区間データにもとづくKolmogorov-Smirnovの検定（2標本問題）

川崎医科大学名誉教授

仮　谷　太　一

（平成13年9月6日受理）

A Generalized Kolmogorov-Smirnov Test Based on 2 Interval Data Samples

Taichi KARIYA

Professor Emeritus,
Kawasaki Medical School,
577 Matsushima, Kurashiki, Okayama, 701-0192, Japan
(Received on September 6, 2001)

概　　要

急速に変転する自然環境・生活環境の下では，新奇な病気の発生も多く，その治療法の開発に必要な研究期間は出来るだけ短くしなければならない。十分なデータ数が得られず，データの分布に関する情報がよく分からない場合においても，さらに一歩前進するためには，その時点で何等かの判断を要請されることが少なくない。こうした場合における統計的判断の一手法として，さきに一般化ラページ検定[1]について述べたが，今回はコルモゴロフースミルノフ検定を一般化した検定法を提案する。

ここに述べるコルモゴロフースミルノフ検定は，2つの母集団の分布が同一とみなせるか否かを検定するもので，経験的分布関数を用いる。ラページ検定と同じく，母集団分布の連続性だけを仮定し，分布型，位置母数および尺度母数に対して何等の制約も設けないので，最も一般的な仮説および対立仮説の検定に適した検定法と言うことができる。しかしデータ解析の立場からは，普通の点データの場合，順位検定の一つとして処理することができ，データ数が少ないと有意確率のきめが粗くて使いものにならないとか，また，利用できる数表は，同順位データがなく2つの標本サイズが等しい場合に限られるとか，いろいろ不具合が多かった。ところが，医学データなどに，しばしば見られる区間データの場合，Fisherの並べかえ検定の手法[2]を用いることにより，標本サイズがあまり大きくなくても，データ数が同じでなくても，また同順位データがあっても，かなり細かい有意水準で精密な検定が可能になるのである。なお，ここで用いる並べかえ検定での，観測データをこみにして2つの標本に分割するプログラムは，かなり面倒なので，その主要部分を抜き書きすることにする。**キーワード**：区間データ，並べかえ検定

Abstract

Under the natural and living environment changing quickly, novel diseases will occur successively, and so we have to find out remedies for these diseases as quickly as possible. Even if we cannot get plenty of the cases and cannot obtain information about their distributions of observation values, we expect some statistical judgements on the

advantages and disadvantages of these remedies. As an example of statistical test on these cases, we proposed "A generalized Lepage test", and this time we will propose "A generalized Kolmogorov-Smirnov test".

In the Kolmogorov-Smirnov two-sample test here, the comparison is made between the empirical distribution functions of the two samples to test that two population distributions of them are same or not. The null hypothesis is formulated as identical populations with the common distribution completely unspecified except for the assumption that it is a continuous distribution function. And so the Kolmogorov-Smirnov test is very easy to apply and is useful mainly for the general alternatives, since the test statistic is sensitive to all types of differences between the empirical distribution functions.

But from the view-point of data analysis, the Kolmogorov-Smirnov test has not been used so frequently in the case of usual point-data, because the test can be treated as a rank test and the probability values are scattered when the sample sizes are small. Moreover, available statistical tables are restricted by the cases that sizes of two samples are equal and there are no ties.

Now in the case of interval data samples which are often observed in medical and environmental sciences, the Kolmogorov-Smirnov test revived by using the Fisher's permutation technique. The provability values can be calculated minutely by a personal computer regardless of ties of data or defference of sample sizes, even in the small size sample cases. In addition, we will extract the main part of program dividing the combined data into X-group and Y-group. **Key words**: interval data, permutation test

1．序

$(X_1, X_2, \cdots, X_m)$, $(Y_1, Y_2, \cdots, Y_n)$ を，連続な分布関数 F(x)，G(y) をもつ母集団からの，それぞれ大きさm，nの独立な無作為標本とする。問題は $\{X_i\}$，$\{Y_j\}$ の観測値にもとづいて

帰無仮説　Ho：F＝G　　を

対立仮説　(1) Ha：F＞G　または　(2) Hb：F≠G

に対して，検定することである。ここに(1) Ha はYがXより確率的に大きいことを，(2) Hb は F(t)≠G(t) なる t が少なくとも1つ存在することを意味する。

Kolmogorov-Smirnov 検定[3,4,5,6]は，$\{X_i\}$，$\{Y_j\}$ の観測値が通常の観測値（点データ）である場合の検定で，Ho，Ha または Hb から明らかなように，母集団の分布型についても，位置母数，尺度母数についても何の制約も設けない理論的には極めて興味深い検定である。

しかし，後で数値例で示すように，標本サイズが小さいときには，データが少し変わっただけで有意確率が大きく変わり，統計的判断がむずかしいことがある。これはこの検定が観測値の順位だけに基づく検定，すなわち順位検定の一つとして処理されうることからも推測されることである。さらに，利用できる統計数値表は，同順位がなく，m＝n の場合に限られているようである。このような事情から，理論的には極めて面白い検定と言われながら，実際に利用されることは稀であったのではなかろうか。特にデータ数の多くない医学データ解析では尚更で

ある。

ところが，医学データには血圧データにしても，乳歯の萌出データ，手術後ある症状の発生までのデータにしても，区間データ表示が適切と考えられるものが少なくない。この区間データにKolmogorov-Smirnov検定を適用するとき，経験的分布関数は，従来の点データにもとづく経験的分布関数にくらべて格段に細密になり，従って有意確率がきめ細かくなって，Hoの棄却または採択の判断に特別の配慮は不要になるのである。また，同順位データの存在，m，nの等不等にも無関係に，Fisherの並べかえ検定の手法を組み込んだプログラムの活用により，容易に有意確率の算定が可能である。Fisherの並べかえ検定では，$\binom{m+n}{m}$回の繰り返し計算を行わなければならないが，極めて演算速度の速くなったパソコンを用いれば，かなりのデータ・サイズまで実行可能である。

2．区間データにもとづく一般化Kolmogorov-Smirnov検定

2標本$(X_1, X_2, \ldots, X_m)$，$(Y_1, Y_2, \ldots, Y_n)$を，連続な分布関数F(x)，G(y)をもつ母集団からの，それぞれ大きさm，nの独立な無作為標本とし，しかもX_i，Y_jの実現値は，それぞれ，区間データ$[x_{iL}, x_{iU}]$，$[y_{jL}, y_{jU}]$によって特性づけられているとする。すなわち，X_iの実現値は区間$[x_{iL}, x_{iU}]$の1要素，Y_jの実現値は区間$[y_{jL}, y_{jU}]$の1要素であることだけが知られているとする。ここで，m，nは2より大きい整数，x_{iL}，x_{iU}，y_{jL}，y_{jU}は有限な実数とする。

分布関数	標本の大きさ	区間データ
F(x)	m	$[x_{1L}, x_{1U}]$，$[x_{2L}, x_{2U}]$，…，$[x_{mL}, x_{mU}]$
G(y)	n	$[y_{1L}, y_{1U}]$，$[y_{2L}, y_{2U}]$，…，$[y_{nL}, y_{nU}]$

問題は，対立仮説(1) Ha：F＞Gまたは(2) Hb：F≠Gに対し，帰無仮説 Ho：F＝Gを検定することである。

検定統計量は，点データの場合と同じであるが，経験的分布関数は同じではない。点データの場合，同順位データがなければ，Fm(t)はm段，Gn(t)はn段の階段関数であるが，区間データの場合，一般にはそれぞれ，m，nより折れ曲がり点の多い経験的分布曲線になる。

対立仮説(1)，(2)に対する検定統計量は，それぞれ

(1)　$Dm,n=\underset{-\infty<t<\infty}{\mathrm{Max}}\,(Fm(t)-Gn(t))$

(2)　$D'm,n=\underset{-\infty<t<\infty}{\mathrm{Max}}\,|Fm(t)-Gn(t)|$

である。ここに，

$$Fm(x)=\frac{1}{m}\left\{x_{iU}\leq x\text{ をみたす }X_i\text{ の個数}+\sum_{i\,;\,x_{iL}<x<x_{iU}}\int_{X_{iL}}^{X_{iU}} f_i(x\,;\,x_{iL},\ x_{iU})dx\right\},$$

$$Gn(y)=\frac{1}{n}\left\{y_{jU}\leq y\text{ をみたす }Y_j\text{ の個数}+\sum_{j\,;\,y_{jL}<y<y_{jU}}\int_{y_{jL}}^{y_{jU}} g_j(y\,;\,y_{jL},\ y_{jU})dy\right\},$$

$f_i(x;x_{iL}, x_{iU})$，$g_j(y;y_{jL}, y_{jU})$ はそれぞれ X_i，Y_j の確率密度関数である。

以下(1)の場合について考察する。

観測データから計算される Dm,n の値を dm,n とする。帰無仮説 Ho のもとでは，X，Yの母集団はまったく同一であるから，観測されたデータを合併して得られる大きさ N＝m＋n の標本データの中から，どのm個を選んでX標本データとすることもみな同様に確からしい。この Fisher の並べかえ検定の考え方に従い，$\binom{N}{m}$ 通りのすべての場合について，Dm,n の値を計算し，観測データの場合の dm,n と比較して，それと等しいかまたはそれより大きい値をもつ場合の数を k とする。

このとき，有意確率（P-値）＝Pr(Dm,n≥dm,n) は $k/\binom{N}{m}$ である。

有意水準を c とすれば，P-値＜c のとき，Ha：F＞G に対し有意水準 c で Ho：F＝G を棄却することができる。

(2)の場合についても，P-値＝Pr(|Dm,n|≥|dm,n|) を除いて，ほぼ同様である。

3．数値例

手術後，毎週きまった日に，ある症状の発生を検診して，次のデータを得た。X群よりY群の方が，発生までの期間が長いと言えるだろうか。勿論特定の処理を除いて両群は同等であると仮定する。また，有意水準は0.05とする。

X群：［28，45］，［40，47］，［40，47］，［41，48］，［42，49］　　m＝8
　　［44，51］，［45，52］，［48，55］

Y群：［39，46］，［43，50］，［45，52］，［46，53］，［47，54］　　n＝7
　　［49，56］，［50，57］

ある症状の発生までの期間は，連続変量であるが，その分布については，観測された区間データ以外，なにも知られていないとする。従って，両群の母分布間の違いについては，区間データ標本に対する一般化 Kolmogorov-Smirnov 検定（GKS 検定）が適切であると考えられる。この場合，Y群の方がX群より，発生までの期間が確率的に長いことを検定したいのであるから，Ha：F＞G に対して Ho：F＝G を検定する(1)の場合になる。

計算結果は次の通りである。ここでは，$f_i(x)$，$g_j(y)$ は一様分布を仮定して計算を行った。

P-値＝298/6435＝0.0463，k＝298，$\binom{N}{m}$＝6435，$d_{8,7}$＝0.375

従って，P-値＝0.0463＜0.05（有意水準）であるから，Ha：F＞G に対し，Ho：F＝G を棄却することができる。YがXより確率的に大きいことが一応認められたことになる。

ところで，先に点データの場合には，Kolmogorov-Smirnov 検定（KS 検定）は小標本のとき，P-値がとびとびになり，検定のきめが粗に過ぎることを述べたが，上記の区間データを，それぞれの区間中央値で置き換えた場合の，KS 検定の結果，ならびに，X群の［44，51］とY群の［43，50］とを交換した場合の，GKS 検定，KS 検定の結果を示すと表1のようになる。

表1　P-値と（dm,n）

	一般化KS	KS検定
実験データ	0.0463* (0.375)	0.1175 (0.482)
1個交換した場合	0.0331* (0.413)	0.0435* (0.607)

＊印は5％有意を示す。

次に，一般化ラページ検定（GLE検定）と，ここでのGKS検定とを比較するため，文献1）で取り扱った男女学生の最高血圧値について，女子より男子の方が確率的に高いという対立仮説に対する検定結果を示そう。区間データは表2，検定結果のP-値は表3に示されている。女子をX群，男子をY群とすれば，(1)の場合と同様に計算すればよい。$f_i(x)$, $g_j(y)$ は一様分布として計算した。

表2　最高血圧値の区間データ

データ（I）

男子：[115, 124], [106, 114], [108, 112], [117, 128], [100, 119]
[118, 134], [124, 130], [116, 131]

女子：[105, 112], [110, 113], [110, 116], [90, 108], [99, 108]
[112, 127], [108, 142], [103, 106]

データ（II）

男子：[111, 118], [107, 110], [117, 130], [114, 120], [121, 130]
[116, 122], [106, 114], [99, 108]

女子：[92, 101], [84, 95], [98, 102], [99, 109], [110, 113]
[90, 101], [104, 118], [105, 112]

表3　表2のデータについての検定結果P-値

組	標本の大きさ	GKS検定	GLE検定
(I)	8，8	0.0399*	0.1927
(II)	8，8	0.0133*	0.0242*

＊印は5％有意を示す。

表3から，GKS検定は，分布の中心部分に重点が置かれており，GLE検定は分布の中心から離れた値に大きく影響を受け易い検定であることが読み取れる。検定結果の解釈には，このように検定統計量に留意することが肝要である。

4．付記　2群へのデータ割り当てプログラム

```
1400 ''XYiは［IXY(i, 1), IXY(i, 2)］において一様分布するものと仮定する
1410 ''X, Yの累積度数分布を計算する
1420 CN=0:PCN=0:SMAX=0:LPRINT
1430 I=1 :JX(1)=1
1440 'LPRINT:LPRINT"          Data    Number"
1450 'LPRINT"          No          No    of    X-Group":LPRINT
1460 IF I=NX THEN GOTO 1510
1470 IP1=I+1
1480 FOR L=IP1 TO NX
1490 JX(L)=JX(L-1)+1
1500 NEXT L
1510 CN=CN+1
1520 '' ----  データを2つの群に分ける  ----
1530 I=1 : K=1 : II=0
1540 IF I=JX(K) THEN GOTO 1580
1550 IF I < JX(K) THEN GOTO 1570
1560 IF K < NX      THEN K=K+1 : GOTO 1540
1570 II=II+1 : JY(II)=I
1580 IF I < N      THEN      I=I+1 : GOTO 1540
1590 'LPRINT:LPRINT USING"####)";CN;
1600 'FOR KX=1 TO NX : LPRINT USING"#### ";JX(KX);: NEXT KX
1610 'FOR KY=1 TO NY : LPRINT USING"####";JY(KY);: NEXT KY
1620 'LPRINT
1630 '' -------  MAX=(FX(t)-FY(t))の計算  -------


1900 IF CN=1 THEN SMAXO=WMAX
1910 'LPRINT USING"####    ###.####";CN,WMAX
1920 IF SMAXO <= WMAX THEN PCN=PCN+1
1930 '' -----  N個からR個をとり出す  -----
1940 I=NX
1950 IF JX(I) < NY+I THEN GOTO 1980
1960 I=I-1
1970 IF I <=0 THEN GOTO 2000 ELSE GOTO 1950
1980 JX(I)=JX(I)+1
1990 GOTO 1460
```

References

1) 仮谷太一：区間データにもとづく一般化ラベージ検定（2標本問題）．川崎医学会誌一般教養篇，26：11-16，2000
2) 仮谷太一：医歯系・生物系のベーシック統計学，共立出版，1988，第6章
3) Massey, F. J.: The Distribution of the Maximum Deviation between Two Samples Cumulative Step Functions. Ann. Math. Statist., 22: 125-128, 1951
4) Massey, F. J.: Distribution Table for the Deviation between Two Sample Cumulatives. Ann. Math. Statist., 23: 435-441, 1952
5) Drion, E. F.: Some Distribution-free Tests for the Difference between Two Empirical Cumulative Distribution Functions. Ann. Math. Statist., 23: 563-574, 1952
6) 柳川　堯：ノンパラメトリック法．倍風館：161-169，付表Ⅰ-1，Ⅰ-2，1982