第二、三讲

概率基础

- 基本概念：随机事件，排列组合
- 概率的定义（概型）
- 条件概率和贝叶斯公式
- 独立试验概型
- 随机变量及其分布
- 随机变量及其数字特征
- 相关性
- 中心极限定理

随机事件、随机试验

- **概率：随机事件发生的可能性的度量**
  P = 0 ~ 1
  分数或百分比

- **随机事件：可能发生可能不发生，大量重复时表现出一定规律的事件**
  概率：P（A）

- **观察某种商品的日销量，各种福利彩票的摇奖、某地区夏季暴雨的次数等.对随机现象的观察或实验统称为随机试验，简称试验。随机试验的结果称为随机事件（**简称**事件，**用*A,B,C …表示*）
  - **例** 掷一颗骰子，观察出现的点数就是一个随机试验.
  - **例** 抛一枚硬币，观察正、反面出现的情况也是一个随机试验

- **基本事件：随机试验中不能再分解的最简单的随机事件**

- **复合事件：由若干个基本事件组合而成的事件称为复合事件**
  - 如“出现偶数点”

排列和组合

- **1. 不重复的排列：从n个不同的元素中每次抽取m个不同的元素，按照一定的顺序排成一列：m<n，为选排列；m=n为全排列**

选排列和全排列的种数分别用P_n^m和P_n^n表示，

计算公式分别为

$$P_n^n = n!; \quad P_n^m = n(n-1)\cdots(n-m+1) = \frac{n!}{(n-m)!}$$

- **例：10个人选4个人排成一队**

排列和组合

- **2. 可重复的排列：从n个不同的元素中每次抽取m个可以相同的元素，按照一定的顺序排成一列**

$$n^m$$

- **例：** 某城市的电话号码是7位数字，并且首位不能为零，最多可以安装多少台不同号码的电话机？

解　是7个数字的可重复排列问题，因首位不能为0，只有9种选择，其余6个位置都有10种选择方法，由乘法原理，最多可以安装9×10^6台电话机.

排列和组合

- **3. 组合：从n个不同的元素中每次抽取m个不同的元素，不管顺序成为一组，组合方式的数目为**

由乘法原理知其组合种数为

$$C_n^m = \frac{P_n^m}{m!} = \frac{n!}{m!(n-m)!}$$

规定$0!=1$.

例 在一次考试中，某学生应做9道考题中的6道，问他有多少种选法？如果还要求他至少回答前5道题中的3道题，有多少种选法？

解 本题为组合问题. 在9道考题中选6道，有 C_9^6=84种选法.

如果要求至少回答前5道题中的3道题，包括3种情况：

(1) 在前5题中选3个，后4题中选3个，有 $C_5^3C_4^3 = 40$种选法;

(2)在前5题中选4个，后4题中选2个，有 $C_5^4C_4^2 = 30$种选法;

(3)前5题全选，后4题中选1个，有 $C_5^5C_4^1 = 4$种选法;

再由加法原理，共有40+30+4=74种选法.

1. 加法原理 若完成一件事有m种不同的方式，第i中方式中有$n_i \ (i = 1, 2, \cdots, m)$种不同的方法，其中任何一种方法都可以一次完成这件事，则完成这件事共有$n_1 + n_2 + \cdots + n_m$种不同的方法.

2. 乘法原理 若一件事需要经过先后m个不同的步骤才能最后完成，其中第i个步骤有$n_i \ (i = 1, 2, \cdots, m)$种不同方法，则完成这件事共有$n_1 \cdot n_2 \cdot \cdots \cdot n_m$种不同的方法.

小象学院 ChinaHadoop.cn

第二、三讲

概率基础

- 基本概念：随机事件，排列组合
- 概率的定义（概型）
- 条件概率和贝叶斯公式
- 独立试验概型
- 随机变量及其分布
- 随机变量及其数字特征
- 相关性
- 中心极限定理

概率的定义

- **描述事件发生可能性大小的数量指标称为事件发生的概率，记作$P(A)$.**
  - 统计定义，古典定义，公理化定义，几何定义；也称为**概型
- **统计定义：**

记$n(A)$为n次试验中事件A出现的次数，称为A的频数. 记$f_n(A)$为n次试验中事件A出现的次数与试验总次数的比值，称为A的频率，即$f_n(A)=\dfrac{n(A)}{n}$.

  - 频率也可以反映事件发生的可能性大小，它是从多次试验的结果来考察随机事件发生的可能性大小，因而有随机性. 它的数值依赖于试验. 对于同一事件，不仅试验次数不同可以得出不同的频率，就是试验次数相同，得到的频率也可能不同.

概率

- **概率是固有和客观存在的**
- **大量重复试验的条件下，随机事件出现的频率将会随着试验次数*n*的增大而逐渐趋于稳定.**

试验者	试验次数 n	频数 $n(A)$	频率 $\mu(A)$
迪摩根	2048	1061	0.5181
蒲丰	4040	2048	0.5069
皮尔逊	12000	6019	0.5016
皮尔逊	24000	12012	0.5005
维尼	30000	14994	0.4998

– 严格地讲，概率的统计定义只是一种描述性的定义. 在大多数情况下，定义中提到的客观存在的数值p无法具体地确定.一般只是在大量重复试验的条件下，通过频率值或一系列频率的均值作为概率$p(A)$的近似值

概率

- **古典定义：**

定义 在古典概型中，随机事件A发生的概率为

$$P(A)=\frac{\#A}{\#\Omega}$$

其中$\#A$、$\#\Omega$分别表示A包含的基本事件个数和试验的基本事件总数.

- **例一个五位数字的号码锁，每位上都有0,1, …9十个数码，若不知道该锁的号码，问开一次锁就能将锁打开的概率有多大？**

解 设$A=$“开一次就把锁打开”，

则$\#A=1,\#\Omega=10^5$,

于是 $$P(A)=\frac{\#A}{\#\Omega}=\frac{1}{10^5}=0.00001.$$

- **例**12个球中有5个红球，4个白球，3个黑球，从中任取2个球，计算没有取到红球的概率.

解　记$A=$"取到的2个球中没有红球"，则

$$\# A=C_7^2=21, \#\Omega=C_{12}^2=66.$$

$$P(A)=\frac{\# A}{\#\Omega}=\frac{21}{66}=\frac{7}{22}.$$

概率

- **例**：求n个人种至少两个人生日同天的概率（n小于等于365）

- **解**：生日各不相同的概率为

$$\frac{365\times 364\times\cdots\times(365-n+1)}{365^n}$$

- 至少两人同天生日的概率为

$$p=1-\frac{365\times 364\times\cdots\times(365-n+1)}{365^n}$$

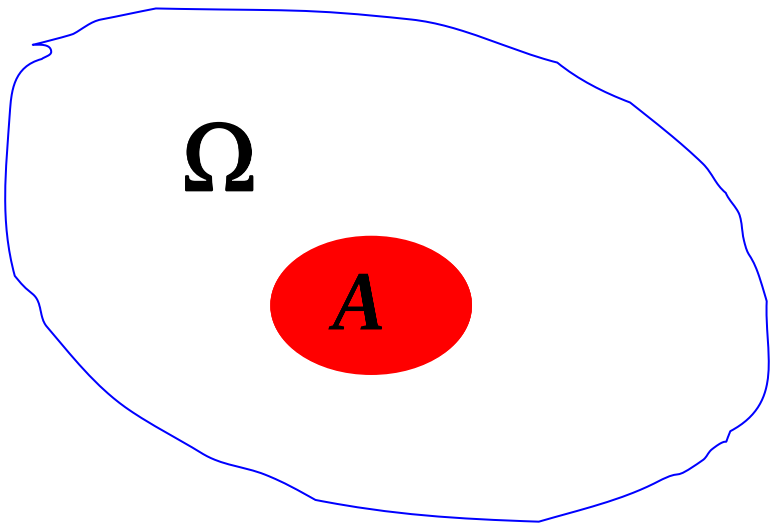
概率

➢ 至少两人生日同天的概率

人数	至少有两人生日相同的概率
10	0.11694817771107765187
20	0.41143838358057998762
30	0.70631624271926865996
40	0.89123180981794898965
50	0.97037357957798839992
60	0.99412266086534794247
70	0.99915957596515709135
80	0.99991433194931349469
90	0.99999384835612360355
100	0.99999969275107214842
110	0.99999998947129430621
120	0.99999999975608521895
130	0.99999999999624032317
140	0.99999999999996210395
150	0.9999999999999997549
160	0.99999999999999999900

概率

- **几何概型**：把有限个样本点推广到无限个样本点的场合,人们引入了**几何概型----等可能随机试验模型**



在面积为$S(\Omega)$ ($S(\Omega) < +\infty$)区域Ω中等可能地随机投点

点落入Ω中任意区域A的可能性大小与A的面积$S(A)$成正比，而与其位置或形状无关。

$$P(A)=\frac{S(A)}{S(\Omega)}$$ ---平面区域Ω上A的几何概率

$$P(A)=\frac{m(A)}{m(\Omega)}$$

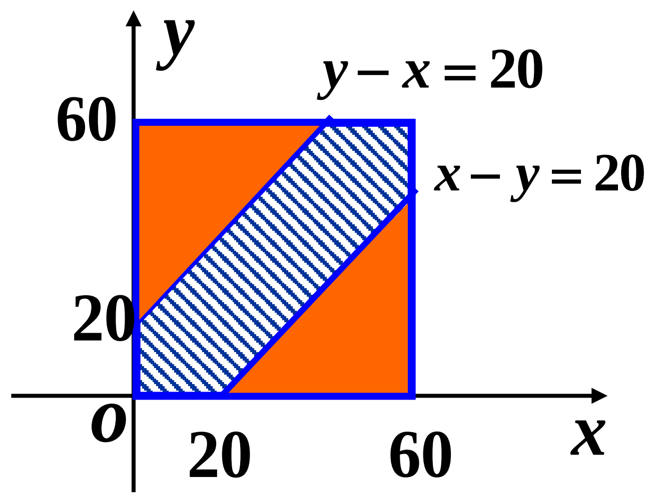
(其中$m(\Omega)$是样本空间的度量$m(A)$是构成事件A的子区域的度量) 这样借助于几何上的度量来合理规定的概率称为几何概率

概率

- **例 某人午觉醒来,发觉表停了,他打开收音机,想听电台报时, 假定电台每小时正点报时一次，求他等待的时间短于10分钟的概率。**

解：因为电台每小时报时一次, 我们自然认为这个人打开收音机时处于两次报时之间, 例如(13:00,14:00), 而且取各点的可能性一样, 要遇到等待时间短于10分钟, 只有当他打开收音机的时间正好处于13:50至14:00之间才有可能, 相应的概率是10/60=1/6.

➢ 例（会面问题）：甲、乙两人相约7点到8点在某地会面，先到者等候另一人20分钟，过时就可离去，试求这两人能会面的概率



解：以x，y分别表示甲、乙两人的到达时刻，则两人能会面的充要条件为$|x-y|\le 20$

$$P(A)=\frac{S(A)}{S(\Omega)}=\frac{60^2-40^2}{60^2}=\frac{5}{9}$$

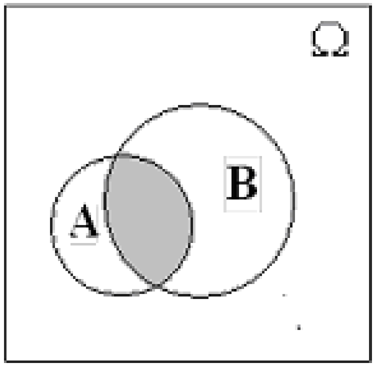
第二、三讲

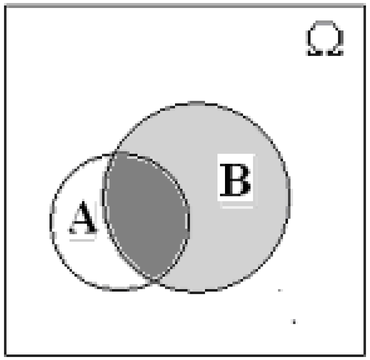
概率基础

- 基本概念：随机事件，排列组合
- 概率的定义（概型）
- 条件概率和贝叶斯公式
- 独立试验概型
- 随机变量及其分布
- 随机变量及其数字特征
- 相关性
- 中心极限定理

概率：条件概率

q是在事件B已经发生的条件下，事件A发生的概率.
一般情况下，它与事件A发生的概率$P(A)$是不相同的.称之为“在事件B发生的条件下，事件A发生的条件概率”，记为$P(A|B)$。事件AB与事件$A|B$可用图表示 .





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

概率：条件概率

小象学院 ChinaHadoop.cn

假设A, B是两个随机事件，若$P(B) > 0$，则有

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

若$P(A) > 0$,则有

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

上两式均称为概率的乘法公式.利用它们可计算两个事件A, B之积的概率$P(AB)$.

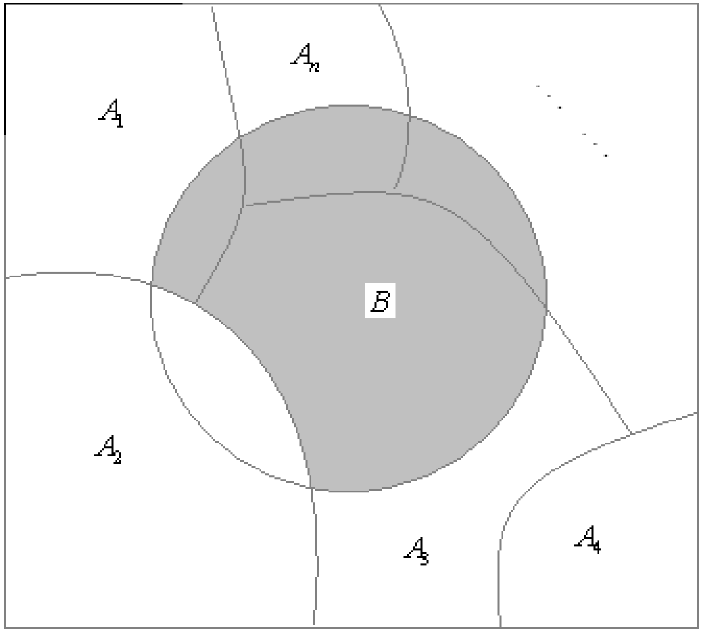
概率：条件概率

全概率公式 设事件$A_1,A_2,\cdots,A_n$是一个完备事件组，并且它们的概率都大于零，则对任意一个事件B，有

$$P(B)=\sum_{i=1}^{n}P(A_i)\cdot P(B|A_i)$$，称之为全概率公式.

$A_1,A_2,\cdots,A_n$是一个完备事件组，即

$A_1,A_2,\cdots,A_n$两两互不相容且$\sum_{i=1}^{n}A_i=\Omega.$



概率：条件概率

- **例：有朋自远方来，乘火车、轮船、汽车、飞机来的概率分别为0.3，0.2，0.1，0.4，迟到的概率分别为0.25，0.3，0.1，0；求他迟到的概率．**

- **解：设A_1 = 乘火车，A_2 = 乘船，A_3 = 乘汽车，A_4 = 乘飞机，B = 迟到. A_1, A_2, A_3, A_4构成一个完备事件组，由全概率公式得**

$$P(B)=\sum_{i=1}^{4}P(A_i)P(B\mid A_i)$$

$$=0.3\times0.25+0.2\times0.3+0.1\times0.1+0.4\times0$$

$$=0.145。$$

概率：贝叶斯公式

设事件 $A_1, A_2, \cdots, A_n$ 构成一个完备事件组，并且它们的概率都大于零，则对于任意一个概率大于零的事件B，有

$$P(A_m \mid B) = \frac{P(A_m B)}{P(B)} = \frac{P(A_m) \cdot P(B \mid A_m)}{\sum_{i=1}^{n} P(A_i) P(B \mid A_i)} \quad (m = 1, 2, \cdots, n).$$

称之为贝叶斯公式. 又称逆概率公式.

$A_i (i = 1, 2, \cdots, n)$视为导致事件发生$B$的"因素",

$P(A_i)$为"因素"的验前概率.

相反，事件B发生了，求各"因素"发生的条件概率$P(A_i|B)$

称$P(A_i|B)$为"因素"A_i的验后概率.

猜性别

小象学院 ChinaHadoop.cn

男/女 男/女 男/女 男/女 男/女 男/女

男/女 男/女 男/女 男/女 男/女 男/女

BEFORE：
性别比例？

1:1

认知模式：古典或贝叶斯

猜性别

男

男/女 男/女 男/女 男/女 男/女

男/女 男/女 男/女 男/女 男/女 男/女

BEFORE：
性别比例？

1:1

AFTER：
性别比例？

认知模式：古典或贝叶斯

留作作业

第二、三讲

概率基础

- 基本概念：随机事件，排列组合
- 概率的定义（概型）
- 条件概率和贝叶斯公式
- 独立试验概型
- 随机变量及其分布
- 随机变量及其数字特征
- 相关性
- 中心极限定理

概率：独立试验概型

设随机事件A, B满足 $P(AB) = P(A)P(B)$

则称事件A与B相互独立，简称A与B独立.

设A, B为两个事件，$P(A)P(B) > 0$，则

A与B独立的充分必要条件是

$P(A|B) = P(A)$或$P(B|A) = P(B)$.

事件 A 与 B 相互独立,是指其中任一事件发生的概率都不受另外一事件 发生的影响.

概率：独立试验

定义 如果试验满足下面两个条件：

（1）这n次试验相互独立；

（2）每次试验只有两种可能结果，即事件A发生或$\overline{A}$发生，且每次试验中，事件A发生的概率都相等，即$P(A)=p$.

则称这n个试验为n重贝努利概型.

定理 如果在n重贝努利概型中，事件A发生的概率为$p(0<p<1)$,则事件A在n次试验中恰好发生$k(0\le k\le n)$的概率$P(B_k)$为

$$P(B_k)=C_n^k p^k q^{n-k}$$

其中$q=1-p$,又称为贝努利公式.

概率：独立概型

小象学院 ChinaHadoop.cn

例 一位医生知道某种疾病患者自然痊愈率为0.25.

为试验一种新药是否有效，他将该药给10个病人服用，且规定若10个病人中至少有4个被医治好，则认为该药有效，反之则认为无效，求：

（1）虽然新药有效，且将痊愈率提高到0.35，但通过实验被认为无效的概率；

（2）新药完全无效，但通过实验被认为有效的概率.

解　将10个病人服用此药视为10次重复试验，每次试验只有两种可能结果：痊愈或不痊愈，而且每个人痊愈与否彼此独立.这是一个10重贝努利概型.

（1）令$A=$"新药有效且将痊愈率提高到0.35但被认为无效"

A发生当且仅当“10个病人中至多有3人被治好”，

$$\text{从而}P(A)=\sum_{k=0}^{3}C_{10}^{k}0.35^{k}(1-0.35)^{10-k}$$

$$=C_{10}^{0}\times 0.35^{0}\times 0.65^{10}+C_{10}^{1}\times 0.35^{1}\times 0.65^{9}+C_{10}^{2}\times 0.35^{2}\times 0.65^{8}$$

$$+C_{10}^{3}\times 0.35^{3}\times 0.65^{7}\approx 0.514.$$

*例 某居民区共有居民n人，设有一个银行，开有c个窗口，设每个窗口都办理所有业务，假定n个人在每一指定时刻是否到银行是独立的，每个人到银行的概率都是p.问：至少要设多少窗口才能以不小于$a(0<a<1)$的概率保证在每一时刻在每个窗口排队人数（包括正在被服务的那个人）不超过m?

概率：独立概型

小象学院 ChinaHadoop.cn

解 设$A=$"每个窗口排队人数不超过m"，为使每个窗口排队人数不超过m，在每一时刻到银行的人数至多为cm，

所以$P(A)=\sum_{k=0}^{cm} C_n^k p^k (1-p)^{n-k} \geq a$

解上述不等式，求出最小自然数c即可.

第二、三讲

概率基础

- 基本概念：随机事件，排列组合
- 概率的定义（概型）
- 条件概率和贝叶斯公式
- 独立试验概型
- 随机变量及其分布
- 随机变量及其数字特征
- 相关性
- 中心极限定理

随机变量的概念

例 掷两颗骰子,观察出现的点数之和.

解　用X 表示掷两颗骰子出现的点数之和，则X的取值可能为2，3，4…，12等这11个值.

例 从某电子元件厂生产的电子元件中，任意抽取一个电子元件，检查它的使用寿命.

解　用X表示其使用寿命，则X的可能取值应为非负实数，即$\{X \mid X \geq 0\}$.

例观察某网站在单位时间内的点击次数.

解 用X表示某网站在单位时间内收到的点击次，则X的可能取值是$0,1,\cdots, n\cdots$.

上述例子可看出，有些随机试验的结果表现为X的取值，有些随机试验的结果不直接表现为数字，但也能和数字建立对应关系.

例 掷一枚硬币，观察正反面出现的情况.

解 令$X=\begin{cases}1, & \text{正面向上}\\ 0, & \text{反面向上}\end{cases}$

随机变量

例 一批产品分为优质品、次品和废品三个等级，从中任取一件检验其质量等级.

解 令 $X = \begin{cases} 2 & \text{抽到优质品} \\ 1 & \text{抽到次品} \\ 0 & \text{抽到废品} \end{cases}$

对于随机试验的每一个基本结果ω，都对应一个实数$X(\omega)$，即$X(\omega)$是试验结果或样本点ω的函数.

定义 称这种依赖于特定随机试验并且由试验结果完全确定的变量为随机变量. 简记为X，常用X、Y、Z或ξ、η、ς等表示.

随机变量

小象学院 ChinaHadoop.cn

- **1. 变异性:** 随试验结果而变的量
- **2. 随机性:** 出现结果随机，试验前无法预测
- **3. 随机变量的每一种取值，就是一个随机事件.**
- **4. 在同一个样本空间可以同时定义多个随机变量**

离散型(***D.r.v.***)：取值有限个或可列个

非离散型(***N.D.r.v.***)

其中一种重要的类型为
连续型 *r.v*.(*C.r.v*.)

离散型随机变量及其分布

定义设X为离散型随机变量，X的一切可能取值为x_1，$x_2,\cdots$(有限个或可列个), X取各个可能值的概率为

$$p_i = P\{X = x_i\} = P(x_i), \quad i = 1, 2, \cdots.$$

称此式为离散型随机变量 X 的概率函数或分布律,简称X的分布.

$$\textbf{(1)}\ p_k \geq 0, \quad k = 1,2,\cdots;$$

$$\textbf{(2)} \sum_{k=1}^{\infty} p_k = 1.$$

- **1. 0-1分布**

定义 如果随机变量X的概率函数为

$$P\{X=1\}=p,\quad P\{X=0\}=1-p=q.$$

或$P\{X=i\}=p^i q^{1-i}, i=0,1.$ （$0<p<1, q=1-p$）

则称 X 服从参数 p 的 0-1分布（或两点分布）.

0-1分布是最简单的离散型分布，常描述只有两种对立结果的随机试验，即贝努利试验. 称贝努利试验的一种结果为“成功”，对立结果为“失败”.

- **二项分布**

定义 如果离散型随机变量X的概率函数为

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

称X服从参数为n，p的二项分布.

记作$X \sim B(n,p) \quad (0<p<1)$

$$\text{二项分布} \xrightarrow{n=1} \text{两点分布}$$

(2) 描述对象：

n重伯努利试验中某事件发生的次数.

离散型随机变量及其分布

➢ 泊松分布

定义如果离散型随机变量X的概率函数为

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

其中$\lambda>0$, 则称X服从参数为λ的泊松分布, 记作$X\sim P(\lambda)$.

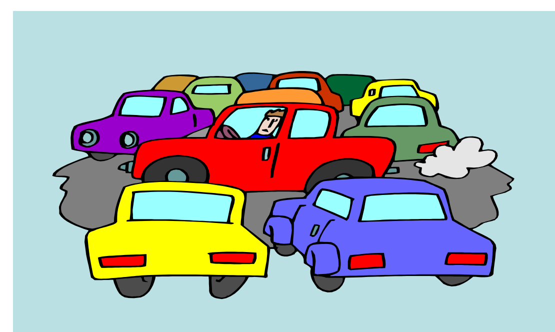
易证：(1) $P\{X=n\}=\frac{\lambda^n}{n!}e^{-\lambda}>0,\quad n=0,1,2,\cdots$

(2) $\sum_{n=0}^{\infty}P\{X=n\}=\sum_{n=0}^{\infty}\frac{\lambda^n}{n!}e^{-\lambda}=e^{-\lambda}\sum_{n=0}^{\infty}\frac{\lambda^n}{n!}=e^{-\lambda}e^{\lambda}=1.$

离散型随机变量及其分布

小象学院 ChinaHadoop.cn

- 在生物学、医学、工业统计、保险科学及公用事业的排队等问题中，泊松分布是常见的分布。例如地震、火山爆发、特大洪水、交换台的电话呼唤次数等，都服从泊松分布.



离散型随机变量及其分布

泊松定理在n重贝努利试验中，事件A发生的次数服从二项分布. 假设每次试验中事件A发生的概率为$p_n(0<p_n<1)$,如果$n\to\infty$时，$np_n\to\lambda$,
则对任意给定的非负整数k,有

$$\lim_{n\to\infty}P\{X=k\}=\lim_{n\to\infty}C_n^k p_n^k(1-p_n)^{n-k}=\frac{\lambda^k}{k!}e^{-\lambda}$$

在实际情况中，n一般为有限数，因此，若$X\sim B(n,p)$,
当n充分大而p相对较小时，有

$$P\{X=k\}=C_n^k p^k(1-p)^{n-k}\approx\frac{(np)^k}{k!}e^{-np}.$$

离散型随机变量及其分布

例某保险公司为了估计企业的利润，需要计算各种各样的概率，下面是较典型的问题之一：若一年中某类保险者里面每个人死亡的概率均等于0.005，现有1600人参加了这类保险，试求未来一年内在这些保险者中，

（1）有15人死亡的概率；（2）死亡人数不超过20的概率.

解 令X表示未来这一年内这些参保者的死亡人数，则

$X \sim B(1600, 0.005), np = 8 = \lambda$,所以由泊松定理，有

$$(1)\ P\{X = 15\} = C_{1600}^{15} 0.005^{15} 0.995^{1585} \approx \frac{8^{15}}{15!} e^{-8} = 0.009026,$$

$$(2) P\{X \le 20\} = \sum_{k=0}^{20} C_{1600}^{k} 0.005^{k} 0.995^{1600-k} \approx \sum_{k=0}^{20} \frac{8^k}{k!} e^{-8} = 0.999907.$$

连续型随机变量及其分布

- **连续型随机变量的概率密度**

定义 设X为一随机变量，若存在非负可积函数$f(x)$，使对于任意实数 $a,b(a<b)$ 有

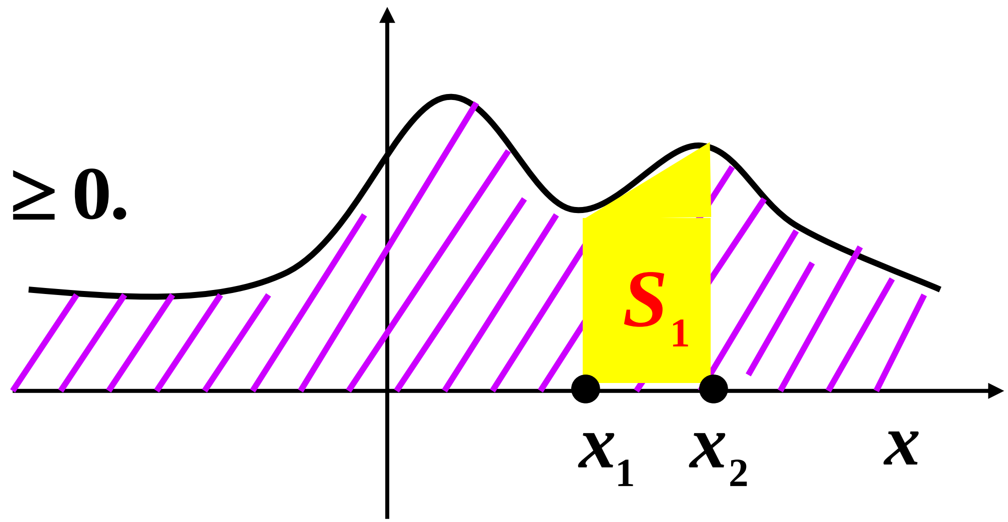
$$P\{a<X<b\}=\int_a^b f(x)dx,$$

则称 X 为连续型随机变量，称 $f(x)$ 为 X的概率密度函数，或分布密度函数，简称概率密度（分布密度），记作$X \sim f(x)$.

连续型随机变量及其分布

(1) 对任意的x， $f(x) \geq 0.$

(2) $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}\,x = 1.$



(3) $P\{x_1 < X \leq x_2\} = \int_{x_1}^{x_2} f(x)dx$

(4) 若 $f(x)$ 在点 x 处连续,则有 $F'(x) = f(x).$

连续型随机变量及其分布

- 1. 均匀分布

定义 如果连续型随即变量X的概率密度函数为

$$f(x)=\begin{cases}\dfrac{1}{b-a} & a\le x\le b\\ 0 & \text{其它}\end{cases}，\ \text{其中}a,b\in(-\infty,+\infty)\text{且}a<b,$$

则称X服从区间$[a,b]$上的均匀分布(等概率分布).

性质：$(1)\ P\{X<a\}=P\{X>b\}=0.$

(2)若$a\le c<d\le b, P\{c<X<d\}=\int_c^d \frac{1}{b-a}dx=\frac{d-c}{b-a}.$

➢ **2. 指数分布**

定义 如果随机变量X的密度函数为

$$f\left(x\right)=\begin{cases}\lambda e^{-\lambda x}, x\geq 0\\ 0, x<0\end{cases}$$

其中$\lambda>0$，则称X服从参数为λ的指数分布.

连续型随机变量及其分布

2. 指数分布

例 一种电子元件的使用寿命X服从参数$\lambda = 0.0005$的指数分布. 试求:

(1)一个电子元件的使用寿命大于2000小时的概率;

(2)独立的对三个这样的电子元件进行检验,至少有两个使用寿命大于2000小时的概率.

解 X的概率分布密度为 $$f(x) = \begin{cases} 0.0005e^{-0.0005x}, x \geq 0 \\ 0, \quad x<0 \end{cases}$$

(1)电子元件的使用寿命大于2000小时的概率

$$P\{X > 2000\} = \int_{2000}^{+\infty} 0.0005e^{-0.0005x}dx = e^{-1} \approx 0.368.$$

- **2. 指数分布**

(2)独立的对三个电子元件进行寿命检验，

令Y表示寿命大于2000小时的元件数，

则$Y \sim B(3, 0.368)$.

所以,

$$P\{Y \geq 2\} = C_3^2 \cdot 0.368^2 \cdot 0.632^1 + C_3^3 \cdot 0.368^3 \approx 0.307$$

连续型随机变量及其分布

小象学院 ChinaHadoop.cn

➢ 3. 正态分布

定义 如果连续型随机变量X的概率密度函数为

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

其中$\sigma>0$，μ为任意常数，则称X服从参数为μ和σ的正态分布. 记作$X\sim N\left(\mu,\ \sigma^2\right)$.

当$\mu=0$，$\sigma=1$时，X的概率密度为,

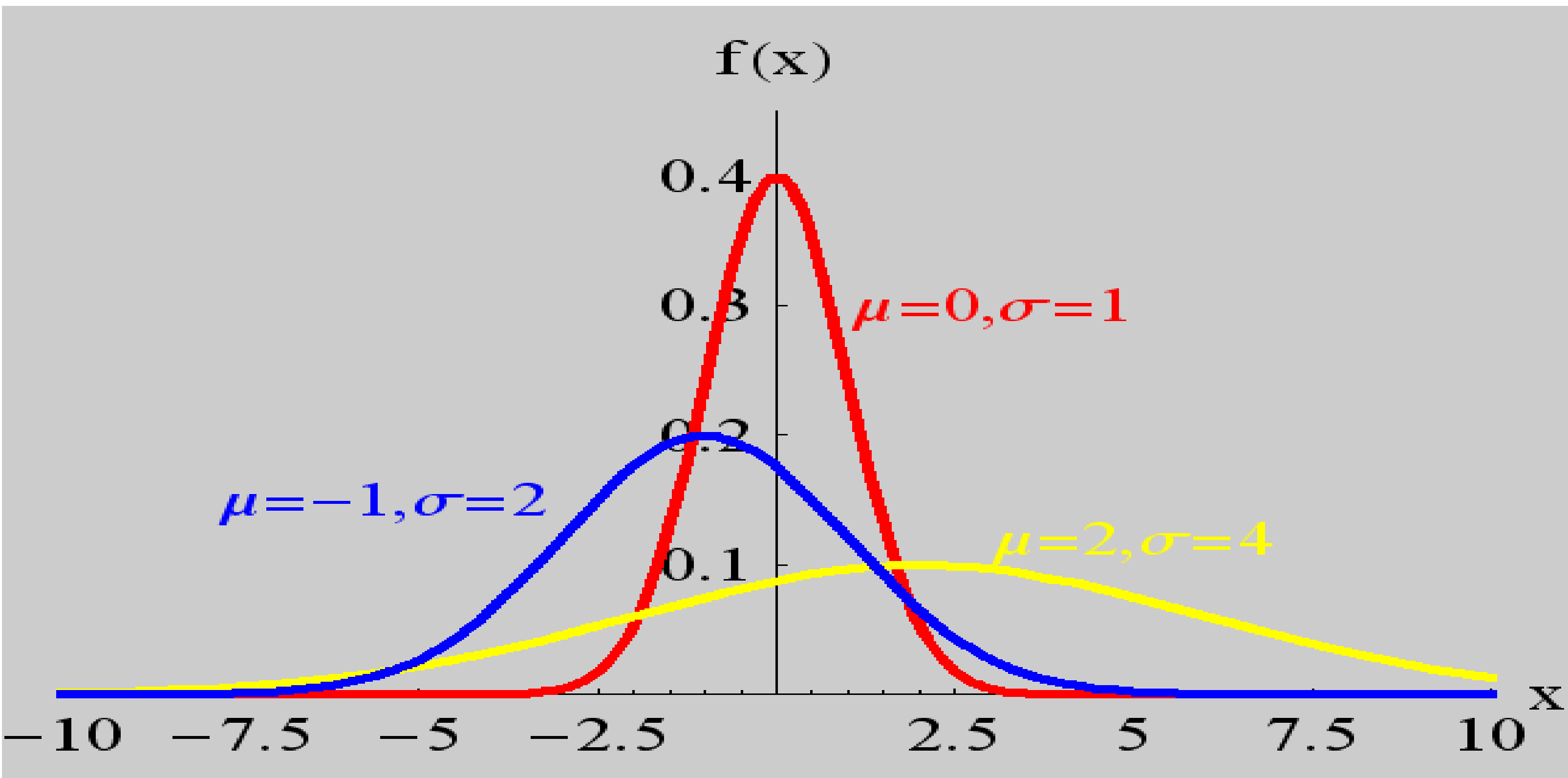
$$\varphi_0(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},\ -\infty<x<+\infty,$$

则称X服从标准正态分布，记作$X\sim N(0,1)$.

连续型随机变量及其分布

小象学院 ChinaHadoop.cn

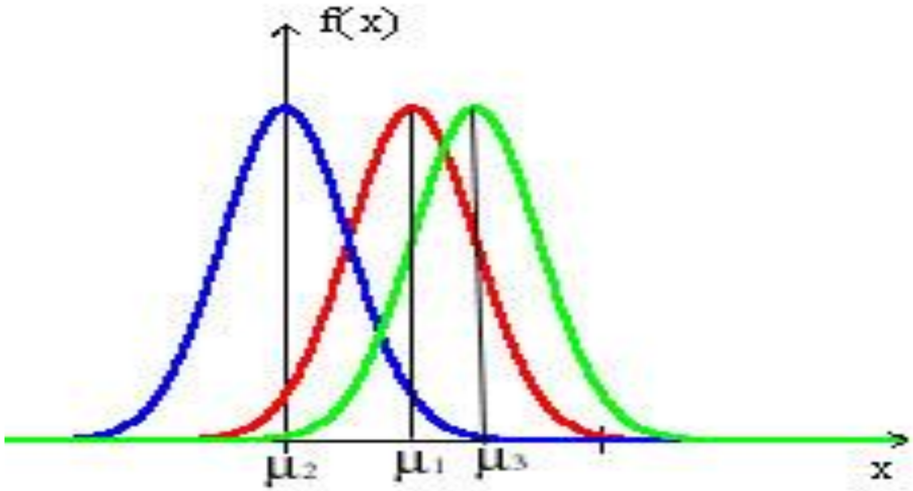
➢ **3. 正态分布**

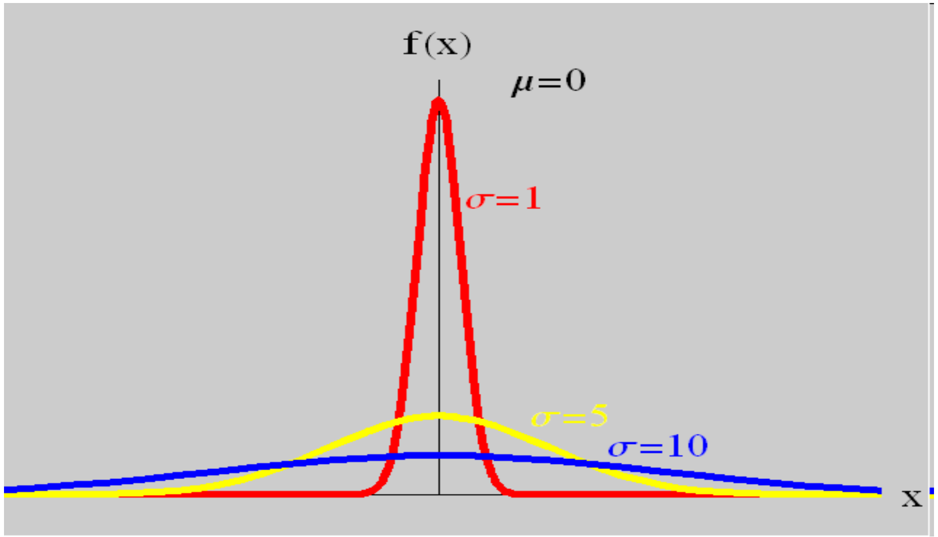


连续型随机变量及其分布

小象学院 ChinaHadoop.cn

- 3. 正态分布





连续型随机变量及其分布

- **3. 正态分布**

正态分布是概率论中最重要的一种分布，一方面，它是自然界中最常见的一种分布，如测量时的误差，人体身高、体重，农作物的收获量，产品的长度、强度，等都近似服从正态分布. 一般地说，这种量都可以看成是由大量的、微小的、相互独立的随机因素作用的结果，而每一种因素都不能起压倒一切的主导作用，则这个指标往往近似服从正态分布. 这可由中心极限定理证明.

正态分布具有良好的性质，许多分布都可用正态分布来近似. 又称高斯分布.

第二、三讲

概率基础

- 基本概念：随机事件，排列组合
- 概率的定义（概型）
- 条件概率和贝叶斯公式
- 独立试验概型
- 随机变量及其分布
- 随机变量及其数字特征
- 相关性
- 中心极限定理

随机变量的数字特征：期望

小象学院 ChinaHadoop.cn

若离散型随机变量X的可能取值为$x_i\,(i=1,2,\ldots)$,

其概率分布为 $P\{X=x_i\}=p_i,\quad i=1,2,\ldots$

则当$\sum_{i=1}^{\infty}x_i p_i$绝对收敛$\left(即\sum_{i=1}^{\infty}|x_i p_i|<\infty\right)$时，

称$\sum_{i=1}^{\infty}x_i p_i$为随机变量$X$的数学期望（简称期望）

常数的数学期望等于这个常数本身，即 $EC=C$.

随机变量X（数学期望存在）与常数C之和的数学期望等于X的数学期望与这个常数的和，即：$E(X+C)=EX+C$

随机变量的数字特征：期望

小象学院 ChinaHadoop.cn

定义若X为连续型随机变量，$f(x)$为其密度函数，

如果$\int_{-\infty}^{+\infty} xf(x)dx$绝对收敛$\left(即\int_{-\infty}^{+\infty} |xf(x)|dx < \infty\right)$，

则称$\int_{-\infty}^{+\infty} xf(x)dx$为随机变量$X$的数学期望。

$$E(X) = EX$$

随机变量的数字特征：期望

设X是一个随机变量，则$Y=g(X)$称为随机变量函数。

(1) 若X是离散型随机变量，概率分布为

则$Eg(X)$存在，且$Eg(X)=\sum_{i=1}^{\infty}g(x_i)p_i$.

(2) 若X是连续型随机变量，$f(x)$是其密度函数，

则$Eg(X)$存在，且$Eg(X)=\int_{-\infty}^{+\infty}g(x)f(x)dx$

随机变量的数字特征：方差

小象学院 ChinaHadoop.cn

设X为一个随机变量，其数学期望EX存在，

如果$E(X-EX)^2$也存在，

则称$E(X-EX)^2$为随机变量X的方差

记作DX或$VarX$，并称$\sqrt{DX}$为标准差或均方差。

注：方差的大小可以衡量随机变量取值的稳定性。

(1) 若X的取值比较集中，则方差较小；

(2) 若X的取值比较分散，则方差较大；

(3) 若$D(X)=0$，则随机变量X以概率1取常数值，

此时，X就不再是随机变量了.

随机变量的数字特征：方差

小象学院 ChinaHadoop.cn

1. 根据定义 $DX=E(X-EX)^2$

(1) 若X是离散型随机变量， 概率分布为$P\{X=x_i\}=p_i$，

$$DX=E(X-EX)^2=\sum_i (x_i-EX)^2 p_i;$$

(2) 若X是连续型随机变量， $f(x)$为其密度函数，则

$$DX=E(X-EX)^2=\int_{-\infty}^{+\infty}(x-EX)^2 f(x)dx$$

随机变量的数字特征：方差

小象学院 ChinaHadoop.cn

2. 根据公式 $$DX = EX^2 - (EX)^2$$

(1) 若X是离散型随机变量， 概率分布为$P\{X = x_i\} = p_i$,

$$DX = \sum_i x_i^2 p_i - \left(\sum_i x_i p_i \right)^2;$$

(2) 若X是连续型随机变量， $f(x)$为其密度函数，则

$$DX = \int_{-\infty}^{+\infty} x^2 f(x)dx - \left(\int_{-\infty}^{+\infty} xf(x)dx \right)^2$$

随机变量的数字特征

- **例：利用大数据平台计算**
  - 全国和各个城市的人均日通话时间，平均值
  - 全国和各个城市的人通话时间的差异性，方差

  - 分布式计算：各个城市的均值和方差相对容易得到
  - 如何计算700多个城市的均值和方差？

随机变量的数字特征

名称	概率分布	期望 EX	方差 DX
0-1 分布	$P\{X=i\}=p^i q^{1-i}$ $i=0,1$	p	pq
二项分布	$P\{X=k\}=C_n^k p^k q^{n-k}$ $k=0,1,2,L\ ,n$	np	npq
泊松分布	$P\{X=n\}=\frac{\lambda^n}{n!}e^{-\lambda}$ $n=0,1,2,L$	λ	λ
超几何分布	$P\{X=m\}=\frac{C_{N_1}^m C_{N_2}^{n-m}}{C_N^n}$ $m=0,1,L\ ,n$	$n\frac{N_1}{N}$.	$n\cdot\frac{N_1}{N}\cdot\frac{N_2}{N}\cdot\frac{N-n}{N-1}$
均匀分布	$f(x)=\begin{cases}\frac{1}{b-a}, & a\le x\le b,\\ 0, & \text{其他.}\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
指数分布	$f(x)=\begin{cases}\lambda e^{-\lambda x}, & x\ge 0,\\ 0, & x<0.\end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
正态分布	$\varphi(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2

定理设随机向量(X,Y)的函数$Z=g(X,Y)$的数学期望存在，则

（1）如果(X,Y)是离散型随机向量，且联合分布为

$P\{X=x_i,Y=y_j\}=p_{ij},i,j=1,2,\cdots$

$$\text{则}\ EZ=Eg(X,Y)=\sum_{i,j}g(x_i,y_j)p_{ij}.$$

（2）如果(X,Y)是$C.r.v.$，且$(X,Y)\sim f(x,y)$

$$\text{则}\ EZ=Eg(X,Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}g(x,y)f(x,y)dxdy.$$

第二、三讲

概率基础

- 基本概念：随机事件，排列组合
- 概率的定义（概型）
- 条件概率和贝叶斯公式
- 独立试验概型
- 随机变量及其分布
- 随机变量及其数字特征
- 相关性
- 中心极限定理

对随机向量，除关心它每个分量的情况外，还要了解其各分量间的联系，即描述各分量间相互联系的数字特征，协方差就是描述分量间的线性关联程度的数字特征

1.协方差

定义 设(X,Y)是二维随机向量, 如果

$$E\left[(X-EX)(Y-EY)\right]$$

存在, 则称为随机向量(X,Y)的协方差, 记为$\operatorname{cov}(X,Y)$,

即$\operatorname{cov}(X,Y)=E\left[(X-EX)(Y-EY)\right]$.

协方差

（1）定义式 $\operatorname{cov}(X,Y)=E[(X-EX)(Y-EY)]$

若$(X,Y)\sim P\{X=x_i,Y=y_j\}=p_{ij},i,j=1,2,\cdots,$

则 $\operatorname{cov}(X,Y)=\sum\limits_{i,j}(x_i-EX)(y_j-EY)p_{ij}$

若$(X,Y)\sim f(x,y),$

则$\operatorname{cov}(X,Y)=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-EX)(y-EY)f(x,y)dxdy$

(2) $\operatorname{cov}(X,Y)=EXY-EX\cdot EY;$

协方差矩阵

定义 设$(\boldsymbol{X},\boldsymbol{Y})$为一个二维随机向量，且$\boldsymbol{X}$与$\boldsymbol{Y}$的方差均存在，称二阶矩阵

$$V=\begin{pmatrix}\operatorname{cov}(X,X) & \operatorname{cov}(X,Y)\\ \operatorname{cov}(Y,X) & \operatorname{cov}(Y,Y)\end{pmatrix}=\begin{pmatrix}DX & \operatorname{cov}(X,Y)\\ \operatorname{cov}(X,Y) & DY\end{pmatrix}$$

为随机向量$(\boldsymbol{X},\boldsymbol{Y})$的协方差矩阵，简称协差阵

多维协方差矩阵

推广　设$(X_1, X_2, \cdots X_n)$为一个n维随机 向量，X_i的方差$DX_i(i=1,2,\cdots,n)$均存在，

则以$\sigma_{ij}=\mathrm{cov}(X_i, X_j)$为第$(i,j)$元素$(i,j=1,2,\cdots,n)$的矩阵（$\sigma_{ij})_{n\times n}$称为随机向量（$X_1, X_2, \cdots X_n$)的协方差矩阵，简称协差阵，记作$DX$，$X=(X_1, X_2, \cdots X_n)$

$$DX=\begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \cdots & \cdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$$

$\sigma_{ij}=\mathrm{cov}(X_i, X_j)=\mathrm{cov}(X_j, X_i)=\sigma_{ji},\quad i,j=1,2,\cdots n.$

相关系数

小象学院 ChinaHadoop.cn

定义 设(X,Y)为二维随机向量,$\mathrm{cov}(X,Y)$存在,

又$DX>0, DY>0$,称

$$\rho_{X,Y}=\frac{\mathrm{cov}(X,Y)}{\sqrt{DX}\sqrt{DY}}$$

为X,Y的相关系数,简记为ρ.

$$\rho_{X,Y}=\frac{\mathrm{cov}(X,Y)}{\sqrt{DX}\sqrt{DY}}=\frac{EXY-EXEY}{\sqrt{DX}\sqrt{DY}}=E\left[\frac{X-EX}{\sqrt{DX}}\cdot\frac{Y-EY}{\sqrt{DY}}\right]$$

相关系数的意义

$$\left|\operatorname{cov}(X,Y)\right| \le \sqrt{DX}\cdot\sqrt{DY}$$

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sqrt{DX}\sqrt{DY}}$$

$$\Downarrow$$

$$0 \le \left|\rho_{XY}\right| \le 1$$

$\rho_{X,Y} > 0$时，称X与Y正相关;

$\rho_{X,Y} < 0$时，称X与Y负相关;

$\rho_{X,Y} = 0$时，称X与Y不相关.

$\left|\rho_{XY}\right|$的大小反映了X,Y之间的线性关系的密切程度

两个随机变量相互独立表明二者之间没有任何联系，而不相关时，仅表明二者之间不存在线性关系，但不能排除存在其他的非线性关系.

不相关的等价条件

如果DX, DY均存在且为正，则下列四个条件等价：

（1）X与Y不相关

（2）$\mathrm{cov}(X,Y)=0$

（3）$EXY=EXEY$

（4）$D(X+Y)=DX+DY$

第二、三讲

概率基础

- 基本概念：随机事件，排列组合
- 概率的定义（概型）
- 条件概率和贝叶斯公式
- 独立试验概型
- 随机变量及其分布
- 随机变量及其数字特征
- 相关性
- 中心极限定理

中心极限定理

小象学院 ChinaHadoop.cn

- 在客观实际中有许多随机变量，他们有大量的相互独立的随机因素综合影响而成，而其中每一个个别因素在总的影响中所起的作用都是微小的.

定理(林德伯格—勒维） 设$\xi_1,\xi_2,\cdots,\xi_n,\cdots$是一列独立同分布的随机变量，且$E\xi_i=\mu$，$D\xi_i=\sigma^2>0$，$i=1,2,\cdots$，则有

$$\lim_{n\to\infty} P\left\{\frac{\sum_{i=1}^{n}\xi_i-n\mu}{\sqrt{n}\sigma}\le x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt=\Phi_0(x)$$

中心极限定理

$$Y_n = \frac{\sum_{i=1}^{n} \xi_i - n\mu}{\sqrt{n}\sigma} \overset{\text{近似}}{\sim} N(0,1)$$

$$E\left(\sum_{i=1}^{n} \xi_i\right)$$

$$\sqrt{D\left(\sum_{i=1}^{n} \xi_i\right)}$$

标准化

$$\sum_{i=1}^{n} \xi_i \overset{\text{近似}}{\sim} N(n\mu, n\sigma^2)$$

$$\frac{1}{n}\sum_{i=1}^{n} \xi_i \overset{\text{近似}}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

中心极限定理

注：

（1）中心极限定理表明大量独立同分布的随机变量之和都近似服从正态分布。

（2）作用

$\xi_1,\xi_2,\cdots,\xi_n,\cdots$独立同分布，$E\xi_i=\mu, D\xi_i=\sigma^2$

令$\xi=\sum_{i=1}^{n}\xi_i$，$E\xi=E(\sum_{i=1}^{n}\xi_i)=n\mu$　$D\xi=D(\sum_{i=1}^{n}\xi_i)=n\sigma^2$

当n充分大时（$n\geq 50$），$\xi \overset{近似}{\sim} N(n\mu,n\sigma^2)$

由此可近似求出由ξ生成的任何事件的概率

中心极限定理

$$P\{a < \xi < b\}$$

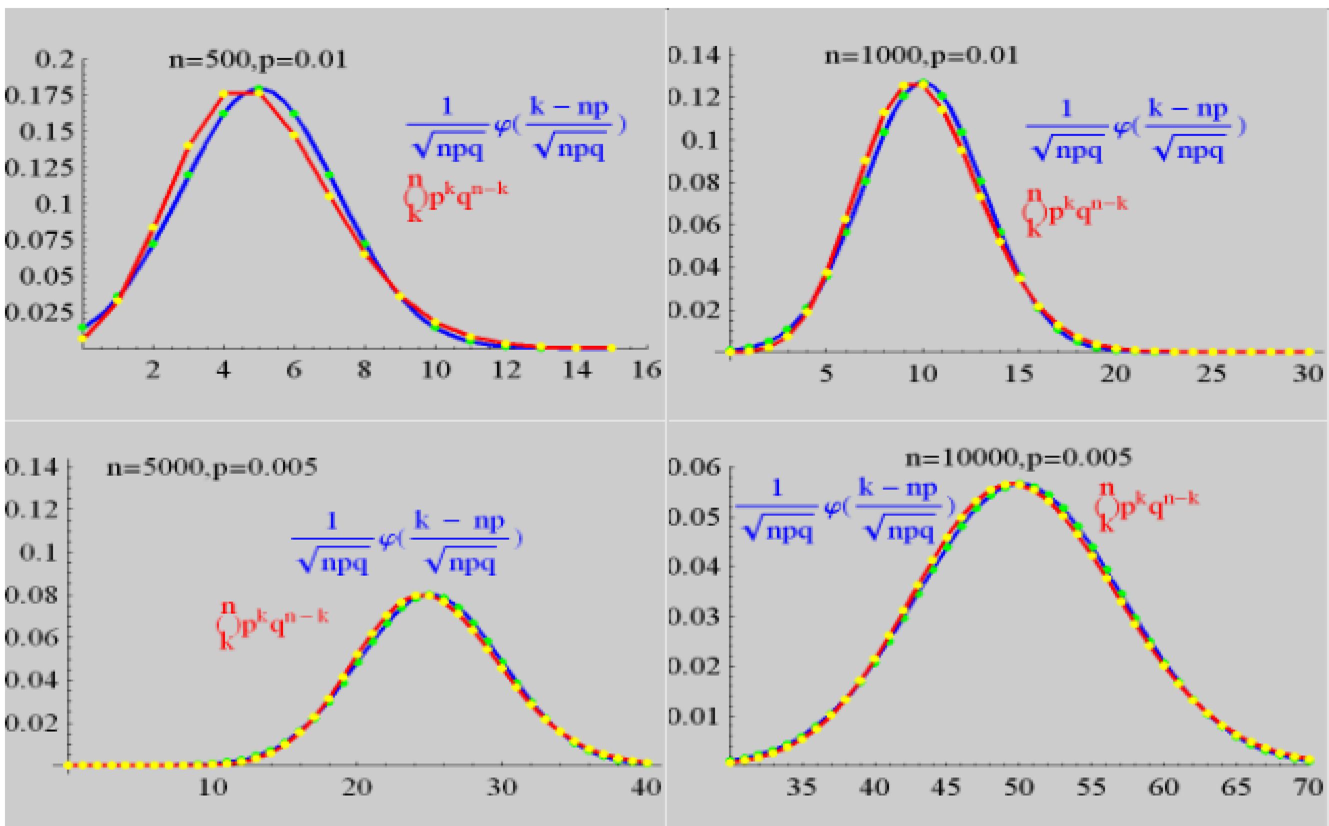
$$= P\left\{\frac{a - n\mu}{\sqrt{n}\,\sigma} < \frac{\xi - n\mu}{\sqrt{n}\,\sigma} < \frac{b - n\mu}{\sqrt{n}\,\sigma}\right\}$$

近似服从$N(0,1)$

$$\approx \Phi_0\left(\frac{b - n\mu}{\sqrt{n}\,\sigma}\right) - \Phi_0\left(\frac{a - n\mu}{\sqrt{n}\,\sigma}\right)$$

中心极限定理

下面的图形表明:正态分布是二项分布的极限分布.



- 基本概念：随机事件，排列组合
- 概率的定义（概型）
- 条件概率和贝叶斯公式
- 独立试验概型
- 随机变量及其分布
- 随机变量及其数字特征
- 相关性
- 中心极限定理

作业：估计性别比例

小象学院
ChinaHadoop.cn

联系我们：

- 新浪微博：ChinaHadoop
- 微信公号：ChinaHadoop
- 网站：http://chinahadoop.cn