如何正确运用*t*检验——两算术均值比较等效性*t*检验及SAS实现

张甜甜[1]，刘媛媛[1]，李长平[1,2]，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】 本文主要介绍临床试验中的等效性检验的概念、原理、作用以及成组设计一元定量资料等效性检验的SAS实现。基于原始的定量数据或者基于给定样本含量、均值、标准差两种数据结构，结合实例展示SAS在等效性检验中的应用，并对结果进行解释、做出结论。

【关键词】 等效性检验；等效性界值；*t*检验；算术均值比较

中图分类号：R195.1 **文献标识码**：A doi：10.11886/scjsws20200717003

How to use *t* test correctly——the equivalence *t* test of the comparison between two arithmetic means and SAS implementation

Zhang Tiantian[1], *Liu Yuanyuan*[1], *Li Changping*[1,2], *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**】 This article mainly introduced the concept, principle and function of equivalence test in clinical trials and the implementation of SAS for equivalence test of unary quantitative data in group design. The application of SAS software in the equivalence test was demonstrated through examples. The operation was based on the original quantitative data or the data structure of the given sample size, mean and standard deviation. And it was also to explain the results and make the conclusion.

【**Keywords**】 Equivalence test; Equivalence limit value; *t* test; Comparison of arithmetic mean

根据目的不同，临床试验设计可以分为常见的差异性设计、标准阳性对照试验中的等效性和非劣效性设计以及以安慰剂或阳性药物为对照药物试验中的优效性设计，其假设检验方法也随之被提出。*t*检验在差异性检验、等效性检验、非劣效性检验和优效性检验中都有应用[1]。本文介绍临床试验中与等效性检验有关的内容，包括等效性检验的概念、假设检验的原理、样本量的估计和界值的确定，再结合临床实例，介绍两算术均值比较等效性*t*检验及其SAS实现。

1 概 述

1.1 等效性检验简介

在药物研发中，新药在临床疗效上有大的突破变得愈来愈困难，因而新药研发的定位发生了改变：即使新药在疗效上没有提高，但如果在其他方面具有明显优势，研发这类新药也是有价值的。验证新的药物或治疗方法是否与已有的标准药物或治疗方法的临床效应相当，适合采用等效性检验，以说明试验方法的干预效应既不优于也不劣于对照方法[2-3]。

等效性试验是指确认两种或多种药物治疗效果的差别在临床上并无实际意义，即试验药与阳性对照药的疗效相当。“相当”不是“相等”，而是对疗效差别的一种临床可接受的允许范围的概念。因为强调的是“等效”，故意味着试验药可以比阳性对照药好一些，但不可能好很多，这种“好”必须在临床上是微不足道的；当然试验药也可以比阳性对照药差一些，但不应差很多，这种“差”必须是临床上可以容忍的[4]。显然，此处所谓的“好得微不足道”或“差得可以容忍”都是有限度的，在临床实践中即

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）

表现为两种药物疗效差值的最大允许值，这两个方向上的临床最大允许差值则被称为等效性界值。可见，等效性界值有两个，包括上界值和下界值，它们的绝对值既可以相等，也可以不等。等效性界值的确定一般应以临床知识为主要依据，应在试验设计阶段确定下来，并连同样本量估计等内容在临床试验方案中详细说明。有了等效性界值，等效性的统计学推断问题也就容易理解了，即在“好”和“差”两个不同的方向上分别进行两次统计推断，若“不好于”和“不差于”能同时满足，则可获得等效的结论。

1.2 等效性设计下两算术均值比较的假设检验及参数解释

1.2.1 检验假设

无效假设和备择假设分别用 H_0 和 H_1 表示，以 α 作为检验水准。设 T 为试验组效应指标的参数，C 为阳性对照组效应指标的参数（假定评价指标为高优指标，设 δ_L 为负值，δ_U 为正值）。等效性检验的检验假设有以下两对。

第 1 对，无效假设 $H_{01}: T - C \geqslant \delta_U$，备择假设 $H_{11}: T - C < \delta_U$；

第 2 对，无效假设 $H_{02}: T - C \leqslant \delta_L$，备择假设 $H_{12}: T - C > \delta_L$。

1.2.2 假设检验的方法

对于等效性检验的统计推断，需要在两个方向上同时进行单侧检验，即双单侧检验。欲得出等效性结论，两个零假设均需要在总的检验水准 α 上被拒绝。由于通常设两个单侧检验总的犯第 I 类错误的概率为“α”，故每个单侧检验的水准 $\alpha'=\alpha/2$，按照 $\alpha=0.05$，$\alpha'=0.025$，即只有当 $P_1<\alpha'$ 和 $P_2<\alpha'$ 同时成立（注意每次检验的水准均为 α'），前者推论 T 不好于 C，后者推论 T 不差于 C，方可综合推断 T 和 C 具有等效性；若 P_1 和 P_2 中的任何一个大于 α'，则不可得出等效性结论。这里 α' 的含义是当 T 与 C 的疗效差值大于 δ_U 或小于 δ_L 时，错误地得出 T 和 C 等效结论的概率[5]。

1.2.3 等效性试验统计推断的检验统计量

算术均值的等效性检验需进行双单侧 t 检验，一个是对“劣”方向上的检验，另一个是对“优”方向上的检验，其检验统计量计算公式分别为（每次检验的水准均为 α'）：

$$t_1 = \frac{\left(\bar{x}_T - \bar{x}_c\right) - \delta_L}{S_{\bar{x}_T - \bar{x}_C}}, \text{自由度} \nu = n_c + n_T - 2 \quad (1)$$

拒绝域为：$t_1 > t_{\left(1-\alpha/2, n_T + n_C - 2\right)}$

$$t_2 = \frac{\left(\bar{x}_T - \bar{x}_c\right) - \delta_U}{S_{\bar{x}_T - \bar{x}_C}}, \text{自由度} \nu = n_c + n_T - 2 \quad (2)$$

拒绝域为：$t_2 < t_{\left(\alpha/2, n_T + n_C - 2\right)}$

在置信区间法中，按双侧 $100(1-\alpha)\%$ 置信度，计算出 $T-C$ 置信区间的下限 C_L 和上限 C_U，若 $[C_L, C_U]$ 完全在 (δ_L, δ_U) 的范围内，或者 $\delta_L < C_L < C_U < \delta_U$，可下等效性的结论。计算两算术均值差值双侧 $100(1-\alpha)\%$ 置信区间下限和上限的公式分别为：

$$C_L = \left(\bar{x}_T - \bar{x}_C\right) - t_{(1-\alpha/2,\nu)} S_{\left(\bar{x}_T - \bar{x}_C\right)} \quad (3)$$

$$C_U = \left(\bar{x}_T - \bar{x}_C\right) + t_{(1-\alpha/2,\nu)} S_{\left(\bar{x}_T - \bar{x}_C\right)} \quad (4)$$

其中，$t_{(1-\alpha/2,\nu)}$ 为自由度为 ν、检验水准为 α 时的单侧 t 分布界值（左侧累积概率为 $1-\alpha/2$ 时的 t 分布分位数），自由度为 $\nu = n_c + n_T - 2$。对于均值的等效性比较，按假设检验和按置信区间方法得到的结论是等价的[6]。

1.3 等效性界值 δ_L、δ_U 的设定

等效性界值的确定可参考文献[7]中非劣效界值的确定原理和方法获得一侧的界值，然后再参考该界值大小确定另一侧的界值。理论上等效性界值的下界值和上界值是可以不等的，但实际中一般取相等数值，只是代数符号相反。

1.4 两算术均值比较临床等效性的样本量估计

当已知对照组与试验组的总体差值为 $\Delta\left(\Delta = \mu_T - \mu_C\right)$，两组的合并方差为 σ^2 时，在检验水准 α' 下，按照等效性界值 δ_L、δ_U，在一定的样本量下，双单侧检验的把握度可从总的Ⅱ类错误概率算得（$power=1-\beta$），而总的Ⅱ类错误概率 β 可分解为下单侧检验的Ⅱ类错误概率（β_L）及上单侧检验的Ⅱ类错误概率（β_U）两部分［$power=1-(\beta_L+\beta_U)$］。经理论推导，可获得把握度和样本量之间的函数关系式：

$$Power=1-probt\left[t_{1-\alpha, n_C(r+1)}, n_C(r+1)-2, \tau_1\right] - probt\left[t_{1-\alpha, n_C(r+1)}, n_C(r+1)-2, \tau_2\right] \quad (5)$$

式中，$probt[\cdot]$ 为非中心 t 分布的分布函数；τ_1 和 τ_2 为非中心 t 分布的参数；r 为试验组与对照组的分配比例；S 为两组的合并标准差，定义如下：

$$\tau_1 = \frac{(-\delta_L - \Delta)\sqrt{rn_C}}{S\sqrt{r+1}} \quad (6)$$

$$\tau_2 = \frac{(\delta_U + \Delta)\sqrt{rn_C}}{S\sqrt{r+1}} \quad (7)$$

由上述公式可见，在给定把握度与相关参数后，只有样本量n_C是未知数，但由这些公式无法直接计算得到样本量，需要通过迭代运算求得。

得到对照组的样本量n_C后，则不难获得试验组所需的样本量($n_T = rn_C$)。对于等效性试验而言，当δ_L与δ_U绝对值相等时，试验组与对照组的样本量应该是相同的[4]。

2 实例分析

2.1 基于“样本含量、均值和标准差”进行等效性检验

【例1】观察氯沙坦与伊贝沙坦对伴高尿酸血症的原发性高血压患者血清尿酸水平的影响并评价其降压效果。采用多中心、随机、双盲、平行对照设计，随机抽取320例受试者，治疗6周后，患者收缩压改变值见表1。根据临床经验，设定等效性界值为5 mmHg，试评价两种药物的降压效果是否等效。

表1 两组患者治疗6周后收缩压下降幅度(mmHg)

药物种类	n	$\bar{x}$	s
氯沙坦	160	13.29	6.10
伊贝沙坦	160	14.87	5.84

该资料属于成组设计一元定量资料，目的是评价两种药物的降压效果是否等效，应采用双单侧检验进行等效性检验，设定等效性界限：L=−5 mmHg，U=5 mmHg。

SAS程序如下：

```
data example1;
n1=160;n2=160;
mean1=13.29;mean2=14.87;
s1=6.10;s2=5.84;
L=-5.00; U=5.00;
ss1=s1**2*(n1-1);
ss2=s2**2*(n2-1);
sc=(ss1+ss2)/(n1+n2-2);
se=sqrt(sc*(1/n1+1/n2));
/*第1步*/
t1=((mean1-mean2)-L)/se;
/*第2步*/
t2=((mean1-mean2)-U)/se;
/*第3步*/
p1=1-probt(t1, n1+n2-2);
/*第4步*/
P2=probt(t2, n1+n2-2);
ods html;
PROC PRINT;
var t1 p1 t2 p2;
run;
ods html close;
```

【程序说明】第一步，对$H_{0(1)}$，即等效性界值的下限L进行假设检验，计算t_1；第二步，对$H_{0(2)}$，即等效性界值的上限U进行假设检验，计算t_2；第三步，计算t_1对应的t分布右侧的累计概率；第四步，计算t_2对应的t分布左侧的累计概率。

【SAS主要输出结果及解释】

Obs	t_1	p_1	t_2	p_2
1	5.12264	0.000000262	−9.85584	1.7939E−20

统计与专业结论：t_1=5.12264，p_1=0.000000262，按照α'=0.025，拒绝$H_{0(1)}$，接受$H_{1(1)}$；t_2=−9.85584，p_2<0.0001，按照α'=0.025，拒绝$H_{0(2)}$，接受$H_{1(2)}$。两个单侧检验均拒绝H_0，可以认为氯沙坦和伊贝沙坦的降压效果是等效的。

2.2 基于原始定量数据

沿用例1的信息，根据样本均值和标准差模拟出与例1中样本含量相同的随机数。评价氯沙坦和伊贝沙坦的降压效果是否等效(设定等效性界限：L=−5 mmHg，U=5 mmHg)。

SAS程序如下：

```
/*第一步*/
%let seed=12345;
data a;
mu=13.29;sigma=6.1;
group="group1";
do_i_=1 to 160;
x=mu+sigma*normal(&seed);
output;
end;
run;
data b;
mu=14.87;sigma=5.84;
group="group2";
```

```
do _i_=1 to 160;
x=mu+sigma*normal(&seed);
output;
end;
run;
data c;
seta b;
run;
/*以下两步计算结果完全一样,即求97.5%单侧置信区间*/
/*第二步*/
proct test data=c alpha=0.025 sides=u h0=-5;
class group;
var x;
run;
/*第三步*/
procttest data=c alpha=0.025 sides=l h0=5;
class group;
var x;
run;
/*第四步:将前面的两步合并成一步来完成*/
/*两算术均值比较等效性检验,一次求出双侧95.0%置信区间*/
proct test data=c tost(-5, 5) alpha=0.025;
class group;
var x;
run;
```

【程序说明】第一步,根据例 1 中的样本均值、标准差使用 normal 函数产生两组服从各组对应样本均值、标准差的正态分布的随机数,通过 seed 设置种子数,使数据能够重现;第二步,在 TTEST 过程中,sides=u 表示采用上单侧检验,h0=-5 为设定的非劣效性界值;第三步,在 TTEST 过程中,sides=l 表示采用下单侧检验,h0=5 为设定的优效性界值。

【SAS 主要输出结果及解释】

因篇幅所限,第二、三步 SAS 输出结果从略。第四步 SAS 输出结果如下。

group	方法	均值	97.5% 置信限均值		标准差	97.5% 置信限标准差	
group1		12.9828	11.9171	14.0486	5.9574	5.2904	6.8084
group2		14.5759	13.5556	15.5962	5.7035	5.0649	6.5182
差(1-2)	汇总	-1.5931	-3.0615	-0.1247	5.8319	5.3548	6.3985
差(1-2)	Satterthwaite	-1.5931	-3.0615	-0.1247			

TOST 水平 0.025 等效性分析

group	方法	均值	下限		95% 置信限均值			上限	评估
差(1-2)	汇总	-1.5931	-5	<	-2.8759	-0.3103	<	5	相等
差(1-2)	Satterthwaite	-1.5931	-5	<	-2.8759	-0.3103	<	5	相等

方法	方差	检验	Null	自由度	t 值	P 值
汇总	等于	上限	-5	318	5.23	<0.0001
汇总	等于	下限	5	318	-10.11	<0.0001
汇总	等于	总体				<0.0001
Satterthwaite	不等于	上限	-5	317.4	5.23	<0.0001
Satterthwaite	不等于	下限	5	317.4	-10.11	<0.0001
Satterthwaite	不等于	总体				<0.0001

方差等价

方法	分子自由度	分母自由度	F 值	$Pr > F$
折叠的F	159	159	1.09	0.5835

第二步与第三步 SAS 输出结果(因篇幅所限,已省略)显示,方差齐性检验的结果为 F=1.09,P>0.05,认为两总体方差相等。对应的 t 检验结果中,应参照汇总方法(Pooled),对应方差相等时的计算结果。

统计与专业结论:第四步输出结果显示,t=5.23,P<0.0001,按照 α'=0.025,拒绝 H_0,接受 H_1,可以认为氯沙坦的降压效果非劣于伊贝沙坦;第三步输出结果显示,t=-10.11,P<0.0001,可以认为氯沙坦不

优于伊贝沙坦。综上，可以认为氯沙坦的降压效果等效于伊贝沙坦。

方差齐性检验结果为 F=1.09，P>0.05，认为两总体方差相等。对应的 t 检验结果中，应该参照汇总方法（Pooled），对应方差相等时的计算结果。

由“TOST 水平 0.025 等效性分析”的结果可知：求得的两算术均值之差的95%等效性置信区间[-2.8759，-0.3103]完全落在等效性界限[-5，5]之间，表明等效性成立。若基于两次单侧检验的结果（即 t=5.23，P<0.0001；t=-10.11，P<0.0001），也可以得出同样的结论。总之，可综合推断出氯沙坦和伊贝沙坦的降压效果具有等效性。

3 讨论与小结

3.1 讨论

至今，有关“等效性”和“非劣效性”之间在概念上仍有混淆，有的认为只要试验药不比阳性对照药差都可以笼统称为“等效性”。但事实上，二者是有严格界定的。与生物等效性一样，新药的生物利用度比参照药不能低太多，也不能高太多，低了达不到参照药效果，高了可能有更多毒性。然而，临床疗效的等效性如果也追求这种“等效性”似无实际意义，因为人们对阳性对照临床试验通常只会关注试验药的疗效是否“不差于”对照药，而往往不关心试验药是否“好于”对照药。当然如果确实要关注试验药是否“好于”对照药，则可按优效性试验进行设计和分析。

此外，关于等效性试验的检验水准问题，在双单侧检验中明确指出，若假设检验的水准是 α，则每次单侧检验的水准都是 α'（校正后的结果）[4]。

3.2 小结

本文详细介绍了等效性检验的相关内容。等效性检验的目的是检验两种或多种药物或医疗器械治疗效果差别大小在临床上有无实际意义，即试验药与阳性对照药在疗效上是否相当。在试验设计时，需注意等效性界值的确定、样本量估计、效应指标定义等的科学性、严谨性和实用性。

参考文献

[1] 陈卫，徐利娜，迭敏，等. 差异性、等效性、非劣效性和优效性设计中的t检验[J]. 成都医学院学报，2009，4(3)：211-213.

[2] 李雪迎. 等效性检验的统计学分析[J]. 中国介入心脏病学杂志，2015，23(12)：716.

[3] Kumbhare D，Alavinia M，Furlan J. Hypothesis testing in superiority，noninferiority，and equivalence clinical trials：implications in physical medicine and rehabilitation[J]. Am J Phys Med Rehabil，2019，98(3)：226-230.

[4] 陈锋，夏结来. 临床试验统计学[M]. 北京：人民卫生出版社，2018：166-167.

[5] 谷恒明，胡良平. 新药临床试验设计中的比较类型[J]. 四川精神卫生，2017，30(4)：317-322.

[6] 王静，胡镜清. 对临床试验中显著性检验、区间检验及置信区间检验之间关系一致性的认识[J]. 中国临床药理学与治疗学，2011，16(3)：281-286.

[7] 陈阳，刘媛媛，李长平，等. 如何正确运用t检验——两算术均值比较非劣效性t检验及SAS实现[J]. 四川精神卫生，2020，33(3)：226-230.

（收稿日期：2020-07-17）

（本文编辑：陈 霞）