χ^2適合度検定の初等的証明
An Elementary Proof of K.Pearson's Chi-square Test for Goodness of Fit

中嶋眞澄
Masumi NAKAJIMA
Department of Economics
International University of Kagoshima
Kagoshima 891-0191, JAPAN
e-mail: nakajima@eco.iuk.ac.jp

概 要

Abstract

We prove here K.Pearson's chi-square test for goodness of fit elementarily.

Key words ; K.Pearson's chi-square test for goodness of fit, χ^2 distribution.
Mathematics Subject Classification 2010;　62E20, 60E99.

χ^2適合度検定とは，ドイツの測地学者兼天文学者Helmertが[3]で発見し，その主要性質を[4]で求めたχ^2分布[2][12]を使って，K.Pearson(1857-1936, 息子のE.S.Pearson(1895-1980)も統計学者)が，1900年[10]で提案した，想定する確率分布とデータが適合しているか否かをある確率で判定する，今日の統計学で広く使用されている検定法の一つで以下のようなものである[1]([1]は[6]の後，著者の講義「統計学II」,「確率と統計II」で教科書として使い続けているバランスの取れた良書である。)。([7]ではχ^2分布とは呼ばずに，その発見者に敬意を表してHelmert分布と呼んでいる。)：
未知のr項分布母集団Zに対して，Zの値域$R \subset \mathbf{R}$を，$R = \bigcup_{i=1}^{r} A_i$

$(A_i \cap A_j = \emptyset \text{ for } i \neq j)$ と分解し，$Z \in A_j$ となる確率 $\mathrm{P}\{Z \in A_j\} =: p_j$，$(0 \leq p_j \leq 1,\ \sum_{j=1}^{r} p_j = 1)$ と想定し (帰無仮説)，実際に測定した n 個の Z の実現値 z で $z \in A_j$ となる度数を n_j とするとき $(\sum_{j=1}^{r} n_j = n)$:

事象	A_1	A_2	$\cdots$	A_r
確率	p_1	p_2	$\cdots$	p_r
頻度	$\frac{n_1}{n}$	$\frac{n_2}{n}$	$\cdots$	$\frac{n_r}{n}$

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$$

が，この仮説の下で $n \to \infty$ のとき，自由度 $(r-1)$ の χ^2 分布をすることを利用して χ^2 検定するものである。
この検定は次の定理を使っている。

定理 1　Z_i が r 項分布 $\mathrm{Mul}(n; p_1, p_2, \cdots, p_r)$ に従っているとき (このとき $Z_i \sim \mathrm{B}(n, p_i)$，即ち各 Z_i は 2 項分布 $\mathrm{B}(n, p_i)$ に従っている)，もしくは $(Z_1, Z_2, \cdots, Z_r) \sim \mathrm{Mul}(n; p_1, p_2, \cdots, p_r)$ のとき ($r = 2$ のときは，$\mathrm{Mul}(n; p_1, p_2) = \mathrm{B}(n, p_1)$ である)，

$$\chi^2 = \sum_{i=1}^{r} \frac{(Z_i - np_i)^2}{np_i}$$

は，$n \to \infty$ のとき，自由度 $(r-1)$ の χ^2 分布に従う。(実際の適用では，$n \to \infty$ ではなく $np_i \geq 5$ $(i = 1, 2, \cdots, r)$ で十分である。)

この定理の言明は簡単であるのにもかかわらず，その証明は難しいせいか，或いは最近の大学生の学力低下の為か，最近の殆どの統計学の書物にはその証明が記載されていない。又，誤った記述がされているものさへある (例えば，過去私が「統計学 II」の教科書として使った [6] の p.130,132,134 では，「$n \to \infty$ では」或いは「近似的に」という語が脱落している)。昨今の新聞紙上をにぎわした血圧降下剤の臨床試験論文のデータ改竄事件を始め，統計学を使用している論文等に統計学の使用上の誤用不正が多く見受けられる。このような御時世からも，判断力のない初心者に対する教科書では，些細な誤りさへ禁忌である。19 世紀英国の首相 Disraeli が述べ，今日英国では諺の様になっていると言われる「嘘には 3 種類ある。普通の嘘，見え透いた嘘，そして統計」が今日の日本で体現されているかのようである。この為か，この定理 1 の証明は大学生の水準の高かっ

た時代の教科書 [5] [9] か，統計学のユーザーである研究者ではなく，統計学自体の研究者となるべき人達が読む限られた専門書にしか現在掲載されていない。このような所からも，上記のような教科書に誤記が現れる理由もあるかのようである。
その肝腎の現在ある証明は，特性関数 (Fourier 変換)，Lévy の定理の多次元版多次元正規分布等を使うもので，数学が専門でなければ初等的なものとは云えない。この論文では，このような高度な道具立てを一切使わず，上記定理を初等的に証明する。これは「統計学 II」の講義準備の際考え付いたものである。この程度に初等的なものならば，文系大学での統計学の講義でも使用できると思われる。
定理の証明　次の補題を使う。

補題 1 *(de Moivre-Laplace の中心極限定理)*
n が十分大きいとき，2 項分布 $\mathrm{B}(n,p)$ は，正規分布 $\mathrm{N}(np, np(1-p))$ で近似できる。即ち，$X \sim \mathrm{B}(n,p)$ ならば

$$\lim_{n\to\infty} \frac{X-np}{\sqrt{np(1-p)}} \sim \mathrm{N}(0,1)$$

である。

この補題を証明するには，de Moivre-Stirling の公式を使用するのが普通であるが，そうでない簡単な「証明」もある。 [8] [11]
r 個の確率変数 Z_i $(i=1,2,\cdots,r)$ を r 項分布:$\mathrm{Mul}(n;p_1,p_2,\cdots,p_r)$ に従っているとする。即ち，$Z_i = q_i$ $(q_i = 0,1,2,\cdots,n,\ q_1+\cdots+q_r = n,\ i = 1,2,\cdots,r)$ となる確率を

$$\mathrm{P}\{Z_1 = q_1,\cdots,Z_r = q_r\} = \frac{n!}{q_1!\cdots q_r!} {p_1}^{q_1}\cdots {p_r}^{q_r}$$

$p_1+p_2+\cdots+p_r = 1$，$0 \le p_1,p_2,\cdots,p_r \le 1$ とする。$(Z_j(j \neq i)$ の値は上記制限内でどのような値でも構わないとして) 周辺分布を考えれば，$Z_i \sim \mathrm{B}(n,p_i)$ である。これは

$$\begin{aligned}\mathrm{P}\{Z_i = q_i\} &= \sum_{\substack{q_j,j\neq i \\ \sum q_j = n-q_i}} \frac{n!}{q_1!\cdots q_i!\cdots q_n!} p_1^{q_1}\cdots p_i^{q_i}\cdots p_n^{q_n} \\ &= \frac{n!}{q_i!} p_i^{q_i} \sum_{\substack{q_j,j\neq i \\ \sum q_j = n-q_i}} \frac{1}{q_1!\cdots 1\cdots q_n!} p_1^{q_1}\cdots 1\cdots p_n^{q_n}\end{aligned}$$

$$= \frac{n!}{q_i!(n-q_i)!}{p_i}^{q_i} \sum_{q_j,\ j\neq i:\sum_{j\neq i} q_i = n-q_i} \frac{(n-q_i)!}{\prod_{j\neq i} q_j!} \prod_{j\neq i} {p_j}^{q_j}$$

$$= \frac{n!}{q_i!(n-q_i)!}{p_i}^{q_j}(1-p_i)^{n-q_i}$$

より明らかである。従って，2項分布の性質から Z_i の平均，分散はそれぞれ

$$\mathrm{E}[Z_i] = np_i,\ \mathrm{V}[Z_i] = np_i(1-p_i)$$

となる。これと同様な計算を $\mathrm{E}[Z_iZ_j]$ $(i \neq j)$ に行うと

$$\begin{aligned}
&\mathrm{E}[Z_iZ_j] \\
=& \sum_{\substack{q_i,q_j=0 \\ 0\leq q_i+q_j\leq n}}^{n} q_iq_j \frac{n!}{q_i!q_j!(n-q_i-q_j)!}{p_i}^{q_i}{p_j}^{q_j} \times \\
& \times \sum_{q_k,k\neq i,j:\sum_{k\neq i,j} q_k = n-q_i-q_j} \frac{(n-q_i-q_j)!}{\prod_{k\neq i,j} q_k!} \prod_{k\neq i,j} {p_k}^{q_k} \\
=& \sum_{\substack{q_i,q_j=0 \\ 0\leq q_i+q_j\leq n}}^{n} q_iq_j \frac{n!}{q_i!q_j!(n-q_i-q_j)!}{p_i}^{q_i}{p_j}^{q_j} \Big(\sum_{k\neq i,j} p_k\Big)^{n-q_i-q_j} \\
=& \sum_{\substack{q_i,q_j=0 \\ 0\leq q_i+q_j\leq n}}^{n} q_iq_j \frac{n!}{q_i!q_j!(n-q_i-q_j)!}{p_i}^{q_i}{p_j}^{q_j}(1-p_i-p_j)^{n-q_i-q_j}
\end{aligned}$$

となる。ここで次の補題を使う。

補題 2

$$\begin{aligned}
\sum_{\substack{q_i,q_j=0 \\ 0\leq q_i+q_j\leq n}}^{n} q_iq_j \frac{n!}{q_i!q_j!(n-q_i-q_j)!}x^{q_1}y^{q_2}z^{n-q_i-q_j} &= xy\frac{\partial^2}{\partial x\partial y}(x+y+z)^n \\
&= n(n-1)xy(x+y+z)^{n-2}
\end{aligned}$$

証明 明らか。□

この補題より，先の続きは

$$\begin{aligned}
\mathrm{E}[Z_iZ_j] &= n(n-1)p_ip_j\{p_i+p_j+(1-p_i-p_j)\}^{n-2} \\
&= n(n-1)p_ip_j
\end{aligned}$$

となる。従って $Z_1, Z_2, \cdots, Z_r$ は独立ではない ($\sum_{k=1}^{r} Z_k = n$)。ここで X_i を次で定義する。

$$X_i := \frac{Z_i - np_i}{\sqrt{np_i}} \ (i = 1, 2, \cdots, r).$$

X_i は

$$\mathrm{E}[X_i] = 0, \ \mathrm{V}[X_i] = \mathrm{E}[X_i{}^2] = \mathrm{E}[\frac{(Z_i - np_i)^2}{np_i}] = \frac{\mathrm{V}[Z_i]}{np_i} = 1 - p_i$$

を満たし，束縛条件：

$$\sum_{k=1}^{r} \sqrt{p_k} X_k = 0$$

を，これ1つのみ持つ。これが $X_1, \cdots, X_r$ が独立でない理由である。補題1より，

$$\lim_{n\to\infty} X_i \sim \mathrm{N}(0, 1 - p_i) \ (i = 1, 2, \cdots, r)$$

である。又，$\mathrm{E}[Z_i Z_j] = n(n-1)p_i p_j$ を使って $\mathrm{E}[X_i X_j]$ を計算すると

$$\begin{aligned} \mathrm{E}[X_i X_j] &= \mathrm{E}[\frac{Z_i - np_i}{\sqrt{np_i}} \cdot \frac{Z_j - np_j}{\sqrt{np_j}}] \\ &= \frac{1}{n\sqrt{p_i p_j}} \{n(n-1)p_i p_j - np_i np_j - np_j np_i + n^2 p_i p_j\} \\ &= -\sqrt{p_i p_j} \end{aligned}$$

である。

次に先の $X_1, \cdots, X_r$ を直交変換 $A = (a_{ij})$ with $a_{rk} = \sqrt{p_k}, (k = 1, 2, \cdots, r)$, ${}^tAA = \mathrm{I}$ で $X_1', \cdots, X_r'$ に次で変換する：

$$X_j' = \sum_{k=1}^{r} a_{jk} X_k.$$

$j = 1, 2, \cdots, r-1$ に対して，$\sum_{k=1}^{r} \sqrt{p_k} X_k = 0$ を使って X_j' を計算すると

$$X_j' = \sum_{k=1}^{r-1} (a_{jk} - a_{jr} \frac{\sqrt{p_k}}{\sqrt{p_r}}) X_k$$

となり，

$$\mathrm{E}[X_j'] = \sum_{k=1}^{r} a_{jk} \mathrm{E}[X_k] = \sum_{k=1}^{r} a_{jk} \cdot 0 = 0 \ (j = 1, 2, \cdots, r),$$

$$
\begin{aligned}
\mathrm{V}[X_j'] &= \sum_{k=1}^{r} a_{jk}^2 \mathrm{V}[X_k] + \sum_{k\neq l} a_{jk}a_{jl}\mathrm{Cov}[X_k X_l] \\
&= \sum_{k=1}^{r} a_{jk}^2 \mathrm{V}[X_k] + \sum_{k\neq l} a_{jk}a_{jl}\mathrm{E}[X_k X_l] \\
&= \sum_{k=1}^{r} a_{jk}^2 (1-p_k) - \sum_{k\neq l} a_{jk}a_{jl}\sqrt{p_k p_l} \\
&= \sum_{k=1}^{r} a_{jk}^2 - \sum_{k=1}^{r} a_{jk}^2 p_k - \sum_{k\neq l} a_{jk}a_{jl}\sqrt{p_k p_l} \\
&= 1 - (\sum_{k=1}^{r} a_{jk}\sqrt{p_k})^2 \\
&= 1 - (\sum_{k=1}^{r} a_{jk}a_{rk})^2 = 1 - 0 = 1 \text{ for } j = 1, 2, \cdots, r-1
\end{aligned}
$$

である。ここで a_{jk} が直交行列の成分であることを使った。
X_r' については，

$$
\begin{aligned}
V[X_r'] &= \sum_{k=1}^{r} a_{rk}^2 V[X_k] + \sum_{k\neq l} a_{rk}a_{rl}\mathrm{Cov}[X_k X_l] \\
&= \sum_{k=1}^{r} a_{rk}^2 (1-p_k) - \sum_{k\neq l} a_{rk}a_{rl}\sqrt{p_k p_l} \\
&= \sum_{k=1}^{r} a_{rk}^2 - \sum_{k=1}^{r} a_{rk}^2 p_k - \sum_{k\neq l} a_{rk}a_{rl}\sqrt{p_k p_l} \\
&= 1 - (\sum_{k=1}^{r} a_{rk}\sqrt{p_k})^2 \\
&= 1 - (\sum_{k=1}^{r} a_{rk}a_{rk})^2 = 1 - 1 = 0,
\end{aligned}
$$

従って，$n \to \infty$ で $X_r', X_i'(i = 1, \cdots, r-1)$ は正規分布をしているから確率1で $\lim_{n\to\infty} X_r' = 0$ となる。

$$
X_j' = \sum_{k=1}^{r-1} (a_{jk} - a_{jr}\frac{\sqrt{p_k}}{\sqrt{p_r}}) X_k
$$

により，$n \to \infty$ で X_j' $(j = 1, \cdots, r-1)$ は正規分布の再生性から，正規分布に従うが，$\mathrm{E}[X_j'] = 0$，$\mathrm{V}[X_j'] = 1$ for $j = 1, 2, \cdots, r-1$ であったことから，更に

$$
\lim_{n\to\infty} X_j' \sim \mathrm{N}(0,1) \quad \text{for } j = 1, 2, \cdots, r-1
$$

となる。
又，$i \neq j$に対して

$$\begin{aligned}
\mathrm{E}[X_i' X_j'] &= \sum_{k=1}^{r}\sum_{l=1}^{r} a_{ik}a_{jl}\mathrm{E}[X_k X_l] \\
&= -\sum_{k=1}^{r}\sum_{l=1}^{r} a_{ik}a_{jl}\sqrt{p_k p_l} \\
&= -\sum_{k=1}^{r}\sum_{l=1}^{r} a_{ik}a_{jl}a_{rk}a_{rl} \\
&= -\sum_{k=1}^{r} a_{ik}a_{rk}\sum_{l=1}^{r} a_{jl}a_{rl} \quad (i,j \neq r) \\
&= -0 \cdot 0 = 0
\end{aligned}$$

となるので

補題 3 *X, Yが正規分布に従う場合，X, Yが無相関である事と，独立である事は同値である。即ち*

$$\textit{In case of } X \sim \mathrm{N}(\mu_X, \sigma_X^2),\ Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2),$$
$$X, Y \textit{ are uncorrelated} \Leftrightarrow X, Y \textit{ are independent.}$$

より，$X_1', \cdots, X_{r-1}'$は独立であり，$n \to \infty$のとき，以下の補題 4 により

$$X_1'^2 + X_2'^2 + \cdots + X_{r-1}'^2 \sim \chi^2(r-1)$$

となる。従って$A = (a_{ij})$が直交変換であることと$X_r' = 0$ a.s. であることから

$$X_1^2 + X_2^2 + \cdots + X_r^2 = X_1'^2 + X_2'^2 + \cdots + X_{r-1}'^2 \text{ with probability 1}$$

となる。これより

$$X_1^2 + X_2^2 + \cdots + X_r^2 = \sum_{i=1}^{r} \frac{(Z_i - np_i)^2}{np_i}$$

が，$n \to \infty$で自由度$(r-1)$のχ^2分布に従うこと，
即ち$n \to \infty$のとき

$$\begin{aligned}
\sum_{i=1}^{r} \frac{(Z_i - np_i)^2}{np_i} &= X_1^2 + X_2^2 + \cdots + X_r^2 \\
&= X_1'^2 + X_2'^2 + \cdots + X_{r-1}'^2 \sim \chi^2(r-1) \\
&\quad \text{with probability 1}
\end{aligned}$$

が証明された。ここで，

補題 4 $X_i \sim \mathrm{N}(0,1)$ $(i=1,2,\cdots,n)$ が独立ならば，

$$X_1{}^2 + X_2{}^2 + \cdots + X_n{}^2 \sim \chi^2(n)$$

即ち，$X_1{}^2 + X_2{}^2 + \cdots + X_n{}^2$ は自由度 n の χ^2 分布に従う。

注 補題4を自由度 n の χ^2 分布の定義 definition としても良い。
を使った。□

参考文献

[1] 儀我真理子 Giga,M.:『確率・統計の基礎』ムイスリ出版，2014, (vii+213)pp.. *Foundation of Probability and Statistics*, Muicuri-Shuppan, Tokyo, 2014.

[2] 林周二 Hayashi,Sh.『統計学講義第2版』丸善, *The Lecture on Staistics*, 2nd ed., Maruzen, Tokyo, 1973, (x+378)pp..

[3] Helmert,F.R.: *Z. Math. und Phys.*, **214**(1876), 192-218.

[4] Helmert,F.R.: *Astron. Nachr.*, **88**(1876), 113-132.

[5] 河田敬義，丸山文行 Kawada,Y., Maruyama,F.『数理統計』裳華房, *Mathematical Statistics*, Shouka-Bo, Tokyo, 1951, (vi+230)pp..

[6] 松井敬 Matsui,T.『統計的推測』共立出版，*Statistical Inference*, Kyouritsu-Shuppan, Tokyo, 2012, (viii+207)pp..

[7] von Mises,R. (edited and complemented by H. Geiringer.): *Mathematical Theory of Probability and Statistics*, Academic Press, 1964, (xiv+694)pp..

[8] Neyman,J.: *First Course in Probability and Statistics*, Henry Holt and Company, 1950.
邦訳：砂田吉一『ネイマン統計学』白桃書房, Tokyo, 1978, (ix+298)pp..

[9] 小川潤次郎 Ogawa,J.『近代数理統計学序説』恵文堂，*Introduction to Modern Mathematical Statistics*, Keibun-Dou, Osaka, 1954, (iii+396)pp..

[10] Pearson,K.: On the Criterion that a given system of deviations is such that it can be reasonably supposed to have arisen from random sampling, *Phil. Mag.*, **50**(1900), p.157.

[11] 田島一郎 Tajima,I.: 二項分布から正規分布へ From Binomical Distribution to Normal Distribution, 数学セミナー, *Seminar for Mathematics*, 1974年3月号, March 1974, p.24-26, 日本評論社, Tokyo.

[12] 吉田伸生 Yoshida,N.『確率の基礎から統計へ』*From Foundation of Probability to Statistics*, 遊星社, Yuusei-sha, Tokyo, 2012, (o+189)pp..

(receiced 23 August 2020.)