第一节　　随机变量

- 随机变量概念的产生
- 引入随机变量的意义
- 随机变量的分类

一、随机变量概念的产生

在实际问题中，随机试验的结果可以用数量来表示，由此就产生了随机变量的概念.

Random variable: A variable that can take different values of the sample space randomly.

随机变量通常用大写字母 X,Y,Z,W,N 等表示

而表示随机变量所取的值时，一般采用小写字母 x, y, z, w, n 等.

二、引入随机变量的意义

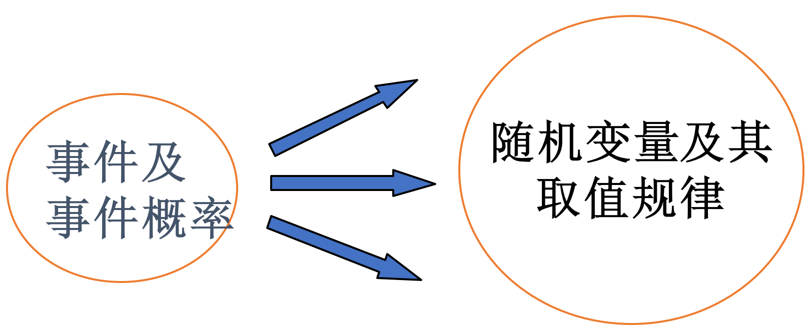
有了随机变量,随机试验中的各种事件，就可以通过随机变量的关系式表达出来.

如：单位时间内某电话交换台收到的呼叫次数用X表示，它是一个随机变量.

事件{收到不少于1次呼叫} $\boxed{\Leftrightarrow}$ $\{X>=1\}$

{没有收到呼叫} $\Leftrightarrow$ $\{X=0\}$

随机变量概念的产生是概率论发展史上的重大事件. 引入随机变量后，对随机现象统计规律的研究，就由对事件及事件概率的研究扩大为对随机变量及其取值规律的研究.



三、随机变量的分类

我们将研究两类随机变量：

随机变量

离散型随机变量

如“取到次品的个数”，
“收到的呼叫数”等.

连续型随机变量

例如，“电视机的寿命”，实际中
常遇到的“测量误差”等.

第二节 离散型随机变量及其分布律

- 离散型随机变量表示方法
- 三种常见分布

一、离散型随机变量表示方法

（1）公式法

$$P\{X=x_k\}=p_k, k=1,2,\cdots$$

（2）列表法

X	x_1	x_2	$\cdots$	x_k	$\cdots$
p_k	p_1	p_2	$\cdots$	p_k	$\cdots$

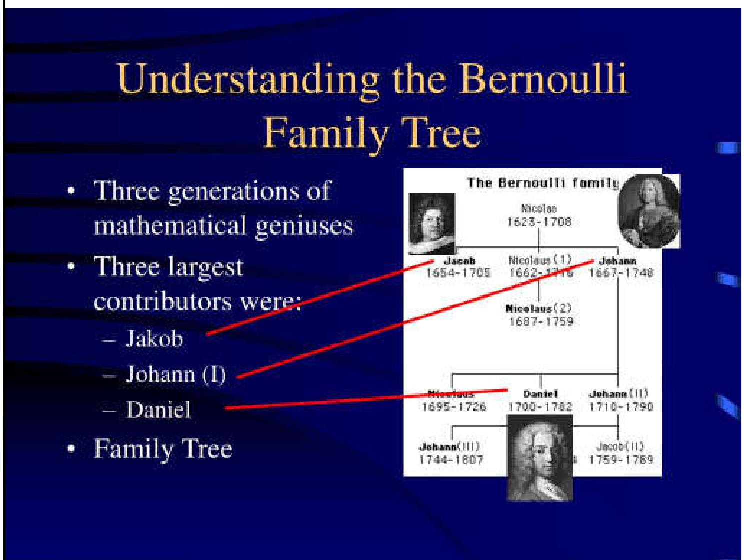
二、三种常见分布

1、（0-1）分布：（也称两点分布或伯努利分布）

随机变量X只可能取0与1两个值，其分布律为：

$$P\{X=k\}=p^k(1-p)^{1-k}, \quad k=0,1 \quad (0<p<1)$$

或 $X \sim \begin{vmatrix} 0 & 1 \\ 1-p & p \end{vmatrix}$



在科学史上，父子科学家、兄弟科学家并不鲜见，然而，在一个家族跨世纪的几代人中，众多父子兄弟都是科学家的较为罕见，其中，瑞士的伯努利(也译作贝努力)家族最为突出。

伯努利家族3代人中产生了8位科学家，出类拔萃的至少有3位；而在他们一代又一代的众多子孙中，至少有一半相继成为杰出人物。伯努利家族的后裔有不少于120位被人们系统地追溯过，他们在数学、科学、技术、工程乃至法律、管理、文学、艺术等方面享有名望，有的甚至声名显赫。

2. 二项分布

将伯努利试验E独立地重复地进行n次，则称这一串重复的独立试验为n重伯努利试验.

“重复”是指这 n 次试验中$P(A)=p$ 保持不变.

“独立”是指各次试验的结果互不影响.

i.i.d RVs

用X表示n重伯努利试验中事件A发生的次数，则

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

易证：（1）$P(X=k) \geq 0$

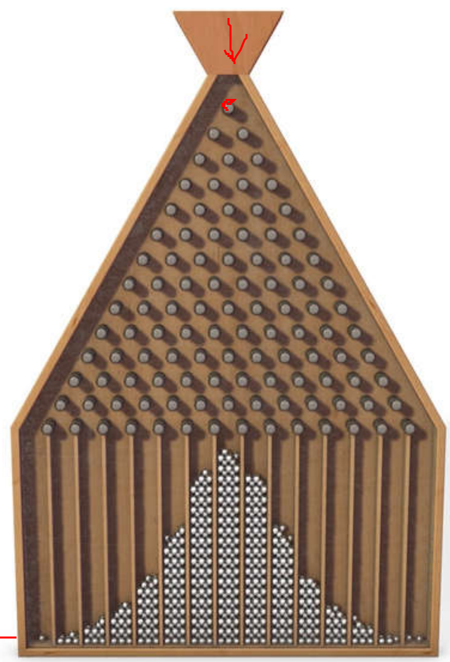
（2）$\sum_{k=0}^{n} P(X=k) = 1$

称 r.v X服从参数为n和p的二项分布，记作

$$X \sim b(n,p)$$

$$E(X) = np, \; D(X) = np(1-p)$$

Galton knocked boards 高尔顿钉板



3. 泊松分布 Poisson distribution

西莫恩·德尼·泊松（Simeon-Denis Poisson 1781～1840）法国数学家、几何学家和物理学家。

设随机变量*X*所有可能取的值为0 , 1 , 2 , … , 且概率分布为：

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

其中 $\lambda>0$ 是常数,则称 *X*服从参数为λ的泊松分布,记作$X\sim\pi(\lambda)$

泊松分布是二项分布**n**很大而**p**很小时的一种极限形式

二项分布是说，已知某件事情发生的概率是**p**，那么做**n**次试验，事情发生的次数就服从于二项分布。

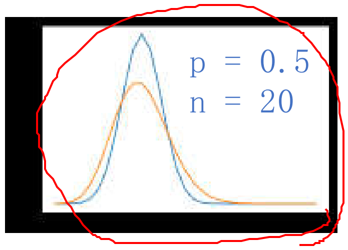
泊松分布是指某段连续的时间内某件事情发生的次数，而且“某件事情”发生所用的时间是可以忽略的。例如，在五分钟内，电子元件遭受脉冲的次数，就服从于泊松分布。（其它例子：一天内，**110**接到报警的数量；一天内，医院来挂号的病人数）

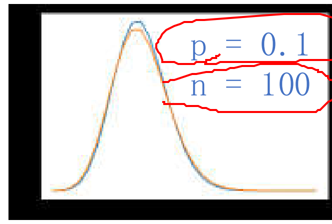
假如你把“连续的时间”分割成无数小份，那么每个小份之间都是相互独立的。在每个很小的时间区间内，电子元件都有可能“遭受到脉冲”或者“没有遭受到脉冲”，这就可以被认为是一个**p**很小的二项分布。而因为“连续的时间”被分割成无穷多份，因此**n**(试验次数)很大。所以，泊松分布可以认为是二项分布的一种极限形式。

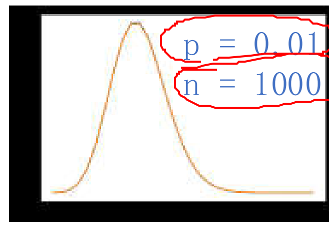
因为二项分布其实就是一个最最简单的“发生”与“不发生”的分布，它可以描述非常多的随机的自然界现象，因此其极限形式泊松分布自然也是非常有用的。

来源：知乎

```
# 泊松分布是二项分布n很大而p很小时的一种极限形式

p = 0.5
n = 20
#p = 0.1
#n = 100
#p = 0.01
#n = 1000

print ("binomial({},{})".format(n,p))
x = scipy.linspace(0,n*p*3,n*p*3+1)
plt.plot(x,scipy.stats.binom.pmf(x,n,p))

print ("poisson({})".format(n*p))
plt.plot(x,scipy.stats.poisson.pmf(x, p*n))
plt.show()
```





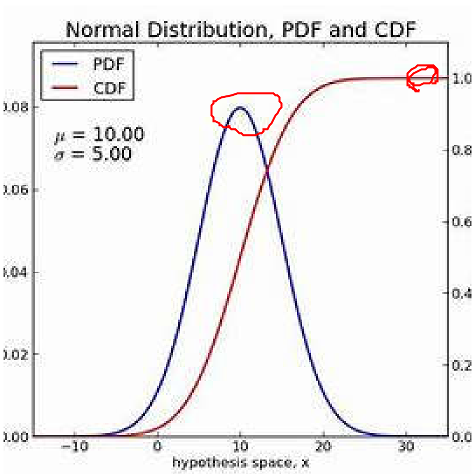


蓝色曲线为b(n,p)，黄色曲线为π(np)

第三节 连续型随机变量及其概率密度

- 连续型随机变量及其概率密度的定义
- 概率密度的性质
- 三种重要的连续型随机变量

一、 连续型随机变量及其概率密度的定义

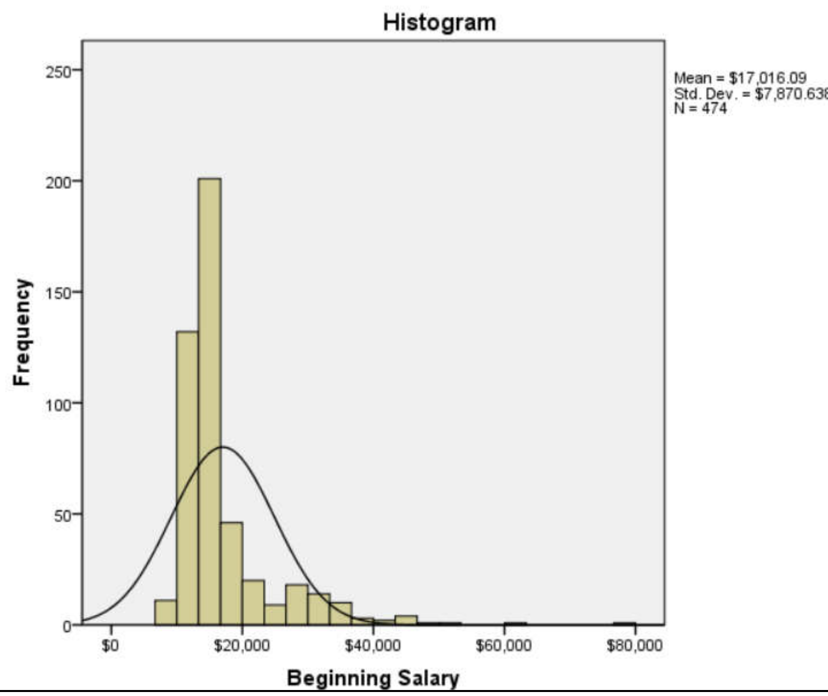


$$F(x) = \int_{-\infty}^{x} f(t)dt = P(X \le x)$$

X为连续型随机变量,称 $f(x)$ 为 X 的概率密度函数，简称为概率密度（PDF）.
$F(x)$ 为 累积密度函数（CDF）

SPSS：分析 ＞ 描述统计 ＞ 频率

频率直方图

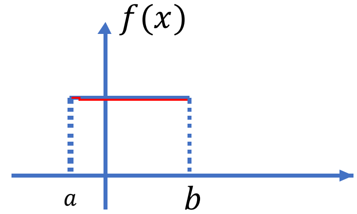


三、三种重要的连续型随机变量

1. 均匀分布

若 $r.v\ X$的概率密度为：

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



则称X在区间(a, b)上服从均匀分布(uniform)，记作

$$X\sim U(a,b)$$

若$X\sim U(a,b)$,

1°.对于长度l为的区间$(c,c+l), a\le c<c+l\le b$,有

$$P\{c<X\le c+l\}=\int_{c}^{c+l}\frac{1}{b-a}dx=\frac{l}{b-a}$$

2°.X的分布函数为：

$$F(x)=P\{X\le x\}=\begin{cases}0, & x<a\\ \dfrac{x-a}{b-a}, & a\le x<b\\ 1 & x\ge b\end{cases}$$

2. 指数分布

若 r .v X具有概率密度

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

其中 $\theta>0$ 为常数，则称 X 服从参数为θ 的指数分布.

X~Expo(θ)

指数分布常用于可靠性统计研究中，如元件的寿命.

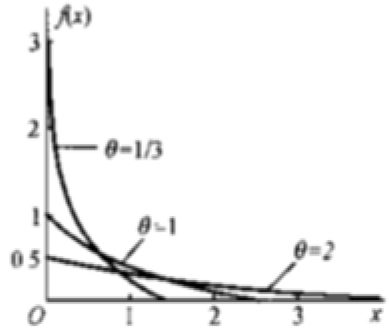
若X服从参数为θ的指数分布,则其分布函数为

$$F(x)=P\{X\le x\}=\begin{cases}1-e^{-x/\theta}, & x>0\\ 0, & 其它\end{cases}$$

服从指数分布的随机变量 X 具有以下有趣的性质：

对于任意 $s,t>0$，有

$$P\{X>s+t \mid X>s\} = P\{X>t\}. \tag{4.9}$$



事实上

$$P\{X>s+t \mid X>s\} = \frac{P\{(X>s+t)\cap(X>s)\}}{P\{X>s\}}$$

$$= \frac{P\{X>s+t\}}{P\{X>s\}} = \frac{1-F(s+t)}{1-F(s)} = \frac{e^{-(s+t)/\theta}}{e^{-s/\theta}} = e^{-t/\theta}$$

$$= P\{X>t\}.$$

性质(4.9)称为**无记忆性**．如果 X 是某一元件的寿命，那么(4.9)式表明：已知元件已使用了 s 小时，它总共能使用至少 $s+t$ 小时的条件概率，与从开始使用时算起它至少能使用 t 小时的概率相等．这就是说，元件对它已使用过 s 小时没有记忆．具有这一性质是指数分布有广泛应用的重要原因．

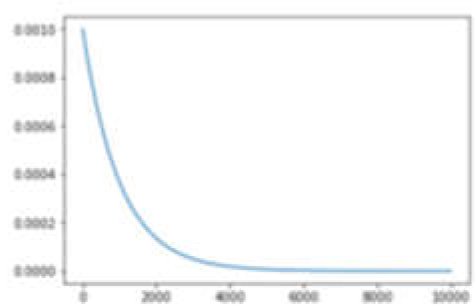
指数分布在可靠性理论与排队论中有广泛的应用．

（选学）寿命为什么符合指数分布？

首先我们考虑一个最简单的事情，一个游戏是每分钟扔一次硬币，扔到正面就继续，扔到反面就**game over.** 这样的话，游戏进行**n**分钟的概率是，这个并不难算对吧。其实这个就是指数分布离散的情况。当时间划分越来越细的时候，这个分布也就趋向于连续的指数分布了。所以指数分布之所以能够刻画寿命，其实是把“死亡”看成了一种均匀的随机事件，比方说一个零件，随时都有一个固定的概率出故障（**e.g.**电容被击穿），那么这个零件的寿命就服从指数分布。类似的，我们举一些其他的例子，比如说打电话的时间服从指数分布，是因为把挂电话看成一个以固定概率随时发生的事情；等公共汽车的时间服从指数分布，是因为把车来看成一个以固定概率随时发生的事情。从这个角度，指数分布的无记忆性也能够解释了。
那么回来看寿命，指数分布并不是描述了我们通常意义下所说的寿命，而是说的是那种以固定概率随机出故障或者出事故的寿命。

来源：知乎

```
from matplotlib import pyplot as plt

# survival game. It has survived n rounds; dies in n+1 round
def survival_dist(n, p):
    return pow((1-p),n)*p;

N = 10000
P = 0.001
x = scipy.linspace(0,N,N+1)
plt.plot(x, survival_dist(x, P))
plt.show()
```



以上代码定义了一个survival game（即每回合有p的死亡率；或电容在单位时间内被击穿的概率）的概率计算函数survival_dist。
取p = 0.001（每回合很小的死亡率），绘制出pmf曲线（离散、等比数组）

3. 正态分布

若连续型 r.v X 的概率密度为

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

其中 μ 和 σ ($\sigma>0$)都是常数,则称X服从参数为μ和 σ 的正态分布或高斯分布.记作

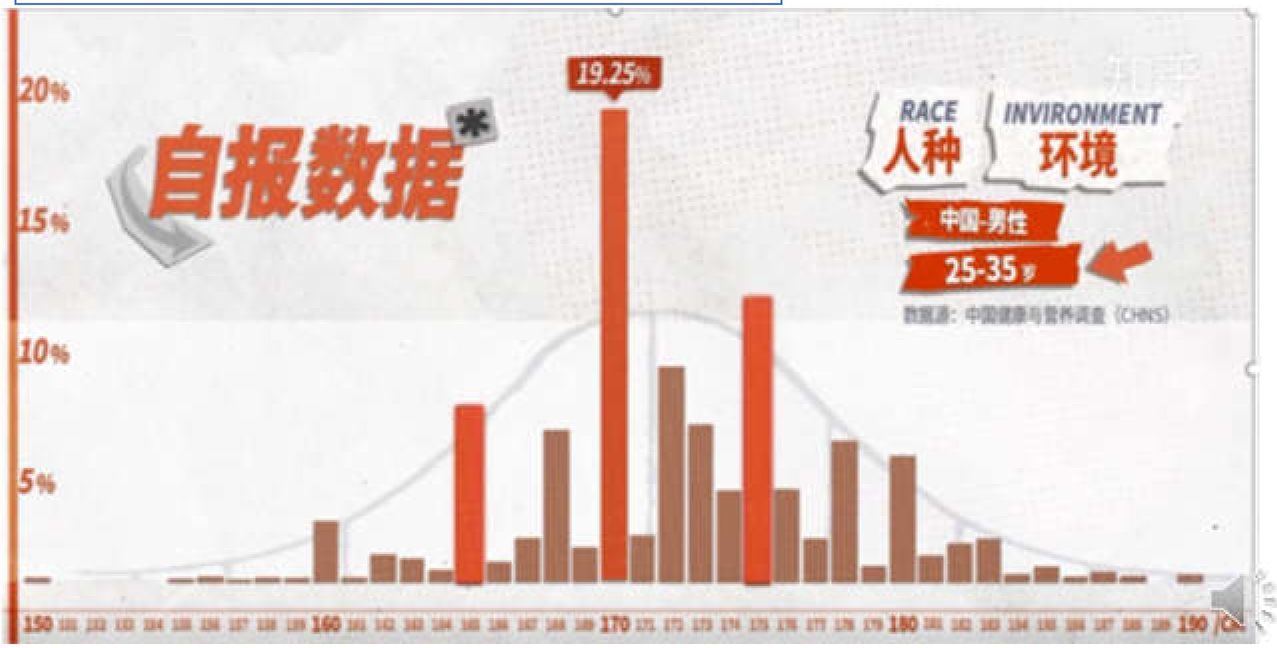
$$X\sim N(\mu,\sigma^2)$$

高斯年冠十九，就在数学上有登峰造极的表现：他突破数学史上两千多年的沉寂， 以直尺与圆规作出正十七边形的图形来。而且他维持如此杰出的研究质量达半个世纪之久。 他的研究范围广泛， 遍及纯数学与应用数学，研究内容新颖、深入。 这使他成为十九世纪科学领域上最突出的人物。 他在曲面学上的研究， 更是导引黎曼创造黎曼几何学

1. 高斯21岁博士毕业
2. 高斯博士毕业后失业三年
3. 高斯因为发明最小二乘法找到了谷神星，才因为这个找到了工作
4. 高斯一代股神，去世时留下200年的薪水
5. 高斯年轻时不教学，因为觉得教学生就是浪费时间，牛逼的学生天然啥都会
6. 黎曼创立了黎曼几何，才拿到个讲师职位。
7. 阿贝尔创立了近世代数，饥寒交迫中死去，死后三天拿到了柏林大学的聘书。
8. 黎曼一生贫困，和导师高斯死在前后脚，他是39岁就饥寒交迫死了。

按照现在的说法，当年真是内卷出银河系了。最后，高斯不愿意儿子学科学，逼着儿子学法律，后来他儿子气不过到美国种地了。

案例：身高调查；结果不符合正态分布？



标准正态分布

$\mu = 0, \sigma = 1$ 的正态分布称为标准正态分布.

其密度函数和分布函数常用$\phi(x)$ 和$\Phi(x)$ 表示：

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

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \quad , \quad -\infty < x < \infty$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \quad (-\infty < x < \infty)$$ 的性质：

(1) $\Phi(0) = \frac{1}{2}$ ；

(2) $\forall x \in R$ ，$\Phi(-x) = 1 - \Phi(x)$ ；

定理1 若 $X \sim N(\mu, \sigma^2)$，则 $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.

标准正态分布的重要性在于，任何一个一般的正态分布都可以通过线性变换转化为标准正态分布.

3 σ 准则

由标准正态分布的查表计算可以求得，

当$X \sim N(0, 1)$时，

$$P(|X| \leq 1) = 2\Phi(1) - 1 = 0.6826$$

$$P(|X| \leq 2) = 2\Phi(2) - 1 = 0.9544$$

$$P(|X| \leq 3) = 2\Phi(3) - 1 = 0.9974$$

这说明，X的取值几乎全部集中在$[-3,3]$区间内，超出这个范围的可能性仅占不到0.3%.

将上述结论推广到一般的正态分布，

$X \sim N(\mu, \sigma^2)$ 时， $Y = \dfrac{X-\mu}{\sigma} \sim N(0,1)$

$$P(|Y-\mu| \leq \sigma) = 0.6826$$

$$P(|Y-\mu| \leq 2\sigma) = 0.9544$$

$$P(|Y-\mu| \leq 3\sigma) = 0.9974$$

可以认为，Y的取值几乎全部集中在

$[\mu - 3\sigma, \mu + 3\sigma]$ 区间内.

这在统计学上称作“3σ 准则”.

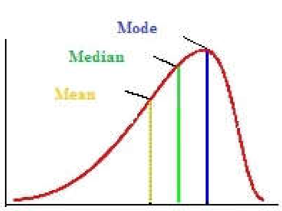
Interval	P(include)	P(exclude)
(-1,1)	0.682689	0.317311
(-2,2)	0.9545	0.0455003
(-3,3)	0.9973	0.0026998
(-4,4)	0.999937	6.33425e-05
(-5,5)	0.999999	5.73303e-07
(-6,6)	1	1.97318e-09

六西格玛(six sigma)是一种改善企业质量流程管理的技术，以“零缺陷”的完美商业追求，带动质量成本的大幅度降低，最终实现财务成效的提升与企业竞争力的突破。

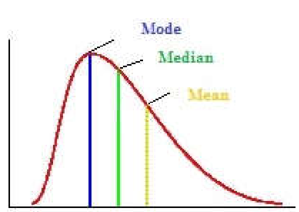
矩(moment)

the zeroth moment is the total probability (i.e. one), the first moment is the mean, the second central moment is the variance, the third central moment is the skewness（偏度）, and the fourth central moment (with normalization and shift) is the kurtosis （峰度）.

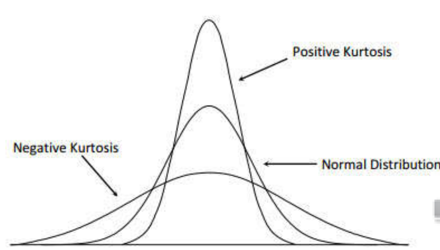
三阶中心矩：偏度（**skewness**）
四阶中心矩：峰度（**kurtosis**）



Left-Skewed (Negative Skewness)



Right-Skewed (Positive Skewness)



类似定义n维随机变量$(X_1, X_2, \ldots, X_n)$的协方差矩阵.

若 $c_{ij} = Cov(X_i, X_j)$

$$= E\{[X_i - E(X_i)][X_j - E(X_j)]\}$$

$$(\, i, j=1,2,\ldots,n \,)$$

都存在, 称

SVD

$$\text{矩阵} \quad C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{pmatrix}$$

为$(X_1, X_2, \ldots, X_n)$的协方差矩阵

小结

离散型随机变量3大分布
连续型随机变量3大分布

TO BE CONTINUED:
3大抽样分布