数理统计与参数估计

七月算法 邹博

2015年10月17日

计算概率

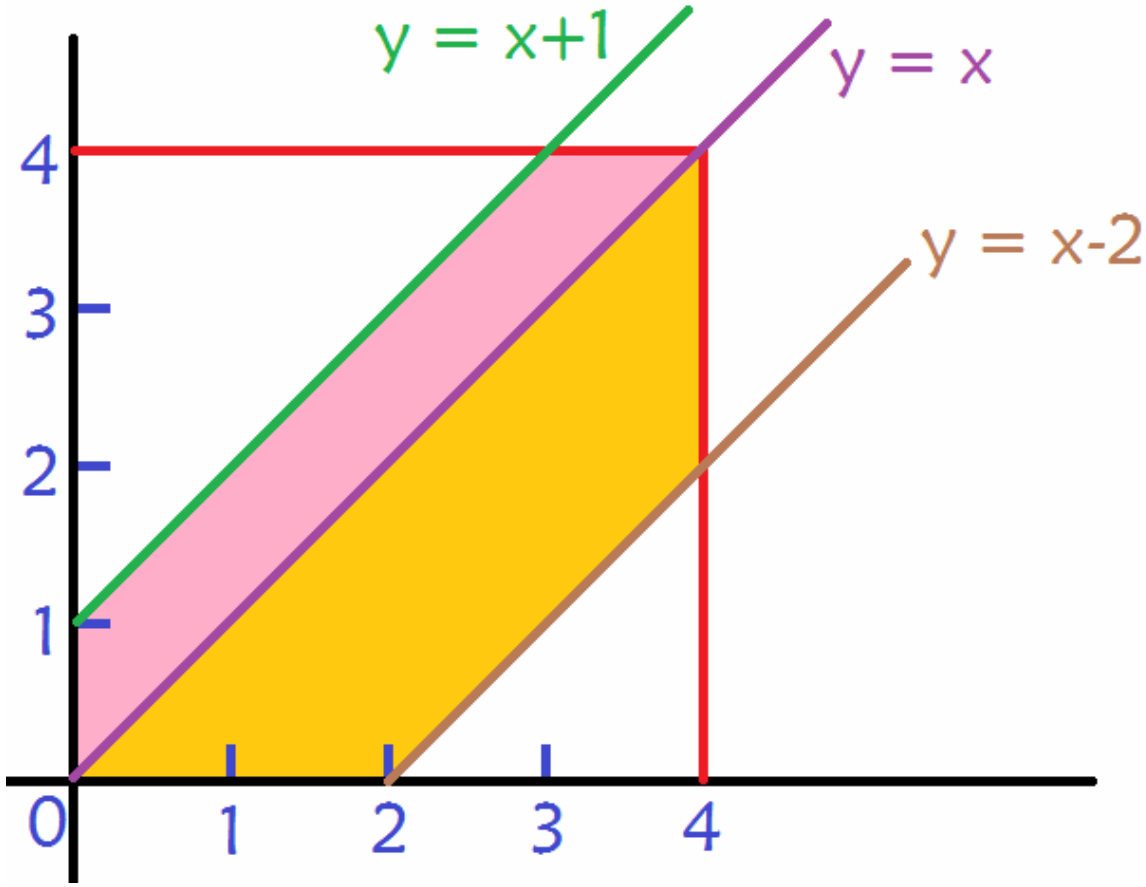
- A、B两国元首相约在首都机场晚20点至24点交换一份重要文件。如果A国的飞机先到，A会等待1个小时；如果B国的飞机先到了，B会等待2个小时。假设两架飞机在20点至24点降落机场的概率是均匀分布，试计算能够在20点至24点完成交换的概率。
  - 假设交换文件本身不需要时间。

事件的形式化表达

- 假定A到达的时刻为x，B达到的时刻为y，则完成交接需满足0<y-x<1或者0<x-y<2；
- 同时要求20<x<24，20<y<24；
- 由于x，y系数都为1，为作图方便，可以将24<x<20，24<y<20平移成4<x<0，4<y<0。

计算面积

- 三角形面积
  - 9/2、2
- 矩形面积
  - 16
- 概率
  - 19/32



一定接受率下的采样

- 已知有个rand7()的函数，返回1到7随机自然数，让利用这个rand7()构造rand10() 随机1~10。
- 解：因为rand7仅能返回1~7的数字，少于rand10的数目。因此，多调用一次，从而得到49种组合。超过10的整数倍部分，直接丢弃。

Code

```
int rand10()
{
    int a1, a2, r;
    do
    {
        a1 = rand7() - 1;
        a2 = rand7() - 1;
        r = a1 * 7 + a2;
    } while (r >= 40);
    return r / 4 + 1;
}
```

期望

- 离散型 $$E(X)=\sum_i x_i p_i$$
- 连续型 $$E(X)=\int_{-\infty}^{\infty} xf(x)dx$$
- 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。

面试题

- 从1,2,3,.....,98,99,2015这100个数中任意选择若干个数(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：

$$E\left(P\left\{X_i=1\right\}\right)=\frac{2^m\cdot\sum_{k\in odd}C_n^k}{2^{m+n}}=\frac{\sum_{k\in odd}C_n^k}{2^n}=\frac{1}{2}$$

总期望

- 11位二进制数中，每个位取1的期望都是0.5

$$
\begin{aligned}
E(X) &= E\left(\sum_{i=1}^{11}\left(X_i \cdot P\{X_i\}\right)\right) \\
&= E\left(\sum_{i=1}^{11}\left(2^i \cdot P\{X_i=1\}+0\cdot P\{X_i=0\}\right)\right) \\
&= E\left(\sum_{i=1}^{11}\left(2^i \cdot P\{X_i=1\}\right)\right) \\
&= \sum_{i=1}^{11} E\left(2^i \cdot P\{X_i=1\}\right) = \sum_{i=1}^{11} 2^i \cdot E\left(P\{X_i=1\}\right) \\
&= \sum_{i=1}^{11} 2^i \cdot \frac{1}{2} = \frac{1}{2}\sum_{i=1}^{11} 2^i = \frac{(11111111111)_2}{2} \\
&= 1023.5
\end{aligned}
$$

采样模拟

```
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;
}
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;
}
```

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;
}
```

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。

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

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：

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

问题分析

- 从而n个字符串的映射，即二项分布：

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

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

- 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 $$Var(X) = E\{[X - E(X)]^2\}$$
- 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$Cov(X,Y)>0$，它们的变化趋势相同，若$Cov(X,Y)<0$，它们的变化趋势相反；若$Cov(X,Y)=0$，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 $Var(X)=\sigma_1^2 \quad Var(Y)=\sigma_2^2$
- 则 $|Cov(X,Y)| \le \sigma_1\sigma_2$
- 当且仅当X和Y之间有线性关系时，等号成立。

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“线性独立”。
- 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。

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维随机向量(X_1,X_2...X_n)，任意两个元素Xi和Xj都可以得到一个协方差，从而形成n*n的矩阵；显然，协方差矩阵是对称阵。

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

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

联想与思考

- 对称阵的不同特征值对应的特征向量，是否一定正交？

矩

- 对于随机变量X，X的k阶原点矩为

$$E\left(X^k\right)$$

- X的k阶中心矩为

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

统计参数的总结

- 均值(期望，一阶)
- 方差(标准差，二阶)
- 变异系数(Coefficient of Variation)
  - 标准差与平均数的比值称为变异系数，记为C·V
- 偏度Skew(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，是概率密度曲线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(负偏态)意味着在概率密度函数左侧的尾部比右侧的长，绝大多数的值(包括中位数在内)位于平均值的右侧。
- 偏度为正(正偏态)意味着在概率密度函数右侧的尾部比左侧的长，绝大多数的值(包括中位数在内)位于平均值的左侧。
- 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。

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

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

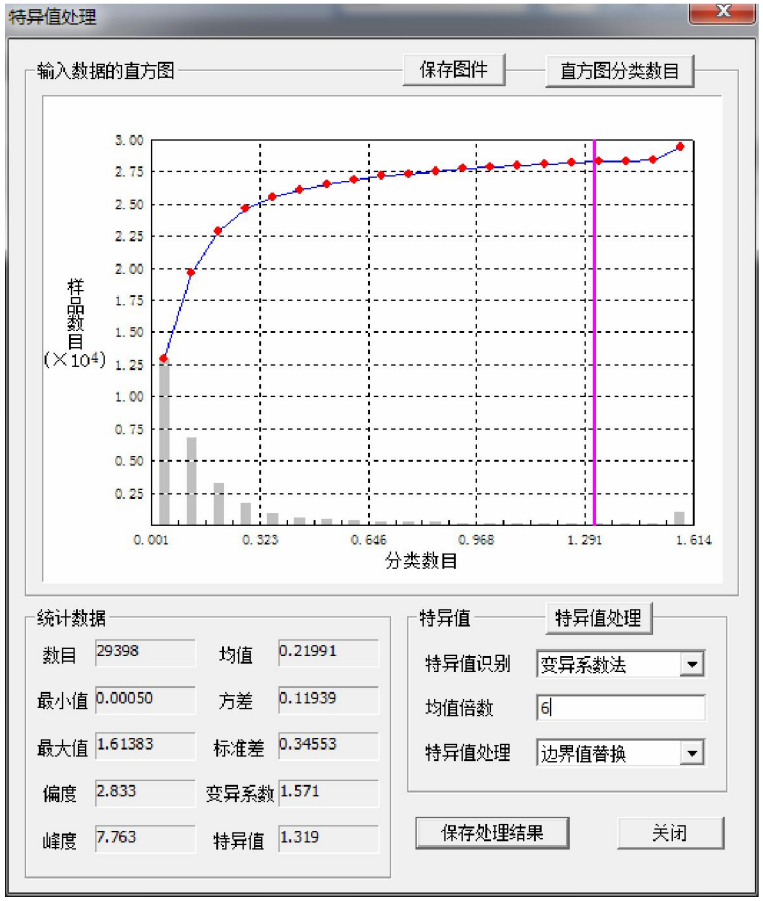
峰度 $\frac{\mu_4}{\sigma^4}$

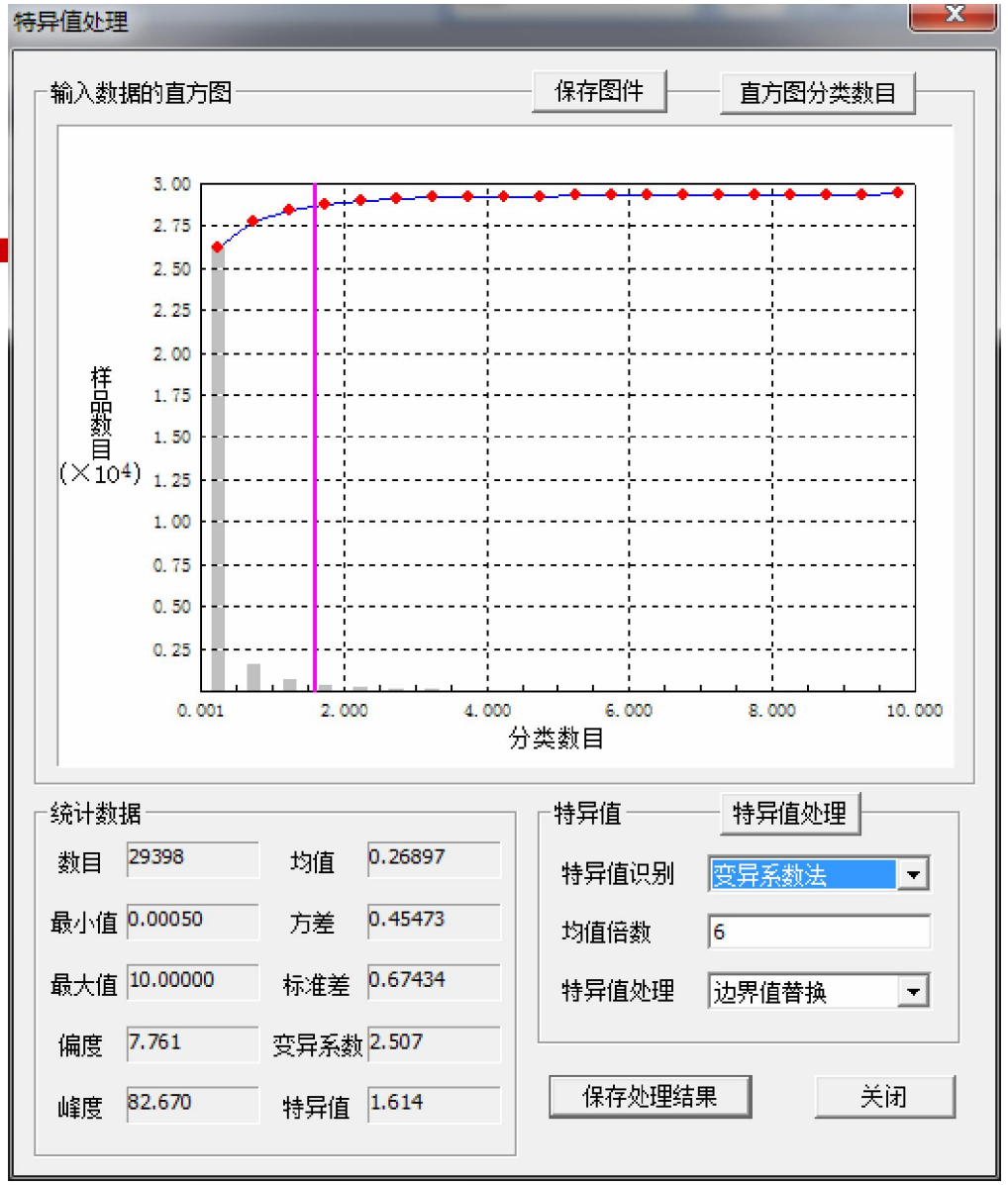
- 峰度是概率密度在均值处峰值高低的特征，通常定义四阶中心矩除以方差的平方减3。

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

- $\frac{\mu_4}{\sigma^4}$ 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)

实践中的例子


特异值处理
输入数据的直方图
保存图件
直方图分类数目
样品数目 (×10⁴)
分类数目
统计数据
数目 29398
均值 0.21991
最小值 0.00050
方差 0.11939
最大值 1.61383
标准差 0.34553
偏度 2.833
变异系数 1.571
峰度 7.763
特异值 1.319
特异值
特异值处理
特异值识别 变异系数法
均值倍数 6
特异值处理 边界值替换
保存处理结果
关闭


特异值处理
输入数据的直方图
保存图件
直方图分类数目
样品数目 (×10⁴)
分类数目
统计数据
数目 29398
均值 0.26897
最小值 0.00050
方差 0.45473
最大值 10.00000
标准差 0.67434
偏度 7.761
变异系数 2.507
峰度 82.670
特异值 1.614
特异值
特异值处理
特异值识别 变异系数法
均值倍数 6
特异值处理 边界值替换
保存处理结果
关闭

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)

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

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

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- μ|＜ε}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$X_1,X_2...X_n...$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$\varepsilon$，有

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

大数定理的意义

- 当n很大时，随机变量$X_1,X_2...X_n$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，

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

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。
  - 朴素贝叶斯做垃圾邮件分类
  - 正态分布的参数估计

中心极限定理

- 设随机变量$X_1, X_2 \ldots X_n \ldots$互相独立，服从同一分布，并且具有相同的期望$\mu$和方差$\sigma^2$，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$N(n\mu, n\sigma^2)$

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：

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

$$\Rightarrow \Phi\left(\frac{\sqrt{n}(b-\mu)}{\sigma}\right)-\Phi\left(\frac{\sqrt{n}(a-\mu)}{\sigma}\right)\ge 1-\alpha$$

- 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

样本的统计量

- 设X_1，X_2，…，X_n为一组样本，则
- 样本均值 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$
- 样本方差 $S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$
  - 样本方差的分母使用n-1而非n，是为了无偏。

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

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

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...，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...，X_n方便的计算出样本的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：

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

- 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：

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

- 由于是根据样本求得的估计结果，根据记号习惯，写作：

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

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：$X_1, X_2 \ldots X_n$，估计该总体的均值和方差。

- 解：直接使用矩估计的结论

$$\begin{cases} \hat{\mu} = \overline{X} \\ \hat{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \end{cases}$$

例：均匀分布的矩估计

- 设X1,X2,…,Xn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：

已知均匀分布的均值和方差为 $\begin{cases} E(X)=\dfrac{a+b}{2} \\ Var(X)=\dfrac{(b-a)^2}{12} \end{cases}$

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

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

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.....A_n出现的概率，即P(A_i|D)

$$\max P(A_i|D)=\max\frac{P(D|A_i)P(A_i)}{P(D)}=\max(P(D|A_i)P(A_i))\rightarrow\max P(D|A_i)$$

$$\Rightarrow\max P(A_i|D)\rightarrow\max P(D|A_i)$$

  - 第一个等式：贝叶斯公式；
  - 第二个等式：样本给定，则对任何A_i,P(D)是常数；
  - 第三个箭头：若这些结论A_1、A_2.....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。

极大似然估计

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

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

- 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$x_1, x_2 \ldots x_n$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，使得似然函数取极大值，这种方法就是极大似然估计。

极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

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

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

极大似然估计

- 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：

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

极大似然估计MLE

- 目标函数：$\max P = \max_{0 \le p \le 1} p^7 (1-p)^3$
- 最优解是：p=0.7
  - 思考：如何求解？

- 一般形式：$L_{\bar{p}} = \prod_x p(x)^{\bar{p}(x)}$

$p(x)$模型是估计的概率分布

$\bar{p}(x)$是实验结果的分布

正态分布的极大似然估计

- 若给定一组样本$X_1,X_2...X_n$，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

$$l(x)=\log\prod_i \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$=\sum_i \log\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$=\left(\sum_i \log\frac{1}{\sqrt{2\pi}\sigma}\right)+\left(\sum_i -\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

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

参数估计的结论

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

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

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

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。

贝叶斯公式 $P(A|B)=\frac{P(B|A)P(A)}{P(B)}$

- 给定某系统的若干样本x，计算该系统的参数，即

$$P(\theta|x)=\frac{P(x|\theta)P(\theta)}{P(x)}$$

  - P(θ)：没有数据支持下，θ发生的概率：先验概率。
  - P(θ |x)：在数据x的支持下，θ发生的概率：后验概率。
  - P(x| θ)：给定某参数θ的概率分布：似然函数。
- 例如：
  - 在没有任何信息的前提下，猜测某人姓氏：先猜李王张刘.....猜对的概率相对较大：先验概率。
  - 若知道某人来自“牛家村”，则他姓牛的概率很大：后验概率——但不排除他姓郭、杨等情况。

思考题：概率计算

- 随机抽查发现“七月题库APP”的用户实际年龄，调查结果显示年龄均值25岁，标准差2，那么实际用户年龄在21-29岁的概率至少是多少？

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

- 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”？
- 对称阵的不同特征值对应的特征向量，是否一定正交？
- 如何证明大数定理？
- 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。

$$f(x)=ax^{-r},\ a,r\text{为正常数}$$

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

- 七月算法 http://www.julyedu.com/
  - 视频/课程/社区
- 七月题库APP：Android/iOS
  - http://www.julyapp.com/
- 微博
  - @研究者July
  - @七月题库
  - @邹博_机器学习
- 微信公众号
  - julyedu

感谢大家！

恳请大家批评指正！