Calculation of the information function with incomplete samples

Khanimkulov Bakhrom Rokhmonkulovich

Associate Professor, Chirchik State Pedagogical University, Uzbekistan

e-mail: xanimkulovbaxrom@gmail.com

Annotation. Abstract. In this paper we derive formulas for the information function under Cramér–Rao regularity conditions for a density *F* in both continuous and discrete cases when the parameter $\boldsymbol{\theta}$ is unknown. We prove that the regularity conditions remain valid even when the sample is incomplete. Based on the stated theorems, two main results are obtained: (1) an explicit expression for the information function and (2) the applicability of the Kramér–Rao bound in the presence of incomplete samples. Examples illustrating the results and applications are provided.

Keywords: Density function, information function, Kramér–Rao regularity conditions, incomplete sample, mean values, parameter estimation.

The problems of determining and estimating parameters in the process of statistical observation are one of the main research areas of probability theory and mathematical statistics. One of the most important concepts used in assessing the quality of estimates is the *information function* (Fisher information), which is of fundamental importance in measuring the accuracy of the selected parameters of a statistical model. The information function is closely related to the Cramér–Rao limit and theoretically determines the ability of estimators to achieve the smallest variance.

Many theoretical results work when complete observations or complete samples are available. However, in applied statistics, data are often observed in the form of incomplete samples due to various reasons - technical errors, interruptions in observation, incompletely recorded values, measurement limitations, or unfavorable experimental conditions. In such conditions, the exact calculation of the information function and the study of its properties are one of the important theoretical and practical issues. In particular, the question of the fulfillment of the regularity conditions and the existence of Fisher information in cases of incomplete samples has not been sufficiently studied.

In the literature, the Kramer–Rao regularity conditions are usually given in terms of complete sampling. Therefore, how to generalize these conditions to the incomplete sampling situation and to determine the calculation formulas for the information function in continuous and discrete forms of the density function remain one of the pressing issues.

In this article, we present general formulas for the information function with respect to an unknown parameter θ when the density function $\boldsymbol{F(x;\theta)}$ is continuous and discrete. We also prove new results on the validity of the Kramer–Rao type regularity conditions even in cases of incomplete sampling.

In statistics, a complete sample is a data set that contains all possible values of the population, from which the parameters of the distribution can be fully determined.

An incomplete sample represents a situation where some observations are missing or the available data are only partially known. In such cases, the density function (i.e., the probability density) is constructed based on the available (incomplete) observations, rather than on the basis of the complete data.

If $\boldsymbol{X=(X_1,X_2,\dots,X_n)}$ is a complete sample and its probability density function is represented by $\boldsymbol{f(x_1,x_2,\dots,x_n;\ \theta)}$, then the density function for an incomplete sample $\boldsymbol{Y=(Y_1,Y_2,\dots,Y_m), m<n}$ is defined as follows:

$$g(y_1,y_2,\dots,y_m;\theta)=\int f(x_1,x_2,\dots,x_n;\ \theta)dx_{m+1}\dots dx_n$$

i.e., obtained by integrating over the missing data. This approach allows us to determine the probability density function for incomplete samples.

Suppose that the random variables $\boldsymbol{X_1,X_2,X_3\sim N(\mu,\sigma^2)}$ – that is, they have the same normal distribution. The probability density function for the complete sample is:

$$\boldsymbol{f(x_1, x_2, x_3;\ \mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{\frac{3}{2}}} exp\left(-\frac{(x_1-\mu)^2+(x_2-\mu)^2+(x_3-\mu)^2}{2\sigma^2}\right)}$$

Suppose we only have $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ observed, meaning that the value of $\boldsymbol{X_3}$ is missing (incomplete sample). So, the incomplete sample is: $\boldsymbol{Y = (X_1, X_2)}$.

For this case, the density function:

We integrate over $\boldsymbol{X_3}$, which is not complete from the density function:

$$g(x_1, x_2;\ \mu, \sigma) = \int_{-\infty}^{\infty} f(x_1, x_2, x_3;\ \mu, \sigma) dx_3$$

By the normality property, this integral gives the following result:

$$g(x_1, x_2;\ \mu, \sigma) = \frac{1}{2\pi\sigma^2} exp\left(-\frac{(x_1-\mu)^2+(x_2-\mu)^2}{2\sigma^2}\right).$$

It is clear that by integrating over the missing X_3, our density function changes, but it remains in the form of a normal distribution. This approach is the basis for statistical analysis, probability, and Fisher information calculations for incomplete samples.

Incomplete sample density functions:

F- with absolute discontinuity

$$g(x;\theta) = \begin{cases} F(a;\theta) = \int_{-\infty}^{a} f(u;\theta)du, agar\ x = a \\ f(x;\theta), agar\ a < x < b \\ 1 - F(b;\theta) = \int_{b}^{\infty} f(u;\theta)du, agar\ x = b \end{cases} \quad (1)$$

by the formula and in the case of *F*- being discrete,

$$g(x;\theta) = \begin{cases} \sum_{(i;x_i<a)} f(x_i;\theta)\,, agar\ x = a \\ f(x_i;\theta), agar\ x = x_i \in (a;b) \\ \sum_{(i;x_i>b)} f(x_i;\theta)\,, agar\ x = b \end{cases} \quad (2)$$

$$g_n(x^n;\theta) = [F(a;\theta)]^{m_1} * [1 - F(b;\theta)]^{m_2} * \prod_{i=1}^{n} [f(x_i;\theta)]^{E_i} \quad (3)$$

let us consider the formulas for calculating Fisher's information function for an unknown parameter θ when given by the formulas. If we apply this directly (i.e., with certain regularity conditions met) to formulas (1) and (2), then in the case of incomplete sampling, the information function is continuous

$$I^*(\theta) = \left(\frac{\partial logF(a;\theta)}{\partial\theta}\right)^2 \cdot F(a;\theta) + \int_a^b \left(\frac{\partial logF(x;\theta)}{\partial\theta}\right)^2 \cdot f(x;\theta)dx +$$

$$+ \left(\frac{\partial \log(1-F(b;\theta))}{\partial\theta}\right)^2 \cdot (1 - F(b;\theta)) \quad (4)$$

using the formula and in discrete form

$$I^*(\theta) = \left(\frac{\partial logF(a;\theta)}{\partial\theta}\right)^2 \cdot F(a;\theta) + \sum_{\{i;a<x_i<b\}} \left(\frac{\partial logF(x;\theta)}{\partial\theta}\right)^2 \cdot f(x;\theta) +$$

$$+ \left(\frac{\partial \log(1-F(b;\theta))}{\partial\theta}\right)^2 \cdot (1 - F(b;\theta)) \quad (5)$$

that is, it is calculated using a formula similar to (4).

Now let us recall the regularity conditions mentioned above:

$(\boldsymbol{R1})$ Let the set $\{\boldsymbol{x; f(x;\theta) > 0}\}$ be independent of θ;

(R2) The derivative $\frac{\boldsymbol{\partial f(x;\theta)}}{\boldsymbol{\partial \theta}}$ exists and is finite for all *θ*;
(R3) For *i*=1,2 and for all *θ*

$$\begin{cases} \int\limits_{-\infty}^{\infty} \left|\frac{\partial^i f(x;\theta)}{\partial \theta^i}\right| dx < \infty, \qquad F - if\ it\ is\ absolutely\ continuous; \\ \sum\limits_{i} \left|\frac{\partial^i f(x;\theta)}{\partial \theta^i}\right| dx < \infty, \qquad F - if\ it\ is\ discrete; \end{cases}$$

(R4) The condition $\boldsymbol{I_x(\theta)} < \infty$is satisfied for all *θ*.

Usually, in mathematical statistics, the conditions *R(1)-R(4)* are called the Cramer–Rao regularity conditions.

Theorem 1: If the regularity conditions *R(1)-R(4)* are satisfied, then the derivatives $\frac{\boldsymbol{\partial F(a;\theta)}}{\boldsymbol{\partial \theta}}$ and $\frac{\boldsymbol{\partial F(b;\theta)}}{\boldsymbol{\partial \theta}}$ exist for all *θ*, which are finite and this

$$\left(\frac{\partial logF(a;\theta)}{\partial \theta}\right)^2 \cdot F(a;\theta) \leq I_x(\theta), \tag{6}$$

$$\left(\frac{\partial \log\left(1-F(b;\theta)\right)}{\partial \theta}\right)^2 \cdot (1 - F(b;\theta)) \ \leq I_x(\theta), \tag{7}$$

$$I^*(\theta) \leq 3I_x(\theta) \tag{8}$$

Proof. We restrict ourselves only to the case where F-absolutely continuous. The claim for the discrete case is proved in a similar way. By conditions *R(1)-R(3),*

$$\underset{\theta \in H}{Sup}\left[\frac{\partial F(a;\theta)}{\partial \theta}\right] = \underset{\theta \in H}{Sup}\left|\frac{\partial}{\partial \theta}\int_{-\infty}^{a} f(x;\theta)dx\right|.$$

It is sufficient to take into account inequalities (7) and the following, that is, for all *θ*:

$$I^*(\theta) \leq 2I_x(\theta) + \int_a^b \left(\frac{\partial logf(x;\theta)}{\partial \theta}\right)^2 \cdot f(a;\theta)\, dx \leq 3I_x(\theta) < \infty.$$

The theorem is proved.

It is known that as *a*→-∞ and *b*→+∞, the approximation $\boldsymbol{I^*(\theta)} \rightarrow \boldsymbol{I(\theta)}$ is reasonable, that is, as the interval that does not allow the observation results to be incomplete disappears, the Fisher information corresponding to the incomplete sample should approach the Fisher information corresponding to the complete sample. In Theorem 2 below, we prove this claim rigorously. As above, we prove it only for the continuous case.

Theorem 2: Assuming that conditions R(1)-R(4) hold, if a→-∞ and b→+∞, then

$$I^*(\theta) \rightarrow I(\theta) \tag{9}$$

Proof. The proof of this claim is trivial, since $\int_{-\infty}^{\infty} \boldsymbol{f(x;\theta)dx = 1}$ is an equality for all *θ* and the regularity conditions *R(1)-R(3)*

$$\frac{\partial}{\partial \theta}\int\limits_{-\infty}^{\infty} f(x;\theta)dx = \frac{\partial}{\partial \theta}$$

$$\text{yoki} \int_{-\infty}^{\infty} \frac{\partial f(x;\theta)}{\partial \theta}dx = 0 \tag{10}$$

It follows that. We know that from the *R(4)*-regularity condition

$$\lim_{\substack{a\downarrow -\infty \\ b\uparrow +\infty}} \int_a^b \left(\frac{\partial logf(x;\theta)}{\partial \theta}\right)^2 \cdot f(a;\theta)\, dx \uparrow \int_{-\infty}^{\infty} \left(\frac{\partial logf(x;\theta)}{\partial \theta}\right)^2 \cdot f(a;\theta)\, dx = I_x(\theta) < \infty; \tag{11}$$

So, to prove the claim

$$\lim_{a\downarrow-\infty}\left(\frac{\partial logF(a;\theta)}{\partial\theta}\right)^2 F(a;\theta)=0 \tag{12}$$

$$\lim_{b\uparrow\infty}\left(\frac{\partial\log(1-F(b;\theta))}{\partial\theta}\right)^2(1-F(b;\theta))=0 \tag{13}$$

it is enough to show that the equations are valid.

$$0\le\lim_{a\downarrow-\infty}\left(\frac{\partial logF(a;\theta)}{\partial\theta}\right)^2 F(a;\theta)\le\lim_{a\downarrow+\infty}\int_{-\infty}^{a}\left(\frac{\partial logf(x;\theta)}{\partial\theta}\right)^2\cdot f(x;\theta)\,dx=0$$

$$0\le\lim_{b\uparrow\infty}\left(\frac{\partial\log(1-F(b;\theta))}{\partial\theta}\right)^2(1-F(b;\theta))\le\lim_{b\uparrow+\infty}\int_{b}^{+\infty}\left(\frac{\partial logf(x;\theta)}{\partial\theta}\right)^2\cdot f(x;\theta)\,dx=0$$

So, (9) is valid.

The theorem is proved. It should be noted that so far we have only considered the Fisher information for the case where the result of the observation is incomplete. In the case of a complete sample, according to the above additivity property $I_{X^{(n)}}(\theta)=nI_x(\theta)$, that is, to calculate the information of the sample $X^{(n)}=(x_1,\dots,x_n)$, it was enough to multiply the information of $I_x(\theta)$ by n. Now, for the case of an incomplete sample, if we take into account formula (6),

$$I_n^x(\theta)=M_\theta\left[\frac{\partial logg_n(x^n;\theta)}{\partial\theta}\right]^2=$$

$$=M_\theta\left[m_1\frac{\partial logF(a;\theta)}{\partial\theta}+m_2\frac{\partial\log(1-F(b;\theta))}{\partial\theta}+\sum_{i=1}^{n}E_i\frac{\partial logf(X_i;\theta)}{\partial\theta}\right]^2$$

we make sure that it is determined by a complex expression. This, in turn, shows that calculating the information function with incomplete samples is not a trivial matter, unlike with complete samples. It should also be noted that so far we have only considered the case when the unknown parameter is a scalar, that is, one-dimensional. If the parameter is multidimensional, that is, a vector $\theta=(\theta_1,\dots,\theta_k)$, then we have to deal with Fisher information matrices, not information functions. In this case, the elements of the Fisher information matrix $J_x(\theta)=\left\|I_x^{ij}(\theta)\right\|_{i,j=\overline{1,k}}$ corresponding to the random variable X are:

$$I_n^{ij}(\theta)=M_\theta\left[\left(\frac{\partial logf(X;\theta)}{\partial\theta_i}\right)\left(\frac{\partial logf(X;\theta)}{\partial\theta_j}\right)\right]$$

$$=\begin{cases}\displaystyle\int_{-\infty}^{\infty}\left(\frac{\partial logf(x;\theta)}{\partial\theta_i}\right)\left(\frac{\partial logf(x;\theta)}{\partial\theta_j}\right)f(x;\theta)dx;\\ \displaystyle\sum_{\{x;a<x_k<b\}}\left(\frac{\partial logf(x_k;\theta)}{\partial\theta_i}\right)\left(\frac{\partial logf(x_k;\theta)}{\partial\theta_j}\right)f(x_k;\theta)\,;\end{cases}$$

is determined by the formulas. If we consider the case where the observation of X is incomplete, then the elements of the Fisher information matrix $J_x(\theta)=\left\|I_*^{ij}(\theta)\right\|_{i,j=\overline{1,k}}$ are similar to (7), (8)

$$I_n^{ij}(\theta)=\left(\frac{\partial logF(a;\theta)}{\partial\theta_i}\right)\left(\frac{\partial logF(a;\theta)}{\partial\theta_j}\right)F(a;\theta)+\left(\frac{\partial\log(1-F(b;\theta))}{\partial\theta_i}\right)\left(\frac{\partial\log(1-\mathrm{F}(b;\theta)}{\partial\theta_j}\right)$$
$$+$$
$$+I^{ij}*(a;b;\theta);$$

determined by formulas, here

$$I^{ij} * (a;b;\theta) = \begin{cases} \int_a^b \left(\frac{\partial logf(x;\theta)}{\partial\theta_i}\right)\left(\frac{\partial logf(x;\theta)}{\partial\theta_j}\right) f(x;\theta)dx; \\ \sum_{\{x;a<x_k<b\}} \left(\frac{\partial logf(x_k;\theta)}{\partial\theta_i}\right)\left(\frac{\partial logf(x_k;\theta)}{\partial\theta_j}\right) f(x_k;\theta); \end{cases}$$

will be.

References

1. Gmurman, V. E. (2012). Theory of probability and mathematical statistics. M.: Vysshaya school.

2. Korolyuk, V. S., & Pugachyov, V. S. (2004). Osnovy statisticheskoy teorii informatsii. M.: Nauka.

3. Makhmudova D.M, Khanimkulov BR Probability theory and mathematical statistics. Zebo print publishing house 2025.

4. Fisher, R. A. (1925). *Theory of Statistical Estimation.* Proceedings of the Cambridge Philosophical Society, **22**, 700–725.

5. Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press.

6. Rao, C. R. (1945). *Information and the Accuracy Attainable in the Estimation of Statistical Parameters.* Bulletin of the Calcutta Mathematical Society, **37**, 81–91.

7. Lehmann, E. L., & Casella, G. (1998). *Theory of Point Estimation.* Springer.

8. Kendall, M., & Stuart, A. (1979). *The Advanced Theory of Statistics.* Vol. 2. Griffin & Co.

9. Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* Wiley-Interscience.

10. Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman & Hall.

11. Tikhonov, A. N., & Arsenin, V. Y. (1977). *Solutions of Ill-Posed Problems.* Winston & Sons.

12. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society, **39**, 1–38.