基于 R 软件实现随机分组及其应用

胡 完[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 软件实现几种随机分组的方法。通过借助 R 软件中实现随机抽样的 sample( ) 函数,间接地实现简单随机分组和分层随机分组的目的。事实表明:R 软件易于获取、易学易用;R 软件功能强大、适用面宽,可以方便快捷地解决试验设计中的随机分组问题。

【关键词】 R 软件;简单随机分组;分层随机分组;非试验因素
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2016.06.003

The realization of the random grouping and its application based on R software

Hu Wan[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 purpose of this paper is to help readers to implement random grouping by using R software. The simple random grouping and stratified random grouping are realized with the help of the function of sample( ), that is used to implement random sampling. In fact, R software is very easy to obtain, learn and use; and R software is very powerful, wide application, can solve problems about random grouping in the experimental designs conveniently and quickly.

【Key words】 R software; Simple random grouping; Stratified random grouping; Non-experimental factor

1 随机分组的概述[1]

从总体中抽取了规定数目的个体或样品或受试对象,通常还需要将他们随机分到若干个组里去。各组需要分配多少名个体呢?应根据具体情况确定,各组的个体数目可以按一定比例来分配,但最好各组的个体数目相等(在结果上将产生的误差最小化)。

如何将规定数目的个体按等比例或规定的某种比例随机分配到两组或多组中去的方法被称为随机分组。其具体方法有:完全或简单随机分组、分层或区组随机分组(注:在本质上,分层因素与区组因素都是指重要的非试验因素)、分层区组随机分组(事实上,就是同时考察两个重要非试验因素的随机分组)。

值得一提的是,目前在 R 软件中尚未找到直接用于随机化分组的函数,本文暂且借用随机抽样的 sample( ) 函数来代替。不可避免的问题可能会出现,即各组的样本含量可能不相等。补救措施是改变随机数的种子数,直到所生成的随机分组结果符合各组样本含量相等的要求时,固定此随机数种子数,并保留程序。以后再运行此程序时,就能再现以前的符合要求的随机化分组结果。

项目基金:国家高技术研究发展计划课题资助(02015AA020102)

2 几种常见随机分组方法的 R 实现[2-5]

2.1 简单随机分组

2.1.1 将全部受试对象随机均分成两组

【例 1】现有编号为 1~24 的 24 位受试对象,希望将他们随机地均分到“试验组”与“对照组”中去,并显示出随机分组的结果。

解答:在 R 中使用下面的语句就可实现前述的目的。

```
> a<-c(1:24)
> b<-rep(c("试验组","对照组"),12)
> c<-sample(b,24,rep=F)
> d<-cbind(a,c);d
```

以上句的含义如下:

第一句:生成一个名为 a 的向量,其元素为 1~24。

第二句:生成一个名为 b 的向量,其元素由“试验组”和“对照组”交替组成,共重复 12 次,故元素个数为 24。

第三句:生成一个向量 c,其元素是从向量 b 中无放回随机抽样的结果,抽出 24 个元素,本质上,相当于对向量 b 中的 24 个元素进行随机化排列。

第四句:将向量 a 与 c 按列进行合并,生成一个向量 d,并将其输出。

```
      a          c
 [1,] "1"        "对照组"
 [2,] "2"        "对照组"
 [3,] "3"        "对照组"
 [4,] "4"        "对照组"
 [5,] "5"        "对照组"
 [6,] "6"        "试验组"
 [7,] "7"        "试验组"
 [8,] "8"        "试验组"
 [9,] "9"        "试验组"
[10,] "10"       "对照组"
[11,] "11"       "试验组"
[12,] "12"       "对照组"
[13,] "13"       "试验组"
[14,] "14"       "对照组"
[15,] "15"       "试验组"
[16,] "16"       "试验组"
[17,] "17"       "对照组"
[18,] "18"       "对照组"
[19,] "19"       "试验组"
[20,] "20"       "对照组"
[21,] "21"       "对照组"
[22,] "22"       "试验组"
[23,] "23"       "试验组"
[24,] "24"       "试验组"
```

以上是将编号为 1~24(见 a 列)的受试对象随机均分入试验组与对照组(见 c 列)。

2.1.2 将全部受试对象随机均分成三组

【例 2】现有编号为 1~24 的 24 位受试对象,希望将他们随机地均分为“试验组”与“对照组 1”和“对照组 2”三组中去(即按 1∶2 分配样品),并显示出随机分组的结果。

解答:在 R 中使用下面的语句就可实现前述的目的。

```
> a <- c(1:24)
> b <- rep(c("试验组","对照组 1","对照组 2"),8)
> c <- sample(b,24,rep=F)
> d <- cbind(a,c);d
```

第一句:生成一个名为 a 的向量,其元素为 1~24。

第二句:生成一个名为 b 的向量,其元素由“试验组”、“对照组 1”和“对照组 2”交替组成,共重复 8 次,故元素个数为 24。

第三句:生成一个向量 c,其元素是从向量 b 中无放回随机抽样的结果,抽出 24 个元素,本质上相当于对向量 b 中的 24 个元素进行随机化排列。

第四句:将向量 a 与 c 按列进行合并,生成一个向量 d,并将其输出。

```
      a          c
 [1,] "1"        "对照组 2"
 [2,] "2"        "对照组 2"
 [3,] "3"        "试验组"
 [4,] "4"        "对照组 1"
 [5,] "5"        "对照组 2"
 [6,] "6"        "试验组"
 [7,] "7"        "试验组"
 [8,] "8"        "试验组"
 [9,] "9"        "对照组 1"
[10,] "10"       "对照组 2"
[11,] "11"       "对照组 1"
[12,] "12"       "对照组 1"
[13,] "13"       "试验组"
[14,] "14"       "对照组 2"
[15,] "15"       "对照组 2"
[16,] "16"       "对照组 2"
[17,] "17"       "对照组 2"
[18,] "18"       "试验组"
[19,] "19"       "对照组 1"
[20,] "20"       "对照组 1"
[21,] "21"       "对照组 1"
[22,] "22"       "试验组"
[23,] "23"       "对照组 1"
[24,] "24"       "试验组"
```

以上是将编号从 1~24 的 24 个样品(见 a 列)按 1∶2 的比例随机化分组的输出结果,试验组、对照组 1、对照组 2 中各包含 8 个样品。

如何使某次随机化分组结果具有重现性呢?若想使前面刚出现过的随机化分组结果再次出现,就很困难了。只有事先设置好随机化种子[使用

set.seed(n)，n 为非 0 的正整数]，然后生成随机化分组结果；再使用相同的随机化种子，第二次运行相同的生成随机化分组的程序，才能得到相同的随机化分组结果。若在未来的某个时间，还想生成相同的随机化分组结果，事先还得运行与以前使用过的相同的随机化种子。

```
> set.seed(2)
> a <- c(1:24)
> b <- rep(c("试验组","对照组1","对照组2"),8)
> c <- sample(b,24,rep=F)
> d <- cbind(a,c);d
```

以上的第一句为设置随机化种子，后面四句的内容与前面程序相同，已解释过了。

```
      a      c
[1,] "1"    "对照组1"
[2,] "2"    "对照组1"
[3,] "3"    "试验组"
[4,] "4"    "试验组"
[5,] "5"    "试验组"
[6,] "6"    "对照组2"
[7,] "7"    "对照组2"
[8,] "8"    "对照组2"
[9,] "9"    "对照组1"
[10,] "10"  "对照组2"
[11,] "11"  "试验组"
[12,] "12"  "对照组2"
[13,] "13"  "试验组"
[14,] "14"  "对照组1"
[15,] "15"  "对照组2"
[16,] "16"  "对照组1"
[17,] "17"  "对照组1"
[18,] "18"  "对照组1"
[19,] "19"  "对照组1"
[20,] "20"  "试验组"
[21,] "21"  "对照组2"
[22,] "22"  "试验组"
[23,] "23"  "试验组"
[24,] "24"  "对照组2"
```

```
> set.seed(2)
> a <- c(1:24)
> b <- rep(c("试验组","对照组1","对照组2"),8)
> c <- sample(b,24,rep=F)
> d <- cbind(a,c);d
```

以上的第一句所设置的随机化种子与前面相同，后面四句的内容与前面程序也相同，已解释过了。

```
      a      c
[1,] "1"    "对照组1"
[2,] "2"    "对照组1"
[3,] "3"    "试验组"
[4,] "4"    "试验组"
[5,] "5"    "试验组"
[6,] "6"    "对照组2"
[7,] "7"    "对照组2"
[8,] "8"    "对照组2"
[9,] "9"    "对照组1"
[10,] "10"  "对照组2"
[11,] "11"  "试验组"
[12,] "12"  "对照组2"
[13,] "13"  "试验组"
[14,] "14"  "对照组1"
[15,] "15"  "对照组2"
[16,] "16"  "对照组1"
[17,] "17"  "对照组1"
[18,] "18"  "对照组1"
[19,] "19"  "对照组1"
[20,] "20"  "试验组"
[21,] "21"  "对照组2"
[22,] "22"  "试验组"
[23,] "23"  "试验组"
[24,] "24"  "对照组2"
```

以上两批生成的随机化分组结果完全相同，这是由于两批运行的随机化种子数和程序都相同。

2.2 分层或区组随机分组

【例 3】现有编号为 1～24 的 24 位受试对象，男性和女性各 12 例。试将性别视为分层因素，将各层中的受试对象随机地均分入试验组与对照组中去。

解答: 在 R 中使用下面的语句就可实现前述的目的。

```
> set.seed(20160407)
>a <- c(1:24)
>b1 <- rep(c("男"),12)
>b2 <- rep(c("女"),12)
> c <- c(b1,b2)
> d1 <- rep(c("试验组"),12)
> d2 <- rep(c("对照组"),12)
> e <- c(d1,d2)
> f <- sample(e,24,rep=F)
> g <- cbind(a,c,f);g
```

以上有 10 行独立的 R 语句,其中,最后一行有两句,用";"(即英文输入状态下分号)隔开。

第 1 行: 固定随机数的种子数为 20160407。

第 2 行: 生成受试对象的编号 1~24 号,作为向量 a 的 24 个元素。

第 3、4 两行: 生成两个各有 12 个元素(或长度为 12)的向量 b1 与 b2,其内容分别由 12 个"男"、12 个"女"组成。

第 5 行: 将 b1 与 b2 两个向量左右拼接成一个长度为 24 的向量 c。

第 6、7 两行: 生成两个各有 12 个元素(或长度为 12)的向量 d1 与 d2,其内容分别由 12 个"试验组"、12 个"对照组"组成。

第 8 行: 将 d1 与 d2 两个向量左右拼接成一个长度为 24 的向量 e。

第 9 行: 调用 sample( ) 函数,从长度为 24 的向量 e(其元素中,前 12 个为"试验组"、后 12 个为"对照组")中无放回地随机抽取 24 个元素(注意: 不能保证各层一定能抽取相同数目的"试验组"与"对照组",如果要保证各层"试验组"与"对照组"数目相同,可分别按层进行随机分组),生成向量 f。

第 10 行: 前一句是将向量 a、c 和 f 按列进行合并生成一个数据框 g; 后一句要求输出数据框 g 的内容。

```
      a     c      f
[1,] "1"  "男"  "试验组"
[2,] "2"  "男"  "试验组"
[3,] "3"  "男"  "对照组"
[4,] "4"  "男"  "试验组"
[5,] "5"  "男"  "对照组"
[6,] "6"  "男"  "对照组"
[7,] "7"  "男"  "对照组"
[8,] "8"  "男"  "对照组"
[9,] "9"  "男"  "试验组"
[10,] "10"  "男"  "对照组"
[11,] "11"  "男"  "试验组"
[12,] "12"  "男"  "试验组"
[13,] "13"  "女"  "试验组"
[14,] "14"  "女"  "对照组"
[15,] "15"  "女"  "试验组"
[16,] "16"  "女"  "试验组"
[17,] "17"  "女"  "对照组"
[18,] "18"  "女"  "对照组"
[19,] "19"  "女"  "试验组"
[20,] "20"  "女"  "对照组"
[21,] "21"  "女"  "对照组"
[22,] "22"  "女"  "试验组"
[23,] "23"  "女"  "试验组"
[24,] "24"  "女"  "对照组"
```

以上是输出结果,第 1 列为 R 软件自动生成的行号; 第 2 列为受试对象的 1~24 个编号; 第 3 列为受试对象的性别(注意: 开始应将 12 位男性受试对象作为第一层编号为 1~12 号、将 12 位女性受试对象作为第二层编号为 13~24 号); 第 4 列相当于[因为是借用随机抽样函数(即 sample( ) 函数)而不是真正的随机分组函数(此种函数在 R 中是否存在,笔者暂时尚不能确定)]将各层内相同性别的受试对象随机地均分入"试验组"与"对照组"。

参考文献

[1] 胡良平. 课题设计与数据分析—关键技术与标准模板[M]. 北京: 军事医学科学出版社, 2014: 93-103.

[2] 黄文, 王正林. 数据挖掘: R 语言实战[M]. 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]. 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]. 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. 电子工业出版社, 2015: 417-421.

(收稿日期: 2016-12-05)

(本文编辑: 唐雪莉)