R 语言基础春令营（2）：基本统计分析和绘图

陈堰平

统计之都
http://cos.name

April 14, 2013

大纲

Section 1

读写数据

键盘输入

- scan() 读入数值
- readline() 输入单行数据
- edit()

显示到屏幕

- `print()`
- `cat()`

读数据

- `scan()`
- `read.table()`
- `read.csv()`

read.table() 和 scan()，可以用网站地址（URL）作为参数

从 Excel 中读取

```
# 第一行包含变量名
# 名为 mysheet 的 sheet 有数据

library(RODBC)
channel <- odbcConnectExcel("c:/myexel.xls")
mydata <- sqlFetch(channel, "mysheet")
odbcClose(channel)
```

从 SPSS 文件中读数据

```
# 从 SPSS 中把数据导成 trasport 格式
# 用 SPSS 脚本写就是
get file='c:\mydata.sav'.
export outfile='c:\mydata.por'.

# 在 R 里 如下操作
library(Hmisc)
mydata <- spss.get("c:/mydata.por",
                   use.value.labels=TRUE)
# 最后一条选项把数据标签转化成 R 的水平
```

从 SAS 文件中读数据

```
# 把SAS数据库转成trasport格式
# 用SAS语句写就是
libname out xport 'c:/mydata.xpt';
data out.mydata;
set sasuser.mydata;
run;

# 在R里 如下操作
library(Hmisc)
mydata <- sasxport.get("c:/mydata.xpt")
# 字符型变量会转化成水平
```

从 Stata 文件中读数据

```
library(foreign)
mydata <- read.dta("c:/mydata.dta")
```

从数据库读取数据

RODBC包

函数	描述
odbcConnect(dsn, uid="", pwd="")	打开 ODBC 数据库的链接
sqlFetch(channel, sqtable)	从数据库中读一张表，转成数据框
sqlQuery(channel, query)	提交一条 SQL 查询语句，返回结果
sqlSave(channel, mydf, tablename = sqtable, append = FALSE)	把数据框写入到数据库的表中
sqlDrop(channel, sqtable)	从数据库中删除一张表
close(channel)	关闭链接

导出数据

- `foreign`包: SPSS、SAS、Stata
- 要转成 Excel 格式
  - `xlsReadWrite`包经常出问题
  - `WriteXLS`包基于 perl 的 Spreadsheet::WriteExcel 包写的
  - `dataframes2xls`包基于 python 写的
  - `xlsx`包基于 java 写的

获取数据集信息

- `ls()`
- `names()`
- `str()`
- `levels()`
- `dim()`
- `class()`
- `head(mydata, n=10)`
- `tail(mydata, n=5)`

缺失值

- `is.na()` 检测是否为缺失
- 用索引操作来重编码
- 在计算中对 NA 的剔除
  - na.rm 选项
  - complete.cases()
  - na.omit()

在高级课程中将介绍数据插补方法

Section 2

数据管理

变量操作

- 创建新变量
- 数据编码
- 给变量重命名

内置数学函数

- exp( ): 以自然常数 e 为底的指数函数
- log( ): 自然对数
- log10( ): 以 10 为底的常用对数
- sqrt( ): 平方根
- abs( ): 绝对值
- sin( ), cos( ) 等: 三角函数
- min( ), max( ): 向量的最小、最大值
- which.min( ), which.max( ): 向量的最小、最大元素的位置索引
- pmin( ), pmax( ): 把多个等长度的向量按元素逐个对比，返回所有向量的第 k 个元素中最小 (最大) 的值。
- sum( ), prod( ): 把一个向量的所有元素求和 (求积)。
- cumsum( ), cumprod( ): 把一个向量的前 k 个元素累计求和 (求积)。
- round( ), floor( ), ceiling( ): 分别是四舍五入去整、向下去整和向上去整

apply 系列函数

- `apply()`
- `tapply()`
- `sapply()`
- `lapply()`

数据合并

数据汇总

比如有个测试数据集如下：

group	value
a	10
a	20
a	30
b	100
b	200

想对不同 group 的 value 求和（或均值）

Section 3

基本统计分析

描述统计

```
sapply(mydata, mean, na.rm=TRUE)
summary(mydata)
fivenum(x)

library(Hmisc)
describe(mydata)

library(epicalc)
des(mydata)
```

频数表

table()函数

随机抽样

sample()函数

回归

```
lm(formula, data, subset)
```

举例

```
x <- 1:20
y <- x + rnorm(20, 0, 0.1)
lm.y <- lm(y ~ x)
lm.y
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##      0.0629       0.9933
```

```
names(lm.y)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

提取模型信息

函数	意义
summary()	拟合模型的摘要
coef()	模型系数
resid()	残差
fitted()	拟合值
confint()	模型参数的置信区间
deviance()	残差平方和
anova()	方差分析表
predict()	预测
plot()	回归诊断图
influence()	回归诊断

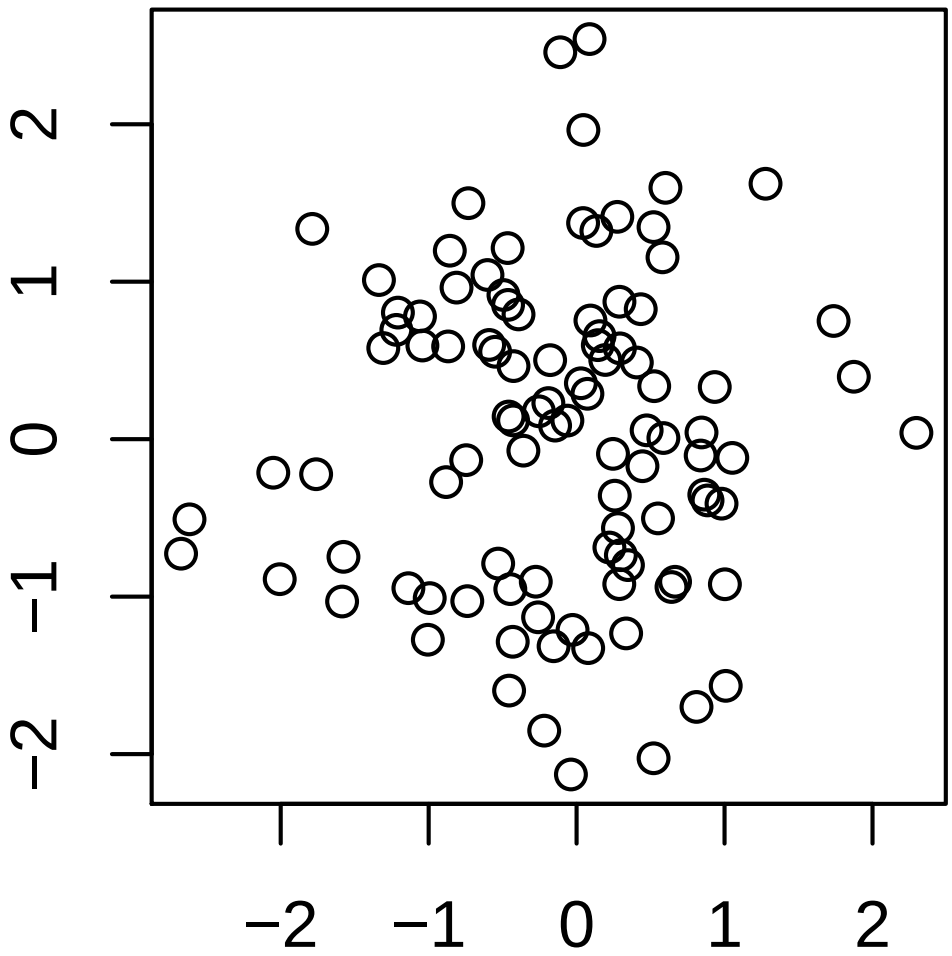
Section 4

基本绘图

散点图

```
x = rnorm(100)
y = rnorm(100)
plot(x, y)
```

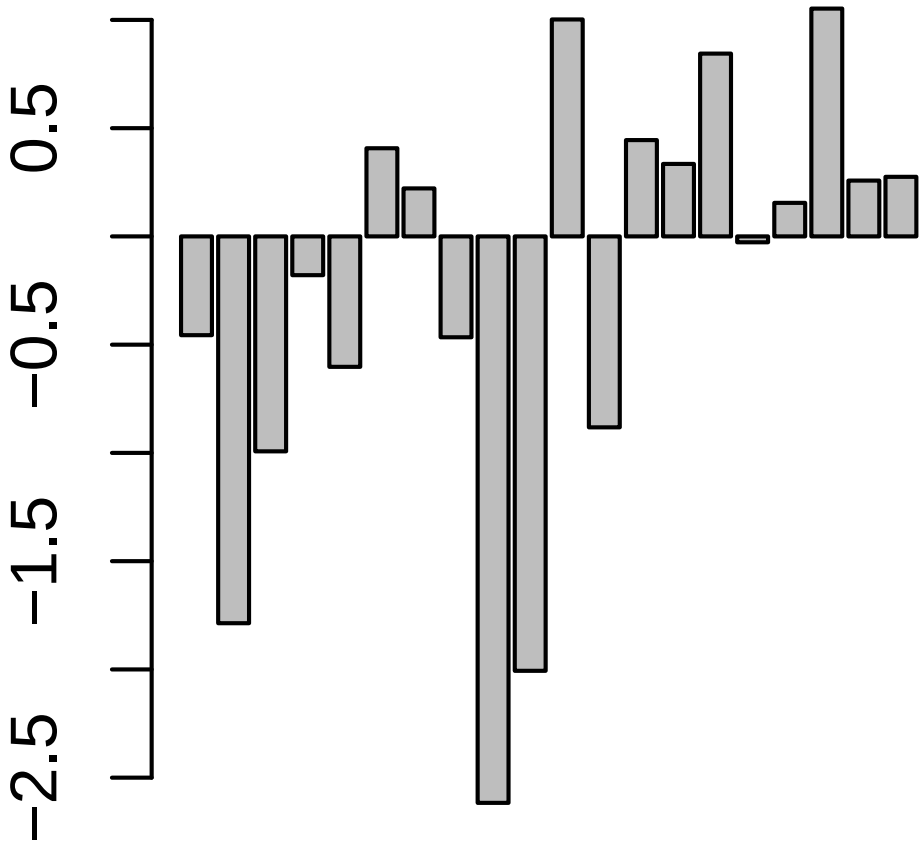
散点图



柱状图

```
barplot(x[1:20])
```

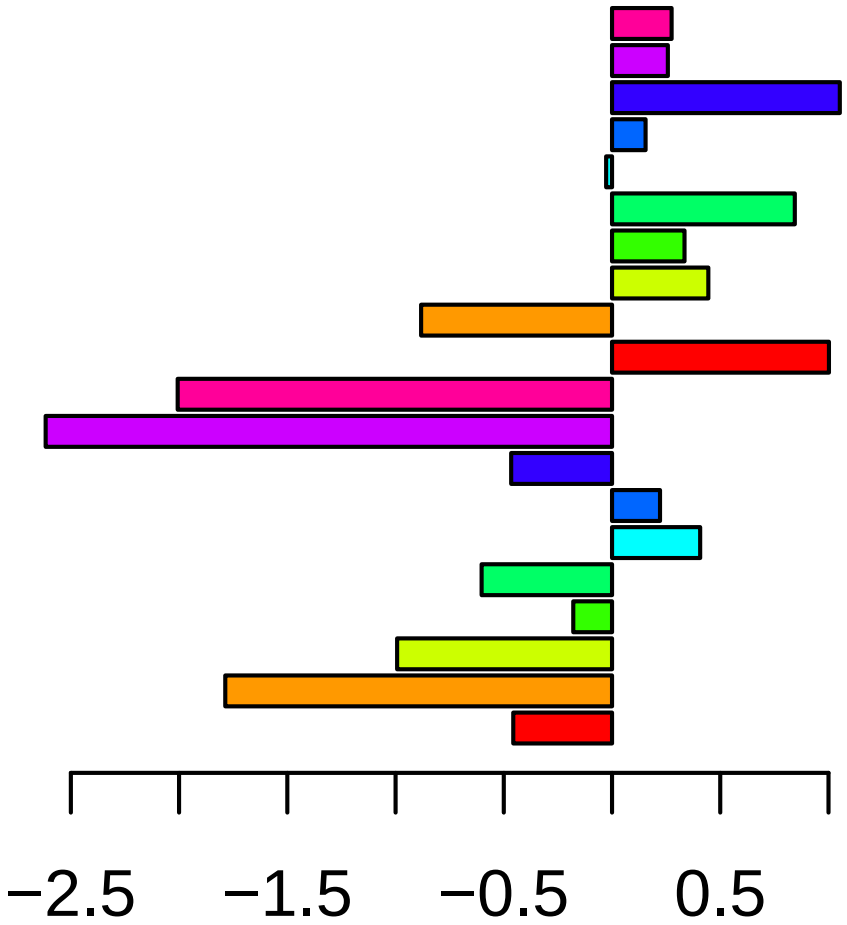
柱状图



柱状图

```
barplot(x[1:20], width=2, horiz=T,
        col=rainbow(10))
```

柱状图



饼图

```
pie(c(10,10,10,20,30,20),
  c("Nature","Science","Cell","NG",
  "Nature Cancer","Other"),col=2:7)
```

饼图