

法律声明

□ 本课件包括演示文稿、示例、代码、题库、视频和声音等内容，小象学院和主讲老师拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意及内容，我们保留一切通过法律手段追究违反者的权利。

□ 课程详情请咨询

■ 微信公众号：小象

■ 新浪微博：ChinaHadoop



数理统计与参数估计



小象学院
ChinaHadoop.cn

邹博

主要内容

□ 统计量

- 期望/方差/偏度/峰度
- 协方差和相关系数
- 独立和不相关

□ 矩估计

- 中心矩和原点矩

□ 最大似然估计

- 深刻理解

事件的独立性

□ 给定A和B是两个事件，若有 $P(AB) = P(A)P(B)$ 则称事件A和B相互独立。

□ 说明：

■ A和B独立，则 $P(A|B) = P(A)$

■ 实践中往往根据两个事件是否相互影响而判断独立性：如给定M个样本、若干次采样等情形，往往假定它们相互独立。

□ 思考：试给出A，B相互包含的信息量的定义 $I(A, B)$ ，要求：如果A、B独立，则 $I(A, B) = 0$

期望

□ 离散型 $E(X) = \sum_i x_i p_i$

□ 连续型 $E(X) = \int_{-\infty}^{\infty} x f(x) dx$

□ 即：概率加权下的“平均值”

期望的性质

□ 无条件成立

$$E(kX) = kE(X)$$

$$E(X + Y) = E(X) + E(Y)$$

□ 若X和Y相互独立

$$E(XY) = E(X)E(Y)$$

■ 反之不成立。事实上，若 $E(XY) = E(X)E(Y)$ ，只能说明X和Y不相关。

■ 关于不相关和独立的区别，稍后马上给出。

例1：计算期望

- 从1,2,3,.....,98,99,2015这100个数中任意选择若干个(可能为0个数)求异或，试求异或的期望值。

计算每一位的期望

- 针对任何一个二进制位：取奇数个1异或后会得到1，取偶数个1异或后会得到0；与取0的个数无关。
- 给定的最大数 $2015=(11111011111)_2$ ，共11位
- 针对每一位分别计算，考虑第 i 位 X_i ，假定给定的100个数中第 i 位一共有 N 个1， M 个0，某次采样取到的1的个数为 k 。则有：

$$P\{X_i = 1\} = \frac{2^m \cdot \sum_{k \in \text{odd}} C_n^k}{2^{m+n}} = \frac{\sum_{k \in \text{odd}} C_n^k}{2^n} = \frac{1}{2}$$

总期望

□ 11位二进制数中，每个位取1的期望都是0.5

$$\begin{aligned} E(X) &= E\left(\sum_{i=0}^{10} (X_i \cdot P\{X_i\})\right) \\ &= E\left(\sum_{i=0}^{10} (2^i \cdot P\{X_i = 1\} + 0 \cdot P\{X_i = 0\})\right) \\ &= E\left(\sum_{i=0}^{10} (2^i \cdot P\{X_i = 1\})\right) \\ &= \sum_{i=0}^{10} E(2^i \cdot P\{X_i = 1\}) = \sum_{i=0}^{10} 2^i \cdot E(P\{X_i = 1\}) \\ &= \sum_{i=0}^{10} 2^i \cdot \frac{1}{2} = \frac{1}{2} \sum_{i=0}^{10} 2^i = \frac{(1111111111)_2}{2} \\ &= 1023.5 \end{aligned}$$

采样模拟1021.18

```
int _tmain(int argc, _TCHAR* argv[])
{
    const int N = 100;
    int a[N];
    bool f[N];
    int i;
    for(i = 0; i < N-1; i++)
        a[i] = i+1;
    a[N-1] = 2015;

    int sampleSize = 10000000;
    double s = 0;
    for(i = 0; i < sampleSize; i++)
    {
        s += Sample(a, N, f);
    }
    cout << s << endl;
    s /= sampleSize;
    cout << s << endl;
    return 0;
}
```

```
int Sample(const int* a, int size, bool* f)
{
    memset(f, 0, sizeof(bool)*size);
    int N = rand() % (size+1); //取多少个数据
    int n = 0; //实际取了多少数据
    while(n < N)
    {
        int t = rand() % size;
        if(!f[t])
        {
            f[t] = true;
            n++;
        }
    }

    n = 0; //当前的异或值
    for(int i = 0; i < size; i++)
    {
        if(f[i])
        {
            n ^= a[i];
        }
    }
    return n;
}
```

进一步思考

- 将原题中的2015改成1024，结论应该是多少呢？
 - 从1,2,3,.....,98,99,1024这100个数中任意选择若干个(可能为0个数)求异或，试求异或的期望值。
- 答：575.5
 - 为什么？

例2：集合Hash问题

- 某Hash函数将任一字符串非均匀映射到正整数 k ，概率为 2^{-k} ，如下所示。现有字符串集合 S ，其元素经映射后，得到的最大整数为10。试估计 S 的元素个数。

$$P\{Hash(< string >) = k\} = 2^{-k}, \quad k \in \mathbb{Z}^+$$

问题分析 $P\{Hash(< string >) = k\} = 2^{-k}, k \in Z^+$

- 由于Hash映射成整数是指数级衰减的，“最大整数为10”这一条件可近似考虑成“整数10曾经出现”，继续近似成“整数10出现过一次”。
- 字符串被映射成10的概率为 $p = 2^{-10} = 1/1024$ ，从而，一次映射即两点分布：

$$\begin{cases} P(X=1) = \frac{1}{1024} \\ P(X=0) = \frac{1023}{1024} \end{cases}$$

问题分析

□ 从而n个字符串的映射，即二项分布：

$$P\{X = k\} = C_n^k p^k (1-p)^{n-k}, \text{ 其中 } p = \frac{1}{1024}$$

□ 二项分布的期望为： $E(P\{X = k\}) = np$ ，其中 $p = \frac{1}{1024}$

□ 而期望表示n次事件发生的次数，当前问题中发生了1次，从而：

$$np = 1 \Rightarrow n = \frac{1}{p} \Rightarrow n = 1024$$

方差

□ 定义 $Var(X) = E\{[X - E(X)]^2\} = E(X^2) - E^2(X)$

■ $E\{[X - E(X)]^2\} \geq 0 \Rightarrow E(X^2) \geq E^2(X)$, 当X为定值时, 取等号

□ 无条件成立 $Var(c) = 0$

$$Var(X + c) = Var(X)$$

$$Var(kX) = k^2 Var(X)$$

□ X和Y独立

$$Var(X + Y) = Var(X) + Var(Y)$$

■ 此外, 方差的平方根, 称为标准差

协方差

□ 定义 $Cov(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$

□ 性质：

$$Cov(X, Y) = Cov(Y, X)$$

$$Cov(aX + b, cY + d) = acCov(X, Y)$$

$$Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$$

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

协方差和独立、不相关

- X 和 Y 独立时, $E(XY) = E(X)E(Y)$
- 而 $Cov(X, Y) = E(XY) - E(X)E(Y)$
- 从而, 当 X 和 Y 独立时, $Cov(X, Y) = 0$

- 但 X 和 Y 独立这个前提太强, 我们定义: 若 $Cov(X, Y) = 0$, 称 X 和 Y 不相关。

协方差的意义

- 协方差是两个随机变量具有相同方向变化趋势的度量；
 - 若 $\text{Cov}(X, Y) > 0$ ，它们的变化趋势相同；
 - 若 $\text{Cov}(X, Y) < 0$ ，它们的变化趋势相反；
 - 若 $\text{Cov}(X, Y) = 0$ ，称 X 和 Y 不相关。
- 思考：两个随机变量的协方差，是否有上界？

协方差的上界

- 若 $Var(X) = \sigma_1^2$ $Var(Y) = \sigma_2^2$
- 则 $|Cov(X, Y)| \leq \sigma_1 \sigma_2$
- 当且仅当 X 和 Y 之间有线性关系时，等号成立。

试分析该证明过程？

$$\begin{aligned} \text{Cov}^2(X, Y) &= E^2((X - E(X))(Y - E(Y))) \quad \dots\dots \text{协方差定义} \\ &\leq E((X - E(X))^2(Y - E(Y))^2) \quad \dots\dots\dots \text{方差性质} \\ &\leq E((X - E(X))^2)E((Y - E(Y))^2) \quad \dots\dots\dots \text{期望性质} \\ &= \text{Var}(X)\text{Var}(Y) \quad \dots\dots\dots \text{方差定义} \end{aligned}$$

□ 注：第三行“期望性质”的不等号不一定成立，即： $E(XY) - E(X)E(Y)$ 符号不定。

协方差上界定理的证明

□ 取任意实数 t ，构造随机变量 Z ，

$$Z = (X - E(X)) \cdot t + (Y - E(Y))$$

□ 从而：
$$\begin{cases} E(Z^2) = \sigma_1^2 t^2 + 2Cov(X, Y) + \sigma_2^2 \\ E(Z^2) \geq 0 \end{cases}$$

$$\Rightarrow \sigma_1^2 t^2 + 2Cov(X, Y) \cdot t + \sigma_2^2 \geq 0$$

$$\Rightarrow \Delta = 4Cov^2(X, Y) - 4\sigma_1^2 \sigma_2^2 \leq 0$$

$$\Rightarrow |Cov(X, Y)| \leq \sigma_1 \sigma_2$$

再谈独立与不相关

- 因为上述定理的保证，使得“不相关”事实上即“二阶独立”。
- 即：若 X 与 Y 不相关，说明 X 与 Y 之间没有线性关系(但有可能存在其他函数关系)，不能保证 X 和 Y 相互独立。
- 但对于二维正态随机变量， X 与 Y 不相关等价于 X 与 Y 相互独立。

Pearson相关系数

- 定义 $\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$
- 由协方差上界定理可知, $|\rho| \leq 1$
- 当且仅当X与Y有线性关系时, 等号成立
- 容易看到, 相关系数是标准尺度下的协方差。
上面关于协方差与XY相互关系的结论, 完全适用于相关系数和XY的相互关系。

协方差矩阵

- 对于n个随机向量 $(X_1, X_2 \dots X_n)$ ，任意两个元素 X_i 和 X_j 都可以得到一个协方差，从而形成 $n \times n$ 的矩阵；协方差矩阵是**对称阵**。

$$c_{ij} = E\{[X_i - E(X_i)][X_j - E(X_j)]\} = \text{Cov}(X_i, X_j)$$

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix}$$

联想与思考

□ 若 X 、 Y 独立，则：

$$\text{Var}(XY) = \text{Var}(X)\text{Var}(Y) + \text{Var}(X)E^2(Y) + \text{Var}(Y)E^2(X)$$

□ 对称阵的不同特征值对应的特征向量，是否一定正交？

■ 对称阵和正交阵是否能够建立联系？

矩

□ 对于随机变量 X ， X 的 k 阶原点矩为

$$E(X^k)$$

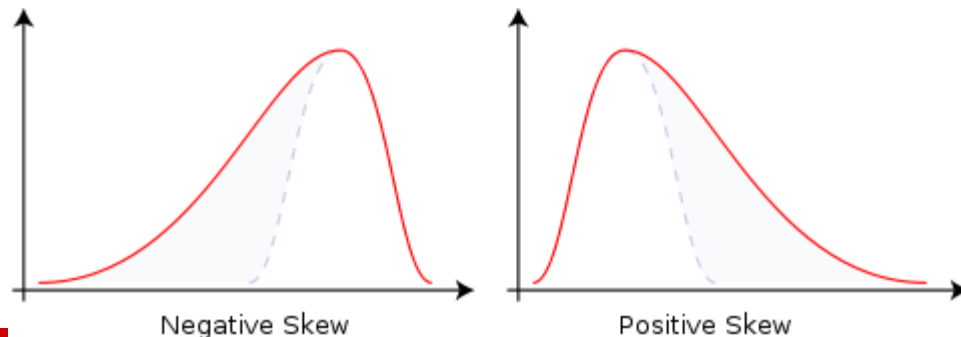
□ X 的 k 阶中心矩为

$$E\{[X - E(X)]^k\}$$

统计参数的总结

- 期望(一阶原点矩)
- 方差(标准差, 二阶中心矩)
- 变异系数(Coefficient of Variation)
 - 标准差与均值的比值称为变异系数, 记为 $C.V$
- 偏度Skewness(三阶)
- 峰度Kurtosis(四阶)

偏度



- 偏度衡量随机变量概率分布的不对称性，是相对于平均值不对称程度的度量。
- 偏度的值可以为正，可以为负或者无定义。
- 偏度为负(负偏)/正(正偏)表示在概率密度函数左侧/右侧的尾部比右侧的长，长尾在左侧/右侧。
- 偏度为零表示数值相对均匀地分布在平均值的两侧，但不一定意味着一定是对称分布。

偏度公式

□ 三阶累积量与二阶累积量的1.5次方的比率。

■ 偏度有时用Skew[X]来表示。

$$\gamma_1 = E\left[\left(\frac{X - \mu}{\sigma}\right)^3\right] = \frac{E[(X - \mu)^3]}{(E[(X - \mu)^2])^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}$$

$$\gamma_1 = E\left[\left(\frac{X - \mu}{\sigma}\right)^3\right] = \frac{E[X^3] - 3\mu E[X^2] + 2\mu^2}{\sigma^3} = \frac{E[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3}$$

峰度 $\frac{\mu_4}{\sigma^4}$

- 峰度是概率密度在均值处峰值高低的特征，通常定义四阶中心矩除以方差的平方减3。

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \frac{\mu_4}{\sigma^4} - 3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2} - 3$$

- $\frac{\mu_4}{\sigma^4}$ 也被称为超值峰度(excess kurtosis)。

- “减3”是为了让正态分布的峰度为0。
- 超值峰度为正，称为尖峰态(leptokurtic)
- 超值峰度为负，称为低峰态(platykurtic)

Code

□ 均值/方差

□ 偏度/峰度

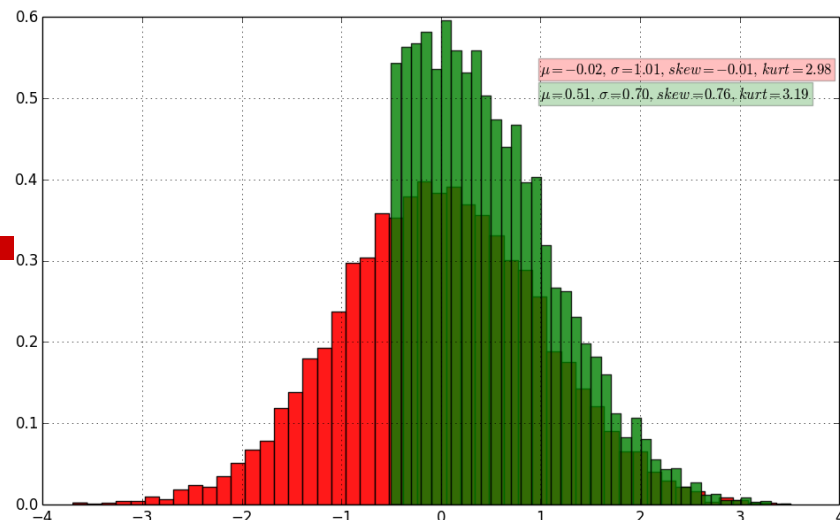
```
def calc_stat(data):  
    [niu, sigma, niu3] = calc(data)  
    n = len(data)  
    niu4 = 0.0  
    for a in data:  
        a -= niu  
        niu4 += a ** 4  
    niu4 /= n  
  
    skew = (niu3 - 3*niu*sigma**2 - niu**3) / (sigma**3)  
    kurt = niu4 / (sigma**4)  
    return [niu, sigma, skew, kurt]
```

```
import matplotlib.pyplot as plt  
import math  
import numpy as np  
  
def calc(data):  
    n = len(data)  
    niu = 0.0  
    niu2 = 0.0  
    niu3 = 0.0  
    for a in data:  
        niu += a  
        niu2 += a**2  
        niu3 += a**3  
    niu /= n  
    niu2 /= n  
    niu3 /= n  
    sigma = math.sqrt(niu2 - niu*niu)  
    return [niu, sigma, niu3]
```

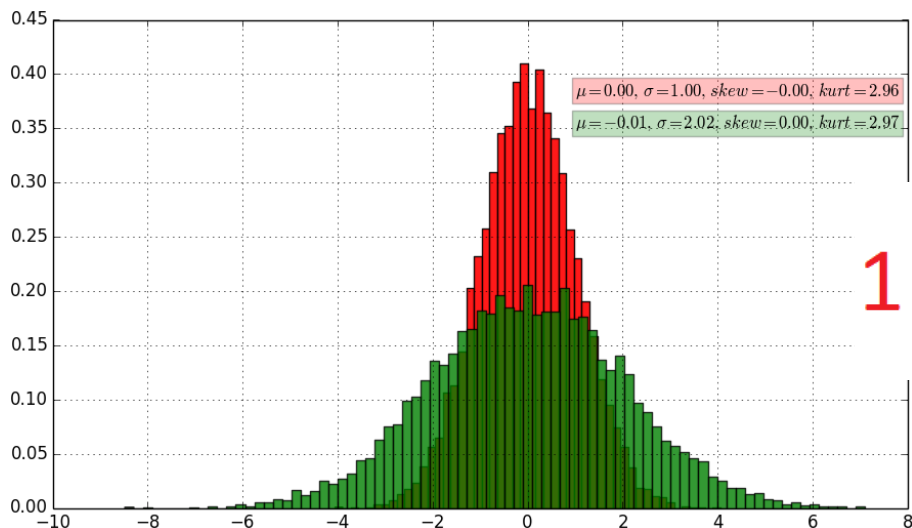
统计量Code

```
if __name__ == "__main__":
    data = list(np.random.randn(10000))
    data2 = list(2*np.random.randn(10000))
    data3 = [x for x in data if x > -0.5]
    data4 = list(np.random.uniform(0, 4, 10000))
    [niu, sigma, skew, kurt] = calc_stat(data)
    [niu2, sigma2, skew2, kurt2] = calc_stat(data2)
    [niu3, sigma3, skew3, kurt3] = calc_stat(data3)
    [niu4, sigma4, skew4, kurt4] = calc_stat(data4)
    print niu, sigma, skew, kurt
    print niu2, sigma2, skew2, kurt2
    print niu3, sigma3, skew3, kurt3
    print niu4, sigma4, skew4, kurt4

    info = r'$\mu=%.2f, \sigma=%.2f, skew=%.2f, kurt=%.2f$' % (niu, sigma, skew, kurt)
    info2 = r'$\mu=%.2f, \sigma=%.2f, skew=%.2f, kurt=%.2f$' % (niu2, sigma2, skew2, kurt2)
    plt.text(1, 0.38, info, bbox=dict(facecolor='red', alpha=0.25))
    plt.text(1, 0.35, info2, bbox=dict(facecolor='green', alpha=0.25))
    plt.hist(data, 50, normed=True, facecolor='r', alpha=0.9)
    plt.hist(data2, 80, normed=True, facecolor='g', alpha=0.8)
    plt.grid(True)
    plt.show()
```



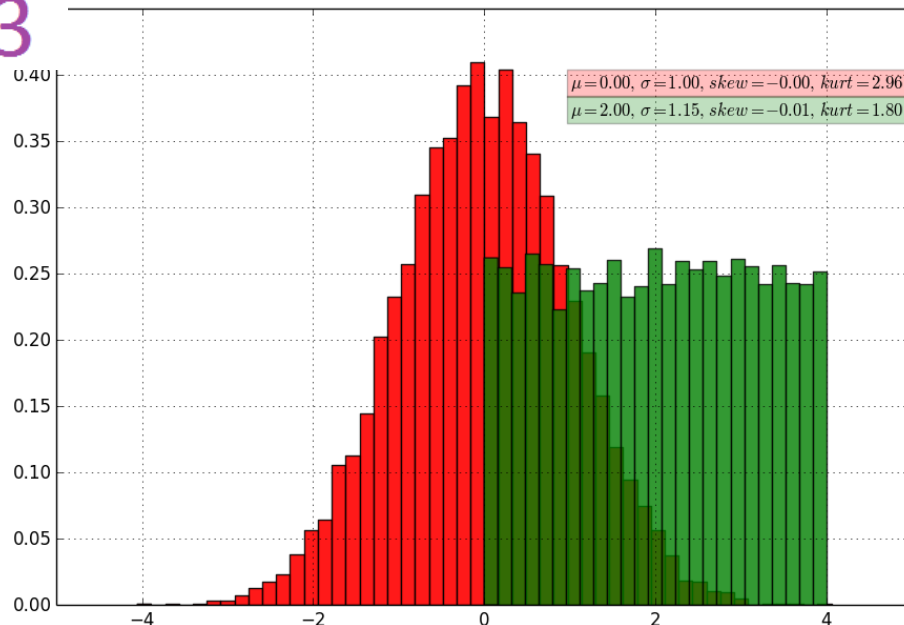
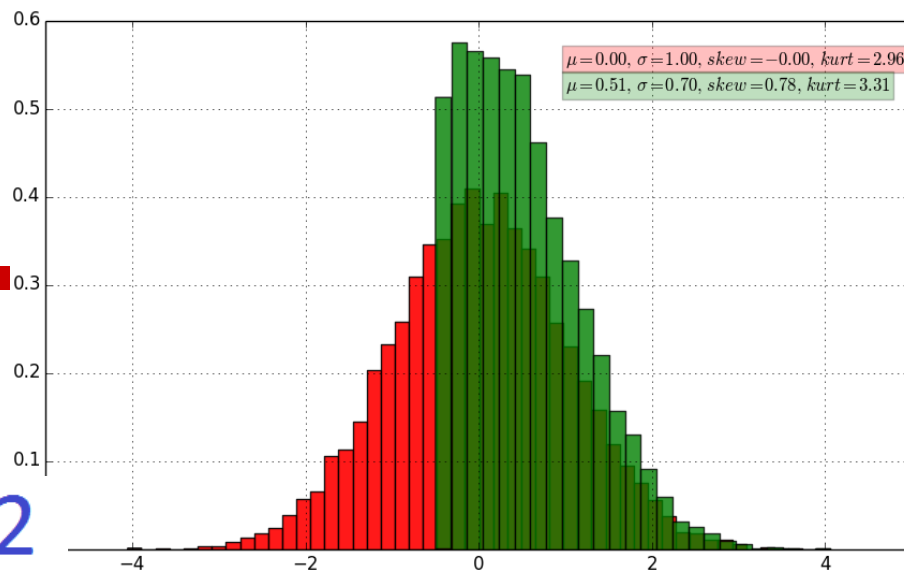
数据分析



1

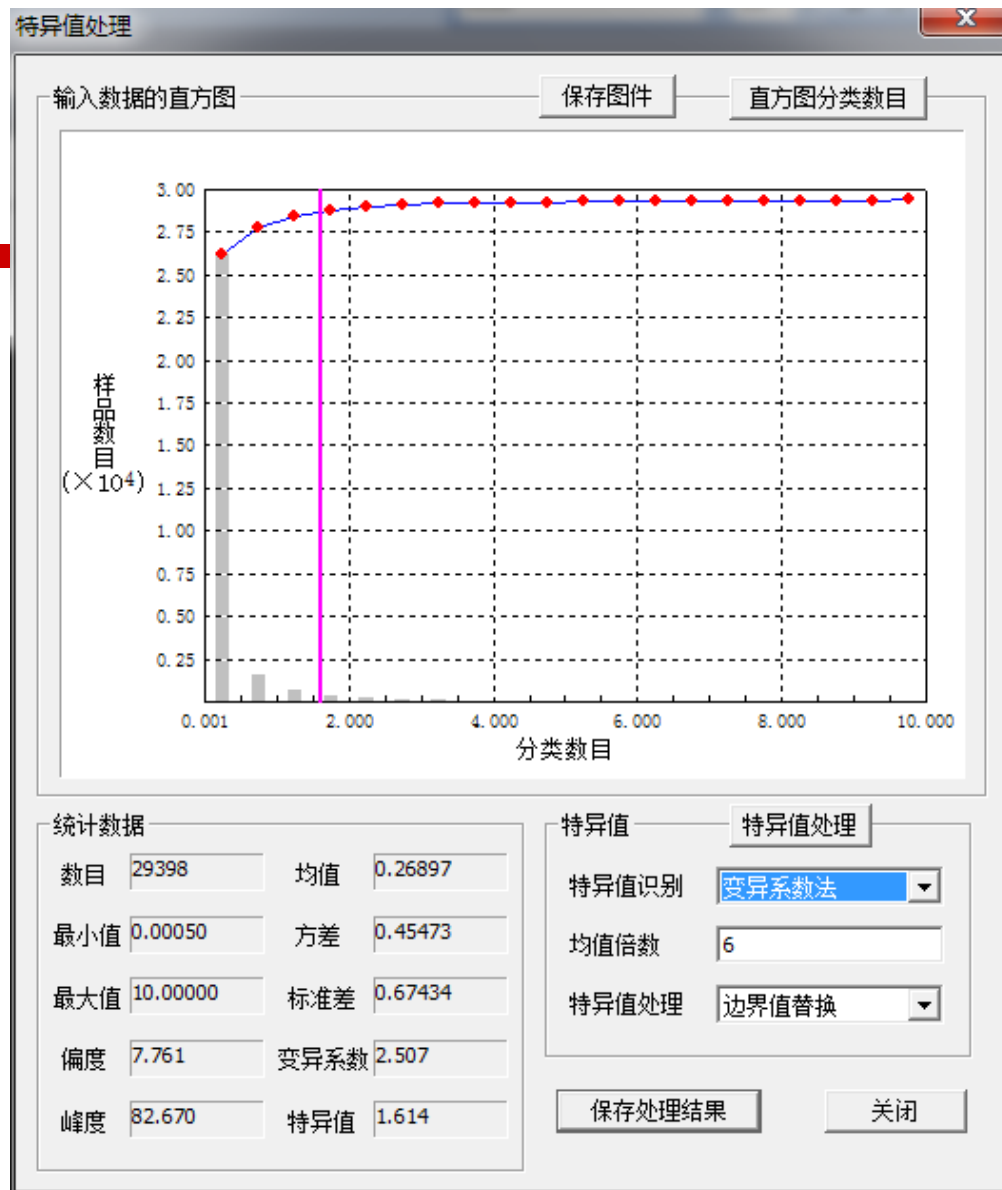
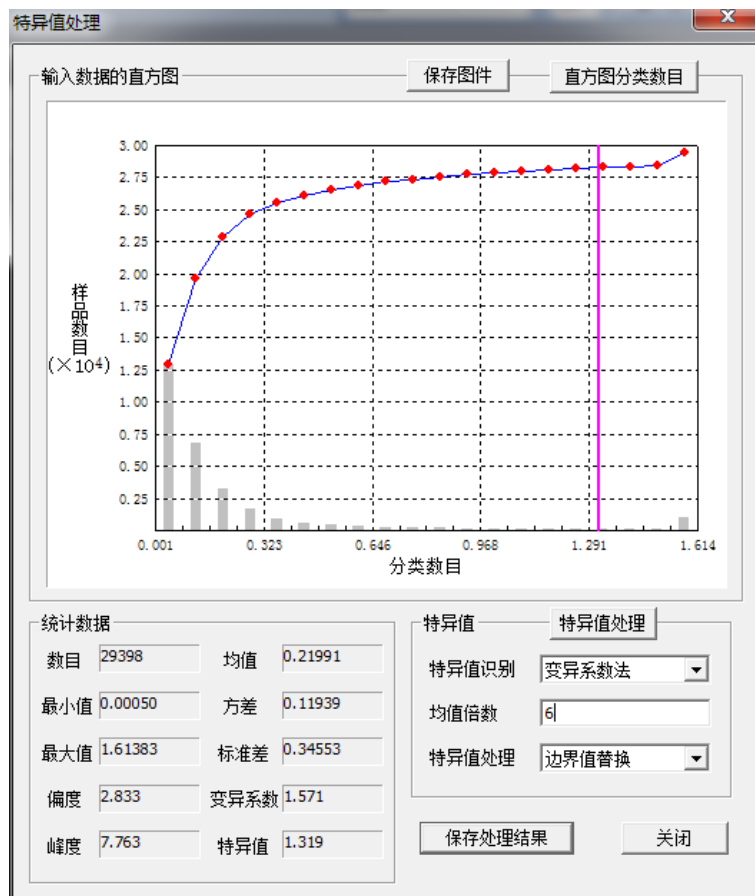
2

3



0.0029	1.0001	-0.0040	2.9643
-0.0137	2.0208	0.0046	2.9748
0.5134	0.6962	0.7825	3.3052
2.0016	1.1548	-0.0107	1.8012

实践中的例子



思考

- 1、给定两个随机变量 X 和 Y ，如何度量这两个随机变量的“距离”？

- 2、设随机变量 X 的期望为 μ ，方差为 σ^2 ，对于任意正数 ε ，试估计概率 $P\{|X-\mu| < \varepsilon\}$ 的下限。
 - 即：随机变量的变化值落在期望值附近的概率

解(以连续型随机变量为例)

$$\begin{aligned} & P\{|X - \mu| \geq \varepsilon\} \\ &= \int_{|X - \mu| \geq \varepsilon} f(x) dx \\ &\leq \int_{|X - \mu| \geq \varepsilon} \frac{|X - \mu|^2}{\varepsilon^2} f(x) dx \\ &= \frac{1}{\varepsilon^2} \int_{|X - \mu| \geq \varepsilon} (X - \mu)^2 f(x) dx \\ &\leq \frac{1}{\varepsilon^2} \int_{-\infty}^{+\infty} (X - \mu)^2 f(x) dx \\ &= \frac{\sigma^2}{\varepsilon^2} \end{aligned}$$

$$\begin{aligned} & P\{|X - \mu| < \varepsilon\} \\ &= 1 - P\{|X - \mu| \geq \varepsilon\} \\ &\geq 1 - \frac{\sigma^2}{\varepsilon^2} \end{aligned}$$

切比雪夫不等式

- 设随机变量 X 的期望为 μ ，方差为 σ^2 ，对于任意正数 ε ，有：

$$P\{|X - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}$$

- 切比雪夫不等式说明， X 的方差越小，事件 $\{|X - \mu| < \varepsilon\}$ 发生的概率越大。即： X 取的值基本上集中在期望 μ 附近。
- 该不等式进一步说明了方差的含义
 - 该不等式可证明大数定理。

大数定理

□ 设随机变量 $X_1, X_2, \dots, X_n, \dots$ 互相独立，并且具有相同的期望 μ 和方差 σ^2 。作前 n 个随机变量的平均 $Y_n = \frac{1}{n} \sum_{i=1}^n X_i$ 则对于任意正数 ε ，有

$$\lim_{n \rightarrow \infty} P\{|Y_n - \mu| < \varepsilon\} = 1$$

大数定理的意义

- 当 n 很大时，随机变量 $X_1, X_2 \dots X_n$ 的平均值 Y_n 在概率意义下无限接近期望 μ 。
- 出现偏离是可能的，但这种可能性很小，当 n 无限大时，这种可能性的概率为0。

思考题

□ 如何证明大数定理？

■ 提示：根据 Y 的定义，求出它的期望和方差，带入切比雪夫不等式即可。

重要推论

- 一次试验中事件A发生的概率为p；重复n次独立试验中，事件A发生了 n_A 次，则p、n、 n_A 的关系满足：
对于任意正数 ε ，

$$\lim_{n \rightarrow \infty} P \left\{ \left| \frac{n_A}{n} - p \right| < \varepsilon \right\} = 1$$

伯努利定理

- 上述推论是最早的大数定理的形式，称为伯努利定理。该定理表明事件A发生的频率 n_A/n 以概率收敛于事件A的概率p，以严格的数学形式表达了频率的稳定性。
- 上述事实为我们在实际应用中用频率来估计概率提供了一个理论依据。
 - 正态分布的参数估计
 - 朴素贝叶斯做垃圾邮件分类
 - 隐马尔可夫模型有监督参数学习

中心极限定理

□ Central Limit Theorem

- 设随机变量 $X_1, X_2, \dots, X_n, \dots$ 互相独立，服从同一分布，并且具有相同的期望 μ 和方差 σ^2 ，则随机变量

$$Y_n = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$$

的分布收敛到标准正态分布。

- 容易得到： $\sum_{i=1}^n X_i$ 收敛到正态分布 $N(n\mu, n\sigma^2)$

例：标准的中心极限定理的问题

- 有一批样本(字符串), 其中a-z开头的比例是固定的, 但是量很大, 需要从中随机抽样。样本量 n , 总体中a开头的字符串占比1%, 需要每次抽到的a开头的字符串占比(0.99%, +1.01%), 样本量 n 至少是多少?
- 问题可以重新表述一下: 大量存在的两点分布 $Bi(1, p)$, 其中, Bi 发生的概率为0.01, 即 $p=0.01$ 。取其中的 n 个, 使得发生的个数除以总数的比例落在区间(0.0099, 0.0101), 则 n 至少是多少?

解：

- 首先，两点分布B的期望为 $\mu=p$ ，方差为 $\sigma^2=p(1-p)$ 。
- 其次，当n较大时，随机变量 $Y = \sum_{i=1}^n B_i$ 近似服从正态分布，事实

上， $X = \frac{Y - n\mu}{\sqrt{n\sigma}} = \frac{\sum_{i=1}^n B_i - n\mu}{\sqrt{n\sigma}}$ 近似服从标准正态分布。

- 从而：

$$P\left\{a \leq \frac{\sum_{i=1}^n B_i}{n} \leq b\right\} \geq 1 - \alpha \Rightarrow P\left\{\frac{\sqrt{n}(a - \mu)}{\sigma} \leq \frac{\sum_{i=1}^n B_i - n\mu}{\sqrt{n\sigma}} \leq \frac{\sqrt{n}(b - \mu)}{\sigma}\right\} \geq 1 - \alpha$$
$$\Rightarrow \Phi\left(\frac{\sqrt{n}(b - \mu)}{\sigma}\right) - \Phi\left(\frac{\sqrt{n}(a - \mu)}{\sigma}\right) \geq 1 - \alpha$$

- 上式中， $\mu=0.01$ ， $\sigma^2=0.0099$ ， $a=0.0099$ ， $b=0.0101$ ， $\alpha=0.05$ 或 0.01 (显著性水平的一般取值)，查标准正态分布表，很容易计算得到n的最小值。
- 注：直接使用二项分布，也能得到结论。

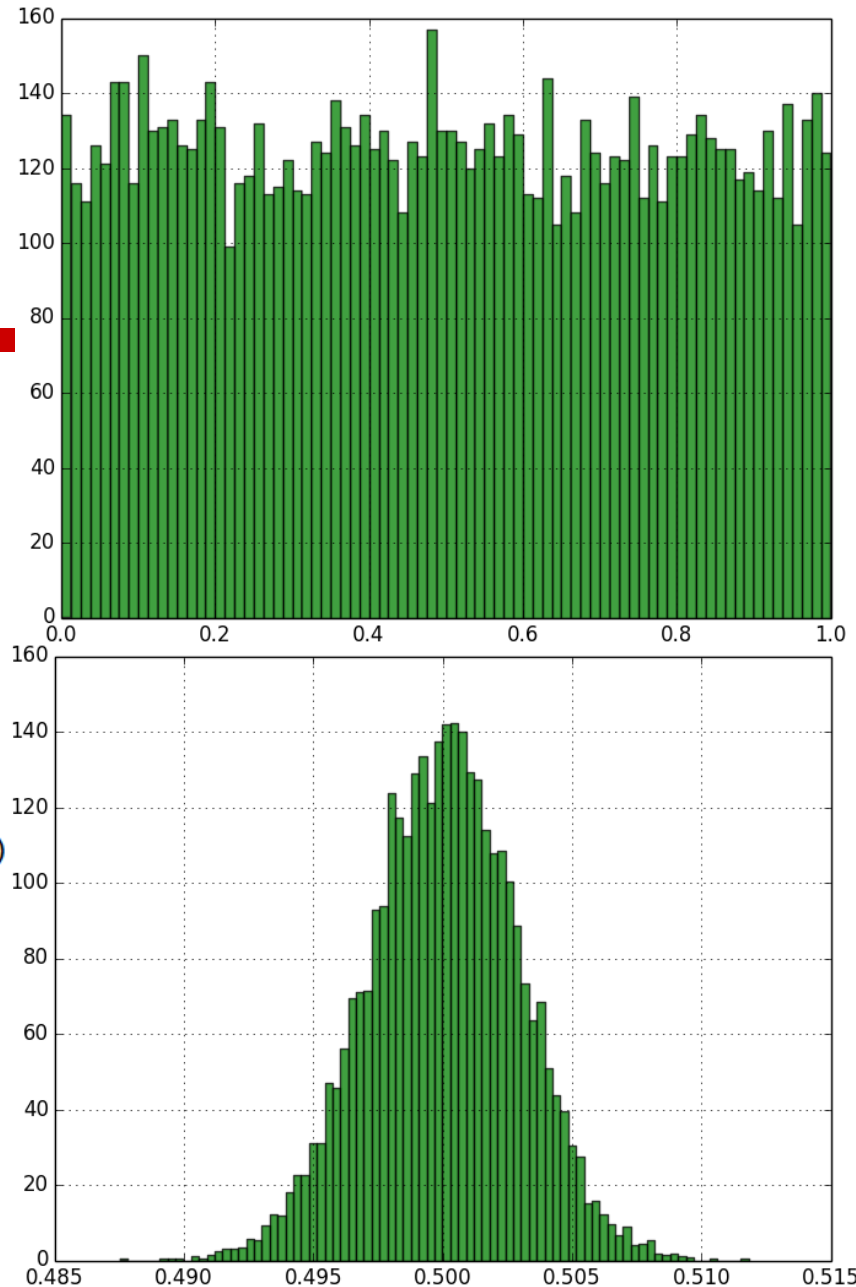
中心极限定理的意义

- 实际问题中，很多随机现象可以看做许多因素的独立影响的综合反应，往往近似服从正态分布。
- 城市耗电量：大量用户的耗电量总和
- 测量误差：许多观察不到的、微小误差的总和
 - 注意：是多个随机变量的和才可以，有些问题是乘性误差，则需要鉴别或者取对数后再使用。
- 线性回归中，将使用该定理论证最小二乘法的合理性

CLT实验

```
if __name__ == "__main__":
    u = numpy.random.uniform(0.0, 1.0, 10000)
    plt.hist(u, 80, facecolor='g', alpha=0.75)
    plt.grid(True)
    plt.show()

    times = 10000
    for time in range(times):
        u += numpy.random.uniform(0.0, 1.0, 10000)
    print len(u)
    u /= times
    print len(u)
    plt.hist(u, 80, facecolor='g', alpha=0.75)
    plt.grid(True)
    plt.show()
```



样本的统计量

□ 设随机变量X的N个样本为 X_1, X_2, \dots, X_n

则样本均值为：
$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

样本方差为：
$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

□ 样本方差的分母使用n-1而非n，是为了无偏。

■ 思考：如何证明？

样本的矩

□ k阶样本原点矩

$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

□ k阶样本中心矩

$$M_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$$

思考

□ 随机变量的矩和样本的矩，有什么关系？

□ 换个提法：

- 假设总体服从某参数为 θ (存在且未知，有可能是值或者向量)的分布，从总体中抽出一组样本 X_1, X_2, \dots, X_n ，如何估计参数 θ ？
- 样本是独立同分布的
- 可以通过 X_1, X_2, \dots, X_n 方便的计算样本的 k 阶矩
- 假设样本的 k 阶矩等于总体的 k 阶矩，可估计出总体的参数。

矩估计

- 设总体的均值为 μ ，方差 σ^2 ，(μ 和 σ 未知，待求)则有原点距表达式：

$$\begin{cases} E(X) = \mu \\ E(X^2) = \text{Var}(X) + [E(X)]^2 = \sigma^2 + \mu^2 \end{cases}$$

- 根据该总体的一组样本，求得原点距：

$$\begin{cases} A_1 = \frac{1}{n} \sum_{i=1}^n X_i \\ A_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \end{cases}$$

矩估计的结论

□ 根据各自阶的中心矩相等，计算得到：

$$\begin{cases} \mu = \bar{X} \\ \sigma^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \end{cases}$$

□ 由于是根据样本求得的估计结果，根据记号习惯，写作：

$$\begin{cases} \hat{\mu} = \bar{X} \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \end{cases}$$

例：正态分布的矩估计

□ 在正态分布的总体中采样得到n个样本：
 X_1, X_2, \dots, X_n ，估计该总体的均值和方差。

□ 解：直接使用矩估计的结论
$$\begin{cases} \hat{\mu} = \bar{X} \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \end{cases}$$

例：均匀分布的矩估计

□ 设 X_1, X_2, \dots, X_n 为定义在 $[a, b]$ 上的均匀分布的总体采样得到的样本，求 a, b 。

□ 解：
均匀分布的均值和方差为
$$\begin{cases} E(X) = \frac{a+b}{2} \\ \text{Var}(X) = \frac{(b-a)^2}{12} \end{cases}$$

矩估计要求满足
$$\begin{cases} \hat{\mu} = \bar{X} \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \end{cases}$$

从而：
$$\begin{cases} \frac{a+b}{2} = \hat{\mu} \\ \frac{(b-a)^2}{12} = \hat{\sigma}^2 \end{cases} \Rightarrow \begin{cases} a = \hat{\mu} - \sqrt{3}\hat{\sigma} \\ b = \hat{\mu} + \sqrt{3}\hat{\sigma} \end{cases}$$

贝叶斯公式带来的思考 $P(A|D) = \frac{P(D|A)P(A)}{P(D)}$

□ 给定某些样本D，在这些样本中计算某结论 A_1 、 $A_2 \dots A_n$ 出现的概率，即 $P(A_i|D)$

$$\begin{aligned} \max P(A_i | D) &= \max \frac{P(D | A_i)P(A_i)}{P(D)} = \max (P(D | A_i)P(A_i)) \xrightarrow{P(A_i) \text{ 近似相等}} \max P(D | A_i) \\ &\Rightarrow \max P(A_i | D) \rightarrow \max P(D | A_i) \end{aligned}$$

- 第一个等式：贝叶斯公式；
- 第二个等式：样本给定，则 $P(D)$ 是常数；
- 第三个箭头：若这些结论 A_1 、 $A_2 \dots A_n$ 的先验概率相等（或近似），则得到最后一个等式：即第二行的公式。

最大似然估计

- 设总体分布为 $f(x, \theta)$ ， $X_1, X_2 \dots X_n$ 为该总体采样得到的样本。因为 $X_1, X_2 \dots X_n$ 独立同分布，于是，它们的联合密度函数为：

$$L(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_k) = \prod_{i=1}^n f(x_i; \theta_1, \theta_2, \dots, \theta_k)$$

- 这里， θ 被看做固定但未知的参数；反过来，因为样本已经存在，可以看成 $X_1, X_2 \dots X_n$ 是固定的， $L(x, \theta)$ 是关于 θ 的函数，即似然函数。
- 求参数 θ 的值，使得似然函数取最大值，这种方法就是最大似然估计。

最大似然估计的具体实践操作

- 在实践中，由于求导数的需要，往往将似然函数取对数，得到对数似然函数；若对数似然函数可导，可通过求导的方式，解下列方程组，得到驻点，然后分析该驻点是极大值点

$$\log L(\theta_1, \theta_2, \dots, \theta_k) = \sum_{i=1}^n \log f(x_i; \theta_1, \theta_2, \dots, \theta_k)$$

$$\frac{\partial L(\theta)}{\partial \theta_i} = 0, \quad i = 1, 2, \dots, k$$

最大似然估计

□ 找出与样本的分布最接近的概率分布模型。

□ 简单的例子

■ 10次抛硬币的结果是：正正反正正正反反正正

□ 假设 p 是每次抛硬币结果为正的概率。则：

□ 得到这样的实验结果的概率是：

$$\begin{aligned} P &= pp(1-p)ppp(1-p)(1-p)pp \\ &= p^7(1-p)^3 \end{aligned}$$

■ 最优解是： $p=0.7$

二项分布的最大似然估计

- 投硬币试验中，进行N次独立试验，n次朝上，N-n次朝下。
- 假定朝上的概率为p，使用对数似然函数作为目标函数：

$$f(n | p) = \log(p^n (1-p)^{N-n}) \xrightarrow{\Delta} h(p)$$

$$\frac{\partial h(p)}{\partial p} = \frac{n}{p} - \frac{N-n}{1-p} \xrightarrow{\Delta} 0 \Rightarrow p = \frac{n}{N}$$

正态分布的最大似然估计

- 若给定一组样本 X_1, X_2, \dots, X_n ，已知它们来自于高斯分布 $N(\mu, \sigma)$ ，试估计参数 μ, σ 。

按照MLE的过程分析

□ 高斯分布的概率密度函数：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

□ 将 X_i 的样本值 x_i 带入，得到：

$$L(x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

化简对数似然函数

$$\begin{aligned}l(x) &= \log \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \\&= \sum_i \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \\&= \left(\sum_i \log \frac{1}{\sqrt{2\pi}\sigma} \right) + \left(\sum_i -\frac{(x_i - \mu)^2}{2\sigma^2} \right) \\&= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2\end{aligned}$$

参数估计的结论

□ 目标函数
$$l(x) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2$$

□ 将目标函数对参数 μ, σ 分别求偏导，很容易得到 μ, σ 的式子：

$$\mu = \frac{1}{n} \sum_i x_i$$

$$\sigma^2 = \frac{1}{n} \sum_i (x_i - \mu)^2$$

符合直观想象

$$\mu = \frac{1}{n} \sum_i x_i$$

$$\sigma^2 = \frac{1}{n} \sum_i (x_i - \mu)^2$$

- 上述结论和矩估计的结果是一致的，并且意义非常直观：样本的均值即高斯分布的均值，样本的**伪方差**即高斯分布的方差。
 - 注：经典意义下的方差，分母是n-1；在似然估计的方法中，求的方差是n
- 该结论将在EM(期望最大化算法)、GMM高斯混合模型中将继续使用。

最大似然估计与过拟合

□ 在校门口统计一定时间段内出入的男女生数目分别为 N_B 和 N_G ，估算该校男女生比例。

$$\begin{cases} P_B = \frac{N_B}{N_B + N_G} \\ P_G = \frac{N_G}{N_B + N_G} \end{cases}$$

□ 若观察到4个女生和1个男生，可以得出该校女生比例是80%吗？

□ 修正公式：

$$\begin{cases} P_B = \frac{N_B + 5}{N_B + N_G + 10} \\ P_G = \frac{N_G + 5}{N_B + N_G + 10} \end{cases} \Rightarrow \begin{cases} P_B = \frac{1 + 5}{1 + 4 + 10} = 40\% \\ P_G = \frac{4 + 5}{1 + 4 + 10} = 60\% \end{cases}$$

偏差与方差

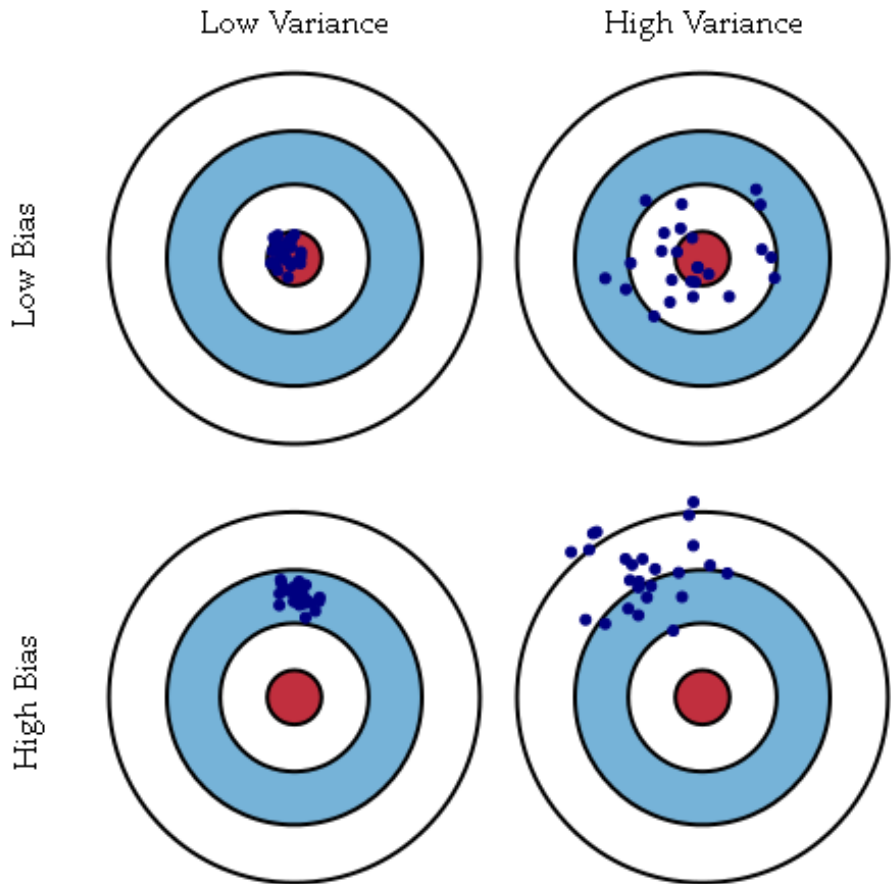
□ 给定数据D，自变量x的相应真实值为 $y(x,D)$ ，预测值为 $h(x)$ ，使用平方误差作为目标函数：

□ 得， $E_D[y(x,D)-h(x)]^2$

$$\begin{aligned} & E_D[y(x,D)-h(x)]^2 \\ &= E_D[y(x,D)-E_D[y(x,D)]+E_D[y(x,D)]-h(x)]^2 \\ &= E_D[\{y(x,D)-E_D[y(x,D)]\}^2 + \{E_D[y(x,D)]-h(x)\}^2 + 2\{y(x,D)-E_D[y(x,D)]\}\{E_D[y(x,D)]-h(x)\}] \\ &= E_D[\{y(x,D)-E_D[y(x,D)]\}^2 + \{E_D[y(x,D)]-h(x)\}^2] \\ &= \underbrace{E_D[\{y(x,D)-E_D[y(x,D)]\}^2]}_{\text{方差Var}} + \underbrace{\{E_D[y(x,D)]-h(x)\}^2}_{\text{偏差Bias}^2} \end{aligned}$$

Bias-variance dilemma

- 圆心为完美预测的模型，蓝色点代表某个模型的学习结果。离靶心越远，准确率越低。
- 低Bias表示离圆心近，高Bias表示离圆心远；
- 高Variance表示学习结果分散，低Variance表示学习结果集中。



思考题：概率计算

- 统计小象学院www.chinahadoop.cn注册用户的实际年龄，均值25岁，标准差2，试估计用户年龄在21-29岁的概率至少是多少？

概率计算

□ 统计小象学院www.chinahadoop.cn注册用户的实际年龄，均值25岁，标准差2，试估计用户年龄在21-29岁的概率至少是多少？

□ 解： $\mu=25, \sigma=2$ ，计算 $P\{21 < X < 29\}$

□ 使用切比雪夫不等式： $P\{|X - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}$

$$\Rightarrow 1 - P\{|X - \mu| \geq \varepsilon\} \geq \frac{\sigma^2}{\varepsilon^2}$$

$$\Rightarrow 1 - P\{|X - 25| \geq 4\} \geq 1 - \frac{2^2}{4^2} = 75\%$$

思考：随机变量无法直接(完全)观察

□ 随机挑选10000位志愿者，测量他们的身高：
若样本中存在男性和女性，身高分别服从
 $N(\mu_1, \sigma_1)$ 和 $N(\mu_2, \sigma_2)$ 的分布，试估计 $\mu_1, \sigma_1, \mu_2, \sigma_2$ 。

□ 无监督分类：聚类/EM

思考

- 分析样本方差除以 $n-1$ 而非 n 的原因。
- 仿照指数分布的概率密度函数 $f(x) = \lambda e^{-\lambda x}$ ，猜测相对应的幂分布的概率密度函数，查阅关于幂律分布的相关文献。

$$f(x) = ax^{-r}, \quad a, r \text{ 为正常数}$$

参考文献

- 王松桂，程维虎，高旅端编，概率论与数理统计，科学出版社，2000

我们在这里

□ <http://wenda.ChinaHadoop.cn>

■ 视频/课程/社区

□ 微博

■ @ChinaHadoop

■ @邹博_机器学习

□ 微信公众号

■ 小象

■ 大数据分析挖掘



课程资源

- 直播课的入口
- 录播视频和讲义资料



搜索课程 搜索

首页 选课中心 小象问答 机器学习实训营 小象训练营 小象公开课

机器学习

算法推导+代码实现+参数调试+应用场景

开课时间：5月23日

主讲人：邹博

我要参团



《机器学习》第三期

★★★★★ (0评价)

承诺服务

试 问 疑 练 动

介绍 课程(2) 评价 话题 笔记



《机器学习算法基础》每周直播课

★★★★★



《机器学习》三期录屏回放与资料

★★★★★

感谢大家！

恳请大家批评指正！