let $Z_n(X_1, \ldots, X_n, \gamma)$ be some function of the data and the parameter of interest, defined for each sample size n. We call $Z_n(X_1, \ldots, X_n, \gamma)$ an *asymptotic* pivot if for all $\theta \in \Theta$,

$$\lim_{n \to \infty} \mathbb{P}_\theta(Z_n(X_1, \ldots, X_n, \gamma) \le \cdot) = G(\cdot),$$

at all continuity points of G, where the limit G does not depend on θ.

Example 6.2.1 The location model
As example, consider again the location model (Section 1.3). Let

$$\Theta := \{\theta = (\mu, F_0), \ \mu \in \mathbb{R}, \ F_0 \in \mathcal{F}_0\},$$

with $\mathcal{F}_0$ a subset of the collection of symmetric distributions (see (1.2)). Let $\hat{\mu}$ be an equivariant estimator, that is: the distribution of $\hat{\mu} - \mu$ does not depend on μ (see Chapter 9 for the formal definition of equivariance).

- *If $\mathcal{F}_0 := \{F_0\}$ is a single distribution (i.e., the distribution F_0 is known), we take $Z(\mathbf{X}, \mu) := \hat{\mu} - \mu$ as pivot. By the equivariance, this pivot has distribution G depending only on F_0.*

- *If $\mathcal{F}_0 := \{F_0(\cdot) = \Phi(\cdot/\sigma) : \ \sigma > 0\}$, we choose $\hat{\mu} := \bar{X}_n$ where $\bar{X}_n = \sum_{i=1}^n X_i/n$ is the sample mean. As pivot, we take*

$$Z(\mathbf{X}, \mu) := \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n},$$

where $S_n^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ is the sample variance. Then G is the Student distribution with $n-1$ degrees of freedom.

- *If $\mathcal{F}_0 := \{F_0$ symmetric and continuous at $x = 0\}$, we let the pivot be the sign test statistic:*

$$Z(\mathbf{X}, \mu) := \sum_{i=1}^n \mathrm{l}\{X_i \ge \mu\}.$$

Then G is the Binomial(n, p) *distribution, with parameter $p = 1/2$.*

- *Suppose now that $X_1, \ldots, X_n$ are the first n of an infinite sequence of i.i.d. random variables, and that*

$$\mathcal{F}_0 := \{F_0 : \ \int x dF_0(x) = 0, \ \int x^2 dF_0(x) < \infty\}.$$

Then

$$Z_n(X_1, \ldots, X_n, \mu) := \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n}$$

is an asymptotic pivot, with limiting distribution $G = \Phi$.

6.3 Equivalence confidence sets and tests

Definition 6.3.1 *A subset $I = I(\mathbf{X}) \subset \Gamma$, depending (only) on the data $\mathbf{X} = (X_1, \ldots, X_n)$, is called a* confidence set *(*Vertrauensbereich*) for γ, at level $1-\alpha$, if*

$$\mathbb{P}_\theta(\gamma \in I) \ge 1 - \alpha, \ \forall \ \theta \in \Theta.$$

A confidence interval *is of the form*

$$I := [\underline{\gamma}, \bar{\gamma}],$$

where the boundaries $\underline{\gamma} = \underline{\gamma}(\mathbf{X})$ and $\bar{\gamma} = \bar{\gamma}(\mathbf{X})$ depend (only) on the data $\mathbf{X}$.

Let for each $\gamma_0 \in \mathbb{R}$, $\phi(\mathbf{X}, \gamma_0) \in \{0, 1\}$ be a test at level α for the hypothesis $H_{\gamma_0} : \ \gamma = \gamma_0$.
Thus, we reject H_{γ_0} if and only if $\phi(\mathbf{X}, \gamma_0) = 1$, and

$$\mathbb{P}_{\theta : \gamma = \gamma_0}(\phi(\mathbf{X}, \gamma_0) = 1) \le \alpha.$$

Then

$$I(\mathbf{X}) := \{\gamma : \phi(\mathbf{X}, \gamma) = \mathbf{0}\}$$

is a $(1-\alpha)$-confidence set for γ.

Conversely, if $I(\mathbf{X})$ is a $(1-\alpha)$-confidence set for γ, then, for all γ_0, the test $\phi(\mathbf{X}, \gamma_0)$ defined as

$$\phi(\mathbf{X}, \gamma_0) = \begin{cases} 1 & \text{if } \gamma_0 \notin I(\mathbf{X}) \\ 0 & \text{else} \end{cases}$$

is a test at level α of H_{γ_0}.

6.4 Comparison of confidence intervals and tests

When comparing confidence intervals, the aim is usually to take the one with smallest length on average (keeping the level at $1-\alpha$). In the case of tests, we look for the one with maximal power. Recall that the power is of a test $\phi(X, \gamma_0)$ at a value θ with $g(\theta) \neq \gamma_0$ is $P_\theta(\phi(\mathbf{X}, \gamma_0) = 1)$.

6.5 An illustration: the two-sample problem

Consider the following data, concerning weight gain/loss. The control group x had their usual diet, and the treatment group y obtained a special diet, designed for preventing weight gain. The study was carried out to test whether the diet works.

control	treatment		
group	group	rank(x)	rank(y)
x	y		
5	6	*7*	*8*
0	-5	*3*	*2*
16	-6	*10*	*1*
2	1	*5*	*4*
9	4	*9*	*6*
—— +	—— +		
32	0		

Table 2

Let n (m) be the sample size of the control group x (treatment group y). The mean in group x (y) is denoted by $\bar{x}$ $(\bar{y})$. The sums of squares are $SS_x := \sum_{i=1}^n (x_i - \bar{x})^2$ and $SS_y := \sum_{j=1}^m (y_j - \bar{y})^2$. So in this study, one has $n = m = 5$ and the values $\bar{x} = 6.4$, $\bar{y} = 0$, $SS_x = 161.2$ and $SS_y = 114$. The ranks, rank(x) and rank(y), are the rank-numbers when putting all $n + m$ data together (e.g., $y_3 = -6$ is the smallest observation and hence rank$(y_3) = 1$).

We assume that the data are realizations of two independent samples, say $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} = (Y_1, \ldots, Y_m)$, where $X_1, \ldots, X_n$ are i.i.d. with distribution function F_X, and $Y_1, \ldots, Y_m$ are i.i.d. with distribution function F_Y. The distribution functions F_X and F_Y may be in whole or in part unknown. The testing problem is:

$H_0: \ F_X = F_Y$
against a one- or two-sided alternative.

6.5.1 Student's test

The classical two-sample student test is based on the assumption that the data come from a normal distribution. Moreover, it is assumed that the variance of F_X and F_Y are equal. Thus,

$$(F_X, F_Y) \in$$

$$\left\{ F_X = \Phi\left(\frac{\cdot - \mu}{\sigma}\right), \ F_Y = \Phi\left(\frac{\cdot - (\mu + \gamma)}{\sigma}\right) : \ \mu \in \mathbb{R}, \ \sigma > 0, \ \gamma \in \Gamma \right\}.$$

Here, $\Gamma \supset \{0\}$ is the range of shifts in mean one considers, e.g. $\Gamma = \mathbb{R}$ for two-sided situations, and $\Gamma = (-\infty, 0]$ for a one-sided situation. The testing problem reduces to
$H_0: \ \gamma = 0$.

We now look for a pivot $Z(\mathbf{X}, \mathbf{Y}, \gamma)$. Define the sample means

$$\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i, \ \bar{Y} := \frac{1}{m} \sum_{j=1}^m Y_j,$$

and the pooled sample variance

$$S^2 := \frac{1}{m+n-2} \left\{ \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2 \right\}.$$

Note that $\bar{X}$ has expectation μ and variance σ^2/n, and $\bar{Y}$ has expectation $\mu + \gamma$ and variance σ^2/m. So $\bar{Y} - \bar{X}$ has expectation γ and variance

$$\frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \sigma^2 \left(\frac{n+m}{nm}\right).$$

The normality assumption implies that

$$\bar{Y} - \bar{X} \text{ is } \mathcal{N}\left(\gamma, \sigma^2 \left(\frac{n+m}{nm}\right)\right) - \text{distributed.}$$

Hence

$$\sqrt{\frac{nm}{n+m}} \left(\frac{\bar{Y} - \bar{X} - \gamma}{\sigma}\right) \text{ is } \mathcal{N}(0,1) - \text{distributed.}$$

To arrive at a pivot, we now plug in the estimate S for the unknown σ:

$$Z(\mathbf{X}, \mathbf{Y}, \gamma) := \sqrt{\frac{nm}{n+m}} \left(\frac{\bar{Y} - \bar{X} - \gamma}{S}\right).$$

Indeed, $Z(\mathbf{X}, \mathbf{Y}, \gamma)$ has a distribution G which does not depend on unknown parameters. The distribution G is Student$(n+m-2)$ (the Student-distribution with $n+m-2$ degrees of freedom). As test statistic for $H_0: \ \gamma = 0$, we therefore take

$$T = T^{\rm Student} := Z(\mathbf{X}, \mathbf{Y}, 0).$$

The one-sided test at level α, for $H_0: \ \gamma = 0$ against $H_1: \ \gamma < 0$, is

$$\phi(\mathbf{X}, \mathbf{Y}) := \begin{cases} 1 & \text{if } T < -t_{n+m-2}(1-\alpha) \\ 0 & \text{if } T \ge -t_{n+m-2}(1-\alpha) \end{cases},$$

where, for $\nu > 0$, $t_\nu(1-\alpha) = -t_\nu(\alpha)$ is the $(1-\alpha)$-quantile of the Student(ν)-distribution.

Let us apply this test to the data given in Table 2. We take $\alpha = 0.05$. The observed values are $\bar x = 6.4$, $\bar y = 0$ and $s^2 = 34.4$. The test statistic takes the value -1.725 which is bigger than the 5% quantile $t_8(0.05) = -1.9$. Hence, we cannot reject H_0. The p-value of the observed value of T is

$$p\text{−value} := \mathbb{P}_{\gamma=0}(T < -1.725) = 0.06.$$

So the p-value is in this case only a little larger than the level $\alpha = 0.05$.

6.5.2 Wilcoxon's test

In this subsection, we suppose that F_X and F_Y are continuous, but otherwise unknown. The model class for both F_X and F_Y is thus

$$\mathcal{F} := \{\text{all continuous distributions}\}.$$

The continuity assumption ensures that all observations are distinct, that is, there are no ties. We can then put them in strictly increasing order. Let $N = n + m$ and $Z_1, \ldots, Z_N$ be the pooled sample

$$Z_i := X_i, \ i = 1, \ldots, n, \ Z_{n+j} := Y_j, \ j = 1, \ldots, m.$$

Define

$$R_i := \text{rank}(Z_i), \ i = 1, \ldots, N.$$

and let

$$Z_{(1)} < \cdots < Z_{(N)}$$

be the order statistics of the pooled sample (so that $Z_i = Z_{(R_i)}$ $(i = 1, \ldots, n)$). The Wilcoxon test statistic is

$$T = T^{\rm Wilcoxon} := \sum_{i=1}^n R_i.$$

One may check that this test statistic T can alternatively be written as

$$T = \#\{Y_j < X_i\} + \frac{n(n+1)}{2}.$$

For example, for the data in Table 2, the observed value of T is 34, and

$$\#\{y_j < x_i\} = 19, \ \frac{n(n+1)}{2} = 15.$$

Large values of T mean that the X_i are generally larger than the Y_j, and hence indicate evidence against H_0.

To check whether or not the observed value of the test statistic is compatible with the null-hypothesis, we need to know its null-distribution, that is, the distribution under H_0. Under $H_0 : F_X = F_Y$, the vector of ranks $(R_1, \ldots, R_n)$ has the same distribution as n random draws without replacement from the numbers $\{1, \ldots, N\}$. That is, if we let

$$\mathbf{r} := (r_1, \ldots, r_n, r_{n+1}, \ldots, r_N)$$

denote a permutation of $\{1, \ldots, N\}$, then

$$\mathbb{P}_{H_0}\biggl((R_1, \ldots, R_n, R_{n+1}, \ldots R_N) = \mathbf{r} \biggr) = \frac{1}{N!},$$

(see Theorem 6.5.1 below), and hence

$$\mathbb{P}_{H_0}(T = t) = \frac{\#\{\mathbf{r} : \ \sum_{i=1}^n r_i = t\}}{N!}.$$

This can also be written as

$$\mathbb{P}_{H_0}(T = t) = \frac{1}{\binom{N}{n}} \#\{r_1 < \cdots < r_n < r_{n+1} < \cdots < r_N : \ \sum_{i=1}^n r_i = t\}.$$

So clearly, the null-distribution of T does not depend on F_X or F_Y. It does however depend on the sample sizes n and m. It is tabulated for n and m small or moderately large. For large n and m, a normal approximation of the null-distribution can be used.

Theorem 6.5.1 formally derives the null-distribution of the test, and actually proves that the order statistics and the ranks are independent. The latter result will be of interest in Example 4.1.4.

For two random variables X and Y, use the notation

$$X \overset{\mathcal{D}}{=} Y$$

when X and Y have the same distribution.

Theorem 6.5.1 *Let $Z_1, \ldots, Z_N$ be i.i.d. with continuous distribution F on $\mathbb{R}$. Then $(Z_{(1)}, \ldots, Z_{(N)})$ and $\mathbf{R} := (R_1, \ldots, R_N)$ are independent, and for all permutations $\mathbf{r} := (r_1, \ldots, r_N)$,*

$$\mathbb{P}(\mathbf{R} = \mathbf{r}) = \frac{1}{N!}.$$

Proof. Let $Z_{Q_i} := Z_{(i)}$, and $\mathbf{Q} := (Q_1, \ldots, Q_N)$. Then

$$\mathbf{R} = \mathbf{r} \ \Leftrightarrow \ \mathbf{Q} = \mathbf{r}^{-1} := \mathbf{q},$$

where $\mathbf{r}^{-1}$ is the inverse permutation of $\mathbf{r}$.[1] For all permutations $\mathbf{q}$ and all measurable maps f,

$$f(Z_1, \ldots, Z_N) \overset{\mathcal{D}}{=} f(Z_{q_1}, \ldots, Z_{q_N}).$$

Therefore, for all measurable sets $A \subset \mathbb{R}^N$, and all permutations $\mathbf{q}$,

$$\mathbb{P}\biggl((Z_1, \ldots, Z_N) \in A, \ Z_1 < \ldots < Z_N \biggr)$$

$$= \mathbb{P}\biggl((Z_{q_1} \ldots, Z_{q_N}) \in A, \ Z_{q_1} < \ldots < Z_{q_N} \biggr).$$

Because there are $N!$ permutations, we see that for any $\mathbf{q}$,

$$\begin{aligned} \mathbb{P}\biggl((Z_{(1)}, \ldots, Z_{(n)}) \in A \biggr) &= N! \mathbb{P}\biggl((Z_{q_1} \ldots, Z_{q_N}) \in A, \ Z_{q_1} < \ldots < Z_{q_N} \biggr) \\ &= N! \mathbb{P}\biggl((Z_{(1)}, \ldots, Z_{(N)}) \in A, \ \mathbf{R} = \mathbf{r} \biggr), \end{aligned}$$

where $\mathbf{r} = \mathbf{q}^{-1}$. Thus we have shown that for all measurable A, and for all $\mathbf{r}$,

$$\mathbb{P}\biggl((Z_{(1)}, \ldots, Z_{(N)}) \in A, \ \mathbf{R} = \mathbf{r} \biggr) = \frac{1}{N!} \mathbb{P}\biggl((Z_{(1)}, \ldots, Z_{(n)}) \in A \biggr). \tag{6.1}$$

Take $A = \mathbb{R}^N$ to find that (6.1) implies

$$\mathbb{P}\biggl(\mathbf{R} = \mathbf{r} \biggr) = \frac{1}{N!}.$$

Plug this back into (6.1) to see that we have the product structure

$$\mathbb{P}\biggl((Z_{(1)}, \ldots, Z_{(N)}) \in A, \ \mathbf{R} = \mathbf{r} \biggr) = \mathbb{P}\biggl((Z_{(1)}, \ldots, Z_{(n)}) \in A \biggr) \mathbb{P}\biggl(\mathbf{R} = \mathbf{r} \biggr),$$

which holds for all measurable A. In other words, $(Z_{(1)}, \ldots, Z_{(N)})$ and $\mathbf{R}$ are independent. □

[1] Here is an example, with $N = 3$:

$$\begin{aligned} (z_1, z_2, z_3) &= (\ 5\ ,\ 6\ ,\ 4\ ) \\ (r_1, r_2, r_3) &= (\ 2\ ,\ 3\ ,\ 1\ ) \\ (q_1, q_2, q_3) &= (\ 3\ ,\ 1\ ,\ 2\ ) \end{aligned}$$

6.5.3 Comparison of Student's test and Wilcoxon's test

Because Wilcoxon's test is ony based on the ranks, and does not rely on the assumption of normality, it lies at hand that, when the data are in fact normally distributed, Wilcoxon's test will have less power than Student's test. The loss of power is however small. Let us formulate this more precisely, in terms of the relative efficiency of the two tests. Let the significance α be fixed, and let β be the required power. Let n and m be equal, $N = 2n$ be the total sample size, and N^{Student} (N^{Wilcoxon}) be the number of observations needed to reach power β using Student's (Wilcoxon's) test. Consider shift alternatives, i.e. $F_Y(\cdot) = F_X(\cdot - \gamma)$, (with, in our example, $\gamma < 0$). One can show that $N^{\text{Student}}/N^{\text{Wilcoxon}}$ is approximately .95 when the normal model is correct. For a large class of distributions, the ratio $N^{\text{Student}}/N^{\text{Wilcoxon}}$ ranges from .85 to ∞, that is, when using Wilcoxon one generally has very limited loss of efficiency as compared to Student, and one may in fact have a substantial gain of efficiency.

Chapter 7

The Neyman Pearson Lemma and UMP tests

Let $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ be a family of probability measures. Let $\Theta_0 \subset \Theta$, $\Theta_1 \subset \Theta$, and $\Theta_0 \cap \Theta_1 = \emptyset$. Based on observations $X \in \mathcal{X}$, with distribution $P \in \mathcal{P}$, we consider the general testing problem for
$H_0 : \ \theta \in \Theta_0$,
against
$H_1 : \ \theta \in \Theta_1$.

A (possibly randomized) test is some function $\phi : \mathcal{X} \to [0, 1]$. Fix some $\alpha \in [0, 1]$. We say that ϕ is a test at level α if

$$\sup_{\theta \in \Theta_0} E_\theta \phi(X) \le \alpha.$$

Definition 7.0.1 *A test ϕ is called* Uniformly Most Powerful *(*UMP*, (German: gleichmässig mächtigst)) if*
- *ϕ has level α,*
- *for all tests ϕ' with level α, it holds that $E_\theta \phi'(X) \le E_\theta \phi(X) \ \forall \ \theta \in \Theta_1$.*

7.1 The Neyman Pearson Lemma

We consider testing
$H_0 : \ \theta = \theta_0$
against the alternative
$H_1 : \ \theta = \theta_1$.

Define the risk $R(\theta, \phi)$ of a test ϕ as the probability of error of first and second kind:

$$R(\theta, \phi) := \begin{cases} E_\theta \phi(X), & \theta = \theta_0 \\ 1 - E_\theta \phi(X), & \theta = \theta_1 \end{cases} .$$

We let p_0 (p_1) be the density of P_{θ_0} (P_{θ_1}) with respect to some dominating measure ν (for example $\nu = P_{\theta_0} + P_{\theta_1}$). A Neyman Pearson test is

$$\phi_{\rm NP} := \begin{cases} 1 & \text{if } p_1/p_0 > c \\ q & \text{if } p_1/p_0 = c \\ 0 & \text{if } p_1/p_0 < c \end{cases} .$$

Here $0 \le q \le 1$, and $0 \le c < \infty$ are given constants.

Lemma 7.1.1 Neyman Pearson Lemma *Let ϕ be some test. We have*

$$R(\theta_1, \phi_{\rm NP}) - R(\theta_1, \phi) \le c[R(\theta_0, \phi) - R(\theta_0, \phi_{\rm NP})].$$

Proof.

$$\begin{aligned} & R(\theta_1, \phi_{\rm NP}) - R(\theta_1, \phi) \\ = & \int (\phi - \phi_{\rm NP}) p_1 \\ = & \int_{p_1/p_0 > c} (\phi - \phi_{\rm NP}) p_1 + \int_{p_1/p_0 = c} (\phi - \phi_{\rm NP}) p_1 + \int_{p_1/p_0 < c} (\phi - \phi_{\rm NP}) p_1 \\ \le & c \int_{p_1/p_0 > c} (\phi - \phi_{\rm NP}) p_0 + c \int_{p_1/p_0 = c} (\phi - \phi_{\rm NP}) p_0 + c \int_{p_1/p_0 < c} (\phi - \phi_{\rm NP}) p_0 \\ = & c[R(\theta_0, \phi) - R(\theta_0, \phi_{\rm NP})]. \end{aligned}$$

□

7.2 Uniformly most powerful tests

7.2.1 An example

Let $X_1, \ldots, X_n$ be i.i.d. copies of a Bernoulli random variable $X \in \{0, 1\}$ with success parameter $\theta \in (0, 1)$:

$$P_\theta(X = 1) = 1 - P_\theta(X = 0) = \theta.$$

We consider three testing problems. The chosen level in all three problems is $\alpha = 0.05$.

Problem 1

We want to test, at level α, the hypothesis

$H_0 : \ \theta = \frac{1}{2} =: \theta_0,$

against the alternative

$H_1 : \ \theta = \frac{1}{4} =: \theta_1.$

Let $T := \sum_{i=1}^n X_i$ be the number of successes (T is a sufficient statistic), and consider the randomized test

$$\phi(T) := \begin{cases} 1 & \text{if } T < t_0 \\ q & \text{if } T = t_0 \\ 0 & \text{if } T > t_0 \end{cases} ,$$

where $q \in (0,1)$, and where t_0 is the critical value of the test. The constants q and $t_0 \in \{0, \ldots, n\}$ are chosen in such a way that the probability of rejecting H_0 when it is in fact true, is equal to α:

$$P_{\theta_0}(H_0 \text{ rejected}) = P_{\theta_0}(T \le t_0 - 1) + q P_{\theta_0}(T = t_0) := \alpha.$$

Thus, we take t_0 in such a way that

$$P_{\theta_0}(T \le t_0 - 1) \le \alpha, \; P_{\theta_0}(T \le t_0) > \alpha,$$

(i.e., $t_0 - 1 = q_{\text{inf}}^G(\alpha)$ with q_{inf}^G the quantile function defined in Section 6.1 and G the distribution function of T) and

$$q = \frac{\alpha - P_{\theta_0}(T \le t_0 - 1)}{P_{\theta_0}(T = t_0)}.$$

Because $\phi = \phi_{\text{NP}}$ is the Neyman Pearson test, it is the most powerful test (at level α) (see the Neyman Pearson Lemma in Section 7.1). The power of the test is $\beta(\theta_1)$, where

$$\beta(\theta) := E_\theta \phi(T).$$

`Numerical Example`

Let $n = 7$. Then

$$\begin{aligned} P_{\theta_0}(T = 0) &= \left(\frac{1}{2}\right)^7 = 0.0078, \\ P_{\theta_0}(T = 1) &= \binom{7}{1}\left(\frac{1}{2}\right)^7 = 0.0546, \\ P_{\theta_0}(T \le 1) &= 0.0624 > \alpha, \end{aligned}$$

so we choose $t_0 = 1$. Moreover

$$q = \frac{0.05 - 0.0078}{0.0546} = \frac{422}{546}.$$

The power is now

$$\begin{aligned} \beta(\theta_1) &= P_{\theta_1}(T = 0) + q P_{\theta_1}(T = 1) \\ &= \left(\frac{3}{4}\right)^7 + \frac{422}{546}\binom{7}{1}\left(\frac{3}{4}\right)^6 \left(1/4\right) \\ &= 0.1335 + \frac{422}{546} 0.3114. \end{aligned}$$

Problem 2

Consider now testing
$H_0 :\ \theta_0 = \frac{1}{2}$,
against
$H_1 :\ \theta < \frac{1}{2}$.

In Problem 1, the construction of the test ϕ is independent of the value $\theta_1 < \theta_0$. So ϕ is most powerful for all $\theta_1 < \theta_0$. We say that ϕ is *uniformly most powerful* for the alternative $H_1 :\ \theta < \theta_0$.

Problem 3

We now want to test
$H_0 : \theta \geq \frac{1}{2}$,
against the alternative
$H_1 : \theta < \frac{1}{2}$.

Recall the function

$$\beta(\theta) := E_\theta \phi(T).$$

The level of ϕ is defined as

$$\sup_{\theta \geq 1/2} \beta(\theta).$$

We have

$$\begin{aligned} \beta(\theta) &= P_\theta(T \leq t_0 - 1) + q P_\theta(T = t_0) \\ &= (1-q) P_\theta(T \leq t_0 - 1) + q P_\theta(T \leq t_0). \end{aligned}$$

Observe that if $\theta_1 < \theta_0$, small values of T are more likely under P_{θ_1} than under P_{θ_0}:

$$P_{\theta_1}(T \leq t) > P_{\theta_0}(T \leq t),\ \forall\ t \in \{0, 1, \ldots, n\}.$$

Thus, $\beta(\theta)$ is a decreasing function of θ. It follows that the level of ϕ is

$$\sup_{\theta \geq \frac{1}{2}} \beta(\theta) = \beta(\tfrac{1}{2}) = \alpha.$$

Hence, ϕ is uniformly most powerful for $H_0 :\ \theta \geq \frac{1}{2}$ against $H_1 :\ \theta < \frac{1}{2}$.

7.3 UMP tests and exponential families

We now study the situation where Θ is an interval in $\mathbb{R}$, and the testing problem is
$H_0 :\ \theta \leq \theta_0$,
against
$H_1 :\ \theta > \theta_0$.

We suppose that $\mathcal{P}$ is dominated by a σ-finite measure ν.

Theorem 7.3.1 *Suppose that $\mathcal{P}$ is a one-dimensional exponential family*

$$p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x).$$

Assume moreover that $c(\theta)$ is a strictly increasing function of θ. Then a UMP test ϕ is

$$\phi(T(x)) := \begin{cases} 1 & \text{if } T(x) > t_0 \\ q & \text{if } T(x) = t_0 \\ 0 & \text{if } T(x) < t_0 \end{cases},$$

where q and t_0 are chosen in such a way that $E_{\theta_0}\phi(T) = \alpha$.

Proof. The Neyman Pearson test for $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ is

$$\phi_{\rm NP}(x) = \begin{cases} 1 & \text{if } p_{\theta_1}(x)/p_{\theta_0}(x) > c_0 \\ q_0 & \text{if } p_{\theta_1}(x)/p_{\theta_0}(x) = c_0 \\ 0 & \text{if } p_{\theta_1}(x)/p_{\theta_0}(x) < c_0 \end{cases},$$

where q_0 and c_0 are chosen in such a way that $E_{\theta_0}\phi_{\rm NP}(X) = \alpha$. We have

$$\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} = \exp\biggl[(c(\theta_1) - c(\theta_0))T(X) - (d(\theta_1) - d(\theta_0))\biggr].$$

Hence

$$\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} \begin{array}{c} > \\ = \\ < \end{array} c \Leftrightarrow T(x) \begin{array}{c} > \\ = \\ < \end{array} t ,$$

where t is some constant (depending on c, θ_0 and θ_1). Therefore, $\phi = \phi_{\rm NP}$. It follows that ϕ is most powerful for $H_0 : \ \theta = \theta_0$ against $H_1 : \ \theta = \theta_1$. Because ϕ does not depend on θ_1, it is therefore UMP for $H_0 : \ \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

We will now prove that $\beta(\theta) := E_\theta\phi(T)$ is increasing in θ. Let

$$\bar{p}_\theta(t) = \exp[c(\theta)t - d(\theta)]$$

be the density of T with respect to dominating measure $\bar{\nu}$. For $\vartheta > \theta$

$$\frac{\bar{p}_\vartheta(t)}{\bar{p}_\theta(t)} = \exp\biggl[(c(\vartheta) - c(\theta))t - (d(\vartheta) - d(\theta))\biggr],$$

which is increasing in t. Moreover, we have

$$\int \bar{p}_\vartheta d\bar{\nu} = \int \bar{p}_\theta d\bar{\nu} = 1.$$

Therefore, there must be a point s_0 where the two densities cross:

$$\begin{cases} \frac{\bar{p}_\vartheta(t)}{\bar{p}_\theta(t)} \le 1 & \text{for } t \le s_0 \\ \frac{\bar{p}_\vartheta(t)}{\bar{p}_\theta(t)} \ge 1 & \text{for } t \ge s_0 \end{cases}.$$

But then

$$\begin{aligned} \beta(\vartheta) - \beta(\theta) &= \int \phi(t)[\bar{p}_\vartheta(t) - \bar{p}_\theta(t)]d\bar{\nu}(t) \\ &= \int_{t \le s_0} \phi(t)[\bar{p}_\vartheta(t) - \bar{p}_\theta(t)]d\bar{\nu}(t) + \int_{t \ge s_0} \phi(t)[\bar{p}_\vartheta(t) - \bar{p}_\theta(t)]d\bar{\nu}(t) \\ &\ge \phi(s_0) \int [\bar{p}_\vartheta(t) - \bar{p}_\theta(t)]d\bar{\nu}(t) = 0. \end{aligned}$$

So indeed $\beta(\theta)$ is increasing in θ.

But then

$$\sup_{\theta \le \theta_0} \beta(\theta) = \beta(\theta_0) = \alpha.$$

Hence, ϕ has level α. Because any other test ϕ' with level α must have $E_{\theta_0}\phi'(X) \le \alpha$, we conclude that ϕ is UMP. □

Example 7.3.1 Test for the variance of the normal distribution
Let $X_1, \ldots, X_n$ be an i.i.d. sample from the $\mathcal{N}(\mu_0, \sigma^2)$-distribution, with μ_0 known, and $\sigma^2 > 0$ unknown. We want to test
$H_0 : \sigma^2 \le \sigma_0^2$,
against
$H_1 : \sigma^2 > \sigma_0^2$.

The density of the sample is

$$\mathbf{p}_{\sigma^2}(x_1, \ldots, x_n) = \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu_0)^2 - \frac{n}{2}\log(2\pi\sigma^2) \right].$$

Thus, we may take

$$c(\sigma^2) = -\frac{1}{2\sigma^2},$$

and

$$T(\mathbf{X}) = \sum_{i=1}^n (X_i - \mu_0)^2.$$

The function $c(\sigma^2)$ is strictly increasing in σ^2. So we let ϕ be the test which rejects H_0 for large values of $T(\mathbf{X})$. Note that under H_0, the statistic $T(\mathbf{X})/\sigma_0^2$ has a χ^2-distribution with n degrees of freedom, the χ_n^2-distribution (see Section 12.2 for a definition). So we can find the critical value from the quantile of the χ_n^2-distribution.

7.4 One- and two-sided tests: an example with the Bernoulli distribution

Let $X_1, \ldots, X_n$ be an i.i.d. sample from the Bernoulli(θ)-distribution, $0 < \theta < 1$. Then

$$\mathbf{p}_\theta(x_1, \ldots, x_n) = \exp\left[\log\left(\frac{\theta}{1-\theta} \right) \sum_{i=1}^n x_i + n\log(1-\theta) \right].$$

We can take

$$c(\theta) = \log\biggl(\frac{\theta}{1-\theta}\biggr),$$

which is strictly increasing in θ. Then $T(\mathbf{X}) = \sum_{i=1}^n X_i$.

Right-sided alternative
$H_0 : \ \theta \le \theta_0$,
against
$H_1 : \ \theta > \theta_0$.

The UMP test is

$$\phi_R(T) := \begin{cases} 1 & T > t_R \\ q_R & T = t_R \\ 0 & T < t_R \end{cases} .$$

The function $\beta_R(\theta) := E_\theta \phi_R(T)$ is strictly increasing in θ.

Left-sided alternative
$H_0 : \ \theta \ge \theta_0$,
against $\quad H_1 : \ \theta < \theta_0$.

The UMP test is

$$\phi_L(T) := \begin{cases} 1 & T < t_L \\ q_L & T = t_L \\ 0 & T > t_L \end{cases} .$$

The function $\beta_L(\theta) := E_\theta \phi_L(T)$ is strictly decreasing in θ.

Two-sided alternative
$H_0 : \ \theta = \theta_0$,
against
$H_1 : \ \theta \ne \theta_0$.

The test ϕ_R is most powerful for $\theta > \theta_0$, whereas ϕ_L is most powerful for $\theta < \theta_0$. Hence, a UMP test does not exist for the two-sided alternative.

7.5 Unbiased tests

Consider again the general case: $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ is a family of probability measures, the spaces Θ_0, and Θ_1 are disjoint subspaces of Θ, and the testing problem is
$H_0 : \ \theta \in \Theta_0$,
against
$H_1 : \ \theta \in \Theta_1$.

The significance level is α (< 1).

As we have seen in Section 7.4, uniformly most powerful tests do not always exist. We therefore restrict attention to a smaller class of tests, and look for uniformly most powerful tests in the smaller class.

Definition 7.5.1 *A test ϕ is called* unbiased *(German* unverfälscht*) if for all $\theta \in \Theta_0$ and all $\vartheta \in \Theta_1$,*

$$E_\theta \phi(X) \le E_\vartheta \phi(X).$$

Definition 7.5.2 *A test ϕ is called* Uniformly Most Powerful Unbiased *(*UMPU*) if*
- *ϕ has level α,*
- *ϕ is unbiased,*
- *for all unbiased tests ϕ' with level α, one has $E_\theta \phi'(X) \le E_\theta \phi(X)$ $\forall$ $\theta \in \Theta_1$.*

We return to the special case where $\Theta \subset \mathbb{R}$ is an interval. We consider testing
$H_0 : \ \theta = \theta_0$,
against
$H_1 : \ \theta \ne \theta_0$.

The following theorem presents the UMPU test. We omit the proof (see e.g. Lehmann (1986)).

Theorem 7.5.1 *Suppose $\mathcal{P}$ is a one-dimensional exponential family:*

$$\frac{dP_\theta}{d\nu}(x) := p_\theta(x) = \exp[c(\theta)T(x) - d(\theta)]h(x),$$

with $c(\theta)$ strictly increasing in θ. Then a UMPU test is

$$\phi(T(x)) := \begin{cases} 1 & \textit{if } T(x) < t_L \textit{ or } T(x) > t_R \\ q_L & \textit{if } T(x) = t_L \\ q_R & \textit{if } T(x) = t_R \\ 0 & \textit{if } t_L < T(x) < t_R \end{cases},$$

where the constants t_R , t_L, q_R and q_L are chosen in such a way that

$$E_{\theta_0}\phi(X) = \alpha, \ \frac{d}{d\theta} E_\theta \phi(X) \Big|_{\theta=\theta_0} = 0.$$

Note Let ϕ_R a right-sided test as defined Theorem 7.3.1 with level at most α and ϕ_L be the similarly defined left-sided test. Then $\beta_R(\theta) = E_\theta \phi_R(T)$ is strictly increasing, and $\beta_L(\theta) = E_\theta \phi_L(T)$ is strictly decreasing. The two-sided test ϕ of Theorem 7.5.1 is a superposition of two one-sided tests. Writing

$$\beta(\theta) = E_\theta \phi(T),$$

the one-sided tests are constructed in such a way that

$$\beta(\theta) = \beta_R(\theta) + \beta_L(\theta).$$

Moreover, $\beta(\theta)$ should be minimal at $\theta = \theta_0$, whence the requirement that its derivative at θ_0 should vanish. Let us see what this derivative looks like. With the notation used in the proof of Theorem 7.3.1, for a test $\tilde{\phi}$ depending only on the sufficient statistic T,

$$E_\theta \tilde{\phi}(T) = \int \tilde{\phi}(t) \exp[c(\theta)t - d(\theta)] d\bar{\nu}(t).$$

Hence, assuming we can take the differentiation inside the integral,

$$\begin{aligned}\frac{d}{d\theta}E_\theta\tilde{\phi}(T) &= \int \tilde{\phi}(t)\exp[c(\theta)t - d(\theta)](\dot{c}(\theta)t - \dot{d}(\theta))d\bar{\nu}(t) \\ &= \dot{c}(\theta)\mathrm{cov}_\theta(\tilde{\phi}(T), T).\end{aligned}$$

The UMPU test sets this to zero. We leave the interpretation to the reader...

Example 7.5.1 Two-sided test for the mean of the normal distribution

Let $X_1, \ldots, X_n$ be an i.i.d. sample from the $\mathcal{N}(\mu, \sigma_0^2)$-distribution, with $\mu \in \mathbb{R}$ unknown, and with σ_0^2 known. We consider testing
$H_0: \ \mu = \mu_0$,
against
$H_1: \ \mu \neq \mu_0$.

A sufficient statistic is $T := \sum_{i=1}^n X_i$. We have, for $t_L < t_R$,

$$\begin{aligned}E_\mu\phi(T) &= \mathbb{P}_\mu(T > t_R) + \mathbb{P}_\mu(T < t_L) \\ &= \mathbb{P}_\mu\biggl(\frac{T - n\mu}{\sqrt{n}\sigma_0} > \frac{t_R - n\mu}{\sqrt{n}\sigma_0}\biggr) + \mathbb{P}_\mu\biggl(\frac{T - n\mu}{\sqrt{n}\sigma_0} < \frac{t_L - n\mu}{\sqrt{n}\sigma_0}\biggr) \\ &= 1 - \Phi\biggl(\frac{t_R - n\mu}{\sqrt{n}\sigma_0}\biggr) + \Phi\biggl(\frac{t_L - n\mu}{\sqrt{n}\sigma_0}\biggr),\end{aligned}$$

where Φ is the standard normal distribution function. To avoid confusion with the test ϕ, we denote the standard normal density in this example by $\dot{\Phi}$. Thus,

$$\frac{d}{d\mu}E_\mu\phi(T) = \frac{n}{\sqrt{n}\sigma_0}\dot{\Phi}\biggl(\frac{t_R - n\mu}{\sqrt{n}\sigma_0}\biggr) - \frac{n}{\sqrt{n}\sigma_0}\dot{\Phi}\biggl(\frac{t_L - n\mu}{\sqrt{n}\sigma_0}\biggr),$$

So putting

$$\frac{d}{d\mu}E_\mu\phi(T)\bigg|_{\mu=\mu_0} = 0,$$

gives

$$\dot{\Phi}\biggl(\frac{t_R - n\mu_0}{\sqrt{n}\sigma_0}\biggr) = \dot{\Phi}\biggl(\frac{t_L - n\mu_0}{\sqrt{n}\sigma_0}\biggr),$$

or

$$(t_R - n\mu_0)^2 = (t_L - n\mu_0)^2.$$

We take the solution $(t_L - n\mu_0) = -(t_R - n\mu_0)$, (because the solution $(t_L - n\mu_0) = (t_R - n\mu_0)$ leads to a test that always rejects, and hence does not have level α, as $\alpha < 1$). Plugging this solution back in gives

$$\begin{aligned}E_{\mu_0}\phi(T) &= 1 - \Phi\biggl(\frac{t_R - n\mu_0}{\sqrt{n}\sigma_0}\biggr) + \Phi\biggl(-\frac{t_R - n\mu_0}{\sqrt{n}\sigma_0}\biggr) \\ &= 2\left(1 - \Phi\biggl(\frac{t_R - n\mu_0}{\sqrt{n}\sigma_0}\biggr)\right).\end{aligned}$$

The requirement $E_{\mu_0}\phi(T) = \alpha$ gives us

$$\Phi\biggl(\frac{t_R - n\mu_0}{\sqrt{n}\sigma_0}\biggr) = 1 - \alpha/2,$$

and hence

$$t_R - n\mu_0 = \sqrt{n}\sigma_0\Phi^{-1}(1-\alpha/2),\ t_L - n\mu_0 = -\sqrt{n}\sigma_0\Phi^{-1}(1-\alpha/2).$$

7.6 Conditional tests $\star$

We now study the case where Θ is an interval in $\mathbb{R}^2$. We let $\theta = (\beta, \gamma)$, and we assume that γ is the parameter of interest. We aim at testing
$H_0 :\ \gamma \le \gamma_0$,
against the alternative
$H_1 :\ \gamma > \gamma_0$.

We assume moreover that we are dealing with an exponential family in canonical form:

$$p_\theta(x) = \exp[\beta T_1(x) + \gamma T_2(x) - d(\theta)]h(x).$$

Then we can restrict ourselves to tests $\phi(T)$ depending only on the sufficient statistic $T = (T_1, T_2)$.

Lemma 7.6.1 *Suppose that $\{\beta :\ (\beta, \gamma_0) \in \Theta\}$ contains an open interval. Let*

$$\phi(T_1, T_2) = \begin{cases} 1 & \textit{if } T_2 > t_0(T_1) \\ q(T_1) & \textit{if } T_2 = t_0(T_1) \\ 0 & \textit{if } T_2 < t_0(T_1) \end{cases},$$

where the constants $t_0(T_1)$ and $q(T_1)$ are allowed to depend on T_1, and are chosen in such a way that

$$E_{\gamma_0}\left(\phi(T_1, T_2)\bigg| T_1\right) = \alpha.$$

Then ϕ is UMPU.

Sketch of proof.

Let $\bar{p}_\theta(t_1, t_2)$ be the density of (T_1, T_2) with respect to dominating measure $\bar{\nu}$:

$$\bar{p}_\theta(t_1, t_2) := \exp[\beta t_1 + \gamma t_2 - d(\theta)]\bar{h}(t_1, t_2).$$

We assume $\bar{\nu}(t_t, t_2) = \bar{\nu}_1(t_1)\bar{\nu}_2(t_2)$ is a product measure. The conditional density of T_2 given $T_1 = t_1$ is then

$$\begin{aligned} \bar{p}_\theta(t_2|t_1) &= \frac{\exp[\beta t_1 + \gamma t_2 - d(\theta)]\bar{h}(t_1, t_2)}{\int_{s_2} \exp[\beta t_1 + \gamma s_2 - d(\theta)]\bar{h}(t_1, s_2) d\bar{\nu}_2(s_2)} \\ &= \exp[\gamma t_2 - d(\gamma|t_1)]\bar{h}(t_1, t_2), \end{aligned}$$

where

$$d(\gamma|t_1) := \log\left(\int_{s_2} \exp[\gamma s_2]\bar{h}(t_1, s_2) d\bar{\nu}_2(s_2)\right).$$

In other words, the conditional distribution of T_2 given $T_1 = t_1$
- does not depend on β,
- is a one-parameter exponential family in canonical form.
This implies that given $T_1 = t_1$, ϕ is UMPU.

Result 1 *The test ϕ has level α, i.e.*

$$\sup_{\gamma \le \gamma_0} E_{(\beta.\gamma)} \phi(T) = E_{(\beta,\gamma_0)} \phi(T) = \alpha, \ \forall \ \beta.$$

Proof of Result 1.

$$\sup_{\gamma \le \gamma_0} E_{(\beta,\gamma)} \phi(T) \ge E_{(\beta,\gamma_0)} \phi(T) = E_{(\beta,\gamma_0)} E_{\gamma_0}(\phi(T)|T_1) = \alpha.$$

Conversely,

$$\sup_{\gamma \le \gamma_0} E_{(\beta,\gamma)} \phi(T) = \sup_{\gamma \le \gamma_0} E_{(\beta,\gamma)} \underbrace{E_\gamma(\phi(T)|T_1)}_{\le \alpha} \le \alpha.$$

Result 2 *The test ϕ is unbiased.*

Proof of Result 2. If $\gamma > \gamma_0$, it holds that $E_\gamma(\phi(T)|T_1) \ge \alpha$, as the conditional test is unbiased. Thus, also, for all β,

$$E_{(\beta,\gamma)} \phi(T) = E_{(\beta,\gamma)} E_\gamma(\phi(T)|T_1) \ge \alpha,$$

i.e., ϕ is unbiased.

Result 3 *Let ϕ' be a test with level*

$$\alpha' := \sup_\beta \sup_{\gamma \le \gamma_0} E_{(\beta,\gamma)} \phi'(T) \le \alpha,$$

and suppose moreover that ϕ' is unbiased, i.e., that

$$\sup_{\gamma \le \gamma_0} \sup_\beta E_{(\beta,\gamma)} \phi'(T) \le \inf_{\gamma > \gamma_0} \inf_\beta E_{(\beta,\gamma)} \phi'(T).$$

Then, conditionally on T_1, ϕ' has level α'.

Proof of Result 3. As

$$\alpha' = \sup_\beta \sup_{\gamma \le \gamma_0} E_{(\beta,\gamma)} \phi'(T)$$

we know that

$$E_{(\beta,\gamma_0)} \phi'(T) \le \alpha', \ \forall \ \beta.$$

Conversely, the unbiasedness implies that for all $\gamma > \gamma_0$,

$$E_{(\beta,\gamma)} \phi'(T) \ge \alpha', \forall \ \beta.$$

A continuity argument therefore gives

$$E_{(\beta,\gamma_0)} \phi'(T) = \alpha', \ \forall \ \beta.$$

In other words, we have

$$E_{(\beta,\gamma_0)}(\phi'(T)-\alpha')=0, \forall \ \beta.$$

But then also

$$E_{(\beta,\gamma_0)}E_{\gamma_0}\left((\phi'(T)-\alpha')\Big|T_1\right)=0, \ \forall \ \beta,$$

which we can write as

$$E_{(\beta,\gamma_0)}h(T_1)=0, \forall \ \beta.$$

The assumption that $\{\beta : \ (\beta,\gamma_0)\in\Theta\}$ contains an open interval implies that T_1 is complete for (β,γ_0). So we must have

$$h(T_1)=0, \ P_{(\beta,\gamma_0)}\text{–a.s.}, \ \forall \ \beta,$$

or, by the definition of h,

$$E_{\gamma_0}(\phi'(T)|T_1)=\alpha', \ P_{(\beta,\gamma_0)} - \text{a.s.}, \ \forall \ \beta.$$

So conditionally on T_1, the test ϕ' has level α'.

Result 4 *Let ϕ' be a test as given in Result 3. Then ϕ' can not be more powerful than ϕ at any (β,γ), with $\gamma>\gamma_0$.*

Proof of Result 4. By the Neyman Pearson lemma, conditionally on T_1, we have

$$E_\gamma(\phi'(T)|T_1)\le E_\gamma(\phi(T)|T_1), \ \forall \ \gamma>\gamma_0.$$

Thus also

$$E_{(\beta,\gamma)}\phi'(T)\le E_{(\beta,\gamma)}\phi(T), \ \forall \ \beta, \ \gamma>\gamma_0.$$

□

Example 7.6.1 Comparing the means of two Poissons
Consider two independent samples $\mathbf{X}=(X_1,\ldots,X_n)$ and $\mathbf{Y}=(Y_1,\ldots,Y_m)$, where $X_1,\ldots,X_n$ are i.i.d. Poisson(λ)*-distributed, and $Y_1,\ldots,Y_m$ are i.i.d.* Poisson(μ)*-distributed. We aim at testing*
$H_0 : \ \lambda\le\mu$,
against the alternative
$H_1 : \ \lambda>\mu$.

Define

$$\beta:=\log(\mu), \ \gamma:=\log(\lambda/\mu).$$

The testing problem is equivalent to
$H_0 : \ \gamma\le\gamma_0$,
against the alternative
$H_1 : \ \gamma>\gamma_0$,
where $\gamma_0:=0$.

The density is

$$\mathbf{p}_\theta(x_1,\ldots,x_n,y_1,\ldots,y_m)$$

$$
\begin{aligned}
&= \exp\Big[\log(\lambda)\sum_{i=1}^n x_i + \log(\mu)\sum_{j=1}^m y_j - n\lambda - m\mu\Big]\prod_{i=1}^n \frac{1}{x_i!}\prod_{j=1}^m \frac{1}{y_j!} \\
&= \exp\Big[\log(\mu)(\sum_{i=1}^n x_i + \sum_{j=1}^m y_j) + \log(\lambda/\mu)\sum_{i=1}^n x_i - n\lambda - m\mu\Big]h(\mathbf{x},\mathbf{y}) \\
&= \exp[\beta T_1(\mathbf{x},\mathbf{y}) + \gamma T_2(\mathbf{x}) - d(\theta)]h(\mathbf{x},\mathbf{y}),
\end{aligned}
$$

where

$$T_1(\mathbf{X},\mathbf{Y}) := \sum_{i=1}^n X_i + \sum_{j=1}^m Y_j,\ T_2(\mathbf{X}) := \sum_{i=1}^n X_i,$$

and

$$h(\mathbf{x},\mathbf{y}) := \prod_{i=1}^n \frac{1}{x_i!}\prod_{j=1}^m \frac{1}{y_j!}.$$

The conditional distribution of T_2 *given* $T_1 = t_1$ *is the* Binomial(t_1, p)*-distribution, with*

$$p = \frac{n\lambda}{n\lambda + m\mu} = \frac{\mathrm{e}^\gamma}{1+\mathrm{e}^\gamma}.$$

Thus, conditionally on $T_1 = t_1$*, using the observation* T_2 *from the* Binomial(t_1, p)*-distribution, we test*
$H_0 :\ p \le p_0$,
against the alternative
$H_1 :\ p > p_0$,
where $p_0 := n/(n+m)$. *This test is UMPU for the unconditional problem.*

Chapter 8

Comparison of estimators

One can compare estimators in terms of their *risk*. This is done in Sections 8.1, 8.2 and 8.3. Section 8.4 briefly addresses sensitivity and robustness and Section 8.5 discusses computational aspects. These last two sections do not present the details.

8.1 Definition of risk

Consider a random variable X with distribution P_θ, $\theta \in \Theta$. Let $T = T(X)$ be an estimator of a parameter of interest $\gamma = g(\theta)$. A risk function $R(\cdot, \cdot)$ measures the loss due to the error of the estimator. The *risk* depends on the unknown parameter θ and on the estimator. We define the risk as

$$R(\theta, T) := \mathbb{E}_\theta(L(\theta, T(X))$$

where $L(\cdot, \cdot)$ is a given so-called *loss function*[1]. A more detailed description is given in Chapter 10.

Example 8.1.1 Risk of a test
Consider the testing problem
$H_0 : \theta = \theta_0$,
against
$H_1 : \theta = \theta_1$.
Let $\phi(X) \in [0, 1]$ *be a test.*
The risk of the test can then be defined as the probability of an error, i.e.

$$R(\theta, \phi) = \begin{cases} \mathbb{E}_{\theta_0}\phi(X) & \theta = \theta_0 \\ 1 - \mathbb{E}_{\theta_1}\phi(X) & \theta = \theta_1 \end{cases} .$$

Example 8.1.2 Risk of an estimator
In the case $\gamma \in \mathbb{R}$ *an important risk measure is the mean square error*

$$R(\theta, T) := \mathbb{E}_\theta(T(X) - g(\theta))^2 =: \mathrm{MSE}_\theta(T).$$

[1] Note that the quantity $L(\theta, T(X))$ is random. Note also that in the notation of risk $R(\theta, T)$, the symbol T stands for the *map* T.

8.2 Risk and sufficiency

Let $S = S(X)$ be sufficient. Knowing the sufficient statistic S one can forget about the original data X without loosing information. Indeed, the following lemma says that any decision based on the original data X can be replaced by a randomized one which depends only on S and which has the same risk.

Lemma 8.2.1 *Suppose S is sufficient for θ. Let $d : \mathcal{X} \to \mathcal{A}$ be some decision. Then there is a randomized decision $\delta(S)$ that only depends on S, such that*

$$R(\theta, \delta(S)) = R(\theta, d), \ \forall \ \theta.$$

Proof. Let X_s^* be a random variable with distribution $P(X \in \cdot | S = s)$. Then, by construction, for all possible s, the conditional distribution, given $S = s$, of X_s^* and X are equal. It follows that X and X_S^* have the same distribution. Formally, let us write Q_θ for the distribution of S. Then

$$\begin{aligned} P_\theta(X_S^* \in \cdot) &= \int P(X_s^* \in \cdot | S = s) dQ_\theta(s) \\ &= \int P(X \in \cdot | S = s) dQ_\theta(s) = P_\theta(X \in \cdot). \end{aligned}$$

The result of the lemma follows by taking $\delta(s) := d(X_s^*)$. □.

8.3 Rao-Blackwell

The Lemma of Rao-Blackwell says that in the case of convex loss an estimator based on the original data X can be replaced by one based only on S without increasing the risk. Randomization is not needed here.

Lemma 8.3.1 *(Rao Blackwell) Suppose that S is sufficient for θ. Suppose moreover that the action space $\mathcal{A} \subset \mathbb{R}^p$ is convex, and that for each θ, the map $a \mapsto L(\theta, a)$ is convex. Let $d : \mathcal{X} \to \mathcal{A}$ be a decision, and define $d'(s) := E(d(X)|S = s)$ (assumed to exist). Then*

$$R(\theta, d') \le R(\theta, d), \ \forall \ \theta.$$

Proof. Jensen's inequality says that for a convex function g,

$$E(g(X)) \ge g(EX).$$

Hence, $\forall \ \theta$,

$$\begin{aligned} E\biggl(L\Bigl(\theta, d(X)\Bigr) \biggm| S = s \biggr) &\ge L\biggl(\theta, E\Bigl(d(X) | S = s\Bigr)\biggr) \\ &= L(\theta, d'(s)). \end{aligned}$$

By the iterated expectations lemma, we arrive at

$$\begin{aligned} R(\theta, d) &= E_\theta L(\theta, d(X)) \\ &= E_\theta E\Big(L\Big(\theta, d(X)\Big)\Big| S\Big) \\ &\geq E_\theta L(\theta, d'(S)). \end{aligned}$$

□

Example 8.3.1 Mean square error
Let T be an estimator of $g(\theta) \in \mathbb{R}$ and let

$$R(\theta, T) := \mathbb{E}_\theta (T(X) - g(\theta))^2 =: \mathrm{MSE}_\theta(T).$$

Let S be sufficient and $\tilde{T} := \mathbb{E}(T|S)$. Then by the Rao-Blackwell Lemma

$$R(\theta, \tilde{T}) \leq R(\theta, T) \ \forall \ \theta.$$

The mean square error can be decomposed in the variance term and the squared bias term. Since $\mathbb{E}\tilde{T} = \mathbb{E}T$ by the iterated expectations lemma, we thus have

$$\mathrm{var}_\theta(\tilde{T}) \leq \mathrm{var}_\theta(T), \ \forall \ \theta.$$

Compare with Lemma 5.2.2 and the result of Lehmann-Scheffé in Section 5.3.

8.4 Sensitivity and robustness

We can compare estimators with respect to their sensitivity to large errors in the data. Let $X_1, \ldots, X_n$ be i.id. copies of a random variable X. Let $T_n = T_n(X_1, \ldots, X_n)$ be a real-valued estimator defined for each n, and symmetric in $X_1, \ldots, X_n$.

Influence of a single additional observation
The influence function is

$$l(x) := T_{n+1}(X_1, \ldots, X_n, x) - T_n(X_1, \ldots, X_n), \ x \in \mathbb{R}.$$

Break down point
Let for $m \leq n$,

$$\epsilon(m) := \sup_{x_1^*, \ldots, x_m^*} |T(x_1^*, \ldots, x_m^*, X_{m+1}, \ldots, X_n)|.$$

If $\epsilon(m) := \infty$, we say that with m outliers the estimator can break down. The break down point is defined as

$$\epsilon^* := \min\{m : \ \epsilon(m) = \infty\}/n.$$

An estimator is called robust if it has a bounded influence function and/or a large breakdown point.

8.5 Computational aspects

Today, the data are often high-dimensional and the number of parameters p is also very large. Maximum likelihood estimation for example requires maximization of a function of p variables and this can be very hard if p is large. The more so if the likelihood is not concave, or if there are e.g. some parameters are integer valued etc. We will moreover examine in Chapter 10 Bayesian theory. Then on needs to find so-called "posterior distributions", which is typically computationally very hard (this is where MCMC (Monte Carlo Markov Chain) algorithms come in). Clearly, an estimator which cannot be computed (say in polynomial time) is of little practical value.

Chapter 9

Equivariant statistics

As we have seen in Chapter 5 for instance, it can be useful to restrict attention to a collection of statistics satisfying certain desirable properties. In Chapter 5, we restricted ourselves to unbiased estimators. In this chapter, equivariance will be the key concept.

The data consists of i.i.d. real-valued random variables $X_1, \ldots, X_n$. We write $\mathbf{X} := (X_1, \ldots, X_n)$. The density w.r.t. some dominating measure ν, of a single observation is denoted by p_θ. The density of $\mathbf{X}$ is $\mathbf{p}_\theta(\mathbf{x}) = \prod_i p_\theta(x_i)$, $\mathbf{x} = (x_1, \ldots, x_n)$.

Location model
Then $\theta \in \mathbb{R}$ is a location parameter, and we assume

$$X_i = \theta + \epsilon_i, \ i = 1, \ldots, n.$$

We are interested in estimating θ. Both the parameter space Θ, as well as the action space $\mathcal{A}$, are the real line $\mathbb{R}$. We assume $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. with a known density $p_0(\cdot)$.

Location-scale model
Here $\theta = (\mu, \sigma)$, with $\mu \in \mathbb{R}$ a location parameter and $\sigma > 0$ a scale parameter. We assume

$$X_i = \mu + \sigma\epsilon_i, \ i = 1, \ldots, n.$$

The parameter space Θ and action space $\mathcal{A}$ are both $\mathbb{R} \times (0, \infty)$. We assume $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. with a known density $p_0(\cdot)$.

9.1 Equivariance in the location model

Definition 9.1.1 *A statistic $T = T(\mathbf{X})$ is called* location equivariant *if for all constants $c \in \mathbb{R}$ and all $\mathbf{x} = (x_1, \ldots, x_n)$,*

$$T(x_1 + c, \ldots, x_n + c) = T(x_1, \ldots, x_n) + c.$$

Examples

$$T = \begin{cases} \bar{X} \\ X_{(\frac{n+1}{2})} \quad (n \text{ odd}) \\ \cdots \end{cases} .$$

Definition 9.1.2 *A loss function $L(\theta, a)$ is called* location invariant *if for all $c \in \mathbb{R}$,*

$$L(\theta + c, a + c) = L(\theta, a), \ (\theta, a) \in \mathbb{R}^2.$$

In this section we abbreviate location equivariance (invariance) to simply equivariance (invariance), and we assume throughout that the loss $L(\theta, a)$ is invariant.

Corollary 9.1.1 *If T is equivariant (and $L(\theta, a)$ is invariant), then*

$$\begin{aligned} R(\theta, T) &= E_\theta L(\theta, T(\mathbf{X})) = E_\theta L(0, T(\mathbf{X}) - \theta) \\ &= E_\theta L(0, T(\mathbf{X} - \theta)) = E_\theta L_0[T(\varepsilon)], \end{aligned}$$

where $L_0[a] := L(0, a)$ and $\varepsilon := (\epsilon_1, \ldots, \epsilon_n)$. Because the distribution of ε does not depend on θ, we conclude that the risk does not depend on θ. We may therefore omit the subscript θ in the last expression:

$$R(\theta, T) = EL_0[T(\varepsilon)].$$

Since for $\theta = 0$, we have the equality $\mathbf{X} = \varepsilon$ we may alternatively write

$$R(\theta, T) = E_0 L_0[T(\mathbf{X})] = R(0, T).$$

Definition 9.1.3 *An equivariant statistic T is called* uniform minimum risk equivariant (UMRE) *if*

$$R(\theta, T) = \min_{d \text{ equivariant}} R(\theta, d), \ \forall \ \theta,$$

or equivalently,

$$R(0, T) = \min_{d \text{ equivariant}} R(0, d).$$

9.1.1 Construction of the UMRE estimator

Lemma 9.1.1 *Let $Y_i := X_i - X_n$, $i = 1, \ldots, n$, and $\mathbf{Y} := (Y_1, \ldots, Y_n)$. We have*

$$T \text{ equivariant } \Leftrightarrow T(\mathbf{X}) = T(\mathbf{Y}) + X_n.$$

Proof.
($\Rightarrow$) Trivial.
($\Leftarrow$) Replacing $\mathbf{X}$ by $\mathbf{X} + c$ leaves $\mathbf{Y}$ unchanged (i.e. $\mathbf{Y}$ is invariant). So $T(\mathbf{X} + c) = T(\mathbf{Y}) + X_n + c = T(\mathbf{X}) + c$. □

Theorem 9.1.1 *Let $Y_i := X_i - X_n$, $i = 1, \ldots, n$, $\mathbf{Y} := (Y_1, \ldots, Y_n)$, and define*

$$T^*(\mathbf{Y}) := \arg\min_v E\left[L_0\left(v + \epsilon_n \right) \middle| \mathbf{Y} \right].$$

Moreover, let

$$T^*(\mathbf{X}) := T^*(\mathbf{Y}) + X_n.$$

Then T^* is UMRE.

Proof. First, note that $\mathbf{Y}$ and its distribution does not depend on θ, so that T^* is indeed a statistic. It is also equivariant, by the previous lemma.

Let T be an equivariant statistic. Then $T(\mathbf{X}) = T(\mathbf{Y}) + X_n$. So

$$T(\mathbf{X}) - \theta = T(\mathbf{Y}) + \epsilon_n.$$

Hence

$$R(0, T) = EL_0\left(T(\mathbf{Y}) + \epsilon_n \right) = E\ E\left[L_0\left(T(\mathbf{Y}) + \epsilon_n \right) \middle| \mathbf{Y} \right].$$

But

$$\begin{aligned} E\left[L_0\left(T(\mathbf{Y}) + \epsilon_n \right) \middle| \mathbf{Y} \right] &\geq \min_v E\left[L_0\left(v + \epsilon_n \right) \middle| \mathbf{Y} \right] \\ &= E\left[L_0\left(T^*(\mathbf{Y}) + \epsilon_n \right) \middle| \mathbf{Y} \right]. \end{aligned}$$

Hence,

$$R(0, T) \geq E\ E\left[L_0\left(T^*(\mathbf{Y}) + \epsilon_n \right) \middle| \mathbf{Y} \right] = R(0, T^*).$$

□

9.1.2 Quadratic loss: the Pitman estimator

Corollary 9.1.2 *If we take quadratic loss*

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

we get $L_0[a] = a^2$, and so, for $\mathbf{Y} = \mathbf{X} - X_n$,

$$T^*(\mathbf{Y}) = \arg\min_v E\left[\left(v + \epsilon_n \right)^2 \middle| \mathbf{Y} \right]$$

$$= -E(\epsilon_n | \mathbf{Y}),$$

and hence

$$T^*(\mathbf{X}) = X_n - E(\epsilon_n | \mathbf{Y}).$$

This estimator is called the Pitman estimator.

To investigate the case of quadratic risk further, we:

Note If (X, Z) has density $f(x, z)$ w.r.t. Lebesgue measure, then the density of $Y := X - Z$ is

$$f_Y(y) = \int f(y + z, z) dz.$$

Lemma 9.1.2 *Consider quadratic loss. Let* $\mathbf{p}_0$ *be the density of* $\varepsilon = (\epsilon_1, \ldots, \epsilon_n)$ *w.r.t. Lebesgue measure. Then a UMRE statistic is*

$$T^*(\mathbf{X}) = \frac{\int z\mathbf{p}_0(X_1 - z, \ldots, X_n - z)dz}{\int \mathbf{p}_0(X_1 - z, \ldots, X_n - z)dz}.$$

Proof. Let $\mathbf{Y} = \mathbf{X} - X_n$. The random vector $\mathbf{Y}$ has density

$$f_{\mathbf{Y}}(y_1, \ldots, y_{n-1}, 0) = \int \mathbf{p}_0(y_1 + z, \ldots, y_{n-1} + z, z)dz.$$

So the density of ϵ_n given $\mathbf{Y} = \mathbf{y} = (y_1, \ldots, y_{n-1}, 0)$ is

$$f_{\epsilon_n}(u) = \frac{\mathbf{p}_0(y_1 + u, \ldots, y_{n-1} + u, u)}{\int \mathbf{p}_0(y_1 + z, \ldots, y_{n-1} + z, z)dz}.$$

It follows that

$$E(\epsilon_n|\mathbf{y}) = \frac{\int u\mathbf{p}_0(y_1 + u, \ldots, y_{n-1} + u, u)du}{\int \mathbf{p}_0(y_1 + z, \ldots, y_{n-1} + z, z)dz}.$$

Thus

$$\begin{aligned}
E(\epsilon_n|\mathbf{Y}) &= \frac{\int u\mathbf{p}_0(Y_1 + u, \ldots, Y_{n-1} + u, u)du}{\int \mathbf{p}_0(Y_1 + z, \ldots, Y_{n-1} + z, z)dz} \\
&= \frac{\int u\mathbf{p}_0(X_1 - X_n + u, \ldots, X_{n-1} - X_n + u, u)du}{\int \mathbf{p}_0(X_1 - X_n + z, \ldots, X_{n-1} - X_n + z, z)dz} \\
&= X_n - \frac{\int z\mathbf{p}_0(X_1 - z, \ldots, X_{n-1} - z, X_n - z)dz}{\int \mathbf{p}_0(X_1 - z, \ldots, X_{n-1} - z, X_n - z)dz}.
\end{aligned}$$

Finally, recall that $T^*(\mathbf{X}) = X_n - E(\epsilon_n|\mathbf{Y})$. □

Example 9.1.1 Uniform distribution with unknown midpoint *Suppose* $X_1, \ldots, X_n$ *are i.i.d.* Uniform$[\theta - 1/2, \theta + 1/2]$, $\theta \in \mathbb{R}$. *Then*

$$p_0(x) = \mathrm{l}\{|x| \le 1/2\}.$$

We have

$$\max_{1 \le i \le n} |x_i - z| \le 1/2 \ \Leftrightarrow\ x_{(n)} - 1/2 \le z \le x_{(1)} + 1/2.$$

So

$$\mathbf{p}_0(x_1 - z, \ldots, x_n - z) = \mathrm{l}\{x_{(n)} - 1/2 \le z \le x_{(1)} + 1/2\}.$$

Thus, writing

$$T_1 := X_{(n)} - 1/2,\ T_2 := X_{(1)} + 1/2,$$

the UMRE estimator T^* *is*

$$T^* = \left(\int_{T_1}^{T_2} zdz\right) \Big/ \left(\int_{T_1}^{T_2} dz\right) = \frac{T_1 + T_2}{2} = \frac{X_{(1)} + X_{(n)}}{2}.$$

9.1.3 Invariant statistics

We now consider more general invariant statistics $\mathbf{Y}$.

Definition 9.1.4 *A map* $\mathbf{Y} : \mathbb{R}^n \to \mathbb{R}^n$ *is called* maximal invariant *if*

$$\mathbf{Y}(\mathbf{x}) = \mathbf{Y}(\mathbf{x}') \ \Leftrightarrow \ \exists \, c : \ \mathbf{x} = \mathbf{x}' + c.$$

(The constant c may depend on $\mathbf{x}$ *and* $\mathbf{x}'$*.)*

Example The map $\mathbf{Y}(\mathbf{x}) := \mathbf{x} - x_n$ is maximal invariant:
($\Leftarrow$) is clear
($\Rightarrow$) if $\mathbf{x} - x_n = \mathbf{x}' - x_n'$, we have $\mathbf{x} = \mathbf{x}' + (x_n - x_n')$.

More generally:

Example Let $d(\mathbf{X})$ be equivariant. Then $\mathbf{Y} := \mathbf{X} - d(\mathbf{X})$ is maximal invariant.

Theorem 9.1.2 *Suppose that* $d(\mathbf{X})$ *is equivariant. Let* $\mathbf{Y} := \mathbf{X} - d(\mathbf{X})$*, and*

$$T^*(\mathbf{Y}) := \arg\min_v E\left[L_0\Big(v + d(\varepsilon) \Big) \Big| \mathbf{Y} \right].$$

Then

$$T^*(\mathbf{X}) := T^*(\mathbf{Y}) + d(\mathbf{X})$$

is UMRE.

Proof. Let T be an equivariant estimator. Then

$$\begin{aligned} T(\mathbf{X}) &= T(\mathbf{X} - d(\mathbf{X})) + d(\mathbf{X}) \\ &= T(\mathbf{Y}) + d(\mathbf{X}). \end{aligned}$$

Hence

$$\begin{aligned} E\left[L_0\Big(T(\varepsilon) \Big) \Big| \mathbf{Y} \right] &= E\left[L_0\Big(T(\mathbf{Y}) + d(\varepsilon) \Big) \Big| \mathbf{Y} \right] \\ &\geq \min_v E\left[L_0\Big(v + d(\varepsilon) \Big) \Big| \mathbf{Y} \right]. \end{aligned}$$

Now, use the iterated expectation lemma. □

9.1.4 Quadratic loss and Basu's Lemma

For quadratic loss ($L_0[a] = a^2$), the definition of $T^*(\mathbf{Y})$ in the above theorem is

$$T^*(\mathbf{Y}) = -E(d(\varepsilon)|\mathbf{Y}) = -E_0(d(\mathbf{X})|\mathbf{X} - d(\mathbf{X})),$$

so that

$$T^*(\mathbf{X}) = d(\mathbf{X}) - E_0(d(\mathbf{X})|\mathbf{X} - d(\mathbf{X})).$$

So for a equivariant estimator T, we have

$$T \text{ is UMRE} \ \Leftrightarrow \ E_0(T(\mathbf{X})|\mathbf{X} - T(\mathbf{X})) = 0.$$

From the right hand side, we conclude that $E_0 T = 0$ and hence $E_\theta(T) = \theta$ $\forall\ \theta$. Thus, in the case of quadratic loss, an UMRE estimator is unbiased. Conversely, suppose we have an equivariant and unbiased estimator T. If $T(\mathbf{X})$ and $\mathbf{X} - T(\mathbf{X})$ are independent, it follows that

$$E_0(T(\mathbf{X})|\mathbf{X} - T(\mathbf{X})) = E_0 T(\mathbf{X}) = 0.$$

So then T is UMRE.

To check independence, Basu's lemma can be useful.

Basu's lemma *Let X have distribution P_θ, $\theta \in \Theta$. Suppose T is sufficient and complete, and that $Y = Y(X)$ has a distribution that does not depend on θ. Then, for all θ, T and Y are independent under P_θ.*

Proof. Let A be some measurable set, and

$$h(T) := P(Y \in A|T) - P(Y \in A).$$

Notice that indeed, $P(Y \in A|T)$ does not depend on θ because T is sufficient. Because

$$E_\theta h(T) = 0,\ \forall\ \theta,$$

we conclude from the completness of T that

$$h(T) = 0,\ P_\theta\text{–a.s.},\ \forall\ \theta,$$

in other words,

$$P(Y \in A|T) = P(Y \in A),\ P_\theta\text{–a.s.},\ \forall\ \theta.$$

Since A was arbitrary, we thus have that the conditional distribution of Y given T is equal to the unconditional distribution:

$$P(Y \in \cdot|T) = P(Y \in \cdot),\ P_\theta\text{–a.s.},\ \forall\ \theta,$$

that is, for all θ, T and Y are independent under P_θ. □

Basu's lemma is intriguing: it proves a probabilistic property (independence) via statistical concepts.

Example 9.1.2 UMRE estimator for the mean of the normal distribution: σ^2 known
Let $X_1, \ldots, X_n$ be independent $\mathcal{N}(\theta, \sigma^2)$, with σ^2 known. Then $T := \bar{X}$ is sufficient and complete, and moreover, the distribution of $\mathbf{Y} := \mathbf{X} - \bar{X}$ does not depend on θ. So by Basu's lemma, $\bar{X}$ and $\mathbf{X} - \bar{X}$ are independent. Hence, $\bar{X}$ is UMRE.
Remark *Indeed, Basu's lemma is peculiar: $\bar{X}$ and $\mathbf{X} - \bar{X}$ of course remain independent if the mean θ is known and/or the variance σ^2 is unknown!*
Remark *As a by-product, one concludes the independence of $\bar{X}$ and the sample variance $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$, because S^2 is a function of $\mathbf{X} - \bar{X}$.*

9.2 Equivariance in the location-scale model ⋆

Location-scale model
We assume

$$X_i = \mu + \sigma \epsilon_i, \ i = 1, \ldots, n.$$

The unknown parameter is $\theta = (\mu, \sigma)$, with $\mu \in \mathbb{R}$ a location parameter and $\sigma > 0$ a scale parameter. The parameter space Θ and action space $\mathcal{A}$ are both $\mathbb{R} \times \mathbb{R}_+$ ($\mathbb{R}_+ := (0, \infty)$). The distribution of $\varepsilon = (\epsilon_1, \ldots, \epsilon_n)$ is assumed to be known.

Definition 9.2.1 *A statistic $T = T(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X}))$ is called* location-scale equivariant *if for all constants $b \in \mathbb{R}$, $c \in \mathbb{R}_+$, and all $\mathbf{x} = (x_1, \ldots, x_n)$,*

$$T(b + cx_1, \ldots, b + cx_n) = b + cT(x_1, \ldots, x_n)$$

and

$$T_2(b + cx_1, \ldots, b + cx_n) = cT_2(x_1, \ldots, x_n).$$

Definition 9.2.2 *A loss function $L(\mu, \sigma, a_1, a_2)$ is called* location-scale invariant *if for all $(\mu, a_1, b) \in \mathbb{R}^3$, $(\sigma, a_2, c) \in \mathbb{R}^3_+$*

$$L(b + c\mu, c\sigma, b + ca_1, ca_2) = L(\mu, \sigma, a_1, a_2).$$

In this section we abbreviate location-scale equivariance (invariance) to simply equivariance (invariance), and we assume throughout that the loss $L(\theta, a)$ is invariant.

Corollary 9.2.1 *If T is equivariant (and $L(\theta, a)$ is invariant), then*

$$\begin{aligned} R(\theta, T) &= E_\theta L(\mu, \sigma, T_1(\mathbf{X}), T_2(\mathbf{X})) \\ &= E_\theta L\left(0, 1, \frac{T_1(\mathbf{X}) - \mu}{\sigma}, \frac{T_2(\mathbf{X})}{\sigma}\right) \\ &= E_\theta L\Big(0, 1, T_1(\epsilon), T_2(\epsilon)\Big) = E_\theta L_0(T(\varepsilon)), \end{aligned}$$

where $L_0(a_1, a_2) := L(0, 1, a_1, a_2)$. We conclude that the risk does not depend on θ. We may therefore omit the subscript θ in the last expression:

$$R(\theta, T) = EL_0(T(\varepsilon)).$$

Definition 9.2.3 *An equivariant statistic T is called* uniform minimum risk equivariant (UMRE) *if*

$$R(\theta, T) = \min_{d \text{ equivariant}} R(\theta, d), \ \forall \ \theta,$$

or equivalently,

$$R(0, 1, T_1, T_2) = \min_{d \text{ equivariant}} R(0, 1, d_1, d_2).$$

9.2.1 Construction of the UMRE estimator ⋆

Theorem 9.2.1 *Suppose that $d(\mathbf{X})$ is equivariant. Let*

$$\mathbf{Y} := \frac{\mathbf{X} - d_1(\mathbf{X})}{d_2(\mathbf{X})},$$

and

$$T^*(\mathbf{Y}) := \arg \min_{a_1 \in \mathbb{R},\ a_2 \in \mathbb{R}_+} E\biggl[L_0 \biggl(d_1(\varepsilon) + d_2(\varepsilon) a_1, d_2(\varepsilon) a_2 \biggr) \biggg| \mathbf{Y} \biggr].$$

Then

$$T^*(\mathbf{X}) := \begin{pmatrix} d_1(\mathbf{X}) + d_2(\mathbf{X}) T_1^*(\mathbf{Y}) \\ d_2(\mathbf{X}) T_2^*(\mathbf{Y}) \end{pmatrix}$$

is UMRE.

Proof. We have

$$\mathbf{Y} = \frac{\mathbf{X} - d_1(\mathbf{X})}{d_2(\mathbf{X})} = \frac{\varepsilon - d_1(\varepsilon)}{d_2(\varepsilon)}.$$

So

$$\varepsilon = d_1(\varepsilon) + d_2(\varepsilon)\mathbf{Y}.$$

Let T be an equivariant estimator. Then

$$EL_0 \biggl(T_1(\varepsilon), T_2(\varepsilon) \biggr)$$

$$\begin{aligned}
&= EL_0 \biggl(T_1(d_1(\varepsilon) + d_2(\varepsilon)\mathbf{Y}), T_2(d_1(\varepsilon) + d_2(\varepsilon)\mathbf{Y}) \biggr) \\
&= EL_0 \biggl(d_1(\varepsilon) + d_2(\varepsilon) T_1(\mathbf{Y}), d_2(\varepsilon) T_2(\mathbf{Y}) \biggr) \\
&= EE\biggl[L_0 \biggl(d_1(\varepsilon) + d_2(\varepsilon) T_1(\mathbf{Y}), d_2(\varepsilon) T_2(\mathbf{Y}) \biggr) \biggg| \mathbf{Y} \biggr] \\
&\geq E \min_{a_1 \in \mathbb{R},\ a_2 \in \mathbb{R}_+} E\biggl[L_0 \biggl(d_1(\varepsilon) + d_2(\varepsilon) a_1, d_2(\varepsilon) a_2 \biggr) \biggg| \mathbf{Y} \biggr] \\
&= EE\biggl[L_0 \biggl(d_1(\varepsilon) + d_2(\varepsilon) T_1^*(\mathbf{Y}), d_2(\varepsilon) T_2^*(\mathbf{Y}) \biggr) \biggg| \mathbf{Y} \biggr].
\end{aligned}$$

□

9.2.2 Quadratic loss ⋆

For quadratic loss ($L_0(a_1, a_2) := a_1^2$), the definition of $T^*(\mathbf{Y})$ in the above theorem is

$$T^*(\mathbf{Y}) = \arg \min_{a_1 \in \mathbb{R}} E\biggl[\biggl(d_1(\varepsilon) + d_2(\varepsilon) a_1 \biggr)^2 \biggg| \mathbf{Y} \biggr].$$

We then have:

Lemma 9.2.1 *Suppose that d is equivariant, and sufficient and complete. Then*

$$T^*(\mathbf{X}) := d_1(\mathbf{X}) - d_2(\mathbf{X}) \frac{Ed_1(\varepsilon)d_2(\varepsilon)}{Ed_2^2(\varepsilon)}$$

is UMRE.

Proof. By Basu's lemma, d and $\mathbf{Y}$ are independent. Hence

$$E\left[\left(d_1(\varepsilon) + d_2(\varepsilon)a_1\right)^2 \middle| \mathbf{Y}\right] = E\left(d_1(\varepsilon) + d_2(\varepsilon)a_1\right)^2.$$

Moreover

$$\arg\min_{a_1 \in \mathbb{R}} E\left(d_1(\varepsilon) + d_2(\varepsilon)a_1\right)^2 = -\frac{Ed_1(\varepsilon)d_2(\varepsilon)}{Ed_2^2(\varepsilon)}.$$

□

Example 9.2.1 UMRE of the mean of the normal distribution: σ^2 unknown

Let $X_1, \ldots, X_n$ be i.i.d. and $\mathcal{N}(\mu, \sigma^2)$-distributed. Define

$$d_1(\mathbf{X}) := \bar{X}, \ d_2(\mathbf{X}) := S,$$

where S^2 is the sample variance

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

It is easy to see that d is equivariant. We moreover know from Example 4.3.6 that d is sufficient, and an application of Lemma 5.4.1 shows that d is also complete. We furthermore have

$$Ed_1(\varepsilon) = E\bar{\epsilon} = 0,$$

and, from Example 9.1.2 (a consequence of Basu's lemma), we know that $d_1(\mathbf{X}) = \bar{X}$ and $d_2(\mathbf{X}) = S$ are independent. So

$$Ed_1(\varepsilon)d_2(\varepsilon) = Ed_1(\varepsilon)Ed_2(\varepsilon) = 0.$$

It follows that $T^*(\mathbf{X}) = \bar{X}$ is UMRE.

Chapter 10

Decision theory

In this chapter, we again denote the observable random variable (the data) by $X \in \mathcal{X}$, and its distribution by $P \in \mathcal{P}$. The probability model is $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$, with θ an unknown parameter.

In particular cases, we apply the results with X being replaced by a vector $\mathbf{X} = (X_1, \ldots, X_n)$, with $X_1, \ldots, X_n$ i.i.d. with distribution $P \in \{P_\theta : \theta \in \Theta\}$ (so that $\mathbf{X}$ has distribution $\mathbb{P} := \prod_{i=1}^n P \in \{\mathbb{P}_\theta = \prod_{i=1}^n P_\theta : \ \theta \in \Theta\}$).

10.1 Decisions and their risk

We give a definition of *risk*, but now somewhat more formal than in Chapter 8.

Let $\mathcal{A}$ be the *action space*.

- $\mathcal{A} = \mathbb{R}$ corresponds to estimating a real-valued parameter.
- $\mathcal{A} = \{0, 1\}$ corresponds to testing a hypothesis.
- $\mathcal{A} = [0, 1]$ corresponds to randomized tests.
- $\mathcal{A} = \{\text{intervals}\}$ corresponds to confidence intervals.

Given the observation X, we decide to take a certain action in $\mathcal{A}$. Thus, an action is a map $d : \mathcal{X} \to \mathcal{A}$, with $d(X)$ being the decision taken. If $\mathcal{A} = \mathbb{R}$ a decision is often called an estimator (denoted by T for instance). If $\mathcal{A} = \{0, 1\}$ or $\mathcal{A} = [0, 1]$, a decision is often called a test (denoted by ϕ for instance).

A *loss function* (*Verlustfunktion*) is a map

$$L : \Theta \times \mathcal{A} \to \mathbb{R},$$

with $L(\theta, a)$ being the loss when the parameter value is θ and one takes action a.

The risk of decision $d(X)$ is defined as

$$R(\theta, d) := E_\theta L(\theta, d(X)), \ \theta \in \Theta.$$

Example 10.1.1 Estimation *In the case of estimating a parameter of interest $g(\theta) \in \mathbb{R}$, the action space is $\mathcal{A} = \mathbb{R}$ (or a subset thereof). Important loss functions are then*

$$L(\theta, a) := w(\theta)|g(\theta) - a|^r,$$

where $w(\cdot)$ are given non-negative weights and $r \geq 0$ is a given power. The risk is then

$$R(\theta, d) = w(\theta) E_\theta |g(\theta) - d(X)|^r.$$

A special case is taking $w \equiv 1$ and $r = 2$. Then $L(\theta, a) = (g(\theta) - a)^2$ is called quadratic loss and

$$R(\theta, d) = E_\theta |g(\theta) - d(X)|^2$$

is called the mean square error.

Example 10.1.2 Tests *Consider testing the hypothesis*
$H_0: \ \theta \in \Theta_0$
against the alternative
$H_1: \ \theta \in \Theta_1$.

Here, Θ_0 and Θ_1 are given subsets of Θ with $\Theta_0 \cap \Theta_1 = \emptyset$. As action space, we take $\mathcal{A} = \{0, 1\}$, and as loss

$$L(\theta, a) := \begin{cases} 1 & \text{if } \theta \in \Theta_0 \text{ and } a = 1 \\ c & \text{if } \theta \in \Theta_1 \text{ and } a = 0 \\ 0 & \text{otherwise} \end{cases}.$$

Here $c > 0$ is some given constant. Then

$$R(\theta, d) = \begin{cases} P_\theta(d(X) = 1) & \text{if } \theta \in \Theta_0 \\ c P_\theta(d(X) = 0) & \text{if } \theta \in \Theta_1 \\ 0 & \text{otherwise} \end{cases}.$$

Thus, the risks correspond to the error probabilities (type I and type II errors).

Note
The best decision d is the one with the smallest risk $R(\theta, d)$. However, θ is not known. Thus, if we compare two decision functions d_1 and d_2, we may run into problems because the risks are not comparable: $R(\theta, d_1)$ may be smaller than $R(\theta, d_2)$ for some values of θ, and larger than $R(\theta, d_2)$ for other values of θ.

Example 10.1.3 Estimating the mean
We revisit Example 5.2.1. Let $X \in \mathbb{R}$ and let $g(\theta) = E_\theta X := \mu$. We take quadratic loss

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

Assume that $\text{var}_\theta(X) = 1$ *for all θ. Consider the collection of decisions*

$$d_\lambda(X) := \lambda X,$$

where $0 \leq \lambda \leq 1$. Then

$$\begin{aligned} R(\theta, d_\lambda) &= \text{var}(\lambda X) + \text{bias}_\theta^2(\lambda X) \\ &= \lambda^2 + (\lambda - 1)^2 \mu^2. \end{aligned}$$

The "optimal" choice for λ would be

$$\lambda_{\text{opt}} := \frac{\mu^2}{1+\mu^2},$$

because this value minimizes $R(\theta, d_\lambda)$. However, λ_{opt} depends on the unknown μ, so $d_{\lambda_{\text{opt}}}(X)$ is not an estimator.

Various optimality concepts
We will consider three optimality concepts: *admissibility* (*zulässigkeit*), *minimax* and *Bayes*.

10.2 Admissible decisions

Definition 10.2.1 *A decision d' is called strictly better than d if*

$$R(\theta, d') \le R(\theta, d), \ \forall\ \theta,$$

and

$$\exists\ \theta: \ R(\theta, d') < R(\theta, d).$$

When there exists a d' that is strictly better than d, then d is called inadmissible.

Example 10.2.1 Using only one of the observations
Let, for $n \ge 2$, $X_1, \ldots, X_n$ be i.i.d., with $g(\theta) := E_\theta(X_i) := \mu$, and $\text{var}(X_i) = 1$ *(for all i). Take quadratic loss $L(\theta, a) := |\mu - a|^2$. Consider $d'(X_1, \ldots, X_n) := \bar{X}_n$ and $d(X_1, \ldots, X_n) := X_1$. Then, $\forall\ \theta$,*

$$R(\theta, d') = \frac{1}{n}, \ R(\theta, d) = 1,$$

so that d is inadmissible.

Note
We note that to show that a decision d is inadmissible, it suffices to find a strictly better d'. On the other hand, to show that d is admissible, one has to verify that there is no strictly better d'. So in principle, one then has to take all possible d' into account.

10.2.1 Not using the data at all is admissible

Let $L(\theta, a) := |g(\theta) - a|^r$ and $d(X) := g(\theta_0)$, where θ_0 is some fixed given value.

Lemma 10.2.1 *Assume that P_{θ_0} dominates P_θ* [1] *for all θ. Then d is admissible.*

[1] Let P and Q be probability measures on the same measurable space. Then P *dominates* Q if for all measurable B, $P(B) = 0$ implies $Q(B) = 0$ (Q *is absolut stetig bezüglich* P).

Proof.

Suppose that d' is better than d. Then we have

$$E_{\theta_0}|g(\theta_0) - d'(X)|^r \le 0.$$

This implies that

$$d'(X) = g(\theta_0), \ P_{\theta_0}-\text{almost surely}. \tag{10.1}$$

Since by (10.1),

$$P_{\theta_0}(d'(X) \ne g(\theta_0)) = 0$$

the assumption that P_{θ_0} dominates P_θ, $\forall\ \theta$, implies now

$$P_\theta(d'(X) \ne g(\theta_0)) = 0, \ \forall\ \theta.$$

That is, for all θ, $d'(X) = g(\theta_0)$, P_θ-almost surely, and hence, for all θ, $R(\theta, d') = R(\theta, d)$. So d' is not strictly better than d. We conclude that d is admissible. □

10.2.2 A Neyman Pearson test is admissible

Consider testing
$H_0: \ \theta = \theta_0$
against the alternative
$H_1: \ \theta = \theta_1$.

Define the risk $R(\theta, \phi)$ of a test ϕ as the probability of error of first and second kind:

$$R(\theta, \phi) := \begin{cases} E_\theta \phi(X), & \theta = \theta_0 \\ 1 - E_\theta \phi(X), & \theta = \theta_1 \end{cases}.$$

We let p_0 (p_1) be the density of P_{θ_0} (P_{θ_1}) with respect to some dominating measure ν (for example $\nu = P_{\theta_0} + P_{\theta_1}$). A Neyman Pearson test is (see Section 7.1)

$$\phi_{\rm NP} := \begin{cases} 1 & \text{if } p_1/p_0 > c \\ q & \text{if } p_1/p_0 = c \\ 0 & \text{if } p_1/p_0 < c \end{cases}.$$

Here $0 \le q \le 1$, and $0 \le c < \infty$ are given constants.

Lemma 10.2.2 *A Neyman Pearson test is admissible if and only if one of the following two cases hold:*
i) its power is strictly less than 1,
or
ii) it has minimal level among all tests with power 1.

Proof. Suppose $R(\theta_0, \phi) < R(\theta_0, \phi_{\rm NP})$. Then from the Neyman Pearson Lemma, we know that either $R(\theta_1, \phi) > R(\theta_1, \phi_{\rm NP})$ (i.e., then ϕ is not better then $\phi_{\rm NP}$), or $c = 0$. But when $c = 0$, it holds that $R(\theta_1, \phi_{\rm NP}) = 0$, i.e. then $\phi_{\rm NP}$ has power one.

Similarly, suppose that $R(\theta_1, \phi) < R(\theta_1, \phi_{\rm NP})$. Then it follows from the Neyman Pearson Lemma that $R(\theta_0, \phi) > R(\theta_0, \phi_{\rm NP})$, because we assume $c < \infty$.

□

10.3 Minimax decisions

Definition 10.3.1 *A decision d is called* minimax *if*

$$\sup_\theta R(\theta, d) = \inf_{d'} \sup_\theta R(\theta, d').$$

Thus, the minimax criterion concerns the best decision in the worst possible case.

10.3.1 Minimax Neyman Pearson test

Lemma 10.3.1 *A Neyman Pearson test $\phi_{\rm NP}$ is minimax, if and only if $R(\theta_0, \phi_{\rm NP}) = R(\theta_1, \phi_{\rm NP})$.*

Proof. Let ϕ be a test, and write for $j = 0, 1$,

$$r_j := R(\theta_j, \phi_{\rm NP}), \ r_j' = R(\theta_j, \phi).$$

Suppose that $r_0 = r_1$ and that $\phi_{\rm NP}$ is not minimax. Then, for some test ϕ,

$$\max_j r_j' < \max_j r_j.$$

This implies that both

$$r_0' < r_0, \ r_1' < r_1$$

and by the Neyman Pearson Lemma, this is not possible.

Let $S = \{(R(\theta_0, \phi), R(\theta_1, \phi)) : \ \phi : \mathcal{X} \to [0, 1]\}$. Note that S is convex. Thus, if $r_0 < r_1$, we can find a test ϕ with $r_0 < r_0' < r_1$ and $r_1' < r_1$. So then $\phi_{\rm NP}$ is not minimax. Similarly if $r_0 > r_1$.

□

10.4 Bayes decisions

Suppose the parameter space Θ is a measurable space. We can then equip it with a probability measure Π. We call Π the *a priori* distribution.

Definition 10.4.1 *The* Bayes *risk (with respect to the probability measure Π) is*

$$r(\Pi, d) := \int_\Theta R(\vartheta, d) d\Pi(\vartheta).$$

A decision d is called Bayes *(with respect to Π) if*

$$r(\Pi, d) = \inf_{d'} r(\Pi, d').$$

Let Π have density $w := d\Pi/d\mu$ with respect to some dominating measure μ. We may then write

$$r(\Pi, d) = \int_\Theta R(\vartheta, d) w(\vartheta) d\mu(\vartheta) := r_w(d).$$

Thus, the Bayes risk may be thought of as taking a weighted average of the risks. For example, one may want to assign more weight to "important" values of θ.

10.4.1 Bayes test

Consider again the testing problem
$H_0 : \ \theta = \theta_0$
against the alternative
$H_1 : \ \theta = \theta_1$.

Let $L(\theta_0, a) := a$ and $L(\theta_1, a) := 1 - a$, $w(\theta_0) =: w_0$ and $w(\theta_1) =: w_1 = 1 - w_0$. Then

$$r_w(\phi) := w_0 R(\theta_0, \phi) + w_1 R(\theta_1, \phi).$$

We take $0 < w_0 = 1 - w_1 < 1$.

Lemma 10.4.1 *Bayes test is*

$$\phi_{\rm Bayes} = \begin{cases} 1 & \text{if } p_1/p_0 > w_0/w_1 \\ q & \text{if } p_1/p_0 = w_0/w_1 \\ 0 & \text{if } p_1/p_0 < w_0/w_1 \end{cases} .$$

Proof.

$$\begin{aligned} r_w(\phi) &= w_0 \int \phi p_0 + w_1 (1 - \int \phi p_1) \\ &= \int \phi (w_0 p_0 - w_1 p_1) + w_1. \end{aligned}$$

So we choose $\phi \in [0, 1]$ to minimize $\phi(w_0 p_0 - w_1 p_1)$. This is done by taking

$$\phi = \begin{cases} 1 & \text{if } w_0 p_0 - w_1 p_1 < 0 \\ q & \text{if } w_0 p_0 - w_1 p_1 = 0 \\ 0 & \text{if } w_0 p_0 - w_1 p_1 > 0 \end{cases} ,$$

where for q we may take any value between 0 and 1. □

Note that

$$2 r_w(\phi_{\rm Bayes}) = 1 - \int |w_1 p_1 - w_0 p_0|.$$

In particular, when $w_0 = w_1 = 1/2$,

$$2 r_w(\phi_{\rm Bayes}) = 1 - \int |p_1 - p_0|/2,$$

i.e., the risk is large if the two densities are close to each other.

10.5 Construction of Bayes estimators

Let X have distribution $P \in \mathcal{P} := \{P_\theta : \ \theta \in \Theta\}$. Suppose $\mathcal{P}$ is dominated by a (σ-finite) measure ν, and let $p_\theta = dP_\theta / d\nu$ denote the densities. Let Π be an a priori distribution on Θ, with density $w := d\Pi / d\mu$. We now think of p_θ as the density of X *given* the value of θ. We write it as

$$p_\theta(x) = p(x|\theta), \ x \in \mathcal{X}.$$

Moreover, we define the marginal density

$$p(\cdot) := \int_\Theta p(\cdot|\vartheta) w(\vartheta) d\mu(\vartheta).$$

Definition 10.5.1 *The* a posteriori *density of θ is*

$$w(\vartheta|x) = p(x|\vartheta) \frac{w(\vartheta)}{p(x)}, \ \vartheta \in \Theta, \ x \in \mathcal{X}.$$

Lemma 10.5.1 *Given the data $X = x$, consider θ as a random variable with density $w(\vartheta|x)$. Let*

$$l(x,a) := E[L(\theta, a)|X = x] = \int_\Theta L(\vartheta, a) w(\vartheta|x) d\mu(\vartheta),$$

and

$$d(x) := \arg\min_a l(x,a).$$

Then d is Bayes decision d_{Bayes}.

Proof.

$$\begin{aligned} r_w(d') &= \int_\Theta R(\vartheta, d') w(\vartheta) d\mu(\vartheta) \\ &= \int_\Theta \left[\int_{\mathcal{X}} L(\vartheta, d'(x)) p(x|\vartheta) d\nu(x) \right] w(\vartheta) d\mu(\vartheta) \\ &= \int_{\mathcal{X}} \left[\int_\Theta L(\vartheta, d'(x)) w(\vartheta|x) d\mu(\vartheta) \right] p(x) d\nu(x) \\ &= \int_{\mathcal{X}} l(x, d'(x)) p(x) d\nu(x) \\ &\geq \int_{\mathcal{X}} l(x, d(x)) p(x) d\nu(x) \\ &= r_w(d). \end{aligned}$$

□

Corollary 10.5.1 *The Bayes decision is*

$$d_{\text{Bayes}}(X) = \arg\min_{a \in \mathcal{A}} l(X, a),$$

where

$$l(x,a) = E(L(\theta,a)|X=x) = \int L(\vartheta,a)w(\vartheta|x)d\mu(\vartheta)$$
$$= \int L(\vartheta,a)g_\vartheta(S(x))w(\vartheta)d\mu(\vartheta)h(x)/p(x).$$

So

$$d_{\text{Bayes}}(X) = \arg\min_{a\in\mathcal{A}} \int L(\vartheta,a)g_\vartheta(S)w(\vartheta)d\mu(\vartheta),$$

which only depends on the sufficient statistic S.

10.5.1 Bayes test revisited

For the testing problem
$H_0: \ \theta=\theta_0$
against the alternative
$H_1: \ \theta=\theta_1$,
with loss function

$$L(\theta_0,a) := a,\ L(\theta_1,a) := 1-a,\ a\in\{0,1\},$$

we have

$$l(x,\phi) = \phi w_0 p_0(x)/p(x) + (1-\phi)w_1 p_1(x)/p(x).$$

Thus,

$$\arg\min_\phi l(\cdot,\phi) = \begin{cases} 1 & \text{if } w_1p_1 > w_0p_0 \\ q & \text{if } w_1p_1 = w_0p_0 \\ 0 & \text{if } w_1p_1 < w_0p_0 \end{cases}.$$

10.5.2 Bayes estimator for quadratic loss

In the next result, we shall use:

Lemma 10.5.2 *Let Z be a real-valued random variable. Then*

$$\arg\min_{a\in\mathbb{R}} E(Z-a)^2 = EZ.$$

Proof.

$$E(Z-a)^2 = \text{var}(Z) + (a-EZ)^2.$$

□

Consider the case $\mathcal{A}=\mathbb{R}$ and $\Theta\subseteq\mathbb{R}$. Let $L(\theta,a) := |\theta-a|^2$. Then

$$d_{\text{Bayes}}(X) = E(\theta|X).$$

For quadratic loss, and for $T=E(\theta|X)$, the Bayes risk of an estimator T' is

$$r_w(T') = E\text{var}(\theta|X) + E(T-T')^2$$

(compare with Lemma 5.2.2). This follows from straightforward calculations:

$$r_w(T') = \int R(\vartheta, T') w(\vartheta) d\mu(\vartheta)$$

$$= ER(\theta, T') = E(\theta - T')^2 = E\biggl[E\biggl((\theta - T')^2 \bigg| X \biggr) \biggr]$$

and, with θ being the random variable,

$$\begin{aligned} E\biggl((\theta - T')^2 \bigg| X \biggr) &= E\biggl((\theta - T)^2 \bigg| X \biggr) + (T - T')^2 \\ &= \mathrm{var}(\theta|X) + (T - T')^2. \end{aligned}$$

10.5.3 Bayes estimator and the maximum a posteriori estimator

Consider again the case $\Theta \subseteq \mathbb{R}$, and $\mathcal{A} = \Theta$, and now with loss function $L(\theta, a) := \mathrm{l}\{|\theta - a| > c\}$ for a given constant $c > 0$. Then

$$l(x, a) = \Pi(|\theta - a| > c | X = x) = \int_{|\vartheta - a| > c} w(\vartheta|x) d\vartheta.$$

We note that for $c \to 0$

$$\frac{1 - l(x, a)}{2c} = \frac{\Pi(|\theta - a| \le c | X = x)}{2c} \approx w(a|x) = p(x|a) \frac{w(a)}{p(x)}.$$

Thus, for c small, Bayes rule is approximately $d_0(x) := \arg\max_{a \in \Theta} p(x|a) w(a)$. The estimator $d_0(X)$ is called the *maximum a posteriori estimator* (MAP). If w is the uniform density on Θ (which only exists if Θ is bounded), then $d_0(X)$ is the maximum likelihood estimator.

10.5.4 Three worked-out examples

In many examples, it saves a lot of work not to write out complete expressions for the posterior $w(\vartheta|x)$. The main interest (at first) is how it depends on ϑ and all the other expressions can (at least theoretically, it may be difficult computationally) be found later by using that a density integrates to one. We therefore recall the symbol $\propto$ (see Section 4.9). For example, we may write

$$\begin{aligned} w(\vartheta|x) &= p(x|\vartheta) w(\vartheta) / p(x) \\ &\propto p(x|\vartheta) w(\vartheta) \end{aligned}$$

because the marginal density $p(x)$ does not depend on ϑ.

As we will see, in the following three examples the posterior is in the same family as the prior. We call them conjugate priors for the distribution concerned.

Example 10.5.1 Poisson with Gamma prior
Suppose that given θ, X has Poisson distribution with parameter θ, and that θ has the Gamma(k, λ)*-distribution. The density of θ is then*

$$w(\vartheta) = \lambda^k \vartheta^{k-1} \mathrm{e}^{-\lambda\vartheta}/\Gamma(k),$$

where

$$\Gamma(k) = \int_0^\infty \mathrm{e}^{-z} z^{k-1} dz.$$

The Gamma(k, λ) *distribution has mean*

$$E\theta = \int_0^\infty \vartheta w(\vartheta) d\vartheta = \frac{k}{\lambda}.$$

The a posteriori density is then

$$\begin{aligned} w(\vartheta|x) &= p(x|\vartheta)\frac{w(\vartheta)}{p(x)} \\ &= \mathrm{e}^{-\vartheta}\frac{\vartheta^x}{x!}\frac{\lambda^k \vartheta^{k-1}\mathrm{e}^{-\lambda\vartheta}/\Gamma(k)}{p(x)} \\ &= \mathrm{e}^{-\vartheta(1+\lambda)}\vartheta^{k+x-1} c(x, k, \lambda), \end{aligned}$$

where $c(x, k, \lambda)$ is such that

$$\int w(\vartheta|x) d\vartheta = 1.$$

With the $\propto$ notation

$$\begin{aligned} w(\vartheta|x) &\propto p(x|\vartheta) w(\vartheta) \\ &\propto \mathrm{e}^{-\vartheta(1+\lambda)}\vartheta^{k+x-1}. \end{aligned}$$

You see this saves a lot of writing. We recognize $w(\vartheta|x)$ as the density of the Gamma$(k + x, 1 + \lambda)$*-distribution. Bayes estimator with quadratic loss is thus*

$$E(\theta|X) = \frac{k + X}{1 + \lambda}.$$

The maximum a posteriori estimator is

$$\frac{k + X - 1}{1 + \lambda}.$$

Example 10.5.2 Binomial distibution with Beta prior
Suppose given θ, X has the Binomial(n, θ)*-distribution, and that θ is uniformly distributed on* $[0, 1]$*. Then*

$$w(\vartheta|x) = \binom{n}{x}\vartheta^x (1 - \vartheta)^{n-x}/p(x).$$

This is the density of the $\mathrm{Beta}(x+1, n-x+1)$*-distribution. Thus, with quadratic loss, Bayes estimator is*

$$E(\theta|X) = \frac{X+1}{n+2}.$$

More generally, suppose that X *is* $\mathrm{binomial}(n, \theta)$ *and that* θ *has the* $\mathrm{Beta}(r, s)$*-prior*

$$w(\vartheta) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\vartheta^{r-1}(1-\vartheta)^{s-1},\ 0 < \vartheta < 1.$$

Here r *and* s *are given positive numbers. The prior expectation is*

$$E\theta = \frac{r}{r+s}.$$

Bayes estimator under quadratic loss is the posterior expectation

$$E(\theta|X) = \frac{X+r}{n+r+s}.$$

Example 10.5.3 Normal distribution with normal prior
Let $X \sim \mathcal{N}(\theta, 1)$ *and* $\theta \sim \mathcal{N}(c, \tau^2)$ *for some* c *and* τ^2*. We have*

$$\begin{aligned} w(\vartheta|x) &= \frac{p(x|\vartheta)w(\vartheta)}{p(x)} \\ &\propto \phi(x-\vartheta)\phi\left(\frac{\vartheta-c}{\tau}\right) \\ &\propto \exp\left[-\frac{1}{2}\left\{(x-\vartheta)^2 + \frac{(\vartheta-c)^2}{\tau^2}\right\}\right] \\ &\propto \exp\left[-\frac{1}{2}\left\{\vartheta - \frac{\tau^2 x + c}{\tau^2+1}\right\}^2 \frac{1+\tau^2}{\tau^2}\right]. \end{aligned}$$

We conclude that Bayes estimator for quadratic loss is

$$T_{\mathrm{Bayes}} = E(\theta|X) = \frac{\tau^2 X + c}{\tau^2+1}.$$

10.6 Discussion of Bayesian approach

A main objection against the Bayesian approach is that it is generally *subjective*. The final estimator depends strongly on the choice of the prior distribution. On the other hand, Bayesian methods are very powerful and often quite natural. The prior may be inspired by or estimated from previous data sets, in which case the above subjectivity problem becomes less pregnant. Furthermore, in complicated models with many unknown parameters, Bayesian methods are a welcome tool for developing sensible algorithms.

Credibility sets. A (frequentist) confidence set for a parameter of interest can be hard to find, and is also less easy to explain to "non-experts". The Bayesian version of a confidence set is called a *credibility set*, which generally is seen as an intuitively much clearer concept. For example, in the case of a real-valued parameter θ, a $(1-\alpha)$-credibility interval is defined as

$$I := [\hat{\theta}_L(X), \hat{\theta}_R(X)],$$

where the endpoints $\hat{\theta}_L$ and $\hat{\theta}_R$ are chosen in such a way that

$$\int_{\hat{\theta}_L(X)}^{\hat{\theta}_R(X)} w(\vartheta|X)d\vartheta = (1-\alpha).$$

Thus, it is the set which has posterior probability $(1-\alpha)$. A $(1-\alpha)$-credibility set is generally not a $(1-\alpha)$-confidence set, i.e., from a frequentist point of view, its properties are not always clear.

Pragmatic point of view. The Bayesian approach is fruitful for the construction of estimators. One can then proceed by studying the frequentist properties of the Bayesian procedure. For example, in the Binomial(n, θ)-model with a uniform prior on θ, the Bayes estimator is

$$\hat{\theta}_{\text{Bayes}}(X) = \frac{X+1}{n+2}.$$

Given this estimator, one can "forget" we obtained it by Bayesian arguments, and study for example its (frequentist) mean square error.

Complexity regularization. Here is a "toy" example, where a Bayesian method helps constructing a useful procedure. Let $X_1, \ldots, X_n$ be independent random variables, where X_i is $\mathcal{N}(\theta_i, 1)$- distributed. The n parameters θ_i are all unknown. Thus, there are as many observations as unknowns, a situation where *complexity regularization* is needed. Complexity regularization (see Chapter 16) means that in principle, one allows for any parameter value, but that one pays a price for choosing "complex" values. What "complexity" means depends on the situation at hand. We consider in this example the situation where complexity is the opposite of *sparsity*, where the *sparseness* of a vector ϑ is defined as its number of non-zero entries. Consider the estimator

$$\hat{\theta} := \arg\min_{\vartheta \in \mathbb{R}^n} \sum_{i=1}^n (X_i - \vartheta_i)^2 + 2\lambda \sum_{i=1}^n |\vartheta_i|,$$

where $\lambda > 0$ is a regularization parameter. Note that when $\lambda = 0$, one has $\hat{\theta}_i = X_i$ for all i, whereas on the other extreme, when $\lambda = \infty$, one has $\hat{\theta} \equiv 0$. The larger λ, the more sparse the estimator will be. In fact, it is easy to verify that for $i = 1, \ldots, n$,

$$\hat{\theta}_i = \begin{cases} X_i - \lambda & X_i > \lambda \\ 0 & |X_i| \le \lambda \\ X_i + \lambda & X_i < -\lambda \end{cases}.$$

This is called the *soft thresholding* estimator. The procedure corresponds to Bayesian maximum a posteriori estimation, with double-exponential (also called Laplace) prior . Indeed, suppose that the prior is $\theta_1, \ldots, \theta_n$ i.i.d. with density

$$w(z) = \frac{1}{\tau\sqrt{2}} \exp\left[-\frac{\sqrt{2}|z|}{\tau}\right], \ z \in \mathbb{R},$$

where $\tau > 0$ is the prior scale parameter (τ^2 is the variance of this distribution). Given $X_1, \ldots, X_n$, the posterior distribution of the vector θ is then

$$w(\vartheta | X_1, \ldots, X_n) \propto$$
$$(2\pi)^{-n/2} \exp\left[- \frac{\sum_{i=1}^n (X_i - \vartheta_i)^2}{2} \right] \times (2\pi\tau)^{-n/2} \exp\left[- \frac{\sqrt{2} \sum_{i=1}^n |\vartheta_i|}{\tau} \right].$$

Thus, $\hat{\theta}$ with regularization parameter $\lambda = \sqrt{2}/\tau$ is the maximum a posteriori estimator.

Bayesian methods as theoretical tool. In Chapter 11 we will illustrate the fact that Bayesian methods can be exploited as a tool for proving for example frequentist lower bounds. We will see for instance that the Bayesian estimator with constant risk is also the minimax estimator. The idea in such results is to look for "worst possible priors".

10.7 Integrating parameters out ⋆

Striving at flexible prior distributions one can model them depending on another "hyper-parameter", say τ, i.e., in formula

$$w(\vartheta) := w(\vartheta | \tau).$$

Keeping τ fixed and integrating ϑ out, the density of X is then

$$\tilde{p}(x|\tau) := \int p(x|\vartheta) w(\vartheta|\tau) d\mu(\vartheta).$$

One can proceed by estimating τ, using for instance maximum likelihood (this is generally computationally quite hard), or the methods of moments. One then obtains a prior $w(\vartheta|\hat{\tau})$ with estimated parameter $\hat{\tau}$. The prior is thus based on the data. The whole procedure is called *empirical Bayes.*

Example 10.7.1 Poisson with Gamma prior with hyperparameters *Suppose $X_1, \ldots, X_n$ are independent and X_i has a* Poisson(θ_i)*-distribution, $i = 1, \ldots, n$. Assume moreover that $\theta_1, \ldots, \theta_n$ are i.i.d. with* Gamma(k, λ)*-distribution, i.e., each has prior density*

$$w(z|k, \lambda) = \mathrm{e}^{-\lambda z} z^{k-1} \lambda^k / \Gamma(k), \ z > 0.$$

Both k and λ are considered as hyper-parameters. Then the density of $X_1, \ldots, X_n$ is

$$\tilde{\mathbf{p}}(x_1, \ldots, x_n | k, \lambda)$$
$$\propto \int \left(\mathrm{e}^{-\sum_{i=1}^n \vartheta_i} \prod_{i=1}^n \vartheta_i^{x_i} \mathrm{e}^{-\lambda \sum_{i=1}^n \vartheta_i} \prod_{i=1}^n \vartheta_i^{k-1} \frac{\lambda^k}{\Gamma(k)} \right) d\vartheta_1 \cdots d\vartheta_n$$
$$= \prod_{i=1}^n \frac{\Gamma(x_i + k)}{\Gamma(k)} p^k (1-p)^{x_i + k - 1},$$

where $p := \lambda/(1+\lambda)$. *Thus, under* $\tilde{\mathbf{p}}(\cdot|k,\lambda)$, *the observations* $X_1, \ldots, X_n$ *are independent and* X_i *has a negative binomial distribution with parameters* k *and* p *(check the formula for the negative binomial distribution, see e.g. Example 2.4.1. The mean and variance of the negative binomial distribution can be calculated directly or looked up in a textbook. We then find (for* $i = 1, \ldots, n$*),*

$$E(X_i|k,\lambda) = \frac{k(1-p)}{p} = \frac{k}{\lambda}$$

and

$$\text{var}(X_i|k,\lambda) = \frac{k(1-p)}{p^2} = \frac{k(1+\lambda)}{\lambda^2}.$$

We use the method of moments to estimate k *and* λ. *Let* $\bar{X}_n$ *be the sample mean and* $S_n^2 := \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ *be the sample variance. We solve*

$$\frac{\hat{k}}{\hat{\lambda}} = \bar{X}_n, \ \frac{\hat{k}(1+\hat{\lambda})}{\hat{\lambda}^2} = S_n^2.$$

This yields

$$\hat{k} = \frac{\bar{X}_n^2}{S_n^2 - \bar{X}_n}, \ \hat{\lambda} = \frac{\bar{X}_n}{S_n^2 - \bar{X}_n}.$$

For given k *and* λ, *the Bayes estimator of* θ_i *is given in Example 10.5.1. We now insert the estimated values of* k *and* λ *to get an empirical Bayes estimator*

$$\hat{\theta}_i = \frac{X_i + \hat{k}}{1+\hat{\lambda}} = X_i(1 - \bar{X}_n/S_n^2) + \bar{X}_n^2/S_n^2, \ i = 1, \ldots, n.$$

The MLE of θ_i *is* X_i *itself* $(i = 1, \ldots, n)$. *We see that the empirical Bayes estimator uses all observations to estimate a particular* θ_i. *The empirical Bayes estimator* $\hat{\theta}_i$ *is a convex combination* $(1-\alpha)X_i + \alpha\bar{X}_n$ *of* X_i *and* $\bar{X}_n$, *with* $\alpha = \bar{X}_n/S_n^2$ *generally close to one if the pooled sample has mean and variance approximately equal, i.e., if the pooled sample is "Poisson-like".*

Chapter 11

Proving admissibility and minimaxity

Bayes estimators are quite useful, also for obdurate frequentists. They can be used to construct estimators that are minimax (admissible), or for verification of minimaxity (admissibility).

Let us first recall the definitions. Let $X \in \mathcal{X}$ have distribution P_θ, $\theta \in \Theta$. Let $T = T(X)$ be a statistic (estimator, decision), $L(\theta, a)$ be a loss function, and $R(\theta, T) := E_\theta L(\theta, T(X))$ be the risk of T.

∘ T is *minimax* if $\forall\ T'$ $\sup_\theta R(\theta, T) \le \sup_\theta R(\theta, T')$.

∘ T is *inadmissible* if $\exists\ T'$: $\{\forall\ \theta\ R(\theta, T') \le R(\theta, T)$ and $\exists\ \theta\ R(\theta, T') < R(\theta, T)\}$.

∘ T is *Bayes* (for the prior density w on Θ) if $\forall\ T'$, $r_w(T) \le r_w(T')$.

Recall also that Bayes risk for w is

$$r_w(T) = \int R(\vartheta, T) w(\vartheta) d\mu(\vartheta).$$

Whenever we say that a statistic T is Bayes, without referring to an explicit prior on Θ, we mean that there exists a prior for which T is Bayes. Of course, if the risk $R(\theta, T) = R(T)$ does not depend on θ, then Bayes risk of T does not depend on the prior.

Especially in cases where one wants to use the uniform distribution as prior, but cannot do so because Θ is not bounded, the notion *extended* Bayes is useful.

Definition 11.0.1 *A statistic T is called* extended Bayes *if there exists a sequence of prior densities $\{w_m\}_{m=1}^\infty$ (w.r.t. dominating measures that are allowed to depend on m), such that $r_{w_m}(T) - \inf_{T'} r_{w_m}(T') \to 0$ as $m \to \infty$.*

11.1 Minimaxity

Lemma 11.1.1 *Suppose T is a statistic with risk $R(\theta, T) = R(T)$ not depending on θ. Then*
(i) T admissible $\Rightarrow T$ minimax,
(ii) T Bayes $\Rightarrow T$ minimax,
and in fact more generally,
(iii) T extended Bayes $\Rightarrow T$ minimax.

Proof.
(i) T is admissible, so for all T', either there is a θ with $R(\theta, T') > R(T)$, or $R(\theta, T') \geq R(T)$ for all θ. Hence $\sup_\theta R(\theta, T') \geq R(T)$.
(ii) Since Bayes implies extended Bayes, this follows from (iii). We nevertheless present a separate proof, as it is somewhat simpler than (iii).
Note first that for any T',

$$\begin{aligned} r_w(T') &= \int R(\vartheta, T') w(\vartheta) d\mu(\theta) && (11.1) \\ &\leq \int \sup_\vartheta R(\vartheta, T') w(\vartheta) d\mu(\theta) && (11.2) \\ &= \sup_\vartheta R(\vartheta, T'), && (11.3) \end{aligned}$$

that is, Bayes risk is always bounded by the supremum risk. Suppose now that T' is a statistic with $\sup_\theta R(\theta, T') < R(T)$. Then

$$r_w(T') \leq \sup_\vartheta R(\vartheta, T') < R(T) = r_w(T),$$

which is in contradiction with the assumption that T is Bayes.
(iii) Suppose for simplicity that a Bayes decision T_m for the prior w_m exists, for all m, i.e.

$$r_{w_m}(T_m) = \inf_{T'} r_{w_m}(T'), \ m = 1, 2, \ldots.$$

By assumption, for all $\epsilon > 0$, there exists an m sufficiently large, such that

$$R(T) = r_{w_m}(T) \leq r_{w_m}(T_m) + \epsilon \leq r_{w_m}(T') + \epsilon \leq \sup_\theta R(\theta, T') + \epsilon,$$

because, as we have seen in (11.1), the Bayes risk is bounded by supremum risk. Since ϵ can be chosen arbitrary small, this proves (iii). □

Example 11.1.1 Minimax estimator for Binomial distribution
Consider a Binomial(n, θ) *random variable X. Let the prior on $\theta \in (0, 1)$ be the* Beta(r, s) *distribution. Then Bayes estimator for quadratic loss is*

$$T = \frac{X + r}{n + r + s}$$

(see Example 10.5.2). Its risk is

$$R(\theta, T) \ = \ E_\theta (T - \theta)^2$$

$$
\begin{aligned}
&= \mathrm{var}_\theta(T) + \mathrm{bias}_\theta^2(T)\\
&= \frac{n\theta(1-\theta)}{(n+r+s)^2} + \left[\frac{n\theta + r}{n+r+s} - \frac{(n+r+s)\theta}{n+r+s}\right]^2\\
&= \frac{[(r+s)^2 - n]\theta^2 + [n - 2r(r+s)]\theta + r^2}{(n+r+s)^2}.
\end{aligned}
$$

This can only be constant in θ if the coefficients in front of θ^2 and θ are zero:

$$(r+s)^2 - n = 0, \; n - 2r(r+s) = 0.$$

Solving for r and s gives

$$r = s = \sqrt{n}/2.$$

Plugging these values back in the estimator T gives

$$T = \frac{X + \sqrt{n}/2}{n + \sqrt{n}}$$

is minimax. The minimax risk is

$$R(T) = \frac{1}{4(\sqrt{n}+1)^2}.$$

We can compare this with the supremum risk of the unbiased estimator X/n:

$$\sup_\theta R(\theta, X/n) = \sup_\theta \frac{\theta(1-\theta)}{n} = \frac{1}{4n}.$$

So for large n, this does not differ much from the minimax risk.

11.1.1 Minimaxity of the Pitman estimator ⋆

We consider again the Pitman estimator (see Lemma 9.1.2)

$$T^* = \frac{\int z\mathbf{p}_0(X_1 - z, \ldots, X_n - z)dz}{\int \mathbf{p}_0(X_1 - z, \ldots, X_n - z)dz}.$$

Lemma 11.1.2 *T^* is extended Bayes (for quadratic loss).*

Proof. Let w_m be (the density of) the uniform distribution on the interval $[-m, m]$:

$$w_m = \mathrm{l}_{[-m,m]}/2m.$$

The posterior density is then

$$w_m(\vartheta|x) = \frac{p_0(x-\vartheta)\mathrm{l}_{[-m,m]}(\vartheta)}{\int_{-m}^m p_0(x-\vartheta)d\vartheta}.$$

Bayes estimator is thus

$$T_m = \frac{\int_{-m}^m \vartheta p_0(x-\vartheta)d\vartheta}{\int_{-m}^m p_0(x-\vartheta)d\vartheta}.$$

We now compute $R(\theta, T_m) = E_\theta(T_m - \theta)^2$. Let

$$T_{a,b}(x) := \frac{\int_a^b z p_0(x-z)dz}{\int_a^b p_0(x-z)dz}.$$

Then for all x, $T_{a,b}(x) \to T(x)$ as $a \to -\infty$ and $b \to \infty$. One can easily verify that also

$$\lim_{a\to -\infty,\ b\to\infty} E_0 T_{a,b}^2(X) \to E_0 T^2(X).$$

(Note that, for any prior w, $E_0 T^2(X)$ is the Bayes risk $r_w(T)$ since the risk $R(\theta, T) = E_0 T^2(X)$ does not depend on θ.) Moreover

$$T_{a,b}(X) - \theta = \frac{\int_a^b (z-\theta) p_0(X-z)dz}{\int_a^b p_0(x-z)dz} = \frac{\int_{a-\theta}^{b-\theta} v p_0(X-\theta-v)dv}{\int_{a-\theta}^{b-\theta} p_0(X-\theta-v)dv}.$$

It follows that

$$E_\theta(T_{a,b}(X)-\theta)^2 = E_0 T_{a-\theta,b-\theta}^2(X).$$

Hence,

$$R(\theta, T_m) = E_0 T_{-m-\theta,m-\theta}^2(X).$$

The Bayes risk is

$$r_{w_m}(T_m) = E_{\theta\sim w_m} R(\theta, T_m) = \frac{1}{2m}\int_{-m}^{m} E_0 T_{-m-\vartheta,m-\vartheta}^2(X) d\vartheta.$$

Hence, for any $0 < \epsilon < 1$, we have

$$\begin{aligned} r_{w_m}(T_m) &\geq \inf_{|\vartheta|\leq m(1-\epsilon)} (1-\epsilon) E_0 T_{-m-\vartheta,m-\vartheta}^2(X) \\ &\geq \inf_{a\leq -m\epsilon,\ b\geq m\epsilon} (1-\epsilon) E_0 T_{a,b}^2(X). \end{aligned}$$

It follows that for any $0 < \epsilon < 1$,

$$\liminf_{m\to\infty} r_{w_m}(T_m) \geq \liminf_{m\to\infty} \inf_{a\leq -m\epsilon,\ b\geq m\epsilon} (1-\epsilon) E_0 T_{a,b}^2(X) = (1-\epsilon) E_0 T^2(X).$$

Hence we have $r_{w_m}(T_m) \to E_0 T^2(X)$, i.e., $r_{w_m}(T_m) - r_{w_m}(T) \to 0$. □

Corollary 11.1.1 *T^* is minimax (for quadratic loss).*

11.2 Admissibility

In this section, the parameter space is assumed to be an open subset of a topological space, so that we can consider open neighborhoods of members of Θ, and continuous functions on Θ. We moreover restrict ourselves to statistics T with $R(\theta, T) < \infty$.

Lemma 11.2.1 *Suppose that the statistic T is Bayes for the prior density w. Then (i) or (ii) below are sufficient conditions for the admissibility of T.*
(i) The statistic T is the unique Bayes decision (i.e., $r_w(T) = r_w(T')$ implies that $\forall\ \theta$, $T = T'$ P_θ-almost surely),
(ii) For all T', $R(\theta, T')$ is continuous in θ, and moreover, for all open $U \subset \Theta$, the prior probability $\Pi(U) := \int_U w(\vartheta) d\mu(\vartheta)$ of U is strictly positive.

Proof.
(i) Suppose that for some T', $R(\theta, T') \le R(\theta, T)$ for all θ. Then also $r_w(T') \le r_w(T)$. Because T is Bayes, we then must have equality:

$$r_w(T') = r_w(T).$$

So then, $\forall\ \theta$, T' and T are equal P_θ-a.s., and hence, $\forall\ \theta$, $R(\theta, T') = R(\theta, T)$, so that T' can not be strictly better than T.
(ii) Suppose that T is inadmissible. Then, for some T', $R(\theta, T') \le R(\theta, T)$ for all θ, and, for some θ_0, $R(\theta_0, T') < R(\theta_0, T)$. This implies that for some $\epsilon > 0$, and some open neighborhood $U \subset \Theta$ of θ_0, we have

$$R(\vartheta, T') \le R(\vartheta, T) - \epsilon,\ \vartheta \in U.$$

But then

$$\begin{aligned} r_w(T') &= \int_U R(\vartheta, T') w(\vartheta) d\nu(\vartheta) + \int_{U^c} R(\vartheta, T') w(\vartheta) d\nu(\vartheta) \\ &\le \int_U R(\vartheta, T) w(\vartheta) d\nu(\vartheta) - \epsilon \Pi(U) + \int_{U^c} R(\vartheta, T) w(\vartheta) d\nu(\vartheta) \\ &= r_w(T) - \epsilon \Pi(U) < r_w(T). \end{aligned}$$

We thus arrived at a contradiction. □

Lemma 11.2.2 *Suppose that T is extended Bayes, and that for all T', $R(\theta, T')$ is continuous in θ. In fact assume, for all open sets $U \subset \Theta$,*

$$\frac{r_{w_m}(T) - \inf_{T'} r_{w_m}(T')}{\Pi_m(U)} \to 0,$$

as $m \to \infty$. Here $\Pi_m(U) := \int_U w_m(\vartheta) d\mu_m(\vartheta)$ is the probability of U under the prior Π_m. Then T is admissible.

Proof. We start out as in the proof of (ii) in the previous lemma. Suppose that T is inadmissible. Then, for some T', $R(\theta, T') \le R(\theta, T)$ for all θ, and, for some θ_0, $R(\theta_0, T') < R(\theta_0, T)$, so that for some $\epsilon > 0$, and some open neighborhood $U \subset \Theta$ of θ_0, we have

$$R(\vartheta, T') \le R(\vartheta, T) - \epsilon,\ \vartheta \in U.$$

This would give that for all m,

$$r_{w_m}(T') \le r_{w_m}(T) - \epsilon \Pi_m(U).$$

Suppose for simplicity that a Bayes decision T_m for the prior w_m exists, for all m, i.e.

$$r_{w_m}(T_m) = \inf_{T'} r_{w_m}(T'),\ m = 1, 2, \ldots .$$

Then, for all m,

$$r_{w_m}(T_m) \le r_{w_m}(T') \le r_{w_m}(T) - \epsilon \Pi_m(U),$$

or

$$\frac{r_{w_m}(T) - r_{w_m}(T_m)}{\Pi_m(U)} \ge \epsilon > 0,$$

that is, we arrived at a contradiction. □

11.2.1 Admissible estimators for the normal mean

Let X be $\mathcal{N}(\theta, 1)$-distributed, $\theta \in \Theta := \mathbb{R}$ and $R(\theta, T) := E_\theta(T - \theta)^2$ be the quadratic risk. We consider estimators of the form

$$T = aX + b,\ a > 0,\ b \in \mathbb{R}.$$

Lemma *T is admissible if and only if one of the following cases hold*
(i) $a < 1$,
(ii) $a = 1$ and $b = 0$.

Proof.
($\Leftarrow$) (i)
First, we show that T is Bayes for some prior. It turns out that this works with a normal prior, i.e., we take $\theta \sim \mathcal{N}(c, \tau^2)$ for some c and τ^2 to be specified. In Example 10.5.3 we have seen that Bayes estimator is

$$T_{\text{Bayes}} = E(\theta|X) = \frac{\tau^2 X + c}{\tau^2 + 1}.$$

Taking

$$\frac{\tau^2}{\tau^2 + 1} = a,\ \frac{c}{\tau^2 + 1} = b,$$

yields $T = T_{\text{Bayes}}$.
Next, we check (i) in Lemma 11.2.1, i.e. that T is unique. Since in view of the calculations in Subsection 10.5.2

$$r_w(T') = E\text{var}(\theta|X) + E(T - T')^2$$

we conclude that if $r_w(T') = r_w(T)$, then

$$E(T - T')^2 = 0.$$

Here, the expectation is with θ integrated out, i.e., with respect to the measure P with density

$$p(x) = \int p_\vartheta(x) w(\vartheta) d\mu(\vartheta).$$

Now, we can write $X = \theta + \epsilon$, with θ $\mathcal{N}(c, \tau^2)$-distributed, and with ϵ a standard normal random variable independent of θ. So X is $\mathcal{N}(c, \tau^2 + 1)$, that is, P is the $\mathcal{N}(c, \tau^2 + 1)$-distribution. Now, $E(T - T')^2 = 0$ implies $T = T'$ P-a.s.. Since P dominates all P_θ, we conclude that $T = T'$ P_θ-a.s., for all θ. So T is unique, and hence admissible.

(⇐) (ii)

In this case, $T = X$. We use Lemma 11.2.2. Because $R(\theta, T) = 1$ for all θ, also $r_w(T) = 1$ for any prior. Let w_m be the density of the $\mathcal{N}(0, m)$-distribution. As we have seen in Example 10.5.3 and also in the previous part of the proof, the Bayes estimator is

$$T_m = \frac{m}{m+1} X.$$

By the bias-variance decomposition, it has risk

$$R(\theta, T_m) = \frac{m^2}{(m+1)^2} + \left(\frac{m}{m+1} - 1\right)^2 \theta^2 = \frac{m^2}{(m+1)^2} + \frac{\theta^2}{(m+1)^2}.$$

As $E\theta^2 = m$, its Bayes risk is

$$r_{w_m}(T_m) = \frac{m^2}{(m+1)^2} + \frac{m}{(m+1)^2} = \frac{m}{m+1}.$$

It follows that

$$r_{w_m}(T) - r_{w_m}(T_m) = 1 - \frac{m}{m+1} = \frac{1}{m+1}.$$

So T is extended Bayes. But we need to prove the more refined property of Lemma 11.2.2. It is clear that here, we only need to consider open intervals $U = (u, u + h)$, with u and $h > 0$ fixed. We have

$$\begin{aligned} \Pi_m(U) &= \Phi\left(\frac{u+h}{\sqrt{m}}\right) - \Phi\left(\frac{u}{\sqrt{m}}\right) \\ &= \frac{1}{\sqrt{m}} \phi\left(\frac{u}{\sqrt{m}}\right) h + o(1/\sqrt{m}). \end{aligned}$$

For m large,

$$\phi\left(\frac{u}{\sqrt{m}}\right) \approx \phi(0) = \frac{1}{\sqrt{2\pi}} > \frac{1}{4} \text{ (say)},$$

so for m sufficiently large (depending on u)

$$\phi\left(\frac{u}{\sqrt{m}}\right) \geq \frac{1}{4}.$$

Thus, for m sufficiently large (depending on u and h), we have

$$\Pi_m(U) \geq \frac{1}{4\sqrt{m}} h.$$

We conclude that for m sufficiently large

$$\frac{r_{w_m}(T) - r_{w_m}(T_m)}{\Pi_m(U)} \leq \frac{4}{h\sqrt{m}}.$$

As the right hand side converges to zero as $m \to \infty$, this shows that X is admissible.
$(\Rightarrow)$
We now have to show that if (i) or (ii) do not hold, then T is not admissible. This means we have to consider two cases: $a > 1$ and $a = 1$, $b \neq 0$. In the case $a > 1$, we have $R(\theta, aX + b) \geq \text{var}(aX + b) > 1 = R(\theta, X)$, so $aX + b$ is not admissible. When $a = 1$ and $b \neq 0$, it is the bias term that makes $aX + b$ inadmissible:

$$R(\theta, X + b) = 1 + b^2 > 1 = R(\theta, X).$$

. $\square$

11.3 Admissible estimators in exponential families ⋆

Lemma 11.3.1 *Let $\theta \in \Theta = \mathbb{R}$ and $\{P_\theta : \ \theta \in \Theta\}$ be an exponential family in canonical form:*

$$p_\theta(x) = \exp[\theta T(x) - d(\theta)]h(x).$$

Then T is an admissible estimator of $g(\theta) := \dot{d}(\theta)$, under quadratic loss (i.e., under the loss $L(\theta, a) := |a - g(\theta)|^2$).

Proof. Recall that

$$\dot{d}(\theta) = E_\theta T, \ \ddot{d}(\theta) = \text{var}_\theta(T) = I(\theta).$$

(see Section 4.8). Now, let T' be some estimator, with expectation

$$E_\theta T' := q(\theta).$$

the bias of T' is

$$b(\theta) = q(\theta) - g(\theta),$$

or

$$q(\theta) = b(\theta) + g(\theta) = b(\theta) + \dot{d}(\theta).$$

This implies

$$\dot{q}(\theta) = \dot{b}(\theta) + I(\theta).$$

By the Cramer Rao lower bound

$$\begin{aligned} R(\theta, T') &= \text{var}_\theta(T') + b^2(\theta) \\ &\geq \frac{[\dot{q}(\theta)]^2}{I(\theta)} + b^2(\theta) = \frac{[\dot{b}(\theta) + I(\theta)]^2}{I(\theta)} + b^2(\theta). \end{aligned}$$

Suppose now that

$$R(\theta, T') \leq R(\theta, T), \forall \ \theta.$$

Because $R(\theta, T) = I(\theta)$ this implies

$$\frac{[\dot{b}(\theta) + I(\theta)]^2}{I(\theta)} + b^2(\theta) \leq I(\theta),$$

or

$$I(\theta)\{b^2(\theta) + 2\dot{b}(\theta)\} \le -[\dot{b}(\theta)]^2 \le 0.$$

This in turn implies

$$b^2(\theta) + 2\dot{b}(\theta) \le 0,$$

and hence, $b(\theta)$ is decreasing and when $b(\theta) \neq 0$,

$$\frac{\dot{b}(\theta)}{b^2(\theta)} \le -\frac{1}{2},$$

so

$$\frac{d}{d\theta}\left(\frac{1}{b(\theta)}\right) - \frac{1}{2} \ge 0,$$

or

$$\frac{d}{d\theta}\left(\frac{1}{b(\theta)} - \frac{\theta}{2}\right) \ge 0.$$

In other words, $1/b(\theta) - \theta/2$ is an increasing function.

We will now show that this gives a contradiction, implying that $b(\theta) = 0$ for all θ.

Suppose instead $b(\theta_0) < 0$ for some θ_0. Then also $b(\vartheta) < 0$ for all $\vartheta > \theta_0$ since $b(\cdot)$ is decreasing. It follows that

$$\frac{1}{b(\vartheta)} \ge \frac{1}{b(\theta_0)} + \frac{\vartheta - \theta_0}{2} \to \infty, \ \vartheta \to \infty$$

i.e.,

$$b(\vartheta) \to 0, \ \vartheta \to \infty.$$

This is not possible, as $b(\theta)$ is a decreasing function.

Similarly, if $b(\theta_0) > 0$, take $\theta_0 \ge \vartheta \to -\infty$, to find again

$$b(\vartheta) \to 0, \ \vartheta \to -\infty,$$

which is not possible.

We conclude that $b(\theta) = 0$ for all θ, i.e., T' is an unbiased estimator of θ. By the Cramer Rao lower bound, we now conclude

$$R(\theta, T') = \text{var}_\theta(T') \ge R(\theta, T) = I(\theta).$$

□

Example 11.3.1 Admissibility of the sample average for estimating the mean of a normal distribution
Let X be $\mathcal{N}(\theta, 1)$-distributed, with $\theta \in \mathbb{R}$ unknown. Then X is an admissible estimator of θ.

Example 11.3.2 Inadmissibility in the normal distribution with σ^2 unknown
Let X be $\mathcal{N}(0, \sigma^2)$, with $\sigma^2 \in (0, \infty)$ unknown. Its density is

$$\begin{aligned} p_\theta(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{x^2}{2\sigma^2} \right] \\ &= \exp[\theta T(x) - d(\theta)] h(x), \end{aligned}$$

with

$$T(x) = -x^2/2, \ \theta = 1/\sigma^2,$$

$$\begin{aligned} d(\theta) &= (\log \sigma^2)/2 = -(\log \theta)/2, \\ \dot{d}(\theta) &= -\frac{1}{2\theta} = -\frac{\sigma^2}{2}, \\ \ddot{d}(\theta) &= \frac{1}{2\theta^2} = \frac{\sigma^4}{2}. \end{aligned}$$

Observe that $\theta \in \Theta = (0, \infty)$, which is not the whole real line. So Lemma 11.3.1 cannot be applied. We will now show that T is not admissible. Define for all $a > 0$,

$$T_a := -aX^2.$$

so that $T = T_{1/2}$. We have

$$\begin{aligned} R(\theta, T_a) &= \text{var}_\theta(T_a) + \text{bias}_\theta^2(T_a) \\ &= 2a^2\sigma^4 + [a - 1/2]^2 \sigma^4. \end{aligned}$$

Thus, $R(\theta, T_a)$ is minimized at $a = 1/6$ giving

$$R(\theta, T_{1/6}) = \sigma^4/6 < \sigma^4/2 = R(\theta, T).$$

11.4 Inadmissibility in higher-dimensional settings ⋆

Let (for $i = 1, \ldots, p$) $X_i \sim \mathcal{N}(\theta_i, 1)$ and let $X_1, \ldots, X_p$ be independent. The vector $\theta := (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$ is unknown. For an estimator $T = (T_1, \ldots, T_p) \in \mathbb{R}^p$, we define the risk

$$R(\theta, T) := \sum_{i=1}^p E_\theta (T_i - \theta_i)^2.$$

Note that $R(\theta, X) = p$ where $X := (X_1, \ldots, X_p)$. One can moreover show (in a similar way as for the case $p = 1$) that X is minimax, extended Bayes, UMRE and that is reaches the Cramer-Rao lower bound. But for $p > 2$, X is inadmissible. This follows from the lemma below, which shows that X can be improved by Stein's estimator. We use the notation $\|X\|^2 := \sum_{i=1}^p X_i^2$.

Definition 11.4.1 *Let $p > 2$ and let $0 < b < 2(p-2)$ be some constant. Stein's estimator is*

$$T^* := \biggl(1 - \frac{b}{\|X\|^2}\biggr)X.$$

Lemma 11.4.1 *We have*

$$R(\theta, T^*) = p - \biggl[2b(p-2) - b^2\biggr] E_\theta \frac{1}{\|X\|^2}.$$

Proof. We first calculate

$$\begin{aligned} E_\theta (T_i^* - \theta_i)^2 &= E_\theta \biggl[\biggl(1 - \frac{b}{\|X\|^2}\biggr)X_i - \theta_i\biggr]^2 \\ &= E_\theta \biggl[(X_i - \theta_i) - \frac{b}{\|X\|^2} X_i\biggr]^2 \\ &= E_\theta \biggl[(X_i - \theta_i)^2 + b^2 \frac{X_i^2}{\|X\|^4} - 2b \frac{X_i(X_i - \theta_i)}{\|X\|^2}\biggr] \\ &= 1 + b^2 E_\theta \frac{X_i^2}{\|X\|^4} - 2b E_\theta \frac{X_i(X_i - \theta_i)}{\|X\|^2}. \end{aligned}$$

Consider now the expectation in the last term, with $i = 1$ (say):

$$\begin{aligned} E_\theta \frac{X_1(X_1 - \theta_1)}{\|X\|^2} &= \int \frac{x_1(x_1 - \theta_1)}{\|x\|^2} \prod_{i=1}^p \biggl\{\phi(x_i - \theta_i) dx_i\biggr\} \\ &= \int \frac{x_1(x_1 - \theta_1)}{\|x\|^2} \phi(x_1 - \theta_1) dx_1 \prod_{i=2}^p \biggl\{\phi(x_i - \theta_i) dx_i\biggr\} \\ &= -\int \frac{x_1}{\|x\|^2} d\phi(x_1 - \theta_1) \prod_{i=2}^p \biggl\{\phi(x_i - \theta_i) dx_i\biggr\} \\ &= \int \phi(x_1 - \theta_1) d\biggl(\frac{x_1}{\|x\|^2}\biggr) \prod_{i=2}^p \biggl\{\phi(x_i - \theta_i) dx_i\biggr\} \\ &= \int \phi(x_1 - \theta_1) \biggl(\frac{1}{\|x\|^2} - 2\frac{x_1^2}{\|x\|^4}\biggr) dx_1 \prod_{i=2}^p \biggl\{\phi(x_i - \theta_i) dx_i\biggr\} \\ &= \int \biggl(\frac{1}{\|x\|^2} - 2\frac{x_1^2}{\|x\|^4}\biggr) \prod_{i=1}^p \biggl\{\phi(x_i - \theta_i) dx_i\biggr\} \\ &= E_\theta \biggl[\frac{1}{\|X\|^2} - 2\frac{X_1^2}{\|X\|^4}\biggr]. \end{aligned}$$

The same calculation can be done for all other i. Inserting the result in our formula for $E_\theta (T_i^* - \theta_i)^2$ gives

$$\begin{aligned} E_\theta (T_i^* - \theta_i)^2 &= 1 + b^2 E_\theta \frac{X_i^2}{\|X\|^4} - 2b E_\theta \biggl[\frac{1}{\|X\|^2} - 2\frac{X_i^2}{\|X\|^4}\biggr] \\ &= 1 + (b^2 + 4b) E_\theta \frac{X_i^2}{\|X\|^4} - 2b E_\theta \frac{1}{\|X\|^2}. \end{aligned}$$

It follows that

$$\begin{aligned} R(\theta, T^*) &= p + (b^2 + 4b) E_\theta \frac{\sum_{i=1}^p X_i^2}{\|X\|^4} - 2bp E_\theta \frac{1}{\|X\|^2} \\ &= p - \left[2b(p-2) - b^2 \right] E_\theta \frac{1}{\|X\|^2}. \end{aligned}$$

□

We thus have the surprising fact that Stein's estimator of θ_i uses also the observations X_j with $j \neq i$, even though these observations are independent of X_i and have a distribution which does not depend on θ_i.

Note that $[2b(p-2) - b^2]$ is maximized for $b = p - 2$. So the value $b = p - 2$ gives the maximal improvement over X. Stein's estimator is then

$$T^* = \left[1 - \frac{p-2}{\|X\|^2} \right] X.$$

Remark It turns out that Stein's estimator is also inadmissible!

Remark Let $g(\theta) := E_\theta 1/\|X\|^2$. One can show that $g(0) = 1/(p-2)$. Moreover, $g(\theta) \downarrow 0$ as $\|\theta\| \uparrow \infty$, so $R(\theta, T^*) \approx R(\theta, X)$ for $\|\theta\|$ large.

Remark Let us take an empirical Bayesian point of view. Suppose $\theta_1, \ldots, \theta_p$ are i.i.d. with the $\mathcal{N}(0, \tau^2)$-distribution. If τ^2 is known, Bayes estimator is

$$T_{i,\text{Bayes}} = \frac{\tau^2}{1 + \tau^2} X_i, \ i = 1, \ldots, p$$

(see Example 5.2.1). Given θ_i, $X_i \sim \mathcal{N}(\theta_i, 1)$ $(i = 1, \ldots, p)$. So unconditionally, $X_i \sim \mathcal{N}(0, 1 + \tau^2)$ $(i = 1, \ldots, p)$. Thus, unconditionally, $X_1, \ldots, X_p$ are identically distributed, each having the $\mathcal{N}(0, \sigma^2)$-distribution with $\sigma^2 = 1 + \tau^2$. As estimator of the variance σ^2 we may use the the sample version $\hat{\sigma}^2 := \sum_{i=1}^p X_i^2 / p = \|X\|^2 / p$ (we need not center with the sample average as the unconditional mean of the X_i is known to be zero). That is, we estimate τ^2 by

$$\hat{\tau}^2 := \hat{\sigma}^2 - 1 = \|X\|^2 / p - 1.$$

This leads to the empirical Bayes estimator

$$T_{i,\text{emp. Bayes}} := \frac{\hat{\tau}^2}{1 + \hat{\tau}^2} X = \left[1 - \frac{p}{\|X\|^2} \right] X.$$

This shows that when $p > 4$, then Stein's estimator with $b = p$ is an empirical Bayes estimator.

Chapter 12

The linear model

Consider n independent observations $Y_1, \ldots, Y_n$. This time we do not assume that they are identically distributed. Let $X \in \mathbb{R}^{n \times p}$ be a given matrix with (non-random) entries $\{x_{i,j} : \ i = 1, \ldots, n, \ j = 1, \ldots, p\}$. The matrix X is considered as (fixed) input and the vector $Y = (Y_1, \ldots, Y_n)^T$ as (random) output. One also calls the columns of X the co-variables. The matrix X is the *design matrix*. We assume it to be non-random, that is, we consider the case of *fixed design*.

12.1 Definition of the least squares estimator

We aim at predicting Y given X and decide to do this by linear approximation: we look for the best linear approximation of Y_i given $x_{i,1}, \ldots, x_{i,p}$. We measure the fit using the residual sum of squares. This means that we minimize

$$\sum_{i=1}^{n} \biggl(Y_i - a - \sum_{j=1}^{p} x_{i,j} b_j \biggr)^2 .$$

over $a \in \mathbb{R}$ and $b = (b_1, \ldots, b_p)^T \in \mathbb{R}^p$.

To simplify the expressions, we rename the quantities involved as follows. Define for all i, $x_{i,p+1} := 1$ and define $b_{p+1} := a$. Then for all i $a + \sum_{j=1}^{p} x_{i,j} b_j = \sum_{j=1}^{p+1} x_{i,j} b_j$. In other words, if we put in the matrix X a column containing only 1's then we may omit the constant a. Thus, putting the column of only 1's in front and replacing $p+1$ by p, we let

$$X := \begin{pmatrix} 1 & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}$$

Then we minimize

$$\sum_{i=1}^{n} \biggl(Y_i - \sum_{j=1}^{p} x_{i,j} b_j \biggr)^2 .$$

over $b = (b_1, \ldots, b_p)^T \in \mathbb{R}^p$.

Let us denote the Euclidean norm of a vector $v \in \mathbb{R}^n$ by[1]

$$\|v\|_2 := \sqrt{\sum_{i=1}^n v_i^2}.$$

Write

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

Then

$$\sum_{i=1}^n \Bigl(Y_i - \sum_{j=1}^p x_{i,j} b_j \Bigr)^2 = \|Y - Xb\|_2^2.$$

Definition 12.1.1 *Suppose X has rank p. One calls*

$$\hat\beta := \arg\min_{b \in \mathbb{R}^p} \|Y - Xb\|_2^2$$

the least squares estimator.

We say that the least squares $\hat\beta$ is obtained by (linear) *regression* of Y on X.

The distance between Y and the space $\{Xb : \ b \in \mathbb{R}^p\}$ spanned by the columns of X is minimized by projecting Y on this space. In fact, one has

$$\frac{1}{2} \frac{\partial}{\partial b} \|Y - Xb\|_2^2 = -X^T(Y - Xb).$$

It follows that $\hat\beta$ is a solution of the so-called *normal equations*

$$X^T(Y - X\hat\beta) = 0$$

or

$$X^T Y = X^T X \hat\beta.$$

If X has rank p, the matrix $X^T X$ has an inverse $(X^T X)^{-1}$ and we get

$$\hat\beta = (X^T X)^{-1} X^T Y.$$

The projection of Y on $\{Xb : \ b \in \mathbb{R}^p\}$ is

$$\underbrace{X(X^T X)^{-1} X^T}_{\text{projection}} Y.$$

Recall that a projection is a linear map of the form PP^T such that $P^T P = I$. We can write $X(X^T X)^{-1} X^T := PP^T$.[2]

[1] We sometimes omit the subscript "2"

[2] Write the singular value decomposition of X as $X = P\phi Q^T$, where $\phi = \text{diag}(\phi_1, \ldots, \phi_p)$ contains the singular values and where $P^T P = I$ and $Q^T Q = I$.

Example with $p = 1$
For $p = 1$

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Then

$$X^T X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix},$$

$$\begin{aligned} (X^T X)^{-1} &= \left(n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2 \right)^{-1} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \\ &= \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^{-1} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}. \end{aligned}$$

Moreover

$$X^T Y = \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n x_i Y_i \end{pmatrix}.$$

We now let (changing notation: $\hat{\alpha} := \hat{\beta}_1$, $\hat{\beta} := \hat{\beta}_2$)

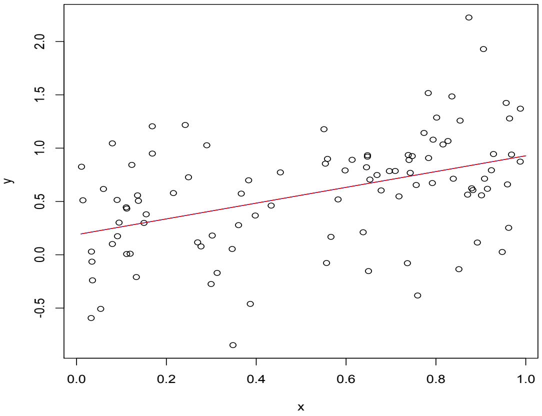
$$\begin{aligned} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} &= (X^T X)^{-1} X^T Y \\ &= \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^{-1} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n x_i Y_i \end{pmatrix} \\ &= \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^{-1} \begin{pmatrix} \sum_{i=1}^n x_i^2 \bar{Y} - \bar{x} \sum_{i=1}^n x_i Y_i \\ -n\bar{x}\bar{Y} + \sum_{i=1}^n x_i Y_i \end{pmatrix} \\ &= \left(\sum_{i=1}^n (x_i - \bar{x})^2 \right)^{-1} \begin{pmatrix} \sum_{i=1}^n (x_i - \bar{x})^2 \bar{Y} - \bar{x}(\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}) \\ \sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y} \end{pmatrix}. \end{aligned}$$

Here we used that $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n\bar{x}^2$. We can moreover write

$$\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}).$$

Thus

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} \bar{Y} - \hat{\beta}\bar{x} \\ \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{pmatrix}.$$



Simulated data with $Y = .3 + .6 \times x + \epsilon$, $\epsilon \sim \mathcal{N}(0, \frac{1}{4})$, $\hat{\alpha} = .19$, $\hat{\beta} = .740$

12.2 Intermezzo: the χ^2 distribution

Let $Z_1, \ldots, Z_p$ be i.i.d. $\mathcal{N}(0,1)$-distributed. Define the p-vector

$$Z := \begin{pmatrix} Z_1 \\ \vdots \\ Z_p \end{pmatrix} .$$

Then Z is $\mathcal{N}(0, I)$-distributed, with I the $p \times p$ identity matrix. The χ^2-distribution with p degrees of freedom is defined as the distribution of

$$\|Z\|_2^2 := \sum_{j=1}^p Z_j^2 .$$

Notation: $\|Z\|_2^2 \sim \chi_p^2$.

For a symmetric positive definite matrix Σ, one can define the square root $\Sigma^{1/2}$ as a symmetric positive definite matrix satisfying

$$\Sigma^{1/2}\Sigma^{1/2} = \Sigma.$$

Its inverse is denoted by $\Sigma^{-1/2}$ (which is the square root of Σ^{-1}). If $Z \in \mathbb{R}^p$ is $\mathcal{N}(0, \Sigma)$-distributed, the transformed vector

$$\tilde{Z} := \Sigma^{-1/2} Z$$

is $\mathcal{N}(0, I)$-distributed. It follows that

$$Z^T \Sigma^{-1} Z = \tilde{Z}^T \tilde{Z} = \|\tilde{Z}\|_2^2 \sim \chi_p^2 .$$

12.3 Distribution of the least squares estimator

Definition 12.3.1 *For $f = EY$ we let $\beta^* := (X^TX)^{-1}X^Tf$ and we call $X\beta^*$ the* best linear approximation *of f.*

Lemma 12.3.1 *Suppose $E\epsilon\epsilon^T = \sigma^2 I$ where $\epsilon := Y - f$. Then*
i) $E\hat\beta = \beta^*$, $\text{Cov}(\hat\beta) = \sigma^2(X^TX)^{-1}$,
ii) $E\|X(\hat\beta - \beta^*)\|_2^2 = \sigma^2 p$,
iii) $E\|X\hat\beta - f\|_2^2 = \underbrace{\sigma^2 p}_{\substack{\text{estimation}\\ \text{error}}} + \underbrace{\|X\beta^* - f\|_2^2}_{\substack{\text{misspecification}\\ \text{error}}}$.

Proof.
i) By straightforward computation

$$\hat\beta - \beta^* = \underbrace{(X^TX)^{-1}X^T}_{:=A}\epsilon.$$

We therefore have

$$E(\hat\beta - \beta^*) = AE\epsilon = 0,$$

and the covariance matrix of $\hat\beta$ is

$$\begin{aligned} \text{Cov}(\hat\beta) &= \text{Cov}(A\epsilon) \\ &= A\underbrace{\text{Cov}(\epsilon)}_{=\sigma^2 I}A^T \\ &= \sigma^2 AA^T = \sigma^2(X^TX)^{-1}. \end{aligned}$$

ii) Define the projection $PP^T := X(X^TX)^{-1}X^T$. Then

$$\|X(\hat\beta - \beta^*)\|_2^2 = \|PP^T\epsilon\|_2^2 =: \sum_{j=1}^p V_j^2,$$

where $V := P^T\epsilon$,

$$EV = P^T E\epsilon = 0,$$

and

$$\text{Cov}(V) = P^T\text{Cov}(\epsilon)P = \sigma^2 I.$$

It follows that

$$E\sum_{j=1}^p V_j^2 = \sum_{j=1}^p EV_j^2 = \sigma^2 p.$$

iii) It holds by Pythagoras' rule for all b

$$\|Xb - f\|_2^2 = \|X(b - \beta^*)\|_2^2 + \|X\beta^* - f\|_2^2$$

since $X\beta^* - f$ is orthogonal to X. □

Lemma 12.3.2 *Suppose $\epsilon := Y - f \sim \mathcal{N}(0, \sigma^2 I)$. Then we have*
i) $\hat\beta - \beta^* \sim \mathcal{N}(0, \sigma^2(X^TX)^{-1})$,
ii) $\frac{\|X(\hat\beta - \beta^*)\|_2^2}{\sigma^2} \sim \chi_p^2$ where χ_p^2 is χ^2-distributed with p degrees of freedom (see Section 12.2 for a definition).

Proof.
i) Since $\hat{\beta}$ is a linear function of the multivariate normal ϵ, the least squares estimator $\hat{\beta}$ is also multivariate normal.
ii) Define the projection $PP^T := X(X^TX)^{-1}X^T$. Then

$$\|X(\hat{\beta}-\beta^*)\|_2^2 = \|PP^T\epsilon\|_2^2 := \sum_{j=1}^p V_j^2.$$

Now $V := P^T\epsilon$ has i.i.d. $\mathcal{N}(0,\sigma)^2$ entries. □

Remark The misspecification error $\|X\beta^* - f\|_2^2$ comes from the possible misspecification of the linear model. That is, f need not be a linear combination of the columns of X. One sometimes also calls $\|X\beta^* - f\|_2^2$ the approximation error. The estimation error is here the variance term $\sigma^2 p$.

Remark More generally, many estimators are approximately normally distributed (for example the sample median) and many test statistics have approximately a χ^2 null-distribution (for example the χ^2 goodness-of-fit statistic). This phenomenon occurs because many models can in a certain sense be approximated by the linear model and many minus log-likelihoods resemble the least squares loss function (see Chapter 14). Understanding the linear model is a first step towards understanding a wide range of more complicated models.

Corollary 12.3.1 *Suppose the linear model is well-specified: for some $\beta \in \mathbb{R}^p$*

$$EY = X\beta.$$

Assume $\epsilon := Y - EY \sim \mathcal{N}(0,\sigma^2)$. where $\sigma^2 := \sigma_0^2$ is known. Then a test for
$H_0: \ \beta = \beta_0$,
is:
reject H_0 when $\|X(\hat{\beta}-\beta^0)\|_2^2/\sigma_0^2 > G_p^{-1}(1-\alpha)$,
where G_p is the distribution function of a χ_p^2-distributed random variable.

Remark When σ^2 is unknown one may estimate it using the estimator

$$\hat{\sigma}^2 = \frac{\|\hat{\epsilon}\|_2^2}{n-p},$$

where $\hat{\epsilon} := Y - X\hat{\beta}$ is the vector of residuals. Under the assumptions of the previous corollary (but now with possibly unknown σ^2) the test statistic $\|X(\hat{\beta}-\beta^0)\|_2^2/(p\hat{\sigma}^2)$ has a so-called F-distribution with p and $n-p$ degrees of freedom.

12.4 Intermezzo: some matrix algebra

Let $z \in \mathbb{R}^p$ be a vector and $B \in \mathbb{R}^{q\times p}$ be a $q \times p$-matrix, $(p \geq q)$ with rank q. Moreover, let $V \in \mathbb{R}^{p\times p}$ be a positive definite $p \times p$-matrix.

Lemma 12.4.1 *We have*

$$\max_{a\in\mathbb{R}^p:\ Ba=0} \{2a^Tz - a^Ta\} = z^Tz - z^TB^T(BB^T)^{-1}Bz.$$

Proof. We use Lagrange multipliers $\lambda \in \mathbb{R}^p$. We have

$$\frac{\partial}{\partial a}\{2a^T z - a^T a + 2a^T B^T \lambda\} = z - a + B^T \lambda.$$

Hence for

$$a_* := \arg \max_{a \in \mathbb{R}^p:\ Ba=0} \{2a^T z - a^T a\},$$

we have

$$z - a_* + B^T \lambda = 0,$$

or

$$a_* = z + B^T \lambda.$$

The restriction $Ba_* = 0$ gives

$$Bz + BB^T \lambda = 0.$$

So

$$\lambda = -(BB^T)^{-1} Bz.$$

Inserting this in the solution a^* gives

$$a_* = z - B^T (BB^T)^{-1} Bz.$$

Now

$$\begin{aligned} a_*^T a_* &= \left(z^T - z^T B^T (BB^T)^{-1} B\right)\left(z - B^T (BB^T)^{-1} Bz\right) \\ &= z^T z - z^T B^T (BB^T)^{-1} Bz. \end{aligned}$$

So

$$2a_*^T z - a_*^T a_* = z^T z - z^T B^T (BB^T)^{-1} Bz.$$

□

Lemma 12.4.2 *We have*

$$\max_{a \in \mathbb{R}^p:\ Ba=0} \{2a^T z - a^T V a\} = z^T V^{-1} z - z^T V^{-1} B^T \left(BV^{-1} B^T\right)^{-1} BV^{-1} z.$$

Proof. Make the transformation $b := V^{1/2} a$, and $y := V^{-1/2} z$, and $C = BV^{-1/2}$. Then

$$\begin{aligned} \max_{a:\ Ba=0} \{2a^T z - a^T V a\} &= \max_{b:\ Cb=0} \{2b^T y - b^T b\} \\ &= y^T y - y^T C^T (CC^T)^{-1} Cy \\ &= z^T V^{-1} z - z^T V^{-1} B^T \left(BV^{-1} B^T\right)^{-1} BV^{-1} z. \end{aligned}$$

□

Corollary 12.4.1 *Let $L(a) := 2a^T z - a^T V a$. The difference between the unrestricted maximum and the restricted maximum of $L(a)$ is*

$$\max_a L(a) - \max_{a:\ Ba=0} L(a) = z^T V^{-1} B^T \left(BV^{-1} B^T\right)^{-1} BV^{-1} z.$$

12.5 Testing a linear hypothesis

In this section we assume the model

$$Y = X\beta + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. We want to test the hypothesis
$H_0 : \ B\beta = 0 \ \ ,$
where $B \in \mathbb{R}^{q\times p}$ is a given $q \times p$ matrix.

Let

$$\hat\beta_0 := \arg \min_{b\in\mathbb{R}^p :\ Bb=0} \|Y - Xb\|_2^2$$

be the least squares estimator under the restriction $B\hat\beta = 0$.

Lemma 12.5.1 *Under H_0*

$$\frac{\|Y - X\hat\beta_0\|_2^2 - \|Y - X\hat\beta\|_2^2}{\sigma^2}$$

has a χ^2_q-distribution.

Proof. Since $\|Y - Xb\|^2 = \|\epsilon\|^2 - 2\epsilon^T X(b-\beta) + (b-\beta)^T X^T X(b-\beta)$ we have under H_0 (write $\tilde b := b - \beta$)

$$\hat\beta_0 - \beta = \arg \max_{\tilde b\in\mathbb{R}^p :\ B\tilde b=0} \biggl\{ 2\epsilon^T X\tilde b - \tilde b^T X^T X \tilde b \biggr\}.$$

Therefore, invoking Corollary 12.4.1

$$\|Y - X\hat\beta_0\|_2^2 - \|Y - X\hat\beta\|_2^2$$

$$\begin{aligned} &= \underbrace{\epsilon^T X(X^TX)^{-1}B^T}_{:=Z^T} \biggl(B(X^TX)^{-1}B^T \biggr)^{-1} \underbrace{B(X^TX)^{-1}X^T\epsilon}_{:=Z} \\ &= Z^T \biggl(B(X^TX)^{-1}B^T \biggr)^{-1} Z. \end{aligned}$$

The q-vector

$$Z := B(X^TX)^{-1}X^T\epsilon$$

has a multivariate normal distribution with mean zero and covariance matrix

$$\sigma^2 \biggl(B(X^TX)^{-1}X^T \biggr) \biggl(B(X^TX)^{-1}X^T \biggr)^T = \sigma^2 B(X^TX)^{-1}B^T.$$

It follows that under H_0

$$\frac{\|Y - X\hat\beta_0\|_2^2 - \|Y - X\hat\beta\|_2^2}{\sigma^2}$$

has a χ^2_q-distribution. □

Chapter 13

Asymptotic theory

In this and subsequent chapters, the observations $X_1, \ldots, X_n$ are considered as the first n of an infinite sequence of i.i.d. random variables $X_1, \ldots, X_n, \ldots$ with values in $\mathcal{X}$ and with distribution P. We say that the X_i are i.i.d. *copies*, of some random variable $X \in \mathcal{X}$ with distribution P. We let $\mathbb{P} = P \times P \times \cdots$ be the distribution of the whole sequence $\{X_i\}_{i=1}^{\infty}$.

The model class for P is

$$\mathcal{P} := \{P_\theta : \ \theta \in \Theta\}.$$

When $P = P_\theta$, we write $\mathbb{P} = \mathbb{P}_\theta = P_\theta \times P_\theta \times \cdots$. The parameter of interest is

$$\gamma := g(\theta) \in \mathbb{R}^p,$$

where $g : \Theta \to \mathbb{R}^p$ is a given function. We let

$$\Gamma := \{g(\theta) : \ \theta \in \Theta\}$$

be the parameter space for γ.

An estimator

$$T_n(X_1, \ldots, X_n)$$

based on the data $X_1, \ldots, X_n$, is some function $T_n(\cdot)$ evaluated at the data $X_1, \ldots, X_n$. We often write shorthand

$$T_n = T_n(X_1, \ldots, X_n).$$

We assume the estimator T_n is defined for all n, i.e., we actually consider a sequence of estimators $\{T_n\}_{n=1}^{\infty}$. We are interested in estimators $T_n \in \Gamma$ of γ.

Remark Under the i.i.d. assumption, it is natural to assume that each T_n is a symmetric function of the data, that is

$$T_n(X_1, \ldots, X_n) = T_n(X_{\pi_1}, \ldots X_{\pi_n})$$

for all permutations π of $\{1, \ldots, n\}$. In that case, one can write T_n in the form $T_n = Q(\hat{P}_n)$, where $\hat{P}_n$ is the empirical distribution (see also Subsection 2.4.1).

13.1 Types of convergence

Definition 13.1.1 *Let $\{Z_n\}_{n=1}^\infty$ and Z be $\mathbb{R}^p$-valued random variables defined on the same probability space[1]. We say that Z_n* converges in probability *to Z if for all $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(\|Z_n - Z\| > \epsilon) = 0.$$

Notation: $Z_n \xrightarrow{\mathbb{P}} Z$.

Remark Chebyshev's inequality can be a tool to prove convergence in probability. It says that for all increasing functions $\psi : [0, \infty) \to [0, \infty)$, one has

$$\mathbb{P}(\|Z_n - Z\| \ge \epsilon) \le \frac{\mathbb{E}\psi(\|Z_n - Z\|)}{\psi(\epsilon)}.$$

Definition 13.1.2 *Let $\{Z_n\}_{n=1}^\infty$ and Z be $\mathbb{R}^p$-valued random variables. We say that Z_n* converges in distribution *to Z, if for all continuous and bounded functions f,*

$$\lim_{n\to\infty} \mathbb{E}f(Z_n) = \mathbb{E}f(Z).$$

Notation: $Z_n \xrightarrow{\mathcal{D}} Z$.

Remark Convergence in probability implies convergence in distribution, but not the other way around.

Example 13.1.1 The central limit theorem (CLT)
Let $X_1, X_2, \ldots$ be i.i.d. real-valued random variables with mean μ and variance σ^2. Let $\bar{X}_n := \sum_{i=1}^n X_i/n$ be the average of the first n. Then by the central limit theorem (CLT),

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

that is

$$\mathbb{P}\left(\sqrt{n}\frac{(\bar{X}_n - \mu)}{\sigma} \le z\right) \to \Phi(z), \ \forall\ z.$$

The following theorem says that for convergence in distribution, one actually can do with one-dimensional random variables. We omit the proof.

Theorem 13.1.1 *(Cramér-Wold device) Let $(\{Z_n\}, Z)$ be a collection of $\mathbb{R}^p$-valued random variables. Then*

$$Z_n \xrightarrow{\mathcal{D}} Z \ \Leftrightarrow\ a^T Z_n \xrightarrow{\mathcal{D}} a^T Z \ \forall\ a \in \mathbb{R}^p.$$

[1] Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and $X : \Omega \to \mathbb{R}^p$ and $Y : \Omega \to \mathbb{R}^q$ be two measurable maps. Then X and Y are called random variables, and they are defined on the same probability space Ω.

Example 13.1.2 Multivariate CLT
Let $X_1, X_2, \ldots$ be i.i.d. copies of a random variable $X = (X^{(1)}, \ldots, X^{(p)})^T$ in $\mathbb{R}^p$. Assume $EX := \mu = (\mu_1, \ldots, \mu_p)^T$ and $\Sigma := \mathrm{Cov}(X) := EXX^T - \mu\mu^T$ exist. Then for all $a \in \mathbb{R}^p$,

$$Ea^TX = a^T\mu, \ \mathrm{var}(a^TX) = a^T\Sigma a.$$

Define

$$\bar{X}_n = (\bar{X}_n^{(1)}, \ldots, \bar{X}_n^{(p)})^T.$$

By the 1-dimensional CLT, for all $a \in \mathbb{R}^p$,

$$\sqrt{n}(a^T\bar{X}_n - a^T\mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, a^T\Sigma a).$$

The Cramér-Wold device therefore gives the p-dimensional CLT

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma).$$

We state the *Portmanteau Theorem*:

Theorem 13.1.2 *Let $(\{Z_n\}, Z)$ be a collection of $\mathbb{R}^p$-valued random variables. Denote the distribution of Z by Q and let $G = Q(Z \le \cdot)$ be its distribution function. The following statements are equivalent:*
(i) $Z_n \xrightarrow{\mathcal{D}} Z$ (i.e., $\mathbb{E}f(Z_n) \to \mathbb{E}f(Z)$ $\forall$ f bounded and continuous).
(ii) $\mathbb{E}f(Z_n) \to \mathbb{E}f(Z)$ $\forall$ f bounded and Lipschitz.[2]
(iii) $\mathbb{E}f(Z_n) \to \mathbb{E}f(Z)$ $\forall$ f bounded and Q-a.s. continuous.
(iv) $\mathbb{P}(Z_n \le z) \to G(z)$ for all G-continuity points z.

13.1.1 Stochastic order symbols

Let $\{Z_n\}$ be a collection of $\mathbb{R}^p$-valued random variables, and let $\{r_n\}$ be strictly positive random variables. We write

$$Z_n = \mathcal{O}_{\mathbf{P}}(1)$$

(Z_n is bounded in probability) if

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{P}(\|Z_n\| > M) = 0.$$

This is also called *uniform tightness* of the sequence $\{Z_n\}$. We write $Z_n = \mathcal{O}_{\mathbf{P}}(r_n)$ if $Z_n/r_n = \mathcal{O}_{\mathbf{P}}(1)$.

If Z_n converges in probability to zero, we write this as

$$Z_n = o_{\mathbf{P}}(1).$$

Moreover, $Z_n = o_{\mathbf{P}}(r_n)$ (Z_n is of small order r_n in probability) if $Z_n/r_n = o_{\mathbf{P}}(1)$.

[2] A real-valued function f on (a subset of) $\mathbb{R}^p$ is *Lipschitz* if for a constant C_L and all $(z, \tilde{z})$ in the domain of f, $|f(z) - f(\tilde{z})| \le C_L\|z - \tilde{z}\|$.

13.1.2 Some implications of convergence

Lemma 13.1.1 *Suppose that Z_n converges in distribution. Then $Z_n = \mathcal{O}_{\mathbf{P}}(1)$.*

Proof. To simplify, take $p = 1$ (Cramér-Wold device). Let $Z_n \xrightarrow{\mathcal{D}} Z$, where Z has distribution function G. Then for every G-continuity point M,

$$\mathbb{P}(Z_n > M) \to 1 - G(M),$$

and for every G-continuity point $-M$,

$$\mathbb{P}(Z_n \le -M) \to G(-M).$$

Since $1 - G(M)$ as well as $G(-M)$ converge to zero as $M \to \infty$, the result follows. □

Example 13.1.3 Averages differ from their means by an order $1/\sqrt{n}$ in probability
Let $X_1, X_2, \ldots$ be i.i.d. copies of a random variable $X \in \mathbb{R}$ with $EX = \mu$ and $\text{var}(X) < \infty$. Then by the CLT,

$$\bar{X}_n - \mu = \mathcal{O}_{\mathbf{P}}\left(1/\sqrt{n}\right).$$

Theorem 13.1.3 *(Slutsky's Theorem) Let $(\{Z_n, A_n\}, Z)$ be a collection of $\mathbb{R}^p$-valued random variables, and $a \in \mathbb{R}^p$ be a vector of constants. Assume that $Z_n \xrightarrow{\mathcal{D}} Z$, $A_n \xrightarrow{\mathbb{P}} a$. Then*

$$A_n^T Z_n \xrightarrow{\mathcal{D}} a^T Z.$$

Proof. Take a bounded Lipschitz function f, say

$$|f| \le C_B, \ |f(z) - f(\tilde{z})| \le C_L \|z - \tilde{z}\|.$$

Then

$$\left| \mathbb{E} f(A_n^T Z_n) - \mathbb{E} f(a^T Z) \right|$$

$$\le \left| \mathbb{E} f(A_n^T Z_n) - \mathbb{E} f(a^T Z_n) \right| + \left| \mathbb{E} f(a^T Z_n) - \mathbb{E} f(a^T Z) \right|.$$

Because the function $z \mapsto f(a^T z)$ is bounded and Lipschitz (with Lipschitz constant $\|a\| C_L$), we know that the second term goes to zero. As for the first term, we argue as follows. Let $\epsilon > 0$ and $M > 0$ be arbitrary. Define $S_n := \{\|Z_n\| \le M, \|A_n - a\| \le \epsilon\}$. Then

$$\begin{aligned} \left| \mathbb{E} f(A_n^T Z_n) - \mathbb{E} f(a^T Z_n) \right| &\le \mathbb{E} \left| f(A_n^T Z_n) - f(a^T Z_n) \right| \\ &= \mathbb{E} \left| f(A_n^T Z_n) - f(a^T Z_n) \right| \mathrm{l}\{S_n\} \\ &+ \mathbb{E} \left| f(A_n^T Z_n) - f(a^T Z_n) \right| \mathrm{l}\{S_n^c\} \end{aligned}$$

$$\leq C_L \epsilon M + 2C_B \mathbb{P}(S_n^c). \tag{13.1}$$

Now

$$\mathbb{P}(S_n^c) \leq \mathbb{P}(\|Z_n\| > M) + \mathbb{P}(\|A_n - a\| > \epsilon).$$

Thus, both terms in (13.1) can be made arbitrary small by appropriately choosing ϵ small and n and M large. □

13.2 Consistency and asymptotic normality

Definition 13.2.1 *A sequence of estimators* $\{T_n\}$ *of* $\gamma = g(\theta)$ *is called* consistent *if*

$$T_n \xrightarrow{\mathbb{P}_\theta} \gamma.$$

Definition 13.2.2 *A sequence of estimators* $\{T_n\}$ *of* $\gamma = g(\theta)$ *is called* asymptotically normal *with asymptotic covariance matrix* V_θ*, if*

$$\sqrt{n}(T_n - \gamma) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, V_\theta).$$

Example 13.2.1 Consistency and asymptotic normality of the average *Suppose* $\mathcal{P}$ *is the location model*

$$\mathcal{P} = \left\{ P_{\mu, F_0}(X \leq \cdot) := F_0(\cdot - \mu), \ \mu \in \mathbb{R}, \ F_0 \in \mathcal{F}_0 \right\}.$$

The parameter is then $\theta = (\mu, F_0)$ *and* $\Theta = \mathbb{R} \times \mathcal{F}_0$*. We assume for all* $F_0 \in \mathcal{F}_0$

$$\int x dF_0(x) = 0, \ \sigma_{F_0}^2 := \int x^2 dF_0(x) < \infty.$$

Let $g(\theta) := \mu$ *and* $T_n := (X_1 + \cdots + X_n)/n =: \bar{X}_n$*. Then, by the law of large numbers,* T_n *is a consistent estimator of* μ *and, by the central limit theorem,*

$$\sqrt{n}(T_n - \mu) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, \sigma_{F_0}^2).$$

13.3 Asymptotic linearity

As we will show, for many estimators asymptotic normality is a consequence of asymptotic linearity, that is, the estimator is approximately an average, to which we can apply the CLT.

Definition 13.3.1 *The sequence of estimators* $\{T_n\}$ *of* $\gamma = g(\theta) \in \mathbb{R}^p$ *is called* asymptotically linear *if for a function* $l_\theta : \mathcal{X} \to \mathbb{R}^p$*, with* $E_\theta l_\theta(X) = 0$ *and*[3]

$$E_\theta l_\theta(X) l_\theta^T(X) =: V_\theta < \infty,$$

it holds that

$$T_n - \gamma = \frac{1}{n} \sum_{i=1}^n l_\theta(X_i) + o_{\mathbf{P}_\theta}(1/\sqrt{n}).$$

[3] In the one-dimensional case ($p = 1$) we thus have $V_\theta = E_\theta l_\theta^2(X)$.

Remark. We then call l_θ the *influence function* of (the sequence) T_n. Roughly speaking, $l_\theta(x)$ approximately measures the influence of an additional observation x (compare with the influence function as defined in Section 8.4).

Example 13.3.1 Influence function of the sample average
Assuming the entries of X have finite variance, the estimator $T_n := \bar{X}_n$ is a linear and hence asymptotically linear estimator of the mean μ, with influence function

$$l_\theta(x) = x - \mu.$$

Example 13.3.2 Influence function of the sample variance
Let X be real-valued, with $E_\theta X =: \mu$, $\mathrm{var}_\theta(X) =: \sigma^2$ and $\kappa =: E_\theta(X-\mu)^4$ (assumed to exist). The sample variance is

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

Let $\hat{\sigma}_n^2$ be the estimator

$$\hat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

One sees that

$$S^2 - \hat{\sigma}^2 = \mathcal{O}_{\mathbf{P}}(1/n) = o_{\mathbf{P}}(1/\sqrt{n}).$$

We rewrite $\hat{\sigma}^2$ as

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

Because by the CLT, $\bar{X}_n - \mu = \mathcal{O}_{\mathbf{P}_\theta}(n^{-1/2})$, we get

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2 + \mathcal{O}_{\mathbf{P}_\theta}(1/n).$$

So asymptotically one does not notice that μ is estimated, and $\hat{\sigma}_n^2$ (and also S^2) is asymptotically linear with influence function

$$l_\theta(x) = (x-\mu)^2 - \sigma^2.$$

The asymptotic variance is

$$V_\theta = E_\theta\biggl((X-\mu)^2 - \sigma^2\biggr)^2 = \kappa - \sigma^4.$$

13.4 The δ-technique

Theorem 13.4.1 *Let $(\{T_n\}, Z)$ be a collection of random variables in $\mathbb{R}^p$, $c \in \mathbb{R}^p$ be a nonrandom vector, and $\{r_n\}$ be a nonrandom sequence of positive numbers, with $r_n \downarrow 0$. Moreover, let $h : \mathbb{R}^p \to \mathbb{R}$ be differentiable at c, with derivative $\dot{h}(c) \in \mathbb{R}^p$. Suppose that*

$$(T_n - c)/r_n \xrightarrow{\mathcal{D}} Z.$$

Then

$$\Big(h(T_n) - h(c)\Big)/r_n \xrightarrow{\mathcal{D}} \dot{h}(c)^T Z$$

and in fact

$$h(T_n) - h(c) = \dot{h}(c)^T (T_n - c) + o_{\mathbf{P}}(r_n).$$

Proof. By Slutsky's Theorem,

$$\dot{h}(c)^T (T_n - c)/r_n \xrightarrow{\mathcal{D}} \dot{h}(c)^T Z.$$

Since $(T_n - c)/r_n$ converges in distribution, we know that $\|T_n - c\|/r_n = \mathcal{O}_{\mathbf{P}}(1)$. Hence, $\|T_n - c\| = \mathcal{O}_{\mathbf{P}}(r_n)$. The result follows now from

$$h(T_n) - h(c) = \dot{h}(c)^T (T_n - c) + o(\|T_n - c\|) = \dot{h}(c)^T (T_n - c) + o_{\mathbf{P}}(r_n).$$

□

Corollary 13.4.1 *Let T_n be an asymptotically normal estimator of $\gamma = g(\theta) \in \mathbb{R}^p$ with asymptotic covariance matrix*

$$V_\theta.$$

Suppose h is differentiable at γ. Then $h(T_n)$ is an asymptotically normal estimator of $h(\gamma)$ with asymptotic variance[4]

$$\dot{h}(\gamma)^T V_\theta \dot{h}(\gamma).$$

If moreover T_n is an asymptotically linear estimator of γ, with influence function

$$l_\theta$$

then $h(T_n)$ is an asymptotically linear estimator of $h(\gamma)$ with influence function

$$\dot{h}(\gamma)^T l_\theta.$$

Example 13.4.1 Asymptotic linear estimator of the parameter of the exponential distribution
Let $X_1, \ldots, X_n$ be a sample from the Exponential(θ) *distribution, with $\theta > 0$.*

[4] For $p = 1$ the asymptotic variance of $h(T_n)$ is thus $\dot{h}^2(\gamma)V_\theta$.

Then $\bar{X}_n$ is a linear estimator of $E_\theta X = 1/\theta := \gamma$, with influence function $l_\theta(x) = x - 1/\theta$. The variance of $\sqrt{n}(T_n - 1/\theta)$ is $1/\theta^2 = \gamma^2$. By Theorem 13.4.1, $1/\bar{X}_n$ is an asymptotically linear estimator of θ. In this case, $h(\gamma) = 1/\gamma$, so that $\dot{h}(\gamma) = -1/\gamma^2$. The influence function of $1/\bar{X}_n$ is thus

$$\dot{h}(\gamma) l_\theta(x) = -\frac{1}{\gamma^2}(x - \gamma) = -\theta^2(x - 1/\theta).$$

The asymptotic variance of $1/\bar{X}_n$ is

$$[\dot{h}(\gamma)]^2 \gamma^2 = \frac{1}{\gamma^2} = \theta^2.$$

So

$$\sqrt{n}\biggl(\frac{1}{\bar{X}_n} - \theta\biggr) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, \theta^2).$$

Example 13.4.2 Two-dimensional asymptotic linearity of the sample average and sample variance

Consider again Example 13.3.2. Let X be real-valued, with $E_\theta X := \mu$, $\mathrm{var}_\theta(X) := \sigma^2$ and $\kappa := E_\theta(X - \mu)^4$ (assumed to exist). Define moreover, for $r = 1, 2, 3, 4$, the r-th moment $\mu_r := E_\theta X^r$. We again consider the estimator

$$\hat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

We have

$$\hat{\sigma}_n^2 = h(T_n),$$

where $T_n = (T_{n,1}, T_{n,2})^T$, with

$$T_{n,1} = \bar{X}_n, \ T_{n,2} = \frac{1}{n}\sum_{i=1}^n X_i^2,$$

and

$$h(t) = t_2 - t_1^2, \ t = (t_1, t_2)^T.$$

The estimator T_n has influence function

$$l_\theta(x) = \begin{pmatrix} x - \mu_1 \\ x^2 - \mu_2 \end{pmatrix}.$$

By the 2-dimensional CLT,

$$\sqrt{n}\biggl(T_n - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}\biggr) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, V_\theta),$$

with

$$V_\theta = \begin{pmatrix} \mu_2 - \mu_1^2 & \mu_3 - \mu_1\mu_2 \\ \mu_3 - \mu_1\mu_2 & \mu_4 - \mu_2^2 \end{pmatrix}.$$

It holds that

$$\dot{h}\biggl(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}\biggr) = \begin{pmatrix} -2\mu_1 \\ 1 \end{pmatrix},$$

so that $\hat{\sigma}_n^2$ has influence function

$$\begin{aligned}\dot{h}^T\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}\right)l_\theta(x) &= \begin{pmatrix}-2\mu_1\\ 1\end{pmatrix}^T\begin{pmatrix}x-\mu_1\\ x^2-\mu_2\end{pmatrix}\\ &= (x-\mu)^2-\sigma^2\end{aligned}$$

(invoking $\mu_1 = \mu$). After some calculations, one finds moreover that

$$\begin{pmatrix}-2\mu_1\\ 1\end{pmatrix}^T V_\theta \begin{pmatrix}-2\mu_1\\ 1\end{pmatrix} = \kappa - \sigma^4,$$

i.e., the δ-method gives the same result as the ad hoc method in Example 13.3.2, as it of course should.

Chapter 14

M-estimators

Recall the maximum likelihood estimator as defined in Section 2.4.3. In this chapter we introduce a general class of estimators of which the MLE is a special case. They are defined as minimizers of some empirical risk function.

Let, for each $\gamma \in \Gamma$, be defined some loss function $\rho_\gamma(X)$. These are for instance constructed as in Chapter 10: we let $L(\theta, a)$ be the loss when taking action a. Then, we fix some decision $d(x)$, and rewrite

$$L(\theta, d(x)) := \rho_\gamma(x),$$

assuming the loss L depends only on θ via the parameter of interest $\gamma = g(\theta)$.

We now require that the *theoretical risk*

$$\mathcal{R}(c) := E_\theta \rho_c(X)$$

is minimized at the value $c = \gamma$ i.e.,

$$\gamma = \arg\min_{c \in \Gamma} E_\theta \rho_c(X) = \arg\min_{c \in \Gamma} \mathcal{R}(c). \tag{14.1}$$

Alternatively, given ρ_c, one may view (14.1) as the *definition* of γ.

If $c \mapsto \rho_c(x)$ is differentiable for all x, we write

$$\psi_c(x) := \dot{\rho}_c(x) := \frac{\partial}{\partial c} \rho_c(x).$$

Then, assuming we may interchange differentiation and taking expectations [1] , we have

$$\dot{\mathcal{R}}(\gamma) = 0,$$

where $\dot{\mathcal{R}}(c) = E_\theta \psi_c(X)$.

Define now the *empirical risk*

$$\hat{\mathcal{R}}_n(c) := \frac{1}{n} \sum_{i=1}^n \rho_c(X_i), \ c \in \Gamma.$$

[1] If $|\partial \rho_c / \partial c| \leq H(\cdot)$ where $E_\theta H(X) < \infty$, then it follows from the dominated convergence theorem that $\partial [E_\theta \rho_c(X)] / \partial c = E_\theta [\partial \rho_c(X) / \partial c]$ or otherwise put, $\dot{\mathcal{R}}(c) = E_\theta \psi(X)$.

Definition 14.0.1 *The* M-estimator $\hat{\gamma}_n$ *of* γ *is defined as*

$$\hat{\gamma}_n := \arg\min_{c \in \Gamma} \frac{1}{n} \sum_{i=1}^n \rho_c(X_i) = \arg\min_{c \in \Gamma} \hat{\mathcal{R}}_n(c).$$

The "M" in "M-estimator" stands for Minimizer (or - take minus signs - Maximizer).

If $\rho_c(x)$ is differentiable in c for all x, we generally can define $\hat{\gamma}_n$ as the solution of putting the derivatives

$$\dot{\hat{\mathcal{R}}}_n(c) = \frac{\partial}{\partial c} \frac{1}{n} \sum_{i=1}^n \rho_c(X_i) = \frac{1}{n} \sum_{i=1}^n \psi_c(X_i)$$

to zero. This is called the Z-estimator. The "Z" in "Z-estimator" stands for Zero.

Definition 14.0.2 *The* Z-estimator $\hat{\gamma}_n$ *of* γ *is defined as a solution of the equations*

$$\dot{\hat{\mathcal{R}}}_n(\hat{\gamma}_n) = 0$$

where $\dot{\hat{\mathcal{R}}}_n(c) = \frac{1}{n} \sum_{i=1}^n \psi_c(X_i)$.

Remark A solution $\hat{\gamma}_n \in \Gamma$ is then assumed to exist.

Example 14.0.1 The least squares estimator
Let $X \in \mathbb{R}$, *and let the parameter of interest be the mean* $\mu = E_\theta X$. *Assume* X *has finite variance* σ^2 *Then*

$$\mu = \arg\min_c E_\theta (X - c)^2,$$

as (recall), by the bias-variance decomposition

$$E_\theta (X - c)^2 = \sigma^2 + (\mu - c)^2.$$

So in this case, we can take

$$\rho_c(x) = (x - c)^2.$$

Clearly

$$\frac{1}{n} \sum_{i=1}^n (X_i - c)^2$$

is minimized at $c = \bar{X}_n := \sum_{i=1}^n X_i / n$. *See also Section 2.3.*

14.1 MLE as special case of M-estimation

Suppose $\Theta \subset \mathbb{R}^p$ and that the densities $p_\theta = dP_\theta / d\nu$ exist w.r.t. some σ-finite measure ν.

Definition 14.1.1 *The quantity*

$$K(\tilde\theta|\theta) = E_\theta \log\left(\frac{p_\theta(X)}{p_{\tilde\theta}(X)}\right)$$

is called the Kullback Leibler information*, or the* relative entropy*.*

Remark Some care has to be taken, not to divide by zero! This can be handled e.g., by assuming that the support $\{x : \ p_\theta(x) > 0\}$ does not depend on θ (see also Condition I in the CRLB of Chapter 5).

Take for all $\tilde\theta \in \Theta$

$$\rho_{\tilde\theta}(x) = -\log p_{\tilde\theta}(x).$$

Then

$$\mathcal{R}(\tilde\theta) = -E_\theta \log p_{\tilde\theta}(X).$$

One easily sees that

$$K(\tilde\theta|\theta) = \mathcal{R}(\tilde\theta) - \mathcal{R}(\theta).$$

Let us restate Lemma 2.4.1 and reprove it in a slightly different manner.

Lemma 14.1.1 *The function $\mathcal{R}(\tilde\theta) = -E_\theta \log p_{\tilde\theta}(X)$ is minimized at $\tilde\theta = \theta$:*

$$\theta = \arg\min_{\tilde\theta} \mathcal{R}(\tilde\theta).$$

Proof. We will show that

$$K(\tilde\theta|\theta) \ge 0.$$

This follows from Jensen's inequality. Since the log-function is concave,

$$\begin{aligned} K(\tilde\theta|\theta) &= -E_\theta \log\left(\frac{p_{\tilde\theta}(X)}{p_\theta(X)}\right) \\ &\ge -\log\left(E_\theta\left(\frac{p_{\tilde\theta}(X)}{p_\theta(X)}\right)\right) \\ &= -\log 1 = 0. \end{aligned}$$

□

With $\rho_{\tilde\theta}(x) = -\log p_{\tilde\theta}(x)$ we find $\psi_{\tilde\theta}(x) := \dot\rho_{\tilde\theta}(x) = -s_{\tilde\theta}(x)$. Recall that s_θ is the score function

$$s_\theta = \dot p_\theta / p_\theta,$$

see Definition 4.7.1. We have seen moreover in Lemma 4.7.1 that $E_\theta s_\theta(X) = 0$. This is just another way to see that θ is a solution of the equation

$$\dot{\mathcal{R}}(\theta) = 0,$$

wehere $\dot{\mathcal{R}}(\tilde\theta) = E_\theta \psi_{\tilde\theta}(X)$ with $\psi_{\tilde\theta} = -\dot p_{\tilde\theta}/p_{\tilde\theta}$.

With $\rho_{\tilde{\theta}}(x) = -\log p_{\tilde{\theta}}(x)$ the M-estimator is the maximum likelihood estimator

$$\begin{aligned} \hat{\theta} &= \arg\min_{\tilde{\theta}\in\Theta} \mathcal{L}_{\mathbf{X}}(\tilde{\theta}) \\ &= \arg\min_{\tilde{\theta}\in\Theta} \frac{1}{n}\sum_{i=1}^{n}\Big(-\log p_{\tilde{\theta}}(X_i)\Big). \end{aligned}$$

14.2 Consistency of M-estimators

Note that γ minimizes a theoretical expectation, whereas the M-estimator $\hat{\gamma}_n$ minimizes the empirical average. Likewise, γ is a solution of putting a theoretical expectation to zero, whereas the Z-estimator $\hat{\gamma}_n$ is the solution of putting an empirical average to zero.

By the law of large numbers, averages converge to expectations. So the M-estimator (Z-estimator) does make sense. However, consistency and further properties are not immediate, because we actually need convergence the averages to expectations over a range of values $c \in \Gamma$ simultaneously. This is the topic of *empirical process theory.*

We will borrow the notation from empirical process theory. For a function $f : \mathcal{X} \to \mathbb{R}$, we let

$$P_\theta f := E_\theta f(X),\ \hat{P}_n f := \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Then, by the law of large numbers, if $P_\theta|f| < \infty$,

$$|(\hat{P}_n - P_\theta)f| \to 0,\ \mathbb{P}_\theta\text{−a.s.}.$$

With this new notation we have

$$\hat{\mathcal{R}}_n(c) = \hat{P}_n\rho_c,\ \mathcal{R}(c) = P\rho_c.$$

Theorem 14.2.1 *Suppose the uniform convergence*

$$\sup_{c\in\Gamma} |\hat{\mathcal{R}}_n(c) - \mathcal{R}(c)| \to 0,\ \mathbb{P}_\theta\text{−a.s.}.$$

Then

$$\mathcal{R}(\hat{\gamma}_n) \to \mathcal{R}(\gamma),\ \mathbb{P}_\theta\text{−a.s.}.$$

Proof. The uniform convergence implies

$$\begin{aligned} 0 &\le P_\theta(\rho_{\hat{\gamma}_n} - \rho_\gamma) \\ &= -(\hat{P}_n - P_\theta)(\rho_{\hat{\gamma}_n} - \rho_\gamma) + \hat{P}_n(\rho_{\hat{\gamma}_n} - \rho_\gamma) \\ &\le -(\hat{P}_n - P_\theta)(\rho_{\hat{\gamma}_n} - \rho_\gamma) \\ &\le |(\hat{P}_n - P_\theta)\rho_{\hat{\gamma}_n}| + |(\hat{P}_n - P_\theta)\rho_\gamma| \\ &\le \sup_{c\in\Gamma} |(\hat{P}_n - P_\theta)\rho_c| + |(\hat{P}_n - P_\theta)\rho_\gamma| \\ &\le 2\sup_{c\in\Gamma} |(\hat{P}_n - P_\theta)\rho_c|. \end{aligned}$$

□

We will need that convergence of to the minimum value also implies convergence of the arg min, i.e., convergence of the location of the minimum. To this end, we present the following definition.

Definition *The minimizer* γ *of* $\mathcal{R}(c)$ *is called* well-separated *if for all* $\epsilon > 0$,

$$\inf\left\{ \mathcal{R}(c) : \ c \in \Gamma, \ \|c - \gamma\| > \epsilon \right\} > \mathcal{R}(\gamma).$$

If γ is well-separated, $\mathcal{R}(\hat{\gamma}_n) \to \mathcal{R}(\gamma)$ $\mathbb{P}_\theta$-a.s.. implies $\hat{\gamma}_n \to \gamma$ $\mathbb{P}_\theta$-a.s..

In the next lemma, we give sufficient conditions for the uniform in c convergence of the empirical risk $\hat{\mathcal{R}}_n(c)$ to the theoretical risk $\mathcal{R}(c)$. Consistency of the M-estimator is then a consequence, as was shown in Theorem 14.2.1. (For consistency the assumption of a compact parameter space Γ can often be omitted if $c \mapsto \rho_c$ is convex. We skip the details.)

Lemma 14.2.1 *Suppose that* Γ *is compact, that* $c \mapsto \rho_c(x)$ *is continuous for all* x*, and that*

$$P_\theta \left(\sup_{c \in \Gamma} |\rho_c| \right) < \infty.$$

Then we have the uniform convergence

$$\sup_{c \in \Gamma} |(\hat{P}_n - P_\theta)\rho_c| \to 0, \ \mathbb{P}_\theta\text{−a.s.}. \tag{14.2}$$

Proof. Define for each $\delta > 0$ and $c \in \Gamma$,

$$w(\cdot, \delta, c) := \sup_{\tilde{c} \in \Gamma : \ \|\tilde{c} - c\| < \delta} |\rho_{\tilde{c}} - \rho_c|.$$

Then for all x, as $\delta \downarrow 0$,

$$w(x, \delta, c) \to 0.$$

So also, by dominated convergence

$$P_\theta w(\cdot, \delta, c) \to 0.$$

Hence, for all $\epsilon > 0$, there exists a δ_c such that

$$P_\theta w(\cdot, \delta_c, c) \le \epsilon.$$

Let

$$B_c := \{\tilde{c} \in \Gamma : \|\tilde{c} - c\| < \delta_c\}.$$

Then $\{B_c : \ c \in \Gamma\}$ is a covering of Γ by open sets. Since Γ is compact, there exists finite sub-covering

$$B_{c_1} \ldots B_{c_N}.$$

For $c \in B_{c_j}$,

$$|\rho_c - \rho_{c_j}| \le w(\cdot, \delta_{c_j}, c_j).$$

It follows that

$$\begin{aligned}\sup_{c\in\Gamma}|(\hat{P}_n - P_\theta)\rho_c| &\le \max_{1\le j\le N}|(\hat{P}_n - P_\theta)\rho_{c_j}| \\ &+ \max_{1\le j\le N}\hat{P}_n w(\cdot,\delta_{c_j},c_j) + \max_{1\le j\le N}P_\theta w(\cdot,\delta_{c_j},c_j) \\ &\to 2\max_{1\le j\le N}P_\theta w(\cdot,\delta_{c_j},c_j) \le 2\epsilon,\ \mathbb{P}_\theta\text{−a.s..}\end{aligned}$$

□

Example 14.2.1 Consistency of the MLE in the logistic location family

The above theorem directly uses the definition of the M-estimator, and does not rely on having an explicit expression available. Here is an example where an explicit expression is indeed not possible. Consider the logistic location family, where the densities are

$$p_\theta(x) = \frac{\mathrm{e}^{x-\theta}}{(1+\mathrm{e}^{x-\theta})^2},\ x\in\mathbb{R},$$

where $\theta\in\Theta\subset\mathbb{R}$ is the location parameter. Take

$$\rho_\theta(x) := -\log p_\theta(x) = \theta - x + 2\log(1+\mathrm{e}^{x-\theta}).$$

Then $\hat{\theta}_n$ is the MLE. It is a solution of

$$\frac{2}{n}\sum_{i=1}^n \frac{\mathrm{e}^{X_i-\hat{\theta}_n}}{1+\mathrm{e}^{X_i-\hat{\theta}_n}} = 1.$$

This expression cannot be made into an explicit expression for $\hat{\theta}_n$. However, we do note the caveat that in order to be able to apply the above consistency theorem, we need to assume that Θ is compact. This problem can be circumvented by using the result below for Z-estimators.

To prove consistency of a Z-estimator of a one-dimensional parameter is relatively easy.

Recall that $\psi_c = \dot{\rho}_c$ and $\dot{\mathcal{R}}(c) = P_\theta\psi_c := E_\theta\psi_c(X)$, $c\in\Gamma$. Recall further that $\dot{\mathcal{R}}(\gamma) = 0$ since γ is defined as the minimizer of $\mathcal{R}(\cdot)$.

Theorem 14.2.2 *Suppose that $\Gamma\subset\mathbb{R}$ and that $\psi_c(x)$ is continuous in c for all x. Assume moreover that*

$$P_\theta|\psi_c| < \infty,\ \forall c,$$

and that $\exists\ \delta > 0$ such that

$$\dot{\mathcal{R}}(c) > 0,\ \gamma < c < \gamma+\delta,$$

$$\dot{\mathcal{R}}(c) < 0,\ \gamma-\delta < c < \gamma.$$

Then for n large enough, $\mathbb{P}_\theta$-a.s., there is a solution $\hat{\gamma}_n$ of $\dot{\hat{\mathcal{R}}}_n(\hat{\gamma}_n) = 0$, and this solution $\hat{\gamma}_n$ is consistent.

Proof. Let $0 < \epsilon < \delta$ be arbitrary. By the law of large numbers, $\mathbb{P}_\theta$-a.s. for n sufficiently large,

$$\dot{\hat{\mathcal{R}}}_n(\gamma + \epsilon) > 0, \ \dot{\hat{\mathcal{R}}}_n(\gamma - \epsilon) < 0.$$

The continuity of $c \mapsto \psi_c$ implies that then $\dot{\hat{\mathcal{R}}}_n(\hat{\gamma}_n) = 0$ for some $|\gamma_n - \gamma| < \epsilon$.

□

14.3 Asymptotic normality of M-estimators

For a function $f : \mathcal{X} \to \mathbb{R}^p$ we let $P_\theta f := E_\theta f(X) \in \mathbb{R}^p$ (whenever it exists). Moreover, we let

$$P_\theta f f^T = E_\theta f(X) f^T(X) \in \mathbb{R}^{p \times p}$$

(whenever is exists). The covariance matrix of the vector $f(X)$ is thus

$$\Sigma := P_\theta f f^T - (P_\theta f)(P_\theta f)^T$$

The CLT says that

$$\sqrt{n}(\hat{P}_n - P_\theta) f \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, \Sigma).$$

With this notation we can translate Definition 13.3.1 as: T_n is an asymptotically linear estimator of γ if

$$T_n - \gamma = \hat{P}_n l_\theta + o_{\mathbf{P}_\theta}(1/\sqrt{n}),$$

with $P_\theta l_\theta = 0$ and $V_\theta := P_\theta l_\theta l_\theta^T < \infty$.

Denote now

$$\nu_n(c) := \sqrt{n}(\hat{P}_n - P_\theta)\psi_c = \sqrt{n}\biggl(\dot{\hat{\mathcal{R}}}_n(c) - \dot{\mathcal{R}}(c)\biggr), \ c \in \Gamma.$$

Definition 14.3.1 *The stochastic process*

$$\{\nu_n(c) : \ c \in \Gamma\}$$

is called the empirical process *indexed by c. The empirical process is called* asymptotically continuous *at γ if for all (possibly random) sequences $\{\gamma_n\}$ in Γ, with $\|\gamma_n - \gamma\| = o_{\mathbf{P}_\theta}(1)$, we have*

$$|\nu_n(\gamma_n) - \nu_n(\gamma)| = o_{\mathbf{P}_\theta}(1).$$

For verifying asymptotic continuity, there are various tools, which involve complexity assumptions on the map $c \mapsto \psi_c$. This goes beyond the scope of these notes. But let us see what asymptotic continuity can bring us.

Recall that $\dot{\mathcal{R}}(c) = P_\theta \psi_c$. We assume that

$$M_\theta := \frac{\partial}{\partial c^T} \dot{\mathcal{R}}(c) \bigg|_{c=\gamma}$$

exists. It is a $p \times p$ matrix. We require it to be of full rank, which amounts to assuming that γ, as a solution to $\dot{\mathcal{R}}(\gamma) = 0$, is well-identified.

Theorem 14.3.1 *Let $\hat{\gamma}_n$ be the Z-estimator of γ. Suppose that $\hat{\gamma}_n$ is a consistent estimator of γ, and that ν_n is asymptotically continuous at γ. Suppose moreover M_θ^{-1} exists, and also*

$$J_\theta := P_\theta \psi_\gamma \psi_\gamma^T.$$

Then $\hat{\gamma}_n$ is asymptotically linear, with influence function

$$l_\theta = -M_\theta^{-1} \psi_\gamma.$$

Proof. By definition,

$$\dot{\hat{\mathcal{R}}}_n(\hat{\gamma}_n) = 0, \ \dot{\mathcal{R}}(\gamma) = 0.$$

So we have

$$\begin{aligned} 0 &= \dot{\hat{\mathcal{R}}}_n(\hat{\gamma}_n) \\ &= \nu_n(\hat{\gamma}_n)/\sqrt{n} + \dot{\mathcal{R}}(\hat{\gamma}_n) \\ &= \nu_n(\hat{\gamma}_n)/\sqrt{n} + \dot{\mathcal{R}}(\hat{\gamma}_n) - \dot{\mathcal{R}}(\gamma) \\ &=: (i) + (ii). \end{aligned}$$

For the first term, we use the asymptotic continuity of ν_n at γ:

$$\begin{aligned} (i) &= \nu_n(\hat{\gamma}_n)/\sqrt{n} \\ &= \nu_n(\gamma)/\sqrt{n} + o_{\mathbf{P}_\theta}(1/\sqrt{n}) \\ &= \dot{\hat{\mathcal{R}}}_n(\gamma) \qquad + o_{\mathbf{P}_\theta}(1/\sqrt{n}). \end{aligned}$$

For the second term, we use the differentiability of $\dot{\mathcal{R}}(c) = P_\theta \psi_c$ at $c = \gamma$:

$$\begin{aligned} (ii) &= \dot{\mathcal{R}}(\hat{\gamma}_n) - \dot{\mathcal{R}}(\gamma) \\ &= M_\theta(\hat{\gamma}_n - \gamma) + o(\|\gamma_n - \gamma\|). \end{aligned}$$

So we arrive at

$$0 = \dot{\hat{\mathcal{R}}}_n(\gamma) + o_{\mathbf{P}_\theta}(1/\sqrt{n}) + M_\theta(\hat{\gamma}_n - \gamma) + o(\|\gamma_n - \gamma\|).$$

Because, by the CLT, $\dot{\hat{\mathcal{R}}}_n(\gamma) = \mathcal{O}_{\mathbf{P}_\theta}(1/\sqrt{n})$, this implies $\|\hat{\gamma}_n - \gamma\| = \mathcal{O}_{\mathbf{P}_\theta}(1/\sqrt{n})$. Hence

$$0 = \dot{\hat{\mathcal{R}}}_n(\gamma) + M_\theta(\hat{\gamma}_n - \gamma) + o_{\mathbf{P}_\theta}(1/\sqrt{n}),$$

or

$$\begin{aligned} M_\theta(\hat{\gamma}_n - \gamma) &= -\dot{\hat{\mathcal{R}}}_n(\gamma) + o_{\mathbf{P}_\theta}(1/\sqrt{n}) \\ &= -\hat{P}_n \psi_\gamma + o_{\mathbf{P}_\theta}(1/\sqrt{n}) \end{aligned}$$

or

$$(\hat{\gamma}_n - \gamma) = -\hat{P}_n M^{-1} \psi_\gamma + o_{\mathbf{P}_\theta}(1/\sqrt{n}).$$

□

Corollary 14.3.1 *Under the conditions of Theorem 14.3.1*

$$\sqrt{n}(\hat{\gamma}_n - \gamma) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, V_\theta),$$

with

$$V_\theta = M_\theta^{-1} J_\theta M_\theta^{-1}.$$

Asymptotic linearity can also be established directly, under rather restrictive assumptions, see Theorem 14.3.2 coming up next. We assume quite a lot of smoothness for the functions ψ_c (namely, derivatives that are Lipschitz), so that asymptotic linearity can be proved by straightforward arguments. We stress however that such smoothness assumptions are by no means necessary.

Theorem 14.3.2 *Let $\hat{\gamma}_n$ be the Z-estimator of γ, and suppose that $\hat{\gamma}_n$ is a consistent estimator of γ. Suppose that, for all c in a neighborhood $\{c \in \Gamma : \|c - \gamma\| < \epsilon\}$, the map $c \mapsto \psi_c(x)$ is differentiable for all x, with derivative*

$$\dot{\psi}_c(x) = \frac{\partial}{\partial c^T} \psi_c(x)$$

(a $p \times p$ matrix). Assume moreover that, for all c and $\tilde{c}$ in a neighborhood of γ, and for all x, we have, in matrix-norm[2]*,*

$$\|\dot{\psi}_c(x) - \dot{\psi}_{\tilde{c}}(x)\| \leq H(x)\|c - \tilde{c}\|,$$

where $H : \mathcal{X} \to \mathbb{R}$ satisfies

$$P_\theta H < \infty.$$

Then

$$M_\theta := \frac{\partial}{\partial c^T} \dot{\mathcal{R}}(c)\Big|_{c=\gamma} = P_\theta \dot{\psi}_\gamma. \tag{14.3}$$

Assuming M_θ^{-1} and $J_\theta := E_\theta \psi_\gamma \psi_\gamma^T$ exist, the influence function of $\hat{\gamma}_n$ is

$$l_\theta = -M_\theta^{-1} \psi_\gamma.$$

Proof. Result (14.3) follows from the dominated convergence theorem.

By the mean value theorem,

$$\begin{aligned} 0 &= \dot{\hat{\mathcal{R}}}_n(\hat{\gamma}) \\ &= \hat{P}_n \psi_{\hat{\gamma}_n} \\ &= \hat{P}_n \psi_\gamma + \hat{P}_n \dot{\psi}_{\tilde{\gamma}_n(\cdot)}(\hat{\gamma}_n - \gamma) \\ &= \dot{\hat{\mathcal{R}}}_n(\gamma) + \hat{P}_n \dot{\psi}_{\tilde{\gamma}_n(\cdot)}(\hat{\gamma}_n - \gamma) \end{aligned}$$

where for all x, $\|\tilde{\gamma}_n(x) - \gamma\| \leq \|\hat{\gamma}_n - \gamma\|$. Thus

$$0 = \dot{\hat{\mathcal{R}}}_n(\gamma) + \hat{P}_n \dot{\psi}_\gamma(\hat{\gamma}_n - \gamma) + \hat{P}_n(\dot{\psi}_{\tilde{\gamma}_n(\cdot)} - \dot{\psi}_\gamma)(\hat{\gamma}_n - \gamma),$$

[2] For a matrix A, $\|A\| := \sup_{v \neq 0} \|Av\|/\|v\|$.

so that

$$\left\| \dot{\hat{\mathcal{R}}}_n(\gamma) + \hat{P}_n \dot{\psi}_\gamma (\hat{\gamma}_n - \gamma) \right\| \le \left(\hat{P}_n H \right) \| \hat{\gamma}_n - \gamma \|^2 = \mathcal{O}_{\mathbf{P}_\theta}(1) \| \hat{\gamma}_n - \gamma \|^2,$$

where in the last inequality, we used $P_\theta H < \infty$. Now, by the law of large numbers,

$$\hat{P}_n \dot{\psi}_\gamma = P_\theta \dot{\psi}_\gamma + o_{\mathbf{P}_\theta}(1) = M_\theta + o_{\mathbf{P}_\theta}(1).$$

Thus

$$\left| \dot{\hat{\mathcal{R}}}_n(\gamma) + M_\theta(\hat{\gamma}_n - \gamma) + o_{\mathbf{P}_\theta}(\| \hat{\gamma}_n - \gamma \|) \right| = \mathcal{O}_{\mathbf{P}_\theta}(\| \hat{\gamma}_n - \gamma \|^2).$$

Because $\dot{\hat{\mathcal{R}}}_n(\gamma) = \mathcal{O}_{\mathbf{P}_\theta}(1/\sqrt{n})$ by the CLT, this ensures that $\| \hat{\gamma}_n - \gamma \| = \mathcal{O}_{\mathbf{P}_\theta}(1/\sqrt{n})$. It follows that

$$\left| \dot{\hat{\mathcal{R}}}_n(\gamma) + M_\theta(\hat{\gamma}_n - \gamma) + o_{\mathbf{P}_\theta}(1/\sqrt{n}) \right| = \mathcal{O}_{\mathbf{P}_\theta}(1/n).$$

Hence

$$M_\theta(\hat{\gamma}_n - \gamma) = -\dot{\hat{\mathcal{R}}}_n(\gamma) + o_{\mathbf{P}_\theta}(1/\sqrt{n})$$

and so

$$\begin{aligned} \hat{\gamma}_n - \gamma &= -M_\theta^{-1} \dot{\hat{\mathcal{R}}}_n(\gamma) + o_{\mathbf{P}_\theta}(1/\sqrt{n}) \\ &= -\hat{P}_n M_\theta^{-1} \psi_\gamma + o_{\mathbf{P}_\theta}(1/\sqrt{n}). \end{aligned}$$

□

Note The asymptotic normality follows again from the asymptotic linearity established in Theorem 14.3.2.

14.4 Asymptotic normality of the MLE

In this section, we show that, under regularity conditions, the MLE is asymptotically normal with asymptotic covariance matrix the inverse of the Fisher-information matrix $I(\theta)$. We use that maximum likelihood estimation is a special case of M-estimation and apply the results of the previous section. In order to do so we need to assume regularity conditions.

Let $\mathcal{P} = \{P_\theta : \ \theta \in \Theta\}$ be dominated by a σ-finite dominating measure ν, and write the densities as $p_\theta = dP_\theta / d\nu$. Suppose that $\Theta \subset \mathbb{R}^p$. Assume that the support of p_θ does not depend on θ (Condition I in Section 5.5). As loss we take minus the log-density:

$$\rho_\theta := -\log p_\theta.$$

The MLE is

$$\hat{\theta}_n := \arg\max_{\tilde{\theta} \in \Theta} \hat{P}_n \log p_{\tilde{\theta}}.$$

We suppose that the score function

$$s_\theta = \frac{\partial}{\partial\theta} \log p_\theta = \frac{\dot{p}_\theta}{p_\theta}$$

exists, and that we may interchange differentiation and integration, so that the score has mean zero:

$$P_\theta s_\theta = \int \dot{p}_\theta d\nu = \frac{\partial}{\partial\theta} \int p_\theta d\nu = \frac{\partial}{\partial\theta} 1 = 0.$$

Recall that the Fisher-information matrix is

$$I(\theta) := P_\theta s_\theta s_\theta^T.$$

Now, it is clear that $\psi_\theta = -s_\theta$, and, assuming derivatives exist and that again we may change the order of differentiation and integration,

$$M_\theta = P_\theta \dot{\psi}_\theta = -P_\theta \dot{s}_\theta,$$

and (see also Lemma 4.7.1)

$$\begin{aligned} P_\theta \dot{s}_\theta &= P_\theta \left(\frac{\ddot{p}_\theta}{p_\theta} - \frac{\dot{p}_\theta}{p_\theta} \frac{\dot{p}_\theta^T}{p_\theta} \right) \\ &= \left(\frac{\partial^2}{\partial\theta\partial\theta^T} 1 \right) - P_\theta s_\theta s_\theta^T \\ &= 0 - I(\theta) = -I(\theta). \end{aligned}$$

Hence, in this case, $M_\theta = -I(\theta)$, and the influence function is

$$l_\theta = I(\theta)^{-1} s_\theta.$$

So the asymptotic covariance matrix of the MLE $\hat{\theta}_n$ is

$$I(\theta)^{-1} \left(P_\theta s_\theta s_\theta^T \right) I(\theta)^{-1} = I(\theta)^{-1}.$$

It follows that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, I^{-1}(\theta)).$$

14.5 Two further examples of M-estimation

In this section we examine the α-quantile and the Huber estimator.

Example 14.5.1 Asymptotic normality of the α-quantile
In this example, the parameter of interest is the α-quantile. We will consider a loss function which does not satisfy regularity conditions, but nevertheless leads to an asymptotically linear, and hence asymptotically normal, estimator.

Let $\mathcal{X} := \mathbb{R}$. The distribution function of X is denoted by F. Let $0 < \alpha < 1$ be given. The α-quantile of F is $\gamma = F^{-1}(\alpha)$ (assumed to exist). We moreover assume that F has density f with respect to Lebesgue measure, and that $f(x) > 0$ in a neighborhood of γ. As loss function we take

$$\rho_c(x) := \rho(x - c),$$

where

$$\rho(x) := (1-\alpha)|x|\mathrm{l}\{x < 0\} + \alpha|x|\mathrm{l}\{x > 0\}.$$

We now first check that for $\mathcal{R}(c) := P_\theta \rho_c$

$$\arg\min_c \mathcal{R}(c) = F^{-1}(\alpha) := \gamma.$$

We have

$$\dot{\rho}(x) = \alpha\mathrm{l}\{x > 0\} - (1-\alpha)\mathrm{l}\{x < 0\}.$$

Note that $\dot{\rho}$ does not exist at $x = 0$. This is one of the irregularities in this example.

It follows that

$$\psi_c(x) = -\alpha\mathrm{l}\{x > c\} + (1-\alpha)\{x < c\}.$$

Hence

$$\dot{\mathcal{R}}(c) = P_\theta \psi_c = -\alpha + F(c)$$

(the fact that ψ_c is not defined at $x = c$ can be shown not to be a problem, roughly because a single point has probability zero, as F is assumed to be continuous). So

$$\dot{\mathcal{R}}(\gamma) = 0, \text{ for } \gamma = F^{-1}(\alpha).$$

We now derive M_θ, which is a scalar in this case:

$$\begin{aligned} M_\theta &= \frac{d}{dc}\dot{\mathcal{R}}(c)\Big|_{c=\gamma} \\ &= \frac{d}{dc}(-\alpha + F(c))\Big|_{c=\gamma} \\ &= f(\gamma) = f(F^{-1}(\alpha)). \end{aligned}$$

The influence function is thus [3]

$$l_\theta(x) = -M_\theta^{-1}\psi_\gamma(x) = \frac{1}{f(\gamma)}\Big\{ -\mathrm{l}\{x < \gamma\} + \alpha \Big\}.$$

[3] Note that in the special case $\alpha = 1/2$ (where γ is the median), this becomes

$$l_\theta(x) = \begin{cases} -\frac{1}{2f(\gamma)} & x < \gamma \\ +\frac{1}{2f(\gamma)} & x > \gamma \end{cases}.$$

We conclude that, for $\hat{\mathcal{R}}_n(c) := \hat{P}_n \rho_c$ *and*

$$\hat{\gamma}_n := \arg\min_c \hat{\mathcal{R}}_n(c),$$

which we write as the sample quantile $\hat{\gamma}_n = \hat{F}_n^{-1}(\alpha)$ *(or an approximation thereof up to order* $o_{\mathbf{P}_\theta}(1/\sqrt{n})$*), one has*

$$\sqrt{n}(\hat{F}_n^{-1}(\alpha) - F^{-1}(\alpha)) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{f^2(F^{-1}(\alpha))}\right).$$

Example 14.5.2 Asymptotic normality of the Huber estimator
In this example, we illustrate that the Huber-estimator is asymptotically linear and hence asymptotically normal. Let again $\mathcal{X} = \mathbb{R}$ *and* F *be the distribution function of* X*. We let the parameter of interest be the a location parameter. The Huber loss function is*

$$\rho_c(x) = \rho(x - c),$$

with

$$\rho(x) = \begin{cases} x^2 & |x| \le k \\ k(2|x| - k) & |x| > k \end{cases}.$$

We define γ *as*

$$\gamma := \arg\min_c P_\theta \rho_c.$$

It holds that

$$\dot{\rho}(x) = \begin{cases} 2x & |x| \le k \\ +2k & x > k \\ -2k & x < -k \end{cases}.$$

Therefore,

$$\psi_c(x) = \begin{cases} -2(x-c) & |x-c| \le k \\ -2k & x - c > k \\ +2k & x - c < -k \end{cases}.$$

One easily derives that

$$\begin{aligned} \dot{\mathcal{R}}(c) := P_\theta \psi_c = & - 2\int_{-k+c}^{k+c} x dF(x) + 2c[F(k+c) - F(-k+c)] \\ & - 2k[1 - F(k+c)] + 2kF(-k+c). \end{aligned}$$

So

$$M_\theta = \frac{d}{dc}\dot{\mathcal{R}}(c)\Big|_{c=\gamma} = 2[F(k+\gamma) - F(-k+\gamma)].$$

The influence function of the Huber estimator is

$$l_\theta(x) = \frac{1}{[F(k+\gamma) - F(-k+\gamma)]} \begin{cases} x - \gamma & |x - \gamma| \le k \\ +k & x - \gamma > k \\ -k & x - \gamma < -k \end{cases}.$$

For $k \to 0$*, this corresponds to the influence function of the median.*

14.6 Asymptotic relative efficiency

In this section, we assume that the parameter of interest is real-valued:

$$\gamma \in \Gamma \subset \mathbb{R}.$$

Definition 14.6.1 *Let $T_{n,1}$ and $T_{n,2}$ be two estimators of γ, that satisfy*

$$\sqrt{n}(T_{n,j} - \gamma) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, V_{\theta,j}), \ j = 1, 2.$$

Then

$$\mathrm{e}_{2:1} := \frac{V_{\theta,1}}{V_{\theta,2}}$$

is called the asymptotic relative efficiency *of $T_{n,2}$ with respect to $T_{n,1}$.*

If $\mathrm{e}_{2:1} > 1$, the estimator $T_{n,2}$ is asymptotically more efficient than $T_{n,1}$. An asymptotic $(1-\alpha)$-confidence interval for γ based on $T_{n,2}$ is then narrower than the one based on $T_{n,1}$.

Example 14.6.1 Asymptotic relative efficiency of sample mean and sample median
Let $\mathcal{X} = \mathbb{R}$, and F be the distribution function of X. Suppose that F is symmetric around the parameter of interest μ. In other words,

$$F(\cdot) = F_0(\cdot - \mu),$$

where F_0 is symmetric around zero. We assume that F_0 has finite variance σ^2, and that is has density f_0 w.r.t. Lebesgue measure, with $f_0(0) > 0$. Take $T_{n,1} := \bar{X}_n$, the sample mean, and $T_{n,2} := \hat{F}_n^{-1}(1/2)$, the sample median. Then $V_{\theta,1} = \sigma^2$ and $V_{\theta,2} = 1/(4f_0^2(0))$ (the latter being derived in Example 14.5.1). So

$$\mathrm{e}_{2:1} = 4\sigma^2 f_0^2(0).$$

Whether the sample mean is the winner, or rather the sample median, depends thus on the distribution F_0. Let us consider three cases.

Case i *Let F_0 be the standard normal distribution, i.e., $F_0 = \Phi$. Then $\sigma^2 = 1$ and $f_0(0) = 1/\sqrt{2\pi}$. Hence*

$$\mathrm{e}_{2:1} = \frac{2}{\pi} \approx 0.64.$$

So $\bar{X}_n$ is the winner. Note that $\bar{X}_n$ is the MLE in this case.

Case ii *Let F_0 be the Laplace distribution, with variance σ^2 equal to one. This distribution has density*

$$f_0(x) = \frac{1}{\sqrt{2}} \exp[-\sqrt{2}|x|], \ x \in \mathbb{R}.$$

So we have $f_0(0) = 1/\sqrt{2}$, and hence

$$\mathrm{e}_{2:1} = 2.$$

Thus, the sample median, which is the MLE for this case, is the winner.

Case iii *Suppose*

$$F_0 = (1-\eta)\Phi + \eta\Phi(\cdot/3).$$

This means that the distribution of X is a mixture, with mixing probabilities $1-\eta$ and η, of two normal distributions, one with unit variance, and one with variance $3^2 = 9$. Otherwise put, associated with X is an unobservable label $Y \in \{0,1\}$. Given $Y = 1$, the random variable X is $\mathcal{N}(\mu, 1)$-distributed. Given $Y = 0$, the random variable X has a $\mathcal{N}(\mu, 3^2)$ distribution. Moreover, $P(Y = 1) = 1 - P(Y = 0) = 1 - \eta$. Hence

$$\sigma^2 := \mathrm{var}(X) = (1-\eta)\mathrm{var}(X|Y=1) + \eta\mathrm{var}(X|Y=0) = (1-\eta) + 9\eta = 1 - 8\eta.$$

It furthermore holds that

$$f_0(0) = (1-\eta)\phi(0) + \frac{\eta}{3}\phi(0) = \frac{1}{\sqrt{2\pi}}\left(1 - \frac{2\eta}{3}\right).$$

It follows that

$$\mathrm{e}_{2:1} = \frac{2}{\pi}\left(1 - \frac{2\eta}{3}\right)^2 (1 + 8\eta).$$

Let us now further compare the results with the α-trimmed mean. Because F is symmetric, it turns out that the α-trimmed mean has the same influence function as the Huber-estimator with $k = F^{-1}(1-\alpha)$:

$$l_\theta(x) = \frac{1}{F_0(k) - F(-k)} \begin{cases} x - \mu, & |x-\mu| \le k \\ +k, & x - \mu > k \\ -k, & x - \mu < -k \end{cases}.$$

(This can be seen from Example 15.3.2 ahead which is not part of the exam). The influence function is used to compute the asymptotic variance $V_{\theta,\alpha}$ of the α-trimmed mean:

$$V_{\theta,\alpha} = \frac{\int_{F_0^{-1}(\alpha)}^{F_0^{-1}(1-\alpha)} x^2 dF_0(x) + 2\alpha(F_0^{-1}(1-\alpha))^2}{(1-2\alpha)^2}.$$

From this, we then calculate the asymptotic relative efficiency of the α-trimmed mean w.r.t. the mean. Note that the median is the limiting case with $\alpha \to 1/2$.

Table: Asymptotic relative efficiency of α-trimmed mean over mean

	α = *0.05*	*0.125*	*0.5*
η = *0.00*	*0.99*	*0.94*	*0.64*
0.05	*1.20*	*1.19*	*0.83*
0.25	*1.40*	*1.66*	*1.33*

14.7 Asymptotic pivots

Again throughout this section, enough regularity is assumed, such as existence of derivatives and interchanging integration and differentiation.

Recall the definition of an asymptotic pivot (see Section 6.2). It is a function $Z_n(\gamma) := Z_n(X_1, \ldots, X_n, \gamma)$ of the data $X_1, \ldots, X_n$ and the parameter of interest $\gamma = g(\theta) \in \mathbb{R}^p$, such that its asymptotic distribution does not depend on the unknown parameter θ, i.e., for a random variable Z, with distribution Q not depending on θ,

$$Z_n(\gamma) \xrightarrow{\mathcal{D}_\theta} Z, \ \forall \ \theta.$$

An asymptotic pivot can be used to construct approximate $(1-\alpha)$-confidence intervals for γ, and tests for $H_0 : \ \gamma = \gamma_0$ with approximate level α.

Consider now an asymptotically normal estimator T_n of γ, which is asymptotically unbiased and has asymptotic covariance matrix V_θ, that is

$$\sqrt{n}(T_n - \gamma) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, V_\theta), \ \forall \ \theta.$$

(assuming such an estimator exists). Then, depending on the situation, there are various ways to construct an asymptotic pivot.

1$^{\text{st}}$ **asymptotic pivot**
If the asymptotic covariance matrix V_θ is non-singular, and depends only on the parameter of interest γ, say $V_\theta = V(\gamma)$ (for example, if $\gamma = \theta$), then an asymptotic pivot is

$$Z_{n,1}(\gamma) := n(T_n - \gamma)^T V(\gamma)^{-1} (T_n - \gamma).$$

The asymptotic distribution is the χ^2-distribution with p degrees of freedom.

2nd **asymptotic pivot**
If, for all θ, one has a consistent estimator $\hat{V}_n$ of V_θ, then an asymptotic pivot is

$$Z_{n,2}(\gamma) := n(T_n - \gamma)^T \hat{V}_n^{-1} (T_n - \gamma).$$

The asymptotic distribution is again the χ^2-distribution with p degrees of freedom. This follows from Slutsky's lemma.

Estimators of the asymptotic variance
∘ If $\hat{\theta}_n$ is a consistent estimator of θ and if $\theta \mapsto V_\theta$ is continuous, one may insert $\hat{V}_n := V_{\hat{\theta}_n}$.
∘ If $T_n = \hat{\gamma}_n$ is the M-estimator of γ, γ being the solution of $P_\theta \psi_\gamma = 0$, then (under regularity) the asymptotic covariance matrix is

$$V_\theta = M_\theta^{-1} J_\theta M_\theta^{-1},$$

where

$$J_\theta = P_\theta \psi_\gamma \psi_\gamma^T,$$

and

$$\begin{aligned} M_\theta &:= \ddot{\mathcal{R}}(\gamma) \\ &:= \frac{\partial}{\partial c^T} \dot{\mathcal{R}}(c)\Big|_{c=\gamma} \\ &= \frac{\partial}{\partial c^T} P_\theta \psi_c \Big|_{c=\gamma} \\ &= P_\theta \dot{\psi}_\gamma . \end{aligned}$$

Then one may estimate J_θ and M_θ by

$$\hat{J}_n := \hat{P}_n \psi_{\hat{\gamma}_n} \psi_{\hat{\gamma}_n}^T = \frac{1}{n} \sum_{i=1}^n \psi_{\hat{\gamma}_n}(X_i) \psi_{\hat{\gamma}_n}^T(X_i),$$

and

$$\begin{aligned} \hat{M}_n &:= \hat{\ddot{\mathcal{R}}}_n(\hat{\gamma}_n) \\ &= \hat{P}_n \dot{\psi}_{\hat{\gamma}_n} \\ &= \frac{1}{n} \sum_{i=1}^n \dot{\psi}_{\hat{\gamma}_n}(X_i), \end{aligned}$$

respectively. Under some regularity conditions,

$$\hat{V}_n := \hat{M}_n^{-1} \hat{J}_n \hat{M}_n^{-1}.$$

is a consistent estimator of V_θ[4].

14.8 Asymptotic pivot based on the MLE

Suppose that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ has $\Theta \subset \mathbb{R}^p$, and that $\mathcal{P}$ is dominated by some σ-finite measure ν. Let $p_\theta := dP_\theta / d\nu$ denote the densities, and let

$$\hat{\theta}_n := \arg \max_{\vartheta \in \Theta} \sum_{i=1}^n \log p_\vartheta(X_i)$$

be the MLE.

Assume enough regularity such as existence of derivatives and interchanging integration and differentiation. Recall that $\hat{\theta}_n$ is an M-estimator with loss function $\rho_\vartheta = -\log p_\vartheta$, and hence under regularity conditions, $\psi_\vartheta = \dot{\rho}_\theta$ is minus

[4] From most algorithms used to compute the M-estimator $\hat{\gamma}_n$, one easily can obtain $\hat{M}_n$ and $\hat{J}_n$ as output. Recall e.g. that the Newton-Raphson algorithm is based on the iterations

$$\hat{\gamma}_{\text{new}} = \hat{\gamma}_{\text{old}} - \left(\hat{\ddot{\mathcal{R}}}_n(\hat{\gamma}_{\text{old}}) \right)^{-1} \hat{\dot{\mathcal{R}}}_n(\hat{\gamma}_{\text{old}}).$$

the score function $s_\vartheta := \dot p_\vartheta / p_\vartheta$. The asymptotic variance of the MLE is $I^{-1}(\theta)$, where $I(\theta) := P_\theta s_\theta s_\theta^T$ is the Fisher information:

$$\sqrt{n}(\hat\theta_n - \theta) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, I^{-1}(\theta)),\ \forall\ \theta$$

(see Section 14.4). Thus, in this case

$$Z_{n,1}(\theta) = n(\hat\theta_n - \theta)^T I(\theta)(\hat\theta_n - \theta),$$

and, with $\hat I_n$ being a consistent estimator of $I(\theta)$

$$Z_{n,2}(\theta) = n(\hat\theta_n - \theta)^T \hat I_n(\hat\theta_n - \theta).$$

Note that one may take

$$\hat I_n := -\frac{1}{n}\sum_{i=1}^n \dot s_{\hat\theta_n}(X_i) = -\frac{\partial^2}{\partial\vartheta\partial\vartheta^T}\frac{1}{n}\sum_{i=1}^n \log p_\vartheta(X_i)\Big|_{\vartheta=\hat\theta_n}$$

as estimator of the Fisher information[5].

3rd **asymptotic pivot**
Define now the twice log-likelihood ratio

$$2\mathcal{L}_n(\hat\theta_n) - 2\mathcal{L}_n(\theta) := 2\sum_{i=1}^n\Big[\log p_{\hat\theta_n}(X_i) - \log p_\theta(X_i)\Big].$$

It turns out that the log-likelihood ratio is indeed an asymptotic pivot. A practical advantage is that it is self-normalizing: one does not need to explicitly estimate asymptotic (co-)variances.

Lemma 14.8.1 *Under regularity conditions, $2\mathcal{L}_n(\hat\theta_n) - 2\mathcal{L}_n(\theta)$ is an asymptotic pivot for θ. Its asymptotic distribution is again the χ^2-distribution with p degrees of freedom:*

$$2\mathcal{L}_n(\hat\theta_n) - 2\mathcal{L}_n(\theta) \xrightarrow{\mathcal{D}_\theta} \chi^2_p\ \forall\ \theta.$$

Sketch of the proof. We have by a two-term Taylor expansion

$$2\mathcal{L}_n(\hat\theta_n) - 2\mathcal{L}_n(\theta) = 2n\hat P_n\Big[\log p_{\hat\theta_n} - \log p_\theta\Big]$$

$$\approx 2n(\hat\theta_n - \theta)^T \hat P_n s_\theta + n(\hat\theta_n - \theta)^T \hat P_n \dot s_\theta(\hat\theta_n - \theta)$$

$$\approx 2n(\hat\theta_n - \theta)^T \hat P_n s_\theta - n(\hat\theta_n - \theta)^T I(\theta)(\hat\theta_n - \theta),$$

where in the second step, we used $\hat P_n \dot s_\theta \approx P_\theta \dot s_\theta = -I(\theta)$. (If you like you may compare this two-term Taylor expansion with the one in the sketch of proof of Le Cam's 3$^{\text{rd}}$ Lemma (which is not part of the exam)). With the remainder terms of the two-term Taylor expansion being asymptotically negligible, we are

[5] In other words (as for general M-estimators), the algorithm (e.g. Newton Raphson) for calculating the maximum likelihood estimator $\hat\theta_n$ generally also provides an estimator of the Fisher information as by-product.

mathematically speaking dealing with a situation as in Lemma 12.3.2 where the least squares estimator is studied[6]. The MLE $\hat{\theta}_n$ is asymptotically linear with influence function $l_\theta = I(\theta)^{-1} s_\theta$:

$$\hat{\theta}_n - \theta = I(\theta)^{-1} \hat{P}_n s_\theta + o_{\mathbf{P}_\theta}(n^{-1/2}).$$

Hence,

$$2\mathcal{L}_n(\hat{\theta}_n) - 2\mathcal{L}_n(\theta) \approx n(\hat{P}_n s_\theta)^T I(\theta)^{-1} (\hat{P}_n s_\theta).$$

The result now follows from

$$\sqrt{n}\hat{P}_n s_\theta \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, I(\theta)).$$

□

14.9 MLE for the multinomial distribution

Let $X_1, \ldots, X_n$ be i.i.d. copies of X, where $X \in \{1, \ldots, k\}$ is a label, with

$$P_\theta(X = j) := \pi_j, \ j = 1, \ldots, k.$$

where the probabilities π_j are positive and add up to one: $\sum_{j=1}^k \pi_j = 1$, but are assumed to be otherwise unknown. Then there are $p := k - 1$ unknown parameters, say $\theta = (\pi_1, \ldots, \pi_{k-1})$. Define $N_j := \#\{i : X_i = j\}$. Note that $(N_1, \ldots, N_k)$ has a multinomial distribution with parameters n and $(\pi_1, \ldots, \pi_k)$.

Lemma 14.9.1 *For each $j = 1, \ldots, k$, the MLE of π_j is*

$$\hat{\pi}_j = \frac{N_j}{n}.$$

Proof. The log-densities can be written as

$$\log p_\theta(x) = \sum_{j=1}^k \mathrm{l}\{x = j\} \log \pi_j,$$

so that

$$\sum_{i=1}^n \log p_\theta(X_i) = \sum_{j=1}^k N_j \log \pi_j.$$

Putting the derivatives with respect to $\theta = (\pi_1, \ldots, \pi_{k-1})$, (with $\pi_k = 1 - \sum_{j=1}^{k-1} \theta_j$) to zero gives,

$$\frac{N_j}{\hat{\pi}_j} - \frac{N_k}{\hat{\pi}_k} = 0.$$

Hence

$$\hat{\pi}_j = N_j \frac{\hat{\pi}_k}{N_k}, \ j = 1, \ldots, k,$$

[6] $\hat{P}_n s_\theta$ takes the role of $X^T \epsilon / n$ and $I(\theta)$ takes the role of $X^T X / n$

and thus

$$1 = \sum_{j=1}^{k} \hat{\pi}_j = n \frac{\hat{\pi}_k}{N_k},$$

yielding

$$\hat{\pi}_k = \frac{N_k}{n},$$

and hence

$$\hat{\pi}_j = \frac{N_j}{n}, \ j = 1, \ldots, k.$$

□

We now first calculate $Z_{n,1}(\theta)$. For that, we need to find the Fisher information $I(\theta)$.

Lemma 14.9.2 *The Fisher information is*

$$I(\theta) = \begin{pmatrix} \frac{1}{\pi_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\pi_{k-1}} \end{pmatrix} + \frac{1}{\pi_k} \iota \iota^T, \ ^7$$

where ι is the $(k-1)$-vector $\iota := (1, \ldots, 1)^T$.

Proof. We have

$$s_{\theta,j}(x) = \frac{1}{\pi_j} \mathrm{l}\{x = j\} - \frac{1}{\pi_k} \mathrm{l}\{x = k\}.$$

So

$$\begin{aligned} (I(\theta))_{j_1,j_2} &= E_\theta \left(\frac{1}{\pi_{j_1}} \mathrm{l}\{X = j_1\} - \frac{1}{\pi_k} \mathrm{l}\{X = k\} \right) \left(\frac{1}{\pi_{j_2}} \mathrm{l}\{X = j_2\} - \frac{1}{\pi_k} \mathrm{l}\{X = k\} \right) \\ &= \begin{cases} \frac{1}{\pi_k} & j_1 \neq j_2 \\ \frac{1}{\pi_j} + \frac{1}{\pi_k} & j_1 = j_2 = j \end{cases} . \end{aligned}$$

□

We thus find

$$\begin{aligned} Z_{n,1}(\theta) &= n(\hat{\theta}_n - \theta)^T I(\theta) (\hat{\theta}_n - \theta) \\ &= n \begin{pmatrix} \hat{\pi}_1 - \pi_1 \\ \vdots \\ \hat{\pi}_{k-1} - \pi_{k-1} \end{pmatrix}^T \left[\begin{pmatrix} \frac{1}{\pi_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\pi_{k-1}} \end{pmatrix} \right. \\ &\qquad \left. + \frac{1}{\pi_k} \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \right] \begin{pmatrix} \hat{\pi}_1 - \pi_1 \\ \vdots \\ \hat{\pi}_{k-1} - \pi_{k-1} \end{pmatrix} \\ &= n \sum_{j=1}^{k-1} \frac{(\hat{\pi}_j - \pi_j)^2}{\pi_j} + n \frac{1}{\pi_k} (\sum_{j=1}^{k-1} (\hat{\pi}_j - \pi_j))^2 \end{aligned}$$

[7] To invert such a matrix, one may apply the formula $(A + bb^T)^{-1} = A^{-1} - \frac{A^{-1} b b^T A^{-1}}{1 + b^T A^{-1} b}$.

$$\begin{aligned} &= n\sum_{j=1}^{k}\frac{(\hat{\pi}_j-\pi_j)^2}{\pi_j} \\ &= \sum_{j=1}^{k}\frac{(N_j-n\pi_j)^2}{n\pi_j}. \end{aligned}$$

This is called the Pearson's chi-square

$$\sum\frac{(\text{observed}-\text{expected})^2}{\text{expected}}.$$

A version of $Z_{n,2}(\theta)$ is to replace, for $j=1,\dots k$, π_j by $\hat{\pi}_j$ in the expression for the Fisher information. This gives

$$Z_{n,2}(\theta)=\sum_{j=1}^{k}\frac{(N_j-n\pi_j)^2}{N_j}.$$

This is called the Pearson's chi-square

$$\sum\frac{(\text{observed}-\text{expected})^2}{\text{observed}}.$$

Finally, the log-likelihood ratio pivot is

$$2\mathcal{L}_n(\hat{\theta}_n)-2\mathcal{L}_n(\theta)=2\sum_{j=1}^{k}N_j\log\left(\frac{\hat{\pi}_j}{\pi_j}\right).$$

The approximation $\log(1+x)\approx x-x^2/2$ shows that $2\mathcal{L}_n(\hat{\theta}_n)-2\mathcal{L}_n(\theta)\approx Z_{n,2}(\theta)$:

$$\begin{aligned} 2\mathcal{L}_n(\hat{\theta}_n)-2\mathcal{L}_n(\theta) &= -2\sum_{j=1}^{k}N_j\log\left(1+\frac{\pi_j-\hat{\pi}_j}{\hat{\pi}_j}\right) \\ &\approx -2\sum_{j=1}^{k}N_j\left(\frac{\pi_j-\hat{\pi}_j}{\hat{\pi}_j}\right)+\sum_{j=1}^{k}N_j\left(\frac{\pi_j-\hat{\pi}_j}{\hat{\pi}_j}\right)^2 \\ &= Z_{n,2}(\theta). \end{aligned}$$

The three asymptotic pivots $Z_{n,1}(\theta)$, $Z_{n,2}(\theta)$ and $2\mathcal{L}_n(\hat{\theta}_n)-2\mathcal{L}_n(\theta)$ are each asymptotically χ^2_{k-1}-distributed under $\mathbb{P}_\theta$.

14.10 Likelihood ratio tests

For the simple hypothesis
$H_0:\ \theta=\theta_0$,
we can use $2\mathcal{L}_n(\hat{\theta}_n)-2\mathcal{L}_n(\theta_0)$ as test statistic: reject H_0 if

$$2\mathcal{L}_n(\hat{\theta}_n)-2\mathcal{L}_n(\theta_0)>G_p^{-1}(1-\alpha),$$

where G_p is the distribution function of the χ^2_p-distribution. This test has approximately level α as was shown in Lemma 14.8.1.

Consider now the hypothesis
$H_0 : \ R(\theta) = 0$,
where

$$R(\theta) = \begin{pmatrix} R_1(\theta) \\ \vdots \\ R_q(\theta) \end{pmatrix}$$

are q restrictions on θ (with $R : \ \mathbb{R}^p \to \mathbb{R}^q$ a given function[8]).

Let $\hat{\theta}_n$ be the unrestricted MLE, that is

$$\hat{\theta}_n = \arg\max_{\vartheta \in \Theta} \sum_{i=1}^n \log p_\vartheta(X_i).$$

Moreover, let $\hat{\theta}_n^0$ be the restricted MLE, defined as

$$\hat{\theta}_n^0 = \arg \max_{\vartheta \in \Theta: \ R(\vartheta)=0} \sum_{i=1}^n \log p_\vartheta(X_i).$$

Define the $(q \times p)$-matrix

$$\dot{R}(\theta) = \frac{\partial}{\partial \vartheta^T} R(\vartheta)|_{\vartheta=\theta}.$$

We assume $\dot{R}(\theta)$ has rank q.

Let

$$\mathcal{L}_n(\hat{\theta}_n) - \mathcal{L}_n(\hat{\theta}_n^0) = \sum_{i=1}^n \left[\log p_{\hat{\theta}_n}(X_i) - \log p_{\hat{\theta}_n^0}(X_i) \right]$$

be the log-likelihood ratio for testing $H_0 : \ R(\theta) = 0$.

Lemma 14.10.1 *Under regularity conditions, and if $H_0 : \ R(\theta) = 0$ holds, we have*

$$2\mathcal{L}_n(\hat{\theta}_n) - 2\mathcal{L}_n(\hat{\theta}_n^0) \xrightarrow{\mathcal{D}_\theta} \chi^2_q.$$

Sketch of the proof. Let

$$\mathbf{Z}_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n s_\theta(X_i).$$

As in the sketch of the proof of Lemma 14.8.1, we can use a two-term Taylor expansion to show for any sequence ϑ_n satisfying $\vartheta_n = \theta + \mathcal{O}_{\mathbf{P}_\theta}(n^{-1/2})$, that

$$2 \sum_{i=1}^n \left[\log p_{\vartheta_n}(X_i) - \log p_\theta(X_i) \right]$$

[8] The notation R is used here for the restrictions. It has nothing to do with the risk $\mathcal{R}$

$$= 2\sqrt{n}(\vartheta_n - \theta)^T \mathbf{Z}_n - n(\vartheta_n - \theta)^2 I(\theta)(\vartheta_n - \theta) + o_{\mathbf{P}_\theta}(1).$$

Here, we also again use that $\sum_{i=1}^n \dot{s}_{\vartheta_n}(X_i)/n = -I(\theta) + o_{\mathbf{P}_\theta}(1)$. Moreover, by a one-term Taylor expansion, and invoking that $R(\theta) = 0$,

$$R(\vartheta_n) = \dot{R}(\theta)(\vartheta_n - \theta) + o_{\mathbf{P}_\theta}(n^{-1/2}).$$

Insert Corollary 12.4.1 with $z := \mathbf{Z}_n$, $B := \dot{R}(\theta)$, and $V = I(\theta)$. This gives

$$2\mathcal{L}_n(\hat{\theta}_n) - 2\mathcal{L}_n(\hat{\theta}_n^0)$$

$$\begin{aligned}
&= 2\sum_{i=1}^n \left[\log p_{\hat{\theta}_n}(X_i) - \log p_\theta(X_i) \right] - 2\sum_{i=1}^n \left[\log p_{\hat{\theta}_n^0}(X_i) - \log p_\theta(X_i) \right] \\
&= \mathbf{Z}_n^T I(\theta)^{-1} \dot{R}^T(\theta) \Big(\dot{R}(\theta) I(\theta)^{-1} \dot{R}(\theta)^T \Big)^{-1} \dot{R}(\theta) I(\theta)^{-1} \mathbf{Z}_n + o_{\mathbf{P}_\theta}(1) \\
&:= \mathbf{Y}_n^T W^{-1} \mathbf{Y}_n + o_{\mathbf{P}_\theta}(1),
\end{aligned}$$

where $\mathbf{Y}_n$ is the q-vector

$$\mathbf{Y}_n := \dot{R}(\theta) I(\theta)^{-1} \mathbf{Z}_n,$$

and where W is the $(q \times q)$-matrix

$$W := \dot{R}(\theta) I(\theta)^{-1} \dot{R}(\theta)^T.$$

We know that

$$\mathbf{Z}_n \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, I(\theta)).$$

Hence

$$\mathbf{Y}_n \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(0, W),$$

so that

$$\mathbf{Y}_n^T W^{-1} \mathbf{Y}_n \xrightarrow{\mathcal{D}_\theta} \chi_q^2.$$

□

Corollary 14.10.1 *From the sketch of the proof of Lemma 14.10.1, one sees that moreover (under regularity),*

$$2\mathcal{L}_n(\hat{\theta}_n) - 2\mathcal{L}_n(\hat{\theta}_n^0) \approx n(\hat{\theta}_n - \hat{\theta}_n^0)^T I(\theta)(\hat{\theta}_n - \hat{\theta}_n^0),$$

and also

$$2\mathcal{L}_n(\hat{\theta}_n) - 2\mathcal{L}_n(\hat{\theta}_n^0) \approx n(\hat{\theta}_n - \hat{\theta}_n^0)^T I(\hat{\theta}_n^0)(\hat{\theta}_n - \hat{\theta}_n^0).$$

14.11 Contingency tables

Let X be a bivariate label, say $X \in \{(j,k) : \; j = 1, \ldots, r, \; k = 1, \ldots, s\}$. For example, the first index may correspond to sex ($r = 2$) and the second index to the color of the eyes ($s = 3$). The probability of the combination (j,k) is

$$\pi_{j,k} := P_\theta \Big(X = (j,k) \Big).$$

Let $X_1, \ldots, X_n$ be i.i.d. copies of X, and

$$N_{j,k} := \#\{X_i = (j,k)\}.$$

From Section 14.9, we know that the (unrestricted) MLE of $\pi_{j,k}$ is equal to

$$\hat{\pi}_{j,k} := \frac{N_{j,k}}{n}.$$

We now want to test whether the two labels are independent. The null-hypothesis is
$H_0 : \; \pi_{j,k} = (\pi_{j,+}) \times (\pi_{+,k}) \; \forall \; (j,k)$.
Here

$$\pi_{j,+} := \sum_{k=1}^{s} \pi_{j,k}, \; \pi_{+,k} := \sum_{j=1}^{r} \pi_{j,k}.$$

One may check that the restricted MLE is

$$\hat{\pi}^0_{j,k} = (\hat{\pi}_{j,+}) \times (\hat{\pi}_{+,k}),$$

where

$$\hat{\pi}_{j,+} := \sum_{k=1}^{s} \hat{\pi}_{j,k}, \; \hat{\pi}_{+,k} := \sum_{j=1}^{r} \hat{\pi}_{j,k}.$$

The log-likelihood ratio test statistic is thus

$$\begin{aligned} 2\mathcal{L}_n(\hat{\theta}_n) - 2\mathcal{L}_n(\hat{\theta}^0_n) &= 2 \sum_{j=1}^{r} \sum_{k=1}^{s} N_{j,k} \left[\log \left(\frac{N_{j,k}}{n} \right) - \log \left(\frac{N_{j,+} N_{+,k}}{n^2} \right) \right] \\ &= 2 \sum_{j=1}^{r} \sum_{k=1}^{s} N_{j,k} \log \left(\frac{n N_{j,k}}{N_{j,+} N_{+,k}} \right). \end{aligned}$$

Its approximation as given in Corollary 14.10.1 is

$$2\mathcal{L}_n(\hat{\theta}_n) - 2\mathcal{L}_n(\hat{\theta}^0_n) \approx n \sum_{j=1}^{r} \sum_{k=1}^{s} \frac{(N_{j,k} - N_{j,+} N_{+,k}/n)^2}{N_{j,+} N_{+,k}}.$$

This is Pearson's chi-squared test statistic for testing independence.

To find out what the value of q is in this example, we first observe that the unrestricted case has $p = rs - 1$ free parameters. Under the null-hypothesis, there remain $(r-1)+(s-1)$ free parameters. Hence, the number of restrictions is

$$q = \Big(rs-1\Big) - \Big((r-1)+(s-1)\Big) = (r-1)(s-1).$$

Thus, under $H_0: \ \pi_{j,k} = (\pi_{j,+}) \times (\pi_{+,k}) \ \forall \ (j,k)$, we have

$$n\sum_{j=1}^{r}\sum_{k=1}^{s}\frac{(N_{j,k}-N_{j,+}N_{+,k}/n)^2}{N_{j,+}N_{+,k}} \xrightarrow{\mathcal{D}_\theta} \chi^2_{(r-1)(s-1)}.$$

Chapter 15

Abstract asymptotics $\star$

In Subsection 2.4.1 we discussed so-called plug-in estimators. The idea is that and estimator $\hat{\gamma}_n$ can (typically) be written as some functional Q of the empirical distribution $\hat{P}_n$. When P is the "true" distribution and $\gamma = Q(P)$, then the point is the studying how close $\hat{\gamma}_n(\hat{P}_n)$ is to $\gamma = Q(P)$ when $\hat{P}_n$ is close to P. This is the topic of the first part of this chapter.

In Section 5.5 we obtained the Cramér Rao lower bound. A somewhat disappointing result was that it can only be reached within exponential families (see Lemma 5.6.1). We have also seen in in Section 14.4 that the MLE is *asymptotically unbiased* and reaches *asymptotically* the CRLB (its asymptotic covariance matrix is $I(\theta)^{-1}$, where $I(\theta)$ is the Fisher information for estimating θ). The topic of the second part of this chapter is to show (for the one-dimensional case for simplicity) that $I(\theta)$ is indeed asymptotically the efficient variance.

One may now wonder why the inverse of $I(\theta)$ is there. Think of it in this way. The Fisher information was obtained by looking at derivatives of the map

$$\theta \mapsto \log p_\theta.$$

But what actually plays a role in the inverse map

$$P \mapsto \theta$$

or, in case $\gamma = g(\theta)$ is the parameter of interest, the map

$$P \mapsto \gamma.$$

Then remember that the derivate of the inverse of a function (say $f : \mathbb{R} \mapsto \mathbb{R}$) is the inverse of its derivative. In our case the mapping $P \mapsto \gamma$ is rather abstract, so studying its derivatives, as done in the first part of this chapter, requires some new notions.

15.1 Plug-in estimators ⋆

When $\mathcal{X}$ is Euclidean space, one can define the distribution function $F(x) := P_\theta(X \le x)$ and the empirical distribution function

$$\hat{F}_n(x) = \frac{1}{n} \#\{X_i \le x, \ 1 \le i \le n\}.$$

This is the distribution function of a probability measure that puts mass $1/n$ at each observation. For general $\mathcal{X}$, we define likewise the empirical distribution $\hat{P}_n$ as the distribution that puts mass $1/n$ at each observation, i.e., more formally

$$\hat{P}_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i},$$

where δ_x is a point mass at x. Thus, for (measurable) sets $A \subset \mathcal{X}$,

$$\hat{P}_n(A) = \frac{1}{n} \#\{X_i \in A, \ 1 \le i \le n\}.$$

For (measurable) functions $f : \mathcal{X} \to \mathbb{R}^r$ (for some $r \in \mathbb{N}$), we write, as in previous sections,

$$\hat{P}_n f := \frac{1}{n} \sum_{i=1}^n f(X_i) = \int f d\hat{P}_n.$$

Thus, for sets,

$$\hat{P}_n(A) = \hat{P}_n \mathrm{l}_A.$$

Again, as in previous sections, we use the same notations for expectations under P_θ:

$$P_\theta f := E_\theta f(X) = \int f dP_\theta,$$

so that

$$P_\theta(A) = P_\theta \mathrm{l}_A.$$

The parameter of interest is denoted as

$$\gamma = g(\theta) \in \mathbb{R}^p.$$

It can often be written in the form

$$\gamma = Q(P_\theta),$$

where Q is some functional on (a supset of) the model class $\mathcal{P}$. Assuming Q is also defined at the empirical measure $\hat{P}_n$, the plug-in estimator of γ is now

$$T_n := Q(\hat{P}_n).$$

Conversely,

Definition 15.1.1 *If a statistic T_n can be written as $T_n = Q(\hat{P}_n)$, then it is called a* Fisher-consistent *estimator of $\gamma = g(\theta)$, if $Q(P_\theta) = g(\theta)$ for all $\theta \in \Theta$.*

We will also encounter modifications, where

$$T_n = Q_n(\hat{P}_n),$$

and for n large,

$$Q_n(P_\theta) \approx Q(P_\theta) = g(\theta).$$

Example 15.1.1 Plug-in estimator of (functions of) the mean
Consider a given $f: \mathcal{X} \to \mathbb{R}^r$ and a given $h: \mathbb{R}^r \to \mathbb{R}^p$. Let $\gamma := h(P_\theta f)$. The plug-in estimator is then $T_n = h(\hat{P}_n f)$.

Example 15.1.2 M- and Z-estimators are plug-in estimators
The M-estimator

$$\hat{\gamma}_n := \arg\min_{c \in \Gamma} \hat{P}_n \rho_c$$

is a plug-in estimator of

$$\gamma = \arg\min_{c \in \Gamma} P_\theta \rho_c.$$

Similarly, the Z-estimator $\hat{\gamma}_n$ as solution of

$$\hat{P}_n \psi_c \Big|_{c = \hat{\gamma}_n} = 0$$

is a plug-in estimator of

$$P_\theta \psi_c \Big|_{c = \gamma} = 0.$$

Example 15.1.3 The α-trimmed mean as plug-in estimator
Let $\mathcal{X} = \mathbb{R}$ and consider the α-trimmed mean

$$T_n := \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} X_{(i)}.$$

What is its theoretical counterpart? Because the i-th order statistic $X_{(i)}$ can be written as

$$X_{(i)} = \hat{F}_n^{-1}(i/n),$$

and in fact

$$X_{(i)} = \hat{F}_n^{-1}(u), \ i/n \le u < (i+1)/n,$$

we may write, for $\alpha_n := [n\alpha]/n$,

$$\begin{aligned} T_n &= \frac{n}{n - 2[n\alpha]} \frac{1}{n} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} \hat{F}_n^{-1}(i/n) \\ &= \frac{1}{1 - 2\alpha_n} \int_{\alpha_n + 1/n}^{1-\alpha_n} \hat{F}_n^{-1}(u) du := Q_n(\hat{P}_n). \end{aligned}$$

Replacing $\hat{F}_n$ by F gives

$$\begin{aligned} Q_n(F) &= \frac{1}{1-2\alpha_n}\int_{\alpha_n+1/n}^{1-\alpha_n} F^{-1}(u)du \\ &\approx \frac{1}{1-2\alpha}\int_{\alpha}^{1-\alpha} F^{-1}(u)du \\ &= \frac{1}{1-2\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} xdF(x) \\ &:= Q(P_\theta). \end{aligned}$$

Example 15.1.4 Histogram as plug-in estimator of a density
Let $\mathcal{X} = \mathbb{R}$, and suppose X has density f w.r.t., Lebesgue measure. Suppose f is the parameter of interest. We may write

$$f(x) = \lim_{h\to 0}\frac{F(x+h)-F(x-h)}{2h}.$$

Replacing F by $\hat{F}_n$ here does not make sense. Thus, this is an example where $Q(P) = f$ is only well defined for distributions P that have a density f. We may however slightly extend the plug-in idea, by using the estimator

$$\hat{f}_n(x) := \frac{\hat{F}_n(x+h_n)-\hat{F}_n(x-h_n)}{2h_n} := Q_n(\hat{P}_n),$$

with h_n "small" ($h_n \to 0$ as $n \to \infty$).

15.2 Consistency of plug-in estimators ⋆

We first present the uniform convergence of the empirical distribution function to the theoretical one.

Such uniform convergence results hold also in much more general settings (see also (14.2) in the proof of consistency for M-estimators).

Theorem 15.2.1 *(Glivenko-Cantelli) Let $\mathcal{X} = \mathbb{R}$. We have*

$$\sup_x |\hat{F}_n(x) - F(x)| \to 0,\ \mathbb{P}_\theta - a.s..$$

Proof. We know that by the law of large numbers, for all x

$$|\hat{F}_n(x) - F(x)| \to 0,\ \mathbb{P}_\theta\text{−a.s.},$$

so also for all finite collection $a_1, \ldots, a_N$,

$$\max_{1\le j\le N} |\hat{F}_n(a_j) - F(a_j)| \to 0,\ \mathbb{P}_\theta\text{−a.s..}$$

Let $\epsilon > 0$ be arbitrary, and take $a_0 < a_1 < \cdots < a_{N-1} < a_N$ in such a way that

$$F(a_j) - F(a_{j-1}) \le \epsilon,\ j = 1, \ldots, N$$

where $F(a_0) := 0$ and $F(a_N) := 1$. Then, when $x \in (a_{j-1}, a_j]$,

$$\hat{F}_n(x) - F(x) \leq \hat{F}_n(a_j) - F(a_{j-1}) \leq F_n(a_j) - F(a_j) + \epsilon,$$

and

$$\hat{F}_n(x) - F(x) \geq \hat{F}_n(a_{j-1}) - F(a_j) \geq \hat{F}_n(a_{j-1}) - F(a_{j-1}) - \epsilon,$$

so

$$\sup_x |\hat{F}_n(x) - F(x)| \leq \max_{1 \leq j \leq N} |\hat{F}_n(a_j) - F(a_j)| + \epsilon \to \epsilon, \ \mathbb{P}_\theta\text{–a.s..}$$

□

Example 15.2.1 Consistency of the sample median
Let $\mathcal{X} = \mathbb{R}$ and let F be the distribution function of X. We consider estimating the median $\gamma := F^{-1}(1/2)$. We assume F to continuous and strictly increasing. The sample median is

$$T_n := \hat{F}_n^{-1}(1/2) := \begin{cases} X_{((n+1)/2)} & n \textit{ odd} \\ [X_{(n/2)} + X_{(n/2+1)}]/2 & n \textit{ even} \end{cases}.$$

So

$$\hat{F}_n(T_n) = \frac{1}{2} + \begin{cases} 1/(2n) & n \textit{ odd} \\ 0 & n \textit{ even} \end{cases}.$$

It follows that

$$\begin{aligned} |F(T_n) - F(\gamma)| &\leq |\hat{F}_n(T_n) - F(T_n)| + |\hat{F}_n(T_n) - F(\gamma)| \\ &= |\hat{F}_n(T_n) - F(T_n)| + |\hat{F}_n(T_n) - \frac{1}{2}| \\ &\leq |\hat{F}_n(T_n) - F(T_n)| + \frac{1}{2n} \to 0, \ \mathbb{P}_\theta\text{–a.s..} \end{aligned}$$

So $\hat{F}_n^{-1}(1/2) = T_n \to \gamma = F^{-1}(1/2)$, $\mathbb{P}_\theta$–a.s., i.e., the sample median is a consistent estimator of the population median.

15.3 Asymptotic normality of plug-in estimators ⋆

Let $\gamma := Q(P) \in \mathbb{R}^p$ be the parameter of interest. The idea in this subsection is to apply a δ-method, but now in a nonparametric framework. The parametric δ-method says that if $\hat{\theta}_n$ is an asymptotically linear estimator of $\theta \in \mathbb{R}^p$, and if $\gamma = g(\theta)$ is some function of the parameter θ, with g being differentiable at θ, then $\hat{\gamma}$ is an asymptotically linear estimator of γ. Now, we write $\gamma = Q(P)$ as a function of the probability measure P (with $P = P_\theta$, so that $g(\theta) = Q(P_\theta)$). We let P play the role of θ, i.e., we use the probability measures themselves as parameterization of $\mathcal{P}$. We then have to redefine differentiability in an abstract setting, namely we differentiate w.r.t. P.

Definition 15.3.1
○ *The* influence function *of Q at P is*

$$l_P(x) := \lim_{\epsilon \downarrow 0} \frac{Q((1-\epsilon)P + \epsilon\delta_x) - Q(P)}{\epsilon}, \ x \in \mathcal{X},$$

whenever the limit exists.
∘ *The map Q is called* Gâteaux differentiable *at P if for all probability measures $\tilde{P}$, we have*

$$\lim_{\epsilon \downarrow 0} \frac{Q((1-\epsilon)P + \epsilon \tilde{P}) - Q(P)}{\epsilon} = E_{\tilde{P}} l_P(X).$$

∘ *Let d be some (pseudo-)metric on the space of probability measures. The map Q is called* Fréchet differentiable *at P, with respect to the metric d, if*

$$Q(\tilde{P}) - Q(P) = E_{\tilde{P}} l_P(X) + o(d(\tilde{P}, P)).$$

Remark 1 In line with the notation introduced previously, we write for a function $f : \mathcal{X} \to \mathbb{R}^r$ and a probability measure $\tilde{P}$ on $\mathcal{X}$

$$\tilde{P} f := E_{\tilde{P}} f(X).$$

Remark 2 If Q is Fréchet or Gâteaux differentiable at P, then

$$P l_P (:= E_P l_P(X)) = 0.$$

Remark 3 If Q is Fréchet differentiable at P, and if moreover

$$d((1-\epsilon)P + \epsilon \tilde{P}, P) = o(\epsilon), \ \epsilon \downarrow 0,$$

then Q is Gâteaux differentiable at P:

$$\begin{aligned} Q((1-\epsilon)P + \epsilon \tilde{P}) - Q(P) &= ((1-\epsilon)P + \epsilon \tilde{P}) l_P + o(\epsilon) \\ &= \epsilon \tilde{P} l_P + o(\epsilon). \end{aligned}$$

The following result says that Fréchet differentiable functionals are generally asymptotically linear.

Lemma 15.3.1 *Suppose that Q is Fréchet differentiable at P with influence function l_P, and that*

$$d(\hat{P}_n, P) = \mathcal{O}_{\mathbf{P}}(n^{-1/2}). \tag{15.1}$$

Then

$$Q(\hat{P}_n) - Q(P) = \hat{P}_n l_P + o_{\mathbf{P}}(n^{-1/2}).$$

Proof. This follows immediately from the definition of Fréchet differentiability. □

Corollary 15.3.1 *Assume the conditions of Lemma 15.3.1, with influence function l_P satisfying $V_P := P l_P l_P^T < \infty$. Then*

$$\sqrt{n}(Q(\hat{P}_n) - Q(P)) \xrightarrow{\mathcal{D}_P} \mathcal{N}(0, V_P).$$

An example where (15.1) **holds**
Suppose $\mathcal{X} = \mathbb{R}$ and that we take

$$d(\tilde{P}, P) := \sup_x |\tilde{F}(x) - F(x)|.$$

Then indeed $d(\hat{P}_n, P) = O_{\mathbf{P}}(n^{-1/2})$. This follows from Donsker's theorem, which we state here without proof:

Theorem 15.3.1 *(Donsker's theorem) Suppose F is continuous. Then*

$$\sup_x \sqrt{n} |\hat{F}_n(x) - F(x)| \xrightarrow{\mathcal{D}} Z,$$

where the random variable Z has distribution function

$$G(z) = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j+1} \exp[-2j^2 z^2], \ z \geq 0.$$

Fréchet differentiability is generally quite hard to prove, and often not even true. We will sketch how to establish Gâteaux differentiability in two examples.

Example 15.3.1 Asymptotic linearity of the Z-estimator
We consider again the asymptotic linearity of the Z-estimator. Throughout in this example, we assume enough regularity. Let γ be defined by the equation

$$P\psi_\gamma = 0.$$

Let $P_\epsilon := (1 - \epsilon)P + \epsilon \tilde{P}$, and let γ_ϵ be a solution of the equation

$$P_\epsilon \psi_{\gamma_\epsilon} = 0.$$

We assume that as $\epsilon \downarrow 0$, also $\gamma_\epsilon \to \gamma$. It holds that

$$(1 - \epsilon) P\psi_{\gamma_\epsilon} + \epsilon \tilde{P} \psi_{\gamma_\epsilon} = 0,$$

so

$$P\psi_{\gamma_\epsilon} + \epsilon(\tilde{P} - P)\psi_{\gamma_\epsilon} = 0,$$

and hence

$$P(\psi_{\gamma_\epsilon} - \psi_\gamma) + \epsilon(\tilde{P} - P)\psi_{\gamma_\epsilon} = 0.$$

Assuming differentiabality of $c \mapsto P\psi_c$, we obtain

$$\begin{aligned} P(\psi_{\gamma_\epsilon} - \psi_\gamma) &= \left(\frac{\partial}{\partial c^T} P\psi_c \Big|_{c=\gamma} \right) (\gamma_\epsilon - \gamma) + o(|\gamma_\epsilon - \gamma|) \\ &:= M_P(\gamma_\epsilon - \gamma) + o(|\gamma_\epsilon - \gamma|). \end{aligned}$$

Moreover, again under regularity

$$\begin{aligned} (\tilde{P} - P)\psi_{\gamma_\epsilon} &= (\tilde{P} - P)\psi_\gamma + (\tilde{P} - P)(\psi_{\gamma_\epsilon} - \psi_\gamma) \\ &= (\tilde{P} - P)\psi_\gamma + o(1) = \tilde{P}\psi_\gamma + o(1). \end{aligned}$$

It follows that

$$M_P(\gamma_\epsilon - \gamma) + o(|\gamma_\epsilon - \gamma|) + \epsilon(\tilde{P} - P)\psi_\gamma + o(\epsilon) = 0,$$

or, assuming M_P to be invertible,

$$(\gamma_\epsilon - \gamma)(1 + o(1)) = -\epsilon M_P^{-1}\tilde{P}\psi_\gamma + o(\epsilon),$$

which gives

$$\frac{\gamma_\epsilon - \gamma}{\epsilon} \to -M_P^{-1}\tilde{P}\psi_\gamma.$$

The influence function is thus (as already seen in Subsection 14.3)

$$l_P = -M_P^{-1}\psi_\gamma.$$

Example 15.3.2 Asymptotic linearity of the α-trimmed mean
The α-trimmed mean is a plug-in estimator of

$$\gamma := Q(P) = \frac{1}{1-2\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x dF(x).$$

Using partial integration, may write this as

$$(1-2\alpha)\gamma = (1-\alpha)F^{-1}(1-\alpha) - \alpha F^{-1}(\alpha) - \int_\alpha^{1-\alpha} v dF^{-1}(v).$$

The influence function of the quantile $F^{-1}(v)$ is

$$q_v(x) = -\frac{1}{f(F^{-1}(v))}\Big(\mathrm{l}\{x \le F^{-1}(v)\} - v\}\Big)$$

(see Example 14.5.1), i.e., for the distribution $P_\epsilon = (1-\epsilon)P + \epsilon\tilde{P}$, with distribution function $F_\epsilon = (1-\epsilon)F + \epsilon\tilde{F}$, we have

$$\begin{aligned} \lim_{\epsilon \downarrow 0} \frac{F_\epsilon^{-1}(v) - F^{-1}(v)}{\epsilon} &= \tilde{P}q_v \\ &= -\frac{1}{f(F^{-1}(v))}\Big(\tilde{F}(F^{-1}(v)) - v\Big). \end{aligned}$$

Hence, for $P_\epsilon = (1-\epsilon)P + \epsilon\tilde{P}$,

$$\begin{aligned} &(1-2\alpha)\lim_{\epsilon \downarrow 0}\frac{Q((1-\epsilon)P + \epsilon\tilde{P}) - Q(P)}{\epsilon} \\ &= (1-\alpha)\tilde{P}q_{1-\alpha} - \alpha\tilde{P}q_\alpha - \int_\alpha^{1-\alpha} v d\tilde{P}q_v \\ &= \int_\alpha^{1-\alpha} \frac{1}{f(F^{-1}(v))}\Big(\tilde{F}(F^{-1}(v)) - v\Big)dv \\ &= \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} \frac{1}{f(u)}\Big(\tilde{F}(u) - F(u)\Big)dF(u) \\ &= \int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} \Big(\tilde{F}(u) - F(u)\Big)du \\ &= (1-2\alpha)\tilde{P}l_P, \end{aligned}$$

where

$$l_P(x) = -\frac{1}{1-2\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} \Big(\mathrm{l}\{x \le u\} - F(u) \Big) du.$$

We conclude that, under regularity conditions, the α-trimmed mean is asymptotically linear with the above influence function l_P, and hence asymptotically normal with asymptotic variance Pl_P^2.

15.4 Asymptotic Cramer Rao lower bound ⋆

Let X have distribution $P \in \{P_\theta : \theta \in \Theta\}$. We assume for simplicity that $\Theta \subset \mathbb{R}$ and that θ is the parameter of interest. Let T_n be an estimator of θ.

Throughout this section, we take certain, sometimes unspecified, regularity conditions for granted.

In particular, we assume that $\mathcal{P}$ is dominated by some σ-finite measure ν, and that the Fisher-information

$$I(\theta) := E_\theta s_\theta^2(X)$$

exists for all θ. Here, s_θ is the score function

$$s_\theta := \frac{d}{d\theta} \log p_\theta = \dot{p}_\theta / p_\theta,$$

with $p_\theta := dP_\theta / d\nu$.

Recall now that if T_n is an unbiased estimator of θ, then by the Cramer Rao lower bound, $1/I(\theta)$ is a lower bound for its variance (under regularity Conditions I and II, see Section 5.5).

Definition 15.4.1 *Suppose that*

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}_\theta} \mathcal{N}(b_\theta, V_\theta), \ \forall \ \theta.$$

Then b_θ is called the asymptotic bias*, and V_θ the* asymptotic variance*. The estimator T_n is called* asymptotically unbiased *if $b_\theta = 0$ for all θ. If T_n is asymptotically unbiased and moreover $V_\theta = 1/I(\theta)$ for all θ, and some regularity conditions holds, then T_n is called* asymptotically efficient*.*

Remark 1 The assumptions in the above definition, are *for all* θ. Clearly, if one only looks at one fixed given θ_0, it is easy to construct a super-efficient estimator, namely $T_n = \theta_0$. More generally, to avoid this kind of super-efficiency, one does not only require conditions to hold *for all* θ, but in fact *uniformly in* θ, or for all *sequences* $\{\theta_n\}$. The regularity one needs here involves the idea that one actually needs to allow for sequences θ_n the form $\theta_n = \theta + h/\sqrt{n}$. In fact, the regularity requirement is that also, for all h,

$$\sqrt{n}(T_n - \theta_n) \xrightarrow{\mathcal{D}_{\theta_n}} \mathcal{N}(0, V_\theta).$$

To make all this mathematically precise is quite involved. We refer to van der Vaart (1998). A glimps is given in Le Cam's 3^{rd} Lemma, see the next subsection.

Remark 2 Note that when $\theta = \theta_n$ is allowed to change with n, this means that distribution of X_i can change with n, and hence X_i can change with n. Instead of regarding the sample $X_1, \ldots, X_n$ are the first n of an infinite sequence, we now consider for each n a new sample, say $X_{1,1}, \ldots, X_{n,n}$.

Remark 3 We have seen that the MLE $\hat{\theta}_n$ generally is indeed asymptotically unbiased with asymptotic variance V_θ equal to $1/I(\theta)$, i.e., under regularity assumptions, the MLE is asymptotically efficient.

For asymptotically linear estimators, with influence function l_θ, one has asymptotic variance $V_\theta = E_\theta l_\theta^2(X)$. The next lemma indicates that generally $1/I(\theta)$ is indeed a lower bound for the asymptotic variance.

Lemma 15.4.1 *Suppose asymptotic linearity:*

$$T_n - \theta = \frac{1}{n}\sum_{i=1}^n l_\theta(X_i) + o_{\mathbf{P}_\theta}(n^{-1/2}),$$

where $E_\theta l_\theta(X) = 0$, $E_\theta l_\theta^2(X) := V_\theta < \infty$. Assume moreover that

$$E_\theta l_\theta(X) s_\theta(X) = 1. \tag{15.2}$$

Then

$$V_\theta \ge \frac{1}{I(\theta)}.$$

Proof. This follows from the Cauchy-Schwarz inequality:

$$1 = |\text{cov}_\theta(l_\theta(X), s_\theta(X))|^2$$

$$\le \text{var}_\theta(l_\theta(X))\text{var}_\theta(s_\theta(X)) = V_\theta I(\theta).$$

□

It may look like a coincidence when in a special case, equality (15.2) indeed holds. But actually, it is true in quite a few cases. This may at first seem like magic.

We consider two examples. To simplify the expressions, we again write shorthand

$$P_\theta f := E_\theta f(X).$$

Example 15.4.1 Equation (15.2) for Z-estimators
This example examines the Z-estimator of θ. Then we have, for $P = P_\theta$,

$$P\psi_\theta = 0.$$

The influence function is

$$l_\theta = -\psi_\theta / M_\theta,$$

where

$$M_\theta := \frac{d}{d\theta} P\psi_\theta.$$

Under regularity, we have

$$M_\theta = P\dot\psi_\theta = \int \dot\psi_\theta p_\theta d\nu, \ \dot\psi_\theta = \frac{d}{d\theta}\psi_\theta.$$

We may also write

$$M_\theta = -\int \psi_\theta \dot p_\theta d\nu, \ \dot p_\theta = \frac{d}{d\theta} p_\theta.$$

This follows from the chain rule

$$\frac{d}{d\theta}\psi_\theta p_\theta = \dot\psi_\theta p_\theta + \psi_\theta \dot p_\theta,$$

and (under regularity)

$$\int \frac{d}{d\theta}\psi_\theta p_\theta d\nu = \frac{d}{d\theta}\int \psi_\theta p_\theta d\nu = \frac{d}{d\theta} P\psi_\theta = \frac{d}{d\theta} 0 = 0.$$

Thus

$$P l_\theta s_\theta = -M_\theta^{-1} P\psi_\theta s_\theta = -M_\theta^{-1}\int \psi_\theta \dot p_\theta d\nu = 1,$$

that is, (15.2) holds.

Example 15.4.2 Equation (15.2) for plug-in estimators
We consider now the plug-in estimator $Q(\hat P_n)$. Suppose that Q is Fisher consistent (i.e., $Q(P_\theta) = \theta$ for all θ). Assume moreover that Q is Fréchet differentiable with respect to the metric d, at all P_θ, and that

$$d(P_{\tilde\theta}, P_\theta) = \mathcal{O}(|\tilde\theta - \theta|).$$

Then, by the definition of Fréchet differentiability

$$h = Q(P_{\theta+h}) - Q(P_\theta) = P_{\theta+h} l_\theta + o(|h|) = (P_{\theta+h} - P_\theta) l_\theta + o(|h|),$$

or, as $h \to 0$,

$$\begin{aligned} 1 &= \frac{(P_{\theta+h} - P_\theta) l_\theta}{h} + o(1) = \frac{\int l_\theta (p_{\theta+h} - p_\theta) d\nu}{h} + o(1) \\ &\to \int l_\theta \dot p_\theta d\nu = P_\theta(l_\theta s_\theta). \end{aligned}$$

So (15.2) holds.

15.5 Le Cam's 3^{rd} Lemma $\star$

The following example serves as a motivation to consider sequences θ_n depending on n. It shows that pointwise asymptotics can be very misleading.

Example 15.5.1 Hodges-Lehmann example of super-efficiency
Let $X_1, \ldots, X_n$ be i.i.d. copies of X, where $X = \theta + \epsilon$, and ϵ is $\mathcal{N}(0,1)$-distributed. Consider the estimator

$$T_n := \begin{cases} \bar{X}_n, & \text{if } |\bar{X}_n| > n^{-1/4} \\ \bar{X}_n/2, & \text{if } |\bar{X}_n| \le n^{-1/4} \end{cases} .$$

Then

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}_\theta} \begin{cases} \mathcal{N}(0,1), & \theta \ne 0 \\ \mathcal{N}(0, \frac{1}{4}), & \theta = 0 \end{cases} .$$

So the pointwise asymptotics show that T_n can be more efficient than the sample average $\bar{X}_n$. But what happens if we consider sequences θ_n? For example, let $\theta_n = h/\sqrt{n}$. Then, under $\mathbb{P}_{\theta_n}$, $\bar{X}_n = \bar{\epsilon}_n + h/(\sqrt{n}) = \mathcal{O}_{\mathbf{P}_{\theta_n}}(n^{-1/2})$. Hence, $\mathbb{P}_{\theta_n}(|\bar{X}_n| > n^{-1/4}) \to 0$, so that $\mathbb{P}_{\theta_n}(T_n = \bar{X}_n) \to 0$. Thus,

$$\begin{aligned} \sqrt{n}(T_n - \theta_n) &= \sqrt{n}(T_n - \theta_n)\mathrm{l}\{T_n = \bar{X}_n\} + \sqrt{n}(T_n - \theta_n)\mathrm{l}\{T_n = \bar{X}_n/2\} \\ &\xrightarrow{\mathcal{D}_{\theta_n}} \mathcal{N}\biggl(-\frac{h}{2}, \frac{1}{4}\biggr). \end{aligned}$$

The asymptotic mean square error $\mathrm{AMSE}_\theta(T_n)$ is defined as the asymptotic variance + asymptotic squared bias:

$$\mathrm{AMSE}_{\theta_n}(T_n) = \frac{1+h^2}{4}.$$

The $\mathrm{AMSE}_\theta(\bar{X}_n)$ of $\bar{X}_n$ is its normalized non-asymptotic mean square error, which is

$$\mathrm{AMSE}_{\theta_n}(\bar{X}_n) = \mathrm{MSE}_{\theta_n}(\bar{X}_n) = 1.$$

So when h is large enough, the asymptotic mean square error of T_n is larger than that of $\bar{X}_n$.

Le Cam's 3^{rd} lemma shows that asymptotic linearity for all θ implies asymptotic normality, now also for sequences $\theta_n = \theta + h/\sqrt{n}$. The asymptotic variance for such sequences θ_n does not change. Moreover, if (15.2) holds for all θ, the estimator is also asymptotically unbiased under $\mathbb{P}_{\theta_n}$.

Lemma 15.5.1 *(Le Cam's 3^{rd} Lemma) Suppose that for all θ,*

$$T_n - \theta = \frac{1}{n}\sum_{i=1}^n l_\theta(X_i) + o_{\mathbf{P}_\theta}(n^{-1/2}),$$

where $P_\theta l_\theta = 0$, and $V_\theta := P_\theta l_\theta^2 < \infty$. Then, under regularity conditions,

$$\sqrt{n}(T_n - \theta_n) \xrightarrow{\mathcal{D}_{\theta_n}} \mathcal{N}\biggl(\{P_\theta(l_\theta s_\theta) - 1\}h, V_\theta\biggr).$$

We will present a sketch of the proof of this lemma. For this purpose, we need the following auxiliary lemma.

Lemma 15.5.2 *(Auxiliary lemma) Let $Z \in \mathbb{R}^2$ be $\mathcal{N}(\mu, \Sigma)$-distributed, where*

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \ \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix}.$$

Suppose that

$$\mu_2 = -\sigma_2^2/2.$$

Let $Y \in \mathbb{R}^2$ be $\mathcal{N}(\mu + a, \Sigma)$-distributed, with

$$a = \begin{pmatrix} \sigma_{1,2} \\ \sigma_2^2 \end{pmatrix}.$$

Let ϕ_Z be the density of Z and ϕ_Y be the density of Y. Then we have the following equality for all $z = (z_1, z_2) \in \mathbb{R}^2$:

$$\phi_Z(z)\mathrm{e}^{z_2} = \phi_Y(z).$$

Proof. The density of Z is

$$\phi_Z(z) = \frac{1}{2\pi\sqrt{\det(\Sigma)}} \exp\left[-\frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu)\right].$$

Now, one easily sees that

$$\Sigma^{-1}a = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

So

$$\begin{aligned} \frac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu) &= \frac{1}{2}(z-\mu-a)^T\Sigma^{-1}(z-\mu-a) \\ &+ a^T\Sigma^{-1}(z-\mu) - \frac{1}{2}a^T\Sigma^{-1}a \end{aligned}$$

and

$$\begin{aligned} a^T\Sigma^{-1}(z-\mu) - \frac{1}{2}a^T\Sigma^{-1}a &= \begin{pmatrix} 0 \\ 1 \end{pmatrix}^T (z-\mu) - \frac{1}{2}\begin{pmatrix} 0 \\ 1 \end{pmatrix}^T a \\ &= z_2 - \mu_2 - \frac{1}{2}\sigma_2^2 = z_2. \end{aligned}$$

□

Sketch of proof of Le Cam's 3rd Lemma. Set

$$\Lambda_n := \sum_{i=1}^n \left[\log p_{\theta_n}(X_i) - \log p_\theta(X_i)\right].$$

Then under $\mathbb{P}_\theta$, by a two-term Taylor expansion,

$$\begin{aligned} \Lambda_n &\approx \frac{h}{\sqrt{n}}\sum_{i=1}^n s_\theta(X_i) + \frac{h^2}{2}\frac{1}{n}\sum_{i=1}^n \dot{s}_\theta(X_i) \\ &\approx \frac{h}{\sqrt{n}}\sum_{i=1}^n s_\theta(X_i) - \frac{h^2}{2}I(\theta), \end{aligned}$$

as

$$\frac{1}{n}\sum_{i=1}^n \dot{s}_\theta(X_i) \approx E_\theta \dot{s}_\theta(X) = -I(\theta).$$

We moreover have, by the assumed asymptotic linearity, under $\mathbb{P}_\theta$,

$$\sqrt{n}(T_n - \theta) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^n l_\theta(X_i).$$

Thus,

$$\begin{pmatrix} \sqrt{n}(T_n - \theta) \\ \Lambda_n \end{pmatrix} \xrightarrow{\mathcal{D}_\theta} Z,$$

where $Z \in \mathbb{R}^2$, has the two-dimensional normal distribution:

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ -\frac{h^2}{2}I(\theta) \end{pmatrix}, \begin{pmatrix} V_\theta & hP_\theta(l_\theta s_\theta) \\ hP_\theta(l_\theta s_\theta) & h^2 I(\theta) \end{pmatrix} \right).$$

Thus, we know that for all bounded and continuous $f : \mathbb{R}^2 \to \mathbb{R}$, one has

$$\mathbb{E}_\theta f(\sqrt{n}(T_n - \theta), \Lambda_n) \to \mathbb{E} f(Z_1, Z_2).$$

Now, let $f : \mathbb{R} \to \mathbb{R}$ be bounded and continuous. Then, since

$$\prod_{i=1}^n p_{\theta_n}(X_i) = \prod_{i=1}^n p_\theta(X_i) \mathrm{e}^{\Lambda_n},$$

we may write

$$\mathbb{E}_{\theta_n} f(\sqrt{n}(T_n - \theta)) = \mathbb{E}_\theta f(\sqrt{n}(T_n - \theta))\mathrm{e}^{\Lambda_n}.$$

The function $(z_1, z_2) \mapsto f(z_1)\mathrm{e}^{z_2}$ is continuous, but not bounded. However, one can show that one may extend the Portmanteau Theorem to this situation. This then yields

$$\mathbb{E}_\theta f(\sqrt{n}(T_n - \theta))\mathrm{e}^{\Lambda_n} \to \mathbb{E} f(Z_1)\mathrm{e}^{Z_2}.$$

Now, apply the auxiliary Lemma, with

$$\mu = \begin{pmatrix} 0 \\ -\frac{h^2}{2}I(\theta) \end{pmatrix}, \ \Sigma = \begin{pmatrix} V_\theta & hP_\theta(l_\theta s_\theta) \\ hP_\theta(l_\theta s_\theta) & h^2 I(\theta) \end{pmatrix}.$$

Then we get

$$\mathbb{E} f(Z_1)\mathrm{e}^{Z_2} = \int f(z_1)\mathrm{e}^{z_2}\phi_Z(z)dz = \int f(z_1)\phi_Y(z)dz = \mathbb{E} f(Y_1),$$

where

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} hP_\theta(l_\theta s_\theta) \\ \frac{h^2}{2}I(\theta) \end{pmatrix}, \begin{pmatrix} V_\theta & hP_\theta(l_\theta s_\theta) \\ hP_\theta(l_\theta s_\theta) & h^2 I(\theta) \end{pmatrix} \right),$$

so that

$$Y_1 \sim \mathcal{N}(hP_\theta(l_\theta s_\theta), V_\theta).$$

So we conclude that

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}_{\theta_n}} Y_1 \sim \mathcal{N}(hP_\theta(l_\theta s_\theta), V_\theta).$$

Hence

$$\sqrt{n}(T_n - \theta_n) = \sqrt{n}(T_n - \theta) - h \xrightarrow{\mathcal{D}_{\theta_n}} \mathcal{N}(h\{P_\theta(l_\theta s_\theta) - 1\}, V_\theta).$$

□

Chapter 16

Complexity regularization

Suppose again the framework where we have i.i.d. copies $X_1, \ldots, X_n$ of a population random variable X, and that we model the distribution P of X as $P \in \mathcal{P} = \{P_\theta : \ \theta \in \Theta\}$. If Θ is finite-dimensional, one can regard its dimension, p say, as a measure of the complexity of the parameter space Θ. If p is larger than n, there are more parameters than observations and one may intuitively already see that this is not a statistically favourable situation. It is in a way comparable to a system with more unknowns then equations: an ill-posed system. A model with very many parameters may fit the data very well, but will have little predictive power. With too many parameters, there is a danger of overfitting. Complexity regularization can be roughly seen as a way to deal with a complex parameter space by letting the data decide which sub-model yields a good trade-off between the approximation error and estimation error.

We note that even when the parameter space is ∞-dimensional one need not always apply complexity regularization. The dimension of Θ is in itself not always the best description of complexity. If Θ is a metric space, one can apply its so-called *entropy* as a measure of complexity. The details go beyond the scope of these lecture notes. We instead will give the non-parametric and the high-dimensional regression problem as prototype examples.

16.1 Non-parametric regression

Consider n real-valued response variables $Y_1, \ldots, Y_n$ which depend on some fixed co-variables $x_1, \ldots, x_n$ (with x_i in some space $\mathcal{X}$ for all i) in the following manner:

$$Y_i = f(x_i) + \epsilon_i, \ i = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ is unobservable noise, and where f is an unknown function. Suppose we think of using the least squares estimator for estimating f. Let $Y := (Y_1, \ldots, Y_n)^T$. With some abuse of notation, we identify a function f :

$\mathcal{X} \to \mathbb{R}$ with the vector

$$\mathrm{f} := (\mathrm{f}(x_1), \ldots, \mathrm{f}(x_n))^T \in \mathbb{R}^n$$

The least squares estimator is

$$\hat{f}_{\mathrm{LS}} := \arg \min_{\mathrm{f} \in \mathbb{R}^n} \|Y - \mathrm{f}\|_2^2.$$

Clearly, if all x_i are distinct, then $\hat{f}_{\mathrm{LS}} = Y$ and one has a perfect fit

$$\|Y - \hat{f}_{\mathrm{LS}}\|_2^2 = 0.$$

This is a typical instance of *overfitting.* The estimator $\hat{f}_{\mathrm{LS}}$ just reproduces the data and has no predictive power.

We need a model for f, say $f \in \mathcal{F}$ with $\mathcal{F}$ some class of functions.

16.2 Smoothness classes

Suppose $\mathcal{X}$ is some interval in $\mathbb{R}$, say $\mathcal{X} = [0, 1]$. It depends on the situation but a reasonable assumption on f may be that it is not too wiggly. One may formulate that mathematically for example by saying that f is differentiable with derivative f' with $|f'|$ within bounds. One may for example measure the roughness of f by the (Sobolev) semi-norm of its derivative

$$\sqrt{\int_0^1 |f'(x)|^2 dx}.$$

One way to go is now to do least squares over all f under the restriction that $\int_0^1 |\mathrm{f}'(x)|^2 dx \le M^2$ where M is a given constant. It turns out that a more flexible approach is to apply the Lagrangian version of this. Then, choose a tuning parameter $\lambda \ge 0$ and define the estimator $\hat{f}$ as

$$\hat{f} := \arg \min_{\mathrm{f}} \biggl\{ \|Y - \mathrm{f}\|_2^2 + \lambda^2 \int_0^1 |\mathrm{f}'(x)|^2 dx \biggr\}.$$

This is called (a version of) *Tikhonov regularization.* One may view

$$\lambda^2 \int_0^1 |\mathrm{f}'(x)|^2 dx$$

as a penalty for choosing a too wiggly function. The penalty regularizes the function. The tuning parameter λ controls the amount of regularization: the larger λ the more regular the estimator will be. The choice of λ is not an easy point. From theoretical point of view there are some guidelines. (Choosing λ^2 of order $n^{1/3}$ trades off approximation error and estimation error. One may alternatively invoke Bayesian arguments to choose λ.) In practice one may apply cross-validation.

So far we described smoothness in terms of first derivatives being bounded. One may also use higher order derivatives if one believes the unknown function f has these. Let $\mathrm{f}^{(m)}$ denote the derivative of order m of the function $\mathrm{f}: [0,1] \to \mathbb{R}$. A possible penalty is then

$$\lambda^2 \int_0^1 |\mathrm{f}^{(m)}(x)|^2 dx.$$

The tuning parameter λ can be chosen smaller for higher values of m. (A value of order $n^{\frac{1}{2m+1}}$ gives a trade-off between approximation error and estimation error.) The resulting estimator is called a *smoothing spline*.

With a quadratic penalty, the penalized least squares estimator $\hat{f}$ is not difficult to compute as it is a minimizer of a quadratic function. Nevertheless, explicit expressions are typically not available. In the next section, we provide explicit expressions for the continuous version, as a "curiosity".

16.3 A continuous version with explicit solution ⋆

We examine the continuous version of the previous section. The problem can be explicitly solved. Suppose we observe a function $y : [0,1] \to \mathbb{R}$ and we aim at smoothing it using the penalty of the previous section. We let

$$\hat{f} = \arg\min_{\mathrm{f}} \left\{ \int_0^1 |y(x) - \mathrm{f}(x)|^2 dx + \lambda^2 \int_0^1 |\mathrm{f}'(x)|^2 dx \right\}. \tag{16.1}$$

Lemma 16.3.1 *Let $\hat{f}$ be given in (16.1). Then*

$$\hat{f}(x) = \frac{C}{\lambda} \cosh\left(\frac{x}{\lambda}\right) + \frac{1}{\lambda} \int_0^x y(u) \sinh\left(\frac{u-x}{\lambda}\right) du,$$

where

$$C = Y(1) - \left\{ \frac{1}{\lambda} \int_0^1 Y(u) \sinh\left(\frac{1-u}{\lambda}\right) du \right\} / \sinh\left(\frac{1}{\lambda}\right),$$

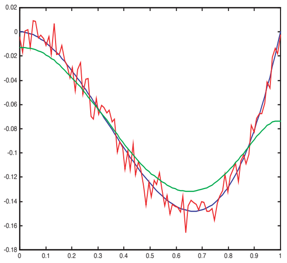
with

$$Y(x) = \int_0^x y(u) du.$$

The proof is calculus of variations and will be skipped.

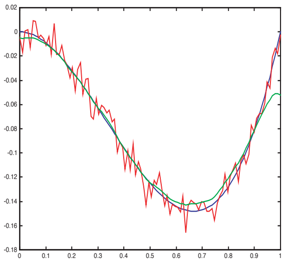
Yet, with this explicit expression it is very easy to carry out the estimation procedure. Here is an example.

`Numerical example`



Denoised, lambda=0.1
Error=2.8119e-04

In black is the unknown function f (as this is a simulation we know the unknown f). The red wiggly function is the observed function y. The green function is the estimator $\hat{f}$. We see in the second figure that by decreasing the tuning parameter λ the estimator $\hat{f}$ is closer to the unknown truth f.



Denoised, lambda=0.05
Error=7.8683e-05

16.4 Estimating a function of bounded variation

Suppose again $\mathcal{X} = [0, 1]$. Instead of taking the penalty

$$\lambda^2 \int_0^1 |\mathrm{f}'(x)|^2 dx$$

one may alternatively choose

$$\lambda^2 \int_0^1 |\mathrm{f}'(x)| dx$$

i.e. without the square. This seems like a minor modification but it makes a huge difference. One may further relax the assumption of differentiability and define the total variation of the function f as

$$\mathrm{TV}(\mathrm{f}) := \sum_{i=2}^{n} |\mathrm{f}(x_i) - \mathrm{f}(x_{i-1})|,$$

where we assume that the x_i are ordered: $x_1 < \ldots < x_n$. This leads to the estimator

$$\hat{f} := \arg\min_{\mathrm{f}} \biggl\{ \|Y - \mathrm{f}\|_2^2 + \lambda^2 \mathrm{TV}(\mathrm{f}) \biggr\}.$$

In analogy, if Y denotes the heights of mountains at locations where a road is to be built, one tries to even out the mountains and valleys using the total variation penalty (so that the road will not have steep slopes) and at the same time move little earth measured in terms of the least squares loss. This can be done in an iterative manner by filling the valley where the slope is steepest, or digging away the mountain with steepest slope. The estimator is locally adaptive: by increasing λ only local changes are made. This is in contrast with the estimator of Section 16.2 or its continuous version in Section 16.3. One can see in the numerical example of Section 16.3 that changing λ has a non-local effect.

One may write the estimator as solution of a linear least squares problem with an ℓ_1-penalty on the coefficients. This will relate the problem with that of Section 16.5. The results there will help to understand why removing the square in the penalty drastically changes the estimator.

Let us define $\mathrm{f}(x_0) =: 0$. Then for $i = 1, \ldots, n$

$$\begin{aligned} \mathrm{f}(x_i) &= \sum_{j=1}^{i} \mathrm{f}(x_j) - \mathrm{f}(x_{j-1}) \\ &= \sum_{i=1}^{n} \underbrace{\biggl(\mathrm{f}(x_j) - \mathrm{f}(x_{j-1}) \biggr)}_{:= b_j} \underbrace{\mathrm{l}\{j \le i\}}_{:= \xi_{i,j}} \\ &= \sum_{i=1}^{n} b_j \xi_{i,j}. \end{aligned}$$

Putting the coefficients b_j, $j = 1, \ldots, n$ in a vector $b = (b_1, \ldots, b_n)^T$ and the $\xi_{i,j}$ in a matrix

$$X := \begin{pmatrix} \xi_{1,1} & \cdots & \xi_{1,n} \\ \vdots & \ddots & \vdots \\ \xi_{n,1} & \cdots & \xi_{n,n} \end{pmatrix}$$

we see that the total variation penalized estimator is

$$\hat{f} = \arg\min_{\mathrm{f} = Xb} \biggl\{ \|Y - \mathrm{f}\|_2^2 + \lambda^2 \|b\|_1 \biggr\}$$

where $\|b\|_1 = \sum_{j=1}^n |b_j|$ denotes the ℓ_1-norm of the vector $b \in \mathbb{R}^n$. This rewriting makes clear that there are as many parameters as there are observations, namely n. The penalty takes care that despite this fact one will not overfit the data (when λ is not too small).

We end this section with a small trip to the case where $\mathcal{X}$ is two-dimensional, say $\mathcal{X} = [0,1]^2$. Then the unknown f is an *image* say, and the observations are

$$Y_{i_1,i_2} = f(x_{i_1,i_2}) + \epsilon_{i_1,i_2}.$$

Consider $n = m^2$ observations and say they are on a regular grid

$$x_{i_1,i_2} = \left(\frac{i_1}{m}, \frac{i_2}{m}\right),\ i_1 = 1, \ldots, m,\ i_2 = 1, \ldots, m.$$

Now, how can we model smoothness of an image? Again, one may opt to work with the squares of derivatives, a counterpart of $\int |\mathrm{f}'(x)|^2 dx$ for the one-dimensional case. However, as in the one-dimensional case, taking squares has non-local effects. The penalized least squares reconstruction of the image will then look *blurred*. For example, if the image is a landscape with lakes and rivers, there are sharp boundaries which will be blurred by taking a penalty based on squared derivatives. An alternative is again a total variation penalty. There are several definitions around for total variation in dimension larger than 1. One possibility in our 2-dimensional case is to define

$$\mathrm{TV}(\mathbf{f}) := \sum_{i_1=2}^{m} \sum_{i_2=2}^{m} |\Delta \mathrm{f}(x_{i_1,i_2})|,$$

where

$$\Delta \mathrm{f}(x_{i_1,i_2}) := \mathrm{f}\left(\frac{i_1}{n}, \frac{i_2}{m}\right) - \mathrm{f}\left(\frac{i_1 - 1}{m}, \frac{i_2}{m}\right) - \mathrm{f}\left(\frac{i_1}{m}, \frac{i_2 - 1}{m}\right) + \mathrm{f}\left(\frac{i_1 - 1}{m}, \frac{i_2 - 1}{m}\right).$$

The image reconstruction algorithm is

$$\hat{f} := \arg\min_{\mathrm{f}} \left\{ \sum_{i_1=1}^{m} \sum_{i_2=1}^{m} \left(Y_{i_1,i_2} - \mathrm{f}(x_{i_1,i_2}) \right)^2 + \lambda^2 \mathrm{TV}(\mathrm{f}) \right\}.$$

This estimator can again be written as a least squares estimator of a linear function with an ℓ_1-penalty on the coefficients.

16.5 The ridge and Lasso penalty

In the linear model one has data $(x_1, Y_1), \ldots, (x_n, Y_n)$ with $x_i \in \mathbb{R}^p$ a p-dimensional row vector and $Y_i \in \mathbb{R}$ $(i = 1, \ldots, n)$ and one wants to find a good linear approximation using the least squares loss function

$$b \mapsto \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} x_{i,j} b_j \right)^2,$$

Define the design matrix

$$X := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}$$

and the vector of responses

$$Y := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

Set the vector of coefficients to $b := (b_1, \ldots, b_p)^T$. Then

$$\sum_{i=1}^n \biggl(Y_i - \sum_{j=1}^p x_{i,j} b_j \biggr)^2 = \|Y - Xb\|_2^2.$$

If $p \ge n$ and X has rank n, minimizing this over all $b \in \mathbb{R}^p$ gives a "perfect" solution $\hat\beta_{\rm LS}$ with $X\hat\beta_{\rm LS} = Y$. This solution just reproduces the data and is therefore of no use. We say that it *overfits*.

Definition 16.5.1 *The ridge regression estimator is*

$$\hat\beta_{\rm ridge} := \arg\min_{b \in \mathbb{R}} \biggl\{ \|Y - Xb\|_2^2 + \lambda^2 \|b\|_2^2 \biggr\},$$

where $\lambda > 0$ is a regularization parameter.

Definition 16.5.2 *The Lasso*[1] *estimator is*

$$\hat\beta_{\rm Lasso} := \arg\min_{b \in \mathbb{R}} \biggl\{ \|Y - Xb\|_2^2 + 2\lambda \|b\|_1 \biggr\},$$

where $\lambda > 0$ is a regularization parameter and $\|b\|_1 := \sum_{j=1}^p |b_j|$ is the ℓ_1-norm of b .

Note Recall the definition of the Bayesian MAP estimator as given in Subsection 10.5.3. Consider the model $Y = X\beta + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. The ridge regression estimator is the MAP estimator using as prior $\beta_1, \ldots, \beta_p$ i.i.d. $\sim \mathcal{N}(0, \tau^2)$. with $\tau = \sigma/\lambda$. The Lasso estimator is the MAP using as prior $\beta_1, \ldots, \beta_p$ i.i.d. $\sim$ Laplace$(0, \tau^2)$ where the standard deviation τ is $\tau = \sqrt{2}\sigma^2/\lambda$. See also Section 10.6.

Remark As λ grows the ridge estimator shrinks the coefficients. They will however not be set exactly to zero. The coefficients of the Lasso estimator shrink as well, and some - or even many - are set exactly to zero. The ridge estimator can be useful if p is moderately large. For very large p the Lasso is often preferred. The idea is that one should not try to estimate something when the signal is below the noise level. Instead, then one should simply put it to zero.

Remark Both ridge estimator and Lasso are biased. As λ increases the bias increases, but the variance decreases.

Remark The regularization parameter λ is for example chosen by using "cross validation" or (information) theoretic or Bayesian arguments. Below, we will see that for the Lasso a choice of order $\sqrt{n \log p}$ is theoretically justified.

[1] This is relatively recent methodology, introduced as Lasso by Tibshirani, R., 1996: Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* 58, 267-288

We now investigate further expressions for both ridge estimator and Lasso.

Lemma 16.5.1 *The ridge estimator $\hat{\beta}_{\text{ridge}}$ is given by*

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda^2 I)^{-1} X^T Y.$$

Proof. We have

$$\begin{aligned} \tfrac{1}{2}\frac{\partial}{\partial b}\Bigl\{ \|Y - Xb\|_2^2 + \lambda^2 \|b\|_2^2 \Bigr\} &= -X^T(Y - Xb) + \lambda^2 b \\ &= -X^T Y + \Bigl(X^T X + \lambda^2 I \Bigr) b. \end{aligned}$$

The estimator $\hat{\beta}_{\text{ridge}}$ puts this to zero. □

Corollary 16.5.1 *Suppose orthonormal design: $X^T X = nI$ (thus $p \le n$ necessarily). Then*

$$\hat{\beta}_{\text{ridge}} = X^T Y/(n + \lambda^2).$$

After some calculations, as in Example 5.2.1, one sees that when $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. with mean zero and variance σ^2, then

$$\mathbb{E}\|X\hat{\beta}_{\text{ridge}} - f\|_2^2 = \underbrace{\left[\frac{\lambda^2/n}{1 + \lambda^2/n}\right]^2 \|X\beta^*\|_2^2}_{(\text{bias})^2} + \underbrace{\left[\frac{1}{1 + \lambda^2/n}\right]^2 p\sigma^2}_{\text{variance}} + \underbrace{\|X\beta^* - f\|_2^2}_{\substack{\text{misspecification}\\ \text{error}}},$$

where $X\beta^*$ is the projection of f on the space spanned by the columns of X. We see that in order to trade off bias and variance, we have to know the variance σ^2, but what is worse, we also have to know the (bias)^2 $\|X\beta^*\|_2^2$. *But the bias is unknown as f is unknown. Thus, we are facing the same problems as in Example 5.2.1.*

For the Lasso estimator there is no simple expression in general. We therefore only consider the special case of orthonormal design.

Lemma 16.5.2 *Suppose $X^T X = nI$ (thus $p \le n$ necessarily). Define $Z := X^T Y$. Then for $j = 1, \ldots, p$*

$$\hat{\beta}_{\text{Lasso},j} = \begin{cases} Z_j/n - \lambda/n & Z_j \ge \lambda \\ 0 & |Z_j| \le \lambda \\ Z_j/n + \lambda/n & Z_j \le -\lambda \end{cases}.$$

Proof. Write $\hat{\beta}_{\text{Lasso}} =: \hat{\beta}$ for short. We can write

$$\|Y - Xb\|_2^2 = \|Y\|_2^2 - 2b^T X^T Y + nb^T b = -2b^T Z + nb^T b.$$

Thus for each j we minimize

$$-2b_j Z_j + nb_j^2 + 2\lambda|b_j|.$$

If $\hat{\beta}_j > 0$ it must be a solution of putting the derivative of the above expression to zero:

$$-Z_j + n\hat{\beta}_j + \lambda = 0,$$

or

$$\hat{\beta}_j = Z_j/n - \lambda/n.$$

Similarly, if $\hat{\beta}_j < 0$ we must have

$$-Z_j + n\hat{\beta}_j - \lambda = 0.$$

Otherwise $\hat{\beta}_j = 0$. □

Some notation
○ For a vector $z \in \mathbb{R}^p$ we let $\|z\|_\infty := \max_{1 \le j \le p} |z_j|$ be its ℓ_∞-norm.
○ We let $X_1, \ldots, X_p$ denote the columns of X.
○ For a subset $S \subset \{1, \ldots, p\}$ we let $X\beta_S^*$ be the best linear approximation of $f := EY$ using the variables in S, i.e., $X\beta_S^*$ is the projection in $\mathbb{R}^n$ of f on the linear space $\{\sum_{j \in S} X_j b_{S,j} : b_S \in \mathbb{R}^{|S|}\}$.

In the next theorem we again assume orthogonal design. For general design, one needs so-called *restricted eigenvalues* or *compatibility conditions* between ℓ_2-norms and ℓ_1-norms.

Theorem 16.5.1 *Consider again fixed design with $X^TX = nI$. Let $f = EY$ and $\epsilon = Y - f$. Fix some level $\alpha \in (0,1)$ and suppose that for some λ_α it holds that*

$$\mathbb{P}(\|X^T\epsilon\|_\infty > \lambda_\alpha) \le \alpha.$$

Then for $\lambda > \lambda_\alpha$ we have with probability at least $1 - \alpha$

$$\|X\hat{\beta}_{\text{Lasso}} - f\|_2^2 \le \min_S \biggl\{ \underbrace{\frac{(\lambda + \lambda_\alpha)^2}{n}|S|}_{\substack{\text{estimation}\\ \text{error}}} + \underbrace{\|X\beta_S^* - f\|_2^2}_{\substack{\text{approximation}\\ \text{error}}} \biggr\}.$$

Proof. Write $\hat{\beta} := \hat{\beta}_{\text{Lasso}}$ and $f = X\beta$. On the set where $\|X^T\epsilon\|_\infty \le \lambda_\alpha$ we have
- $n|\beta_j| > \lambda + \lambda_\alpha \Rightarrow n|\hat{\beta}_j - \beta_j| \le \lambda + \lambda_\alpha$,
- $n|\beta_j| \le \lambda + \lambda_\alpha \Rightarrow \ \ |\hat{\beta}_j - \beta_j| \le |\beta_j|$.
So with probability at least $(1 - \alpha)$,

$$\begin{aligned} \|X\hat{\beta}_{\text{Lasso}} - f\|_2^2 &\le \frac{(\lambda + \lambda_\alpha)^2}{n}\biggl(\#\{j : \ n|\beta_j| > \lambda + \lambda_\alpha\}\biggr) + \sum_{n|\beta_j| \le \lambda + \lambda_\alpha} n\beta_j^2 \\ &= \min_S \biggl\{ \frac{(\lambda + \lambda_\alpha)^2}{n}|S| + \|X\beta_S^* - f\|_2^2 \biggr\}. \end{aligned}$$

□

If we compare the result of Theorem 16.5.1 with result iii) of Lemma 12.3.1 concerning the ordinary least squares estimator, we see that the Lasso exhibits an automatic trade-off between approximation error and estimation error. This is called *adaptation.* As we will see below, the parameter λ can typically be chosen of order $\sqrt{n \log p}$. Therefore the price paid for not knowing a priori which subset of the coefficients is relevant is of order $\log p$. This is generally considered as being a relatively low price.

Note One may write

$$\|X\beta_S^* - f\|_2^2 = \|X\beta_S^* - X\beta^*\|_2^2 + \|X\beta^* - f\|_2^2$$

where $X\beta^*$ is the projection of f on the space spanned by the columns of X. Thus, the "approximation error" actually consists of two terms. The second term is the misspecification error: it vanishes when the linear model is well-specified.

Corollary 16.5.2 *Suppose that $f = X\beta$ where β has $s := \#\{j : \ \beta_j \neq 0\}$ non-zero components. Then under the conditions of the above theorem, with probability at least $1 - \alpha$*

$$\|X(\hat\beta_{\text{Lasso}} - \beta)\|_2^2 \leq \frac{(\lambda + \lambda_\alpha)^2}{n} s.$$

The above corollary tells us that the Lasso estimator adapts to favourable situations where β has many zeroes (i.e. where β is *sparse*).

To complete the story, we need to study a bound for λ_α. It turns out that for many types of error distributions, one can take λ_α of order $\sqrt{n \log p}$. We show this for the case of i.i.d. $\mathcal{N}(0, \sigma^2)$ noise. For that purpose, we start with a bound for the tails of a standard normal random variable.

Lemma 16.5.3 *Suppose $Z \sim \mathcal{N}(0,1)$. Then for all $t > 0$*

$$\mathbb{P}(Z \geq \sqrt{2t}) \leq \exp[-t].$$

Proof. First check that for all $u > 0$

$$E \exp[uZ] = \exp[u^2/2].$$

Then by Chebyshevs inequality

$$\mathbb{P}(Z > \sqrt{2t}) \leq \exp[u^2/2 - u\sqrt{2t})].$$

Now choose $u = \sqrt{2t}$. □

Corollary 16.5.3 *Let $Z_1, \ldots, Z_p$ be p (possibly dependent) standard normal random variables. Then for all t*

$$\begin{aligned} \mathbb{P}\left(\max_{1 \leq j \leq p} |Z_j| \geq \sqrt{2(\log(2p) + t)} \right) &\leq \sum_{j=1}^p \mathbb{P}(|Z_j| \geq \sqrt{2(\log(2p) + t)}) \\ &\leq 2p \exp[-(\log(2p) + t)] = \exp[-t]. \end{aligned}$$

Remark In the above corollary we used $\mathbb{P}(A \cup B) \le \mathbb{P}(A) + \mathbb{P}(B)$. This is called the *union bound.*

Corollary 16.5.4 *Let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. $\mathcal{N}(0, \sigma^2)$ and let $X = (X_1, \ldots, X_p)$ with $X_1, \ldots, X_p$ be fixed vectors in $\mathbb{R}^n$ with $\|X_j\|_2 = n$ for all j. Then for $0 < \alpha < 1$*

$$\mathbb{P}\left(\|X^T \epsilon\|_\infty \ge \sigma \sqrt{2n(\log(2p/\alpha))} \right) \le \alpha.$$

Remark. The value $\alpha = \frac{1}{2}$ thus gives a bound for the median of $\|X\hat{\beta}_{\text{Lasso}} - f\|_2^2$. In the case of Gaussian errors one may use "concentration of measure" to deduce that $\|X\hat{\beta}_{\text{Lasso}} - f\|_2^2$ is "concentrated" around its median.

16.6 Conclusion

We have seen in this chapter that several concepts from classical statistics also play their role in the modern version of statistics where the parameter is possibly high dimensional. The classical least squares methodology keeps its prominent place, but now it is equipped with a regularization penalty. More generally, M-estimators (e.g. maximum likelihood estimators) can also be used in high dimensions, applying again some regularization technique. The bias-variance decomposition continues to play its role too, for instance leading to guidelines for choosing tuning parameters.

Shrinkage estimators play an important role in high-dimensional statistics. This is also related to the result of Section 11.4 where we have seen that the sample average in dimension higher than 2 is inadmissible as it can be improved by a shrinkage estimator.

Complexity regularization can typically be seen as a Bayesian MAP approach. One may also use the a posteriori mean as estimator etc. Today, Bayesian approaches to high-dimensional and non-parametric problems are very important and successful.

Complexity regularization is typically invoked for the construction of adaptive estimators. An adaptive estimator mimics the situation where we knew beforehand the complexity of the underlying target to be estimated. To evaluate the performance of an adaptive estimator, one uses as benchmark the case where the (hopefully low) complexity of the target is indeed known. Thus the benchmark comes from classical statistical theory.

Chapter 17

Literature

- J.O. Berger (1985) *Statistical Decision Theory and Bayesian Analysis* Springer
  A fundamental book on Bayesian theory.

- P.J. Bickel, K.A. Doksum (2001) *Mathematical Statistics, Basic Ideas and Selected Topics* Volume I, 2nd edition, Prentice Hall
  Quite general, and mathematically sound.

- D.R. Cox and D.V. Hinkley (1974) *Theoretical Statistics* Chapman and Hall
  Contains good discussions of various concepts and their practical meaning. Mathematical development is sketchy.

- A. DasGupta (2011) *Probability for Statistics and Machine Learning*, Springer
  Contains all the probability theory background needed. (Look out for the upcoming book *Statistical Theory, a Comprehensive Course* by the same author.)

- J.G. Kalbfleisch (1985) *Probability and Statistical Inference* Volume 2, Springer
  Treats likelihood methods.

- L.M. Le Cam (1986) *Asymptotic Methods in Statistical Decision Theory* Springer
  Treats decision theory on a very abstract level.

- E.L. Lehmann (1983) *Theory of Point Estimation* Wiley
  A “klassiker”. The lecture notes partly follow this book

- E.L. Lehmann (1986) *Testing Statistical Hypothesis* 2nd edition, Wiley
  Goes with the previous book.

- J.A. Rice (1994) *Mathematical Statistics and Data Analysis* 2nd edition, Duxbury Press
  A more elementary book.

- M.J. Schervish (1995) *Theory of Statistics* Springer
  Mathematically exact and quite general. Also good as reference book.

- R.J. Serfling (1980) *Approximation Theorems of Mathematical Statistics* Wiley
  Treats asymptotics.

- A.W. van der Vaart (1998) *Asymptotic Statistics* Cambridge University Press
  Treats modern asymptotics and e.g. semiparametric theory

- L. Wasserman (2004) *All of Statistics. A Concise Course in Statistical Inference* Springer.
  Contains a wide range of topics in mathematical statistics and machine learning.