如何正确运用χ^2检验——秩和检验与SAS实现

胡纯严[1]，胡良平[1,2*]

（1. 军事科学院研究生院，北京 100850；
2. 世界中医药学会联合会临床科研统计学专业委员会，北京 100029
*通信作者：胡良平，E-mail：lphu927@163.com）

【摘要】 本文目的是介绍秩和检验及其SAS实现，具体内容包括以下三个方面：①两样本资料的简单线性秩检验；②多样本资料的单因素ANOVA秩和检验；③前述两种情形下的“评分方法”。在前述提及的第三方面内容中，包含十种具体的评分方法。本文结合一个实例并借助SAS软件实现前述提及的第一类秩和检验，对输出结果做出解释，并给出统计结论和专业结论。

【关键词】 定量资料；秩和检验；评分；标准正态分布；χ^2分布

中图分类号：R195.1 **文献标识码**：A doi：10.11886/scjsws20210915003

How to use χ^2 test correctly——the rank sum tests and the implementation of the SAS software

Hu Chunyan[1], *Hu Liangping*[1,2*]

(1. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*
**Corresponding author: Hu Liangping, E-mail: lphu927@163.com*)

【**Abstract**】 The purpose of this article was to introduce the rank sum test and its SAS implementation. The concrete contents included the following three aspects: ①the simple linear rank test of two-sample data, ②the one-way ANOVA rank sum test of the multi-sample data, ③the "scoring methods" in the aforementioned two situations. In the third aspect above, there were 10 concrete scoring approaches. Based on an example and with the aid of the SAS software, the paper implemented the first-type rank sum test previously, explained the output results, and made statistical and professional conclusions.

【**Keywords**】 Quantitative data; Rank sum test; Score; Standard normal distribution; χ^2 distribution

在对单因素资料进行差异性分析时，若发现资料不满足参数检验的前提条件，宜选用适用面宽的非参数检验。其中，秩和检验可能是常被选用的方法之一。本文将介绍两样本资料秩和检验[1-6]、多样本资料秩和检验[1-6]以及前述两种情形下都可能会用到的10种评分方法[1]。

1 两样本资料秩和检验

1.1 简单线性秩检验

1.1.1 未分层资料的简单线性秩检验

未分层资料的简单线性秩检验的检验统计量[1]见式(1)：

$$z=\frac{S-E_0(S)}{Var_0(S)}\sim N(0,1) \tag{1}$$

在式(1)中，z是一个服从标准正态分布的检验统计量(即随机变量)，S、$E_0(S)$和$Var_0(S)$分别代表未分层资料的“简单线性秩统计量”“简单线性秩统计量的期望值”和“简单线性秩统计量的方差”，定义分别见式(2)、式(3)、式(4)：

$$S=C_j a\left(R_j\right) \tag{2}$$

$$E_0(S)=\frac{n_1}{n}\sum_{j=1}^{n}a\left(R_j\right) \tag{3}$$

$$Var_0(S)=\frac{n_1 n_2}{n(nS-1)}\sum_{j=1}^{n}\left[a\left(R_j\right)-\bar{a}\right]^2 \tag{4}$$

在上述三式中，各变量的含义如下：R_j是第j个观测或个体的秩；$a(R_j)$是基于第j个观测的秩的得分(注意：评分的具体方法有多种，将在后文介绍)；C_j是指示变量代表第j个观测所在的组；n代表总观测数(即总样本含量)；n_1代表第1组(较小样本含量)样本含量；n_2代表第2组样本含量；$\bar{a}=\frac{1}{n}\sum_{j=1}^{n}a\left(R_j\right)$是平均得分。

基于标准正态分布理论和式(1)计算所得到的z值，可计算出与z对应的正态分布曲线下尾端的概率P值，右单侧概率、左单侧概率和双侧概率分别见式(5)、式(6)、式(7)：

$$P_1(z)=\text{Prob}(Z>z)\text{, 如果}z>0 \quad (5)$$

$$P_1(z)=\text{Prob}(Z<z)\text{, 如果}z\leqslant 0 \quad (6)$$

$$P_2(z)=\text{Prob}(|Z|>|z|) \quad (7)$$

连续性校正：当基于 Wilcoxon 和 Siegel-Tukey 评分且进行渐近双侧检验时，SAS/STAT 的 NPAR1WAY 过程将默认进行连续性校正，即当式(1)分子的计算结果大于0时，分子减掉0.5；当式(1)分子的计算结果小于0时，分子加上0.5。若想取消连续性校正，需要在“PROC NPAR1WAY”语句中增加选项“CORRECT=NO”。

1.1.2 分层资料的简单线性秩检验

若资料中有一个分层因素(通常称其为重要非试验因素)，设其有 K(K>2)个水平，在分层因素的每个水平下，试验因素均有两个水平(即两个对比组)。于是，分层资料的简单线性秩检验的检验统计量[1]见式(8)：

$$z=\frac{T-E_0(T)}{Var_0(T)}\sim N(0,1) \quad (8)$$

式(8)中，z是一个服从标准正态分布的检验统计量(即随机变量)，T、$E_0(T)$和$Var_0(T)$分别代表分层资料的“简单线性秩统计量”“简单线性秩统计量的期望值”和“简单线性秩统计量的方差”，其定义分别见式(9)、式(10)、式(11)：

$$T=\sum_{k=1}^{K}\frac{S_k}{w_k} \quad (9)$$

$$E_0(T)=\sum_{k=1}^{K}\frac{E_0(S_k)}{w_k} \quad (10)$$

$$Var_0(T)=\sum_{k=1}^{K}\frac{Var_0(S_k)}{w_k^2} \quad (11)$$

在以上三式中，S_k为第k层的“简单线性秩统计量”，w_k为第k层的权重，其定义见式(12)：

$$w_k=\frac{1}{n_k+1} \quad (12)$$

在式(12)中，n_k为第k层的样本含量。如果在“STRATA”语句中，指定“WEIGHTS=STRATUM”，则$w_k=1/(n_k+1)$；如果指定“WEIGHTS=EQUAL”，则w_k=1。

1.2 Fligner-Policello 检验

1.2.1 概述

Fligner 和 Policello 于 1981 年提出的比较两组定量资料中位数的检验方法[1]，被称为“Fligner-Policello 检验法”。该法假定每组定量资料服从对称分布，但不要求两组定量资料具有相同的分布，也不要求两组定量资料的方差相等。该法是基于 Orban 和 Wolfe 于1979年提出的“配置得分”而构建。设有X与Y两个组，来自X组的观测X_i的配置得分记为$P(X_i)$，其取值定义如下：$P(X_i)$=Y组中取值小于X_i的数据个数；如果遇到相等的数值，需要对$P(X_i)$进行校正，即在已知$P(X_i)$的基础上增加Y组中取值等于X_i的数据个数的一半。对来自Y组的观测Y_j的配置得分记为$P(Y_j)$，其取值定义与$P(X_i)$相同。

1.2.2 配置得分的定义

$$P(X_i)=\sum_{j=1}^{n_Y}\left[\text{I}(Y_j<X_i)+0.5\text{I}(Y_j=X_i)\right] \quad (13)$$

$$P(Y_j)=\sum_{i=1}^{n_X}\left[\text{I}(X_i<Y_j)+0.5\text{I}(X_i=Y_j)\right] \quad (14)$$

在式(13)、式(14)中，n_X和n_Y分别代表X组与Y组的样本含量；I(·)是指示函数。于是，两组各自的平均配置得分的计算公式分别见式(15)、式(16)：

$$\bar{P}_X=\frac{1}{n_X}\sum_{i=1}^{n_X}P(X_i) \quad (15)$$

$$\bar{P}_Y=\frac{1}{n_Y}\sum_{j=1}^{n_Y}P(Y_j) \quad (16)$$

1.2.3 Fligner-Policello 检验统计量

Fligner-Policello 检验统计量见式(17)：

$$z=\frac{\sum_{j=1}^{n_Y}P(Y_j)-\sum_{i=1}^{n_X}P(X_i)}{2\sqrt{V_X+V_Y+\bar{P}_X\bar{P}_Y}}\sim N(0,1) \quad (17)$$

在式(17)中，z是一个服从标准正态分布的检验统计量(即随机变量)；V_X和V_Y的计算分别见式(18)、式(19)：

$$V_X=\sum_{i=1}^{n_X}\left[P(X_i)-\bar{P}_X\right]^2 \quad (18)$$

$$V_Y=\sum_{j=1}^{n_Y}\left[P(Y_j)-\bar{P}_Y\right]^2 \quad (19)$$

X和Y与两个组的配置得分的标准差分别见式(20)、式(21)：

$$SD_X=\sqrt{\frac{V_X}{n_X-1}} \quad (20)$$

$$SD_Y=\sqrt{\frac{V_Y}{n_Y-1}} \quad (21)$$

【说明】P值的定义与式(5)、式(6)、式(7)相同，此处从略。

2 多样本资料秩和检验

2.1 概述

对多组定量资料进行比较的秩和检验法常有下面两个名称，第一个叫做“单因素 ANOVA 检验”；第二个叫做“Kruskal-Wallis 检验（采取 Wilcoxon 评分法）”。其实，它们本质上都属于“χ^2检验”。当对多组定量资料进行整体比较时，其检验假设为“H_0：各组之间没有差别”。

2.2 多组之间的整体比较

设有一个具有 r 个水平的试验因素，对定量资料进行 r 组之间的整体比较时，所需要的检验统计量[1]见式（22）：

$$C=\frac{\sum_{i=1}^{r}\left[T_i-E_0\left(T_i\right)\right]^2\Big/n_i}{S^2}\sim\chi^2_{r-1} \tag{22}$$

在式（22）中，C 是一个服从自由度为 $df=r-1$ 的 χ^2 分布的检验统计量；n_i 是第 i 个水平组的样本含量；T_i 是第 i 个水平组的得分之和；$E_0(T_i)$ 是在 H_0 成立的条件下第 i 个水平组的期望秩和；S^2 是得分的样本方差。T_i、$E_0(T_i)$ 和 S^2 的计算公式分别见式（23）、式（24）、式（25）：

$$T_i=\sum_{j=1}^{n}C_{ij}a\left(R_j\right) \tag{23}$$

$$E_0\left(T_i\right)=\frac{n_i}{n}\sum_{j=1}^{n}a\left(R_j\right) \tag{24}$$

$$S^2=\frac{1}{(n-1)}\sum_{j=1}^{n}\left[a\left(R_j\right)-\bar{a}\right]^2 \tag{25}$$

在式（25）中，$\bar{a}=\frac{1}{n}\sum_{j=1}^{n}a\left(R_j\right)$ 是平均得分。

2.3 多组之间的两两比较

2.3.1 概述

由 Dwass、Steel、Critchlow 和 Fligner 提出的多重比较方法，简称为“DSCF 检验法”。此法从 $r(r>2)$ 个组中每次抽取两组进行比较，故总共需要比较 $r\times(r-1)/2$ 次。每次比较都基于标准化的威尔科克森检验统计量，即采取威尔科克森法评分，并采用式（1）计算 z 统计量。

2.3.2 DSCF 检验统计量

基于标准化的威尔科克森 z 检验统计量构造出 DSCF 检验统计量见式（26）：

$$\mathrm{DSCF}=\sqrt{2}\,z \tag{26}$$

在式（26）中，z 是采取威尔科克森法评分，并采用式（1）计算的结果（注意：每次比较只涉及两组定量资料）；而 DSCF 是一个近似服从于“r 个标准正态变量的学生化极差分布”的检验统计量。两样本 DSCF 比较的 P 值可以通过下面的方法求出，即将 DSCF 统计量的值视为学生化极差分布的百分位数，从而，基于学生化极差分布下特定百分位数计算出分布曲线下尾端的概率，即为所求的 P 值。

3 秩和检验中的评分方法

3.1 概述

秩和检验的一个特点就是不直接利用原始数据，而是先根据原始数据的大小给它们编秩。所谓编秩，就是给每个原始数据赋予一个自然数，代表每个原始数据在一组和整个资料中的“相对位置”。然后再依据不同的数学原理，对每个“秩”进行“评分”或“赋值”。SAS/STAT 的 NPAR1WAY 过程[1]中介绍了十多种评分方法，现呈现其主要内容。

3.2 用于位置比较的评分方法

3.2.1 威尔科克森（Wilcoxon）评分法

威尔科克森评分是观测的秩，可用公式表示如下：

$$a\left(R_j\right)=R_j \tag{27}$$

在式（27）中，R_j 是第 j 个观测的秩，而 $a(R_j)$ 是第 j 个观测的评分。

【说明】在两样本资料的线性秩统计量中采用威尔科克森评分产生 Mann-Whitney-Wilcoxon 检验的秩和统计量；在多样本资料的单因素 ANOVA 统计量中采用威尔科克森评分产生 Kruskal-Wallis 检验的秩和统计量；对于 logistic 分布的位置改变来说，威尔科克森评分是局部最有效能的。

3.2.2 中位数（Median）评分法

当资料中的观测值大于中位数时，则该观测的中位数评分等于 1；否则，中位数评分等于 0。依据观测的秩，中位数评分的定义见下式：

$$a\left(R_j\right)=\begin{cases}1 & \text{如果}R_j>(n+1)/2\\ 0 & \text{如果}R_j\leqslant(n+1)/2\end{cases} \tag{28}$$

【说明】在两样本资料的线性秩统计量中采用中位数评分产生两样本中位数检验的秩和统计量；在多样本资料的单因素 ANOVA 统计量中采用中位

数评分产生 Brown-Mood 检验的秩和统计量；中位数评分用于尾部抬高且对称分布时，效能特别高。

3.2.3 Van der Waerden（正态）评分

Van der Waerden 评分是标准正态分布的分位数，也被称为分位数正态评分。该评分的计算公式见式（29）：

$$a\left(R_j\right)=\Phi^{-1}\left(\frac{R_j}{n+1}\right) \tag{29}$$

在式（29）中，Φ 是标准正态分布的累计分布函数。对于正态分布而言，这些评分的效能极高。

3.2.4 Savage 评分

Savage 评分是来自指数分布的顺序统计量的期望值，通过减掉 1 使评分的中心位于 0 附近。该评分的计算公式见式（30）：

$$a\left(R_j\right)=\sum_{i=1}^{R_j}\left(\frac{1}{n-i+1}\right)-1 \tag{30}$$

Savage 评分在以下两种情形中具有高效能，其一，在指数分布中比较尺度差异；其二，在极值分布中比较位置变化。

3.3 用于尺度比较的评分方法

3.3.1 Siegel-Tukey 评分

Siegel-Tukey 评分的定义如下：

$a(1)=1, a(n)=2, a(n-1)=3, a(2)=4$

$a(3)=5, a(n-2)=6, a(n-3)=7, a(4)=8, \cdots$

这里得分值按此模式朝着中间连续增加，直到全部观测中的每个观测都被赋予一个得分值为止。

【说明】当进行 Siegel-Tukey 两样本检验的计算时，SAS/STAT 中 NPAR1WAY 过程默认需要进行校正；如果用户不想进行校正，需要在“PROC NPAR1WAY”语句中增加选项“CORRECT=NO”。

3.3.2 Ansari-Bradley 评分

Ansari-Bradley 评分为对应的极端秩赋予相同的得分，其定义如下：

$a(1)=1, a(n)=1, a(2)=2, a(n-1)=2$

$a(3)=3, a(n-2)=3, a(4)=4, a(n-3)=4, \cdots$

等价地，Ansari-Bradley 评分可用如下通式表示：

$$a\left(R_j\right)=\frac{n+1}{2}-\left|R_j-\frac{n+1}{2}\right| \tag{31}$$

3.3.3 Klotz 评分

Klotz 评分是 Van der Waerden 评分的平方，其定义如下：

$$a\left(R_j\right)=\left[\Phi^{-1}\left(\frac{R_j}{n+1}\right)\right]^2 \tag{32}$$

在式（32）中，Φ 是标准正态分布的累计分布函数。

3.3.4 Mood 评分

Mood 评分按照观测的秩与平均秩之差量的平方进行计算，其定义如下：

$$a\left(R_j\right)=\left(R_j-\frac{n+1}{2}\right)^2 \tag{33}$$

3.4 用于位置和尺度比较的评分方法

Conove 评分是基于观测值与其样本算术平均值之离差绝对值的秩的平方，对于第 j 个观测而言，其定义如下：

$$a\left(U_j\right)=\left[\mathrm{Rank}\left(U_j\right)\right]^2 \tag{34}$$

在式（34）中，U_j 的计算见式（35）：

$$U_j=\left|X_{j(i)}-\bar{X}_i\right| \tag{35}$$

在式（35）中，i 代表第 i 个样本（组）；j 代表第 i 个样本中第 j 个观测；$X_{j(i)}$ 代表第 i 样本中第 j 个观测的观测值；$\bar{X}_i$ 代表第 i 个样本的算术平均值；U_j 代表第 i 个样本中第 j 个观测的秩。

【说明】Conove 于 1999 年提出，若在第 i 个样本的全部 U_j 中出现了相同的数值（称为“ties”），则先按无相同数据编秩（即给予编号），再求那几个相同数据所对应秩的算术平均值，并以此平均值作为它们的“秩”。Conove 评分检验也被称为“方差的平方秩检验”。

4 实例与 SAS 实现

4.1 问题与数据

【例 1】某地 59 例女性类风湿性关节炎患者参加了一项临床试验[1]，她们被随机分配进入试验组（n=27）与安慰剂对照组（n=32）。结果变量有 5 种不同的取值，即“疗效特好=5”“疗效尚好=4”“疗效中等=3”“疗效一般=2”和“疗效差=1”。记录每位患者所接受的处理和疗效的具体取值，临床试验结果以频数表形式呈现，详见后面的 SAS 程序，此处从略。试对两组有序资料进行秩和检验，以评价两种治疗方法的效果差异是否有统计学意义。

4.2 SAS实现

4.2.1 对例1的SAS实现

【分析与解答】设所需要的SAS程序如下：

```
data Arthritis;
input Treatment $ Response Freq @@;
datalines;
Active 5 5 Active 4 11 Active 3 5 Active 2 1 Active 1 5
Placebo 5 2 Placebo 4 4 Placebo 3 7 Placebo 2 7 Placebo 1 12
;
run;
proc npar1way data=Arthritis ab conover
klotz median mood
savage st vw
wilcoxon fp;
class Treatment;
var Response;
freq Freq;
run;
```

【程序说明】“从ab到fp”这10个选项是秩和检验中的10种评分方法；其中，“fp”是前文介绍的Fligner-Policello检验法。

【SAS输出结果及解释】因篇幅所限，以下仅呈现“威尔科克森检验结果”，其他检验方法输出的结果从略。

Treatment	数目	评分汇总	H0之下的期望值	H0之下的标准差	均值评分
Active	27	999.0	810.0	63.972744	37.000000
Placebo	32	771.0	960.0	63.972744	24.093750

以上是威尔科克森检验输出的第1部分结果，即输出两组描述性统计量的计算结果，治疗组的平均秩为37.00分，对照组的平均秩为24.09分。

Wilcoxon 双样本检验

计量	Z	Pr>Z	Pr>|Z|	t近似值	
				Pr>Z	Pr>|Z|
999.000	2.9466	0.0016	0.0032	0.0023	0.0046

以上是威尔科克森检验输出的第2部分结果，即以标准正态分布为理论根据计算得到的结果，双侧检验的P=0.0046<α=0.05。

Kruskal-Wallis 检验

卡方	自由度	Pr>卡方
8.7284	1	0.0031

以上是威尔科克森检验输出的第3部分结果，即以χ^2分布为理论根据计算得到的结果，P=0.0031<α=0.05。

【统计结论与专业结论】由输出结果可知，前4种检验方法和第10种检验方法给出的检验结果均为P<0.05，说明治疗组与对照组疗效的“平均秩”或“中位数”之间差异有统计学意义；由于治疗组的“平均秩”或“中位数”大于对照组的“平均秩”或“中位数”，又因评分值越大标志着疗效越好，故可以认为“active”治疗方法的效果优于安慰剂。

第5~8种检验方法的检验结果均为P>0.05，说明治疗组与对照组疗效的“离散度（即尺度参数）”之间差异无统计学意义，即两组有序资料的变化范围接近一致。

第9种检验方法的检验结果为P>0.05，说明治疗组与对照组疗效的“位置参数”和“离散度（即尺度参数）”综合的指标之间差异无统计学意义。

5 讨论与小结

5.1 讨论

秩和检验有两个优点：其一，对资料的要求不高；其二，选择不同的评分方法可以分别实现“位置参数（如平均值、平均秩、中位数）”“尺度参数（如标准差、分位数间距）”和“位置参数以及尺度参数”的比较。其缺点在于：不适合分析多因素资料。为了保留对资料要求不高的优点，又能够处理多因素资料，需要选择复杂的非参数统计分析方法[6-8]。

5.2 小结

本文介绍了适用于分析单因素资料的秩和检验方法，包括分析单因素两水平设计资料的“简单线性秩检验”和单因素多水平设计资料的“单因素多水平ANOVA检验”。详细介绍了在前述两类检验中都不可缺少的10种评分方法。通过一个实例并借助SAS软件，实现了单因素两水平设计资料的简单线性秩检验，呈现了10种评分方法计算所得到的结果，对输出结果作出了解释，并给出了统计结论和专业结论。

参考文献

[1] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 9119-9194.

[2] 王广义, 薛禾生. 中国医学统计百科全书非参数统计分册[M]. 北京: 人民卫生出版社, 2004: 40-81.

[3] 茆诗松. 统计手册[M]. 北京: 科学出版社, 2003: 184-226.

[4] 罗斯纳, 孙尚拱. 生物统计学基础[M]. 北京:科学出版社, 2004: 317-339.

[5] Davison AC, Hinkley DV. Bootstrap methods and their application[M]. Cambridge University Press, 1997: 136-190.

[6] 日本数学会. 数学百科辞典[M]. 北京: 科学出版社, 1984: 1210-1214.

[7] 沃塞曼. 现代非参数统计[M]. 吴喜之, 译. 北京: 科学出版社, 2008: 22-186.

[8] 方积乾, 陆盈. 现代医学统计学[M]. 北京: 人民卫生出版社, 2002: 577-607.

(收稿日期:2021-09-15)

(本文编辑:陈 霞)

科研方法专题策划人——胡良平教授简介

胡良平,男,1955年8月出生,教授,博士生导师,曾任军事医学科学院研究生部医学统计学教研室主任和生物医学统计学咨询中心主任、国际一般系统论研究会中国分会概率统计系统专业理事会常务理事、中国生物医学统计学会副会长、北京大学口腔医学院客座教授和《中华医学杂志》等10余种杂志编委;现任世界中医药学会联合会临床科研统计学专业委员会会长、国家食品药品监督管理局评审专家和3种医学杂志编委;主编统计学专著48部、参编统计学专著10部;发表第一作者和通信作者学术论文300余篇、发表合作论文130余篇;获军队科技成果和省部级科技成果多项;参加并完成三项国家标准的撰写工作、参加三项国家科技重大专项课题研究工作。在从事统计学工作的30年中,为几千名研究生、医学科研人员、临床医生和杂志编辑讲授生物医学统计学,在全国各地作统计学学术报告100余场,举办数十期全国统计学培训班,培养20多名统计学专业硕士和博士研究生。近几年来,参加国家级新药和医疗器械项目评审数十项、参加100多项全军重大重点课题的统计学检查工作。归纳并提炼出有利于透过现象看本质的"八性"和"八思维"的统计学思想,独创了逆向统计学教学法和三型理论。擅长于科研课题的研究设计、复杂科研资料的统计分析和SAS与R软件实现、各种层次的统计学教学培训和咨询工作。