

# Statistics Introduction

Thomas Bonald  
Institut Polytechnique de Paris

2024 – 2025

Statistics is the science of data. The objective is to extract useful information from data so as to make predictions. The approach consists in viewing the data samples as a sequence of independent observations of some random variable  $X$ . The exact probability distribution of  $X$  is unknown and must be learned.

## 1 Parametric model

A statistical model is *parametric* if the probability distribution of  $X$  belongs to some family of distributions indexed by some parameter  $\theta$  of finite dimension.

### Parametric model

A parametric model is a set of probability distributions  $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$  with  $\Theta \subset \mathbb{R}^d$  for some finite dimension  $d$ .

The parameter  $\theta$  is unknown and must be learned, using the observations. This is possible only if each probability distribution  $P_\theta \in \mathcal{P}$  is defined by a unique parameter  $\theta$ .

### Identifiability

The parametric model  $\mathcal{P}$  is *identifiable* if the mapping  $\theta \mapsto P_\theta$  is a bijection.

In most practical cases, the probability distribution  $P_\theta$  has a density with respect to some measure  $\nu$  (e.g., the Lebesgue measure).

### Dominated model

The parametric model  $\mathcal{P}$  is *dominated* if there exists some measure  $\nu$  such that  $P_\theta$  has a density  $p_\theta$  with respect to  $\nu$  for all  $\theta \in \Theta$ .

**Example 1** The model  $\mathcal{P} = \{P_\theta \sim \mathcal{N}(\mu, \sigma^2), \theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+\}$  of Gaussian distributions is dominated by the Lebesgue measure  $\nu = \lambda$ . The density is  $p_\theta(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ .

**Example 2** The model  $\mathcal{P} = \{P_\theta \sim \mathcal{B}(\theta), \theta \in [0, 1]\}$  of Bernoulli distributions is dominated by the counting measure  $\nu = \delta_0 + \delta_1$ . The density is  $p_\theta(x) = \theta^x(1-\theta)^{1-x}$ .

When  $n$  observations are available, say  $x_1, \dots, x_n$ , it is common to assume that these observations are independent and identically distributed. If the model is dominated, the probability density function of the vector of observations  $x = (x_1, \dots, x_n)$  is the product of the probability density functions. We use the same notations  $P_\theta$  and  $p_\theta$  for convenience.

#### Independent observations

For a dominated model with  $n$  independent observations  $x = (x_1, \dots, x_n)$ ,  $p_\theta(x) = p_\theta(x_1) \dots p_\theta(x_n)$ .

## 2 Decision function

The role of the statistician is to take *actions* based on the observations.

#### Decision function

A decision function  $\delta$  is a mapping from the set of observations  $\mathcal{X}$  to the set of actions  $\mathcal{A}$ .

The set of actions  $\mathcal{A}$  depends on the task:

Estimation  $\rightarrow$  The objective is to find  $\theta$  and thus  $\mathcal{A} = \Theta$ .

Hypothesis testing  $\rightarrow$  The objective is to answer a binary question and thus  $\mathcal{A} = \{0, 1\}$ .

Confidence region  $\rightarrow$  The objective is to find a confidence region for  $\theta$  and thus  $\mathcal{A} = \mathcal{P}(\Theta)$ .

## 3 Loss function

A loss function is used to assess the quality of an action  $a \in \mathcal{A}$ . The quality of this action depends on the parameter  $\theta$ .

#### Loss function

A loss function  $L$  is a mapping from  $\Theta \times \mathcal{A}$  to  $\mathbb{R}_+$  so that  $L(\theta, a)$  is the cost of selecting action  $a$  for the parameter  $\theta$ .

Usual loss functions are:

Estimation  $\rightarrow$  For  $\theta, a \in \mathbb{R}$ , the square error  $L(\theta, a) = (\theta - a)^2$ .

Hypothesis testing  $\rightarrow$  Given a partition  $\Theta = \Theta_0 \cup \Theta_1$  for hypotheses 0 and 1, the error  $L(\theta, a) = 1_{\{\theta \notin \Theta_a\}}$ .

Confidence region  $\rightarrow$  For  $\theta \in \mathbb{R}$ ,  $a \subset \mathbb{R}$ , the error  $L(\theta, a) = 1_{\{\theta \notin a\}}$ .

## 4 Risk

To assess the quality of a decision function  $\delta$ , it is necessary to consider all possible actions taken by this decision function, depending on the observations. This can be done by averaging the loss function over all observations (and thus all actions taken for these observations).

## Risk

The risk  $R$  is the expectation of the loss function  $L$  over all observations,  $R(\theta, \delta) = E_{\theta}(L(\theta, \delta(X)))$ .

Note that the risk  $R$  depends on the parameter  $\theta$ , which is unknown. Several strategies exist to select the best decision function:

Min-max  $\rightarrow$  Minimize the maximum risk,  $\max_{\theta \in \Theta} R(\theta, \delta)$ .

Bayesian approach  $\rightarrow$  Minimize the expected risk  $E(R(\theta, \delta))$ , assuming some prior distribution on  $\theta$ .

# Statistics

## 1. Statistics and Information

Thomas Bonald, Philippe Ciblat  
Institut Polytechnique de Paris

2024 – 2025

Consider a statistical model with unknown parameter  $\theta$ . Our objective is to get information on  $\theta$  on observing  $X$ , a sequence of random samples drawn independently from the probability distribution  $P_\theta$ . In this lecture, we introduce the notion of *information* contained in  $X$  on  $\theta$ . This is a key theoretical tool to assess how difficult it is to estimate  $\theta$  on observing  $X$ . It will be used in the next lecture to identify efficient estimators of  $\theta$ .

### 1 Notion of statistic

Let  $P_\theta$  be a probability distribution that depends on some unknown parameter  $\theta \in \Theta$  with  $\Theta \subset \mathbb{R}^d$  for some finite  $d$ . We assume that the model is dominated in the sense that  $P_\theta$  has a density with respect to some probability measure  $\nu$  (typically the counting measure or the Lebesgue measure). We denote by  $X = (X_1, \dots, X_n)$  a sequence of  $n$  i.i.d. samples drawn from  $P_\theta$ , and by  $x = (x_1, \dots, x_n)$  the values taken by  $X$ , observed by the statistician.

#### Statistic

A statistic is any function  $T$  of the observations  $x$  only. It does not depend on  $\theta$ .

The statistician must take a decision based on the observations. This decision is typically based on some statistic (or some set of statistics). Some statistics are better than others in the sense that they bring more information on the unknown parameter  $\theta$ . In particular, some statistics do not bring *any* information on  $\theta$  (free statistics) while some other bring *all* the information on  $\theta$  contained in the observations (sufficient statistics).

#### Free statistic

A statistic  $T$  is free if the distribution of  $T(X)$  is independent of  $\theta$ .

#### Sufficient statistic

A statistic  $T$  is sufficient if the distribution of  $X$  given  $T(X)$  is independent of  $\theta$ .

Sufficient statistics, when they exist, can be used to take optimal decisions. Unfortunately, it is generally not obvious to check whether a statistic is sufficient. The following result provides a criterion based on the form of the distribution.

**Theorem 1 (Fisher Factorization Theorem)** *A statistic  $T$  is sufficient if and only if there exist some functions  $g_\theta$  and  $h$  (independent of  $\theta$ ) such that*

$$\forall x, \quad p_\theta(x) = g_\theta(T(x))h(x).$$

**Example 1** *In the Bernoulli model  $\mathcal{P} = \{P_\theta \sim \mathcal{B}(\theta), \theta \in [0, 1]\}$ , the statistic  $T(x) = \sum_{i=1}^n x_i$  is sufficient.*

## 2 Entropy

Before measuring the information on  $\theta$  contained in  $X$ , we consider the information contained in  $X$  itself, using the notion of entropy [1]. In this and the following section, we omit the parameter  $\theta$  as the considered notions of information are general and applicable to non-parametric models.

### Entropy

The entropy of the random variable  $X$  is:

$$H(X) = - \int p(x) \log p(x) d\nu(x) = -E(\log p(X)).$$

**Example 2** *The entropy of a Bernoulli random variable  $X \sim \mathcal{B}(\theta)$  is:*

$$H(X) = -\theta \log \theta - (1 - \theta) \log(1 - \theta).$$

Observe that independent random variables  $X, Y$  bring their own information in the sense that:

$$H(X, Y) = H(X) + H(Y).$$

More generally, for two random variables  $X, Y$  with joint probability density function  $p(x, y)$ , it is useful to introduce the notion of conditional entropy of  $Y$  given  $X$ , which is the entropy of the distribution of  $Y$  given  $X = x$ , denoted by  $p(y|x)$ , expected over the distribution of  $X$ :

$$\begin{aligned} H(Y|X) &= \int p(x) H(Y|X = x) d\nu(x), \\ &= - \int p(x) \left( \int p(y|x) \log p(y|x) d\nu(y) \right) d\nu(x), \\ &= - \int p(x, y) \log p(y|x) d\nu(x) d\nu(y), \\ &= -E(\log p(Y|X)). \end{aligned}$$

Observe that  $H(Y|X) = H(Y)$  if  $X$  and  $Y$  are independent.

### Chain rule

For any random variables  $X, Y$ ,

$$H(X, Y) = H(X) + H(Y|X).$$

An important consequence of the chain rule is that a statistic cannot contain more information than the original data. This is known as the data processing inequality.

#### Data processing inequality

For any statistic  $T$ ,  $H(T(X)) \leq H(X)$ .

The chain rule extends to  $n$  random variables:

$$H(X_1, \dots, X_n) = H(X_1) + H(X_2|X_1) + \dots + H(X_n|X_1, \dots, X_{n-1}).$$

In particular, we have for  $n$  i.i.d. random variables:

$$H(X_1, \dots, X_n) = H(X_1) + \dots + H(X_n) = nH(X_1).$$

### 3 Kullback-Leibler divergence

We now introduce a notion of information gap between two probability distributions  $P, Q$  related to entropy. This will provide useful to quantify the impact on  $P_\theta$  of errors on the estimation of  $\theta$ . We denote by  $p, q$  the respective probability density functions of  $P, Q$  with respect to some measure  $\nu$ .

#### Kullback-Leibler divergence

The Kullback-Leibler divergence between  $P$  and  $Q$  is defined by:

$$D(P||Q) = \int p(x) \log \frac{p(x)}{q(x)} d\nu(x) = \mathbb{E} \left( \log \frac{p(X)}{q(X)} \right)$$

with  $X \sim P$ .

**Example 3** For two Bernoulli distributions  $P \sim \mathcal{B}(p), Q \sim \mathcal{B}(q)$ ,

$$D(P||Q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$$

Observe that  $D$  is not a metric as it is not symmetric and it does not satisfy the triangle inequality. However, it follows from Jensen's inequality that  $D(P||Q) \geq 0$  with equality if and only if  $P = Q$ .

### 4 Fisher information

Finally, we consider the information on  $\theta$  contained in  $X$  when  $X \sim P_\theta$ . This is known as Fisher information. Although it shares some properties with entropy (chain rule, data processing inequality), it is fundamentally different as the focus is on the parameter  $\theta$ . In particular, the standard notation is  $I(\theta)$  (or  $I_X(\theta)$  when necessary) to highlight the fact that the information is related to the parameter  $\theta$ .

Some regularity conditions are required:

- $\frac{\partial \log p_\theta}{\partial \theta}$  exists,
- the support of  $p_\theta$  is independent of  $\theta$ ,
- the expectation and the partial derivative  $\frac{\partial}{\partial \theta}$  can be inverted.

We first consider the scalar case  $\Theta \subset \mathbb{R}$ . The score is defined by:

$$S(X) = \frac{\partial \log p_\theta(X)}{\partial \theta}.$$

### Fisher information

The Fisher information is the variance of the score:

$$I(\theta) = \text{var} \left( \frac{\partial \log p_\theta(X)}{\partial \theta} \right)$$

**Example 4** For a Bernoulli random variable  $X \sim \mathcal{B}(\theta)$  with  $\theta \in (0, 1)$ , we have:

$$p_\theta(x) = \theta^x(1 - \theta)^{1-x}, \quad x = 0, 1.$$

The score is:

$$S(X) = \frac{\partial \log p_\theta(X)}{\partial \theta} = \frac{X}{\theta} - \frac{1-X}{1-\theta}.$$

The Fisher information is:

$$I(\theta) = \frac{1}{\theta(1-\theta)}.$$

The score is centered,  $E(S(X)) = 0$ , so that the Fisher information satisfies:

$$I(\theta) = \text{var}(S(X)) = E(S(X)^2).$$

We have the following equivalent definition of Fisher information.

### Fisher information

The Fisher information is the opposite of the curvature of the log-likelihood:

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

Like entropy, independent random variables  $X, Y$  bring their own information on  $\theta$  in the sense that:

$$I_{X,Y}(\theta) = I_X(\theta) + I_Y(\theta).$$

Now let for two random variables  $X, Y$  with joint probability density function  $p_\theta(x, y)$ . We define the notion of Fisher information of  $Y$  given  $X$  on  $\theta$  by:

$$\begin{aligned} I_{Y|X}(\theta) &= \int p_\theta(x) \text{var} \left( \frac{\partial \log p_\theta(Y|X=x)}{\partial \theta} \right) d\nu(x), \\ &= \int p_\theta(x) \left( \int p_\theta(y|x) \left( \frac{\partial \log p_\theta(y|x)}{\partial \theta} \right)^2 d\nu(y) \right) d\nu(x), \\ &= \int p_\theta(x, y) \left( \frac{\partial \log p_\theta(y|x)}{\partial \theta} \right)^2 d\nu(x) d\nu(y), \\ &= E \left( \left( \frac{\partial \log p_\theta(Y|X)}{\partial \theta} \right)^2 \right). \end{aligned}$$

Observe that  $I_{Y|X}(\theta) = I_Y(\theta)$  if  $X$  and  $Y$  are independent.

#### Chain rule

For any random variables  $X, Y$ ,

$$I_{X,Y}(\theta) = I_X(\theta) + I_{Y|X}(\theta).$$

The data processing inequality follows.

#### Data processing inequality

For any statistic  $T$ ,  $I_{T(X)}(\theta) \leq I_X(\theta)$  for all  $\theta \in \Theta$ .

Observe that  $0 \leq I_{T(X)}(\theta) \leq I_X(\theta)$  with:

- $I_{T(X)}(\theta) = 0$  if  $T$  is free,
- $I_{T(X)}(\theta) = I_X(\theta)$  if  $T$  is sufficient (by the chain rule).

The chain rule extends to  $n$  random variables:

$$I_{X_1, \dots, X_n}(\theta) = I_{X_1}(\theta) + I_{X_2|X_1}(\theta) + \dots + I_{X_n|X_1, \dots, X_{n-1}}(\theta).$$

In particular, we have for  $n$  i.i.d. random variables:

$$I_{X_1, \dots, X_n}(\theta) = I_{X_1}(\theta) + \dots + I_{X_n}(\theta) = nI_{X_1}(\theta).$$

**Entropy and Fisher information.** The link between both notions of information is provided by the following result:

$$\forall h, \quad D(P_\theta || P_{\theta+h}) = \frac{1}{2} I(\theta) h^2 + o(h^2).$$

**Example 5** For the Bernoulli model  $P_\theta \sim \mathcal{B}(\theta)$  with  $\theta \in (0, 1)$ , we have:

$$D(P_\theta || P_{\theta+h}) = \frac{h^2}{\theta(1-\theta)} + o(h^2).$$

**Vectorial case.** When  $\theta$  is a vector, that is  $\Theta \subset \mathbb{R}^d$ , the score is the gradient:

$$S(X) = \nabla_\theta \log p_\theta(X).$$

The Fisher information becomes a matrix:

#### Fisher information

The Fisher information is the covariance matrix of the score:

$$I(\theta) = \text{cov}(\nabla_\theta \log p_\theta(X))$$

The score is a centered,  $E(S(X)) = 0$ , so that the Fisher information matrix can be written:

$$I(\theta) = \text{cov}(S(X)) = E(S(X)S(X)^T).$$

The chain rule and the data processing inequality hold. The link with entropy is given by:

$$\forall h, \quad D(P_\theta || P_{\theta+h}) = \frac{1}{2} h^T I(\theta) h + o(\|h\|^2).$$

## References

- [1] T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley, 1999.