数理解析研究所講究録
第 498 巻 1983 年 172-181

グラフィカルな立場からの適合度検定法

岡山大学　脇本和昌

1．はじめに

生成された乱数列を $u_1, u_2, \cdots\cdots, u_N$ とし、はじめから大きさ $l_1, l_2, \cdots\cdots, l_n$ の n 個の組にわけて，くり返し連続したシミュレーションに使用する場合，各組ごとに算出した n 個の検定量の値が，その検定量のもつ確率分布に従っているかどうかを検定することが主たる目的である。

いま、母集団の分布関数を $F(x)$，ある定められた連続分布関数を $F_0(x)$ とする。この母集団から大きさ n の独立な random sample を $X_1, X_2, \cdots\cdots, X_n$ とするとき

$$H_0: \ F(x) = F_0(x)$$

$$H_1: \ F(x) \neq F_0(x)$$

を検定する問題に帰着される。このように、母集団の分布 $F(x)$ に何も仮定せず，仮説 H_0 を検定する場合には普通はつぎに示す Kolmogorov-Smirnov 検定（略して K-S 検定と書くことにする）が用いられる。

$X_1, X_2, \cdots\cdots, X_n$を大きさの順に並べたものを$X_{(1)} \leqq X_{(2)} \leqq \cdots \leqq X_{(n)}$とする。ここで標本累積分布関数$F_n(x)$をつぎのようにつくる。

$$F_n(x) = \begin{cases} 0, & x < X_{(1)} \\ \dfrac{i}{n}, & X_{(i)} \leqq x < X_{(i+1)}, \quad i = 1, 2, \cdots, n-1 \\ 1, & X_{(n)} \leqq x \end{cases}$$

このとき、検定量として

$$D_n = \max\left(F_0(x) - F_n(x)\right) \quad (\text{片側検定})$$

$$d_n = \max\left|F_0(x) - F_n(x)\right| \quad (\text{両側検定})$$

を用いるのがK-S検定である。

D_n, d_nの確率分布は$n = 100$までは正確に求められて表にされており、$n > 100$についてはつぎのように漸近分布とパーセント点が求められている。

$P(D_n \leqq \varepsilon) = P_n(\varepsilon) = 1 - \alpha$　なるεの値ε_αは

$\varepsilon_\alpha \fallingdotseq \sqrt{\dfrac{1}{2n} \log \dfrac{1}{\alpha}}$　となり、

$P_n(\varepsilon) \fallingdotseq 1 - e^{-2n\varepsilon^2}$　となることが知られている。

さて、ここではH_0を検定するために、linked line chartによって囲まれる面積を検定量とする方法を提案し、D_nとの

比較をグラフィカルな立場からおこなう。

2、提案する検定法

$Z_i=F_0(X_i)$，$Z_{(i)}=F_0(X_{(i)})$，$i=1,2,\cdots,n$ とおき，原点Oを始点としてx軸となす角が $Z_{(1)}\pi$，$\frac{\pi}{n+1}$ ラジアンである長さが $1/\binom{n}{2}$ の線分 $\overline{OP_1}$，$\overline{OQ_1}$ を引き、順次長さが $1/\binom{n}{2}$ の線分 $\overline{P_{i-1}P_i}$，$\overline{Q_{i-1}Q_i}$ をx軸となす角が $Z_{(i)}\pi$，$\frac{i\pi}{n+1}$ ラジアンであるように引き，順次結んでいく。つぎに点 $P_1,\cdots,P_{n-1}$，$Q_1,\cdots,Q_{n-1}$ をそれぞれ $\overline{OP_n}$，$\overline{OQ_n}$ の中点M，M'に対称に移した点を P_1'，$\cdots$，P_{n-1}'，$Q_1',\cdots,Q_{n-1}'$ とし、それらを順次結んで図1のような多角形Pと多角形Qをつくる。この両多角形を Linked Line Chart（LLC−n）といい、多角形Pの多角形Qからのずれ具合によって検定するものである。

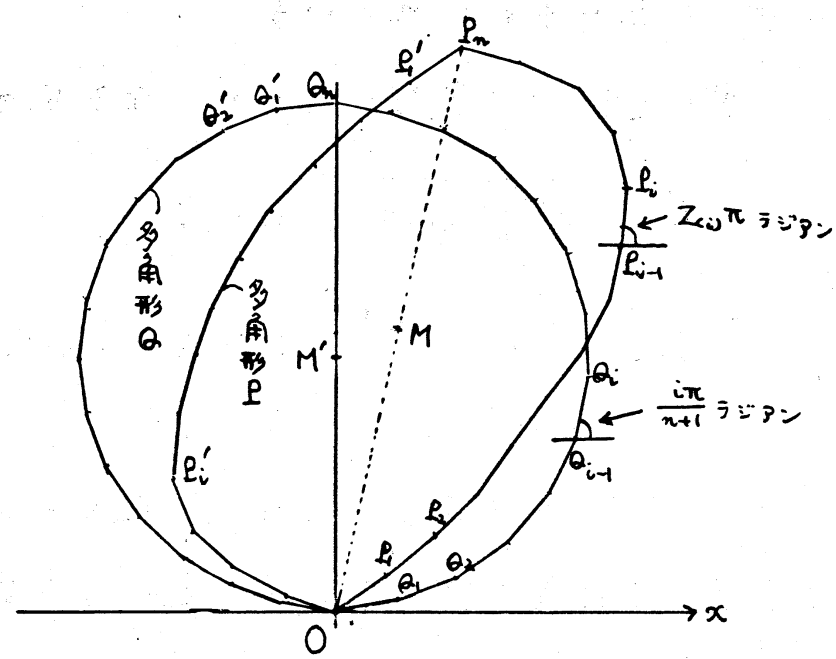


図1.　LLC−n

多角形P，Qの面積をそれぞれU_n，S_nとすると

(2, 1) $$U_n=\frac{1}{\binom{n}{2}}\sum_{1\leqq i<j\leqq n}\sin|Z_i-Z_j|\pi \quad (\text{U統計量})$$

(2, 2) $$S_n=\frac{1}{\binom{n}{2}}\sum_{1\leqq i<j\leqq n}\sin\frac{j-i}{n+1}\pi$$

となる。

U_nの平均と分散はつぎのようになる。ただし、Z_1，Z_2，Z_3，Z_4は［0，1］上の一様分布をもつ互いに独立な確率変数とする。

(2, 3) $$\begin{aligned}E(U_n)&=\frac{1}{\binom{n}{2}}\sum_{1\leqq i<j\leqq 1}E(\sin|Z_i-Z_j|\pi)\\&=E(\sin|Z_1-Z_2|\pi)\\&=\int_0^1\int_0^1\sin|z_1-z_2|\pi\,dz_1dz_2 \quad =\frac{2}{\pi}\end{aligned}$$

(2, 4) $$\begin{aligned}V(U_n)&=E(U_n-E(U_n))^2=E[U_n^2]-(E(U_n))^2\\&=\frac{1}{\binom{n}{2}^2}\Big[{}_nP_3\cdot E\{\sin|Z_1-Z_2|\pi\sin|Z_1-Z_3|\pi\}\\&\quad+\binom{n}{2}\cdot E\{(\sin|Z_1-Z_2|\pi)^2\}\\&\quad+\binom{n}{2}\binom{n-2}{2}\cdot E(\sin|Z_1-Z_2|\pi\sin|Z_3-Z_4|\pi)\Big]\\&\quad-\frac{4}{\pi^2}\\&=\frac{1}{\binom{n}{2}^2}\Big\{{}_nP_3\int_0^1\Big(\int_0^1\sin|z_1-z_2|\pi\,dz_2\int_0^1\sin|z_1-z_3|\pi\,dz_3\Big)dz_1\\&\quad+\binom{n}{2}\int_0^1\int_0^1\sin^2|z_1-z_2|\pi\,dz_1dz_2\end{aligned}$$

$$+\binom{n}{2}\binom{n-2}{2}\int_0^1\int_0^1 \sin|z_1-z_2|\pi\, dz_1 dz_2 \cdot \int_0^1\int_0^1 \sin|z_3-z_4|\pi\, dz_3 dz_4 \Big\}$$

$$-\frac{4}{\pi^2}$$

$$=\frac{1}{\binom{n}{2}^2}\left\{n(n-1)(n-2)\times\frac{4}{\pi^2}+\frac{1}{2}\binom{n}{2}+\binom{n}{2}\binom{n-2}{2}\left(\frac{2}{\pi}\right)^2\right\}-\frac{4}{\pi^2}$$

$$=\frac{\pi^2-8}{n(n-1)\pi^2}$$

また、S_nは次のように変形できる。

$$(2,5)\qquad S_n=\frac{1}{\binom{n}{2}}\sum_{k=1}^{n-1} k \sin\frac{n-k}{n+1}\pi$$

$$=\frac{2}{n-2}\left(\sum_{k=1}^{n-1}\sin\frac{k\pi}{n+1}-\frac{1}{n}\sum_{k=1}^{n-1}k\sin\frac{k\pi}{n+1}\right)$$

$$=\frac{2}{n-2}\left\{\sin\frac{n\pi}{2(n+1)}\sin\frac{(n-1)\pi}{2(n+1)}\Big/\sin\frac{\pi}{2(n+1)} -\frac{1}{n}\left(\frac{\sin\frac{\pi}{n+1}}{4\sin^2\frac{\pi}{2(n+1)}}-\frac{n\cos\frac{(2n-1)\pi}{2(n+1)}}{2\sin\frac{\pi}{2(n+1)}}\right)\right\}$$

$$\therefore\ \lim_{n\to\infty} S_n=\frac{2}{\pi}$$

$\boxed{注1}$

n	S_n
10	0.6955
20	0.6672
30	0.6573
40	0.6522
50	0.6491

n	S_n
100	0.6429
500	0.6378
1000	0.6372
2000	0.6369
5000	0.6367
∞	$2/\pi=0.6366$

$\boxed{注2}$ $n \leqq 100$ についてのU_nの確率分布のパーセント点はシミュレーションにより計算している。

$\boxed{注3}$ U_nの漸近分布については現在白旗慎吾氏（大阪大学）と共同研究している。

3、K-S検定，LLC検定のグラフ表示例と比較

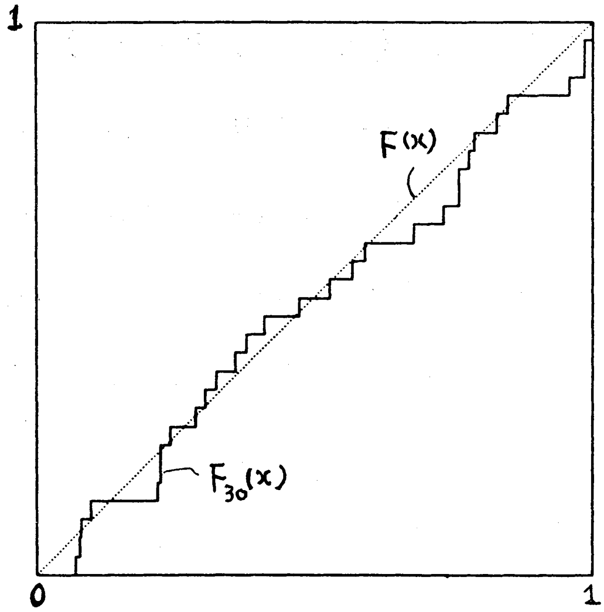
検定量U_nを用いる検定をLLC検定とよぶことにする。さて、発生させた［0, 1］上の一様乱数をはじめから500個ずつに区切り、区間［0, 1］上を10等分して、各区間に入る度数によりχ^2値を順次30個算出した値をつぎに示す。

(a)	19.92	3.68	9.24	5.64	5.84	5.64	8.24
	12.12	13.24	3.52	12.84	7.08	6.72	6.32
	27.96	4.00	6.52	8.80	17.04	19.76	3.64
	9.48	5.60	11.64	11.64	11.92	7.28	10.52
	7.60	11.24					
(b)	9.92	6.30	9.24	4.68	5.84	5.64	8.24
	8.12	10.24	8.52	11.84	7.08	10.72	8.32
	10.20	9.00	11.52	8.80	7.04	9.76	8.64
	9.48	5.60	6.64	10.64	11.92	7.28	10.52
	7.60	9.24					

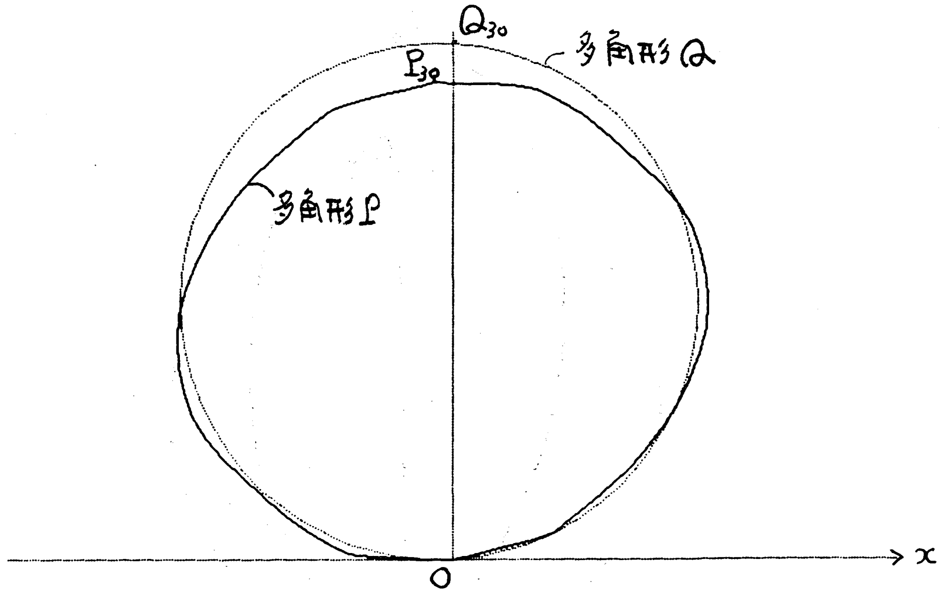
グラフ表示

(a), (b) の場合にK-S検定については累積グラフ$F_n(x)$を描き$F(x)$とのずれ度合をみる。この場合、x軸を$x \to F(x)$と変換して直線とのずれを見るようにする。また、LLC-30も(a), (b)について描き比較する。なおLLC-30は、直線からのずれ具合でなく円からのずれ具合を見るところに特徴がある。

(a)の場合

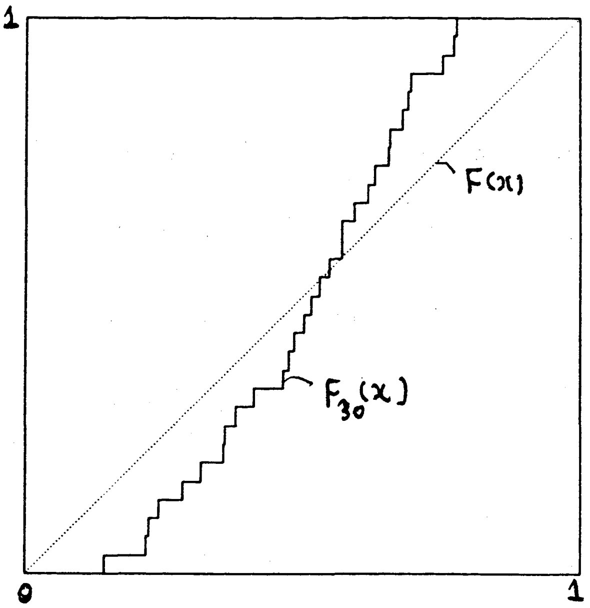


K-S検定の累積グラフ

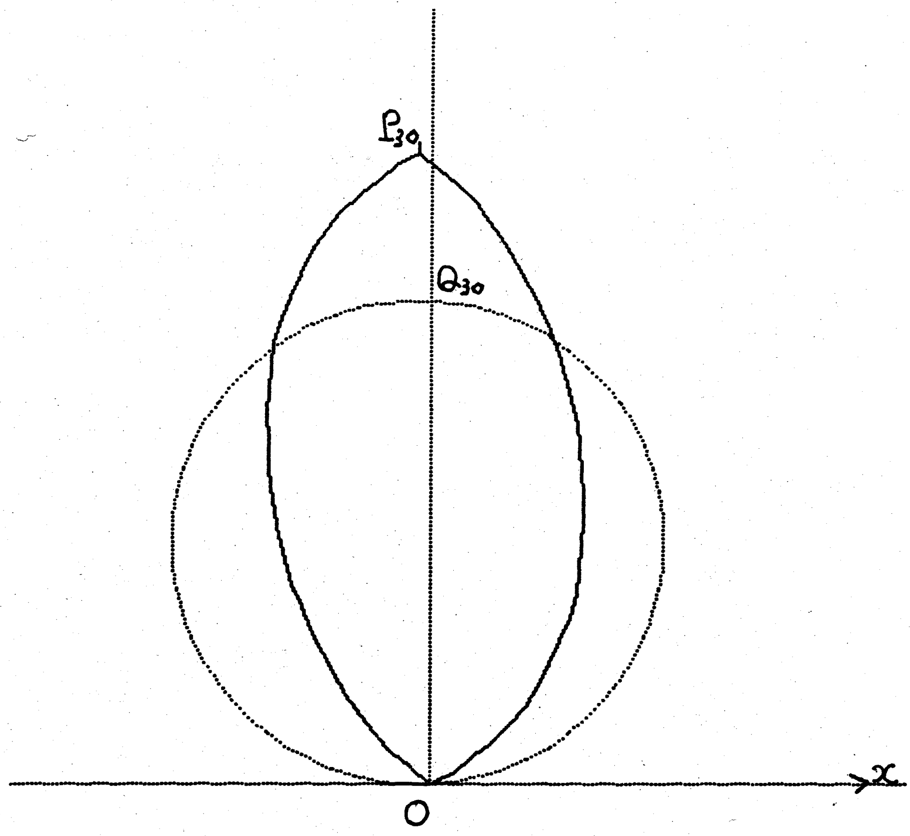


LLC − 30

(b)の場合



K−S 検定の累積グラフ



LLC-30

考察

(a)と(b)の場合，累積グラフとＬＬＣ－30はよく似た傾向を示すが、ＬＬＣ－30の方が円からのずれで、対称性がはっきりする。また、分布の裾のずれ具合はＬＬＣ－30の方がはっきりわかり、確率プロットとして知られるQ－Qプロットの特徴を備えているように思われる。

現在，検定量としてＬＬＣ検定のK-S検定に対する効率を研究しているが、いずれにせよ、K-S検定，ＬＬＣ検定，そ

れにグラフ表示を合わせて用いると$F_n(x)$の$F_0(x)$への適合度の確率の評価，また視覚処理により適合度合の様子がよく把握できる。

参考文献

[1] Capon, J.(1965). On the asymptotic efficiency of the Kolmogorov-Smirnov test, JASA, 60, 843-853.

[2] Feller, W.(1948). On the Kolmogorov-Smirnov limit theorems for empirical distributions, A.M.S., 19, 177-189.

[3] Hoeffding, W.(1948). A class of statistics with asymptotically normal distribution, A.M.S., 19, 293-325.

[4] Miller, L.H.(1956). Table of percentage points of Kolmogorov statistics, JASA, 51, 111-121.