基于 R 软件估计样本含量与检验效能及其应用

郭春雪[1]，胡良平[1,2*]

（1. 军事医学科学院生物医学统计学咨询中心，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：lphu812@sina.com）

【摘要】 本文目的是使读者快速掌握用 R 软件估计样本含量和检验效能的方法。通过 R 软件中的 stats 包中的三个函数，即 power.t.test()、power.prop.test() 和 power.anova.test()，可以很方便地估计若干种场合下的样本含量或检验效能。事实表明：R 软件易于获取、易学易用、功能强大、适用面宽，可以方便快捷地解决试验设计中的样本含量与检验效能估计问题。

【关键词】 R 软件；样本含量；检验效能；假设检验；均值；率

中图分类号：R195.1 文献标识码：A doi：10.11886/j.issn.1007-3256.2016.06.004

The estimation of sample size and power and its application based on R software

Guo Chunxue[1], *Hu Liangping*[1,2*]

(1. *Consulting Center of Biomedical Statistics, Academy of Military Medical Sciences, 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: lphu812@sina.com*)

【Abstract】 The paper aims to help the readers to grasp the method of estimating the sample size and power with R software. By using the three functions [power.t.test(), power.prop.test() and power.anova.test()] of stat in R software, it is convenient for readers to realize the estimation of sample size and power by using R software under the different situations. The methods of estimating the sample size and power by R software were introduced through several real examples in this article. Since that R is very easy for people to learn and use, and has the advantages of powerful functions and wide application, the users can solve the concrete problems concerned with the estimation of sample size and power in experimental designs conveniently and easily.

【Key words】 R software; Sample size; Power; Hypothesis testing; Mean value; Rate

1 估计样本含量与检验效能的概述

1.1 估计样本含量与检验效能的前提条件[1]

关于估计样本含量与检验效能的概念和前提条件，尽管在文献[1]中已作了介绍，为了便于读者阅读本文，此处仍以概要的形式总结如下。在试验设计中，拟对定量指标的平均值或定性指标的率进行假设检验时，常需提供与结果精确度、评价指标、设计类型和比较类型有关的前提条件。

1.1.1 与结果精确度有关的前提条件

①定出检验水准：即事先规定本批试验允许犯Ⅰ型（或假阳性）错误的概率 α，通常规定 $\alpha=0.05$，同时应明确单双侧检验。α 定得越小，研究所需样本含量就越大。

②提出期望的检验效能（或称把握度）$1-\beta$：即在指定的 α 水准下，若比较的总体之间确实存在着差别，该试验可以发现差别的概率。检验效能越大，所需样本含量越多。在科研设计时，检验效能一般取 0.8 或以上比较适宜。

③估计实验过程中的样本损耗。假设研究者估计本批实验过程中将有 10% 的受试者脱落而无法完成实验，则应将通过计算得到的样本量除以 0.9，将此时得到的结果作为该实验最终需要的样本量。

1.1.2 与评价指标有关的前提条件

必须知道由样本推断总体的一些信息。在比较两总体均数或概率之间的差别是否具有统计学意义时，需要知道总体参数间差值 δ 的信息。如两总体均数间的差值 $\delta=\mu_1-\mu_2$ 的信息（或有关于 μ_1 和 μ_2 的估计值），两总体概率间的差值 $\delta=\pi_1-\pi_2$ 的信息（或有关于 π_1 和 π_2 的估计值）。此外，确定两均数比较的样本含量时，还需要有关总体标准差 σ 的信息（或有关于总体标准差 σ 的估计值）。若希望进行非劣效性检验、等效性检验或优效性检验时，需要提供在临床上有意义的界值 δ（此界值一般应由多位不同地区且学术权威性高、经验丰富的临床和统计学专家共同讨论来商定）。这些信息可以通过查阅资料、借鉴前人的经验或进行预试验寻找参考值。

项目基金：国家高技术研究发展计划课题资助（2015AA020102）

1.1.3 与设计类型和比较类型有关的前提条件

前面提到“两总体”，其真实含义是指所采用的是“单因素两水平设计（常简称为成组设计）”。换句话说，拟采用什么试验设计类型（除了单因素两水平设计之外，还有单组设计、配对设计、单因素多水平设计、某种特定的多因素设计等设计类型）是估计样本含量的重要前提条件之一；而拟采用的比较类型（包括差异性检验、非劣效性检验、等效性检验或优效性检验）也是估计样本含量的重要前提条件之一。

1.2 R 软件中可用于估计样本含量与检验效能的程序包及函数[2-5]

在 R 软件的 stats 包中，有三个函数，即 power.t.test()、power.prop.test() 和 power.anova.test()，可用于估计样本含量或检验效能。

在 R 软件的 sample.size 包中，n.ttest()、samplesize-package() 这两个函数可用于估计样本含量。其中，n.ttest() 函数可用于配对设计和非配对设计（即成组设计）一元定量资料 t 检验时，估计样本含量；而 samplesize-package() 函数可用于多种成组设计场合下假设检验时估计样本含量（注：前述提及的“多种成组设计场合”指“成组设计与配对设计一元定量资料 t 检验时”、“Welch 近似 t 检验时”、“有序资料中带有或不带有结的 Wilcoxon-Mann-Whitney 检验时”，对最后的场合，R 软件的 samplesize 包中，还有 n.wilcox.ord() 函数可用于估计样本含量）。

值得一提的是，在 R 软件的 samplesize4surveys 包中，有 12 个函数，即 b4ddm()、b4ddp()、b4dm()、b4dp()、b4m()、b4p()；ss4ddmH()、ss4ddpH()、ss4dmH()、ss4dpH()、ss4dH()、ss4pH()，其中，前 6 个函数用于 6 种场合下估计检验效能，后 6 个函数用于前述 6 种场合下估计样本含量。

以上提及的 6 种场合分别为：成组设计一元定量资料均值的双侧差异性检验和单侧差异性检验、单组设计一元定量资料均值的假设检验；成组设计一元定性资料比例或率的双侧差异性检验和单侧差异性检验、单组设计一元定性资料比例或率的假设检验。

值得注意的是，R 软件中的内容十分丰富，其“程序包、小插件和函数”多如牛毛，而且每项内容放置在何处，可能的确无人精准知晓。只能是发现什么，调用什么，很难穷尽！

特别提示：当用户加载了程序包“samplesize”后，就可在 R 软件环境中，使用如下命令：> help(samplesize)，就可进入有关样本大小估计的帮助窗口，此窗口内列示出了几十个以“samplesize”或“sample.size”开头的程序包或函数；以“power”开头的用于估计样本大小和检验效能的函数被放置在“stats”程序包中，使用命令“>help(stats)”后可进入此程序包的帮助信息查询窗口，选择其中的“index”，就可按 26 个字母顺序去查看相应字母开头的函数（例如，选择字母 P，就可迅速显示此程序包中以字母 P 开头的全部函数名及其功能的解说信息）。

2 几种场合下估计样本含量与检验效能的实例[2-6]

2.1 评价指标为定量变量的场合

【例 1】某研究者观察氯沙坦与伊贝沙坦治疗对伴高尿酸血症的原按发性高血压患者血清尿酸水平的影响并评价其降压疗效。采用多中心、随机、双盲、平行对照设计。预试验的结果表明，收缩压改变值的情况见表 1。使用双侧差异性检验评价两种药物的降压效果的差别是否具有统计学意义。取 $\alpha=0.05$，$\beta=0.20$，试估计该试验所需的样本含量。

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

解答：需要获得两样本均值之差量 delta 的数值和两样本标准差之均值 sd 的数值。然后，调用 power.t.test() 函数。

```
> delta = 14.87 - 13.29; delta
[1] 1.58
```

以上是求出两样本均值之差量 delta 的数值为 1.58。

```
> sd = (6.10 + 5.84) /2; sd
[1] 5.97
```

以上是求出两样本标准差之均值 sd 的数值为 5.97。

```
> power.t.test(power = 0.80, sig.level = 0.05, delta = 1.58, sd = 5.97)
```

以上语句的目的是调用 power.t.test() 函数，其中的四个参数分别给定了具体的数值。事实上，还有三个参数取默认值，第一个为设计类型：type = c（“two.sample”，“one.sample”，“paired”），默认值为“two.sample”，即成组设计；第二个为备择假设：

alternative = c("two.sided", "one.sided"),默认值为"two.sided";第三个为"strict = T or F",等号后面只能选定一个,其默认值为"strict = F"或"strict = FALSE",其含义是:指定在双侧检验时是否使用严格解释。还剩下一个参数(即 n)的值未给定,需要 R 软件计算。

以下是输出结果:

```
Two - sample t test power calculation
          n = 225.08
      delta = 1.58
         sd = 5.97
  sig.level = 0.05
      power = 0.8
alternative = two.sided
NOTE: n is number in * each* group
```

由输出结果可知:每组应选取约 226 例。

【例 2】某研究者观察氯沙坦与伊贝沙坦治疗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疗效。采用多中心、随机、双盲、平行对照设计。随机选取 320 例受试者,治疗 6 周后收缩压改变值的情况见表 2。使用双侧差异性检验评价两种药物的降压效果的差别是否具有统计学意义。取 $\alpha = 0.05$,$\beta = 0.20$,试估计该试验的检验效能。

表 2 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

解答:需要获得两样本均值之差量 delta 的数值和两样本标准差之均值 sd 的数值。然后,调用 power.t.test() 函数。

```
> delta = 14.87 - 13.29; delta
[1] 1.58
```

以上是求出两样本均值之差量 delta 的数值为 1.58。

```
> sd = (6.10 + 5.84) /2; sd
[1] 5.97
```

以上是求出两样本标准差之均值 sd 的数值为 5.97。

```
> power.t.test(n = 160, sig.level = 0.05, delta = 1.58, sd = 5.97)
```

以上语句的目的是调用 power.t.test() 函数,其中的四个参数分别给定了具体的数值。三个默认参数前已述及,不再赘述。还剩下一个参数(即 power)的值未给定,需要 R 软件计算。

以下是输出结果:

```
Two - sample t test power calculation
          n = 160
      delta = 1.58
         sd = 5.97
  sig.level = 0.05
      power = 0.6554376
alternative = two.sided
NOTE: n is number in * each* group
```

以上结果表明:每组用 160 例,其检验效能仅为 65.54% < 80.0%(常规的要求),犯假阴性错误的概率(34.46%)过大。

【例 3】为观察神经功能康复情况,使用三种方法分别治疗脑卒中抑郁患者,估计治疗后三种方法的 SSS 评分均值分别为 11.0、10.0、9.0,组间方差相等且都为 9,组内方差分别为 4、5、6、9 四种取值条件下,取 $\alpha = 0.05$,$\beta = 0.10$,要求得到三组间差别有统计学意义的结论,每组各需要患者多少例(三组所需患者人数相等)?

解答:情形一,在组内方差为 4 的条件下;

```
> power.anova.test(group = 3, between.var = 9, within.var = 4, sig.level = 0.05, power = 0.90)
```

此条件下输出的结果如下:

```
Balanced one - way analysis of variance power calculation
     groups = 3
          n = 4.017349
between.var = 9
 within.var = 4
  sig.level = 0.05
      power = 0.9
NOTE: n is number in each group
```

以上结果表明:每组只需要 5 例。

情形二,在组内方差为 5 的条件下;

```
> power.anova.test(group = 3, between.var = 9, within.var = 5, sig.level = 0.05, power = 0.90)
```

此条件下输出的结果如下:

```
Balanced one - way analysis of variance power calculation
     groups = 3
          n = 4.688307
between.var = 9
 within.var = 5
  sig.level = 0.05
      power = 0.9
NOTE: n is number in each group
```

以上结果表明：每组只需要 5 例。

情形三，在组内方差为 6 的条件下；

> power.anova.test(group = 3, between.var = 9, within.var = 6, sig.level = 0.05, power = 0.90)

此条件下输出的结果如下：

Balanced one - way analysis of variance power calculation

groups = 3
n = 5.36743
between.var = 9
within.var = 6
sig.level = 0.05
power = 0.9

NOTE: n is number in each group

以上结果表明：每组只需要 6 例。

情形四，在组内方差为 9 的条件下；

> power.anova.test(group = 3, between.var = 9, within.var = 9, sig.level = 0.05, power = 0.90)

此条件下输出的结果如下：

Balanced one - way analysis of variance power calculation

groups = 3
n = 7.431865
between.var = 9
within.var = 9
sig.level = 0.05
power = 0.9

NOTE: n is number in each group

以上结果表明：每组只需要 8 例。

2.2 评价指标为定性变量的场合

【例 4】一个新的抗肿瘤药物 A 与临床有效药物 B 对照进行临床试验，选取一定数目且符合要求的患者随机均分成两组，分别接受 A 药和 B 药治疗。预试验结果为 A 药的有效率是 58.0%，B 药的有效率是 46.0%。欲使用双侧差异性检验评价两种药物的降压效果的差别且希望得出具有统计学意义的结果，取 $\alpha = 0.05$，$\beta = 0.20$，试估计该试验中各组至少需要多大的样本含量？

解答：需要获得估计成组设计两比例或率差异性检验时样本含量所需要的基本信息：$p_1 = 0.58$、$p_2 = 0.46$、sig.level = 0.05、power = 0.80，将各组样本含量 n 留作待估计的参数。然后，调用 power.prop.test() 函数。

> power.prop.test($p_1 = 0.58$, $p_2 = 0.46$, sig.level = 0.05, power = 0.80)

以上语句的目的是调用 power.prop.test() 函数，其中的四个参数分别给定了具体的数值。两个默认参数如下。第一个为备择假设：alternative = c('two.sided', 'one.sided')，默认值为"two.sided"；第二个为"strict = T or F"，等号后面只能选定一个，其默认值为"strict = F"或"strict = FALSE"，其含义是：指定在双侧检验时是否使用严格解释；还剩下一个参数(即 power)的值未给定，需要 R 软件计算。

以下是输出结果：

Two - sample comparison of proportions power calculation

n = 270.9126
p_1 = 0.58
p_2 = 0.46
sig.level = 0.05
power = 0.8
alternative = two.sided

NOTE: n is number in * each * group

以上结果表明：各组需要约 271 例。

【例 5】在一个Ⅱ期临床试验中，已知对照组有效率 $p_1 = 30\%$、试验组有效率 $p_2 = 60\%$、三个限制条件分别为：伽玛 C = 15%、伽玛 E = 15%、伽玛 Delta = 15%、对照组与试验组样本含量的比例分别为 1∶3 与 1∶1 两种条件下，试估计各组的样本含量。

【解答】在临床试验中，有所谓的"双臂试验"，即一个试验组与一个对照组比较。在 R 软件的 sample.size 包中，有一个 Sample.Size() 函数可用于Ⅱ期临床试验且采用双臂优化设计中比较两个比例或率(此法由 Mayo 等于 2010 年提出，可采用固定或灵活的分配比例，可以基于多种限制条件下进行优化设计)时，估计样本含量或检验效能。

情形一，对照组与试验组样本含量的比例为1∶3条件下估计各组的样本含量

> Sample.Size(0.3, 0.6, 0.15, 0.15, 0.15, Allratio_c = 1, Allratio_e = 3)

以上语句的目的是调用 Sample.Size() 函数，其中七个参数的数值均被给定，最后两个参数若不出现，就取默认值，即 1∶1。

以下是输出的结果：

Specified values for parameters:

Response rates:

control = 0.3 experiment = 0.6

Upper bounds for constriants:

gammaC = 0.15 gammaE = 0.15 gammaDelta = 0.15

以上内容实际上是给定的前提条件。

Required sample sizes:

[1] Optimal Design:

nc = 20 ne = 20 n = 40

[2] 1 to 1 Allocation Design:

nc = 20 ne = 20 n = 40

[3] 1 to 3 Allocation Design:

nc = 13 ne = 39 n = 52

第一部分“见上面的[1]”给出了优化设计下的两组各需要 20 例;第二部分“见上面的[2]”给出了 1∶1 条件下的两组各需要 20 例;第三部分“见上面的[3]”给出了 1∶3 条件下的对照组需要 13 例、试验组需要 39 例。

说明:由以上给定条件和输出结果可知,即便给定的前提条件是两组样本含量之比为 1∶3,但也将默认的前提条件 1∶1 多对应的样本含量估计出来了。

情形二,对照组与试验组样本含量的比例为 1∶1 条件下估计各组的样本含量

> Sample.Size(0.3,0.6,0.15,0.15,0.15)

此语句与前面的语句相比,最后两个参数取默认值,即对照组与试验组样本含量之比为 1∶1。

以下是输出的结果:

Specified values for parameters:

Response rates:

control = 0.3 experiment = 0.6

Upper bounds for constriants:

gammaC = 0.15 gammaE = 0.15 gammaDelta = 0.15

以上内容实际上是给定的前提条件。

Required sample sizes:

[1] Optimal Design:

nc = 20 ne = 20 n = 40

[2] 1 to 1 Allocation Design:

nc = 20 ne = 20 n = 40

第一部分“见上面的[1]”给出了优化设计下的两组各需要 20 例;第二部分“见上面的[2]”给出了 1∶1 条件下的两组各需要 20 例。

值得注意的是:此设计并没有交代清楚:sig.level = ? power = ?

笔者认为:这种设计要慎用!

参考文献

[1] 张效嘉,胡良平. 精神卫生科研如何严格遵守试验设计四原则之重复原则[J]. 四川精神卫生,2016,29(4):303-306.

[2] 黄文,王正林. 数据挖掘:R 语言实战[M]. 北京:电子工业出版社,2015:34-39.

[3] 李诗羽,张飞,王正林. 数据分析:R 语言实战[M]. 北京:电子工业出版社,2015:88-156.

[4] 方匡南,朱建平,姜叶飞. R 数据分析:方法与案例详解[M]. 北京:电子工业出版社,2015:54-168.

[5] Joseph Adler. R 语言核心技术手册[M]. 2 版. 刘思喆,李舰,陈钢,等译. 北京:电子工业出版社,2015:417-421.

[6] 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,2013:101-134.

(收稿日期:2016-12-06)

(本文编辑:吴俊林)