前　言

1982年美国大学生数学建模竞赛开始，1992年我国全国大学生数学建模竞赛也开赛，之后数学模型、数学建模、数学实验课程应运而生，且越来越普及；国家统计局于2008年开展面向全国政府统计系统的职业技能竞赛活动——统计建模大赛。2009年全国大学生统计建模竞赛也开赛了，2011年国务院学位办将统计学提升为一级学科。这不仅仅是统计学的重要性和必要性所致，更是学科划分的科学性使然。理所当然统计模型、统计建模、统计实验课程也应该设置起来。

目前国内在统计专业开设统计模型课程和统计实验课程的很少，其中统计模型的部分内容在数学模型课程中有所涉及，统计实验的部分内容在数学实验课程中有所涉及，主要是数学软件应用。随着计算机的普及和统计软件的大力发展，SAS、SPSS、MINITAB、JMP、R、DPS、马克威等统计软件被广泛应用。2009年，中国统计教育学会、中国现场统计研究会、中国数学会概率统计学会联合举办首届全国大学生统计建模大赛，共有来自全国57所大学的334支队伍参赛，其中包括北京大学、中国人民大学、南开大学等全国知名高校，之后的5年，我们在各类统计课程、课外科技活动中融入统计建模的思想和方法，2013年第三届全国大学生统计建模大赛共有来自全国103所大学的508支队伍参赛；以多年从事统计专业教学实践来看，为使学生具备良好的应用统计的能力，我们仍感到有开设一门相关课程的必要，为此南京财经大学在新的《统计学专业培养方案（2010级）》中，设立了《统计模型与统计实验》课程，并将它列为新的专业主干课之一（在每届二年级下学期开设），同时它也是一门创新课程。我校以前从未开过，国内极少学校有所涉及，可供参考模式少，因此这是一项具有挑战性的统计课程体系的教学改革，也必将促进其他许多相关课程的改革，特别是统计学专业课程的内容与体系结构改革和调整。由此看来，建设面向21世纪的新型课程《统计模型与统计实验》势在必行。

主讲、主编人王庚1994—2007年从事数学模型教学与应用，做大学生数学

包含很多 R 的入门资料、实用函数列表和更广泛的应用案例等。编程能力的高低取决于两个方面，一是对程序本身的认识；另一个是算法设计的技巧。对于前者，如果读者想更方便地了解 R 语言全貌，建议遵循如下步骤：(1)了解这个软件长什么样子；(2)找几个简单案例，知道这些程序大概长什么样子，大概是怎么运行的；(3)了解 R 语言的各种数据类型及其注意事项；(4)把一些常用的、实用的函数整理成列表，打印出来放在手边，作为参考手册之一；(5)有机会的话多处理一些实际问题。

Ross Ihaka

Robert Gentleman

图 1.1.14　R 开发者

R 的下载和安装

1. R 的下载

(1)进入网站：http://www.r-project.org，点击其中的“download R”；

(2)如果是第一次进入 R 官网，并按上述点击，则网站会提示您选择一个下载镜像(Mirrors)。如果在国内的话，选择“China”下面的某个镜像就可以了；

(3)点击 Download R for windows，并按照提示进行后续操作，即可下载得到在 windows 操作系统上运行的 R 语言软件。如果需要其他操作系统的 R 软件，如 MacOS X 或 Linux 等，读者可以选择相应的链接。

2. R 的安装

这里以 R-2.14.0-win.exe 为例讲解 R 安装的过程。下载得到 R-2.14.0-win.exe 以后，双击它。在下图 1.1.15 中选择需要的语言，这里使用“中文(简体)”，点击“确定”，根据提示完成安装。

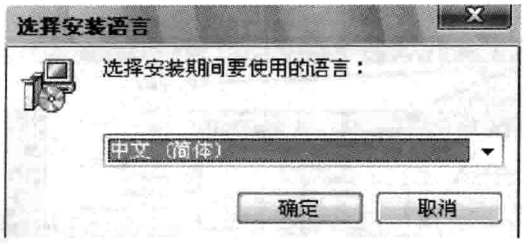


图 1.1.15　R 的安装

R 的进入与界面认识

1. 双击桌面上的 R 图标(图 1.1.16),进入 R 软件界面。

图 1.1.16　R 开始图标

2. 认识 R 界面

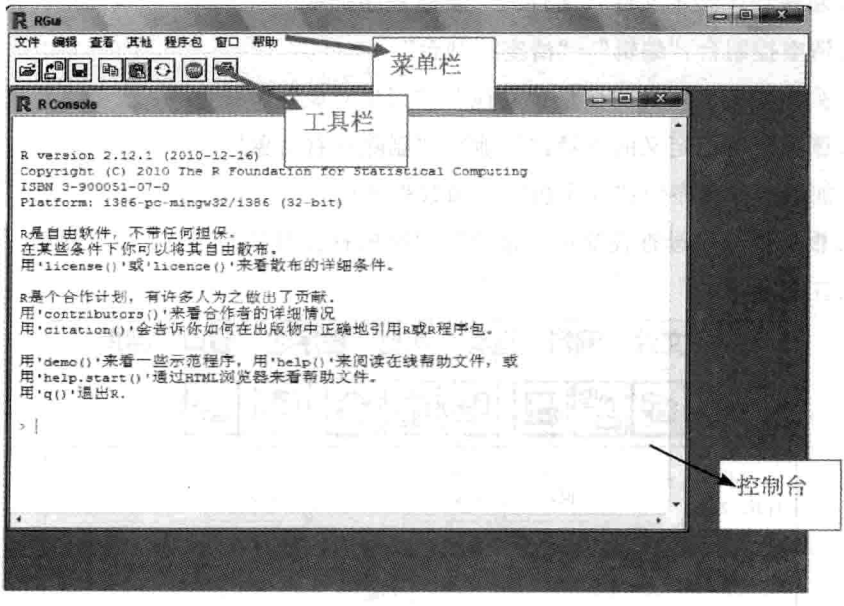


图 1.1.17　R 界面认识

特色，不过目前大家用得比较多的应算 Rstudio 软件[①]，这也是本书推荐的辅助工具。

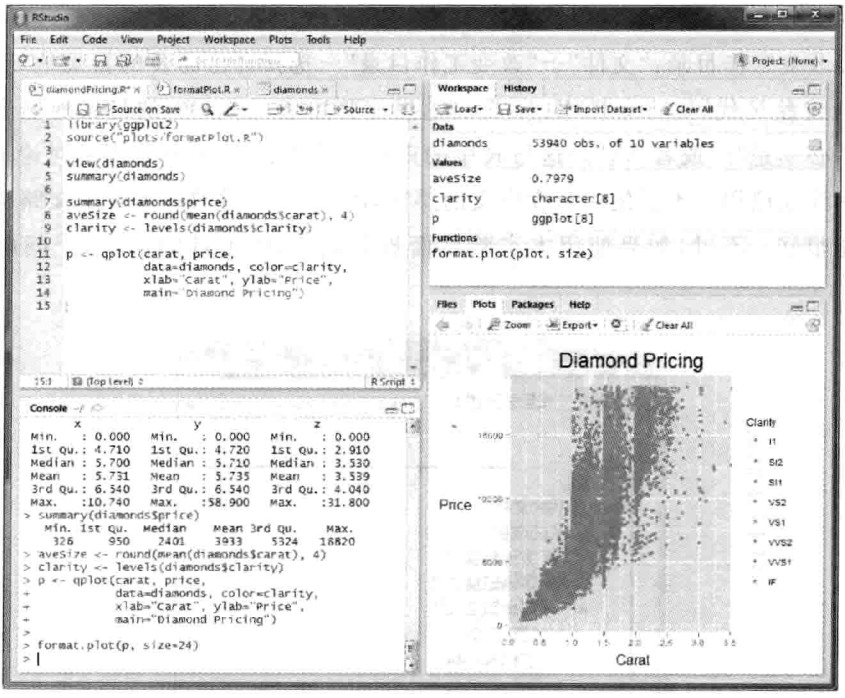


图 1.1.21　Rstudio 软件界面

按照其官网的介绍，Rstudio 是辅助用户进行 R 语言编程的一个集成开发环境(IDE，Integrated Development Environment)。与 R 软件类似，它是开源的、免费的，针对 Windows、Mac 或 Linux 不同操作系统有不同的版本。它可以直接在其官网上(http://www.rstudio.com/)下载。下载完毕并安装以后，它会自动识别计算机中已经装好的 R 软件，并自动与最新版本的 R 软件进行连接。当读者在 Rstudio 中输入命令，并点击 RUN 按钮，或者按 Ctrl＋R 或按 Ctrl＋Enter 以后，读者的命令就可以运行得到结果了。图 1.1.21 是 Rstudio 的一个界面。

在使用 Rstudio 过程中需要注意几个问题：(1)windows 系统的账户名称最好不要包含中文，否则软件自带的图形(Plot)显示区域可能不能正常工作。(2)windows 系统下可以手动将“.R”文件关联到 Studio。操作方法是，右击任意“.R”文

① 请读者注意，不是 R-studio，而是 Rstudio。这两个软件的用途是截然不同的。

件，选择“打开方式”，选择“默认程序”，找到 Rstudio，确定即可。使用 Rtudio 直接打开 R 程序文件以后，工作路径会自动指向该文件所在文件夹，无须再设置工作路径了。

1.3　统计科学基础

R 语言案例

下面借用几个例子简单熟悉一下统计软件的编程过程；这些例子都是采用最基本的命令。由于读者可能对编程不是很熟悉，因此这里仔细讲解了每个程序的设计思路。

【例 1.2】 计算常用统计量

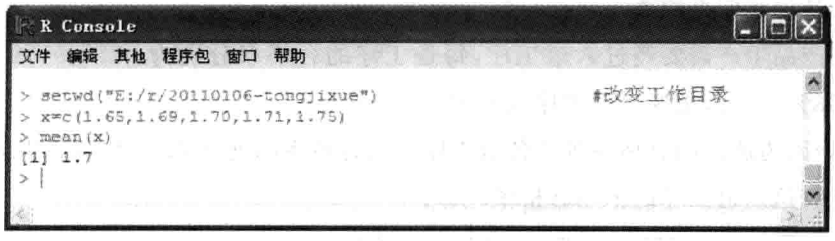
1. 计算平均身高

(1)计算平均身高主要运用了 R 语言中求平均数的函数 mean(x)。其中的 x 是包含一系列数值的向量，在 R 中可根据 c()函数生成向量。关于 R 语言的数据结构和其他常用统计函数的介绍，读者可以阅读本书的附录部分，或者访问本书对应的网站。

(2)新建一个脚本程序窗口(“问卷”—“新建脚本程序”)，在打开的窗口中输入以下源程序，就可以计算得到他们的平均身高。

```
x=c(1.65,1.69,1.70,1.71,1.75)
mean(x)
```

(3)全选以上程序，按快捷键 Ctrl+R 运行以上程序，得到如下结果，均值为 1.7。



2. 计算调和平均数

假如某种蔬菜在早、中、晚市的每市斤的单价分别为 0.5 元、0.4 元、0.2 元，若

rep()函数的标准结构为：

rep(x,times/each)

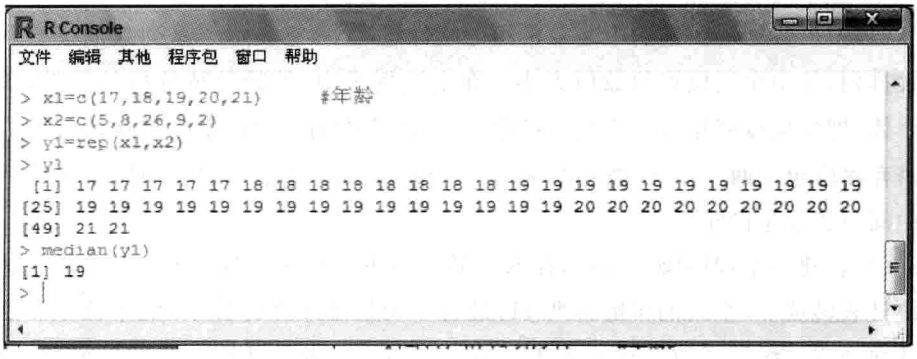
其中 x 表示一个数据向量，each 和 times 参数只能有一个。each 的值只能是一个非负数值，表示对 x 中每个元素重复 each 次以后，再把重复以后的片段连接起来。times 参数稍复杂点，如果 times 的值和 x 包含的元素个数相同，则表示对 x 的每个元素重复相应次数以后，再把片段连接起来；如果 times 的值只包含一个数值，则表示把 x 这个向量整体重复 times 次以后，再把每个片段连接起来。通常默认是 times 参数。

具体可以看下面的一个例子。

```
x1=c(17,18,19,20,21)    #年龄
x2=c(5,8,26,9,2)        #学生人数
y1=rep(x1,x2)
y1
median(y1)              #直接用函数 median

#关于 rep()的一个例子
x=c(1,2,3)
rep(x,times=c(3,2))     #times 长度和 x 不同,会出错哦
rep(x, each=3)
rep(x,each=c(3,2,1))    #只识别第一个元素 3
rep(x,times=3)          #把整体重复 3 遍
rep(x,times=c(3,2,1))   #对 x 的每个元素重复不同次数
```

(3)全选以上程序，按快捷键 Ctrl+R 运行以上程序，得到如下结果。



b. 次数统计。对于一个数据向量，首先对其进行分组，再对每个组的个数进行统计。cut(x,breaks)表示按照 breaks 的划分标准，对 x 向量进行分组，并标记 x 中每个元素的组别。table(x)统计向量 x 中相同元素出现的次数。

c. 绘制直方图。采用 hist()函数可以直接对一个数据向量进行分组统计，并绘图。所以这一步可以不依赖于 b 步骤。如果想要直方图画的漂亮点，可以对其中的多个参数进行设定。hist( )的常用参数包括 hist(x, breaks, freq, right, density, col, angle, xaxt, yaxt, xlab, ylab, main,…)

其中：x 表示数据向量。

breaks 表示分割点的位置向量。

freq 是逻辑参数，如果为真(T)，表示直方图的纵坐标是概率密度；如果为假(F)或空(NULL)，表示直方图的纵坐标为个数统计。

right 是一个逻辑参数，真(right＝T)表示每个分组区间是左开右闭的；假(right＝F)表示每个分组区间是左闭右开的。这里采用后者的设定方法。

density 默认情况下为空(NULL)，表示对直方图每个条柱的填充是纯色的。当 density 参数为非负数值时，直方图的每个条柱由直线纹理构成。density 数值的大小表示一英寸宽度内的直线条数，可控制纹理的密度。

angel 表示以上纹理的倾斜度数，默认是 45 度。

col 是颜色参数，表示每个直方条柱的填充颜色。注意的是，如果 density 不是 NULL 的时候，表示其中纹理线条的颜色。读者可以把下述程序中的 density＝50 去掉以后看看是什么效果。

xaxt 表示是否绘制横坐标的刻度，xaxt＝"n"表示不绘制横坐标刻度；yaxt 表示是否绘制纵坐标的刻度，与 xaxt 类似。

xlab 表示横坐标的标签；ylab 表示纵坐标的标签。

main 表示图形的主标题。

其他参数的介绍可以查看 hist 的帮助，即在命令窗口中输入？hist 即可。由于默认情况下的直方图横坐标刻度比较松散，在这里的绘图中首先不绘制横坐标的刻度，然后采用 axis(side,at,label)添加刻度。

axis(side,at,labels)

其中，side 的取值为 1 到 4 的数值之一，分别表示在下、左、上、右添加相应的刻度。

at 是一个数据向量，表示在哪几个位置上添加刻度。

labels 是与 at 相同长度的向量，表示在以上位置分别添加什么刻度，可以是字

```
#本程序用于实现“排列图”
#输入数据
x=c(104,42,20,10,6,4,14)
names(x)=c("变形","刮花","针眼","裂缝","斑点","有沟","其他")
x                                          #看看 x 是什么样子的
x3=sort(x,decreasing=T)                    #排序
#绘制条件图
barplot(x3,col=rgb(0/255,1,77/255),las=2,angle=c(45,135),density=50,
    main="排列图(帕累托图)")               #条形图命令
#运算过程
累计比例刻度位置=c(0, max(x3)/10*(1:10))
                                           #累计比例曲线纵坐标刻度的位置
累计比例刻度值=0:10/10                     #累积比例曲线纵坐标刻度
累计比例=cumsum(x3/sum(x3))
累计比例纵坐标=累计比例*max(累计比例刻度位置)
累计比例横坐标=1:length(累计比例纵坐标)+0.2*0:(length(累计比例纵坐标)-
1)-0.25
##添加图线
#辅助线
for(i in 1:11){lines(c(-100,100), c(累计比例刻度位置[i],累计比例刻度位置[i]),
      lty=3,col=3,lwd=1)}
#累积比例为 80%的线
lines(c(-100,100),c(累计比例刻度位置[9],累计比例刻度位置[9]),
    lty=3,lwd=3,col=3)
#累计比例曲线
lines(x4a,x4,type="o",lwd=2,col=4,pch=15)
                                           #添加累积比例曲线
#右侧刻度
axis(4,at=累计比例刻度位置,labels=累计比例刻度值,las=1)
```

(因子类型),在无特别要求情况下建议读者尽量避免使用它。不过,R 默认是将字符转成因子类型的,所以在每次程序的开头加上 options(stringsAsFactors=F)就可以避免字符类型向因子类型的转换。

```
setwd("E:/r/20120203-book/programs_in_book")
dat=read.csv("02.1.csv",header=T)
head(dat)                    #查看前六行
tail(dat)                    #查看后六行
#考核结果是分类数据,这里按照指定顺利重新编码
x1=factor(dat$X2002考核, levels=c("不合格","合格","中","良","优"),
          labels=c("不合格","合格","中","良","优"))
x2=factor(dat$X2003考核, levels=c("不合格","合格","中","良","优"),
          labels=c("不合格","合格","中","良","优"))
x3=factor(dat$X2004考核, levels=c("不合格","合格","中","良","优"),
          labels=c("不合格","合格","中","良","优"))
#进行统计,并显示结果
( y1=table(x1) )
( y2=table(x2) )
( y3=table(x3) )
# 将三个图画在一个图形中
ol=par(mfrow=c(1,3))      #将图形区域切成1行3列,可以包含三个图
barplot(y1,col=rgb(178/255,204/255,51/255),las=2,angle=c(45,135),density
        =50,
main="2002年职工考核情况")
barplot(y2,col=rgb(178/255,204/255,51/255),las=2,angle=c(45,135),density
        =50,
main="2003年职工考核情况")
barplot(y3,col=rgb(178/255,204/255,51/255),las=2,angle=c(45,135),density
        =50,
main="2004年职工考核情况")
par(ol) # 恢复默认设置,以免干扰后续程序
```

实验练习

1. 安装 Minitab15.0 中文版，进入系统，学会利用帮助运行基本模式。

2. 安装 R3.0.1 版，进入系统，熟悉命令交互式和输入程序运行基本模式。学会利用帮助。逐句运行例 1.3～例 1.6，并改写其中 1～2 个程序。

参考文献

[1] 汤银才：《R 语言与统计分析》，北京，高等教育出版社，2008 年.

[2] 洪楠等：《Minitab 统计分析教程》，北京，电子工业出版社，2007 年.

[3] [美]R. I. Kabacaff，高涛等译：《R 语言实践》，北京，人民邮电出版社，2012 年.

[4] 薛毅、陈立萍：《统计建模与 R 软件》，北京，清华大学出版社，2007 年.

续表

分布名称	R 中的表达	参数 1	参数 2	参数 3
t 分布	t	df	ncp	
均匀分布	unif	min	max	
Weibull 分布	weibull	shape	scale	
Wilcoxon 分布	wilcox	m	n	

每种分布有四种表示形式：d－、q－、p－和 r－，分别表示概率密度，累计概率密度的反函数，累计概率密度，生成随机数等。以正态分布为例，它的四种形式为：

```
dnorm(x, mean = 0, sd = 1, log = FALSE)   #计算 x 值对应的概率密度
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
                                          #计算 q 对应的累计概率密度值
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
                                          #计算累计概率密度为 p 的位置
rnorm(n, mean = 0, sd = 1)                #生成 n 个正态分布随机数
```

2. 部分统计函数

```
+,－,×,÷,?,%%,%/%
< > <= >= ==..! =..
sin,cos,tan,asin,acos,atan,atan2,log,log10,exp
max(x)   最大值
min(x)   最小值
range(x)   返回 x 中的最大值和最小值组成的向量
sum(x)   求和
diff(x)   进行差分运算，后面减前面
prod(x)   x 中元素全部相乘
mean(x)   均值
abs(x)   绝对值
sqrt(x)   开根号
median(x)   中位数
quantile(x,probs=)   分位数，默认返回最小值、第一分位数、第三分位数、最
```

choose(n,k)　组合数=n! /[(n−k)! k!]

sign　大于 0 则返回 1,小于 0 返回−1,等于 0 返回 0

duplicated(x)　返回向量中重复元素的逻辑值

4. 部分高级数据处理函数

apply(X,INDEX,FUN)　x 为矩阵或者数据框。index=1 时,按行调用 FUN 函数;Index=2 时,按列调用 FUN 函数。

lapply、vapply、sapply、tapply 等均类似,可以提高运算速度。

5. 几个高级绘图函数

plot(x)或 plot(x,y)　最基本的绘图函数

hist(x)　直方图

barplot(x) 条形图

boxplot(x)　箱线图

6. 几个低水平绘图函数

points(x,y)　加点

lines(x,y)　加线

segments()　线段

arrows()　箭头

rect()　矩形

polygon()　多边形

legend()　添加标签

text(x,y,labels,...)　添加文字

7. 补充说明

(1)具体函数的使用,可以参照 R 的帮助。运行"? 函数名称"可以打开帮助文档。

(2)写出函数的前面几个字母,按 tab 键会自动匹配能够匹配的函数。这在 Rstudio 中尤其实用。

实验举例

频率与概率

【例 3.1】(高尔顿钉板试验) 自高尔顿钉板上端放一个小球,任其自由下落。在其下落过程中,当小球碰到钉子时从左边落下的概率为 p,从右边落下的概率为 $1-p$,碰到下一排钉子又是如此,最后落到底板中的某一格子。因此任意放入

输出：

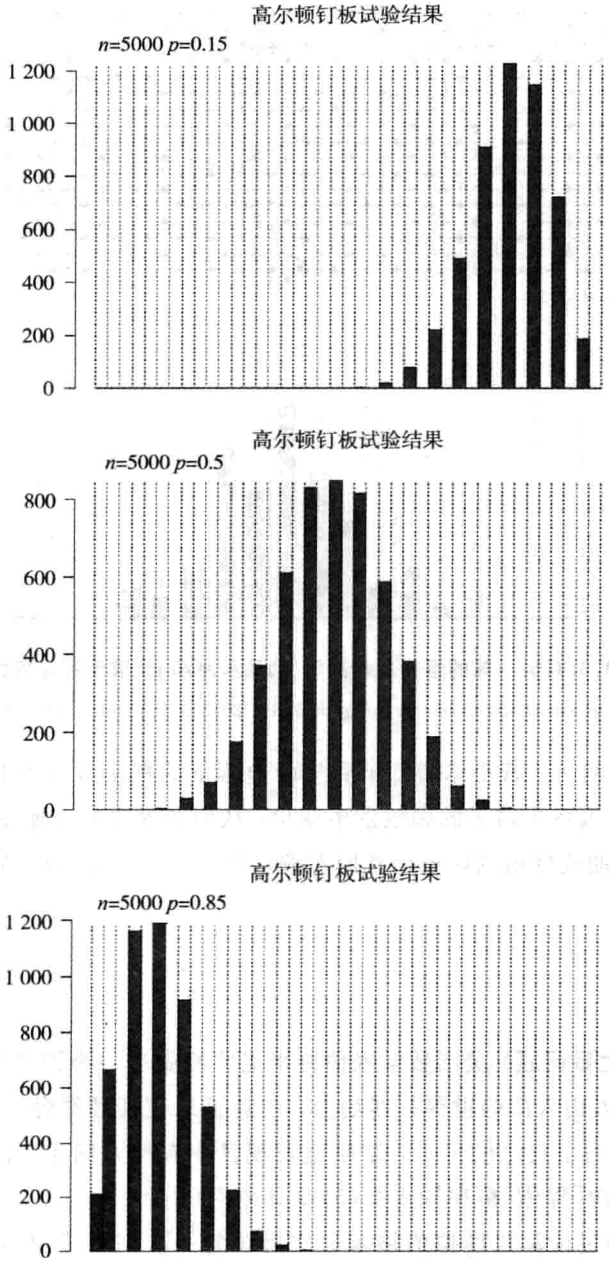


图 1.3.4(a)　最终落点的直方图

(2) 重复(1)1000 次，统计试验结果并填入下表(表 1.3.2)中；

(3) 产生 40 个，50 个，64 个随机数，重复(1)，(2).

表 1.3.2　人群规模与出现生日相同的概率

$n=1000$	r			
	$r=22$	$r=40$	$r=50$	$r=64$
出现同生日次数	506	878	969	999
出现同生日频率	0.506	0.878	0.969	0.999
$f(r)$	0.476	0.891	0.970	0.997

事实上，设随机选取 r 人，$A=\{$至少有两人同生日$\}$，则

$$\overline{A}=\{\text{生日全不相同}\},P(\overline{A})=\frac{P_{365}^{r}}{(365)^{r}},$$

而

$$P(A)=1-P(\overline{A})=1-\frac{P_{365}^{r}}{(365)^{r}}\overset{\Delta}{=}f(r).$$

【问题分析】

本程序的设计思路与前面相似。在前面的基础上，这里应用了 any()、diff()和 sort()函数配合判断向量中是否有相同的数值。sort()是对数据进行排序的函数，默认按小到大排序。diff()是计算相邻元素之间的差值，在时间序列分析的差分过程经常使用。如果向量排序以后相邻元素的差值存在 0，那么显然就是有重复的值了，也即年龄相同的人。any()的输入向量是逻辑值，其取值只有真或假。any()判断其中是否有真的，只要有一个是真的，那么输出真。当然，向量中是否存在重复值这个问题，也可以直接使用 duplicated()函数判断。读者可以自己试试。

【输入命令】(程序 ch010302.R)

```
#1-初始参数
N=1000   #重复次数
n=c(22,40,50,64)   #样本数
y1=y2=y3=c()   #空向量，用于保存频数、频率和理论频率
#2-计算
#随机抽取 n 个数并进行判断，是则返回 TRUE，否则返回 FALSE
fun0305=function(n){
  t1=round(runif(n,1,365),digits=0) #在 1-365 中随机抽取 n 个数，并四舍五入
```

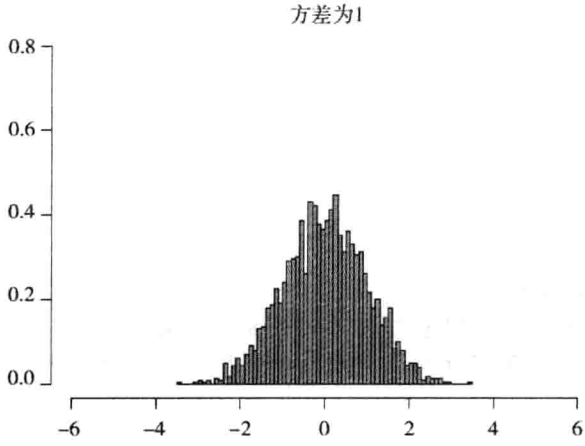


图 1.3.14　方差对正态分布的影响

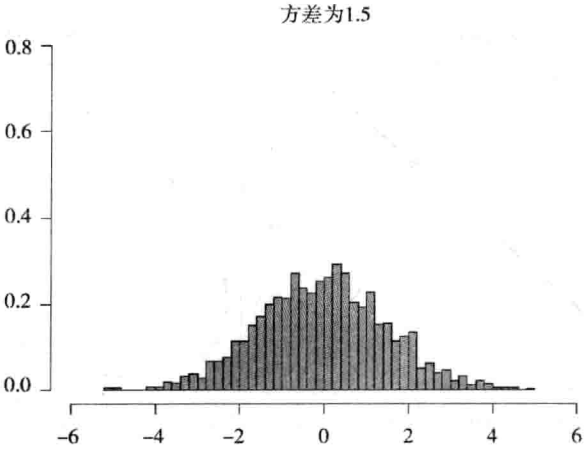


图 1.3.15　方差对正态分布的影响

协方差与相关系数

【例 3.9】 设 B 服从$[0,2\pi]$上的均匀分布，$X=\cos(B)$，$Y=\cos(A+B)$(A 为常数)，X 和 Y 的相关系数为 $\rho=\cos(A)$. 产生服从 $U[0,2\pi]$ 的 N 个随机数，取 $N=100$，对应 $A=0$，$A=\frac{\pi}{3}$，$A=\frac{\pi}{2}$，$A=\pi$ 分别绘出 X 和 Y 的散点图，观察 ρ 对散点图的影响.

输入命令(程序 ch010309.R)

5	270	9	0.09
6	810	0	0.00

将上述结果整理成下表(表 1.3.5)形式：

表 1.3.5　伯努利定理的演示

n	$\left\|\frac{n_A}{n}-p\right\|\geqslant\varepsilon$ 出现的次数	$\left\|\frac{n_A}{n}-p\right\|\geqslant\varepsilon$ 出现的频率
10	76	0.76
30	63	0.63
50	50	0.50
90	35	0.35
270	9	0.09
810	0	0.00

从上表可见，随着 n 的增大，伯努利实验中事件 A 的频率与概率的偏差不小于 ε 的概率越来越接近于 0，即当 n 很大时，事件的频率与概率有较大偏差的可能性很小，由实际推断原理，在实际应用中，当试验次数很大时，便可以用事件发生的频率来代替概率.

中心极限定理的直观演示

【例 3.11】本例旨在直观演示中心极限定理的基本结论："大量独立同分布随机变量的和的分布近似服从正态分布".

按以下步骤设计程序：

(1) 产生服从二项分布 $b(10,p)$ 的 n 个随机数，取 $p=0.2,n=50$，计算 n 个随机数之和 y 以及 $\dfrac{y-10np}{\sqrt{10np(1-p)}}$；

(2) 将(1)重复 $m=1000$ 组，并用这 m 组 $\dfrac{y-10np}{\sqrt{10np(1-p)}}$ 的数据作频率直方图进行观察.

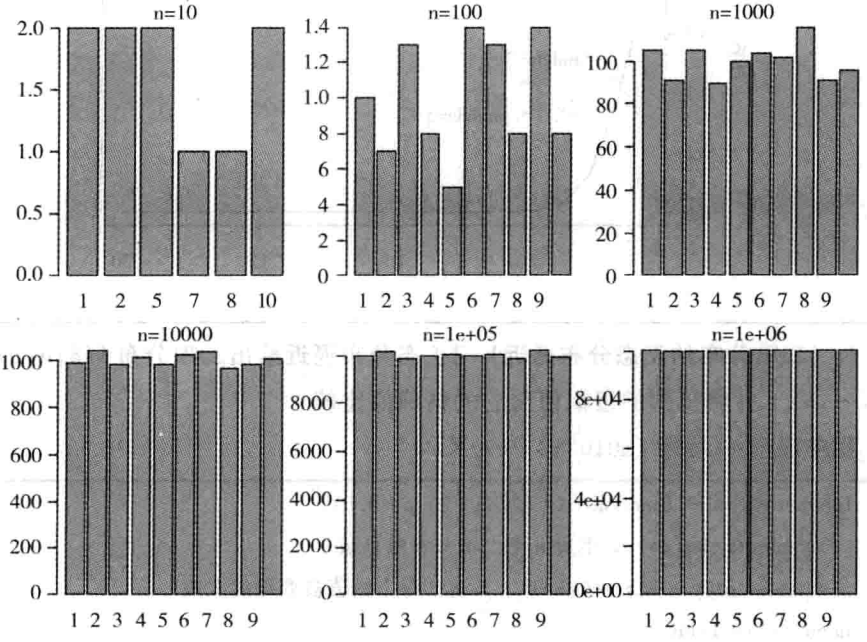
输入(程序 ch010311.R)

```
fun03.ex02=function(n){
    x=replicate(n,{which.max(runif(10))})
    y=table(x)    #统计10所在位置
    barplot(y,main=paste("n=",n,sep="") )    #绘图
}
fun03.ex02(10)
fun03.ex02(100)
fun03.ex02(1000)
fun03.ex02(10000)
fun03.ex02(100000)
fun03.ex02(1000000)
```

则分别输出模拟实验10次、100次、1000次、10000次等的结果，将实验结果进行统计分析，给出分析结果.

注：理论上，易证明每个人抽到大王的机会均等.

结果：



3.（**泊松分布**）利用R语言在同一坐标系下绘出λ取不同值时泊松分布$\pi(\lambda)$的概率分布曲线，通过观察输出的图形，进一步理解泊松分布的概率分布的性质.