第5章

数 据 分 析

5.1 数据的特征量

5.1.1 随机变量的数字特征

数据分析是科学研究中的常用方法。MATLAB 中提供的数据分析常用函数如表 5-1 所示。其中部分函数的功能是求取随机变量的数字特征，即与随机变量有关的某些数值，它们虽然不能完整地描述随机变量，但是能描述随机变量在某些方面的重要特征。

表 5-1 数据分析常用函数

函数	功 能	函数	功 能
max	求最大元素	mean	求算术平均值
min	求最小元素	median	求中值
sum	求和	cumsum	求累加和
prod	求积	cumprod	求累乘积
var	求方差	std	求标准差
cov	求协方差	corrcoef	求相关系数

求向量中的最大元素可调用函数 max(x)，求向量中的最小元素可调用函数 min(x)，具体调用格式如下（x 为向量）：

```
y=max(x)                %返回 x 中的最大元素给 y
[y,k]=max(x)            %返回 x 中的最大元素给 y，所在位置为 k
y=min(x)                %返回 x 中的最小元素给 y
[y,k]=min(x)            %返回 x 中的最小元素给 y，所在位置为 k
```

求矩阵各行或各列中的最大元素可调用函数 max(A)，求矩阵各行或各列中的最小元素可调用函数 min(A)，具体调用格式如下（A 为矩阵）：

```
Y=max(A)                %返回矩阵 A 每列中的最大元素给 Y，Y 是一个行向量
[Y,k]=max(A)            %返回矩阵 A 每列中的最大元素给 Y，k 记录每列最大元素的行号
[Y,k]=max(A,[],dim)     %dim=2 时，返回每行中的最大元素；dim=1 时，与 max(A)完全
                        %相同
Y=min(A)                %返回矩阵 A 每列中的最小元素给 Y，Y 是一个行向量
[Y,k]=min(A)            %返回矩阵 A 每列中的最小元素给 Y，k 记录每列最小元素的行号
```

```
[Y,k]=min(A,[],dim)     %dim=2 时，返回每行中的最小元素；dim=1 时，与 min(A)完全
                        %相同
```

对于相同维度的向量或矩阵，也可以用 max 函数和 min 函数求所有对应位置的最大值和最小值，具体调用方法参考例 5-1。

【例 5-1】 给定维度相等的矩阵 A 和 B（矩阵中元素值见下面的命令行输入），求其对应位置元素最大值和最小值。

```
>> A=[1 9 8;6 3 2;2 7 6];
>> B=[3 7 3;5 6 1;7 9 5];
>> C=max(A,B)
C =
   3   9   8
   6   6   2
   7   9   6
>> D=min(A,B)
D =
   1   7   3
   5   3   1
   2   7   5
```

mean 函数可以用来求向量或矩阵元素的算术平均值，具体调用格式如下（X 为向量，A 为矩阵）：

```
Y=mean(X)           %返回向量元素的算术平均值
B=mean(A)           %返回矩阵每列元素的算术平均值的行向量
B=mean(A,dim)       %dim=2 时，返回矩阵每行元素的算术平均值的列向量；dim=1 时，
                    %与 mean(A)完全相同
```

median 函数可以用来求向量或矩阵元素的中值，具体调用格式如下（X 为向量，A 为矩阵）：

```
y=median(X)         %返回向量元素的中值
B=median(A)         %返回矩阵每列元素的中值的行向量
B=median(A,dim)     %dim=2 时，返回矩阵每行元素的中值的列向量；dim=1 时，
                    %与 median(A)完全相同
```

sum 函数可以用来对向量或矩阵元素求和，具体调用格式如下（X 为向量，A 为矩阵）：

```
Y = sum(X)          %返回向量元素之和
B = sum(A)          %返回矩阵各列元素之和的行向量
B = sum(A,dim)      %dim=2 时，返回矩阵各行元素之和的列向量；dim=1 时，与 sum(A)
                    %完全相同
```

prod 函数可以用来对向量或矩阵元素求积，具体调用格式如下（X 为向量，A 为矩阵）：

```
Y = prod(X)         %返回向量元素之积
B = prod(A)         %返回矩阵各列元素之积的行向量
B = prod(A,dim)     %dim=2 时，返回矩阵各行元素之积的列向量；dim=1 时，与 prod(A)
                    %完全相同
```

cumsum 函数可以用来对向量或矩阵元素求累加和，具体调用格式如下（X 为向量，A 为矩阵）：

```
Y=cumsum(X)          %返回向量元素累加和
B=cumsum(A)          %返回矩阵各列元素累加和的行向量
B=cumsum(A,dim)      %dim=2 时，返回矩阵各行元素累加和的列向量；dim=1 时，
                     %与 cumsum(A)完全相同
```

cumprod 函数可以用来对向量或矩阵元素求累乘积，具体调用格式如下（X 为向量，A 为矩阵）：

```
Y=cumprod(X)         %返回向量元素累乘积
B=cumprod(A)         %返回矩阵各列元素累乘积的行向量
B=cumprod(A,dim)     %dim=2 时，返回矩阵各行元素累乘积的列向量；dim=1 时，与
                     %cumsum(A)完全相同
```

【例 5-2】 求向量 X 和矩阵 A 中元素的累加和与累乘积（向量和矩阵中的元素值见下面的命令行输入）。

```
>> A=[1 9 8;6 3 2;2 7 6];
>> X=[1 5 -2 3 6];
>> Y=cumsum(X)
Y =
1    6    4    7   13
>> Z=cumprod(X)
Z =
1    5   -10   -30  -180
>> B=cumsum(A)
B =
1    9    8
7   12   10
9   19   16
>> C=cumprod(A)
C =
 1    9    8
 6   27   16
12  189   96
>> D=cumsum(A,2)
D =
1   10   18
6    9   11
2    9   15
>> E=cumprod(A,2)
E =
1    9   72
6   18   36
2   14   84
```

var 函数可以用来返回向量或矩阵元素的方差，具体调用格式如下（X 为向量，A 为矩阵）：

```
Y=var(X)              %采用无偏估计式计算向量元素的方差，即前置因子为 1/(n-1)
Y=var(X,1)            %采用有偏估计式计算向量元素的方差，即前置因子为 1/n
B=var(A)              %采用无偏估计式计算矩阵中各列向量的方差，组成行向量
B=var(A,flag,dim)     %计算矩阵中指定维度的方差。dim 用来指定在行方向或列方向计算；
                      %flag=0 时采用无偏估计式，flag=1 时采用有偏估计式
```

std 函数可以用来返回向量或矩阵元素的标准差，具体调用格式如下（X 为向量，A 为矩阵）：

```
Y=std(X)              %采用无偏估计式计算向量元素的标准差
Y=std(X,1)            %采用有偏估计式计算向量元素的标准差
B=std(A)              %采用无偏估计式计算矩阵中各列向量的标准差，组成行向量
B=std(A,flag,dim)     %计算矩阵中指定维度的标准差。dim 用来指定在行方向或列方向计算；
                      %flag=0 时采用无偏估计式，flag=1 时采用有偏估计式
```

对于随机变量 x 和 y，其协方差的定义如下式所示：

$$\mathrm{cov}(x,y)=E\{[x-E(x)][y-E(y)]\}$$

对于 n 维随机变量$(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$，定义其协方差矩阵如下：

$$\boldsymbol{C}=\begin{bmatrix}\mathrm{cov}(x_1,x_1) & \mathrm{cov}(x_1,x_2) & \cdots & \mathrm{cov}(x_1,x_n)\\ \mathrm{cov}(x_2,x_1) & \mathrm{cov}(x_2,x_2) & \cdots & \mathrm{cov}(x_2,x_n)\\ \vdots & \vdots & \ddots & \vdots\\ \mathrm{cov}(x_n,x_1) & \mathrm{cov}(x_n,x_2) & \cdots & \mathrm{cov}(x_n,x_n)\end{bmatrix}$$

使用协方差，还可以计算随机变量 x 和 y 的相关系数，其公式如下：

$$\mathrm{cof}(x,y)=\frac{\mathrm{cov}(x,y)}{\sqrt{D(x)}\sqrt{D(y)}}$$

在 MATLAB 中，使用 cov 函数计算向量的协方差或矩阵的协方差矩阵，具体调用格式如下（X 和 Y 为维数相同的向量，A 和 B 为维数相同的矩阵）：

```
Z=cov(X)              %采用无偏估计式计算向量元素的协方差
Z=cov(X,1)            %采用有偏估计式计算向量元素的协方差
Z=cov (X,Y)           %采用无偏估计式得到两个向量的协方差矩阵
Z=cov(X,Y,1)          %采用有偏估计式得到两个向量的协方差矩阵
C=cov(A)              %采用无偏估计式得到矩阵列向量的协方差矩阵，即矩阵每行表示一组
                      %观察值，每列表示一个随机向量
C=cov(A,1)            %计算方法与 C=cov(A)一致，但采用有偏估计式
C=cov(A,B)            %采用无偏估计式，通过计算所有对应元素得到两个矩阵的协方差矩阵
C=cov(A,B,1)          %采用有偏估计式，通过计算所有对应元素得到两个矩阵的协方差矩阵
```

使用 corrcoef 函数可以计算向量的相关系数或矩阵的相关系数矩阵，具体调用格式如下（X 和 Y 为维数相同的向量，A 为矩阵）：

```
Z=corrcoef(X,Y)       %返回两个向量的相关系数
B=corrcoef(A)         %返回矩阵列向量的相关系数矩阵
```

除了以上常用的特征量计算函数以外，MATLAB 还提供了丰富的函数库用于计算数据特征量。

例如，对于平均值的计算，除了常用的 mean 函数外，还有以下函数：geomean 函数用

于计算样本的几何平均值，harmmean 函数用于计算样本的调和平均值，nanmean 函数用于在计算样本的平均值时忽略样本中的非数值型输入，trimmean 函数用于在计算样本的平均值时忽略数值变化过大的值。

对向量或矩阵进行排序，可以调用 sort 函数，具体调用格式如下（X 为向量，A 为矩阵）：

```
Y=sort(X)                 %返回一个按升序排列的向量
[B,I]=sort(A,dim,mode)    %dim=1 时按列排序，dim=2 时按行排序；mode 为'ascend'时
                          %升序，为'descend'时降序；I 记录 B 中元素在 A 中的位置
```

5.1.2 随机变量的分布

有些随机变量服从特殊的数学分布，如均匀分布、正态分布等，MATLAB 工具箱提供了丰富的函数对这些特殊分布进行描述。

在 MATLAB 中，提供了 pdf 函数作为通用的概率密度计算方法，具体调用格式如下（X 为向量，A 为矩阵）：

```
Y=pdf(name,X,v1,v2)       %返回概率密度向量，即输入的样本向量各元素在此分布中对应的
                          %概率密度。name 为分布名称，v1、v2 为此分布的参数
B=pdf(name,A,V1,V2,V3)    %返回输入样本矩阵在此分布中对应的概率密度矩阵。name 为分布
                          %名称，V1、V2、V3 为此分布的参数矩阵
```

除了 pdf 函数以外，MATLAB 也为很多特殊分布提供了独立的实现函数，在实际应用中它们往往可以与 pdf 函数互相替代。

1. 均匀分布

均匀分布就是在一个大的区域内，数据出现在任何一个小的区域的概率都是相同的。其概率密度函数见下式：

$$p(z)=\begin{cases}\dfrac{1}{b-a}, & a\leqslant z<b\\ 0, & \text{其他}\end{cases}$$

式中，a、b 分别为均匀分布的下界、上界。

当 μ 表示 z 的平均值或期望值，σ 表示 z 的标准差，而标准差的平方 σ^2 称为 z 的方差。均匀分布概率密度的均值和方差分别为

$$\mu=\frac{a+b}{2}$$

$$\sigma^2=\frac{(b-a)^2}{12}$$

均匀分布在 MATLAB 中的实现方式如下（X 为向量）：

```
Y=pdf('unif',X,v1,v2)     %返回概率密度向量，X 为样本向量，v1、v2 分别为均匀分布
                          %的下界、上界
```

```
Y=unifpdf(X,v1,v2)      %返回值与各参数意义均与 pdf 函数相同
```

另外，均匀分布作为产生模拟随机数的工具是非常有用的。

2. 正态分布

正态分布也称高斯分布或常态分布，呈钟形，两头低，中间高，是一个在数学、物理学及工程等领域都非常重要的概率分布，在统计学的许多方面有着重大的影响力。其概率密度函数见下式：

$$p(z)=\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{\frac{-(z-\mu)^2}{2\sigma^2}}$$

式中，μ 和 σ 为正态分布的均值和标准差。

正态分布在 MATLAB 中的实现方式如下（X 为向量）：

```
Y=pdf('norm',X,v1,v2)   %返回概率密度向量，X 为样本向量，v1、v2 分别为正态分布的
                        %均值和标准差
Y=normpdf(X,v1,v2)      %返回值与各参数意义均与 pdf 函数相同
```

当随机变量满足正态分布时，其值约 70%落在$[\mu-\sigma, \mu+\sigma]$范围内，且有约 95%落在$[\mu-2\sigma, \mu+2\sigma]$范围内。

3. 伽马分布

伽马分布是一种连续概率函数，指数分布和χ^2分布都是伽马分布的特例。其概率密度函数见下式：

$$p(z)=\begin{cases}\dfrac{a^n z^{n-1}}{(n-1)!}\mathrm{e}^{-az}, & z\geqslant 0\\ 0, & z<0\end{cases}$$

式中，$a>0$，n 为正整数。

伽马分布概率密度的均值和方差分别为

$$\mu=\frac{n}{a}$$

$$\sigma^2=\frac{n}{a^2}$$

伽马分布在 MATLAB 中的实现方式如下（X 为向量）：

```
Y=pdf('gam',X,v1,v2)    %返回概率密度向量，X 为样本向量，v1、v2 分别为伽马分布的
                        %形状参数和逆尺度参数
Y=gampdf(X,v1,v2)       %返回值与各参数意义均与 pdf 函数相同
```

4. 瑞利分布

瑞利分布是最常见的用于描述平坦衰落信号接收包络或独立多径分量接收包络统计时变特性的一种分布类型。两个正交高斯噪声信号之和的包络服从瑞利分布。其概率密度函数见下式：

$$p(z)=\begin{cases}\dfrac{z}{b^2}\mathrm{e}^{\frac{-z^2}{2b^2}}, & z\geqslant 0\\ 0, & z<0\end{cases}$$

式中，b 为尺度参数。

瑞利分布概率密度的均值和方差分别为

$$\mu=\sqrt{\frac{\pi}{2}}b$$

$$\sigma^2=\frac{b^2(4-\pi)}{2}$$

瑞利分布在 MATLAB 中的实现方式如下（X 为向量）：

```
Y=pdf('rayl',X,v1)        %返回概率密度向量，X 为样本向量，v1 为尺度参数
Y=raylpdf(X,v1)           %返回值与各参数意义均与 pdf 函数相同
```

5. 指数分布

指数分布又称负指数分布。泊松事件流的等待时间（相继两次出现之间的间隔）服从指数分布。通常假定排队系统中服务器的服务时间和 Petri 网中变迁的实施速率服从指数分布。其概率密度函数见下式：

$$p(z)=\begin{cases}a\mathrm{e}^{-az}, & z\geqslant 0\\ 0, & z<0\end{cases}$$

式中，a 为形状参数。

指数分布概率密度的均值和方差分别为

$$\mu=\frac{1}{a}$$

$$\sigma^2=\frac{1}{a^2}$$

指数分布在 MATLAB 中的实现方式如下（X 为向量）：

```
Y=pdf('exp',X,v1)         %返回概率密度向量，X 为样本向量，v1 为形状参数
Y=exppdf(X,v1)            %返回值与各参数意义均与 pdf 函数相同
```

实践中，指数分布常被用于描述非老化性元件的寿命（非老化性元件不老化，仅由于突然故障而毁坏）。

6. 伪随机数

对应于各种分布并仿照随机数发生的规律所计算出来的随机数称为伪随机数。

不同于真正意义上的随机数，伪随机数是由数学公式计算出来的。

在 MATLAB 中，提供了 random 函数作为通用的伪随机数产生方法，具体调用格式如下：

```
Y=random (name,v1,v2,v3,v4)     %返回伪随机数矩阵,name 为伪随机数服从的分布名称,
```

```
                                        %v1、v2 为此分布的参数，返回的矩阵有 v3 行、v4 列
```

除了 random 函数以外，MATLAB 也针对很多特殊分布提供了独立的伪随机数生成函数，在实际应用中它们往往可以与 random 函数互相替代。

例如，unifrnd 函数、normrnd 函数和 raylrnd 函数分别可以生成服从均匀分布、正态分布和瑞利分布的伪随机数，具体调用格式如下：

```
y=unifrnd(v1,v2)      %生成服从均匀分布的伪随机数，v1、v2 分别为均匀分布的下界、上界
y=normrnd(v1,v2)      %生成服从正态分布的伪随机数，v1、v2 分别为正态分布的均值、标准差
y=raylrnd(v1)         %生成服从瑞利分布的伪随机数，v1 为瑞利分布的尺度参数
```

另外，比较常用的随机数生成函数还有 rand 函数和 randn 函数，常被用于生成伪随机数矩阵，具体调用格式如下：

```
B=rand(v1,v2)         %生成服从[0,1]区间内均匀分布的伪随机数矩阵，该矩阵有 v1 行、
                      %v2 列
B=randn(v1,v2)        %生成服从标准正态分布（均值为 0、标准差为 1）的伪随机数矩阵，
                      %该矩阵有 v1 行、v2 列
```

5.1.3 参数估计

参数估计是统计推断的一种，是根据从总体中抽取的随机样本来估计总体分布中未知参数的过程。从估计形式看，参数估计分为点估计与区间估计：从构造估计量的方法看，参数估计有矩法估计、最小二乘估计、似然估计、贝叶斯估计等。参数估计要处理两个问题：第一，求出未知参数的估计量；第二，在一定信度（可靠程度）下指出所求的估计量的精度。信度一般用概率表示，如信度为 95%；精度用估计量与被估参数之间的接近程度或误差来度量。

MATLAB 在工具箱中针对常用的多种随机变量分布，提供了对应的极大似然估计函数。

调用 unifit 函数可以对均匀分布参数进行极大似然估计，具体调用格式如下（X 为向量，为一组样本值）：

```
[y1,y2]=unifit(X)                 %返回值 y1、y2 分别为均匀分布的下界、上界的极大
                                  %似然估计
[y1,y2,z1,z2]=unifit(X,alpha)     %返回值 z1、z2 分别为 y1、y2 两个极大似然估计结果
                                  %在显著性水平 alpha 下的置信区间估计
[y1,y2,z1,z2]=unifit(X)           %返回值 z1、z2 分别为 y1、y2 两个极大似然估计结果
                                  %在显著性水平 0.05 下的置信区间估计
```

调用 normfit 函数可以对正态分布参数进行极大似然估计，具体调用格式如下（X 为向量，为一组样本值）：

```
[y1,y2]=normfit(X)                %返回值 y1、y2 分别为正态分布的均值、标准差的极大
                                  %似然估计
[y1,y2,z1,z2]=normfit(X,alpha)    %返回值 z1、z2 分别为 y1、y2 两个极大似然估计结果
                                  %在显著性水平 alpha 下的置信区间估计
```

```
[y1,y2,z1,z2]=normfit(X)        %返回值 z1、z2 分别为 y1、y2 两个极大似然估计结果
                                %在显著性水平 0.05 下的置信区间估计
```

调用 gamfit 函数可以对伽马分布参数进行极大似然估计，具体调用格式如下（X 为向量，为一组样本值）：

```
[y]=gamfit(X)                   %返回值 y 为伽马分布参数的极大似然估计，y(1)和 y(2)
                                %分别是形状参数和尺度参数的估计
[y,za,zb]=gamfit(X,alpha)       %返回值 za、zb 分别为极大似然估计结果在显著性水平 alpha
                                %下的置信区间估计
[y,za,zb]=gamfit(X,alpha)       %返回值 zb、zb 分别为极大似然估计结果在显著性水平 0.05
                                %下的置信区间估计
```

调用 raylfit 函数可以对瑞利分布参数进行极大似然估计，具体调用格式如下（X 为向量，为一组样本值）：

```
[y]=raylfit(X)                  %返回值 y 为瑞利分布参数的极大似然估计，y(1)和 y(2)
                                %分别是形状参数和尺度参数的估计
[y,za,zb]=raylfit(X,alpha)      %返回值 za、zb 分别为极大似然估计结果在显著性水平 alpha
                                %下的置信区间估计
[y,za,zb]=raylfit(X,alpha)      %返回值 za、zb 分别为极大似然估计结果在显著性水平 0.05
                                %下的置信区间估计
```

调用 expfit 函数可以对指数分布参数进行极大似然估计，具体调用格式如下（X 为向量，为一组样本值）：

```
[y]=expfit(X)                   %返回值 y 为指数分布参数的极大似然估计
[y,za,zb]=expfit(X,alpha)       %返回值 za、zb 分别为极大似然估计结果在显著性水平 alpha
                                %下的置信区间估计
[y,za,zb]=expfit(X,alpha)       %返回值 za、zb 分别为极大似然估计结果在显著性水平 0.05
                                %下的置信区间估计
```

【例 5-3】 假设数据 X 服从均匀分布，现采集到 X 的一组样本值，即向量 X1，试估计均匀分布的上下界以及 95%置信区间。

```
>> X1=[3.9 1.3 2.8 1.9 2.1 3.5];
>> [y1,y2,z1,z2]=unifit(X1)
y1 =
      1.3000
y2 =
      3.9000
z1 =
     -0.3836
      1.3000
z2 =
      3.9000
      5.5836
```

5.2 数据统计处理

5.2.1 假设检验

假设检验又称统计假设检验，用来判断样本与样本、样本与总体的差异是由抽样误差引起的还是本质差别造成的。显著性检验是假设检验中最常用的一种方法，也是一种最基本的统计推断形式。

显著性检验的基本原理是：先对总体的特征做出某种假设，然后通过抽样研究的统计推理，对此假设应该被拒绝还是被接受做出推断。

常用的假设检验方法有 z 检验、t 检验等。z 检验的前提条件是样本数据服从正态分布，而实际应用中总体方差往往是未知的，需要用大样本数据的方差作为总体方差的估计值。因此，z 检验主要适用于总体方差未知的大样本数据。t 检验是指在未知标准差的情况下对于服从正态分布的样本均值的检验。

在 MATLAB 中，用 ztest 函数实现 z 检验，具体调用格式如下（X 为向量，为一组样本值）：

```
h=ztest(X,m,sigma)          %已知标准差为 sigma 的情况下，在显著性水平 0.05 下
                            %对均值是否为 m 的检验
h=ztest(X,m,sigma,alpha)    %已知标准差为 sigma 的情况下，在显著性水平 alpha 下
                            %对均值是否为 m 的检验
```

在实际应用中，t 检验还可分为双侧检验、左尾检验和右尾检验。三者所检验的假设命题 H_0 均为“样本均值为 m”，但区别在于对立假设命题 H_1 不同，分别为“样本均值不为 m”“样本均值小于 m”和“样本均值大于 m”。左尾检验和右尾检验又可统称为单侧检验。

在 MATLAB 中，用 ttest 函数实现 t 检验，具体调用格式如下（X 为向量，为一组样本值）：

```
h=ttest(X)                  %检验正态分布样本均值是否为 0
h=ttest(X,m)                %检验正态分布样本均值是否为 m
h=ttest(X,m,Name,Value)     %当 Name 为'Alpha'时，可指定显著性水平；当 Name 为
                            %'Tail'时，可根据 Value 为'both'、'left'或'right'
                            %指定假设检验为双侧检验、左尾检验或右尾检验，默认值为
                            %双侧检验
```

5.2.2 方差分析

方差分析又称变异数分析，常用于两个及两个以上样本均值差别的显著性检验。由于各种因素的影响，研究所得的数据往往呈现出波动。造成波动的原因可分成两类：一是不可控的随机因素；二是研究中施加的对结果形成影响的可控因素。

方差分析返回多组原假设样本来自具有相同均值总体的 p 值，根据数据设计类型的不同，有不同的方差分析方法：