•医学统计学讲座•

临床试验常用的组间比较统计方法分析

冯国双※

（首都医科大学附属北京儿童医院大数据中心，北京　100045）

摘要：组间比较是临床试验中很常见的分析方式，方法很多，需要根据结局类型、比较组数、数据分布等各种因素综合选择。尽管组间比较方法相对简单，但仍存在应用错误。本文主要介绍临床试验中常见的组间比较方法，总结常见应用错误，给出正确应用组间比较方法的建议。

关键词：临床试验；组间比较；统计方法

中图分类号：R737.14　　**文献标识码：**A　　**文章编号：**1674-7410(2022)01-0118-03

临床试验是以人（患者或健康人）为研究对象，评估某种医学治疗的效果的试验。临床试验不同于临床研究，后者是一个更为广泛的概念，包括观察性研究和实验性研究。临床试验属于实验性研究，目的通常是评价某种药物（或器械、手术方式等）的治疗效果，实际中绝大多数情况下是通过随机对照试验来实施。常规的组间比较所用的统计分析方法并不复杂，如t检验、卡方检验等，但临床科研工作者在选择和应用时仍然存在各种问题，进而影响结果的可靠性[1-2]。本研究主要针对临床试验中常见的组间比较方法，介绍这些方法的正确选择与应用，并对其常见的错误应用情况进行总结，为临床科研工作者提供参考和借鉴。

1　组间比较方法选择前需明确的几个概念

1.1 *定量资料*　定量资料的特点是其值有实际意义、有单位、可以比较大小。如身高170 cm，有单位cm，有明确的实际意义，170 cm>160 cm，可以比较。定量资料可以是连续的，也可以是离散的，连续的是指（理论上）可以取任意值，如体质量80.6 kg、血压值126.7 mmHg (1 mmHg=0.133 kPa)；离散的定量资料只能取整数，如呕吐3次、癫痫发作4次。

1.2 *分类资料*　分类资料的特点是其值无实际意义、无单位、不可用于比较大小。如性别的男女，分别取值1和2，这两个值并无单位，也不能说2>1，因为1和2并无实际意义，只是个数字代码。分类资料可以是二分类和多分类，如死亡结局是二分类（死亡或存活），而疗效则可以是多分类（如痊愈、有效、无效）。多分类变量根据是否有序还分为无序分类资料和有序分类资料，如职业、血型这类分类资料都是无序的，其顺序的置换并不影响分析；而疗效（如痊愈、有效、无效）、严重程度（如轻、中、重）都是有固定的顺序，不能随意置换，否则会影响结果。

1.3 *正态性*　正态分布是定量资料分析中非常重要的一个概念，定量资料组间比较方法的选择几乎都需要考虑数据的正态性。正态分布是一种钟型曲线，以均值为中心，中间高，两边逐渐降低。正态性简单来说是要求数据大致呈正态分布，其直方图应大致为中间高两边低的对称形状。

1.4 *方差齐性*　定量资料的组间比较还需要考虑组间方差是否一致，即方差齐性问题。如果两组（或多组）的方差大致相等，说明方差齐，否则说明方差不齐。

2　组间比较常用方法

2.1 *t检验*　t检验主要用于两组均值的比较，它反映了相对抽样误差（用标准误表示）而言，两组均值差是否有较大差异，如果差异超出了抽样误差所能解释的范围，便认为差异有统计学意义。t检验要求数据满足正态性和方差齐性。

2.2 *方差分析*　方差分析可用于两组或多组均值的比较，它通过组间变异和组内变异（随机误差）的比较，观察组间变异是否足够大，如果组间变异远远大于随机误差，提示组间差异有统计学意义。方差分析要求数据满足正态性和方差齐性。

2.3 *秩和检验*　秩和检验主要是利用数据的秩次而不是原始数据，比较两组（或多组）的秩次分布情

※**通信作者：**冯国双，E-mail: glxfgsh@163.com

况，如果差异达到了一定界值，便可认为差异有统计学意义。秩和检验可用于偏态数据的分析，因为它用的是秩次，无论数据偏态严重与否，均可使用。

2.4 卡方检验　卡方检验主要用于两组或多组率（或比例）的比较，它主要是通过列联表中的实际频数和理论频数的差异来实现，如果实际频数与理论频数差异太大，超出了既定界值，提示组间率（或比例）的差异有统计学意义。

3 组间比较方法的选择思路

首先要明确数据类型和比较组数，是定量资料还是分类资料，是两组比较还是多组比较。定量资料需要检验数据是否服从正态分布，即正态性检验。正态性检验通常可借助于统计分析软件，比较常用的检验方法是Shapiro-Wilk检验和Kolmogorov-Smirnov检验，当检验结果显示P值<0.05，提示样本数据可能不满足正态分布。定量资料如果满足正态性，还需要进一步检验方差是否齐性。方差齐性检验常用的方法有Bartlett检验和Levene检验。如果检验结果显示P值<0.05，提示组间方差的差异有统计学意义，可认为不满足方差齐性。分类资料需要明确结局是无序的还是有序的。基于上述的条件和思路，临床试验常用组间比较方法总结。见图1。

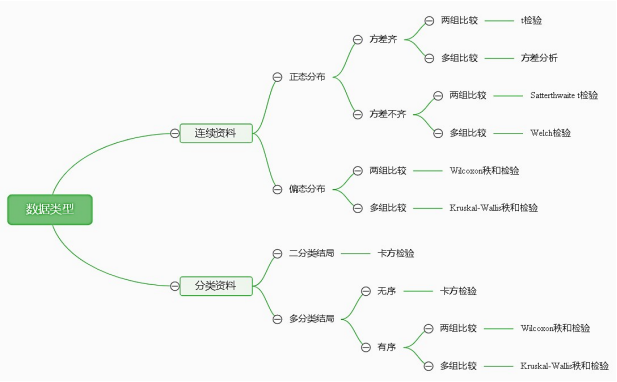

图1　临床试验组间比较统计方法思维导图

4 组间比较方法的常见应用错误

组间比较方法尽管简单，却需要综合多个条件来判断选择，然而这些前提条件在实际中却很容易被忽略，从而导致论文中仍存在不少错误应用，主要的几种情形如下。

4.1 定量资料的比较未考虑数据是否满足正态性，两组比较直接用t检验　这是比较常见的一种错误，不少论文中在统计方法部分直接写“采用t检验进行组间比较”，但却未提及数据是否服从正态分布。严格来说，应该先说明满足使用方法的前提条件，然后再使用该方法进行组间比较。

4.2 多组定量资料比较时采用t检验进行两两比较，增加假阳性错误　从图1可以看出，多组定量资料的比较如果满足条件，可采用方差分析。如果方差分析发现多组间有差异，想进行两两比较，需采用特定的两两比较方法，而不是直接用t检验[3-4]。两两比较方法有很多，常用的有Tukey法、Bonferroni法、Dunnett法等。Tukey法可用于各组例数相等的情形，该法检验效率最高，当各组例数相等时，建议可首选。Bonferroni法的思想是校正检验水准，将原检验水准除以比较次数得到新的检验水准，并根据新的检验水准做出结论。例如，假定原检验水准为0.05，如需对四组进行两两比较，共需比较6次，此时可将检验水准调整为0.05/6=0.0083，即两两比较时只有P值小于0.0083才认为有统计学意义。Bonferroni法应用场景广泛，可用于各组例数相等或不等的情形，不仅可以用于均数的两两比较，也可用于率、比例或等级资料的两两比较。Dunnett法主要用于多个试验组与一个对照组的比较。实际中经常会有这种情况，即设置一个对照组和多个试验组，研究者的目的主要是想比较各个试验组分别与对照组的差异，而各试验组之间不做比较。这种情况下可采用Dunnett法，指定对照组为参照，分别比较各个试验组与对照组的差异。

4.3 有序分类资料的组间比较采用卡方检验，而未采用秩和检验　临床中有不少资料存在自然的等级顺序，如疗效、疾病严重程度等。这类资料如果采用卡方检验，无法体现等级顺序，因为对于卡方检验而言，将等级（如轻、中、重）顺序调换，其结果并无变化。如果要体现等级严重程度，需要采用秩和检验。

4.4 分类资料的两两比较直接采用卡方检验，增加假阳性错误　分类资料的两两比较也需要采用两两比较方法，不能直接用卡方检验分别比较，否则也会犯假阳性错误。分类资料的比较指标通常是率（或比例），率的两两比较可采用Bonferroni法，即根据比较次数重新调整检验水准，根据新的检验水准进行统计决策。

5 组间比较分析的注意事项

尽管组间比较在临床试验中应用广泛，但有一些细节问题仍需注意。目前仍有医学论文中存在对关键概念的理解问题，以至于影响结果的解释，主要有以下几个问题。

5.1 关于*P*值的理解　绝大多数专业论文中都会报告*P*值，但不少临床科研工作者对*P*值的解释却存在明显问题。*P*值是一种概率，反映了在零假设成立的前提下，基于样本数据计算出当前（甚至更极端）的统计量，有多大可能是随机误差造成的。例如，比较试验药和对照药的疗效，计算两组的有效率分别为93%和88%，基于此得到的*P*值为0.03，其含义是在假定试验药和对照药疗效相等（即两组有效率相等）的条件下，根据样本数据计算得到现有的结果（93% *vs.* 88%），甚至差异更大的结果（如94% *vs.* 87%、95% *vs.* 87%等），只有不超过3%的可能性是随机误差造成的。当这一概率较大时，如*P*值为23%，意思是这种差异有较大可能性是随机误差造成的，而不是两组疗效真的存在差异，这种情况下通常不能认为两组疗效差异有统计学意义。

5.2 关于统计学意义（significance）　当组间比较结果显示*P*值低于事先设定的检验水准（通常为0.05），此时一般称为有统计学意义（statistically significant），不建议写成“差异显著”。因为*P*值是对已有结果的概率判断，并不反映实际差异大小。目前有些文章仍然存在着关于统计学意义的不规范用语，如*P*≤0.05认为“差异显著”，*P*≤0.01认为“差异非常显著”等，将*P*值大小与实际差异大小联系起来，这是不严谨的说法。差异有统计学意义并不等同于差异有实际意义，统计学意义是一个统计学的术语，主要与样本量大小有关。例如比较两种药物的降压效果，如果每组只有10例，两组差值3 mm Hg时差异并无统计学意义；但每组50例的时候，两组差值3 mm Hg差异便有统计学意义。实际差值并无改变，但*P*值在大样本时会更低。有的临床科研工作者对此很不理解，认为统计学像是“数字游戏”，实际上，这恰恰反映了统计学的严谨性。因为*P*值与抽样误差有关，样本量越大，抽样误差越小，从统计学的角度越有把握认为这种差异是“真实的”而不是“偶然的”。而在小样本情况下，即便得到了一个较大的差值，但由于样本量太小，抽样误差较大，因此统计学通过一个较大的*P*值来提醒科研工作者不要太过于相信该结果。

5.3 关于正态性的判断　对于定量资料的组间比较，尽管可以通过Shapiro-Wilk检验等方法来判断正态性，但不能仅依靠这些统计方法。正态性检验的零假设是“数据服从正态分布”。也就是说，Shapiro-Wilk检验等方法是计算偏离正态的程度，然后看能否推翻无效假设。与其他假设检验一样，当数据越多时，越容易推翻无效假设。在大样本的时候，即使数据轻微偏离正态，正态性检验也很容易得到一个较小的*P*值。因此当样本量较大时（如数百例，甚至数千例），不要简单依靠统计学检验的方法来判断是否满足正态性，建议结合图形来判断，如直方图、正态分位数图等。

6 小结

本研究介绍了临床试验中组间比较的常见方法以及应用错误和注意事项。进行组间比较是临床试验中十分常见的方法，绝大多数的临床试验都会用到。组间比较的方法都不难，然而依然存在不少错误应用的情形，这些错误应用大多数不是方法的实现问题，而是是否细心的问题，如正态性检验，只要能想到，其软件实现很容易。因此，本研究重点介绍了常见组间比较方法应用的错误和注意事项，希望对临床科研工作者能够起到提醒的作用，以减少论文中的错误应用。

参考文献：

[1] 蔡思雨,冯国双.两组定量资料比较的分析方法及常见错误辨析[J].慢性病学杂志,2016,17(5):477-478.

[2] 冯国双.分类资料组间比较的思路及误区分析[J].中华全科医师杂志,2017,16(6):490-492.

[3] LAWSON M T,HERRING A H,HEMMING K. Multiple comparisons: a tutorial. Part 2: Understanding multiple comparisons [J]. BJOG,2021,128(9):1432.

[4] BENSKEN W P,HO V P,PIERACCI F M. Basic Introduction to Statistics in Medicine, Part 2: Comparing Data [J]. Surgical Infections,2021,22(6):597-603.

（上接第114页）

危险因素相关性分析[J].广东医学,2019,40(14):2063-2066.

[9] 方震,诸靖宇,侯宝生,等.2型糖尿病患者泌尿系结石形成的相关影响因素分析[J].现代生物医学进展,2017,17(24):4660-4663.

[10] 陈城,李翔翔,胡林昆,等.血脂异常与泌尿系结石形成的相关性研究[J].中华泌尿外科杂志,2016,37(9):698-702.

[11] 邓华,杨义,陆丽兰,等.结石类型与血脂代谢的相关性研究[J].临床泌尿外科杂志,2020,35(4):287-290,296.

[12] 李宇斯,曾国华,麦赞林,等.基于两水平Logistic回归分析模型分析我国成人尿石症影响因素[J].中华疾病控制杂志,2019,23(7):866-870.

[13] 胡正委,曹全富,王洛夫,等.超重肥胖泌尿系结石患者133例结石成分分析[J].现代泌尿外科杂志,2017,22(1):25-28,36.