December 16, 1947

Remarks on Estimation

by George W. Rasch

1. In the problem of estimating α in the stochastic difference equation

$$x_t = \alpha x_{t-1} + u_t$$

where x_0 is fixed and u_t normally and independently distributed about 0 with a constant variance σ^2 we get 3 statistics which are together sufficient for the 2 parameters. Similar difference equations of second order contain 3 parameters for which we derive a set of 6 sufficient statistics. And so on.

Thus we are faced with the problem: How to utilize k statistics for the estimation of $l < k$ parameters? The following is an attempt to formulate this problem for $l = 1$ and solve it approximately under the assumption that the variances of the statistics are fairly small as is more often the case for medium-sized samples. It should be noted, however, that the principle may be extended to more parameters.

The problem we are going to solve approximately may be stated in the following terms: Provided the vector $x = (x_1, \dots x_k)$ follows some k-dimensional distribution with one parameter θ:

(1) $$\varphi\{x_1, \dots x_k \mid \theta\}$$

we want an unbiased estimate of θ with a minimum variance.

Denote the mean vector and the variance matrix of x by

(2) $$E\{x\} = \xi(\theta), \text{ resp. } V\{x\} = M(\theta).$$

Then if $f(x)$ is some reasonable function and if φ is some reasonable distribution we have with a tolerable approximation

(3) $$E\{f(x)\} = f(\xi(\theta))$$

and

(4) $$V\{f(x)\} = \left(\frac{\partial f}{\partial \xi_1}, \dots \frac{\partial f}{\partial \xi_k}\right) M(\theta) \begin{pmatrix} \frac{\partial f}{\partial \xi_1} \\ \dots \\ \frac{\partial f}{\partial \xi_k} \end{pmatrix}$$

$$= \sum \mu_{ij}(\theta) \frac{\partial f}{\partial \xi_i} \cdot \frac{\partial f}{\partial \xi_j}.$$

For

(5) $$t = f(x)$$

to be an unbiased estimate of θ we therefore must have

(6) $$f(\xi(\theta)) = \theta$$

which on differentiation gives

(7) $$\sum \frac{\partial f}{\partial \xi_i} \cdot \xi_i'(\theta) = 1.$$

To the degree of approximation we are working with our problem is a minimization of (4) under the condition (7). Putting for a moment

(8) $$\frac{\partial f}{\partial \xi_i} = \lambda_i$$

we differentiate

(9) $$\sum \mu_{ij} \lambda_i \lambda_j - 2\kappa \left(\sum \xi_i' \lambda_i - 1 \right)$$

with respect to the λ_i' and equate to 0:

(10) $$\sum \mu_{ij} \lambda_j = \kappa \cdot \xi_i'$$

or[1]

(11) $$M \lambda^* = \kappa \xi'^*$$

i.e.

(12) $$\lambda^* = \kappa \xi' M^{-1}$$

where κ from

(13) $$\xi' \lambda^* = 1$$

is determined as

(14) $$\kappa = 1 / \xi' M^{-1} \xi'^*.$$

The minimum value of (4) is found to be

(15) $$V\{t\} = \lambda M \lambda^* = \kappa^2 \xi' M^{-1} \xi'^* = \kappa.$$

Now we are faced with the problem of determining such a function $f(\cdot)$ that

(16) $$f(\xi(\theta)) = \theta$$

and

(17) $$f_i'(\theta) = \left(\frac{\partial f}{\partial x_i} \right)_{x = \xi} = \lambda_i(\theta)$$

---

1 -- To avoid confusion with the differentiation sign we shall in this paper use * for transposition.

where

$$\lambda_i(\theta) = \frac{\sum \mu_{ij}^{(-1)}(\theta)\,\xi_j'(\theta)}{\sum\sum \mu_{kj}^{(-1)}(\theta)\,\xi_k'(\theta)\,\xi_j'(\theta)} \tag{18}$$

are known functions.

A solution is readily obtained if we consider the first terms of the Taylor expansion of $f(x)$ about $\xi(z)$:

$$\begin{cases} f(x) = f(\xi(z)) + (x - \xi(z))\left(\frac{f_i(z)}{f_k(z)}\right) + \text{terms of second order} \\ \quad = z + \sum (x_j - \xi_j(z))\lambda_j(z) + \text{" " "} \end{cases} \tag{19}$$

Take in this expansion z as some function of x :

$$z = h(x) \tag{20}$$

and differentiate (19) with respect to x_i. Then we get

$$\frac{\partial f}{\partial x_i} = \frac{\partial h}{\partial x_i} + \lambda_i(z) - \sum \xi_j'(z)\lambda_j(z)\frac{\partial z}{\partial x_i} + \text{terms of first order} \tag{21}$$

which according to (13) reduces to

$$\frac{\partial f}{\partial x_i} = \lambda_i(z) + \text{terms of first order.} \tag{22}$$

If we now choose $h(x)$ such that

$$h(\xi(z)) = z \tag{23}$$

and put

$$f(x) = x + \sum (x_j - \xi_j(z))\lambda_j(z), \quad z = h(x), \tag{24}$$

then we have a solution of our problem.

One way of choosing $h(x)$ is to invert the functions $\xi_1(\theta), \ldots \xi_k(\theta)$:

$$\eta_i(\xi_i(\theta)) = \theta \tag{25}$$

and take a linear combination of $\eta_i(x_i)$:

$$z = \sum \alpha_i \eta_i(x_i), \quad \sum \alpha_i = 1. \tag{26}$$

More generally we may take any function of x :

$$y = g(x), \tag{27}$$

find out which function

$$g(\xi(\theta)) = \psi(\theta) \tag{28}$$

is of θ, invert that function

(29) $$\varphi(\psi(\theta)) = \theta$$

and take

(30) $$h(x) = \varphi(g(x)).$$

With such a multitude of solutions as given by (30) and (24) -- all of which are fairly easily obtainable -- it seems natural to ask: What can be gained by a sensible choice of h? Could the bias still present, due to our approximation, be reduced? Or could we secure a very nearly normal distribution of t? These questions still have to be investigated.

2. It is generally realized that maximum likelihood estimates are often hard to handle in practice. In the following is suggested the application of a class of estimates which - after a once for all tabulation of some functions - seem to be really computable and nevertheless possess a high degree of efficiency.

For simplicity we shall in this note only consider the case of n independent observations $x_1, \ldots x_n$ and one parameter θ:

$$(1) \qquad g_n\{x_1, \quad , x_n \mid \theta\} = \prod_{\nu=1}^{n} g_1\{x_\nu \mid \theta\}$$

We consider the computation of the mean of some tabulated function - $\varphi(x)$ - of the observation as fairly easy:

$$(2) \qquad a = \frac{1}{n} \sum_{\nu=1}^{n} \varphi(x_\nu)$$

If φ is a sensible function and the distribution of $y = \psi(x)$ not too bad a is fairly nearly normally distributed about the mean

$$(3) \qquad \alpha = E\{\varphi(x)\} = \int \varphi(x)\, g_1\{x \mid \theta\}\, dx = \lambda(\theta), \text{ say}$$

and with the variance

$$(4) \qquad V\{a\} = \frac{1}{n} V\{\varphi(x)\}$$

This will often hold for medium sized and large samples.

Now denote by $k(\alpha)$ the inverse function to $\lambda(\theta)$:

$$(5) \qquad \lambda(K(\alpha)) = \alpha$$

and take

$$(6) \qquad t = K(a)$$

as as estimate of θ which in any particular case has particular value θ_0.

If $V(a)$ is not too large - how large it may be depends on the function we are going to choose and has got to be checked in each particular case - we can safely apply the approximations

$$(7) \qquad E\{t\} \simeq K(\alpha) = \theta_0$$

and

$$(8) \qquad V\{t\} \simeq V\{a\} K'^2(\alpha) = \frac{1}{n} \frac{V\{\varphi(x)\}}{\lambda'^2(\theta_0)} = \frac{1}{n} \mu(\theta_0)$$

The question we are going to consider is this: How to choose $\varphi(x)$ in order to get "the best computable estimate" of θ_0? Under certain safeguards t may be assumed to be approximately normally distributed about θ_0 with the variance (8) and our question therefore boils down to a minimization of $V\{t\}$.

Now, let $\varphi(x)$ be the minimizing function and consider the variation of (8) if $\varphi(x)$ is replaced by some neighbour function

$$(9)\qquad \varphi(x) + \varepsilon\,\psi(x)$$

It is found that

$$\frac{1}{n}\,\delta\mu(\theta_0) = \frac{\int_{-\infty}^{\infty}\varphi(x)\psi(x)p\{x|\theta_0\}\,dx - \int_{-\infty}^{\infty}\varphi(x)p(x|\theta_0)\,dx\int_{-\infty}^{\infty}\psi(x)p(x|\theta)\,dx}{\left(\int_{-\infty}^{\infty}\varphi(x)\frac{\partial p\{x|\theta_0\}}{\partial\theta_0}\,dx\right)^2}$$

$$(10)\qquad -\frac{\int_{-\infty}^{\infty}\varphi^2(x)p\{x|\theta_0\}\,dx - \left(\int_{-\infty}^{\infty}\varphi(x)p\{x|\theta_0\}\,dx\right)^2}{\left(\int_{-\infty}^{\infty}\varphi(x)\frac{\partial p\{x|\theta_0\}}{\partial\theta_0}\,dx\right)^3}\int_{-\infty}^{\infty}\psi(x)\frac{\partial p\{x|\theta_0\}}{\partial\theta_0}\,dx$$

On equating to 0 and choosing for $\psi(x)$ the particular function

$$(11)\qquad \psi(x) = \begin{cases} 0, & x < z \\ 1, & x > z \end{cases}$$

where z is any real number we find that $\varphi(x)$ must satisfy the equation

$$(12)\qquad \int_z^{\infty}\varphi(x)p\{x|\theta_0\}\,dx - c_1\int_z^{\infty}p\{x|\theta_0\}\,dx = c_2\int_z^{\infty}\frac{\partial p\{x|\theta_0\}}{\partial\theta_0}\,dx$$

where c_1 and c_2 are constants:

$$(13)\qquad c_1 = \int_{-\infty}^{\infty}\varphi(x)p\{x|\theta_0\}\,dx$$

$$c_2 = \frac{\int_{-\infty}^{\infty}\varphi^2(x)p\{x|\theta_0\}\,dx - \left(\int_{-\infty}^{\infty}\varphi(x)p\{x|\theta_0\}\,dx\right)^2}{\int_{-\infty}^{\infty}\varphi(x)\frac{\partial p\{x|\theta_0\}}{\partial\theta_0}\,dx}$$

when differentiating and replacing z by x we get

$$(14)\qquad \varphi(x) = c_1 + c_2\,\frac{\partial \log p\{x|\theta_0\}}{\partial\theta_0}$$

for which function actually

$$(15)\quad E\{\varphi(x)\} = c_1 + c_2 \int_{-\infty}^{\infty} \frac{\partial \log p\{x|\theta_0\}}{\partial \theta_0} p\{x|\theta_0\}\,dx = c_1$$

$$(16)\quad V\{\varphi(x)\} = c_2^2 \int_{-\infty}^{\infty} \left(\frac{\partial \log p\{x|\theta_0\}}{\partial \theta_0}\right)^2 p\{x|\theta_0\}\,dx = c_2^2 \int_{-\infty}^{\infty} \left(\frac{\partial p\{x|\theta_0\}}{\partial \theta_0}\right)^2 \frac{dx}{p\{x|\theta_0\}}$$

$$(17)\quad \int_{-\infty}^{\infty} \varphi(x) \frac{\partial p\{x|\theta_0\}}{\partial \theta_0}\,dx = c_2 \int_{-\infty}^{\infty} \frac{\partial \log p\{x|\theta_0\}}{\partial \theta_0} \frac{\partial p\{x|\theta_0\}}{\partial \theta_0}\,dx$$

and consequently

$$= c_2 \int_{-\infty}^{\infty} \left(\frac{\partial p\{x|\theta_0\}}{\partial \theta_0}\right)^2 \frac{dx}{p\{x|\theta_0\}}$$

$$(18)\quad V\{t\} = \frac{1}{n} \frac{1}{\int_{-\infty}^{\infty} \left(\frac{\partial p\{x|\theta_0\}}{\partial \theta_0}\right)^2 \frac{dx}{p\{x|\theta_0\}}}$$

which is the maximum likelihood variance.

The conclusion of this development seems rather sad. On the assumption that a "best computable estimate" of θ_0 exists we find that such an estimate depends on θ_0 itself - and in such a way that part of the argument, in the form presented, breaks down since under this condition

$$(19)\quad \int_{-\infty}^{\infty} \varphi(x) \frac{\partial p\{x|\theta_0\}}{\partial \theta_0}\,dx$$

cannot be taken as $\lambda'(\theta_0)$ -- but that if it did exist, then it would be just as good as the maximum likelihood estimate itself.

In practice these difficulties can however be pretty nearly avoided: Select a "suitable" --to be specified below--sequence of point on the θ-axis: $\theta_0, \theta_{\pm 1}, \theta_{\pm 2}, \ldots\ldots\ldots$ and take them as say midpoints of/to some extent overlapping intervals: $(\theta_i', \theta_{i+1}')$ and tabulate the functions

$$(20)\quad \varphi_i(x) = \frac{\partial \log p\{x|\theta_i\}}{\partial \theta_i}$$

Then

$$(21)\quad a_i = \frac{1}{n} \sum \varphi_i(x_\nu)$$

has the expectation

$$(22)\quad \alpha_i = \int \frac{\partial \log p_x\{x|\theta_i\}}{\partial \theta_i} p_x\{x|\theta\}\,dx = \lambda(\theta_i, \theta), \text{ say}$$

which vanishes for $\theta = \theta_i$:

$$(23)\quad \lambda(\theta_i, \theta_i) = 0$$

Invert the function $\lambda(\theta_i, \theta)$:

$$(24)\quad \theta = K(\alpha_i, \theta_i)$$

and take

$$(25)\quad t_i = K(a_i, \theta_i)$$

as an estimate of θ :

$$(26)\quad E\{t_i\} \simeq K(\alpha_i, \theta_i) = \theta$$

The variance of t_i becomes

$$(27)\quad V\{t_i\} = \frac{V\{a_i\}}{\left(\frac{\partial \lambda(\theta_i, \theta)}{\partial \theta}\right)^2} = \frac{1}{n}\mu(\theta, \theta_i)$$

where

$$(28)\quad \mu(\theta, \theta_i) = \frac{\int_{-\infty}^{\infty}\left(\frac{\partial \log p_x\{x|\theta_i\}}{\partial \theta_i}\right)^2 p_x\{x|\theta\}\,dx}{\left(\int_{-\infty}^{\infty}\frac{\partial \log p_x\{x|\theta_i\}}{\partial \theta_i}\,\frac{\partial \log p_x\{x|\theta\}}{\partial \theta}\,dx\right)^2}$$

For $\theta = \theta_i$ this reduces to the maximum likelihood estimate and in practice the intervals $(\theta_i', \theta_{i+1}')$ should be such chosen that $\mu(\theta, \theta_i)$ does not deviate too much from $\mu(\theta_i, \theta_i)$ such that the function φ_i covers an interval in which the efficiency of a_i is not smaller than wanted. For the method to be really practicable it should be possible to choose such intervals farily large.

The method leads inevitably to more than one estimate of θ in several situations, but as each estimate gives rise to a confidence interval this should present no real difficulty in practice.

It may be mentioned that the method has actually been used in a case which otherwise would seem rather difficult to handle (Hald, Jersild & Rasch, "On the

Determination of the Phagocytic Power of Leucocytes," Acta path. et microbiol. 20, 1942.)

One point more may be worth mentioning: The characteristic function of is readily available. When von Neuman's machine starts working, it may therefore be possible to solve the sampling problems even for small samples. Which of course does not imply that such estimates are "best" in any sense for small