应用统计

第4讲　参数的矩估计和极大似然估计

参数点估计

二战中的德国坦克数

整个二战期间：

马克1型，1493辆；马克2型，1856辆；马克3型，5774辆；马克4型，8800辆；以上也就是我们常说的1号到4号坦克

马克5型，也就是豹式坦克，6000辆；马克6型，也就是虎1式坦克，1347辆；马克6型改，也就是虎王坦克，492辆

二战中的德国坦克数

德国人生产的坦克按照出厂的先后顺序编号1, 2, 3⋯

盟军估计德国的坦克数量，用缴获坦克的最大编号乘以1加上缴获数分之1 $max\{x_1, x_2, \cdots, x_n\} \times \left(1 + \frac{1}{n}\right)$

1940.6—1942.9 平均每月生产246辆，估计值为245

而通过间谍、解码等的得到的信息的大约1400辆

一组学生身高数据处理

50名17岁城市男生身高（单位：cm）：

170.1 179.0 171.5 173.1 174.1 177.2 170.3 176.2 163.7 175.4
163.3 179.0 176.5 178.4 165.1 179.4 176.3 179.0 173.9 173.7
173.2 172.3 169.3 172.8 176.4 163.7 177.0 165.9 166.6 167.4
174.0 174.3 184.5 171.9 181.4 164.6 176.4 172.4 180.3 160.5
166.2 173.5 171.7 167.9 168.7 175.6 179.6 171.6 168.1 172.2

假设学生的身高服从正态分布 $N(\mu,\sigma^2)$

即样本来自正态总体 $N(\mu,\sigma^2)$，估计参数 μ,σ^2

由此甚至可以推测男生身高高于2m的比例，即便样本里面没有2m的

参数点估计

$X_1, X_2, \cdots, X_n$是来自某总体的样本，利用这些样本估计总体分布的参数θ。构造适当的统计量

$\hat{\theta}=\theta\left(X_1, X_2, \cdots, X_n\right)$ 称为参数θ 的点估计。

$\hat{\theta}$ 的构造不一定是唯一的，但需要满足一定的合理性。

我们首先学习两种最常用的点估计方法：

矩估计和极大似然估计，

以及判断估计统计量合理性的基本方法。

估计盒子中黑球的比例 $p \approx 0.3$

- 设一个盒子里装有一定量的白球和黑球，试估计其中黑球比例。 a个白球，b个黑球：$p = \frac{b}{a+b}$

- 假定进行10次有放回的抽取，抽到3个黑球。

$$X \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$$

$$E(X) = p \approx \bar{x} = \frac{x_1 + \cdots + x_{10}}{10} = 0.3$$

矩估计法

$$E(X)=p$$

$$E(X)\approx \bar{x}$$

$$p\approx \bar{x}=\frac{x_1+\cdots+x_{10}}{10}=0.3$$

$$Var(X)=p(1-p)$$

$$Var(X)\approx S^2=\sum_{k=1}^{n}\frac{(x_k-\bar{x})^2}{n-1}$$

$$p(1-p)\approx S^2=\frac{3\times 0.7^2+7\times 0.3^2}{9}$$

- 替换原理

- 矩的理论表达式为参数的函数

  用样本矩替换理论上的矩，解方程得到参数的近似。

- 这里的“矩”可以是原点矩、中心矩，以及样本方差等。

- 为了计算简单，尽可能用低阶矩

$$E(X^k)=f_k(\theta_1,\theta_2,\cdots,\theta_m)\Rightarrow f_k(\hat{\theta}_1,\hat{\theta}_2,\cdots,\hat{\theta}_m)=\frac{x_1^{\ k}+x_2^{\ k}+\cdots+x_n^{\ k}}{n}\quad k=1,\cdots,m$$

例题：对于均匀分布总体$U(a,b)$,估计参数a,b

$$E(X)=\frac{a+b}{2},\quad Var(X)=\frac{(b-a)^2}{12}\ \Rightarrow\ \begin{cases}\dfrac{\hat{a}+\hat{b}}{2}=\bar{x}\\ \dfrac{(\hat{b}-\hat{a})^2}{12}=s^2\end{cases}$$

得到参数$\theta=(a,b)$的点估计$\hat{\theta}=(\hat{a},\hat{b})=\left(\bar{x}-\sqrt{3}s,\bar{x}+\sqrt{3}s\right)$。

或利用 $E(X)=\dfrac{a+b}{2},\quad E(X^2)=\dfrac{a^2+ab+b^2}{3}\ \Rightarrow\ \begin{cases}\dfrac{\hat{a}+\hat{b}}{2}=\bar{x}\\ \dfrac{\hat{a}^2+\hat{a}\hat{b}+\hat{b}^2}{3}=\dfrac{x_1^2+\cdots x_n^2}{n}\end{cases}$

解得参数$\theta=(a,b)$的另一种点估计$\hat{\theta}=(\hat{a},\hat{b})$。

$X_1, X_2, \cdots, X_n$ 是来自均匀总体 $X \sim \mathrm{U}(-a,a)$ 的样本，用矩估计法估计参数 a

$$E(X)=0, \qquad Var(X)=\frac{a^2}{3}$$

$$\frac{a^2}{3}=S^2 \quad \Rightarrow \quad \hat{a}=\sqrt{3}S$$

$$E(X^2)=\frac{a^2}{3}$$

$$\frac{a^2}{3}=\frac{X_1{}^2+X_2{}^2+\cdots+X_n{}^2}{n} \quad \Rightarrow \quad \bar{a}=\sqrt{3}\sqrt{\frac{X_1{}^2+X_2{}^2+\cdots+X_n{}^2}{n}}$$

$X_1, X_2, \cdots, X_m$ 是来自二项分布总体 $X \sim b(n,p)$ 的样本，用矩估计法估计参数n, p

$$E(X) = np, \qquad Var(X) = np(1-p)$$

$$\begin{cases} \bar{X} = np \\ S^2 = np(1-p) \end{cases} \Rightarrow \hat{p} = 1 - \frac{S^2}{\bar{X}}, \quad \hat{n} = \frac{\bar{X}^2}{\bar{X} - S^2}$$

其中 $\begin{cases} \bar{X} = \dfrac{X_1 + X_2 + \cdots + X_m}{m} \\ S^2 = \dfrac{\sum_{k=1}^{m}\left(X_k - \bar{X}\right)^2}{m-1} \end{cases}$

$$E(X) = np, \qquad E(X^2) = np(1-p) + n^2p^2$$

$$\begin{cases} \bar{X} = np \\ \dfrac{X_1^2 + X_2^2 + \cdots + X_n^2}{n} = np(1-p) + n^2p^2 \end{cases} \Rightarrow \ ?$$

$X_1, X_2, \cdots, X_m$ 是来自二项分布总体 $X \sim b(n,p)$ 的样本，用矩估计法估计参数n, p

$E(X) = np,\qquad Var(X) = np(1-p)$

$$\begin{cases} \bar{X} = np \\ S^2 = np(1-p) \end{cases} \Rightarrow \hat{p} = 1 - \frac{S^2}{\bar{X}},\quad \hat{n} = \frac{\bar{X}^2}{\bar{X} - S^2}$$

其中 $\begin{cases} \bar{X} = \dfrac{X_1 + X_2 + \cdots + X_m}{m} \\ S^2 = \dfrac{\sum_{k=1}^{m}(X_k - \bar{X})^2}{m-1} \end{cases}$

设想实验

有一盒规格完全相同的不均匀的硬币，总数 n 和抛到正面概率 p 都未知，对 n 和 p 进行估计，

将这盒硬币一个一个拿出抛掷，抛出正面时加1，这盒硬币都抛完，并没有数一共有多少枚

但知道总共掷出 x_1 次正面。然后将所有硬币装回盒子，再抛一遍，记录到总共掷出 x_2 次正面。

将上述过程重复 m 次，得到 m 个观测值：$x_1, x_2, \cdots, x_m$。

通过这 m 个观测值，得到参数 n 和 p 的一组估计值。

$\hat{p} = 1 - \frac{S^2}{\bar{X}},\quad \hat{n} = \frac{\bar{X}^2}{\bar{X} - S^2}$ 给出了 n 和 p 的估计方法，$\hat{p}$ 和 $\hat{n}$ 是统计估计量，是随机变量

估计的基本想法（矩估计与极大似然估计）

- 设一个盒子里装有一定量的白球和黑球，试估计其中黑球比例。
- 假定进行10次有放回的抽取，抽到3个黑球。
- 用抽出的球中黑球的比例近似盒子中黑球的比例。（以样本矩近似理论矩）
- 盒子中黑球的比例为多少时，实际抽取的结果最有可能发生。（什么样的黑球比例以最大概率解释样本数据）

$$P\left(X_1+X_2+\cdots+X_{10}=3\right)=\binom{10}{3}p^3(1-p)^7$$

$p=0.3$: $\binom{10}{3}p^3(1-p)^7=0.267$

$p=0.2$: $\binom{10}{3}p^3(1-p)^7=0.201$

$p=0.6$: $\binom{10}{3}p^3(1-p)^7=0.042$

极大似然估计

- 极大似然估计（MLE）
  maximum likelihood estimation

- 以最大概率解释样本数据。
- 相对于其他参数，所考虑的样本数据更像（more likely）是来自于这组参数。

$$X\sim\begin{pmatrix}0 & 1\\ 1-p & p\end{pmatrix}$$

最大化: $P\left(X_1+X_2+\cdots+X_{10}=3\right)=\binom{10}{3}p^3(1-p)^7$

最大化: $p^3(1-p)^7$

极大似然估计方法

设总体的概率函数为 $p(x;\theta)$，$\theta\in\Theta$，其中 θ 是一组未知参数，Θ 称为参数空间，即参数 θ 可能取值的集合。$x_1,x_2,\cdots,x_n$ 是来自该总体的样本，则样本的联合概率函数是关于 θ 的函数，用 $L(\theta;x_1,x_2,\cdots,x_n)$ 表示，简记为 $L(\theta)$。$L(\theta)=L(\theta;x_1,x_2,\cdots,x_n)=\prod_{k=1}^{n}p(x_k;\theta)$，称为样本的似然函数，如某统计量 $\hat{\theta}=\hat{\theta}(x_1,x_2,\cdots,x_n)$ 满足 $L(\hat{\theta})=\max_{\theta\in\Theta}L(\theta)$，则称 $\hat{\theta}$ 为 θ 的极大似然估计(maximum likelihood estimation, MLE)。

对数似然函数 $\ln(L(\theta))$，由于 $\ln x$ 是 x 的单调函数，使得 $\ln(L(\theta))$ 与 $L(\theta)$ 达到最大的 θ 相同，常利用对数似然函数求解极大似然估计。

极大似然估计方法

$p(x_k;\theta)$为概率函数，对连续型是密度函数，
对离散型是分布律，样本的似然函数为

$$L(\theta)=L(\theta;x_1,x_2,\cdots,x_n)=\prod_{k=1}^{n}p(x_k;\theta)$$

$x_1,x_2,\cdots,x_n$是样本值

$$\max L(\theta;x_1,x_2,\cdots,x_n)$$

对数似然函数

$p(x_k;\theta)$对θ可导时，对数似然函数$\ln L(\theta;x_1,x_2,\cdots,x_n)$

$$\max \ln L(\theta;x_1,x_2,\cdots,x_n)$$

极大似然估计例题

例 1. 总体服从指数分布 $Exp(\lambda)$, 估计参数 λ。 $p(x)=\lambda\cdot e^{-\lambda x}\ (x>0)$, $x_1,x_2,\cdots,x_n$ 为来自该总体的样本，则似然函数 $L(\lambda)=\lambda^n\cdot e^{-\lambda(x_1+\cdots+x_n)}$, 利用对数似然函数求关于 λ 的极值，得到 $\hat{\lambda}=\dfrac{1}{\bar{x}}$。

例 2. 总体服从泊松分布 $P(\lambda)$, 估计参数 λ。 $x_1,x_2,\cdots,x_n$ 为来自该总体的样本，则似然函数 $L(\lambda)=\prod_{k=1}^{n}P(X=x_k)=\prod_{k=1}^{n}\dfrac{\lambda^{x_k}}{x_k!}\cdot e^{-\lambda}$, 利用对数似然函数求关于 λ 的极值，得到 $\hat{\lambda}=\bar{x}$。

例 3 设元件的寿命服从指数分布 $p(x)=\lambda\cdot e^{-\lambda x}\ (x>0)$。指定某一个时刻 $T>0$，抽样试验进行到元件失效或时刻 T，

$$y_k=\begin{cases}\text{元件}k\text{的寿命}, & \text{当此寿命小于}T\text{时}\\ T, & \text{当时刻}T\text{元件}k\text{还没有失效}\end{cases},$$

求参数λ的极大似然估计。

解： 若 n 个样本值中有 r 个是元件的实际寿命，则

似然函数 $L(x;\lambda)=\lambda^r\cdot e^{-\lambda\cdot(x_1+\cdots+x_r)}\cdot e^{-\lambda\cdot(n-r)\cdot T}$

$$\Rightarrow\hat{\lambda}=r\Big/\left[x_1+\cdots+x_r+(n-r)T\right]$$

思考：除了极大似然的思想，也可以对一问题尝试矩估计的想法，试基于以下几种思路对元件寿命参数进行估计：

（1）利用取值小于T的样本数占总样本数的比例；

（2）利用所有取值小于T的样本的均值；

（3）利用样本的平均值。

例 4 $X_1, X_2, \cdots, X_n$ 为来自均匀分布总体 $U(a,b)$ 的样本，试利用极大似然估计给出参数 a,b 的估计量。

解：设 $x_1, x_2, \cdots, x_n$ 为样本观测值，

总体分布的密度函数为 $f(x;a,b)=\dfrac{1}{b-a}, a \le x \le b$

似然函数为 $L(a,b)=\prod_{k=1}^{n} f(x_k;a,b)=(\dfrac{1}{b-a})^n$， $a \le x_k \le b, k=1,2,\cdots,n$，

显然 $L(a,b)$ 关于 a 是单调递增函数，关于 b 是单调递减函数。使得 $L(a,b)$ 最大，必须使 $b-a$ 达到最小，即使 b 尽可能小，a 尽可能大。

考虑到 $a \le x_1, x_2, \cdots, x_n \le b$， $a \le \min\{x_1, x_2, \cdots, x_n\}$，$b \ge \max\{x_1, x_2, \cdots, x_n\}$，

参数 a,b 的极大似然估计 $\hat{a}=\min(X_1, X_2 \cdots, X_n)$， $\hat{b}=\max(X_1, X_2 \cdots, X_n)$。

例 5　估计 Cauchy 分布 $f(x;\theta)=\dfrac{1}{\pi\left[1+(x-\theta)^2\right]}$ （$x\in R$）的参数θ。

$$\int_{-\infty}^{+\infty}|x|\cdot f(x)dx=2\int_0^{+\infty}x\cdot f(x)dx=2\int_0^{+\infty}\frac{x}{\pi\left[1+(x-\theta)^2\right]}dx$$

$$=2\int_0^{+\infty}\frac{x-\theta}{\pi\left[1+(x-\theta)^2\right]}dx+2\int_0^{+\infty}\frac{\theta}{\pi\left[1+(x-\theta)^2\right]}dx$$

$$=\frac{1}{\pi}\int_0^{+\infty}\frac{d\left[1+(x-\theta)^2\right]}{\left[1+(x-\theta)^2\right]}+\theta=\frac{1}{\pi}\ln\left[1+(x-\theta)^2\right]\Big|_0^{+\infty}+\theta$$

Cauchy 分布随机变量的期望不存在，因此不能用矩估计法对参数θ进行估计。

例 5 估计 Cauchy 分布 $f(x;\theta)=\dfrac{1}{\pi\left[1+(x-\theta)^2\right]}$ （$x\in R$）的参数 θ。

解：设 $x_1,x_2,\cdots,x_n$ 为来自该总体的样本观测值，

似然函数 $L(\theta)=\prod_{k=1}^{n} f(x_k;\theta)=\prod_{k=1}^{n}\dfrac{1}{\pi\left[1+(x_k-\theta)^2\right]}$

对数似然函数 $\ln L(\theta)=\sum_{k=1}^{n}-\left(\ln\pi+\ln\left(1+(x_k-\theta)^2\right)\right)$

将上式对 θ 求导，并令其等于 0， $\dfrac{d\ln L(\theta)}{d\theta}=-\sum_{k=1}^{n}\dfrac{x_k-\theta}{1+(x_k-\theta)^2}=0$

此方程无法的到解析解，需要用一定的计算方法近似求解。

例6. $X_1, X_2, \cdots, X_n$ 是来自二项分布总体 $X \sim b(k, p)$ 的样本，p已知，用极大似然估计法估计参数k

例如我们抛一枚均匀的硬币，并观察到正面的次数为x_i，但我们不知道抛了多少次。此时，似然函数为

$$L(k; x, p) = \prod_{i=1}^{n} \binom{k}{x_i} p^{x_i} (1-p)^{k-x_i}$$

因为含有阶乘，同时 k 必须为整数，所以无法用微分法求极值。所求MLE应该是一个整数，满足 $k \geq \max_i x_i$ ，$\frac{L(k;x,p)}{L(k-1;x,p)} \geq 1$，$\frac{L(k+1;x,p)}{L(k;x,p)} < 1$，

$$\left(k(1-p)\right)^n \geq \prod_{i=1}^{n}(k - x_i) \text{ 与 } \left((k+1)(1-p)\right)^n < \prod_{i=1}^{n}(k+1-x_i)$$

$$(1-p)^n \geq \prod_{i=1}^{n}\left(1 - \frac{1}{k}x_i\right) \text{ 与 } (1-p)^n < \prod_{i=1}^{n}\left(1 - \frac{1}{k+1}x_i\right)$$

$$(1-p)^n = \prod_{i=1}^{n}(1 - zx_i), \quad \text{其中} 0 \leq z \leq \frac{1}{\max\limits_i x_i}, \quad \text{解出}\hat{z}, \quad k\text{的MLE}\left[\frac{1}{\hat{z}}\right]$$

例6的注记

Feldman, D., and Fox, M. (1968). Estimation of the Parameter n in the Binomail Distribution. *J. Amer. Statist. Assoc.* **69** 990-996

Olkin, Petkau and Zidek (1981). A Comparison of n Estimators for the Binomial Distribution. *J. Amer. Statist. Assoc.* **76** 637-642

B(k, p)，k, p均未知

(16, 18, 22, 25, 27)　$\hat{k} = 99$

(16, 18, 22, 25, 28)　$\hat{k} = 190$

问题的病态性

--- 数值计算的基本概念

- 考虑如下的问题

f(x)=(x-1)(x-2)…….(x-20)

显然方程 f(x)=0 的解是

1 2 3 4 ……… 19 20

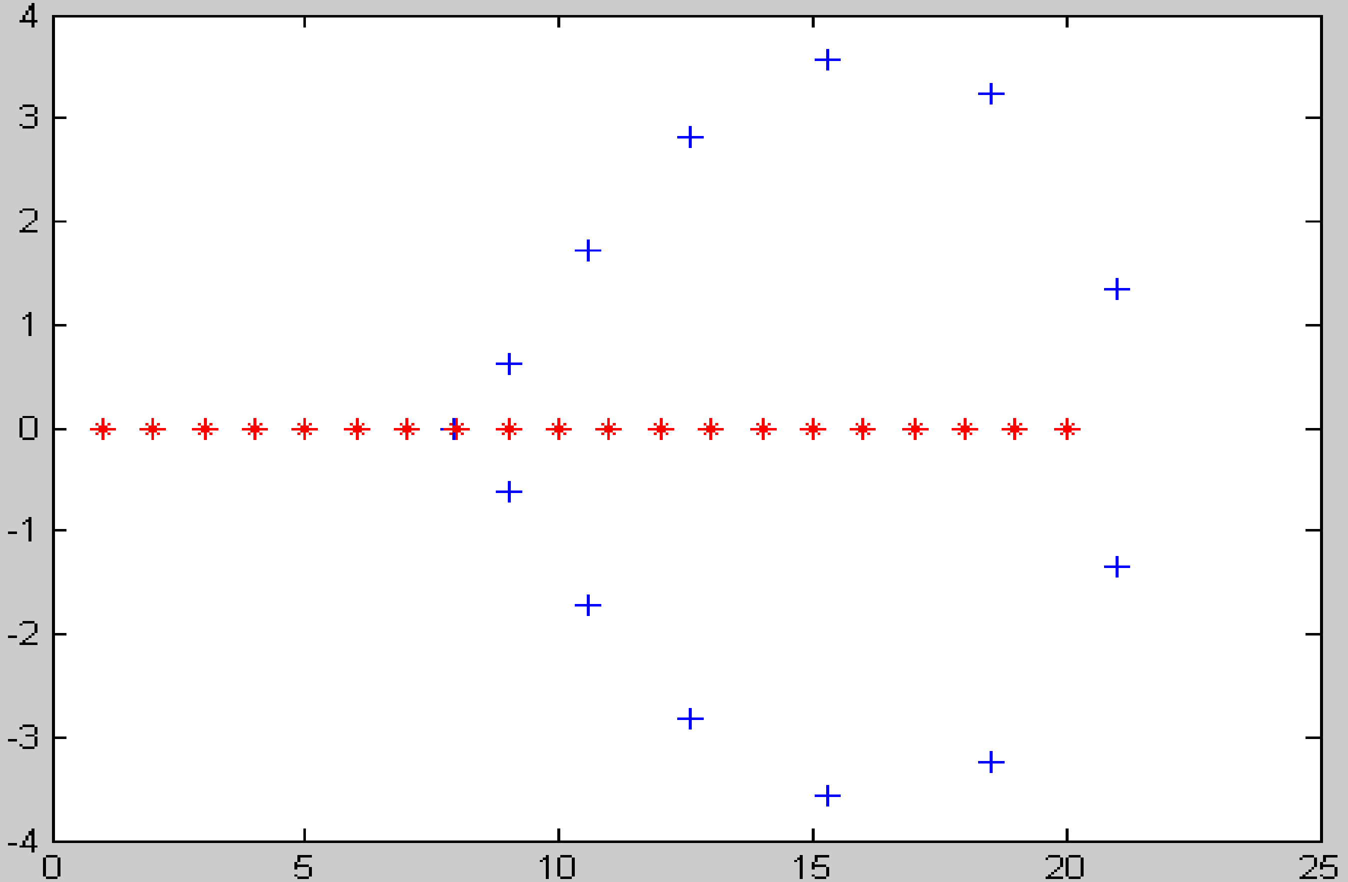
请问：如下方程的解是什么？

$$f_{\varepsilon}(x) = f(x) + \varepsilon x^{18} = 0$$

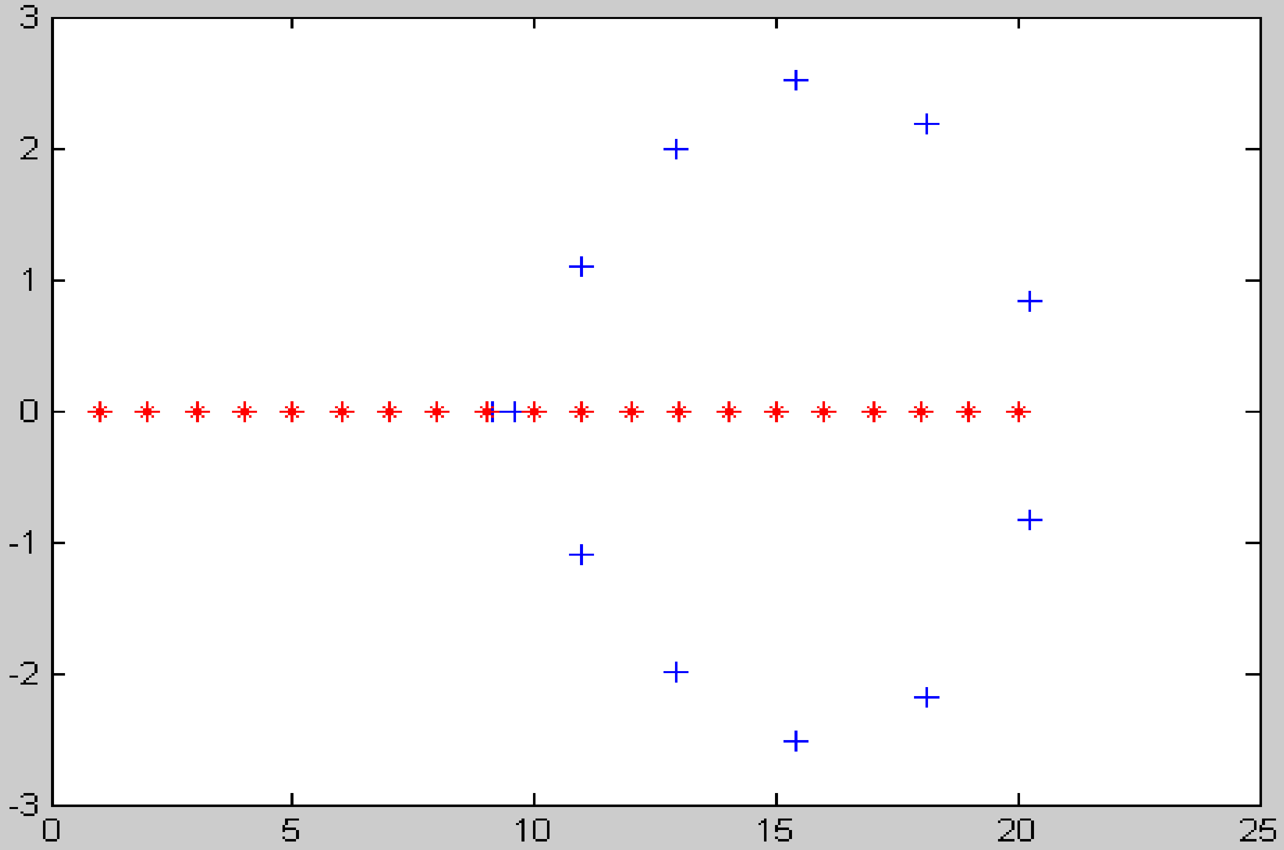
Matlab program

```
p=poly(1:20);
ep=zeros(1,21);
ep(3)=1.0e-7;
re=roots(p+ep)
plot(re,‘g+');
hold on
plot(1:20,0,'r*');
hold off
```

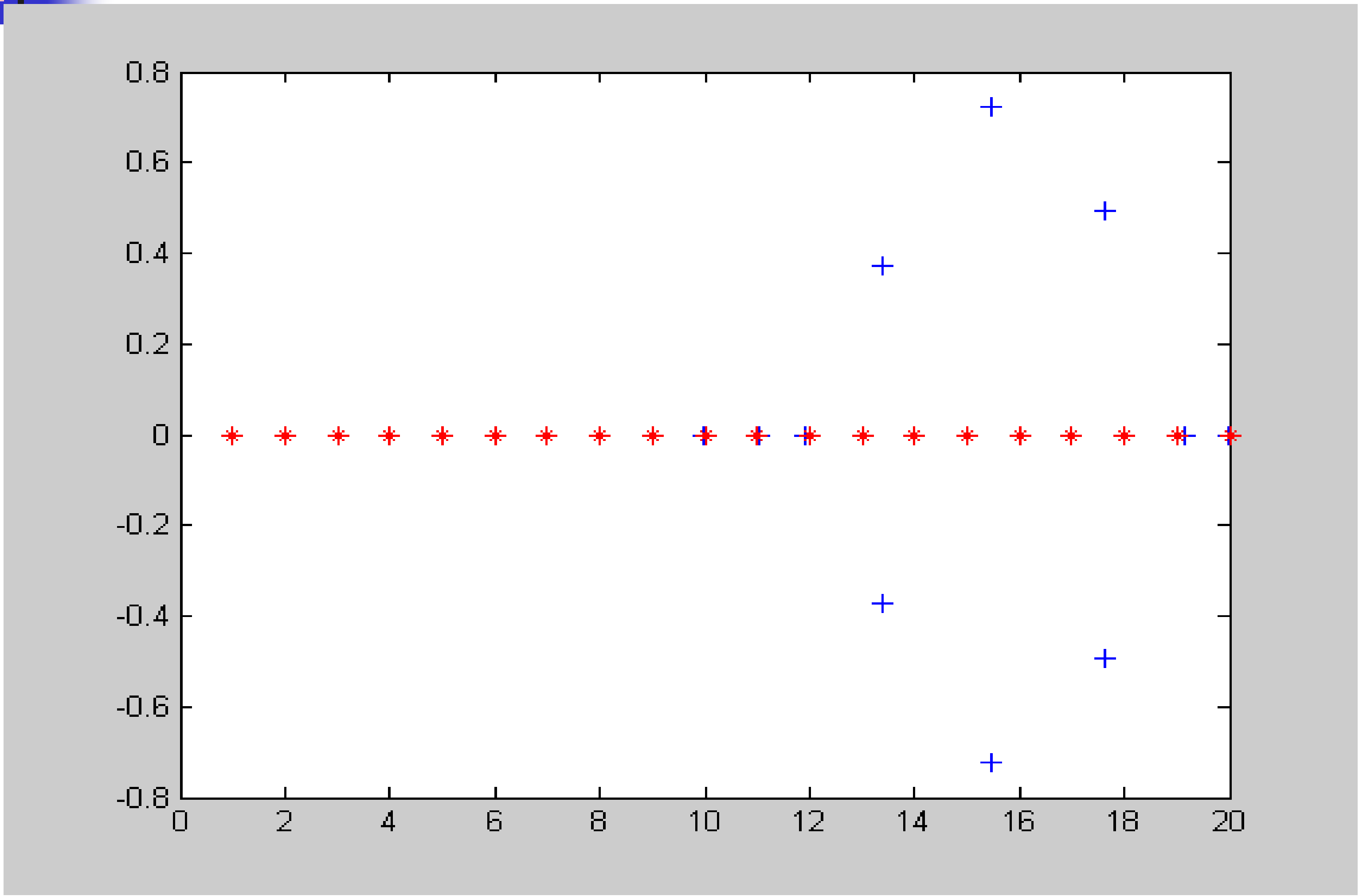
$\varepsilon = 10^{-5}$



$\varepsilon = 10^{-6}$



$\varepsilon = 10^{-8}$



算法的稳定性

--- 数值计算的基本概念

- 考虑如下的序列

$$E_n = \int_0^1 x^n e^x \, dx, \qquad n = 1, 2, \cdots$$

- 可以证明

$$E_n = e - nE_{n-1}$$

$$0 < E_n < e/(n+1)$$

两个算法

----有什么差别，哪个可以用??

- Algorithm 1

$$E_1 = 1$$

$$E_n = e - nE_{n-1}$$

$$n = 2, 3, \cdots$$

- Algorithm 2

$$E_N = 0,$$

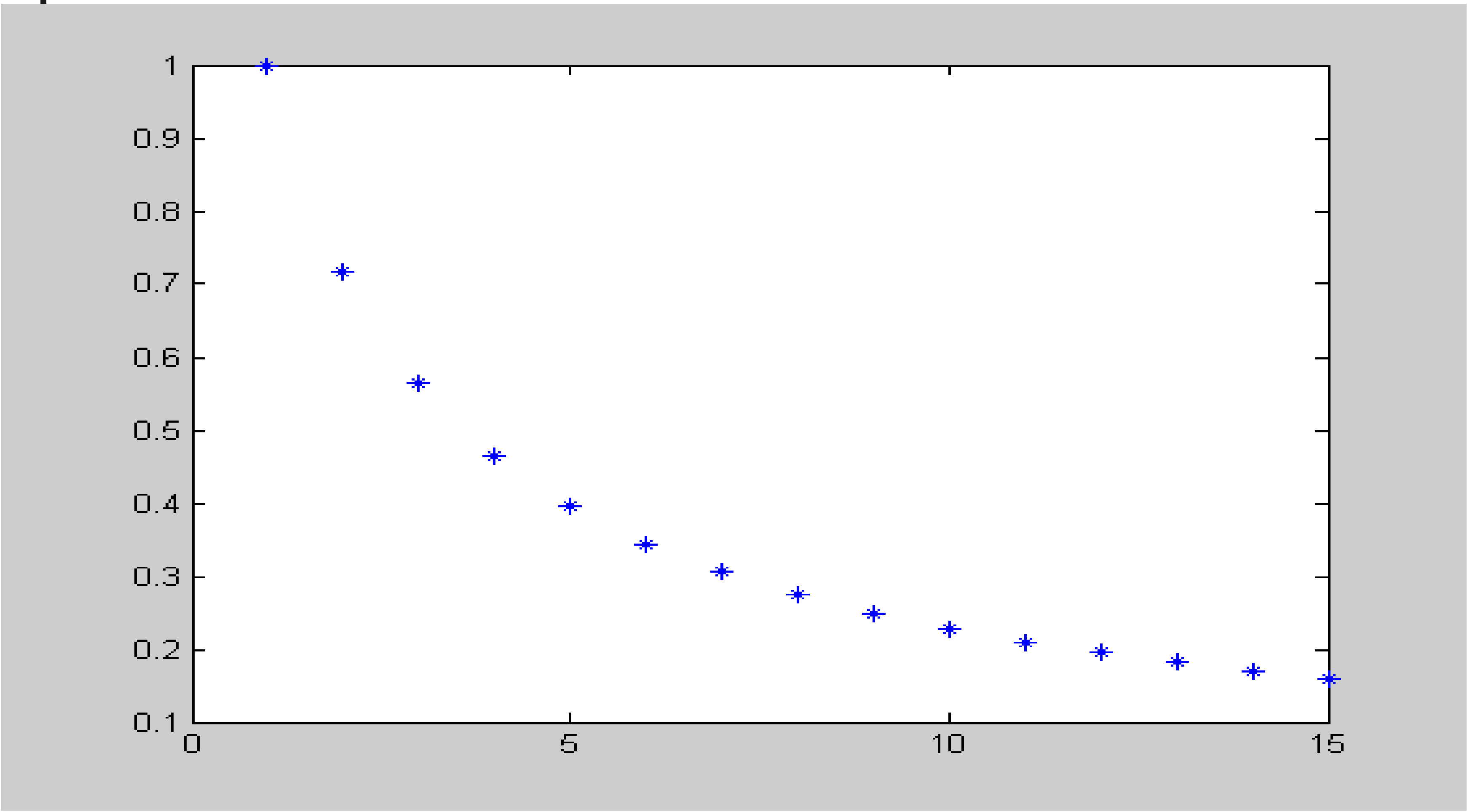
$$E_{n-1} = (e - E_n) / n$$

$$n = N, N-1, \cdots, 2, 1$$

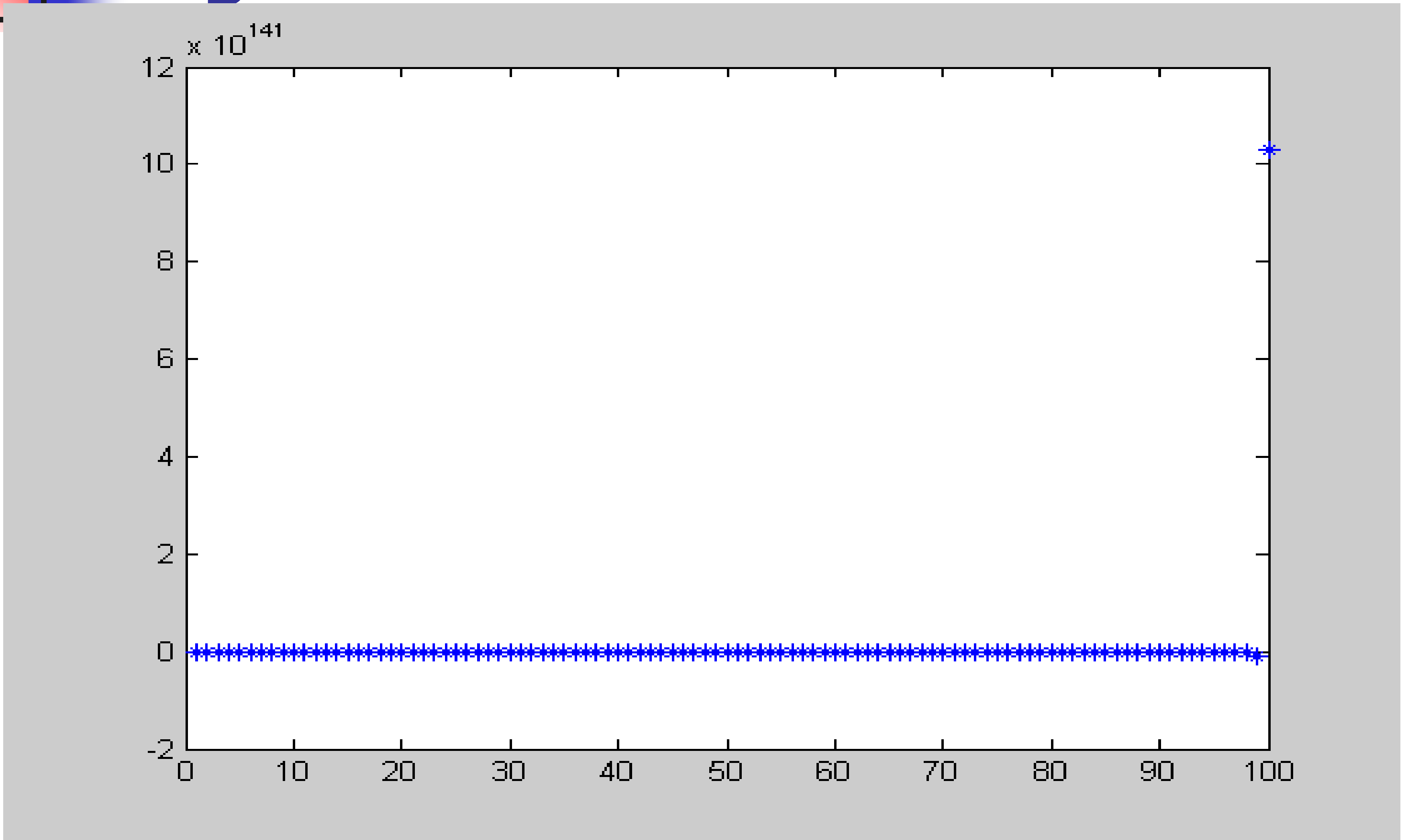
Program of algorithm 1

```
clear
ep(1)=1
for n=2:100
    ep(n)=exp(1.0)-n*ep(n-1)
end
plot(ep,'b*');
```

Algorithm 1 with n=15



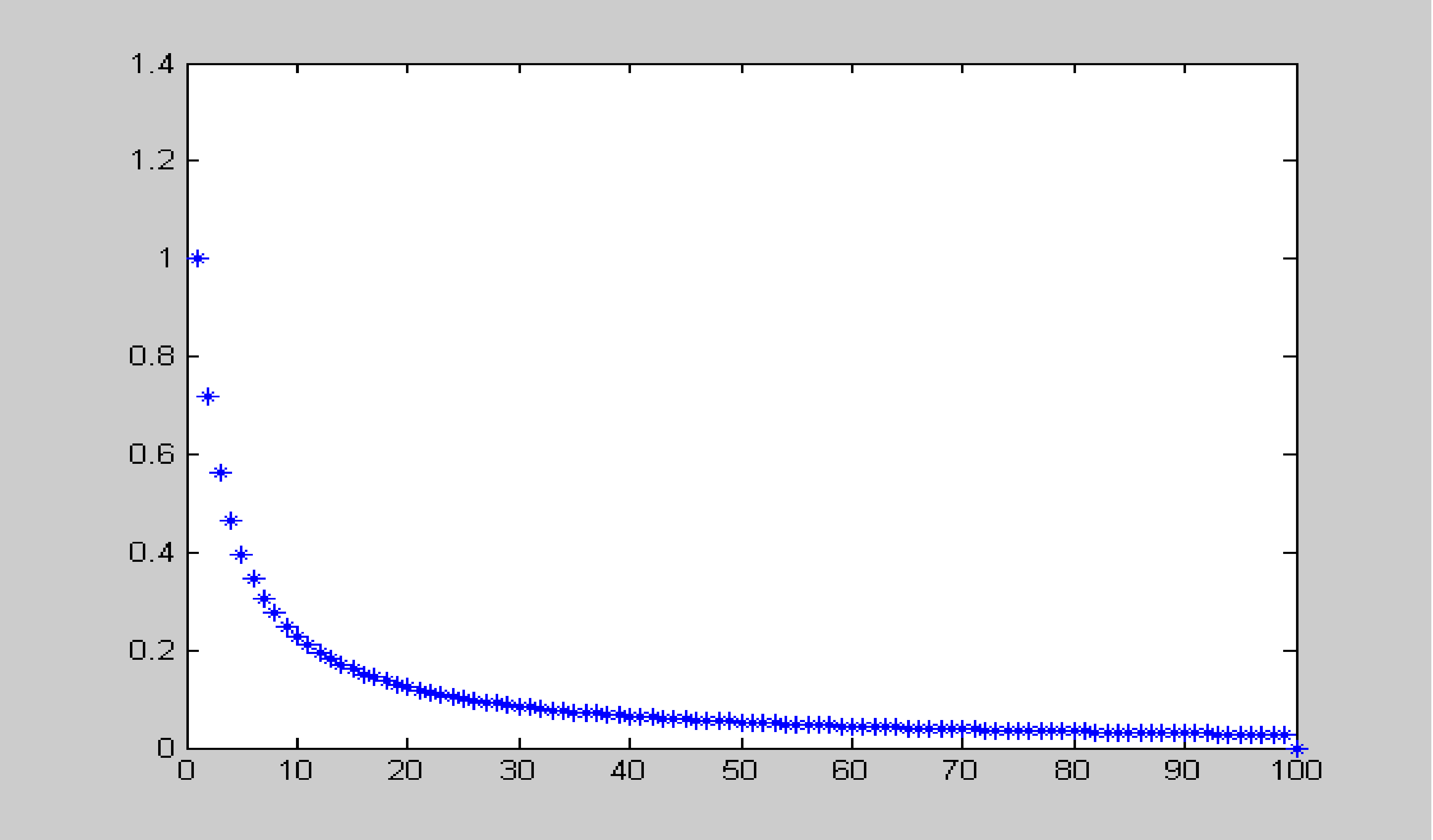
Algorithm 1 with n=100



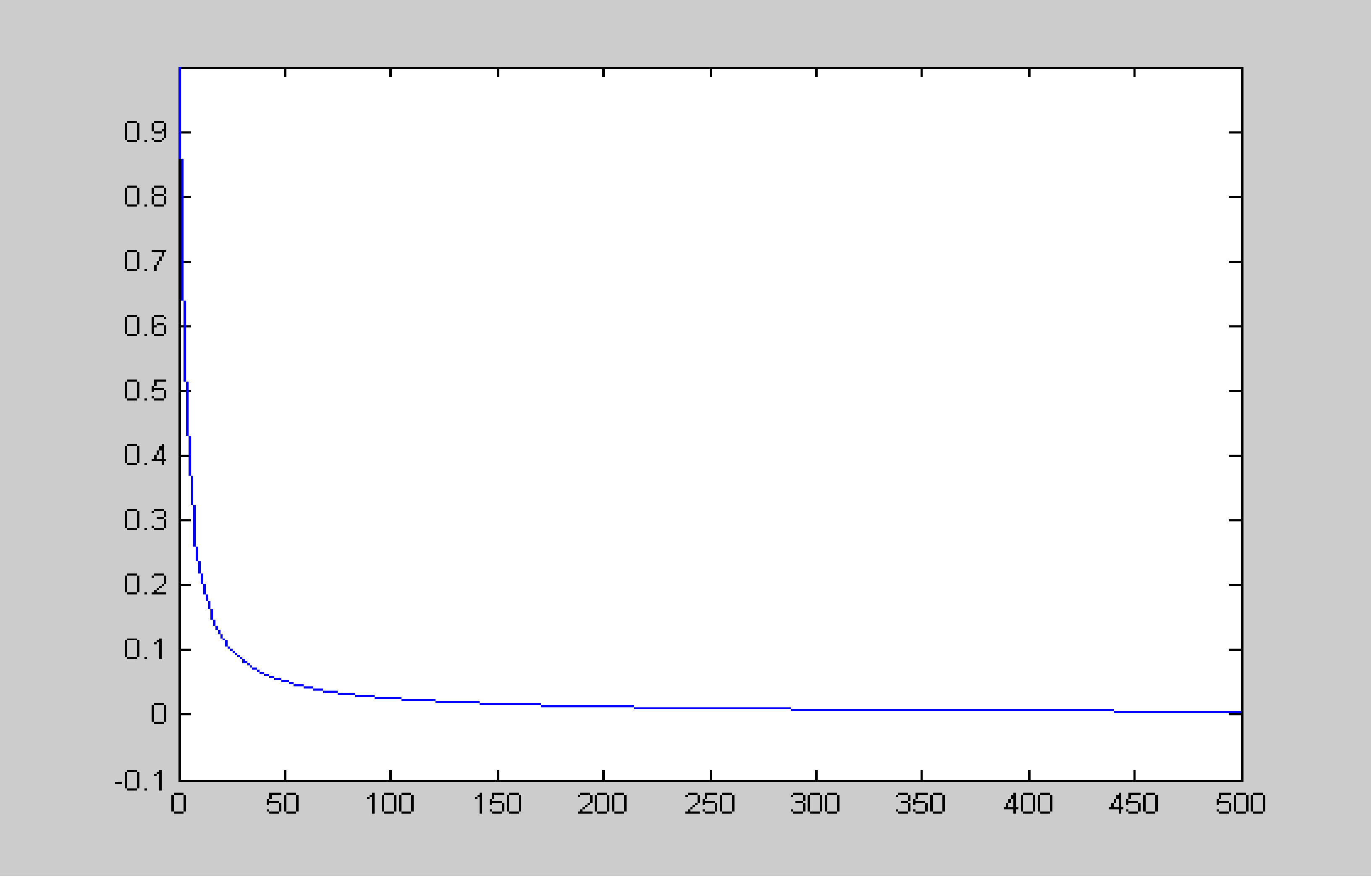
Program of algorithm 2

```
clear
ep(100)=0
for n=100:-1:2
    ep(n-1)=(exp(1.0)-ep(n))/n;
end
plot(ep,'b*');
```

Algorithm 2 with n=100



Algorithm 2 with n=500



正态分布导出 ： 误差分布函数
高斯的极大似然思想

正态分布的导出

设对某个量（如长度、浓度等）X 进行 n 次独立测量，得到测量值 $x_i(i=1,2,\cdots,n)$，

因而误差 $\varepsilon_i = x_i - x(i=1,2,\cdots,n)$，其中 ε_i， x_i 为随机变量， x 为未知常数。设 ε_i

的密度函数为 $f(x_i - x)$。 【注： x_i 同时表示随机变量及其取值】

则 $\varepsilon_1, \varepsilon_2, \cdots \varepsilon_n$ 的联合密度函数（即同时发生的概率密度）为 $L = \prod_{i=1}^{n} f(x_i - x)$，

x 的值应该使 L 极大（其含义为最可能使测量值 $x_i(i=1,2,\cdots,n)$ 出现的 x ）。

故 x 应满足 $\dfrac{dL}{dx}=0$， $\dfrac{-d\ln L}{dx} = \sum_{i=1}^{n} \dfrac{f'(x_i - x)}{f(x_i - x)} = 0$。 此方程没有办法求解。

则 $\varepsilon_1,\varepsilon_2,\cdots\varepsilon_n$ 的联合密度函数（即同时发生的概率密度）为 $L=\prod_{i=1}^{n} f(x_i-x)$，

x 的值应该使 L 极大（其含义为最可能使测量值 $x_i\,(i=1,2,\cdots,n)$ 出现的 x）。

$$\frac{-d\ln L}{dx}=\sum_{i=1}^{n}\frac{f'(x_i-x)}{f(x_i-x)}=0$$ 。 此方程没有办法求解。

f, x 均未知

假设 $x=\frac{1}{n}\sum_{i=1}^{n}x_i=a$，即假设平均值最接近真实值。

记 $g(y)=\frac{f'(y)}{f(y)}$，令 $y_i=x_i-x$，可设 $G\equiv\sum_{i=1}^{n}g(x_i-x)=\sum_{i=1}^{n}g(y_i)=0$

因为 $\sum_{i=1}^{n}y_i=0$，故此 n 个量只有 $n-1$ 个是相互独立的，令 $y_n=-\sum_{i=1}^{n-1}y_i$，

对于 $i=1,2,\cdots,n-1$，$\frac{\partial G}{\partial y_i}=\frac{dg}{dy_i}+\frac{dg}{dy_n}\frac{\partial y_n}{\partial y_i}=0\Rightarrow\frac{dg}{dy_i}=\frac{dg}{dy_n}=c$(常数)

对于$i=1,2,\cdots,n-1$，$\dfrac{\partial G}{\partial y_i}=\dfrac{dg}{dy_i}+\dfrac{dg}{dy_n}\dfrac{\partial y_n}{\partial y_i}=0\Rightarrow\dfrac{dg}{dy_i}=\dfrac{dg}{dy_n}=c$(常数)

$g(y)=cy+b$，由$\sum_{i=1}^{n}g(y_i)=c\sum_{i=1}^{n}y_i+nb=0\Rightarrow b=0$

$$\frac{f'(y)}{f(y)}=cy\Rightarrow f(y)=k\cdot e^{\frac{1}{2}cy^2}$$。

因为$\int_{-\infty}^{+\infty}f(t)dt=1$，所以$c<0$。

令$c=-\dfrac{1}{\sigma^2}$，则$f(y)=\dfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{y^2}{2\sigma^2}}\sim N(0,\sigma^2)$。

作业

- 习题二
- 2.(3)(4)，3，4(2)(7)，7