⑧ 実験データの棄却について

阪大理学部数学教室

小川潤次郎

1組の実験データがあるとき，その中に異質的なものが混入しているならば，これを摘出するとか，又わ既に得られている1組のデータに対して新たに得られたデータが等質的なものと見られるかどうかを判定する場合に所謂"棄却検定法"（Rejection Test）が用いられる。

この棄却検定法については幾多の文献があるが(1)，その中にわ概念の混乱も見受けられるようである。(2) この棄却検定法従って棄却限界の概念を明確にするには矢張り統計的仮説検定論（Theory of Testing Statistical Hypotheses）(3) の立場に立つのが最も良いが，一般にわ複合仮説の検定となるので，対立仮説が複雑で仲々うまく数学的取扱いが出来ないのは E.S. Pearson 及び C. Chandra Sekar(4) の指摘する通りである。

こゝでは先づ最も単純な Situation について考えよう。未知なる真値 m をもつ量 X を n 回測定して，n コの実測値 $x_1, x_2, \cdots, x_n$ を得たとする。但しこの測定にわ Systematic Error わないものとする。そうすれば，この実測値 $x_1, x_2, \cdots, x_n$ わ真値 m に夫々偶発誤差 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ が加わったものと考えてよいであろう。

このとき $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ わ互に独立に平均 0 なる正規分布に従うものと考える。

このような Situation の下で，例えば x_1 が残餘の $x_2, \cdots, x_n$ に対して異質的であるかどうかという問題は，一応次のように Formulate されるであろう。

問題： x_1 わ正規母集団 $N(m, \lambda^2\sigma^2)$ から抽出された Random

Sample, $x_2, \cdots, x_n$ が同一の正規母集団 N (m, σ^2) から抽出された Size $(n-1)$ の Random Sample であり，従って $x_1, x_2, \cdots, x_n$ の尤度函数が

$$p(x_1, x_2, \cdots, x_n ; \lambda, m, \sigma)$$

$$= \frac{1}{\lambda}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n exp\left\{-\frac{1}{2\sigma^2}\left(\frac{(x_1-m)^2}{\lambda^2}+(x_2-m)^2+\cdots+(x_n-m)^2\right)\right\}$$

であるとき，パラメーター λ に関する統計的仮説

$$H.\quad \lambda = 1$$

を検定する Optium な Critical Region を求めよ。

この場合には，パラメーターは三つあって，その admissible values わ

$$-\infty < m < +\infty,\quad 0 < \sigma < +\infty,\quad 0 < \lambda < +\infty$$

であるから，仮説 H わ明かに複合仮説である。

吾々わ以下に於てわ，λ の admissible values としてわ

$$\lambda \geqq 1$$

のみ，従って H の対立仮説の指定する λ の値としてわ

$$\lambda > 1$$

のみ考えることにするが，棄却検定の意味を考えるとき，これわ不都合でわないであろう。

先づ初めに，σ が既知の場合を考える。これは森口氏(5)が取扱った場合である。このときわ $\sigma = 1$ として一般性を失わない。

J. Neyman & E.S. Pearson の一般的方法(6)を適用する爲に

$$p_0(x_1, x_2, \cdots, x_n) = p(x_1 \cdots x_n ; 1, m, 1)$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2}\left((x_1-m)^2 + \cdots + (x_n-m)^2\right)\right\}$$

とおくと，

$$\phi = \frac{\partial \log p_0}{\partial m} = n(\bar{x} - m), \quad \phi' = \frac{\partial \varphi}{\partial m} = -n$$

であるから，Neyman－Pearson の條件[7]

$$\phi' = A + B\phi$$

わ満される。このとき

$$\phi = \text{Const.}$$

なる Hyper Surface わ，超平面

$$\bar{x} = \text{Const.}$$

となる。　今求める Critical Region を w とし，これと上の超平面との交わりを $w(\bar{x})$ とすると，これわ

$$p_t(x_1, \cdots, x_n) \geq k(\bar{x}) \cdot p_0(x_1, \cdots, x_n)$$

で定められる。　但し，

$$p_t(x_1, \cdots, x_n) = p(x_1 \cdots x_n ; \lambda_t \cdot m_t \cdot 1) \qquad \lambda_t > 1.$$

で定数 $k(\bar{x})$ わ，1rst kind の Error の確率を ε（即ち Critical region の Size ε）とすれば

$$\iint_{w(\bar{x})} p_0(x_1 \cdots x_n)\, dx_1 \cdots dx_n$$

$$= \varepsilon \iint_{\bar{x}} p_0(x_1 \cdots x_n)\, dx_1 \cdots dx_n$$

より定められる。　領分領域 $\bar{x}$ というのわ，考える超平面 $\bar{x} = Const.$ の全体の意味である.

ところで $w(\bar{x})$ を書き直すと

$$(x_1 - m_1)^2 \geqq c'(\bar{x})$$

又は

$$|x_1| \geqq c(\bar{x})$$

であつて，x_1 と $\bar{x}$ の同時分布の密度は，仮説 H の下では

$$\text{Const. } e^{-\frac{1}{2}(x_1-\bar{x})^2 - \frac{n}{2}\bar{x}^2}$$

であるから，$\bar{x}$ given のときの x_1 の Conditional frequency function は

$$\text{const. } e^{-\frac{1}{2}(x_1-\bar{x})^2}$$

である。　よつて $w(\bar{x})$ は，

$$|x_1 - \bar{x}| \geq \lambda_p$$

但し

$$1 - 2\int_0^{\lambda_p} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = p/100 \equiv \varepsilon.$$

によつて定められる。

よつて，この場合の最良棄域は，

$$|x_1 - \bar{x}| \geqq \lambda_p \sigma$$

である。 従って棄却限界は

$$\bar{x} - \lambda_p \sigma, \quad \bar{x} + \lambda_p \sigma$$

である。 これは森口氏が何んとなく作ったものである。

次には σ を unknown とする。この場合は

$$p_0(x_1, \cdots, x_n) = p(x_1, \cdots, x_n; 1, m, \sigma)$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2\right\}$$

$$\log p_0 = \text{Const} - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2$$

$$\phi_1 = \frac{\partial \log p_0}{\partial m} = \frac{n}{2\sigma^2}(\bar{x} - m). \quad \phi_1' = \frac{\partial \phi_1}{\partial m} = -\frac{n}{2\sigma^2}$$

$$\phi_2 = \frac{\partial \log p_0}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \Sigma (x_i m)^2$$

$$\phi_2' = \frac{\partial \phi_2}{\partial \sigma} = \frac{n}{\sigma^2} - \frac{3}{\sigma^4} \Sigma (x_i - m)^2$$

だから、Neyman-Pearson の Condition (5)

$$\phi_1' = A_1 + B_1 \phi_1$$

$$\phi_2' = A_2 + B_2 \phi_2$$

を満たす。 そしてこのときは

$$\phi_1 = \text{Const} \qquad s^2 = \text{Const}$$

となる。 よって仮説 H の Best Critical Region とこれらの Hyper surfaces との交わりを $\omega(\bar{x}, s^2)$ とすれば、それは、

$$p_t\,(x_1 - x_n) \geqq k\,(\bar{x}, s^2)\,p_0(x_1, \cdots, x_n)$$

で定められる。 但し，定数 $k\,(\bar{x}, s^2)$ は，

$$\underset{\omega(\bar{x}, s^2)}{\iint} p_0\,(x_1 \cdots x_n)\,dx_1 \cdots dx_n = \varepsilon \underset{\bar{x},\, s^2}{\iint} p_0\,(x_1 \cdots x_n)\,dx_1 \cdots dx_n$$

によつて定められる．従つて

$$\omega\,(\bar{x}, s^2): \qquad (x_1 - m)^2 \geqq c\,(\bar{x}, s^2).$$

今仮説 H が真なるときの $x_1, \cdots\cdots x_n$ の分布は

$$\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i - m)^2\right\} dx_1 \cdots dx_n$$

$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left\{-\frac{n}{2\sigma^2}\left(s^2 + (\bar{x} - m)^2\right)\right\} dx_1 \cdots dx_n$$

$$y_1 = \sqrt{n}\,\bar{x} = \frac{1}{\sqrt{n}}x_1 + \frac{1}{\sqrt{n}}x_2 + \cdots\cdots + \frac{1}{\sqrt{n}}x_n$$

$$y_2 = \sqrt{\frac{n}{n-1}}\,(x_1 - \bar{x}) = \sqrt{\frac{n-1}{n}}\,x_1 - \frac{1}{\sqrt{n(n-1)}}x_2 - \cdots - \frac{1}{\sqrt{n(n-1)}}x_n$$

$$y_3 = \qquad\qquad \sqrt{\frac{n-2}{n-1}}\,x_2 - \cdots - \frac{1}{\sqrt{(n-1)(n-2)}}x_n$$

$$\cdots\cdots\cdots\cdots$$

$$y_n = \qquad\qquad \sqrt{\frac{1}{2}}\,x_{n-1} - \frac{1}{\sqrt{2\cdot 1}}x_n$$

なる直交変換をすると，

$$\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp.\left\{-\frac{1}{2\sigma^2}\left(y_2^2 + \cdots + y_n^2 + (y_1 - \sqrt{n}\,m)^2\right)\right\} dy_1\,dy_2 \cdots dy_n$$

$$n s^2 = y_2^2 + \cdots\cdots + y_n^2$$

であるが，こゝで，$\bar{x}$，s^2 given の conditional prob を考える爲に

$$y_2 = \sqrt{n}\, s\, Z_2, \cdots\cdots, y_n = \sqrt{n}\, s\, Z_n$$

とおくと

$$Z_2^2 + \cdots\cdots + Z_n^2 = 1$$

こゝで変数を $\bar{x}, s, Z_1, \cdots\cdots, Z_{n-2}$ に変換すると

$$\frac{2n^{\frac{n}{2}}}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{n}{2\sigma^2}(s^2+(\bar{x}-m)^2)} \frac{s^{n-2}}{\sqrt{1-Z_2^2-\cdots\cdots-Z_{n-2}^2}} d\bar{x}\, ds\, dZ_1 \cdots dZ_{n-2}.$$

こゝで，$\bar{x}, s$ given のときの Z_2 の分布を考える。それ結局

$$\text{Const.} \underbrace{\iint}_{n-4} \frac{dZ_3 \cdots dZ_{n-2}}{\sqrt{1-Z_2^2-Z_3^2-\cdots-Z_{n-2}^2}}\, dZ_2$$

$$= \text{Const.} \times (1-Z_2^2)^{\frac{n-4}{2}} dZ_2$$

$$= \text{Const.} \left(1 - \frac{\tau^2}{n-1}\right)^{\frac{n-4}{2}} d\tau.$$

但 $$\tau = \frac{x_1 - \bar{x}}{s}$$

これはつまり，Thompson の棄却検定である。

このような簡単な場合には Thompson の棄却限界は Optimum であることが証明された訳である。

もつと複雑なときには，Thompson の方法は必ずしも良くない。これについては，E.S. Pearson 及 C. Chandra Sekar の前出の論文参

照

Thompson Test の Efficiency については別に述べる積りである。

E. S. Pearson 及 C. Chandra Sekar の前記論文の指唆に従つて，Order を考慮した。棄却検法が Grubb の最近の論文[9]で扱われていることを指摘し度い。

しかし，いづれも Power という見地からの研究はしていない。

参考文献

(1) W.R.Thompson: On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation. Ann. of Math. Statist. Vol. 6 (1935). p. 214.

E.S. Pearson and C. Chandra Sekar: The efficiency of statistical tools and a criterion for rejection of outlying observations:

Biometrika. Vol. 28 (1936). p. 308.

森口 繁一：実験データの棄却について，

数学，Vol. 2, No. 1 (1949). p. 65

(2) 森口，loc. cit. 参照、尤も森口氏の論文もはつきりと棄却域の概念がつかめているとは思われない点がある。

(3) J. Neyman and E.S. Pearson: On the problem of the most efficient tests of statistical hypotheses. Phil. Trans. of the Roy. Society. London. A 231. (1933). p. 289

(4) E.S. Pearson and C. Chandra Sekar. loc. cit. § 2, (2) 参照

(5) 森口，loc. cit.

(6) J. Neyman and E. S. Pearson, loc. cit.
J. Neyman : On a statistical problem arising in routine analyses and in sampling inspection in man production : Ann. of Math. Statist. Vol. 12. (1941)
(7) J. Neyman and E. S. Pearson, loc. cit.
(8) J. Neyman and E. S. Pearson, loc. cit.
(9) Frank E. Grubb: Sample Criteria For Testing Outlying Observations Ann. of Math. Statist. Vol. 21. NO. 1. March. (1950) pp. 27-58.

(受付 1950. 10. 10)