数据挖掘与R软件实战演练
中级课程

主讲人：程豪

SSstudy.com
科学软件学习网

第三章 R软件数据描述性分析

SSstudy.com 科学软件学习网

第一节 R软件数据描述性分析（一）

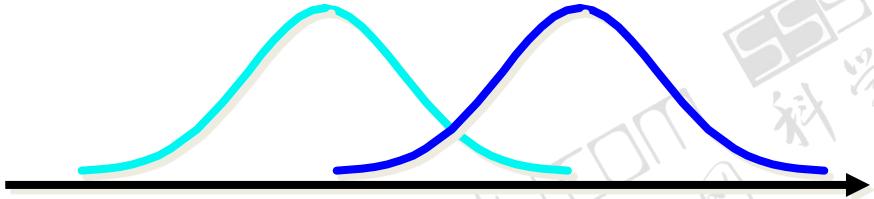
- 何为统计描述性分析
- R数据描述性分析
- R语言绘图

SSstudy.com 科学软件学习网

数据特征度量

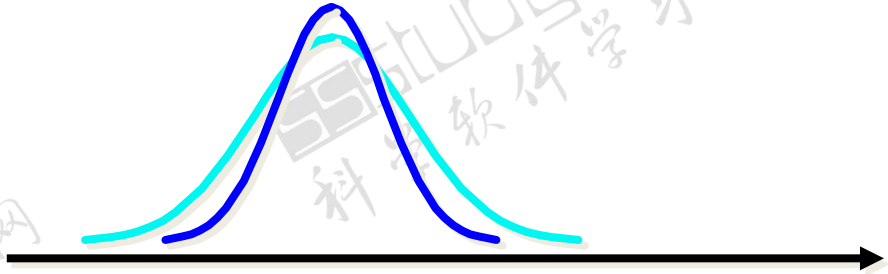
集中趋势

(位置)



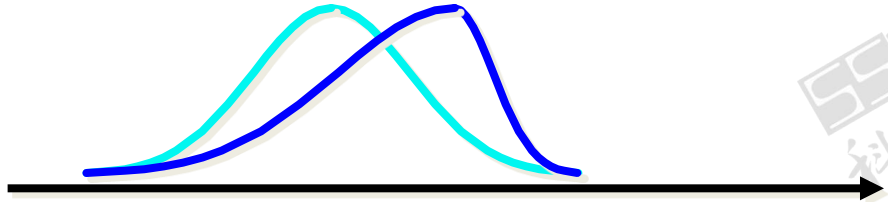
离中趋势

(分散程度)



偏态和峰态

（形状）



集中趋势(central tendency)

1. 一组数据向其中心值靠拢的倾向和程度。
2. 测度集中趋势就是寻找数据水平的代表值或中心值。
3. 不同类型的数据用不同的集中趋势测度值。

顺序数据：中位数和分位数

数值型数据：平均数

集中趋势(central tendency)

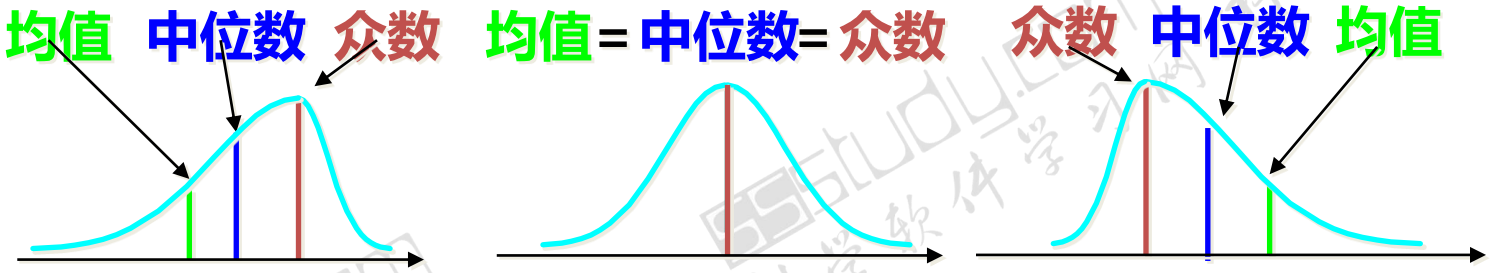
4. 低层次数据的测度值适用于高层次的测量数据，但高层次数据的测度值并不适用于低层次的测量数据。

即：集中趋势是平均数或中位数或众数,低层次数据是指定性数据,如性别,它的集中趋势只能是众数或中位数.高层次数据是定量数据,如身高,它的集中趋势是平均数.不能用平均数去测度性别的集中趋势.

分类数据：众数

1. 一组数据中出现次数最多的变量值；
2. 适合于数据量较多时使用；
3. 不受极端值的影响；
4. 一组数据可能没有众数或有几个众数；
5. 主要用于分类数据，也可用于顺序数据和数值型数据。

众数、中位数和平均数的比较



左偏分布　　对称分布　　右偏分布

众数、中位数、平均数的特点和应用

1. 众数
   - 不受极端值影响
   - 具有不惟一性
   - 数据分布偏斜程度较大且有明显峰值时应用
2. 中位数
   - 不受极端值影响
   - 数据分布偏斜程度较大时应用
3. 平均数
   - 易受极端值影响
   - 数学性质优良
   - 数据对称分布或接近对称分布时应用

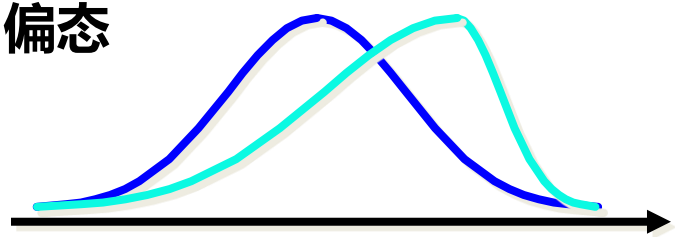
众数、中位数、平均数的特点和应用

数据类型和所适用的集中趋势测度值				
数据类型	分类数据	顺序数据	间隔数据	比率数据
适用的测度值	※众数	※中位数	※平均数	※平均数
	—	四分位数	众数	几何平均数
	—	众数	中位数	中位数
	—	—	四分位数	四分位数
	—	—	—	众数

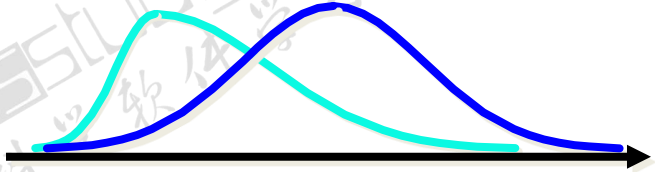
峰态与偏态的度量

与标准正态分布比较！

偏态

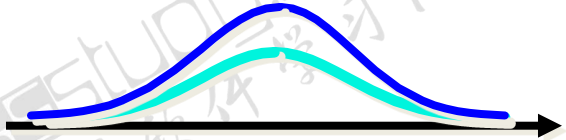


左偏分布

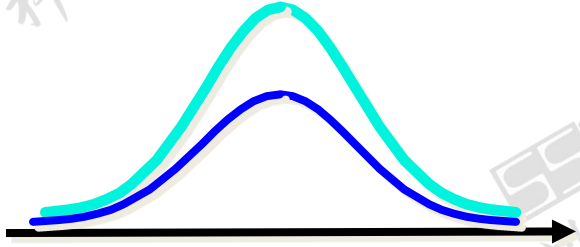


右偏分布

峰态



扁平分布



尖峰分布

离散程度的度量

分类数据：

异众比率

顺序数据：

四分位差

数值型数据：

方差和标准差

相对离散程度：

离散系数

1. 数据分布的另一个重要特征；
2. 反映各变量值远离其中心值的程度(离散程度)；
3. 从另一个侧面说明了集中趋势测度值的代表程度；
4. 不同类型的数据有不同的离散程度测度值。

第二节 R软件数据描述性分析（二）

- 何为统计描述性分析
- R数据描述性分析
- R语言绘图

SSstudy.com 科学软件学习网

R数据描述性分析

- R的统计分析分为统计描述和统计推断两部分。统计描述是通过绘制统计图形、编制统计表、计算统计量等方法来表述数据的分布特征。
- 描述统计量包括了均值、中位数、次序统计量、百分数、方差、标准差、变异系数、极差、偏度系数等，是数据的位置度量、分散程度和分布形状的体现。
- 还包括分布函数、直方图、经验分布图、QQ图、茎叶图、箱线图等等。

- 均值(mean)的基本用法是
- mean(x, trim=0, na.rm=FALSE);
- 其中x是要计算均值的那个量，trim是计算均值前，去掉x两端观测值的比例，na.rm如果是TRUE，则表示删除NA再计算均值，允许缺失数据。
- x<-c(75, 76, 77); mean(x)

  [1] 76
- 若x是个矩阵，则mean(x)返回矩阵所有元素的均值。
- x<-1:12
- dim(x)=c(3,4)

  mean(x)

  [1] 6.5

- 若要分别求矩阵的行和列，则要用apply()函数
- apply(x, 1, mean)

  [1] 5.5 6.5 7.5
- apply(x, 2, mean)

  [1] 2 5 8 11
- 若x是个date frame, 则返回的是各列的平均值
- mean(as.data.frame(x))

  V1 V2 V3 V4

  2 5 8 11
- 因此多元数据的输入采用数据框的形式，会便于后期数据的处理

- w<-c(75, 64, 47.4, 66.9, 62.2, 62.2, 58.7, 63.5, 66.6, 64,57, 69, 56.9, 50, 72)
- w.mean <- mean(w, trim=0.1); w.mean

  [1] 62.53846
- trim的取值在0.1-0.5之间，可以消除极端值对均值的影响。
- 若数据当中含有缺失值NA时，可以加na.rm来处理
- w.na<-c(75, 64, 47.4, 66.9, 62.2, 62.2, 58.2, 63.5, 66.6, 64,57, 69, 56.9, 50, 72, NA);

  mean(w.na);

  NA

  w.na.mean<-mean(w.na, na.rm=TRUE);

  [1] 62.36

- 若要计算数据的加权平均，可以用weighted.mean()函数，其基本格式为：

  weighted.mean(x, w, na.rm=FALSE)

- 其中w是数据x的权重系数，其维数与x相同，基本用法与mean()相同，唯一有区别的地方是：

  weighted.mean()不适用于数据框，它作用在数据框的时候，和作用于矩阵的时候，结果是一样的，返回全部数据的加权平均

- 另外，对向量就平均等价于sum(x)/length(x), sum()的用法和mean类似，只不过前者是求和

- 顺序统计量sort()的基本用法如下：

  sort(x, partial=NULL, na.last=NA, decreasing=FALSE, method=c("shell", "quick"), index.return=FALSE)

- partial是部分排序的指标向量；na.last是处理缺失数据的，默认值为NA，表示不处理缺失数据，若na.last=TRUE，则将缺失数据放在最后；decreasing是升序或降序的选择，默认值FALSE表示按从小到大的升序排列，否则降序排列；method是排序方法，默认值是"shell"; index.return返回排序下标(只有na.last=NA才可以用index.return这个参数)，默认值是FALSE，若TRUE，返回的是一个列表，列表的第一个变量$x是排序的顺序，第二个变量$ix是排序顺序对应的下标值。

- x<-c(75, 64, 47.4, 66.9, 62.2, 62.2, 58.7, 63.5)
- sort(x, decreasing=TRUE)

  [1] 75 66.9 64.0 63.5 62.2 62.2 58.7 47.4
- 若数据当中含有缺失值NA时，可以加na.rm来处理
- x.na<- c(75, 64, 47.4, NA, 66.9, 62.2, 62.2, 58.7, 63.5)

  sort(x.na);

  [1] 47.4 58.7 62.2 62.2 63.5 64.0 66.9 75.0

  sort(x.na, na.last=T)

  [1] 47.4 58.7 62.2 62.2 63.5 64.0 66.9 75.0 NA

  sort(x.na, na.last=F)

  [1] NA 47.4 58.7 62.2 62.2 63.5 64.0 66.9 75.0

- x<-c(75, 64, 47.4, 66.9, 62.2, 62.2, 58.7, 63.5)
- sort(x, index.return=T)

```
$x
[1] 47.4 58.7 62.2 62.2 63.5 64.0 66.9 75.0
$ix
[1] 3 7 5 6 8 2 4 1
```

- 中位数函数median()的基本格式为
- median(x, na.rm=FALSE)
- median(x.na)

```
[1] NA
median(x.na, na.rm=T)
[1] 62.85
```

- 百分数是中位数的推广，将数据按从小到大的顺序排列后，取p分位数，若np是整数，则取第np和第np加一个数的平均；若np不是整数，取第[np]+1那个数。
- 计算百分数要用到的函数是quantile()，基本格式为：

  quantile(x, probs=seq(0,1,0.25), na.rm=FALSE, names=TRUE, type=7, ...)
- probs给出相应的百分位数，默认值是0，0.25，0.5，0.75，1；na.rm是处理缺失数据的，na.rm=TRUE时，NA和NaN将从数据中移走，向量取值中若有NA或NaN，要添加这一参数，否则会出错；names若为TRUE，返回值当中有names这个属性"; type是取值1-9的整数，选择了九种分位数算法(具体算法见帮助文件)中的一种。

- w.quantile <- quantile(w); w.quantile

```
  0%     25%     50%     75%    100%
47.40  57.85    65.50    66.75   75.00
```

- attributes(w.quantile)

```
$names
[1] "0%"   "25%"   "50%"   "75%"   "100%"
```

- quantile(w, probs=seq(0,1, 0.2))

```
    0%    20%    40%    60%    80%    100%
47.4   56.98   62.20   64.00   67.32   75.00
```

数据的分布

- 数据的分布主要考察分布函数(p), 密度函数(d), 分位数函数(q)及产生随机数(r).
- 以正态分布为例：

$z_{\alpha/2} = \text{qnorm}(1-0.025,\ 0,1) = 1.959964.$

- data<-rnorm(100, mean=0, sd=1);

dnorm(data, mean=0, sd=1, log=F);

pnorm(data, mean=0, sd=1, lower.tail=T, log.p=F);

p<-c(0.975, 0.95)

qnorm(p, mean=0, sd=1, lower.tail=T, log.p=F);

[1] 1.959964 1.644854 # 0.05/2, 0.1/2分位数

分布	R 中的名称	附加参数
beta	beta	shape1, shape2, ncp
binomial	binom	size, prob
Cauchy	cauchy	location, scale
chi-squared	chisq	df, ncp
exponential	exp	rate
F	f	df1, df2, ncp
gamma	gamma	shape, scale
geometric	geom	prob
hypergeometric	hyper	m, n, k
log-normal	lnorm	meanlog, sdlog
logistic	logis	location, scale
negative binomial	nbinom	size, prob
normal	norm	mean, sd
Poisson	pois	lambda
Student's t	t	df, ncp
uniform	unif	min, max
Weibull	weibull	shape, scale
Wilcoxon	wilcox	m, n

第三节 R软件数据描述性分析（三）

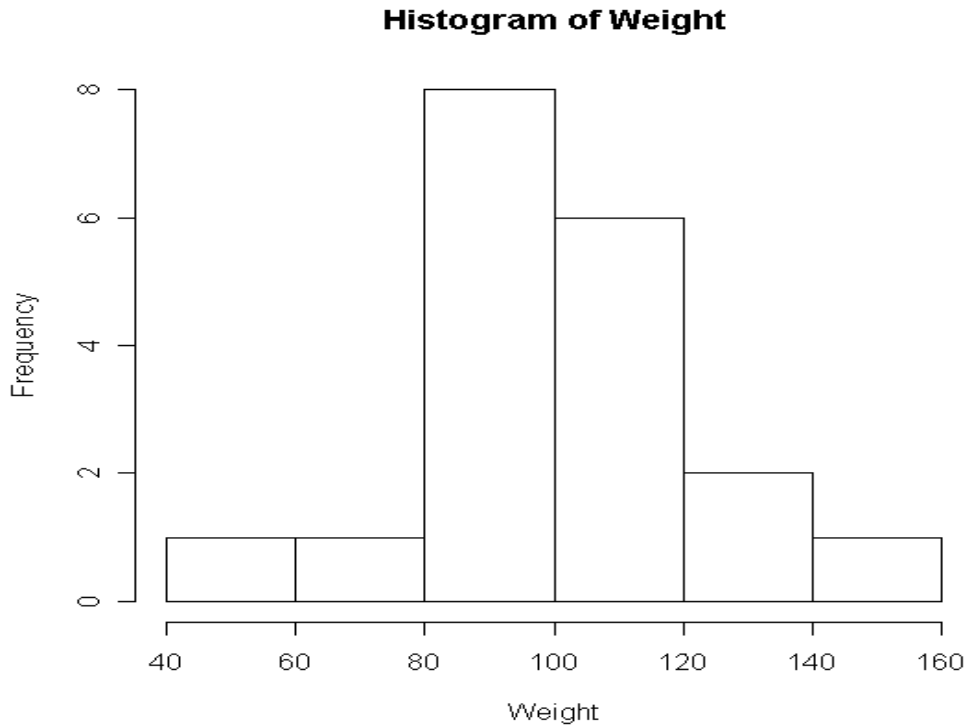
- 何为统计描述性分析
- R数据描述性分析
- R语言绘图

SSstudy.com 科学软件学习网

直方图、经验分布图与QQ图

- cl<-read.table("chapter4-cl.txt", header=T);
- 用hist()函数可以绘制直方图。
- hist的一般用法为：
- hist(x, breaks="Sturges", freq=NULL, probability=!freq,...)
- break规定了直方图的组距(必须覆盖数据的范围)；freq是逻辑变量，TRUE是频率直方图，FALSE是密度直方图；probability和freq相反，TRUE是密度直方图，FALSE是频率直方图。其他参数详见帮助文档。

> hist(Weight)

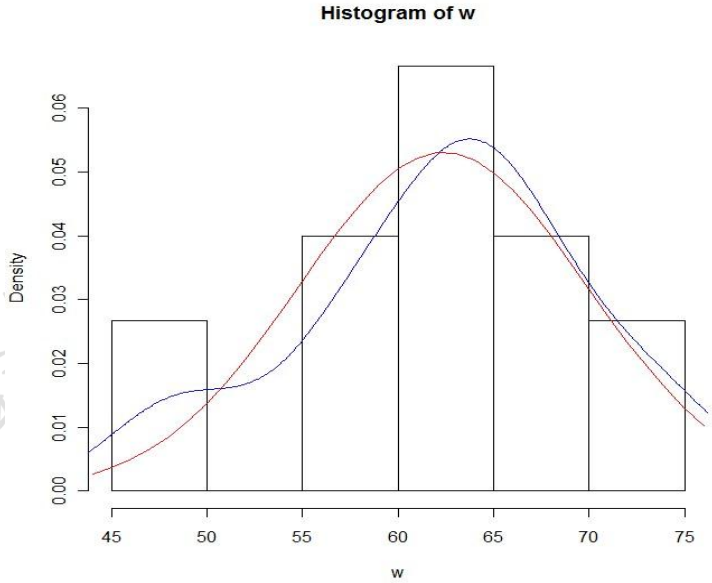

Histogram of Weight
Frequency
0
2
4
6
8
40
60
80
100
120
140
160
Weight

- 用density()函数可以绘制与直方图配套的核密度估计。
- density的一般用法为：
- density(x, bw="nrd0", adjust=1, kernel=c("gaussian", "..."), window=kernel, width...)
- bw是带宽，默认值R画出光滑图形；kernel是核函数；adjust表示实际带宽是adjust*bw。其他参数详见帮助文档。

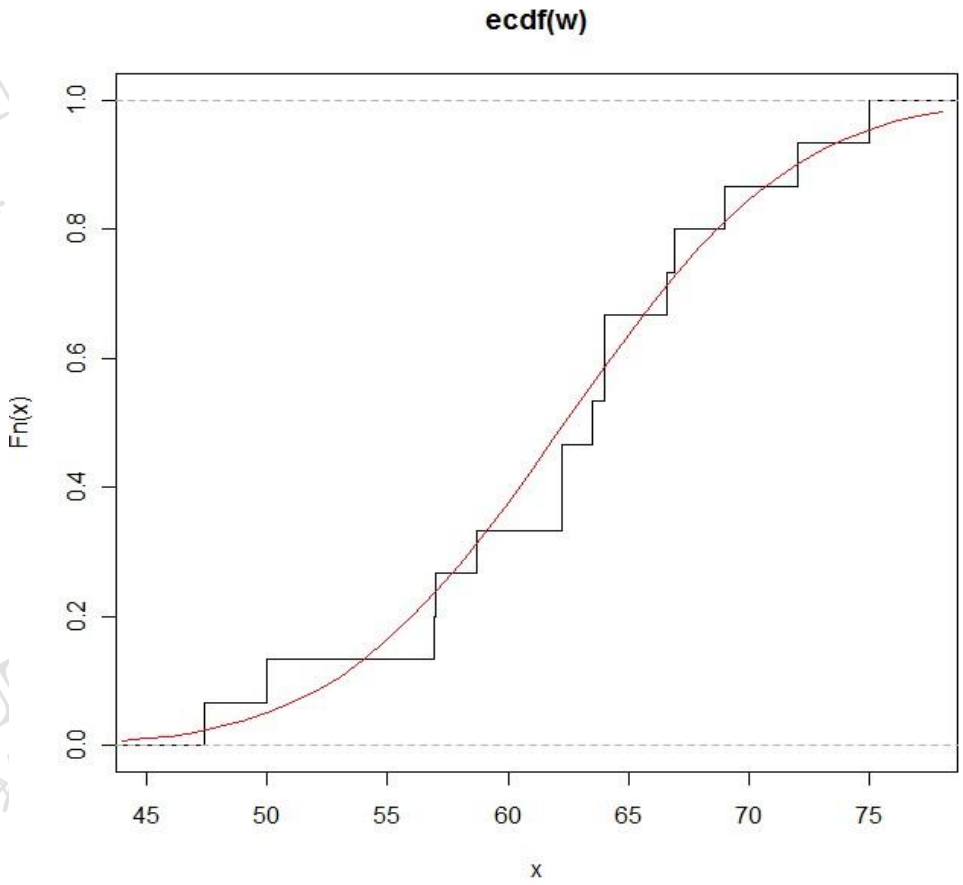
- w<-c(75, 64, 47.4, 66.9, 62.2, 62.2, 58.7,
  63.5, 66.6, 64,57, 69, 56.9, 50, 72)
- hist(w, freq=F);
- w.density <- density(w); w.density

```
Call:
        density.default(x = w)
Data: w (15 obs.);      Bandwidth 'bw' = 3.478
       x                y
 Min.   :36.97   Min.   :9.044e-05
 1st Qu.:49.08   1st Qu.:4.402e-03
 Median :61.20   Median :1.603e-02
 Mean   :61.20   Mean   :2.061e-02
 3rd Qu.:73.32   3rd Qu.:3.409e-02
 Max.   :85.43   Max.   :5.518e-02
```

- lines(w.density, col="blue");
- x<- 44:76;
- lines(x, dnorm(x, mean(w), sd(w)), col="red");

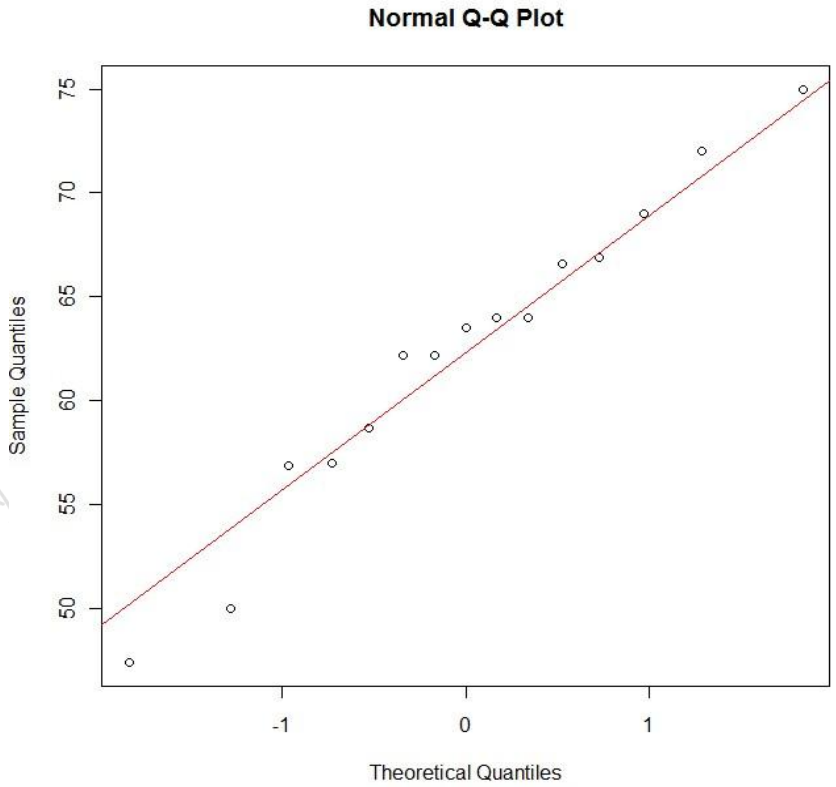


- 经验分布函数ecdf()可以估计总体的分布函数，一般用法为：
- ecdf(x)
- 若要在R中画出经验分布函数，则用plot函数：
- plot(ecdf(x), ylab="Fn(x)", verticals=FALSE, col.01line="gray70")
- verticals是逻辑变量，TRUE时表示画竖线，否则不画竖线；col.01line是0-1线的颜色。
- plot(ecdf(w), verticals=TRUE, do.p=FALSE)；
  x<-44:78;
  lines(x, pnorm(x, mean(w), sd(w)))


ecdf(w)
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
45
50
55
60
65
70
75
x

- QQ图是用来鉴别样本的分布是否近似于某种类型的分布
- qqnorm()和qqline()提供了画正态QQ图和相应直线的方法
- qqnorm(y, ylim, xlab=" ", ylab=" ", plot.it=TRUE, datax=FALSE)
- plot.it是逻辑变量，TRUE时将结果画出来；datax是将样本数据放x轴，默认值是FALSE。
- qqplot(x, y, plot.it=TRUE)；
- 比较x和y的分布接近程度

qqnorm(w)
qqline(w)



茎叶图、箱线图及五数总括

- 茎叶图stem()可以细致地看出数据分布的结构。
- stem()的一般用法为：
- > stem(x, scale=1, width=80, atom=1e-08)
- scale控制了茎叶图的长度，默认值是1，如果scale=2，则表示将0-9这10个个位数分成两段，0~4为一段，5~9为一段；width是绘图的宽度；atom是容差，一般选择默认值即可。

```
stem(x, scale=2)
 The decimal point is 1 digit(s) to the right of the |
  2 | 5
  3 |
  3 |
  4 |
  4 | 5
  5 | 04
  5 | 5
  6 | 14
  6 | 8
  7 | 2
  7 | 5589
  8 | 13444
  8 | 5667999
  9 | 0112
  9 |
 10 | 0
```

- x<-c(25, 45, 50, 54, 55, 61, 64, 68, 72, 75, 75, 78, 79,
  81, 83, 84, 84, 84, 85, 86, 86, 87, 89, 89, 89, 90,
  91, 91, 92, 100)
- stem(x);

The decimal point is 1 digit(s)
to the right of the |

2 | 5
3 |
4 | 5
5 | 045
6 | 148
7 | 25589
8 | 134445667999
9 | 0112
10 | 0

- stem(x, scale=0.5); # scale也可以是小数，等于0.5时，表示将0-9这10个个位数分成1/2段，即20个数为一段

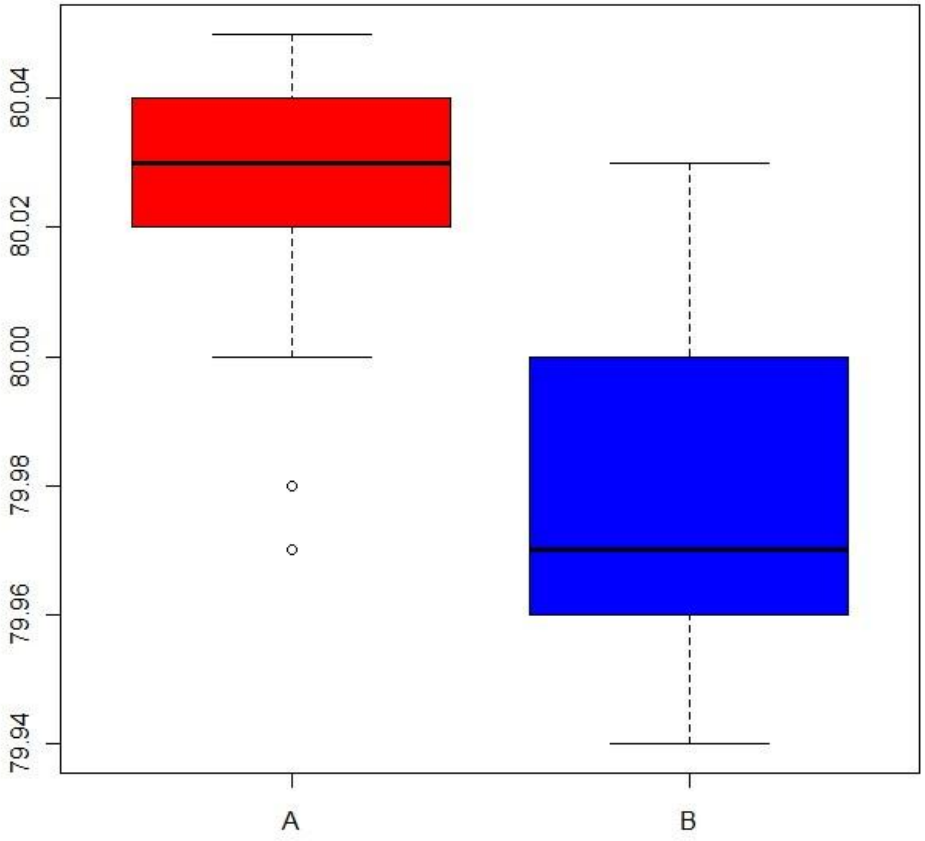
The decimal point is 1 digit(s) to the right of the |

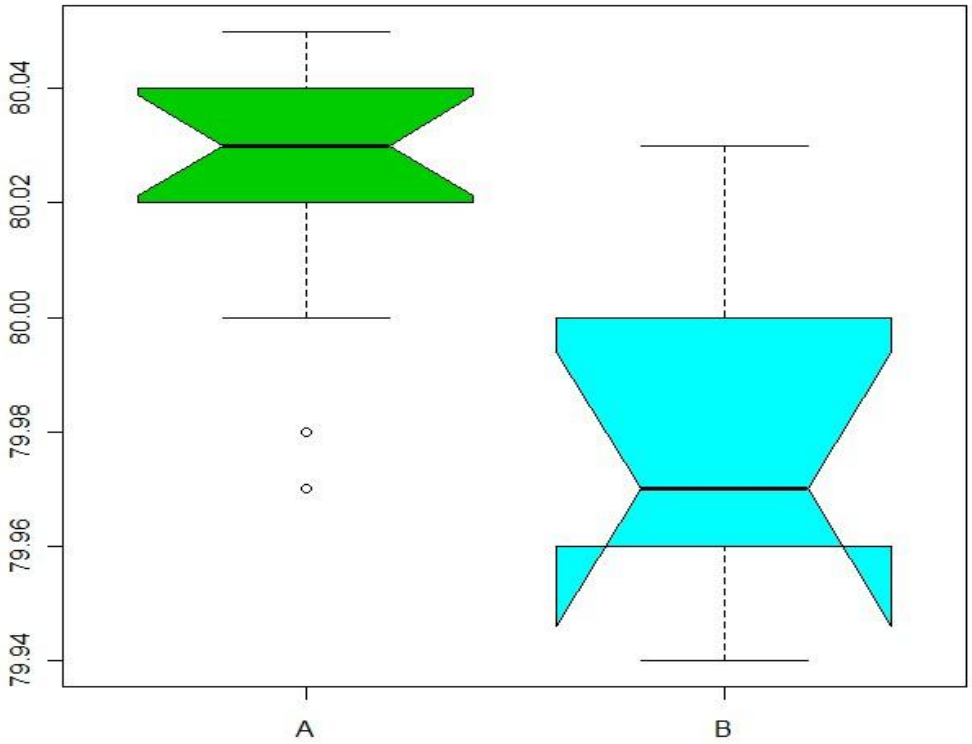
```
 2 | 5
 4 | 5045
 6 | 14825589
 8 | 1344456679990112
10 | 0
```

- 箱线图boxplot()直观地展现数据分布的主要特征。
- boxplot()有三种基本的用法：
- > boxplot(x, ...)
- > boxplot(x, ..., range=1.5, width, varwidth, notch=FALSE, outline=TRUE, ..., horizontal=FALSE, add=FALSE, at=NULL)
- > boxplot(formula, data, ..., subset, na.action=NULL);

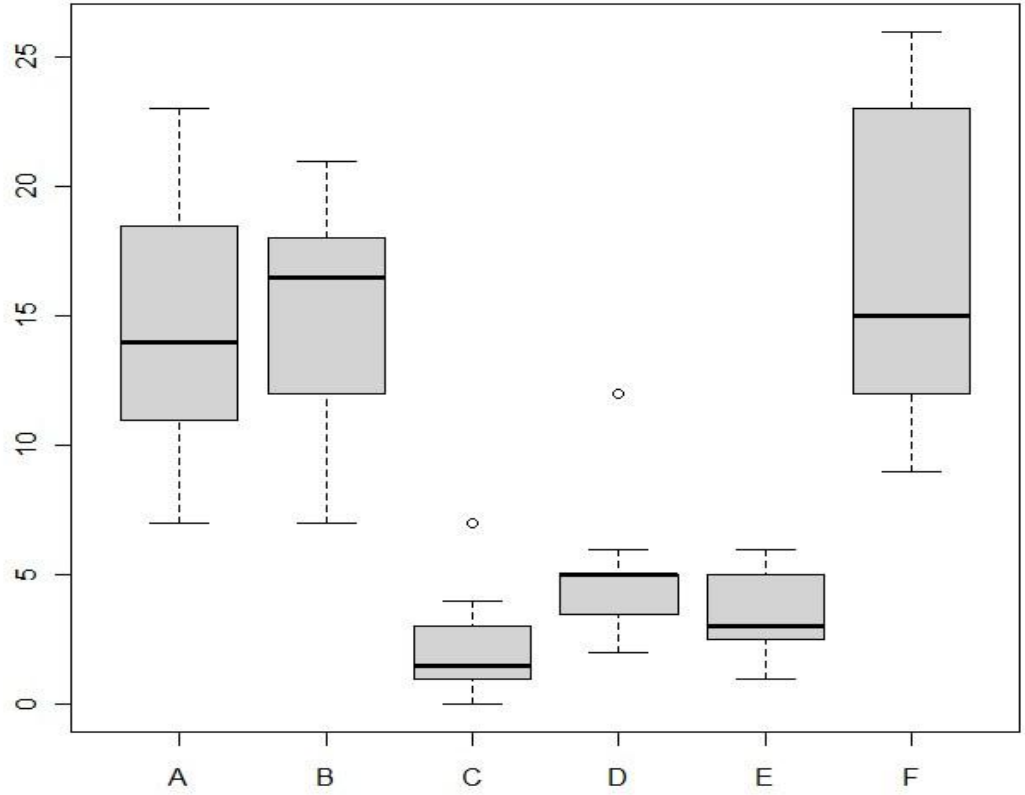
- x是数据构成的数值型向量；range控制了“触须”的范围(默认值1.5)；notch=TRUE时，箱线图带有切口； outline是逻辑变量，TRUE时标出异常点；horizontal是逻辑变量，TRUE表示把箱线图绘制成水平状(默认值为FALSE)； add是逻辑变量，TRUE时表示在原图上画图，否则替换一张图(默认值为FALSE)。
- 若用最后一种形式，fomular是公式；data给出了公式作用的对象；subset是可选参数，可以给定要绘制的数据子集；na.action表示对NA数据作出处理，默认值为NULL，即忽略NA数据。

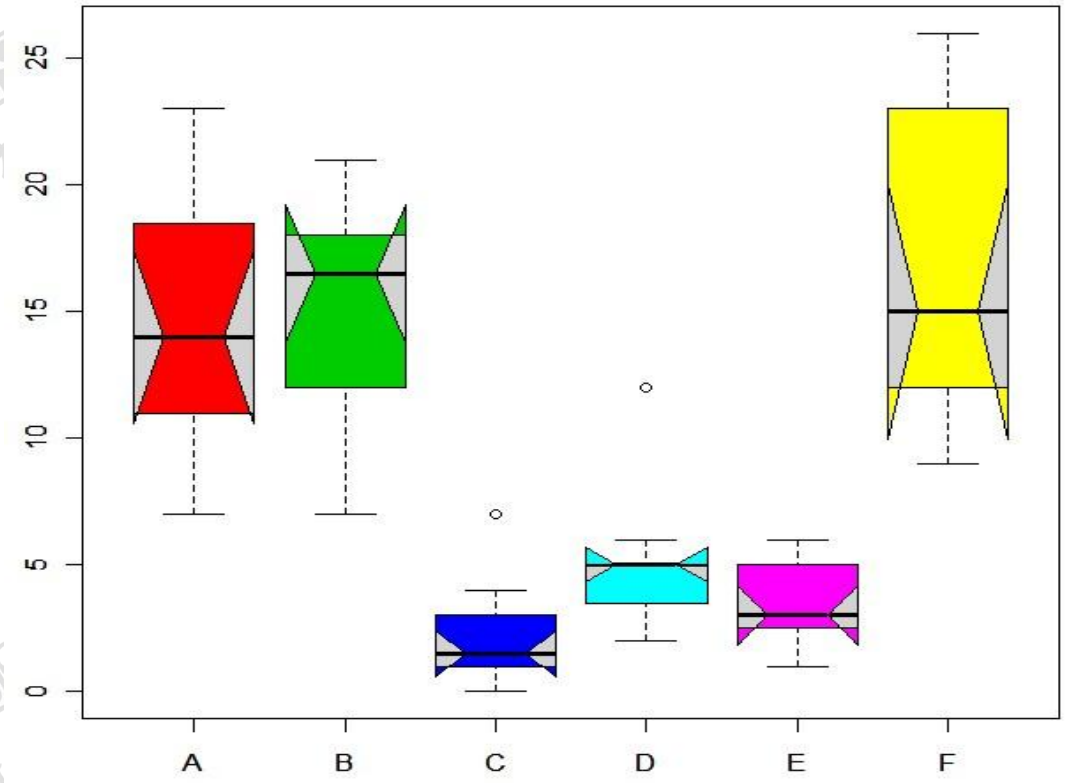
- A<-c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97, 80.05, 80.03, 80.02, 80.00, 80.02);
- B<-c(80.02, 79.94, 79.98, 79.97, 80.03, 79.95, 79.97)
- boxplot(A, B, names=c("A", "B"), col=c("red", "blue"));
- boxplot(A, B, notch=T, outline=T, names=c("A", "B"), col=c(3,5));


80.04
80.02
80.00
79.98
79.96
79.94
A
B


80.04
80.02
80.00
79.98
79.96
79.94
A
B

- InsectSprays;
- boxplot(count~spray, data=InsectSprays, col="lightgray")
- boxplot(count~spray, data=InsectSprays, notch=T, col=2:7, add=T)


0
5
10
15
20
25
A
B
C
D
E
F


0
5
10
15
20
25
A
B
C
D
E
F

正态性检验与分布拟合

- 前面介绍的茎叶图、箱线图直观地现实了数据的分布情况，直方图、经验分布图、QQ图配合了总体的分布或者密度曲线，可以考察总体和正态分布的接近程度。
- 然而，究竟所配的曲线是否合适，这种接近程度如何给出统计分析，还需要做正态性检验。
- 这里将简单地介绍两种正态性检验方法：Shapiro-Wilk W检验和Kolmogorov-Smirnov经验分布拟合检验。

Shapiro-Wilk W检验

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

where

- $x_{(i)}$ (with parentheses enclosing the subscript index i) is the ith order statistic, i.e., the ith-smallest number in the sample;
- $\overline{x} = (x_1 + \ldots + x_n) / n$ is the sample mean;
- the constants a_i are given by[2]

$$(a_1, \ldots, a_n) = \frac{m^\top V^{-1}}{(m^\top V^{-1} V^{-1} m)^{1/2}}$$

where

$$m = (m_1, \ldots, m_n)^\top$$

and $m_1, \ldots, m_n$ are the expected values of the order statistics of independent and identically-distributed random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics.

- R中调用shapiro.test()函数即可使用该检验，使用格式为：

shapiro.test(x)

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/936012033011010132