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

小象学院 ChinaHadoop.cn 邹博

主要内容

- 概率论基础
  - 概率与直观
  - 频率学派与贝叶斯学派
  - 常见概率分布
  - Sigmoid/Logistic函数的引入
- 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

统计数字的概率

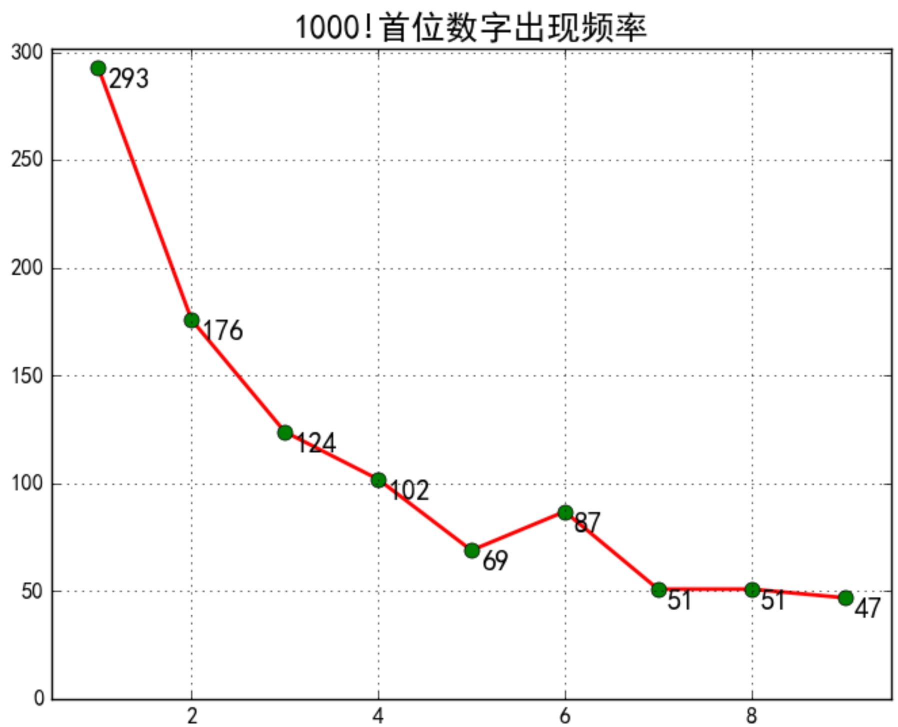
- 给定某正整数N，统计从1到N!的所有数中，首位数字出现1的概率。
- 进而，可以计算首位数字是2的概率，是3的概率，从而得到一条“九点分布”。

Code

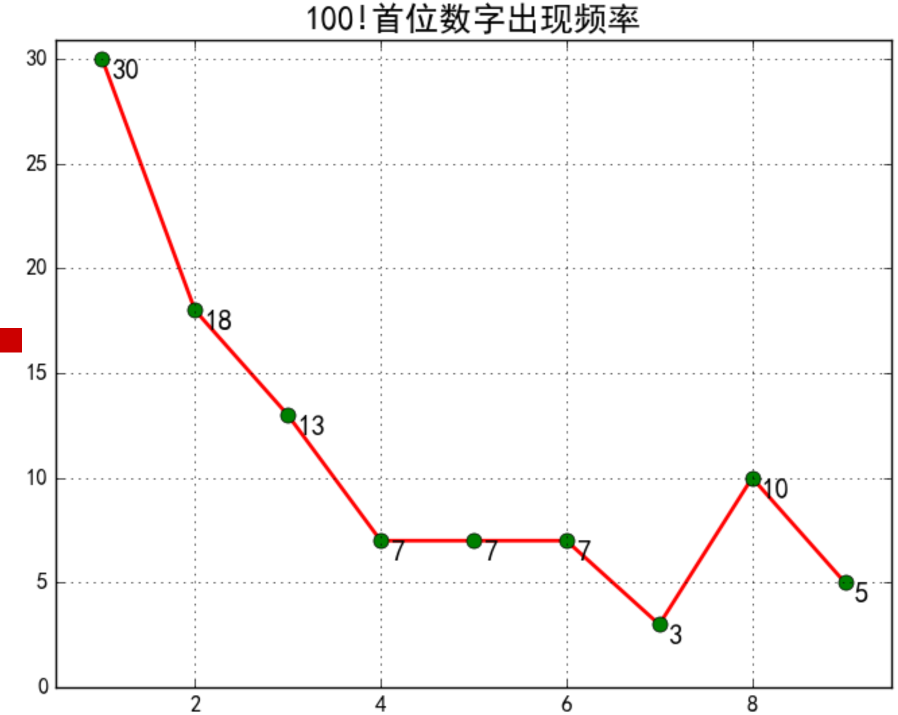
```python
def first_digital(x):
    while x >= 10:
        x /= 10
    return x


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()
```

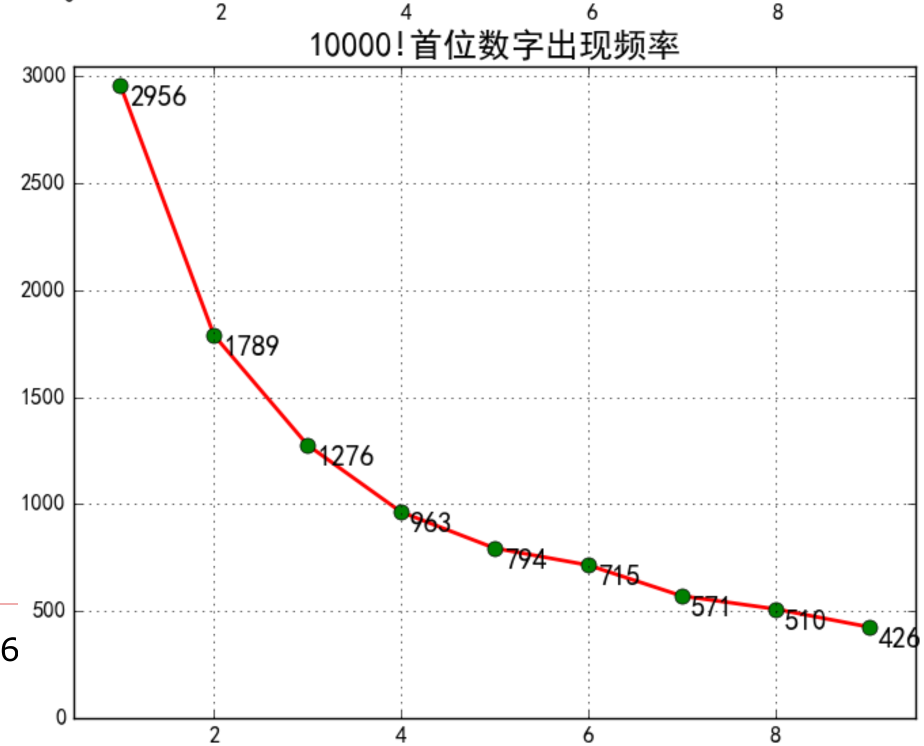
数字的概率



1000!首位数字出现频率
300
250
200
150
100
50
0
2
4
6
8
293
176
124
102
69
87
51
51
47



100!首位数字出现频率
30
25
20
15
10
5
0
2
4
6
8
30
18
13
7
7
7
3
10
5



10000!首位数字出现频率
3000
2500
2000
1500
1000
500
0
2
4
6
8
2956
1789
1276
963
794
715
571
510
426

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

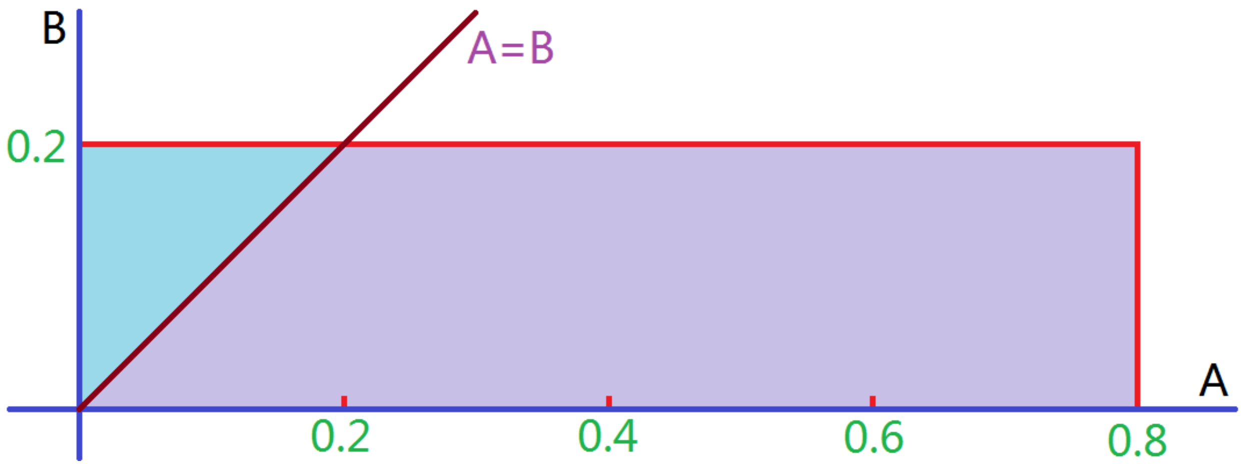
数字	出现概率
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%

商品推荐

- 商品推荐场景中过于聚焦的商品推荐往往会损害用户的购物体验，在有些场景中，系统会通过一定程度的随机性给用户带来发现的惊喜感。
- 假设在某推荐场景中，经计算A和B两个商品与当前访问用户的匹配度分别为0.8分和0.2分，系统将随机为A生成一个均匀分布于0到0.8的最终得分，为B生成一个均匀分布于0到0.2的最终得分，试计算最终B的分数大于A的分数的概率。

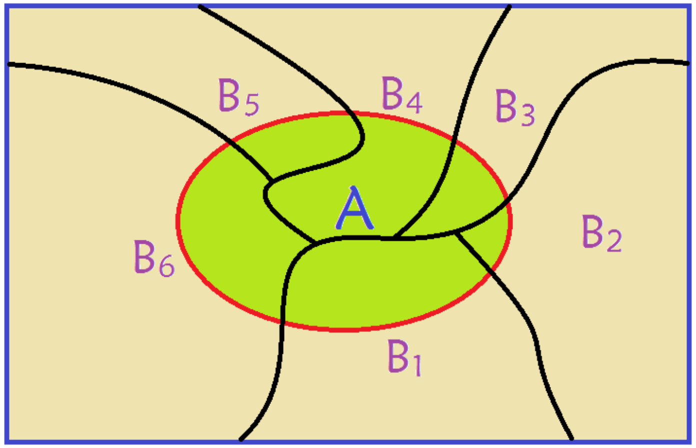
商品推荐

- A=B的直线上方区域，即为B>A的情况。
- $S_{蓝色}=0.02$　　$S_{矩形}=0.16$
- $p=0.02/0.16=0.125$



概率公式

- 条件概率：$P(A|B)=\dfrac{P(AB)}{P(B)}$
- 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。

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

- 解：

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

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

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

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

两种认识下的两个学派

- 给定某系统的若干样本，求该系统的参数。
- 矩估计/MLE/MaxEnt/EM等：
  - 假定参数是某个/某些未知的定值，求这些参数如何取值，能够使得某目标函数取极大/极小。
  - 频率学派
- 贝叶斯模型：
  - 假定参数本身是变化的，服从某个分布。求在这个分布约束下使得某目标函数极大/极小。
  - 贝叶斯学派

频率学派和贝叶斯学派

- 无高低好坏之分，只是认识自然的手段。只是在当前人们掌握的数学工具和需解决的实践问题中，贝叶斯学派的理论体系往往能够比较好的解释目标函数、分析相互关系等。
  - 前半段的内容，大多是频率学派的思想；后半段的内容，使用贝叶斯学派的观点。
- 思考：大数据
  - 频率学派对于贝叶斯学派一次强有力逆袭。

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

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 Bernoulli 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

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

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

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

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

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

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

$$=np(1-p)$$

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

设 $X \sim \pi(\lambda)$, 且分布律为

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

则有

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

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

泊松分布Poisson distribution

- 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

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

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

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

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

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

设 $X \sim U(a,b)$, 其概率密度为

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

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

设随机变量 X 服从指数分布，其概率密度为

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

则有

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

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

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

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

指数分布

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

- 其中$\lambda>0$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$[0,\infty)$。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：

$$P\left(x > s + t \middle| x > s\right) = P\left(x > t\right)$$

  - 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

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

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

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

正态分布

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

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

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

$$=\mu.$$

正态分布

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

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

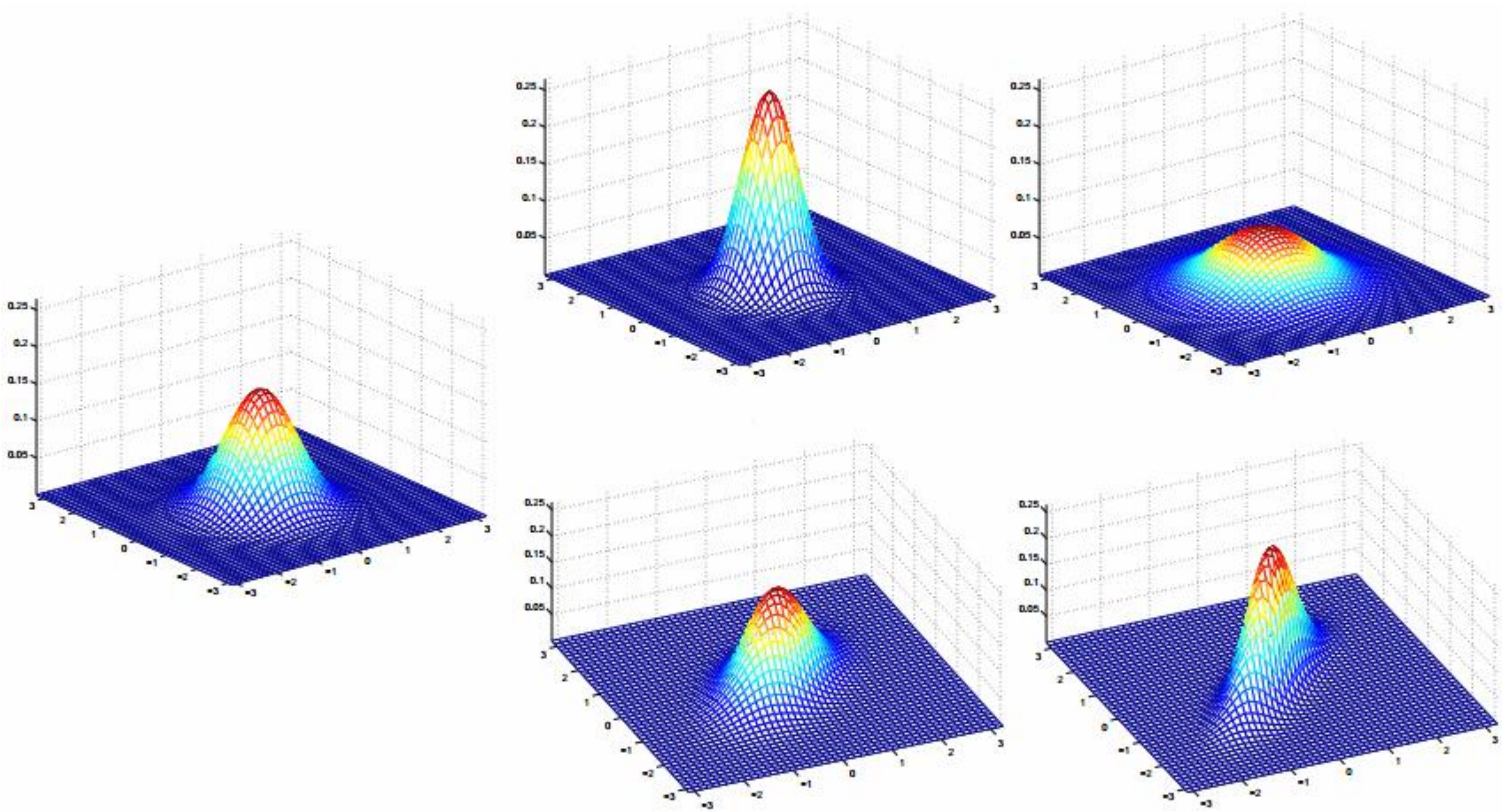
令 $\dfrac{x-\mu}{\sigma}=t$,得

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

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

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

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：

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

- 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：

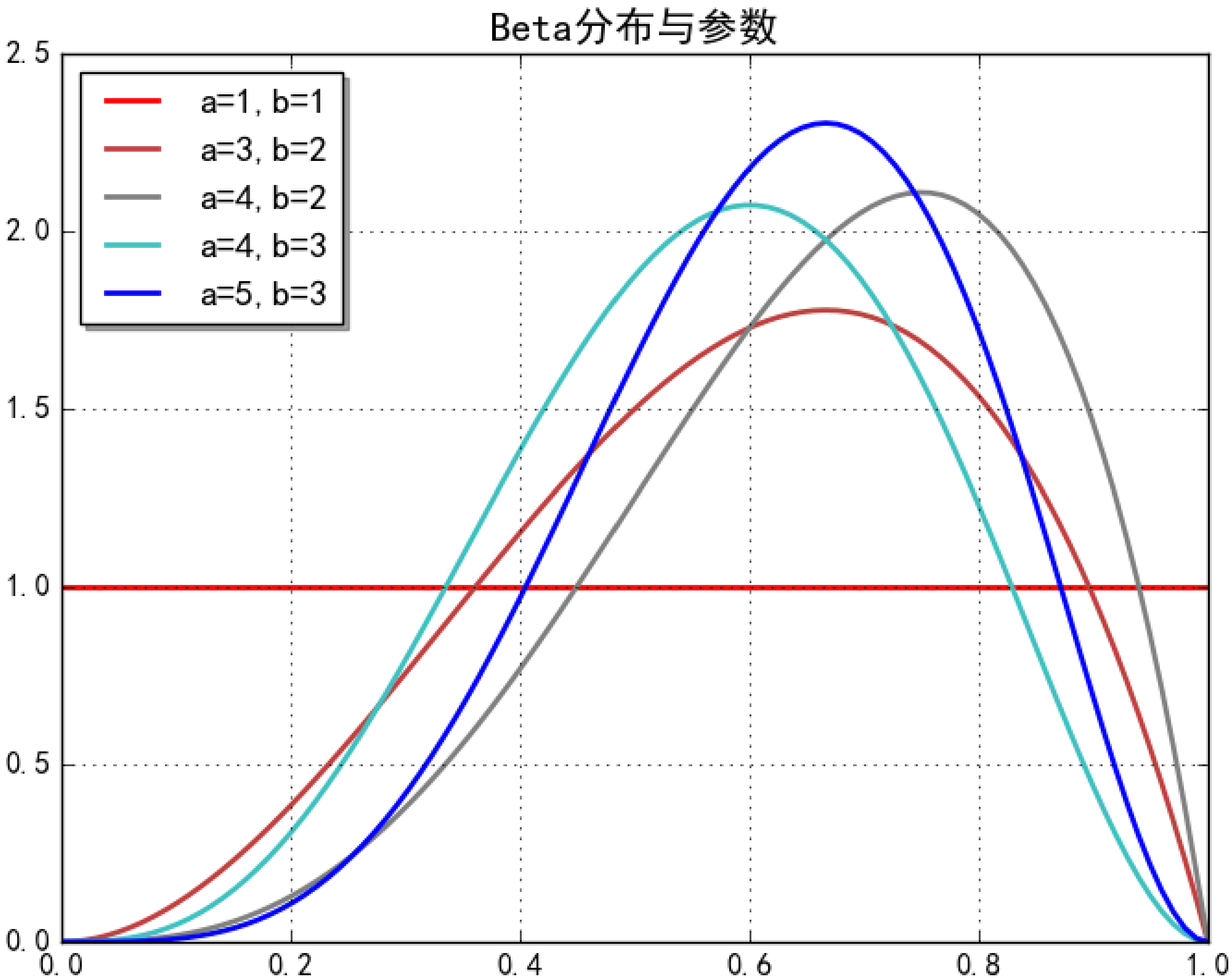
$$E(X)=\int_0^1 x\cdot\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}dx$$

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

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

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

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

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

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

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

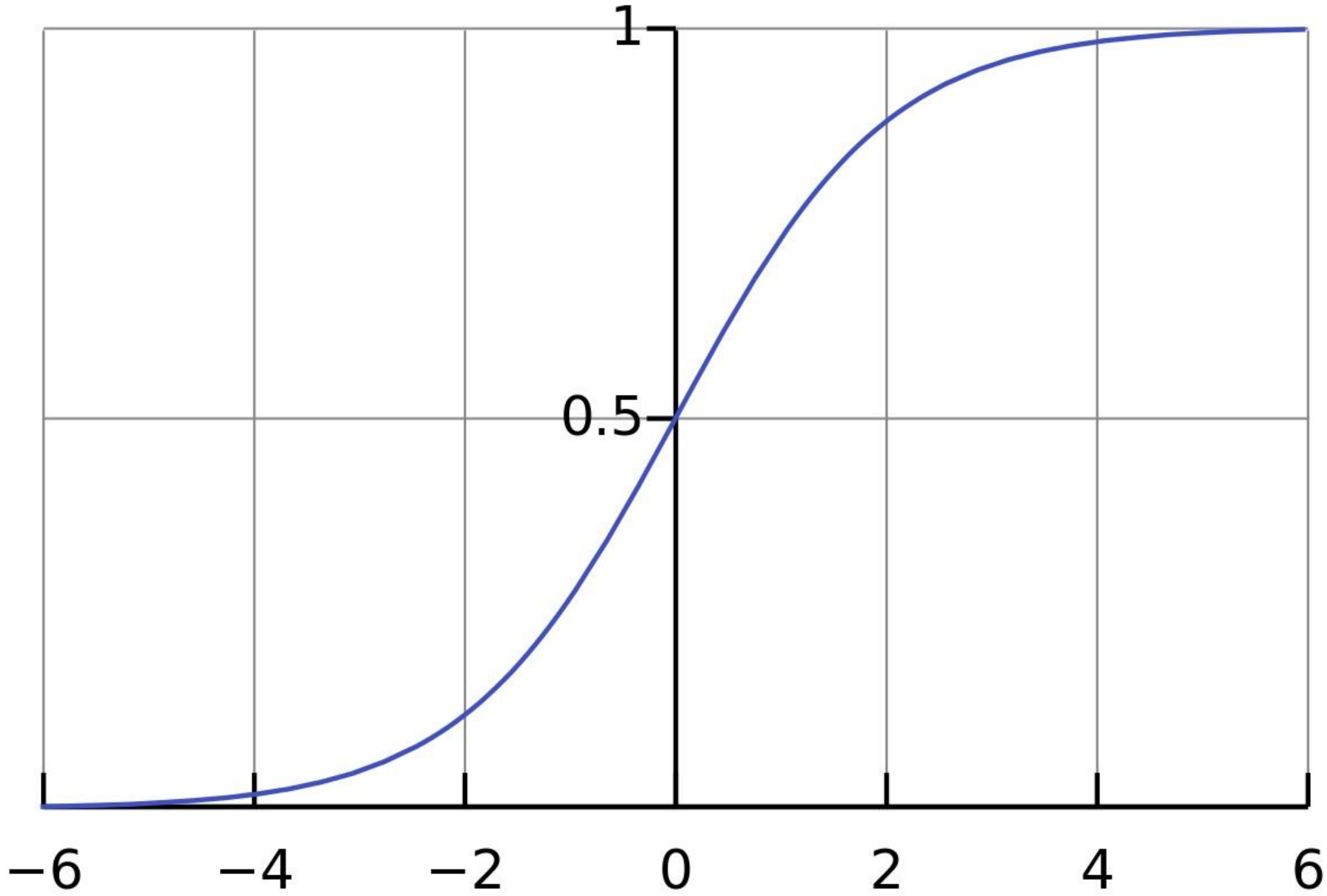
考察参数Φ

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

$$
\begin{aligned}
f'(x) &= \left(\frac{1}{1+e^{-x}}\right)' \\
&= \frac{e^{-x}}{\left(1+e^{-x}\right)^2} \\
&= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} \\
&= \frac{1}{1+e^{-x}} \cdot \left(1-\frac{1}{1+e^{-x}}\right) \\
&= f(x)\cdot\left(1-f(x)\right)
\end{aligned}
$$

- 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:

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

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

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：

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

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

```cpp
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;
}
```

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

- 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：

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

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

- 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 $Var(c)=0$

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

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

- 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？

$$Cov^2(X,Y)=E^2\left(\left(X-E(X)\right)\left(Y-E(Y)\right)\right)$$ ……协方差定义

$$\le E\left(\left(X-E(X)\right)^2\left(Y-E(Y)\right)^2\right)$$ ……………………方差性质

$$\le E\left(\left(X-E(X)\right)^2\right)E\left(\left(Y-E(Y)\right)^2\right)$$ ………………期望性质

$$=Var(X)Var(Y)$$ ………………………………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，

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

- 从而：

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

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

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

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

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，$|\rho| \le 1$
- 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…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？

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

感谢大家！

恳请大家批评指正！