川崎医会誌一般教，29号：1−6(2003)

区間データ標本へのKruskal-Wallis検定の拡張

川崎医科大学　名誉教授

仮　谷　太　一

(平成15年9月3日受理)

An Extension of the Kruskal-Wallis Test to Interval Data Samples

Taichi KARIYA

Professor Emeritus,
Kawasaki Medical School,
577 Matsushima, Kurashiki, Okayama, 701-0192, Japan
(Received on September 3, 2003)

概　　要

Kruskal-Wallisの一元配置分散分析検定の，区間データ標本への拡張である検定法を提案する。この検定法は，変量の分布型に影響を受けないノンパラメトリック検定である。区間データ標本は，医学実験では，しばしば得られる。すなわち，各個体の観測値が区間によってのみ表示可能な場合である。例えば，ある外科的手術から特定の症状の発現までの期間，特定乳歯の萌出年齢，眼底出血の発生時点など，正確な観測値を得ることは困難である。このような場合には，区間データによって表示することが望ましいからである。

さて，2つの観測値の差の一般化符号を，それぞれの区間データにもとづいて定義し，それらを用いて各区間データの一般化順位を定める。こうしておいて，Kruskal-Wallisの検定法における普通の順位の代わりに，この一般化順位を用いて検定統計量を構成する。

一般化順位を用いるこの検定法では，実験毎に各区間データの一般化順位は，通常違ってくるから，あらかじめ有意水準の表を作っておいて，利用するというわけにはいかない。従って計算はコンピュータに頼らなければならないが，基本的なプログラムさえ作って置けば，かなりの標本数および標本サイズまで，処理は可能である。キーワード：Kruskal-Wallis検定，区間データ，一般化符号，一般化順位

Abstract

A nonparametric test is proposed that is an extension of the Kruskal-Wallis one-way ANOVA test to interval data samples. Interval data samples are often obtained in medical experiments, namely, in the cases where the observation for each subject is specified only by an interval. We cannot observe exactly a period of time from some surgical operation to expression of specified symptom, an eruptive age of a specified deciduous tooth, or a point of time when hemorrhage happened in one's eyes and so on. In these cases, it is preferable that the observation should be specified by interval.

We define a generalized sign of difference between two observations based on their interval data. Then using these generalized signs, we define the generalized rank

of each interval datum, and organize a test statistic using the generalized ranks instead of ordinary ranks in the Kruskal-Wallis test one.

In this test using the generalized ranks, we cannot use the table of critical values, because the generalized rank of each interval datum changes usually at every experiment. Therefore we have to calculate by computer, but if we compose and keep the fundamental program, we can deal with considerable sizes of samples. Key words: Kruskal-Wallis test, interval data, generalized sign, generalized rank

1．序

分布関数 $F_1(X), \ldots, F_i(X), \ldots, F_s(X)$ をもつ母集団からの，それぞれ大きさ $n_1, \ldots, n_i, \ldots, n_s$ の，独立な無作為標本を，$(X_{11}, X_{12}, \ldots, X_{1n_1}), \ldots, (X_{i1}, X_{i2}, \ldots, X_{in_i}), \ldots, (X_{s1}, X_{s2}, \ldots, X_{sn_s})$ とする。

問題は，X_{ij} の観察値が，それぞれ区間データ (x_{ijL}, x_{ijU}) として表示されているとき，

帰無仮説　Ho：$F_1(X)=F_2(X)=\cdots=F_s(X)$　を

対立仮説　Ha：Ho　の否定

に対して検定することである。

Kruskal-Wallis 検定[1]は，X_{ij} の観測値が通常の観測値（点データ）であるとき，各観察値について，他のすべての観察値に対する相対的な大きさを考えて順位を与え，それら順位を用いて検定を行う。このとき，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，何等制約を設けないノンパラメトリック検定である。このように，ほとんど制約がなく，適用範囲の広い検定であるが，計算量が膨大で，同順位がなく標本数 s が 3，データ数合計が 6 ～15の場合に対しても，有意確率の表が 5 ページ（E. L. レーマン）も必要となるという実用上致命的ともいえる欠点があった。

さらに，医歯学データには，乳歯の萌出日時，手術後ある症状が発生するまでの期間，あるいは血圧データにしても，区間表示が適切と考えられるものが少なくない。区間データは互いに重なりあうことが多いから，それぞれにその大きさに応じて，普通の順位を与えることはできない。しかし，区間データにも，次節のようにして，小数点のついた一般化順位を定義しておけば，同順位データの存在にも無関係に，Kruskal-Wallis の考えにしたがった検定が，極めて高速化したコンピューターの活用により，容易に行うことが出来るようになった。

2．Kruskal-Wallis one-way ANOVA 検定[1]の区間データ標本への拡張

s 個の標本 $(X_{11}, X_{12}, \ldots, X_{1n_1}), (X_{21}, X_{22}, \ldots, X_{2n_2}), \ldots, (X_{s1}, X_{s2}, \ldots, X_{sn_s})$ を，分布関数 $F_1(X)$, $F_2(X), \ldots, F_s(X)$ をもつ母集団からの，それぞれ大きさ $n_1, n_2, \ldots, n_s$ の独立な無作為標本とし，しかも X_{ij} の実現値は，それぞれ区間データ[2-8] (X_{ijL}, X_{ijU}) によって特性づけられているとする。すなわち，X_{ij} の実現値は，区間 (X_{ijL}, X_{ijU}) の一要素であることだけが知られているとする。ここで，n_i は 2 以上の整数，X_{ijL}，X_{ijU} は有限な正の実数とする。

分布関数	標本の大きさ	区　間　デ　ー　タ
$F_1(X)$	n_1	$(X_{11L}, X_{11U}), (X_{12L}, X_{12U}), \ldots, (X_{1n_1L}, X_{1n_1U})$
$F_2(X)$	n_2	$(X_{21L}, X_{21U}), (X_{22L}, X_{22U}), \ldots, (X_{2n_2L}, X_{2n_2U})$
⋮	⋮	⋮
$F_s(X)$	n_s	$(X_{s1L}, X_{s1U}), (X_{s2L}, X_{s2U}), \ldots, (X_{sn_sL}, X_{sn_sU})$

問題は，帰無仮説 Ho：$F_1(X)=F_2(X)=\cdots=F_s(X)$ を，対立仮説 Ha：Ho の否定に対し，検定することである。

そのために，まず $X_{i\alpha_i}-X_{j\alpha_j}$ の，それらの区間データ $(X_{i\alpha_iL}, X_{i\alpha_iU}), (X_{j\alpha_jL}, X_{j\alpha_jU})$ にもとづく，一般化符号[2-8] $D_{i\alpha_i j\alpha_j}$ を次のように定義する。

$$D_{i\alpha_i j\alpha_j}=E(\mathrm{sgn}(x_{i\alpha_i}-x_{j\alpha_j}))$$
$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{sgn}(x_{i\alpha_i}-x_{j\alpha_j})f_{i\alpha_i}(x_{i\alpha_i}; x_{i\alpha_iL}, x_{i\alpha_iU})\cdot f_{j\alpha_j}(x_{j\alpha_j}; x_{j\alpha_jL}, x_{j\alpha_jU})dx_{i\alpha_i}\cdot dx_{j\alpha_j} \quad \cdots\cdots(1)$$

ここに，$f_{i\alpha_i}(x_{i\alpha_i}; x_{i\alpha_iL}, x_{i\alpha_iU})$ は，第 i 標本の α_i 番目の区間における $x_{i\alpha_i}$ の確率密度関数で，$F_i(X)$ とは独立であり，また $f_{j\alpha_j}(x_{j\alpha_j}; x_{j\alpha_jL}, x_{j\alpha_jU})$ は，第 j 標本の α_j 番目の区間における $x_{j\alpha_j}$ の確率密度関数で，$F_j(X)$ とは独立である。

なお，

$$\mathrm{sgn}(y_i-y_j)=\begin{cases}1 & y_i>y_j \text{ のとき},\\ 0 & y_i=y_j \text{ のとき},\\ -1 & y_i<y_j \text{ のとき},\end{cases}$$

このとき，$D_{i\alpha_i i\alpha_i}=0$，$D_{i\alpha_i j\alpha_j}=-D_{j\alpha_j i\alpha_i}$ $(i\neq j)$ が成り立つことは見やすい。

このとき，(1)により

$$D_{i\alpha_i j\alpha_j}=\begin{cases}-1 & x_{i\alpha_iU}\leqq x_{j\alpha_jL} \text{ のとき},\\ +1 & x_{i\alpha_iL}\geqq x_{j\alpha_jU} \text{ のとき},\\ -1 \text{ より大きく } +1 \text{ より小さい実数} & \text{その他のとき}\end{cases}$$

を得る。

次に，$x_{i\alpha_i}$ の一般化順位[2-8] $R_{i\alpha_i}$ を次式によって定義する。

$$R_{i\alpha_i}=\sum_{j=1}^{s}\sum_{\alpha_j=1}^{n_j}\frac{D_{i\alpha_i j\alpha_j}}{2}+\frac{N+1}{2} \quad \cdots\cdots(2)$$

ただし $N=n_1+n_2+\cdots+n_s$.

このとき区間データ標本にもとづく一般化 Kruskal-Wallis の検定統計量 GH は

$$GH=\left(\frac{1}{\sigma^2}\right)\left(\frac{N-1}{N}\right)\sum_{i=1}^{s}\frac{(R_{i.}-n_i\mu)^2}{n_i} \quad \cdots\cdots(3)$$

となる。帰無仮説 Ho は GH が大きいとき，対立仮説 Ha に対し，棄却される。

ここに $R_{i.}=\sum_{\alpha_i=1}^{n_i}R_{i\alpha_i}$ は第 i 標本に属する変量の一般化順位である。また，μ，σ^2 は Ho の下での R_{ij} $(i=1,2,\ldots,s;\ j=\alpha_1,\alpha_2,\ldots,\alpha_{nj})$ 全体(N 個)の平均と分散を表す。

ところで(3)式は，重みつきの偏差平方和で，重みとしてはそれぞれの標本サイズの逆数を用いている。また定数 $\frac{N-1}{N}$ は，GH に，漸近分布として近似的に χ^2 分布に従うようにするための工夫である。

3．帰無仮説の下における統計量 GH の確率分布と漸近分布

N 個の一般化順位からなる有限母集団から，重複を許さずにランダム抽出して s 個の標本に配分分割する方法は $N!/\prod_{i=1}^{s} n_i!$ 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，これらの分割のそれぞれに対して算定される統計量 GH の，Ho のもとにおける確率分布を求めることが出来る。このとき観測された標本データの有意確率 (P-値) は，次式によって与えられる。対立仮説が Ha であるから

$$\text{P-値} = (\text{GHo 以上の値の GH の個数}) \cdot \prod_{i=1}^{s} \frac{n_i!}{N!} \qquad \cdots\cdots(4)$$

ここに，GHo は観察された標本データに対する GH 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 は対立仮説 Ha に対して棄却される。

次に，帰無仮説 Ho の下で，n_i 個の $R_{i\alpha_i}$ $(\alpha_i = 1,2,\ldots,n_i)$ は，1 から N までの間の N 個の一般化順位（実数）から，重複を許さずにランダムに抽出された，サイズ n_i のランダム標本を構成している。従ってその平均及び分散について，

$$E(\overline{R_{i.}}) = \frac{N+1}{2} = \mu ,$$

$$V(\overline{R_{i.}}) = \sigma^2\left(\frac{N-n_i}{n_i(N-1)}\right)$$

が成り立つ。

$\overline{R_{i.}} = \frac{R_{i.}}{n_i}$ は第 i 群の標本平均であるから，もし n_i があまり小さくないならば，中心極限定理により

$$Z_i = \left(\overline{R_{i.}} - \frac{N+1}{2}\right) \Big/ \sqrt{\sigma^2\left(\frac{N-n_i}{n_i(N-1)}\right)}$$

は近似的に標準正規分布にしたがう。故に $Z_i^2 \sim \chi^2$（自由度 1）。

ところで，このことは Z_i $(i=1,2,\ldots,s)$ の個々に対して成り立つが，Z_i $(i=1,2,\ldots,s)$ について，$\sum_{i=1}^{s} n_i \overline{R_{i.}} = \frac{N(N+1)}{2}$ が成り立っているから，互いに独立ではない。従って，もし n_i $(i=1,2,\ldots,s)$ が非常に小さくはないならば，確率変数

$$\sum_{i=1}^{s}\left(\frac{N-n_i}{N}\right)Z_i^2 = \left(\frac{1}{\sigma^2}\right)\left(\frac{N-1}{N}\right)\sum_{i=1}^{s} n_i\left(\overline{R_{i.}} - \frac{N+1}{2}\right)^2$$

$$=\left(\frac{1}{\sigma^2}\right)\left(\frac{N-1}{N}\right)\sum_{i=1}^{s}\left(\frac{1}{n_i}\right)\left(R_{i\cdot}-\frac{n_i(N+1)}{2}\right)^2$$

$$=GH$$

は近似的に自由度 s－1 の χ^2 分布に従うことが導かれる。

4．数値例

実験対象を無作為に 3 群に分け，第 i 群に第 i 処理を施して，それらの処理効果を検討する一元配置モデルにおいて，各個体についての観測値が第 1 表のように，区間データとして表示されているとする。

表 1　区間データとその一般化順位

群	サイズ	区　間　デ　ー　タ	一　般　化　順　位
1	4	(18,21), (22,26), (28,33), (37,42)	1.000, 2.150, 3.958, 7.000
2	4	(24,29), (33,38), (41,46), (45,49)	2.906, 5.389, 8.664, 10.067
3	4	(34,39), (40,45), (47,52), (57,62)	5.778, 8.239, 10.850, 12.000

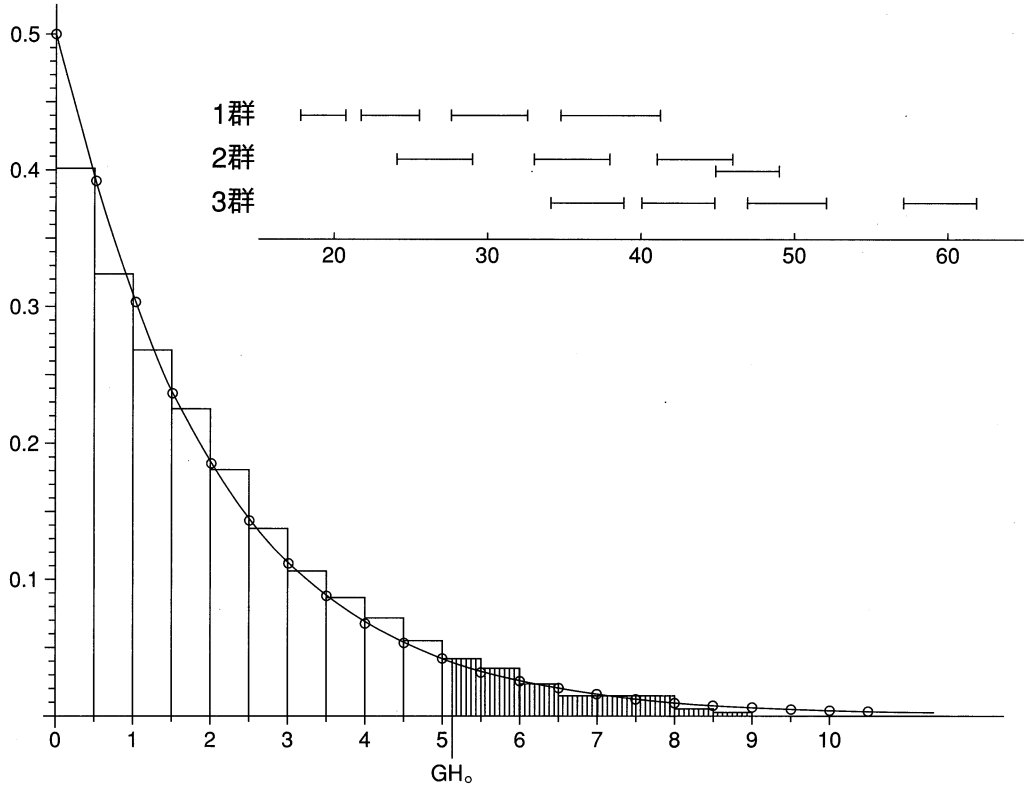

図 1　区間データおよび DH 値のヒストグラムと自由度 2 の χ^2 曲線

このとき，帰無仮説 Ho：各群の処理効果はすべて等しい　を　Ha：処理効果は等しくないに対して一般化 Kruskal-Wallis 検定を実施した。

一般化順位の算定に当たっては，区間データの密度関数 $f_i(x)$ はすべて一様分布を仮定した。その結果は第1表の通りである。

3群への配分分割の仕方は34650通り。12個の一般化順位母集団の平均と分散は $\mu=6.5$，$\sigma^2=11.6799$。また，GHo＝5.112 である。そして最後に有意確率 (P-値)＝$\frac{2400}{34650}=0.0693$ を得た。

従って，有意水準を0.05とすれば，Ho を Ha に対して，棄却することは出来なかった。ちなみに，区間データ，および GH 値のヒストグラムと自由度 s－1＝2 の χ^2 曲線は第1図の通りで，ヒストグラムと近似分布の χ^2 曲線はかなりよく一致している。

References

1) Kruskal, W. H. and W. A. Wallis: Use of Ranks in One Criterion Analysis of Variance, JASA 47:583-621, 1952　errate, ibid, JASA 48:907-911, 1953
2) Kariya, T.: A generalized sign test based on paired interval data. The second Japan-China symposium on statistics:125-128, 1986
3) 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
4) Kariya, T.: A generalized Wilcoxon test for comparing interval data samples, Kawasaki Med J 14:187-192, 1988
5) Kariya, T.: A bivariate permutation test for analysis of three interval data samples, Kawasaki Med J 18:25-30,1992
6) Kariya, T.: A permutation S-sample test against ordered alternatives based on interval data samples, Kawasaki Med J 21:19-24, 1995
7) 仮谷太一：区間データにもとづく一般化ラページ検定（2標本問題），川崎医学会誌一般教養篇，26：11-16，2000
8) 仮谷太一：区間データにもとづく Kolomogorov-Smirnov の検定（2標本問題），川崎医学会誌一般教養篇，27：11-16，2001