科研方法专题

快速学会使用 R 软件的方法

谷恒明[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】 本文目的是使用户快速了解 R 软件的概况和 R 语言基础，并能快速采用 R 软件解决常用试验设计与统计分析问题。通过扼要介绍 R 软件的概况、R 语言的基础知识、在 R 环境中读入和存储数据文件的方法以及用 R 软件解决九个与试验设计和统计分析有关的实际问题，使用户能方便快捷地实现前述目的。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，能解决与试验设计、数据可视化和各种统计分析有关的问题。

【关键词】 R 软件；函数；向量；矩阵；数组；数据框；列表

中图分类号：R195.1 文献标识码：A doi：10.11886/j.issn.1007-3256.2016.06.001

How to learn the usage of R software quickly

Gu Hengming[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 enable users to know the general situation of R software and the foundation of R language, and use R software to solve the general problems of the experimental design and statistical analysis rapidly. Users easily and quickly achieve the aim mentioned above through the brief introduction of R software, the basic knowledge of R language, the method of accessing data in the R environment, and using R software to solve the actual problems related to nine experimental designs and statistical analyses. The fact shows that R software is easy to obtain, learn and use; R software can solve the experimental designs, data visualization and various issues related to statistical analyses due to its the powerful function and wide application scope.

【**Key words**】 R software; Function; Vector; Matrix; Array; Data Frame; List

1 R 软件概述[1-3]

1.1 R 软件的历史

新西兰奥克兰大学的 Ross Ihaka 与 Rontleman 一道基于 S 语言开发了一个面向对象的编程环境，简称为“R 软件”，它是一个免费且开源的计算机运行环境。

1.2 R 软件的功能

R 软件的主要功能是可视化、数学计算与统计分析。所谓可视化，就是不仅可以绘制一般的统计图，还可绘制很多复杂且精细的图形；所谓数学计算与统计分析，其所涉猎的范围也是非常宽泛的。R 软件是迄今为止全球统计软件中内容最丰富的，几乎涵盖了全世界已经发现的各种算法，有些虽然暂时可能还没有加入进来，但 R 软件包每天都由全世界的学术志愿者在追加新的内容，而且增加的数量和速度相当可观。

1.3 R 软件的获取

由于 R 软件是一个免费且开源的统计分析软件，所以 R 软件的获取是极其方便的。建议将 R 软件安装在一直可以上网的计算机（因为保密计算机是不允许与国际互联网连接的）上，这样便于 R 软件的更新。

R 软件官网地址：http://www.r-project.org/；与 R 软件配套使用的 IDE（Integrated Development Environment，集成开发环境，简单可理解为编程工具）的地址：http://www.rstudio.com/。

从前述第一个网址可以下载到最新版本的

项目基金：国家高技术研究发展计划课题资助（2015AA020102）

R 软件;从前述第二个网址可以下载到最新版本的 rstudio 软件。

在 R 软件环境中,每次只能输入一行代码。当然,一行中可以只输入一个 R 语言的语句或函数或命令,也可以输入多个 R 语言的语句或函数或命令,但各语句之间必须用分号隔开。只要一键入回车键,就立即执行。

在 rstudio 软件环境中,一次可以输入一个文本,还可以将其存储起来。于是,用户可以把整个文本选中以后,再一次性提交给系统去执行。

1.4 R 软件的界面

在下载并成功安装了 R 软件包后,就会在计算机桌面上出现一个 R 软件快捷方式图标,其形状就是一个笔画很粗的大写英文字母“R”。用鼠标左键双击此图标,就可以启动 R 软件。在其窗口的左上角上有四行内容,从上到下依次为:

第 1 行为“R RGui(64 - bit)”,指明系统为 64 - 位;第 2 行为菜单栏,分别为“文件”、“编辑”、“查看”、“其他”、“程序包”、“窗口”、“帮助”;第 3 行为八个快捷图标,分别为“打开文件”、“加载工作空间”、“保存工作空间”、“复制”、“粘贴”、“复制并粘贴”、“终止当前计算”和“打印”;第 4 行为“R R console”,该行表明在大窗口内嵌套着的这个小窗口名叫“R 软件运行环境中与用户交流信息的控制台,简称控制台,具体录入信息或命令的位置在此小窗口的最下面,以‘>’为标志,它被称为‘提示符’”。

若选中菜单栏上的“程序包”,再选择其中的第一行“加载程序包”,就会弹出一个长方形的窗口。在这个长方形窗口内共呈现了约 30 个对象,它们当中有些是“程序包”,还有一些是“函数”。事实上,它们之间的区别就在于:“程序包”中包含多个不同的“函数”,而“函数”就是一个完成某项任务的“独立程序”。由此可知,所谓用 R 软件包来实现统计计算或绘制图形,实际上就是调用某个程序包中的某个函数或调用一个独立的函数。

1.5 R 软件包中的函数

1.5.1 R 软件包中程序包和函数的数量

R 软件的一切操作几乎都是通过“对象(注:函数是最常被使用的一种对象)”来实现的。从表面上看,R 软件包中有许多“程序包”。粗略地说,迄今为止 R 软件包中约有近万个程序包,其数目是不确定的。R 软件官方网站(www.r-project.org)上显示,到 2015 年 12 月 2 日止,已有 7562 个程序包。但仔细查看,还有一些程序包(如 base、graphic、stats、stats4 等)不在列表之中。其实,在很多所谓的“程序包”中,类似 base 和 stats 这样的程序包(其内包含很多函数)并非很多,大多数被冠以“程序包”的东西其实就是一个“函数”,例如 boot、class 等。正因如此,当你看到如下的递增数目才不会感到惊讶:从 R 软件官网上按时间排序的结果可知:2015 年 12 月 1 日就增加了大约 17 个所谓的“程序包”,2015 年 12 月 2 日就增加了大约 23 个所谓的“程序包”。事实上,它们很可能都是“函数”。

在 R 软件的所谓“程序包”中,除了少量真正的程序包与函数外,还有一些“对象”介于它们二者之间。例如 foreign。它既不是一个真正的程序包,也不是一个函数,而是一个 R 软件与其他统计软件(如 SAS、SPSS)进行数据格式互相交换的“界面”,更确切地说,它是一个不同统计软件信息交换的“接口”。这一点可以通过在 R 软件的控制台发送下面两条命令来证实:

>help(foreign) 回车

>?? foreign 回车

第一条命令为启用“帮助功能”寻求帮助信息。然而,此时会显示出错信息;第二条命令为启用“检索功能”寻求检索结果。检索的结果会弹出一个窗口,因篇幅所限,展示窗口的图片从略。此窗口中的内容表明:foreign 是一个 R 软件与其他统计软件数据格式转换的一个界面或接口(Foreign Function Interface)。

1.5.2 往已安装的 R 软件包中追加新程序包的方法

通过 R 软件窗口上的菜单栏“程序包”弹出的长方形窗口可知:在用户所安装的 R 软件包中,真正的“程序包(其中有些是函数)”的数量是十分有限的。由此可知,用户是经常需要往已安装的 R 软件包中追加新程序包。具体方法如下:

若需要某些其他程序包时,当用户正在使用的计算机是与网络连接的,可以随时安装或更新程序包。例如,现在在线安装一个新的程序包

AggregateR(其实,它是一个函数。值得注意的是:这个函数在已出版的有关 R 语言的书中是 aggregate,后来被更新成 AggregateR;原先有 tapply 函数,现在它似乎被取消掉了,这些都需要在连网计算机上选定了"镜像(参见下面的一个段落)"后,从"程序包"窗口中查找才能发现)。具体加载方法如下:

> install.packages("AggregateR") 回车

回车后,会立即弹出一个名为"HTTPS CRAN mirror"长方形窗口(就是让用户指定"镜像",用通俗的语言表达,就是指定从哪个国家的哪个服务器上下载用户所需要的程序包。一般来说,在此窗口中所列出的许多国家的具体服务器中,R 软件的全部内容是基本相同的,只是最近更新时间不同,可能在"新包或函数"的数量上略有区别。通常用户选择离自己最近的服务器,例如中国就有 4 个服务器,分别分布在北京、武汉、广州等地点)。

笔者选定"China(Beijing 4) [https]",然后按"OK"按钮,计算机就会从指定"镜像"上自动下载并安装所要求的程序包或函数。

若不使用 install.packages() 函数,可通过 R 软件的菜单栏中"程序包"弹出的窗口,先选定"设定 CRAN 镜像",再选定"安装程序包"来完成与前面类似的任务。

1.6 R 软件的工作目录

1.6.1 如何知道 R 软件启动后的工作目录

启动 R 软件之后,系统会将当前工作目录(即存储信息的文件夹)自动设置为安装 R 软件的目录,通过在控制台键入函数 getwd() 回车后显示的信息,便可得知。

> getwd() 回车

例如,在计算机上键入上述函数回车后就会显示如下的信息:

[1] "C:/Users/hu/Documents"

这个输出结果表明:在此 R 软件中,一旦产生了需要存储的信息,都将被存储在 C 盘上的三级文件夹内:第一级为"Users",第二级为"hu",第三级为"Documents"。

1.6.2 如何改变 R 软件的工作目录

若用户想改变工作目录,需要事先创建一个工作目录。通常就是事先在指定盘上创建一个文件夹,以便在本次启动 R 软件后且在退出 R 软件之前将需要存储的信息存入指定的文件夹中去。例如,在计算机 G 盘上创建了一个名为 studyr 的文件夹,然后选中刚启动 R 软件控制台界面上的菜单"文件",再选中弹出的小窗口内的"改变工作目录",便弹出一个大窗口,再根据窗口中的提示信息,可以选定用户的"工作目录"。

事实上,改变工作目录的最简单方法是在控制台上输入函数"setwd()"并回车即可。具体方法如下:

> setwd("G:\studyr") 回车

1.7 如何使用 R 软件的帮助功能

R 软件包中有强大的帮助功能,当启动 R 软件之后,在其操作窗口的第二行最后位置有菜单"帮助",选中它后会弹出一个小窗口。根据此小窗口内的提示信息,逐一选中并阅读,就可找到用户关心的很多内容及其解决方法。

2 使用 R 软件需要了解的基本知识[1-3]

2.1 在 R 中数据的组织形式的概述

使用任何统计软件时,首先需要提供待分析的数据。在 R 中,有如下多种方法可将数据按某种形式组织在一起。然后,R 软件就可对其实现可视化或统计分析。

2.2 用 c() 函数生成向量

2.2.1 生成自然顺序的向量

```
> x1 <- c(12:24); x1
 [1] 12 13 14 15 16 17 18 19 20 21 22 23 24
```

以上是生成一个名为 x1 的向量,其元素分别为 12~24 的自然数。

2.2.2 生成指定分量的向量

```
> x2 <- c(13,21,45,19,28,79); x2
[1] 13 21 45 19 28 79
```

以上是生成一个名为 x2 的向量,其元素分别为"13、21、45、19、28、79"。

2.3 用 matrix() 函数生成矩阵

2.3.1 生成以自然数为元素的矩阵

```
> A <- matrix(1:30,nrow=6,ncol=5); A
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    7   13   19   25
[2,]    2    8   14   20   26
[3,]    3    9   15   21   27
[4,]    4   10   16   22   28
[5,]    5   11   17   23   29
[6,]    6   12   18   24   30
```

以上是生成一个名为 A 的 6×5 矩阵，其元素分别为“1～30”的自然数，按“列”的顺序放置。

```
> B <- matrix(1:30,nrow=6,ncol=5,byrow=T);B
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    2    3    4    5
[2,]    6    7    8    9   10
[3,]   11   12   13   14   15
[4,]   16   17   18   19   20
[5,]   21   22   23   24   25
[6,]   26   27   28   29   30
```

以上是生成一个名为 B 的 6×5 矩阵，其元素分别为“1～30”的自然数，按“行”的顺序放置。

2.3.2 生成指定元素的矩阵

```
> C <- matrix(c(2,4,6,8,7,5,3,1),nrow=2,ncol=4,byrow=T);C
     [,1] [,2] [,3] [,4]
[1,]    2    4    6    8
[2,]    7    5    3    1
```

以上是生成一个名为 C 的 2×4 矩阵，其元素分别为“2、4、6、8、7、5、3、1”的自然数，按“行”的顺序放置。

2.4 用 array( ) 函数生成数组

2.4.1 生成二维数组举例

```
>D <- array(data=1:12,dim=c(3,4));D
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
```

以上就是所创建的二维数组 D 的具体内容。

2.4.2 生成三维数组举例

```
>E <- array(data=1:24,dim=c(4,3,2));E
, , 1
     [,1] [,2] [,3]
[1,]    1    5    9
[2,]    2    6   10
[3,]    3    7   11
[4,]    4    8   12
```

以上显示的是第 1 层上的二维数组。

```
, , 2
     [,1] [,2] [,3]
[1,]   13   17   21
[2,]   14   18   22
[3,]   15   19   23
[4,]   16   20   24
```

以上显示的是第 2 层上的二维数组。

2.5 用 factor( ) 函数生成因子

2.5.1 何为因子

所谓因子，实际上就是在进行试验设计时需要考察的影响因素。例如，性别(sex) 通常只有男(M) 与女(F) 两个水平。即使研究者观测了一万个人的性别，似乎 sex 有一万个取值，但完全不同的取值只有两个，它们就是性别(sex) 这个“因子”的两个水平。

2.5.2 举例说明

```
> sex <- c("M","F","F","M","M","M","F","F","M","F")
```

以上语句生成一个叫 sex 的向量，代表已观测到的 10 名受试者的性别。

```
> sex_factor <- factor(sex);sex_factor
```

以上语句的目的是调用 factor( ) 函数，将 sex 变量中的所有观测归纳成一个叫 sex_factor 的因子，并输出该因子。

```
[1] M F F M M M F F M F
Levels: F M
```

以上输出的结果表明，该因子的原始观测值有 10 个(列在第一行)，所形成的水平被列在第二行

上，其两个水平分别为“F”和“M”。

2.6 用 data.frame() 函数生成数据框

2.6.1 何为数据框

在 R 中，所谓的数据框，实际上就是表达“数据库结构”的一种表格。此种表格中，假定有 n 行 m 列，这 n 行代表 n 个受试对象，而这 m 列代表从每位受试对象身上观测的 m 个变量的具体取值。

注意：数据框中各列的长度应相同，即各列中元素的个数相同。各列变量的性质（数值型或字符型）可以不同，但同一列中元素的性质必须相同。

2.6.2 举例说明

```
> sex < - c( " M"," F"," F"," M"," M"," M"," F"," F"," M"," F")
> height < - c( 165,171,163,184,169,192,158,167,183,176)
> data < - data. frame( sex,height) ; data
```

前两句创建两个向量，分别为 sex 和 height；第三句创建一个名为 data 的数据框并显示其内容（即将前两行的内容按“列”呈现出来，因篇幅所限，此处从略）。

2.7 用 list() 函数生成列表

2.7.1 何为列表

R 中的列表（list）是一个有序的对象集合。用户可以通过列表的位置来引用列表中的元素。在一个列表中，可以放置多种不同的数据对象。

2.7.2 列表举例

```
> a < - c( 1:5)          （创建一个向量 a）
> b < - matrix           （创建一个矩阵 b）
( 1:25,5,5)
> c < - list( a,b)       （创建一个列表 c，使其包含 a 和 b）
> c                      （要求输出列表 c 的内容）
[ [1] ]
[1] 1 2 3 4 5
（说明：这里显示了列表中的第一部分内容）
[ [2] ]
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    6   11   16   21
[2,]    2    7   12   17   22
[3,]    3    8   13   18   23
[4,]    4    9   14   19   24
[5,]    5   10   15   20   25
（说明：这里显示了列表中的第二部分内容）
```

3 以文件的方式在 R 中输入和输出数据[1-3]

3.1 不同数据格式的概述

前面讲的内容都属于直接将数据放在 R 程序语句中，以某种组织形式呈现出来。然而，在实际使用中，用户经常需要从外部文件中读取以第三方格式（如文本文件、EXCEL 文件、SAS 数据集或 SPSS 数据集等）存储的数据，使其成为能被 R 软件识别和调用的某种格式（如数据框或矩阵等形式）。当然，有时也需要将 R 中已创建的数据集按某种第三方数据格式存储到外部设备上去。

3.2 向 R 环境中读入和输出几种不同格式数据的方法

3.2.1 如何用 read.table( ) 函数在 R 软件环境中以文本格式输入数据

假定我们已将包含“name、drug、blood、age、height、weight、effect”7 个变量 10 个观测的资料以文本格式的数据文件存储在 G 盘的 studyr 的文件夹中，数据文件的名称为 raw_data.txt。当启动 R 软件后，首先改变工作目录，使其成为“G\studyr”。然后在控制台键入如下内容：

```
>x < - read. table( " raw_data. txt",header = TRUE)  回车
```

就可将原先的 10 行 7 列数据读入 R 运行环境中，其数据集名称为 x。值得注意的是：若“name、drug、blood”是字符型变量，并假定“name”的具体取值为“带空格的字符串”（如：ZhangSan）时，创建文本文件时，所有的“带空格的字符串”必需被放置在英文双引号之内，即“"Zhang San"”。

3.3.2 如何用 write.table( ) 函数在 R 软件环境中以文本格式输出数据

第一步：在控制台上使用下面的命令，可为已创建的数据集 total_data 指定一个输出文件名 new_output1：

> write.table(total_data,file = "new_output1")

第二步:在控制台上使用下面的命令,可将已创建的数据集 total_data **以文本文件的格式且文件名为** new_output1 **存入当前工作目录中:**

> write.table(total_data,file = "new_output1", quote = FALSE, row.names = FALSE, col.names = TRUE)

其中,某些参数的含义如下:

quote = FALSE **要求各列变量及其取值不加引号。**

row.names = FALSE **要求各行前不加行名称(注:行名称就是行号)。**

col.names = TRUE **要求各列头上应该保留变量名。**

说明:上面仅介绍了"文本格式"的数据的读入和存储方法,因篇幅所限,其他格式的数据文件的读入与存储方法,参见文献[1-3],此处从略。

4 用 R 软件解决试验设计与统计分析问题举例[1-4]

4.1 用 R 产生随机数并绘制直方图举例

【例 1】试生成 10 000 个服从均值为 165 cm、标准差为 20 cm 的正态分布的随机数,并用直方图展示它们(模拟某地 10 000 个正常成年人的身高数据)。

解答:为实现题中的目标,可使用下面的语句(说明:每个语句开始的">"为 R 软件的提示符,下同,不再赘述)。

> x <- rnorm(10000,165,20)

> hist(x,prob = T,main = "normal distribution (mean = 165 cm,sigma = 20 cm)")

第一句的目的是生成 10 000 个服从均值为 165 cm、标准差为 20 cm 的正态分布的随机数。

第二句的目的是绘制这 10 000 个随机数的直方图。

产生的结果从略。

4.2 用 R 进行随机抽样举例

【例 2】假定掷一枚质地均匀的骰子(有 6 个面,每个面上分别有 1、2、3、4、5、6 个点),现重复抛骰子 20 次,显示有放回随机抽样的结果。

解答:在 R 中使用下面的语句就可实现前述的目的。

> sample(c(1:6),20,rep = T)

[1] 3 2 4 1 5 1 6 1 6 4 3 6 3 6 5 6 4 4 2 2

以上是 20 次试验的结果,每次试验的结果是 1~6六个数字中的一个出现。这批试验的结果表明,1 出现了 3 次、2 出现了 3 次、3 出现了 3 次、4 出现了 4 次、5 出现了 2 次、6 出现了 4 次。

4.3 用 R 进行随机分组举例

【例 3】现有编号为 1 到 24 号的 24 位受试对象,希望将他们随机地均分为"试验组"与"对照组"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

以上语句的含义如下:

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(从略)。**

输出结果显示:将编号为 1~24 **号(见** a **列)的受试对象随机均分入试验组与对照组(见** c **列)。**

4.4 用 R 估计样本含量举例

【例 4】假定有两种处理方法,有一个定量评价指标。已知两组的均值之差量为 1.58**、合并标准差为** 5.97**、显著性水准为** 0.05**、检验效能为** 80%**。试估计各组至少需要多大的样本含量。**

解答:这是一个单因素两水平设计一元定量资料假设检验之前估计样本含量的问题,在 R **的控制台上发送下面的一条命令,就可实现。**

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

输出结果显示:每组应选取约 226 **例。**

4.5 用 R 估计检验效能举例

【例 5】假定有两种处理方法,有一个定量评价指标。已知两组的均值之差量为 1.58**、合并标准差为** 5.97**、显著性水准为** 0.05**、各组的样本含量均为**

226 例。试估计能达到的检验效能是多大。

解答: 这是一个单因素两水平设计一元定量资料假设检验之前估计检验效能的问题,在 R 的控制台上发送下面的一条命令,就可实现。

```
> power.t.test(n = 226,sig.level = 0.05,delta = 1.58,sd = 5.97)
```

输出结果显示: 检验效能大约为 80.16%。

4.6 用 R 分析定性资料举例

【例 6】试分析文献[4]中第 374 页的"表 13-3"的资料。

解答: 该资料属于横断面研究设计的四格表资料,对该资料进行 χ^2 检验所需要的 R 程序如下。

```
> rownum <- c(41,32)
> colnum <- c(43,8)
> chisq.test(rbind(rownum,colnum))
```

R 结果解读: 采用 Pearson 卡方检验并进行耶茨校正,$\chi_c^2 = 9.637$,$df = 1$,$P = 0.001907$。

【结论】该校男、女学生英语六级考试通过率不等,由数据显示,女生通过率高于男生。

4.7 用 R 分析定量资料举例

【例 7】试分析文献[4]中第 268 页的"例 8-5"的资料。

解答: 该资料属于单因素两水平设计一元定量资料,用 R 对此资料进行差异性分析(包括前提条件的检验)所需要的程序如下。

```
>x1 <- c(66,65,66,68,62,65,63,66,68,62)
>x2 <- c(64,61,57,65,65,63,62,63,64,60)
>shapiro.test(x1)
>shapiro.test(x2)
>var.test(x1,x2)
>t.test(x1,x2,alternative = "two.sided",var.equal = TRUE)
>t.test(x1,x2,alternative = "two.sided",var.equal = FALSE)
>wilcox.test(x1,x2)
```

输出结果表明: 两组定量资料均满足正态性和方差齐性要求(故近似 t 检验与秩和检验的结果都可以不看了);可以采用单因素两水平设计一元定量资料 t 检验分析资料。其结果为: $t = 2.5705$,$df = 18$,$P = 0.01926$,说明两个平均值(65.1 与 62.4)之间差异有统计学意义。

【结论】此小麦品种第 5 代平均株高高于第 6 代平均株高,株高性状没有达到稳定状态。

4.8 用 R 实现简单相关分析举例

【例 8】采用文献[4]中第 419 页的"表 16-1"的资料,进行简单相关分析。

解答: 该资料属于单组设计二元定量资料,用 R 对此资料进行直线相关分析所需要的程序如下。

第一步,将资料按文本格式存储在指定位置上,例如,存储在 G:\study 文件夹中,文件名为 xuejiaxinjijia.txt,其第一行可以为变量名,第一列与第二列分别为变量 x、y 及其取值。

第二步,编写 R 语句,读入数据。

```
> dataset <- read.table("G:/studyr/xuejiaxinjijia.txt",header = T)
```

第三步,编写 R 语句,调用 plot() 函数,考察两变量之间的散布图是否呈现线性变化趋势。

```
> plot(y ~ x,data = dataset)
```

绘出的散布图(此处从略)显示,散点呈先线性变化趋势,可以进行直线相关分析。

第四步,编写 R 语句,调用 cor.test() 函数,对两定量变量进行 Pearson 直线相关分析(包括求出相关系数及其置信区间和假设检验)。

```
> cor.test(~ y + x,data = dataset)
```

输出结果表明: $r = 0.8463629$,总体相关系数 ρ 的 95% 置信区间为 [0.6456651,0.9376883];采用 t 检验考察总体相关系数为零的零假设是否成立,得: $t = 6.742$,$df = 18$,$P < 0.0001$,说明总体相关系数不为零。

【结论】正常人血液中钾元素含量(mmol/L)与心肌中钾元素平均含量(mg/g)之间呈正向线性变化趋势。

4.9 用 R 实现直线回归分析举例

【例 9】采用文献[4]中第 419 页的"表 16-1"的资料,以心肌中钾元素平均含量为因变量 y、以血液中钾元素含量为自变量 x,进行直线回归分析。

解答: 这是一个简单直线回归分析问题,创建文本文件和绘制散布图同前例(此处从略),下面只写出与读取数据、构建直线回归方程和输出计算结果有关的语句。

```
> dataset <- read.table("G:/studyr/xuejiaxinjijia.txt",header = T)
> equation <- lm(y ~ x,data = dataset)
> summary(equation)
```

输出结果表明: 截距、斜率与 0 之间的差别均有统计学意义。

【结论】依据正常人血液中钾元素含量(mmol/L)推测心肌中钾元素平均含量(mg/g)的直线回归方程为: $\hat{Y}=25.7246+4.8387x$。

参考文献

[1] 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,2015:8-15,134-156.

[2] 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,2015:1-12,54-73,126-168.

[3] 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,2015:3-49,389-421.

[4] 胡良平. SAS 常用统计分析教程[M]. 2 版. 北京:电子工业出版社,2015:264-391,419-434.

(收稿日期:2016-12-03)

(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和北京大学口腔医学院客座教授;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、中国生物医学统计学会副会长,《中华医学杂志》等 10 余种杂志编委和国家食品药品监督管理局评审专家。主编统计学专著 45 部,参编统计学专著 10 部;发表第一作者学术论文220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与 SAS 实现、各种层次的统计学教学培训和咨询工作。