数理解析研究所講究録
第 777 巻 1992 年 1-11

一様分布の特徴付けに基づく適合度検定

大阪大学 基礎工学部　橋本隆正(Takamasa Hashimoto)
大阪大学 教養部　白旗慎吾(Shingo Shirahata)

§1. はじめに

$Z_1, \cdots, Z_n$ は連続な分布関数 $F(z)$ に従うランダムサンプルとし, 次の仮説を検定したいとする.

$$H_0 : F(z) = F_0(z)$$

ここで $F_0(z)$ は完全に specify される.
対立仮説 H_1 は H_0 でないものを考える. このとき変換 $X_i = F_0(Z_i)$ により問題は X_i が (0,1) 上の一様分布に従うか否かを検定するものになる. そこで我々は一様分布の特徴付けに注目して上の検定を行なう.

§2. 一様分布の特徴付け

Papathanasiou(1990) は次の様な一様分布の特徴付けを行なった.

定理 2.1(Papathanasiou)
$X_{(1)}, X_{(2)}$ を平均 μ, 分散 σ^2 の絶対連続な分布 F (密度 f) に従うサイズ 2 の標本からの順序統計量とする. この時

$$\mathrm{Cov}(X_{(1)}, X_{(2)}) \leq \frac{1}{3}\sigma^2$$

を得る. さらに等号が成り立つ必要十分条件は, F が一様分布であることである.

(証明)

$$\mathrm{E}(X_{(1)}X_{(2)}) = \mathrm{E}^2(X).$$

ここで X は 確率密度関数 f をもつ確率変数である. 故に,

$$\mathrm{Cov}(X_{(1)}, X_{(2)}) = \left(\int_{-\infty}^{\infty} xf(x)dx\right)^2 - 4\int_{-\infty}^{\infty} xf(x)F(x)dx \int_{-\infty}^{\infty} xf(x)(1-F(x))dx.$$

最初の積分の項に $F(x)+(1-F(x))$ を掛けて式を整理すると

$$\mathrm{Cov}(X_{(1)}, X_{(2)}) = \left(\int_{-\infty}^{\infty} x[2F(x)-1]f(x)dx\right)^2.$$

さて, $\int_{-\infty}^{\infty}[2F(x)-1]f(x)dx=0$ より

$$\mathrm{Cov}(X_{(1)},X_{(2)})=\left(\int_{-\infty}^{\infty}(x-\mu)[2F(x)-1]f(x)dx\right)^2$$

シュワルツの不等式を用いて

$$\begin{aligned}\mathrm{Cov}(X_{(1)},X_{(2)}) &\le \int_{-\infty}^{\infty}(x-\mu)^2f(x)dx\int_{-\infty}^{\infty}(2F(x)-1)^2f(x)dx\\ &=\frac{1}{3}\sigma^2.\end{aligned}$$

等号成立は, $2F(x)-1=c(x-\mu)$ の場合に限る. これは一様分布の分布関数.

(証明終り)

§3. 特徴付けに基づく検定統計量と帰無仮説での性質

定理 2.1 で確率変数 X が一様分布に従う必要十分条件は

$$\frac{1}{3}\sigma^2-\mathrm{Cov}(X_{(1)},X_{(2)})=0$$

であることが分った. そこで我々は $\frac{1}{3}\sigma^2-\mathrm{Cov}(X_{(1)},X_{(2)})$ の推定量 C_n を考え $C_n\approx 0$ なら X は一様分布に従っていると思いたい. そこで $\frac{1}{3}\sigma^2-\mathrm{Cov}(X_{(1)},X_{(2)})$ の推定量として次のものを提案する.

$$C_n(X_1,\cdots,X_n)=\frac{1}{\binom{n}{4}}\sum_{i<j<k<l}h(X_i,X_j,X_k,X_l).$$

ここで

$$\begin{aligned}h(X_i,X_j,X_k,X_l)=\frac{1}{36}\{&(X_i-X_j)^2+(X_i-X_l)^2+(X_i-X_k)^2\\ &+(X_j-X_k)^2+(X_j-X_l)^2+(X_k-X_l)^2\}\\ -\frac{1}{6}[&\{\max(X_i,X_j)-\max(X_k,X_l)\}\{\min(X_i,X_j)-\min(X_k,X_l)\}\\ &+\{\max(X_i,X_k)-\max(X_j,X_l)\}\{\min(X_i,X_k)-\min(X_j,X_l)\}\\ &+\{\max(X_i,X_l)-\max(X_j,X_k)\}\{\min(X_i,X_l)-\min(X_j,X_k)\}].\end{aligned}$$

今, C_n は degree が 4 の U-統計量であり, $\mathrm{E}[C_n(X_1,\cdots,X_n)]=\frac{1}{3}\sigma^2-\mathrm{Cov}(X_{(1)},X_{(2)})$ となる. 即ち C_n は $\frac{1}{3}\sigma^2-\mathrm{Cov}(X_{(1)},X_{(2)})$ の不偏推定量である.

U-統計量は Hoeffding(1948) により非常に緩い条件のもとで漸近正規性が示された. 又 UMVU 推定量でもあり非常によい推定量であることが分かっている.

我々が提案した検定統計量は漸近正規性のない退化したU-統計量である事が後に分かる. 退化したU-統計量の漸近挙動は, Gregory(1977) により degree が 2 の場合に求められた. Eagleson(1979) はさらに高次の degree の場合や 2 標本の場合等の退化したU-統計量の漸近挙動を求めた. ここで Eagleson の結果をまとめておく.

(定義)

カーネル関数 $\Phi(X_1,\cdots,X_r)$ をもつ U-統計量は, 一般性を失うことなく $\mathrm{E}(\Phi(X_1,\cdots,X_r))=0$ とする. この時もし

$$\mathrm{E}(\Phi(X_1,....,X_r)|X_1)=0 \qquad \text{a. s.}$$

なら U-統計量とそのカーネル Φ は, 一次の退化 (first-order degeneracy) と呼ぶ. さらに,

$$\mathrm{E}(\Phi(X_1,\cdots,X_r)|X_1,X_2)=0 \qquad \text{a. s.}$$

なら二次の退化 (second-order degeneracy) と呼ぶ.

以下, 三次の退化, 四次の退化, $\cdots$ も同様に定義する.

定理 3.1 (Eagleson)

U_n は, degree 2 の退化した U-統計量とし, そのカーネル関数 h は二乗可積分とする. さらに一般性を失うことなく, $\mathrm{E}(h)=0$ とする. その時

$$nU_n \to \sum_{k=1}^{\infty}\lambda_k({Z_k}^2-1) \qquad \text{in law.}$$

ここで, $\{\lambda_k\}$ は次の積分方程式の固有値である.

$$\int \psi_k(x)h(x,y)dF(x)=\lambda_k\psi_k(y).$$

$\{\psi_k(x)\}$ は完全正規直交系.

$\{Z_k\}: N(0,1)$ の独立な確率変数の列.

(略証)

$h\in L^2$ かつ対称より次の固有関数展開が可能である.

$$h(x,y)=\sum_{k=1}^{\infty}\lambda_k\psi_k(x)\psi_k(y).$$

ここで

$$\int \psi_k(x)h(x,y)dF(x) = \lambda_k\psi_k(y)$$

かつ

$$\sum_{k=1}^{\infty}\lambda_k^2 < \infty \quad , \quad \{\psi_k\}: \text{完全正規直交系}.$$

今

$$h_N(x,y) = \sum_{k=1}^{N}\lambda_k\psi_k(x)\psi_k(y)$$

とおき

$$U_n^{(N)} = \frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n} h_N(X_i,X_j)$$

とおく. このとき

$$nU_n^{(N)} \to nU_n \qquad \text{in p} \quad (N\to\infty)$$
$$\sum_{k=1}^{N}\lambda_k(Z_k^2-1) \to \sum_{k=1}^{\infty}\lambda_k(Z_k^2-1) \qquad \text{in p} \quad (N\to\infty)$$
$$nU_n^{(N)} \to \sum_{k=1}^{N}\lambda_k(Z_k^2-1) \qquad \text{in law} \quad (n\to\infty)$$

が示せる. 以上より

$$nU_n \to \sum_{k=1}^{\infty}\lambda_k(Z_k^2-1) \qquad \text{in law.}$$

(証明終り)

定理 3.2 (Eagleson)

U_n は, degree 4 の退化した U-統計量とし, そのカーネル関数 h は二乗可積分とする. その時

$$nU_n \to Z \qquad \text{in law.}$$

ここで,

$$Z = \sum_{0<k,l}\lambda_{kl00}Z_kZ_l + \sum_{0<k,m}\lambda_{k0m0}Z_kZ_m + \sum_{0<k,n}\lambda_{k00n}Z_kZ_n$$
$$+ \sum_{0<l,m}\lambda_{0lm0}Z_lZ_m + \sum_{0<l,n}\lambda_{0l0n}Z_lZ_n + \sum_{0<m,n}\lambda_{00mn}Z_mZ_n.$$

さらに

$$\sum_{k,l,m,n} \lambda_{klmn}^2 < \infty \quad , \quad \{Z_k\} : \text{独立なN(0,1)確率変数の列.}$$

(略証)

$h \in L^2$ で一次の退化より任意の正規直交系 $\{\psi(\cdot)\}$ で h は次の様にかける.

$$\begin{aligned} h(x,y,z,w) = & \sum_{0<k,l,m,n} \lambda_{klmn}\psi_k(x)\psi_l(y)\psi_m(z)\psi_n(w) \\ & + \sum_{0<k,l,m} \lambda_{klm0}\psi_k(x)\psi_l(y)\psi_m(z) + \sum_{0<k,m,n} + \sum_{0<k,l,n} + \sum_{0<l,m,n} \\ & + \sum_{0<k,l} \lambda_{kl00}\psi_k(x)\psi_l(y) + \sum_{0<k,m} + \sum_{0<k,n} + \sum_{0<l,m} + \sum_{0<l,n} + \sum_{0<m,n} . \end{aligned}$$

定理 3.1 と同様に $h_N(x,y,z,w)$ と $U_n^{(N)}$ を考えて

$$nU_n \to Z \qquad \text{in law.}$$

(証明終り)

さて定理 3.2 より U_nの漸近挙動の確率表現は求まったが, $\{\lambda_k\}$ を探すことは不可能に等しいだろう. そこで我々はさらに近似を進める.

補題 3.3

定理 3.2 の条件の下で

$$nU_n \to \sum_{k=1}^{\infty} 6\eta_k(Z_k^2 - 1) \qquad \text{in law.}$$

ここで $\{\eta_k\}$ は以下の積分方程式の固有値である.

$$\int \phi_k(x)\Psi(x,y)dF(x) = \eta_k\phi_k(y).$$

又,

$$\{\phi_k(x)\} : \text{完全正規直交系.}$$

$$\Psi(x,y) = \mathrm{E}(h(X,Y,Z,W)|X=x, Y=y).$$

(証明)

定理 3.2 より $h(x,y,z,w)$ は任意の完全正規直交系 $\{\psi(\cdot)\}$ を用いて展開できる. このとき $\psi(\cdot)$ の 3 次形式と 4 次形式は漸近的に無視できる. 従って漸近的にはカーネル関数 h は

$$h(x,y,z,w)=\sum_{0<k,l}\lambda_{kl00}\psi_k(x)\psi_l(y)+\sum_{0<k,m}+\sum_{0<k,n}+\sum_{0<l,m}+\sum_{0<l,n}+\sum_{0<m,n}$$

とみてよい. 又,

$$\mathrm{E}(h(x,y,z,w)|X=x,Y=y)=\sum_{0<k,l}\psi_k(x)\psi_l(y)$$

に注意して, これを $\Psi(x,y)$ とおくと $\Psi\in L^2$ かつ対称より

$$\Psi(x,y)=\sum_{k=1}^{\infty}\eta_k\phi_k(x)\phi_k(y)$$

とかける. ここで η_k, ϕ_k は次の積分方程式をみたす.

$$\int\phi_k(x)\Psi(x,y)dF(x)=\eta_k\phi_k(y).$$

あとは h の対称性と定理 3.1 により示せる.

(証明終り)

次にこれらの結果を使って我々の提案した統計量の帰無仮説の下での漸近挙動を調べる.

命題 3.4

C_n は次の性質をもつ.

(1) C_nは(1次の)退化したU－統計量である.

(2) $$\mathrm{E}(h(X,Y,Z,W)|X=x,Y=y)$$
$$=\frac{1}{6}x^2y+\frac{1}{6}xy^2+\frac{1}{18}\min(x,y)-\frac{2}{9}xy-\frac{1}{6}x^2y^2.$$

(3) $$nC_n\to\sum_{k=1}^{\infty}6\lambda_k(Z_k^2-1)\qquad\text{in law.}$$

但し $\{\lambda_k\}$ は

$$\sin\frac{1}{6\sqrt{2\lambda}}=0\quad\text{又は,}\quad\tan\frac{1}{6\sqrt{2\lambda}}=\frac{1}{6\sqrt{2\lambda}}$$

の解で $\lambda_1>\lambda_2>\cdots>0$ とする.

(証明)

$\phi(x,y)=\frac{1}{6}x^2y+\frac{1}{6}xy^2+\frac{1}{18}\min(x,y)-\frac{2}{9}xy-\frac{1}{6}x^2y^2$ とおき $\int_0^1\psi_k(x)\phi(x,y)dx=\lambda_k\psi_k(y)$ を解く. これは, 微分方程式

$$\lambda_k\psi_k'''(y)+\frac{1}{18}\psi_k'(y)=0$$

を $\psi_k(0)=\psi_k(1)=0,\quad \int_0^1\psi_k(x)dx=0,\quad \int_0^1\psi_k^2(x)dx=1$ という条件のもとで解くことになる. これを解いて補題 3.3 を使って (3) を得る.

(証明終り)

次に棄却域などを決める為に C_n の帰無仮説の下でのパーセント点を求める必要がある. 命題 3.4(3) より C_n の漸近分布はカイ二乗確率変数の重み付き和である. このタイプの分布関数は, 白旗 (1988) により求められてるのでここではその結果をまとめておく.

定理 3.5(白旗)

$X_1,X_2,\cdots$ は独立に自由度 1 のカイ二乗分布に従うとする. 又,$c_1>c_2>\cdots>0$ とおく. この時,

$$X=\sum_{i=1}^{\infty}\frac{c_i}{2}X_i$$

の分布関数は,

$$F(x)=\lim_{n\to\infty}\sum_{\substack{j=1\\ j:odd}}^{n}\frac{(-1)^{\frac{j+1}{2}}}{\pi}\left(\frac{1}{c_{j+1}}-\frac{1}{c_j}\right)\int_0^1\frac{e^{-h_j(t)x}-1}{h_j(t)}\prod_{k=1}^{n}|1-c_kh_j(t)|^{-\frac{1}{2}}dt$$

となる. ここで $h_j(t)=\dfrac{t}{c_j}+\dfrac{1-t}{c_{j+1}}\quad(0<t<1)$

統計量 C_n のパーセント点を求めるには定理 3.5 より $\sum_{k=1}^{\infty}6\lambda_kZ_k^2$ の分布関数を求めそのパーセント点を調べそこから location のずれ $\sum_{k=1}^{\infty}6\lambda_k$ を引けば求まる. しかし今, 定理 3.5 の積分をうまく解くことは不可能である. 従って数値積分によりパーセント点を求める.

まず固有値 λ を求める事を考える. $\sin\frac{1}{6\sqrt{2\lambda}}=0$ からはうまく求まるが $\tan\frac{1}{6\sqrt{2\lambda}}=\frac{1}{6\sqrt{2\lambda}}$ からはうまく求まらないのでニュートン法で数値的に求める.

又数値積分は白旗のアルゴリズムに従って求める.

$n\to\infty$ の操作は n を適当に動かして決めて値が落着けば収束したものと見なす. 以下 C_n の漸近分布のパーセント点の表をあげる. (但し実際の値を 100 倍してある.)

$N\backslash\%$	90	80	70	60	50	40
4	-1.255	-1.047	-.845	-.629	-.385	-.093
8	-1.286	-1.057	-.844	-.622	-.373	-.079
16	-1.292	-1.059	-.844	-.620	-.371	-.076
32	-1.293	-1.059	-.844	-.620	-.371	-.076
64	-1.293	-1.059	-.844	-.620	-.371	-.076
$N\backslash\%$	30	20	10	8	5	3
4	.280	.808	1.733	2.037	2.689	3.414
8	.295	.824	1.748	2.052	2.704	3.428
16	.298	.827	1.751	2.054	2.707	3.431
32	.299	.828	1.752	2.056	2.707	3.431
64	.299	.828	1.752	2.056	2.707	3.431
$N\backslash\%$	2	1				
4	3.999	5.019				
8	4.009	5.019				
16	4.016	5.035				
32	4.016	5.036				
64	4.017	5.036				

次に n を 20 から 100 まで 5 刻みで各々10000 回繰返して nC_n を計算してそのパーセント点を求める. そして $\frac{1}{n}$ で補間を行ない縦軸にパーセント点をプロットする. 今 $n \to \infty$ 即ち縦軸上の値は前表より求まっている. そこで $y_n = \frac{a}{n} + b_0 + \epsilon_n$ とおいて線形回帰により a を推定して smoothing を行なう. 以下 5% 点と 10% 点の推定値について表をあげる.

Sample Size	5% Point	10% Point
20	0.0299	0.0196
25	0.0293	0.0192
30	0.0289	0.0189
35	0.0287	0.0187
40	0.0285	0.0186
45	0.0283	0.0185
50	0.0282	0.0184
55	0.0281	0.0183
60	0.0280	0.0182
65	0.0279	0.0182
70	0.0279	0.0181
75	0.0278	0.0181
80	0.0278	0.0180
85	0.0277	0.0180
90	0.0277	0.0180
95	0.0277	0.0180
100	0.0276	0.0180
∞	0.0271	0.0175

§4. パワーの計算

ここでは, パワーの計算と他の検定統計量との比較を行なう.

Stephens(1974) は一様性の検定を行なう時, 次の 3 つの対立仮説を考えた.(但し $k > 1$)

Alternative ; A

$$F(z) = 1 - (1-z)^k \qquad 0 \le z \le 1$$

Alternative ; B

$$F(z) = 2^{k-1} z^k \qquad 0 \le z \le 0.5$$
$$F(z) = 1 - 2^{k-1}(1-z)^k \qquad 0.5 \le z \le 1$$

Alternative ; C

$$F(z) = 0.5 - 2^{k-1}(0.5-z)^k \qquad 0 \le z \le 0.5$$
$$F(z) = 0.5 + 2^{k-1}(z-0.5)^k \qquad 0.5 \le z \le 1$$

我々は $k = 2$ の時,A,B,C でのパワーをシミュレーションによって求めた. 棄却域は上側 5 % ,10 % とする. さらに他の検定統計量とのパワーの比較を試みる. C_n 以外の統計量に関しては $n = 20, 40$ のときに帰無仮説のもとで各々10000 回繰返して各統計量を計算し,Stephens の有限修正に従い棄却域をもとめる. そしてそのもとでパワーを計算する. C_n の $n = 20, 40$ のときの上側 5 % 点 10 % 点は §3 で求めてあるのでそれを棄却域として用いる. 以下に具体的な検定統計量の形とシミュレーション結果をあげる. 但し以下では $Z_1 < Z_2 < \cdots < Z_n$ とする.

Simulation Results.

K-S : Kolmogorov-Smirnov Statistics.

$$D = \max\left(\max_{1 \le i \le n} \left[\frac{i}{n} - Z_i \right] , \max_{1 \le i \le n} \left[Z_i - \frac{i-1}{n} \right] \right)$$

CvM : Cramer-von Mises.

$$W^2 = \sum_{i=1}^{n} \left[Z_i - \frac{2i-1}{2n} \right]^2 + \frac{1}{12n}$$

Kui : Kuiper.

$$V = \max_{1 \le i \le n} \left[\frac{i}{n} - Z_i \right] + \max_{1 \le i \le n} \left[Z_i - \frac{i-1}{n} \right]$$

Wat : Watson.

$$U^2 = W^2 - n \left(\frac{1}{n} \sum_{i=1}^{n} Z_i - \frac{1}{2} \right)^2$$

A-D : Anderson-Darling.

$$A^2 = -\frac{1}{n}\left\{\sum_{i=1}^{n}(2i-1)\left[\ln Z_i + \ln(1-Z_{n+1-i})\right]\right\} - n$$

C_n : Our Statistics.

Alternative ; A

N	%	K-S	CvM	Kui	Wat	A-D	C_n
20	5	.6506	.7142	.4386	.4306	.8102	.4730
	10	.7594	.8032	.5570	.5462	.8728	.5852
40	5	.9180	.9466	.7800	.7492	.9744	.7840
	10	.9538	.9690	.8564	.8376	.9854	.8578

Alternative ; B

N	%	K-S	CvM	Kui	Wat	A-D	C_n
20	5	.2762	.2648	.5628	.5962	.6574	.2928
	10	.4174	.4340	.6706	.7064	.7766	.3880
40	5	.4978	.5372	.8644	.8986	.9078	.5362
	10	.6754	.7370	.9214	.9424	.9578	.6484

Alternative ; C

N	%	K-S	CvM	Kui	Wat	A-D	C_n
20	5	.2146	.1232	.5854	.6206	.1024	.2714
	10	.3774	.3168	.6986	.7278	.2750	.4842
40	5	.4822	.4520	.8710	.9012	.4314	.7066
	10	.6684	.6918	.9208	.9406	.6534	.8542

上で比べた 6 つの統計量のうち C_n 以外は直観的に作られたものだと思われるので今回我々は一様分布の特徴付けに基づいて統計量を構成してみた. パワーのシミュレーション結果はあまり良いとは言い難いが他のどの検定統計量と比べても一様に悪いわけではないのでまずまずの結果がでたと思う.

参考文献

Eagleson, G. K., *Orthogonal expansions and U-statistics*, Austral. J. Statist. **21** (1979), 221–237.

Gregory, G. G., *Large sample theory for U-statistics and tests of fit*, Ann. Statist. **5** (1977), 110–123.

Hoeffding, W., *A class of statistics with asymptotically normal distribution*, Ann. Math. Statist. **19** (1948), 293–325.

Papathanasiou, V., *Some characterizations of distributions based on order statistics*, Statist. & Prob. Letters **9** (1990), 145–147.

白旗慎吾, カイ二乗確率変数の重み付き和の分布関数の計算, 計算機統計学 **1** (1988), 37–44.

Stephens, M. A., *EDF statistics for goodness of fit and some comparisons*, J. Amer. Statist. Assoc. **69** (1974), 730–737.