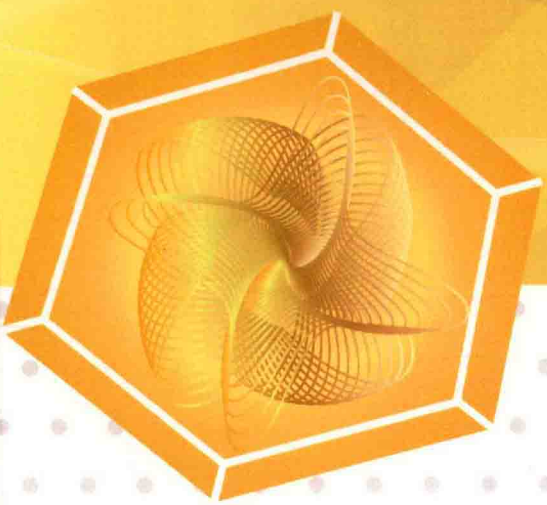
浙江省级重点学科应用数学教学改革与科学研究丛书

数据分析与R软件

（第二版）

李素兰 著



科学出版社

浙江省级重点[redacted]科学研究丛书

数据分析与R软件

（第二版）

李素兰 著

科学出版社
北京

内 容 简 介

本书是《数据分析与 R 软件》第二版，在第一版的基础上，结合信息与计算科学及相关专业的最新培养目标及 R 软件的更新与发展进行了修订.

本次修订对第一版的主要内容进行了调整. 第 1 章只介绍简单一元数据的探索性分析，学生新接触 R 软件，比较容易接受. 第 3 章增加了多元统计分析的相关基础知识，为后面学习多元统计分析方法(回归分析、主成分分析、因子分析、聚类分析、判别分析、相关分析等)奠定理论基础，便于学生理解或证明相关结论. 对 R 软件的使用也增加了一些必要的函数或功能介绍. 同时删除了非参数统计的多样本问题等内容.

本书可作为信息与计算科学、数据科学与大数据技术、应用数学、统计学等专业数据分析类课程的基础教材，也可作为工科硕士研究生统计类课程的基础教材，也可供科研和工程技术人员参考.

图书在版编目(CIP)数据

数据分析与 R 软件/李素兰著. —2 版. —北京: 科学出版社, 2017.8

(浙江省级重点学科应用数学教学改革与科学研究丛书)

ISBN 978-7-03-053158-2

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 (2017) 第 125530 号

责任编辑: 石 悦 李淑丽 / 责任校对: 郭瑞芝

责任印制: 吴兆东 / 封面设计: 华路天然工作室

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

北京教图印刷有限公司印刷

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

2017 年 8 月第 二 版 印张: 17 1/2

2017 年 8 月第二次印刷 字数: 340 000

定价: 42.00 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

第二版前言

R 软件由 2013 年的版本 R2.12.1 升级到版本 R3.3.2, 经历了 R2.12.2, R2.13.0, R2.13.1, R2.13.2,$\cdots$, R3.3.1 等二十六个版本, 可见 R 软件更新之快、发展之快. 这与 R 软件的开源、分享和贡献的理念息息相关. R 软件本身也是一款十分优秀的统计分析软件, 其应用已经变得十分广泛, 从最初的在学术界流行发展到在商业分析领域中的非统计专业的其他分析人员中流行. 近几年, R 软件的书籍和文档也逐渐增多, 信息与计算科学、应用数学、统计学等专业都在为本科生开设相关课程. 本书第一版深受广大读者的欢迎, 也被多所高校选用为数据分析等课程的教材. 有授课教师提出相当宝贵的意见, 部分热心读者也提出很多建设性意见, 本人欣然接受. 结合自己授课过程的经验积累, 决定推出第二版. 这里着重介绍第一版与第二版的不同之处.

首先, 第 1 章只介绍一元数据的探索性分析, 把多元数据的探索性分析放在第 3 章. 因为第 1 章是入门知识, 所以只介绍较简单的一元数据, 学生更容易接受.

其次, 增加了第 3 章多元统计分析相关基础. 后面几章介绍常用的多元统计分析方法, 包括回归分析、主成分分析、因子分析、聚类分析、判别分析和相关分析等. 这些方法理论基础是多维随机向量或多个随机向量放在一起组成的随机矩阵, 以及各种多元统计量的分布及由其导出的分布, 特别是多元正态分布及多元正态分布的导出分布. 为此, 第 3 章介绍多元统计分析相关基础知识. 有了这些知识基础, 学生更容易理解或推导后面涉及的多元统计方法.

再次, 一些具体内容做了一些细节修改. 例如, p 分位数的定义由原来的一种改为常用四种定义, 计算 p 分位数的函数 quantile() 由原来只介绍默认值改为介绍参数的具体用法, 随机模拟增加了非常必要的设置种子函数 set.seed() 等. 还有, 一般统计意义上的定义和 R 中函数运行结果的差异也尽可能增加了说明.

最后, 对 R 软件使用部分也增加了一些必要的函数或功能的介绍, 增加了下载和安装及运行 R 一节, 还增加了获取 R 函数的帮助文档一节, 还有使用 R 的内置数据集等内容. 因为 R 在处理大型数据集方面发展迅速, 所以增加了读取大数据文件和处理大数据 R 包的相关内容.

由于篇幅所限, 删减了非参数的多样本问题等内容.

由于编者水平有限, 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

2017 年 1 月于浙江工业大学

第一版前言

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(exploratory data analysis) 的基本思想是从数据本身出发, 介绍一元数据分析的基本方法, 不拘泥于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, 也可以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, 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.

1.1　数 字 特 征

1.1.1　一元数据的数字特征

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 个数据构成一组样本容量为 n 的样本观测值. $x_1, x_2, \cdots, x_n$ 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 $\overline{x}$, 即

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

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, 则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(quantile of order p) 又称为**百分位数**(percentile), 是中位数的推广, 是反映统计数据一定比例的数据集中位置的数学特征. 大体上整个样本容量的 $(100p)\%$ 观测值不超过 p 分位数, 虽然也是由次序统计量给出, 但给定的方式有多种. 较常用的定义, 譬如:

对于 $0 \leqslant p < 1$, 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 $M_1 = x_{(n)}$.

其他较常见的定义还有

$$M_p = x_{([np]+1)}, \tag{1.1.4}$$

$$M_p = x_{([(n+1)p])}, \tag{1.1.5}$$

$$M_p = x_{([np])} + (n+1)\left(p - \frac{[np]}{n+1}\right)\left(x_{([np]+1)} - x_{([np])}\right). \tag{1.1.6}$$

其他分位数的定义, 在此不一一列举. 虽然有所不同, 但这些分位数相差都不大, 当样本容量 n 无限增大时, 这些差别是微不足道的.

0.5 分位数 $M_{0.5}$ 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排列并分成四等份, 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

$$\frac{1}{4}Q_1 + \frac{1}{2}M + \frac{1}{4}Q_3. \tag{1.1.7}$$

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

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 10 章.

例 1.1.1 调查 20 名男婴的出生体重 (单位: 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
> 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
```

输出结果为

```
> 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
```

因为刚接触 R 语句实现, 所以详细地给出注释和全部输出结果, 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, 3.55400; 三均值计算得 3.106188. 下面详细介绍函数 quantile().

在 R 中, 函数 quantile() 是计算样本分位数的函数. 使用格式如下:

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

其中, x为样本组成数值向量, probs为数值向量, 给定要计算的分位数, 默认值为 0, 0.25, 0.5, 0.75, 1. na.rm为逻辑变量, 当取值为TRUE时, 可处理缺失数据, 默认值为FALSE, 不可处理缺失数据. names为逻辑变量, 当取值为TRUE(默认值) 时, 返回值有百分位数作为向量元素的名称, 即生成一个带有元素名字的向量. type为 1~9 中任何一个整数, 表示计算分位数的算法, 默认值为 7, 具体算法详见 R 函数帮助. 例如

```
> x<-c(2,8,3,7,9,4,1,10,11,12,18)
> quantile(x)
```

输出结果为

```
> quantile(x)
 0%  25%  50%  75% 100%
1.0  3.5  8.0 10.5 18.0
```

修改一些参数, 如type=2 就是式 (1.1.3) 定义的分位数计算方法执行 R 命令.

```
> quantile(x,probs=c(0.25,0.75),names=FALSE,type=2)
```

输出结果为下、上四分位数.

```
> quantile(x,probs=c(0.25,0.75),names=FALSE,type=2)
[1]  3 11
```

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. 它是描述数据分散性的一个重要的数字特征, 计算公式为

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{1.1.8}$$

样本方差作为数据分散程度的度量, 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, 使用样本方差的算术平方根. 样本方差的算术平方根称为**样本标准差**(sample 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. 如果度量单位和均值不同, 比较其变异程度就不能采用标准差. 变异系数可以消除度量单位和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

$$\mathrm{CV} = \left(100 \times \frac{s}{\overline{x}}\right)\%. \tag{1.1.9}$$

(3) 极差.

极差(range) 也称**全距**, 计算公式为

$$R = x_{(n)} - x_{(1)}, \tag{1.1.10}$$

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

$$R_1 = Q_3 - Q_1. \tag{1.1.11}$$

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.

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)

$$\mathrm{CSS} = \sum_{i=1}^{n} (x_i - \overline{x})^2. \tag{1.1.12}$$

样本未校正平方和(uncorrected sum of squares)

$$\mathrm{USS} = \sum_{i=1}^{n} x_i^2. \tag{1.1.13}$$

例 1.1.2 已知例 1.1.1 中 20 名男婴的出生体重 (单位: 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
> Qd<-Q1-1.5*R1; Qd
```

输出结果为

```
> m<-mean(w)
> v<-var(w); v
[1] 0.395825
> s<-sd(w); s
[1] 0.6291462
> R<-max(w)-min(w); R
[1] 2.382
> cv<-(s/m); cv
[1] 0.1973204
> q.quantile=quantile(w)
> Q1=q.quantile[2]
> Q3=q.quantile[4]
> R1<-Q3-Q1; R1
    75%
0.76525
> Qu<-Q3+1.5*R1; Qu
    75%
4.701875
> Qd<-Q1-1.5*R1; Qd
    25%
1.640875
```

从输出结果可以看出, 样本方差 v=0.395825, 标准差 s=0.6291462, 变异系数 cv=0.1973204 或 19.73204%, 计算得四分位极差 R_1=0.76525, 上、下截断点分别为 Q_u=4.701875, Q_d=1.640875.

3. 数据形状的数字特征

偏度系数和峰度系数是刻画数据不对称程度或尾重程度的指标, 定义如下.

(1) 偏度系数.

偏度系数(skewness) 是统计数据分布偏斜方向和程度的度量, 是统计数据分布非对称程度的数字特征. 偏度系数的计算公式为

$$g_1 = \frac{n}{(n-1)(n-2)} \frac{1}{s^3} \sum_{i=1}^{n} (x_i - \overline{x})^3 \tag{1.1.14}$$

$$= \frac{n^2 u_3}{(n-1)(n-2)s^3}, \tag{1.1.15}$$

其中 $u_3 = \dfrac{1}{n} \sum\limits_{i=1}^{n} (x_i - \overline{x})^3$ 是三阶中心矩, s 是标准差. 当数据分布关于均值对称时, 它的所有奇数阶中心矩均为 0, 所以关于均值对称的数据其偏度系数为 0(图 1.1(c)); 右侧更分散的数据偏度系数为正 (图 1.1(b)); 左侧更分散的数据偏度系数为负 (图 1.1(a)).

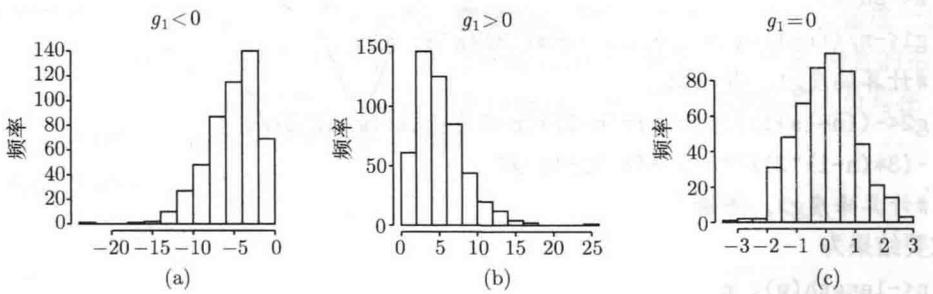


图 1.1 样本偏度系数比较

(2) 峰度系数.

峰度系数(kurtosis) 是用来反映频数分布曲线顶端尖峭或扁平程度的指标. 有时两组数据的算术平均数、标准差和偏度系数都相同, 但它们分布曲线顶端的高耸程度却不同, 峰度系数计算公式为

$$g_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{1}{s^4} \sum_{i=1}^{n} (x_i - \overline{x})^4 - \frac{3(n-1)^2}{(n-2)(n-3)}. \tag{1.1.16}$$

以正态分布为标准, 比较两侧极端数据分布情况. 若数据来自正态总体, 则峰度系数 $g_2 = 0$; 当数据的总体较正态分布尾部更分散时, 峰度系数 $g_2 > 0$; 否则 $g_2 < 0$. 也就是说, 当峰度系数为正时, 数据两侧的极端数据较多 (与正态分布比较), 当峰度系数为负时, 数据两侧的极端数据较少 (与正态分布比较). 从图 1.2 可以看出峰度系数大于零 (图 1.2(a)) 和等于零 (图 1.2(b)) 时极端数据的分布情况的不同.