

循证医学与中医

临床研究中的样本量估算：(1) 临床试验

万霞 李赞华 刘建平

北京中医药大学基础医学院循证医学中心, 100029

【摘要】 临床试验中的样本含量估算是临床医生进行科研设计所关注的重要问题之一。介绍了临床试验研究中样本量估算的重要性、常用的参数以及常见的两个样本比较及分层设计的样本含量的估算,并附有实例说明。同时,针对在样本含量估算过程中存在的常见问题,提出了几点注意事项。

【关键词】 临床试验;样本量;均数;率;分层

1 临床研究样本含量估算的重要性

目前在临床科研设计时,临床医生最关注,也最头痛的就是样本含量的估算。为什么在临床研究中如此强调样本含量的估算呢?理论上来说,样本含量越小,所需的经费越少,实际操作也越简单,这正是所有临床医生所期望的;但是如果样本含量太小,则研究结果的可重复性及代表性较差,研究结果容易得出假阴性或假阳性的结论,即很难获得两组之间存在的真实差异。那么是否样本量越大就越好呢?样本量越大,则所需经费和需要的研究资源越多,实际操作的难度就越大。因此,样本的代表性和执行成本之间存在着不可调和的矛盾。所以,如果对样本含量进行正确的估算,获得适当的样本含量,既可以节省大量的人力、物力和财力,同时也可以减少对受试者造成的潜在伤害,使研究结果真实可靠。一般情况下,两组间效应差异越小,则所需的样本含量越大。在临床中,需要根据研究的目的和研究设计的方法,对样本含量进行正确的估算。

本文介绍临床试验研究中常见的两个样本比较及分层设计的样本含量的估算,下期将介绍观察性研究的样本量估算。

2 决定临床试验样本含量大小的参数

在正确估计样本含量时,应考虑以下几个重要的统计学参数:(1)总体平均数(μ)、标准差(σ)或总体率(π)等。 μ 、 σ 、 π 一般未知,通常以样本的 $\bar{x}$ 、 S 、 P 作为估计值,可以从预试验、查阅文献、经验估计而获得;(2)处理组间的差别 δ 所比较的两个总体参数间的差别

δ 如 $\delta = \mu_1 - \mu_2$ 或 $\delta = \pi_1 - \pi_2$ 。由于研究者无法得到总体参数的信息,可以通过预试验来估计或用临床上认为有意义的差值(假设)来代替;(3)统计检验的水准 α (即第一类错误的概率), α 规定越小,则所需样本量越多。一般取值为0.05。 $U_{1-0.05/2} = 1.96$ (双侧), $U_{1-0.05} = 1.64$ (单侧);为了计算简便,以下所用公式 α 均取0.05。(4)统计检验的效能($1 - \beta$ 其中 β 为第二类错误的概率),即在特定的 α 水准下,若总体间确实存在差异,该次试验能发现此差异的概率。统计检验效能越大,所需样本例数越多,通常取 $\beta = 0.1$ 或 $\beta = 0.2$ 。此时的检验效能分别为90%或80%。一般临床试验的检验效能不能低于75%,否则可能出现非真实的阴性结果。为了计算简便,以下所用公式 β 均取0.10。 $U_{1-0.10/2} = 1.64$ (双侧), $U_{1-0.10} = 1.28$ (单侧)。(5)单侧还是双侧的问题,需要根据专业知识来定。以未知总体与已知总体均数比较为例,通常的假设检验的目的是两个总体是否相等,备择假设是 $\mu \neq \mu_0$,即 μ 可以大于 μ_0 ,也可以小于 μ_0 ,因此是双侧检验。但是如果从专业的角度判断 μ 不可能大于 μ_0 或者是 μ 不可能小于 μ_0 ,这就是单侧的检验,此时备择假设为 $\mu < \mu_0$ 或者是 $\mu > \mu_0$ 。

在试验方案设计中应交代以上参数的设置要求,并通过相应的样本含量计算公式,估计出每组的样本例数。

3 假设检验的类型

临床研究的目的不同,所采用的样本含量估算方法也不同。在临床试验过程,需要区分是做显著性检验(significance test),还是区间假设检验(interval

hypotheses test)。

显著性检验 (significance test)用于推断两个样本是否来自同一总体,它的检验假设为两组相等的零假设,即样本来自同一总体。其无效假设为 $\mu_1 = \mu_2$, 备择假设为 $\mu_1 \neq \mu_2$ 。临床试验中,对于两组疗效的评价,显著性检验结果不能评价差别的实际大小,更不能说明差别是否有临床实际意义,只能说明两组的疗效是否来自不同的总体。

目前,在临床中往往是要确认新药是否不差于或相当于甚至优于标准的有效药物,所以非劣效等效效优效检验也就应运而生。它们的检验假设不再是一个点,而是一个区间,所以又可称之为“区间假设”(interval hypothesis)或“区间检验”(interval test)^[1]。区间假设检验包括了等效性检验 (equivalence test)、非劣效性检验 (noninferiority test) 与 优 效 性 检 验 (superiority test)。以等效性检验为例,其检验假设为① $\mu_1 - \mu_2 \leq -\delta$ ② $\mu_1 - \mu_2 \geq \delta$ ③ $\mu_1 - \mu_2 > -\delta$ ④ $\mu_1 - \mu_2 < \delta$ (δ 为等效界值),只有两组假设同时成立,才认为等效。

4 常用的样本含量估算方法

4.1 两样本均数比较时样本含量的估算方法

4.1.1 两个均数之间的显著性检验

计算公式^[2]:

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 \sigma^2 (1 + \frac{1}{r})}{(\mu_1 - \mu_2)^2} \quad (公式 1)$$

上式中,通常令 $\alpha = 0.05$, 则 $U_{1-\alpha/2} = 1.96$, σ 为总体标准差,它可以从样本或预调查资料通过合并样本方差 $\bar{s}^2$ 和 $\bar{s}^2$ 得到合并方差 $\bar{s}^2$ 来估计。 $\bar{s}^2 = \frac{(n_1 - 1)\bar{s}_1^2 + (n_2 - 1)\bar{s}_2^2}{n_1 + n_2 - 2}$, 其中 n_1 和 n_2 是预调查的样本大小, S_1 和 S_2 是预调查中两个样本的标准差; d 表示不同方向离开总体差别 $\mu_1 - \mu_2$ 的距离,即容许误差。常根据文献得到或通过预调查的样本值来估计。 c 为两样本例数之比,当样本例数相同时, $c = 1$, $n_1 = n_2$ 。

例 1: 欲了解某中成药与某西药对糖尿病患者空腹血糖值的影响有无差别。将 60 例符合纳入标准的糖尿病患者,按照随机分配的原则进入中成药和西药组进行预试验。结果,中成药组的空腹血糖均数 $\bar{x}_1 = 5.60 \text{ mmol/L}$, 标准差 $S_1 = 0.25 \text{ mmol/L}$, 西药组的空腹血糖均数 $\bar{x}_2 = 5.70 \text{ mmol/L}$, 标准差 $S_2 = 0.30 \text{ mmol/L}$, 要求两组例数相同,问需要多大的样本量?

解答: 本例只需比较两药物的疗效上是否有差别, 并不关心中成药的疗效是否不差于或相当于甚至优于标准的西药, 因此, 属于显著性检验。

按照上述公式:

$$\bar{s}^2 = \frac{(30-1)0.25^2 + (30-1)0.30^2}{(30-1) + (30-1)} = 0.07625$$

$$5.70)^2 = 0.01$$

$$n = \frac{1.96^2 \times 0.07625 \times (1+1)}{0.01} = 58.58 \approx 59$$

因此, 每组样本含量为 59 例。

4.1.2 两个均数之间的区间假设检验

计算公式:

① 优效性或非劣效性试验:

$$n = \left[\frac{(Z_{1-\alpha} + Z_{1-\beta})^2 \sigma^2}{(\mu_1 - \mu_2)^2} \right] \times \frac{(1+r)}{r} \quad (公式 2)$$

② 等效性试验:

$$n = \left[\frac{(Z_{1-\alpha} + Z_{1-\beta})^2 \sigma^2}{(\mu_1 - \mu_2)^2} \right] \times \frac{(1+r)}{r} \quad (公式 3)$$

以上两个公式中, 通常令当 $\alpha = 0.05$, 则 $U_{1-\alpha} = 1.64$, $\beta = 0.10$, 则 $U_{1-\beta} = 1.28$, $U_{1-\alpha/2} = 1.64$, α , c 含义及计算同公式 1; δ 为等效界值, 根据经验或专业知识得到。

例 2: 欲了解某中药治疗糖尿病的疗效, 采用西药作对照, 观察其对血糖水平的影响。根据预调查结果, 可知一般糖尿病患者的空腹血糖水平为 9.7 mmol/L (标准差为 2.1 mmol/L), 若两药物降糖均数相差不超过 1 mmol/L , 则认为该中药的疗效不差于某西药。要求两组例数相同, 问需要多大的样本量?

解答: 本例设计为非劣效性试验, 将有关的数值代入公式 2

$$n = \left[\frac{(1.64 + 1.28)^2 \times 2.1^2}{1} \right] \times \frac{(1+1)}{1} = 75.20 \approx 76$$

因此, 每组样本含量为 76 例。

例 3: 假定例 2 各参数不变, 只是按等效性试验来设计, 则每组样本含量可用公式 3 来估计:

$$n = \left[\frac{(1.64 + 1.64)^2 \times 2.1^2}{1} \right] \times \frac{(1+1)}{1} = 94.89 \approx 95$$

因此, 每组样本含量为 95 例。

4.2 两样本率比较时样本含量估计方法

4.2.1 两个率之间的显著性检验

计算公式:

$$n = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2 [p(1-p) + \frac{1}{r}p(1-p)]}{(p_1 - p_2)^2} \quad (公式 4)$$

上式中, 通常令 $\alpha = 0.05$, 则 $U_{1-\alpha/2} = 1.96$

$\bar{p}$ 和 $\bar{p}$ 为两样本的阳性率, 可通过文献或预试验获得。若文献中未报告过, 且又没有做过预试验, 一般可取最为保守的估计值 $\bar{p} = \bar{p} = 0.5$, d 表示与总体率差的距离, 即容许误差。常根据文献得到或通过预调查的样本值来估计。 c 的含义同前。

例 4: 欲了解某中药与某西药治疗糖尿病患者的疗效有无差别。假设某中药治疗糖尿病的有效率 $\bar{p}_1 = 0.40$, 某西药的有效率 $\bar{p}_2 = 0.32$, 则两组估计的率差 $\bar{p}_1 - \bar{p}_2 = 0.08$ 。如果认为估计的总体率差与真实率差

相差不超过 10%，则认为两组的疗效无差别。要求两组例数相同，问需要多大的样本量？

解答：本例只需比较两药物的疗效上是否有差别，并不关心中药的疗效是否不差于或相当于甚至优于标准的西药，因此，属于显著性检验。

按照上述公式：

$$n = n_1 = n_2 = \frac{1.96^2 [0.40(1-0.40) + 0.32(1-0.32)]}{(0.10)^2} = 175.19 \approx 176$$

因此，每组样本含量为 176 例。

4.2.2 两个率之间的区间假设检验

计算公式：

(1) 优效性或非劣效性试验：

$$n = n_1 = n_2 = \frac{[\frac{Q_1 - Q_2}{\lambda}]^2 (1 + \frac{Q_1 + Q_2}{2})}{\lambda^2}, \quad \text{公式 5}$$

(2) 等效性试验：

$$n = n_1 = n_2 = \frac{[\frac{Q_1 - Q_2}{\lambda}]^2 (1 + \frac{Q_1 + Q_2}{2})}{\lambda^2}, \quad \text{公式 6}$$

以上两个公式中， Q_1 为总体率，常用样本率来估计；其余各指标含义同前。

例 5 欲了解某中药与治疗糖尿病的疗效不差于某西药。假设两种方案的有效率约为 0.80， $\lambda = 0.15$ ，要求两组例数相同，问需要多大的样本量？

解答：本例属于非劣效性试验，将有关数值代入公式 5

$$n = n_1 = n_2 = 0.80(1-0.80) \left[\frac{(1.64+1.28)}{0.15} \right]^2 \frac{(1+1)}{1} = 121.26 \approx 122$$

因此，每组样本含量为 122 例。

例 6 假定例 2 各参数不变，只是按等效性试验来设计，则每组样本含量可用公式 3 来估计：

$$n = n_1 = n_2 = 0.80(1-0.80) \left[\frac{(1.64+1.64)}{0.15} \right]^2 \frac{(1+1)}{1} = 153$$

因此，每组样本含量为 153 例。

4.3 分层设计的样本含量估算

由于分层设计的样本含量估算比较复杂，本文仅以两个率比较的显著性检验为例。

例 7 判断某种中药治疗糖尿病肾病三期和四期是否有效，采用西药作对照，中、西药病例数按 1:1，糖尿病肾病三期和四期也按 1:1。根据现有文献资料可知，糖尿病肾病三期治疗组有效率 85%，对照组 70%，糖尿病肾病四期治疗组有效率 75%，对照组 60%，问用何种计算公式能估算最后的样本量？

该例属于分层的随机对照试验。设所需总病例数为 N ，糖尿病肾病各期病例数为 $Q_i N$ ， Q_i 是层样本分数， $i = 1, \dots, R$ ， $\sum Q_i = 1$ 。设第 i 层每组样本量为 $Q_i q_i N$ ， q_i 是组样本分数， $j = 1, 2$ ， $q_1 + q_2 = 1$ 。设第 i 层

平均期望有效率， $\xi_i = q_1 \pi_{i1} + q_2 \pi_{i2}$ 。

计算公式^[3]：

$$n = \frac{\lambda^2 (1 + \frac{Q_1 + Q_2}{2})}{\lambda^2} \quad \text{公式 7}$$

式中 λ 值为非中心参数表达式，通常， $\alpha = 0.05$ ， $\beta = 0.10$ ，则 $\lambda(1, 0.05, 0.10) = 10.507$

因此，代入公式 7：

由于三期和四期病例数的比例为 1:1，故 $Q_1 = Q_2 = 0.5$ ，中、西药病例数的比例也为 1:1，故 $q_1 = q_2 = 0.5$ ， $\pi_{11} = 0.85$ ， $\pi_{12} = 0.70$ ， $\pi_{21} = 0.75$ ， $\pi_{22} = 0.60$ ，则 $\xi_1 = 0.5 \times (0.85 + 0.70) = 0.775$ ， $\xi_2 = 0.5 \times (0.75 + 0.60) = 0.675$ ，因此可求得：

$$n = \frac{10.507^2 \times 0.5 \times [0.5 \times (0.775 \times 0.225 + 0.675 \times 0.325)]}{0.5 \times 0.5 \times [0.5 \times (0.85 - 0.70) + 0.5 \times (0.75 - 0.60)]^2} \approx 368$$

因此，三期和四期的病例数分别为 164 例。

这种分层设计的样本含量估算的方法在临床中其实最为常用。临床试验最常见的是多中心两组临床试验。各院条件不同，治疗同一疾病的有效率可能不同，所得病例资料不宜作简单相加，宜按分层样本处理。

5 注意事项

通过上述介绍的样本含量的公式，我们可以看出决定样本含量估算主要的要素有容许误差 (d 或 δ 值)、 α 、 β 变异度 (σ)、设计类型、两组的分配比例等。在实际的试验中，还有一些其他的影响因素影响着样本含量的估算，包括依从性、失访、设计方法、组数、研究的变量个数等。

依从性 (compliance)：也称顺从性、顺应性，指患者执行医疗措施的程度，即患者执行医嘱的程度。在一些试验中，如果存在比较大的副作用，则多数患者不会长期忍耐，在开始治疗后不久就可能会发生不依从性。或者由于观察的时间较长，患者无法长期坚持，也会发生不依从性。如果估计不依从的病例数过多，则需要对估算的样本含量进行校正。校正的公式^[4]为：

$$n_{\text{校正}} = \frac{n}{(1 - \lambda)^2}, \quad \text{其中 } \lambda \text{ 为估计不依从的比例。例如，在例 4 中每组要求样本含量为 176，在试验前预估计每组的不依从性可能为 10%，则 } \lambda = 0.05 \text{，则经过依从性的校正，每组所需样本含量 } n_{\text{校正}} = \frac{176}{(1 - 0.05)^2} \approx 196$$

失访 (loss to follow up)：是指研究对象因迁移或与本病无关的其他疾病死亡等而造成失访，是截尾 (censoring) 的一种类型，泛指分析时未能获得最后的终点结果。假定失访的原因与研究的终点结局事件无关，如果失访的概率为 λ 则可对样本含量进行调整^[4]：

牌_新 = $\frac{\text{牌}}{(1-\text{牌})}$ 。为了简便计算,按照失访率为10%~15%^[5],对样本量进行偏大估计:牌_新 = 1.15牌。例如,上述例4中,经过不依从性的校正,每组样本量为196例,再按照10%的失访率进行偏大估计,每组样本量则为牌_新 = 1.15 × 196 ≈ 216。

研究设计不同,样本含量的估算也不同。本文所介绍的两样本比较,均指临床中常见的完全随机设计(或称为成组设计)。若临床实际中出现的是配对设计,则需要选用其它公式进行估算。具体公式可参考相关书籍^[5]或文献^[6]。

由于篇幅的限制,本文仅介绍了两个样本的比较。有时在临床中会遇到3个甚至多个样本间的比较,有些研究者会按照两个样本的比较方法,直接选取其中两个样本的统计量值进行估算,这样做是不对的。应选用多个样本比较的样本含量公式,具体公式可参考相关书籍^[5]或文献^[6]。利用公式计算比较复杂,但所得样本例数较精确。有时,为了方便,也可以采用线性算图法^[7],该法应用简便,只需计算差别参数,但所得样本例数不够精确。

多组设计时,由于在各对比组例数相等时进行统计推断效能最高,因此一般要求各组间的样本含量相等。只有在某些特殊情况下,才考虑各组的样本含量不等。

本文所介绍的都是单因素的样本含量的估算。不

可简单用来对多因素的样本含量的估算。对于多因素有一个粗略的估计方法是,观测数至少是变量数目的5~10倍。例如:欲了解治疗高血压病的疗效除了受到药物的疗效,是否还受到年龄、饮食、锻炼及遗传等许多因素的影响。这时属于多因素分析,样本含量则可粗略地估计为250~500例。

6 结语

在临床试验过程中,样本量既不是越大越好,也不是越小越好,需要根据研究的目的、研究的设计方案选择正确的方法正确地进行样本含量的估算。

参考文献

- [1]夏结来,于莉莉,蒋红卫.新药疗效评价的区间检验.第四军医大学学报. 2003, 24 (24): 2209~2212
- [2]周得锋,高尔生主译.卫生研究中样本含量的确定.上海:复旦大学出版社,上海医科大学出版社. 2001: 31~33
- [3]赵国龙,王曙炎.多中心两组临床试验所需样本量的测定.中国医院统计. 2003, 10 (1): 27~29
- [4]刘玉秀,姚晨,陈峰,等.非劣性等效性试验的样本含量估计及把握度分析.中国卫生统计. 2004, 21 (1): 31~35
- [5]王家良主编.临床流行病学——临床科研设计、衡量与评价. 2版.上海:上海科学技术出版社, 1990: 138~156
- [6]刘勘.临床试验设计与数据处理(四).疑难病杂志, 2003, 2 (1): 55~57.
- [7]郭静,徐勇勇,何大卫.多组比较样本含量及检验效能的线性算图估计.中国卫生统计. 2002, 19 (2): 94~95

收稿日期: 2007-05-15



东华原医疗
DONGHUAYUAN MEDICAL

YZN-50多功能真空(提取 分离 蒸馏)浓缩机

留驻精华, 留驻健康

适用范围

广泛适用于诊所、药店、中医院、综合医院、药物研发、临床应用、研究机构、食品饮料加工业、化工行业等及其分支行业的制剂室、实验室、中试阶段研究、小批量生产以及为终端客户服务。

主要功能

- 真空低温煎药;
- 蒸馏浓缩中药, 制作膏剂、粉剂;
- 有机溶剂提取(酒精、乙醚等);
- 果酱、凝胶体(成冻体)等的低温浓缩和提取;
- 芳香剂的提取、各种精油成分的分离;
- 红参制造功能(蒸熟和炖)等。



北京东华原医疗设备有限公司

电话: 010-89718013

传真: 010-89718021

地址: 北京市中关村昌平科技园园区振超路1号

邮编: 102200

网址: www.donghuagroup.com



全国免费垂询热线: 800-810-0359

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net