·基础理论与方法·

测量的可靠性及其估计方法

周艺彪　赵根明

R19 A

【摘要】 目的　探讨测量可靠性的估计方法及其局限。**方法**　结合实例进行可靠性的计算和分析测量可靠性方法的不足。**结果**　2 名病理学者间可靠性 Kappa 值为0.793；观察者对具有不同患病率的人群分类时，其 Kappa 值分别为0.800和0.137，变化较大；A 型行为综合指标的克朗巴哈 α 系数为0.552。**结论**　Kappa 指数和 α 系数都具有内在人群特性，在推广到不同人群，都需进行可靠性测定。

【关键词】 Kappa 指数；组内相关系数；克朗巴哈 α 系数

Reliability of measurement and the methods of estimating reliability　*ZHOU Yi-biao, ZHAO Gen-ming.*

The Department of Epidemiology, School of Public Health, Fudan University, Shanghai 200032, China

【Abstract】 Objective　To explore the methods of estimating measurement reliability and their limitations. **Methods**　According to the given examples, reliabilities of measurement were calculated and limitations of the methods of estimating reliability were analyzed. **Results**　The Kappa value of interobserver reliability was 0.793 between two pathologists. Between the two populations with different prevalence rates, the values of Kappa were 0.800 and 0.137 respectively, and with big difference. Cronbach's alpha coefficient of compositive index for A type behavior was 0.55. **Conclusion**　The Kappa index and alpha coefficient were both inherently population-specific. Before generalizing to different populations, the reliability needs to be measured.

【Key words】 Kappa index; Intra-class correlation coefficient; Cronbach's alpha coefficient

测量是依据优先次序的规则，对测量对象（分析单位）的变量分类或数值进行赋值的过程，其目的是去描绘测量对象基本概念（或因素）的分类或数量。在测量过程中不可避免地会出现测量误差，其误差来源有多种，不仅有来源于观测者、观测对象和测量或收集信息的工具等，而且还能来自于疾病的分类系统[1]。因此有必要对测量的质量进行评价，测量的质量可用两条标准进行评价：可靠性和有效性，本文主要讨论测量的可靠性及其估计方法。

测量的可靠性是指用特殊的试验或工具获得的测量值（赋值或分类）可重复测量的程度。例如：2 名临床医师对某些病例诊断的一致性程度；某些调查对象在不同时间报告同一行为、事件或信仰的一致性程度（基于重复调查）。评价测量可靠性的方法有 3 种：①时间或观察者内可靠性：某种测量方法在不同时间对同一调查对象进行测量产生相同结果的稳定程度；②一致性或观察者间可靠性：不论谁作观测，某种测量方法对每个调查对象进行测量产生相同观察的稳定程度；③内部一致性：构成一个组合测量（指标）的所有条目或检验反映同一基本概念的程度。事实上，这三种方法在量化一个给定的、应用于一特殊人群的测量方法的相对可靠性时，所有这三种可靠性的概念是相关的；都反映测量的某种相似性或同一性：①在时间上相似性；②在观察者之间的相似性；和③在组成一个组合测量的条目之间的相似性。

作者单位：200032 上海，复旦大学公共卫生学院流行病学教研室

一、时间（观察者内）可靠性

为了评价观察者内可靠性的水平，研究者经常在不同时间（t_1、t_2）对同一组研究对象进行观察，然后计算两组观察结果的相关系数，即观测与重复观测的相关系数（r_{tt}）。

实例：在残疾儿童综合功能评定法的研究中[2]，由专科医生对住院的 84 例患儿进行间隔7～14 天的两次评测以评价测量的时间可靠性，两次评定结果的 Spearman 的相关为 $r_{tt}=0.997$，$P<0.01$，结论该量表具有高度的时间可靠性。

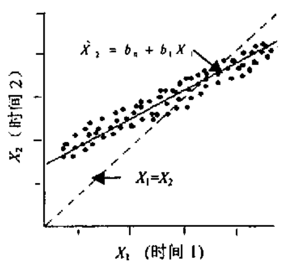
评价：$r_{tt}=1$ 似乎说明测量是完美可靠的，相反地，$r_{tt}=0$ 的测量是完全不可靠的。但这样解释存

在以下三个问题:

(1)如果同一个问题(条目)在太短的时间间隔内被重复,第一次应答的记忆也许会影响第二次的应答。这样,在 t_2 次的测量误差将与 t_1 次的报告值存在阳性联系,这会增大 r_{tt} 的值,导致对可靠性的歪曲评价。注意对于生物学的测量,这问题不会发生。为了防止这个问题的出现,我们可延长两次调查的时间间隔,然而,这一策略将恶化第二个问题。

(2)不仅误差的可变性(不可靠性)可以导致个体观察测量值的变化,而且被测量的基本概念真实值的变化(不稳定性)也可引起观测值的变化。这样 r_{tt} 反映的是测量误差和真实值系统变化的一个未知组合。可以分别做3次相同时间间隔的观察,如果三次的测量误差是没有相关的,这样能够从稳定性估计值中分离出时间可靠性的估计值。用不稳定性对重复检测的可靠性系数进行校正,其公式(r_{tt}')是 $r_{12}r_{23}/r_{13}$(r_{ij} 为在时间 i 和时间 j 观察值的相关系数),这样如果对每一个对象同时作两次观测,则 r_{tt}' 是 r_{tt} 的期望值。

(3)此外,r_{tt} 和 r_{tt}' 的解释都受到 Pearson 相关系数的内在特性的干扰。假设对某一组人群在时间 t_1 和 t_2 测量变量 X,分别记为 X_1 和 X_2。如图1实线所示,重复检测之间的相关显示 X_2 对 X_1 的线性回归模型具有理想的拟合优度,但这种拟合模型并不是测量之间完美的一致性的必要条件,这种完美一致可用虚线表示($X_1=X_2$)。这样,既是 r_{tt} 接近1,也可显示观测结果之间存在明显的差别,这也许反映的是测量误差。



因此乘积矩阵相关系数也许不适合去描述时间可靠性。对于连续变量,一个适宜的描述方法是计算每一个对象两次测量(X_2、X_1)的差值,然后检验这个差值的分布和描述它与测量值的平均数($[X_1+X_2]/2$)或其他因素的关系。另一种测量时间可靠性的方法是用定量测量观察者间可靠性的方法,即 Kappa 指数法(将在下部分进行讨论)。

二、观察者间可靠性

观察者间的可靠性(一致性)可以用 Kappa 指数和组内相关系数(ICC)两个指标进行描述:

1. Kappa 指数:对于分类资料可用 Kappa 指数测量观察者间可靠性。定量测量观察者间一致性水平最简单的方法是计算观察者之间的符合率 P_0,然而,这种测量方法并不能提供完全的信息,这是因为可能由于机遇而使一些对象在观察者之间判断一致,即 P_0 的无效值不是0,故 P_0 的解释是不明确的。为了弥补这个问题,需对 P_0 和 P_e 进行比较,P_e 为观测者由于机遇而测量一致的对象所占的比例(机遇一致率)。Kappa(K)指数为校正机遇后的一致性的统计量[3],即:$K=(P_0-P_e)/(1-P_e)$,这样,K 可以解释为观测者实际获得的一致性(P_0-P_e)占排除机遇后一致性($1-P_e$)的比例。其中 P_0、P_e 和统计检验公式可参考方积乾主编的《医学统计学与电脑实验》[4]进行。

不同数值范围的 K 表明不同程度的一致性,表1为 Kappa 量判断表。K 值的下限在0到-1之间。

表1 Kappa 量判断表

K 值	一致性强度
1	完全一致性
>0.80	几乎完全一致性
0.60~0.80	很强一致性
0.40~0.60	中等一致性
0.00~0.40	弱一致性
0	机遇一致性(完全由机遇所致的一致率 $P_a=P_0$)
<0	一致性程度极差(实际观察的一致率小于机遇一致率)

2. ICC:对于计量资料,观察者间的一致性可以用 ICC 进行测量,ICC 能被认为是加权 Kappa 的一种特殊类型[5]。其计算公式为

$$R=\frac{n(MS_P-MS_E)}{nMS_P+(k-1)MS_R+(n-1)(k-1)MS_E}$$

式中 MS_P:观察对象间均方;MS_R:观察者间均方;MS_E:误差均方;n:观察对象数;k:观察者个数。观察者间变异越少,均方 MS_R 也越少,R 越大,说明观察者间的一致性越强,反之亦然,具体实例计算见方积乾主编的《医学统计学与电脑实验》[4]。

实例:200例活检病例由2名病理学者同时独立读片,并将他们划分为恶性(+)肿瘤或良性(-)

肿瘤,结果见表 2。

表2　2 名病理学者独立将活检病例划分为恶性或良性的结果

病理学(1)	病理学(2)		合计
	+	−	
+	85	15	100
−	10	90	60
合计	95	105	200

(1)一致性检验:病理学者 1 把活检病例判为恶性的比例为 100/200 = 50%;病理学者 2 把同一批活检病例判为恶性的比例为 95/200 = 47.5%,提示着他们之间的诊断差异不大。这种差别的 McNemar[6] 双侧检验是:

$$X_{McN}=\frac{(15-10)}{\sqrt{15+10}}=1.00;P>0.25$$

这样,2 名病理学者的一致性的概率>25%,因此这 2 名病理学者对样本的分类具有一致性。一致性检验也可用一般的卡方进行检验[4]

(2)一致性程度:2 名病理学者对观察对象诊断一致的比例为(85+90)/200=0.875;相应的由机遇所致的期望一致性为$[100(95)+60(105)]/200^2=0.395$。因此,相应的 K 是:

$$K=\frac{0.875-0.395}{1-0.395}=0.793$$

这样,可下一个结论:2 名病理学者在排除机遇之后,两者的诊断具有很强的一致性,与一致检验相符。

评价:Kappa 指数的值不仅依赖于观测者间的一致性水平,也依赖于样本中变量的实际分布。例如,有两个人群 A 和 B,每个人群分别有1 000人,其实际患病率分别为 50% 和 1%。假设在两个人群,观测者 1 能正确分类所有对象,观测者 2 在每个人群中错误分类的比例都为 10%,如表 3 和表 4。

表3　A 人群的分类结果

观测者(1)	观测者(2)		合计
	+	−	
+	450	50	500
−	50	450	500
合计	500	500	1 000

表4　B 人群的分类结果

观测者(1)	观测者(2)		合计
	+	−	
+	9	1	10
−	99	891	990
合计	108	892	1 000

虽然表 3 和表 4 仅在患病率上不同,但这两个 K 值(A 人群 $K=0.800$,B 人群 $K=0.137$)有很大差别,它易误导提示在两个人群中观测者间的一致性是不同的。Kappa 对实际频数的依赖性在非常少(<5%)和非常大(>80%)的患病率人群中是一个很大的问题。这个问题不只是限制了 Kappa 本身对一致性的测量,而且它是所有可靠性估计的一个内在的特性。可靠性的程度是固定具有人群特殊性的,不能越过具有不同实际患病率或不同测量因素的人群而推广。

三、内部一致性和克朗巴哈 α 系数

可靠性的另一种测量方法是评估组成一个综合指标的一组条目或检验之间的内在一致性。也就是说,测量组成这个综合指标的所有条目反映同一基本概念的程度。

假设分开对 K 个条目($S_i, i=1,\cdots,K$)进行评分,然后对它们求和得到每个对象的综合指标的评分(S)。理论上,综合指标评分的总方差(S_S^2)可分成两部分:一部分归因于基本概念的真实可变性(S_T^2);另一部分归因于误差可变性(S_E^2),即:

$$S_S^2=S_T^2+S_E^2$$

内部一致性是真实方差在总方差中所占的比例:

$$S_T^2/S_S^2=(S_S^2-S_E^2)S_S^2=1-S_E^2/S_S^2$$

为了估计一个给定数据集的可靠性,可以通过计算总指标评分的样本方差来估计 S_T^2 和求 K 个条目评分的方差之和来估计 S_E^2。则综合指标的可靠性系数是

$$r_{kk}=(1-\sum_{i=1}^{K}S_i^2/S_S^2)[k/(k-1)]$$

$k/(k-1)$是一个校正因子,用来说明指标中的条目数。注意当 k 取值非常大时,校正因子接近 1。可靠性系数的这种形式就是克朗巴哈 α 可靠性系数[7]。一般来说,条目方差(S_i^2)相对于综合指标评分的方差来说越少,指标越可靠。r_{kk} 接近于 1,意味着在一个条目得分高的对象在别的条目得分也高,提示所有条目都能对同一基本概念进行测量,即该综合指标具有内在一致性。

克朗巴哈 α 系数也可以解释为两个 K-条目综合指标之间的期望相关性,这两个指标都是在同一内容领域(与被测量概念有关的所有可能条目的集合)中所抽取不同的条目上所建立起的。这样,如果创造两个独立 K-条目综合指标去测量同一个概念和对同一人群应用双方的尺度,则这两个指标之间的期望相关性能被解释为两个尺度内部一致性的一个平均测量。

可以使用与前面建议的与 Kappa 相同的量判表去解释克朗巴哈 α 系数的数值范围的意义

(表 5)。

表5　克朗巴哈 α 系数量判表

r_{kk} 值	内部一致性强度
>0.80	几乎完全内部一致性
0.60~0.80	很强内部一致性
0.40~0.60	中等内部一致性
<0.40	弱内部一致性

如果指标中每个条目是二分的(如,是/否),则克朗巴哈 α 系数为

$$rkk = [1 - \sum_{i=1}^{k} P_i(1-P_i)/S_T^2][k/(k-1)]$$

式中 P_i 为第 i 条目"阳性"(如,是)反应者在所有 n 个对象中的比例;r_{kk} 与哪个反应设为"阳性"没有关系,这是因为每个 P_i 都与 $(1-P_i)$ 相乘。这种 α 系数形式在心理测量的文献中是 Kuder-Richardson 公式。

实例:衡量 A 型行为的指标是由下面 3 个条目综合而成,每个问题都有是或否两种反应。对 80 个调查对象进行调查,结果问题 A 回答"是"有 42 个,问题 B 回答"是"有 25 个,问题 C 回答"是"有 45 个,样本综合指标评分总方差为1.124。

A 当人们正在谈话时,你经常打断他们吗?(时间依赖性)

B 当你和小孩玩游戏时,你很少让他们赢吗?(竞争性)

C 当紧张时,你一般会立即对它做一些事吗?(好斗性)

则该综合指标的可靠性 α 系数

$$r_{33} = \left(1 - \frac{(42/80)(38/80) + (25/80)(55/80) + (45/80)(35/80)}{1.124}\right)\left(\frac{3}{3-1}\right) = 0.552$$

这样由三个条目组成的综合指标在此样本中仅有中等的内在一致性,这种 A 型行为的测量不是非常可靠的。

评价:一般可以采用下列一个或多个的对策去改善内在一致性:①删除与其他条目没有高度相关的条目;②识别条目的聚集性(如使用因子分析),每一个类衡量较宽概念的一个不同的方面,然后对每一个方面产生一个分开的亚尺度;③增加合成综合指标的条目,该条目选自于原始条目的同一内容领域。

为了理解这三个对策的合理性,可参考计算可靠性系数的另一种形式 Spearman-Brown 公式[5]:$r_{sb} = k\bar{r}_{ij}/[1+(k-1)\bar{r}_{ij}$。式中 $\bar{r}_{ij}$ 为所有可能条目对相关系数(r_{ij})的平均数。注意存在可能的对子数为 $k(k-1)/2$。例如,如果 $k=10$,就可能有 10(9)/2=45 个条目对。在得到每个对象指标总评分之前,将所有条目评分经标化(即减去均数再除以标准差后转化为 z 评分),这时得到的 Spearman-Brown 表达式就是 α 系数。

Spearman-Brown 公式的一个主要优点是能表明可靠性怎样依赖于指标中条目的数目。对于一个给定的 $\bar{r}_{ij}$ 值,可靠性系数 r_{sb} 随着条目数目(k)的增加而增大。这样,可以通过增加来自同一内容领域的条目,来提高由多个条目组成的指标的可靠性,然而条目超过一定数目,可靠性的增加会加大资料收集的时间和花费,这是得不偿失的。例如,如果 $\bar{r}_{ij}=0.25$,构建一个超过10~15 个条目组成的指标是值得的。而且,一般在调查中都要使调查手段尽可能短少,以减少调查对象的负担,特别是在有多个调查变量或概念的研究中尤其如此。

α 系数是测量理论中最重要的,它不仅是由多个条目组成的指标的内在一致性的测量方法,而且也与其他可靠性测量方法相关。但系数也有 Kappa 指数所注意的那样,内在一致性的评估也是具有人群特性的,仅因为一个给定的综合指标在某个人群中是具有内在一致性的,但并不意味它在另一个人群中也具有内在一致性。这样无论什么时候使用综合指标,都应该对其内在一致性进行评估,即使在别的研究中显示它们是可靠的,亦如此。

参考文献

1 Rothman KJ, Greenland S. Modern epidemiology. Boston: Lippincott-Raven, 1998. 507.

2 吴卫红,刘建军,胡莹媛,等. 残疾儿童综合功能评定法的研究:(三)可靠性. 中国康复理论与实践, 2002, 8:531.

3 Grabam Dunn. Design and analysis of reliability studies. Oxford University Press, 1989. 37.

4 徐勇勇,赫元涛. 测量手段的效度和信度. 见:方积乾,主编. 医学统计学与电脑实验. 第 2 版. 上海:上海科学技术出版社, 2001. 238-251.

5 Fleiss JL, Nee JCM, Landis, JR. Large sample variance of Kappa in the case of different sets of raters. Psychological Bulletin, 1979, 86: 974-977.

6 Dixen WJ. BMDP statistical software manual. Vol 1. Unversity of California Press, 1988. 554-555.

7 Streiner D, Norman G. Health measurement scales: a practical guide their development and use. Oxford University Press, 1993. 86-88.

(收稿日期:2003-03-15)

(本文编辑:张林东)