The Estimate of variance and the Precision of Sampling

C. Hayashi
M.D. Ichida

In this paper it is considered to decide the sample size estimatig the variance step by step of a sampling process. First drawing n random samples, we estimate the variance and calculate the sampling variance using it. Then if this variance is over a criterion, we draw more samples and repeat the some process. As the variance is estimated, the sampling variance is not that of population. This is in question. But if we can properly make t the criterion, concerning with the estimated sampling variance the last result will be obtained under the desired level of confidence. Thus we continue the sampling following the criterion untill we secure the desired precision from the point of view of the theory of probability.

In order to secure the desired sampling precision in the sense of level of confidence, the criterion must be decided.

The descision of this criterion will be described as follows and the example for the population normally distributed is given.

④

分散の推定とサムプリングの精度

林　知己夫
石田　正次

母集団平均推定のためにサムプリング調査を行ふ場合分散の判明してゐないのが通例である。　此の様な時あらかじめ準備調査を行つて分散を推定してゆくやり方があるが，準備調査としてどの位のサムプル数をとつて分散を推定し，サムプル数を決定してゆくならば，本当の調査の時欲する丈の精度をどの位の信頼度を以て確保することが出来るであらうか。

此の問題を考へる場合当然推定分散の精度を考慮に入れなければならない。

こゝではあらかじめ準備調査を行ふと云ふ方法ではなく逐次サムプルをとりながら，そのサムプルから母集団分散を推定し，此の推定分散を用いて推定平均（サムプル平均）の信頼巾（信頼度一定）を求め此についてサムプル数は十分か否か，少ければさらにサムプルを増加さす等の事を行ふ方法を考へてみよう。

此がある値より小になるならば欲する丈の精度（母集団分散を用ひた時の信頼巾）をある信頼度の下に得てゐ筈である。　その値より大ならば欲する精度が得られてゐると言へぬからさらに又サムプルをとり，ここで同様の手続きをくりかへしてゆく様なことになる。こゝで又云々…以上の様なサムプリングの方法をとるときある値は如何にきめられるべきであらうか。　此の問題を少しく考へてゆく事にしよう。

実際問題でこの様な場合はよくであふものである。　一例をあげよう。　手許に莫大な資料がある。　ここからいくつかのサムプルをとり此をしらべある標識についての総平均を一定の精度で推定したいと言ふ様な場合（その事についての分散不明）がさうである。

この様なとき莫大な資料から母集団を構成し，まづ少しのサムプルをとり此より分散を推定し，サムプル平均の精度を計算する。不十分なら又サムプルをとり又精度を計算し又不十分なら－－－－－－－－。

本論にうつらう。

調査対象の標識を x とする。　此の各に等しい抽出確率をあたへて母集団を構成する。　此の大さは N とする。　しかし N は十分大であり

$\dfrac{N-1}{N-n} \fallingdotseq 1$　の程度であるとする。

n はサムプル数である。　母集団平均を $\overline{X}$，分散を σ^2

$\dfrac{\mu_4}{\sigma^4}$ を β_2 とする。　ここに $\mu_4 = \sum_{i=1}^{N} (X_i - \overline{X})^4$

である。　即ち平均のまはりの4次のモーメントである。

母集団から n 個のサムプルを抽出し，母集団平均の推定のためのサムプル平均 $\overline{x}$ をつくるとき，　$\overline{x}$ の信頼巾（信頼度一定）は

$$\alpha\sqrt{\frac{\sigma^2}{n}} = \gamma$$

によつてあたへられる。

$$\frac{\gamma}{\alpha} = \varepsilon \quad \text{とおくと}$$

$$\frac{\sigma^2}{n} = \varepsilon \quad \text{となる。}$$

ε を相対精度とかりに名づけよう。　この ε は欲する精度をきめておけば一定なものである。

$\dfrac{\sigma^2}{n}$ が ε より小になる様に n を決めねばならぬのである。

さて今 n' 個のサムプルを先づ抽出したとしよう。　此から σ^2 の

偏りのない推定値 $S_{n'}^2$ をつくらう。　此から $\overline{x}_{n'}$（サムプル平均）の相対精度の推定 $\dfrac{S_{n'}^2}{n'}$ を考へる。

此がどの位小であるならば $\dfrac{\sigma^2}{n'} < \varepsilon$ なることが言はれるであらうか。

今 $\dfrac{S_{n'}^2}{n'}$ が $\varepsilon(n')$ より小ならばある一度の信頼度の下に $\dfrac{\sigma^2}{n'} < \varepsilon$，が言はれるものとしよう。

それならば $\varepsilon(n')$ は如何に定められるであらうか。

まづ $S_{n'}^2$ の分散を考へてみよう。

此が

$$\sigma_{S_{n'}^2}^2 = \frac{\sigma^4}{n}\left(\beta_2 - \frac{n'-3}{n'-1}\right)$$ であることは

よく知られてゐる所である。

$$\sigma_{S_{n'}^2} = \sigma^2\sqrt{\frac{1}{n'}\left(\beta_2 - \frac{n'-3}{n'-1}\right)}$$

さて，σ^2 の推定のための信頼区間の問題を考へてみよう。此れには単なるシエブイシエフの定理でなく高能化されたものを用ひよう。

つまり

$$P_r\left\{|y - E(y)| < k\sigma_y\right\} \geqq 1 - \frac{\mu_{2\lambda}}{k^{2\lambda}\sigma_y^{2\lambda}}$$

λは正の整数 σ_y^2，$\mu_{2\lambda}$ は y の分散平均のまはりの２次のモーメント

特に $y = \dfrac{y_1 + \text{------} + y_n}{n}$ と考へるならば

一般は n がさう小でない時近似的に

$$P_r\left\{|y - E(y)| < k\sigma_y\right\} \geqq 1 - \frac{1.3\text{----}(2\lambda - 1)}{k^{2\lambda}}$$

が言へる。　此よりみるときは信頼度95%ならば $k=2.4$ 程度でよいことが言へる。　安全をみても $k=3$ ととれば十分信頼出来る。　なほ，一般の場合でも $k=3$ と考へれば相当信頼ある結論である事が了解せられる。

我々の場にもどらう。

yとして，$S_{n'}^2$ を考へ $\sigma_y^2=\sigma^2$ と考へよう。

かうするならばサムプルたる $S_{n'}^2$ が

$$\sigma^2 \pm k\sigma^2\sqrt{\frac{1}{n'}\left(\beta_2-\frac{n'-3}{n'-1}\right)} \qquad (k=2.5\sim 3)$$

なる巾を逸脱する確率は十分小になつてくるのである。

ましてや此の下限 $\underline{S}_{n'}^2$

$$\underline{S}_{n'}^2=\sigma^2\left(1-k\sqrt{\frac{1}{n'}\left(\beta_2-\frac{n'-3}{n'-1}\right)}\right) \text{ よりも}$$

サムプルの $S_{n'}^2$ が，小となつて現れる確率はますます小となつてきて「此より小なることはない」と言ふ Proposition は十分安全なものとなつてくるであらう。

此の下限 $\underline{S}'_{n'}$ を用いて考へをすすめよう。

$$\frac{\underline{S}_{n'}^2}{n'}=\frac{\sigma^2}{n'}\left(1-k\sqrt{\frac{1}{n'}\left(\beta_2-\frac{n'-3}{n'-1}\right)}\right)$$

を考へる。　n'をもつて欲する精度のサムプルが得られるためには即ち $\dfrac{\sigma^2}{n'}=\varepsilon$ であるためにはその條件の下で

$$\frac{\underline{S}'_{n'}}{n'}=\varepsilon(n')$$ ときめてあればよいであらう。　つまり我々のサムプルからつくつた $\dfrac{S_{n'}^2}{n'}$ が此の $\varepsilon(n')$ より大でなければサムプリングをやめてよいのである。　即ち

$$\varepsilon(n')=\frac{\sigma^2}{n'}\left(1-k\sqrt{\frac{1}{n'}\left(\beta_2-\frac{n'-3}{n'-1}\right)}\right)$$

$$= \varepsilon \left(1 - k \sqrt{\frac{1}{n'} \left(\beta_2 - \frac{n'-3}{n'-1} \right)} \right)$$

が，サムプリングの続行、中止を決定する規準となるのである。

なんとならば

$$\frac{S_{n'}^2}{n'} < \varepsilon \left(1 - k \sqrt{\frac{1}{n'} \left(\beta_2 - \frac{n'-3}{n'-1} \right)} \right)$$

とすれば，　$S_{n'}^2$　は十分安全に（信頼度高く）

$$S_{n'}^2 \geqq \sigma^2 \left(1 - k \sqrt{\frac{1}{n'} \left(\beta_2 - \frac{n'-3}{n'-1} \right)} \right)$$

であるから

$$\frac{\sigma^2}{n'} \left(1 - k \sqrt{\frac{1}{n'} \left(\beta_2 - \frac{n'-3}{n'-1} \right)} \right) < \varepsilon \left(1 - k \sqrt{\frac{1}{n'} \left(\beta_2 - \frac{n'-3}{n'-1} \right)} \right)$$

となり，　$\frac{\sigma^2}{n'} < \varepsilon$　となるからである。

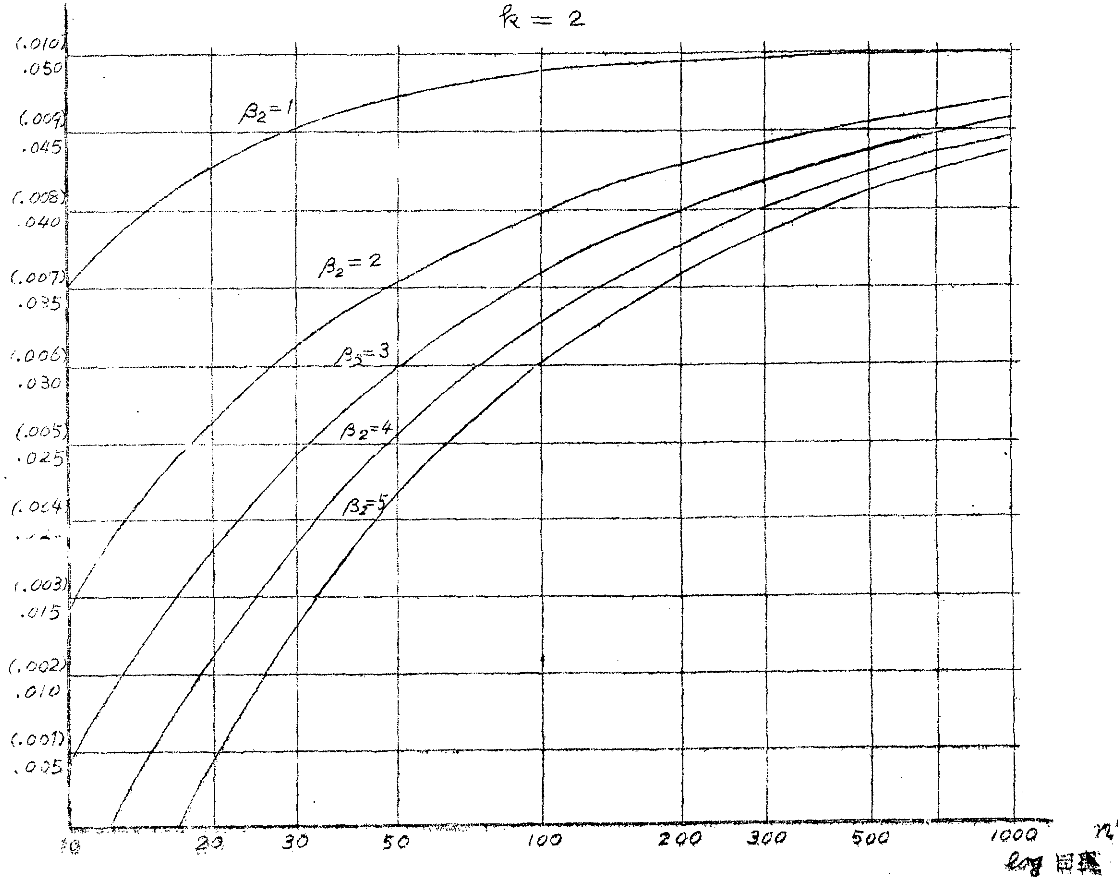
此の $\varepsilon(n')$ の曲線を n' の値についてあらかじめつくつておき，一サムプルを n' 個とり出して $\frac{S_{n'}^2}{n'}$ をつくり此の値が上の曲線の上にあるか下にあるかによつてサムプリングの逐次試行を考へればよいであらう。　一般に β_2 は同様未知であるがサムプルの値から一応推定されると考へられるから実際のサムプリングの用に供す爲め次に k の値，β_2 の値，ε の値について $\varepsilon(n')$ の曲線を描いてみよう。

実際の時は k, ε の値はきまつてゐるから β_2 の値を推定し其に応ずる $\varepsilon(n')$ をつかへばよいのである。

なほ，グラフは n' について log. 目盛によつて描いておいた。

$$\varepsilon(n') = \varepsilon\left(1 - k\sqrt{\frac{1}{n'}\left(\beta_2 - \frac{n'-3}{n'-1}\right)}\right)$$

$k = 2$



括弧 $\varepsilon = 0.01$

$\varepsilon = 0.05$

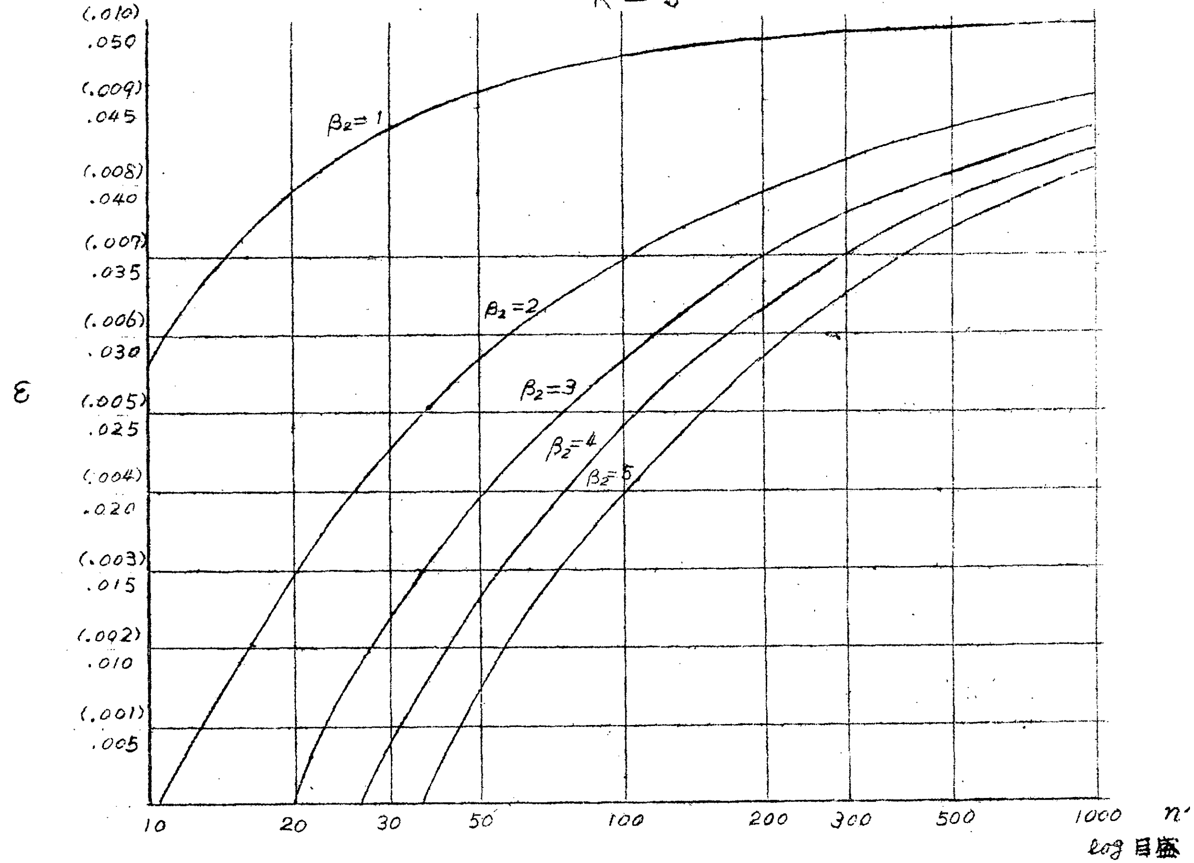
ε

$$\varepsilon(n') = \varepsilon\left(1 - k\sqrt{\frac{1}{n'}\left(\beta_2 - \frac{n'-3}{n'-1}\right)}\right)$$

$$k = 3$$

括弧 $\varepsilon = 0.01$

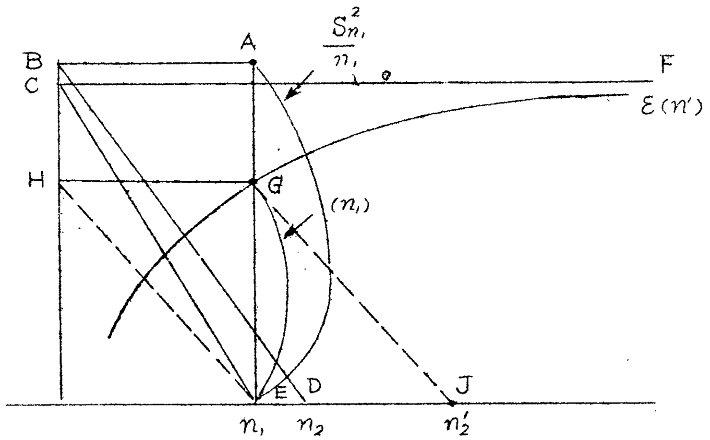
$\varepsilon = 0.05$



n' 個のサムプルをとり $\dfrac{S_{n'}^2}{n'}$ をつくり此が $\varepsilon(n')$ の曲線より上にあつたとする時次に何個のサムプルをとるべきかの目安のつけ方を示しておかう。

まづ n_1 個のサムプルをとる。

A点は $\dfrac{S_{n_1}^2}{n_1}$ をあらはす。



AB // FC // DE // GH とする。

此のとき次にとるべきサムプル数は（$n_2 - n_1$）と目安がつけられる。　ここに EC // DB である。何とならば，近似的にみて次の様なことが考へられるからである。

$$\frac{\sigma^2}{n_1} \doteqdot \frac{S_{n_1}^2}{n_1}$$

$$\frac{\sigma^2}{n_2} = \varepsilon$$

$$n_2 : n_1 = \frac{S_{n_1}^2}{n_1} : \varepsilon$$

此の方法によつて逐次求めてゆくのであるが或は HE // GJ によつて（$n_2' - n_1$）個とる目安をつけることもよい。後者の方がより安全な仕組みと考へられる。

次に此の様にしてサンプル数 n' をつくつてゆくとき，σ^2を知つてゐる場合にくらべて同一の精度をうるためにどの位どう言ふ形でサンプル数は増大するものであらうか，又 n' の分布（ n' が確率変数となることに注意）はどうなるであらうか。

一つの簡単な実例について此の様態をしめしてみよう。

母集団の分布が $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ であたへられてゐると考へる。

この時は $\sum_{i=1}^{n} x_i^2$ は χ^2分布にしたがふ事に注意されたい。

此の時 $S^2 = \frac{1}{n'} \sum_{i=1}^{n'} x_i^2$ が母集団分散の偏りのない推定値となる。

$$\sigma_{S^2}^2 = \frac{\sigma^4}{n'}(\beta_2 - 1) = 2\frac{\sigma^4}{n'}$$ となる。

此の様な値をつかひ $\varepsilon(n')$ が決定されるのである。

数値的には χ^2 分布の表を用ひて容易に決定されるのである。

1個サムプルをとり出し $x_1^2 = y_1$ をつくる。

次に，又一個とり出す。 云々、 n' 個をとり出し此でサムプリングのおはる確率如何、

y_i の分布は $$\frac{1}{\sqrt{2\pi}}(y)^{-\frac{1}{2}} e^{-\frac{y}{2}} dy = p(y_i)\,dy_i$$

である。 一般に

$$\frac{S_{n'}^2}{n'} \leqq \varepsilon(n') \quad 即ち \quad S_{n'}^2 < n'\varepsilon(n')$$

ならばサムプリングがおはる。

$$即ち \quad x_1^2 + \cdots\cdots + x_{n'}^2 \leqq n'^2 \varepsilon(n') = \tau(n')$$

$$y_1 + \cdots\cdots + y_{n'} \leqq \tau(n')$$

ならおはるのである。

したがつて n' 個サムプルを抽出しサムプリングのおはる確率 q_n

は

$$q_{n'} = \underset{\substack{\infty > y_1 \geqq \tau(1) \\ \infty > y_1 + y_2 \geqq \tau(2) \\ \vdots \\ \infty > y_1 + \cdots + y_{n'-1} \geqq \tau(n'-1) \\ \tau(n) \geqq y_1 + \cdots + y_{n'} \geqq 0}}{\int\int\cdots\int} p_1(y_1)p_2(y_2)\cdots p_n(y_n)dy_1dy_2\quad dy_n$$

である。

今 $$\underset{\substack{\infty > y_1 \geqq \tau(1) \\ \vdots \\ \infty > y_1 + \cdots + y_{n'-1} \geqq \tau(n'-1) \\ \infty > y_1 + \cdots + y_{n'-1} + y_{n'} \geqq \tau(n')}}{\int\int\cdots\int} p_1\cdots p_n dy_1 \cdots\quad dy_n = r_{n'}$$

とおくと　$q_{n'} = r_{n'-1} - r_{n'}$　が成立する。（$n' \geqq 1$, $r_0 = 1$）

さて，此の r_n を，ぢりくれ変換によつて変数の変換をやれば

$$r_{n'} = \int_{\tau(n')}^{\infty} p(\chi_n^2)\,d\chi_n^2$$

となる。　此を用いて $q_{n'}$ を出し n' のモーメントを出さう。

$$E(n) = 1 + \sum_{i=1}^{\infty} \int_{\tau(i)}^{\infty} p(\chi_i^2)\,d\chi_i^2$$

$$E(n^2) = 1 + \sum_{i=1}^{\infty} \left((i+1)^2 - i^2\right) r_i$$

$$E(n^3) = 1 + \sum_{i=1}^{\infty} \left((i+1)^3 - i^3\right) r_i$$

$$E(n^4) = 1 + \sum_{i=1}^{\infty} \left((i+1)^4 - i^4 \right) \gamma_i$$

τ(n') が求められ定められてゐるから，$\gamma_{n'}$ は χ^2 分布の表から計算することができしたがつてモーメントを数値的に計算することができる。 此の計算は相当面倒であり且つ χ^2 分布の表が階差をとつて行つてみるとすこぶる数値にあやしい所が多いのできはめて困難を感じた。 （χ^2 分布の表の計算を正しくしなほし，且つ詳細なものにする必要が痛感せられる。） したがつて我々の場合 n' のモーメントを決定するのに相当近似的な計算を行つた。

今假りに $\varepsilon = 0.2$ （$\sigma^2 = 1$ であるから $n = 5$ の場合）とつて試みに計算を行つてみると次の様になつた。

平　均	11.5
分　散 σ^2	12.5
三次のモーメント μ_3	−16.8
四次のモーメント μ_4	512.7

を得た。

此より
$$\beta_1 = \frac{(\mu_3)^2}{(\sigma^2)^3} = 0.14$$

$$\beta_2 = \frac{\mu_4}{\sigma^4} = 3.28$$

を得た。 n' の分布の形をペアソン系によつて考へてみるに判別條件によると

$$\text{IV 型} \qquad y = y_0 \left(1 + \frac{x^2}{a^2}\right)^{m} e^{-\nu \tan^{-1}\frac{x}{a}}$$

と言ふことになる。 しかしペアソン系のあてはめには相当疑问があると考へられるから実際に実験せざるかぎり n' の分布の形がこれの型であると言ふのはきはめて危険である。しかし平均，分散 β_1，β_2 よりみて n' の様子はほぼ察せられることであらう。