第一章 R语言基础知识

复旦大学附属肿瘤医院 周支瑞

1.1 R语言在医学科研与论文写作中的应用前景

典型的数据分析步骤

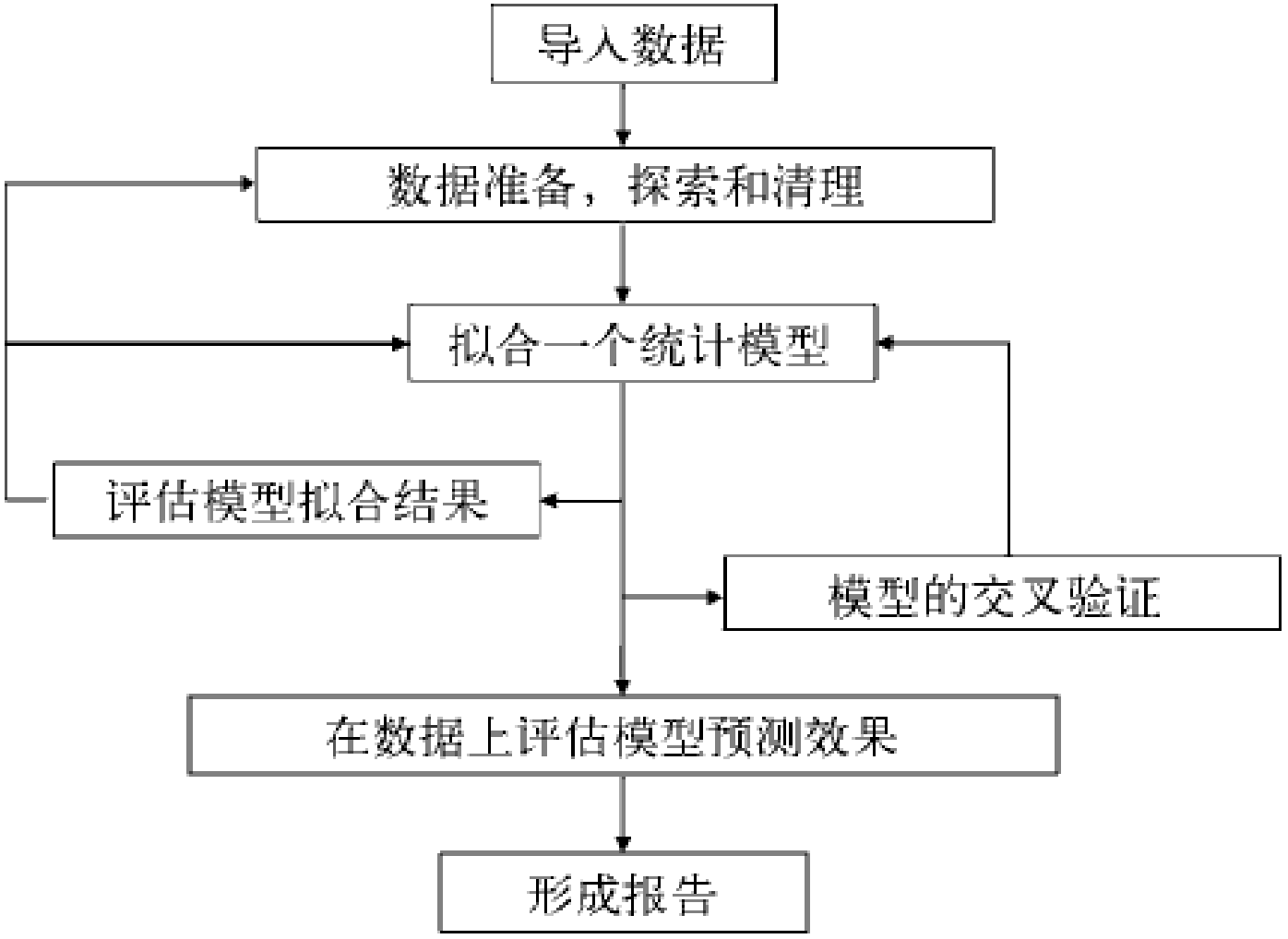


图1-1　典型的数据分析步骤

为何要使用R?

丁香公开课
CLASS.DXY.CN

- 市面上也有许多其他流行的统计和制图软件，如Microsoft Excel、SAS、IBM SPSS、Stata以及Minitab。为何偏偏要选择R?

四种经典统计学软件的比较 -- 软件仅仅是工具，会用一种即可

丁香公开课
CLASS.DXY.CN

软件	优点	不足	推荐指数
Stata	半开源，小巧轻便，功能齐全，医学统计分析的全才	付费使用，命令操作，统计图有时不够美观	*****
IBM SPSS	菜单操作，简单易学，适合初学者，可满足医学统计学需求	固定模块，非开源；付费试用，功能有限，一些高级方法无法实现	*****
R	**完全开源，免费使用，功能强大，统计分析的全才，绘图功能强大**	**命令操作，需要一定的R语言基础**	*****
SAS	统计功能强大，绘图功能强大，FDA指定软件	固定模块，付费使用，命令操作，需要SAS编程基础	****

R语言在医学科研中的应用

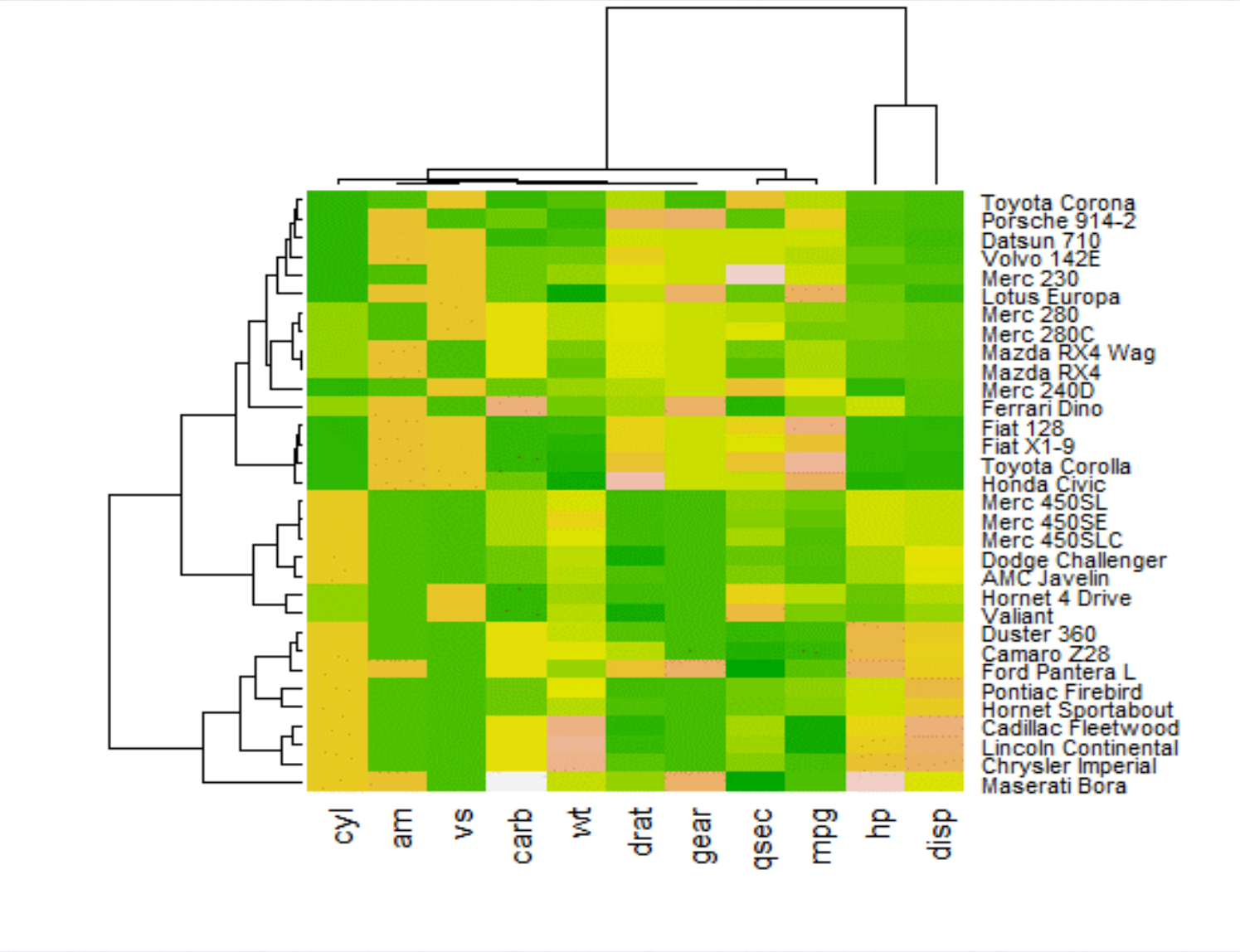
丁香公开课
CLASS.DXY.CN

- 数据清洗：SAS，R等
- 统计分析：SAS，Stata, SPSS, R等
- 画图功能：SAS，Stata, R等

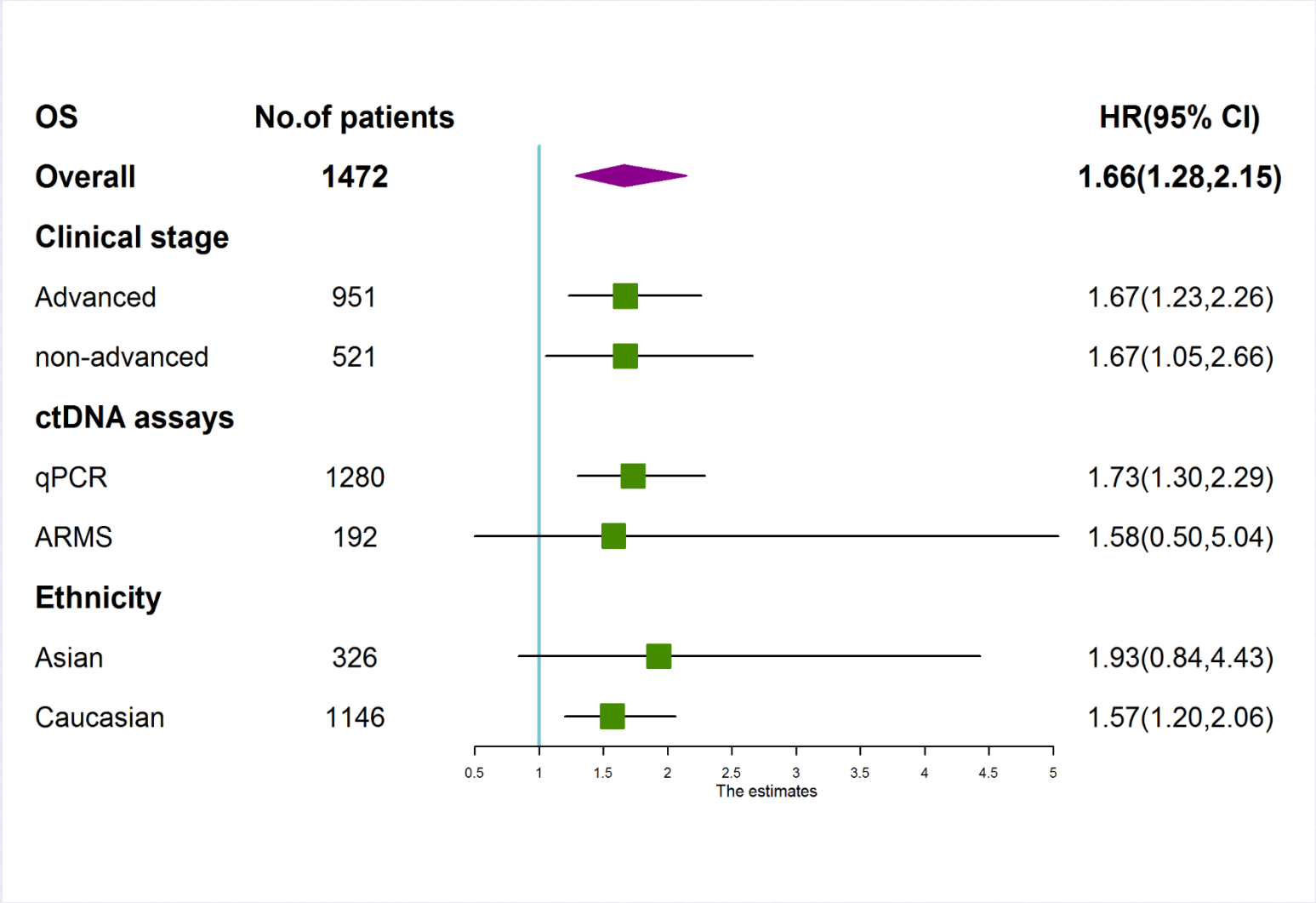
- 从近十年发表的医学文献来看，使用到R语言作为统计分析与绘图的文献逐年增多；R语言方法学文献逐年增多

热图

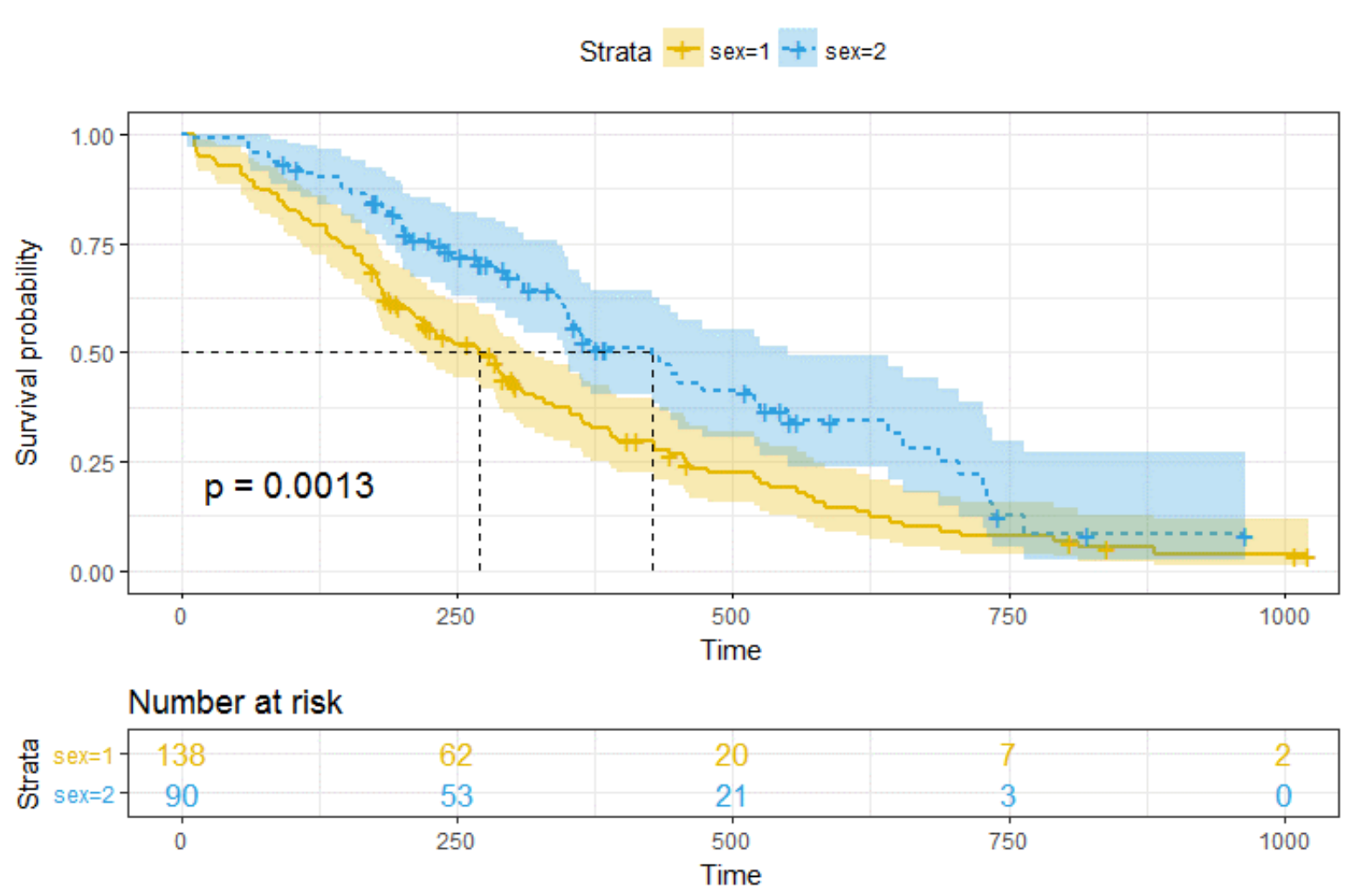
丁香公开课
CLASS.DXY.CN



亚组分析森林图


OS
No.of patients
HR(95% CI)
Overall
1472
1.66(1.28,2.15)
Clinical stage
Advanced
951
1.67(1.23,2.26)
non-advanced
521
1.67(1.05,2.66)
ctDNA assays
qPCR
1280
1.73(1.30,2.29)
ARMS
192
1.58(0.50,5.04)
Ethnicity
Asian
326
1.93(0.84,4.43)
Caucasian
1146
1.57(1.20,2.06)
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
The estimates

生存分析曲线


Strata
sex=1
sex=2
1.00
0.75
0.50
0.25
0.00
Survival probability
p = 0.0013
0
250
500
750
1000
Time
Number at risk
Strata
sex=1
sex=2
138
62
20
7
2
90
53
21
3
0
0
250
500
750
1000
Time

中国知网 医学相关R语言方法学文献概览

丁香公开课
CLASS.DXY.CN

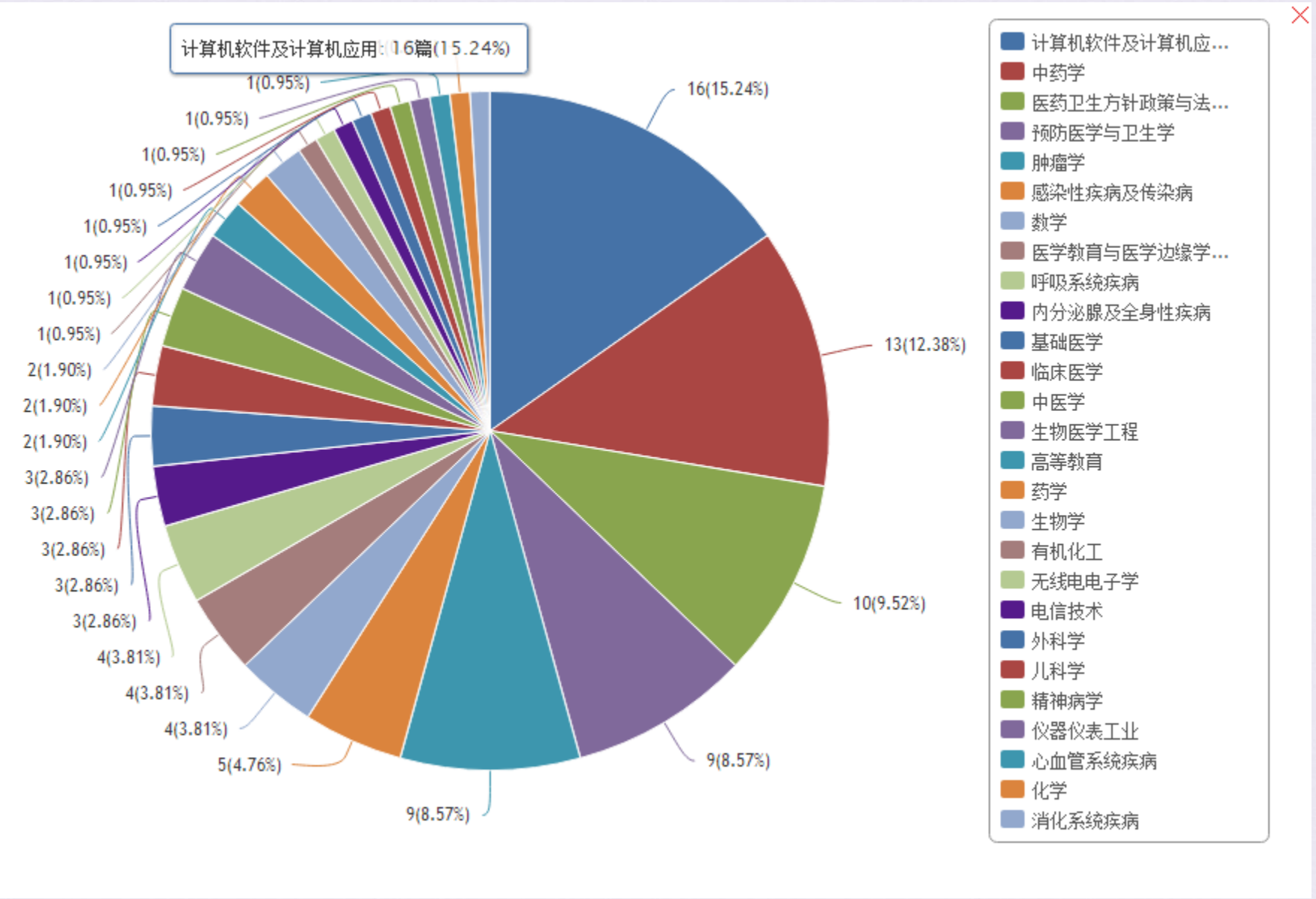
学科分布

数据库：文献库

检索条件：指定条件

分布项：学科

显示数量： 10 20 30



中国知网 医学相关R语言方法学文献概览

丁香公开课
CLASS.DXY.CN

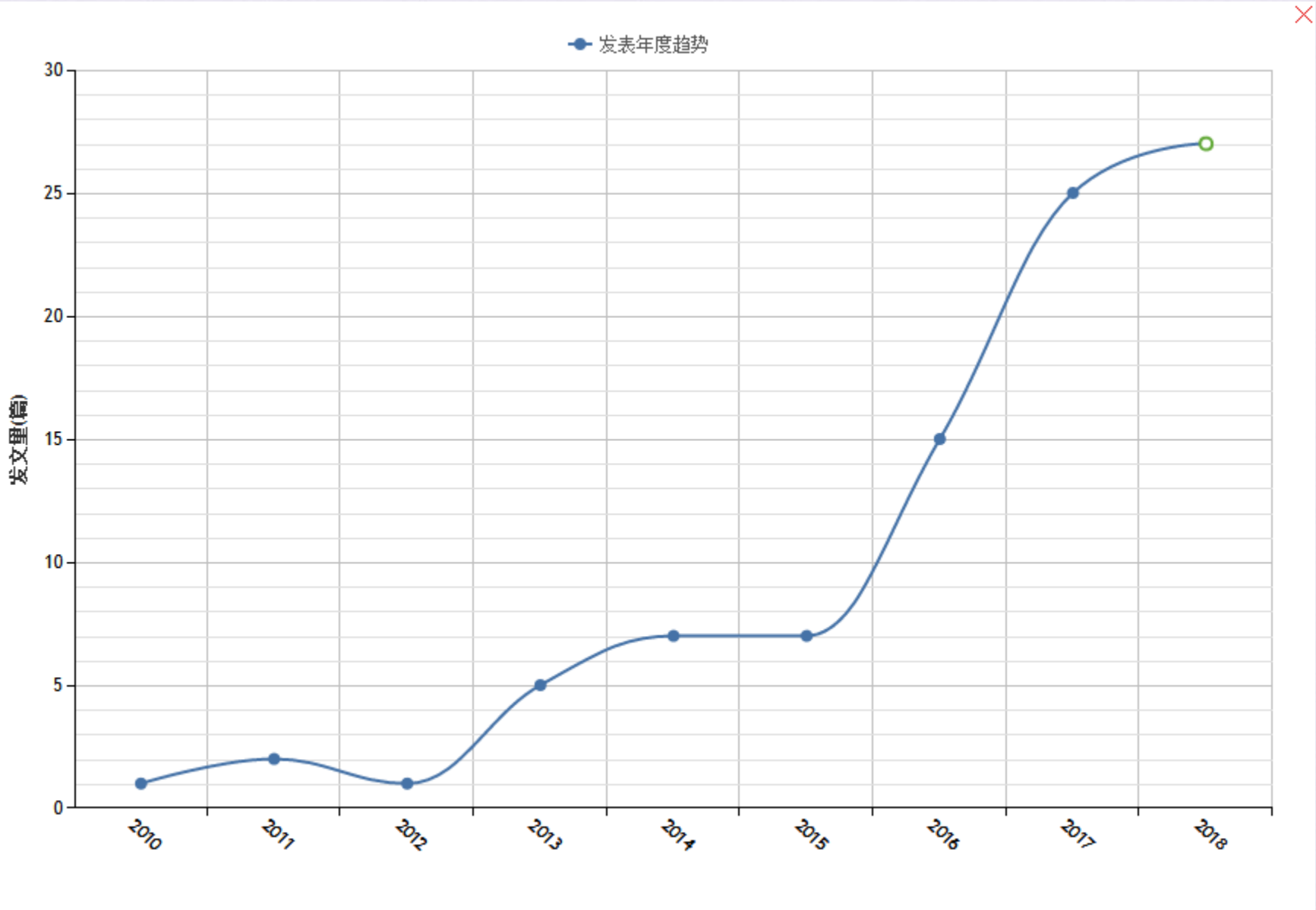
发表年度分布

数据库：文献库

检索条件：指定条件

分布项：发表年度

显示数量：全部



1.2 R 软件简介与安装

R语言官方下载地址

https://www.r-project.org/

Rstudio 官方下载地址

丁香公开课
CLASS.DXY.CN

https://www.rstudio.com/

工作空间

丁香公开课
CLASS.DXY.CN

- 工作空间（workspace）就是当前R的工作环境，它存储着所有用户定义的对象（向量、矩阵、函数、数据框、列表）

表1 用于管理R工作空间的函数

函　数	功　能
getwd()	显示当前的工作目录
setwd("*mydirectory*")	修改当前的工作目录为 *mydirectory*
ls()	列出当前工作空间中的对象
rm(*objectlist*)	移除（删除）一个或多个对象
help(*options*)	显示可用选项的说明
options()	显示或设置当前选项
history(#)	显示最近使用过的#个命令（默认值为 25）
savehistory("*myfile*")	保存命令历史到文件 *myfile* 中（默认值为.Rhistory）
loadhistory("*myfile*")	载入一个命令历史文件（默认值为.Rhistory）
save.image("*myfile*")	保存工作空间到文件 *myfile* 中（默认值为.RData）
save(*objectlist*, file="*myfile*")	保存指定对象到一个文件中
load("*myfile*")	读取一个工作空间到当前会话中（默认值为.RData）
q()	退出 R。将会询问你是否保存工作空间

什么是packages包?

丁香公开课
CLASS.DXY.CN

- 包是R函数、数据、预编译代码以一种定义完善的格式组成的集合。计算机上存储包的目录称为库（library）。函数.libPaths()能够显示库所在的位置， 函数library()则可以显示库中有哪些包。
- R自带了一系列默认包（包括base、datasets、utils、grDevices、graphics、stats以及methods），它们提供了种类繁多的默认函数和数据集。其他包可通过下载来进行安装。安装好以后，它们必须被载入到会话中才能使用。命令search()可以告诉你哪些包已加载并可使用。

packages包的安装

丁香公开课
CLASS.DXY.CN

- 使用命令 install.packages(“gclus”), 下载和安装包
- 使用命令 update.packages(), 更新已经安装的包
- 使用命令 installed.packages(), 列出已安装的包

包的载入

丁香公开课
CLASS.DXY.CN

- 要在R会话中使用包，还需要使用library()命令载入这个包。例如，要使用gclus包，执行命令library(gclus)即可。当然，在载入一个包之前必须已经安装了这个包。在一个会话中，包只需载入一次。

获取帮助

丁香公开课
CLASS.DXY.CN

表 2 常用获得帮助的命令

函　数	功　能
help.start()	打开帮助文档首页
help("*foo*")或?*foo*	查看函数 *foo* 的帮助（引号可以省略）
help.search("*foo*")或??*foo*	以 *foo* 为关键词搜索本地帮助文档
example("*foo*")	函数 *foo* 的使用示例（引号可以省略）
RSiteSearch("*foo*")	以 *foo* 为关键词搜索在线文档和邮件列表存档
apropos("*foo*", mode="function")	列出名称中含有 *foo* 的所有可用函数
data()	列出当前已加载包中所含的所有可用示例数据集
vignette()	列出当前已安装包中所有可用的 vignette 文档
vignette("*foo*")	为主题 *foo* 显示指定的 vignette 文档

R语言编程中的常见错误

丁香公开课
CLASS.DXY.CN

- 使用了错误的大小写。help()、Help() 和 HELP() 是三个不同的函数（只有第一个是正确的）。
- 忘记使用必要的引号。install.packages("gclus") 能够正常执行，然而Install.packages(gclus)将会报错。
- 在函数调用时忘记使用括号。例如，要使用help()而非help。即使函数无需参数，仍需加上()。
- 在Windows上，路径名中使用了\。R将反斜杠视为一个转义字符。setwd("c:\mydata")会报错。正确的写法是setwd("c:/mydata")或setwd("c:\\mydata")。
- 使用了一个尚未载入包中的函数。函数order.clusters()包含在包gclus中。如果还没有载入这个包就使用它，将会报错。

1.3 R 语言中数据集的创建

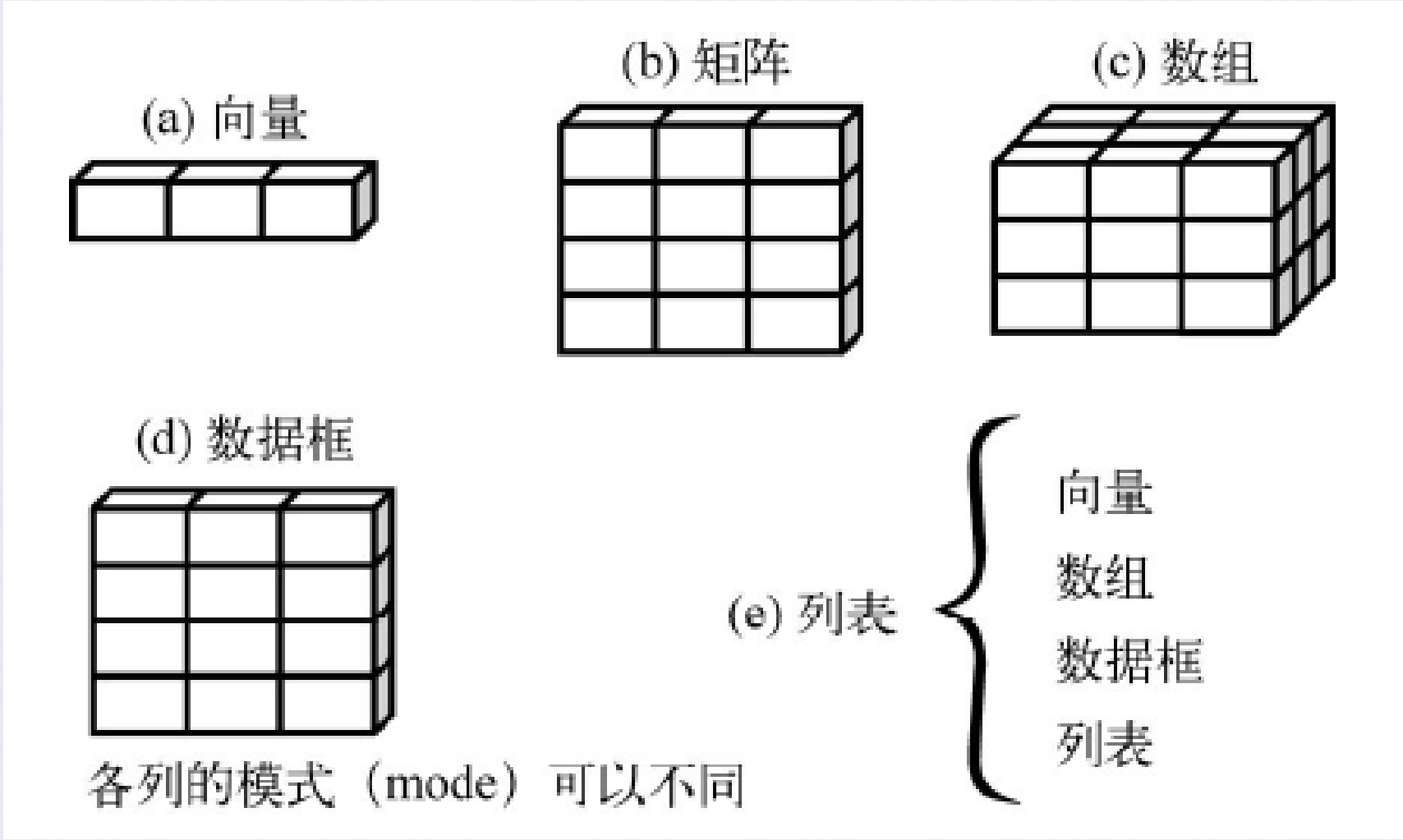
数据集的概念

丁香公开课
CLASS.DXY.CN

表 3 病例数据集

病人编号（PatientID）	入院时间（AdmDate）	年龄（Age）	糖尿病类型（Diabetes）	病情（Status）
1	10/15/2009	25	Type 1	Poor
2	11/01/2009	34	Type 2	Improved
3	10/21/2009	28	Type 1	Excellent
4	10/28/2009	52	Type 1	Poor

数据结构



R语言中的数据结构

标量与向量

丁香公开课
CLASS.DXY.CN

- 标量是只含一个元素的向量，例如：f <- 3、g <- “US”和 h <- TRUE。它们用于保存常量。
- 向量是用于存储数值型、字符型或逻辑型数据的一维数组。执行组合功能的函数c()可用来创建向量。各类向量如下例所示：

```
a <- c(1, 2, 5, 3, 6, -2, 4)
b <- c("one", "two", "three")
c <- c(TRUE, TRUE, TRUE, FALSE, TRUE, FALSE)
```

矩阵

丁香公开课
CLASS.DXY.CN

- 矩阵是一个二维数组，只是每个元素都拥有相同的模式（数值型、字符型或逻辑型），可通过函数 matrix() 创建矩阵，一般使用格式如下：

```
myymatrix <- matrix(vector, nrow=number_of_rows,
ncol=number_of_columns,byrow=logical_value,
dimnames=list(char_vector_rownames, char_vector_colnames))
```

数组

丁香公开课
CLASS.DXY.CN

- 数组（array）与矩阵类似，但是维度可以大于2。数组可通过array函数创建，形式如下：

  myarray <- array(vector, dimensions, dimnames)

- 其中vector包含了数组中的数据，dimensions是一个数值型向量，给出了各个维度下标的最大值，而dimnames是可选的、各维度名称标签的列表。

数据框

丁香公开课
CLASS.DXY.CN

- 由于不同的列可以包含不同模式（数值型、字符型等）的数据，数据框的概念较矩阵来说更为一般。它与通常在SAS、SPSS和Stata中看到的数据集类似。数据框将是R中最常处理的数据结构。数据框可通过函数 data.frame() 创建：

  mydata <- data.frame(col1, col2, col3,...)
- 其中的列向量col1、col2、col3等可为任何类型（如字符型、数值型或逻辑型）。每一列的名称可由函数names指定。

因子

丁香公开课
CLASS.DXY.CN

- 类别（名义型）变量和有序类别（有序型）变量在R中称为因子（factor）。因子在R中非常重要，因为它决定了数据的分析方式以及如何进行视觉呈现。函数factor()以一个整数向量的形式存储类别值，整数的取值范围是[1...k]（其中k是名义型变量中唯一值的个数），同时一个由字符串（原始值）组成的内部向量将映射到这些整数上。

列表

丁香公开课
CLASS.DXY.CN

- 列表（list）是R的数据类型中最为复杂的一种。一般来说，列表就是一些对象（或成分，component）的有序集合。列表允许你整合若干（可能无关的）对象到单个对象名下。
- 例如，某个列表中可能是若干向量、矩阵、数据框，甚至其他列表的组合。可以使用函数list()创建列表：

  mylist <- list(object1, object2, ...)
- 其中的对象可以是目前为止讲到的任何结构。你还可以为列表中的对象命名：

  mylist <- list(name1=object1, name2=object2, ...)

编程注意事项

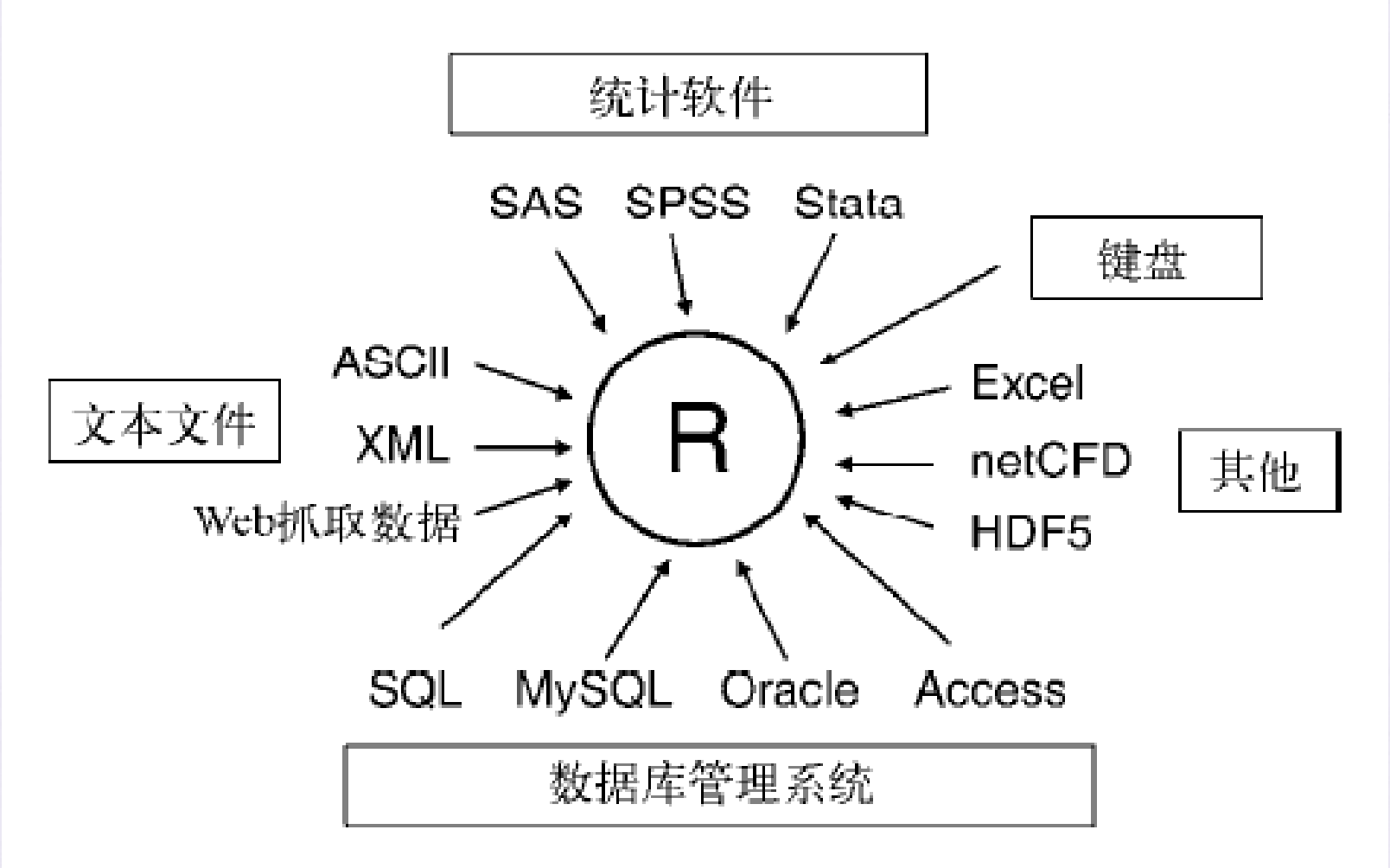
丁香公开课
CLASS.DXY.CN

❑对象名称中的句点（.）没有特殊意义，但美元符号（$）却有着和其他语言中的句点类似的含义，即指定一个数据框或列表中的某些部分。例如，A$x是指数据框A中的变量x。
❑R不提供多行注释或块注释功能。你必须以#作为多行注释每行的开始。出于调试目的，你也可以把想让解释器忽略的代码放到语句if(FALSE){... }中。将FALSE改为TRUE即允许这块代码执行。
❑将一个值赋给某个向量、矩阵、数组或列表中一个不存在的元素时，R将自动扩展这个数据结构以容纳新值。举例来说，考虑以下代码：

```
> x <- c(8, 6, 4)
> x[7] <- 10
> x
```

通过赋值，向量x由三个元素扩展到了七个元素。x <- x[1:3]会重新将其缩减回三个元素。
❑R中没有标量。标量以单元素向量的形式出现。
❑R中的下标不从0开始，而从1开始。在上述向量中，x[1]的值为8。
❑变量无法被声明。它们在首次被赋值时生成。

可供R导入的数据源


统计软件
SAS
SPSS
Stata
键盘
ASCII
XML
Web抓取数据
文本文件
R
Excel
netCFD
HDF5
其他
SQL
MySQL
Oracle
Access
数据库管理系统

1.4 基本数据管理方法

举例

丁香公开课
CLASS.DXY.CN

表4-1　领导行为的性别差异

经理人	日　期	国　籍	性　别	年　龄	q1	q2	q3	q4	q5
1	10/24/14	US	M	32	5	4	5	5	5
2	10/28/14	US	F	45	3	5	2	5	5
3	10/01/14	UK	F	25	3	5	5	5	2
4	10/12/14	UK	M	39	3	3	4		
5	05/01/14	UK	F	99	2	2	1	2	1

创建新变量

丁香公开课
CLASS.DXY.CN

- 变量名 <- 表达式

表4-2　算术运算符

运　算　符	描　　述
+	加
–	减
*	乘
/	除
^或**	求幂
x%%y	求余（x mod y）。5%%2 的结果为 1
x%/%y	整数除法。5%/%2 的结果为 2

变量的重编码

丁香公开课
CLASS.DXY.CN

- 重编码涉及根据同一个变量和/或其他变量的现有值创建新值的过程

表4-3　逻辑运算符

运　算　符	描　　述
<	小于
<=	小于或等于
>	大于
>=	大于或等于
==	严格等于①
!=	不等于
!x	非x
x | y	x或y
x & y	x和y
isTRUE(x)	测试x是否为TRUE

变量的重命名

丁香公开课
CLASS.DXY.CN

- fix() 函数
- names()函数
- plyr包rename()函数

缺失值

丁香公开课
CLASS.DXY.CN

- 函数is.na()允许你检测缺失值是否存在
- 重编码某些值为缺失值
- 在分析中排除缺失值

日期值

丁香公开课
CLASS.DXY.CN

表4-4　日期格式

符　号	含　义	示　例
%d	数字表示的日期（0~31）	01~31
%a	缩写的星期名	Mon
%A	非缩写星期名	Monday
%m	月份（00~12）	00~12
%b	缩写的月份	Jan
%B	非缩写月份	January
%y	两位数的年份	07
%Y	四位数的年份	2007

类型转换

丁香公开课
CLASS.DXY.CN

表4-5　类型转换函数

判　断	转　换
is.numeric()	as.numeric()
is.character()	as.character()
is.vector()	as.vector()
is.matrix()	as.matrix()
is.data.frame()	as.data.frame()
is.factor()	as.factor()
is.logical()	as.logical()

数据集的合并

丁香公开课
CLASS.DXY.CN

➢ 向数据框添加列

要横向合并两个数据框（数据集），请使用merge()函数。在多数情况下，两个数据框是通过一个或多个共有变量进行联结的（即一种内联结，inner join）。例如：

```
total <- merge(dataframeA, dataframeB, by="ID")
```

➢ 向数据框添加行

要纵向合并两个数据框（数据集），请使用rbind()函数：

```
total <- rbind(dataframeA, dataframeB)
```

数据集取子集

丁香公开课
CLASS.DXY.CN

- 从一个大数据集中选择有限数量的变量来创建一个新的数据集是常有的事。通过dataframe[row indices, column indices]这样的记号来访问选择变量。例如：

  newdata <- leadership[, c(6:10)]

主要参考文献

【1】Robert I. Kabacoff 著, 《R语言实战 》(第2版), 人民邮电出版社, 2016

【2】Peter Dalgaard 著, 《R语言统计入门》》(第2版), 人民邮电出版社, 2014

【3】薛毅 陈立萍 著,《R语言实用教程》, 清华大学出版社, 2014

【4】张铁军 陈兴栋 刘振球 著,《R语言与医学统计图形》, 人民卫生出版社, 2018

丁香公开课
CLASS.DXY.CN

Thanks!

感谢您的观看！