

# 1 Information and its Main Quantitative Properties

## 1.1 Definition and Generalities

This chapter constitutes the base of the next theoretical formalism to be developed in this part of the book. The connection between the Bayesian rule and Kullback information divergence is first envisaged. This will permit a better understanding of Shannon-Jaynes-Kullback cross-entropy. The next section will deal with the connection between Shannon-Gibbs entropy and non-extensive (Tsallis) entropy. Finally, the generalized non-extensive cross-entropy will be presented for further applications in the remaining parts of the work.

Many forms and measures of information exist. As far as parameters linked to data observations are concerned, one well-known measure of information was provided by R.A. Fisher in 1929. As will be clear below, the next can be  $\log(n)$ , explaining the sum of  $n$  hypotheses  $H_i$ , all uniformly distributed and known as Hartley's information measure, Hartley (1928). Information theory has its mathematical roots in the concept of disorder or entropy in statistical mechanics. Kullback (1959) provides an extensive literature on the form and mathematics linking entropy and information theory. As mentioned, the next formal definition will be followed by theoretical and empirical extensions arising from the entropy principle.

Let us now develop a workable measure of information obtained through observation of an event having probability  $p$ . Our first problem is to ignore any particular features of the event and focus only on whether or not it happened. Thus we will think of an event as the observance of a symbol whose probability of occurring is  $p$ . Thus, the information will be defined in terms of the probability  $p$ .

Let us consider the probability spaces  $(\chi, \mathcal{G}, \mu_i)$ ,  $i = 1, 2$  as a basic set of elements  $x \in \chi$  (sample space) and the  $\sigma$  – algebra  $\mathcal{G}$ , a collection of all possible sets of events from  $\chi$  with the probability measure  $\mu_i$ . Under general assumptions of the above probability measures, in particular those stating their absolute continuity with respect to one another, let  $\lambda \equiv \mu_i$ . By the Radon-Nikodym theorem (e.g., Loeve, 1955), there exist functions  $f_i(x)$ ,  $i = 1, 2$ , called generalized probability densities,  $0 < f_i(x) < \infty$  [ $\lambda$ ] such that:

$$\mu_i(E) = \int_E f_i(x) d\lambda(x), \quad i = 1, 2, \quad (2.1)$$

for all  $E$  belonging to the  $\sigma$  – algebra  $\mathcal{G}$ . Following Kullback (1959) and Halmos & Savage (1949), the symbol [ $\lambda$ ], pronounced “modulo  $\lambda$ ”, means that the assertion is true along with all the support space of events  $E$  except the case for  $E \in \mathcal{G}$  and  $\lambda(E) = 0$ .

In (2.1), the function  $f_i(x)$  is also referred to as the Radon-Nikodym derivative. If the probability measure  $\mu$  is absolutely continuous with respect to the probability measure  $\lambda$  and the probability measure  $\nu$  is absolutely continuous with respect to  $\mu$ ,

then the probability measure  $\nu$  is also absolutely continuous with respect to  $\lambda$ , and the Radon-Nikodym derivatives satisfy:

$$\frac{d\nu}{d\lambda} = \frac{d\nu}{d\mu} \cdot \frac{d\mu}{d\lambda} [\lambda]$$

The defined symbols above allow us to better derive the conceptual definition of information below as it will be understood in the coming chapters of this book.

Next, let  $H_i$ ,  $i = 1, 2$ , be the hypothesis that a variable is  $X$  from the statistical population with probability measure  $\mu_i$ . Then, by applying Bayes's theorem, it follows that:

$$P(H_i | x) = \frac{P(H_i)f_i(x)}{P(H_1)f_1(x) + P(H_2)f_2(x)} [\lambda], \quad i = 1, 2, \quad (2.2)$$

After transformations with respect to logarithms of relative function densities  $f_i(x)$ , we obtain:

$$\log \frac{f_1(x)}{f_2(x)} = \log \frac{P(H_1 | x)}{P(H_2 | x)} - \log \frac{P(H_1)}{P(H_2)} [\lambda], \quad (2.3)$$

where:  $x$  is an element of  $X$ ;  $P(H_i)$  is the prior probability of  $H_i$  and  $P(H_i | x)$  is the posterior probability of  $H_i$ . The logarithm in (2.3) stands for an information measure base unit (Hartley, 1928). The right-hand side of (2.3) is an informative measure resulting from the difference (positive or negative) between the logarithm of the odds in favour of  $H_i$  once observation of  $x$  has occurred and before it occurred.

Thus, following Kullback, one defines the logarithm of the likelihood ratio,

$$\log \frac{f_1(x)}{f_2(x)},$$

as the information in  $X = x$  for discrimination in favour of  $H_1$  against  $H_2$ . An interesting alternative definition of information after (2.3) is the weight of evidence for  $H_1$  given  $x$  (Kullback, 1959), (Good, 1963). Next, most informative is the mean information for discrimination in favour of  $H_1$  against  $H_2$  given  $x \in E \in \mathcal{D}$ , for  $\mu_i$ , which is defined as follows:

$$\begin{aligned} I(\mu_1 : \mu_2) &= \int \log \frac{f_1(x)}{f_2(x)} d\mu_1(x) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)} d\lambda(x) = \\ &= \int \log \frac{P(H_1 | x)}{P(H_2 | x)} d\mu_1(x) - \log \frac{P(H_1)}{P(H_2)} \end{aligned} \quad (2.4)$$

with  $d\mu_i(x) = f_i(x) d\lambda(x)$ .

Here one has treated the general case when  $E$  represents the entire sample space  $\chi$  and then must not appear as support space for integration (see 2.1). The last member in (2.4) is the difference between the mean value, with respect to  $\mu_1$ , of the logarithms of the posterior and prior odds of the hypotheses. Following Savage (1954), Kullback (1959),  $I(1:2)$  could be referred to as the information of  $\mu_1$  with respect to  $\mu_2$ .

Let us extend the above general definition of information to some known cases. Suppose we have a set (categories) of hypotheses,  $H_i = 1, 2, \dots, n$  and that from observation, we can infer with certainty which hypothesis is true. Then the mean information in an observation about  $H$  is the mean value of  $-\log P(H_i)$ , that is,

$$P(H_1) \log P(H_1) - P(H_2) \log P(H_2) - \dots - P(H_n) \log P(H_n). \tag{2.5}$$

The expression in (2.5) above is called entropy of the  $\hat{a}_j$  (e.g., Khinchin, 1957; Shannon, 1948). When hypotheses  $H_i$  are uniformly distributed (then equally probable) so that

$$P(H_i) = 1/n, i = 1 \dots n; \text{ this leads to } - \sum_{i=1}^n P(H_i) \log P(H_i) = \log n,$$

which turns out to be Hartley’s information measure.

As shown below, an interesting applicability of (2.4) may concern the analysis of hypotheses  $H_i, i = 1, 2$ , on dependency between variables  $x$  and  $y$  ( $\forall q > 0$ ) or on the measure of divergence between given hypotheses  $H_i$ . Presenting relationships between information discriminating measure and dependency between variables will be useful when we introduce an inferential approach for entropy econometrics models. In particular, measure of divergence constitutes, once again, the cornerstone of the present work in which *a priori* and *a posteriori* hypotheses will be recalled in many applicable analyses.

Suppose we have the entire sample space  $\chi$  being the Euclidean space of two dimensions  $R^2$  with elements  $X = (x, y)$ . Let us consider that under  $H_1$  variables  $x$  and  $y$  ( $\forall q > 0$ ) are dependent with probability density  $f(x, y)$  and that, under the alternative hypothesis  $H_2$ , both variables are independent with probabilities  $g(x)$  and  $h(y)$ . In this case, we rewrite (2.4) as follows:

$$I(\mu_1 : \mu_2) = \iint f(x, y) \log \frac{f(x, y)}{g(x)h(y)} dx dy \tag{2.6}$$

Information measure  $I(\mu_1 : \mu_2)$  is nonnegative (Kullback, 1959) and equal to zero if and only if  $f(x, y) = g(x) h(y)$  [  $\lambda$  ]. As such, it constitutes an informative indicator on dependency degree between  $x$  and  $y$  ( $\forall q > 0$ ). Note that in the case of a bivariate normal density

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y(1-\rho^2)^{\frac{1}{2}}} \exp \left[ -\frac{1}{2(1-\rho^2)} \left( \frac{x^2}{\sigma_x^2} - 2\rho \frac{xy}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2} \right) \right]$$

where hypothesis  $H_2$  then represents the product of the normal densities as explained in (2.6), and finally one obtains:

$$I(\mu_1 : \mu_2) = -\frac{1}{2} \log(1 - \rho^2), \quad (2.7)$$

which indicates that in the case of bivariate normal distribution, as expected, the mean information is discriminatory in favour of  $H_1$  (dependence) against  $H_2$  (independence); that is  $I(\mu_1 : \mu_2)$  is a function of the correlation coefficient  $\rho$  alone.

Following (Jeffreys, 1946), (Kullback, Information theory and statistics, 1959), if we define  $I(2: 1)$  as

$$I(\mu_1 : \mu_2) = \int f_2(x) \log \frac{f_2(x)}{f_1(x)} d\lambda(x) \quad (2.8)$$

as the mean information from  $\mu_2$  for discrimination in favour of  $H_2$  against  $H_1$ , one can define the divergence (noted  $\nabla$ ) by:

$$\begin{aligned} \nabla(H_1, H_2) &= I(\mu_1 : \mu_2) + I(\mu_2 : \mu_1) = \int (f_1(x) - f_2(x)) \log \frac{f_1(x)}{f_2(x)} d\lambda(x) = \\ &= \int \log \frac{P(H_1 | x)}{P(H_2 | x)} d\mu_1(x) - \int \log \frac{P(H_1 | x)}{P(H_2 | x)} d\mu_2(x) \text{ divergence between hypotheses.} \end{aligned} \quad (2.9)$$

Thus,  $\nabla(H_1, H_2)$  measures the divergence between  $H_1$  and  $H_2$  or between  $\mu_1$  and  $\mu_2$ . As such, it constitutes a measure of the difficulty of discriminating between them.

## 1.2 Main Quantitative Properties of Statistical Information

The approach undertaken here is axiomatic (Carter, 2011). It is worthwhile to note that we can apply this axiomatic system in any context where we have an available set of non-negative real numbers. This can be the case, for instance, when we dispose of non-negative coefficients (noted  $p$ ) of a given set and target the estimation of the related model parameters through their reparametrization (Golan, Judge & Miller, 1996). Naturally, we will come back to such applications, and an estimation approach using probabilities and support space simultaneously will be presented. This underscores an important role to be assigned to the probability form of numbers, which motivated the selection of the axioms below. We will want our information measure  $I(p)$  to have several properties:

1. Information is a non-negative quantity, i.e.,  $I(p) \geq 0$ . Following what has been presented above on information definition (see 2.4), one may generalize this property to convexity in the next theorem:

*Theorem:*  $I(p_1 : p_2)$  is almost positive defined, that is  $I(p_1 : p_2) \geq 0$  with equality if and only if  $f_1(x) = f_2(x) [\lambda]$ .

We will not demonstrate this theorem (see Kullback, 1959, pp. 14–15); we just provide the reader with the essence channelled through it. The above theorem explains that in the mean, discrimination information from statistical observations is positive. It follows from what has been previously said that no discrimination information will result if the distribution of observations is the same  $[\lambda]$  under hypothesis one and two. A typical example—as we will see later—may constitute maximum entropy and cross-entropy principles. In that case, when non-informative consistency moments from observations are not provided, minimum cross-entropy declines into maximum entropy.

2. If an event has probability 1, certainty follows, and we get no information from the occurrence of the event:  $I(p = 1) = 0$ .
3. If two independent events occur (whose joint probability is the product of their individual probabilities), then the information we get from observing the events is the sum of the two pieces of information:

$I(p_1, p_2) = I(p_1) + I(p_2)$ . This property is referred to as additivity. Note that this property presents a valuable feature; it represents the basis of the logarithmic form of information. Intuitively, that means that a sample of  $n$  independent observations from the same population provides  $n$  times the mean information in a single observation.

In the case of non-independent events, the additive property is retained, but in terms of conditional information.

4. Finally, as already stipulated in the preceding section, we will want our information measure to be a continuous (and, in fact, monotonic) function of the probability—slight changes in probability should result in slight changes in information. For consistency with the properties above, it can be useful to show the logarithmic feature of statistical information in the following way:

$$1. I(p^2) = I(pp) = I(p) + I(p) = 2I(p) \tag{2.10}$$

$$2. \text{ Through inductive reasoning, one can generalize (2.10) and write, } I(pn) = nI(p)$$

$$3. I(p) = I((p^{1/m})^m) = mI(p^{1/m})$$

and we have

$$I(p^{1/m}) = \frac{1}{m} I(P)$$

and, once again, we can generalize in the following way :

$$I(p^{n/m}) = \frac{n}{m} I(p)$$

4. The property of continuity allows us to write, for  $0 < p \leq 1$  and a real number  $\alpha$ :

$$I(p^\alpha) = \alpha I(p).$$

From (2.10), one can observe that an operator transforming the probability  $p$  at the power  $n/m$ , (that is,  $p^{n/m}$ ) into an information measure  $I(p^{n/m})$  displays a logarithmic property of additivity. This allows us to write a general, useful relation:

$$I(p) = -\log_b(p) = \log_b\left(\frac{1}{p}\right) \text{ for base } b. \quad (2.11)$$

For other information properties not directly connected with the aim of this work, such as invariance or sufficiency, which will not be presented here, see Jaynes (1994), Kullback (1959). Furthermore, in the coming chapters, additional properties for different forms of entropy will be presented, such as concavity and stability (common for both Shannon-Gibbs and Tsallis entropies) or extensivity (common for both Shannon-Gibbs and Renyi (1961) entropies).

As a final remark of this section, it is important to note that the above logarithmic nature of information as explained in (2.11)—for the case of independent events—is limited to ergodic systems which convey additive-extensive properties of information in the case of independent events.