素分布による適合度の検定法

石　川　栄　助

Test of Goodness of Fit with the Basic Set

Eisuke ISHIKAWA

§1 はしがき

筆者は嘗つて少数標本の精密検定表を述べ，その精密検定法を公にしたが[1]，この方法を拡張しN個の資料が，連続確率密度 $f(x)$ をもつ分布に適合するか否かを，$f(x)$ による n 個の等確率区間に分布するN個の資料の素分布によつてその適合度の検定を工夫する．なお精密検定表を少しく拡げる．

§2 素分布

連続確率密度 $f(x)$ をもつ分布の n 個の等確率区間 $(-\infty,\ c_1), (c_1,\ c_2),\ \cdots\cdots, (c_{n-1},\ \infty)$ を $R_1,\ R_2,\ \cdots\cdots, R_n$ とおき，N個の資料 $x_1,\ x_2,\ \cdots\cdots, x_N$ が区間 R_i に含まれる度数を a_i とし，度数分布 $(a_1,\ a_2,\ \cdots\cdots, a_n)$ を考える．これを資料 $x_1,\ x_2,\ \cdots\cdots, x_N$ の確率密度函数 $f(x)$ に関する n 区の素分布又は単に**一般素分布**とよぶことにする．記号的に $(a_1,\ a_2,\ \cdots\cdots, a_n)$ で表わす．ここに $c_1,\ c_2,\ \cdots\cdots,\ c_{n-1}$ は次の式を満足する限界である．

$$\int_{-\infty}^{c_1} f(x)\,dx = \int_{c_1}^{c_2} f(x)\,dx = \cdots\cdots\cdots\cdots = \int_{c_{n-1}}^{\infty} f(x)\,dx = \frac{1}{n} \quad \cdots\cdots(1)$$

特に $a_1 = a_2 = \cdots\cdots = a_n = \frac{N}{n}$ なる素分布 $\left(\frac{N}{n},\ \frac{N}{n},\ \cdots\cdots, \frac{N}{n}\right)$ を，**一般標準素分布**とよぶことにする．

扨て，一般素分布 $(a_1,\ a_2,\ \cdots\cdots, a_n)$ に於て零ならざるものが順に $b_1,\ b_2,\ \cdots\cdots b_s$ である時，これを $[b_1,\ b_2,\ \cdots\cdots, b_s]$ で表わすことがある．ここに $s \leqq n$ である．

一般素分布 $(a_1,\ a_2,\ \cdots\cdots, a_n)$ 又は $[b_1,\ b_2,\ \cdots\cdots, b_s]$ に於いて要素 $a_1,\ a_2,\ \cdots\cdots, a_n$ 又は $b_1,\ b_2, \cdots\cdots, b_s$ の排列の順序のみ異なる素分布を同種の分布とみなしうるから，一般素分布 $(a_1,\ a_2,\ \cdots\cdots, a_n)$ の出現する確率 p は

$$p = {}_n\mathrm{C}_s \left(\frac{1}{n}\right)^N \frac{N\,!}{b_1!\ b_2!\ \cdots\cdots b_s!} \cdot \frac{s\,!}{\alpha_1!\ \alpha_2!\ \cdots\cdots \alpha_t!} \quad \cdots\cdots\cdots\cdots(2)$$

ここに $b_1,\ b_2, \cdots\cdots, b_s$ のうち同一文字がそれぞれ $\alpha_1,\ \alpha_2, \cdots\cdots, \alpha_t\ (t \leqq s)$ 個あるものとする．

§3　素分布検定法

確率密度 $f(x)$ をもつ分布の n 個の等確率区間による資料 x_1, x_2, ……, x_N の実現素分布 $(a_1, a_2, \cdots\cdots, a_n)$ と一般標準分布 $\left(\frac{N}{n}, \frac{N}{n}, \cdots\cdots, \frac{N}{n}\right)$ とのくいちがいの程度を χ^2—量で表わせば,

$$\chi_0^2 = \frac{\sum_i^n \left(a_i - \frac{N}{n}\right)^2}{\frac{N}{n}} = \frac{n}{N}\sum_j^s b_j^2 - N \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(3)$$

従つて　$P = P_r\left(\chi^2 \geqq \chi_0^2\right)$　である確率Pは

$$P = \sum_{\chi^2 \geqq \chi_0^2} {}_nC_s \left(\frac{1}{n}\right)^N \frac{N!}{b_1!\ b_2!\cdots\cdots b_s!} \cdot \frac{s!}{\alpha_1!\ \alpha_2!\cdots\cdots\alpha_t!} \quad \cdots\cdots(4)$$

この確率Pが有意水準 α 以下であれば，この資料 x_1, x_2, ……, x_N は確率密度 $f(x)$ である母集団からの標本でないと断定する．よつて予めNと n およびP＝α とを与えて(4)式より χ_α^2 を求めておけば，$\chi_0^2 \geqq \chi_\alpha^2$ ならば $P \leqq \alpha$ となり，容易に適合度が検定出来る（この表は計算中）．

§4　単純素分布による検定法

素分布 $(a_1, a_2, \cdots\cdots, a_N)$ に於いて $N = n$ なる時，これを単純素分布とよび一般素分布と区別する．又この場合の一般標準素分布は（1，1，……，1）となる．これを単に標準素分布とよぶことにする[2]．然る時，(3)，(4) 式は次のようになる．

$$\chi_0^2 = \sum_i (a_i - 1)^2 = \sum_j^s b_j^2 - n \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(5)$$

$$P = \sum_{\chi^2 \geqq \chi_0^2} {}_nC_s \left(\frac{1}{n}\right)^n \frac{n!}{b_1!\, b_2!\cdots\cdots b_s!} \cdot \frac{s!}{\alpha_1!\, \alpha_2!\cdots\cdots\alpha_t!} \quad \cdots\cdots(6)$$

ここに　$P = P\left(\chi^2 \geqq \chi_0^2\right)$　である．

依つて $P \leqq \alpha$ ならば，この資料は確率密度 $f(x)$ をもつ母集団からの標本でないとする．次表は n と α とを与えての χ_α^2 の値である．

第1表　単純素分布による χ^2 の表

n \ α	0.05	0.01	0.001	n \ α	0.05	0.01	0.001
3	—	—	—	12	22	26	36
4	12	—	—	13	22	28	36
5	12	20	—	14	24	30	38
6	14	20	30	15	26	32	40
7	14	18	30	16	26	32	42
8	16	20	24	17	28	34	44
9	16	22	32	18	30	36	46
10	18	24	32	19	30	38	48
11	20	26	34	20	32	38	48

§5 少数標本に於ける正規性の検定

少数資料 $x_1, x_2, \cdots\cdots, x_n$ が正規分布 (m, σ^2) の標本であるか，否かは分散分析その他の検定の時，度々問題となる．これの第1表を利用しての検定を考える．

今上の資料 $x_1, x_2, \cdots\cdots, x_n$ が 正規分布 (m, σ^2) をなすとして，次の様に標準化する．

$$u_i = \frac{x_i - m}{\sigma} \qquad \cdots\cdots(7)$$

然る時，u の分布は標準正規分布をなす．よつてあらかじめ標準正規分布のn個の等確率区間 R_1 $R_2, \cdots\cdots, R_n$ の限界点 $c_1, c_2, \cdots\cdots, c_{n-1}$ を求めておく時は素分布$(a_1, a_2, \cdots\cdots, a_n)$は容易に求まる．従つて$\chi_0^2$ を(5)式によつて計算し，n と α を目安に χ_α^2 を第1表より見出し，

$\chi_0^2 \geqq \chi_\alpha^2$ ならば，この資料は正規でない．

$\chi_0^2 < \chi_\alpha^2$ ならば，この資料は正規とする．

なお，標準正規分布のn個の等確率区間の限界点の値は次の通りである．

第2表　標準正規分布のn等積分點

（負の値及び∞點を除く）

n	上	の	限	界	値					
2	0									
3	0.430									
4	0	0.674								
5	0.253	0.841								
6	0	0.430	0.967							
7	0.180	0.565	1.067							
8	0	0.318	0.674	1.150						
9	0.139	0.430	0.764	1.220						
10	0	0.253	0.524	0.841	1.281					
11	0.114	0.348	0.604	0.908	1.335					
12	0	0.210	0.430	0.674	0.967	1.382				
13	0.096	0.293	0.502	0.736	1.020	1.426				
14	0	0.180	0.366	0.565	0.791	1.067	1.465			
15	0.083	0.253	0.430	0.622	0.841	1.110	1.501			
16	0	0.157	0.318	0.488	0.674	0.887	1.150	1.534		
17	0.073	0.223	0.377	0.541	0.721	0.928	1.186	1.564		
18	0	0.137	0.282	0.430	0.589	0.764	0.967	1.220	1.593	
19	0.066	0.199	0.336	0.479	0.633	0.804	1.003	1.252	1.619	
20	0	0.125	0.253	0.385	0.524	0.674	0.841	1.036	1.281	1.644

例えば5等分点は，上表の（0.253, 0.841）を用ひ（—∞, —0.842），（—0.841, —0.254),（—0.253, +0.253),（0.254, 0.841),（0.842, ∞）の様に使用する．

§6 使 用 例

資料 9，9，9，9，12, 12, は，正規分布をなす母集団からの標本か．

解 R. A. FISHER による k — 統計量を用いて検定すれば，

$$k_2 = 2.40 \qquad k_3 = 3.60 \qquad k_4 = 10.80$$

依つて $g_1 = 0.968$ 　 $g_2 = 1.875$ ， $\sigma_{g1} = 0.845$ 　 $\sigma_{g2} = 1.740$

従つて 歪度について $u_0 = \dfrac{0.968}{0.845} = 1.146$

故に $P = P(|u| \geqq 1.146) > 0.25$

又， 尖度について $u_0 = \dfrac{1.875}{1.740} = 1.078$

故に $P = P(|u| \geqq 1.078) > 0.28$

即ち，歪度によるも尖度によるも，この資料は正規分布をなす母集団からの標本とみなしてよい．

これを上の素分布検定法にかければ，

上の資料から $m = 10$, 　 $\sigma = 1.414$

依つて（7）式によつて標準化すれば，

$$u_i = \frac{x_i - 10}{1.414}, \qquad \therefore \ u_1 = u_2 = u_3 = u_4 = -0.707, \quad u_5 = u_6 = 1.414,$$

又，第2表から等確率分点を求めれば，

（—∞, —0.967, —0.430, 0, 0.430, 0.967, +∞）従つて，この資料の素分布は（0，4，0，0，0，2）又は〔4，2〕となる．故に，

$$\chi_0^2 = \sum_j^2 b_j^2 - n = 16 + 4 - 6 = 14$$

第1表より $n = 6$ の時の $\chi^2_{0 \cdot 05} = 14$

即ち $\chi_0^2 = \chi^2_{0 \cdot 05}$ 故に この資料は正規分布（10, 1.414^2）からの標本といわれない．

文 献

1) 石川榮助 : 適合度の精密検定表（科學 vol 23. no. 11. 1953)

2) 所 一夫 : 小標本の場合の適合度検定について（科學 vol 21. no. 8. 1951)

3) 所 一夫 : 小標本の場合の正規性の検定表（科學 vol 22. no. 11. 1952)

4) KENDALL : The advanced theory of statistics vol. 2. (1951)

Summary

Let $x_1, x_2, \cdots\cdots, x_N$ be a random small sample from the population with the $d.f. f(x)$. and let $c_1, c_2, \cdots\cdots, c_{n-1}$ be limit values of the following integrations,

$$\int_{-\infty}^{c_1} f(x)\,dx = \int_{c_1}^{c_2} f(x)\,dx = \cdots\cdots = \int_{c_{n-1}}^{\infty} f(x)\,dx = \frac{1}{n} \quad .$$

And let a_i, be the frequency of x's that fall into the interval (c_{i-1}, c_i).

Then we will have a set $(a_1, a_2 \cdots\cdots, a_n)$, which is named "basic set" by the writer.

In such a $(a_1, a_2 \cdots\cdots a_n)$, when $a_1 = a_2 = \cdots\cdots = a_n = \frac{N}{n}$, it is named the "standard set".

If $(a_1\ a_2 \cdots\cdots a_n)$ and $\left(\frac{N}{n}, \frac{N}{n}, \cdots\cdots, \frac{N}{n}\right)$ are two sets, the significance of the difference between their sets, it may be tested by calculating the following statistics.

$$\chi_0^2 = \frac{\sum_i \left(a_i - \frac{N}{n}\right)^2}{\frac{N}{n}}$$

$$P = P\left(\chi^2 \geqq \chi_0^2\right) = \sum_{\chi^2 \geqq \chi_0^2} {}_nC_s \left(\frac{1}{n}\right)^N \frac{N!}{b_1!\, b_2! \cdots\cdots b_s!} \cdot \frac{s!}{\alpha_1!\, \alpha_2! \cdots\cdots \alpha_t!}$$

where $b_1, b_2, \cdots, b_s$ represent natural numbers which are not zero in the set $(a_1, a_2 \cdots\cdots, a_n)$ $(s \leqq n)$, and $\alpha_1, \alpha_2, \cdots\cdots, \alpha_t$ are the frequency of the same natural numbers in the set $(b_1, b_2, \cdots\cdots, b_s)$, $(t \leqq s)$.

In the set $(a_1, a_2 \cdots\cdots, a_n)$ if $N = n$, then $\chi_0^2 = \sum b_j^2 - n$, and

$$P = P\left(\chi^2 \geqq \chi_0^2\right) = \sum_{\chi^2 \geqq \chi_0^2} {}_nC_s \left(\frac{1}{n}\right)^n \frac{n!}{b_1!\, b_2! \cdots\cdots b_s!} \cdot \frac{s!}{\alpha_1!\, \alpha_2! \cdots\cdots \alpha_t!}$$

And values of P is shown on the following table.

Table of χ_α^2

n	0.05	0.01	0.001	n	0.05	0.01	0.001
3	—	—	—	12	22	25	36
4	12	—	—	13	22	28	36
5	12	20	—	14	24	30	38
6	14	20	30	15	26	32	40
7	14	18	30	16	26	32	42
8	16	20	24	17	28	34	44
9	16	22	32	18	30	36	46
10	18	24	32	19	30	38	48
11	20	26	34	20	32	38	48

Example Can the data : 9, 9, 9, 9, 12, 12 be a sample from the population of N(10, 2) ?

Solution Let u_i be normal deviate as follows;

$$u_i = \frac{x_i - m}{\sigma} = \frac{x_i - 10}{\sqrt{2}}$$

Then $u_1 = u_2 = u_3 = u_4 = -0.707$, $u_5 = u_6 = 1.414$

and $c_1 = -0.967$, $c_2 = -0.430$, $c_3 = 0$, $c_4 = 0.430$, $c_5 = 0.967$

Thus we have basic set (0, 4, 0, 0, 0, 2), Therefore $\chi_0^2 = \sum_i \left(a_i - 1\right)^2 = 14$

By the table of χ_α^2, value of χ_α^2 $(n = 6\ ,\ \alpha = 0.05)$ is 14.

since $P = 0.05$

The null hypothesis may be rejected.