

· 科研方法专题 ·

# 如何正确运用 $t$ 检验——两算术均值比较 优效性 $t$ 检验及 SAS 实现

姚婷婷<sup>1</sup>, 李长平<sup>1,2</sup>, 胡良平<sup>2,3</sup>, 马 骏<sup>1</sup>, 刘媛媛<sup>1\*</sup>

(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: 刘媛媛, E-mail: ivyuan10@126.com)

**【摘要】** 本文主要介绍临床试验中优效性检验的概念、假设检验、界值设定和样本量的估计以及成组设计一元定量资料优效性  $t$  检验的 SAS 实现。通过实例展示 SAS 在优效性  $t$  检验中的应用, 分别基于两种数据结构(原始定量数据和样本含量、均值、标准差数据)进行操作, 对结果进行解释并作出结论。

**【关键词】** 优效性检验; 优效性界值;  $t$  检验; 算术均值比较

中图分类号: R195.1

文献标识码: A

doi: 10.11886/scjsws20200717002

## How to use $t$ test correctly——the superiority $t$ test of the comparison between two arithmetic means and SAS implementation

Yao Tingting<sup>1</sup>, Li Changping<sup>1,2</sup>, Hu Liangping<sup>2,3</sup>, Ma Jun<sup>1</sup>, Liu Yuanyuan<sup>1\*</sup>

(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: Liu Yuanyuan, E-mail: ivyuan10@126.com)

**【Abstract】** This article mainly introduced the concepts, hypothesis testing, cut-off setting and sample size estimation of the superiority test in clinical trials and the implementation of SAS for superiority  $t$  test of unary quantitative data in group design. The application of SAS software in the superiority  $t$  test was demonstrated through examples. Operating steps based on two data structures (based on raw quantitative data and based on sample size, mean and standard deviation data) were given. And it was also to explain the results and make the conclusion.

**【Keywords】** Superiority test; Superiority cut-off value;  $t$  test; Comparison of arithmetic mean

在进行成组设计一元定量资料两总体均值之间的比较时, 传统差异性检验的零假设为两个总体均值相等, 其统计推断往往仅限于两者的差异是否有统计学意义。基于给定的显著性水平  $\alpha$ , 若  $P > \alpha$ , 意味着统计上“不能拒绝零假设”, 但并非说明零假设(两均值相等)一定成立; 若  $P \leq \alpha$ , 虽然可“拒绝零假设”, 但也只能推断两者之间的差异在统计学上是认可的, 而不能评价其差异的大小。在临床上, 新药临床试验的开展越来越多, 由于差异性检验不能准确区分两种药物疗效差异的方向性并体现差异大小所代表的临床实际意义, 于是, 优效性检验、

等效性检验和非劣效性检验被提出并广泛应用<sup>[1-2]</sup>。本文主要介绍临床试验中优效性检验的相关内容, 包括优效性检验的概念、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优效性  $t$  检验及 SAS 实现。

## 1 概 述

### 1.1 优效性检验简介

优效性检验的主要目的是检验试验药的治疗效果是否优于对照药(安慰剂对照或阳性对照)。对于尚未上市的药品, 无论是创新药还是仿制药, 以安慰剂作为对照时, 应证实其实效性<sup>[2]</sup>; 选择已上市同一治疗领域的药物作为阳性对照药时(阳性药物的选择应考虑阳性对照药有效性的既有证据

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)

和阳性对照药物效应的稳定性<sup>[3]</sup>),当预试验的结果表明试验药的效果不仅比对照药效果好,而且两者之间效应指标的差量在数量上相当可观,结合临床专业知识可知,此差量具有临床上的实际意义,此时为了通过正式临床试验,显示试验药的治疗效果优于对照药,可选用优效性检验<sup>[4]</sup>。在试验设计阶段需要设定一个界值 $\delta_U$ 来界定试验药的优效性。 $\delta_U$ 取正值,表示“方向”,即当疗效指标为高优指标时,试验药的效应指标的取值大于阳性药的效应指标的取值,这是一个单侧检验。

## 1.2 优效性设计下两均值比较的假设检验

对两均值比较的优效性试验,若评价指标为高优指标,如生存时间,即均值越大表明疗效越好,其相应的检验假设如下。

无效假设 $H_0: \mu_T - \mu_R \leq \delta_U$ (两种药物疗效相等或试验药劣于对照药)

备择假设 $H_1: \mu_T - \mu_R > \delta_U$ (试验药优于对照药)

对两均值比较的优效性试验,若评价指标为低优指标,如疾病持续时间,即均值越小表明疗效越好,其相应的检验假设为:

无效假设 $H_0: \mu_R - \mu_T \leq \delta_U$ (两种药物疗效相等或试验药劣于对照药)

备择假设 $H_1: \mu_R - \mu_T > \delta_U$ (试验药优于对照药)

其中, $\mu_T$ 为试验组总体均值, $\mu_R$ 为对照组总体均值, $\delta_U$ 为优效性界值,当 $\delta_U$ 为0时,分析方法与差异性检验相同。

以下基于“高优指标”的前提,给出成组设计一元定量资料两均值比较时优效性检验所需要的 $t$ 检验统计量:

$$t = \frac{(\bar{X}_T - \bar{X}_R) - \delta_U}{s_{\bar{X}_T - \bar{X}_R}} = \frac{(\bar{X}_T - \bar{X}_R) - \delta_U}{\sqrt{S_c^2 \left( \frac{1}{n_T} + \frac{1}{n_R} \right)}}, \quad df = (n_T + n_R - 2) \quad (1)$$

$$s_c^2 = \frac{(n_T - 1)S_T^2 + (n_R - 1)S_R^2}{n_T + n_R - 2} = \frac{\sum (X_T - \bar{X}_T)^2 + \sum (X_R - \bar{X}_R)^2}{n_T + n_R - 2} \quad (2)$$

拒绝域为: $t > t_{(1-\alpha, n_T + n_R - 2)}$

在式(1)和式(2)中, $\bar{X}_T$ 和 $\bar{X}_R$ 分别为试验组和对照组的样本均值; $n_T$ 和 $n_R$ 分别为试验组和对照组的例数; $df$ 为 $t$ 分布的自由度; $s_T^2$ 和 $s_R^2$ 分别为试验组

和对照组的样本方差; $s_{\bar{X}_T - \bar{X}_R}$ 为两样本均值之差的标准误; $s_c^2$ 为两组合并方差。

采用单侧检验,当 $t < t_{(1-\alpha, n_T + n_R - 2)}$ ,即 $P > \alpha$ 时,不能拒绝 $H_0$ ,说明试验组的总体平均值与对照组的总体平均值之差小于或等于 $\delta_U$ ,还不能认为试验组的疗效优于对照组;当 $t > t_{(1-\alpha, n_T + n_R - 2)}$ ,即 $P < \alpha$ 时,拒绝 $H_0$ ,说明试验组的总体平均值与对照组的总体平均值之差大于 $\delta_U$ ,认为试验组的疗效优于对照组。

在实际应用中,还可采用置信区间法检验,它是分析时最直接的方法,且相比于假设检验可提供更多的信息。对于两个均值比较的优效性检验,按单侧 $(1-\alpha)\%$ 的置信度,计算 $\bar{X}_T - \bar{X}_R$ 单侧置信区间下限 $C_L$ ,公式如下:

$$C_L = (\bar{X}_T - \bar{X}_R) - t_{(1-\alpha, n_T + n_R - 2)} S_{(\bar{X}_T - \bar{X}_R)} \quad (3)$$

若 $(C_L, \infty)$ 不包括 $\delta_U$ ,即 $C_L > \delta_U$ ,可以得出试验组疗效优于对照组的结论<sup>[5]</sup>。

## 1.3 优效性界值 $\delta_U$ 的设定

从临床意义上确认药物的疗效,界值的确定不可缺少。在优效性试验中,界值是指试验药与对照药之间相差的临床认可的的最小值<sup>[6]</sup>。优效性试验中仅用一个界值,界值的确定必须在试验设计阶段完成并在试验方案中阐明,若有修订,必须在揭盲之前进行并陈述理由。

在确定界值时,若 $\delta_U$ 太小,将可能把疗效远不如对照药的药物判断为优效或等效;若 $\delta_U$ 太大,则可能将本来可推广应用的比对照药更有效的药物误判为无效而得不到及时上市。但在实践中界值的确定往往较复杂,需根据已有的文献资料、设计类型及数据的分布类型,从临床认识水平及成本效益等多方面综合考虑,将统计学推理和临床判断相结合<sup>[2]</sup>。

## 1.4 优效性检验样本含量的估计

两均值比较的优效性临床试验样本量大小与两组均值、标准差、优效性界值、I类错误和II类错误大小以及两组例数分配比例有关。两组均值相差越小,所需样本量越大;优效性界值越大,所需样本量越大;所设定的允许犯的I类错误和II类错误越小,所需样本量越大<sup>[7]</sup>。对于两均值比较的优效性试验,对照组样本量估计公式如下:

$$n_c = \frac{(Z_{1-\alpha} + Z_{1-\beta})^2 \sigma^2 \left(1 + \frac{1}{K}\right)}{(\mu_T - \mu_C - \delta_U)^2} \quad (4)$$

上式中,  $\mu_T$  和  $\mu_C$  分别为试验组和对照组均值,  $\sigma$  为标准差(假设两组标准差相同);  $\alpha$  和  $\beta$  分别为犯 I 类错误和 II 类错误的概率;  $K$  为试验组与对照组例数分配比, 可根据估计的对照组样本量和分配比例  $K$  获得试验组样本量。

试验中, 还应考虑资料质量、患者依从性等问题, 若受试者脱落, 应在实际临床试验操作中扩充一定的病例数来校正估计实际的样本含量。

## 2 实例分析

### 2.1 基于“样本含量、均值和标准差”进行优效性检验

【例 1】比较泮托拉唑与奥美拉唑对消化性溃疡出血患者止血的效果<sup>[8]</sup>。将 80 例消化性溃疡出血患者随机分为泮托拉唑组和奥美拉唑组, 记录两组患者的出血停止时间, 评价泮托拉唑对消化性溃疡出血患者止血的效果是否优于奥美拉唑(假定: 经临床专家商定, 优效性界值取 0.6 d)。数据见表 1。

表 1 泮托拉唑组与奥美拉唑组的出血停止时间(d)

组别	<i>n</i>	$\bar{x}$	<i>s</i>
泮托拉唑组	40	1.5	0.4
奥美拉唑组	40	2.4	0.6

该资料为成组设计一元定量资料, 目的是评价泮托拉唑对消化性溃疡出血患者止血的效果是否优于奥美拉唑, 并设定了优效性界值, 这时采用优效性检验。本例中, 评价指标为出血停止时间, 数值越小表明疗效越好, 故评价指标为低优指标。

SAS 程序如下<sup>[9]</sup>:

```
data example1;
/*第 1 步*/
n1=40;n2=40;
m1=1.5;m2=2.4;
s1=0.4;s2=0.6;
u=0.6;
/*第 2 步*/
ss1=s1**2*(n1-1);
ss2=s2**2*(n2-1);
/*第 3 步*/
sc=(ss1+ss2)/(n1+n2-2);
se=sqrt(sc*(1/n1+1/n2));
```

```
/*第 4 步*/
t=((m2-m1)-u)/se;
p=1-probt(t,n1+n2-2);
/*第 5 步*/
df=n1+n2-2;
up=1-0.05;
utc=tinv(up,df);
cl=(m2-m1)-utc*se;
run;
proc print;
var t p utc cl;
run;
```

【程序说明】在以上 SAS 程序中, 下标为“2”的代表对照组; 由于评价指标为低优指标, 故在计算  $t$  检验统计量时, 公式(1)的分子上两算术均值的顺序要互换一下。第 1 步定义样本含量、两组样本均值、样本标准差和优效性界值; 第 2 步计算两组样本均值的离均差平方和; 第 3 步计算两组的合并方差和两样本平均值之差的标准误; 第 4 步进行假设检验, 计算检验统计量  $t$  值、 $P$  值; 第 5 步计算均值差值单侧置信区间下限  $C_L$ ; PRINT 过程步输出计算结果。

#### 【SAS 主要输出结果及解释】

Obs	<i>t</i>	<i>p</i>	utc	cl
1	2.63117	0.005125328	1.66462	0.71020

统计与专业结论:  $t=2.63117, P=0.005125328$ , 按照  $\alpha=0.05$ (单侧检验), 拒绝  $H_0$ , 接受  $H_1$ , 可以认为泮托拉唑对消化性溃疡出血患者止血的效果优于奥美拉唑。从 95% 置信区间下限来看,  $C_L=0.71020 > 0.6$ , 可以认为泮托拉唑对消化性溃疡出血患者止血的效果优于奥美拉唑, 该结论与假设检验结果一致。

### 2.2 基于原始定量数据

【例 2】沿用例 1 中的信息, 根据样本均值、标准差模拟出与例 1 中样本含量相同的随机数。评价泮托拉唑对消化性溃疡出血患者止血的效果是否优于奥美拉唑(假定: 经临床专家商定, 优效性界值取 0.6 d)。

SAS 程序如下:

```
/*第一步, 下面的 SAS 程序产生的模拟数据具有重现性*/
%let seed=12345;
data a;
mu=1.5;sigma=0.4;
```

```

group="group2";
do _i_=1 to 40;
x=mu+sigma*normal(&seed);
output;
end;
run;
data b;
mu=2.4;sigma=0.6;
group="group1";
do _i_=1 to 40;
x=mu+sigma*normal(&seed);
output;
end;
run;
data c;
set a b;

```

```

run;
/*第二步*/
proc ttest data=c alpha=0.10 sides=u h0=0.6;
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,使模拟数据具有重现性;第二步,在 TTEST 过程中,“sides=u”表示采用上单侧检验,“h0=0.6”为设定的优效性界值;“alpha=0.10”代表求均值之差的双侧 90.0% 置信区间,相当于单侧 95.0% 置信区间。由于评价指标为低优指标,故将试验组设为第二组(group2),对照组设为第一组(group1)。

【SAS 主要输出结果及解释】

group	N	Mean	Std Dev	Std Err	Minimum	Maximum
group1	40	2.3707	0.5248	0.0830	1.4662	3.5054
group2	40	1.4805	0.3499	0.0553	0.8774	2.2370
Diff (1-2)		0.8902	0.4460	0.0997		

group	Method	Mean	90% CL Mean	Std Dev	90% CL Std Dev
group1		2.3707	2.2309	2.5105	0.5248
group2		1.4805	1.3873	1.5737	0.3499
Diff (1-2)	Pooled	0.8902	0.7613	Infty	0.4460
Diff (1-2)	Satterthwaite	0.8902	0.7612	Infty	

以上结果分别为两组出血停止时间的均值以及它们差值的均值、标准差、标准误、双侧 90.0% (相当于单侧 95.0%) 置信区间等信息。

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	78	2.91	0.0024
Satterthwaite	Unequal	67.948	2.91	0.0024

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	39	39	2.25	0.0130

根据方差齐性检验的结果(Equality of Variances),  $F=2.25, P<0.05$ , 认为两总体方差不等。对应的  $t$  检验结果中,应该参照 Satterthwaite 方法。

统计与专业结论: $t=2.91, P=0.0024$ , 按照  $\alpha=0.05$  (单侧检验), 拒绝  $H_0$ , 接受  $H_1$ , 可以认为泮托拉唑对消化性溃疡出血患者止血的效果优于奥美拉唑。从 95% 置信区间下限来看,  $C_L=0.7612>0.6$ , 可以认为泮托拉唑的效果优于奥美拉唑, 该结论与假设检

验结果一致, 即可以认为泮托拉唑对消化性溃疡出血患者止血的效果优于奥美拉唑。

### 3 讨论与小结

#### 3.1 讨论

基于不同的研究目的, 临床试验有不同的类型, 而且试验类型必须在试验设计阶段确定。然而, 实际工作中有些情况完全确定下来并不容易, 例如对新产品的研发定位如果在疗效上有足够的信心认为试验药优于阳性对照药, 则可直接采用优效性试验; 如果新产品的疗效与阳性对照药比较并无明显优势, 但有足够的信心认为疗效相差处于临床可接受范围, 关键是有其他方面的明显优势, 这时可选择非劣效性试验。如果对试验药疗效的优势把握不大, 一般不适合贸然设计为优效性试验, 可考虑先设计成非劣效性试验, 如果试验结果经过统计分析(前提是基于临床认可的优效性界值)能显示出优效性, 则可按优效性下结论, 具体见文献[10-11]。

若按照优效性试验设计,而没有得出优效性结论,则通常不再考虑进行非劣效推断,毕竟在优效性试验设计时一般不会考虑非劣效界值的问题,如果事后确定非劣效界值将会引入偏倚,导致 I 类错误膨胀<sup>[11]</sup>。

### 3.2 小结

本文详细介绍了优效性检验的相关内容。当检验一种药物的疗效是否优于另一种药物时,采用优效性试验,在进行试验设计时应注意阳性对照药的选择、界值的确定及效应指标定义等。当对新药的研发定位不明确时,即对试验药的疗效优势把握不大时,建议采用非劣效试验设计。

### 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] 黄钦,赵明.对临床试验统计学假设检验中非劣效、等效和优效性设计的认识[J].中国临床药理学杂志,2007,23(1):63-67.

- [3] 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.
- [4] 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.
- [5] 方积乾.卫生统计学[M].7版.北京:人民卫生出版社,2012:324-325.
- [6] 姚晨,刘玉秀.临床试验确证疗效时应注意的统计学问题[J].中国循证医学杂志,2005,5(3):242-245,261.
- [7] 吴兴,李婵娟,丁伯福,等.两均数比较的优效性临床试验样本量估计[J].数理医药学杂志,2013,26(5):517-519.
- [8] 叶鑫,周伊燕,姜秋.比较泮托拉唑与奥美拉唑对消化性溃疡出血患者止血效果、不良反应及复发率的影响[J].世界华人消化杂志,2016,24(28):4008-4012.
- [9] 胡良平.SAS常用统计分析教程[M].2版.北京:电子工业出版社,2015:273-274.
- [10] Committee for Proprietary Medicinal Products. Points to consider on switching between superiority and non-inferiority [J]. Br J Clin Pharmacol, 2001, 52(3): 223-228.
- [11] CCTS工作小组,夏结来.非劣效临床试验的统计学考虑[J].中国卫生统计,2012,29(2):270-274.

(收稿日期:2020-07-17)

(本文编辑:陈霞)



## 科研方法专题策划人——胡良平教授简介

胡良平,男,1955年8月出生,教授,博士生导师,曾任军事医学科学院研究生部医学统计学教研室主任和生物医学统计学咨询中心主任、国际一般系统论研究会中国分会概率统计系统专业理事会常

务理事、中国生物医学统计学学会副会长、北京大学口腔医学院客座教授和《中华医学杂志》等10余种杂志编委;现任世界中医药学会联合会临床科研统计学专业委员会会长、国家食品药品监督管理局评审专家和3种医学杂志编委;主编统计学专著48部、参编统计学专著10部;发表第一作者和通信作者学术论文300余篇、发表合作论文130余篇;获军队科技

成果和省部级科技成果多项;参加并完成三项国家标准的撰写工作、参加三项国家科技重大专项课题研究工作。在从事统计学工作的30年中,为几千名研究生、医学科研人员、临床医生和杂志编辑讲授生物医学统计学,在全国各地作统计学学术报告100余场,举办数十期全国统计学培训班,培养20多名统计学专业硕士和博士研究生。近几年来,参加国家级新药和医疗器械项目评审数十项、参加100多项全军重大重点课题的统计学检查工作。归纳并提炼出有利于透过现象看本质的“八性”和“八思维”的统计学思想,独创了逆向统计学教学法和三型理论。擅长于科研课题的研究设计、复杂科研资料的统计分析和SAS与R软件实现、各种层次的统计学教学培训和咨询工作。