

台が有界な分布に対する分布の台の幅の逐次区間 推定

筑波大・数理物質系 小池 健一 (Ken-ichi Koike)

(Faculty of Pure and Applied Sciences, University of Tsukuba)

筑波大・数理物質科学研究科 平山 令士 (Reishi Hirayama)

(Graduate School of Pure and Applied Sciences, University of Tsukuba)

1 はじめに

一般に統計的推測においては、抽出する標本数を事前に決めてから、その標本に基づいて母数を推定する。しかし実際の統計調査では標本の抽出(観測)自体にもコストがかかり、妥当な標本数(観測回数)を決定することも問題になってくる。そこで、事前に標本数を固定する代わりに適当な停止則(stopping rule)を定めておき、それに基づいて逐次的に標本を抽出していき、各観測後に標本抽出を停止するか、または標本抽出を継続するかを決定するのが望ましい。

この問題に関して、Chow and Robbins(1965)は、平均と分散が未知の分布族における、平均に対する固定幅の逐次信頼区間を構成した。この逐次信頼区間は、区間の幅を0に近づけたときに漸近一致性をもち、分散が既知のときの最小標本数と逐次推定に必要な最適標本数との比が1に収束する。この推定方法では母集団の分布型を仮定しなくてもよいので、その2次の漸近展開を求めるなど、多くの研究がなされている(Woodroffe(1977)など)。特に区間 $(0, \theta)$ 上の一様分布については比較的取り扱いやすく、Graybill and Connell(1964), Akahira(1993), Bose(2001), Mukhopadhyay and Cicconttei(2002)など、多くの研究がある。

また一方では、Akahira and Koike(2005)が、位置尺度母数をもつ一様分布族における平均に対する固定幅の逐次信頼区間を構成した。さらにKoike(2007)は、有界な台(support)上の位置尺度分布族における位置母数に対する固定幅の逐次信頼区間を構成した。その結果、分布の台の両端点で密度関数が急激に0に近づくような場合には、この新しい推定方法の方がChow and Robbins(1965)の方法よりも、標本数の決定において漸的に優れていることが確認された。すなわち、分布の台の端点に関する情報を勘案して推定することで、抽出する標本数の節約をすることができるので、実際の統計調査においてはコストダウンを図ることができる。

本論文では、以上の先行研究を踏まえて、有界な台をもつ分布族の確率密度関数の台の幅に対する逐次信頼区間を構成する。台の幅を推定するための統計量としては標本範囲を用いて、信頼区間の幅を0に近づけたとき、この逐次信頼区間が漸近一致性、漸近有効性などの優れた性質をもつことを示す。このことは、分布に関する情報をうまくとり入れることでより効率の良い推定が可能になることを意味している。また、この逐次信頼区間の被覆確率を具体的なモデルに当てはめてコンピュータで数値シミュレーションを行い、この逐次推定方式の性能を検証する。

2 分布の台の幅の逐次区間推定

本章では、有界な台をもつ分布族の確率密度関数の台の幅に対する逐次信頼区間を構成する。

2.1 標本範囲の漸近分布

本節では、Koike(2007)に沿って、有界な台上の確率分布に対する極値の漸近分布を求める。

確率変数列 $X_1, X_2, \dots$ は、互いに独立にいずれも確率密度関数 $f_X(x)$ をもつ確率分布に従うとする。ここで $f_X(x)$ に対して、以下の条件 (A1)~(A3) を仮定する。

(A1) $f_X(x)$ は有界な台 (a, b) をもつ。すなわち、

$$f_X(x) \begin{cases} > 0 & (a < x < b), \\ = 0 & (\text{他}). \end{cases}$$

(A2) $f_X(x)$ は区間 (a, b) で連続微分可能で、以下を満たすような定数 $c_1, c_2 > 0$ が存在する；

$$\lim_{x \rightarrow a+0} f_X(x) = c_1, \quad \lim_{x \rightarrow b-0} f_X(x) = c_2.$$

(A3) $\gamma > -1$ に対して、 $f_X(x)$ が以下を満たすような定数 $g_1, g_2 > 0$ が存在する；

$$\begin{aligned} f_X(x) &\approx g_1(x-a)^\gamma & (x \rightarrow a+0), \\ f_X(x) &\approx g_2(b-x)^\gamma & (x \rightarrow b-0). \end{aligned}$$

ただし、“ $\approx$ ”は x についての左辺と右辺の比が1に収束することを意味する。また (A3) は (A2) を一般化したものであり、 $\gamma = 0$ のとき、(A2) と (A3) は同等である。ここで以下

のように記号を定義すると、次の補題がわかる；

$$\begin{aligned} X_{(1)} &:= \min_{1 \leq i \leq n} X_i, & X_{(n)} &:= \max_{1 \leq i \leq n} X_i, \\ S' &:= \frac{n\{(b+a) - (X_{(n)} + X_{(1)})\}}{2}, \\ T' &:= \frac{n\{(b-a) - (X_{(n)} - X_{(1)})\}}{2}. \end{aligned}$$

補題 2.1 (Koike(2007)). $f_X(x)$ に条件 (A1), (A2) を仮定する. (S', T') の同時密度関数 $f_{S', T'}^{(n)}(s, t)$ は $n \rightarrow \infty$ とすると、以下が成り立つ；

$$f_{S', T'}^{(n)}(s, t) \rightarrow \begin{cases} 2c_1 c_2 \exp\{(c_1 - c_2)s - (c_1 + c_2)t\} & (t > |s|), \\ 0 & (\text{他}). \end{cases} \quad (1)$$

補題 2.1 より、 T' の漸近密度関数は

$$\begin{aligned} f_{T'}(t) &= \int_{-\infty}^{\infty} f_{S', T'}(s, t) ds \\ &= 2c_1 c_2 \exp\{-(c_1 + c_2)t\} \int_{-t}^t \exp\{(c_1 - c_2)s\} ds \\ &= \begin{cases} 4c_1^2 t e^{-2c_1 t} & (t > 0, c_1 = c_2), \\ \frac{2c_1 c_2}{c_2 - c_1} (e^{-2c_1 t} - e^{-2c_2 t}) & (t > 0, c_1 \neq c_2), \\ 0 & (t \leq 0) \end{cases} \end{aligned} \quad (2)$$

となる.

また、 $f_X(x)$ に対して、条件 (A1) と (A2) の代わりに (A1) と (A3) を仮定し、

$$U' := n^{\frac{1}{\gamma+1}}(Y_{(1)} - a), \quad V' := n^{\frac{1}{\gamma+1}}(Y_{(n)} - b)$$

とおけば、 (U', V') の漸近同時分布は次の補題のようになる.

補題 2.2 (Koike(2007)). $f_X(x)$ に条件 (A1), (A3) を仮定する. (U', V') の同時密度関数 $f_{U', V'}^{(n)}(u, v)$ は、 $n \rightarrow \infty$ とすると以下が成り立つ；

$$f_{U', V'}^{(n)}(u, v) \rightarrow \begin{cases} g_1 g_2 (-uv)^\gamma \exp\left\{-\frac{g_2(-v)^{\gamma+1}}{\gamma+1} - \frac{g_1 u^{\gamma+1}}{\gamma+1}\right\} & (v < 0 < u), \\ 0 & (\text{他}). \end{cases} \quad (3)$$

これ以降、標本範囲 (range) を

$$R_n := X_{(n)} - X_{(1)}$$

と表すことにする.

2.2 信頼区間の構成

～台の端点で密度関数が正定数に収束する場合～

この節では、台の端点で密度関数が正定数に収束する場合における、分布の台の幅の信頼区間を構成する。 $f_X(x)$ に (A1), (A2) を仮定する。ただし式 (2) より, $c_1 = c_2$ の場合と $c_1 \neq c_2$ の場合に分けて考える必要がある。

2.2.1 台の両端点で密度関数の値が等しい場合

$f_X(x)$ に (A1), (A2) を仮定し, $c_1 = c_2$ とする。ただし, $c_1 = c_1(b-a)$ と表すことができ, $c_1(x)$ は $x = b-a$ で右連続で, $0 < x \leq b-a$ に対して $c_1(x) \geq c_1(b-a)$ となる関数であると仮定する。このことは台の幅が大きくなると端点での確率密度関数の値が小さくなることを意味し, 不自然な仮定ではない。式 (2) より, T' の漸近確率密度関数は

$$f_{T'}(t) = 4c_1^2 te^{-2c_1 t} \quad (t > 0)$$

である。標本範囲 R_n によって台の幅 $b-a$ を推定することを考える。任意の $n \geq 2$ に対して $0 \leq R_n \leq b-a$ だから,

$$\begin{aligned} P(|b-a-R_n| \leq d) &= P(b-a-R_n \leq d) \\ &= P\left(T' \leq \frac{dn}{2}\right) \\ &\approx \int_0^{\frac{dn}{2}} f_{T'}(t) dt \\ &= 4c_1^2 \int_0^{\frac{dn}{2}} te^{-2c_1 t} dt \\ &= 1 - (c_1 dn + 1)e^{-c_1 dn} =: \kappa_1(n) \quad (\text{say}) \end{aligned}$$

となる。ここで, $0 < \alpha < 1$ に対して $\alpha = (l+1)e^{-l}$ を満たす l を l_0 とし^{*1}, $n_1 := l_0/c_1(b-a)d$ とおく。 $\kappa_1(n)$ は n について単調増加だから, $n \geq n_1$ に対して,

$$\kappa_1(n) \geq \kappa_1(n_1) = 1 - \alpha.$$

したがって, $n \geq n_1$ となる最小の n を停止則として定めたいが, n_1 が $c_1(b-a)$ の関数なのでこれを用いることができない。そこで, $b-a$ を R_n で置き換えることによって, 以下のように停止則を定める;

$$\tau_1 := \inf \left\{ n \geq n_0 \mid \frac{1}{c_1(R_n)} \leq \frac{dn}{l_0} \right\}. \quad (4)$$

^{*1} このような l_0 は明示的に求められないが, 一意的である。また, $\alpha = 0.05, 0.01$ に対する近似値は, それぞれ 4.7438..., 6.6383... となる。

ここで $n_0 \geq 2$ は初期標本数である。

定理 2.3. 上記の停止則 τ_1 に対して以下が成り立つ；

- (i) $\frac{\tau_1}{n_1} \xrightarrow{a.s.} 1 \quad (d \rightarrow 0+),$
- (ii) $\lim_{d \rightarrow 0+} P(|b - a - R_{\tau_1}| \leq d) = 1 - \alpha \quad (\text{漸近一貫性}),$
- (iii) $\frac{E(\tau_1)}{n_1} \rightarrow 1 \quad (d \rightarrow 0+) \quad (\text{漸近有効性}).$

例 2.4. 確率変数列 $X_1, X_2, \dots$ がそれぞれ独立に一様分布 $U(a, b)$ に従うとする。確率密度関数は

$$f_0(x) = \begin{cases} \frac{1}{b-a} & (a \leq x \leq b), \\ 0 & (\text{他}) \end{cases}$$

であり、このとき $c_1(b-a) = 1/(b-a)$ である。この $c_1(b-a)$ は、 $x = b-a$ で右連続かつ、 $0 < x \leq b-a$ に対して $c_1(x) \geq c_1(b-a)$ であるという $c_1(x)$ の条件を満たしている。このとき、(3) の τ_1 は

$$\tau_1 := \inf \left\{ n \geq n_0 \mid R_n \leq \frac{dn}{l_0} \right\}.$$

となる。

2.2.2 台の両端点で密度関数の値が異なる場合

$f_X(x)$ に条件 (A1), (A2) を仮定し、 $c_1 \neq c_2$ とする。式 (2) より、 T' の漸近密度関数は、

$$f_{T'}(t) = \frac{2c_1c_2}{c_2 - c_1} (e^{-2c_1t} - e^{-2c_2t}) \quad (t > 0)$$

である。ここでも R_n によって $b-a$ を推定することを考える。任意の $n \geq 2$ に対して $0 \leq R_n \leq b-a$ だから、

$$\begin{aligned} P(|b-a-R_n| \leq d) &= P(b-a-R_n \leq d) \\ &= P\left(T' \leq \frac{dn}{2}\right) \\ &\approx \int_0^{\frac{dn}{2}} f_{T'}(t) dt \\ &= \frac{2c_1c_2}{c_2 - c_1} \int_0^{\frac{dn}{2}} (e^{-2c_1t} - e^{-2c_2t}) dt \\ &= 1 - \frac{c_1c_2}{c_2 - c_1} \left(\frac{e^{-c_1dn}}{c_1} - \frac{e^{-c_2dn}}{c_2} \right) =: \kappa_2(n) \end{aligned}$$

となる. ここで, $0 < \alpha < 1$ に対して

$$\alpha = \frac{c_1 c_1}{c_2 - c_1} \left(\frac{e^{-c_1 l}}{c_1} - \frac{e^{-c_2 l}}{c_2} \right)$$

を満たす l を l_0 とする*2. ここで $l_0 = l_0(b-a)$ と表すことができ, さらに $l_0(x)$ は $x = b-a$ で右連続かつ, $0 < x \leq b-a$ に対して $l_0(x) \leq l_0(b-a)$ となる関数であると仮定する. ここで $n_2 := l_0(b-a)/d$ とおくと, $\kappa_2(n)$ は n について単調増加だから, $n \geq n_2$ に対して,

$$\kappa_2(n) \geq \kappa_2(n_2) = 1 - \alpha.$$

したがって $n \geq n_2$ となる最小の n を停止則として定めたいが, n_2 が $l_0(b-a)$ の関数なのでこれを用いることができない. そこで, $b-a$ を R_n で置き換えることによって以下のよう停止則を定める;

$$\tau_2 := \inf\{n \geq n_0 \mid l_0(R_n) \leq dn\}. \quad (5)$$

ただし, ここで $n_0 \geq 2$ は初期標本数である.

定理 2.5. 上記の停止則 τ_2 に対して以下が成り立つ;

- (i) $\frac{\tau_2}{n_2} \xrightarrow{a.s.} 1 \quad (d \rightarrow 0+),$
- (ii) $\lim_{d \rightarrow 0+} P(|b-a-R_{\tau_2}| \leq d) = 1 - \alpha \quad (\text{漸近一貫性}),$
- (iii) $\frac{E(\tau_2)}{n_2} \rightarrow 1 \quad (d \rightarrow 0+) \quad (\text{漸近有効性}).$

例 2.6. 確率密度関数が以下の図 1 のような台形の形をした分布 ($c_2 = 2c_1$) を考える. 確率密度関数 $f_X(x)$ は以下のようになる;

$$f_X(x) = \begin{cases} \frac{2}{3(b-a)^2}x + \frac{2(b-2a)}{3(b-a)^2} & (a \leq x \leq b), \\ 0 & (\text{他}). \end{cases} \quad (6)$$

確率変数列 $X_1, X_2, \dots$ がこの分布に従うとすると, $T' = n(b-a-R_n)/2$ の漸近密度関数は

$$f_{T'}(t) = \begin{cases} 4c_1(e^{-2c_1 t} - e^{-4c_1 t}) & (t > 0), \\ 0 & (\text{他}) \end{cases}$$

となる. $\alpha = 2e^{-cl} - e^{-2cl}$ を満たす l を $l_0(b-a)$ とすれば,

$$l_0(b-a) = -\frac{3}{2}(b-a) \log(1 - \sqrt{1 - \alpha})$$

*2 このような l_0 は明示的に求められないが, 一意的である. また, α, c_1, c_2 が具体的に与えられれば, l_0 の近似値を数値的に求めることが可能である.

となり, これは $x = b - a$ で右連続かつ, $0 < x \leq b - a$ に対して $l_0(x) \leq l_0(b - a)$ であるという $l_0(x)$ の条件を満たしている. このとき, 停止則は

$$\tau_2 = \inf \left\{ n \geq n_0 \mid -\frac{3}{2} R_n \log(1 - \sqrt{1 - \alpha}) \leq dn \right\} \quad (7)$$

となる.

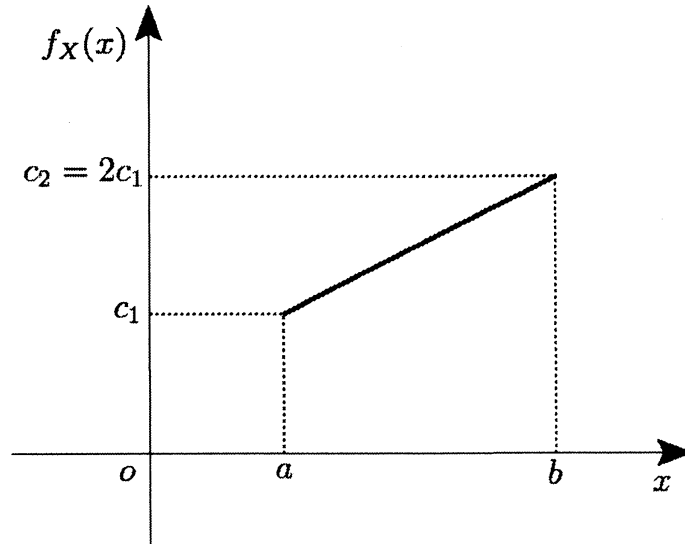


図 1. 例 2.6 における確率密度関数

2.3 信頼区間の構成

～台の端点で密度関数が 0 に収束する場合～

ここからは $f_X(x)$ に条件 (A1), (A3) を仮定して考える.

$$U' := n^{\frac{1}{\gamma+1}}(X_{(1)} - a), \quad V' := n^{\frac{1}{\gamma+1}}(X_{(n)} - b)$$

とおくと, 補題 2.2 より, (U', V') の漸近分布は

$$f_{U', V'}(u, v) = g_1 g_2 (-uv)^\gamma \exp \left\{ -\frac{g_1}{\gamma+1} u^{\gamma+1} - \frac{g_2}{\gamma+1} (-v)^{\gamma+1} \right\} \quad (v < 0 < u)$$

となる. ここで U', V' は漸近的に独立な確率変数であり, その漸近周辺密度関数は, それぞれ

$$\begin{aligned} f_{U'}(u) &= g_1 u^\gamma \exp \left\{ -\frac{g_1}{\gamma+1} u^{\gamma+1} \right\} & (u > 0), \\ f_{V'}(v) &= g_2 (-v)^\gamma \exp \left\{ -\frac{g_2}{\gamma+1} (-v)^{\gamma+1} \right\} & (v < 0) \end{aligned}$$

である. U', V' の漸近分布をそれぞれ Φ, Ψ とおくと,

$$\begin{pmatrix} U' \\ V' \end{pmatrix} = n^{\frac{1}{\gamma+1}} \left\{ \begin{pmatrix} X_{(1)} \\ X_{(n)} \end{pmatrix} - \begin{pmatrix} a \\ b \end{pmatrix} \right\} \xrightarrow{\mathcal{L}} \begin{pmatrix} \Phi \\ \Psi \end{pmatrix}$$

と表すことができるので、関数 $h(x, y) = x - y$ に対して Cramér の定理を用いると、

$$n^{\frac{1}{\gamma+1}} \left\{ h \left(\frac{X_{(1)}}{X_{(n)}} \right) - h \left(\frac{a}{b} \right) \right\} \xrightarrow{\mathcal{L}} \dot{h} \left(\frac{a}{b} \right) \begin{pmatrix} \Phi \\ \Psi \end{pmatrix} = \Phi - \Psi =: \Lambda.$$

ここで、

$$\int_0^l f_\Lambda(\lambda) d\lambda = 1 - \alpha$$

の l についての解を l_0 と表す。ここで、 $l_0 = l_0(b-a)$ と表すことができ、 $l_0(x)$ は $x = b-a$ で右連続かつ、 $0 < x \leq b-a$ に対して $l_0(x) \leq l_0(b-a)$ となる関数であると仮定する。任意の $n \geq 2$ に対して $0 \leq R_n \leq b-a$ だから、

$$\begin{aligned} P(|b-a-R_n| \leq d) &= P(b-a-R_n \leq d) \\ &= P\left(\Lambda \leq dn^{\frac{1}{\gamma+1}}\right) \\ &\approx \int_0^{dn^{\frac{1}{\gamma+1}}} f_\Lambda(\lambda) d\lambda. \end{aligned}$$

ここで、 $n_3 := \{l_0(b-a)/d\}^{\gamma+1}$ とおいて停止則を定めたいが、この n_3 は $b-a$ の関数なので用いることができない。したがって、 $b-a$ を R_n で置き換えることによって、以下のように停止則を定める；

$$\tau_3 := \inf \left\{ n \geq n_0 \mid l_0(R_n) \leq dn^{\frac{1}{\gamma+1}} \right\}. \quad (8)$$

ここで $n_0 \geq 2$ は初期標本数である。

定理 2.7. 上記の停止則 τ_3 に対して以下が成り立つ；

- (i) $\frac{\tau_3}{n_3} \xrightarrow{a.s.} 1 \quad (d \rightarrow 0+),$
- (ii) $\lim_{d \rightarrow 0+} P(|b-a-R_{\tau_3}| \leq d) = 1 - \alpha \quad (\text{漸近一貫性}),$
- (iii) $\frac{E(\tau_3)}{n_3} \rightarrow 1 \quad (d \rightarrow 0+) \quad (\text{漸近有効性}).$

3 数値実験

この章では、本論で得られた逐次推定方式の性能について、例 2.6 の確率分布を使って具体的な数値で検証する。式 (6) において $a = 0$ としても一般性を失わないので、確率密度関数を

$$f_X(x) = \begin{cases} \frac{2}{3b^2}x + \frac{2}{3b} & (0 \leq x \leq b), \\ 0 & (\text{他}) \end{cases}$$

とする. このとき, $R_n \xrightarrow{a.s.} b (d \rightarrow 0+)$ となり, 停止則は式 (7) と同じ

$$\tau_2 = \inf \left\{ n \geq n_0 \mid -\frac{3}{2} R_n \log(1 - \sqrt{1 - \alpha}) \leq dn \right\}$$

である. この分布から擬似乱数を 10000 回発生させた上で

$$b = 1(1)5, \quad d = 0.01(0.01)0.05, \quad \alpha = 0.10$$

として, この逐次推定方式 $(\tau_2, [R_{\tau_2}, R_{\tau_2} + d])$ の被覆確率, 平均標本数, 標本数の標準偏差を, コンピュータで数値計算した. その結果が下記の表 1, 表 2, 表 3 である. これより, この逐次推定方式の被覆確率が信頼係数 0.90 にほぼ等しくなることが確認できる.

表 1. 逐次推定方式 $(\tau_2, [R_{\tau_2}, R_{\tau_2} + d])$ の被覆確率

$b \setminus d$	0.01	0.02	0.03	0.04	0.05
1	0.9206	0.9218	0.9147	0.9237	0.9183
2	0.9061	0.9142	0.9178	0.9242	0.9213
3	0.8999	0.9050	0.9082	0.9194	0.9203
4	0.9022	0.9122	0.9086	0.9180	0.9080
5	0.8975	0.9101	0.9085	0.9079	0.9148

表 2. 逐次推定方式 $(\tau_2, [R_{\tau_2}, R_{\tau_2} + d])$ による平均標本数

$b \setminus d$	0.01	0.02	0.03	0.04	0.05
1	45.4752	23.3445	15.1497	11.6608	8.5145
2	90.4238	45.4998	30.4826	23.3901	18.2748
3	134.35	67.5023	45.4834	34.496	27.4434
4	179.125	90.3894	60.5016	45.5736	36.4304
5	222.949	112.406	75.361	56.4904	45.5039

表 3. 逐次推定方式 $(\tau_2, [R_{\tau_2}, R_{\tau_2} + d])$ による標本数の標準偏差

$b \setminus d$	0.01	0.02	0.03	0.04	0.05
1	4.1596	3.38929	2.9604	2.93225	2.2652
2	6.75096	4.46452	3.6223	3.23083	3.02523
3	10.4473	5.81749	4.13977	3.94497	3.46983
4	13.0597	5.75635	5.43249	4.64447	4.06446
5	15.752	8.20156	6.28596	4.74606	4.40903

参考文献

- Akahira, M. (1993). Two-stage sequential estimation procedures for the uniform distribution. 京都大学数理解析研究所講究録 842: 151-156.
- Akahira, M. and Koike, K. (2005). Sequential interval estimation of a location parameter with the fixed width in the uniform distribution with an unknown scale parameter. *Sequential Analysis* 24: 63-75.
- Anscombe, F. J. (1952). Large sample theory of sequential estimation. *Proc. Camb. Phil. Soc.* 48: 600-607.
- Bose, A. (2001). A boundary crossing problem with application to sequential estimation. *Sequential Analysts* 20: 65-76.
- Chow, Y. S. and Robbins, H. (1965). On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *Ann. Math. Statist.* 36: 457-462.
- Graybill, F. A. and Connel, T. L. (1964). Sample size required to estimate the parameter in the uniform density within d units of the true value. *J. Amer. Statist. Assoc.* 59: 550-556.
- Koike, K. (2007). Sequential interval estimation of a location parameter with the fixed width in the non-regular case. *Sequential Analysis* 26: 63-70.
- Mukhopadhyay, N. and Cicconetti, G. (2002). Second-order properties of a two-stage point estimation procedure for the range in a power family distribution. *Calcutta Statist. Assoc. Bull.* 52: 219-234.
- Woodroffe, M. (1977). Second order approximations for sequential point and interval estimation. *Ann. Statist.* 5: 984-995.