Miloš Pavlík

Approximate construction of a two-dimensional confidence region

Aplikace matematiky, Vol. 15 (1970), No. 5, 305–309

Persistent URL: http://dml.cz/dmlcz/103301

SVAZEK 15 (1970) APLIKACE MATEMATIKY ČÍSLO 5

APPROXIMATE CONSTRUCTION OF A TWO-DIMENSIONAL CONFIDENCE REGION

Miloš Pavlík

(Received November 11, 1968)

It is, as a rule, not very difficult to construct an interval estimate of a single parameter characterizing some known distribution of an unidimensional random variate. However, the exact solution of an analogous problem with $m > 1$ parameters, viz., the construction of a confidence region in m-dimensional Euclidean space $\mathfrak{E}_m$, may represent a rather difficult and cumbersome task. Approximate constructions were proposed for large samples, based upon the fact that certain functions of sample parameters are asymptotically normally distributed ([3], [4], [5]); there are, however, cases in which even these constructions involve a great deal of toilsome work. In some such cases the limit distribution of the χ^2 statistic, calculated with sample values, may provide an useful basis for an advantageous approximate construction. In the present paper the principle of this construction with $m = 2$ will be briefly exposed; theoretically it may be extended easily to any number of parameters, although the usefulness of constructing confidence regions in multidimensional spaces is, as to its applicability in practice, reduced considerably by the abstractness of multidimensional objects.

Let X be a random variate the distribution of which is specified by the distribution function $F(x, \zeta)$; the unknown population value of parameter couple $\zeta = \{\zeta_1, \zeta_2\}$ be $\zeta_0 = \{\zeta_{01}, \zeta_{02}\}$. Let us have a system of $n + 1$ intervals $\mathfrak{a}_i = \langle a_i, a_{i+1})$, $i \in \in \{0, 1, 2, \ldots, n\}$, such that $\bigcup_{i=0}^{n} \mathfrak{a}_i = \mathfrak{E}_1$ and $\mathsf{P}\{X \in \mathfrak{a}_i \mid \zeta\} = F(a_{i+1}, \zeta) - F(a_i, \zeta) = \pi_i(\zeta)$, and two open intervals $\mathfrak{g}_1$, $\mathfrak{g}_2$; let $\mathfrak{g}_1 \times \mathfrak{g}_2 = \mathfrak{G}$ and

$$\zeta \in \mathfrak{G} \Rightarrow \pi_i(\zeta) \in (0; 1)\,, \tag{1}$$

$$\zeta \notin \mathfrak{G} \Rightarrow \pi_i(\zeta) = 0\,, \tag{2}$$

$$\lim_{\zeta_j \to g_j} \pi_i(\zeta) \in \{0; 1\} \tag{3}$$

for all j's ($j \in \{1; 2\}$) and i's, if $g_j \in \{\inf \mathfrak{g}_j, \sup \mathfrak{g}_j\}$. The function $\pi_i(\zeta)$ let be continuous on $\mathfrak{G}$ for all i's. Furthermore, let $\boldsymbol{X}$ be a sample set of h values of X; on $\boldsymbol{X}$ a system of $n + 1$ disjunct subsets $\boldsymbol{X}_i$ let be defined, h_i elements in each, such that $\boldsymbol{X}_i \subset \mathfrak{a}_i$. The vector $\boldsymbol{h} = \{h_i\}_{i=0}^n$ is the value of a random vector $\boldsymbol{H} = \{H_i\}_{i=0}^n$ which is polynomially distributed, the parameters being $\pi_i(\zeta)$.

Now let $\varphi_\alpha(\boldsymbol{H}, \zeta)$ be some random function of ζ such that $\mathfrak{c}_\alpha(\boldsymbol{H}) = \{\boldsymbol{\kappa} : \varphi_\alpha(\boldsymbol{H}, \boldsymbol{\kappa}) = 0\}$ is a continuous closed curve in $\mathfrak{G}$ the inner region $\mathfrak{Z}_\alpha(\boldsymbol{H})$ of which satisfies

$$(4) \qquad \lim_{h\to\infty} \mathsf{P}\{\zeta_0 \in \mathfrak{Z}_\alpha(\boldsymbol{H})\} = 1 - \alpha$$

if $1 - \alpha$ denotes the confidence level chosen; thus $\mathfrak{Z}_\alpha(\boldsymbol{H})$ represents, if h is large enough, an approximately $100(1 - \alpha)$ per cent confidence region for ζ_0 and any sample defines one realization $\varphi_\alpha(\boldsymbol{h}, \zeta)$ or $\mathfrak{c}_\alpha(\boldsymbol{h})$ or $\mathfrak{Z}_\alpha(\boldsymbol{h})$ of random function $\varphi_\alpha(\boldsymbol{H}, \zeta)$ or random curve $\mathfrak{c}_\alpha(\boldsymbol{H})$ or random region $\mathfrak{Z}_\alpha(\boldsymbol{H})$, respectively. Let the space of all admissible $\mathfrak{Z}_\alpha(\boldsymbol{H})$'s be denoted with $\mathscr{Z}_\alpha(\boldsymbol{H})$; now let us seek for a function $\varphi_\alpha(\boldsymbol{H}, \zeta)$ such that the construction of any realization of $\mathfrak{c}_\alpha(\boldsymbol{H})$ and thus of $\mathfrak{Z}_\alpha(\boldsymbol{H})$, would not be very difficult.

Let $P_r(\chi^2)$ denote the χ^2 distribution function with r degrees of freedom and let $P_r(\chi_r^2[\alpha]) = 1 - \alpha$. Let us define on $\mathfrak{G}$ a random function

$$(5) \qquad \chi^2(\boldsymbol{H}, \zeta) = \sum_{i=0}^{n} [h\, \pi_i(\zeta)]^{-1}\, [H_i - h\, \pi_i(\zeta)]^2 = h^{-1} \sum_{i=0}^{n} [\pi_i(\zeta)]^{-1}\, H_i^2 - h$$

of ζ; it is well known that

$$(6) \qquad \lim_{h\to\infty} \mathsf{P}\{\chi^2(\boldsymbol{H}, \zeta) \geqq \chi^2(\boldsymbol{h}, \zeta)\} = 1 - P_n(\chi^2(\boldsymbol{h}, \zeta))\,,$$

or, what is the same,

$$(6a) \qquad \lim_{h\to\infty} \mathsf{P}\{\chi^2(\boldsymbol{H}, \zeta) < \chi_n^2[\alpha]\} = 1 - \alpha\,.$$

The construction of a confidence region for ζ_0 has no practical significance unless an estimate $\boldsymbol{z} = \{z_1, z_2\} \in \mathfrak{G}$ of ζ_0 exists, constructed with $\boldsymbol{X}$ by the χ^2 minimum or with $\boldsymbol{h}$ by the maximum likelihood procedure, such that $\chi^2(\boldsymbol{h}, \boldsymbol{z}) < \chi_{n-2}^2[\alpha]$, what will be assumed throughout the paper (for otherwise the $\boldsymbol{h}$ value implies the conclusion that the distribution of X is not adequately described by $F(x, \zeta)$). From (1), (2) and (3) it follows that any relization $\chi^2(\boldsymbol{h}, \zeta)$ of (5) is continuous on $\mathfrak{G}$ and that for all j's

$$(7) \qquad \lim_{\zeta_j\to g_j} \chi^2(\boldsymbol{h}, \zeta) = \infty$$

holds; therefore

1. $\chi^2(\boldsymbol{h}, \zeta)$ possesses a minimum value in $\mathfrak{G}$ and, since at least one $\zeta \in \mathfrak{G}$ (viz., $\zeta = \boldsymbol{z}$) exists which fulfils $\chi^2(\boldsymbol{h}, \zeta) < \chi_{n-2}^2[\alpha] < \chi_n^2[\alpha]$, the relation

$$(8) \qquad \min_{\zeta\in\mathfrak{G}} \chi^2(\boldsymbol{h}, \zeta) < \chi_n^2[\alpha]$$

holds;

2. finite closed intervals $\mathfrak{z}_j \subset \mathfrak{g}_j$ exist such that $\zeta_j \notin \mathfrak{z}_j \Leftrightarrow \chi^2(\boldsymbol{h}, \boldsymbol{\zeta}) > \chi^2_n[\alpha]$ with all ζ_k's ($j \neq k \in \{1; 2\}$).

Therefore a continuous closed curve $\mathfrak{b}_\alpha(\boldsymbol{h}) = \{\boldsymbol{\lambda} : \chi^2(\boldsymbol{h}, \boldsymbol{\lambda}) = \chi^2_n[\alpha]\}$ exists in $\mathfrak{G}$, enclosing the region $\mathfrak{Y}_\alpha(\boldsymbol{h}) = \{\boldsymbol{\mu} : \chi^2(\boldsymbol{h}, \boldsymbol{\mu}) < \chi^2_n[\alpha]\}$ which satisfies $\boldsymbol{\zeta}_0 \in \mathfrak{Y}_\alpha(\boldsymbol{h}) \Leftrightarrow \chi^2(\boldsymbol{h}, \boldsymbol{\zeta}_0) < \chi^2_n[\alpha]$, and, therefore, according to (6a),

$$\lim_{h \to \infty} \mathsf{P}\{\boldsymbol{\zeta}_0 \in \mathfrak{Y}_\alpha(\boldsymbol{H})\} = 1 - \alpha . \tag{9}$$

Thus $\mathfrak{Y}_\alpha(\boldsymbol{H}) \in \mathscr{L}_\alpha(\boldsymbol{H})$; the curve enclosing the approximately $100(1 - \alpha)$ per cent confidence region for $\boldsymbol{\zeta}_0$ is given, with any $\boldsymbol{h}$ fixed, by $\varphi_\alpha(\boldsymbol{h}, \boldsymbol{\zeta}) = \chi^2(\boldsymbol{h}, \boldsymbol{\zeta}) - \chi^2_n[\alpha])$, i.e., by the equation

$$h^{-1} \sum_{i=0}^{n} [\pi_i(\boldsymbol{\zeta})]^{-1} h_i^2 - h = \chi^2_n[\alpha] . \tag{10}$$

The conditions under which the limit relations (6), (6a) and (9) may, in practice, be replaced by equalities, are commonly known; they are identical with the conditions of a legitimate use of Pearson χ^2 goodness of fit test.

The curve $\mathfrak{b}_\alpha(\boldsymbol{h})$ can, for any $\boldsymbol{h}$, be constructed graphically if the $\pi_i(\boldsymbol{\zeta})$ values in a sufficiently large neighbourhood of $\boldsymbol{z}$ in $\mathfrak{G}$ are known. Fix some values ζ_j^* of ζ_j in the neighbourhood of z_j and construct, for each ζ_j^*, the curve $\mathfrak{p}_j$ given by

$$\chi_j^2 = h^{-1} \sum_{i=0}^{n} [\pi_i(\{\zeta_j^*, \zeta_k\})]^{-1} h_i^2 - h \tag{11}$$

in the Cartesian coordinate system (ζ_k, χ_j^2). If $\zeta_j^* \in \operatorname{Int} \mathfrak{z}_j$, then $\mathfrak{p}_j$ cuts the line $\chi_j^2 = \chi_n^2[\alpha]$ in two points the abscissae of which let be denoted by $\zeta_k^{(s)}$, $s \in \{1; 2\}$; then, for all s's, $\{\zeta_j^*, \zeta_k^{(s)}\} \in \mathfrak{b}_\alpha(\boldsymbol{h})$ holds. Thus, by constructing, for all ζ_j^*'s, the points $\{\zeta_j^*, \zeta_k^{(s)}\}$ in the Cartesian coordinate system (ζ_j, ζ_k) and connecting them by a smooth closed curve, an approximate graphic representation is obtained of the confidence region having been sought.

As it has been observed above, this procedure may be of some advantage if $F(x, \boldsymbol{\zeta})$ is of such type that other approximate methods are difficult to apply. Let X have, e.g., a Pólya distribution the probability function of which is

$$f \equiv f(x, \boldsymbol{\zeta}) = \binom{n}{x} [\mathrm{B}(\zeta_1, \zeta_2)]^{-1} \, \mathrm{B}(x + \zeta_1, n - x + \zeta_2) = \binom{-\zeta_1 - \zeta_2}{n}^{-1} \binom{-\zeta_1}{x} \binom{-\zeta_2}{n - x} \tag{12}$$

where $x \in \{0, 1, 2, \ldots, n\}$, $\boldsymbol{\zeta} \in (0, \infty)^2$,

$$\mathrm{B}(a, b) = \int_0^1 t^{a-1}(1 - t)^{b-1} \, \mathrm{d}t .$$

The approximate construction of optimum confidence region as proposed by Wilks ([3], [4], [5]) is based upon the fact that the random vector $\{\partial \log L/\sqrt{(h)}\,\partial\zeta_1, \partial \log L/\sqrt{(h)}\,\partial\zeta_2\}$ where $L \equiv L(\boldsymbol{H}, \zeta)$ is the likelihood function of $\boldsymbol{H}$, has, assuming $h \to \infty$, an approximately normal bivariate distribution with zero means and variance-covariance matrix

$$\mathbf{V} = \left\| \mathsf{E}[(\partial \log f/\partial\zeta_j)\,(\partial \log f/\partial\zeta_k)] \right\| ;$$

thus the curve enclosing the approximately $100(1-\alpha)$ per cent confidence region is, for any $\boldsymbol{h}$ fixed, given by

$$h^{-1} \sum_{j,k=1}^{2} w_{jk}[\partial \log L(\boldsymbol{h}, \zeta)/\partial\zeta_j]\,[\partial \log L(\boldsymbol{h}, \zeta)/\partial\zeta_k] = \chi_2^2[\alpha] \tag{13}$$

where $\|w_{jk}\| = \mathbf{V}^{-1}$. In the case of Pólya distribution we have, putting $\psi(u) = \mathrm{d} \log \Gamma(u)/\mathrm{d}u$, $\Gamma(u) = \int_0^\infty \mathrm{e}^{-t} t^{u-1}\,\mathrm{d}t$ [1], and $\psi_j = \psi(n + \zeta_1 + \zeta_2) + \psi(\zeta_j) - \psi(\zeta_1 + \zeta_2)$,

$$\partial \log f/\partial\zeta_1 = \psi(x + \zeta_1) - \psi_1\,, \tag{14}$$

$$\partial \log f/\partial\zeta_2 = \psi(n - x + \zeta_2) - \psi_2\,, \tag{14a}$$

$$\partial \log L(\boldsymbol{h}, \zeta)/\partial\zeta_1 = \sum_{i=0}^{n} h_i\, \psi(i + \zeta_1) - h\psi_1\,, \tag{15}$$

$$\partial \log L(\boldsymbol{h}, \zeta)/\partial\zeta_2 = \sum_{i=0}^{n} h_i\, \psi(n - i + \zeta_2) - h\psi_2\,; \tag{15a}$$

thus it is clear that the equation (13) becomes, if the expressions (14) to (15a) are substituted, decidedly more complicated than the relation

$$h^{-1} \sum_{i=0}^{n} [f(i, \zeta)]^{-1}\, h_i^2 - h = \chi_n^2[\alpha]\,. \tag{16}$$

Therefore in this case it would be essentially easier to proceed in the manner suggested above, although the $\psi(u)$ function is tabulated [2]. The values of Pólya probability function which, as far as the writer is aware, are not tabulated, may be found, e.g., by using the combinatorical form of (12), either directly, by the aid of common logarithmic tables, or by converting factorials into polynomials and applying Horner scheme; a table of these values being once available for given n and an adequate set of ζ values, the construction of $\mathfrak{p}_j$ curves and, thus, of an appropriate number of $\mathfrak{b}_\alpha(\boldsymbol{h})$ curve points is not difficult at all. An additional advantage consists in the fact that the suggested procedure, unlike the construction of the confidence ellipse based upon the asymptotic normality of the distribution of maximum likelihood parameter estimates, does without any exact maximum likelihood or χ^2 minimum estimate constructions which in themselves are rather cumbersome with Pólya distribution; it is sufficient to construct an estimate $\mathbf{z}^*$ of ζ_0 by any easy method, such as that of moments, for the validity of $\chi^2(\boldsymbol{h}, \mathbf{z}^*) < \chi^2_n[\alpha]$ is a sufficient warrant for the existence of $\mathfrak{Y}_\alpha(\boldsymbol{h})$. The ζ point corresponding to the minimum of all χ^2_j values found by computation or by graphic interpolation when constructing the $\mathfrak{p}_j$ curves may, too, be considered mostly to be a fairly good (approximately χ^2 minimum) ζ_0 estimate for many practical purposes.

References

[1] *Bateman, H.* - *A. Erdélyi:* Higher transcendental functions, vol. 1. McGraw-Hill, New York—Toronto—London, 1953.

[2] Таблицы логарифмической производной гамма-функции и ее производных в комплексной области. Вычислительный центр АН СССР, Москва, 1965.

[3] *Wilks, S. S.:* Shortest average confidence intervals for large samples. Ann. Math. Stat. *9*, 166—175, 1938.

[4] *Wilks, S. S.:* Optimum fiducial regions for simultaneous estimation of several population parameters from large samples. (Abstract.) Ann. Math. Stat. *10*, 85—86, 1939.

[5] *Wilks, S. S.* - *J. F. Daly:* An optimum property of confidence regions associated with the likelihood function. Ann. Math. Stat. *10*, 225—235, 1939.

Acknowledgment. The writer is much indebted to Ing. JOSEF MACHEK, of the Faculty of Mathematics and Physics, Charles University, Prague, for much helpful advice and many valuable instructions as to the final arrangement of this paper.

Souhrn

PŘIBLIŽNÁ KONSTRUKCE DVOJROZMĚRNÉ KONFIDENČNÍ OBLASTI

MILOŠ PAVLÍK

Jestliže výběrový soubor hodnot dané náhodné veličiny o známém rozložení s dvojicí ζ parametrů je dosti rozsáhlý, aby shoda pozorovaných a teoretických četností mohla být ověřena Pearsonovým χ^2-testem, a jestliže χ^2 chápeme jako náhodnou funkci dvojice ζ, definovanou na prostoru $\mathfrak{G}$ všech přípustných hodnot této dvojice, potom množina všech ζ, pro která χ^2 se rovná kritické hodnotě na zvolené hladině α statistické významnosti, je uzavřená křivka v $\mathfrak{G}$, jejíž vnitřní oblast je přibližně $100(1 - \alpha)$ % konfidenční oblastí pro skutečnou hodnotu ζ v základním souboru. Tuto křivku lze snadno sestrojit, známe-li hodnoty dané pravděpodobnostní funkce v dostatečně širokém okolí bodového odhadu dvojice ζ. Konstrukce je výhodná zejména v případech, kdy vzhledem ku povaze daného rozložení je použití jiných přibližných konstručních metod obtížné.

Author's address: RNDr. *Miloš Pavlík,* OHS Dolný Kubín, laboratórne oddelenie v Trstenej.