# Chapter 4

# Frequentist approach

We adopt the frequentist approach here, and consider a family of probability distributions $p_\theta(x)$ parameterized by an unknown parameter $\theta$. For simplicity of presentation we first focus on single-parameter estimation and will generalize our results to multi-parameter case in section 4.3.

## 4.1 Optimal unbiased estimator

In order to provide an intuition into the problem of determining the optimal estimator let us start with a simple example.

**Example 4.1** Consider $N$ identically and independently distributed (i.i.d.) random variables: $\boldsymbol{x} = (x_1, \ldots, x_N)$, where $x_i = \theta + w_i$ and $w_i \sim \mathcal{N}(0, \sigma^2)$ is a normally distributed random variable with mean 0 and variance $\sigma^2$. As a result $x_i \sim \mathcal{N}(\theta, \sigma^2)$. More explicitly, we can write the joint probability of observing measurement events $\boldsymbol{x}$ as

$$p_\theta(\boldsymbol{x}) = p_\theta(x_1) \cdot \cdots \cdot p_\theta(x_N), \tag{4.1}$$

where

$$p_\theta(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}}. \tag{4.2}$$

Assume we observe a given sequence of events: $(x_1, \ldots, x_N)$. What will be the optimal way to estimate $\theta$? Natural guess is that we should take the average $\tilde{\theta}(\boldsymbol{x}) = \sum_i x_i/N$, but can we prove this is the optimal choice?

In what follows, we will quantify the optimality of an estimator $\tilde{\theta}$ using its mean squared distance from the true value of the parameter:

$$\Delta^2\tilde{\theta} = \int \mathrm{d}x \, \Big(\tilde{\theta}(x) - \theta\Big)^2 p_\theta(x). \tag{4.3}$$

Since within the frequentist framework the parameter $\theta$ is unknown but fixed, we have to specify some constraints on the class of estimators we will be considering. Notice, that otherwise there is no fundamental limit on precision of estimator as we might define the estimator $\tilde{\theta}(x) = \theta_0$ to be a constant function and if we are lucky and $\theta_0 = \theta$ we have an estimator with zero uncertainty. Of course, it is clear that such estimators are useless in practice. We will therefore require form our estimators to satisfy the *unbiasedness* condition which excludes the above mentioned pathological cases.

**Definition 4.1** (Unbiased estimator)**.** We say that an estimator $\tilde{\theta}$ is unbiased, if and only if for all $\theta$:

$$\langle\tilde{\theta}\rangle = \int \mathrm{d}x \, \tilde{\theta}(x) p_\theta(x) = \theta, \tag{4.4}$$

which is equivalent to saying that on average the estimator returns the true value for all values of parameter $\theta$.

The goal of estimation theory can now be formulated as the task of determining the unbiased estimator that provides the minimum variance—*the minimum variance unbiased estimator*. Interestingly it might happen that such an estimator does not exist, in the sense

that there is no single estimator that is optimal for the whole range of parameters $\theta$ (see Problem **??**).

Recalling that the frequentist approach assumes a fixed but unknown parameter, it is typical that we deal with situation where we known roughly the parameter value to be around some value $\theta_0$ and want to estimate it precisely staying within some small interval around it. It is therefore useful to introduce a weaker condition of local unbiasedness, which will actually be sufficient to derive all the bounds that will follow, and moreover there will be no issue of nonexistence of minimal variance locally unbiased estimator.

**Definition 4.2** (locally unbiased estmator)**.** We say that an estimator $\tilde{\theta}$ is locall unbiased at $\theta = \theta_0$, if and only if

$$\langle \tilde{\theta} \rangle_{\theta=\theta_0} = \int \mathrm{d}x\, \tilde{\theta}(x) p_{\theta_0}(x) = \theta_0, \tag{4.5}$$

$$\left.\frac{\mathrm{d}\langle \tilde{\theta} \rangle}{\mathrm{d}\theta}\right|_{\theta=\theta_0} = \int \mathrm{d}x\, \tilde{\theta}(x)\, \left.\frac{\mathrm{d}p_{\theta_0}(x)}{\mathrm{d}\theta}\right|_{\theta=\theta_0} = 1, \tag{4.6}$$

which means that we only expect the estimator to track the true parameter up to the first order around a given value of parameter $\theta = \theta_0$.

**Example 4.1** (continued) Considering the same gaussian example as before, we see that indeed the proposed estimator $\tilde{\theta}(\boldsymbol{x}) = \sum_i x_i/N$ is unbiased, whereas its uncertainty reads:

$$\Delta^2\tilde{\theta} = \left\langle \left( \frac{1}{N}\sum_i x_i - \theta \right)^2 \right\rangle = \frac{\sigma^2}{N}. \tag{4.7}$$

The question remains if this is the minimal possible variance?

## 4.2 Cramér-Rao bound

We would like now to derive a lower bound on variance of any unbiased (locally) estimator, the so called Cramér-Rao (CR) bound. Thanks to this once we are able to show that a given estimator saturates the bound we will be sure that it is optimal.

**Theorem 4.1** (Cramér-Rao bound)**.** Let $p_\theta(x)$ be a family of probability distributions. Provided $p_\theta(x)$ satisfies some regularity conditions (see the proof), precision of any locally unbiased estimator $\tilde{\theta}$ is lower bounded by:

$$\Delta^2\tilde{\theta} \geq \frac{1}{F},\ F = \int \mathrm{d}x\, \frac{\dot{p}_\theta(x)^2}{p_\theta(x)}, \tag{4.8}$$

where $\dot{p}_\theta(x) = \frac{\mathrm{d}p_\theta(x)}{\mathrm{d}\theta}$, and $F$ is called the *Fisher Information* (FI). For simplicity of notation we have replaced $\theta_0$ with $\theta$.

*Proof.* We assume

$$\int \mathrm{d}x\, \tilde{\theta}(x)\dot{p}_\theta(x) = 1, \tag{4.9}$$

$$\int \mathrm{d}x\, \dot{p}_\theta(x) = 0, \tag{4.10}$$

where the first condition is the local unbiasedness condition, while the second is the formal requirement for regularity of $p_\theta(x)$ (if $p_\theta(x)$ is regular we may enter with the integral under the derivative and trivially satisfy this condition)—see Problem **??** to see an example of the model where this regularity assumption is not satisfied and there is no lower bound on uncertainty of the estimator.

Consider the following chain of inequalities

$$\Delta^2\tilde{\theta}\cdot F = \int \mathrm{d}x\, p_\theta(x) \left(\tilde{\theta}(x) - \theta\right)^2 \cdot \int \mathrm{d}x\, \frac{\dot{p}_\theta(x)^2}{p_\theta(x)} =$$

$$\int \mathrm{d}x\, \left[\sqrt{p_\theta(x)}\left(\tilde{\theta}(x) - \theta\right)\right]^2 \cdot \int \mathrm{d}x\, \left(\frac{\dot{p}_\theta(x)}{\sqrt{p_\theta(x)}}\right)^2 \overset{\text{C-S}}{\geq}$$

$$\left(\int \mathrm{d}x\, \left(\tilde{\theta}(x) - \theta\right)\dot{p}_\theta(x)\right)^2 = 1, \tag{4.11}$$

where we have used the Cauchy-Schwarz (C-S) inequality and utilized the local unbiasedness and regularity conditions in the last step. ■

**Remark.** One can encounter different but equivalent formulas for the FI:

$$F = \left\langle \left(\tfrac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(x)\right)^2 \right\rangle = -\left\langle \tfrac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log p_\theta(x) \right\rangle. \tag{4.12}$$

**Additivity of FI.** The FI is additive for product distributions. Let $p_\theta^{(12)}(x_1, x_2) = p_\theta^{(1)}(x_1)p_\theta^{(1)}(x_2)$, then $F^{(12)} = F^{(1)} + F^{(2)}$.

This is the justification for referring to this quantity as information. In particular, given $N$ i.i.d. random variables $x_i$, $F^{(N)} = NF$, where $F$ is the FI for single random variable, and in such cases the CR bound yields

$$\Delta^2\tilde{\theta} \geq \frac{1}{NF}, \tag{4.13}$$

showing the expected $1/N$ decrease in estimation variance as the number of repetitions of experiment increases.

**Example 4.1** (continued) Let us calculate the FI for the Gaussian example studied in this chapter. Since we deal with $N$ i.i.d. random variables, we can immediately say that $F^{(N)} = NF$, where $F$ is the FI for the Gaussian $p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\theta)^2/2\sigma^2}$, and equals $F = 1/\sigma^2$. Hence we obtain

$$\Delta^2\tilde{\theta} \geq \frac{\sigma^2}{N} \tag{4.14}$$

demonstrating that indeed the estimator considered before is optimal. In general an estimator that saturates the CR bound is called *efficient*.

**Condition for saturability of the CR bound.** Recalling the derivation of the CR bound, we see that the saturation of the CR bound is equivalent to saturation of the Cauchy-Schwarz inequality which is equivalent to:

$$\lambda(\theta)\sqrt{p_\theta(x)}(\tilde{\theta}(x) - \theta) = \frac{\dot{p}_\theta(x)}{\sqrt{p_\theta(x)}} \tag{4.15}$$

or equivalently

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(x) = \lambda(\theta)\left(\theta(x)\tilde{} - \theta\right), \tag{4.16}$$

where $\lambda(\theta)$ is arbitrary function. One can check the the above condition indeed holds for the exemplary Gaussian model we discussed in this section, provided we set $\tilde{\theta}(x) = \sum_i x_i/N$, $\lambda(\theta) = N/\sigma^2$.

## 4.3 Multi-parameter case

We now consider a general situation where we want to estimate multiple parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_P)$. The object which is a natural generalization of the estimator variance is the estimator covariance matrix $\mathbb{C}$:

$$\mathbb{C}_{ij} = \int \mathrm{d}x\, p_{\boldsymbol{\theta}}(x)\left(\tilde{\theta}_i(x) - \theta_i\right)\left(\tilde{\theta}_j(x) - \theta_j\right). \tag{4.17}$$

Diagonal elements represent the variances of estimators of a particular parameter, while off-diagonal terms represent potential correlations between estimation of different parameters. The multi-parameter generalization of the CR bound is a matrix inequality bounding the $\mathbb{C}$ matrix with the FI matrix.

**Theorem 4.2** (Multi-parameter CR bound)**.**

$$\mathbb{C} \geq \mathbb{F}^{-1},\ \mathbb{F}_{ij} = \int \mathrm{d}x\, \frac{\partial_i p_{\boldsymbol{\theta}}(x)\partial_j p_{\boldsymbol{\theta}}(x)}{p_{\boldsymbol{\theta}}(x)}, \tag{4.18}$$

where $\mathbb{F}$ is the FI matrix and $\partial_i$ denote differentiation with respect to $\theta_i$ parameter. The above matrix inequality should be understood in the sense that $\mathbb{C} - \mathbb{F}^{-1}$ is a positive semi-definite matrix.

*Proof.* We assume regularity and local unbiasedness conditions, which in the multiparameter case ammount to:

$$\int \mathrm{d}x\, \tilde{\theta}_i(x)\partial_j p_{\boldsymbol{\theta}}(x) = \delta_{ij}, \tag{4.19}$$

$$\int \mathrm{d}x\, \partial_i p_{\boldsymbol{\theta}}(x) = 0. \tag{4.20}$$

Let us choose some vectors $w$ and $v$ of length $P$ and write

$$w^T\mathbb{C}w\cdot v^T\mathbb{F}v = \int \mathrm{d}x \sum_{ij} w_i p_{\boldsymbol{\theta}}(x)(\tilde{\theta}_i(x)-\theta)(\tilde{\theta}_j(x)-\theta)w_j$$
$$\cdot \int \mathrm{d}x' \sum_{i'j'} v_{i'}\frac{\partial_{i'}p_{\boldsymbol{\theta}}(x')\partial_{j'}p_{\boldsymbol{\theta}}(x')}{p_{\boldsymbol{\theta}}(x')}v_{j'} =$$
$$\int \mathrm{d}x \sum_i w_i\sqrt{p_{\boldsymbol{\theta}}(x)}(\tilde{\theta}_i(x)-\theta)\sum_j \sqrt{p_{\boldsymbol{\theta}}(x)}(\tilde{\theta}_j(x)-\theta)w_j\cdot$$
$$\int \mathrm{d}x' \sum_{i'} v_{i'}\frac{\partial_{i'}p_{\boldsymbol{\theta}}(x')}{\sqrt{p_{\boldsymbol{\theta}}(x')}}\sum_{j'}\frac{\partial_{j'}p_{\boldsymbol{\theta}}(x')}{\sqrt{p_{\boldsymbol{\theta}}(x')}}v_{j'}$$
$$\overset{\text{C-S}}{\geq}\left[\int \mathrm{d}x\left(\sum_i w_i(\tilde{\theta}_i(x)-\theta)\right)\left(\sum_{i'} v_{i'}\partial_{i'}p_{\boldsymbol{\theta}}(x)\right)\right]^2 =$$
$$(w^Tv)^2, \tag{4.21}$$

where in the last step we have used the local unbiasedness as well as regularity conditions. Choosing

$w = Fv$, we get:

$$v^T \mathbb{F}\mathbb{C}\mathbb{F} v \cdot v^T \mathbb{F} v \geq (v^T \mathbb{F} v)^2, \tag{4.22}$$

$$v^T \mathbb{F}\mathbb{C}\mathbb{F} v \geq v^T \mathbb{F} v. \tag{4.23}$$

Since the above inequality is valid for arbitrary $v$, this implies

$$\mathbb{F}\mathbb{C}\mathbb{F} \geq \mathbb{F} \Rightarrow \mathbb{C} \geq \mathbb{F}^{-1}, \tag{4.24}$$

where the final result we have obtained by acting on both sides with $\mathbb{F}^{-1}$. ■

**Remark.** From the derived bound it follows in particular that: $\Delta^2 \tilde{\theta}_i \geq (\mathbb{F}^{-1})_{ii} \geq (\mathbb{F}_{ii})^{-1}$, and the last inequality is in general strict if $\mathbb{F}$ contains nonzero off-diagonal elements. To see this consider: $1 = e_i^T \sqrt{\mathbb{F}} \sqrt{\mathbb{F}^{-1}} e_i \overset{\text{C-S}}{\leq} e_i^T \mathbb{F} e_o e_i^T \mathbb{F}^{-1} e_i$, where $e_i$ is the basis vector with 1 at $i$-th position and zeros elsewhere. This inequality leads to $(\mathbb{F}^{-1})_{ii} \geq 1/\mathbb{F}_{ii}$.

## 4.4 Maximum likelihood estimator