第十章　数据的统计描述和分析

数理统计研究的对象是受随机因素影响的数据，以下数理统计就简称统计，统计是以概率论为基础的一门应用学科。

数据样本少则几个，多则成千上万，人们希望能用少数几个包含其最多相关信息的数值来体现数据样本总体的规律。描述性统计就是搜集、整理、加工和分析统计数据，使之系统化、条理化，以显示出数据资料的趋势、特征和数量关系。它是统计推断的基础，实用性较强，在统计工作中经常使用。

面对一批数据如何进行描述与分析，需要掌握参数估计和假设检验这两个数理统计的最基本方法。

我们将用 Matlab 的统计工具箱(Statistics Toolbox)来实现数据的统计描述和分析。

§1　统计的基本概念

1.1　总体和样本

总体是人们研究对象的全体，又称母体，如工厂一天生产的全部产品（按合格品及废品分类），学校全体学生的身高。

总体中的每一个基本单位称为个体，个体的特征用一个变量（如 x ）来表示，如一件产品是合格品记 $x=0$，是废品记 $x=1$；一个身高 170（cm）的学生记 $x=170$。

从总体中随机产生的若干个个体的集合称为样本，或子样，如 n 件产品，100 名学生的身高，或者一根轴直径的 10 次测量。实际上这就是从总体中随机取得的一批数据，不妨记作 $x_1,x_2,\cdots,x_n$， n 称为样本容量。

简单地说，统计的任务是由样本推断总体。

1.2　频数表和直方图

一组数据（样本）往往是杂乱无章的，做出它的频数表和直方图，可以看作是对这组数据的一个初步整理和直观描述。

将数据的取值范围划分为若干个区间，然后统计这组数据在每个区间中出现的次数，称为频数，由此得到一个频数表。以数据的取值为横坐标，频数为纵坐标，画出一个阶梯形的图，称为直方图，或频数分布图。

若样本容量不大，能够手工做出频数表和直方图，当样本容量较大时则可以借助 Matlab 这样的软件了。让我们以下面的例子为例，介绍频数表和直方图的作法。

例 1　学生的身高和体重

学校随机抽取 100 名学生，测量他们的身高和体重，所得数据如表

表 1　身高体重数据

身高	体重	身高	体重	身高	体重	身高	体重	身高	体重
172	75	169	55	169	64	171	65	167	47
171	62	168	67	165	52	169	62	168	65
166	62	168	65	164	59	170	58	165	64
160	55	175	67	173	74	172	64	168	57
155	57	176	64	172	69	169	58	176	57
173	58	168	50	169	52	167	72	170	57
166	55	161	49	173	57	175	76	158	51
170	63	169	63	173	61	164	59	165	62
167	53	171	61	166	70	166	63	172	53
173	60	178	64	163	57	169	54	169	66
178	60	177	66	170	56	167	54	169	58
173	73	170	58	160	65	179	62	172	50
163	47	173	67	165	58	176	63	162	52

165	66	172	59	177	66	182	69	175	75
170	60	170	62	169	63	186	77	174	66
163	50	172	59	176	60	166	76	167	63
172	57	177	58	177	67	169	72	166	50
182	63	176	68	172	56	173	59	174	64
171	59	175	68	165	56	169	65	168	62
177	64	184	70	166	49	171	71	170	59

（i） 数据输入

数据输入通常有两种方法，一种是在交互环境中直接输入，如果在统计中数据量比较大，这样作不太方便；另一种办法是先把数据写入一个纯文本数据文件 data.txt 中，格式如例 1 的表 1，有 20 行、10 列，数据列之间用空格键或 Tab 键分割，该数据文件 data.txt 存放在 matlab\work 子目录下，在 Matlab 中用 load 命令读入数据，具体作法是：

load data.txt

这样在内存中建立了一个变量 data，它是一个包含有 20×10 个数据的矩阵。

为了得到我们需要的 100 个身高和体重各为一列的矩阵，应做如下的改变：

```
high=data(:,1:2:9);high=high(:)
weight=data(:,2:2:10);weight=weight(:)
```

（ii）作频数表及直方图

求频数用 hist 命令实现，其用法是：

[N,X] = hist(Y,M)

得到数组（行、列均可）Y 的频数表。它将区间[min(Y),max(Y)]等分为 M 份（缺省时 M 设定为 10），N 返回 M 个小区间的频数，X 返回 M 个小区间的中点。

命令

hist(Y,M)

画出数组 Y 的直方图。

对于例 1 的数据，编写程序如下：

```
load data.txt;
high=data(:,1:2:9);high=high(:);
weight=data(:,2:2:10);weight=weight(:);
[n1,x1]=hist(high)
%下面语句与hist命令等价
%n1=[length(find(high<158.1)),...
%   length(find(high>=158.1&high<161.2)),...
%   length(find(high>=161.2&high<164.5)),...
%   length(find(high>=164.5&high<167.6)),...
%   length(find(high>=167.6&high<170.7)),...
%   length(find(high>=170.7&high<173.8)),...
%   length(find(high>=173.8&high<176.9)),...
%   length(find(high>=176.9&high<180)),...
%   length(find(high>=180&high<183.1)),...
%   length(find(high>=183.1))]
[n2,x2]=hist(weight)
subplot(1,2,1), hist(high)
subplot(1,2,2), hist(weight)
```

计算结果略，直方图如图 1 所示。

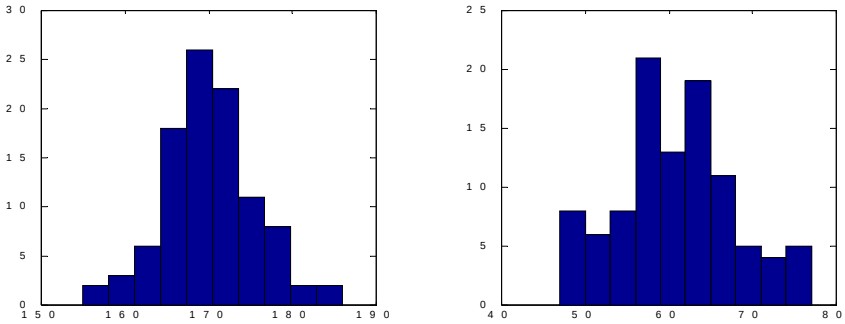

图 1　直方图

从直方图上可以看出，身高的分布大致呈中间高、两端低的钟形；而体重则看不出什么规律。要想从数值上给出更确切的描述，需要进一步研究反映数据特征的所谓“统计量”。直方图所展示的身高的分布形状可看作正态分布，当然也可以用这组数据对分布作假设检验。

例 2　统计下列五行字符串中字符 a、g、c、t 出现的频数

1.aggcacggaaaaacgggaataacggaggaggacttggcacggcattacacggagg
2.cggaggacaaacgggatggcggtattggaggtggcggactgttcgggga
3.gggacggatacggattctggccacggacggaaaggaggacacggcggacataca
4.atggataacggaaacaaaccagacaaacttcggtagaaatacagaagctta
5.cggctggcggacaacggactggcggattccaaaaacggaggaggcggacggaggc

解　把上述五行复制到一个纯文本数据文件 shuju.txt 中，放在 matlab\work 子目录下，编写如下程序：

```
clc
fid1=fopen('shuju.txt','r');
i=1;
while (~feof(fid1))
data=fgetl(fid1);
a=length(find(data==97));
b=length(find(data==99));
c=length(find(data==103));
d=length(find(data==116));
e=length(find(data>=97&data<=122));
f(i,:)=[a  b  c  d  e  a+b+c+d];
i=i+1;
end
f, he=sum(f)
dlmwrite('pinshu.txt',f); dlmwrite('pinshu.txt',he,'-append');
fclose(fid1);
```

我们把统计结果最后写到一个纯文本文件 pinshu.txt 中，在程序中多引进了几个变量，是为了检验字符串是否只包含 a、g、c、t 四个字符。

1.3　统计量

假设有一个容量为 n 的样本（即一组数据），记作 $x=(x_1,x_2,\cdots,x_n)$，需要对它进行一定的加工，才能提出有用的信息，用作对总体（分布）参数的估计和检验。**统计量**就是加工出来的、反映样本数量特征的函数，它不含任何未知量。

下面我们介绍几种常用的统计量。

（i）表示位置的统计量—算术平均值和中位数

算术平均值（简称均值）描述数据取值的平均位置，记作 $\bar{x}$，

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

中位数是将数据由小到大排序后位于中间位置的那个数值。

Matlab 中 mean(x)返回 x 的均值，median(x)返回中位数。

（ii）表示变异程度的统计量—标准差、方差和极差

标准差 s 定义为

$$s=\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{\frac{1}{2}} \tag{2}$$

它是各个数据与均值偏离程度的度量，这种偏离不妨称为变异。

方差是标准差的平方 s^2。

极差是 $x=(x_1,x_2,\cdots,x_n)$ 的最大值与最小值之差。

Matlab 中 std(x)返回 x 的标准差，var(x)返回方差，range(x)返回极差。

你可能注意到标准差 s 的定义（2）中，对 n 个 $(x_i-\bar{x})$ 的平方求和，却被 $(n-1)$ 除，这是出于无偏估计的要求。若需要改为被 n 除，Matlab 可用 std(x,1)和 var(x,1)来实现。

（iii）中心矩、表示分布形状的统计量—偏度和峰度

随机变量 x 的 r 阶**中心矩**为 $E(x-Ex)^r$。

随机变量 x 的偏度和峰度指的是 x 的标准化变量 $(x-Ex)/\sqrt{Dx}$ 的三阶中心矩和四阶中心矩：

$$\nu_1=E\left[\left(\frac{x-E(x)}{\sqrt{D(x)}}\right)^3\right]=\frac{E\left[(x-E(x))^3\right]}{(D(x))^{3/2}},$$

$$\nu_2=E\left[\left(\frac{x-E(x)}{\sqrt{D(x)}}\right)^4\right]=\frac{E\left[(x-E(x))^4\right]}{(D(x))^{2}}.$$

偏度反映分布的对称性，$\nu_1>0$ 称为右偏态，此时数据位于均值右边的比位于左边的多；$\nu_1<0$ 称为左偏态，情况相反；而 ν_1 接近 0 则可认为分布是对称的。

峰度是分布形状的另一种度量，正态分布的峰度为 3，若 ν_2 比 3 大得多，表示分布有沉重的尾巴，说明样本中含有较多远离均值的数据，因而峰度可以用作衡量偏离正态分布的尺度之一。

Matlab 中 moment(x,order)返回 x 的 order 阶中心矩，order 为中心矩的阶数。skewness(x)返回 x 的偏度，kurtosis(x)返回峰度。

在以上用 Matlab 计算各个统计量的命令中，若 x 为矩阵，则作用于 x 的列，返回一个行向量。

对例 1 给出的学生身高和体重，用 Matlab 计算这些统计量，程序如下：

```
clc
load data.txt;
high=data(:,1:2:9);high=high(:);
weight=data(:,2:2:10);weight=weight(:);
```

```
shuju=[high weight];
jun_zhi=mean(shuju)
zhong_wei_shu=median(shuju)
biao_zhun_cha=std(shuju)
ji_cha=range(shuju)
pian_du=skewness(shuju)
feng_du=kurtosis(shuju)
```

统计量中最重要、最常用的是均值和标准差，由于样本是随机变量，它们作为样本的函数自然也是随机变量，当用它们去推断总体时，有多大的可靠性就与统计量的概率分布有关，因此我们需要知道几个重要分布的简单性质。

1.4 统计中几个重要的概率分布

1.4.1 分布函数、密度函数和分位数

随机变量的特性完全由它的（概率）分布函数或（概率）密度函数来描述。设有随机变量 X，其分布函数定义为 $X \le x$ 的概率，即 $F(x)=P\{X \le x\}$。若 X 是连续型随机变量，则其密度函数 $p(x)$ 与 $F(x)$ 的关系为

$$F(x)=\int_{-\infty}^{x} p(x)dx .$$

上 α 分位数是下面常用的一个概念，其定义为：对于 $0<\alpha<1$，使某分布函数 $F(x)=1-\alpha$ 的 x，称为这个分布的上 α 分位数，记作 x_α。

我们前面画过的直方图是频数分布图，频数除以样本容量 n，称为频率，n 充分大时频率是概率的近似，因此直方图可以看作密度函数图形的（离散化）近似。

1.4.2 统计中几个重要的概率分布

（i）正态分布

正态分布随机变量 X 的密度函数曲线呈中间高两边低、对称的钟形，期望（均值）$EX=\mu$，方差 $DX=\sigma^2$，记作 $X \sim N(\mu,\sigma^2)$，σ 称均方差或标准差，当 $\mu=0,\sigma=1$ 时称为标准正态分布，记作 $X \sim N(0,1)$。正态分布完全由均值 μ 和方差 σ^2 决定，它的偏度为 0，峰度为 3。

正态分布可以说是最常见的（连续型）概率分布，成批生产时零件的尺寸，射击中弹着点的位置，仪器反复量测的结果，自然界中一种生物的数量特征等，多数情况下都服从正态分布，这不仅是观察和经验的总结，而且有着深刻的理论依据，即在大量相互独立的、作用差不多大的随机因素影响下形成的随机变量，其极限分布为正态分布。

鉴于正态分布的随机变量在实际生活中如此地常见，记住下面 3 个数字是有用的：

68%的数值落在距均值左右 1 个标准差的范围内，即

$$P\{\mu-\sigma \le X \le \mu+\sigma\}=0.68;$$

95%的数值落在距均值左右 2 个标准差的范围内，即

$$P\{\mu-2\sigma \le X \le \mu+2\sigma\}=0.95;$$

99.7%的数值落在距均值左右 3 个标准差的范围内，即

$$P\{\mu-3\sigma \le X \le \mu+3\sigma\}=0.997 .$$

（ii）χ^2 分布(Chi square)

若 $X_1,X_2,\cdots,X_n$ 为相互独立的、服从标准正态分布 $N(0,1)$ 的随机变量，则它们的平方和 $Y=\sum_{i=1}^{n} X_i^2$ 服从 χ^2 分布，记作 $Y \sim \chi^2(n)$，n 称自由度，它的期望 $EY=n$，

方差 $DY=2n$。

（iii）t 分布

若 $X\sim N(0,1)$，$Y\sim\chi^2(n)$，且相互独立，则 $T=\dfrac{X}{\sqrt{Y/n}}$ 服从 t 分布，记作 $T\sim t(n)$，n 称自由度。t 分布又称学生氏(Student)分布。

t 分布的密度函数曲线和 $N(0,1)$ 曲线形状相似。理论上 $n\to\infty$ 时，$T\sim t(n)\to N(0,1)$，实际上当 $n>30$ 时它与 $N(0,1)$ 就相差无几了。

（iv）F 分布

若 $X\sim\chi^2(n_1)$，$Y\sim\chi^2(n_2)$，且相互独立，则 $F=\dfrac{X/n_1}{Y/n_2}$ 服从 F 分布，记作 $F\sim F(n_1,n_2)$，(n_1,n_2) 称自由度。

1.4.3 Matlab 统计工具箱(Toolbox\Stats)中的概率分布

Matlab 统计工具箱中有 27 种概率分布，这里只对上面所述 4 种分布列出命令的字符：

norm 正态分布； chi2 χ^2 分布；

t t 分布 f F 分布

工具箱对每一种分布都提供 5 类函数，其命令的字符是：

pdf 概率密度； cdf 分布函数； inv 分布函数的反函数；

stat 均值与方差； rnd 随机数生成

当需要一种分布的某一类函数时，将以上所列的分布命令字符与函数命令字符接起来，并输入自变量（可以是标量、数组或矩阵）和参数就行了，如：

p=normpdf(x,mu,sigma) 均值 mu、标准差 sigma 的正态分布在 x 的密度函数(mu=0，sigma=1 时可缺省)。

p=tcdf(x,n) t 分布（自由度 n）在 x 的分布函数。

x=chi2inv(p,n) χ^2 分布（自由度 n）使分布函数 F(x)=p 的 x(即 p 分位数)。

[m,v]=fstat(n1,n2) F 分布（自由度 n1,n2）的均值 m 和方差 v。

几个分布的密度函数图形就可以用这些命令作出，如：

```
x=-6:0.01:6;y=normpdf(x);z=normpdf(x,0,2);
plot(x,y,x,z),gtext('N(0,1)'),gtext('N(0,2^2)')
```

分布函数的反函数的意义从下例看出：

```
x=chi2inv(0.9,10)
x =
    15.9872
```

如果反过来计算，则

```
P=chi2cdf(15.9872,10)
P =
     0.9000
```

1.5 正态总体统计量的分布

用样本来推断总体，需要知道样本统计量的分布，而样本又是一组与总体同分布的随机变量，所以样本统计量的分布依赖于总体的分布。当总体服从一般的分布时，求某个样本统计量的分布是很困难的，只有在总体服从正态分布时，一些重要的样本统计量（均值、标准差）的分布才有便于使用的结果。另一方面，现实生活中需要进行统计推断的总体，多数可以认为服从（或近似服从）正态分布，所以统计中人们在正态总体的

假定下研究统计量的分布，是必要的与合理的。

设总体 $X \sim N(\mu,\sigma^2)$， $x_1,x_2,\cdots,x_n$ 为一容量 n 的样本，其均值 $\bar{x}$ 和标准差 s 由式（1）、（2）确定，则用 $\bar{x}$ 和 s 构造的下面几个分布在统计中是非常有用的。

$$\bar{x} \sim N(\mu,\frac{\sigma^2}{n}) \quad 或 \quad \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{3}$$

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1). \tag{4}$$

$$\frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t(n-1) \tag{5}$$

设有两个总体 $X \sim N(\mu_1,\sigma_2^2)$ 和 $Y \sim N(\mu_2,\sigma_2^2)$，及由容量分别为 n_1， n_2 的两个样本确定的均值 $\bar{x},\bar{y}$ 和标准差 s_1,s_2，则

$$\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \sim N(0,1) \tag{6}$$

$$\frac{(\bar{x}-\bar{y})-(\mu_1-\mu_2)}{s_w\sqrt{1/n_1+1/n_2}} \sim t(n_1+n_2-2) \tag{7}$$

其中 $s_w^2=\dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$，

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(n_1-1,n_2-1) \tag{8}$$

对于（7）式，假定 $\sigma_1=\sigma_2$，但它们未知，于是用 s 代替。在下面的统计推断中我们要反复用到这些分布。

§2 参数估计

利用样本对总体进行统计推断的一类问题是参数估计，即假定已知总体的分布，通常是 $X \sim N(\mu,\sigma^2)$，估计有关的参数，如 μ,σ^2。参数估计分点估计和区间估计两种。

2.1 点估计

点估计是用样本统计量确定总体参数的一个数值。评价估计优劣的标准有无偏性、最小方差性、有效性等，估计的方法有矩法、极大似然法等。

最常用的是对总体均值 μ 和方差 σ^2（或标准差 σ）作点估计。让我们暂时抛开评价标准，当从一个样本按照式（1）、（2）算出样本均值 $\bar{x}$ 和方差 s^2 后，对 μ 和 σ^2（或 σ）一个自然、合理的点估计显然是（在字母上加^表示它的估计值）

$$\hat{\mu}=\bar{x}\,,\ \hat{\sigma}^2=s^2,\quad \hat{\sigma}=s \tag{9}$$

2.2 区间估计

点估计虽然给出了待估参数的一个数值，却没有告诉我们这个估计值的精度和可信程度。一般地，总体的待估参数记作 θ（如 μ,σ^2），由样本算出的 θ 的估计量记作 $\hat{\theta}$，人们常希望给出一个区间 $[\hat{\theta}_1,\hat{\theta}_2]$，使 θ 以一定的概率落在此区间内。若有

$$P\{\hat{\theta}_1<\theta<\hat{\theta}_2\}=1-\alpha\,,\ 0<\alpha<1 \tag{10}$$

则$[\hat{\theta}_1,\hat{\theta}_2]$称为$\theta$的置信区间，$\hat{\theta}_1,\hat{\theta}_2$分别称为置信下限和置信上限，$1-\alpha$称为置信概率或置信水平，$\alpha$称为显著性水平。

给出的置信水平为$1-\alpha$的置信区间$[\hat{\theta}_1,\hat{\theta}_2]$，称为$\theta$的区间估计。置信区间越小，估计的精度越高；置信水平越大，估计的可信程度越高。但是这两个指标显然是矛盾的，通常是在一定的置信水平下使置信区间尽量小。通俗地说，区间估计给出了点估计的误差范围。

2.3 参数估计的 Matlab 实现

Matlab 统计工具箱中，有专门计算总体均值、标准差的点估计和区间估计的函数。对于正态总体，命令是

[mu,sigma,muci,sigmaci]=normfit(x,alpha)

其中 x 为样本（数组或矩阵），alpha 为显著性水平α（alpha 缺省时设定为 0.05），返回总体均值μ和标准差σ的点估计 mu 和 sigma，及总体均值μ和标准差σ的区间估计 muci 和 sigmaci。当 x 为矩阵时，x 的每一列作为一个样本。

Matlab 统计工具箱中还提供了一些具有特定分布总体的区间估计的命令，如 expfit，poissfit，gamfit，你可以从这些字头猜出它们用于哪个分布，具体用法参见帮助系统。

§3 假设检验

统计推断的另一类重要问题是假设检验问题。在总体的分布函数完全未知或只知其形式但不知其参数的情况，为了推断总体的某些性质，提出某些关于总体的假设。例如，提出总体服从泊松分布的假设，又如对于正态总体提出数学期望等于μ_0的假设等。假设检验就是根据样本对所提出的假设做出判断：是接受还是拒绝。这就是所谓的假设检验问题。

3.1 单个总体$N(\mu,\sigma^2)$均值μ的检验

假设检验有三种：

双边检验：$H_0:\mu=\mu_0$，$H_1:\mu\neq\mu_0$；

右边检验：$H_0:\mu\leq\mu_0$，$H_1:\mu>\mu_0$；

左边检验：$H_0:\mu\geq\mu_0$，$H_1:\mu<\mu_0$。

3.1.1 σ^2已知，关于μ的检验（Z检验）

在 Matlab 中Z检验法由函数 ztest 来实现，命令为

[h,p,ci]=ztest(x,mu,sigma,alpha,tail)

其中输入参数 x 是样本，mu 是H_0中的μ_0，sigma 是总体标准差σ，alpha 是显著性水平α（alpha 缺省时设定为 0.05），tail 是对备选假设H_1的选择：H_1为$\mu\neq\mu_0$时用 tail=0（可缺省）；H_1为$\mu>\mu_0$时用 tail=1；H_1为$\mu<\mu_0$时用 tail=-1。输出参数 h=0 表示接受H_0，h=1 表示拒绝H_0，p 表示在假设H_0下样本均值出现的概率，p 越小H_0越值得怀疑，ci 是μ_0的置信区间。

例 3 某车间用一台包装机包装糖果。包得的袋装糖重是一个随机变量，它服从正态分布。当机器正常时，其均值为 0.5 公斤，标准差为 0.015 公斤。某日开工后为检验包装机是否正常，随机地抽取它所包装的糖 9 袋，称得净重为（公斤）：

0.497 0.506 0.518 0.524 0.498 0.511 0.520 0.515 0.512

问机器是否正常？

解　总体σ已知，$x \sim N(\mu,0.015^2)$，μ未知。于是提出假设$H_0:\mu=\mu_0=0.5$和$H_1:\mu\neq 0.5$。

Matlab 实现如下：

```
x=[0.497  0.506  0.518  0.524  0.498...
  0.511  0.520  0.515  0.512];
[h,p,ci]=ztest(x,0.5,0.015)
```

求得 h=1，p=0.0248，说明在 0.05 的水平下，可拒绝原假设，即认为这天包装机工作不正常。

3.1.2　σ^2未知，关于μ的检验（t检验）

在 Matlab 中t检验法由函数 ttest 来实现，命令为

[h,p,ci]=ttest(x,mu,alpha,tail)

例 4　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.现得 16 只元件的寿命如下:

159　280　101　212　224　379　179　264
222　362　168　250　149　260　485　170

问是否有理由认为元件的平均寿命大于 225(小时)？

解　按题意需检验

$$H_0:\mu\le\mu_0=225,\quad H_1:\mu>225,$$

取$\alpha=0.05$。Matlab 实现如下：

```
x=[159  280  101  212  224  379  179  264 ...
  222  362  168  250  149  260  485  170];
[h,p,ci]=ttest(x,225,0.05,1)
```

求得 h=0，p=0.2570，说明在显著水平为 0.05 的情况下，不能拒绝原假设，认为元件的平均寿命不大于 225 小时。

3.2　两个正态总体均值差的检验（t检验）

还可以用t检验法检验具有相同方差的 2 个正态总体均值差的假设。在 Matlab 中由函数 ttest2 实现，命令为：

[h,p,ci]=ttest2(x,y,alpha,tail)

与上面的 ttest 相比，不同处只在于输入的是两个样本 x,y（长度不一定相同），而不是一个样本和它的总体均值；tail 的用法与 ttest 相似，可参看帮助系统。

例 5 在平炉上进行一项试验以确定改变操作方法的建议是否会增加钢的得率,试验是在同一平炉上进行的。每炼一炉钢时除操作方法外,其它条件都可能做到相同。先用标准方法炼一炉,然后用建议的新方法炼一炉,以后交换进行,各炼了 10 炉,其得率分别为

1° 标准方法　78.1　72.4　76.2　74.3　77.4　78.4　76.0　75.6　76.7　77.3
2° 新方法　79.1　81.0　77.3　79.1　80.0　79.1　79.1　77.3　80.2　82.1

设这两个样本相互独立且分别来自正态总体$N(\mu_1,\sigma^2)$和$N(\mu_2,\sigma^2)$，μ_1,μ_2,σ^2均未知，问建议的新方法能否提高得率?(取$\alpha=0.05$。)

解　(i) 需要检验假设

$$H_0:\mu_1-\mu_2\ge 0,\quad H_1:\mu_1-\mu_2<0.$$

(ii)Matlab 实现

```
x=[78.1  72.4  76.2  74.3  77.4  78.4  76.0  75.6  76.7  77.3];
y=[79.1  81.0  77.3  79.1  80.0  79.1  79.1  77.3  80.2  82.1];
[h,p,ci]=ttest2(x,y,0.05,-1)
```

求得 h=1, p=2.2126×10^{-4}。表明在$\alpha=0.05$的显著水平下，可以拒绝原假设，即认为建议的新操作方法较原方法优。

3.3　分布拟合检验

在实际问题中，有时不能预知总体服从什么类型的分布，这时就需要根据样本来检验关于分布的假设。下面介绍χ^2检验法和专用于检验分布是否为正态的“偏峰、峰度检验法”。

3.3.1　χ^2检验法

H_0：总体x的分布函数为$F(x)$，

H_1：总体x的分布函数不是$F(x)$.

在用下述χ^2检验法检验假设H_0时，若在假设H_0下$F(x)$的形式已知，但其参数值未知，这时需要先用极大似然估计法估计参数，然后作检验。

χ^2检验法的基本思想如下：将随机试验可能结果的全体Ω分为k个互不相容的事件$A_1, A_2, A_3, ..., A_k$ $(\sum_{i=1}^{k} A_k = \Omega, A_i A_j = \Phi, i \neq j, i, j = 1,2,\cdots,k)$。于是在假设$H_0$下，我们可以计算$p_i = P(A_i)$（或$\hat{p}_i = \hat{P}(A_i)$），$i = 1,2,\cdots,k$。在$n$次试验中，事件$A_i$出现的频率$f_i/n$与$p_i$（$\hat{p}_i$）往往有差异，但一般来说，若$H_0$为真，且试验的次数又甚多时，则这种差异不应该很大。基于这种想法，皮尔逊使用

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} \quad (\text{或 } \chi^2 = \sum_{i=1}^{k} \frac{(f_i - n\hat{p}_i)^2}{n\hat{p}_i}) \tag{11}$$

作为检验假设H_0的统计量。并证明了以下定理。

定理　若n充分大，则当H_0为真时（不论H_0中的分布属什么分布），统计量（11）总是近似地服从自由度为$k-r-1$的χ^2分布，其中r是被估计的参数的个数。

于是，若在假设H_0下算得（11）有

$$\chi^2 \geq \chi_a^2(k-r-1),$$

则在显著性水平α下拒绝H_0，否则就接受。

注意：在使用χ^2检验法时，要求样本容量n不小于 50，以及每个np_i都不小于 5，而且np_i最好是在 5 以上。否则应适当地合并A_i，以满足这个要求。

例 6　下面列出了 84 个伊特拉斯坎（Etruscan）人男子的头颅的最大宽度（mm），试检验这些数据是否来自正态总体（取$\alpha=0.1$）。

141	148	132	138	154	142	150	146	155	158
150	140	147	148	144	150	149	145	149	158
143	141	144	144	126	140	144	142	141	140
145	135	147	146	141	136	140	146	142	137
148	154	137	139	143	140	131	143	141	149
148	135	148	152	143	144	141	143	147	146

150 132 142 142 143 153 149 146 149 138
142 149 142 137 134 144 146 147 140 142
140 137 152 145

解 编写Matlab程序如下：

```
clc
x=[141 148 132 138 154 142 150 146 155 158 ...
150 140 147 148 144 150 149 145 149 158 ...
143 141 144 144 126 140 144 142 141 140 ...
145 135 147 146 141 136 140 146 142 137 ...
148 154 137 139 143 140 131 143 141 149 ...
148 135 148 152 143 144 141 143 147 146 ...
150 132 142 142 143 153 149 146 149 138 ...
142 149 142 137 134 144 146 147 140 142 ...
140 137 152 145];
mm=minmax(x)   %求数据中的最小数和最大数
hist(x,8)      %画直方图
fi=[length(find(x<135)),...
   length(find(x>=135&x<138)),...
   length(find(x>=138&x<142)),...
   length(find(x>=142&x<146)),...
   length(find(x>=146&x<150)),...
   length(find(x>=150&x<154)),...
   length(find(x>=154))]     %各区间上出现的频数
mu=mean(x),sigma=std(x)     %均值和标准差
fendian=[135,138,142,146,150,154]   %区间的分点
p0=normcdf(fendian,mu,sigma)        %分点处分布函数的值
p1=diff(p0)                          %中间各区间的概率
p=[p0(1),p1,1-p0(6)]                 %所有区间的概率
chi=(fi-84*p).^2./(84*p)
chisum=sum(chi)                      %皮尔逊统计量的值
x_a=chi2inv(0.9,4)     %chi2分布的0.9分位数
```

求得皮尔逊统计量chisum= 2.2654，$\chi^2_{0.1}(7-2-1)=\chi^2_{0.1}(4)=7.7794$，故在水平0.1下接受$H_0$，即认为数据来自正态分布总体。

3.3.2 偏度、峰度检验（留作习题1）

3.4 其它非参数检验

Matlab还提供了一些非参数方法。

3.4.1 Wilcoxon秩和检验

在Matlab中，秩和检验由函数ranksum实现。命令为：

[p,h]=ranksum(x,y,alpha)

其中x，y可为不等长向量，alpha为给定的显著水平，它必须为0和1之间的数量。p返回产生两独立样本的总体是否相同的显著性概率，h返回假设检验的结果。如果x和y的总体差别不显著，则h为零；如果x和y的总体差别显著，则h为1。如果p接近于零，则可对原假设质疑。

例7 某商店为了确定向公司A或公司B购买某种产品，将A,B公司以往各次进货的次品率进行比较，数据如下所示，设两样本独立。问两公司的商品的质量有无显著差异。设两公司的商品的次品的密度最多只差一个平移，取$\alpha=0.05$。

A：7.0　3.5　9.6　8.1　6.2　5.1　10.4　4.0　2.0　10.5

B：5.7　3.2　4.2　11.0　9.7　6.9　3.6　4.8　5.6　8.4　10.1　5.5　12.3

解　分别以μ_A、μ_B记公司A、B的商品次品率总体的均值。所需检验的假设是

$H_0:\mu_A=\mu_B$，$H_1:\mu_A\neq\mu_B$.

Matlab实现如下：

```
a=[7.0  3.5  9.6  8.1  6.2  5.1  10.4  4.0  2.0  10.5];
b=[5.7  3.2  4.2  11.0  9.7  6.9  3.6  4.8  5.6  8.4  10.1  5.5
12.3];
[p,h]=ranksum(a,b)
```

求得p=0.8041，h=0，表明两样本总体均值相等的概率为0.8041，并不很接近于零，且h=0说明可以接受原假设，即认为两个公司的商品的质量无明显差异。

3.5　中位数检验

在假设检验中还有一种检验方法为中位数检验,在一般的教学中不一定介绍,但在实际中也是被广泛应用到的。在Matlab中提供了这种检验的函数。函数的使用方法简单，下面只给出函数介绍。

3.5.1　signrank函数

signrank　Wilcoxon符号秩检验

[p,h]=signrank(x,y,alpha)

其中p给出两个配对样本x和y的中位数相等的假设的显著性概率。向量x，y的长度必须相同，alpha为给出的显著性水平，取值为0和1之间的数。h返回假设检验的结果。如果这两个样本的中位数之差几乎为0，则h=0；若有显著差异，则h=1。

3.5.2　signtest函数

signtest　符号检验

[p,h]= signtest(x,y,alpha)

其中p给出两个配对样本x和y的中位数相等的假设的显著性概率。x和y若为向量，二者的长度必须相同；y亦可为标量，在此情况下，计算x的中位数与常数y之间的差异。alpha和h同上。

习 题 十

1. 试用偏度、峰度检验法检验例6中的数据是否来自正态总体（取$\alpha=0.1$）。

2. 下面列出的是某工厂随机选取的20只部件的装配时间（分）：9.8,10.4,10.6,9.6,9.7,9.9,10.9,11.1,9.6,10.2,10.3,9.6,9.9,11.2,10.6,9.8,10.5,10.1,10.5,9.7。设装配时间的总体服从正态分布，是否可以认为装配时间的均值显著地大于10（取$\alpha=0.05$）？

3.表2分别给出两个文学家马克．吐温(Mark Twain)的八篇小品文及斯诺特格拉斯（Snodgrass）的10篇小品文中由3个字母组成的词的比例。

表2

马克．吐温	0.225	0.262	0.217	0.240	0.230	0.229	0.235	0.217		
斯诺特格拉斯	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。两样本相互独立，问两个作家所写的小品文中包含由3个字母组成的词的比例是否有显著的差异（取$\alpha=0.05$）？