

MATLAB统计分析与应用

——参数估计与假设检验

主讲人：谢中华

ssstudy.com

科学软件学习网

主要内容

- 参数估计
- 正态总体参数的检验
- 常用非参数检验

第一节 参数估计

一、分布参数估计的MATLAB函数

函数名	说 明	函数名	说 明
betafit	β 分布的参数估计	lognfit	对数正态分布的参数估计
binofit	二项分布的参数估计	mle	最大似然估计 (MLE)
dfittool	分布拟合工具	mlecov	最大似然估计的渐进协方差矩阵
evfit	极值分布的参数估计	nbinfit	负二项分布的参数估计
expfit	指数分布的参数估计	normfit	正态 (高斯) 分布的参数估计
fitdist	分布的拟合	poissfit	泊松分布的参数估计
gamfit	Γ 分布的参数估计	raylfir	瑞利 (Rayleigh) 分布的参数估计
gevfit	广义极值分布的参数估计	unifit	均匀分布的参数估计
gmdistribution	高斯混合模型的参数估计	wblfit	威布尔 (Weibull) 分布的参数估计
gpfir	广义 Pareto 分布的参数估计		

二、常见分布的参数估计

【例 9.1-1】从某厂生产的滚珠中随机抽取 10 个，测得滚珠的直径（单位：mm）如下：

15.14 14.81 15.11 15.26 15.08 15.17 15.12 14.95 15.05 14.87

若滚珠直径服从正态分布 $N(\mu, \sigma^2)$ ，其中 μ, σ 未知，求

μ, σ 的最大似然估计和置信水平为 90% 的置信区间。

```
>> x = [15.14,14.81,15.11,15.26,15.08,15.17,15.12,14.95,15.05,14.87];
```

```
>> [muhat,sigmahat,muci,sigmaci] = normfit(x,0.1)
```

```
>> [mu_sigma,mu_sigma_ci] = mle(x,'distribution','norm','alpha',0.1)
```

【例9.1-2】 调用normrnd函数生成100个服从均值为10，标准差为4的正态分布的随机数，然后调用mle函数求均值和标准差的最大似然估计。

```
>> x = normrnd(10,4,100,1);
```

```
>> [phat,pci] = mle(x)
```

```
>> [phat,pci] = mle(x,'distribution','normal')
```

```
>> [phat,pci] = mle(x,'pdf',@normpdf,'start',[0,1])
```

```
>> [phat,pci] = mle(x,'cdf',@normcdf,'start',[0,1])
```

三、自定义分布的参数估计

1. 单参数情形

【例 9.1-3】已知总体 X 的密度函数为

$$f(x; \theta) = \begin{cases} \theta x^{\theta-1}, & 0 < x < 1 \\ 0, & \text{其他} \end{cases}$$

其中 $\theta > 0$ 是未知参数。现从总体 X 中随机抽取容量为 20 的样本，得样本观测值如下：

0.7917	0.8448	0.9802	0.8481	0.7627	0.9013	0.9037	0.7399	0.7843	0.8424
0.9842	0.7134	0.9959	0.6444	0.8362	0.7651	0.9341	0.6515	0.7956	0.8733

试根据以上样本观测值求参数 θ 的最大似然估计和置信水平为 95% 的置信区间。

```
>> x = [0.7917,0.8448,0.9802,0.8481,0.7627  
0.9013,0.9037,0.7399,0.7843,0.8424  
0.9842,0.7134,0.9959,0.6444,0.8362  
0.7651,0.9341,0.6515,0.7956,0.8733];
```

```
>> PdfFun = @(x,theta) theta*x.^(theta-1).*(x>0 & x<1);
```

```
>> [phat,pci] = mle(x(:),'pdf',PdfFun,'start',1)
```

2. 多参数情形

【例 9.1-4】设总体 X 服从由正态分布和 I 型极小值分布（即 Gumbel 分布）混合而成的混合分布，两种分布的比例分别为 0.6 和 0.4。总体 X 的密度函数为：

$$f(x) = \frac{0.6}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) + \frac{0.4}{\sigma_2} \exp\left(\frac{x-\mu_2}{\sigma_2}\right) \exp\left(-\exp\left(\frac{x-\mu_2}{\sigma_2}\right)\right)$$

试根据例 8.2-10 中生成的随机数求参数 $\mu_1, \sigma_1, \mu_2, \sigma_2$ 的最大似然估计和置信水平为 95% 的置信区间。

```
>> rand('seed',1);
>> randn('seed',1);
>> x = normrnd(35,5,1000,1);
>> y = evrnd(20,2,1000,1);
>> z = randsrc(1000,1,[1,2;0.6,0.4]);
>> data = x.*(z==1) + y.*(z==2);
>> pdffun = @(t,mu1,sig1,mu2,sig2)...
    0.6*normpdf(t,mu1,sig1)+0.4*evpdf(t,mu2,sig2);
>> [phat,pci] = mle(data,'pdf',pdffun,'start',[10,10,10,10],...
    'lowerbound',[-inf,0,-inf,0],'upperbound',[inf,inf,inf,inf])
```

第二节 正态总体参数的检验

一、总体标准差已知时的单个正态总体均值的U检验

总体: $X \sim N(\mu, \sigma_0^2)$

样本: $X_1, X_2, \dots, X_n$

假设:

$$H_0: \mu = \mu_0, \quad H_1: \mu \neq \mu_0 .$$

$$H_0: \mu \geq \mu_0, \quad H_1: \mu < \mu_0$$

$$H_0: \mu \leq \mu_0, \quad H_1: \mu > \mu_0$$

➤ ztest函数

调用格式：

$h = \text{ztest}(x, m, \text{sigma})$

$h = \text{ztest}(\dots, \alpha)$

$h = \text{ztest}(\dots, \alpha, \text{tail})$

$h = \text{ztest}(\dots, \alpha, \text{tail}, \text{dim})$

$[h, p] = \text{ztest}(\dots)$

$[h, p, ci] = \text{ztest}(\dots)$

$[h, p, ci, zval] = \text{ztest}(\dots)$

【例 9.2-1】某切割机正常工作时，切割的金属棒的长度服从正态分布 $N(100,4)$ 。从该切割机切割的一批金属棒中随机抽取 15 根，测得它们的长度（单位：mm）如下：

97 102 105 112 99 103 102 94 100 95 105 98
102 100 103.

假设总体方差不变，试检验该切割机工作是否正常，即总体均值是否等于 100mm？取显著性水平 $\alpha = 0.05$ 。

```
>> x = [97 102 105 112 99 103 102 94 100 95  
105 98 102 100 103];  
>> [h,p,muci,zval] = ztest(x,100,2,0.05)  
>> [h,p,muci,zval] = ztest(x,100,2,0.05,'right')
```

二、总体标准差未知时的单个正态总体均值的t检验

总体: $X \sim N(\mu, \sigma^2)$

样本: $X_1, X_2, \dots, X_n$

假设:

$$H_0: \mu = \mu_0, \quad H_1: \mu \neq \mu_0 .$$

$$H_0: \mu \geq \mu_0, \quad H_1: \mu < \mu_0$$

$$H_0: \mu \leq \mu_0, \quad H_1: \mu > \mu_0$$

➤ **ttest函数**

调用格式：

`h = ttest(x)`

`h = ttest(x,m)`

`h = ttest(x,y)`

`h = ttest(...,alpha)`

`h = ttest(...,alpha,tail)`

`[h,p,ci,stats] = ttest(...,alpha,tail,dim)`

【例 9.2-2】化肥厂用自动包装机包装化肥，某日测得 9 包化肥的质量（单位：kg）如下：

49.4 50.5 50.7 51.7 49.8 47.9 49.2 51.4 48.9

设每包化肥的质量服从正态分布，是否可以认为每包化肥的平均质量为 50kg？取显著性水平 $\alpha = 0.05$ 。

```
>> x = [49.4 50.5 50.7 51.7 49.8 47.9 49.2 51.4 48.9];  
>> [h,p,muci,stats] = ttest(x,50,0.05)
```

三、总体标准差未知时的两个正态总体均值的比较 t检验

总体1: $X \sim N(\mu_1, \sigma_1^2)$

样本1: $X_1, X_2, \dots, X_{n_1}$

总体2: $Y \sim N(\mu_2, \sigma_2^2)$

样本2: $Y_1, Y_2, \dots, Y_{n_2}$

假设:

$$H_0: \mu_1 = \mu_2, \quad H_1: \mu_1 \neq \mu_2$$

$$H_0: \mu_1 \geq \mu_2, \quad H_1: \mu_1 < \mu_2$$

$$H_0: \mu_1 \leq \mu_2, \quad H_1: \mu_1 > \mu_2$$

1. 两独立样本的比较 t检验

➤ ttest2函数

调用格式：

`h = ttest2(x,y)`

`h = ttest2(x,y,alpha)`

`h = ttest2(x,y,alpha,tail)`

`h = ttest2(x,y,alpha,tail,vartype)`

`[h,p,ci,stats] =`

`ttest2(x,y,alpha,tail,vartype,dim)`

【例 9.2-3】甲、乙两台机床加工同一种产品，从这两台机床加工的产品中随机抽取若干件，测得产品直径（单位：mm）为：

甲机床：20.1, 20.0, 19.3, 20.6, 20.2, 19.9, 20.0, 19.9, 19.1, 19.9 .

乙机床：18.6, 19.1, 20.0, 20.0, 20.0, 19.7, 19.9, 19.6, 20.2 .

设甲、乙两机床加工的产品的直径分别服从正态分布 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$ ，试比较甲、乙两台机床加工的产品的直径是否有显著差异？取显著性水平 $\alpha = 0.05$.

```
>> x = [20.1, 20.0, 19.3, 20.6, 20.2, 19.9, 20.0,  
19.9, 19.1, 19.9];  
>> y = [18.6, 19.1, 20.0, 20.0, 20.0, 19.7, 19.9,  
19.6, 20.2];  
>> alpha = 0.05;  
>> tail = 'both';  
>> vartype = 'equal';  
>> [h,p,muci,stats] = ttest2(x,y,alpha,tail,vartype)
```

2. 配对样本的比较 t检验

➤ ttest函数

调用格式：

`h = ttest(x,y)`

`h = ttest(...,alpha)`

`h = ttest(...,alpha,tail)`

`h = ttest(...,alpha,tail,dim)`

`[h,p] = ttest(...)`

`[h,p,ci] = ttest(...)`

`[h,p,ci,stats] = ttest(...)`

【例 9.2-4】想要比较某种减肥药的疗效，选定了 10 个人进行试验，收集了每个人服用减肥药前后的体重数据，如下表所列。试分析服药后体重是否有显著的降低。取显著性水平 $\alpha = 0.05$ 。

试验者 前或后	1	2	3	4	5	6	7	8	9	10
服药前	90	79	86	88	92	79	76	87	102	96
服药后	83	70	80	84	87	74	79	83	96	90

```
x = [80.3,68.6,72.2,71.5,72.3,70.1,74.6,73.0,58.7,78.6,85.6,78.0];  
y = [74.0,71.2,66.3,65.3,66.0,61.6,68.8,72.6,65.7,72.6,77.1,71.5];  
Alpha = 0.05;  
tail = 'both';  
[h,p,muci,stats] = ttest(x,y,Alpha,tail)
```

四、总体均值未知时的单个正态总体方差的卡方检验

总体: $X \sim N(\mu, \sigma^2)$

样本: $X_1, X_2, \dots, X_n$

假设:

$$H_0: \sigma^2 = \sigma_0^2, \quad H_1: \sigma^2 \neq \sigma_0^2$$

$$H_0: \sigma^2 \geq \sigma_0^2, \quad H_1: \sigma^2 < \sigma_0^2$$

$$H_0: \sigma^2 \leq \sigma_0^2, \quad H_1: \sigma^2 > \sigma_0^2$$

➤ **vartest函数**

调用格式：

$H = \text{vartest}(X,V)$

$H = \text{vartest}(X,V,\alpha)$

$H = \text{vartest}(X,V,\alpha,\text{tail})$

$[H,P] = \text{vartest}(\dots)$

$[H,P,CI] = \text{vartest}(\dots)$

$[H,P,CI,STATS] = \text{vartest}(\dots)$

$[\dots] = \text{vartest}(X,V,\alpha,\text{tail},\text{dim})$

【例 9.2-4】根据例 9.2-2 中的样本观测数据检验每包化肥的质量的方差是否等于 1.5？取显著性水平 $\alpha = 0.05$ 。

```
>> x = [49.4 50.5 50.7 51.7 49.8 47.9 49.2 51.4 48.9];  
>> var0 = 1.5;  
>> alpha = 0.05;  
>> tail = 'both';  
>> [h,p,varci,stats] = vartest(x,var0,alpha,tail)
```

五、总体均值未知时的两个正态总体方差的比较 F 检验

总体1: $X \sim N(\mu_1, \sigma_1^2)$

样本1: $X_1, X_2, \dots, X_{n_1}$

总体2: $Y \sim N(\mu_2, \sigma_2^2)$

样本2: $Y_1, Y_2, \dots, Y_{n_2}$

假设:

$$H_0 : \sigma_1^2 = \sigma_2^2, \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

$$H_0 : \sigma_1^2 \geq \sigma_2^2, \quad H_1 : \sigma_1^2 < \sigma_2^2$$

$$H_0 : \sigma_1^2 \leq \sigma_2^2, \quad H_1 : \sigma_1^2 > \sigma_2^2$$

➤ **vartest2函数**

调用格式：

$H = \text{vartest2}(X,Y)$

$H = \text{vartest2}(X,Y,\alpha)$

$H = \text{vartest2}(X,Y,\alpha,\text{tail})$

$[H,P] = \text{vartest2}(\dots)$

$[H,P,CI] = \text{vartest2}(\dots)$

$[H,P,CI,STATS] = \text{vartest2}(\dots)$

$[\dots] = \text{vartest2}(X,Y,\alpha,\text{tail},\text{dim})$

【例 9.2-5】根据例 9.2-3 中的样本观测数据检验甲、乙两台机床加工的产品的直径的方差是否相等？取显著性水平 $\alpha = 0.05$.

```
>> x = [20.1, 20.0, 19.3, 20.6, 20.2, 19.9, 20.0, 19.9,  
19.1, 19.9];
```

```
>> y = [18.6, 19.1, 20.0, 20.0, 20.0, 19.7, 19.9, 19.6,  
20.2];
```

```
>> alpha = 0.05;
```

```
>> tail = 'both';
```

```
>> [h,p,varci,stats] = vartest2(x,y,alpha,tail)
```

第三节 常用非参数检验

一、游程检验

作用：用来检验来自于同一总体样本数据是否随机

1. 游程的定义

➤ 以时间顺序或其他顺序排列的有序数列中，具有相同的事件或符号的连续部分称为一个**游程**，通常用 R 表示游程总个数。

++++-----++++-----++-----
1111 0000 111 000 11 0000

2. 游程检验基本原理

H_0 : 数据出现顺序随机, H_1 : 数据出现顺序不随机

- 求出样本中位数, 将样本观测值分为大于中位数和小于中位数的两个部分。用1, 0 (或 + -) 交错形成的序列的游程个数来检验样本是否随机。
- 在固定样本量之下, 如果游程个数过少, 说明0和1相对比较集中, 如果游程过多, 说明0和1交替周期特征明显, 这都不符合序列随机性的要求。也就是说游程个数过多或过少都应拒绝原假设。

$$W = \{R \leq R_{1,\alpha/2} \text{ 或 } R \geq R_{2,\alpha/2}\}$$

3. 游程检验的MATLAB函数

➤ runtest函数

调用格式：

`h = runtest(x)`

`h = runtest(x,v)`

`h = runtest(x,'ud')`

`h = runtest(...,param1,val1,param2,val2,...)`

`[h,p] = runtest(...)`

`[h,p,stats] = runtest(...)`

【例 9.3-1】中国福利彩票“双色球”开奖号码由 6 个红色球号码和 1 个蓝色球号码组成。红色球号码从 1~33 中随机选择；蓝色球号码从 1~16 中随机选择。现收集了 2012 年 1 月 1 日~2012 年 8 月 19 日共 97 期双色球开奖数据，完整数据保存在文件“2012 双色球开奖数据.xls”中。试根据收集到的 97 组数据研究蓝色球号码出现顺序是否随机？取显著性水平 $\alpha = 0.05$ 。

```
>> x = xlsread('2012双色球开奖数据.xls',1,'I2:I98');  
>> [h,p,stats] = runstest(x,[],'method','approximate')
```

二、符号检验

作用：用来检验中位数是否等于给定常数。

$$H_0 : M_e = M_0, \quad H_1 : M_e \neq M_0$$

1. 符号检验的MATLAB函数

➤ **signtest函数**

调用格式：

`[p,h,stats] = signtest(x)`

`[p,h,stats] = signtest(x,m,param1,val1,...)`

`[p,h,stats] = signtest(x,y,param1,val1,...)`

【例 9.3-2】在某国总统选举的民意调查中，随机询问了 200 名选民，结果显示，69 人支持甲候选人，108 人支持乙候选人，23 人弃权，试分析甲、乙两位候选人的支持率是否有显著差异？取显著性水平 $\alpha = 0.05$ 。

```
>> x = [-ones(69,1);zeros(23,1);ones(108,1)];  
>> p = signtest(x)
```

三、Wilcoxon符号秩检验

作用：用来检验中位数是否等于给定常数。

$$H_0 : M_e = M_0, \quad H_1 : M_e \neq M_0$$

1. Wilcoxon符号秩检验的MATLAB函数

➤ **signrank函数**

调用格式：

`[p,h,stats] = signrank(x)`

`[p,h,stats] = signrank(x,m,param1,val1,...)`

`[p,h,stats] = signrank(x,y,param1,val1,...)`

【例 9.3-3】抽查精细面粉的装包重量，抽查了 16 包，其观测值（单位：kg）

如下：

20.21 19.95 20.15 20.07 19.91 19.99 20.08 20.16

19.99 20.16 20.09 19.97 20.05 20.27 19.96 20.06

试检验平均重量与原来设定的 20kg 是否有显著差别？取显著性水平 $\alpha=0.05$ 。

```
>> x =
```

```
[20.21,19.95,20.15,20.07,19.91,19.99,20.08,20.16,
```

```
...
```

```
19.99,20.16,20.09,19.97,20.05,20.27,19.96,20.06];
```

```
>> [p,h,stats] = signrank(x,20)
```

四、Mann-Whitney秩和检验

作用：对两总体均值作比较检验。

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$$

1. Mann-Whitney秩和检验的MATLAB函数

➤ **ranksum函数**

调用格式：

`[p,h,stats] = ranksum(x,y,param1,val1,...)`

【例 9.3-4】某科研团队研究两种饲料（高蛋白饲料和低蛋白饲料）对雌鼠体重的影响，用高蛋白饲料饲喂 12 只雌鼠，用低蛋白饲料饲喂 7 只雌鼠，记录两组雌鼠在 8 周内体重的增加量，得观测数据如表 6.3-4 所列。试检验不同饲料饲喂的雌鼠的体重增加量是否有显著差异？取显著性水平 $\alpha = 0.05$ 。

饲料	各鼠增加的体重 (g)											
高蛋白	133	112	102	129	121	161	142	88	115	127	96	125
低蛋白	71	119	101	83	107	134	92					

```
>> x = [133,112,102,129,121,161,142,88,115,127,96,125];  
>> y = [71,119,101,83,107,134,92];  
>> [p,h,stats] = ranksum(x,y,'method','approximate')
```

Thank You