第5章

统计性数据分析实战

统计工具箱提供了用于描述数据、分析数据以及为数据建模的函数和 App。可以使用描述性统计量和绘图进行探索性数据分析，对数据进行概率分布拟合，生成进行蒙特卡罗仿真的随机数，以及执行假设检验。

5.1 统计量和统计图

描述性统计分析要对调查总体所有变量的有关数据做统计性描述，主要包括数据的频数分析、数据的集中趋势分析、数据离散程度分析、数据的分布，以及一些基本的统计图形。常见的分析方法包括对比分析法、平均分析法、交叉分析法等。

5.1.1 描述性统计量

描述性统计量是指通过生成汇总统计量，包括集中趋势、散度、形状和相关性方面的度量，以数值方式来探索数据。统计工具箱允许计算包含缺失(NaN)值的样本数据的汇总统计量。利用一元图、二元图和多元图，实现数据的可视化。可用选项包括箱线图、直方图和概率图。

1. 数据管理

在 MATLAB 中，可以使用多种不同的文件格式将数据导入和导出。有效格式包括表格数据、以制表符分隔的文件、Microsoft Excel 电子表格以及 SAS XPORT 文件。统计工具箱还提供了更多数据类型，用于处理分组变量和分类数据。此工具箱还支持 MATLAB 中许多(但不是全部)可用的数据类型。

统计工具箱中包括表 5-1 中的实例数据集。要将数据集加载到 MATLAB 工作区中，可输入：

```
load filename
```

其中，filename 是表中列出的文件之一。

数据集包含单独的数据变量、具有引用的描述变量以及封装数据集及其描述的数据集数组(如果适用)。

表 5-1 常用的实例数据集

文　　件	数据集的描述
acetylene. mat	具有相关预测变量的化学反应数据
arrhythmia. mat	来自 UCI 机器学习存储库的心律失常数据
carbig. mat	汽车的测量值,1970—1982
carsmall. mat	carbig. mat 的子集。汽车的测量值,1970、1976、1982
census1994. mat	来自 UCI 机器学习存储库的成人数据
cereal. mat	早餐谷物成分
cities. mat	美国大都市地区的生活质量评分
discrim. mat	用于判别分析的 cities. mat 版本
examgrades. mat	0～100 分的考试成绩
fisheriris. mat	Fisher 1936 年的鸢尾花数据
flu. mat	Google 流感趋势估计的美国不同地区的 ILI(流感样疾病)百分比,疾病预防控制中心根据哨点提供商报告对 ILI 百分比进行了加权
gas. mat	1993 年马萨诸塞州的汽油价格
hald. mat	水泥发热与原料混合
hogg. mat	牛奶的不同配送方式中的细菌数量
hospital. mat	仿真的医疗数据
humanactivity. mat	5 种活动的人类活动识别数据：坐、站、走、跑和跳舞
imports-85. mat	1985 年来自 UCI 存储库的自动导入数据库
ionosphere. mat	来自 UCI 机器学习存储库的电离层数据集
kmeansdata. mat	四维聚类数据
lawdata. mat	15 所法学院的平均分数和 LSAT 分数
mileage. mat	两家工厂的三种汽车型号的里程数据
moore. mat	关于 5 个预测变量的生化需氧量
morse. mat	非编码人员对摩尔斯电码的识别情况
nlpdata. mat	从 MathWorks 文档中提取的自然语言处理数据
ovarianceancer. mat	关于 4000 个预测变量的分组观测值
parts. mat	36 个圆形零件的大小偏差
polydata. mat	多项式拟合的样本数据
popcorn. mat	爆米花机型和品牌的爆米花产出
reaction. mat	Hougen-Watson 模型的反应动力学
sat. dat	按性别和测验分列的学术能力测验（SAT）平均分(表)
sat2. dat	按性别和测验分列的学术能力测验（SAT）平均分（csv）
spectra. mat	60 份汽油样本的近红外光谱和辛烷值
stockreturns. mat	仿真的股票回报

2. 数据类型

统计工具箱还另外提供了两种数据类型。要处理有序和无序的离散非数值数据,可以使用 nominal 和 ordinal 数据类型。要将多个变量(包括具有不同数据类型的变量)存储到一个对象中,可以使用 dataset 数组数据类型。但是,这些数据类型是统计和机器学习工具箱所独有的。要获得更好的跨产品兼容性,可分别使用 MATLAB 中提供的 categorical 或 table 数据类型。

【例 5-1】 使用数据集数组变量及其数据。

(1) 按名称访问变量。

可以通过使用变量(列)名称和点索引来访问变量数据或选择变量子集。加载样本数据集

数组，显示 hospital 中变量的名称。

```
>> load hospital
hospital.Properties.VarNames(:)
ans =
  7×1 cell 数组
   {'LastName'      }
   {'Sex'           }
   {'Age'           }
   {'Weight'        }
   {'Smoker'        }
   {'BloodPressure' }
   {'Trials'        }
```

数据集数组中有 7 个变量(列)和 100 个观测值(行)。可以在工作区窗口中双击 hospital 以在变量编辑器中查看数据集数组。

(2) 绘制直方图。

绘制变量 Weight 中数据的直方图，如图 5-1 所示。

```
>> figure
histogram(hospital.Weight)
```

图 5-1 中的直方图显示体重呈双峰分布。

(3) 绘制按类别分组的数据。

绘制按 Sex 中的值分组(男性和女性)的 Weight 的箱线图。也就是说，使用变量 Sex 作为分组变量。

```
>> figure
boxplot(hospital.Weight,hospital.Sex)        % 效果如图 5-2 所示
```

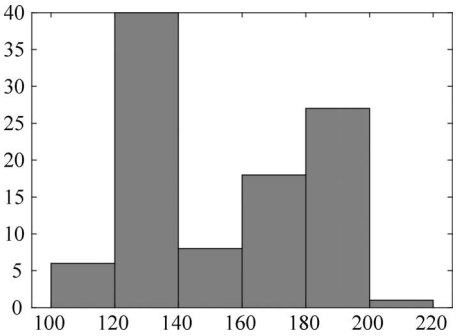

图 5-1　Weight 中的数据直方图

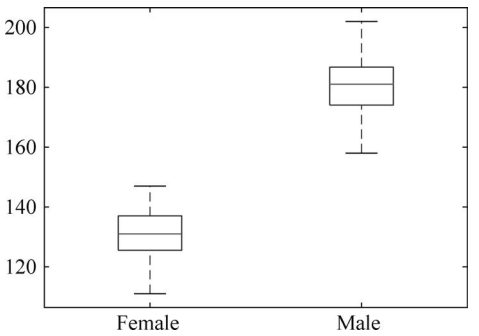

图 5-2　箱线图

图 5-2 的箱线图表明性别是体重呈双峰分布的原因。

(4) 选择一个变量子集。

创建一个新数据集数组，其中仅包含变量 LastName、Sex 和 Weight。可以通过名称或列号访问变量。

```
>> ds1 = hospital(:,{'LastName','Sex','Weight'});
ds2 = hospital(:,[1,2,4]);
```

数据集数组 ds1 和 ds2 是等同的。在对数据集数组进行索引时，使用括号()可保留数据类型；也就是说，基于数据集数组的子集创建一个数据集数组。还可以使用变量编辑器基于变量和观测值的子集创建一个新数据集数组。

(5) 转换变量数据类型。

将变量 Smoker 的数据类型从逻辑值转换为名义值,标签为 No 和 Yes。

```
>> hospital.Smoker = nominal(hospital.Smoker,{'No','Yes'});
class(hospital.Smoker)
ans =
    'nominal'
```

(6) 探查数据。

显示 Smoker 的前 10 个元素。

```
>> hospital.Smoker(1:10)
ans =
  10×1 nominal 数组
    Yes
    No
    No
    No
    No
    No
    Yes
    No
    No
    No
```

如果要更改名义数组中的水平标签,请使用 setlabels。

(7) 添加变量。

变量 BloodPressure 是 100×2 数组。第一列对应于收缩压,第二列对应于舒张压。将此数组分成两个新变量 SysPressure 和 DiaPressure。

```
>> hospital.SysPressure = hospital.BloodPressure(:,1);
hospital.DiaPressure = hospital.BloodPressure(:,2);
hospital.Properties.VarNames(:)
ans =
  9×1 cell 数组
   {'LastName'      }
   {'Sex'           }
   {'Age'           }
   {'Weight'        }
   {'Smoker'        }
   {'BloodPressure' }
   {'Trials'        }
   {'SysPressure'   }
   {'DiaPressure'   }
```

可见,数据集数组 hospital 有两个新变量。

(8) 按名称搜索变量。

使用 regexp 查找 hospital 中变量名称包含'Pressure'的变量。创建只包含这些变量的新数据集数组。

```
>> bp = regexp(hospital.Properties.VarNames,'Pressure');
bpIdx = cellfun(@isempty,bp);
bpData = hospital(:,~bpIdx);
bpData.Properties.VarNames(:)
ans =
  3×1 cell 数组
```

```
    {'BloodPressure'}
    {'SysPressure'  }
    {'DiaPressure'  }
```

可见，新数据集数组 bpData 仅包含血压变量。

(9) 删除变量。

从数据集数组 hospital 中删除变量 BloodPressure。

```
>> hospital.BloodPressure = [];
hospital.Properties.VarNames(:)
ans =
  8×1 cell 数组
    {'LastName'   }
    {'Sex'        }
    {'Age'        }
    {'Weight'     }
    {'Smoker'     }
    {'Trials'     }
    {'SysPressure'}
    {'DiaPressure'}
```

可见，变量 BloodPressure 不再在数据集数组中。

5.1.2　常用的统计量函数

根据样本数据计算描述性统计量，包括有关集中趋势、散度、形状、相关性和协方差的度量。制作数据的一般报表和交叉表，并计算分组数据的汇总统计量。如果数据中包含缺失(NaN)值，MATLAB 算术运算函数将返回 NaN。不过，统计工具箱提供的专用函数可以忽略这些缺失值，并返回使用其余值计算的数值。

下面介绍两个较为常用的统计量函数。

1. prctile 函数

prctile 函数用于计算数据集的百分位数。函数的调用格式如下。

Y＝prctile(X,p)：根据区间[0,100]中的百分比 p 返回数据向量或数组 X 中元素的百分位数。

- 如果 X 是向量，则 Y 是标量或向量，向量长度等于所请求百分位数的个数(length(p))。$Y(i)$包含第 $p(i)$个百分位数。
- 如果 X 是矩阵，则 Y 是行向量或矩阵，其中，Y 的行数等于所请求百分位数的个数(length(p))。Y 的第 i 行包含 X 的每一列的第 $p(i)$个百分位数。
- 对于多维数组，prctile 在 X 的第一个非单一维度上进行运算。

  Y＝prctile(X,p,'all')：返回 X 的所有元素的百分位数。

  Y＝prctile(X,p,dim)：返回运算维度 dim 上的百分位数。

  Y＝prctile(X,p,vecdim)：基于向量 vecdim 所指定的维度返回百分位数。例如，如果 X 是矩阵，则 prctile(X,50,[1 2])返回 X 的所有元素的第 50 个百分位数，因为矩阵的每个元素都包含在由维度 1 和 2 定义的数组切片中。

  Y＝prctile(____,'Method',method)：使用上述任一语法中的输入参数组合，根据 method 的值，返回精确或近似百分位数。

【例 5-2】 计算数组中所有值的百分位数。

```
>> % 创建 3×5×2 数组 X
```

```
X = reshape(1:30,[3 5 2])
X(:,:,1) =
    1     4     7     10    13
    2     5     8     11    14
    3     6     9     12    15
X(:,:,2) =
    16    19    22    25    28
    17    20    23    26    29
    18    21    24    27    30
>> % 计算 X 的元素的第 40 个和第 60 个百分位数
Y = prctile(X,[40 60],'all')
Y =
    12.5000
    18.5000
```

2. corr 函数

corr 函数用于计算线性或秩相关性。函数的调用格式如下。

rho=corr(X)：返回输入矩阵 ***X*** 中各列之间的两两线性相关系数矩阵。

rho=corr(X,Y)：返回输入矩阵 ***X*** 和 ***Y*** 中各列之间的两两相关系数矩阵。

[rho,pval]=corr(X,Y)：还返回 pval,它是一个 p 值矩阵,用于基于非零相关性备择假设来检验无相关性假设。

除了上述语法中的输入参数,[rho,pval]=corr(____,Name,Value)还使用一个或多个名称-值对组参数指定选项。例如,'Type','Kendall'指定计算 Kendall tau 相关系数。

【例 5-3】 计算两个矩阵之间的相关性,并将其与两个列向量之间的相关性进行比较。

```
% 生成样本数据
>> rng('default')
X = randn(30,4);
Y = randn(30,4);
>> % 在矩阵 X 的第二列和矩阵 Y 的第四列之间引入相关性
Y(:,4) = Y(:,4) + X(:,2);
% 计算 X 和 Y 的列之间的相关性
rho,pval] = corr(X,Y)
rho =
      -0.1686   -0.0363    0.2278    0.3245
       0.3022    0.0332   -0.0866    0.7653
      -0.3632   -0.0987   -0.0200   -0.3693
      -0.1365   -0.1804    0.0853    0.0279
pval =
       0.3731    0.8489    0.2260    0.0802
       0.1045    0.8619    0.6491    0.0000
       0.0485    0.6039    0.9166    0.0446
       0.4721    0.3400    0.6539    0.8837
```

5.1.3 统计可视化

在 MATLAB 中,使用一元图(如箱线图和直方图)研究一元分布,使用二元图(如分组散点图和二元直方图)显示变量之间的关系,使用多元图(如 Andrews 图和图形符号图)可视化多个变量之间的关系。通过添加记录名称、最小二乘线条和参考曲线来自定义绘图。

下面介绍两个常用的函数。

1. boxplot 函数

在统计工具箱中，提供了 boxplot 函数绘制箱线图。函数的调用格式如下。

boxplot(***x***)：创建 ***x*** 中数据的箱线图。如果 ***x*** 是向量，boxplot 绘制一个箱子；如果 ***x*** 是矩阵，boxplot 为 ***x*** 的每列绘制一个箱子。

在每个箱子上，中心标记表示中位数，箱子的底边和顶边分别表示第 25 个和 75 个百分位数。须线会延伸到不是离群值的最远端数据点，离群值会以'＋'符号单独绘制。

boxplot(x,g)：使用 g 中包含的一个或多个分组变量创建箱线图。boxplot 为具有相同的一个或多个 g 值的各组 x 值创建一个单独的箱线。

boxplot(ax,____)：使用坐标区图形对象 ax 指定的坐标区和任何上述语法创建箱线图。

boxplot(____,Name,Value)：使用由一个或多个 Name 和 Value 对组参数指定的附加选项创建箱线图。例如，可以指定箱子样式或顺序。

函数的应用可参考例 5-1。

2. refline 函数

在统计工具箱中，提供了 refline 函数实现将参考线添加到绘图中。函数的调用格式如下。

refline(m,b)：在当前坐标区中添加一条具有斜率 m 和截距 b 的参考线。

refline(coeffs)：将由向量 coeffs 的元素定义的线添加到图窗中。

refline(ax,____)：使用上述任一语法中的输入参数，向 ax 所指定坐标区中的图上添加一条参考线。

hline＝refline(____)：使用上述任一语法中的输入参数，返回参考线对象 hline。在创建参考线后，使用 hline 修改其属性。

【例 5-4】 在均值处添加参考线。

```
>> % 为自变量 x 和因变量 y 生成样本数据
x = 1:10;
y = x + randn(1,10);
 % 创建 x 和 y 的散点图，如图 5 - 3 所示
scatter(x,y,25,'b','*')
下面实现在散点图上叠加一条最小二乘线，如图 5 - 4 所示
>> refline
```

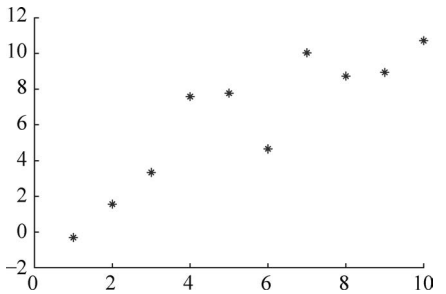

图 5-3　x 和 y 的散点图

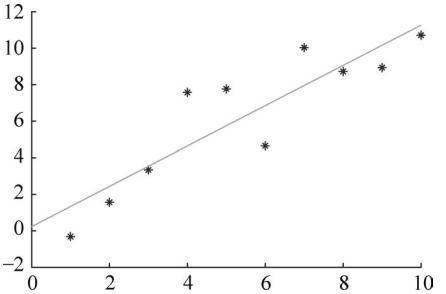

图 5-4　叠加一条最小二乘线

下面实现在散点图的均值处添加一条参考线，如图 5-5 所示。

```
>> mu = mean(y);
hline = refline([0 mu]);
hline.Color = 'r';
```

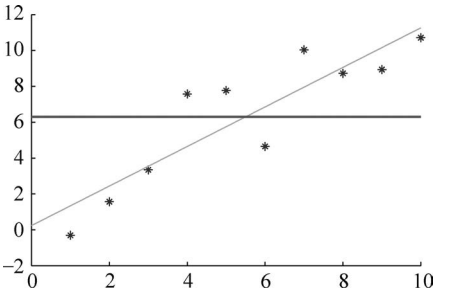

图 5-5　均值处添加一条参考线

5.2　概 率 分 布

概率分布，是概率论中的基本概念之一，主要用于表述随机变量取值的概率规律，可用来计算均值和中值等汇总统计量、可视化样本数据、生成随机数等。在 MATLAB 中可使用概率分布对象、命令行函数或交互式 App 来处理概率分布。

概率分布可分为离散概率分布和连续概率分布，下面针对这两种分布进行介绍。

5.2.1　离散概率分布

离散概率分布是指随机变量只能取有限（或可数无限）数量的值的概率分布。例如，在二项分布中，随机变量 X 只能取值 0 或 1。统计工具箱提供了几种处理离散概率分布的方法，包括概率分布对象、命令行函数和交互式 App。

1. 二项分布

二项分布即重复 n 次独立的伯努利实验。在每次实验中只有两种可能的结果，而且两种结果发生与否互相对立，并且相互独立，与其他各次实验结果无关，事件发生与否的概率在每一次独立实验中都保持不变，则这一系列实验总称为 n 重伯努利实验，当实验次数为 1 时，二项分布服从 0-1 分布。

在统计和机器学习工具箱中提供了如下几种处理二项分布的方法。

（1）通过将概率分布拟合到样本数据（fitdist）或指定参数值（makedist）来创建概率分布对象 BinomialDistribution。然后，使用对象函数来评估分布、生成随机数等。

（2）使用 Distribution Fitter 应用程序以交互方式处理二项分布。可以从应用程序中导出对象并使用对象函数。

（3）使用具有指定分布参数的分布特定函数（binocdf、binopdf、binoinv、binostat、binofit、binornd）。特定于分布的函数可以接受多个二项分布的参数。

（4）使用具有指定分布名称（“二项式”）和参数的通用分布函数（cdf、icdf、pdf、随机）。

二项分布的概率密度函数（pdf）为：

$$f(x \mid N,p)=\binom{N}{x}p^{x}(1-p)^{N-x},\quad x=0,1,2,\cdots,N$$

其中，x 是成功概率为 p 的伯努利过程的 N 次实验中的成功次数。结果是在 N 次实验中恰好 x 次成功的概率。对于离散分布，pdf 也称为概率质量函数（pmf）。

二项分布的累积分布函数（cdf）为：

$$F(x \mid N,p)=\sum_{i=0}^{x}\binom{N}{i}p^{i}(1-p)^{N-i};\quad x=0,1,2,\cdots,N$$

其中，x 是成功概率为 p 的伯努利过程的 N 次实验中的成功次数。结果是在 N 次实验中最多 x 次成功的概率。

下面对几个常用的二项分布函数进行介绍。

1）fitdist 函数

在统计工具箱中，提供了 fitdist 函数对数据进行概率分布对象拟合。函数的调用格式如下。

pd＝fitdist(x,distname)：通过对列向量 $\boldsymbol{x}$ 中的数据进行 distname 指定的分布拟合，创建概率分布对象。

pd＝fitdist(x,distname,Name,Value)：使用一个或多个名称-值对组参数指定的附加选项创建概率分布对象。例如，可以为迭代拟合算法指示删失数据或指定控制参数。

[pdca,gn,gl]＝fitdist(x,distname,'By',groupvar)：基于分组变量 groupvar 对 x 中的数据进行 distname 指定的分布拟合，以创建概率分布对象。它返回拟合后的概率分布对象的元胞数组 pdca、组标签的元胞数组 gn 以及分组变量水平的元胞数组 gl。

[pdca,gn,gl]＝fitdist(x,distname,'By',groupvar,Name,Value)：使用一个或多个名称-值对组参数指定的附加选项返回上述输出参数。

2）pdf 函数

在统计工具箱中，pdf 函数为概率密度函数。函数的调用格式如下。

y＝pdf('name',x,A)：返回由'name'和分布参数 A 指定的单参数分布族的概率密度函数(pdf)，在 x 中的值处计算函数值。

y＝pdf('name',x,A,B)：返回由'name'以及分布参数 A 和 B 指定的双参数分布族的 pdf，在 x 中的值处计算函数值。

y＝pdf('name',x,A,B,C)：返回由'name'以及分布参数 A、B 和 C 指定的三参数分布族的 pdf，在 x 中的值处计算函数值。

y＝pdf('name',x,A,B,C,D)：返回由'name'以及分布参数 A、B、C 和 D 指定的四参数分布族的 pdf，在 x 中的值处计算函数值。

y＝pdf(pd,x)：返回概率分布对象 pd 的 pdf，在 x 中的值处计算函数值。

【例 5-5】 对数据进行正态分布拟合。

```
>> % 加载样本数据。创建包含患者体重数据的向量
load hospital
x = hospital.Weight;
% 通过对数据进行正态分布拟合来创建正态分布对象
pd = fitdist(x,'Normal')
pd =
  NormalDistribution
  正态 分布
       mu =      154   [148.728, 159.272]
    sigma = 26.5714   [23.3299, 30.8674]
```

参数估计值旁边的区间是分布参数的 95%置信区间。

```
% 绘制分布的 pdf，如图 5-6 所示
>> x_values = 50:1:250;
y = pdf(pd,x_values);
plot(x_values,y,'LineWidth',2)
```

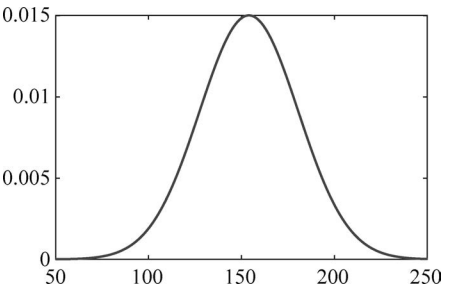

图 5-6　分布的 pdf 图

3）cdf 函数

在统计工具箱中，cdf 函数为累积分布函数。函数的调用格式如下。

y＝cdf('name',x,A)：基于 x 中的值计算并返回由'name'和分布参数 A 指定的单参数分布族的累积分布函数(cdf)值。

y＝cdf('name',x,A,B)：基于 x 中的值计算并返回由'name'以及分布参数 A 和 B 指定的双参数分布族的 cdf。

y＝cdf('name',x,A,B,C)：基于 x 中的值计算并返回由'name'以及分布参数 A、B 和 C 指定的三参数分布族的 cdf。

y＝cdf('name',x,A,B,C,D)：基于 x 中的值计算并返回由'name'以及分布参数 A、B、C 和 D 指定的四参数分布族的 cdf。

y＝cdf(pd,x)：基于 x 中的值计算并返回概率分布对象 pd 的 cdf。

y＝cdf(____,'upper')：使用可更精确计算极值上尾概率的算法返回 cdf 的补函数。'upper' 可以跟在上述语法中的任何输入参数之后。

【例 5-6】 计算正态分布 cdf。

```
% 创建均值 μ 等于 0、标准差 σ 等于 1 的标准正态分布对象
>> mu = 0;
sigma = 1;
pd = makedist('Normal','mu',mu,'sigma',sigma);
>> % 定义输入向量 x 以包含用于计算 cdf 的值
x = [-2,-1,0,1,2];
% 基于 x 中的值计算标准正态分布的 cdf 值
y = cdf(pd,x)
y =
    0.0228    0.1587    0.5000    0.8413    0.9772
```

y 中的每个值对应于输入向量 x 中的一个值。例如，在值 x 等于 1 时，对应的 cdf 值 y 等于 0.8413。

也可以不创建概率分布对象而直接计算同样的 cdf 值。使用 cdf 函数，再使用同样的 μ 和 σ 参数值指定一个标准正态分布。

```
>> y2 = cdf('Normal',x,mu,sigma)
y2 =
    0.0228    0.1587    0.5000    0.8413    0.9772
```

cdf 值与使用概率分布对象计算的值相同。

2. 多项式分布

多项式分布是二项分布的推广。二项分布(也叫伯努利分布)的典型例子是扔硬币，硬币正面朝上概率为 p，重复扔 n 次硬币，k 次为正面的概率即为一个二项分布概率。而多项分布就像扔骰子，有 6 个面对应 6 个不同的点数。二项分布时事件 X 只有两种取值，而多项分布的 X 有多种取值，多项分布的概率公式为：

$$f(x \mid n,p)=\frac{n!}{x_1!\cdots x_k!}p_1^{x_1}\cdots p_k^{x_k}$$

式中，k 是每个实验中可能出现的相互排斥的结果数，n 是实验总数。向量 $x=(x_1\cdots x_k)$ 是每 k 个结果的观察数，包含总和为 n 的非负整数分量。向量 $p=(p_1\cdots p_k)$ 是每 k 个结果的固定概率，包含总和为 1 的非负标量分量。

在 n 次实验中结果 i 的预期观察次数为：

$$E\{x_i\}=np_i$$

其中，p_i 是结果 i 的概率。方差的结果 i 是：

$$\mathrm{var}(x_i)=np_i(1-p_i)$$

结果 i 和 j 的协方差为：

$$\mathrm{cov}(x_i,x_j)=-np_ip_j,\quad i\neq j$$

【例 5-7】 生成随机数，计算并绘制 pdf，以及使用概率分布对象计算多项式分布的描述性统计信息。

(1) 定义分布参数。

创建一个包含每个结果概率的向量 p。结果 1 的概率为 1/2，结果 2 的概率为 1/3，结果 3 的概率为 1/6。每个实验的实验次数 n 为 5 次，重复次数为 8 次。

```
>> p = [1/2 1/3 1/6];
n = 5;
reps = 8;
```

(2) 创建一个多项式概率分布对象。

使用“概率”参数的指定值 p 创建多项式概率分布对象。

```
>> pd = makedist('Multinomial','Probabilities',p)
pd =
  MultinomialDistribution
  Probabilities:
    0.5000    0.3333    0.1667
>> rng('default')                % 重复性
r = random(pd)
r =
    2
```

结果表明，本实验结果为 2。

(3) 生成一个随机数矩阵。

还可以从多项式分布生成一个随机数矩阵，该矩阵报告了包含多个实验的多个实验结果。生成的矩阵，其中包含 $n=5$ 次实验和 8 次重复的实验结果。

```
>> r = random(pd,reps,n)
r =
    3     3     3     2     1
    1     1     2     2     1
    3     3     3     1     2
    2     3     2     2     2
    1     1     1     1     1
    1     2     3     2     3
    2     1     3     1     1
    3     1     2     1     1
```

结果矩阵中的每个元素都是一次实验的结果。列对应于每个实验中的 5 个实验，行对应于 8 个实验。例如，在第一个实验中(对应于第一行)，5 个实验中的一个得出结果 1，5 个实验中的一个得出结果 2，5 个实验中的三个得出结果 3。

(4) 计算并绘制 pdf。

计算分布的 pdf，并绘图，如图 5-7 所示。

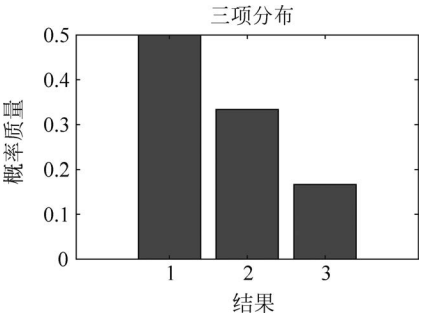

图 5-7　多项分布图

```
>> x = 1:3;
y = pdf(pd,x);
bar(x,y)
xlabel('结果')
ylabel('概率质量')
title('三项分布')
```

该图显示了每 k 个可能结果的概率质量。对于此分布，除 1、2 或 3 之外的任何 x 的 pdf 值均为 0。

(5) 计算描述性统计。

计算分布的均值、中值和标准差。

```
>> m = mean(pd)        % 均值
m =
    1.6667
>> med = median(pd)    % 中值
med =
    1
>> s = std(pd)         % 标准差
s =
    0.7454
```

3. 泊松分布

泊松分布适用于涉及计算在给定的时间段、距离、面积等范围内发生随机事件的次数的应用情形。应用泊松分布的例子包括盖革计数器每秒咔嗒的次数、每小时走入商店的人数，以及每 1000 英尺录像带的瑕疵数。

泊松的概率分布函数为：

$$f(x \mid \lambda)=\frac{\lambda^{x}}{x!}\mathrm{e}^{-\lambda};\quad x=0,1,2,\cdots,\infty$$

泊松分布是接受非负整数值的单参数离散分布。参数 λ 既是分布的均值，也是分布的方差。因此，随着泊松随机数的特定样本中的数字变大，数字的变异性也变大。

泊松分布是二项分布的极限情况，其中，N 趋向无穷大，p 趋向零，而 $N_p=\lambda$。

泊松分布和指数分布是相关的。如果计数的数量遵循泊松分布，则单个计数之间的间隔遵循指数分布。

【例 5-8】 计算并绘制参数 $\lambda=5$ 的泊松分布的 pdf。

```
>> x = 0:15;
y = poisspdf(x,5);
plot(x,y,'+')
```

运行程序，效果如图 5-8 所示。

4. 离散型均匀分布

均匀分布是一种简单的概率分布，分为离散型均匀分布和连续型均匀分布。此处介绍的是离散型均匀分布。

离散型均匀分布是一个简单的分布，它对从 1 到 N 的整数赋予相等的权重。离散型均匀分布的概率公式为：

$$y=f(x \mid N)=\frac{1}{N}I_{(1,\cdots,N)}(x)$$

【例 5-9】 绘制离散型均匀分布 cdf。

对于所有离散分布，cdf 是一个阶跃函数。图 5-9 显示了 $N=10$ 的离散型均匀分布 cdf。

```
>> x = 0:10;
y = unidcdf(x,10);
figure;
stairs(x,y)
h = gca;
h.XLim = [0 11];
```

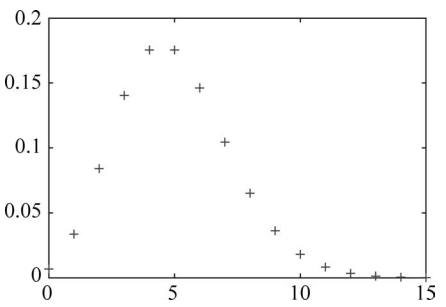

图 5-8　泊松分布的概率分布图

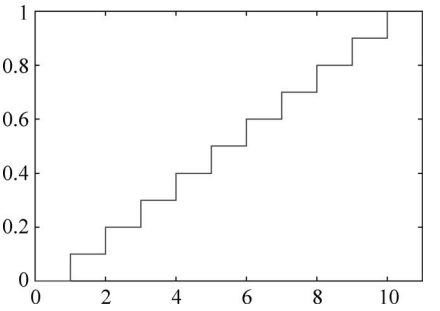

图 5-9　离散型均匀分布的 cdf 图

5.2.2　连续分布

连续型随机变量 X 的分布函数是连续的，它对应的分布为连续分布。常用的连续分布有正态分布、均匀分布、指数分布、伽马分布、贝塔分布等。其中，正态分布是最常用的连续分布，如测量误差、人的身高、年降雨量等都可用正态分布描述。

本节对几个常用的连续分布展开介绍。

1. 正态分布

正态分布，有时也称为高斯分布，是双参数曲线族。使用正态分布建模的通常理由是中心极限定理，该定理(粗略地)指出，随着样本大小趋向无穷大，来自任何具有有限均值和方差的分布的独立样本总和会收敛为正态分布。

统计工具箱提供了以下几种处理正态分布的方法。

(1) 通过对样本数据进行概率分布拟合(fitdist)或通过指定参数值(makedist)来创建概率分布对象 NormalDistribution。然后使用对象函数来计算分布、生成随机数等。

(2) 使用 Distribution Fitter App 以交互方式处理正态分布。

(3) 将分布特定的函数(normcdf、normpdf、norminv、normlike、normstat、normfit、normrnd)与指定的分布参数结合使用。分布特定的函数可以接受多个正态分布的参数。

(4) 将一般分布函数(cdf、icdf、pdf、random)与指定的分布名称('Normal')和参数结合使用。

1) 参数估计

最大似然估计(MLE)是最大化似然函数的参数估计。正态分布的 μ 和 σ^2 的最大似然估计量分别是：

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

和

$$s^2_{\mathrm{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

其中，$\bar{x}$ 是样本 $x_1, x_2, \cdots, x_n$ 的样本均值。样本均值是参数 μ 的无偏估计量。但是，s^2_{MLE} 是

参数 σ^2 的有偏估计量，这意味着其预期值不等于参数。

最小方差无偏估计量(MVUE)通常用于估计正态分布的参数。MVUE 是参数的所有无偏估计量中方差最小的估计量。正态分布的参数 μ 和 σ^2 的 MVUE 分别是样本均值 $\bar{x}$ 和样本方差 s^2。

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

要对数据进行正态分布拟合并求出参数估计值，可使用 normfit、fitdist 或 mle。

(1) 对于未删失数据，normfit 和 fitdist 计算无偏估计值，mle 计算最大似然估计值。

(2) 对于删失数据，normfit、fitdist、mle 计算最大似然估计值。

与返回参数估计值的 normfit 和 mle 不同，fitdist 返回拟合的概率分布对象 NormalDistribution。对象属性 μ 和 σ 存储参数估计值。

2) 概率密度函数

正态概率密度函数是：

$$y=f(x\mid\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{\frac{-(x-\mu)^2}{2\sigma^2}},\quad x\in\Re$$

似然函数是被视为参数函数的 pdf。最大似然估计(MLE)是最大化 x 的固定值的似然函数的参数估计。

3) 累积分布函数

正态累积分布函数表示为：

$$p=F(x\mid\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}\frac{-(t-\mu)^2}{2\sigma^2}\mathrm{d}t,\quad x\in\Re$$

p 是参数为 μ 和 σ 的正态分布中的一个观测值落入 $(-\infty,x]$ 区间的概率。标准正态累积分布函数 $\Phi(x)$ 在功能上与误差函数 erf 相关。

$$\Phi(x)=\frac{1}{2}\left(1-\mathrm{erf}\left(-\frac{x}{\sqrt{2}}\right)\right)$$

其中，

$$\mathrm{erf}(x)=\frac{2}{\sqrt{\pi}}\int_{0}^{x}\mathrm{e}^{-t^2}\mathrm{d}t=2\Phi(\sqrt{2}x)-1$$

【例 5-10】 拟合正态分布对象。

```
>> % 加载样本数据并创建包含学生考试成绩数据的第一列的向量
load examgrades
x = grades(:,1);
 % 通过对数据进行正态分布拟合来创建正态分布对象
pd = fitdist(x,'Normal')
pd =
  NormalDistribution
  正态 分布
      mu = 75.0083  [73.4321, 76.5846]
   sigma =  8.7202  [7.7391, 9.98843]
```

参数估计值旁边的区间是分布参数的 95%置信区间。

2. 指数分布

在概率理论和统计学中，指数分布(也称为负指数分布)是描述泊松过程中的事件之间的时间的概率分布，即事件以恒定平均速率连续且独立地发生的过程。指数分布与分布指数族

的分类不同，后者是包含指数分布作为其成员之一的大类概率分布，也包括正态分布、二项分布、伽马分布、泊松分布等。

指数函数的一个重要特征是无记忆性(Memoryless Property，又称遗失记忆性)。这表示如果一个随机变量呈指数分布，当 $s,t\geqslant 0$ 时，有 $P(T>s+t\mid T>t)=P(T>s)$，即，如果 T 是某一元件的寿命，已知元件使用了 t 小时，它总共使用至少 $s+t$ 小时的条件概率，与从开始使用时算起它使用至少 s 小时的概率相等。

指数分布的概率密度公式为：

$$f(x)=\begin{cases}\lambda e^{-\lambda_x}, & x>0\\ 0, & x\leqslant 0\end{cases}$$

其中，$\lambda>0$ 是分布的一个参数，常被称为率参数(rate parameter)，即每单位时间内发生某事件的次数。指数分布的区间是$[0,\infty)$。如果一个随机变量 X 呈指数分布，则可以写作$X\sim$ Exponential(λ)。

累积分布函数为：

$$F(x;\lambda)=\begin{cases}1-e^{-\lambda x}, & x\geqslant 0\\ 0, & x<0\end{cases}$$

【例 5-11】 将指数分布拟合到数据。

```
>> % 生成 100 个平均值为 700 的指数分布随机数样本
x = exprnd(700,100,1);                                  % 生成样本
 % 使用 fitdist 将指数分布拟合到数据
pd = fitdist(x,'exponential')
pd =
  ExponentialDistribution
  指数 分布
    mu = 661.084   [548.486, 812.502]
```

fitdist 返回一个指数分布对象。参数估计值旁边的区间是分布参数的 95%置信区间。

```
>> % 使用分布函数估计参数
[muhat,muci] = expfit(x)                                % 分布特定函数
muhat =
  661.0843
muci =
  548.4859
  812.5023
>> [muhat2,muci2] = mle(x,'distribution','exponential') % 通用分布函数
muhat2 =
  661.0843
muci2 =
  548.4859
  812.5023
```

5.3　假 设 检 验

统计工具箱中提供参数化假设检验和非参数化假设检验，帮助确定样本数据是否来自具有特定特征的总体。

(1) 分布检验(如 Anderson-Darling 检验和单样本 Kolmogorov-Smirnov 检验)可以检验样本数据是否来自具有特定分布的总体。双样本 Kolmogorov-Smirnov 检验可以检验两组样本数据是否具有相同的分布。

(2) 位置检验(如 z 检验和单样本 t 检验)可以检验样本数据是否来自具有特定均值或中位数的总体。双样本 t 检验或多重比较检验可以检验两组或多组样本数据是否具有相同的位置值。

(3) 散度检验(如卡方方差检验)可以检验样本数据是否来自具有特定方差的总体。双样本 F 检验或多样本检验可以比较两个或多个样本数据集的方差。

5.3.1 K-S 检验

Kolmogorov-Smirnov 检验(K-S 检验)基于累积分布函数,用以检验一个经验分布是否符合某种理论分布或比较两个经验分布是否有显著性差异。

两样本 K-S 检验由于对两样本的经验分布函数的位置和形状参数的差异都敏感而成为比较两样本有用且常规的非参数方法之一。

K-S 检验的累积分布函数为:

$$F_n(x)=\frac{1}{n}\sum_{i=1}^{n}I_{|-\infty,x|}(X_i)$$

其中,$I_{|-\infty,x|}$ 为指示函数:

$$I_{|-\infty,x|}(X_i)=\begin{cases}1, & X_i\leqslant x\\ 0, & X_i>x\end{cases}$$

对于一个样本集的累积分布函数 $F_n(x)$和一个假设的理论分布 $F(x)$,K-S 定义为:

$$D_n=\sup_x|F_n(x)-F(x)|$$

$\sup_x$ 是距离的上确界(supremum),如果 X_i 服从理论分布 $F(x)$,则当 n 趋于无穷时,D_n 趋于 0。

在 MATLAB 中,提供了 kstest 函数用来做单个样本的 Kolmogorov-Smirnov 检验;它可以做双侧检验,检验样本是否服从指定的分布;也可以做单侧检验,检验样本的分布函数是否在指定的分布函数之上或之下。

kstest 函数的调用格式如下。

h=kstest(x):检验样本 x 是否服从标准正态分布,原假设是 x 服从标准正态分布,对立假设是 x 不服从标准正态分布。当输出 $h=1$ 时,在显著性水平 $\alpha=0.05$ 下拒绝原假设;当 $h=0$ 时,则在显著性水平 $\alpha=0.05$ 下接受原假设。

h=kstest(x,CDF):检验样本 x 是否服从由 CDF 定义的连续分布。这里的 CDF 可以是包含两列元素的矩阵,也可以是概率分布对象,如 ProbDistUnivParam 类对象或 ProbDistUnivKernel 类对象。当 CDF 是包含两列元素的矩阵时,它的第 1 列表示随机变量的可能取值,可以是样本 x 中的值,也可以不是,但是样本 x 中的所有值必须在 CDF 的第 1 列元素的最小值与最大值之间。CDF 的第 2 列是指定分布函数 $G(x)$的取值。如果 CDF 为空(即[]),则检验样本 x 是否服从标准正态分布。

h=kstest(x,CDF,alpha):指定检验的显著性水平 alpha,默认值为 0.05。

h=kstest(x,CDF,alpha,type):用 type 参数指定检验的类型(双侧或单侧)。type 参数的可能取值如下。

(1) 当 type='unequal'时即为双侧检验,对立假设是总体分布函数不等于指定的分布函数。

(2) 当 type='larger'时为单侧检验,对立假设是总体分布函数大于指定的分布函数。

（3）当 type='smaller'时为单侧检验，对立假设是总体分布函数小于指定的分布函数。其中，后两种情况下算出的检验统计量不用绝对值。

[h,p,ksstat,cv]=kstest(…)：返回检验的 p 值、检验统计量的观测值 ksstat 和临界值 cv。

【例 5-12】 在 20 天内，从维尼纶正常生活时的生产报表中看到的维尼纶纤度（纤维的粗细程度的一种度量）的情况，有如下 100 个数据。

1.36,1.49,1.43,1.41,1.37,1.40,1.32,1.43,1.47,1.39,
1.41,1.36,1.40,1.34,1.42,1.42,1.45,1.35,1.42,1.39,
1.44,1.42,1.39,1.42,1.42,1.30,1.34,1.42,1.37,1.36,
1.37,1.34,1.37,1.37,1.44,1.45,1.32,1.48,1.40,1.45,
1.39,1.46,1.39,1.53,1.36,1.48,1.40,1.39,1.38,1.40,
1.36,1.45,1.50,1.43,1.38,1.43,1.41,1.48,1.39,1.45,
1.37,1.37,1.39,1.45,1.31,1.41,1.44,1.44,1.42,1.42,
1.35,1.36,1.39,1.40,1.38,1.35,1.42,1.43,1.42,1.42,
1.42,1.40,1.41,1.37,1.46,1.36,1.37,1.27,1.37,1.38,
1.42,1.34,1.43,1.42,1.41,1.41,1.44,1.48,1.55,1.37.

试根据这 100 个样本数据在 0.10 显著性水平下，用 Kolmogorov-Smirnov 检验对维尼纶纤度数据进行正态性检验。（将数据保存到 data.txt 中。）

解析：检验的原假设是维尼纶纤度服从正态分布。

其实现的 MATLAB 代码如下。

```
>> clear all;
X = load('data.txt');          % 加载数据
[mu,sigma] = normfit(X)
x = (X - mu)/sigma;
[h,p,stats,cv] = kstest(x,[],0.10,0)
```

运行程序，输出如下。

```
mu =
     1.4038
sigma =
        0.0474
h =
    logical
    0
p =
    0.2951
stats =
       0.2931
cv =
     0.3687
```

结果表明，接受原假设，即认为维尼纶纤度服从均值为 1.4038、标准差为 0.0474 的正态分布。

此外，MATLAB 还提供了 kstest2 函数用来做两个样本的 Kolmogorov-Smirnov 检验，它可以做双侧检验，检验两个样本是否服从相同的分布，也可以做单侧检验，检验一个样本的分布函数是否在另一个样本的分布函数之上或之下。函数的调用格式如下。

h=kstest2(x1,x2)：检验样本 x_1 与 x_2 是否具有相同的分布，原假设是 x_1 与 x_2 来自相

同的连续分布，对立假设是来自于不同的连续分布。当输出 $h=1$ 时，在显著性水平 $\alpha=0.05$ 下拒绝原假设；当 $h=0$ 时，则在显著性水平 $\alpha=0.05$ 下接受原假设。这里并不要求 x_1 与 x_2 具有相同的长度。

h=kstest2(x1,x2,alpha,type)：指定检验的显著性水平 alpha，默认值为 0.05；并用参数 type 指定检验的类型(双侧或单侧)。type 参数的可能取值如下。

(1) 当 type='unequal'时即为双侧检验，对立假设是两个总体的分布函数不相等。

(2) 当 type='larger'时即为单侧检验，对立假设是第 1 个总体的分布函数大于第 2 个总体的分布函数。

(3) 当 type='smaller'时即为单侧检验，对立假设是第 1 个总体的分布函数小于第 2 个总体的分布函数。

[h,p]=kstest2(…)：返回检验的渐近 p 值，当 p 值小于或等于给定的显著性水平 alpha 时，拒绝原假设。样本容量越大，p 值越精确，通常要求：

$$\frac{n_1 n_2}{n_1+n_2} \geqslant 4$$

其中，n_1，n_2 分别为样本 x_1 和 x_2 的样本容量。

[h,p,ks2stat]=kstest2(…)：返回检验统计量的观测值 ks2stat。

【例 5-13】 利用 kstest2 函数对创建的标准正态随机分布检验是否接受原假设，并绘制其分布曲线图。

```
>> clear all;
x = -1:1:5;
y = randn(20,1);
[h,p,k] = kstest2(x,y)
% 以下代码用于绘制测试统计图
F1 = cdfplot(x);
hold on
F2 = cdfplot(y);
set(F1,'LineWidth',2,'Color','r')
set(F2,'LineWidth',2)
legend([F1 F2],'F1(x)','F2(x)','Location','NW')
```

运行程序，输出如下，效果如图 5-10 所示。

```
h =
     0
p =
    0.2387
k =
    0.4214
```

结果表明，由于 $h=0$，所以在默认显著性下接受原假设。

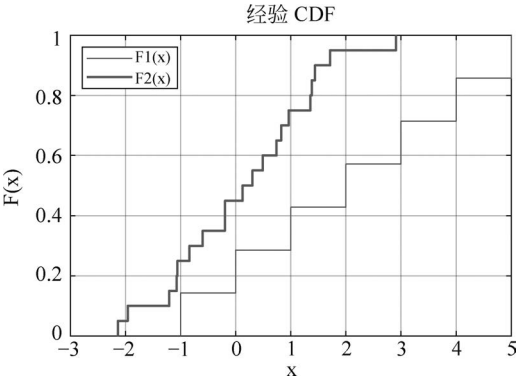

图 5-10　测试统计图

5.3.2　t 检验

t 检验，也称为 student t 检验(Student's t test)，主要用于样本含量较小(例如 $n<30$)、总体标准差 σ 未知的正态分布资料。

单样本 t 检验是当总体标准差未知时位置参数的参数化检验。检验统计量的计算公式为：

$$t=\frac{\bar{x}-\mu}{s/\sqrt{n}}$$

其中，$\bar{x}$ 是样本均值，μ 是假设的总体均值，s 是样本标准差，n 是样本大小。在原假设下，检验统计量具有 Student t 分布和 $n-1$ 个自由度。

在 MATLAB 中，提供了 ttest 函数用于实现单样本 t 检验。函数的调用格式如下。

h=ttest(x)：使用单样本 t 检验返回原假设的检验决策，该原假设假定 x 中的数据来自均值等于零且方差未知的正态分布。备择假设是总体分布的均值不等于零。如果检验在 5% 的显著性水平上拒绝原假设，则结果 h 为 1，否则为 0。

h=ttest(x,y)：使用配对样本 t 检验返回针对原假设的检验决策，该原假设假定 $x-y$ 中的数据来自均值等于零且方差未知的正态分布。

h=ttest(x,y,Name,Value)：返回配对样本 t 检验的检验决策，其中使用由一个或多个名称-值对组参数指定附加选项。

h=ttest(x,m)：返回针对原假设的检验决策，该原假设假定 x 中的数据来自均值为 m 且方差未知的正态分布。备择假设是均值不为 m。

h=ttest(x,m,Name,Value)：返回单样本 t 检验的检验决策，其中使用一个或多个名称-值对组参数指定附加选项。

[h,p]=ttest(____)：还使用上述语法组中的任何输入参数返回检验的 p 值。

[h,p,ci,stats]=ttest(____)：还返回 x（对于配对 t 检验则为 $x-y$）的均值的置信区间 ci，以及包含检验统计量信息的结构体 stats。

【例 5-14】 某种电子元件的寿命 X（以小时计）服从正态分布，μ、σ^2 均未知。现测得 16 只元件的寿命如下。

```
160  278  198  200  236  257  270  167  150  250  194  224  137
185  167  255
```

问是否有理由认为元件的平均寿命大于 220（小时）？

其实现的 MATLAB 代码如下。

```
>> clear all;
X = [160  278  198  200  236  257  270  167  150  250  194  224  137  185  167  255];
[h,p,ci] = ttest(X,220,0.005,1)
```

运行程序，输出如下。

```
h =
     0
p =
    0.8457
ci =
  174.4465         Inf              % 均值 225 在该置信区间内
```

结果表明，$h=0$ 表示在水平 $\alpha=0.05$ 下应该接受原假设 h_0，即认为元件的平均寿命不大于 220 小时。

5.3.3　双样本 t 检验

根据方差齐与不齐两种情况，应用不同的统计量进行检验。

方差不齐时，检验统计量为：

$$t=-\frac{\bar{x}-\bar{y}}{\sqrt{\frac{S_x^2}{m}+\frac{S_y^2}{n}}}$$

式中，$\bar{x}$ 和 $\bar{y}$ 表示样本 1 和样本 2 的均值；S_x^2 和 S_y^2 为样本 1 和样本 2 的方差；m 和 n 为样本 1 和样本 2 的数据个数。

方差齐时，检验统计量为：

$$t=-\frac{\bar{x}-\bar{y}}{S_w\sqrt{\frac{1}{m}+\frac{1}{n}}}$$

式中，S_w 为两个样本的标准差，它是样本 1 和样本 2 的方差的加权平均值的平方根，为：

$$S_W=\sqrt{\frac{(m-1)S_x^2+(n-1)S_y^2}{m+n+1}}$$

在不假设两个数据样本来自具有方差齐性的总体的情况下，原假设下的检验统计量具有近似 Student t 分布，其自由度的数目由 Satterthwaite 逼近给出。此检验有时称为 Welch t 检验。

在 MATLAB 中，提供了 ttest2 函数用于实现双样本 t 检验。函数的调用格式如下。

h=ttest2(x,y)：使用双样本 t 检验返回原假设的检验决策，该原假设假定向量 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 中的数据来自均值相等、方差相同但未知的正态分布的独立随机样本。备择假设是 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 中的数据来自均值不相等的总体。如果检验在 5%的显著性水平上拒绝原假设，则结果 h 为 1，否则为 0。

h=ttest2(x,y,Name,Value)：返回针对双样本 t 的检验决策，该检验使用由一个或多个名称-值对组参数指定的附加选项。例如，可以更改显著性水平或进行无须假设方差齐性的检验。

[h,p]=ttest2(____)：还使用上述语法中的任何输入参数返回检验的 p 值。

[h,p,ci,stats]=ttest2(____)：还返回总体均值差的置信区间 ci，以及包含检验统计量信息的结构体 stats。

【例 5-15】 下面分别给出文学家马克·吐温的 8 篇小品文以及斯诺德格拉斯的 10 篇小品文中的 3 个字母组成的单词的比例。

马克·吐温　0.225 0.262 0.217 0.240 0.230 0.229 0.235 0.217

斯诺德格拉斯　0.209 0.205 0.196 0.210 0.202 0.207 0.224 0.223 0.220 0.201

设两组数据分别来自正态总体，且两总体方差相等，但参数均未知。两样本相互独立，问两个作家所写的小品文中包含由 3 个字母组成的单词的比例是否有显著差异？零假设为两个作家对应的比例没有显著差异。

其实现的 MATLAB 代码如下。

```
>> clear all;
x = [0.225 0.262 0.217 0.240 0.230 0.229 0.235 0.217];
y = [0.209 0.205 0.196 0.210 0.202 0.207 0.224 0.223 0.220 0.201];
[h,signnificance,ci] = ttest2(x,y)
```

运行程序，输出如下。

```
h =
     1
signnificance =
```

```
    0.0013
ci =
    0.0101    0.0343
```

结果表明，$h=1$，拒绝零假设，认为两个作家所写小品文中包含 3 个字母组成的单词的比例有显著差异。

5.4　方 差 分 析

方差分析（ANOVA）是指确定响应变量的变异是出现在总体组内还是出现在不同总体组之间的过程。统计工具箱提供了单因素/双因素/N 因素方差分析（ANOVA）、多元方差分析（MANOVA）、重复测量模型以及协方差分析（ANCOVA）。

5.4.1　方差的基本原理

方差分析的基本原理是认为不同处理组的均数间的差别基本来源有以下两个。

（1）随机误差，如测量误差造成的差异或个体间的差异，称为组内差异，用变量在各组的均值与该组内变量值的偏差平方和的总和表示，记作 SS_w，组内自由度 df_w。

（2）实验条件，即不同的处理造成的差异，称为组间差异。用变量在各组的均值与总均值之偏差平方和表示，记作 SS_b，组间自由度 df_b。

总偏差平方和 $SS_t=SS_w+SS_b$。

组内 SS_w、组间 SS_b 除以各自的自由度（组内 $df_w=n-m$，组间 $df_b=m-1$，其中，n 为样本总数，m 为组数），得到其均方 MS_w 和 MS_b，一种情况是处理没有作用，即各组样本均来自同一总体，$MS_b/MS_w\approx1$。另一种情况是处理确实有作用，组间均方是由于误差与不同处理共同导致的结果，即各样本来自不同总体。那么，$MS_b>MS_w$。

MS_b/MS_w 比值构成 F 分布。用 F 值与其临界值比较，推断各样本是否来自相同的总体。

5.4.2　单因素方差分析

单因素方差分析是指对单因素实验结果进行分析，检验因素对实验结果有无显著性影响的方法。它是用来研究一个控制变量的不同水平是否对观测变量产生了显著影响。这里，由于仅研究单个因素对观测变量的影响，因此称为单因素方差分析。

例如，分析不同施肥量是否给农作物产量带来显著影响，考察地区差异是否影响妇女的生育率，研究学历对工资收入的影响等。这些问题都可以通过单因素方差分析得到答案。

单向方差分析是线性模型的一个简单特例。模型的单向方差分析形式为：

$$y_{ij}=a_j+\varepsilon_{ij}$$

其中，y_{ij} 是一个观察值，i 代表观察值，j 代表预测变量 y 的不同组（水平）。所有 y_{ij} 都是独立的。a_j 代表第 j 组（水平）的总体平均值。ε_{ij} 是独立的正态分布随机误差，具有零均值和常数方差，即 $\varepsilon_{ij}\sim N(0,\sigma^2)$。

这个模型也称为均值模型。该模型假设 y 列为常数 a_j 加上误差分量 ε_{ij}。方差分析有助于确定这些常数是否都相同。

方差分析检验了所有组均数相等的假设，以及至少一个组与其他组不同的替代假设。

$$H_0:a_1=a_2=\cdots=a_k$$

H_1：并非所有组为均组

方差分析基于所有样本总体均为正态分布的假设。众所周知，它对适度违反这一假设具有鲁棒性。可以使用法线图(normclot)直观地检查法线假设。或者，可以使用统计和机器学习工具箱中的一个函数来检查正态性：Anderson-Darling 检验(adtest)、卡方拟合优度检验(chi2gof)、Jarque-Bera 检验(jbtest)或 Lilliefors 检验(Lilliefest)。

在 MATLAB 中，提供了 anova1 函数实现单因素方差分析。anova1 主要是比较多组数据的均值，然后返回这些均值相等的概率，从而判断这一因素是否对实验指标有显著影响。函数的调用格式如下。

p＝anova1(X)：为零假设存在的概率，一般 p 小于 0.05 或 0.01 时，认为结果显著(零假设可疑)。

p＝anova1(X,group)：当 X 为矩阵时，利用 group 变量作为 X 中样本箱线图的标签。

p＝anova1(X,group,displayopt)：displayopt 为 on 时，则激活 anova1 表和箱线图的显示。

[p,table]＝anova1(…)：返回单元数组表中的 anova1 表。

[p,table,stats]＝anova1(…)：返回 stats 结构，用于多元比较检验。

【例 5-16】 有 A、B、C、D、E、F 这 6 个小麦品种产量的比较实验，设置标准品种 CK，采用 3 次重复的对比设计，所得产量结果如表 5-2 所示。

表 5-2　6 个小麦品种产量结果分析

品种	各重复的产量/kg		
	Ⅰ	Ⅱ	Ⅲ
A	560	582	520
B	582	565	525
C	600	600	572
D	525	496	590
E	560	578	615
F	640	662	508
CK	500	510	519

利用 anova1 实现方差分析，代码如下。

```
>> clear al;
X = [560 582 600 525 560 640 500;582 565 600 496 578 662 510;520 525 572 590 615 508 519];
group = {'A','B','C','D','E','F','CK'};
[p,table,stats] = anova1(X,group)
```

运行程序，输出如下，效果如图 5-11 及图 5-12 所示。

```
p =
    0.1602
```

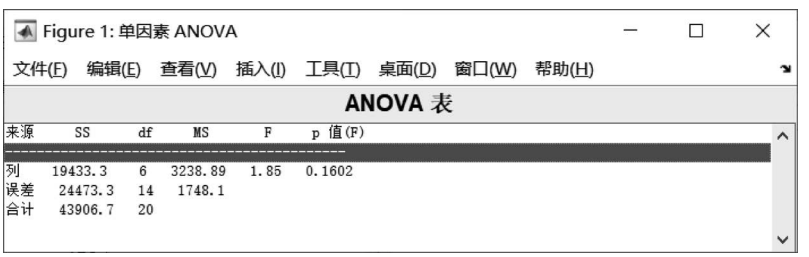

图 5-11　ANOVA 表

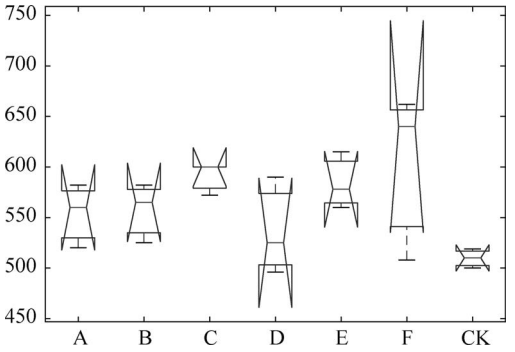

图 5-12　方差盒形图

5.4.3　双因素方差分析

双因素方差分析法是一种统计分析方法，这种分析方法可以用来分析两个因素的不同水平对结果是否有显著影响，以及两因素之间是否存在交互效应。一般运用双因素方差分析法，先对两个因素的不同水平的组合进行设计实验，要求每个组合下所得到的样本的含量都是相同的。

在实际问题的研究中，有时需要考虑两个因素对实验结果的影响。例如饮料销售，除了关心饮料品牌，还想了解销售地区是否影响销售量，如果在不同的地区，销售量存在显著的差异，就需要分析原因。采用不同的销售策略，使该饮料品牌在市场占有率高的地区保持领先地位；在市场占有率低的地区进一步扩大宣传。如果把饮料的品牌看作影响销售量的因素 A，饮料的销售地区则是影响因素 B。对因素 A 和因素 B 同时进行分析，就属于双因素方差分析的内容，双因素方差分析是对影响因素进行检验：究竟是一个因素在起作用，还是两个因素都起作用，或是两个因素的影响都不显著。

1. 双因素方差分析法的类型

双因素方差分析有以下两种类型。

(1) 一个是无交互作用的双因素方差分析，它假定因素 A 和因素 B 的效应之间是相互独立的，不存在相互关系。

(2) 另一个是有交互作用的双因素方差分析，它假定因素 A 和因素 B 的结合会产生出一种新的效应。

例如，如果假定不同地区的消费者对某种颜色有与其他地区消费者不同的特殊偏爱，这就是两个因素结合产生的新效应，属于有交互作用的背景；否则，就是无交互作用的背景。

2. 双因素方差的模型

双向方差分析是线性模型的特例。模型的双向方差分析形式为：

$$y_{ijr}=\mu+a_i+\beta_j+(a\beta)_{ij}+\varepsilon_{ij}$$

式中，

- y_{ijr} 是对响应变量的观察。i 表示行因子 A 的组 i，$i=1,2,\cdots,I$；j 表示列因子 B 的组 j，$j=1,2,\cdots,J$；r 表示复制数，$r=1,2,\cdots,R$，即总共有 $N=I\times J\times R$ 个观测值。
- μ 为总体平均值。
- a_i 是由于行因子 B 导致的行因子 A 中的组与总体平均值 μ 的偏差。a_i 的值的和为 0。

$$\sum_{i=1}^{I} a_i = 0$$

- β_j 是由于行因子 B 导致的列因子 B 中的组与总体平均值 μ 的偏差。给定列 β_j 中的所有值均相同，且 β_j 的值总和为 0。

$$\sum_{j=1}^{J} \beta_j = 0$$

- $(a\beta)_{ij}$ 是由于行因子 B 导致的列因子 B 中的组与总体平均值 μ 的偏差。β_j 的给定列中的所有值都是相同的，并且 β_j 的值总和为 0。

$$\sum_{i=1}^{I} (a\beta)_{ij} = \sum_{j=1}^{J} (a\beta)_{ij} = 0$$

- ε_{ij} 是随机干扰。假设它们是独立的、正态分布的，并且具有恒定的方差。

在 MATLAB 统计工具箱中，提供了 anova2 函数用于实现无交互作用的双因素方差分析。函数的调用格式如下。

p=anova2(X,reps)：根据样本观测值矩阵 ***X*** 进行均衡实验的双因素一元方差分析。***X*** 的每一列对应因素 A 的一个水平，每行对应因素 B 的一个水平，***X*** 还应满足方差分析的基本假定。reps 表示因素 A 和 B 的每一个水平组合下重复实验的次数。

p=anova2(X,reps,displayopt)：当 displayopt 为 on 时，则显示方差分析表和箱线图。

[p,table]=anova2(…)：返回单元数组表中的 ANOVA 表。

[p,table,stats] = anova2(…)：返回 stats 结构，用于多元检验。

【例 5-17】 执行双向 ANOVA 以确定汽车型号和工厂对汽车里程等级的影响。

```
>> % 加载并显示示例数据
load mileage
mileage
mileage =
  33.3000  34.5000  37.4000
  33.4000  34.8000  36.8000
  32.9000  33.8000  37.6000
  32.6000  33.4000  36.6000
  32.5000  33.7000  37.0000
  33.0000  33.9000  36.7000
```

返回的结果中，有三个车型(列)和两个工厂(行)。该数据有六个里程行，因为每个工厂为研究提供了每种型号的三辆汽车(即复制数为三)。第一个工厂的数据在前三行，第二个工厂的数据在后三行，进行双向方差分析。

```
>> nmbcars = 3;                                  % 每个型号的汽车数量,即复制次数
[~,~,stats] = anova2(mileage,nmbcars);           % 效果如图 5-13 所示
```

执行多重比较以找出三种车型中哪一对有显著差异。

```
>> c = multcompare(stats)                        % 效果如图 5-14 所示
```

注意：模型中包含一个交互效应项。当模型中包括交互效应时，主效应检验可能很难解释。

```
c =
    1.0000    2.0000   -1.5865   -1.0667   -0.5469    0.0004
    1.0000    3.0000   -4.5865   -4.0667   -3.5469    0.0000
    2.0000    3.0000   -3.5198   -3.0000   -2.4802    0.0000
```

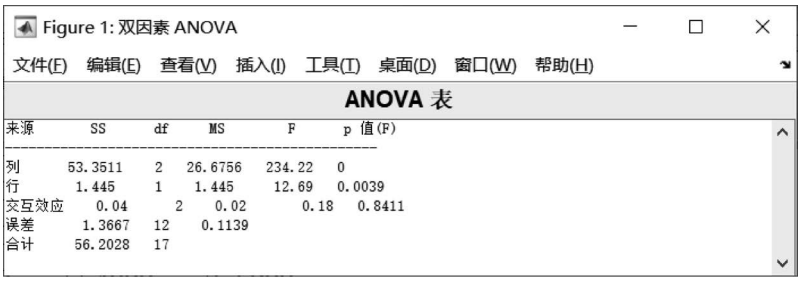

图 5-13　ANOVA 表

在矩阵 c 中，前两列显示了比较的汽车模型对，最后一列显示了检验的 p 值。所有 p 值都很小(0.0004、0 和 0)，这表明所有车型的平均里程彼此之间存在显著差异。

在图 5-14 中，蓝色线为第一款车型平均里程的比较区间，红色条是第二款和第三款车型平均里程的比较区间。第二和第三比较区间均不与第一比较区间重叠，表明第一车型的平均里程与第二和第三车型的平均里程不同。如果单击其他栏之一，可以测试其他车型。如果没有一个比较区间重叠，表明每个车型的平均里程与其他两个有显著差异。

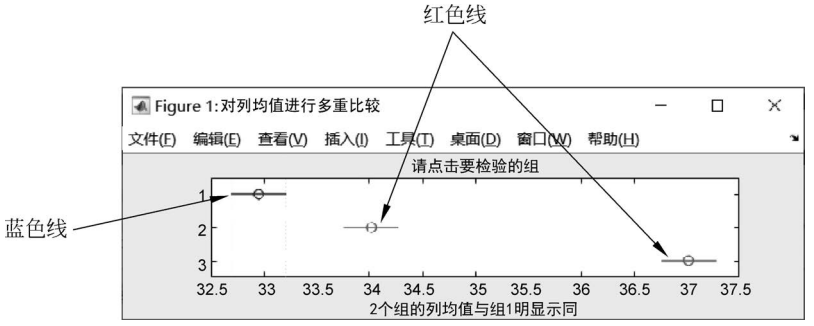

图 5-14　多重比较界面

5.4.4　多因素方差分析

多因素方差分析，用于研究一个因变量是否受到多个自变量(也称为因素)的影响，它检验多个因素取值水平的不同组合之间、因变量的均值之间是否存在显著的差异。多因素方差分析既可以分析单个因素的作用(主效应)，也可以分析因素之间的交互作用(交互效应)，还可以进行协方差分析，以及各个因素变量与协变量的交互作用。

多因素方差分析是两因素方差分析的一般形式，对三个因素的情况，其模型表达式为：

$$y_{ijkl}=\mu+\alpha_{.j.}+\beta_{i..}+\gamma_{..k}+(\alpha\beta)_{ij.}+(\alpha\gamma)_{i.k}+(\beta\gamma)_{.jk}+(\alpha\beta\gamma)_{ijk}+\varepsilon_{ijkl}$$

式中两个连在一起的标记，如$(\alpha\beta)_{ij.}$，表示两个因素之间的交互作用，参数$(\alpha\beta\gamma)_{ijk}$ 表示三个因素之间的交互作用。

在 MATLAB 中，提供了 anovan 函数用于实现多因素方差分析。函数的调用格式如下。

p=anovan(y,group)：根据样本观测值向量 y 进行均衡或非均衡实验的多因素一元方差分析，检验多个因素的主效应是否显著。输入参数 group 为一个元胞数组，它的每一个元素对应一个因素，是该因素的水平列表，与 y 等长，用来标记 y 中每个观测所对应的因素的水平。每个元胞中因素的水平列表可以是一个分类(categorical)数组、数值向量、字符矩阵或单列的字符串元胞数组。输出参数 p 是检验的 p 值向量，p 中的每个元素对应一个主效应。

p=anovan(y,group,param,val)：通过指定一个或多个成对出现的参数名与参数值来控

制多因素一元方差分析。

[p,table]=anovan(y,group,param,val)：同时返回元胞数组形式的方差分析表 table(包含列标签和行标签)。

[p,table,stats]=anovan(y,group,param,val)：同时返回一个结构体变量 stats,用于进行后续的多重比较。当某因素对实验指标的影响显著时,在后续的分析中,可以调用 multcompare 函数,把 stats 作为它的输入,进行多重比较。

[p,table,stats,terms]=anovan(y,group,param,val)：同时返回方差分析计算中的主效应项和交互效应项矩阵 terms。terms 的格式与'model'参数的最后一种取值的格式相同。当'model'参数的取值为一个矩阵时,anovan 函数返回的 terms 就是这个矩阵。

【例 5-18】 显示了如何对 1970—1982 年生产的 406 辆汽车的里程和其他信息的汽车数据进行 N 向方差分析。

```
% 加载样本数据
>> load carbig
```

该实例侧重于四个变量,MPG 是 406 辆汽车每加仑行驶的英里数,其他三个变量是因素：cyl4(是否为四缸汽车)、org(汽车起源于欧洲、日本或美国),以及何时(汽车在该时期的早期、中期或期末)。

```
% 拟合完整模型,要求最多三向交互和类型 3 平方和
>> varnames = {'Origin';'4Cyl';'MfgDate'};
anovan(MPG,{org cyl4 when},3,3,varnames)    % 效果如图 5-15 所示
ans =
    0.0000
       NaN
    0.0000
    0.7032
    0.0001
    0.2072
    0.6990
```

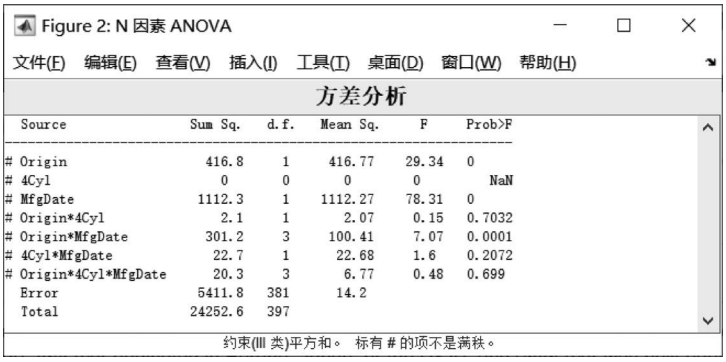

图 5-15 N 因素方差分析表

请注意,许多术语用 # 符号标记为没有满秩,其中一个术语的自由度为零,缺少 p 值。当缺少因子组合且模型具有高阶项时,可能会发生这种情况。在这种情况下,下面的交叉表显示,在这一时期的早期,除了四缸,没有欧洲制造的汽车,如 tbl(2,1,1)中的 0 所示。

```
>> [tbl,chi2,p,factorvals] = crosstab(org,when,cyl4)
tbl(:,:,1) =
    82    75    25
    0     4     3
```

```
    3     3     4
tbl(:,:,2) =
    12    22    38
    23    26    17
    12    25    32
chi2 =
  207.7689
p =
  8.0973e-38
factorvals =
  3×3 cell 数组
    {'USA'}    {'Early'}    {'Other'}
  {'Europe'}    {'Mid'}    {'Four'    }
  {'Japan'}    {'Late'}    {0×0 double}
```

由结果可以看出，不可能估计三向相互作用效应，并且在模型中包含三向相互作用项会使拟合奇异。即使使用 ANOVA 表中有限的可用信息，也可以看到三向交互作用的 p 值为 0.699，因此不显著。

只检查双向互动，如图 5-16 所示。

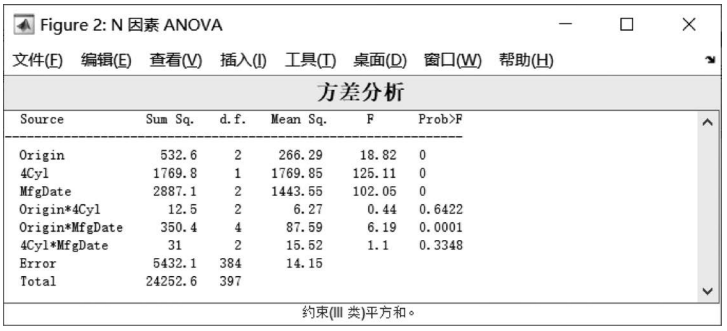

图 5-16　约束平方和方差表

```
>> terms
terms =
    1     0     0
    0     1     0
    0     0     1
    1     1     0
    1     0     1
    0     1     1
```

现在所有的项都是可估计的。相互作用项 4（原点×4Cyl）和相互作用项 6（6Cyl×MfgDate）的 p 值远大于典型的临界值 0.05，表明这些项不重要。可以选择忽略这些项，将它们的影响合并到误差项中。输出 terms 变量返回一个代码矩阵，每个代码都代表一个术语的位模式。通过从术语中删除条目来从模型中省略术语。

```
>> [~,~,stats] = anovan(MPG,{org cyl4 when},terms,3,varnames) % 得到方差分析表与图 5-16 一致
stats =
  包含以下字段的 struct:
       source : 'anovan'
        resid: [1×406 double]
       coeffs : [30×1 double]
          Rtr: [14×14 double]
     rowbasis : [14×30 double]
          dfe: 384
…
```

现在有一个更简洁的模型，表明这些汽车的里程似乎与所有三个因素有关，并且制造日期的影响取决于汽车的制造地点。对 Origin 和 Cylinder 执行多重比较。

```
>> results = multcompare(stats,'Dimension',[1,2]) % 多重比较表如图 5-17 所示
results =
    1.0000    2.0000   -7.5624   -4.0331   -0.5038    0.0144
    1.0000    3.0000   -8.4566   -4.0174    0.4218    0.1024
    1.0000    4.0000  -10.3771   -8.7025   -7.0278    0.0000
    1.0000    5.0000  -14.0054  -12.3561  -10.7069    0.0000
    1.0000    6.0000  -12.8187  -11.1673   -9.5158    0.0000
    2.0000    3.0000   -5.4698    0.0157    5.5012    1.0000
    2.0000    4.0000   -8.3570   -4.6693   -0.9817    0.0042
    2.0000    5.0000  -11.9795   -8.3230   -4.6666    0.0000
    2.0000    6.0000  -10.7977   -7.1341   -3.4706    0.0000
    3.0000    4.0000   -9.3273   -4.6850   -0.0428    0.0464
    3.0000    5.0000  -12.9050   -8.3387   -3.7724    0.0000
    3.0000    6.0000  -11.6841   -7.1498   -2.6156    0.0001
    4.0000    5.0000   -5.5949   -3.6537   -1.7125    0.0000
    4.0000    6.0000   -4.4210   -2.4648   -0.5086    0.0045
    5.0000    6.0000   -0.7652    1.1889    3.1430    0.5092
```

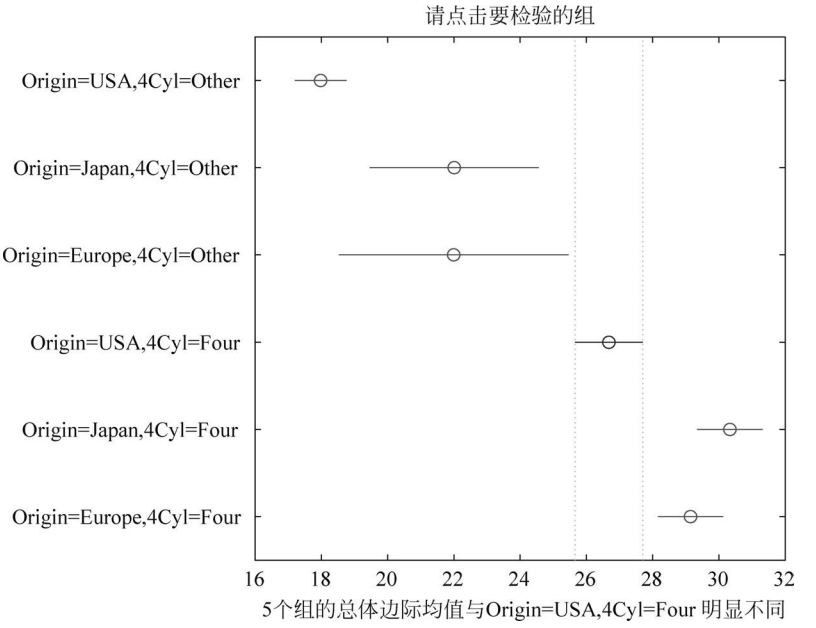

图 5-17　多重比较表