



北京航空航天大學
BEIHANG UNIVERSITY

自然语言处理

人工智能研究院

主讲教师 沙磊



第一课

预备知识

基本概念(1)

- 概率论
 - 研究随机现象的统计规律性的一门科学。
 - 现象有确定性的、有些是随机性的。
 - 大量实验才能发现一定的规律性
- 随机现象
 - 一定条件下，并不总是出现相同结果的现象
- 随机试验：
 - 可以在相同条件下重复进行；
 - 事先能够知道实验的所有可能结果；
 - 事先不知道哪个结果会出现。

基本概念(2)

- 样本空间：随机试验所有可能基本结果组成的集合
 - 抛硬币：{正面, 反面}
 - 掷骰子：{1,2,3,4,5,6}
 - 一天内进入超市的人数：
 - 电视的使用寿命：
- 随机事件：某些基本结果组成的集合
 - 掷骰子出现奇数点：{1,3,5}
- 必然事件： Ω 例：“出现点数不超过6”
- 不可能事件： Φ 例：“出现点数为7”

随机事件示例

例 1.1.3 抛两枚硬币的基本空间 Ω 由下列四个基本结果组成：

$$\omega_1 = (\text{正}, \text{正}) \quad \omega_2 = (\text{正}, \text{反})$$

$$\omega_3 = (\text{反}, \text{正}) \quad \omega_4 = (\text{反}, \text{反})$$

下面几个事件可用集合形式表示,也可用语言形式表示。

$$A = \text{“至少出现一个正面”} = \{\omega_1, \omega_2, \omega_3\}$$

$$B = \text{“最多出现一个正面”} = \{\omega_2, \omega_3, \omega_4\}$$

$$C = \text{“恰好出现一个正面”} = \{\omega_2, \omega_3\}$$

$$D = \text{“出现二面相同”} = \{\omega_1, \omega_4\}$$

例 1.1.4 掷一颗骰子,“出现 6 点”、“出现偶数点”、“出现点数不超过 2”、“出现点数不等于 3”都是事件,若依次记为 A, B, C, D , 那它们都可以用其基本空间 $\Omega = \{1, 2, 3, 4, 5, 6\}$ 的某个子集表示。

$$A = \{6\} \quad B = \{2, 4, 6\}$$

$$C = \{1, 2\} \quad D = \{1, 2, 4, 5, 6\}$$

事件的概率

- $P(A)$: 表示随机事件A发生可能性大小的度量
 - 非负性: >0
 - 正则性: $P(\Omega)=1$
 - 可加性: 互不相容的事件 $P(A_1 \cup A_2)=P(A_1)+P(A_2)$
- 概率的确定
 - 古典方法
 - 有限个基本结果、可能性相同、不需要实验
 - 频率方法
 - 大量重复实验中用频率去获得概率近似值
 - 主观方法
 - 根据经验对该事件可能性所给出的个人信念

例子

例 1.3.2 一批产品共 100 件,其中有 5 件不合格品,现从中随机抽出 10 件,其中最多有 2 件不合格品的概率是多少?

$$P(A_i) = \frac{\binom{5}{i} \binom{95}{10-i}}{\binom{100}{10}}, \quad i=0,1,2$$

表 1.2.2 掷一枚硬币,正面出现的频率

实验者	掷硬币次数	正面出现次数	频率
蒲来	4 040	2 048	0.5069
皮尔逊	12 000	6 019	0.5016
皮尔逊	24 000	12 012	0.5005

注:此表引自赫涅·尼科《概率论教程》,高盛教育出版社 1957

例子

一个企业家认为“一项新产品在未来市场上畅销”的概率是 0.8。这里的 0.8 是根据他自己多年经验和当时的一些市场信息综合而成的个人信念。

一位投资者认为“购买某种股票能获得高收益”的概率是 0.6，这里的 0.6 是投资者根据自己多年股票生意经验和当时股票行情综合而成的个人信念。

一位脑外科医生要对一位病人动手术，他认为成功的概率是 0.9，这是他根据手术的难易程度和自己的手术经验而对“手术成功”所给出的把握程度。

独立性

定义 1.4.1 对任意两个事件 A 与 B ,若有 $P(AB)=P(A)P(B)$,则称事件 A 与 B 相互独立,简称 A 与 B 独立。否则称事件 A 与 B 不独立。

A =“第一颗骰子出现 1 点”

B =“第二颗骰子出现偶数点”

经验事实告诉我们,第一颗骰子出现的点数不会影响第二颗骰子出现的点数,假如规定第二颗骰子出现偶数点可得奖,那末不管第一颗骰子出现什么点都不会影响你得奖的机会,这时就可以说:事件 A 与 B 独立。

从概率角度看,两个事件之间的独立性与这两个事件同时发生的概率有密切关系,譬如在上面掷两颗骰子的试验中,事件 A 与 B 的概率分别是 $P(A)=1/6$, $P(B)=1/2$,而这两个事件同时发生 AB 含有三个基本结果:(1,2),(1,4),(1,6),故 $P(AB)=3/36=1/12$,于是有等式 $P(AB)=P(A)P(B)$ 。这不是偶然的,而是两独立事件的共同特征,即两独立事件同时发生的概率等于它们各自概率的乘积,这就引出两事件独立的一般定义。

独立性

定义 1.4.2 设有 n 个事件 A_1, A_2, \dots, A_n , 假如对所有可能的 $1 \leq i < j < k < \dots < n$, 以下等式均成立

$$P(A_i A_j) = P(A_i)P(A_j)$$

$$P(A_i A_j A_k) = P(A_i)P(A_j)P(A_k)$$

⋮

$$P(A_1 A_2 \dots A_n) = P(A_1)P(A_2) \dots P(A_n)$$

则称此 n 个事件 A_1, A_2, \dots, A_n 相互独立。

试验的独立性

- 设有两个试验 E_1 和 E_2 ，假如试验 E_1 的任一个结果（事件）与试验 E_2 的任一个结果（事件）都是相互独立事件，则称两个试验相互独立。
 - 掷一枚硬币和一颗骰子
- 类似可以定义n个试验 $E_1, E_2 \dots E_n$ 的相互独立性，则称试验 $E_1, E_2 \dots E_n$ 相互独立。
- 假如n个试验是相同的，则称为n重独立重复试验。例如掷n枚硬币，n颗骰子。

贝努里试验

- 只有两个结果（成功与失败， A 与 \bar{A} ）
 - 抛一枚硬币（正与反）
 - 检查产品（合格与不合格）
 - 一次射击打靶（命中与不命中）
 - $P(A) = p$ 与 $P(\bar{A}) = 1-p$

(3) **n 重贝努里试验** 由 n 个(次)相同的、独立的贝努里试验组成的随机试验称为 n 重贝努里试验。譬如,抛 3 枚硬币(或一硬币抛 3 次)、检查 7 个产品、打 10 次靶、诞生 100 个婴儿等都是多重贝努里试验。

贝努里试验

在 n 重贝努里试验中, 人们最关心的是成功次数(或 A 的个数)。因为成功次数是基本结果中所含的最重要信息, 而 A 与 \bar{A} 的排列次序在实际中往往 是不感兴趣的信息。若记

$B_{n,k}$ = “ n 重贝努里试验中 A 出现 k 次”

$$P(B_{n,k}) = \binom{n}{k} p^k (1-p)^{n-k}$$

条件概率

定义 1.5.1 设 A 与 B 是基本空间 Ω 中的两个事件,且 $P(B) > 0$,在事件 B 已发生的条件下,事件 A 的条件概率 $P(A|B)$ 定义为 $P(AB)/P(B)$,即

$$P(A|B) = \frac{P(AB)}{P(B)}$$

其中 $P(A|B)$ 也称为给定事件 B 下事件 A 的条件概率。

条件概率也是概率,满足以下三条:

(1) 非负性: $P(A|B) \geq 0$

(2) 正则性: $P(\Omega|B) = 1$

(3) 可加性: 假如事件 A_1 与 A_2 互不相容,且 $P(B) > 0$,则

$$P(A_1 \cup A_2 | B) = P(A_1 | B) + P(A_2 | B)$$

乘法公式

定理 1.5.1(乘法公式) 对任意两个事件 A 与 B , 有

$$P(AB) = P(A|B)P(B) = P(B|A)P(A)$$

其中第一个等式成立要求 $P(B) > 0$, 第二个等式成立要求 $P(A) > 0$.

定理 1.5.2 假如事件 A 与 B 独立, 且 $P(B) > 0$, 则 $P(A|B) = P(A)$, 反之亦然.

定理 1.5.3(一般乘法公式) 对任意三个事件 A_1 、 A_2 和 A_3 , 有

$$P(A_1 A_2 A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1 A_2)$$

其中 $P(A_1 A_2) > 0$.

全概率公式

- 设A与B为任意两个事件，假如 $0 < P(B) < 1$ ，则

$$P(A) = P(B)P(A|B) + P(\bar{B})P(A|\bar{B})$$

例 1.5.6 设在n张彩票中有一张奖券。求第二人摸到奖券的概率是多少？

解：设 A_i 表示第*i*人摸到奖券

$$\begin{aligned}P(A_2) &= P(A_1)P(A_2|A_1) + P(\bar{A}_1)P(A_2|\bar{A}_1) \\&= \frac{1}{n} \cdot 0 + \frac{n-1}{n} \cdot \frac{1}{n-1} = \frac{1}{n}\end{aligned}$$

全概率公式的一般形式

定义 1.5.2 把基本空间 Ω 分为 n 个事件 B_1, B_2, \dots, B_n (见图 1.5.2), 假如

(1) $P(B_i) > 0, i = 1, 2, \dots, n$

(2) B_1, B_2, \dots, B_n 互不相容

(3) $\bigcup_{i=1}^n B_i = \Omega$

则称事件组 B_1, B_2, \dots, B_n 为基本空间 Ω 的一个分割。

定理 1.5.5 设 B_1, B_2, \dots, B_n 是基本空间 Ω 的一个分割, 则对 Ω 中任一事件 A , 有

$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i)$$

贝叶斯公式

定理 1.5.6(贝叶斯公式) 设事件 B_1, B_2, \dots, B_n 是基本空间 Ω 的一个分割, 且它们各自概率 $P(B_1), P(B_2), \dots, P(B_n)$ 皆已知且为正, 又设 A 是 Ω 中的一个事件, $P(A) > 0$, 且在诸 B_i 给定下事件 A 的条件概率 $P(A|B_1), P(A|B_2), \dots, P(A|B_n)$ 可通过试验等手段获得, 则在 A 给定下, 事件 B_k 的条件概率为

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i=1}^n P(A|B_i)P(B_i)}, \quad k=1, 2, \dots, n$$

理解贝叶斯公式

例 1.5.11 伊索寓言“孩子与狼”讲的是一个小孩每天到山上放羊，山里有狼出没，第一天他在山上喊：“狼来了！狼来了！”山下的村民们闻声便去打狼，可到了山上，发现狼没有来；第二天仍是如此；第三天，狼真的来了，可无论小孩怎么喊叫，也没有人来救他，因为前两次他说了谎话，人们不再相信他了。

《概率论与数理统计》：P64 1.5.11

《MIT Foundation》：P44 example 2

随机变量

- 随机变量：从样本空间到实数域的映射（把随机试验的结果和实数联系起来）

定义 2.1.1 假如一个变量在数轴上的取值依赖于随机现象的基本结果，则称此变量为**随机变量**，常用大写字母 X, Y, Z 等表示，其取值用小写字母 x, y, z 等表示。假如一个随机变量仅取数轴上的有限个或可列个孤立点（见图 2.1.1），则称此随机变量为**离散随机变量**。假如一个随机变量的可能取值充满数轴上的一个区间 (a, b) （见图 2.1.2），则称此随机变量为**连续随机变量**，其中 a 可以是 $-\infty$, b 可以是 $+\infty$ 。

随机变量的概率分布

定义 2.1.2 设 X 为一个随机变量, 对任意实数 x , 事件“ $X \leq x$ ”的概率是 x 的函数, 记为

$$F(x) = P(X \leq x)$$

这个函数称为 X 的累积概率分布函数, 简称分布函数。

- For every random variable X , **Cumulative distribution function** (cdf), denoted by $F_X(x)$ is defined by $F_X(x) = P_X(X \leq x)$
- A random variable X is **continuous** if $F_X(x)$ is a continuous function of x . A random variable X is **discrete** if $F_X(x)$ is a step function of x .

Density and Mass functions

- Probability mass function (**pmf**) of a **discrete** random variable X is given by $f_X(x)=P(X=x)$
- Probability density function (**pdf**) of a **continuous** random variable X is the function that satisfies

$$F_X(x) = \int_{-\infty}^x f_X(t)dt$$

- Both pdfs and pmfs are concerned with “**point probabilities**” of random variables.

数学期望和方差

> mean(x)
> var(x)
> sd(x)

$$E(X) = \begin{cases} \sum_i x_i p(x), & \text{在离散场合} \\ \int_{-\infty}^{\infty} x p(x) dx, & \text{在连续场合} \end{cases}$$

定义 2.4.1 设随机变量 X 的 EX^2 存在, 则称偏差平方的数学期望 $E[X - E(X)]^2$ 为随机变量 X (或相应分布) 的**方差**, 记为

$$\text{Var}(X) = E[X - E(X)]^2 \quad (2.4.3)$$

方差的正平方根 $[\text{Var}(X)]^{1/2}$ 称为随机变量 X (或相应分布) 的**标准差**, 记为 σ_X 或 $\sigma(X)$.

Discrete Distributions

- Discrete Uniform Distribution
- Binomial Distribution
- Poisson Distribution
- Hypergeometric Distribution
- Negative Binomial Distribution
- Geometric Distribution

Discrete Uniform Distribution

- A random variable X has a discrete uniform (1,N) if

$$P(X = x \mid N) = \frac{1}{N}, x = 1, 2, \dots, N,$$

$$EX = \frac{N+1}{2}$$

$$VarX = EX^2 - (EX)^2 = \frac{(N+1)(N-1)}{12}$$

```
> plot(function(x)dunif(x,min=0,max=1),0,1,main="Uniform")
```

Binomial Distribution

Bernoulli trial:

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases} \quad (0 \leq p \leq 1)$$

$$EX = p$$

$$VarX = p(1-p)$$

Binomial(n,p) variable: $Y = \sum_{i=1}^n X_i$

$$P(Y = y | n, p) = \binom{n}{y} p^y (1-p)^{n-y}$$

$$EX = np$$

$$VarX = np(1-p)$$

Binomial Distribution: example

- Throw a pair of dice 24 times and ask for the probability of at least one double 6.
 - $p=1/36$ (role a double 6)
 - $Y = \text{number of double 6s in 24 rolls}$ $Y \sim \text{binomial}(24, 1/36)$
 - $P(\text{at least one double 6}) = P(Y > 0) = 1 - P(Y = 0) = 0.491$

```
> plot(dbinom(c(0:24),24,1/36),main="Binomial")
```

Poisson Distribution

$$P(X = x | \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}, x = 0, 1, \dots$$

- The Poisson distribution is a widely applied discrete distribution and can serve as a model for a number of different types of experiments.
 - a phenomenon waiting for an occurrence (such as waiting for bus, waiting for customers to arrive in a bank)
 - the number of occurrences in a given time interval can be modeled by the Poisson distribution.
 - One basic assumption is that, for small time intervals, the probability of an arrival is proportional to the length of waiting time.
 - It makes sense to assume that the longer we wait, the more likely it is that a customer will enter the bank.

Poisson Distribution: example

- An example of a waiting-for-occurrence application, consider a telephone operator who, on the average, handles five calls every 3 minutes. What is the probability that there will be no calls in the next minute? At least two calls?
 - If we let X =number of calls in a minute, then X has a Poisson distribution with $EX=\lambda=5/3$.
 - $P(\text{no calls in the next minute}) = P(X=0) = 0.189$
 - $P(\text{at least two calls in the next minute}) = P(X \geq 2) = 1 - P(X=0) - P(X=1) = 0.496$

```
> y <- dpois(0, 5/3)
```

```
> z <- 1 - dpois(0, 5/3) - dpois(1, 5/3)
```

Continuous Distributions

- Uniform Distribution
- Normal Distribution
- Gamma Distribution
- Beta Distribution
- Cauchy Distribution
- Lognormal Distribution
- Double Exponential Distribution

Uniform Distribution

- The continuous uniform distribution is defined by spreading mass uniformly over an interval $[a,b]$.

$$f(x | a, b) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in [a, b] \\ 0, & \text{otherwise} \end{cases}$$

$$EX = \frac{b+a}{2}$$

$$VarX = \frac{(b-a)^2}{12}$$

Normal Distribution(1)

- The normal distribution plays a central role in a large body of statistics.
- the normal distribution and distributions associated with it are very tractable analytically.
- Has the familiar bell shape, whose symmetry makes it an appealing choice for many population models.

$$f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, -\infty < x < \infty$$
$$EX = \mu \quad \text{Var}X = \sigma^2$$

Normal Distribution(2)

- The normal pdf has its maximum at $x=\mu$, and inflection points (where the curve changes from concave to convex) at $\mu\pm\sigma$. Furthermore, the probability within 1, 2, or 3 standard deviation of the mean is
 - $P(|X-\mu|\leq\sigma)=P(|Z|\leq 1)=0.6826$
 - $P(|X-\mu|\leq 2\sigma)=P(|Z|\leq 2)=0.9544$
 - $P(|X-\mu|\leq 3\sigma)=P(|Z|\leq 3)=0.9974$

Normal Distribution(3)

- Approximation to other distributions $X \sim \text{binomial}(n, p)$
 $E[X]=np$; $\text{Var}[X]=np(1-p)$
- Under the condition that n is large and p is not extreme (near 0 or 1), X can be approximated by a normal random variable with mean $\mu = np$ and variance $\sigma^2 = np(1 - p)$
 - Example: $X \sim \text{binomial}(25, 0.6)$, $\mu = 25 * 0.6$, $\sigma^2 = 25 * 0.6 * 0.4$
 - $P(X \leq 13) = 0.267$; $P(Y \leq 13) = 0.206$;
 - $P(X \leq 13) = P(X \leq 13.5) \approx P(Y \leq 13.5) = 0.271$
 - $P(X \leq x) \approx P\left(Y \leq x + \frac{1}{2}\right)$
 - $P(X \geq x) \approx P\left(Y \geq x - \frac{1}{2}\right)$

Gamma Distribution

- The gamma family of distributions is a flexible family of distributions on $[0, \infty]$
- Gamma Function

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

$$\Gamma(1) = 1; \Gamma(n) = (n-1)!; \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

$$f(t) = \frac{t^{\alpha-1} e^{-t}}{\Gamma(\alpha)}, 0 < t < \infty \quad X = \beta T$$

$$f(x | \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

Gamma Distribution (cont.)

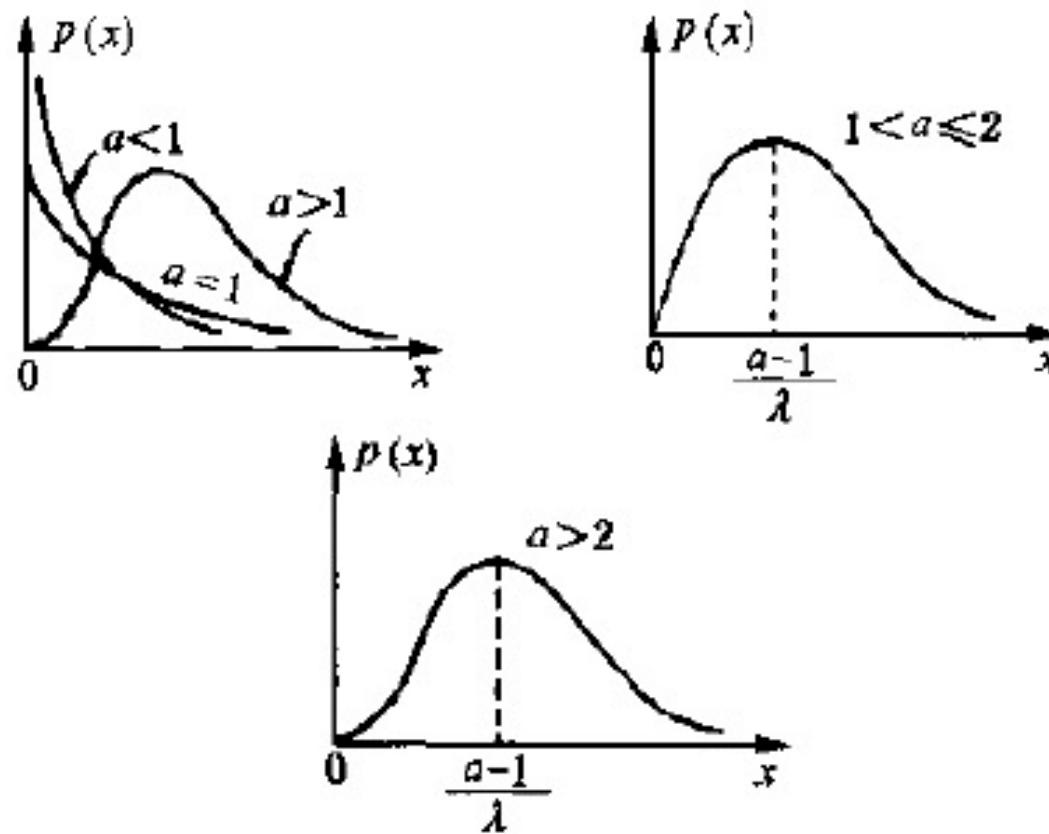
- The parameter α is known as **the shape parameter**, since it most influences the peakedness of the distribution, while the parameter β is called the **scale parameter**, since most of its influence is on the spread of the distribution.

$$f(x | \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, 0 < x < \infty, \alpha > 0, \beta > 0.$$

$$EX = \alpha\beta \quad \quad \quad VarX = \alpha\beta^2$$

Gamma Distribution (cont.)

从图上可见, $\alpha > 1$ 时, 伽玛密度函数是单峰, 峰值位于 $x = (\alpha - 1)/\lambda$; 对 $1 < \alpha \leq 2$, 其密度函数是先上凸, 后下凸; 对 $\alpha > 2$, 其密度是先下凸, 中间上凸, 最后又下凸。



Gamma-Poisson relationship

- Set $\alpha=p/2$ where p is an integer, $\beta=2$, the gamma pdf becomes

$$f(x | p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2}, 0 < x < \infty$$

- The chi squared pdf with p degrees of freedom
- Set $\alpha=1$

$$f(x | \beta) = \frac{1}{\beta} e^{-x/\beta}$$

- The *exponential pdf with scale parameter β*

Exponential distribution

- The exponential distribution shares the “memoryless” property
 - $P(X>s+t|X>s)=P(X>t)$
 - $f(x|\lambda) = \lambda e^{-\lambda x}$
- Play an important role in the analysis of failure time data.
- Very useful for modeling hazard functions.

Beta Distribution(1)

- The Beta family of distributions is a continuous family on (0,1) indexed by two parameters.

$$f(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, 0 < x < 1, \alpha > 0, \beta > 0$$

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

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

The equation is very useful in dealing with the beta function, allowing us to take advantage of the properties of the gamma function.

The beta distribution is the distribution that gives probability 1 to a finite interval, here taken to be (0,1). As such, the beta is often used to **model proportions, which naturally lie between 0 and 1.**

Beta Distribution(2)

$$E[X] = \frac{\Gamma(\alpha + 1)\Gamma(\alpha + \beta)}{\Gamma(\alpha + \beta + 1)\Gamma(\alpha)} = \frac{\alpha}{\alpha + \beta}$$

$$Var[X] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

$$E[X^n] = \frac{\Gamma(\alpha + n)\Gamma(\alpha + \beta)}{\Gamma(\alpha + \beta + n)\Gamma(\alpha)}$$

As the parameters α and β vary, the beta distribution takes on many shapes. The pdf can be strictly increasing ($\alpha > 1, \beta = 1$); strictly decreasing ($\alpha = 1, \beta > 1$), U-shaped ($\alpha < 1, \beta < 1$); unimodal ($\alpha > 1, \beta > 1$)

The pdf becomes more concentrated as α increases, but stays symmetric.

$\alpha = \beta = 1$, the beta distribution reduces to the uniform(0,1), showing that the uniform can be considered to be a member of the beta family.

Exponential Families (*)

- A family of pdfs or pmfs is called an *exponential family* if it can be expressed as

$$f(x | \theta) = h(x)c(\theta)\exp\left(\sum_{i=1}^k w_i(\theta)t_i(x)\right).$$

- $h(x) \geq 0$ and $t_1(x), \dots, t_k(x)$ are real-valued functions of the observation x .
 $c(\theta) \geq 0$ and $w_1(\theta), \dots, w_k(\theta)$ are real-valued functions of the possibly vector-valued parameter θ .
- Many common families introduced above are exponential families.
- Continuous families: Normal, gamma, beta,
- Discrete families: binomial, Poisson, negative binomial.

Binomial Exponential Families

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} (1-p)^n \exp\left(x \log\left(\frac{p}{1-p}\right)\right)$$

$$f(x|\theta) = h(x)c(\theta) \exp\left(\sum_{i=1}^k w_i(\theta)t_i(x)\right)$$

$$E(X) = np$$

Normal Exponential Families

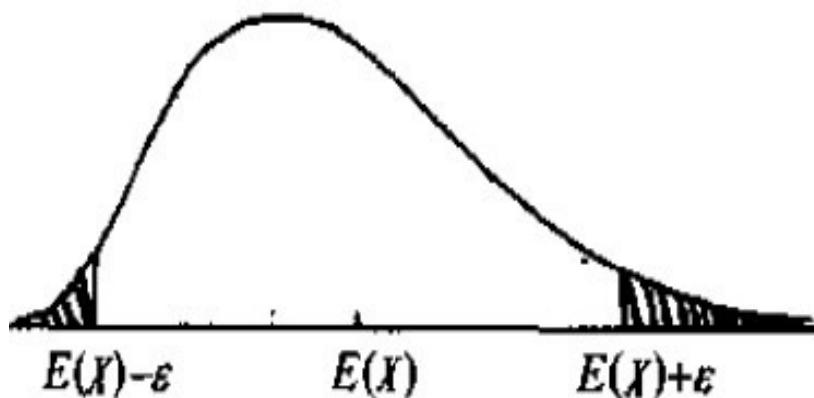
$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}\right)$$

$$f(x|\theta) = h(x)c(\theta) \exp\left(\sum_{i=1}^k w_i(\theta)t_i(x)\right)$$

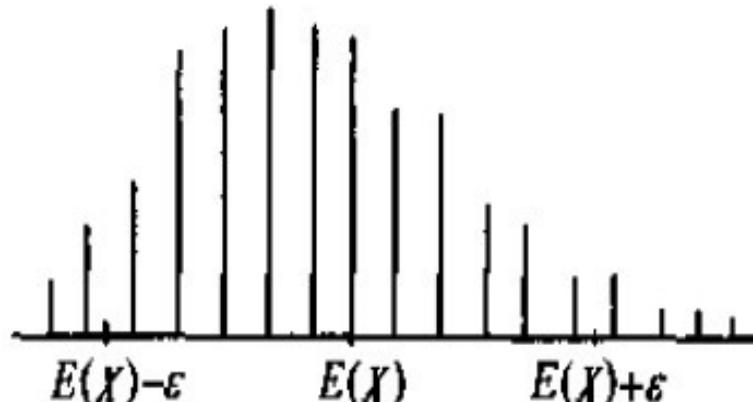
大数定律

定理 2.4.8 (切比雪夫不等式) 对任一随机变量 X , 若 EX^2 存在, 则对任一正数 ϵ , 恒有

$$P(|X-EX| \geq \epsilon) \leq \frac{\text{Var}(X)}{\epsilon^2} \quad (2.4.5)$$



(a) $P(|X-EX| \geq \epsilon) = \text{两尾部面积之和}$



(b) $P(|X-EX| \geq \epsilon) = \text{两尾部线段之和}$

大数定律

定理 2.4.10(贝努里大数定律) 设 X_n 是 n 重贝努里试验中事件 A 发生的次数, 又设事件 A 发生的概率 $P(A) = p$, 则对任意的 $\epsilon > 0$, 有

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{X_n}{n} - p\right| \geq \epsilon\right) = 0 \quad (2.4.8)$$

- 说明: 事件发生的频率与概率有较大偏差的可能性愈来愈小, 但并不意味着较大偏差永远不再发生, 只是说小偏差发生的概率大, 而大偏差的概率小, 小到可以忽略不计, 这就是频率稳定于概率的意义。

随机变量的描述统计量

- 矩
- 集中趋势的统计量
 - 均值、众数、中位数、百分位数、四分位数
- 分散程度
 - 方差、标准差
- 分布形状
 - 偏度系数、峰度系数

矩

定义 2.5.1 设 X 为随机变量, c 为常数, k 为正整数, 则量 $E(X-c)^k$ (假如它存在) 称为 X 分布关于 c 的 k 阶矩。若 $c=0$, 则量 EX^k 称为 X 分布的 k 阶(原点)矩, 记为 μ_k ; 若 $c=EX$, 则量 $E(X-EX)^k$ 称为 X 分布的 k 阶中心矩, 记为 ν_k 。

Moments and Moment Generating Functions

- For each integer n ,
 - the n -th moment of X , is $E(X^n)$
 - the n -th central moment of X , is $E((X - \mu)^n)$, $\mu = E(X)$
- The **variance** of X is the second central moment.
$$Var(X) = E((X - \mu)^2), \mu = E(X)$$
- Standard deviation is the positive square root of $Var X$.

Moment generating function

- The moment generating function(mgf) of X , denoted by $M_X(t)$

$$M_X(t) = E(e^{tX})$$

$$M_X(t) = \int_{-\infty}^{+\infty} e^{tx} f_X(x) dx$$

$$M_X(t) = \sum_x e^{tx} P(X = x)$$

- The function can be used to generate moments.
- The main use is to help in characterizing a distribution.

矩生成函数

- 如果 X 有矩生成函数 $M_X(t)$ 那么 $E[X^n] = M_X^{(n)}(0)$

$$M_X^{(n)}(t) = \frac{d^n}{dt^n} M_X(t)|_{t=0}$$

$$\frac{d}{dt} M_X(t)|_{t=0} = E[X]$$

$$\frac{d^n}{dt^n} M_X(t)|_{t=0} = E[X^n]$$

矩生成函数

- Gamma pdf:

$$f(x) = \frac{1}{\gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

$$M_X(t) = \left(\frac{1}{1 - \beta t} \right)^\alpha$$

$$EX = \alpha\beta$$

矩生成函数

- Binomial mgf:

$$M_X(t) = [pe_t + (1-p)]^n$$

$$EX = np$$

矩生成函数

- The major usefulness of the moment generating function is **not** in its ability to generate moments. Rather, its usefulness stems from the fact that, in many cases, the moment generating function can **characterize a distribution**.
 - Example: binomial probabilities can be approximated by Poisson probabilities.
- Characterizing the set of moments is **not enough** to determine a distribution uniquely because there may be two distinct random variables having the same moments.

众数

定义 2.5.7 假如 X 是离散随机变量, 则 X 最可能取的值(即使概率 $P(X = x)$ 达到最大的 x 值) 称为 X 分布的众数。假如 X 是连续随机变量, 则使其密度函数 $p(x)$ 达到最大的 x 值称为 X 的众数, X 的众数常记为 $\text{Mod}(X)$ 。

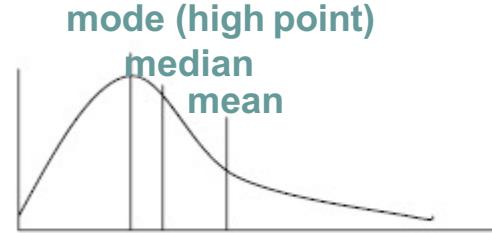
```
> y = c(1, 1, 2, 2, 2, 3, 4)*2
> table(y)
y
2 4 6 8
2 3 1 1
> which(table(y) == max(table(y)))
4
2
```

中位数

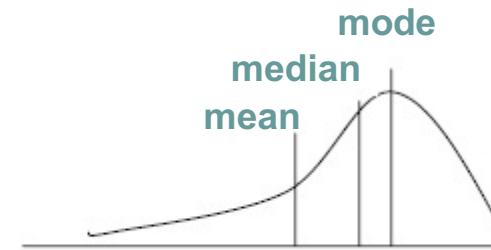
- 中位数：将 X 的取值范围分为概率相等（各为 0.5）的两部分的数值。
- `>median(x)`

Mean or Median?

- Appropriate summary of the center of the data?
 - Mean if the data has a symmetric distribution with light tails (i.e. a relatively small proportion of the observations lie away from the center of the data).
 - Median if the distribution has heavy tails or is asymmetric.
- Extreme values that are far removed from the main body of the data are called outliers
- Large influence on the mean but not on the median. Right and left skewness (asymmetry)



(reverse alphabetic -RIGHT skewed)



(alphabetic -LEFT skewed)

分位数

定义 2.5.6 设连续随机变量 X 的分布函数为 $F(x)$, 密度函数为 $p(x)$, 对任意 $\alpha(0 < \alpha < 1)$, 假如 x_α 满足条件

$$F(x_\alpha) = \int_{-\infty}^{x_\alpha} p(x)dx = \alpha$$

则 x_α 称为 X 分布的 α 分位数, 或称 α 下侧分位数。假如 x'_{α} 满足条件

$$1 - F(x'_{\alpha}) = \int_{x'_{\alpha}}^{\infty} p(x)dx = \alpha$$

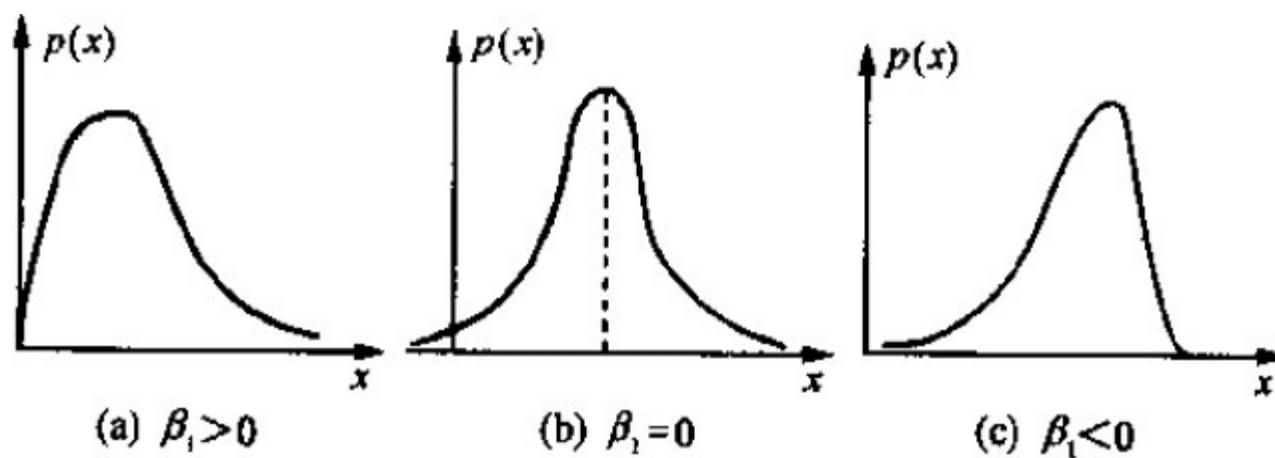
则 x'_{α} 称为 X 分布的 α 上侧分位数, 它们的区别见图 2.5.5。

偏度

定义 2.5.3 设随机变量 X 的三阶矩存在，则比值

$$\beta_1 = \frac{\nu_3}{(\nu_2)^{3/2}} = \frac{E(X - EX)^3}{[E(X - EX)^2]^{3/2}}$$

称为 X 分布的偏度系数，简称偏度。



峰度

定义 2.5.4 设随机变量 X 的四阶矩存在, 则比值减去 3

$$\beta_2 = \frac{\nu_4}{\nu_2^2} - 3 = \frac{E(X - EX)^4}{[E(X - EX)^2]^2} - 3$$

称为 X 分布的峰度系数, 简称峰度。

X 的标准化变量为 $X^* = (X - EX)/\sigma$.

$$\beta_2 = \frac{E(X^*)^4}{[E(X^*)^2]^2} - 3 = E(X^*)^4 - 3$$

- 峰度的含义: $\beta_2 > 0$: X 的标准化变量 X^* 在零附近集中取值的概率要大于标准化正态变量; < 0 , 则小于。

联合分布

- 用联合概率分布来描述样本中多个随机变量的 分布，设两个离散随机变量 X 和 Y ，它们的联合密度函数可写为：
 - $p(x,y)=P(X=x,Y=y)$
- 描述其中单个随机变量的概率密度函数称为边缘概率密度函数，则以上离散随机变量的边缘 概率分布函数可写为：

$$p_X(x) = \sum_Y p(x,y) \quad p_Y(y) = \sum_X p(x,y)$$

随机变量的独立性

- 边缘概率分布不能决定联合概率分布，除非各个随机变量独立。如果两个离散变量X和Y的分布独立，则联合概率分布函数为：

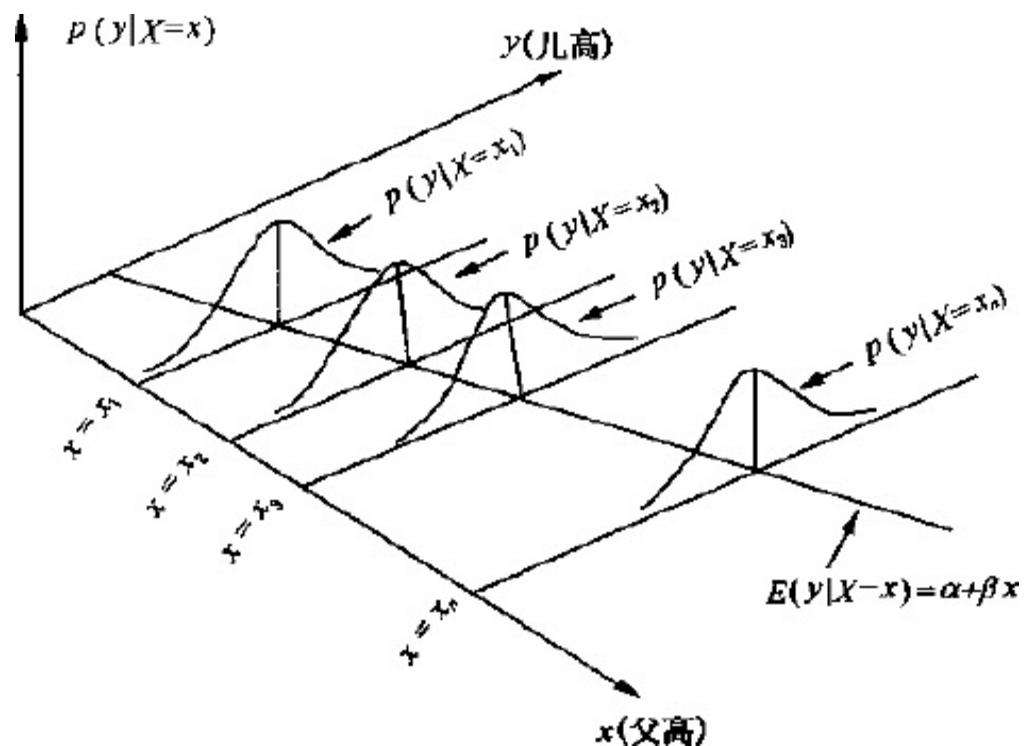
$$p(x, y) = p_X(x) p_Y(y)$$

- 例如：两个骰子点数都为6的概率计算如下：

$$p(Y = 6, Z = 6) = p(Y = 6) p(Z = 6) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

条件分布

$$P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_{\cdot j}}, \quad i = 1, 2, \dots$$



协方差

定义 3.3.1 设 (X, Y) 为二维随机变量, 它的二个方差都存在, 则称 $E[(X - EX)(Y - EY)]$ 为 X 与 Y 的协方差或相关矩, 记为

$$\text{Cov}(X, Y) = E[(X - EX)(Y - EY)] \quad (3.3.8)$$

- 取值:
- 正: 正相关, Y 的取值将会随 X 的取值增加而有增加的趋势 负: 负相关
- 0: 不相关

相关系数

定义 3.3.2 设 (X, Y) 为二维随机变量, 它的两个方差 σ_X^2 和 σ_Y^2 都存在, 且都为正, 则称 $\text{Cov}(X, Y)/\sigma_X\sigma_Y$ 为 X 与 Y 的线性相关系数, 简称相关系数, 记为

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X\sigma_Y} \quad (3.3.13)$$

它与协方差同符号, 当 $\text{Corr}(X, Y) > 0$, 称 X 与 Y 间为正相关; 当 $\text{Corr}(X, Y) < 0$, 称 X 与 Y 间为负相关; 当 $\text{Corr}(X, Y) = 0$, 称 X 与 Y 不相关。

随机变量的关系

随机变量的关系

独立

$$P(X,Y) = P(X)P(Y)$$

不独立

$$P(X) \neq P(X|Y)$$

独立一定不相关
不相关不一定独立

不相关

$$\text{Corr}(X,Y) = 0$$

相关

$$\text{Corr}(X,Y) \neq 0$$

中心极限定理

- 中心极限定理： n 个独立的同分布的随机变量之和的分布近似于正态分布。并且 n 越大，近似程度越好。（小误差的累积是符合正态分布的）

数理统计

- 研究怎样以有效的方式收集、整理和分析带随机性的数据，并在此基础上，对所研究的问题做出统计推断。
- 数理统计是研究处理数据的一门学问，概率论为数据处理提供了理论基础。

总体和样本

- 总体：一个统计问题中，研究对象的全体
- 个体：构成总体的每个成员
- 样本：从总体中抽出的部分个体组成的集合
 - 代表性
 - 独立性

样本统计量

- 样本均值

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

- 样本方差

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

概率函数的估计

- 参数化方法：假设服从了某一类已知的概率分布
 - 已经有了一个数据生成的概率模型，要在一系列可能概率分布中确定一个
 - 只有几个参数需要确定，所需要的训练样本数据的规模不用太大
- 非参数化方法 (**distribution free**)
 - 产生一个离散的概率分布，或用插值得到连续的分布函数
 - 训练数据需要很多

矩估计

- 用样本矩估计总体矩

设 X_1, X_2, \dots, X_n 是来自某总体 X 的一个样本, 则样本的 k 阶原点矩为:

$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k \quad , \quad k = 1, 2, \dots$$

如果总体 X 的 k 阶原点矩 $\mu_k = E(X^k)$ 存在, 则用 A_k 去估计 μ_k , 记为

$$\hat{\mu}_k = A_k$$

- 优点: 不要求知道总体的分布

矩估计-示例

例 5.1.1 设总体 $X \sim b(1, p)$, 从中获得样本 X_1, X_2, \dots, X_n , 由于 $E(X) = p$, 故 p 的矩法估计为

$$\hat{p} = A_1 = \bar{X}$$

设样本的观察值为 x_1, x_2, \dots, x_n , 那么每一个 x_i 不是 0 便是 1, 从而 \hat{p} 的观察值便是

$$\hat{p} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1, \dots, x_n \text{ 中 } 1 \text{ 的个数}}{n}$$

这便是频率。

极大似然估计

- 总体分布已知，设总体含待估参数 θ ，可以取很多值，在一切可能值中选出一个使样本观测值出现概率为最大的 θ 值作为 θ 值的估计，称为极大似然估计。

$$L(\theta) = \prod_{i=1}^n p(x_i; \theta)$$

极大似然估计-示例

例 5.3.2 设某工序生产的产品的不合格品率为 p , 抽 n 个产品作检验, 发现有 T 个不合格, 试求 p 的极大似然估计。

- 写出似然函数
 - 对数似然函数
 - 求导或求偏导
 - 解似然方程求出参数的估计值
- 《概率论与数理统计》: P252 例5.3.3

$$L(p) = \prod_{i=1}^n p^{x_i} (1-p)^{(1-x_i)}$$

贝叶斯统计

- 在重视使用总体信息和样本信息的同时，还注意先验信息的收集、挖掘和加工，使其数量化，形成先验分布，参加到统计推断中，以提高统计推断的质量。
- 贝叶斯学派的基本观点是：任一未知量 θ 都可以看做是随机变量，用一个概率分布去描述，即先验分布。

贝叶斯的具体方法

- 1) $p(x|\theta)$ 表示在随机变量 θ 给定某个值， X 的条件密度函数
- 2) 参数 θ 的先验分布 $\pi(\theta)$
- 3) 贝叶斯观点，设想从 $\pi(\theta)$ 产生一个 θ' ；从 $p(x|\theta')$ 产生样本
- 4) 样本和参数的联合分布 $h(x, \theta) = p(x|\theta)\pi(\theta)$
- 5) 对未知参数 θ 做统计推断

$$\pi(\theta|x) = \frac{h(x, \theta)}{m(x)} = \frac{p(x|\theta)\pi(\theta)}{\int_{\Theta} p(x|\theta)\pi(\theta)d\theta}$$

示例

- 抛硬币，用 $\mu(m)$ 表示先验分布。设 s 表示观察到的事件的特定序列，其中*i*次正面朝上，*j*次正面朝下，若 $P(\text{head})=m$ ，则

$$P(s|\mu_m) = m^i(1-m)^j$$

- 先验分布为： $P(\mu_m) = B(m|a=2, b=2) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}m^{a-1}(1-m)^{b-1} = 6m(1-m)$
- 计算后验分布：

$$\begin{aligned} P(\mu_m|s) &= \frac{P(s|\mu_m)P(\mu_m)}{P(s)} \\ &= \frac{m^i(1-m)^j 6m(1-m)}{P(s)} \end{aligned}$$

- 求出 $\arg \max_m P(\mu_m|s)$

共轭先验

- In [Bayesian probability](#) theory, if the [posterior distribution](#) $p(\theta|x)$ is in the same [probability distribution family](#) as the [prior probability distribution](#) $p(\theta)$, the prior and posterior are then called **conjugate distributions**, and the prior is called a **conjugate prior** for the [likelihood function](#) $p(x|\theta)$
- A conjugate prior is an algebraic convenience, giving a closed-form expression for the posterior; otherwise, numerical integration may be necessary. Further, conjugate priors may give intuition by more transparently showing how a likelihood function updates a prior distribution.

共轭先验

Binomial: $f(x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$

Beta: $g(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$

$$\begin{aligned} p(\theta|x) &= \frac{p(x|\theta)p(\theta)}{\int_{\theta} p(x|\theta)p(\theta)d\theta} \\ &= \frac{\binom{n}{x} \theta^x (1 - \theta)^{n-x} \cdot \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}}{\int_{\theta} \binom{n}{x} \theta^x (1 - \theta)^{n-x} \cdot \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1} d\theta} \\ &= \frac{\frac{C_n^x}{B(\alpha, \beta)} \theta^{x+\alpha-1} (1 - \theta)^{n-x+\beta-1}}{\frac{C_n^x}{B(\alpha, \beta)} \int_0^1 \theta^{x+\alpha-1} (1 - \theta)^{n-x+\beta-1} d\theta} \\ &= Beta(x + \alpha, n - x + \beta) \end{aligned}$$

$$\int_0^1 \theta^{x+\alpha-1} (1 - \theta)^{n-x+\beta-1} d\theta = B(x + \alpha, n - x + \beta)$$

共轭先验

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters ^[note 1]
Bernoulli	p (probability)	Beta	$\alpha, \beta \in \mathbb{R}$	$\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i$
Binomial with known number of trials, m	p (probability)	Beta	$\alpha, \beta \in \mathbb{R}$	$\alpha + \sum_{i=1}^n x_i, \beta + \sum_{i=1}^n N_i - \sum_{i=1}^n x_i$
Negative binomial with known failure number, r	p (probability)	Beta	$\alpha, \beta \in \mathbb{R}$	$\alpha + rn, \beta + \sum_{i=1}^n x_i$
Poisson	λ (rate)	Gamma	$k, \theta \in \mathbb{R}$	$k + \sum_{i=1}^n x_i, \frac{\theta}{n\theta + 1}$
			α, β ^[note 4]	$\alpha + \sum_{i=1}^n x_i, \beta + n$
Categorical	\mathbf{p} (probability vector), k (number of categories; i.e., size of \mathbf{p})	Dirichlet	$\boldsymbol{\alpha} \in \mathbb{R}^k$	$\boldsymbol{\alpha} + (c_1, \dots, c_k)$, where c_i is the number of observations in category i
Multinomial	\mathbf{p} (probability vector), k (number of categories; i.e., size of \mathbf{p})	Dirichlet	$\boldsymbol{\alpha} \in \mathbb{R}^k$	$\boldsymbol{\alpha} + \sum_{i=1}^n \mathbf{x}_i$

最优编码

- 有一个房间中有时没有人，有时甲在房间中，有时乙在房间中，有时甲乙都在房间中，房间状态服从下面的概率分布：

房间状态	房间没有人	甲在房间	乙在房间	甲乙均在房间
概率	0.5	0.125	0.125	0.25

- 定时记录房间状态(消息)，将房间状态编码，并通过通信设备发送出去。如何编码，使得连续发送消息时，编码长度最短？
- 定长编码2个二进制位
- 发送一个消息，平均2个二进制位。

最优编码

- 变长编码：给小概率信息赋以较长的编码，而给大概率消息赋以较短的编码。

消息	编码
房间没有人	0
甲在房间	110
乙在房间	111
甲乙均在房间	10

- 发送一个消息，平均需要1.75个二进制位。
- $0.5 \times 1 + 0.125 \times 3 + 0.125 \times 3 + 0.25 \times 2 = 1.75$

最优编码

- 随机变量 X 服从概率分布 P , 如果消息 x 的概率为 $p(x)$, 则给其分配一个长度为 $\lceil -\log_2 p(x) \rceil$ 个二进制位的编码。发送一个消息平均需要 $-\sum p(x) \log_2 p(x)$ 个二进制位。
- 消息的编码长度大, 可理解为消息所含信息量大。消息的编码长度小, 则消息所含信息量小。
- 平均信息量即为发送一个消息的平均编码长度。
- 信息论中用熵描述随机变量平均信息量。

熵(entropy)

- 定义1 熵设 X 是取有限个值的随机变量，若其概率分布为
 $p(x) = P\{X=x\}$, 且 $x \in X$

- 则， X 的熵定义为

$$H(X) = -\sum p(x) \log_a p(x)$$

- 规定 $0 \log_a 0 = 0$
- 通常 $a=2$, 此时熵的单位为比特。

- 熵的基本性质：

1. $H(X) \geq 0$, 等号表明确定场(无随机性)的熵最小。
2. $H(X) \leq \log|X|$, 等号表明等概场的熵最大。

◆ 熵描述了随机变量的不确定性。

熵

例子 1：假定有一种语言 P 有 6 个字母 p, t, k, a, i, u , 字母的分布密度为：

P	p	t	k	a	i	u
概率	1/8	1/4	1/8	1/4	1/8	1/8

则随机变量 P 的熵为：

语言的字母熵

$$\begin{aligned} H(P) &= - \sum_{i \in \{p, t, k, a, i, u\}} p(i) \log p(i) \\ &= -[4 \times \frac{1}{8} \log \frac{1}{8} + 2 \times \frac{1}{4} \log \frac{1}{4}] \\ &= 2 \frac{1}{2} \text{ bit} \end{aligned}$$

联合熵、条件熵

- 定义2 联合熵设 X 、 Y 是两个离散型随机变量，它们的联合分布密度为 $p(x,y)$ ，则 X,Y 的联合熵定义为：

$$\bullet H(X,Y) = - \sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x,y)$$

- 定义3 条件熵设 X 、 Y 是两个离散型随机变量，它们的联合分布密度为 $p(x,y)$ ，则给定 X 时 Y 的条件熵定义为：

$$\bullet H(Y | X) = - \sum_{x \in X} p(x) H(Y | X=x) = \sum_{x \in X} p(x) \left[- \sum_{y \in Y} p(y | x) \log p(y | x) \right]$$

$$\bullet = - \sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(y | x)$$

- 链式规则 $H(X,Y) = H(X) + H(Y|X)$

熵率 (entropy rate)

- 信息量的大小随着消息长度的增加而增加，为了便于比较，一般使用熵率的概念，熵率一般也称为字符熵 (per-letter entropy) 或词熵 (per-word entropy)。
- 定义4：熵率，对于长度为 n 的消息，熵率定义为：

$$H_{rate} = \frac{1}{n} H(X_{1:n}) = -\frac{1}{n} \sum_{x_{1:n}} p(x_{1:n}) \log p(x_{1:n})$$

- Shannon, 1949 英文字母熵 27 letters (A-Z, space) unigram 4.03bit bigram 3.32bit trigram 3.1bit
- 冯志伟教授最早估算出汉字的熵为 9.65 比特。
- 语言 L 的熵
$$H_{rate}(L) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1, X_2, \dots, X_n)$$

互信息(mutual information)

- 根据链式规则，有：
 - $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$
- 可以推导出：
 - $H(X) - H(X|Y) = H(Y) - H(Y|X)$
- $H(X)$ 和 $H(X|Y)$ 的差称为互信息，一般记作 $I(X;Y)$
- $I(X;Y)$ 描述了包含在 X 中的有关 Y 的信息量，或包含在 Y 中的有关 X 的信息量

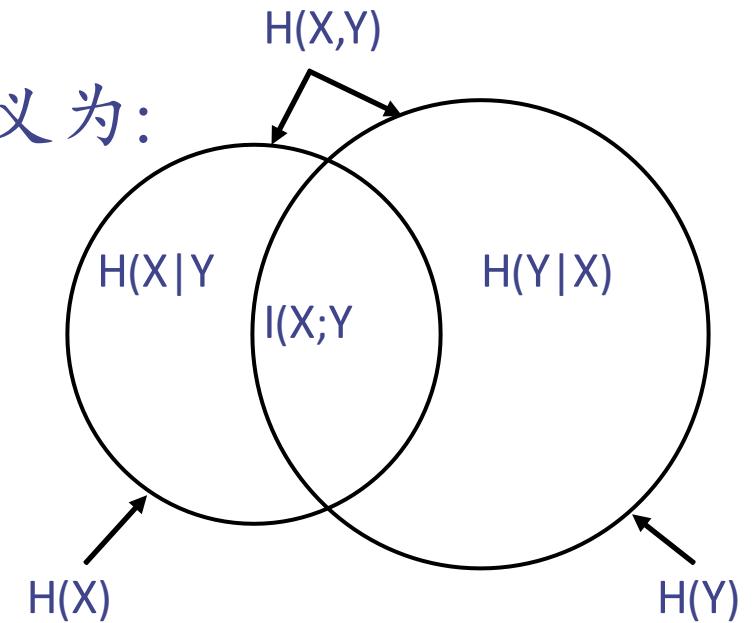
互信息

- 定义5: 互信息, 随机变量 X, Y 之间的互信息定义为:

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

- 互信息的性质:

- $I(X;Y) \geq 0$ 等号成立当且仅当 X 和 Y 相互独立。
- $I(X;Y) = I(Y;X)$ 说明互信息是对称的。



点间互信息(pointwise mutual information)

- 在计算语言学中，更为常用的是两个具体事件之间的互信息，一般称之为点间互信息。
- 定义6：点间互信息，事件 x, y 之间的互信息定义为：

$$I(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

- 点间互信息度量两个具体事件之间的相关程度
 - 当 $I(x, y) >> 0$ 时， x 和 y 高度相关。
 - 当 $I(x, y) = 0$ 时， x 和 y 相互独立。
 - 当 $I(x, y) << 0$ 时， x 和 y 呈互补分布。

相对熵(relative entropy)

- 定义7: 相对熵, 设 $p(x)$ 、 $q(x)$ 是随机变量 X 的两个不同的分布密度, 则它们的相对熵定义为:

$$D(p\|q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

- 相对熵一般也称谓*Kullback-Leibler* 发散度或*Kullback-Leibler* 距离。
- 度量同一个随机变量的不同分布的差异。
- 相对熵描述了因为错用分布密度而增加的信息量。

交叉熵(cross entropy)

- 定义8: 交叉熵, 设随机变量 X 的分布密度为 $p(x)$, 在很多情况下 $p(x)$ 是未知的, 人们通常使用通过统计手段得到的 X 的近似分布 $q(x)$, 则随机变量 X 的交叉熵定义为:

$$\bullet H(X, q) = - \sum_{x \in X} p(x) \log q(x)$$