实践七 参数估计

一、 实践问题

1. 问题背景

参数估计是数理统计中的一个基本概念和重要的基本方法，是指用样本对总体分布中的未知参数做出的估计，这种估计我们常见的有点估计和区间估计两种. 所谓点估计，就是用样本统计量确定总体参数的一个取值. 评价估计优劣的标准有无偏性、最小方差性、有效性等. 点估计的方法有矩法、极大似然法. 区间估计有置信度、可靠度和精确度的问题.

对于一个总体，可以用一些参数来描述其表征，如数学期望、方差等. 如果总体的某个参数未知，统计学提供了一些方法，它们可以用来估计未知参数介于哪个区间内. 参数估计是数理统计中重要的内容，也是计算量很大的问题. 以前在这方面的教学中都是使用计算器和查表，非常麻烦. 下面我们用 MATLAB 来解决这个问题.

2. 实践目的与要求

(1) 掌握参数估计的 MATLAB 的有关命令;

(2) 熟悉区间估计的置信度、可靠度和精确度的三者关系;

(3) 熟练掌握总体数学期望和方差的点估计和区间估计的相关命令.

二、 实践操作过程

基本数学原理:

离散型随机变量的极大似然估计法:

(1) 似然函数

若 X 为离散型，似然函数为

$$L(\theta)=L(x_1,x_2,...,x_n;\ \theta)=\prod_{i=1}^{n}p(x_i;\theta),\quad \theta\in\Theta$$

(2) 求似然函数$L(\theta)$的最大值点$\hat{\theta}$，则$\hat{\theta}$就是未知参数θ的极大似然估计值.

连续型随机变量的极大似然估计法:

(1) 似然函数

若 X 为连续型，似然函数为

$$L(\theta)=L(x_1,x_2,...,x_n;\ \theta)=\prod_{i=1}^{n}f(x_i;\theta),\quad \theta\in\Theta$$

(2) 求似然函数$L(\theta)$的最大值点$\hat{\theta}$，则$\hat{\theta}$就是未知参数θ的极大似然估计值.

下面是 MATLAB 软件提供的一些常用的参数估计函数命令.

1. 均匀分布的参数估计函数

基本数学原理: 均匀分布的概率密度函数见实践二.

在 MATLAB 中用函数 unifit 来计算均匀分布中参数的极大似然估计值和置信区间，其基本调用格式如下:

• [ahat,bhat]=unifit(x) % 根据给出的均匀分布样本数据 x，计算并返回均匀分布的两个参数 a 和 b 极大似然估计值 ahat 和 bhat;

• [ahat,bhat,aci,bci]=unifit(x) % 除返回参数 a 和 b 的极大似然估计值 ahat 和 bhat 外，还要计算返回置信度为 0.95 的两个参数的置信区间 aci 和 bci;

• [ahat,bhat,aci,bci]=unifit(x,alpha) % 返回的置信区间的置信度为 100(1-alpha)%.

例 7-1 产生 100 行 2 列服从区间(10, 12)上的均匀分布的随机数，计算区间端点“a”和“b”的极大似然估计值，求出置信度为 0.95 的这两个参数的置信区间.

解 在命令窗口中输入:

```
r = unifrnd(10, 12, 100, 2);
[ahat, bhat, aci, bci] = unifit(r)
```

回车后显示:

```
ahat =
    10.0154   10.0060
bhat =
    11.9989  11.9743
aci = 9.9551    9.9461
    10.0154 10.0060
bci =
    11.9989   11.9743
    12.0592   12.034
```

结果表明：以第一列随机数为例，区间端点 a 和 b 的极大似然估计值分别是 10.0154(比“10”略大)和 11.9989(比“12”略小)，置信度为 0.95 的两个参数的置信区间分别是(9.9551, 10.0154)和(11.9989, 12.0592).

2. **指数分布的参数估计函数**

基本数学原理: 指数分布的概率密度函数见实践二. 在 MATLAB 中用函数 expfit 来计算指数分布中参数的极大似然估计值及置信区间,其基本调用格式如下:

- lambdahat=expfit(x)　% 根据指数分布样本数据 x，计算并返回指数分布参数 lambda 的极大似然估计值;
- [lambdahat,lambdaci]=expfit(x)　% 返回参数 lambda 的估计值和置信度为 95%的置信区间;
- [lambdahat,lambdaci]=expfit(x,alpha)　% 返回的置信区间的置信度为 100(1-alpha)%.

例 7-2 产生 1 行 100 列的参数为 3 的指数分布的随机数，计算“λ”的极大似然估计值，求出置信度为 0.99 的参数 λ 的置信区间.

解 在命令窗口中输入:

```
r=exprnd(3,1,100);
[lambdahat,lambdaci]=expfit(r,0.01)
```

回车后显示:

```
lambdahat=
    2.6787
lambdaci=
    2.0988
    3.5190
```

例 7-3 设电池的寿命服从指数分布，其概率密度为

$$f(t)=\begin{cases}\frac{1}{\lambda}e^{-\frac{t}{\lambda}} & t>0,\\ 0 & t\le 0.\end{cases}$$

λ >0 未知. 随机抽取 15 只电池进行寿命试验，测得失效时间(单位:小时)为

115, 119, 131, 138, 142, 147, 148, 155, 158, 159, 163, 166, 167, 170, 172.

试求电池的平均寿命 λ 的极大似然估计值和置信水平为 98%的置信区间.

解 在命令窗口输入：

```
x=[115 119 131 138 142 147 148 155 158 159 163 166 167 170 172];
[lambdahat,lambdaci]=expfit(x,0.02)  % 这里是 expfit(x,0.02),不是 expfit(x,0.98).
```

回车后显示：

```
lambdahat=
          150
lambdaci=
          74.7673
          254.4609
```

结果表明:平均寿命 λ 的极大似然估计值为 150 小时，置信水平为 0.98 的平均寿命 λ 的置信区间为[74.7673, 254.4609].

3. **正态分布的参数估计函数**

基本数学原理: 正态分布的概率密度函数见实践二.

正态总体未知参数的区间估计

单个正态总体	均值μ的置信区间	σ^2已知	$\left(\bar{X}-\frac{\sigma}{\sqrt{n}}z_{\alpha/2},\ \bar{X}+\frac{\sigma}{\sqrt{n}}z_{\alpha/2}\right)$
		σ^2未知	$\left(\bar{X}-\frac{S}{\sqrt{n}}t_{\alpha/2}(n-1),\ \bar{X}+\frac{S}{\sqrt{n}}t_{\alpha/2}(n-1)\right)$
	标准差 σ 的置信区间		$\left(\frac{\sqrt{n-1}S}{\sqrt{\chi^2_{\alpha/2}(n-1)}},\ \frac{\sqrt{n-1}S}{\sqrt{\chi^2_{1-\alpha/2}(n-1)}}\right)$

在 MATLAB 中, 用函数 normfit 来计算正态分布中参数的估计值及置信区 间, 其基本调用格式如下:

- [muhat,sigmahat,muci,sigmaci]=normfit(x) % 根据正态分布数据 x 对参数 mu 和 sigma 做出估计, 返回值 muhat 是 x 的均值的点估计, sigmahat 是标准差的点估计, muci 是均值的置信度为 95%的置信区间, sigmaci 是标准差的置信度为 95%的置信区间.
- [muhat,sigmahat,muci,sigmaci]=normfit(x,alpha) % 返回的置信区间的置信度为 100(1-alpha)%.

例 7-4 分别使用金球和铂球测定引力常数, 数据如下:

(1) 用金球测定观察值为 6.683, 6.681, 6.676, 6.678,6.679, 6.672;

(2) 用铂球测定观察值为 6.661, 6.661, 6.667, 6.667,6.664.

设测定值总体为$N(\mu,\sigma^2)$, μ 和 σ 为未知. 对(1)和(2)两种情况分别求 μ 和 σ 的置信度为 0.9 的置信区间, 并比较这两种情况的精确度.

解 在命令窗口中输入:

```
X=[6.683  6.681  6.676  6.678  6.679  6.672];
Y=[6.661  6.661  6.667  6.667  6.664];
[mu,sigma,muci,sigmaci]=normfit(X,0.1);  %金球测定的估计.
[my,sy,myci,syci]=normfit(Y,0.1);        %铂球测定的估计.
```

回车后显示:

```
mu =
        6.6782
sigma =
```

```
        0.0039
muci =
        6.6750
        6.6813
sigmaci =
        0.0026
        0.0081
my =
        6.6640
sy=
        0.0030
myci =
        6.6611
        6.6669
syci=
        0.0019
        0.0071
```

结果说明：金球测定的μ估计值为 6.6782，置信区间为[6.6750，6.6813]；σ的估计值为0.0039，置信区间为[0.0026 ，0.0081]；泊球测定的μ估计值为 6.6640，置信区间为[6.6611，6.6669]；σ的估计值为 0.0030，置信区间为[0.0019，0.0071]. 可见，铂球测定的精确度高(因为标准差小，关于总体期望μ和总体标准差σ置信区间长度相对要小).

例 7-5 用产生正态分布随机数命令生成一组正态分布随机样本，用 normfit 函数给出该正态分布的参数估计，并分析置信度、 可靠度和精确度三者关系.

解 在命令窗口中输入:

```
xx=normrnd(4,2,50,1); % 生成一组(有 50×1 个)μ=4，σ=2 的正态随机样本.
[muhat,sigmahat,muci,sigmaci]=normfit(xx)
```

回车后显示:

```
muhat=
    4.0787
sigmahat=
    1.9519
muci=
    3.5239
    4.6334
sigmaci=
    1.6305
    2.4324
```

结果说明：给出了μ(这里已知μ=4)和σ(这里已知σ=2)的极大似然估计值分别为muhat＝4.0787 和 sigmahat＝1.9519，μ和σ的置信度为 95%(缺省值)的置信区间分别为[3.5239 , 4.6334]和[1.6305, 2.4324].

若再执行命令:

```
[muhat,sigmahat,muci,sigmaci]=normfit(xx,0.01)
```

回车后显示:

```
muhat=
    4.0787
sigmahat=
    1.9519
muci=
    3.3389
    4.8185
sigmaci=
    1.5448
    2.6175
```

结果说明：现在给出μ(这里仍然取μ=4)和σ(这里仍然取σ=2)的极大似然估计值仍分别为 muhat＝4.0787 和 sigmahat=1.9519，但是μ和σ是置信度为 100(1-0.01)％=99%的置信区间就分别为[3.3389, 4.8185]和[1.5448, 2.6175]．可见可靠度增加了(因为置信区间长度加大了).

若再执行命令:

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

得到结果说明：现在给出μ(这里已知μ=4)和σ(这里已知σ=2)的极大似然估计值仍分别为 muhat＝4.0787 和 sigmahat=1.9519，但是μ和σ的置信度为 100(1-0.1)%=90%的置信区间就分别为[3.6159, 4.5415]和[1.6776, 2.3457]，可见精确度提高了(因为置信区间长度减小了).

从上面执行的结果，我们可以看到：要求的置信度越高，给出的置信区间就越宽(区间长度加大)，但是都包含了参数的真实值，即估计的可靠度越大，估计的精确度在降低.

4. 利用通用 mle 函数来进行参数估计

- phat=mle('dist', X); % 返回用 dist 指定分布的极大似然估计值.
- [phat, pci]=mle('dist', X); % 同时进行区间估计，默认置信度为 95%.
- [phat, pci]=mle('dist', X, alpha); %同时进行区间估计，置信度由 alpha 确定.
- [phat, pci]=mle('dist', X, alpha, pl); %仅用于二项分布, pl 为试验次数.

说明 dist 为分布函数名,如 beta(β分布)、bino(二项分布)等，见表 7-1. X 为数据样本,alpha 为显著水平α, 100(1-α)%为置信度.

表 7-1 常见分布函数名称表

dist 的取值	函数说明
'beta' 或 'Beta'	Beta 分布
'bino' 或 'Binomial'	二项分布
'chi2' 或 'Chisquare'	χ^2分布
'exp' 或 'Exponential'	指数分布
'f' 或 'F'	F 分布
'gam' 或 'Gamma'	Γ分布
'geo' 或 'Geometric'	几何分布
'hyge' 或 'Hypergeometric'	超几何分布
'logn' 或 'Lognormal'	对数正态分布
'nbin' 或 'Negative Binomial'	负二项分布
'ncf' 或 'Noncentral F'	非中心 F 分布
'nct' 或 'Noncentral t'	非中心 t 分布
'ncx2' 或 'Noncentral Chi-square'	非中心χ 2 分布

'norm' 或 'Normal'	正态分布
'poiss' 或 'Poisson'	泊松分布
'rayl' 或 'Rayleigh'	瑞利分布
't' 或 'T'	t 分布
'unif' 或 'Uniform'	均匀分布
'unid' 或 'Discrete Uniform'	离散均匀分布
'weib' 或 'Weibull'	Weibull 分布

例 7-6 产生 200 个概率 p 为 0.75 的试验次数 N=50 的二项分布的随机数，求出置信度为 0.95 的参数 p 的置信区间.

解 在命令窗口中输入:

```
X=binornd(50,0.75,200,1);  % 产生题目要求的二项分布的 200 个随机数.
[p,pci]=mle('bino',X,0.05,50);  % 求概率的估计值和置信区间，置信度为95%.
```

回车后显示:

```
p =
    0.7471
pci =
    0.7385
    0.7556
```

关于其它分布的参数估计，读者可以类似地写出，也可以在 MATLAB 工作空间中用 help 命令获得具体用法和命令解释.

三、 实践结论与总结

参数估计是数理统计中一个基本的重要问题，就是用样本对总体的未知参数做出估计，分为点估计和区间估计；区间估计涉及估计的置信水平、可靠度和精确度. 参数估计方法在各个方面具有广泛的应用. 本试验用 MATLAB 软件工具箱中提供的参数估计函数 normfit, expfit, mle 等诸多函数估计出了总体分布类型已知的情况下未知参数的极大似然估计值和在一定的置信水平下 的置信区间.