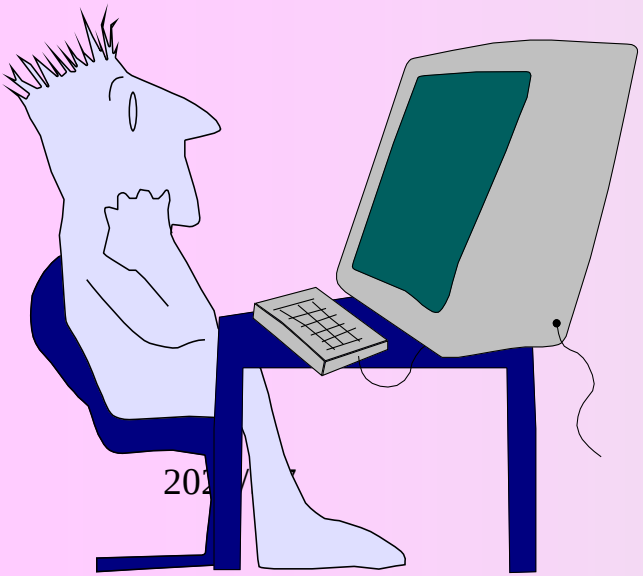
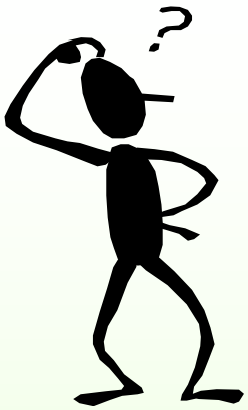
数学建模与数学实验

数据的统计描述和分析





实验目的

1、直观了解统计基本内容。

2、掌握用数学软件包求解统计问题。

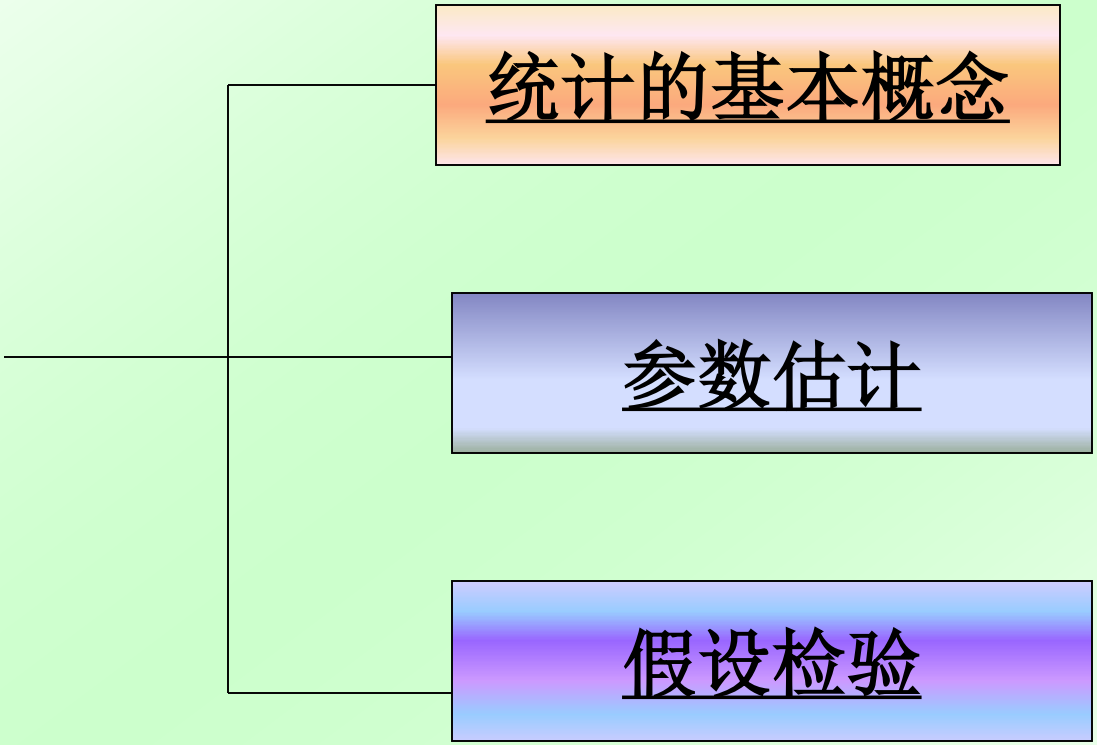
实验内容

1、统计的基本理论。

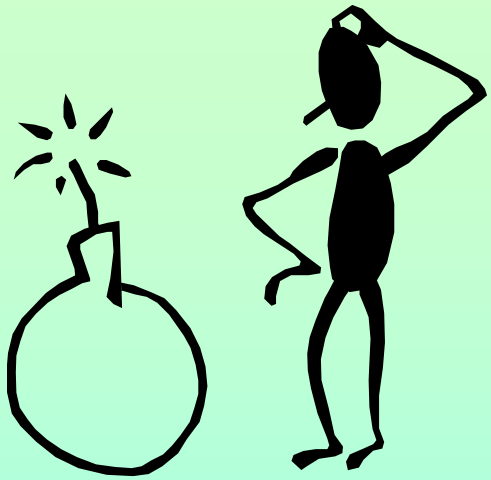
2、用数学软件包求解统计问题。

3、实验作业。

数据的统计描述和分析



一、统计量



1、表示位置的统计量—平均值和中位数

平均值（或均值，数学期望）：$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$

中位数：将数据由小到大排序后位于中间位置的那个数值.

2、表示变异程度的统计量—标准差、方差和极差

标准差：$s=[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2]^{\frac{1}{2}}$

它是各个数据与均值偏离程度的度量.

方差：标准差的平方.

极差：样本中最大值与最小值之差.

3. 表示分布形状的统计量—偏度和峰度

偏度： $g_1 = \frac{1}{s^3}\sum_{i=1}^{n}(X_i - \overline{X})^3$ **峰度：** $g_2 = \frac{1}{s^4}\sum_{i=1}^{n}(X_i - \overline{X})^4$

偏度反映分布的对称性，$g_1 > 0$ 称为右偏态，此时数据位于均值右边的比位于左边的多；$g_1 < 0$ 称为左偏态，情况相反；而 g_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 g_2 比 3 大很多，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，因而峰度可用作衡量偏离正态分布的尺度之一.

4. k 阶原点矩： $V_k = \frac{1}{n}\sum_{i=1}^{n}X_i^k$ **k 阶中心矩：** $U_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k$

二、分布函数的近似求法

1、整理资料： 把样本值 x_1，x_2，…，x_n 进行分组，先将它们依大小次序排列得 $x_1^* \le x_2^* \le \cdots \le x_n^*$.在包含$[x_1^*, x_n^*]$的区间[a，b]内插入一些等分点：$a < x_1' < x_2' < \cdots < x_n' < b$, 注意要使每一个区间$(x_i', x_{i+1}']$（i=1，2，…，n-1）内都有样本观测值 x_i（i=1，2，…，n-1）落入其中.

2、求出各组的频数和频率：统计出样本观测值在每个区间$(x_i', x_{i+1}']$中出现的次数n_i，它就是这区间或这组的频数.计算频率$f_i = \dfrac{n_i}{n}$.

3、作频率直方图：在直角坐标系的横轴上，标出$x_1', x_2', \cdots, x_n'$各点，分别以$(x_i', x_{i+1}']$为底边，作高为$\dfrac{f_i}{\Delta x_i'}$ 的矩形，$\Delta x_i' = x_{i+1}' - x_i', i = 1,2,\cdots,n-1$,即得频率直方图.

三、几个在统计中常用的概率分布

1．正态分布$N(\mu,\sigma^2)$

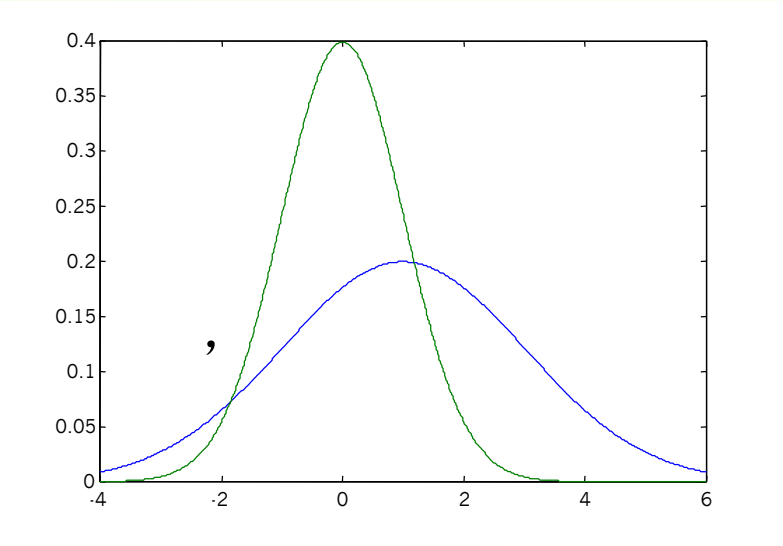
密度函数：$p(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ 分布函数：$F(x)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}e^{-\frac{(y-\mu)^2}{2\sigma^2}}dy$

其中μ为均值，σ^2为方差，$-\infty<x<+\infty$.

标准正态分布：N（0，1）

密度函数 $\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$

分布函数 $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{y^2}{2}}dy$



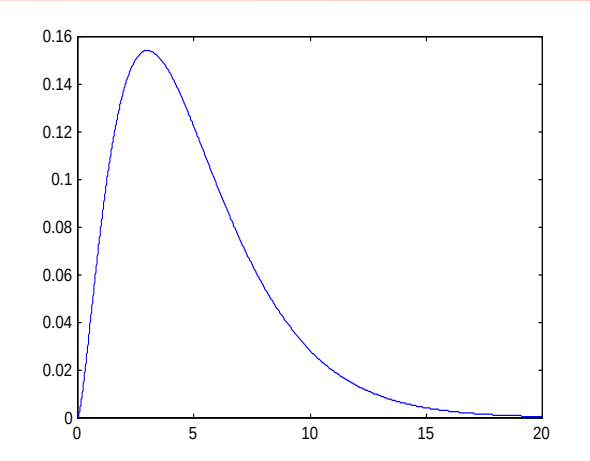
2、χ^2 分布 χ^2 （n）

若随机变量 X_1，X_2，… X_n 相互独立，都服从标准正态分布 N（0，1），则随机变量

$$Y = X_1^2 + X_2^2 + \cdots + X_n^2$$

服从自由度为 n 的 χ^2 分布，记为 $Y \sim \chi^2$ （n）.

Y 的均值为 n，方差为 2n.



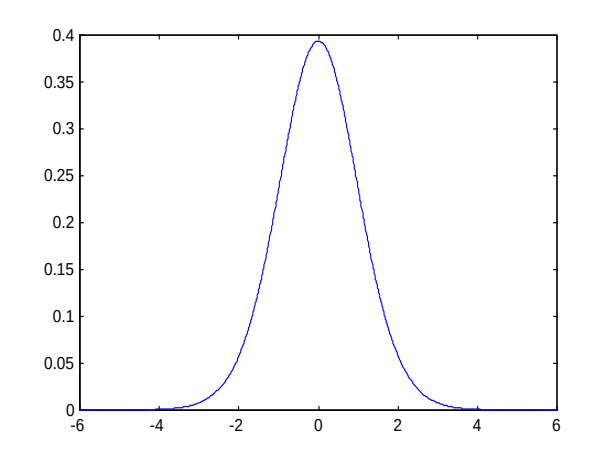
3、**t 分布** t（n）

若 X~N（0，1），Y~χ^2 （n），且相互独立，则随机变量

$$T=\frac{X}{\sqrt{\frac{Y}{n}}}$$

服从自由度为 n 的 t 分布，记为 T~t（n）.

t 分布 t（20）的密度函数曲线和 N（0，1）的曲线形状相似.理论上 n$\to\infty$ 时，T~t（n）$\to$ N（0，1）.



4. F 分布 F（n_1，n_2）

若 X~χ^2（n_1），Y~χ^2（n_2），且相互独立，则随机变量

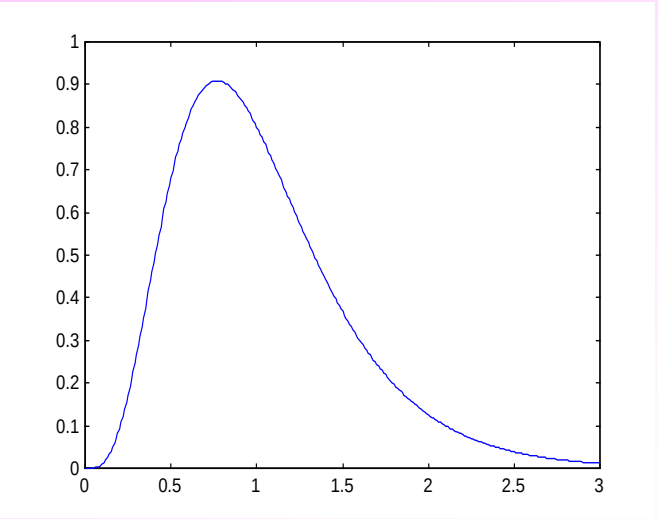
$$F=\frac{\frac{X}{n_1}}{\frac{Y}{n_2}}$$

服从自由度为（n_1，n_2）的 F 分布，记作 F~ F（n_1，n_2）.

由 F 分布的定义可以得到 F 分布的一个重要性质：

若 F~ F（n_1，n_2），则 $\frac{1}{F}\sim F(n_2,n_1)$

F分布F（10，50）的密度函数曲线



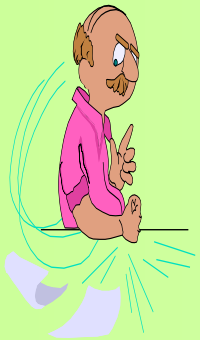
返回

参数估计

无论总体 X 的分布函数 F（x；$\theta_1,\theta_2,\cdots,\theta_k$）的类型已知或未知，我们总是需要去估计某些未知参数或数字特征，这就是参数估计问题.即参数估计就是从样本（X_1，X_2，…，X_n）出发，构造一些统计量$\hat{\theta}_i$（ X_1，X_2，…，X_n）（i=1，2，…，k）去估计总体 X 中的某些参数（或数字特征）θ_i（i=1，2，…，k）.这样的统计量称为**估计量**.

1. 点估计：构造（X_1，X_2，…，X_n）的函数$\hat{\theta}_i$（ X_1，X_2，…，X_n）作为参数θ_i 的点估计量，称统计量$\hat{\theta}_i$ 为总体 X 参数θ_i 的点估计量.

2. 区间估计：构造两个函数θ_{i1}（ X_1，X_2，…，X_n）和θ_{i2}（ X_1，X_2，…，X_n）做成区间，把这（θ_{i1},θ_{i2}）作为参数θ_i 的区间估计.



一、点估计的求法

（一）矩估计法

假设总体分布中共含有 k 个参数，它们往往是一些原点矩或一些原点矩的函数，例如，数学期望是一阶原点矩，方差是二阶原点矩与一阶原点矩平方之差等.因此，要想估计总体的某些参数θ_i（i=1，2，…k），由于 k 个参数一定可以表为不超过 k 阶原点矩的函数，很自然就会想到用样本的 r 阶原点矩去估计总体相应的 r 阶原点矩，用样本的一些原点矩的函数去估计总体的相应的一些原点矩的函数，再将 k 个参数反解出来，从而求出各个参数的估计值.这就是矩估计法，它是最简单的一种参数估计法.

（二）极大似然估计法

极大似然法的想法是：若抽样的结果得到样本观测值 $x_1,x_2,\cdots,x_n$，则我们应当这样选取参数 θ_i 的值，使这组样本观测值出现的可能性最大．即构造似然函数：

$$L(\theta_1,\theta_2,\cdots,\theta_k)=P(X_1=x_1,X_2=x_2,\cdots,X_n=x_n)=P(X_1=x_1)P(X_2=x_2)\cdots P(X_n=x_n)$$

$$=p(x_1,\theta_1,\cdots\theta_k)\,p(x_2,\theta_1,\cdots,\theta_k)\cdots\,p(x_n,\theta_1,\cdots\theta_k)=\prod_{i=1}^{n}p(x_i,\theta_1,\cdots\theta_k)$$

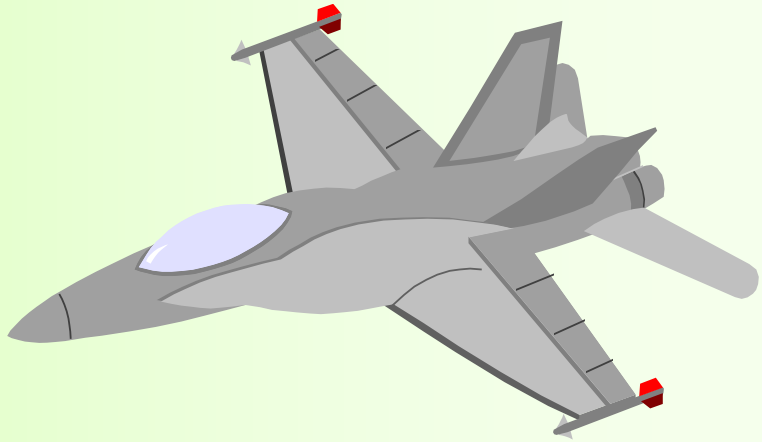
使 $L(\theta_1,\cdots,\theta_k)$ 达到最大，从而得到参数 θ_i 的估计值 $\hat{\theta}_i$.此估计值叫**极大似然估计值**.函数 $L(\theta_1,\cdots,\theta_k)$ 称为**似然函数**.

求极大似然估计值的问题，就是求似然函数 $L(\theta_1,\cdots,\theta_k)$ 的最大值的问题，则

$$\frac{\partial L}{\partial\theta_i}=0\quad i=1,2,\cdots,k$$

即

$$\frac{\partial LnL}{\partial\theta_i}=0\qquad i=1,2,\cdots,k$$



二、区间估计的求法

设总体 X 的分布中含有未知参数θ ，若对于给定的概率$1-\alpha$（$0<\alpha<1$），存在两个统计量$\hat{\theta}_1(X_1, X_2, \cdots, X_n)$和$\hat{\theta}_2(X_1, X_2, \cdots, X_n)$,使得

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1-\alpha$$

则称随机区间$(\hat{\theta}_1, \hat{\theta}_2)$ 为参数θ 的置信水平为$1-\alpha$ 的**置信区间**，$\hat{\theta}_1$ 称为**置信下限**,$\hat{\theta}_2$ 称为**置信上限**.

(一)数学期望的置信区间

1、已知DX，求EX的置信区间

设样本（X_1，X_2，…，X_n）来自正态母体 X，已知方差$DX=\sigma^2$，

EX 在置信水平 1-α 下的置信区间为$[\overline{X}-u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\overline{X}+u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$.

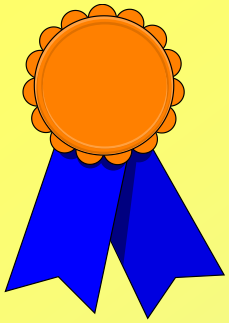
2． 未知方差DX，求EX的置信区间

EX 在置信水平 1-α 下的置信区间为$[\overline{X}-t_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}},\overline{X}+t_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}]$.

（二）方差的区间估计

DX 在置信水平 1-α 下的置信区间为$[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}},\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}]$.

返回



假设检验

对总体X的分布律或分布参数作某种假设，根据抽取的样本观察值，运用数理统计的分析方法，检验这种假设是否正确，从而决定接受假设或拒绝假设.

1.参数检验：如果观测的分布函数类型已知，这时构造出的统计量依赖于总体的分布函数，这种检验称为参数检验. 参数检验的目的往往是对总体的参数及其有关性质作出明确的判断.

2.非参数检验：如果所检验的假设并非是对某个参数作出明确的判断，因而必须要求构造出的检验统计量的分布函数不依赖于观测值的分布函数类型，这种检验叫非参数检验. 如要求判断总体分布类型的检验就是非参数检验.

假设检验的一般步骤是：

1. 根据实际问题提出原假设 H_0 与备择假设 H_1，即说明需要检验的假设的具体内容；
2. 选择适当的统计量，并在原假设 H_0 成立的条件下确定该统计量的分布；
3. 按问题的具体要求，选取适当的显著性水平 α ，并根据统计量的分布查表，确定对应于 α 的临界值.一般 α 取 0.05,0.01 或 0.10
4. 根据样本观测值计算统计量的观测值，并与临界值进行比较，从而在检验水平 α 条件下对拒绝或接受原假设 H_0 作出判断.

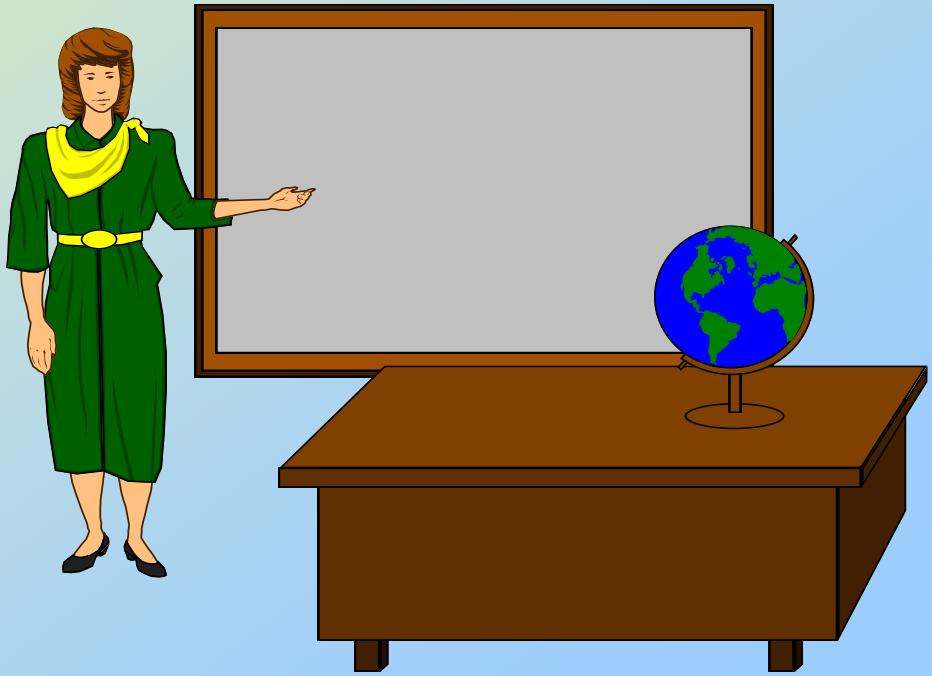
一、参数检验

（一）单个正态总体均值检验

设取出一容量为 n 的样本，得到均值$\overline{X}$ 和标准差 s，现要对总体均值μ 是否等于某给定值 μ_0 进行检验.记

$$H_0:\mu=\mu_0;\quad H_1:\mu\neq\mu_0$$

称 H_0 为**原假设**，H_1 为**备择假设**，两者择其一：接受 H_0；拒绝 H_0，即接受 H_1.



1、总体方差σ^2 已知

用 **u 检验**，检验的拒绝域为

$W=\{|z|>u_{1-\frac{\alpha}{2}}\}$　　即　$W=\{z<-u_{1-\frac{\alpha}{2}}$ 或 $z>u_{1-\frac{\alpha}{2}}\}$

2. 总体方差σ^2 未知

用样本方差 s^2 代替总体方差σ^2 ，这种检验叫 **t 检验**.

	H_0	H_1	总体方差σ^2 已知 统计量 $z=\dfrac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$	总体方差σ^2 未知 统计量 $t=\dfrac{\bar{X}-\mu_0}{s/\sqrt{n}}$
			在显著水平α 下拒绝 H_0，若	
I	$\mu=\mu_0$	$\mu\neq\mu_0$	$\lvert z\rvert>u_{1-\frac{\alpha}{2}}$	$\lvert t\rvert>t_{1-\frac{\alpha}{2}}(n-1)$
II	$\mu=\mu_0$	$\mu>\mu_0$	$z>u_{1-\alpha}$	$t>t_{1-\alpha}(n-1)$
III	$\mu=\mu_0$	$\mu<\mu_0$	$z<-u_{1-\alpha}$	$t<-t_{1-\alpha}(n-1)$

（二）单个正态总体方差检验

设 X_1，X_2，…，X_n 是来自正态总体 $N(\mu,\sigma^2)$ 的样本，欲检验假设：

$H_0:\sigma^2=\sigma_0^2$　　$H_1:\sigma^2\neq\sigma_0^2$（或 $\sigma^2>\sigma_0^2$ 或 $\sigma^2<\sigma_0^2$）

这叫 χ^2 检验.

	H_0	H_1	均值 μ 已知 统计量 $\chi^2=\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i^2-\mu)^2$	均值 μ 未知 统计量 $\chi^2=\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i^2-\bar{X})^2$
			在显著水平 α 下拒绝 H_0，若	
I	$\sigma^2=\sigma_0^2$	$\sigma^2\neq\sigma_0^2$	$\chi^2<\chi^2_{\frac{\alpha}{2}}(n)$ 或 $\chi^2>\chi^2_{1-\frac{\alpha}{2}}(n)$	$\chi^2<\chi^2_{\frac{\alpha}{2}}(n-1)$ 或 $\chi^2>\chi^2_{1-\frac{\alpha}{2}}(n-1)$
II	$\sigma^2=\sigma_0^2$	$\sigma^2>\sigma_0^2$	$\chi^2>\chi^2_{1-\alpha}(n)$	$\chi^2>\chi^2_{1-\alpha}(n-1)$
III	$\sigma^2=\sigma_0^2$	$\sigma^2<\sigma_0^2$	$\chi^2<\chi^2_{\alpha}(n)$	$\chi^2<\chi^2_{\alpha}(n-1)$

（三）两个正态总体均值检验

1、σ_1^2 与σ_2^2 已知时　构造统计量 $z=\dfrac{\bar{X}-\bar{Y}}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$.

2、σ_1^2 与σ_2^2 未知但相等时

构造统计量 $t=\dfrac{\bar{X}-\bar{Y}}{\sqrt{(n_1-1)s_1^2+(n_2-1)s_2^2}}\sqrt{\dfrac{n_1n_2(n_1+n_2-2)}{n_1+n_2}}$，

	H_0	H_1	方差σ_1^2,σ_2^2已知 统计量 z	方差σ_1^2,σ_2^2未知但相等 统计量 t
			在显著水平 α 下拒绝 H_0，若	
Ⅰ	$\mu_1=\mu_2$	$\mu_1\neq\mu_2$	$\lvert z\rvert>u_{1-\frac{\alpha}{2}}$	$\lvert t\rvert>t_{1-\frac{\alpha}{2}}(n_1+n_2-2)$
Ⅱ	$\mu_1=\mu_2$	$\mu_1>\mu_2$	$z>u_{1-\alpha}$	$t>t_{1-\alpha}(n_1+n_2-2)$
Ⅲ	$\mu_1=\mu_2$	$\mu_1<\mu_2$	$z<-u_{1-\alpha}$	$t<-t_{1-\alpha}(n_1+n_2-2)$

（四）两个正态总体方差检验

设样本 $X_1, X_2, \cdots, X_{n1}$ 与 $Y_1, Y_2, \cdots, Y_{n2}$ 分别来自正态总体$N(\mu_1,\sigma_1^2)$ 与 $N(\mu_2,\sigma_2^2)$，检验假设：

$H_0:\sigma_1^2=\sigma_2^2$　　$H_1:\sigma_1^2\neq\sigma_2^2$（或$\sigma_1^2>\sigma_2^2$ 或$\sigma_1^2<\sigma_2^2$）

	H_0	H_1	均值μ_1,μ_2 已知 统计量F_0	均值μ_1,μ_2 未知 统计量F
			在显著水平α 下拒绝 H_0，若	
I	$\sigma_1^2=\sigma_2^2$	$\sigma_1^2\neq\sigma_2^2$	$F_0>F_{1-\frac{\alpha}{2}}(n_1,n_2)$ 或 $F_0<\dfrac{1}{F_{1-\frac{\alpha}{2}}(n_2,n_1)}$	$F>F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)$ 或 $F<\dfrac{1}{F_{1-\frac{\alpha}{2}}(n_2-1,n_1-1)}$
II	$\sigma_1^2=\sigma_2^2$	$\sigma_1^2>\sigma_2^2$	$F_0>F_{1-\alpha}(n_1,n_2)$	$F>F_{1-\alpha}(n_1-1,n_2-1)$
III	$\sigma_1^2=\sigma_2^2$	$\sigma_1^2<\sigma_2^2$	$F_0<\dfrac{1}{F_{1-\alpha}(n_2,n_1)}$	$F<\dfrac{1}{F_{1-\alpha}(n_2-1,n_1-1)}$

$$F_0=\frac{\frac{1}{n_1}\sum_{i=1}^{n_1}(X_i-\mu_1)^2}{\frac{1}{n_2}\sum_{i=1}^{n_2}(Y_i-\mu_2)^2},\quad F=\frac{s_1^2}{s_2^2}\ （设 s_1^2\geq s_2^2）$$

二、非参数检验

（一）皮尔逊χ^2 拟合检验法

（二）概率纸检验法

概率纸是一种判断总体分布的简便工具.使用它们，可以很快地判断总体分布的类型.概率纸的种类很多.

如果一个总体的分布 F（X）是正态的，则（x，F（x））点在正态概率纸上应呈一条直线.设 X_1，X_2，…，X_n 是从正态总体中抽得的样本观测值，将它们按大小排列后，记作 $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$. 则当 n 较大时，样本的经验分布函数 F_n（x）和理论分布 F（x）很接近. 因此，如果用（x，F（x））画图，则必应近似为一条直线.

返回

统计工具箱中的基本统计命令

1.数据的录入、保存和调用

2.基本统计量

3.常见概率分布的函数

4. 频 数 直 方 图 的 描 绘

5.参数估计

6.假设检验

7.综合实例

返回

统计工具箱中的基本统计命令

一、数据的录入、保存和调用

例1 上海市区社会商品零售总额和全民所有制职工工资总额的数据如下

年份	78	79	80	81	82	82	84	85	86	87
职工工资总额（亿元）	23.8	27.6	31.6	32.4	33.7	34.9	43.2	52.8	63.8	73.4
商品零售总额（亿元）	41.4	51.8	61.7	67.9	68.7	77.5	95.9	137.4	155.0	175.0

方法1

1、年份数据以1为增量，用产生向量的方法输入。

命令格式： x=a:h:b

t=78:87

2、分别以x和y代表变量职工工资总额和商品零售总额。

x=[23.8,27.6,31.6,32.4,33.7,34.9,43.2,52.8,63.8,73.4]

y=[41.4,51.8,61.7,67.9,68.7,77.5,95.9,137.4,155.0,175.0]

3、将变量t、x、y的数据保存在文件data中。

save data t x y

4、进行统计分析时，调用数据文件data中的数据。

load data

To MATLAB(txy)

方法2

1、输入矩阵：

```
data=[78,79,80,81,82,83,84,85,86,87,88;
      23.8,27.6,31.6,32.4,33.7,34.9,43.2,52.8,63.8,73.4;
      41.4,51.8,61.7,67.9,68.7,77.5,95.9,137.4,155.0,175.0]
```

2、将矩阵data的数据保存在文件data1中：save data1 data

3、进行统计分析时，先用命令：　　load data1

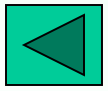
调用数据文件data1中的数据，再用以下命令分别将矩阵data的第一、二、三行的数据赋给变量t、x、y：

```
t=data(1,:)
x=data(2,:)
y=data(3,:)
```

若要调用矩阵data的第j列的数据，可用命令：

```
data(:,j)
```

To MATLAB(data)

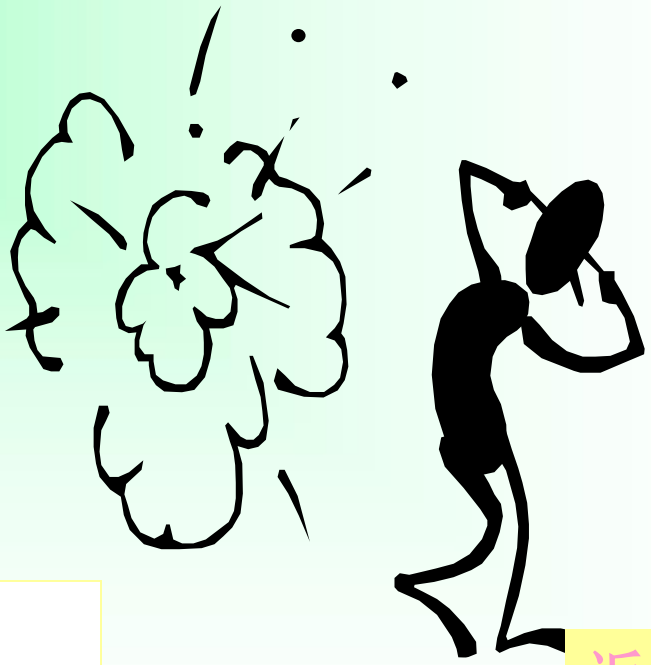


返回

二、基本统计量

对随机变量x，计算其基本统计量的命令如下：

均值：mean(x)
中位数：median(x)
标准差：std(x)
方差：var(x)
偏度：**skewness(x)**
峰度：**kurtosis(x)**



例 对例1中的职工工资总额x，可计算上述基本统计量。

To MATLAB(tjl)

返回

三、常见概率分布的函数

常见的几种分布的命令字符为：

正态分布：norm　　指数分布：exp

帕松分布：poiss　　β 分布：beta

威布尔分布：weib　　χ^2 分布：chi2

t 分布：t　　F 分布：F

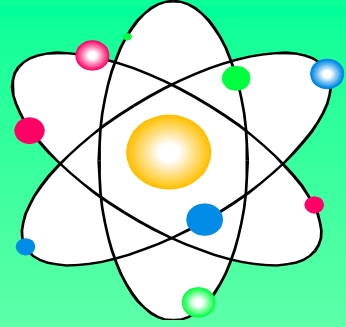
Matlab工具箱对每一种分布都提供五类函数，其命令字符为：

概率密度：pdf　　概率分布：cdf

逆概率分布：inv　　均值与方差：stat

随机数生成：**rnd**

（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）和参数即可.）



如对均值为mu、标准差为sigma的正态分布，举例如下：

1、密度函数： p=normpdf(x,mu,sigma) (当mu=0,sigma=1时可缺省)

例 2 画出正态分布 $N(0,1)$ 和 $N(0,2^2)$ 的概率密度函数图形.

在Matlab中输入以下命令：

```
x=-6:0.01:6;
y=normpdf(x);
z=normpdf(x,0,2);
plot(x,y,x,z)
```

To MATLAB(liti2)

2、概率分布： P=normcdf(x,mu,sigma)

例 3. 计算标准正态分布的概率 P{-1<X<1}.

命令为：P=normcdf(1)-normcdf(-1)

结果为：P =0.6827

To MATLAB(liti3)

3、逆概率分布： x=norminv(P,mu,sigma). 即求出x，使得P{X<x}=P.此命令可用来求分位数.

例 4 取$\alpha = 0.05$，求$u_{1-\frac{\alpha}{2}}$

$u_{1-\frac{\alpha}{2}}$ 的含义是：$X \sim N(0,1)$，$P\{X<u_{1-\frac{\alpha}{2}}\}=1-\frac{\alpha}{2}$

$\alpha = 0.05$ 时，P=0.975, $u_{0.975}$ = norminv(0.975)=1.96

To MATLAB(liti4)

4、均值与方差：[m,v]=normstat(mu,sigma)

例5 求正态分布$N(3,5^2)$的均值与方差.

命令为：[m,v]=normstat(3,5)

结果为：m=3,v=25

To MATLAB(liti5)

5、随机数生成：normrnd(mu,sigma,m,n).产生m n阶的正态分布随机数矩阵.

例6 命令：M=normrnd([1 2 3;4 5 6],0.1,2,3)

结果为：M=0.9567 2.0125 2.8854

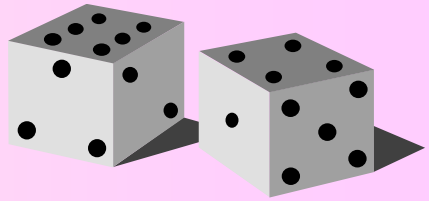
3.8334 5.0288 6.1191

此命令产生了2 3的正态分布随机数矩阵，各数分别服从$N(1,0.1^2)$, $N(2,2^2)$, $N(3,3^2)$, $N(4,0.1^2)$, $N(5,2^2)$,$N(6,3^2)$

返回

To MATLAB（liti6）

四、频数直方图的描绘



1、给出数组data的**频数表**的命令为：

[N,X]=hist(data,k)

此命令将区间[min(data),max(data)]分为k个小区间（缺省为10），返回数组data落在每一个小区间的频数N和每一个小区间的中点X.

2、描绘数组data的**频数直方图**的命令为：

hist(data,k)

返回

五、参数估计

1、正态总体的参数估计

设总体服从正态分布，则其点估计和区间估计可同时由以下命令获得：

[muhat,sigmahat,muci,sigmaci] = normfit(X,alpha)

此命令在显著性水平alpha下估计数据X的参数（alpha缺省时设定为0.05），返回值muhat是X的均值的点估计值，sigmahat是标准差的点估计值，muci是均值的区间估计,sigmaci是标准差的区间估计.

2、其它分布的参数估计

有两种处理办法:

一. 取容量充分大的样本（n>50），按中心极限定理，它近似地服从正态分布;

二.使用Matlab工具箱中具有特定分布总体的估计命令.

（1）[muhat, muci] = expfit(X,alpha)----- 在显著性水平alpha下，求指数分布的数据X的均值的点估计及其区间估计.

（2）[lambdahat, lambdaci] = poissfit(X,alpha)----- 在显著性水平alpha下，求泊松分布的数据X 的参数的点估计及其区间估计.

（3）[phat, pci] = weibfit(X,alpha)----- 在显著性水平alpha下，求Weibull分布的数据X 的参数的点估计及其区间估计.

返回

六、假设检验

在总体服从正态分布的情况下，可用以下命令进行假设检验.

1、总体方差sigma²已知时，总体均值的检验使用 z-检验

[h,sig,ci] = ztest(x,m,sigma,alpha,tail)

检验数据 x 的关于均值的某一假设是否成立，其中sigma 为已知方差， alpha 为显著性水平，究竟检验什么假设取决于 tail 的取值：

tail = 0，检验假设“x 的均值等于 m ”

tail = 1，检验假设“x 的均值大于 m ”

tail =-1，检验假设“x 的均值小于 m ”

tail的缺省值为 0， alpha的缺省值为 0.05.

返回值 h 为一个布尔值，h=1 表示可以拒绝假设，h=0 表示不可以拒绝假设，sig 为假设成立的概率，ci 为均值的 1-alpha 置信区间.

例7 Matlab统计工具箱中的数据文件gas.mat.中提供了美国1993年一月份和二月份的汽油平均价格（price1,price2分别是一，二月份的油价，单位为美分），它是容量为20的双样本.假设一月份油价的标准偏差是一加仑四分币（σ=4），试检验一月份油价的均值是否等于115.

解 作假设：m = 115.

首先取出数据，用以下命令：

load gas

然后用以下命令检验

```
[h,sig,ci] = ztest(price1,115,4)
```

返回：h = 0，sig = 0.8668，ci = [113.3970 116.9030].

To MATLAB（liti7）

检验结果：1. 布尔变量h=0, 表示不拒绝零假设. 说明提出的假设均值115是合理的.

2. sig-值为0.8668, 远超过0.5, 不能拒绝零假设

3. 95%的置信区间为[113.4, 116.9], 它完全包括115, 且精度很高.

.

2、总体方差sigma²未知时，总体均值的检验使用t-检验

[h,sig,ci] = ttest(x,m,alpha,tail)

检验数据 x 的关于均值的某一假设是否成立，其中 alpha 为显著性水平，究竟检验什么假设取决于 tail 的取值：

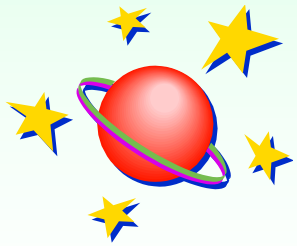
tail = 0，检验假设“x 的均值等于 m ”

tail = 1，检验假设“x 的均值大于 m ”

tail =-1，检验假设“x 的均值小于 m ”

tail的缺省值为 0， alpha的缺省值为 0.05.

返回值 h 为一个布尔值，h=1 表示可以拒绝假设，h=0 表示不可以拒绝假设，sig 为假设成立的概率，ci 为均值的 1-alpha 置信区间.



例8 试检验例8中二月份油价 Price2的均值是否等于115.

解 作假设：m = 115，

price2为二月份的油价，不知其方差，故用以下命令检验

[h,sig,ci] = ttest(price2 ,115)

To MATLAB（liti8）

返回：h = 1，sig = 4.9517e-004，ci =[116.8 120.2].

检验结果: 1. 布尔变量h=1, 表示拒绝零假设. 说明提出的假设油价均值115是不合理的.

2. 95%的置信区间为[116.8 120.2], 它不包括115, 故不能接受假设.

3. sig-值为4.9517e-004, 远小于0.5, 不能接受零假设.

3、两总体均值的假设检验使用 t-检验

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

检验数据 x ，y 的关于均值的某一假设是否成立，其中 alpha 为显著性水平，究竟检验什么假设取决于 tail 的取值：

tail = 0，检验假设“x 的均值等于 y 的均值 ”

tail = 1，检验假设“x 的均值大于 y 的均值 ”

tail =-1，检验假设“x 的均值小于 y 的均值 ”

tail的缺省值为 0， alpha的缺省值为 0.05.

返回值 h 为一个布尔值，h=1 表示可以拒绝假设，h=0 表示不可以拒绝假设，sig 为假设成立的概率，ci 为与x与y均值差的的 1-alpha 置信区间.

例9 试检验例8中一月份油价Price1与二月份的油价Price2均值是否相同.

解 用以下命令检验

```
[h,sig,ci] = ttest2(price1,price2)
```

To MATLAB（liti9）

返回：h = 1，sig = 0.0083，ci =[-5.8,-0.9].

检验结果:1. 布尔变量h=1，表示拒绝零假设. 说明提出的假设“油价均值相同”是不合理的.

2. 95%的置信区间为[-5.8,-0.9],说明一月份油价比二月份油价约低1至6分.

3. sig-值为0.0083, 远小于0.5, 不能接受“油价均相同”假设.

4、非参数检验：总体分布的检验

Matlab工具箱提供了两个对总体分布进行检验的命令：

(1) h = normplot(x)

此命令显示数据矩阵x的正态概率图.如果数据来自于正态分布，则图形显示出直线性形态.而其它概率分布函数显示出曲线形态.

（2）h = weibplot(x)

此命令显示数据矩阵x的Weibull概率图.如果数据来自于Weibull分布，则图形将显示出直线性形态.而其它概率分布函数将显示出曲线形态.

返回

例10 一道工序用自动化车床连续加工某种零件，由于刀具损坏等会出现故障.故障是完全随机的，并假定生产任一零件时出现故障机会均相同.工作人员是通过检查零件来确定工序是否出现故障的.现积累有100次故障纪录，故障出现时该刀具完成的零件数如下：

459 362 624 542 509 584 433 748 815 505
612 452 434 982 640 742 565 706 593 680
926 653 164 487 734 608 428 1153 593 844
527 552 513 781 474 388 824 538 862 659
775 859 755 49 697 515 628 954 771 609
402 960 885 610 292 837 473 677 358 638
699 634 555 570 84 416 606 1062 484 120
447 654 564 339 280 246 687 539 790 581
621 724 531 512 577 496 468 499 544 645
764 558 378 765 666 763 217 715 310 851

试观察该刀具出现故障时完成的零件数属于哪种分布.

解 1、数据输入

To MATLAB（liti101）

2、作频数直方图

To MATLAB（liti102）

```
hist(x,10)
```

（看起来刀具寿命服从正态分布）

3、分布的正态性检验

To MATLAB（liti103）

normplot(x)

（刀具寿命近似服从正态分布）

4、参数估计：

To MATLAB（liti104）

[muhat,sigmahat,muci,sigmaci] = normfit(x)

估计出该刀具的均值为594，方差204，均值的0.95置信区间为[553.4962，634.5038]，方差的0.95置信区间为[179.2276，237.1329].

5、假设检验

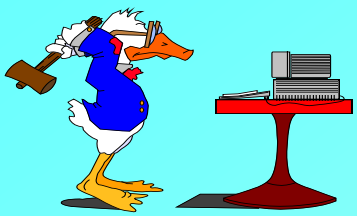
To MATLAB
(liti105)

已知刀具的寿命服从正态分布，现在方差未知的情况下，检验其均值 m 是否等于594.

结果：h = 0， sig = 1， ci =[553.4962， 634.5038].

检验结果: 1. 布尔变量h=0, 表示不拒绝零假设. 说明提出的假设寿命均值594是合理的.
2. 95%的置信区间为[553.5，634.5], 它完全包括594, 且精度很高.
3. sig-值为1, 远超过0.5, 不能拒绝零假设.

返回



作　业

1、某校60名学生的一次考试成绩如下:

93 75 83 93 91 85 84 82 77 76 77 95 94 89 91 88 86
83 96 81 79 97 78 75 67 69 68 84 83 81 75 66 85 70
94 84 83 82 80 78 74 73 76 70 86 76 90 89 71 66 86
73 80 94 79 78 77 63 53 55

1)计算均值、标准差、极差、偏度、峰度，画出直方图;
2)检验分布的正态性;
3)若检验符合正态分布，估计正态分布的参数并检验参数.

2、据说某地汽油的价格是每加仑115美分，为了验证这种说法，一位学者开车随机选择了一些加油站，得到某年一月和二月的数据如下：

一月：119 117 115 116 112 121 115 122 116 118 109 112 119 112 117 113 114 109 109 118

二月：118 119 115 122 118 121 120 122 128 116 120 123 121 119 117 119 128 126 118 125

1）分别用两个月的数据验证这种说法的可靠性；

2）分别给出1月和2月汽油价格的置信区间；

3）给出1月和2月汽油价格差的置信区间.

谢谢大家