川崎医会誌一般教，28号：1—6(2002)

環境汚染の程度を比較するパックデータ・アナリシス

川崎医科大学　名誉教授

仮　谷　太　一

(平成14年9月13日受理)

A Statistical Analysis of Packed Data for Comparing the Pollution Level.

Taichi KARIYA

Professor Emeritus,
Kawasaki Medical School,
577 Matsushima, Kurashiki, Okayama, 701-0192, Japan
(Received on September 13, 2002)

概　　要

環境汚染の問題は，きわめて広範多岐にわたり，汚染度の観測・観測値の解析に問題を限定しても複雑多様に過ぎる。そこでこの小論では，騒音公害・悪臭公害など，それらの観測値の平均値や合計値よりも，有害とみなされる高レベルの観測値の頻度や強度が問題となる場合を取りあげ，各地域間の汚染度の高低を検定する問題に限定することにする。

さて，騒音や悪臭の場合，汚染状況は概ね日・週・季節を周期として変動する。そしてその基本となる最短周期の一日においても，時刻・晴雨・気温・風向きなどにより，観測値は大きく変動する。従って，地域間の汚染度の比較においては，日々の観測値の生起する順番はあまり重要ではない。かくして，このような公害データ解析の手始めとしては，各観測地点ごとに一日間の観測値をパックし，一日ごとに汚染の程度を比較をするのが妥当であろう。そのとき，各観測値のもつ情報を出来るだけ有効に利用する方法として，Wilcoxonの検定法を一般化したWilcoxonの精密検定法を提案する。キーワード：パックデータ，一般化Wilcoxon検定

Abstract

The problem of environmental pollution covers a wide range of subjects, and is too complicated even if we limit the problem to study the methods of observation and statistical analyses of pollution. So in this article, we will take up the noise pollution and the bad smell pollution. In these cases, the frequency and the strength of observations which are considered harmful to the health are more important than their means or totals.

To make the points clear, we limit ourselves to test which of the two areas is higher in the degree of pollution. And furthermore, we assume that the observation system is decided in advance.

Now in the cases of noise pollution and bad smell pollution, the state of pollution fluctuates as a period of a day, week and season approximately. And even in a day which is the shortest period, the observations change by the tine, weather, tempera-

ture, the direction of the wind and so on. Therefore, the time order of observations in one day does not play a vital role, in order to compare the pollution level between areas.

Under these conditions, as the first step of these pollution data analysis, it is appropriate to pack the observations, day by day, every areas, and compare the pollution levels among areas. Then we propose the generalized Wilcoxon exact test, as a method which utilizes the information of every observations effectively. Key words: packed data, the generalized Wilcoxon exact test

1．序

環境汚染データの中には，汚染物質の累積量の比較により，地域間の汚染の程度を比較計量することが出来るものもあるが，騒音・悪臭などの場合は，そのような方法では問題の核心に迫ることは困難である。それらの観測値の平均値や合計値よりも，有害と見なされる高レベルの観測値の頻度や強度などを問題としなければならない。観測地点については，それぞれの地域の状況に応じ，例えば幹線道路沿いとか，工場近辺とか，住宅地区など，環境科学的観点から考慮しなくてはならないし，また観測方法についても，最近では低周波公害が問題になるとか，悪臭濃度ではなく臭気判定技士による臭覚測定とか，極めて複雑になっている。

そこでこの小論では，各地域における観測地点・観測時点および観測方法はあらかじめ定められており，与えられた観測データに基づいて地域間の汚染度に有意な差が認められるか否かを問題とすることにする。

さて，各観測地点における，その観測値変動の周期は，厳密な意味では定め難たいが，概周期としては日，週，季節，年などが考えられるであろう。そして地点ごとに，時刻・晴雨・気温・風向きなどにより，観測値はそれぞれに変動する。従って，地域間の汚染度の違いを問題にするときには，最短概周期の一日の観測値の生起する順番はあまり重要ではない。そこで各観測地点毎に一日の観測値をパックし，各地点間ごとに，その日の汚染度の比較をするのが，このような公害データ解析の手始めとしては適当であると考える。

このとき，各パックデータのもつ情報を出来るだけ有効に利用する方法としては，パックデータの差に対して一般化符号[1,2,3]を与え，Wilcoxon 検定[4]を一般化した Wilcoxon の精密検定法[3]に従うのが妥当であると考える。

2．パックデータにもとづく一般化 Wilcoxon 精密検定

比較したい 2 地域の標本 $(X_1, X_2, \cdots, X_m)$，$(Y_1, Y_2, \cdots, Y_n)$ 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し，X_i，Y_j は，それぞれ p 個の観測値により特性づけられているものとする。これら p 個ずつの観測値は，X_i，Y_j の最短周期内の定められた時刻における観測値である。

分布関数	標本の大きさ	パックデータ
F(x)	m	$\{X_{11}, X_{12}, \cdots, X_{1p}\}, \{X_{21}, X_{22}, \cdots, X_{2p}\}, \cdots, \{X_{m1}, X_{m2}, \cdots, X_{mp}\}$
G(y)	n	$\{Y_{11}, Y_{12}, \cdots, Y_{1p}\}, \{Y_{21}, Y_{22}, \cdots, Y_{2p}\}, \cdots, \{Y_{n1}, Y_{n2}, \cdots, Y_{np}\}$

問題は帰無仮説　Ho：F(t)＝G(t)　を

対立仮説　Ha：F(t)＜G(t)　あるいは

Hb：F(t)＜G(t)　または　F(t)＞G(t)

に対して検定することである。

提案する統計量 W は，次のとおりである。

まず，パックデータ $\{X_{i1}, X_{i2}, \cdots, X_{ip}\}$，$\{Y_{j1}, Y_{j2}, \cdots, Y_{jp}\}$ で表示される確率変数の差 $X_i - Y_j$ の一般化符号 U_{ij}[1,2,3]を次式によって定義する。

$$U_{ij} = \sum_{h=1}^{p}\sum_{k=1}^{p}\frac{\mathrm{sgn}(X_{ih} - Y_{jk})}{p \cdot p} \qquad \cdots\cdots(1)$$

ここに，

$$\mathrm{sgn}(x-y) = \begin{cases} 1 & \text{もし}\quad x > y \quad \text{ならば} \\ 0 & \text{もし}\quad x = y \quad \text{ならば} \\ -1 & \text{もし}\quad x < y \quad \text{ならば} \end{cases}$$

である。

このとき，一般化統計量 W を次式により定義する。

$$W = \sum_{i=1}^{m}\sum_{j=1}^{n} U_{ij} \qquad \cdots\cdots(2)$$

3．帰無仮説 Ho の下における統計量 W の確率分布

帰無仮説 Ho の下では，$(X_1, X_2, \cdots, X_m)$，$(Y_1, Y_2, \cdots, Y_n)$ は，同じ母集団からの独立な標本であるから，合併して大きさ m＋n の独立標本 $(XY_1, XY_2, \cdots, XY_{m+n})$ として考察する。このとき前述した U_{ij} は，確率変数 $XY_i - XY_j$ の一般化符号を表し，$U_{ij} = -U_{ji}$，$U_{ii} = 0$ である。さて，$(XY_1, XY_2, \cdots, XY_{m+n})$ を，それぞれ大きさ m，n の X 群，Y 群に分割する方法は $\frac{(m+n)!}{(m!n!)}$ 通りあり，帰無仮説 Ho の下ではすべて同じ生起確率をもつ。従って，我々はこれらの分割によってそれぞれ決まる統計量 W の，Ho の下における精密分布を求めることが出来る。

このとき，両地域で観測された大きさ m，n の標本パックデータにもとづく有意確率 (P-値) は次式で与えられる。

対立仮説が Ha：F(t)＜G(t) のとき

P-値＝P(W|W≧Wo)

ここに，Wo は観測された標本データに対する W の値である。

このP-値が，あらかじめ定められた有意水準より小さいとき，帰無仮説Hoを棄却することが出来る。なお，この片側検定は，X地域の汚染度がY地域のそれより大きいことが，かなり確かな事前情報としてある場合に用いられる。

対立仮説が F(t)＞G(t) のときは，X，Yを交換して行えばよい。

なお，両地域間の汚染度の差に関する事前情報がない場合には，両側検定（対立仮説Hb）となり，P-値は

$$\text{P 値}=\begin{cases}2\cdot P(W|W\geqq Wo)\cdots W\geqq 0\text{ のとき},\\ 2\cdot P(W|W\leqq Wo)\cdots W\leqq 0\text{ のとき},\end{cases}$$

となる。

つぎに，Wの条件付平均値及び分散を E(W|P，Ho)，V(W|P，Ho) と表示しよう。ここにPは，観測されたパックデータのパターンである。期待値および分散は，観測されたパックデータのパターンPに帰着する全部で ${}_{m+n}C_m$ 通りの等確率標本にわたって計算される。

分割の対称性から，容易に

$$E(W|P,\ Ho)=0 \qquad \cdots\cdots(3)$$

また，分散は次のようになる。

$$V(W|P,\ Ho)=E(W-E(W|P,\ Ho))^2=E(W^2)$$

$$=\sum_{s=1}^{C_1}\frac{W^2(s)}{C_1}=C_2\sum_{i=1}^{m+n}\frac{U_{i\cdot}^2}{C_1}$$

$$=m\cdot n\frac{\sum_{i=1}^{m+n}U_{i\cdot}^2}{(m+n)(m+n-1)} \qquad \cdots\cdots(4)$$

ここに，W(s) はs番目の標本分割に対するWの値であり，

$C_1={}_{m+n}C_m$,

$C_2={}_{m-n-2}C_{m-1}$,

$U_{i\cdot}^2=\sum_{j=1}^{m+n}(j\neq i)\ U_{ij}=\sum_{j=1}^{m+n}U_{ij}\quad(\because U_{ii}=0)$

である。

4．統計量Wの漸近分布

さて，3．で述べたと同様にして，ここでの一般化符号 $U_{ij}\ (i\neq j)$，i，$j=1$，2，…，m＋nは合併した大きさm＋nの標本（XY_1，XY_2，…，XY_{m+n}）からの，2つの確率変数のすべての組み合わせの差に対して計算され，次式の成り立つことは明らかである。すなわち

$$U_{ij}=-U_{ji},\quad -1\leqq U_{ij}\leqq +1,\quad U_{ii}=0$$

これら (m＋n)·(m＋n－1) 個の $U_{ij}\ (i\neq j)$ の中から，各分割に対応してm·n個が選ばれ加算されたのち，Wの総計に加えられる。従って，全部でm·n·C_1 個の U_{ij} がWの総計の中に含まれ，しかも各 U_{ij} の出現回数は相等しい。この故に，$U_{ij}\ (i\neq j)$ はWの総計の中に，それぞ

れ $\frac{m\cdot n\cdot C_1}{(m+n)\cdot(m+n-1)}$ 回ずつ出現することになる。このことが，(4)の計算式の根拠になっている。

さてここで，U_{ij} を測定における根元誤差と考えるならば，容易にガウスの誤差法則と同様にして，m，n→∞のとき W は正規分布に従うことが理解できるであろう。われわれのシミュレーション・スタディーの結果により，W の分布の正規近似は，m，n がともに 7 をこえるならば，非常に良好であることが判明した。すなわち，

$$Z=\frac{W}{\sqrt{V(W|P,\ Ho)}}\ \sim\ N(0,\ 1)$$

5. 数値例（仮想例）

この節では，1．序で述べたように，各地域における観測地点，観測時刻および観測方法は，あらかじめ定められているものとする。そして，そうしたシステムの下で得られたデータにもとづいて，地域間の汚染度に関する検定を Wilcoxon の精密検定法を用いて行う。

地域は，簡単のため X，Y の 2 地域とし，観測地点数はともに 7 (m＝n＝7)，1 日の観測回数は 6 (p＝6) とする。

さて，汚染度について，Y 地域の方が X 地域より高いという，かなり確かな事前情報があるので，帰無仮説 Ho，対立仮説 H′a は次の通りとする。

Ho：F(t)＝G(t)，H′a：F(t)＞G(t)

(X_1，X_2，…，X_7)，(Y_1，Y_2，…，Y_7) のパックデータは次表の通りで，それぞれ 6 個の観測値で構成されている。

X 地域 (m＝7)

(106, 107, 112, 105, 109, 106), (112, 110, 113, 110, 110, 110)
(116, 112, 110, 111, 116, 118), (101, 99, 98, 93, 90, 89)
(108, 99, 99, 101, 106, 96), (121, 127, 123, 127, 112, 117)
(116, 113, 108, 110, 142, 112)

Y 地域 (n＝7)

(124, 124, 117, 124, 115, 119), (113, 106, 111, 114, 111, 112)
(110, 112, 109, 111, 108, 106), (123, 127, 128, 128, 117, 123)
(117, 100, 118, 109, 119, 118), (134, 121, 133, 123, 118, 123)
(130, 124, 124, 128, 130, 119)

計算の結果は

Wo＝－27.0556，C_1＝3432

$$P\text{-値} = P(W|W \leqq Wo) = \frac{89}{3432} = 0.0259$$

従って，有意水準を0.05とすれば，HoはH'aに対し棄却することができる。平たく云えば，この日の汚染度は，Y地域の方がX地域より高かったと云うことが出来る（危険率5％)。こうした計算を7日に亙って行えば，週単位での汚染度の比較が可能になる。

ちなみに，正規近似を用いると，Zo＝－1.9406でP-値＝0.0252となる。

なお，この数値例では，outlier（ど外れ値）は認められなかったので，問題はないが，outlierがある場合には，この検定法による解析の外に，それぞれのoutlierについて，個別の検討が必要である。

こうして最短概周期についての検定結果をもとにして，さらに，週，月，季節などについて，地域間の環境汚染の程度を検討することができる。

Wの精密分布と漸近分布は，次の通りである。

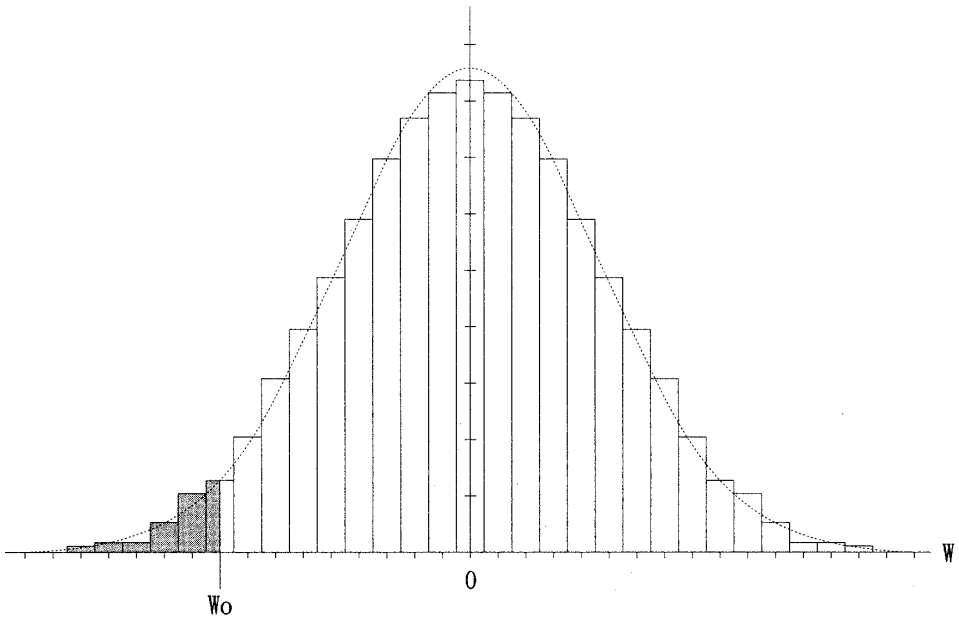

図1　Wの精密分布と漸近分布

References

1) Kariya, T.: A generalized sign test based on paired interval data. The second Japan-China symposium on statistics: 125-128, 1986
2) 仮谷太一：対応のある2標本の区間データに基づく一般化符号検定．応用統計学　16：77-88, 1987
3) Kariya, T.: A generalized Wilcoxon test for comparing interval data samples. Kawasaki Med J 14: 187-192, 1988
4) Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics 1: 80-83, 1945