知乎 ··· 写文章

诊断试验的一致性检验，这么多方法，全都搞清楚了吗？

医小咖

关注她

3 人赞同了该文章

> 作者：李延龙

聊起诊断试验那可是应用相当广泛：评价两种方法或仪器诊断结果是否一致，得用到诊断试验；看看两个大夫对同一群病人诊断是否一致，要用诊断试验；评价同一组患者前后两次诊断结果的一致性，还得用到诊断试验，等等.....

简而言之，诊断试验关注的重点是“一致性”，也就是说同一个体用两种仪器（方法/评价者）或前后两次时间进行观测，其结果在误差允许范围内是一致的。**评价一致性程度的方法很多**，比如说Kappa值、Kendall一致性系数、组内相关系数（ICC）等等，但是选对合适的方法却不容易，接下来就帮大家梳理一番！

配对χ2检验 vs. 一致性检验

配对χ2检验（McNemar检验）和Kappa一致性检验都可以用于**配对设计的列联表分析**（表1），例如，比较超声和CT平扫对于急性阑尾炎的诊断价值，但是两者却各有侧重。

表 1. 配对数据形式——二分类变量

		B 方法		合计
		1	0	
A 方法	1	a	b	a + b
	0	c	d	c + d
合计		a + c	b + d	N

(1) 两者计算方法不同

$$Kappa = (Po - Pe)/(1 - Pe) \quad ①$$

观察一致率：$Po=(a+d)/n$ ②

机遇一致率：$Pe = [(a+b)(a+c) + (b+d)(c+d)]/n^2$ ③

$$\chi^2 = (b-c)^2/(b+c),\ v=1 \quad ④$$

由①②③可知在计算Kappa过程中，会利用到四格表中全部的数据（a、b、c、d），而公式④表明配对χ2检验只利用了四格表中“不一致”的数据（b和c）。

(2) 两者提供的信息不同

一致性检验不仅可以明确两种方法是否存在一致，更重要的是可以计算Kappa值，进而评价一致性的程度。目前认为，Kappa<0，一致性强度极差（实际情况下发生可能性较低）；0-0.20，微弱；0.21-0.40，弱；0.41-0.60，中度；0.61-0.80，高度；0.81-1.00，极强。

配对χ2检验只能给出两种方法阳性（或阴性）检出率的差异是否具有统计学意义，但配对卡方检验掩盖了一个问题，即它对两种方法阳性（或阴性）检出率不区分真阳性（真阴性）和假阳性（假阴性）。事实上我们更想知道两种方法都检出真正病人或者非病人一致性如何，**这里就凸显了Kappa的重要性**。

详细操作戳以下链接：SPSS详细操作：一致性检验和配对卡方检验 / SPSS操作：一致性检验，如何计算kappa值?

加权Kappa系数和Kendall协同系数

除了上面提到的无序分类变量，实际过程中我们还会遇到一些**有序分类资料（等级资料）**的结果（表2），比如化验结果的“-、±、+、++、+++”，这时候就需要用到加权Kappa系数和Kendall协同系数来评价诊断试验的一致性。

表 2. 配对数据形式——等级变量

		B 方法				
		1	2	…	j	合计
A 方法	1	n_{11}	n_{12}	…	n_{1j}	R_1
	2	n_{21}	n_{22}	…	n_{2j}	R_2
	…	…	…	…	…	…
	i	n_{i1}	n_{i2}	…	n_{ij}	R_i
合计		S_1	S_2	…	S_j	N

加权Kappa系数是简单Kappa系数的推广，是用加权的方法对两个评价结果进行量化。较早的时候推送过一篇介绍加权Kappa的文章：SPSS操作：有序分类变量的一致性检验——加权kappa，还不熟悉的伙伴，可以再回去温习一下。

这里着重聊聊**Kendall协同系数**[1]，它是一种非参数检验方法，可实现对评判者的评判标准或结果是否一致的分析。一个比较经典的应用场景：不同研究者会对研究对象的某些特征（比如影像学检查结果）进行评估或者排序，观察这些评估结果的一致性。W的取值为0~1，数值越大，表明评估结果的一致性越高。这还是以“加权Kappa的SPSS操作”的例子介绍一下如何计算Kendall协同系数。

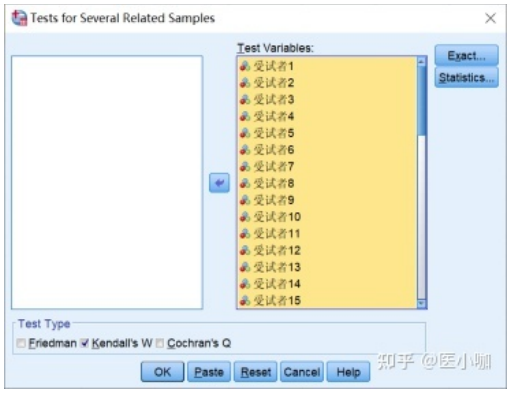
某医院拟分析不同放射科医生对疾病严重程度诊断的一致性。现招募两位放射医生（Radiologist 1和Radiologist 2）分别判断50位受试者的MRI检查结果，并给予Grade I（最轻）到Grade V（最重）五个等级的临床诊断（数据库中Grade I→Grade V分别赋值为1~5）。部分数据如下：

Radiologist1	Radiologist2
2	2
2	2
3	3
3	3
3	5
3	3
3	3

不同于加权kappa的计算，Kendall协同系数需要一个“水平数据库”，即每一列代表一位受试者的MRI检查结果，而每一行分别代表不同放射科医生评估结果，部分数据如下图。

评估者	受试者1	受试者2	受试者3	受试者4	受试者5
Radiologist1	1	1	1	1	1
Radiologist2	1	1	1	1	1

SPSS中依次选择Analyze→Nonparametric Tests→Legacy Dialogs→K Related Samples→显示“Tests for Several Related Samples”主对话框（如下图）→将所有受试者拖入“Test Variables”→“Test Type”框中勾选“Kendall’ s W”→OK



结果显示，W = 0.935（P<0.001），提示两位放射科医生对50位受试者疾病严重程度的诊断具有较高的一致性。

Test Statistics

N	2
Kendall's W[a]	.935
Chi-Square	91.648
df	49
Asymp. Sig.	.000

a. Kendall's Coefficient of Concordance

配对t检验/相关性分析 vs. 组内相关系数（ICC）

上面聊了分类变量的一致性检验，那么**遇到连续变量**（表3）怎么办？多数小伙伴一上来就要用相关分析和配对t检验进行处理，实际上这两种方法都不能对“是否具有一致性”进行判断，为啥呢？且听我慢慢道来。

表 3. 配对数据形式——连续变量

个体	A 方法	B 方法
1	X_{11}	X_{12}
2	X_{21}	X_{22}
⋮	⋮	⋮
n	X_{n1}	X_{n2}

(1) 相关分析

假设将两种方法所得结果看作是两个变量，利用相关分析可以判断变量之间是否具有相关性（还在晕圈的小伙伴戳：SPSS超详细教程：Pearson相关分析），但不能判断两者是否具有一致性。为啥呢？以“SPSS操作：组内相关系数(ICC)”教程中的部分数据来说明。

现假设有2位研究者使用相同的诊断试验分别测量10位受试者的血糖水平。

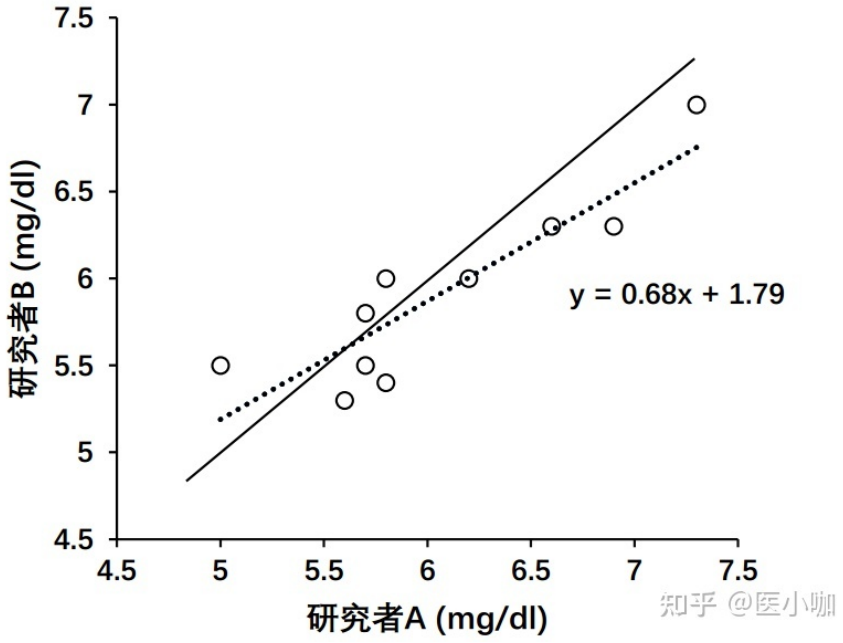


图1. 两名研究者测得血糖水平

首先，看看散点图（相关分析的神器，五星级推荐！），以研究者A和B测得血糖水平分别作为两个坐标，将成对的数据在直角坐标上描点（图1）。

一致性检验意味着分析所有数据到Y=X直线（图1中实线）的误差，而相关（二元相关分析和一元直线回归是等价的）意味着分析数据到Y=aX+b（通常a≠1，b≠0）（图1中虚线）的残差。

其次，相关分析容易受到离群点的影响，如图1所示，两名研究者所测得血糖水平的相关性较好（r=0.89），但是若去掉右上角的点，相关系数会变为r=0.81。显然，通过相关系数来衡量两名研究者血糖水平的关系是不妥当的。

因此，相关分析并不能代替一致性检验。

(2) 配对t检验

配对t检验适用于配对数据，其原理是将两种方法的所得结果之差d看成一个变量，前提条件是该变量服从方差未知的正态分布，目的是用来考察“两种方法平均来讲是否存在显著差异”（**详见：**配对样本t检验，史上最完整SPSS操作教程）。

H0：μd=0，两总体均值无差异；

H1：μd≠0，两总体均值存在差异

如果*P*>0.05，只能说明目前证据尚不能认为两种方法的平均差值不等于0，并不能充分反映两者的一致性。事实上，保持差值的均数和标准差不变，当样本量足够大时，总会得到*P*<0.05的结果。显然，用配对t检验来判断诊断试验的一致性好坏，无疑是不合适的。

(3) 组内相关系数（ICC）

组内相关系数（ICC）[2,3]可用于评价不同测量方法或评价者对同一定量测量结果的一致性或可靠性。

$$ICC=\sigma_T^2 / (\sigma_T^2 + \sigma_B^2 + \sigma_E^2)$$

σ_T^2为被测量者变异，σ_B^2为系统误差(不同测量方法或评价者)造成的变异，σ_E^2为随机误差造成的变异。

一般认为：ICC>0.75—

经过数据模拟分析发现[3]，配对t检验对系统误差敏感（不同测量方法、仪器、评价者），但不能同时兼顾随机误差（研究对象本身变异），而简单相关系数则正好相反。因此，配对t检验与简单相关分析具有明显的片面性，不能同时兼顾随机误差和系统误差，用它们来评价一致性所得的结论可能是错误的。

尽管组内相关系数的计算模型目前尚有争论，但是它同时考虑了系统误差和随机误差的影响，且不受资料类型影响，因而在与配对t检验和简单相关分析的比较中，组内相关系数具有明显的优势。

如何计算ICC，可以点击以下链接：SPSS操作：组内相关系数(ICC)

参考文献

1. Applied nonparametric statistics, 2th Edition. 1990.

2. 中国卫生统计. 2011; 28:497-500.

3. 中国卫生统计. 2011; 28:40-2.

发布于 2021-01-28 10:02

医学　医学统计　数据分析

推荐阅读

诊断试验的一致性检验，方法那么多，你捋清楚了吗?

医小咖

听说过诊断试验的灵敏度、特异度，那你听过符合率吗?

拿到诊断试验结果之后，许多医生会问：“这项试验结果可重复吗？”如果一项诊断试验的结果不能重复，无论有多好的灵敏度和特异度，该诊断试验的实用价值都是非常小的。 真实性，一般使用灵…

医小咖

如何提升诊断的灵敏度或特异度?

在诊断疾病时，可采用多项诊断试验检查同一对象，以提升诊断的灵敏度或特异度，这种方式称为联合试验。根据联合的形式，分为串联与并联。 一、串联试验在进行诊断试验时，我们常常会优先使…

医小咖

诊断准确度评价指标

T-Red

还没有评论

写下你的评论...