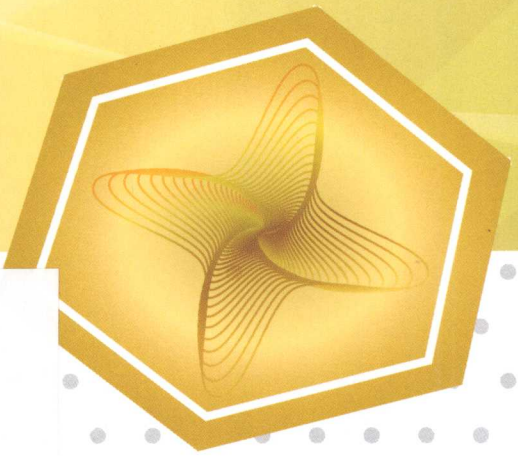
浙江省级重点学科应用数学教学改革与科学研究丛书

数据分析与R软件

李素兰 编



科学出版社

浙江省级重点学科应用数学教学改革与科学研究丛书

数据分析与R软件

李素兰 编

浙江工业大学重点教材建设项目基金资助

科学出版社

内 容 简 介

本书的主要内容包括描述性统计分析、非参数统计及常用的多元统计分析方法，如回归分析、主成分分析、聚类分析、判别分析等.这些统计方法是进行数据处理的必要技术，是进一步深造与统计相关专业的基础，是金融、统计、计算机等行业必不可少的分析处理工具之一. 应用实例通过国际通用统计软件 R 实现. R 软件是完全免费的统计软件，是用于统计分析和制图的优秀软件.

本书可作为信息与计算科学、应用数学、统计学等专业数据分析类课程的基础教材,也可作为对统计数据分析有较高要求的高年级本科学生和工科硕士研究生各专业的选修教材，还可作为统计、管理、经济、金融、生物、心理、医疗等科研和工程技术人员的参考读物.

图书在版编目(CIP)数据

数据分析与 R 软件/李素兰编. —北京: 科学出版社, 2013

(浙江省级重点学科应用数学教学改革与科学研究丛书)

ISBN 978-7-03-038072-2

Ⅰ. ①数… Ⅱ. ①李… Ⅲ. ①统计分析–应用软件 Ⅳ. ①C819

中国版本图书馆 CIP 数据核字 (2013) 第 141204 号

责任编辑: 石 悦 孙翠勤 / 责任校对: 包志虹

责任印制: 阎 磊 / 封面设计: 华路天然设计工作室

科学出版社 出版

北京东黄城根北街 16 号

邮政编码:100717

http://www.sciencep.com

北京市安泰印刷厂印刷

科学出版社发行 各地新华书店经销

*

2013 年 6 月第 一 版 开本: 720×1000 B5

2013 年 6 月第一次印刷 印张: 18

字数: 383 000

定价: 38.00 元

(如有印装质量问题, 我社负责调换)

“浙江省级重点学科应用数学教学改革与科学研究丛书”

编　委　会

总　　序

近年来，关于数学的各种新观点不断出现.

有一种观点认为，随着数学的发展，数学已经从自然科学中分离出来，成为独立的科学门类——数学科学.

持这种观点的学者的依据是：①从现代数学的发展情况可以看出，数学的许多内容和方法的产生，不再是基于研究自然界中存在的物质运动规律的需要，而是基于数学自身的需要. 例如，6=3+3，8=3+5，10=5+5=3+7，等等，即每一个大于等于6的偶数都可以表示为两个奇素数的和，这就是哥德巴赫猜想，至今没有证明. 但是，这样一个在数学中显得十分重要的著名的猜想，其结果的对与否，不会对数学之外的任何学科产生影响，证明它不是自然科学的需要，而仅仅是数学科学的需要. ②数学不仅具有应用功能，而且具有其他学科不能比拟的教育功能. 数学的应用功能表现在：没有数学，现代科技无从谈起；任何一种学科，只有应用了数学，才能成为科学学科. 数学的教育功能表现在：在中国，语文、数学、英语被认为是初等教育中最重要的三门课程；在世界范围内，有不学中文的学生，有不学英语的学生，但没有不学数学的学生.

我同意这种观点，希望在数学教学改革和科学研究中体现这种观点.

数学教学改革，首先需要的是教材的改革，而教材的改革，涉及的只有两个方面：一是内容；二是方法.

如何在一本数学教材中以数学科学的观点选取内容、介绍方法？

我的认识是：无论是选取内容方面，还是介绍方法方面，都要关注数学的应用功能和教育功能的展现.

在内容的选取方面，既不是不管数学的教育功能，狭隘地全部以目前生产生活的实际应用为目的，打乱系统，什么“有用”就选什么，什么“没用”就跳过什么；也不是完全从数学的需要出发，一点也不考虑所选取的内容和实际应用的联系. 本套丛书采取有实际应用背景的内容优先选取的原则. 我们的考虑是：没有迹象表明，没有实际应用背景的内容在体现数学的教育功能时强于有实际应用背景的内容，既然如此，后者更有利于同时展现数学的应用功能和教育功能.

在方法的介绍方面，既不完全采用公理化体系的做法，让读者在接受严格数学训练的基础上自然地受到数学科学的熏陶；也不完全摒弃数学特有的推理过程，以急功近利的方式只讲结果，只讲计算公式. 我们知道，公理化体系的做法是将数学的训练目的不直接说出来，而是藏起来，藏在严密的过程背后，让学生不知不觉得到严格的数学训练. 这种体系在介绍内容时，不交代前因后果，一上来就是莫名其妙的定义、公理，然后一步步以极其严密的方式展开讨论. 这种做法在知识门类相对少的过去是有效的，但在知识爆炸、课程门类不断增加、学生同时要有做学问和实际应用两

手准备的现在, 没有时间这样做. 训练要有, 但训练目的不是藏起来, 而是尽可能直接讲出来. 例如, 数学书籍中一定会用到归纳法、演绎法、反证法, 这些方法不是数学特有的, 但可以被数学最为有效地传授给学生, 这一事实恰好可以说明数学的教育功能的强大. 但是, 如果我们去问一下数学系的毕业生什么是演绎法, 恐怕很少有人能说周全, 究其原因, 是我们的教材没有明确地告诉学生演绎法的基本内容和过程. 本套丛书将致力于改变这种状况.

本套丛书注意到：根据课程和授课对象的不同, 数学的应用功能和教育功能的展现需分层次, 两种功能的展现要有机配合. 例如, 有的数学分支本来就属于应用数学, 对这样的课程, 在选取内容和介绍方法时必须首先保证应用方面的需要, 其次才考虑教育功能的融入; 有的授课对象是文科学生, 对这些学生, 在编写教材时就要充分注意他们的基础、兴趣、思维方式和希望通过数学的学习要达到的目的, 因此要首先考虑数学的教育功能, 其次才考虑应用功能的融入.

现代化的标志是数字化, 也就是要在所有的领域尽最大可能地使用计算机技术, 因此, 在数学教学中, 对数字化的配合和适应是必需的. 为了展现数学的应用功能, 在数学教学的每个环节, 都应该关注计算机技术, 包括有意考虑内容的计算机实现, 如算法问题, 内容与几个成功的数学软件的结合问题. 我们知道, 介绍如何应用数学软件的最好环境, 当为相应的数学课程. 因此, 本套丛书中的教材, 特别注意介绍与主要内容配套的软件的应用, 例如, 介绍相应的MATLAB 软件包的使用.

科学研究成果整理成学术著作, 可以总结和条理化研究问题, 这对于传播研究成果、深化研究工作是有利的, 这些著作还可以作为研究生教材使用.

本套丛书中学术著作的撰写遵循了如下的原则:

首先, 作为介绍学术成果的学术著作要有新内容、新观点, 学术系统应是明显的, 不是杂乱的、拼凑的, 特别是著作中作者的成果应有重要的分量.

其次, 本套丛书中的学术著作特别注意内容的系统性、完备性.

再次, 也是最重要的, 本套丛书中的学术著作, 和教材一样注意展现数学的应用功能和教育功能, 在必要时, 还考虑内容的计算机实现, 如算法问题, 内容与几个成功的数学软件的结合问题.

最后, 在写作细节上, 本套丛书要求作者以严格的科学态度对待自己的著作, 概念和符号应明确, 推导和介绍要细致, 避免突然出现翻遍全书都找不到介绍的概念和符号, 避免用显然、易知等词语掩盖困难的证明过程.

教学改革涉及的问题很多, 有些问题需要一步步解决, 有的还需要根据形势的变化调整解决方案. 我们仅做了初步的尝试, 加之水平有限, 本套丛书中的问题一定很多, 迫切希望读者批评指正.

邸继征

2013 年 3 月 8 日

前　言

本书作为信息与计算科学、应用数学、统计学等专业数据分析类课程的教材, 结合了信息与计算科学及相关专业的专业培养目标, 即掌握数学科学的基本理论与方法,具有运用数学知识、使用计算机解决实际问题的能力, 受到科学研究的初步训练,能在信息与计算科学领域从事科学研究, 解决有关实际问题及设计开发某些软件. 本书的主要内容包括描述性统计分析、非参数统计推断及常用的多元统计分析方法 (回归分析、主成分分析、聚类分析、判别分析、典型相关分析等). 这些统计方法是进行数据处理的必要技术, 是进一步深造与统计相关专业的基础, 是金融、统计、计算机等行业必不可少的分析处理工具之一. 本书应用实例通过国际通用统计软件 R 实现. R 软件是完全免费的统计软件, 是用于统计分析和制图的优秀软件, 具有统计分析功能强大、用户可以编写自己的程序等优点. 熟练掌握本教材的相关内容, 能提高学生解决实际问题的能力、学生的创新能力和开发某些软件的能力, 为学生在科技、教育、经济、统计、金融和计算机等部门从事研究、教学工作或在生产、经营及管理部门从事实际应用、开发研究和管理工作奠定基础.

本书具有如下特点.

(1) 精选教材的支撑内容, 兼顾学生的数学基础和工程应用的实际, 着重介绍在经济、生物、金融、心理等相关领域实用的统计方法.

(2) 介绍具体知识点有侧重, 详尽介绍概念的实际意义, 重点介绍统计思想、数据分析方法、统计模型及分析结果, 避免太复杂的理论推导和证明过程, 力求易懂, 注重实用性!

(3) 内容结构采用模块设置, 内容安排相对独立, 用书单位可根据具体情况选择模块教学.

(4) 与数据处理软件 R 结合. 出于对工科专业数学计算的多样需求、软件的通用性和一举多得等方面的考虑, 本书同步介绍了 R 软件, 使理论学习更容易、更直观,使软件学习更充实, 内容更丰富.

(5) 方法与统计案例结合. 本书精选统计方法的应用实例和统计案例, 涉及经济、生物、金融、心理、医学、气象等领域, 通过 R 软件实现, 便于学生活学活用, 学以致用!

本书可作为工科硕士研究生统计类课程的基础教材, 也可作为对统计数据分析有较高要求的本科各专业高年级学生的选修教材, 还可作为统计、管理、经济、金融、生

物、心理、医疗等科研和工程技术人员的参考读物.

由于编者水平所限, 书中尚存在一些不妥之处, 欢迎读者不吝指正. 读者如果需要编者自编的 R 程序, 可以通过电子邮件索取, 邮箱地址：sulanli@zjut. edu.cn.

编者

2013 年 5 月于浙江工业大学

目　录

第 1 章　探索性数据分析

探索性数据分析(exploratary data analysis) 的基本思想是从数据本身出发, 介绍数据分析的基本方法, 不涉及模型的假设和统计推断, 采用非常灵活的方法来探究数据分布的大致情况, 主要内容包括基本数字特征、绘制直方图、茎叶图和箱线图等. 为进一步结合模型的研究提供线索, 为传统的统计推断提供良好的基础并减少盲目性.

1.1　数 字 特 征

1.1.1　一维数据的数字特征

假设有一组样本数据 $x_1, x_2, \cdots, x_n$, 如果来自总体 X, 则这 n 个数据构成一个样本容量为 n 的样本数据观测值. $x_1, x_2, \cdots, x_n$ 就是所要研究对象的全体, 数据分析的目的就是对 n 个样本观测值进行分析, 提取数据中包含的有用信息. 研究数据的数字特征是主要分析方法之一, 通过数据的数字特征分析, 反映数据的集中位置、分散程度、分布形状等, 进一步可以推断样本中包含的总体信息.

1. 数据位置的数字特征

假设研究对象是 n 个样本数据 $x_1, x_2, \cdots, x_n$. 最常用的描述数据集中位置的数字特征是均值.

(1) 均值.

均值(mean) 是这 n 个数据 $x_1, x_2, \cdots, x_n$ 的样本平均值, 记为 $\bar{x}$, 即

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1.1.1}$$

它描述了数据的集中位置, 是总体均值的矩估计, 更适合来自正态分布的数据分析.

若总体分布未知、数据严重偏态或有若干异常值时, 均值所反映数据的集中位置不是十分合理, 可以采用中位数.

(2) 中位数.

n 个数据 $x_1, x_2, \cdots, x_n$ 从小到大排序后记为

$$x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)},$$

这就是次序统计量的值. **中位数**(median) 的定义为

$$M=\begin{cases} x_{(\frac{n+1}{2})}, & n\text{为奇数}, \\ \dfrac{1}{2}\left\{x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)}\right\}, & n\text{为偶数}, \end{cases} \tag{1.1.2}$$

中位数是描述数据中间位置的数字特征, 对于对称分布的数据, 中位数两侧的数据个数大致相等, 中位数与均值也比较接近; 对于偏态分布的数据, 中位数与均值不同. 中位数不受异常值的影响, 具有稳健性. 在实用上, 中位数用得很多, 有不少社会统计资料, 常用中位数来刻画某个量的代表性数值.

更详细描述数据位置的数字特征还有 p 分位数和三均值.

(3) p 分位数.

p 分位数又称为**百分位数**(percentile), 是中位数的推广, 对于 $0\leqslant p\leqslant 1$, p 分位数定义为:

$$M_p=\begin{cases} x_{([np]+1)}, & np\text{不是整数}, \\ \dfrac{1}{2}\left\{x_{(np)}+x_{(np+1)}\right\}, & np\text{是整数}, \end{cases} \tag{1.1.3}$$

其中 $[np]$ 表示 np 的整数部分, $x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$ 是次序统计量的值.

当 $p=1$ 时, 规定 $M_1=x_{(n)}$.

大体上整个样本的 $(100p)\%$ 观测值不超过 p 分位数, 0.5 分位数就是中位数 M, 在实际应用中, 0.25 分位数和 0.75 分位数比较重要, 分别称为**下、上四分位数**, 记为 $Q_1=M_{0.25}$, $Q_3=M_{0.75}$.

把下四分位数、中位数和上四分位数合称为**四分位数**(quartile), 即 $Q_1=M_{0.25}$, $Q_2=M_{0.5}$, $Q_3=M_{0.75}$ 为四分位数. 将所有数据按从小到大顺序并分成四等份, 处于三个分割点位置的数据就是四分位数.

(4) 三均值.

均值 $\bar{x}$ 包含了样本 $x_1,x_2,\cdots,x_n$ 的全部信息, 但存在异常值时缺乏稳健性. 中位数 M 具有较强的稳健性, 但仅用了数据分布中的部分信息. 考虑到既要充分利用样本信息, 又要具有较强的稳健性, 可以用**三均值**作为数据集中位置的数字特征. 三均值计算公式为

$$\frac{1}{4}Q_1+\frac{1}{2}M+\frac{1}{4}Q_3. \tag{1.1.4}$$

它是 Q_1, M 和 Q_3 的加权平均, 权重分别为 $\dfrac{1}{4}$, $\dfrac{1}{2}$ 和 $\dfrac{1}{4}$.

下面看一个计算上述数字特征的例子, 本书所有实例都通过 R 软件实现, 符号“>”后面的语句是输入的 R 命令, 符号“#”后的内容是上面命令的注释. R 软件的使用说明可参见第 9 章.

例 1.1.1 调查 20 名男婴的出生体重 (kg), 资料如下, 试求位置的数字特征.
2.770, 2.915, 2.795, 2.995, 2.860, 2.970, 3.087, 3.126, 3.125, 4.654, 2.272, 3.503, 3.418, 3.921, 2.669, 4.218, 3.707, 2.310, 2.573, 3.881.

解 输入 R 命令:

```
> w<-c(2.770,2.915,2.795,2.995,2.860,2.970,3.087,3.126,3.125,4.654,
2.272,3.503,3.418,3.921,2.669,4.218,3.707,2.310,2.573,3.881)
  #数据赋值于向量w
> w.mean<-mean(w); w.mean
  #求均值赋值于w.mean,并输出
> w.median<-median(w); w.median
  #求中位数赋值于w.median,并输出
> q.quantile=quantile(w); q.quantile
  #求分位数赋值于q.quantile,输出结果是带有元素名字的向量(详见245页取
  字符型值的下标向量)
> Q1=q.quantile[2]; Q1
  #求下四分位数赋值于Q1,并输出
> Q3=q.quantile[4]; Q3
  #求上四分位数赋值于Q3,并输出
> M3=Q1*(1/4)+q.quantile[3]*(1/2)+Q3*(1/4); M3
  #求三均值赋值于M3,并输出
```

输出结果:

```
> w<-c(2.770,2.915,2.795,2.995,2.860,2.970,3.087,3.126,+3.125,4.654,
2.272,3.503,3.418,3.921,2.669,4.218,3.707,+2.310,2.573,3.881)
> w.mean<-mean(w);  w.mean
[1] 3.18845
> w.median<-median(w); w.median
[1] 3.041
> q.quantile=quantile(w); q.quantile
     0%      25%      50%      75%      100%
2.27200  2.78875  3.04100  3.55400  4.65400
> Q1=q.quantile[2]; Q1
  25%
2.78875
> Q3=q.quantile[4]; Q3
  75%
```

```
3.554
> M3=Q1*(1/4)+q.quantile[3]*(1/2)+Q3*(1/4); M3
  25%
3.106188
```

这是第一个 R 语句实现, 所以详细地写出了输出结果, 以后熟悉了, 只写主要结论. 从输出结果可以看出: 均值为 3.18845; 中位数为 3.041; 下、上四分位数分别为 2.78875, 3.554; 三均值计算得 3.106188.

2. 数据分散性的数字特征

除了关心数据的集中位置, 还需要研究数据在其中心位置附近散布程度的数字特征, 其中最重要的是样本方差.

(1) 样本方差.

样本方差(sample variance) 是样本相对于均值的偏差平方和的平均, 记为 s^2. 是描述数据分散性的一个重要的数字特征, 计算公式为

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{1.1.5}$$

样本方差作为数据分散程度的度量, 它有一个缺点, 它的单位是样本取值单位的平方. 为了使该度量的单位与样本取值的单位相同, 使用样本方差的平方根. 样本方差的平方根称为**样本标准差**(standard deviation) , 记为 s, 即 $s = \sqrt{s^2}$.

(2) 变异系数.

变异系数(coefficient of variance) 又称为标准差系数, 是标准差与均值的比值. 标准差是绝对指标, 其值大小不仅取决于样本数据的分散程度, 还取决于样本数据平均水平的高低, 当进行两个或多个资料变异程度的比较时, 如果度量单位和均值相同, 可以直接利用标准差来比较. 如果单位或平均值不同时, 比较其变异程度就不能采用标准差. 变异系数可以消除单位或平均值不同对两个或多个资料变异程度比较的影响. 变异系数的计算公式为

$$\mathrm{CV} = \left(100 \times \frac{s}{\bar{x}}\right)\%. \tag{1.1.6}$$

(3) 极差.

极差(range) 也称**全距**, 计算公式为

$$R = x_{(n)} - x_{(1)}, \tag{1.1.7}$$

即最大值与最小值的差, 也是描述数据分散性的指标. 数据越分散, 极差越大. 由于极差仅取决于两个极值, 容易受异常值影响, 所以在实际中很少使用.

上、下四分位数之差称为**四分位极差**(quartile range) 或**半极差**, 记为 R_1, 即

$$R_1 = Q_3 - Q_1. \tag{1.1.8}$$

它也是度量样本数据分散性的重要数字特征, 因为具有稳健性, 特别对于有异常值的数据, 在稳健性数据分析中具有重要作用.

判断数据中是否有异常值, 可以用下面的方法.

(4) 上、下截断点和异常值.

定义 $Q_3+1.5R_1$, $Q_1-1.5R_1$ 为数据的**上、下截断点**. 大于上截断点的数据称为特大值, 小于下截断点的数据称为特小值, 特大值和特小值合称为**异常值**(abnormal value). 如果需要, 可以删除异常值后再对数据进行分析.

还有下列数字特征与数据的分散程度有关.

样本校正平方和(corrected sum of squares)

$$\mathrm{CSS}=\sum_{i=1}^{n}(x_i-\overline{x})^2. \tag{1.1.9}$$

样本未校正平方和(uncorrected sum of squares)

$$\mathrm{USS}=\sum_{i=1}^{n}x_i^2. \tag{1.1.10}$$

例 1.1.2 已知例 1.1.1 中 20 名男婴的出生体重 (kg) 资料, 试求分散性的数字特征.

解 输入 R 命令:

```
> w<-c(2.770,2.915,2.795,2.995,2.860,2.970,3.087,3.126,3.125,4.654,
2.272,3.503,3.418,3.921,2.669,4.218,3.707,2.310,2.573,3.881)
> m<-mean(w)
> v<-var(w); v
  #求方差赋值于v, 并输出
> s<-sd(w); s
  #求标准差赋值于s, 并输出
> R<-max(w)-min(w); R
  #计算极差赋值于R, 并输出
> cv<-(s/m); cv
  #计算变异系数赋值于cv, 并输出
> q.quantile=quantile(w)
> Q1=q.quantile[2]
> Q3=q.quantile[4]
> R1<-Q3-Q1; R1
  #计算四分位极差赋值于R1, 并输出
>Qu<-Q3+1.5*R1; Qu
  #计算上截断点赋值于Qu, 并输出
```