

· 科研方法专题 ·

如何正确运用 Z 检验——两总体率比较 非劣效性 Z 检验及 SAS 实现

胡纯严¹, 胡良平^{1,2*}

(1. 军事科学院研究生院, 北京 100850;

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

*通信作者: 胡良平, E-mail: lphu927@163.com)

【摘要】 本文介绍了两总体率比较非劣效性 Z 检验的原理、方法和 SAS 实现。文中涉及两个方面的内容, 包括两总体率比较非劣效性 Z 检验的原理和方法以及基于 SAS 软件实现两总体率比较非劣效性 Z 检验。文中给出了 SAS 输出结果并对其进行解释, 还给出了统计学结论和专业结论。

【关键词】 总体率; 正态分布; 非劣效性检验; 非劣效性界值

中图分类号: R195.1

文献标识码: A

doi: 10.11886/scjsws20201119001

How to use Z test correctly——comparison of two population rates for the non-inferiority Z test and the SAS implementation

Hu Chunyan¹, Hu Liangping^{1,2*}

(1. Graduate School, Academy of Military Sciences PLA China, Beijing 100850, China;

2. Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing 100029, China

*Corresponding author: Hu Liangping, E-mail: lphu927@163.com)

【Abstract】 The paper introduced the principles, methods and SAS implementation of the comparison of two population rates by using the non-inferiority Z test. Two special titles were involved as follows: one was to introduce the principles and methods of the non-inferiority Z test for the comparison of two population rates, the other was to implement the non-inferiority Z test for the comparison of two population rates based on the SAS software. The SAS output results were presented and explained, and also the statistical conclusions as well as the professional conclusions were given in this paper.

【Keywords】 Population rate; Normal distribution; Non-inferiority test; Non-inferiority dividing value

在常规的临床研究中, 研究者常关注的是药物的疗效和安全性, 对两种药物或治疗方案的效果进行比较时, 通常都采用一般差异性检验^[1-2]; 而在新药(或医疗器械)的临床试验研究中, 虽然仍关注疗效和安全性, 但在对疗效进行评价时, 常需要结合临床实际, 提出一些应当满足的前提条件(例如对两种药物的有效率进行比较时, 要求试验药的有效率不应比对照药的有效率低 10%, 才允许试验药投入生产并进入临床使用), 以便使疗效的评价更加科学、合理和实用。也就是除一般差异性检验之外, 还可能采取非劣效性检验或等效性检验或优效性检验^[3]。本文将结合临床实际, 介绍与“两总体率比较的非劣效性检验”有关的内容。

1 两总体率比较非劣效性 Z 检验的基本原理和方法

1.1 基本概念

非劣效性检验是指主要研究目的为显示试验药的治疗效果在临床上不比阳性对照药差的试验。在试验设计阶段需要设定一个界值 δ_L (通常取负值, 代表“差或劣”), 将其代入检验统计量中进行计算。若计算结果为 $P \leq \alpha$ (通常取 $\alpha = 0.05$), 就可以明确判定试验药的疗效非劣效于对照药; 反之亦然^[3-4]。

1.2 问题与数据结构

【例 1】 为评价度洛西汀肠溶胶囊(试验药, 简称 A 药)治疗抑郁症的效果, 采用随机、双盲、氟西汀(对照药, 简称 B 药)平行对照的方法进行试验。受

试者分别口服度洛西汀肠溶胶囊或氟西汀胶囊,共观察6周。其中,试验组111例患者,对照组117例患者。治疗后,试验组和对照组的有效率分别为87.39%(有效例数 $n=97$)和88.89%(有效例数 $n=104$)。根据临床实际,设置非劣效性界限为-10%。试评价度洛西汀肠溶胶囊治疗抑郁症的效果是否非劣效于氟西汀胶囊的治疗效果^[5]。临床资料见表1。

表1 两组患者的治疗效果

药物种类	例数(n)	有效率(%)
度洛西汀肠溶胶囊	111	87.39
氟西汀胶囊	117	88.89

注:对原文中有效率数据做了微调,以便求出有效例数

【对数据结构的分析】该资料的试验因素为“药物种类”,其两个水平分别为“度洛西汀肠溶胶囊”与“氟西汀胶囊”,观测指标为“有效率”,所以该资料为成组设计一元定性资料。

【统计分析方法的选择】研究者的研究目的是评价度洛西汀肠溶胶囊治疗抑郁症的效果是否非劣效于氟西汀胶囊的治疗效果,并且设定了非劣效性界值 $\delta_L = -10\%$,这时应选用非劣效性检验。

1.3 基本原理

1.3.1 检验假设

给出检验假设,并规定检验水准 α 的值。

$$H_0: \pi_T - \pi_R \leq \delta_L;$$

$$H_1: \pi_T - \pi_R > \delta_L;$$

$\alpha=0.05$ (单侧)。

1.3.2 检验统计量

所需要的检验统计量如下式:

$$Z = \frac{P_T - P_R - \delta_L}{\sqrt{\frac{P_T(1-P_T)}{n_T} + \frac{P_R(1-P_R)}{n_R}}} \sim N(0, 1) \quad (1)$$

拒绝域为 $Z > Z_{1-\alpha}$ 。

在式(1)中,检验统计量 Z 服从标准正态分布; (P_T, n_T) 与 (P_R, n_R) 分别为试验组(T)与对照组(R)的“样本率、样本含量”; δ_L 为非劣效性界值(通常取负值)。

1.3.3 基于检验统计量 Z 的取值计算单侧尾端概率

根据检验水准查表或借助相应的SAS函数进行计算,获得检验统计量 Z 的临界值(针对手工计算而言)和(或)标准正态分布尾端的概率(针对统计软件而言)。

1.3.4 得出统计学结论和专业结论

根据拒绝还是不能拒绝 H_0 的结果,得出统计学结论,最后结合专业知识得出专业结论。

2 两总体率比较非劣效性检验的SAS实现

2.1 基于SAS语言按公式编程间接实现计算

基于SAS语言和标准正态分布函数等知识^[6]并按式(1)进行编程,就可间接实现两总体率比较非劣效性 Z 检验。

SAS程序如下:

```
%let n_t=111;
%let n_r=117;
%let p_t=0.873874;
%let p_r=0.888889;
%let delta_L=-0.10;
data a;
nu=((&p_t-&p_r)-&delta_L);
A=&p_t*(1-&p_t)/&n_t;
B=&p_r*(1-&p_r)/&n_r;
nd=sqrt(A+B);
Z=nu/nd;
p=1-probnorm(Z);
run;
proc print noobs data=a;
var Z p;
run;
```

【程序说明】变量名后面为“_t”与“_r”分别代表“试验组”与“对照组”相应的量(包括样本含量与有效率);有效率采用6位小数表示的目的是使计算结果尽可能精确,以助于与后面基于FREQ过程直接计算的结果具有一定的可比性;在评价指标为高优指标时(例如本例中为“有效率”,其值越大越好),非劣效性界值应采用负值来表示。

【SAS输出结果】

Z	P
1.98278	0.023696

【统计学结论】因 $Z=1.983, P=0.023696$ (单侧概率) <0.05 ,所以,应拒绝 H_0 ,接受 H_1 。

【专业结论】在非劣效界值 $\delta_L=-10\%$ 的前提条件下,度洛西汀肠溶胶囊治疗抑郁症的效果非劣效于氟西汀胶囊的治疗效果。

2.2 基于 SAS 中 FREQ 过程直接实现计算

由于 SAS 中 FREQ 过程包含了几乎绝大多数定性资料假设检验方法^[7], 故此处将采用 FREQ 过程来直接实现两总体率比较非劣效性 Z 检验。

SAS 程序如下:

```
data a;
do a=1 to 2;
do b=1 to 2;
input f @@;
output;
end;
end;
cards;
97 14
```

```
104 13
;
run;
proc freq data=a;
tables a*b/riskdiff(noninf MARGIN=0.10);
weight f;
run;
```

【程序说明】各组中的数据应为“有效例数”与“无效例数”, 而不适合直接采用“有效率”; 程序中 noninf 表示进行非劣效性检验; MARGIN=0.10 表示非劣效界值(注意: 因为在 SAS 程序的“tables”语句中的选项“noninf”指明为“非劣效性检验”, 故此处非劣效界值不使用负值)。

【SAS 主要输出结果及解释】

比例(风险)差值的非劣效性分析
H0: P1-P2 ≤ -边际 Ha: P1-P2 > -边际
边际=0.1 Wald 方法

比例差值	ASE(样本)	Z	Pr>Z	非劣效性限值	90% 置信区间
-0.0150	0.0429	1.9828	0.0237	-0.1000	-0.0855 0.0555

因 $Z=1.9828, P=0.0237$, 说明度洛西汀肠溶胶囊治疗抑郁症的效果非劣效于氟西汀胶囊的治疗效果; 若利用置信区间来判断, 由计算结果可知: 两有效率之差的 90% 置信区间为 $[-0.0855, 0.0555]$, 因置信区间位于 $\delta_l = -0.10$ 的右侧, 说明非劣效性结论成立。

3 讨论与小结

3.1 讨论

3.1.1 影响非劣效性检验结果的因素

3.1.1.1 概述

非劣效性检验结果不是绝对的, 随着条件的改变, 最终的检验结果也会发生相应的改变。具体地说, 影响其结果的因素有“显著性水平 α ”“非劣效性界值 δ_l ”和“样本含量”。

3.1.1.2 α 的取值对非劣效性检验结果的影响

非劣效性检验结果与 δ_l 和 α 的取值有密切关系。例如在本文中, 若保持 $\delta_l = -10\%$, 但取 $\alpha = 0.01$, 则无法得出“非劣效”的结论(因 $p=0.0237 > 0.01$); 文献[8]要求, 进行非劣效性检验时取 $\alpha = 0.025$ (单侧)。事实上, 这只是把犯假阳性错误的概

率从常规的 0.05 降低到 0.025; 与此同时, 犯假阴性错误的概率会有所增大。因为非劣效性检验本身就要求进行“单侧检验”, 而检验水准 α 的取值在统计学上习惯取 0.05 或 0.01, 一般来说, 其取值不应随采取“双侧检验”还是“单侧检验”而改变。这一点从本文“2.2”中的 SAS 输出结果中就可看出: 在没有特别设定显著性水平[即采用 SAS 软件中默认的显著性水平 $\alpha = 0.10$ (双侧), 等价于 $\alpha = 0.05$ (单侧)]的前提下, 进行非劣效性检验, SAS 在计算出 Z 值和单侧概率的同时, 还给出了两率之差的“90% 置信区间”。这就意味着在此置信区间的左、右两端各去掉了“5%”, 也就是说, 其单侧尾端概率(即设定非劣效性检验的显著性水平)为 $\alpha = 0.05$ 。

3.1.1.3 δ_l 的取值对非劣效性检验结果的影响

就本文例 1 而言, 若取 $\alpha = 0.05$ 但将 δ_l 取为“-5%”, 得到的结果为 $Z=0.81623, P=0.20718 > 0.05$, 此时, 就无法得出“非劣效”的结论了。由此可知, 非劣效界值的确定是至关重要的。一般来说, 其值的确定应有充分的临床专业知识为依据。通常需要结合已有文献提供的信息、当前临床试验研究预试验的结果和多位从事所研究问题且有丰富临床经验的专家的共识, 综合考虑来确定 δ_l 的取值。

3.1.1.4 样本含量对非劣效性检验结果的影响

在本文例 1 中,若保持取显著性水平 $\alpha=0.05$ 、 $\delta_L=-10\%$ 、A 药的有效率为 87.27% (与表 1 中的 87.39% 接近)、B 药的有效率为 88.14% (与表 1 中的 88.89% 接近),但两组的样本含量都减少一半,即 A 药组有效例数为 48 例、无效例数为 7 例;B 药组有效例数为 52 例、无效例数为 7 例。此时,进行非劣效性检验,其结果为 $Z=1.4832$, $P=0.0689>0.05$,结论就变成了“A 药劣效于 B 药”。

而需要注意的是,在拟开展非劣效性试验研究之初期,就应当按照已知的条件[包括两组估计的有效率、显著性水平 α 的值、检验效能 $(1-\beta)$ 的值和非劣效性界值 δ_L 的值],找到相应的计算公式,估计出合适的样本含量^[3]。

3.1.2 主要疗效评价指标的临床意义

在进行非劣效性检验时,一个常被忽视的问题是评价指标的临床意义。具体地说,就是应明确当前的评价指标究竟属于“高优指标(指标取值越大越好,如治愈率)”还是“低优指标(指标取值越小越好,如死亡率)”。当面对高优指标时,通常情况下,试验组指标的取值低于对照组指标的取值时才适合考虑选用“非劣效性检验”;如果在较大样本含量的预试验中,已获得的试验结果是“试验组指标的取值明显高于对照组指标的取值”,此时,不适合选用“非劣效性检验”,而应当考虑选用“等效性检验”或“优效性检验”^[3]。

3.1.3 δ_L 是否总是取负值

在进行非劣效性检验时,统计学教科书上通常都默认把 δ_L 取为“负值”,代表试验药疗效比对照药疗效“差”。事实上,这样做有一个隐含的前提,即疗效的评价指标为“高优指标”。而当疗效的评价

指标为低优指标时,若拟采用非劣效性检验,此时,其界值 δ_L 应取正值。

3.2 小结

本文介绍了两总体率比较非劣效性 Z 检验的原理、方法和 SAS 实现。在 SAS 实现方面,介绍了基于 SAS 语言编程间接实现两总体率比较非劣效性 Z 检验以及基于 SAS 中 FREQ 过程直接实现与前面提及的相同计算。在后者中,还给出了两总体率之差的置信区间。基于置信区间也可以判断非劣效性结论是否成立。

参考文献

- [1] 孙桂芝,崔明湖,都萌萌,等.阿戈美拉汀治疗首发抑郁障碍效果和安全性 Meta 分析[J].四川精神卫生,2020,33(3):257-262.
- [2] 许国生,车鹏,胡俊,等.氨磺必利和利培酮治疗老年精神分裂症的效果和安全性 Meta 分析[J].四川精神卫生,2020,33(4):342-348.
- [3] 胡良平,陶丽新.临床试验设计与统计分析[M].北京:军事医学科学出版社,2013:71-74,88-100,111-118.
- [4] 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性 t 检验及 SAS 实现[J].四川精神卫生,2020,33(3):226-230.
- [5] 胡良平.面向问题的统计学——(1)科研设计与统计基础[M].北京:人民卫生出版社,2012:504.
- [6] 胡良平,胡纯严.SAS 语言基础与高级编程技术[M].北京:电子工业出版社,2014:37-97.
- [7] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary, NC: SAS Institute Inc, 2018: 2997-3216.
- [8] 邓伟,贺佳.临床试验设计与统计分析[M].北京:人民卫生出版社,2012:134-145.
- [9] 胡良平.现代医学统计学[M].北京:科学出版社,2020:244-257.

(收稿日期:2020-11-19)

(本文编辑:戴浩然)