华章教育

· 华章应用统计系列 ·

Mathematical Modeling with R

数学建模

基于R

- 以R语言为载体介绍数学建模的常用方法，探索R扩展程序包强大的计算与求解能力。
- 非传统的数学建模教材，穿插丰富的案例，重点介绍实际应用广泛的统计模型和优化模型。

薛毅 编著

机械工业出版社
China Machine Press

数学建模

基于R

Mathematical Modeling

薛毅 编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

数学建模：基于R / 薛毅编著 . —北京：机械工业出版社，2017.5
（华章应用统计系列）

ISBN 978-7-111-57068-4

I. 数… II. 薛… III. 数学模型 – 高等学校 – 教材 IV. O141.4

中国版本图书馆CIP数据核字（2017）第129150号

本书以R语言为载体，介绍数学建模常用的统计方法，并着重介绍了如何从CRAN社区下载相关的R扩展程序包，以及如何使用这些程序包中的函数求解线性规划、最优化、图论与网络、数值分析方面的模型 .

本书可作为“数学建模”课程的教材或数学建模竞赛的辅导教材，也可作为理工、经管、生物等专业的本科生、研究生或相关专业技术人员学习R软件的参考书 .

出版发行：机械工业出版社（北京市西城区百万庄大街22号 邮政编码：100037）
责任编辑：和 静　　责任校对：殷 虹
印　刷：北京市荣盛彩色印刷有限公司　　版　次：2017年7月第1版第1次印刷
开　本：186mm×240mm 1/16　　印　张：21.5
书　号：ISBN 978-7-111-57068-4　　定　价：69.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991 88361066　　投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259　　读者信箱：hzjsj@hzbook.com

前　言

R是一款免费软件，主要用于统计分析、绘图和数据挖掘等．但随着R的广泛使用，R软件的求解能力已不仅仅局限于统计计算的内容，特别是R扩展程序包的下载和安装，大大地增强了R软件的计算与求解能力，例如，能够完成优化、图论与网络、数值分析等方面的计算．

本书之所以命名为《数学建模：基于R》，是因为除介绍数学建模常用的统计方法外，还着重介绍了如何从CRAN(Comprehensive R Archive Network)社区下载相关的扩展程序包，如何使用这些程序包中的函数求解线性规划、最优化、图论与网络、数值分析方面的模型．

采用该命名的第二个原因，是在内容的编排和选取方面与传统的数学建模教材不同，基本上不再讲授传统数学建模课程的基本内容，而是将侧重点放在实际应用中使用较为广泛的两类模型——统计模型和优化模型，以及如何使用R软件求解这两类模型上．

本书共有6章．第1章“概率统计模型”和第2章“多元分析模型”属于统计模型的范畴，只需使用R基本库中的函数就可完成相应的求解工作．第3章“线性规划模型”、第4章“最优化模型”和第5章“图论与网络模型”属于运筹学的内容，使用R基本库中的函数无法完成此类模型的求解，需要在CRAN社区下载相关的扩展程序包，使用程序包中的函数完成运筹学模型的求解工作．第6章“数值分析”介绍数值代数和微分方程数值解等内容，这部分内容实际上是数值计算(包括统计计算)的基础，也可以看成前面内容的补充．

作为数学建模教材，本书的每一章都有一至两个数学建模案例分析，其目的有两个：一是让读者了解数学建模的整个过程；二是复习该章所讲授的知识及相关R函数，学会使用R软件求解问题．

本书介绍的模型完全可由其他软件完成求解工作，如SPSS、LINGO或MATLAB等，但这些都是商业软件，而且有的还价格昂贵．而R是一款免费的开源软件，从这一点来说，对读者更有意义．你不但能够享受到他人的工作成果，也能将你的成果放到网上，与他人分享．这正是R的魅力，也是R这些年来发展如此迅速的原因．

从严格意义上讲，本书不能算作数学建模的教材，也不是R软件使用手册，而是希望通过R对数学模型的求解，让读者了解并学会使用R求解统计或非统计模型，以及如何下载程序包来扩展R的计算能力．当然，在学习了这些内容之后，你可以下载其他的程序包[㊀]，帮助你完成工作或科研所需的计算工作．

本书所介绍的R函数均以R-3.1.1版本为基准，所有函数(包括下载程序包中的函数)均通过测试，读者如果需要书中例题的相关程序，以及例题和习题中的数据文件，可以发

㊀ 截至2015年8月1日，CRAN网站共有6 957个R包，涵盖了不同领域的应用．

送电子邮件向作者索取，邮件地址：xueyi@bjut. edu. cn.

本书可作为“数学建模”课程的教材或教学参考书，也可作为数学建模竞赛的辅导教材，还可作为理工、经济、管理、生物等专业的本科生、研究生或者相关专业的技术人员学习 R 软件的参考书.

受编者水平所限，书中难免存在不足甚至错误之处，欢迎读者不吝指正.

在本书出版之际，谨向对本书提供帮助的各位老师和专家表示感谢，对北京工业大学研究生院对于数学建模课程的支持表示感谢，同时对机械工业出版社为本书出版所做的大量工作表示感谢.

目　录

第 1 章　概率统计模型

数学模型是用变量及数学符号建立起来的一系列等式和不等式，是用来描述客观事物的特征、内在联系及其规律性的模型.

客观事物的某些特征的表现形式，往往具有某种不确定性，因此，代表其特征的变量的取值具有随机性. 有时这些变量虽然不具有随机性，但由于观测条件的限制或随机因素的干扰，这些变量的观测值也具有随机性.

如果按变量的取值是否具有随机性来划分数学模型，数学模型就分成了确定性数学模型和不确定性数学模型. 不确定性数学模型又称为随机模型.

本章和下一章将介绍随机模型的有关内容，以及如何使用 R 中的函数求解随机模型. 后面的各章将介绍确定性模型，以及如何下载扩展程序包，完成确定性模型的求解工作.

1.1　数据的描述性分析

在建立随机模型之前，首先要分析数据的主要特征，也就是数据的数字特征. 这些特征通常是均值、方差，或者是数据服从什么分布. 只有在确定了这些特征之后，才能建立起符合实际的模型.

本节介绍数据的描述性分析的统计方法，以及完成此类分析的 R 函数.

1.1.1　数据的数字特征

已知一组试验(或观测)数据为 X_1，X_2，…，X_n，它们可以是从所要研究的对象的全体——总体 X 中取出的，这 n 个观测值就构成一个样本. 数据分析的任务就是要对这 n 个数据进行分析，提取数据中包含的有用信息.

1. 样本均值

样本均值是数据的平均数，也就是样本的算术平均值，其计算公式为

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{1.1}$$

它描述数据取值的平均位置.

在 R 中，mean()函数用于计算样本均值，其使用格式为

```
mean(x, trim = 0, na.rm = FALSE, ...)
```

参数 x 为需要计算均值的对象(如向量、矩阵、数组或数据框).

trim 为(0，0.5)之间的数(默认值为 0)，表示在计算均值之前，去掉两端数据的百分比，即计算截尾均值.

na.rm 为逻辑变量，表示能否处理带有缺失数据(NA)的样本，默认值为 FALSE.

...为附加参数.

例如，

```
> x <- c(0:10, 50); xm <- mean(x)
> c(xm, mean(x, trim = 0.10))
[1] 8.75 5.50
```

程序中的c()为连接函数，将数据连接成向量.

2. 样本方差

方差是描述数据取值分散性的一个度量. 样本方差是样本相对于均值的偏差平方和的平均，记为S^2，即

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{1.2}$$

其中$\overline{X}$为样本均值. 样本方差的开方称为样本标准差，记为S，即

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{1.3}$$

在R中，可用var()函数计算样本方差，其使用格式为

```
var(x, y = NULL, na.rm = FALSE, use)
```

参数x为数值向量、矩阵或数据框. 当y为NULL(默认值)时，计算样本x的方差；当y为数值向量、矩阵或数据框时，计算样本x与y的协方差.

na.rm为逻辑变量，表示能否处理带有缺失数据(NA)的样本，默认值为FALSE.

计算样本标准差的函数是sd()，其使用格式为

```
sd(x, na.rm = FALSE)
```

参数的意义与var()函数相同.

例如，

```
> x<- c(12, 9, 11, 5, 1, 4, 8, 3, 2, 10, 6, 7)
> var(x)
[1] 13
> sd(x)
[1] 3.605551
```

3. 中位数

中位数就是数据排序位于中间位置的值. 如果数据按顺序排列，对于奇数个数据，中位数就是中间位置的数据；对于偶数个数据，中位数就是中间两个数据的平均值.

在R中，可用median()函数计算样本的中位数，其使用格式为

```
median(x, na.rm = FALSE)
```

参数x为需要计算中位数的数值型向量.

na.rm 为逻辑变量，表示能否处理带有缺失数据(NA)的样本，默认值为 FALSE.

例如，

```
> x <- c(12, 9, 11, 5, 1, 4, 8, 3, 2, 10, 6, 7, NA)
> median(x)
[1] NA
> median(x, na.rm = TRUE)
[1] 6.5
```

4. 分位数

分位数可以看成中位数的推广，如中位数就是 0.5 分位数．常用的还有$\frac{1}{4}$分位数和$\frac{3}{4}$分位数，分别称为下四分位数和上四分位数．p 分位数又可称为第 $100p$ 百分位数，如中位数可称为第 50 百分位数，下四分位数可称为第 25 百分位数，上四分位数可称为第 75 百分位数．

在 R 中，quantile()函数计算样本的分位数，其使用格式为

```
quantile(x, probs = seq(0, 1, 0.25), na.rm = FALSE,
         names = TRUE, type = 7, ...)
```

参数 x 为样本构成的数值向量．

probs 为数值向量，数值在 0 与 1 之间，表示需要计算的分位数，默认值为 0，0.25，0.50，0.75 和 1.

na.rm 为逻辑变量，表示能否处理带有缺失数据(NA)的样本，默认值为 FALSE.

names 为逻辑变量，表示是否将百分位数作为返回值中变量的属性，默认值为 TRUE.

type 为 1～9 之间任何一个整数，表示计算分位数的算法，默认值为 7.

例如，

```
> quantile(1:10)
   0%   25%   50%   75%  100%
 1.00  3.25  5.50  7.75 10.00
```

5. 极差与四分位极差

样本极差(记为 R)是一组数据的最大值与最小值之差，其计算公式为

$$R = \max(X) - \min(X) \tag{1.4}$$

其中 X 是由样本构成的向量．样本极差是描述样本分散性的数字特征．数据越分散，其极差越大．由于极差是利用一组数据两端的信息，因此容易受到极端值的影响．

在 R 中，与极差有关的函数有 max()，min()和 range()，其使用格式为

```
max(..., na.rm = FALSE)
min(..., na.rm = FALSE)
range(..., na.rm = FALSE)
```

参数...为数据构成的向量．

max()函数的返回值是最大值，min()函数的返回值是最小值，range()函数的返回

值是由最小值和最大值构成的二维向量.

因此，计算极差的程序为

```
R <- max(x) - min(x)
r <- range(x); R <- r[2] - r[1]
```

其中 x 是由样本构成的向量.

样本的上、下四分位数之差称为四分位极差(或半极差)，记为 R_1，即

$$R_1 = Q_3 - Q_1 \tag{1.5}$$

它也是度量样本分散性的重要数字特征，特别对于具有异常值的数据，它作为分散性度量具有稳健性. 因此，它在稳健性数据分析中具有重要作用.

由半极差的定义(式(1.5))，可得到半极差的计算公式

```
R1 <- quantile(x, 3/4) - quantile(x, 1/4)
```

其中 x 是由样本构成的向量.

例 1.1　某班有 31 名学生，某门课的考试成绩如下：

25　45　50　54　55　61　64　68　72　75　75

78　79　81　83　84　84　84　85　86　86　86

87　89　89　89　90　91　91　92　100

求这门课程考试成绩的平均值、中位数、方差、极差和四分位极差.

解　编写一个统一计算平均值、中位数、方差、极差和四分位极差等统计量的函数(程序名：discript.R).

```
discript <- function(x){
    R1 <- quantile(x, 3/4, names = FALSE)
          - quantile(x, 1/4, names = FALSE)
    data.frame(
        n = length(x), max = max(x), min = min(x),
        R = max(x) - min(x), R1 = R1,
        mean = mean(x), median = median(x),
        sd = sd(x)
    )
}
```

程序中的 length()函数是计算向量的长度，即维数.

将数据写入数据文件(exam0101.data)，然后调用该函数计算：

```
X <- scan("exam0101.data")
source("discript.R"); discript(X)
```

在程序中，scan()函数是从数据文件中读取数据，source()是将编写好的函数调入内存，其计算结果如下：

```
n  max  min   R  R1  mean    median          sd
1   31  100  25  75    18  76.70968  84 16.69769
```

计算结果说明：共有 31 名学生的成绩，最高分为 100 分，最低分为 25 分，极差是 75 分，半极差是 18 分，平均分数为 76.7 分，中位数分数为 84 分，标准差为 16.7 分.

1.1.2　随机变量的分布

1. 随机变量

从一个总体中抽取不同的样本，分析各个样本所获得的点估计往往不尽相同，这种表现出变异性特征的量称为变量.

在进行统计试验以前，一般并不知道某一试验的确切结果，但是可以赋予试验结果以实际数量的一个函数. 因此这一变量称为随机变量. 随机变量常用大写字母表示，如 X，Y，Z. 它们可能出现的具体结果或数值则可用小写字母表示，如 x，y，z.

最常见的随机变量有两类. 一类是以计数形式表示的随机变量，称为离散型随机变量；另一类是取值在某个有限或无限区间的随机变量，称为连续型随机变量.

2. 分布函数

描述一个随机变量 X，不仅要说明它能够取哪些值，而且还要关心它取这些值的概率. 对任意的实数 x，令

$$F(x) = P\{X \leqslant x\}, \quad x \in (-\infty, +\infty) \tag{1.6}$$

则称 $F(x)$为随机变量 X 的分布函数，也称为累积分布函数.

从直观上看，分布函数 $F(x)$是一个定义在$(-\infty, +\infty)$上的实值函数，$F(x)$在点 x 处取值为随机变量 X 落在区间$(-\infty, x]$上的概率.

3. 概率函数与概率密度函数

如果随机变量 X 的全部可能取值只有有限多个或可列无穷多个，则称 X 为离散型随机变量.

对于离散型随机变量 X，可能取值为 x_k 的概率为

$$P\{X = x_k\} = p_k, \quad k = 1,2,\cdots \tag{1.7}$$

则称式(1.7)为离散型随机变量 X 的分布律.

离散型随机变量的分布函数为

$$F(x) = P\{X \leqslant x\} = \sum_{x_k \leqslant x} P\{X = x_k\} = \sum_{x_k \leqslant x} p_k \tag{1.8}$$

对于随机变量 X，如果存在一个定义在$(-\infty, +\infty)$上的非负函数 $f(x)$，使得对于任意实数 x，总有

$$F(x) = P\{X \leqslant x\} = \int_{-\infty}^{x} f(t)\mathrm{d}t, \quad -\infty < x < +\infty \tag{1.9}$$

则称 X 为连续型随机变量，$f(x)$为 X 的概率密度函数，简称概率密度.

4. 分位数

设 X 为随机变量，对任给的 $0<\alpha<1$，若存在 x_α，使得

$$P\{X \leqslant x_\alpha\} \geqslant 1-\alpha, \quad P\{X > x_\alpha\} \geqslant \alpha \tag{1.10}$$

则称 x_α 为 X 的上 α 分位数(或上 α 分位点). 对任给的 $0<p<1$，若存在 x_p，使得

$$P\{X \leqslant x_p\} \geqslant p, \quad P\{X > x_p\} \geqslant 1-p \tag{1.11}$$

则称点 x_p 为 X 的下 p 分位数(或下 p 分位点). 由式(1.10)和式(1.11)可以得到上、下分位数之间的关系：上 α 分位数就是下 $1-\alpha$ 分位数.

由分位数的关系式(1.10)(或式(1.11))可知，分位数不是唯一的. 但对于连续型随机变量，分位数确实是唯一的，此时，可将式(1.11)改写成

$$F(x_p) = p$$

其中 $F(x)$为随机变量 X 的分布函数. 对于连续型随机变量，分布函数是严格单调递增的，所以可将下 p 分位数表示为逆分布函数，即

$$x_p = F^{-1}(p) \tag{1.12}$$

由上、下分位数之间的关系式，上 α 分位数可表示为

$$x_\alpha = F^{-1}(1-\alpha) \tag{1.13}$$

1.1.3　常用的分布

1. 正态分布

若随机变量 X 的概率密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < +\infty \tag{1.14}$$

其中 μ 和 $\sigma(\sigma>0)$为两个常数，则称 X 服从参数为 μ 和 σ^2 的正态分布，也称为 Gauss(高斯)分布，记作 $X\sim N(\mu, \sigma^2)$. 若 $X\sim N(\mu, \sigma^2)$，则

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt, \quad -\infty < x < +\infty \tag{1.15}$$

在 R 中，正态分布的基本名称为 norm，加上不同的前缀表示不同的函数，如 dnorm 表示概率密度函数，pnorm 表示分布函数，qnorm 表示分位函数. 函数的使用格式为

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
```

参数 x 或 q 为纯量或向量，表示概率密度函数或分布函数的自变量.

p 为纯量或向量，描述分位点的概率.

mean 为纯量或向量，描述均值参数(即 μ)的取值，默认值为 0.

sd 为纯量或向量，描述标准差参数(即 σ)的取值，默认值为 1.

log，log.p 为逻辑变量，表示概率值 p 是否由其对数值 log(p)给出，默认值为 FALSE.

lower.tail 为逻辑变量，表示是否作下尾运算. 取 TRUE(默认值)表示 $F(x)=P\{X\leqslant x\}$，对应的分位数是下分位数. 取 FALSE 表示 $F(x)=P\{X>x\}$，对应的分位数是上

分位数.

图 1.1 描绘的是不同参数的正态分布的概率密度函数图，分别是：$\mu=0$，$\sigma=0.5$；$\mu=2$，$\sigma=0.5$；$\mu=0$，$\sigma=1$.

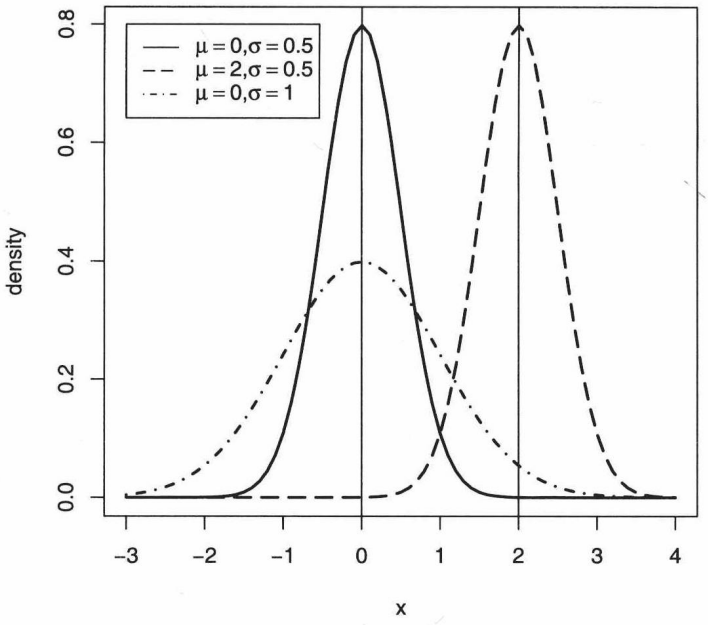

图 1.1　正态分布的概率密度函数

例 1.2　设 $X\sim N(\mu,\ \sigma^2)$，分别计算 $P\{|X-\mu|\leqslant\sigma\}$，$P\{|X-\mu|\leqslant 2\sigma\}$ 和 $P\{|X-\mu|\leqslant 3\sigma\}$，以及正态分布的上 0.025 分位点 $Z_{0.025}$.

解　当 $X\sim N(\mu,\ \sigma^2)$ 时，$Z=\dfrac{X-\mu}{\sigma}\sim N(0,\ 1)$，所以用标准正态分布计算即可.

```
> x <- 1:3; p <- pnorm(x) - pnorm(- x); p
[1] 0.6826895 0.9544997 0.9973002
```

这就是通常所说的 3σ 原则，即在 1σ，2σ 和 3σ 区间内的概率分别为 68.3%，95.4% 和 99.7%.

注意：qnorm()函数提供的是下分位点，计算上 α 分位点等价于计算下 $1-\alpha$ 分位点，所以计算程序为

```
> alpha <- 0.025; z <- qnorm(1- alpha); z
[1] 1.959964
```

2. χ^2 分布

如果 $Z_i\sim N(0,\ 1)(i=1,\ 2,\ \cdots,\ n)$，且 Z_i 是相互独立的，则称

$$X=Z_1^2+Z_2^2+\cdots+Z_n^2 \tag{1.16}$$

为自由度为 n 的 χ^2 分布，记为 $X\sim\chi^2(n)$. 如果 $Z_i\sim N(\delta,\ 1)$，则称 X 为非中心化的 χ^2 分

布，记 $X\sim\chi^2(n,\delta)$，称 δ 为非中心化参数.

在R软件中，用 chisq 表示 χ^2 分布，其调用格式如下：

```
dchisq(x, df, ncp= 0, log = FALSE)
pchisq(q, df, ncp= 0, lower.tail = TRUE, log.p = FALSE)
qchisq(p, df, ncp= 0, lower.tail = TRUE, log.p = FALSE)
```

参数 df 为自由度，ncp 为非中心化参数(即 δ)，其余参数的意义与正态分布函数相同.

图 1.2 描绘的是 χ^2 分布的概率密度函数在不同参数下的图形.

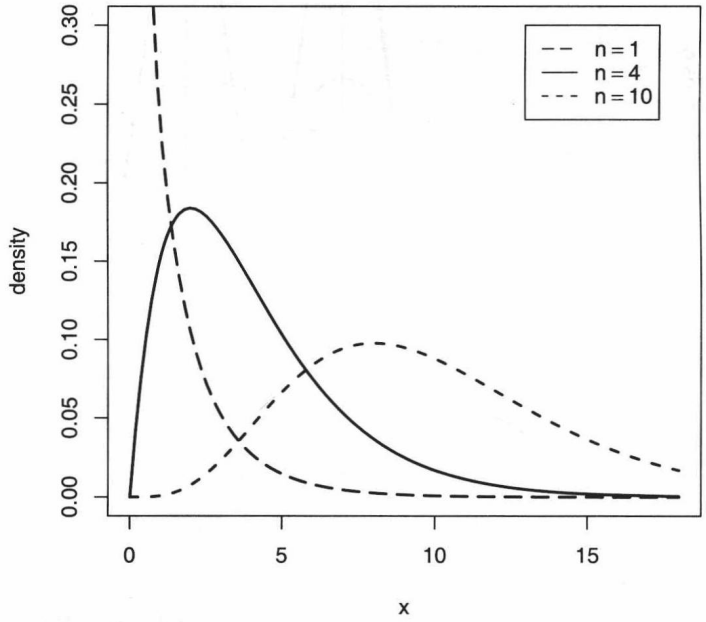

图 1.2　χ^2 分布的概率密度函数

3. t 分布

如果随机变量 $Z\sim N(0,1)$，$X\sim\chi^2(n)$且 X 与 Z 相互独立，则称

$$T=\frac{Z}{\sqrt{X/n}} \tag{1.17}$$

为自由度为 n 的 t 分布，记为 $T\sim t(n)$. 如果 $Z\sim N(\delta,1)$，则称 T 为非中心化 t 分布，记为 $T\sim t(n,\delta)$，称 δ 为非中心化参数.

在R中，t 分布的使用格式是

```
dt(x, df, ncp = 0, log = FALSE)
pt(q, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
qt(p, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
```

参数 df 为自由度，ncp 为非中心化参数(即 δ)，其余参数的意义与正态分布函数相同.

图 1.3 描绘的是 t 分布的概率密度函数在不同参数下的图形.

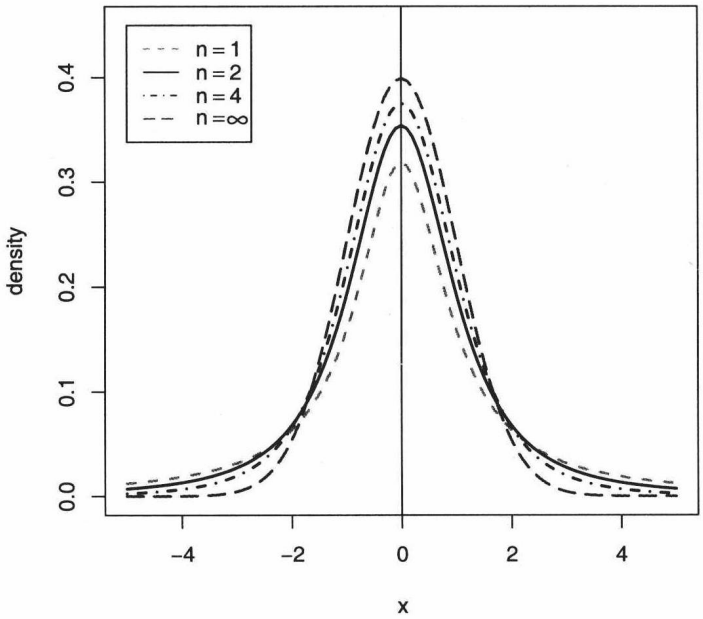

图 1.3　t 分布的概率密度函数

4. F 分布

如果随机变量 $X\sim\chi^2(n_1)$，$Y\sim\chi^2(n_2)$且相互独立，则称

$$F=\frac{X/n_1}{Y/n_2} \tag{1.18}$$

为第 1 自由度为 n_1 和第 2 自由度为 n_2 的 F 分布，记为 $F\sim F(n_1, n_2)$. 如果 $X\sim\chi^2(n_1, \delta)$，则称 F 为非中心化 F 分布，记为 $F\sim F(n_1, n_2; \delta)$，称 δ 为非中心化参数.

在 R 中，F 分布的使用格式为

```
df(x, df1, df2, ncp = 0, log = FALSE)
pf(q, df1, df2, ncp = 0, lower.tail = TRUE, log.p = FALSE)
qf(p, df1, df2, ncp = 0, lower.tail = TRUE, log.p = FALSE)
```

参数 df1 为第 1 自由度，df2 为第 2 自由度，ncp 为非中心化参数(即 δ)，其余参数的意义与正态分布函数相同.

图 1.4 描绘的是 F 分布的概率密度函数在不同参数下的图形.

1.1.4　数据的图形描述

可以通过数据的图形描述判断数据的分布情况，比如是否来自于正态分布等.

1. 直方图

直方图又称柱状图或质量分布图，是一种统计报告图，由一系列高度不等的纵条纹或