Developments in Statistics
Andrej Mrvar and Anuška Ferligoj (Editors)
Metodološki zvezki, 17, Ljubljana: FDV, 2002

Relations among Fisher, Shannon-Wiener and Kullback Measures of Information for Continuous Variables

Anton Cedilnik and Katarina Košmelj[1]

Abstract

In statistics, Fisher was the first to introduce the measure of the amount of information supplied by the data about the unknown parameter. We analyze the disadvantages of Fisher information measure for optimization of sampling designs. To overcome this problem, we modify Fisher information measure and we upgrade it to the multivariate setting. It turns out that a reasonable modification of Fisher information measure leads to a special case of Kullback information measure, both in the univariate and multivariate setting. Using Shannon's and Wiener's concept of information we also show a simple derivation of Kullback information measure for a special case when the prior distribution of the parameter is uniform and the posterior distribution is truncated normal.

1 Introduction

The motivation for our work was to improve the soil sampling design to a new design where maximum information about the unknown parameters is obtained, given a fixed budget. The variables under study were the concentrations of several chemical compounds in soil and the means of these variables were the parameters of interest.

Information about the unknown parameter is defined as Fisher information measure, which is a standard statistical concept of information. It turns out that this concept of information is not applicable in our case. Therefore we review other concepts.

[1] Biotehnical Faculty, University of Ljubljana, Jamnikarjeva 101, 1000 Ljubljana, Slovenia.

The notion of information is very broad. We restricted our work to Fisher's, Shannon-Wiener's and Kullback's concepts of information. First, we analyze the disadvantages of Fisher information measure for our optimization purpose. We modify Fisher information measure to fullfil our purpose and upgrade it to the multivariate setting. It turns out that a reasonable modification of Fisher information measure leads to a special case of Kullback information measure, both in the univariate and multivariate setting. Using Shannon's and Wiener's concept of information we show a simple derivation of Kullback information measure for a special case, when the prior distribution of the unknown parameter is uniform and the posterior distribution is truncated normal.

2 Fisher and Kullback information measure

2.1 Fisher information measure and its modifications

In statistics, Fisher was the first to introduce the measure of the amount of information supplied by data about the unknown parameter. It plays an important role in the theory of statistical estimation and inference. Fisher defined the amount of information in a sample about the unknown parameter θ as the reciprocal of the variance of θ:

$$I_F(\theta) = \frac{1}{\mathrm{Var}(\theta)} \tag{1}$$

For our optimization purpose we ascertain the following two disadvantages of the Fisher information:

- Fisher information is dependent on the measurement unit of variable and therefore can not be compared for different variables.
- Multivariate approach is not straightforward. Some authors suggest some compromise of univariate information (Cochran, 1963). A simplified approach is to specify an importance weight for each variable and to form a linear combination of the univariate informations. This approach is difficult to justify: such a linear combination is meaningless, correlations among the variables are ignored. As an alternative some authors (Arvantis and Afonja, 1971) advocate the use of the generalized variance of the sample means - the determinant of the variance-covariance matrix of the vector parameter $\boldsymbol{\theta}$:

$$I_F(\boldsymbol{\theta}) = \frac{1}{\det \mathbf{Var}(\boldsymbol{\theta})} \tag{2}$$

This approach takes into consideration the magnitude of the correlations among the variables, however the measurement units of the original variables are inherited. For different sets of variables information is not comparable.

To overcome these drawbacks we tried to modify Fisher information. The first attempt was to multiply it by the squared expected value of the estimator:

$$I_F^1(\theta) = \frac{\mathrm{E}(\theta)^2}{\mathrm{Var}(\theta)} = CV(\theta)^{-2}. \tag{3a}$$

Here, information for different variables can be compared. This measure is based on the relative precision of the parameter, it is related to its coefficient of variation. However, (3a) is a suitable measure of information only for a parameter with the range $[0, \max\theta]$. Therefore, a preferable alternative to (3a) is:

$$I_F^2(\theta) = \frac{(b-a)^2}{\mathrm{Var}(\theta)}, \tag{3b}$$

where: $b = \sup(\theta),\ a = \inf(\theta)$.

Supremum and infimum of the parameter are defined according to the prior knowledge of researchers.

For the multivariate case with *K* variables, the following alternative, based on the generalized variance, turns out to be adequate:

$$I_F^2(\boldsymbol{\theta}) = \frac{\prod_{i=1}^{K}(b_i - a_i)^2}{\det \mathbf{Var}(\boldsymbol{\theta})}, \tag{4}$$

because it has a meaningful interpretation in the context of the general information theory. According to this theory information is defined as the difference between the prior uncertainty and the posterior uncertainty. Let us consider the following situation: the a priori distribution of the parameter is a $K-$dimensional uniform distribution with uncorrelated components on the parallelotop $\prod_{i=1}^{K}[a_i, b_i)$, the aposteriori distribution q is the actual distribution of the parameter. Then it is easy to see that (4) is equivalent to:

$$I_F^2(\boldsymbol{\theta}) = 12^{-K}\,\frac{\det \mathbf{Var}(\boldsymbol{\theta}^{uniform})}{\det \mathbf{Var}(\boldsymbol{\theta}^{q})}.$$

This leads to the further idea that information should be defined as a function of the ratio of the generalized variances, as follows:

$$I_F^3(\boldsymbol{\theta}) = f\left(\frac{\det \mathbf{Var}(\boldsymbol{\theta}^{uniform})}{\det \mathbf{Var}(\boldsymbol{\theta}^{q})}\right).$$

To determine function *f*, which, of course, has to be continuous and strictly increasing, we additionally require that information is expressed in bits. It turns out that the logarithm is the unique function which fulfils this condition, which can be seen as follows. Assume *q* is also a *K*-dimensional uniform distribution with uncorrelated components on the parallelotop $\prod_{i=1}^{K}[c_i, d_i)$, where $a_i \le c_i < d_i \le b_i$, and $d_i - c_i = 2^{-n_i}(b_i - a_i)$. Then we gain additional n_i binary digits

for each component θ_i, thus in total $\sum_{i=1}^{K} n_i$ bits of information. Hence, $\sum n_i = f(2^{2\sum n_i})$, and $f(x) = \frac{1}{2}\log_2 x$.

The generalization of Fisher information measure for the multivariate case is then:

$$I_F^3(\boldsymbol{\theta}) = \frac{1}{2}\log_2\left(\frac{\det \mathbf{Var}(\boldsymbol{\theta}^{uniform})}{\det \mathbf{Var}(\boldsymbol{\theta}^q)}\right) = \frac{1}{2}\log_2 \frac{\left(\prod_{i=1}^{K}(b_i - a_i)\right)^2}{\det \mathbf{Var}(\boldsymbol{\theta}^q)} - K \cdot B, \tag{5a}$$

where q is any distribution concentrated to the support of the uniform distribution and

$$B = 1 + \frac{\ln 3}{2\ln 2} = 1.79.$$

Consequently, for the univariate setting, the corresponding formula is:

$$I_F^3(\theta) = \frac{1}{2}\log_2 \frac{(b-a)^2}{\mathrm{Var}(\theta^q)} - B. \tag{5b}$$

It is important to notice that the original Fisher measure (1) does not differ significantly from (5c) which can be rewritten as:

$$I_F^3(\theta) = A - \frac{1}{2}\log_2 \mathrm{Var}(\theta^q).$$

Both functions have a pole at the origin and a similar behavior near it.

2.2 Kullback measure and its special case

Kullback (1968) defined directed divergence for binary hypothesis testing. Hypotheses H_0 and H_1 imply probability distributions p and q, respectively, with a necessary condition about their supports:

$\forall x : [p(x) = 0 \Rightarrow q(x) = 0]$.

The mean information for discrimination in favour of H_1 against H_0 when H_1 is true is defined by:

$$I_K(q : p) = \int_{-\infty}^{+\infty} q(x)\log_2 \frac{q(x)}{p(x)}\,dx \tag{6a}$$

with an additional definition: $0 \cdot \log\frac{0}{p} = 0$. For the multivariate case the corresponding definition is:

$$I_K(q : p) = \int_{-\infty}^{+\infty} q(\mathbf{x})\log_2 \frac{q(\mathbf{x})}{p(\mathbf{x})}\,dx \tag{6b}$$

In another setting, $I_K(q:p)$ is called the relative entropy or Kullback-Leibler distance between two probability distributions. In can be interpreted as the amount of information necessary to change a prior probability distribution p into the posterior probability distribution q (Cover and Thomas, 1991). According to this definition, information is expressed in bits.

Now, let us consider a special case:

- prior distribution p is uniform on the parallelotop $\prod_{i=1}^{K}[a_i, b_i)$;
- posterior distribution q is normal, truncated to this parallelotop.

When a parameter is estimated, this situation can often be assumed. Prior to the study, only the infimum and supremum of the parameter are known. When the sample data is acquired, normal distribution of the parameter can be assumed due to the Central Limit Theorem.

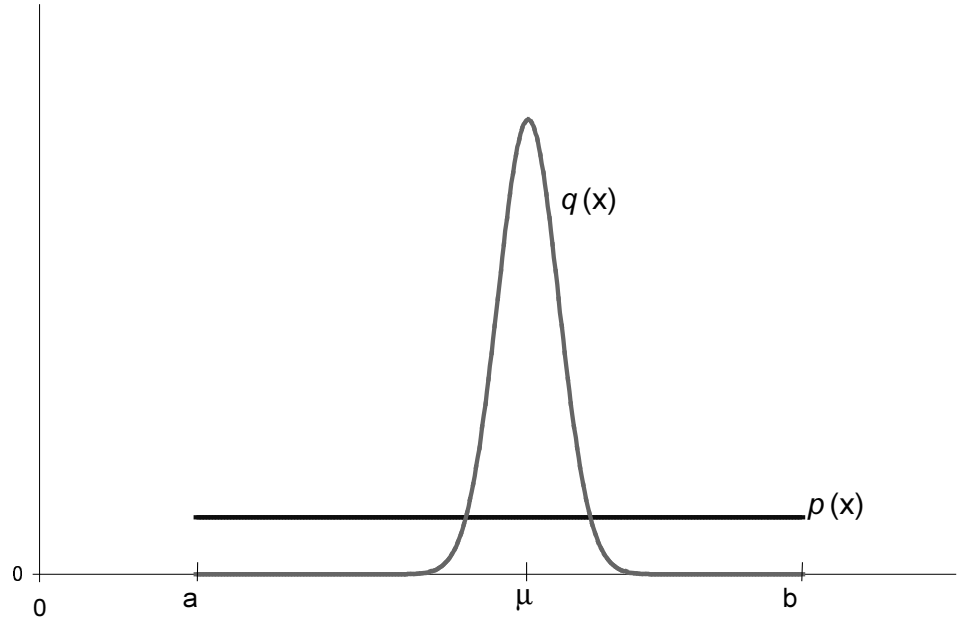

Figure 1: Prior probability distribution p is uniform on $[a,b)$, posterior distribution q is normal truncated to the same interval.

We derived the Kullback-Leibler distance for this case. We have assumed that normal distribution q is truncated to the support of the distribution p. If the truncation of the normal distribution is negligible, e.g. $a < \mu - 3\sigma$, $b > \mu + 3\sigma$, the information is given by:

- multivariate case: K variables

$$I_K(q:p) = \frac{1}{2}\log_2 \frac{\left(\prod_{i=1}^{K}(b_i - a_i)\right)^2}{\det \mathbf{Var}(\boldsymbol{\theta}^q)} - K \cdot \log_2 \sqrt{2\pi e}\,. \tag{7a}$$

- univariate case:

$$I_K(q:p) = \frac{1}{2}\log_2 \frac{(b-a)^2}{\mathrm{Var}(\theta^q)} - \log_2 \sqrt{2\pi e} = \log_2 \frac{b-a}{\sigma} - 2.05\,. \qquad (7b)$$

Derivation of these formulas is long but straightforward. The value of $I_K(q:p)$ in (7a, 7b) is a little bigger if truncation is not negligible.

To sum up: a reasonable modification of the Fisher information measure (5a, 5b) leads to a special case of Kullback information measure (7a, 7b), in the univariate and multivariate case.

3 Shannon-Wiener like derivation of Kullback measure

Suppose that we would like to determine, with some measurements, the value of the unknown parameter θ. The only information about θ we a priori have is that it lies on the interval $[0,1)$. Then, θ can be written in a binary form as follows:

$\theta = 0.d_1d_2d_3... = d_1.2^{-1} + d_2.2^{-2} + d_3.2^{-3} + ...$

Each digit d_i, $(d_i \in \{0,1\})$ can be regarded as a random variable:

$$d_i : \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix},$$

and obtaining its value, we gain the information of 1 bit. Knowing θ exactly, we would have an infinite information.

More realistic is the situation where we want to know the parameter θ to some relevant accuracy. Suppose again that we know only that θ belongs to the interval $[a,b)$. We want to find its value to the accuracy ε (i.e. the distance between the true value θ and its estimate must be above ε). It is reasonable to assume that ε is much smaller than $b-a$. Diminishing ε a little more, we can achieve: $\frac{b-a}{\varepsilon} = n \in \mathbf{N}$.

If we try to guess the value of θ to the required accuracy, the corresponding approximation θ' is a discrete uniformly distributed random variable:

$$\theta' : \begin{pmatrix} [a, a+\varepsilon) & [a, a+2\varepsilon) & \ldots & [a+(n-1)\varepsilon, b) \\ \frac{1}{n} & \frac{1}{n} & \ldots & \frac{1}{n} \end{pmatrix}$$

According to Shannon (Berger, 1981) and Wiener (1948), the entropy of this random variable is:

$$H(\theta') = -\sum_n \frac{1}{n}\log_2 \frac{1}{n} = \log_2 n\,.$$

Now, suppose that we obtained additional information about the parameter θ: the probability density function for its true value is a continuous function $q(x)$

with the support somewhere on the interval $[a,b)$: $x \notin [a,b) \Rightarrow q(x) = 0$. Then, the corresponding approximation θ'' is also represented as a discrete random variable:

$$\theta'' : \begin{pmatrix} [a, a+\varepsilon) & [a+\varepsilon, a+2\varepsilon) & \dots & [a+(n-1)\varepsilon, b) \\ q(z_1)\varepsilon & q(z_2)\varepsilon & \dots & q(z_n)\varepsilon \end{pmatrix},$$

where $z_i \in [a+(i-1)\varepsilon, a+i\varepsilon)$ satisfy the equations

$$q(z_i) = \frac{1}{\varepsilon} \int_{a+(i-1)\varepsilon}^{a+i\varepsilon} q(x)dx \ .$$

The entropy of θ'' is

$$\begin{aligned} H(\theta'') &= -\sum_{i=1}^{n} (q(z_i)\varepsilon) \cdot \log_2 (q(z_i)\varepsilon) \\ &= -\sum_{i=1}^{n} q(z_i) \cdot \log_2 q(z_i) \cdot \varepsilon - \log_2 \frac{b-a}{n} \cdot \sum_{i=1}^{n} q(z_i)\varepsilon \end{aligned}$$

The first sum in the last expression is an integral sum for a continuous function. For small ε (i.e. for big n) its value is close to

$$\int_a^b q(x) \cdot \log_2 q(x) \cdot dx \ .$$

The sum in the second term equals 1. The information gained with $q(x)$ is then, in accordance with Shannon's theory:

$$\begin{aligned} I_p(\theta) = H(\theta') - H(\theta'') &= \log_2 (b-a) + \sum_{i=1}^{n} q(z_i) \cdot \log_2 q(z_i) \cdot \varepsilon \\ &\approx \log_2 (b-a) + \int_a^b q(x) \cdot \log_2 q(x) \cdot dx \qquad (8) \\ &= \int_a^b q(x) \cdot \log_2 \frac{q(x)}{(b-a)^{-1}} \cdot dx \quad . \end{aligned}$$

The last expression in (8) is a form of the expression (6a), with suitable change of symbol *p*. Analogous procedure in the multivariate setting gives us (6b) for *p* uniform. Therefore, Wiener's and Shannon's principles lead us precisely to Kullback information measure.

4 Conclusions

Considering the situation when the prior distribution of the parameter is uniform and the posterior distribution is truncated normal, two different approaches (modified Fisher's and Shannon-Wiener's) lead to a special case of Kullback information measure (6a, 6b), in the univariate as well as in the multivariate case.

This situation can often be assumed when an unknown parameter is estimated: the prior knowledge about the parameter is that it lies in a particular interval, the

posterior knowledge originates from the Central Limit Theorem as a normal distribution with a reasonably small variance.

We find the formula (6a) interpretable in the following sense: the calculated information divided by $\log_2 10 = 3.322$ gives an impression of the number of gained decimal digits of the unknown parameter.

References

[1] Arvantis, L.G. and Afonya, B. (1971): Use of the generelized variance and the gradient projection method in multivariate stratified sampling. *Biometrics,* **27**, 119-27.

[2] Berger, T. (1981): Information theory and coding theory. In S. Kotz and N.L. Johnson (Eds): *Encyclopedia of Statistical Sciences*, **4**, 125-141.

[3] Cochran, W.C. (1953): *Sampling Techniques*. New York: Wiley.

[4] Cover, T.C. and Thomas, J.A. (1991): *Elements of Information Theory*. New York: Wiley.

[5] Kullback, S. (1968): *Information Theory and Statistics*. Dover Publications.

[6] Kullback, S. (1981): Kullback information. In S. Kotz and N.L. Johnson (Eds): *Encyclopedia of Statistical Sciences*, **4**, 421-425.

[7] Wiener, N. (1948). *Cybernetics*. New York: Wiley.