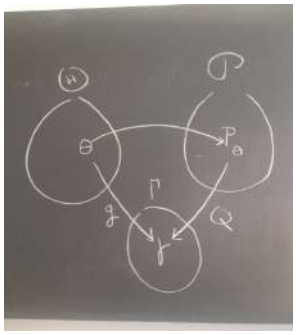
Mathematical Statistics



Sara van de Geer

September 2018

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.

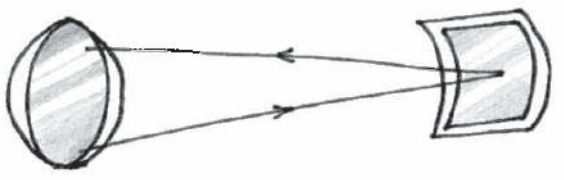


Fig. 1

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.

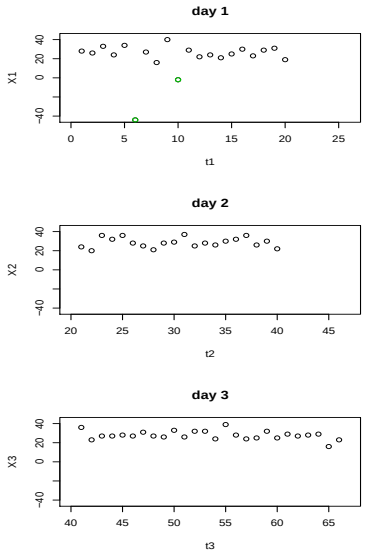
distance: 7.44373 km.

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 $g : \Theta \to \Gamma$ 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 $Z =$ 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

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

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:

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

Stochastic block model, just to make the point of this more complicated statistical model

⇒ **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 $j \neq k$

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

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

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

different modeling depending on whether in same network or not. ⇒ Network modeled by sparse matrix.

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

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

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. (p individuals)

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

If p is large again because of the permutations this problem becomes difficult to solve.

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 = rainfall, Z_2 = tea consumption, Z_3 = number of tall people, Z_4 = 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

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

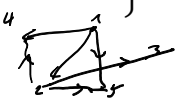
noise

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

p variables

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

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.

We use the "dot" notation e.g.:

$F(x) = P(X \leq x)$ $\quad x \in \mathbb{R}^d$ $\quad$ short cut ⇒ $\quad F = F(\cdot) = P(X \leq x)$

$X \sim P$:= X has distribution P.

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*. → will then come out in chapter 10.

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

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

This is an estimator of the theoretical distribution function

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

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 $X_{(1)} \leq \cdots \leq X(n)$ 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

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

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

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

An estimator of F is the empirical distribution function

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

Note that when knowing only $\hat{F}_n$, one can reconstruct the order statistics $X_{(1)} \leq \ldots \leq X_{(n)}$, 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 $F_\theta := P_\theta(X \leq \cdot)$ 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 $\theta > 1$

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

$$= \left[\frac{\theta}{(1+x)^{1+\theta}} + -(1+\theta)\cdot \frac{\theta}{(1+x)^{2+\theta}} \right]_0^\infty$$

$$- \theta + (1+\theta)\cdot\theta = \cancel{\theta} + \theta^2 - \cancel{\theta}$$

with inverse

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

The method of moments estimator would thus be

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

So not good for every case ! The function for the moment must exist and be possible injective and surjective over the entire parameter space.

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.

Notice: a measure ν is dominated by μ if $\nu \ll \mu$, that is if $\mu(N)=0 \Rightarrow \nu(A)=0$. A family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$

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

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

Notice that you should take the log when the support makes sure you never end up with the untractable log(0).

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 $\log x \leq x - 1$, $x > 0$, for all $\vartheta \in \Theta$

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

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

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

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

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

so that

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

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*

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

so

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

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

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

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

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

Now, since for all $z \neq 0$

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

we have

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

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

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

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

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

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

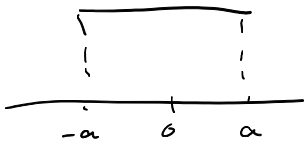
$$2\Big\{\log L_{\mathbf{X}}(\hat{\mu}) - \log L_{\mathbf{X}}(\mu)\Big\} = n(\bar{X} - \mu)^2.$$

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.*

As an exercise do MLE location $\rightrightarrows$ uniform, Laplace, normal dist. with var $\equiv 1$

var $\equiv 1$ for Laplace $\Rightarrow$ $f_0(x) = \frac{1}{\sqrt{2}} \exp\left[\sqrt{2}|x|\right]$

for uniform $f_0(x) = \frac{\mathbb{1}_{[-a,a]}(x)}{2a}$



with $a = \sqrt{3}$

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*

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

[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

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

□

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

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

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

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

Here

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

Example 3.2.1 Histograms

Let $X_1, \dots, X_n$ 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 $j = 1, \dots, k$,

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

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

3.3 The Poisson distribution

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

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

(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*

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

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,

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

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

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

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

look at this post if really interested. I believe the answer is hidden there.

https://stats.stackexchange.com/questions/232085/conditional-distribution-of-uniform-random-variable-given-order-statistic

$$f_{X_1,Z} = \begin{cases} \frac{f(z)}{f(z)} = 1 & \text{if } x_1 \geq z \text{ then} \\ \frac{f(x_1)\cdot f(z)}{2F(z)\cdot f(z)} & \text{if } x_1 < z \end{cases}$$

constant when integrating just

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).

Important Def.

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.

It means that knowing the result of the given data the probability distribution of a Random Variable is not affected by the exact knowledge of the parameter value. So if this means the statistic is sufficient.

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,

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

Let S be sufficient (Given $S=s$). Let $\mathbf{X}_s^*$ be a sample from the distribution of X given $S=s$. Then X_s^* has the same distribution as X.

$$P(X_s^* \in A) = E_\theta P(X \in A | S) = E_\theta P(X \in A) = P_\theta(X \in A)$$

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 $(p_1, \ldots, p_n) = (\frac{1}{n}, \ldots, \frac{1}{n})$:*

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

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$(2, \theta)$-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 $(x_{q_1}, \ldots, x_{q_n}) = s$,

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

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*

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

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 $Z = z \in (0, 2\theta)$ is now

$$f_{X,X_2}(x_1, x_2|Z = z; \theta) = \begin{cases} \frac{1}{z} & 0 \le z \le \theta \\ \frac{1}{2\theta - z} & \theta \le z \le 2\theta \end{cases}.$$

This 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 $a_1, a_2, \ldots$ $\forall$ θ (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

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

(⇒) 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 $g_\theta(s) = Q_\theta(S = s)$.

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

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

comes from inserting in (1) above.

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

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

with

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

and

$$h(x_1, \ldots, x_n) := 1\!\!1\{0 \le \min\{x_1, \ldots, x_n\}\}.$$

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*

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

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

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

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

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

from this and note (1) we see immediately that if $X_1, \ldots, X_n$ i.i.d. Poisson(θ) then $\sum X_i$ sufficient.

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

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

⇒ involved $x \wedge m$. how is it defined? number of successes or in successes.

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 1 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

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

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

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

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

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

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

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

□

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

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

and

$$\begin{aligned} \mathrm{var}_\theta(T(X)) &= \ddot{d}_0(\gamma) = \frac{\ddot{d}(c^{-1}(\gamma))}{[\dot{c}(c^{-1}(\gamma))]^2} - \frac{\dot{d}(c^{-1}(\gamma))\ddot{c}(c^{-1}(\gamma))}{[\dot{c}(c^{-1}(\gamma))]^3} \\ &= \frac{\ddot{d}(\theta)}{[\dot{c}(\theta)]^2} - \frac{\dot{d}(\theta)\ddot{c}(\theta)}{[\dot{c}(\theta)]^3} \\ &= \frac{1}{[\dot{c}(\theta)]^2}\left(\ddot{d}(\theta) - \frac{\dot{d}(\theta)}{\dot{c}(\theta)}\ddot{c}(\theta)\right). \end{aligned} \tag{4.2}$$

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

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

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 $\mathrm{var}_\theta s_\theta(X)$, 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 $\mathrm{var}_\theta(s_\theta(X)) = -E_\theta \dot{s}_\theta(X)$ implies

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

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.

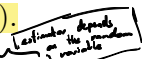
Chapter 5

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*

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

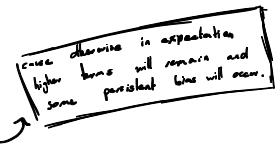
The estimator T *is called* unbiased *if*

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

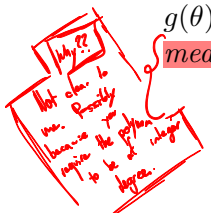
Thus, unbiasedness means that there is no systematic error: $E_\theta T = g(\theta)$. We require this **for all** θ.

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 $X \sim \text{Poisson}(\theta)$. *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 $e^{-\theta}$ *is for instance*

$$T(X) = 1\{X = 0\}.$$

As another example, suppose the parameter of interest is

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

An unbiased estimator is

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

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*

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

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

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

□

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]

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

It follows that

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

Moreover

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

so that

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

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 *(Uniform Minimum Variance Unbiased) if for any other unbiased estimator T,*

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

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:

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

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

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

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

Proof. It holds that

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

and

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

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

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

□

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

$$T(X_1, \ldots, X_n) := 1\{X_1 = 0\}.$$

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,

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

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*

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

Its density is thus

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

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,

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

where

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

and

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

It follows that

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

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*

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

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

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

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

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

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

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

$$\leq \left[E_\theta \left(\frac{p_{\theta+h}(X) - p_\theta(X)}{h p_\theta(X)} - s_\theta(X) \right) \right]^2 \to 0.$$

□

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

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

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

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*

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

Moreover,

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

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

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

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

$$\begin{aligned} & \left| E_\theta \Big(T(X) - g_\theta \Big) \left(\frac{p_{\theta+h}(X) - p_\theta(X)}{h p_\theta(X)} - s_\theta(X) \right) \right|^2 \\ \leq \; & \mathrm{var}_\theta(T) \left[E_\theta \left(\frac{p_{\theta+h}(X) - p_\theta(X)}{h p_\theta(X)} - s_\theta(X) \right) \right]^2 \\ \to \; & 0. \end{aligned}$$

Thus,

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

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

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

□

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

$$\dot{s}_\theta(x) = 1/\theta - x,$$

and hence

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

The CRLB for n observations is thus

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

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 Cramér 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

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

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

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

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

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

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

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

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

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

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 $a \in \mathbb{R}^k$,

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

Combining this with the auxiliary lemma gives

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

□

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*

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

But

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

So for all a,

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

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

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

$$q^F_{\sup}(u) := \sup\{x : \ F(x) \leq u\},$$

and

$$q^F_{\inf}(u) := \inf\{x : \ F(x) \geq u\} := F^{-1}(u).$$

It holds that

$$F(q^F_{\inf}(u)) \geq u$$

and, for all $h > 0$,

$$F(q^F_{\sup}(u) - h) \leq u.$$

Hence

$$F(q^F_{\sup}(u)-) := \lim_{h \downarrow 0} F(q^F_{\sup}(u) - h) \leq u.$$

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 $P_\theta(\phi(X,\gamma_0) = 1) \leq \alpha$ 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

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

so it is a statistic based on a statistic.

The constant c is called the *critical value*.

To test $H_{\gamma_0} : \gamma = \gamma_0$,
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

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

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

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

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 $\gamma_0 = g(\theta_0)$:

$$\begin{aligned} \mathbb{P}_{\theta_0}(\phi(\mathbf{X}, g(\theta_0)) = 1) &= P_{\theta_0}(Z(\mathbf{X}, g(\theta_0)) > q_R) + \mathbb{P}_{\theta_0}(Z(\mathbf{X}), g(\theta_0)) < q_L) \\ &= 1 - G(q_R) + G(q_L-) \leq 1 - \left(1 - \frac{\alpha}{2}\right) + \frac{\alpha}{2} = \alpha. \end{aligned}$$

QED

And notice our pivot does not depend on θ such that the test statistics are valid across methods

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 $\theta \in \Theta$,

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

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:*

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

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*

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

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

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

Then

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

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 group x	treatment group y	rank(x)	rank(y)
5	6	7	8
0	-5	3	2
16	-6	10	1
2	1	5	4
9	4	9	6
32	0	34	

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

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

we assumed the data to be i.i.d.

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

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

No Cov by independence assumption.

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

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

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

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

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

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

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

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

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

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

(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

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

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,

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

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

$$\mathbb{P}\Big((Z_1, \ldots, Z_N) \in A,\ Z_1 < \ldots < Z_N\Big) = \mathbb{P}\Big((Z_{q_1} \ldots, Z_{q_N}) \in A,\ Z_{q_1} < \ldots < Z_{q_N}\Big).$$

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

$$\mathbb{P}\Big((Z_{(1)}, \ldots, Z_{(n)}) \in A\Big) = N!\mathbb{P}\Big((Z_{q_1} \ldots, Z_{q_N}) \in A,\ Z_{q_1} < \ldots < Z_{q_N}\Big) = N!\mathbb{P}\Big((Z_{(1)}, \ldots, Z_{(N)}) \in A,\ \mathbf{R} = \mathbf{r}\Big),$$

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

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

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

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

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

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

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 $N = 3$:

$$(z_1, z_2, z_3) = (5, 6, 4)$$
$$(r_1, r_2, r_3) = (2, 3, 1)$$
$$(q_1, q_2, q_3) = (3, 1, 2)$$

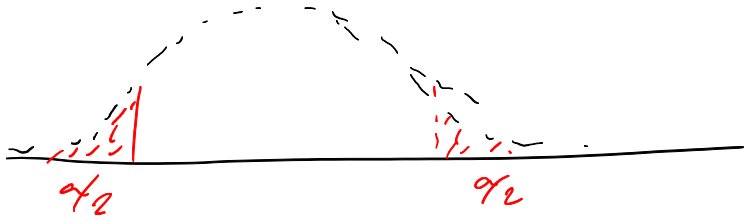
6.5.3 Comparison of Student's test and Wilcoxon's test

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.

Will reappear in section 14.6.

Given the Wilcoxon test it is then possible to prove it by:

(i) Either simulate from the null-distribution of Z.



Reject if $Z \in$

(ii) Calculate

$\text{Var}_{H_0}(Z)$ so that

$$Z_{\text{standard-Norm}} = \frac{Z - E_{H_0}(Z)}{\sqrt{\text{Var}_{H_0}(Z)}} \xrightarrow{D_{H_0}} N(0,1)$$

Notice that despite obs. dependent it is possible to prove that CLT holds

Check at the following Notes:
https://web.stanford.edu/~lmackey/stats300a/doc/stats300a-fall15-lecture13.pdf

Much more exhaustive than the current.

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

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

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 $E_\theta \phi'(X) \leq E_\theta \phi(X)$ $\forall\ \theta \in \Theta_1$.*

7.1 The Neyman Pearson Lemma

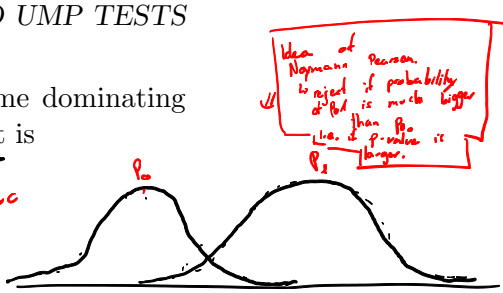
We consider testing
$H_0 : \theta = \theta_0$
against the alternative
$H_1 : \theta = \theta_1$.

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

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



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

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

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

Proof.

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

□

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:

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

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., $t_0 - 1 = q^G_{\inf}(\alpha)$ with $q^G_{\inf}$ 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

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

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

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

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

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

Hence

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

where t is some constant (depending on c, θ_0 and θ_1). Therefore, $\phi = \phi_{\text{NP}}$. 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 $\vartheta > \theta$

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

which is increasing in t. Moreover, we have

$$\text{(1)} \quad \int \bar{p}_\vartheta d\bar{\nu} = \int \bar{p}_\theta d\bar{\nu} = 1.$$

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

$$\text{(2)} \quad \begin{cases} \frac{\bar{p}_\vartheta(t)}{\bar{p}_\theta(t)} \le 1 & \text{for } t \le s_0 \\ \frac{\bar{p}_\vartheta(t)}{\bar{p}_\theta(t)} \ge 1 & \text{for } t \ge s_0 \end{cases}.$$

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

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

which is strictly increasing in θ. Then $T(\mathbf{X}) = \sum_{i=1}^n X_i$.

Right-sided alternative
$H_0 : \theta \leq \theta_0$,
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
$H_0 : \theta \geq \theta_0$,
against $H_1 : \theta < \theta_0$.

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
$H_1 : \theta \neq \theta_0$.

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 $\vartheta \in \Theta_1$,*

$$E_\theta \phi(X) \leq E_\vartheta \phi(X).$$

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 $E_\theta \phi'(X) \leq E_\theta \phi(X)$ $\forall\, \theta \in \Theta_1$.*

We return to the special case where $\Theta \subset \mathbb{R}$ is an interval. We consider testing
$H_0 : \theta = \theta_0$,
against
$H_1 : \theta \neq \theta_0$.

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

$$\phi(T(x)) := \begin{cases} 1 & \text{if } T(x) < t_L \text{ or } T(x) > t_R \\ q_L & \text{if } T(x) = t_L \\ q_R & \text{if } T(x) = t_R \\ 0 & \text{if } t_L < T(x) < t_R \end{cases},$$

where the constants t_R, t_L, q_R and q_L are chosen in such a way that

$$E_{\theta_0}\phi(X) = \alpha, \quad \left.\frac{d}{d\theta} E_\theta \phi(X)\right|_{\theta=\theta_0} = 0.$$

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 $t_L < t_R$,

$$\begin{aligned}E_\mu\phi(T) &= \mathbb{P}_\mu(T > t_R) + \mathbb{P}_\mu(T < t_L)\\ &= \mathbb{P}_\mu\left(\frac{T - n\mu}{\sqrt{n}\sigma_0} > \frac{t_R - n\mu}{\sqrt{n}\sigma_0}\right) + \mathbb{P}_\mu\left(\frac{T - n\mu}{\sqrt{n}\sigma_0} < \frac{t_L - n\mu}{\sqrt{n}\sigma_0}\right)\\ &= 1 - \Phi\left(\frac{t_R - n\mu}{\sqrt{n}\sigma_0}\right) + \Phi\left(\frac{t_L - n\mu}{\sqrt{n}\sigma_0}\right),\end{aligned}$$

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,

$$\frac{d}{d\mu}E_\mu\phi(T) = \frac{n}{\sqrt{n}\sigma_0}\dot{\Phi}\left(\frac{t_R - n\mu}{\sqrt{n}\sigma_0}\right) - \frac{n}{\sqrt{n}\sigma_0}\dot{\Phi}\left(\frac{t_L - n\mu}{\sqrt{n}\sigma_0}\right),$$

So putting

$$\frac{d}{d\mu}E_\mu\phi(T)\Big|_{\mu=\mu_0} = 0,$$

gives

$$\dot{\Phi}\left(\frac{t_R - n\mu_0}{\sqrt{n}\sigma_0}\right) = \dot{\Phi}\left(\frac{t_L - n\mu_0}{\sqrt{n}\sigma_0}\right),$$

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

$$\begin{aligned}E_{\mu_0}\phi(T) &= 1 - \Phi\left(\frac{t_R - n\mu_0}{\sqrt{n}\sigma_0}\right) + \Phi\left(-\frac{t_R - n\mu_0}{\sqrt{n}\sigma_0}\right)\\ &= 2\left(1 - \Phi\left(\frac{t_R - n\mu_0}{\sqrt{n}\sigma_0}\right)\right).\end{aligned}$$

The requirement $E_{\mu_0}\phi(T) = \alpha$ gives us

$$\Phi\left(\frac{t_R - n\mu_0}{\sqrt{n}\sigma_0}\right) = 1 - \alpha/2,$$

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 ⋆

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
$H_0 :\ \gamma \leq \gamma_0$,
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*

$$\phi(T_1, T_2) = \begin{cases} 1 & \text{if } T_2 > t_0(T_1) \\ q(T_1) & \text{if } T_2 = t_0(T_1) \\ 0 & \text{if } T_2 < t_0(T_1) \end{cases},$$

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*

$$E_{\gamma_0}\left(\phi(T_1, T_2)\middle| T_1\right) = \alpha.$$

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.*

$$\sup_{\gamma \leq \gamma_0} E_{(\beta.\gamma)}\phi(T) = E_{(\beta,\gamma_0)}\phi(T) = \alpha, \ \forall \ \beta.$$

Proof of Result 1.

$$\sup_{\gamma \leq \gamma_0} E_{(\beta,\gamma)}\phi(T) \geq E_{(\beta,\gamma_0)}\phi(T) = E_{(\beta,\gamma_0)}E_{\gamma_0}(\phi(T)|T_1) = \alpha.$$

Conversely,

$$\sup_{\gamma \leq \gamma_0} E_{(\beta,\gamma)}\phi(T) = \sup_{\gamma \leq \gamma_0} E_{(\beta,\gamma)} \underbrace{E_\gamma(\phi(T)|T_1)}_{\leq \alpha} \leq \alpha.$$

Result 2 *The test ϕ is unbiased.*

Proof of Result 2. If $\gamma > \gamma_0$, it holds that $E_\gamma(\phi(T)|T_1) \geq \alpha$, as the conditional test is unbiased. Thus, also, for all β,

$$E_{(\beta,\gamma)}\phi(T) = E_{(\beta,\gamma)}E_\gamma(\phi(T)|T_1) \geq \alpha,$$

i.e., ϕ is unbiased.

Result 3 *Let ϕ' be a test with level*

$$\alpha' := \sup_\beta \sup_{\gamma \leq \gamma_0} E_{(\beta,\gamma)}\phi'(T) \leq \alpha,$$

and suppose moreover that ϕ' is unbiased, i.e., that

$$\sup_{\gamma \leq \gamma_0} \sup_\beta E_{(\beta,\gamma)}\phi'(T) \leq \inf_{\gamma > \gamma_0} \inf_\beta E_{(\beta,\gamma)}\phi'(T).$$

Then, conditionally on T_1, ϕ' has level α'.

Proof of Result 3. As

$$\alpha' = \sup_\beta \sup_{\gamma \leq \gamma_0} E_{(\beta,\gamma)}\phi'(T)$$

we know that

$$E_{(\beta,\gamma_0)}\phi'(T) \leq \alpha', \ \forall \ \beta.$$

Conversely, the unbiasedness implies that for all $\gamma > \gamma_0$,

$$E_{(\beta,\gamma)}\phi'(T) \geq \alpha', \forall \ \beta.$$

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

$$E_{(\beta,\gamma_0)} E_{\gamma_0}\left((\phi'(T) - \alpha') \middle| T_1 \right) = 0, \ \forall \ \beta,$$

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

$$E_{\gamma}(\phi'(T)|T_1) \leq E_{\gamma}(\phi(T)|T_1), \ \forall \ \gamma > \gamma_0.$$

Thus also

$$E_{(\beta,\gamma)}\phi'(T) \leq E_{(\beta,\gamma)}\phi(T), \ \forall \ \beta, \ \gamma > \gamma_0.$$

□

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*
$H_0 : \ \lambda \leq \mu$,
against the alternative
$H_1 : \ \lambda > \mu$.

Define

$$\beta := \log(\mu), \ \gamma := \log(\lambda/\mu).$$

The testing problem is equivalent to
$H_0 : \ \gamma \leq \gamma_0$,
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.

$L(\cdot,\cdot): \Theta \times \mathcal{A} \to \mathbb{R}$

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*

$$R(\theta, d') \leq R(\theta, d), \ \forall\ \theta.$$

Proof. Jensen's inequality says that for a convex function g,

$$E(g(X)) \geq g(EX).$$

Hence, $\forall\ \theta$,

$$\begin{aligned} E\Big(L\Big(\theta, d(X)\Big)\Big|S = s\Big) &\geq L\Big(\theta, E\Big(d(X)|S = s\Big)\Big) \\ &= L(\theta, d'(s)). \end{aligned} \tag{1}$$

By the iterated expectations lemma, we arrive at

$$\begin{aligned} R(\theta, d) &= E_\theta L(\theta, d(X)) \\ &= E_\theta E\Big(L\Big(\theta, d(X)\Big)\Big|S\Big) \\ &\geq E_\theta L(\theta, d'(S)). \end{aligned}$$

$R(\theta, d')$

by (1)

□

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

This proves $\mathrm{MSE}_\theta(T) \geq \mathrm{MSE}_\theta(\tilde{T})$ ⇐ $R(\theta, \tilde{T}) \leq R(\theta, T)\ \forall\ \theta.$

"⇒" rela Hoadley given that $E(\tilde{T}) = E(T)$ and therefore bias equality holds

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

Then this proves bias is the same for the two

$$\Rightarrow \mathrm{var}_\theta(\tilde{T}) \leq \mathrm{var}_\theta(T),\ \forall\ \theta.$$

So that $\mathrm{MSE}_\theta(T) \geq \mathrm{MSE}_\theta(\tilde{T})$

$\mathrm{Var}_\theta(T) \geq \mathrm{Var}_\theta(E(T|S))$

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 $m < n$,

$$\epsilon(m) := \sup_{x_1^*, \ldots, x_m^*} |T(x_1^*, \ldots, x_m^*, X_{m+1}, \ldots, X_n)|.$$

If $\epsilon(m) := \infty$, we say that with m outliers the estimator can break down. The break down point is defined as

$$\epsilon^* := \min\{m :\ \epsilon(m) = \infty\}/n.$$

Share of outliers needed in order for the statistic to break down

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

$$T = \begin{cases} \bar{X} \\ X_{(\frac{n+1}{2})} \quad (n \text{ odd}) \\ \dots \end{cases} .$$

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 $\varepsilon := (\epsilon_1, \dots, \epsilon_n)$. 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:

$$R(\theta, T) = E L_0[T(\varepsilon)].$$

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 $Y_i := X_i - X_n$, $i = 1, \dots, n$, and $\mathbf{Y} := (Y_1, \dots, Y_n)$. 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*

$$T^*(\mathbf{Y}) := \arg\min_v E\left[L_0\Big(v + \epsilon_n\Big)\Big|\mathbf{Y}\right].$$

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

$$R(0,T) = EL_0\Big(T(\mathbf{Y}) + \epsilon_n\Big) = E\ E\left[L_0\Big(T(\mathbf{Y}) + \epsilon_n\Big)\Big|\mathbf{Y}\right].$$

But

$$\begin{aligned} E\left[L_0\Big(T(\mathbf{Y}) + \epsilon_n\Big)\Big|\mathbf{Y}\right] &\geq \min_v E\left[L_0\Big(v + \epsilon_n\Big)\Big|\mathbf{Y}\right] \\ &= E\left[L_0\Big(T^*(\mathbf{Y}) + \epsilon_n\Big)\Big|\mathbf{Y}\right]. \end{aligned}$$

Hence,

$$R(0,T) \geq E\ E\left[L_0\Big(T^*(\mathbf{Y}) + \epsilon_n\Big)\Big|\mathbf{Y}\right] = R(0,T^*).$$

□

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 $\mathbf{Y} = \mathbf{X} - X_n$,

$$\begin{aligned} T^*(\mathbf{Y}) &= \arg\min_v E\left[\Big(v + \epsilon_n\Big)^2\Big|\mathbf{Y}\right] \\ &= -E(\epsilon_n|\mathbf{Y}), \end{aligned}$$

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*

$$p_0(x) = 1\{|x| \le 1/2\}.$$

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

$$\mathbf{p}_0(x_1 - z, \ldots, x_n - z) = 1\{x_{(n)} - 1/2 \le z \le x_{(1)} + 1/2\}.$$

Thus, writing

$$T_1 := X_{(n)} - 1/2, \; T_2 := X_{(1)} + 1/2,$$

the UMRE estimator T^* is

$$T^* = \left(\int_{T_1}^{T_2} 1 \cdot z dz\right) \Big/ \left(\int_{T_1}^{T_2} 1 \cdot dz\right) = \frac{T_1 + T_2}{2} = \frac{X_{(1)} + X_{(n)}}{2}.$$

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*

$$T^*(\mathbf{Y}) := \arg\min_v E\Big[L_0\Big(v + d(\varepsilon)\Big)\Big|\mathbf{Y}\Big].$$

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

$$\begin{aligned} E\Big[L_0\Big(T(\varepsilon)\Big)\Big|\mathbf{Y}\Big] &= E\Big[L_0\Big(T(\mathbf{Y}) + d(\varepsilon)\Big)\Big|\mathbf{Y}\Big] \\ &\geq \min_v E\Big[L_0\Big(v + d(\varepsilon)\Big)\Big|\mathbf{Y}\Big]. \end{aligned}$$

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

$$T^*(\mathbf{Y}) := \arg \min_{a_1 \in \mathbb{R},\ a_2 \in \mathbb{R}_+} E\left[L_0\Big(d_1(\varepsilon) + d_2(\varepsilon)a_1, d_2(\varepsilon)a_2 \Big) \Big| \mathbf{Y} \right].$$

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

$$EL_0\Big(T_1(\varepsilon), T_2(\varepsilon) \Big)$$

$$\begin{aligned}
&= EL_0\Big(T_1(d_1(\varepsilon) + d_2(\varepsilon)\mathbf{Y}), T_2(d_1(\varepsilon) + d_2(\varepsilon)\mathbf{Y}) \Big) \\
&= EL_0\Big(d_1(\varepsilon) + d_2(\varepsilon)T_1(\mathbf{Y}), d_2(\varepsilon)T_2(\mathbf{Y}) \Big) \\
&= EE\left[L_0\Big(d_1(\varepsilon) + d_2(\varepsilon)T_1(\mathbf{Y}), d_2(\varepsilon)T_2(\mathbf{Y}) \Big) \Big| \mathbf{Y} \right] \\
&\geq E \min_{a_1 \in \mathbb{R},\ a_2 \in \mathbb{R}_+} E\left[L_0\Big(d_1(\varepsilon) + d_2(\varepsilon)a_1, d_2(\varepsilon)a_2 \Big) \Big| \mathbf{Y} \right] \\
&= EE\left[L_0\Big(d_1(\varepsilon) + d_2(\varepsilon)T_1^*(\mathbf{Y}), d_2(\varepsilon)T_2^*(\mathbf{Y}) \Big) \Big| \mathbf{Y} \right].
\end{aligned}$$

□

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

$$T^*(\mathbf{Y}) = \arg \min_{a_1 \in \mathbb{R}} E\left[\Big(d_1(\varepsilon) + d_2(\varepsilon)a_1 \Big)^2 \Big| \mathbf{Y} \right].$$

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.