如何正确运用*t*检验——相关系数与0比较*t*检验及SAS实现

宋德胜[1],刘媛媛[1],李长平[1,2],崔 壮[1],胡良平[2,3*]

(1. 天津医科大学公共卫生学院,天津 300070;
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;
3. 军事科学院研究生院,北京 100850
*通信作者:胡良平,E-mail:lphu927@163.com)

【摘要】 本文目的是介绍Pearson相关系数、Spearman秩相关系数和Kendall's tau-b秩相关系数的概念及应用场合。首先介绍三种相关系数的计算及其假设检验的基本原理,然后使用SAS程序对实例进行分析,最后对分析结果进行解释和讨论。

【关键词】 相关分析;Pearson相关系数;Spearman秩相关系数;Kendall's tau-b秩相关系数;*t*检验
中图分类号:R195.1 **文献标识码**:A doi:10.11886/scjsws20200717005

How to use *t* test correctly——comparison of correlation coefficient with 0 *t* test and SAS implementation

Song Desheng[1], *Liu Yuanyuan*[1], *Li Changping*[1,2], *Cui Zhuang*[1], *Hu Liangping*[2,3*]

(1. *School of Public Health, Tianjin Medical University, Tianjin* 300070, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*;
3. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*
**Corresponding author: Hu Liangping, E-mail: lphu927@163.com*)

【Abstract】 The purpose of this article was to introduce the concepts and application situation of Pearson's correlation coefficient, Spearman's rank correlation coefficient and Kendall's tau-b rank correlation coefficient. Firstly, this article introduced the basic statistical principles of the three kinds of correlation coefficients, then SAS software was used to perform the correlation analysis of the example, and finally the analyzed results were explained and discussed.

【Keywords】 Correlation analysis; Pearson's correlation coefficient; Spearman's rank correlation coefficient; Kendall's tau-b rank correlation coefficient; *t* test

相关分析的目的是揭示两个变量之间是否存在线性相关性。目前较常见的两个随机变量之间的相关性度量指标是Pearson乘积矩相关系数,简称Pearson相关系数,其公式表示的是实际数据与期望数据的偏离程度[1]。与Pearson相关系数不同,Spearman秩相关系数和Kendall's tau-b秩相关系数利用了秩。这两种秩相关系数一般是当资料不满足Pearson相关系数时的替代方法。从几何学角度看,Pearson相关系数测量的是两个定量变量之间呈线性相关的程度;而两种秩相关系数则不限于线性相关。但三者具有类似的性质:①相关系数的范围都是[-1,1];②都具有对称性。值得注意的是,不管是正相关还是负相关,都不涉及"因果关系"。本文对这三种相关系数的概念、作用以及应用进行介绍。

基金项目:国家自然科学基金项目(项目名称:贝叶斯生存分析方法在肝细胞癌肝移植患者预后预测中的应用研究,项目编号:81803333)

1 基本原理

1.1 Pearson相关系数与0比较*t*检验

Pearson乘积矩相关系数由Karl Pearson提出[2]。它是两个定量变量相关的一种参数化测量,既可计算相关强度,也可得出相关方向。若两个定量变量呈完全正线性相关,则Pearson相关系数为1;若两个定量变量呈完全负线性相关,则Pearson相关系数为-1;若两个定量变量不呈线性相关,则Pearson相关系数为0。因此,Pearson相关系数的取值范围为[-1,1]。此外,Pearson相关系数要求两个定量变量服从二元正态分布、两个定量变量每一对取值应来自同一个个体,且所有受试对象应抽自满足"同质性"的同一个总体。若不满足这些前提条件,则不适合计算任何相关系数;仅当不满足"双变量正态分布"时,可考虑进行秩相关分析。Pearson相关系数计算公式如下:

$$r_{xy}=\frac{Cov(x,y)}{\sqrt{V(x)V(y)}}=\frac{E\{[x-E(x)][y-E(y)]\}}{\sqrt{E[x-E(x)]^2E[y-E(y)]^2}} \quad (1)$$

一般来说，上述统计量来自于样本，若需推论到总体，则需要进行假设检验。经统计学家证明，Pearson 相关系数服从自由度为 $n-2$ 的 t 分布。检验统计量 t 的计算公式如下：

$$t=(n-2)^{1/2}\left(\frac{r^2}{1-r^2}\right)^{1/2} \quad (2)$$

上式中，r 是来自样本的 Pearson 相关系数。

1.2 Spearman 秩相关系数与 0 比较 t 检验

Spearman 秩相关系数是 Charles Spearman 提出的一个一般非参数统计量，通常用 r_s 表示[3]。它使用单调函数度量两个变量的关系。与 Pearson 相关系数类似，Spearman 秩相关系数的取值范围也为[-1，1]。实际上，Spearman 秩相关系数等同于两个变量秩次值的 Pearson 相关系数，既可以用于连续型随机变量，也可用于离散有序随机变量。计算公式如下：

$$\theta=\frac{\sum_i\left[\left(R_i-\overline{R}\right)\left(S_i-\overline{S}\right)\right]}{\sqrt{\sum_i\left(R_i-\overline{R}\right)^2\Sigma\left(S_i-\overline{S}\right)^2}} \quad (3)$$

上式中，R_i 是 x_i 的秩次，S_i 是 y_i 的秩次，$\overline{R}$ 是 R_i 的均值，$\overline{S}$ 是 S_i 的均值。若存在结（即同一个变量的原始数据中存在 2 个或 2 个以上相同的数据），则使用平均秩。

对来自于样本的 Spearman 秩相关系数仍需假设检验才可以推广到总体。Spearman 秩相关系数仍服从自由度为 $n-2$ 的 t 分布。检验统计量 t 的计算公式如下：

$$t=(n-2)^{1/2}\left(\frac{r_s^2}{1-r_s^2}\right)^{1/2} \quad (4)$$

1.3 Kendall's tau-b 秩相关系数与 0 比较 Z 检验

Kendall's tau-b 秩相关系数也称为 Kendall's τ 系数，是用于测量两个观测指标之间秩相关的指标，由 Maurice Kendall 提出，可用于衡量配对设计扩大条件下两属性变量等级之间的相关程度[4]。本质上，两个变量之间秩的分布越相似，则 Kendall's tau-b 秩相关系数越大，它也是一种一般的非参数统计量，因为它不依赖于两个变量的分布。由于 Kendall's tau-b 秩相关系数对结进行了处理，因此，其取值范围为[-1，1]。Kendall's tau-b 秩相关系数计算公式[5]如下：

$$\tau_b=\frac{n_c-n_d}{\sqrt{(n_0-n_1)(n_0-n_2)}} \quad (5)$$

上式中，$n_0=n(n-1)/2$，$n_1=\sum_i t_i(t_i-1)/2$，$n_2=\sum_j \mu_j(\mu_j-1)/2$。$n_c$ 表示变化一致对子数，n_d 表示变化不一致对子数，t_i 和 t_j 分别是两个变量的第 i 组结的值的数量。与 Pearson 相关系数和 Spearman 秩相关系数类似，对 Kendall's tau-b 秩相关系数也需要进行假设检验才可将样本结论推论到总体。Kendall's tau-b 秩相关系数的检验统计量计算公式如下：

$$Z=\frac{s}{\sqrt{V(s)}} \quad (6)$$

Z 服从标准正态分布。其中，

$$s=\sum_{i<j}\left[sgn(x_i-x_j)sgn(y_i-y_j)\right] \quad (7)$$

$$V(s)=\frac{v_0-v_t-v_u}{18}+\frac{v_1}{2n(n-1)}+\frac{v_2}{9n(n-1)(n-2)} \quad (8)$$

在这里，

$$v_0=n(n-1)(2n+5)$$

$$v_t=\sum_k t_k(t_k-1)(2t_k+5)$$

$$v_u=\sum_l u_l(u_l-1)(2u_l+5)$$

$$v_1=\sum_k t_k(t_k-1)\Sigma u_i(u_l-1)$$

$$v_2=\sum_l t_i(t_k-1)(t_k-2)\Sigma u_l(u_l-1)(u_l-2)$$

2 实 例

2.1 Pearson 相关系数 r 与 0 比较 t 检验

【例 1】本例数据为 N. C. State University 关于男性身体健康课程的调查数据。数据集变量包括年龄（岁）、体重（kg）、跑步时长（跑步 1.5 英里所需时间，按分钟计）以及氧摄入量[mL/(kg•min)]。试分析氧摄入量与跑步时长之间的关系。资料见表 1。

表 1 男性身体健康调查数据

id	年龄	体重	氧摄入量	跑步时长
1	44	89.47	44.609	11.37
2	44	85.84	54.297	8.65
3	38	89.02	49.874	-
…	…	…	…	…
29	54	91.63	39.203	12.88
30	57	59.08	50.545	9.93
31	48	61.24	47.920	11.50

2.1.1 创建数据集

```
data example1;
input Age Weight Oxygen RunTime @@;
datalines;
44 89.47 44.609 11.37 40 75.07 45.313 10.07
44 85.84 54.297  8.65 42 68.15 59.571  8.17
38 89.02 49.874  .    47 77.45 44.811 11.63
40 75.98 45.681 11.95 43 81.19 49.091 10.85
……
57 73.37 39.407 12.63 54 79.38 46.080 11.17
52 76.32 45.441  9.63 50 70.87 54.625  8.92
51 67.25 45.118 11.08 54 91.63 39.203 12.88
51 73.71 45.790 10.47 57 59.08 50.545  9.93
49 76.32  .      .    48 61.24 47.920 11.50
52 82.78 47.467 10.50
;
run;
```

2.1.2 绘制氧摄入量与跑步时长的散点图

```
PROC SGPLOT data=example1;
Scatter x=Oxygen y=RunTime;
Run;
```

【程序说明】SGPLOT过程是SAS软件中的绘图过程。Scatter语句表示绘制散点图。之后的“x=”和“y=”分别指定需要绘制散点图的两个变量。运行结果见图1。

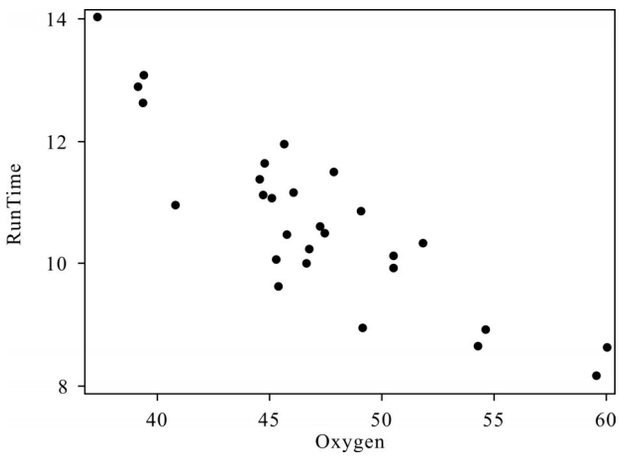

图1 Oxygen(氧摄入量)和Runtime(跑步时长)的散点图

由图1的散点图可知,Oxygen和Runtime之间大致呈线性负相关。也就是说,随着氧摄入量的增加,跑步到1.5英里所需时间逐步减少。因此,可以进行相关分析。

2.1.3 相关分析

```
PROC CORR data=example1 Pearson;
VAR Oxygen Runtime;
Run;
```

【程序说明】本例调用SAS软件中的CORR过程进行相关性分析。CORR过程语句后的data选项指定包含相关分析变量的数据集,pearson表示计算Pearson相关系数(若此处不指定任何相关系数,默认显示Pearson相关系数)。VAR语句指定需要进行相关分析的变量。

【SAS主要输出结果及解释】

简单统计量

变量	数目	均值	标准差	总和	最小值	最大值
Oxygen	29	47.22721	5.47718	1370	37.38800	60.05500
RunTime	29	10.67414	1.39194	309.55000	8.17000	14.03000

Pearson相关系数,Prob>|r|,H0:Rho=0

	Oxygen	RunTime
Oxygen	1.00000	−0.86843
		<0.0001
	29	28
RunTime	−0.86843	1.00000
	<0.0001	
	28	29

输出结果中,首先给出了两个变量的一些简单统计描述结果,接着给出了Pearson相关系数及其假设检验结果。Pearson相关系数分析结果的右上方由上到下分别是样本Pearson相关系数、对Pearson相关系数进行假设检验对应的P值以及排除缺失值后用于分析的样本例数。可以看到,在本例中,氧摄入量和跑步至1.5英里所需时间呈负相关(相关系数为−0.86843),且t检验的P值小于0.0001,因此可认为随着氧摄入量的增加,跑步至1.5英里所需时间呈线性下降趋势。说明:输出结果中未给出检验统计量t的数值。

2.2 Spearman秩相关系数r_s及与0比较t检验

【例2】某研究机构收集了成年人年龄和身体脂肪百分比的数据,本例选取了其中18例数据。分析年龄与身体脂肪百分比的关系。见表2。

表 2 18 名成年人年龄和身体脂肪百分比数据

调查对象	年龄(岁)	身体脂肪百分比
01	23	9.5
02	23	27.9
03	27	7.8
…	…	…
16	58	33.8
17	60	41.1
18	61	34.5

2.2.1 创建数据集

```
data example2;
Input subject age bodyfat_perc;
cards;
01 23 9.5
02 23 27.9
03 27 7.8
04 27 17.8
……
15 58 33.0
16 58 33.8
17 60 41.1
18 61 34.5
;
run;
```

2.2.2 绘制散点图

```
PROC SGPLOT data=example2;
scatter x=age y=bodyfat_perc;
run;
```

【程序说明】与例 1 类似,绘制年龄与身体脂肪百分比的散点图见图 2。

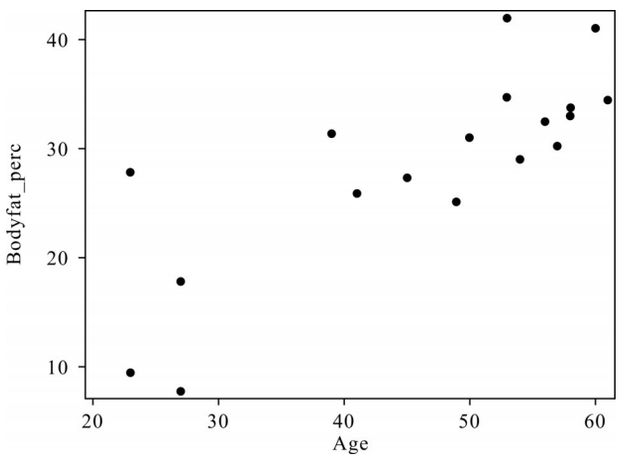

图 2 年龄(Age)与身体脂肪百分比(Bodyfat_perc)散点图

由图 2 可知,年龄与身体脂肪百分比线性趋势不明显,因此,本例将使用 Spearman 秩相关分析。

2.2.3 秩相关分析

```
PROC CORR data=example2 spearman;
VAR age;
with bodyfat_perc;
run;
```

【程序说明】本例也使用 PROC CORR 调用 SAS 软件的 CORR 相关分析过程。过程语句后的 Spearman 选项表示进行 Spearman 秩相关分析。VAR 语句指定要分析的变量,with 指定另一个需要分析的变量。

【SAS 主要输出结果及解释】

简单统计量

变量	数目	均值	标准差	中位数	最小值	最大值
Bodyfat_perc	18	28.61111	9.14439	30.70000	7.80000	42.00000
Age	18	46.33333	13.21764	51.50000	23.00000	61.00000

Spearman 相关系数,N=18 Prob>|r|,H0:Rho=0

	Age
Bodyfat_perc	0.75388
	0.0003

由结果可知,年龄和身体脂肪百分比的 Spearman 秩相关系数为 0.75388,P 值为 0.0003,由此可认为,随着年龄的增加,身体脂肪百分比也在上升。

说明:由于对 Kendall's tau-b 秩相关系数的假设检验是 Z 检验(即以正态分布为理论依据的检验),不是 t 检验,超出了本文的范围,故从略。值得一提的是:适合采用 Kendall's tau-b 秩相关分析的数据结构为"配对设计扩大形式的定性资料",可参阅文献[6]了解其方法及应用,此处从略。

3 讨论与小结

3.1 讨论

Spearman 秩相关系数通常适用于单组设计双变量且资料不符合 Pearson 相关分析的前提条件的场合,由于它是基于"秩次"计算得到的秩相关系数,故其精确度会有所降低。

Pearson 相关分析对资料的要求很高,通常要求资料为单组设计二元定量资料且两变量呈线性变

化趋势。但在实际使用中，前提条件可能会略有偏移，但应注意不能偏离过大。如张美燕等[7]利用其分析了精神科门诊患者使用四种量表评定之间的关系。

Kendall 秩相关系数有三种，分别是 Kendall's tau-a[8]、Kendall's tau-b 和 Kendall's tau-c[9]秩相关系数。本文主要探讨的是 Kendall's tau-b 秩相关系数，它与其他两类 Kendall 秩相关系数的主要区别在于对结的处理方法不同。因篇幅所限，详情从略。

3.2 小结

综上所述，在进行相关分析时，需根据数据的特点（特别是所满足的前提条件）和所取自的设计类型选择合适的相关分析方法。此外，相关分析的结果并不能代表变量之间的因果关系。如需进行因果判断，需要特殊的统计学方法。

参考文献

[1] Rodgers JL, Nicewander WA. Thirteen ways to look at the correlation coefficient[J]. The American Statistician, 1988, 42(1): 59-66.

[2] Pearson K. Notes on the history of correlation[J]. Biometrika, 1920, 13(1): 25-45.

[3] 胡良平. SAS常用统计分析教程[M]. 2版. 北京：电子工业出版社, 2015: 421-424.

[4] Kendall MG. A new measure of rank correlation[J]. Biometrika, 1938, 30: 81-93.

[5] Agresti A. Analysis of ordinal categorical data[M]. 2nd edition. New York: John Wiley & Sons, 2010: 188.

[6] 胡良平. 面向问题的统计学——(1)科研设计与统计基础[M]. 北京：人民卫生出版社, 2012: 549-563.

[7] 张美燕, 李小群, 邓先华, 等. 精神科门诊患者综合使用MMPI、SCL-90、SAS、SDS的相关分析[J]. 四川精神卫生, 2018, 31(4): 356-360.

[8] Kendall MG. The treatment of ties in ranking problems[J]. Biometrika, 1945, 33(3): 239-251.

[9] Stuart A. The estimation and comparison of strengths of association in contingency tables[J]. Biometrika, 1953, 40: 105-110.

（收稿日期：2020-07-17）

（本文编辑：戴浩然）