$$SK_1 = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{n}\sum_i^n (x_i - \overline{x})^3}{\left[\frac{1}{n}\sum_i^n (x_i - \overline{x})^2\right]^{\frac{3}{2}}}$$

式中, m_3 为样本的三阶中心矩; m_2 为样本的二阶中心矩.

算法 2: 对应于 R 的 e1071 包中 skewness 函数的 type=2 是 SPSS, SAS, Excel 软件中的默认算法. 计算公式为:

$$SK_2 = SK_1 \times \frac{\sqrt{n(n-1)}}{n-2} = \frac{m_3}{m_2^{3/2}} \times \frac{\sqrt{n(n-1)}}{n-2}$$

算法 3: 对应于 R 的 e1071 包中 skewness 函数的 type=3 (函数默认算法) 是 Mintab 软件中的默认算法. 计算公式为:

$$SK_3 = \frac{m_3}{s^3} = SK_1\left(\frac{n-1}{n}\right)^{3/2} = \frac{m_3}{m_2^{3/2}}\left(\frac{n-1}{n}\right)^{3/2}$$

式中, s 为样本标准差.

当数据对称分布时, 偏度系数等于 0. 偏度系数越接近 0, 偏斜程度就越低, 也越接近对称分布. 如果偏度系数明显不同于 0, 表示分布是不对称的. 若偏度系数大于 1 或小于 –1, 视为严重偏斜分布; 若偏度系数在 0.5~1 或 –1~–0.5 之间, 视为中等偏斜分布; 若偏度系数小于 0.5 或大于 –0.5, 视为轻微偏斜. 其中负值表示左偏分布 (在分布的左侧有长尾), 正值则表示右偏分布 (在分布的右侧有长尾).

3.3.2 峰度系数

峰度 (kurtosis) 是指数据分布峰值的高低, 这一概念由统计学家 K. Pearson 于 1905 年首次提出. 测度一组数据分布峰值高低的统计量是**峰度系数** (coefficient of kurtosis), 记作 K.

设 $m_r = \frac{1}{n}\sum_i^n (x_i - \overline{x})^r$ 为样本的 r 阶中心距, 峰度系数有以下 3 种算法.

算法 1: 对应于 R 的 e1071 包中 kurtosis 函数的 type=1 也是很多传统教材中的定义. 计算公式为:

$$K_1 = \frac{m_4}{(m_2)^2} - 3$$

算法 2: 对应于 R 的 e1071 包中 kurtosis 函数的 type=2 是 SPSS, SAS, Excel 软件中的默认算法. 计算公式为:

$$K_2 = ((n+1)K_1 + 6) \times \frac{n-1}{(n-2)(n-3)}$$

算法 3: 对应于 R 的 e1071 包中 kurtosis 函数的 type=3 (函数默认算法) 是 Mintab 软件中的默认算法. 计算公式为:

$$K_3 = \frac{m_4}{s^4} - 3 = (K_1 + 3)\left(1 - \frac{1}{n}\right)^2 - 3 \tag{3.17}$$

式中, s 为样本标准差.

峰度通常是与标准正态分布相比较而言的. 由于标准正态分布的峰度系数为 0, 当 $K > 0$ 时, 为尖峰分布, 数据分布的峰值比标准正态分布高, 数据相对集中; 当 $K < 0$ 时, 为扁平分布, 数据分布的峰值比标准正态分布低, 数据相对分散.

文本框 3–11 模拟了不同分布形状对应的偏度系数和峰度系数 (见图 3–2).

文本框 3–11　不同分布形状对应的偏度系数和峰度系数

```
> library(e1071)
> par(mfrow=c(1,3),mai=c(0.7,0.5,0.2,0.1))
> mf<-function(x){
+ hist(x,probability=T,col='lightblue',xlab="x",ylab="Density",
+ main=paste("kurtosis=",round(kurtosis(x),digits=4)),
+ sub=paste("skewness=",round(skewness(x),digits=4)))
+ lines(density(x),col='red',lwd=2)
+ }
> n<-5000
> mf(rchisq(n,10))
> mf(rnorm(n))
> mf(-rchisq(n,10)+36)
```

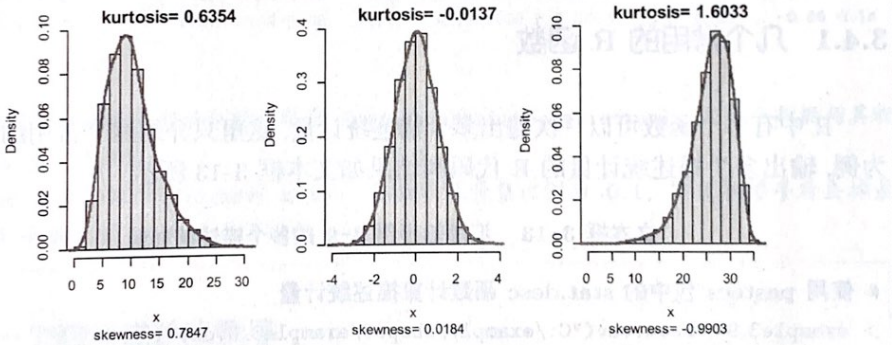

图 3–2　不同分布形状对应的偏度系数和峰度系数

注:每次运行上述代码都会得到略有不同的分布形状和偏度系数及峰度系数, 读者可以反复进行模拟.

【例 3–11】 (数据: example3_1.csv) 沿用例 3–1. 计算 30 名学生考试分数的偏度系数和峰度系数.

解: 计算偏度系数和峰度系数的 R 代码和结果如文本框 3–12 所示.

文本框 3–12　计算 30 名学生考试分数的偏度系数和峰度系数

```
# 计算偏度系数
> example3_1<-read.csv("C:/example/chap03/example3_1.csv")
```

```
> library(e1071)
> skewness(example3_1$分数,type=3)                    # 函数默认 type=3
[1] -0.7500727

# 计算峰度系数
> kurtosis(example3_1$分数,type=3)                    # 默认 type=3
[1] -0.6537093
```

注：偏度系数和峰度系数有不同的计算方法，R 的 e1071 包中提供了各种方法的介绍. 加载 e1071 后查看帮助可得到相关信息. agricolae 包、DescTools 包中也有相应的计算函数，使用时注意，不同函数的默认算法可能是不同的. 更多信息查阅相应函数的帮助.

文本框 3–12 中的结果显示, 30 名学生考试分数的偏度系数为 –0.750 072 7, 表示考试分数的分布为中等程度的左偏分布. 峰度系数为 –0.653 709 3, 表示考试分数分布的峰值比标准正态分布的峰值要略低一些.

3.4 数据的综合描述

本节首先介绍两个常用的综合描述的 R 函数, 然后结合第 2 章的内容给出一个综合描述性分析的例子.

3.4.1 几个常用的 R 函数

R 中有多个函数可以一次输出多个描述统计量, 这里只介绍两个常用的. 以例 3-9 为例, 输出多个描述统计量的 R 代码和结果如文本框 3–13 所示.

文本框 3-13 汇总输出例 3-9 的多个描述统计量

```
# 使用 pastecs 包中的 stat.desc 函数计算描述统计量
> example3_9<-read.csv("C:/example/chap03/example3_9.csv")
> library(pastecs)
> round(stat.desc(example3_9),4)
```

	纳塔利娅·帕杰林娜	郭文珺	卓格巴德拉赫·蒙赫珠勒	妮诺·萨卢克瓦泽
nbr.val	10.0000	10.0000	10.0000	10.0000
nbr.null	0.0000	0.0000	0.0000	0.0000
nbr.na	0.0000	0.0000	0.0000	0.0000
min	8.5000	9.4000	8.3000	9.1000
max	10.6000	10.8000	10.7000	10.8000
range	2.1000	1.4000	2.4000	1.7000
sum	98.1000	102.3000	92.6000	101.4000
median	9.9000	10.3500	9.2000	10.2500
mean	9.8100	10.2300	9.2600	10.1400
SE.mean	0.1946	0.1383	0.2237	0.1727
CI.mean.0.95	0.4403	0.3128	0.5061	0.3907
var	0.3788	0.1912	0.5004	0.2982

```
std.dev          0.6154          0.4373          0.7074          0.5461
coef.var         0.0627          0.0427          0.0764          0.0539
                维多利亚.柴卡  莱万多夫斯卡.萨贡  亚斯娜.舍卡里奇  米拉.内万苏
nbr.val         10.0000         10.0000         10.0000         10.0000
nbr.null         0.0000          0.0000          0.0000          0.0000
nbr.na           0.0000          0.0000          0.0000          0.0000
min              8.6000          8.1000          9.1000          8.7000
max             10.5000         10.7000         10.2000         10.3000
range            1.9000          2.6000          1.1000          1.6000
sum             98.0000         97.3000         96.9000         96.5000
median           9.9500          9.8500          9.8000          9.7500
mean             9.8000          9.7300          9.6900          9.6500
SE.mean          0.2055          0.2319          0.1130          0.1462
CI.mean.0.95     0.4648          0.5246          0.2556          0.3308
var              0.4222          0.5379          0.1277          0.2139
std.dev          0.6498          0.7334          0.3573          0.4625
coef.var         0.0663          0.0754          0.0369          0.0479
```

 # 使用 psych 包中的 describe 函数计算描述统计量

```
> library(psych)
> describe(example3_9)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
纳塔利娅.帕杰林娜	1	10	9.81	0.62	9.90	9.88	0.52	8.5	10.6	2.1	-0.65	-0.45	0.19
郭文珺	2	10	10.23	0.44	10.35	10.26	0.44	9.4	10.8	1.4	-0.50	-1.04	0.14
卓格巴德拉赫.蒙赫珠勒	3	10	9.26	0.71	9.20	9.20	0.59	8.3	10.7	2.4	0.54	-0.65	0.22
妮诺.萨卢克瓦泽	4	10	10.14	0.55	10.25	10.19	0.59	9.1	10.8	1.7	-0.52	-1.10	0.17
维多利亚.柴卡	5	10	9.80	0.65	9.95	9.86	0.82	8.6	10.5	1.9	-0.42	-1.35	0.21
莱万多夫斯卡.萨贡	6	10	9.73	0.73	9.85	9.81	0.67	8.1	10.7	2.6	-0.79	-0.12	0.23
亚斯娜.舍卡里奇	7	10	9.69	0.36	9.80	9.70	0.30	9.1	10.2	1.1	-0.29	-1.46	0.11
米拉.内万苏	8	10	9.65	0.46	9.75	9.69	0.30	8.7	10.3	1.6	-0.65	-0.66	0.15

注:

1. 输出结果中，mad 是中位数绝对离差 (median absolute deviation)，即每个数据与其相应中位数的离差平均数.

2. trimmed 是修整均值 (trimmed mean)，参数默认修整比例为 0.1，即数据排序后去掉最小的 10% 和最大的 10% 后计算均值.

3.4.2　一个综合描述的例子

在实际分析中, 通常要对数据从图表和统计量两个方面同时进行描述. 下面结合第 2 章的内容, 通过一个例子说明对数据进行综合描述的基本思路.

【例 3-12】 (数据: example3_12.csv) 在某大学随机抽取 60 名大学生, 调查得到他们的性别、家庭所在地和月生活费支出 (单位: 元) 数据如表 3-4 所示. 对调查数据进行综合分析.

表 3-4　60 名大学生的调查数据

性别	家庭所在地	月生活费支出	性别	家庭所在地	月生活费支出
女	中小城市	1 500	女	乡镇地区	1 850

文本框 3-15　使用 summary 函数对性别和家庭所在地计数、对月生活费支出计算描述统计量

```
# 使用 summary 函数对性别和家庭所在地计数、对月生活费支出计算统计量
> example3_12<-read.csv("C:/example/chap03/example3_12.csv")
> summary(example3_12)
  性别     家庭所在地   月生活费支出
 男:25   大型城市:26   Min.   :1100
 女:35   乡镇地区:10   1st Qu.:1550
         中小城市:24   Median :1850
                       Mean   :1812
                       3rd Qu.:2000
                       Max.   :2800
# 分别按性别和家庭所在地分类计算描述统计量
> my_summary<-function(x){
+ with(x,data.frame(
+ n=length(月生活费支出),
+ "平均数"=mean(月生活费支出),
+ "中位数"=median(月生活费支出),
+ "标准差"=sd(月生活费支出),
+ "全距"=max(月生活费支出)-min(月生活费支出),
+ "变异系数"=sd(月生活费支出)/mean(月生活费支出),
+ "偏度系数"=e1071::skewness(月生活费支出)))
+ }
>  library(plyr)
>  ddply(example3_12,.(性别),my_summary)
   性别   n    平均数  中位数     标准差  全距    变异系数     偏度系数
1   男  25 1701.200    1780 275.4893   900 0.1619382 -0.4845139
2   女  35 1891.714    1900 331.1521  1500 0.1750539  0.4605462

> ddply(example3_12,.(家庭所在地),my_summary)
   家庭所在地   n    平均数  中位数     标准差  全距    变异系数     偏度系数
1  大型城市  26 1848.846    1850 364.1354  1700 0.1969528  0.2849817
2  乡镇地区  10 1757.000    1860 236.0344   700 0.1343394 -0.7580232
3  中小城市  24 1795.833    1800 308.6565  1060 0.1718737  0.2366286
# 同时按性别和家庭所在地分类计算描述统计量
> library(doBy)
> myfun<-function(x){
+    c(n=length(x),mean=mean(x),median=median(x),
+    sd=sd(x),CV=sd(x)/mean(x),R=(max(x)-min(x)),SK=e1071::skewness(x))}
> summaryBy(月生活费支出~性别+家庭所在地,data=example3_12,FUN=myfun)
```