

On the Measure of Representativeness of a Sample and its Use in Statistical Inference

Sulla Misura della Rappresentatività di un Campione e sul suo Uso nell'Inferenza Statistica

Salvatore Bertino

Dipartimento di Statistica, Probabilità e Statistiche Applicate,
Università "La Sapienza", Roma, e-mail: salvatore.bertino@uniroma1.it

Riassunto: Viene proposta una misura della rappresentatività di un campione casuale estratto da una distribuzione assolutamente continua. Tale misura viene utilizzata per definire la funzione di rappresentatività mediante la quale si possono ricavare procedure inferenziali che riguardano, ad esempio, la stima dei parametri, il confronto tra ipotesi, gli insiemi di stima, il test delle ipotesi, la scelta del modello.

Keywords: Index of the Representativeness, Representativeness function, Parameter estimation, Choice among Hypotheses, Set Estimation, Hypotheses testing, Model choice.

1. The measure of the representativeness of a sample

The representativeness of a sample can be defined as “the degree of capability of the sample to exhibit the characteristics of the population it derives from”.

Let $\mathbf{x} = \{x_{(1)}, x_{(2)}, x_{(3)}, \dots, x_{(n)}\}$ an ordered sample of size n deriving from a random variable (rv) X with an absolutely continuous distribution and let $F(x) = \Pr\{X \leq x\}$ the corresponding cumulative distribution function (cdf). Denote with $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ the vector of sample values transformed according to the relation $y_i = F(x_{(i)})$. As it is known, the vector $\mathbf{y}$ can be seen as an ordered sample deriving from a uniform distribution on $[0,1]$.

Let $D(\mathbf{x}, F) = D(\mathbf{y}, U)$ be the squared distance, measured through the dissimilarity index of degree 2 (Landenna, 1956), between the uniform distribution on $[0,1]$ and the empirical cdf of the transformed sample, that is:

$$D(\mathbf{x}, F) = D(\mathbf{y}, U) = \int_0^1 \int_0^1 (y - z)^2 d \min\{U_c(y), z\}$$

where $U_c(y) = r/n$ for $y_{(r)} \leq y < y_{(r+1)}$, $r = 0, 1, 2, \dots, n$, $y_{(0)} = 0$, $y_{(n+1)} = 1$

Given the sample $\mathbf{x}$ and the cdf F , the value $D(\mathbf{x}, F)$ is a measure of how much the empirical cdf of the transformed sample is close to the cdf it derives from, that is $D(\mathbf{x}, F)$ can be seen as an inverse measure of how much the sample represents its parent population. It can be shown that:

$$D(\mathbf{x}, F) = \frac{1}{12n^2} + \frac{1}{n} \sum_{r=1}^n \left(y_{(r)} - \frac{2r-1}{2n} \right)^2 \quad (1)$$

Observe that $D(\mathbf{x}, F)$ will assume values between a strictly positive minimum ($=1/(12n^2)$) and a fixed maximum ($=1/3$).

In order to have an index $R(\mathbf{x}, F)$ such that $0 \leq R(\mathbf{x}, F) \leq 1$, we propose to measure the representativeness of the sample $\mathbf{x}$ with respect to the cdf F through the value

$$R(\mathbf{x}, F) = \left(\max_{\mathbf{x} \in X} D(\mathbf{x}, F) - D(\mathbf{x}, F) \right) / \left(\max_{\mathbf{x} \in X} D(\mathbf{x}, F) - \min_{\mathbf{x} \in X} D(\mathbf{x}, F) \right) \quad (2)$$

Hence the representativeness of an ordered sample $\mathbf{x} = (x_{(1)}, x_{(2)}, \dots, x_{(n)})$ drawn from a rv with cdf $F(x)$, measured according to (2), will be given by the formula:

$$R(\mathbf{x}, F) = 1 - \frac{12n}{4n^2 - 1} \sum_{r=1}^n \left(F(x_{(r)}) - \frac{2r-1}{2n} \right)^2 \quad (3)$$

The value $R(\mathbf{x}, F)$ will be called *representativeness index*.

2. The Representativeness Function

Let $\mathbf{F}$ a family of absolutely continuous cdf's (which can be the family of all the absolutely continuous distributions) and $\mathbf{X}$ the set of the samples of size n which can be drawn from at least one of the distributions of the family $\mathbf{F}$. Hence the value $R(\mathbf{x}, F)$, as a function of $\mathbf{x}$, will show how the different samples represent in a different degree the same distribution. *Vice versa*, if we read $R(\mathbf{x}, F)$ as a functional on $\mathbf{F}$ for a given $\mathbf{x}$, we can see how differently the sample $\mathbf{x}$ represents the various distributions of the family $\mathbf{F}$.

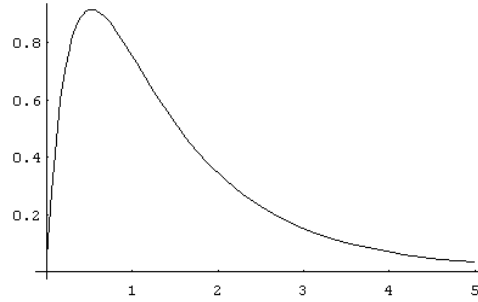


Figure 1: Graph of $R(\mathbf{x}, F)$ for a sample $\mathbf{x}$ simulated from a negative exponential distribution of mean 1 as a function of the parameter θ characterizing the negative exponential distributions (cdf: $F(x) = 1 - \exp(-\theta x)$).

The quantity $R(\mathbf{x}, F)$, as a function of the two variables $\mathbf{x}$ and F , will be called *representativeness function*.

The representativeness function can be used in a statistical framework with different purposes: parameter estimation, hypotheses comparison, model choice and so on. Its potentialities are not yet fully explored but the following examples give relevant hints.

3. Parameter estimation

Let $F_{\theta}(\cdot)$ be the cdf, depending on a vector parameter θ , of a rv X , and $\mathbf{x} = \{x_{(1)}, x_{(2)}, x_{(3)}, \dots, x_{(n)}\}$ an ordered sample of X of size n . An estimate of θ can be obtained choosing the vector which maximises the representativeness of the sample.

For instance, if X is uniformly distributed on $[\mu - \lambda/2, \mu + \lambda/2]$, where λ and μ are unknown parameters, the corresponding estimates turn out to be:

$$\tilde{\mu} = \sum_{r=1}^n x_{(r)} / n = \bar{x}$$

$$\tilde{\lambda} = \frac{n \sum_{r=1}^n (x_{(r)} - \bar{x})^2}{\sum_{r=1}^n r(x_{(r)} - \bar{x})} = \frac{2n^2 s^2}{2 \sum_{r=1}^n r x_{(r)} - n(n+1)\bar{x}}$$

4. Choosing between hypotheses

A choice between two simple hypotheses can be done in a symmetrical way (that is without assigning different roles to the two hypotheses) pointing out the hypothesis which has the maximum representativeness, given the sample.

5. Sampling distribution of the representativeness index

A remarkable property of the representativeness index $R(x, F)$ is that its sampling distribution, under F , does not depend on F . In fact from (3) we get:

$$R(n) = 1 - \frac{12n}{4n^2 - 1} \sum_{r=1}^n \left(Y_{(r)} - \frac{2r-1}{2n} \right)^2$$

where $Y_{(r)} = F(X_{(r)})$, $(r=1, 2, \dots, n)$, are the ordered components of a random sample drawn from an uniform distribution on $[0, 1]$.

The analytical determination of the distribution of R cannot be easily specified. However, a good approximation can be obtained through simulation; in particular it is immediately clear that this distribution depends on n .

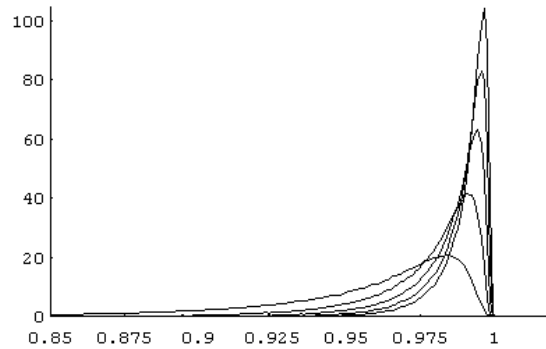


Figure 2: Graphs of empirical distributions of the representativeness index obtained by simulating 1,000,000 samples of size 10, 20, 30, 40, 50.

6. Set estimation

A natural way to define a set estimation is to include in the set all the parameter vectors such that the corresponding representativeness index, for the given sample, is larger or equal to a specified quantile of the sampling distribution, for instance the quantiles of level 5%, 1% and so on.

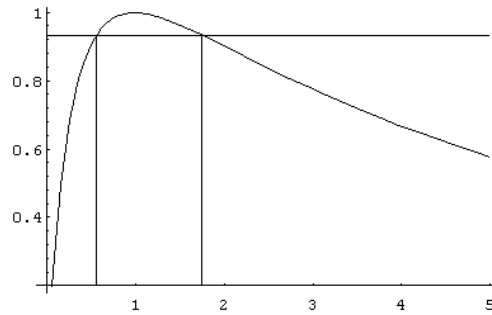


Figure 3: Graph of $R(x, F)$ for the sample x in the Example 1.

Example 1. Let us assume that the sample $x = \{0.363107, 0.751177, 0.557004, 1.06977, 2.63048, 0.861171, 0.158519, 2.29445, 1.93354, 0.552838, 0.0965417, 3.15438, 0.255703, 1.72489, 0.256191, 1.09645, 0.215457, 0.436102, 1.52554, 0.0700184\}$, of size 20, is drawn from a distribution with cdf $F(x) = 1 - \exp(-\lambda x)$, dependent on the parameter λ . The set estimate for λ at the level 5% is obtained through the condition $R(x, F) \geq 0.93211$, since 0.93211 is approximately the quantile at the level 5% of the sampling distribution of the representativeness index for samples of size 20. We get this way the interval $0.553987 \leq \lambda \leq 1.7482$.

7. Hypotheses testing

The testing of hypotheses according to the standard procedure is particularly simple since the representativeness index has a distribution free sampling distribution.

8. Model choice

Given a sample x , any model can be evaluated by means of the maximum of the representativeness index under the model itself. Then the “best” model will be the model for which this maximum is also maximum among all the considered models.

Example 2. Given the sample $x = \{77.388, 23.581, 5.649, 11.565, 5.425, 2.215, 4.157, 5.590, 2.675, 36.276, 19.518, 11.409, 0.204, 1.887, 1.124, 0.301, 1.058, 0.081, 25.597, 3.914\}$ we wish to compare the model consisting of LogNormal distributions with the model consisting of Gamma distributions.

It is easy to verify, numerically, that the maximum of $R(x, F)$ with a LogNormal F is $R(x, LN) = 0.99679774$, while the maximum of $R(x, F)$ with a Gamma F is $R(x, G) = 0.99603921$. Hence the LogNormal will be preferred from this point of view.

References

- Bertino S. (1971) Gli indici di dissomiglianza e la stima dei parametri, *Studi di probabilità, statistica e ricerca operativa in onore di Giuseppe Pompilj*, Oderisi - Gubbio, 1971.
- Dall’Aglio G. (1956) Sugli estremi dei momenti delle funzioni di ripartizione doppia, *Annali della Scuola Normale di Pisa*, ser. III, vol. X, 1-2.
- Landenna G. (1956) La dissomiglianza, *Statistica*, Anno XVI, I.
- Landenna G. (1957) Osservazioni sulla connessione, *Statistica*, Anno XVII, I.