

# 量化投资 Using R

## 第二章了解 R 语言特性

罗智超 *Rokia.org*

### Contents

内容简介	2
R 语言自带数据	2
Anscombe 数据	2
查看数据集	2
查看数据集	3
数据集的基本元素	3
什么是 tidy 数据格式	3
tidy data Rule	3
扩展案例	3
课后练习	3
本周“大牛”	4

## 内容简介

- 通过案例了解 R 语言的基本特点
- 了解数据集的结构
- 通过案例了解简单回归代码

## R 语言自带数据

- R 里面自带了很多数据集，这样方便研究人员验证算法
- 通过 `data()` 可以查阅所有数据集名称
- 通过 `data(package="packagename")` 来查阅 R 包里面自带的数据集名称

```
data()
#sample data faithful,Titanic,iris,mtcars,flight
data(iris)
iris

f <- faithful

View(f)
str(f)
hist(f$eruptions,breaks = 20)
hist(f$waiting,breaks=20)

head(iris)
tail(iris)
```

## Anscombe 数据

- 1973 年，统计学家 F.J. Anscombe 在 *Graphs in Statistical Analysis* 构造出了四组奇特的数据。它告诉人们，在分析数据之前，描绘数据所对应的图像有多么的重要。
- 本章通过对 R 自带的的 Anscombe 数据进行简单的处理，让同学了解 R 语言的基本特征

## 查看数据集

```
anscombe
# 练习：计算所有变量的均值方差，并绘制每个变量的频数分布图、盒形图、以及每组变量的散点图 (x1*y1,x2*y2)
str(anscombe) #x1-x4,y1-y4

m<-apply(anscombe,2,mean)
v<-apply(anscombe,1,var)

hist(anscombe$x1)

plot(anscombe$x1,anscombe$y1)
plot(anscombe$x2,anscombe$y2)
plot(anscombe$x3,anscombe$y3)
plot(anscombe$x4,anscombe$y4)
# 思考，如何转变数据结构使得可以使用下面伪代码方法
plot(x,y,by=group)
```

```
?anscombe
# 思考如何把四张图画在一张图里面

View(anscombe)
```

## 查看数据集

- 看看数据集的结构有什么特征
- 计算下各个变量的统计值
- 基本绘图

## 数据集的基本元素

- 变量 ( variable name )
- 记录 ( column name )
- 变量类型 ( 数值、字符、因子 )

## 什么是 tidy 数据格式

- Hadley 发表在 Journal of statistical software 上的文章Tidy Data
- 该文章的源码地址
- 宽数据 VS 窄数据

## tidy data Rule

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table.

## 扩展案例

- 考试成绩的回归分析

## 课后练习

```
#Forbes2013 top 2000 company
#Statistics by Wuxizhi P13
w<-read.csv("data/Forbes2000.csv")
names(w)
summary(w)
View(w)
w[w[,3]=="China",2]
par(mfrow=c(2,2))
for(i in 4:7)
{
```

```

hist(log(w[,i]),main=paste("Log",names(w)[i]),xlab="")
rug(log(w[,i]))
}
C<-w[w[,3]=="China",]
G<-w[w[,3]=="Germany",]
par(mfrow=c(1,2))
hist(C$Market.Value,20,main="Histogram of Market Value (China)",col=3,prob=T,ylim=c(0,0.07))
lines(density(C$Market.Value),lwd=2)
hist(G$Market.Value,20,,main="Histogram of Market Value (German)",col=2,prob=T,ylim=c(0,0.07))
lines(density(G$Market.Value),lwd=2)

```

## 本周“大牛”

- 弗朗西斯·高尔顿 (Francis Galton, 1822 年 2 月 16 日 - 1911 年 1 月 17 日)，英国科学家和探险家。他曾到西南非洲探险，因树立功绩而知名并被选为英国皇家地理学会会员，三年后又入选英国皇家学会，晚年受封为爵士。他的学术研究兴趣广泛，包括人类学、地理、数学、力学、气象学、心理学、统计学等方面。他是查尔斯·达尔文的表弟，深受其进化论思想的影响，把该思想引入到人类研究。他着重研究个别差异，从遗传的角度研究个别差异形成的原因，开创了优生学。他关于人类官能的研究开辟了个体心理和心理测验研究的新途径。