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

概率论与贝叶斯先验

邹博

主要内容

- 概率论基础
  - 概率与直观
  - 常见概率分布
  - Sigmoid/Logistic函数的引入
- 统计量
  - 期望/方差/协方差/相关系数
  - 独立和不相关
- 大数定律
- 中心极限定理
- 最大似然估计
  - 过拟合

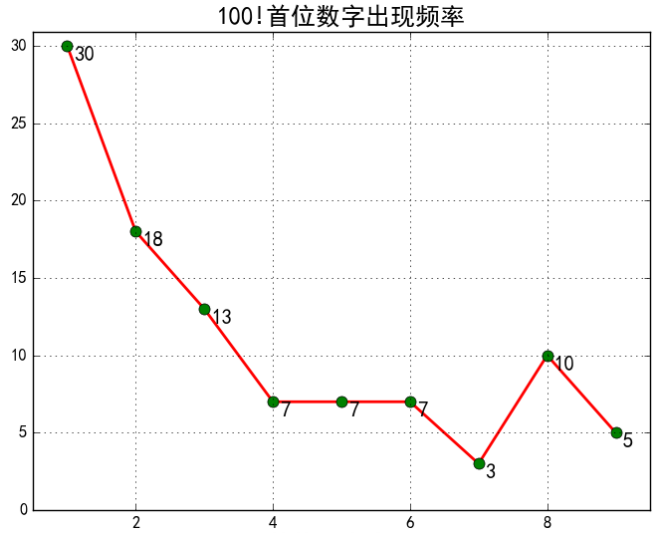
统计数字的概率

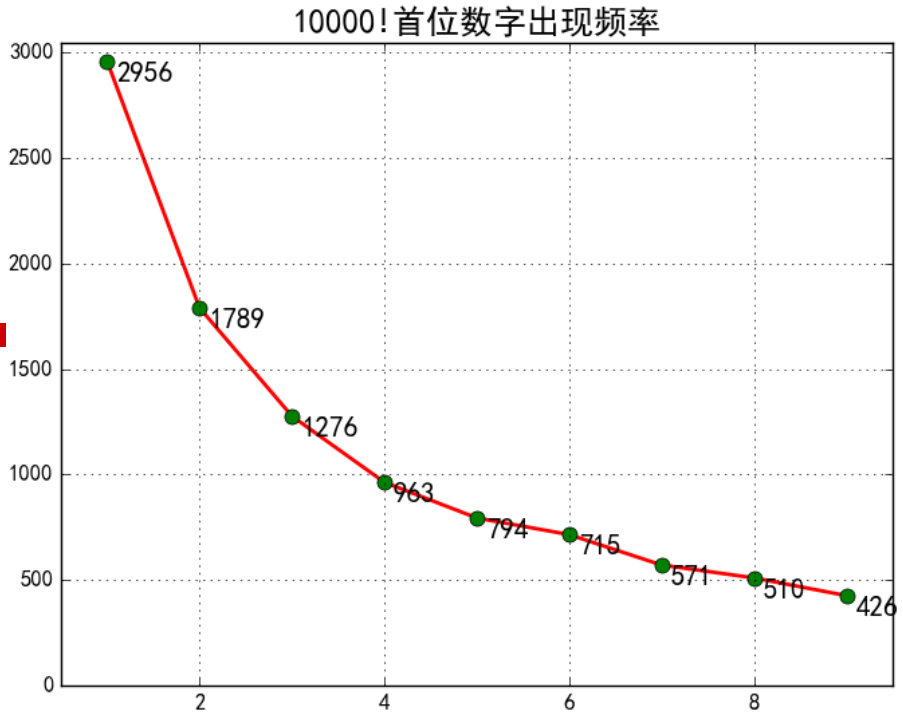
- 给定某正整数N，统计从1到N!的所有数中，首位数字出现1的概率。
- 进而，可以计算首位数字是2的概率，是3的概率，从而得到一条“九点分布”。

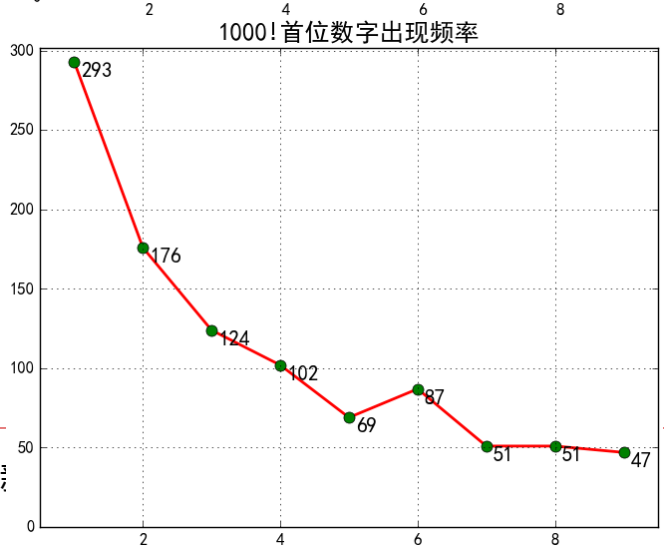
Code

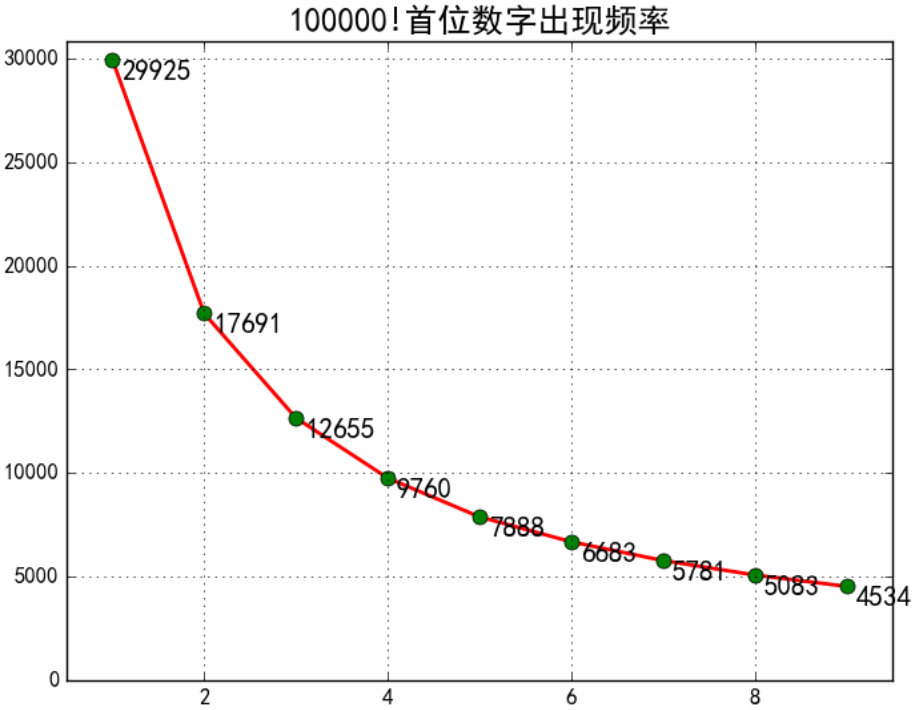
```
def first_digital(x):
    while x >= 10:
        x /= 10
    return x


if __name__ == "__main__":
    n = 1
    frequency = [0] * 9
    for i in range(1, 1000):
        n *= i
        m = first_digital(n) - 1
        frequency[m] += 1
    print frequency
    plt.plot(frequency, 'r-', linewidth=2)
    plt.plot(frequency, 'go', markersize=8)
    plt.grid(True)
    plt.show()
```

数字的概率


100!首位数字出现频率
30
18
13
7
7
7
3
10
5


10000!首位数字出现频率
2956
1789
1276
963
794
715
571
510
426


1000!首位数字出现频率
293
176
124
102
69
87
51
51
47


100000!首位数字出现频率
29925
17691
12655
9760
7888
6683
5781
5083
4534

本福特定律

- 本福特定律(本福德法则, Frank Benford)，又称第一数字定律，是指在实际生活得出的一组数据中，以1为首位数字出现的概率约为总数的三成；是直观想象1/9的三倍。
  - 阶乘/素数数列/斐波那契数列首位
  - 住宅地址号码
  - 经济数据反欺诈
  - 选举投票反欺诈

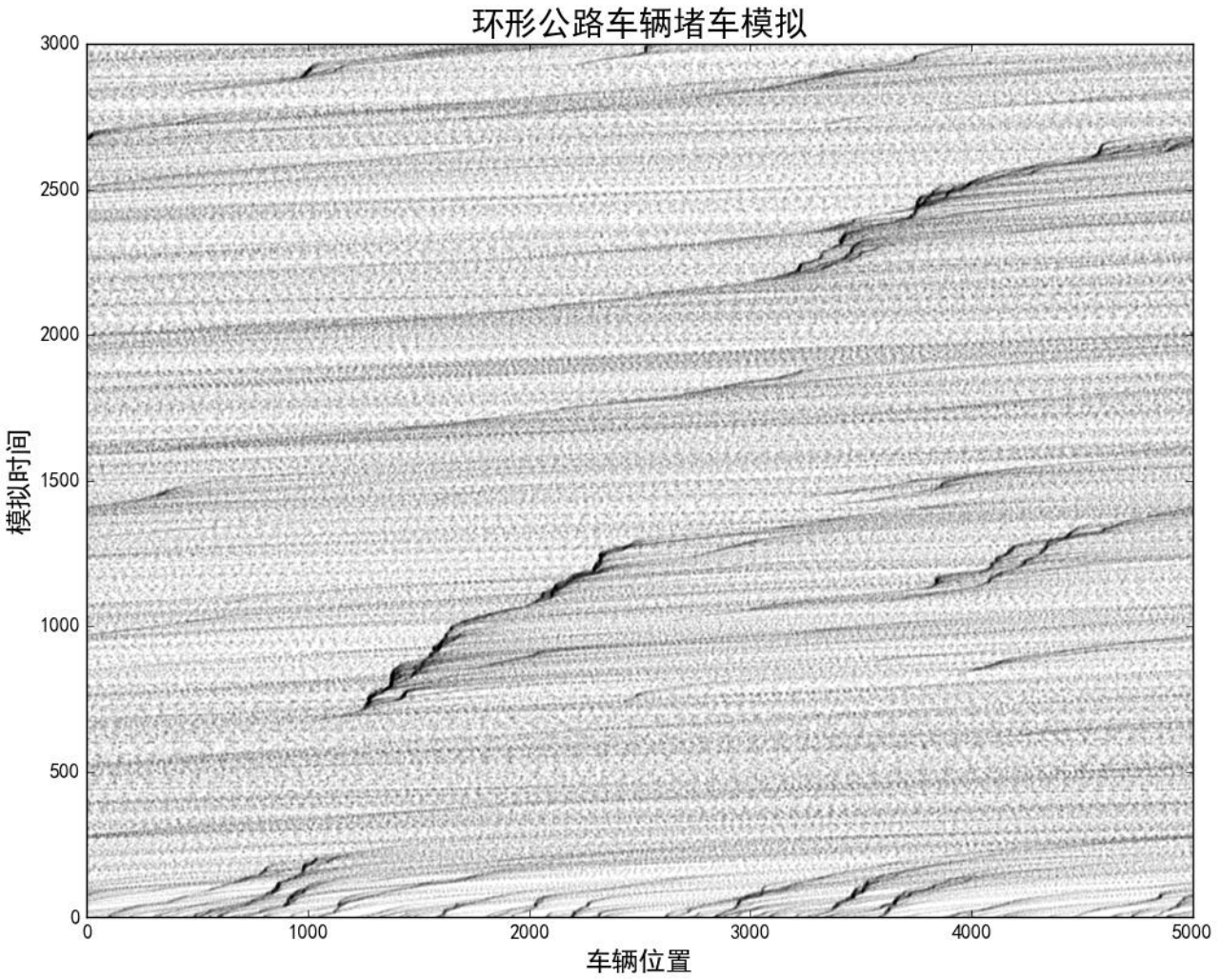
数字	出现概率
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%

公路堵车概率模型

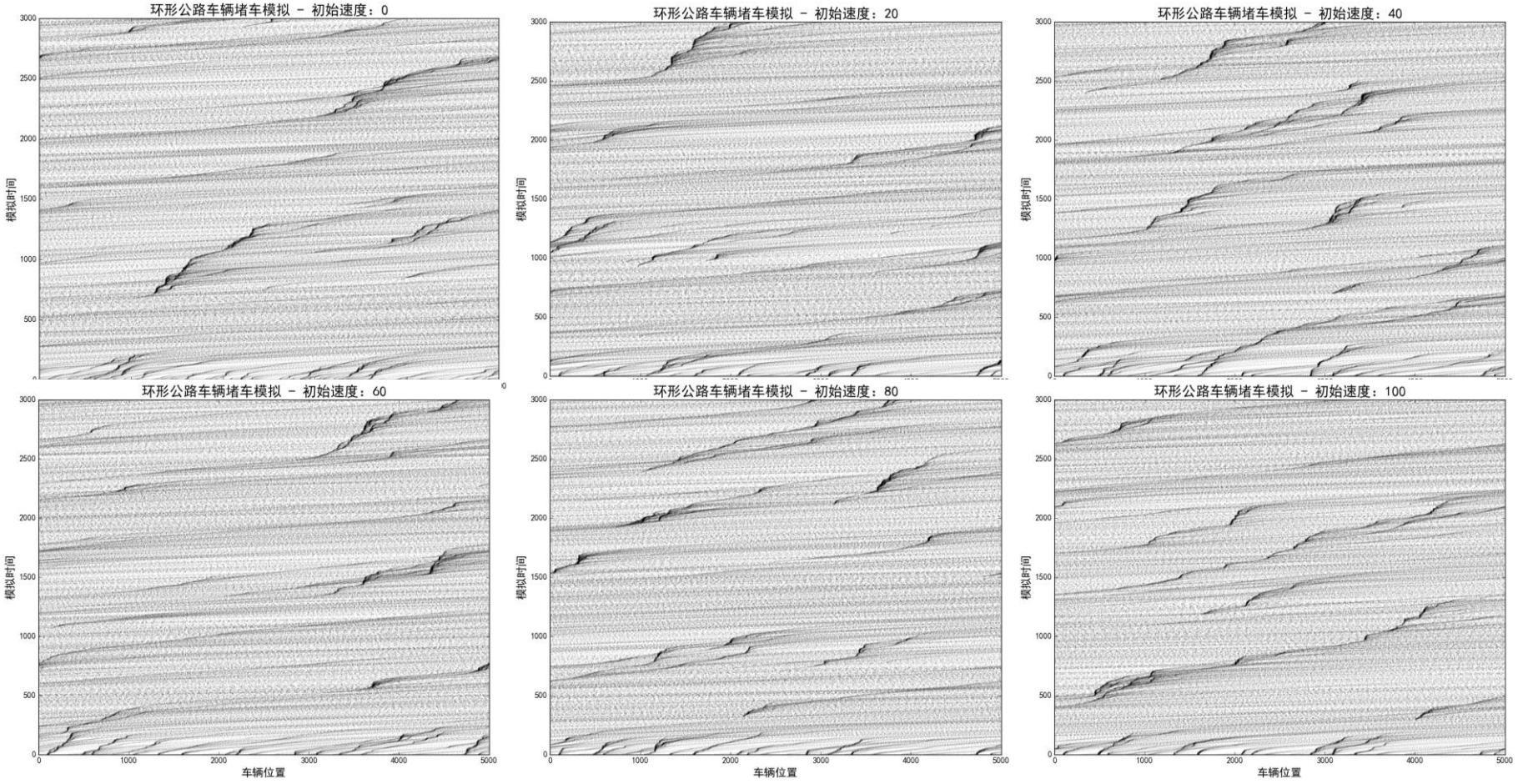
- Nagel-Schreckenberg交通流模型
- 路面上有N辆车，以不同的速度向前行驶，模拟堵车问题。有以下假设：
  - 假设某辆车的当前速度是v。
  - 若前方可见范围内没车，则它在下一秒的车速提高到v+1，直到达到规定的最高限速。
  - 若前方有车，前车的距离为d，且d＜v，则它下一秒的车速降低到d-1。
  - 每辆车会以概率p随机减速v-1。

Nagel-Schreckenberg 模型模拟

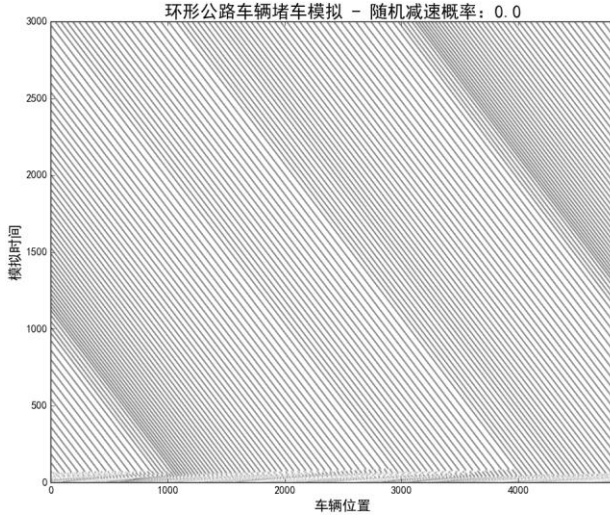
```
path = 5000      # 环形公路的长度
n = 100          # 公路中的车辆数目
v0 = 5           # 车辆的初始速度
p = 0.3          # 随机减速概率
Times = 3000

np.random.seed(0)
x = np.random.rand(n) * path
x.sort()
v = np.tile([v0], n).astype(np.float)

plt.figure(figsize=(10, 8), facecolor=
for t in range(Times):
    plt.scatter(x, [t]*n, s=1, c='k',
    for i in range(n):
        if x[(i+1)%n] > x[i]:
            d = x[(i+1) % n] - x[i]
        else:
            d = path - x[i] + x[(i+1)
        if v[i] < d:
            if np.random.rand() > p:
                v[i] += 1
            else:
                v[i] -= 1
        else:
            v[i] = d - 1
    v = v.clip(0, 150)
    x += v
    clip(x, path)
plt.xlim(0, path)
plt.ylim(0, Times)
plt.xlabel(u'车辆位置', fontsize=16)
plt.ylabel(u'模拟时间', fontsize=16)
plt.title(u'环形公路车辆堵车模拟', font
plt.tight_layout(pad=2)
plt.show()
```

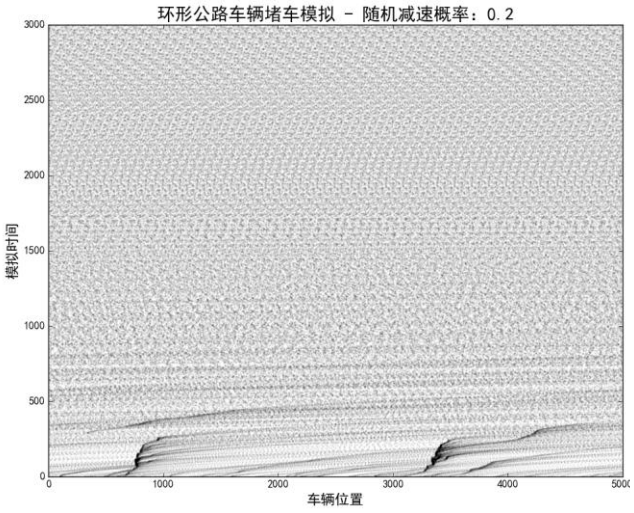


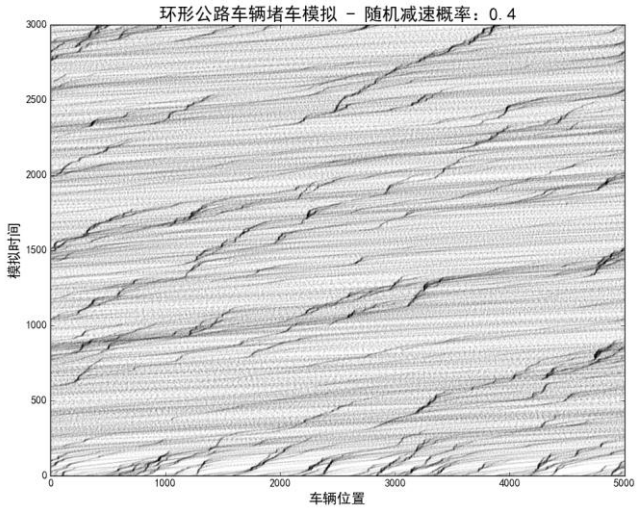
初始车速对NS模型的影响

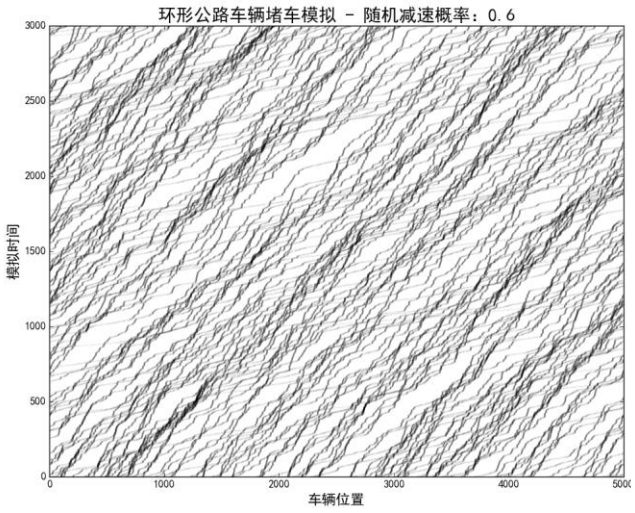


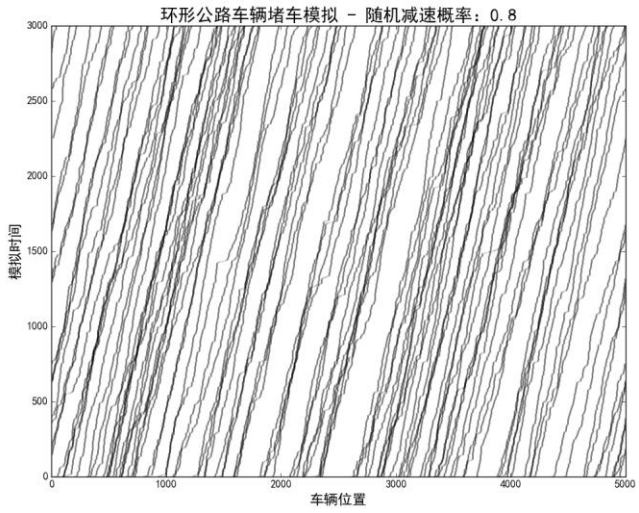
减速概率对NS模型的影响

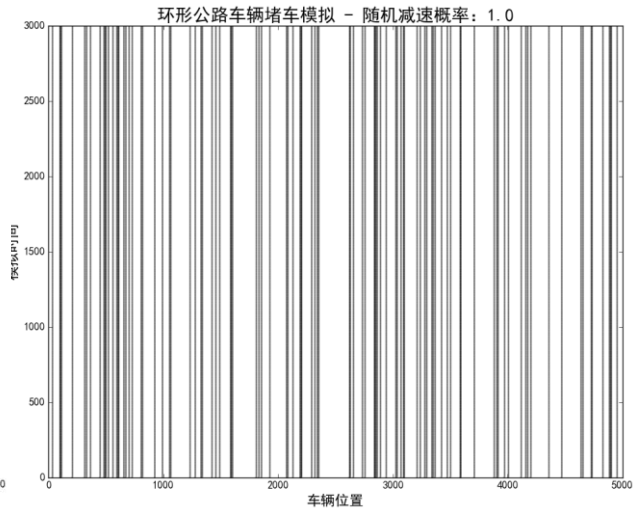






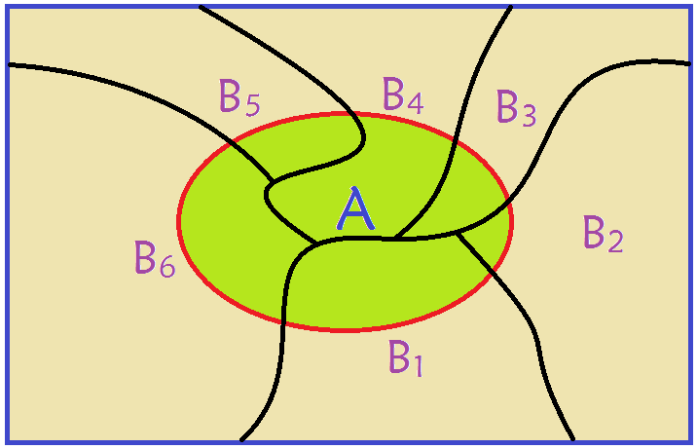






概率公式

- 条件概率：$P(A|B)=\frac{P(AB)}{P(B)}$



- 全概率公式：

$$P(A)=\sum_i P(A|B_i)P(B_i)$$

- 贝叶斯(Bayes)公式：

$$P(B_i|A)=\frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$$

思考题

- 8支步枪中有5支已校准过，3支未校准。一名射手用校准过的枪射击，中靶概率为0.8；用未校准的枪射击，中靶概率为0.3；现从8支枪中随机取一支射击，结果中靶。求该枪是已校准过的概率。

贝叶斯公式的应用

- 8支步枪中有5支已校准过，3支未校准。一名射手用校准过的枪射击，中靶概率为0.8；用未校准的枪射击，中靶概率为0.3；现从8支枪中随机取一支射击，结果中靶。求该枪是已校准过的概率。

- 解：

$$P(G=1)=\frac{5}{8} \qquad P(G=0)=\frac{3}{8}$$

$$P(A=1|G=1)=0.8 \qquad P(A=0|G=1)=0.2$$

$$P(A=1|G=0)=0.3 \qquad P(A=0|G=0)=0.7$$

$$P(G=1|A=1)=?$$

$$P(G=1|A=1)=\frac{P(A=1|G=1)P(G=1)}{\sum_{i\in G}P(A=1|G=i)P(G=i)}=\frac{0.8\times\frac{5}{8}}{0.8\times\frac{5}{8}+0.3\times\frac{3}{8}}=0.8163$$

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。

分布

- 复习各种常见分布本身的统计量
- 在复习各种分布的同时，重温积分、Taylor展式等前序知识
- 常见分布是可以完美统一为一类分布

两点分布Bernoulli distribution

0—1分布

已知随机变量 X 的分布律为

X	1	0
p	p	$1-p$

则有 $E(X)=1\cdot p+0\cdot q=p,$

$$D(X)=E(X^2)-[E(X)]^2$$

$$=1^2\cdot p+0^2\cdot(1-p)-p^2=pq.$$

二项分布Binomial Distribution

设随机变量 X 服从参数为 n, p 二项分布,

(法一) 设X_i为第i 次试验中事件 A 发生的次数，$i=1,2,\cdots,n$

则

$$X=\sum_{i=1}^{n} X_i$$

显然，X_i 相互独立均服从参数为 p 的0—1分布，

所以 $$E(X)=\sum_{i=1}^{n} E(X_i)=np.$$

$$D(X)=\sum_{i=1}^{n} D(X_i)=np(1-p).$$

二项分布

(法二) X 的分布律为

$$P\{X=k\}=\binom{n}{k}p^k(1-p)^{n-k},(k=0,1,2,\cdots,n),$$

则有 $E(X)=\sum_{k=0}^{n}k\cdot P\{X=k\}=\sum_{k=0}^{n}k\binom{n}{k}p^k(1-p)^{n-k}$

$$=\sum_{k=0}^{n}\frac{kn!}{k!(n-k)!}p^k(1-p)^{n-k}$$

$$=\sum_{k=1}^{n}\frac{np(n-1)!}{(k-1)![(n-1)-(k-1)]!}p^{k-1}(1-p)^{(n-1)-(k-1)}$$

$$=np\sum_{k=1}^{n}\frac{(n-1)!}{(k-1)![(n-1)-(k-1)]!}p^{k-1}(1-p)^{(n-1)-(k-1)}$$

$$=np[p+(1-p)]^{n-1}=np$$

二项分布

$$E(X^2)=E[X(X-1)+X] \quad =E[X(X-1)]+E(X)$$

$$=\sum_{k=0}^{n}k(k-1)\binom{k}{n}p^k(1-p)^{n-k}+np$$

$$=\sum_{k=0}^{n}\frac{k(k-1)n!}{k!(n-k)!}p^k(1-p)^{n-k}+np$$

$$=n(n-1)p^2\sum_{k=2}^{n}\frac{(n-2)!}{(n-k)!(k-2)!}p^{k-2}(1-p)^{(n-2)-(k-2)}+np$$

$$=n(n-1)p^2[p+(1-p)]^{n-2}+np=(n^2-n)p^2+np.$$

$$D(X)=E(X^2)-[E(X)]^2=(n^2-n)p^2+np-(np)^2$$

$$=np(1-p)$$

负二项分布

- 对于一系列独立的成败实验，每次实验成功的概率恒为p，持续实验直到r次成功(r为正整数)，则总实验次数X的概率为

$$P(X=x;r,p)=C_{x-1}^{r-1}\cdot p^{r}\cdot(1-p)^{x-r}$$

$$x\in[r,r+1,r+2,\cdots\infty)$$

- 若记X=k为失败的次数，则有：

$$P(X=k;r,p)=C_{k+r-1}^{r-1}\cdot p^{r}\cdot(1-p)^{k},\ k\in N$$

应用：我爱乒乓球

- 福原爱与刘诗雯正在乒乓球比赛，若任何一球刘诗雯赢的概率都是60%。则对于11分制的一局，刘诗雯获胜的概率有多大？
  - 为计算简便，暂不考虑分差必须大于等于2
    - 注：如果考虑分差大于等于2，结果相差非常小
  - 0.825622133638/0.836435199842
- 如果考虑“五局三胜制”或“七局四胜制”，则刘诗雯最终获胜的概率有多大？
  - 0.966274558546
  - 0.983505058096

Code

```python
import numpy as np
from scipy import special

if __name__ == '__main__':
    method = 'strict'

    # 1.暴力模拟
    if method == 'simulation':
        p = 0.6
        a, b, c = 0, 0, 0
        t, T = 0, 1000000
        while t < T:
            a = b = 0
            while (a <= 11) and (b <= 11):
                if np.random.uniform() < p:
                    a += 1
                else:
                    b += 1
            if a > b:
                c += 1
            t += 1
        print float(c) / float(T)
    # 2.直接计算
    elif method == 'simple':
        answer = 0
        p = 0.6      # 每分的胜率
        N = 11       # 每局多少分
        for x in np.arange(N):  # x为对手得分
            answer += special.comb(N + x - 1, x) * ((1-p) ** x) * (p ** N)
        print answer
    # 3.严格计算
    else:
        answer = 0
        p = 0.6  # 每分的胜率
        N = 11  # 每局多少分
        for x in np.arange(N-1):  # x为对手得分：11:9  11:8  11:7  11:6...
            answer += special.comb(N + x - 1, x) * ((1 - p) ** x) * (p ** N)
        p10 = special.comb(2*(N-1), N-1) * ((1-p)*p) ** (N-1)   # 10:10的概率
        t = 0
        for n in np.arange(100):    # {XO}(0,)|OO   思考：可以如何简化？
            t += (2*p*(1-p)) ** n * p * p
        answer += p10 * t
        print answer
```

考察Taylor展式

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^k}{k!} + R_k$$

$$1 = 1 \cdot e^{-x} + x \cdot e^{-x} + \frac{x^2}{2!} \cdot e^{-x} + \frac{x^3}{3!} \cdot e^{-x} + \cdots + \frac{x^k}{k!} \cdot e^{-x} + R_n \cdot e^{-x}$$

$$\frac{x^k}{k!} \cdot e^{-x} \longrightarrow \frac{\lambda^k}{k!} \cdot e^{-\lambda}$$

泊松分布Poisson distribution

设 $X \sim \pi(\lambda)$, 且分布律为

$$P\{X=k\}=\frac{\lambda^k}{k!}e^{-\lambda}, \quad k=0,1,2,\cdots, \quad \lambda>0.$$

则有

$$E(X)=\sum_{k=0}^{\infty}k\cdot\frac{\lambda^k}{k!}e^{-\lambda}=e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}\cdot\lambda$$

$$=\lambda e^{-\lambda}\cdot e^{\lambda}=\lambda$$

泊松分布

- 在实际事例中，当一个随机事件，以固定的平均瞬时速率λ(或称密度)随机且独立地出现时，那么这个事件在单位时间(面积或体积)内出现的次数或个数就近似地服从泊松分布P(λ)。
  - 某一服务设施在一定时间内到达的人数
  - 电话交换机接到呼叫的次数
  - 汽车站台的候客人数
  - 机器出现的故障数
  - 自然灾害发生的次数
  - 一块产品上的缺陷数
  - 显微镜下单位分区内的细菌分布数
  - 某放射性物质单位时间发射出的粒子数

泊松分布

$$E(X^2)=E[X(X-1)+X]$$

$$=E[X(X-1)]+E(X)$$

$$=\sum_{k=0}^{+\infty}k(k-1)\cdot\frac{\lambda^k}{k!}e^{-\lambda}+\lambda$$

$$=\lambda^2e^{-\lambda}\sum_{k=2}^{+\infty}\cdot\frac{\lambda^{k-2}}{(k-2)!}+\lambda=\lambda^2e^{-\lambda}e^{\lambda}+\lambda=\lambda^2+\lambda.$$

所以 $D(X)=E(X^2)-[E(X)]^2=\lambda^2+\lambda-\lambda^2=\lambda$

泊松分布的期望和方差都等于参数λ.

均匀分布Uniform Distribution

设 $X \sim U(a,b)$, 其概率密度为

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a<x<b,\\ 0, & \text{其他.}\end{cases}$$

则有 $E(X)=\int_{-\infty}^{\infty} xf(x)\mathrm{d}x=\int_{a}^{b}\dfrac{1}{b-a}x\mathrm{d}x=\dfrac{1}{2}(a+b).$

$$D(X)=E(X^2)-[E(X)]^2$$

$$=\int_a^b x^2\frac{1}{b-a}\mathrm{d}x-\left(\frac{a+b}{2}\right)^2=\frac{(b-a)^2}{12}$$

指数分布Exponential Distribution

设随机变量 X 服从指数分布，其概率密度为

$$f(x)=\begin{cases}\dfrac{1}{\theta}\mathrm{e}^{-x/\theta}, & x>0,\\ 0, & x\le 0.\end{cases}\quad \text{其中}\ \theta>0.$$

则有

$$E(X)=\int_{-\infty}^{+\infty}xf(x)\,\mathrm{d}x=\int_{0}^{+\infty}x\cdot\frac{1}{\theta}\mathrm{e}^{-x/\theta}\,\mathrm{d}x$$

$$=-x\mathrm{e}^{-x/\theta}\Big|_{0}^{+\infty}+\int_{0}^{+\infty}\mathrm{e}^{-x/\theta}\,\mathrm{d}x\quad=\theta$$

$$D(X)=E(X^2)-[E(X)]^2=\int_{0}^{+\infty}x^2\cdot\frac{1}{\theta}\mathrm{e}^{-x/\theta}\,\mathrm{d}x-\theta^2$$

$$=2\theta^2-\theta^2\quad=\theta^2$$

指数分布

$$f(x)=\begin{cases}\lambda e^{-\lambda x} & x>0\\ 0 & x\le 0\end{cases}$$

- 其中λ＞0是分布的一个参数，常被称为率参数(rate parameter)。即每单位时间内发生某事件的次数。指数分布的区间是[0,∞)。如果一个随机变量X呈指数分布，则可以写作：X~ Exponential(λ)。
- 指数分布可以用来表示独立随机事件发生的时间间隔，比如旅客进机场的时间间隔、软件更新的时间间隔等等。
- 许多电子产品的寿命分布一般服从指数分布。有的系统的寿命分布也可用指数分布来近似。它在可靠性研究中是最常用的一种分布形式。

指数分布的无记忆性

- 指数函数的一个重要特征是无记忆性(遗失记忆性，Memoryless Property)。
  - 如果一个随机变量呈指数分布，当s,t≥0时有：

$$P(x > s+t \mid x > s) = P(x > t)$$

  - 即，如果x是某电器元件的寿命，已知元件使用了s小时，则共使用至少s+t小时的条件概率，与从未使用开始至少使用t小时的概率相等。
- 思考：是否有“半记忆性”？

正态分布Normal/Gaussian distribution

设 $X \sim N(\mu,\sigma^2)$, 其概率密度为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\quad \sigma>0,\quad -\infty<x<+\infty.$$

则有 $$E(X)=\int_{-\infty}^{+\infty}xf(x)\,dx$$

$$=\int_{-\infty}^{+\infty}x\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx.$$

令 $$\frac{x-\mu}{\sigma}=t \Rightarrow x=\mu+\sigma t,$$

正态分布

$$E(X)=\int_{-\infty}^{+\infty}x\cdot\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx$$

$$=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}(\mu+\sigma t)e^{-\frac{t^2}{2}}\,dt$$

$$=\mu\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{t^2}{2}}\,dt+\frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}te^{-\frac{t^2}{2}}\,dt$$

$$=\mu.$$

正态分布

$$D(X)=\int_{-\infty}^{+\infty}(x-\mu)^2 f(x)\,\mathrm{d}x$$

$$=\int_{-\infty}^{+\infty}(x-\mu)^2\cdot\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x.$$

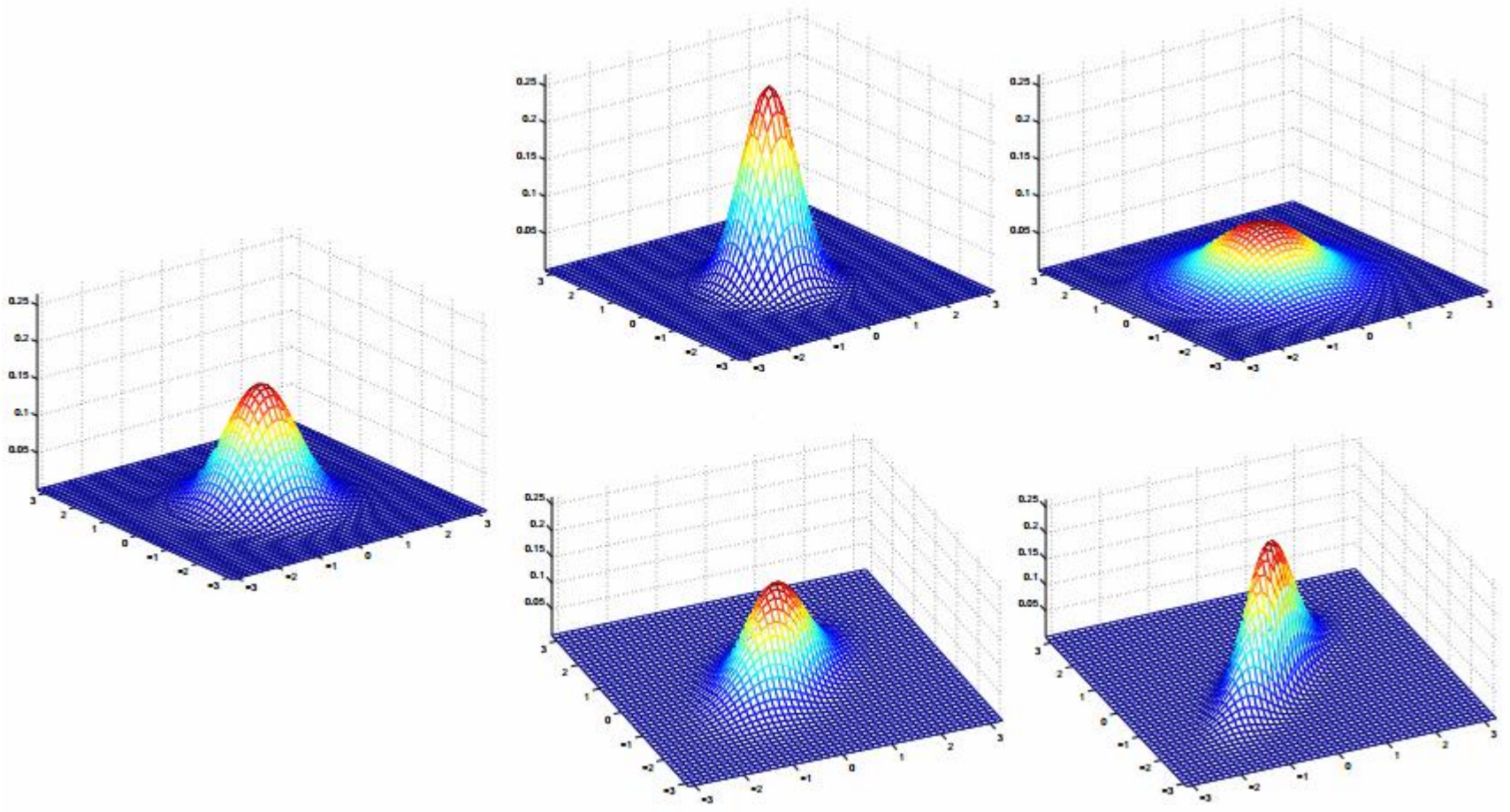
令 $\dfrac{x-\mu}{\sigma}=t$,得

$$D(X)=\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}t^2\mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t$$

$$=\frac{\sigma^2}{\sqrt{2\pi}}\left(-t\mathrm{e}^{-\frac{t^2}{2}}\Big|_{-\infty}^{+\infty}+\int_{-\infty}^{+\infty}\mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t\right)$$

$$=0+\frac{\sigma^2}{\sqrt{2\pi}}\sqrt{2\pi}=\sigma^2.$$

二元正态分布



总结

分　　布	参数	数学期望	方差
两点分布	$0<p<1$	p	$p(1-p)$
二项分布	$n\geq 1,$ $0<p<1$	np	$np(1-p)$
泊松分布	$\lambda>0$	λ	λ
均匀分布	$a<b$	$(a+b)/2$	$(b-a)^2/12$
指数分布	$\theta>0$	θ	θ^2
正态分布	$\mu,\sigma>0$	μ	σ^2

Beta分布

- Beta分布的概率密度：

$$f(x)=\begin{cases}\dfrac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}, & x\in[0,1]\\ 0, & \text{其他}\end{cases}$$

- 其中系数B为：

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- Gamma函数可以看成阶乘的实数域推广：

$$\Gamma(x)=\int_0^\infty t^{x-1}e^{-t}dt$$

$$\Rightarrow\Gamma(n)=(n-1)!\Rightarrow B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Beta分布的期望

$$f(x)=\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1},\ x\in[0,1]$$

$$B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- 根据定义：

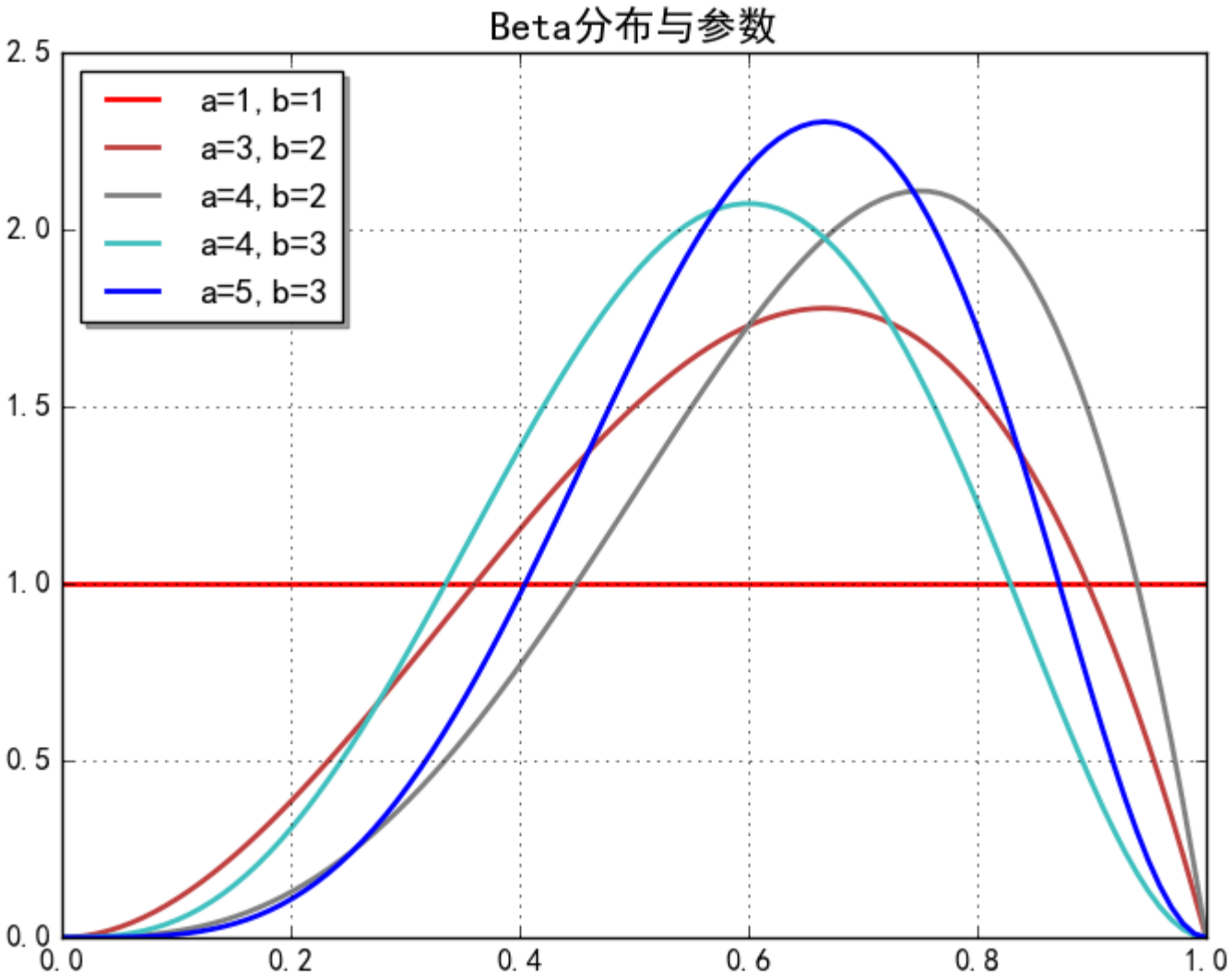
$$E(X)=\int_0^1 x\cdot\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}dx$$

$$=\frac{1}{B(\alpha,\beta)}\int_0^1 x^{(\alpha+1)-1}(1-x)^{\beta-1}dx$$

$$=\frac{B(\alpha+1,\beta)}{B(\alpha,\beta)}=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\Big/\frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)}$$

$$=\frac{\alpha}{\alpha+\beta}$$

Beta分布


Beta分布与参数
a=1, b=1
a=3, b=2
a=4, b=2
a=4, b=3
a=5, b=3
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0

指数族

The exponential family

To work our way up to GLMs, we will begin by defining exponential family distributions. We say that a class of distributions is in the exponential family if it can be written in the form

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta)) \tag{6}$$

Here, η is called the **natural parameter** (also called the **canonical parameter**) of the distribution; $T(y)$ is the **sufficient statistic** (for the distributions we consider, it will often be the case that $T(y) = y$); and $a(\eta)$ is the **log partition function**. The quantity $e^{-a(\eta)}$ essentially plays the role of a normalization constant, that makes sure the distribution $p(y;\eta)$ sums/integrates over y to 1.

A fixed choice of T, a and b defines a *family* (or set) of distributions that is parameterized by η; as we vary η, we then get different distributions within this family.

如：Bernoulli分布和高斯分布

We now show that the Bernoulli and the Gaussian distributions are examples of exponential family distributions. The Bernoulli distribution with mean ϕ, written Bernoulli(ϕ), specifies a distribution over $y \in \{0, 1\}$, so that $p(y = 1; \phi) = \phi$; $p(y = 0; \phi) = 1 - \phi$. As we varying ϕ, we obtain Bernoulli distributions with different means. We now show that this class of Bernoulli distributions, ones obtained by varying ϕ, is in the exponential family; i.e., that there is a choice of T, a and b so that Equation (6) becomes exactly the class of Bernoulli distributions.

Bernoulli分布属于指数族

We write the Bernoulli distribution as:

$$\begin{aligned} p(y;\phi) &= \phi^y(1-\phi)^{1-y} \\ &= \exp(y\log\phi + (1-y)\log(1-\phi)) \\ &= \exp\left(\left(\log\left(\frac{\phi}{1-\phi}\right)\right)y + \log(1-\phi)\right). \end{aligned}$$

Thus, the natural parameter is given by $\eta = \log(\phi/(1-\phi))$. Interestingly, if we invert this definition for η by solving for ϕ in terms of η, we obtain $\phi = 1/(1+e^{-\eta})$. This is the familiar sigmoid function! This will come up again when we derive logistic regression as a GLM. To complete the formulation of the Bernoulli distribution as an exponential family distribution, we also have

$$\begin{aligned} T(y) &= y \\ a(\eta) &= -\log(1-\phi) \\ &= \log(1+e^{\eta}) \\ b(y) &= 1 \end{aligned}$$

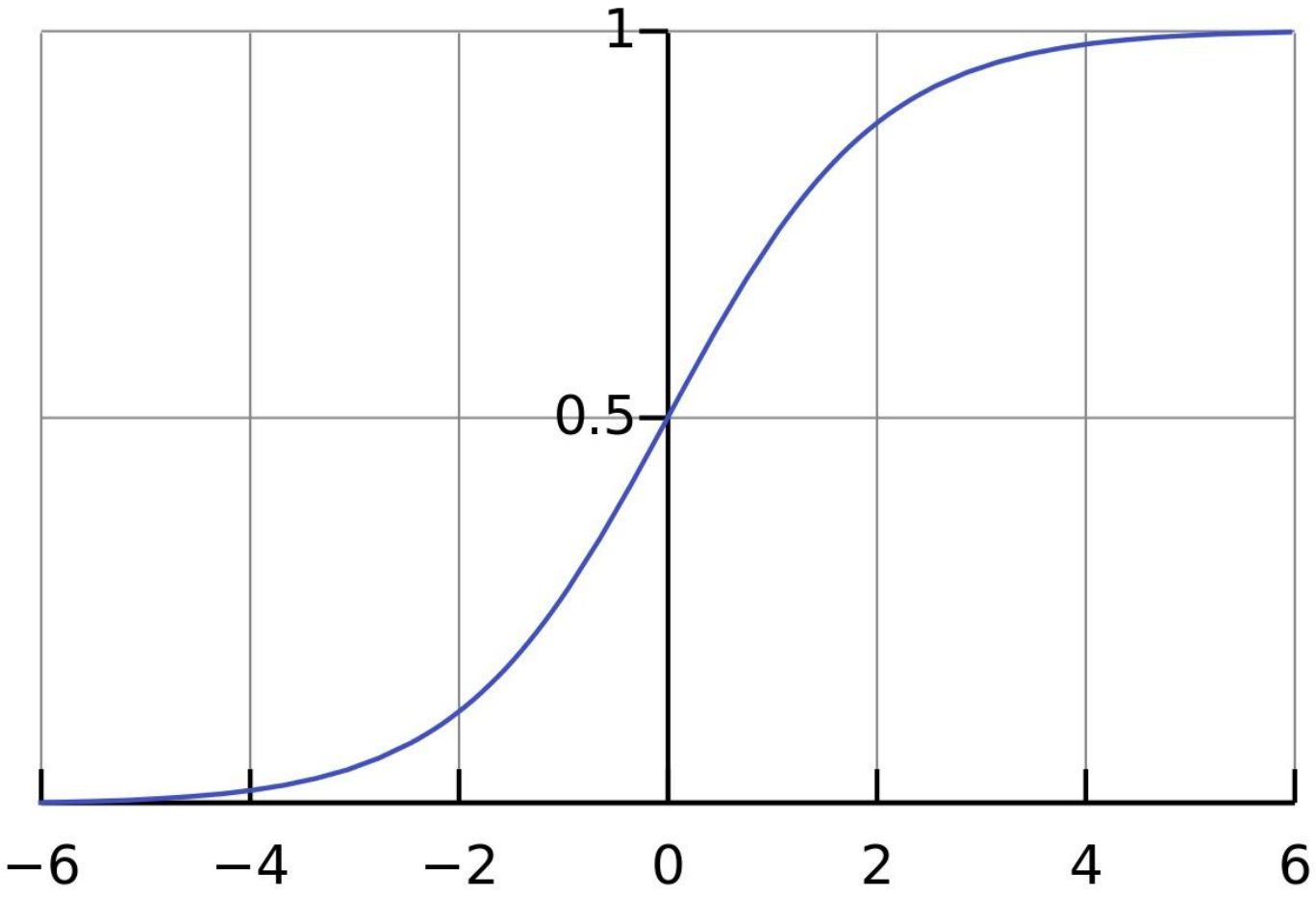
考察参数Φ

- 注意在推导过程中，出现了Logistic方程。

$$\Phi = \frac{1}{1+e^{-\eta}}$$

$$f(x) = \frac{1}{1+e^{-x}}$$

Sigmoid/Logistic函数



Sigmoid函数的导数 $f(x)=\frac{1}{1+e^{-x}}$

$$f'(x)=\left(\frac{1}{1+e^{-x}}\right)'$$

$$=\frac{e^{-x}}{\left(1+e^{-x}\right)^2}$$

$$=\frac{1}{1+e^{-x}}\cdot\frac{e^{-x}}{1+e^{-x}}$$

$$=\frac{1}{1+e^{-x}}\cdot\left(1-\frac{1}{1+e^{-x}}\right)$$

$$=f(x)\cdot(1-f(x))$$

- 该结论后面会用到

Gaussian分布也属于指数族分布

Lets now move on to consider the Gaussian distribution. Recall that, when deriving linear regression, the value of σ^2 had no effect on our final choice of θ and $h_\theta(x)$. Thus, we can choose an arbitrary value for σ^2 without changing anything. To simplify the derivation below, lets set $\sigma^2 = 1$. We then have:

$$\begin{aligned} p(y;\mu) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y-\mu)^2\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right) \cdot \exp\left(\mu y - \frac{1}{2}\mu^2\right) \end{aligned}$$

$$\begin{aligned} \eta &= \mu \\ T(y) &= y \\ a(\eta) &= \mu^2/2 \\ &= \eta^2/2 \\ b(y) &= (1/\sqrt{2\pi}) \exp(-y^2/2) \end{aligned}$$

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 $$E(X)=\int_{-\infty}^{\infty} xf(x)dx$$
- 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。

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：

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

总期望

- 11位二进制数中，每个位取1的期望都是0.5

$$
\begin{aligned}
E(X) &= E\left(\sum_{i=0}^{10}\left(X_i \cdot P\{X_i\}\right)\right) \\
&= E\left(\sum_{i=0}^{10}\left(2^i \cdot P\{X_i=1\}+0\cdot P\{X_i=0\}\right)\right) \\
&= E\left(\sum_{i=0}^{10}\left(2^i \cdot P\{X_i=1\}\right)\right) \\
&= \sum_{i=0}^{10} E\left(2^i \cdot P\{X_i=1\}\right) = \sum_{i=0}^{10} 2^i \cdot E\left(P\{X_i=1\}\right) \\
&= \sum_{i=0}^{10} 2^i \cdot \frac{1}{2} = \frac{1}{2}\sum_{i=0}^{10} 2^i = \frac{(11111111111)_2}{2} \\
&= 1023.5
\end{aligned}
$$

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数(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。

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

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 $Var(X)=E\{[X-E(X)]^2\}=E(X^2)-E^2(X)$
  - $E\left\{[X-E(X)]^2\right\}\geq 0\Rightarrow E(X^2)\geq E^2(X)$，当X为定值时，取等号
- 无条件成立 $Var(c)=0$

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

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

- 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。

试分析该证明过程？

$$Cov^2(X,Y) = E^2((X-E(X))(Y-E(Y))) \quad \cdots\cdots \text{协方差定义}$$

$$\leq E\left((X-E(X))^2(Y-E(Y))^2\right) \quad \cdots\cdots\cdots\cdots\cdots\cdots \text{方差性质}$$

$$\leq E\left((X-E(X))^2\right)E\left((Y-E(Y))^2\right) \quad \cdots\cdots\cdots\cdots \text{期望性质}$$

$$= Var(X)Var(Y) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \text{方差定义}$$

- 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，

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

- 从而：

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

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

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

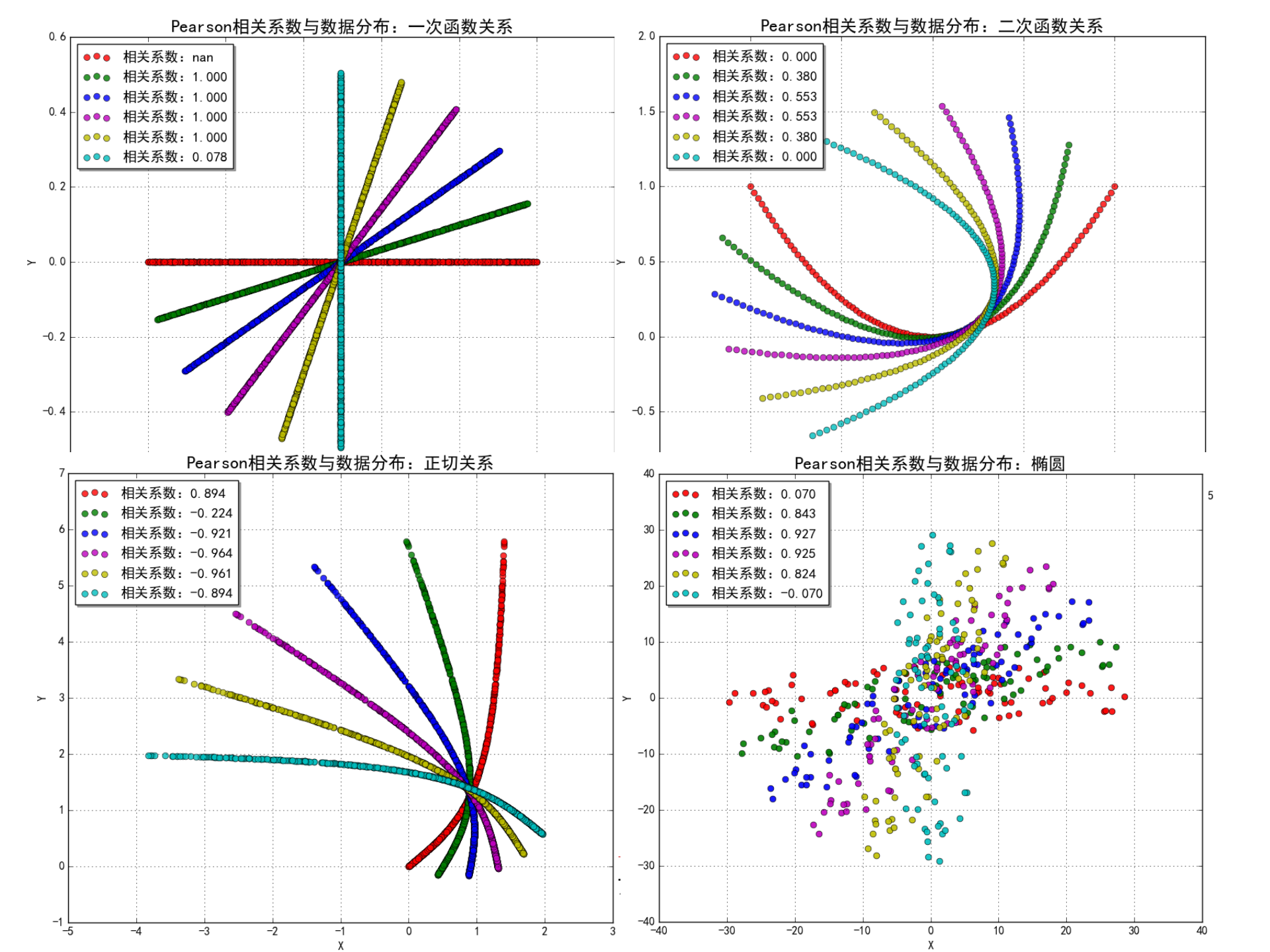
$$\Rightarrow \left|Cov\left(X,Y\right)\right|\leq \sigma_1\sigma_2$$

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。


Pearson相关系数与数据分布：一次函数关系
相关系数：nan
相关系数：1.000
相关系数：1.000
相关系数：1.000
相关系数：1.000
相关系数：0.078
Pearson相关系数与数据分布：二次函数关系
相关系数：0.000
相关系数：0.380
相关系数：0.553
相关系数：0.553
相关系数：0.380
相关系数：0.000
Pearson相关系数与数据分布：正切关系
相关系数：0.894
相关系数：-0.224
相关系数：-0.921
相关系数：-0.964
相关系数：-0.961
相关系数：-0.894
Pearson相关系数与数据分布：椭圆
相关系数：0.070
相关系数：0.843
相关系数：0.927
相关系数：0.925
相关系数：0.824
相关系数：-0.070
X
Y

Code

```
def calc_pearson(x, y):
    std1 = np.std(x)
    # np.sqrt(np.mean(x**2) - np.mean(x)**2)
    std2 = np.std(y)
    cov = np.cov(x, y, bias=True)[0,1]
    return cov / (std1 * std2)


def pearson(x, y, tip):
    clrs = list('rgbmyc')
    plt.figure(figsize=(10, 8), facecolor='w')
    for i, theta in enumerate(np.linspace(0, 90, 6)):
        xr, yr = rotate(x, y, theta)
        p = stats.pearsonr(xr, yr)[0]
        # print calc_pearson(xr, yr)
        print '旋转角度：', theta, 'Pearson相关系数：', p
        str = u'相关系数：%.3f' % p
        plt.scatter(xr, yr, s=40, alpha=0.9, linewidths=0.5, c=clr
    plt.legend(loc='upper left', shadow=True)
    plt.xlabel(u'X')
    plt.ylabel(u'Y')
    plt.title(u'Pearson相关系数与数据分布：%s' % tip, fontsize=18)
    plt.grid(b=True)
    plt.show()
```

协方差矩阵

- 对于n个随机向量(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；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

- 若X、Y独立，则：

$$Var(XY)=Var(X)Var(Y)+Var(X)E^2(Y)+Var(Y)E^2(X)$$

  - 思考：应用？

- 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？

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-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。

大数定律

- 设随机变量$X_1, X_2 \ldots X_n \ldots$互相独立，并且具有相同的期望$\mu$和方差$\sigma^2$。取前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

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

大数定律的意义

- 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。
  - 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$X_1, X_2 \ldots X_n \ldots$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$N(n\mu, n\sigma^2)$

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μ=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：

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

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

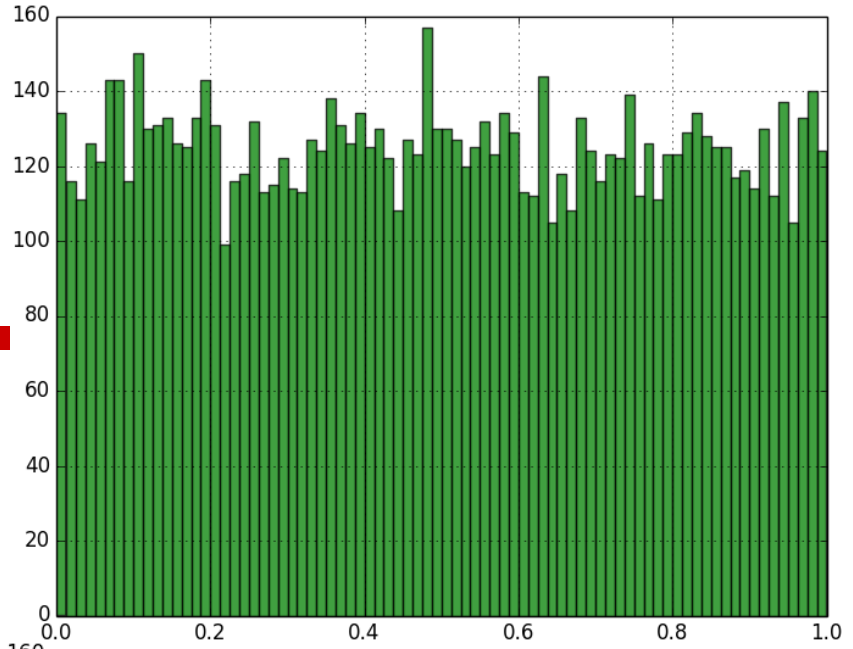
- 上式中，μ=0.01，$\sigma^2=0.0099$，a=0.0099，b=0.0101，α=0.05或0.01(显著性水平的一般取值)，查标准正态分布表，很容易计算得到n的最小值。
  - 注：直接使用二项分布，也能得到结论。

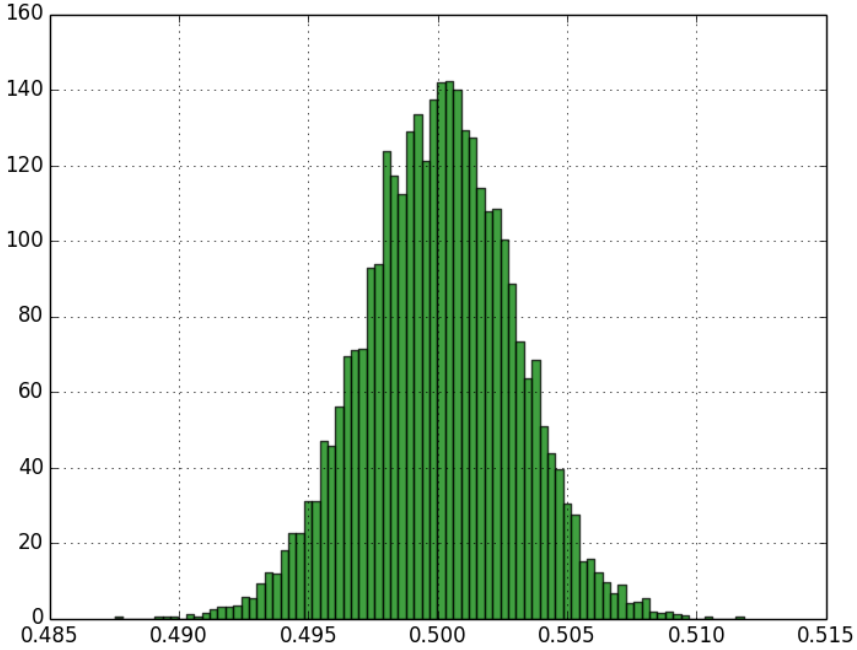
中心极限定理的意义

- 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()
```





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)

$$\max P(A_i|D)=\max\frac{P(D|A_i)P(A_i)}{P(D)}=\max\left(P(D|A_i)P(A_i)\right)\xrightarrow{P(A_i)\text{近似相等}}\max P(D|A_i)$$

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

  - 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.....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：

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

- 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$x_1,x_2 \dots x_n$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

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

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

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：

$$P = pp(1-p)ppp(1-p)(1-p)pp$$
$$= p^7(1-p)^3$$

  - 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：

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

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

正态分布的最大似然估计

- 若给定一组样本$X_1,X_2...X_n$，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期望，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
- 该结论将在期望最大化EM算法、高斯混合模型GMM中将继续使用。

思考：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。

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

- 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：

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

作业：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？

答案：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：μ=25,σ=2，计算P{21<X<29}
- 使用切比雪夫不等式：$P\{|X-\mu|\geq\varepsilon\}\leq\frac{\sigma^2}{\varepsilon^2}$

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

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

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
- Prof. Andrew Ng, *Machine Learning*, Stanford University

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

感谢大家！

恳请大家批评指正！