第三讲 随机变量及其分布

随机变量取哪些值？每个取值对应的概率是多少（即出现的可能性大小）？

一、随机变量

随机现象两大特征：一是发生前不可完全预知，二是多次重复后呈现出统计规律性。

（一）发生前不可预知

你知道下一分钟的股指会涨还是跌吗？你知道自己将来会成为一个什么样的人吗？人生的迷人之处可能就在于，“生活就像一盒巧克力”，不知道下一颗是苦还是甜。要理解这一点，可想象一下，如果一切都是注定的已知的，生活还有什么劲？

案例 1：不能用决定论解释的随机现象

约 200 年前后，Brown（1773-1858）用显微镜观察水中的花粉，发现花粉在水中不停地运动，花粉为什么会不停运动呢？开始，他认为是花粉这种“有机分子”的运动，于是他改用玻璃粉、花岗石、甚至不惜到埃及收集狮身人面像的碎片来做实验。结果发现任何微粒的运动方式本质上都是相同的：（1）向各个方向运行可能性相同，（2）不受过去的影响，（3）不停运动。他认为是水的流动和蒸发导致花粉运行，于是他改为观察在油滴中的运行，结果仍然一样。

当时的科学家认为这种运动受到一个尚未发现的确定性原理支配，如同行星轨道那样存在一个理论解释。物理学家 Maxwell（1831-1879）突破了用牛顿定理的决定论思维来描述每个分子的思维模式，认为气体的性质是整体性质，只能整体进行概率描述，并提出了麦克斯韦-玻尔兹曼分布定律。此后，波兰的 Smoluchowiski（1872-1917）定量解释了布朗运运，认为布朗运动本质上是随机的，任何非随机理论都不能解释它，挑战了因果论。

随机现象被赋予数值就成为随机数。随机数可以这样来理解：设想有一个非常长的数列，想象用一个计算机程序来描述这个数列，如果能描述这个数列的每个可能的程序都至少和数列本身一样长，那么数列是随机的。即随机数列不可压缩。

随机数没有任何规律性，也不可压缩。我们有时需要用到随机数，简单的可以抛硬币或者抓阄，但是复杂一点呢？实际上广泛使用的随机数仍然是程序产生出来的，称为“伪随机数”，看起来毫无规律，不知“内情”的人也无法预期，但与真正的随机数相比，这些数实际上是可以用程序压缩的。

上机 1：伪随机数

```
clear

di   uniform()   //display 显示结果，uniform()产生 0-1 均匀分布

di   uniform()    //重复执行，得到另一个服从均匀分布的随机数

/*如果要生成一位数的随机数（即 0，1，2，3，4，5，6，7，8，9），可以取小数点后第一位数，通常用下面的命令*/

di   int(10*uniform())   //先将原伪随机数乘 10 倍，再取整 int()

*试一试：产生一个骰子，即取值为 1，2，3，4，5，6

/*也可以同时生成多个随机数（相当于抽取样本），然后将该随机数赋给某个变量。要注意的是，伪随机数实质上是按照一定的规律生成的。如果给定基于生成伪随机数的初始数值（即 set seed #），则对相同初始数值，生成的伪随机数序列完全一样。*/

set   obs   5        //指定生成 5 个观察值

g   x1=uniform()      //将 5 个随机数赋值给变量 x1

g   x2=uniform()     //将另 5 个随机数赋值给变量 x2

list    //显示结果，注意到 x1 与 x2 不一样

set   seed   1234        //指定初始值，如果不指定，默认为 123456789

g    y1=uniform()
```

```
set seed 1234
g y2=uniform()
g y3=uniform()
list //注意到 y1 与 y2 一样，但均与 y3 不同
set seed 5634
g z1=uniform()
set seed 1234
g z2=uniform()
list //注意到 z2 与 y1,y2 一样，但 z1 与 z2 不同
```

（二）多次重复后呈现出统计规律性

谁能 100%地事前确定一枚骰子在某次投掷中的点数呢？尽管掷多次，次次不同；但是还是可以在结果里面看到某种规则模式，而且只有在重复许多次后，这个模式才会清楚浮现，这个了不起的事实，就是概率概念的基础。

“短期机遇现象无法预测，但是长期下来，会呈现有规则且可预测的模式。”

这便是随机现象的第二个特点：**在多次重复后会呈现出统计规律性**。

这种在个别试验中其结果呈现出不确定性；在大量重复试验（观察）中其结果又具有统计规律性的现象，便称之为**随机现象**。**随机试验通常**具有下面三个特点：（1）可以在相同的条件下重复地进行;（2）每次试验的可能结果不止一个，并且事先能明确试验的所有可能结果;（3）进行一次试验之前不能确定哪一个结果会出现。

案例 2:数学家掷硬币的记录

法国的布丰掷硬币 4040 次，得出硬币出现正面的频率为 50.69%。英国的皮尔逊做过两组上万次重复试验，得出正面出现的概率分别为 50.16%和 50.05%。

科学家	掷币次数	正面出现频率
布丰	4040	0.5069
棣莫根	4092	0.5005
杰万斯	20480	0.5068
皮尔逊	24000	0.5005
罗曼诺夫斯基	80640	0.4979
费勒	10000	0.4923

统计规律是大量重复后呈现出来的，并不适用于少量观察或小样本的情形。比如在小样本情形下，服从均匀分布的随机现象看起来并不是均匀的。

上机 2：飞镖试验与癌症丛集

真正的随机现象看起来并不随机。癌症丛集这种现象非常有名，假设随机掷出 16 只飞镖到一个正方形，它们插中正方形中任何一个地方的概率相同，现在把这个正方形分成 16 个更小的正方形，我们预期每个小正方形平均会有一支飞镖在上面——但这只是平均值而已。16 只飞镖恰好分别插中 16 个不同的正方形，这样的概率非常低。常见的是一些格子里会有一支以上的飞镖，如果不出现癌症丛集，将是极为罕见的事。可是一些报纸宣称某个地方高压线幅射太强，造成的癌症显著增多。

模拟：生成 1-16 的 16 个自然数，观察最多一个格中有多少个飞镖

```
clear      //将系统空置
set obs 16
set seed 1234
g x=1+int(16*uniform())    //生成 16 个取值 1-16 的随机数
tab x,p    //x 的频率，第 16 格中有 7 支飞镖，有 9 格中没有一支飞镖
set obs 10000    //设定产生 10000 个观察值
g y=1+int(16*uniform())    //生成 10000 个取值 1-16 的随机数
```

tab y,plot //每个格中的飞镖差别不大

中新泽西州彩票两次的概率有多少？1.7*10e-9，但这种事情就发生在亚当期身上。然而，某个人在某个地方，以完全未指明的方式，碰到那么幸运的巧事的概率，居然高达 1/30。

人类不是被设计来理解事物的，我们只是被设计来求生和繁衍，但为了求生存，我们必须夸大某些事情的概率，例如可能影响我们存活的事件发生的概率。大脑对生命危险特别在意的人容易生存下来，因此他们的基因遗传下去。但是偏执狂也不能过头，否则必须付出太高代价，反而成为缺点。

（三）随机模拟

随机变量是随机试验的结果的数量化，是随机试验的结果与实数间的一一映射，比如，设随机变量 x 为世界上下一个出生的婴儿的性别，这个结果只有两个，男和女，当为男时，定义 x=1，当为女时，定义 x=0。

可以这样来想像：x 是实数轴上变幻不定的一个数，一会儿是 1，一会儿是 0。在观察之前知道 x 只可能取值 0 和 1，但特定的某次观察之前不能确定会取哪个结果，即无法预知是男还是女。但多次重复观察，则男孩出现的机会约为 0.51，女孩约为 0.49。

上机 3： 模拟

利用随机数字表或者电脑软件中的随机数发生器，来模仿机遇现象，叫模拟（simulation）。只要你自己试试模拟随机现象几次，就会加强对概率的理解，比读很多页的数理统计和概率论的文章还有用。一旦有了可靠的概率模型，模拟还是找出复杂事件发生概率的有效工具。蒙特卡罗仿真法是研制原子弹时在 Los Alamos 实验室发展出来的，常用来模拟随机现象。

下一个出生的是男孩还是女孩

一个事件在重复结果中发生的比例，迟早会接近它的概率，所以模拟可以对概率做适当的估计。如法国数学家拉普拉斯对伦敦、彼得堡、柏林和全

法国的大量人口资料进行研究，发现男婴出生率总在一个数左右波动，这个数大约是 22/43。另一位统计学家克拉美引用瑞典 1935 年的官方统计资料，发现女婴出生的频率稳定在 0.482 左右。下面的命令可以得到生男还是生女的一个模拟结果

```
di  uniform()<0.482
```

神秘信件能骗到多少人？

元旦时你收到一封匿名信，说这个月股市会上涨，但你不以为意。到了 2 月 1 日，你又接到另一封信，说股市将下跌，这一次，又给那封信说中了；3 月 1 日再接到信，情形一样，7 月，你对那位匿名先生的先见之明很感兴趣，对方邀你投资某个海外基金，于是你把全部积蓄全部拿出来投资，两个月后，那些钱有如肉包子打狗。你伏在邻居的肩膀上号啕大器，他告诉你，他也接过两封这种神秘信，但寄到第二封就停了。他说，第一封信的预测正确，第二封不正确。这是怎么一回事？原来，那些骗子从 1 万个人名中寄出后市看长的信给其中一半的人，看跌的给另一半的人，一个月后，将有 5000 人接到的信预测正确，然后再针对这 5000 人如法炮制，如此直到名单上剩下 500 人，其中有 200 人会上当，骗子只花了几千元的邮资，却能赚进数百万元。假设收信人只有看到 10 次都预测正确才会投资，我们来模拟 10 次均正确的概率。

```
clear
set  obs 10000
g  x=uniform()>0.5     //假设上涨和下跌的概率相等，均为 0.5
keep  if  x==1     //若预测错误，不再寄信
 *将上述程序再重复 9 次，共 10 次
 forvalue   i=1/9    {      //forvalue 为循环命令
 g   x`i'=uniform()>0.5     //如有不能执行，请英文半角重输单引号
 keep   if   x`i'==1        //如有不能执行，请英文半角重输单引号
```

```
 }
count      //计算最后还剩下多少个 10 次均预测成功的收信者
```

射击求圆周率

如何近似计算圆周率π？在一个边长为 2 的正方形内画个圆，然后举枪对其胡乱射击，用圈内的弹孔数除以全部弹孔数再乘以 4 即可。

```
clear
set  obs  100000
g  x=2*uniform()-1      //生成一个（-1，1）的均匀分布随机变量 x
g  y=2*uniform()-1      //生成一个（-1，1）的均匀分布随机变量 y
g  r=((x^2+y^2)<1)*4   //平方和小于 1 时 r=4，点位于单位圆内，否则 r=0
sum  r                  //注意 r 的均值即为π的近似
```

二、分布函数

不确定性事件和风险事件不同，前者取哪些值不确定，不同取值出现的可能性也未知，而风险事件是指取值及其对应的出现概率均已知，但特定试验或观察之前无法确定取哪个值的情形。

（一）频率与概率

案例 3：如果试很多次，会发生什么？

1986 年 1 月 28 日，挑战号航天飞机发射后不久就爆炸了。总统特别委员会开始调查:像这样的发射失败的机会有多大？工程师说，大约是 1%的机会;管理部门说，大概 10 万次才会发生一次。物理学家费曼就问："你们的意思是说，如果连续 300 年每天发射一次火箭，预期只会失败一次？"。300 年约等于 109500 天。费曼简短的对话中做了两件很重要的事:(1)把模糊的个人意见，改用具体的意向来表达，也就是同一件事重复做许多次的概念：如果我们发射了非常多的航天飞机，那失败的频率大概会是多少？(2)通过和真实生

活相联系，让人们对于尝试某件事 10 万次的意义更容易了解，即每天试一次，共试 300 年。

对某随机现象观察 N 次，特定结果出现的总次数记为 n，称 n/N 为频率，频率随着特定的试验或观察而波动，不同的观察得到不同的频率。

上机 4：频率与概率的关系

在人大校门口统计进校园人士的性别 x，假设人大女生（x=1）占 70%，统计 1000 人。

```
clear
set obs 1000
g  N=_n                  //生成一个序列号 N,取值从 1 到 1000
g  x=uniform()<0.7       //性别女＝1，男＝0
g  n=sum(x)              //依次计算前 N 个进校门中的人中女生的总数
g  y=n/N                 //在前 N 个进校门的人中女生出现的频率 n/N
line  y  N,yline(0.7)    //yline(0.7)绘出 y=0.7 的直线
```

可以总结抽象出一个表达式：

$$\hat{\Pr}(x = x_i) = \frac{n_i}{N}$$

频率：特定结果和该结果出现的可能性之间的一一映射（可表达为一种函数）

概率：在同一试验条件下，当独立试验次数趋于无穷时的频率，它是一个常数 p_i，该规律由贝努里大数定律所证明。

$$f(x_i) = \Pr(x = x_i) = \lim_{N\to\infty}\frac{n_i}{N} = p_i$$

概率为常数是由随机现象的内在性质决定的。例孟德尔的豌豆试验，新生儿性别，硬币，骰子（如赌片中的“老千”骰子被做过手脚后与标准的骰子就不一样了）。

案例 4：孟德尔的试验与遗传规律

Mendel 按颜色（黄/青）和形状（圆/有角）把豌头分为四类：他提出的遗传学理论认为，遗传因子有显性和隐性之分，如果从父系和母系接受的显性因子会使隐性因子不起作用。黄色是显性的，圆也是显性的。因此理论上四种豌豆的概率为：

随机结果	黄圆	青圆	黄角	青角
随机变量 X	$X_1=1$	$X_2=2$	$X_3=3$	$X_4=4$
理论概率 p=f(X)	9/16=.5625	3/16=.1875	3/16=.1875	1/16=.0625
实际观察 n_k	315	108	101	32
实际观察频率 n_k/N	0.5665	0.1942	0.1817	0.0576

概率与频率：一个事件的概率是由事件本身特性所决定的客观存在。频率稳定性是大量重复试验的统计结论。

Bernoulli（1654-1705）在《猜度术》一书中提出了大数定理，认为独立随机事件的频率随着试验次数趋于无穷，会逼近概率，已知概率能推测频率，已知频率能推测概率。Venn（1834-1923）是概率的频率观点的发明人之一。

对于数列“565656…”6 的出现概率和频率都是 50%，但如果我们知道了 5，则下一个数字为 6 的概率是 100%而不是 50%，这是条件概率。

（二）经验分布

假设 x 为一个随机变量，并且有一个容量为 n 的随机样本，每个 x_i 都是 x 的一个独立的实现，则对应这个样本的经验分布定义为一个离散分布，给每个点 x_i 以 1/n 的权重。其经验概率密度函数为：

$$\hat{f}(x_i) = \hat{\Pr}(x = x_i) = \frac{1}{n}$$

经验累积分布函数，可以表示为：

$$\hat{F}(x_i) = \hat{\Pr}(x < x_i) = \frac{1}{n}\sum_{i=1}^{n} I(x < x_i)$$

其中 I(.)为示性函数，当自变量为真时取值 1，否则取值 0。EDF 具有阶梯函数的形式。每一阶梯的宽度为相邻的两个 X_i 值的差，阶梯之间的跃度为 1/n。

（三）概率密度

上述经验分布，当 N 趋于无穷时即得到概率密度函数。概率密度函数 f(X)建立起随机变量每个可能值及该值出现可能性之间的映射关系。

对离散随机变量，表现为取特定值的可能性。如生男孩（x=1）与生女孩（x=0）.

对连续随机变量，取特定值的概率为零，概率密度函数可以理解为 X 落在某点 X_0 的一个领域内的可能性。

$$f(x_0) = \lim_{\delta \to 0}(x_0 - \delta < x < x_0 + \delta)$$

连续变量是客观存在的，但实际测量，无论多精细，总是离散的。所以最好把取众多值的随机变量看做是连续的，如价格，高考分数。

对连续性变量，特定值出现的概率为零，讨论它没有意义，我们关注的是取值落在一定范围内的概率。如湖北高考分数的统计，当说到考 600 分的学生时，600 分实际上是一个四舍五入的数，它在区间(599.5,600.5)之间，而且它的概率很小。

案例 5：看盘次数与心痛频率

你去炒股，预期报酬率为 15%，波动性为每年 10%，换算之后，任何一年赚钱的概率为 93%。但在不同的时间尺度下赚钱的机率如下：

时间尺度	一年	一季	一月	一天	一小时	一分钟	一秒
赚钱机率	93%	77%	67%	54%	51.3%	50.17%	50.02%

由于每分钟看投资组合，设一天观察 8 小时，每天会有 241 分钟心情愉快，239 分钟不愉快；一年中有 60688 分钟愉快，60271 分钟不愉快。由于不愉快的程度大于愉快的程度，所以时刻盯着屏幕反而给自己制造了很大的情绪赤字。如果每月看一次，则有 67%的月份赚钱，一年只心痛 4 次，快乐 8 次。如果一年看一次，则在 20 年中，有 19 次惊喜，只有一次不愉快。

（四）累积分布

累积分布函数 F(x)是经验累积分布函数在 N 趋于无穷时的情形。F(x)为随机变量 x 小于给定值 x_0 的可能性，在数轴上看，是 x 落在 x_0 左边的比率。记为

$$F(x_0) = \Pr(x < x_0)$$

连续随机变量的分布函数建立起随机变量小于特定值 X_0 的可能性与特定值之间的一一对应关系。

$$dF(x) = f(x)$$

EDF 称之外经验分布函数，是从一个样本来估计分布密度函数和累积分布函数的方法。可以证明，当样本趋于无穷时，EDF 即为 CDF。

（五）分位数

定义：设随机变量 x 的分布函数 F(x)，对给定的实数 α(0<α<1)，如果实数 F_α 满足：

$$\Pr(x > F_\alpha) = \alpha$$
$$F(F_\alpha) = 1 - \alpha$$
$$F^{-1}(1 - \alpha) = F_\alpha$$

则称为随机变量 x 的分布的水平 α 的上侧分位数。或分布函数 F(x)的水平 α 的上侧分位数。

上机 5：婴儿身高分布

*真实数据：初生婴儿的身高服从正态分布，下面是 3000 个婴儿的身高

```
use "http://wps.pearsoned.co.uk/wps/media/objects/16103/16489878/data3eu/birthweight_smoking.dta",clear

sum birth

hist birth,norm kdensity
```

```
*hist 给出变量 z 直方图，norm 为正态分布，kdensity 为密度函数的核估计
    cumul birth,gen(F)        //生成婴儿身高累积经验分布函数值 F
    line   F   birth,sort          //绘制 ECDF 的图
    su   birth,d                     //注意 5%分位数和中值
    line   F   birth,sort yline(0.05) xline(2410)   //5%的分位点
    line   F   birth,sort yline(0.5) xline(3420)   //中值，即 50%分位点
```

三、随机变量的函数

随机变量 x 的函数 y=g(x)是按一定的运算规则对 x 进行的一种转化，转化后 y 仍然为随机变量，并且 y 与 x 可一一对应（每个 x 取值对应一个 y 值，反过来不成立），不合并可能出现的相同取值，其出现的概率亦可一一对应，但若出现多个 x 对应同一个 y 值时，将相应的概率合并相加后，将导致看似不同的分布函数。

随机变量 x 的取值	-1	0	1	2
x 各取值出现的概率 p=f(x)	0.1	0.3	0.4	0.2
x 的函数变换 $y=g(x)=x^2$	1	0	1	4
Y 取值出现的概率 f(y)=f(x)	0.1	0.3	0.4	0.2

y 与 x 的取值和概率可一一对应（上表），合并相同 y 值后概率可能不同（下表）

$y=g(x)=x^2$	0	1	4
f(y)	0.3	0.1+0.4	0.2

四、数字特征

随机变量 x 是变幻不定的，但它的期望 Ex 必为常数：因为 x 的所有可能取值 x_i 确定，每个取值对应的概率 p_i 也确定，而期望只不过是所有可能取值依概率加权求和。

x	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n
$Ex = \sum_{i=1}^{n} x_i p_i = \mu$	$x_1 p_1$	$x_2 p_2$	…	$x_n p_n$
$y = g(x)$	$y_1 = g(x_1)$	$y_2 = g(x_2)$	…	$y_n = g(x_n)$
$Ey = \sum_{i=1}^{n} g(x_i) p_i$	$y_1 p_1$	$y_2 p_2$	…	$y_n p_n$
$g(x) = (x - Ex)^2$ $Var(x) = E(x - Ex)^2 = \sigma^2$	$(x_1 - \mu)^2 p_1$	$(x_2 - \mu)^2 p_2$		$(x_n - \mu)^2 p_n$
$g(x) = (x - Ex)^k$				

和同学抛硬币赌 1000 元，则要么你的口袋里一毛钱都没有，要么放着 2000 元，能够想象你有 1000 元吗（数学期望值）？

由随机变量 x 经函数 g(.)变换成的 y 也是随机变量，每个 x 的取值均对应有一个 y 的取值，因而出现的概率也相同，对 y 依此概率加权求和得到 y 的期望 Ey。实际上 x 的方差等高阶矩便可视为 x 的函数变换后的期望。方差 Var(X)也为常数：因为 x_i-μ 为确定值，其平方也确定，相对应的概率 p_i 确定，因此依概率加权求和必然得到一个常数。类似地随机变量 x 的 k 阶原点矩和中心矩亦为常数。

定理：随机变量 x 的 t 阶矩存在，则其 s 阶矩存在（0<s<t）.

$$
\begin{aligned}
& E|x|^s = \int |x|^s \ f(x)dx \\
& = \int_{|x|^s \le 1} |x|^s \ f(x)du + \int_{|x|^s > 1} |x|^s \ f(x)dx \\
& \le \int_{|x|^s \le 1} f(x)dx + \int_{|x|^s > 1} |x|^t \ f(x)dx \\
& \le P\{|x|^s \le 1\} + E|x|^t < \infty
\end{aligned}
$$

五、若干重要分布函数

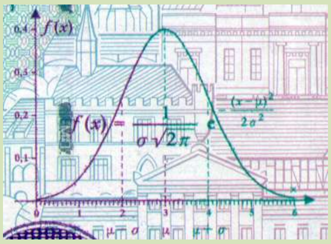
分布函数是无数伟大的数学家如高斯，泊松等在总结客观随机现象的基础上得出来的理论函数公式。这些公式把握了随机现象及其出现可能性之间的内在联系及本质规律性。分布函数理论用少数参数来刻画复杂，不易把握的随机现象。

分布名称	参数	概率函数	期望	方差
0-1 分布	p	$f(x)=p^x(1-p)^{1-x} \quad x=0,1$	p	$p(1-p)$
伯努利分布 B(n,p)			np	np(1-p)
泊松分布	λ	$f(x)=\dfrac{\lambda^{-x}e^{\lambda}}{x!} \qquad x=0,1,2\ldots$	λ	λ
均匀分布	a,b	$f(x)=\begin{cases}\dfrac{1}{b-a}, a<x<b \\ 0,其他\end{cases}$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
指数分布	λ	$f(x)=\begin{cases}\lambda e^{-\lambda x}, x\ge 0 \\ 0, \quad x<0\end{cases}$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$
正态分布	μ,σ	$f(x)=\dfrac{1}{\sqrt{2\pi\sigma}}e^{\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
卡方分布	n	略	n	$2n$
T 分布	n	略	0	$\dfrac{n}{n-2}$
F 分布	m,n	略	$\dfrac{n}{n-2}$	略

案例 6：300 万套军装的快速生产

1917 年美国仓促决定赴欧洲参战，当时面临一个需要解决的例是：三百万参战大军的军装、军鞋应按什么尺寸规格才能在短期内最快地加工出来。美国电话研究所的休哈特（W. A. Shewhart）提出一个方案。他通过抽样调查，发现军衣、军鞋的尺寸规模分布与正态分布曲线形状类似，按照正态分布的统计规律，他提出按照两头小中间大的排列规则，按高矮、胖瘦分十档进行加工制作。美国国防部采用了他的建议，结果与参战军人体形基本吻合，全部分配完毕，及时保证了军需供应。

德国1991年至2001年间发行的的一款10马克的纸币上印着高斯(Carl Friedrich Gauss, 1777-1855)的头像和正态密度曲线，而1977年东德发行的20马克的可流通纪念钢镚上，也印着正态分布曲线和高斯的名字。



正态分布，为众多微小因素累和而成，如误差，身高，体重，学习成绩等。对数正态分布，为众多微小因素累积而成，如工资，收入，财富等。如收入服从对数正态分布，要征收所得税或者要救助贫困人群，能收到多少税或要投入多少财政资金救助，可以从对数正态分布公式中来进行推导计算。

标准正态分布的密度函数为：

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$$

上机 6：各类分布

处理随机变量的几个步骤：（1）确定随机变量 x，（2）确定所有可能取值，（3）确定各个取值出现的可能性，即概率。（4）建立随机变量取值与其出现概率之间的函数关系。

两点分布

所有非此及彼的选择，如新生婴儿的性别、产品的质量是否合格、外出不外出，买不买房子等。

泊松分布

服从泊松分布的随机变量 X 取 0 或正整数，其他实例如每天光顾小店的顾客数，出现车祸次数，到医院就诊次数。

```
clear
set obs 10000
g y= rpoisson(4)     //生成服从泊松分布的随机变量 y,参数 λ 为 4
tab y,plot           //y 的频率和分布
su y                 //均值约为 4
```

正态分布

```
clear
set obs 10000
g z=invnormal(uniform())     //得到服从标准正态分布的随机变量 z
g zs=rnormal()               //与上一个命令等价
g zr=rnormal(10, 4)          //均值 10、方差 4 的正态分布随机变量
hist zr,bin(100) norm        //画出直方图并配上标准正态分布曲线
```

已知分布曲线，可以随机变量在特定区间出现的概率

例：人的智商（I.Q.）得分一般服从均值为 100，标准差为 16 的正态分布，随机抽取一人，他的智商在 100-115 之间的概率是多少（也即占人口多大比例？）

di **normal***((115-100)/16)*- **normal***((100-100)/16)*

反过来，已知随机变量的分布曲线和出现概率，亦可求出区间临界点。

例：设在注册会计师的会计科目考试中，其通过率只有 10%，从历年的经验来看，分数的均值和标准差分别为 72 和 13。如果分数近似正态分布，为了获得顶部 10%的分数并通过考试所需要的最小分数是多少？

di **invnormal***(0.9)*13+72* *//顶部 10%等价于 F(a)=Pr(x<a)=1-0.1*

**inv*normal()与 normal 互为反函数。

*在标准正态分布中，出现小于-1.96 的随机数的概率是.025

di **normal***(-1.96)*

*标准正态分布的水平 α=0.05 的双侧分位数(因对称，单侧水平为 0.025)为 1.96;水平 α=0.05 的上侧分位数为 1.64

di **invnormal***(0.95)*

卡方分布

*卡方分布是若干个独立标准正态分布的平方和。

clear

set **obs** 10000

g xsqr= **(rnormal())^2** //标准正态分布平方和为卡方分布 xsqr

g chi=**rchi2**(1) //直接生成服从卡方分布的随机变量 chi

g chi3=**rchi2**(3) //生成服从自由度为 3 的卡方分布的随机变量 chi3

tw **(kdensity** xsqr**)** **(kdensity** chi **)** **(kdensity** chi3**)**

/*自由度为 10 时，累积分布为 0.95 所对应的随机变量为 18.31，即 10 个独立的标准正态分布随机变量平方和大于 18.31 的可能性为 0.05.*/

```
di chi2(10,18.31)

di invchi2(10,0.95)          //反函数
```

t 分布

```
*t 分布是标准正态分布与卡方分布除自由度并取平方根后两者之比

clear

set obs 10000

g t2=rt(2)      //生成自由度为 2 的 t 分布随机变量

*当自由度 n 趋于无穷时，t 分布近似于标准正态分布

tw (function y=tden(1,x), range(-4 4)) (function y=normalden(x), range(-4 4))(function y=tden(30,x), range(-4 4))

di invnormal(0.95)    //正态分布 1.64

di invttail(1000,0.05)   //自由度较大时，t 分布逼近正态分布 1.64
```

F 分布

```
*F 分布为两个服从卡方分布随机变量分别除以其自由度之后的比

di F(10,5,4.735)    /*服从分子自由度为 10，分母自由度为 5 的 F 分布，
                    小于 4.74 的概率为 0.95*/

di invF(10,5,0.95)          //F 分布的逆函数，得到临界点 4.735

di 1/(invFtail(10,5,0.95))   //交换自由度,相当于交换分子分母
```

定理：X 为连续随机变量，若其累积分布函数 F(x)为严格单调递增的函数，则 Y=F(X)服从[0，1]上的均匀分布。

证明：由累积分布函数定义

$$\Pr(X < x) = F(x)$$

因 $F(.)$ 严格单调递增，其反函数存在，记为 $X = F^{-1}(Y)$

$$\begin{aligned} F(y) = \Pr(Y < y) = \Pr[F(X) < y] \\ &= \Pr\{F^{-1}[F(X)] < F^{-1}(y)\} \\ &= \Pr[X < F^{-1}(y)] \\ &= F[F^{-1}(y)] \\ &= y \end{aligned}$$

上式正是[0，1]均匀分布的累积分布函数。利用该性质，可以生成服从各种分布的随机变量，先生成一个服从[0，1]的均匀分布 Y，然后用种 CDF 函数的反函数作用于 Y，得到 $X=F^{-1}(Y)$.

上机 7：从均匀分布生成各类分布的随机变量

```
clear
set  obs  1000
g  y=uniform()            //生成[0,1]均匀分布
g  x=invnorm(y)        //求累积标准正态分布反函数得标准正态随机变量
line  y  x,sort           //得到标准正态分布曲线
kdensity  x
```

六、 随机向量

（一）定义

❖ K 维随机向量的定义

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$$

上机 8：随机向量

如一个人的特征，性别、年龄、教育水平、工作经验、收入等。

反复执行

```
clear
mata
uniform(5,1)
end
```

/*每执行一次，得到一个服从均匀分布的 5 维随机向量的实现。

如果对该服从均匀分布的 5 维随机向量观察 9 次，得到 9 个样本值（实现），可以同时得到 9 个 5 维随机向量的实现值*/

```
clear
mata
uniform(5,9)
end
```

/*还可以定义随机矩阵，矩阵中的每一个元素都是一个随机变量，不同的实现要用不同的矩阵来表达。*/

```
clear
mata
for    (i=1;i<=5;i++)    {
X`i'=uniform(2,3)      //若不执行，请在英文半角状态重新输入单引号
X`i'                   //若不执行，请在英文半角状态重新输入单引号
}
end
```

- ❖ 二维随机向量

二维随机变量（x ,y）’，如观察下一个从教室门口经过的人，定义他的身高为 x，体重为 y，经过的人不同，其身高体重变来变去。再如门卫查看证件，是否人民大学学生，定义 x 定义为是否人大学生，y 定义为性别，则有

x\y	女	男
人大学生	0．4	0．3
非人大学生	0．1	0．2

对离散的二维随机变量，其概率密度函数为一个常数，表示两个维度分别取特定值时的概率，为 f(x,y)=P(x=女，y=人大学生)=0.4

离散的二维随机变量可以被想象成：在水平地面上打上方格，然后把沙子一颗颗扔上去，堆积起来，最后，数一数落到每一格中的沙子的个数，除以总的沙子数，得到落在每一格的概率。

相应地连续随机变量的概率密度函数为落在平面上的一个点(x,y)的无穷小领域内的概率。

二维随机变量的累积分布函数类似地定义，F(x,y)=P(X<x, Y<y)。是落在平面上点（x,y）的右下方的概率。

$$F(x,y)=\int_{-\infty}^{y}\int_{-\infty}^{x} f(s,t)dsdt$$

（二）随机向量的期望与方差

随机向量的每个元素不仅有其自身的方差，两两之间也可能存在相关性，通常用协方差表示。将方差和协方差组合到一个矩阵之中，称为随机向量的方差阵。定义：

$$E\boldsymbol{x} = \begin{bmatrix} Ex_1 \\ Ex_2 \\ \vdots \\ Ex_k \end{bmatrix}$$

$$\operatorname{var}(\boldsymbol{x}) = E[\boldsymbol{x} - E\boldsymbol{x})(\boldsymbol{x} - E\boldsymbol{x})'] = \begin{bmatrix} \operatorname{var}(x_1) & \operatorname{cov}(x_1, x_2) & \cdots & \operatorname{cov}(x_1, x_k) \\ \operatorname{cov}(x_2, x_1) & \operatorname{var}(x_2) & \cdots & \operatorname{cov}(x_2, x_k) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(x_k, x_1) & \operatorname{cov}(x_k, x_2) & \cdots & \operatorname{var}(x_k) \end{bmatrix}$$

若 x 的各个元素都不相关，则 Cov(x)是一个对角阵。而且各个元素方差相同，则

$$\operatorname{var}(\boldsymbol{x}) = \sigma^2 I_k$$

上机 9：均值、方差

```
*生成特定相关结构的随机向量
clear
set    obs    10000
mat   v=(4,2.4\2.4, 4)          //协方差阵
mat   m=(5,10)                  //均值阵
corr2data   x1   x2, means(m) cov(v) cstorage(full)    //生成相关向量 x,y
mata
x=st_data(.,.)        //数据变矩阵
mean(x)               //样本均值
variance(x)           //样本方差阵
st_matrix("v")
end
```

（三）随机向量的函数

设 x 为某随机向量，y=Ax，其中 A 为非随机矩阵，则 y 也为随机向量，其期望和方差为

$$
\begin{aligned}
y &= Ax \\
Ey &= AEx \\
Var(y) &= Var(Ax) = AVar(X)A' \\
Var(y) &= E\{E[Ax - E(Ax)][(Ax - E(Ax)]'\} \\
&= E\{AE[x - E(x)][A(x - Ex)]'\} \\
&= AE[x - E(x)](x - Ex)'A \\
&= AVar(x)A'
\end{aligned}
$$

随机向量的方差阵恒为半正定对称阵，由于上式中 y 的任何一个元素均为 x 的某种线性组合，设 c 为任意的非随机向量,y=c’x

$$0 \leq Var(\mathbf{y}) = Var(c'\mathbf{x}) = c'Var(\mathbf{X})c$$

上机 10：随机向量的函数

```
clear
mata
X=uniform(5,9)        //服从均匀（0，1）分布的五维随机向量，取 9 个样本
EX=mean(X')           //样本均值，默认列为维度，行为观察值，故需转置
DX=variance(X')            //样本方差阵
A=(1,1,1,1,0\2,0,3,5,1)        //线性变换阵（加权）
Y=A*X            //Y 为 A 和 X 的线性组合，Y 为二维随机向量，有 9 个观察值
Y
(mean(Y'))'            //Y 的均值,由于系统默认列为维度，故需两次转置
A*EX'            //利用公式 EY=A*EX 计算出的均值
variance(Y')          //Y 的协方差阵
A*DX*A'            //利用公式 D(Y)=AD(X)A’计算出来的协方差阵
end
```

（四）二次型

运用矩阵乘法，将下面的矩阵积表示为二次型

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2x_1^2 + 2x_1x_2 + 2x_1x_2 - x_2^2$$

由此可见，二次型是一个数，随机变量的二次型则是一个随机数

$$z = \mathbf{x}'\mathbf{A}\mathbf{x}$$
$$Ex = \boldsymbol{\mu}$$
$$Var(\mathbf{x}) = \Sigma$$
$$Ez = \boldsymbol{\mu A \mu} + tr(A\Sigma)$$

证明：

$$\boldsymbol{x}'A\boldsymbol{x} = (\boldsymbol{x}-\boldsymbol{\mu}+\boldsymbol{\mu})'A(\boldsymbol{x}-\boldsymbol{\mu}+\boldsymbol{\mu})' = (\boldsymbol{x}-\boldsymbol{\mu})'A(\boldsymbol{x}-\boldsymbol{\mu}) + (\boldsymbol{x}-\boldsymbol{\mu})'A\boldsymbol{\mu} + \boldsymbol{\mu}'A(\boldsymbol{x}-\boldsymbol{\mu}) + \boldsymbol{\mu}'A\boldsymbol{\mu}$$

$$E[(\boldsymbol{x}-\boldsymbol{\mu})'A\boldsymbol{\mu}] = E[\boldsymbol{\mu}'A(\boldsymbol{x}-\boldsymbol{\mu})] = 0$$

$$\begin{aligned} E[(\boldsymbol{x}-\boldsymbol{\mu})'A(\boldsymbol{x}-\boldsymbol{\mu})] &= E\{tr[(\boldsymbol{x}-\boldsymbol{\mu})'A(\boldsymbol{x}-\boldsymbol{\mu})]\} \\ &= E\{tr[A(\boldsymbol{x}-\boldsymbol{\mu})'(\boldsymbol{x}-\boldsymbol{\mu})]\} \\ &= tr\{E[A(\boldsymbol{x}-\boldsymbol{\mu})'(\boldsymbol{x}-\boldsymbol{\mu})]\} \\ &= tr\{A[E[(\boldsymbol{x}-\boldsymbol{\mu})'(\boldsymbol{x}-\boldsymbol{\mu})]\} \\ &= E(A\Sigma) \end{aligned}$$

七、多元分布

（一）二元正态分布

二元正态分布的概率密度函数为

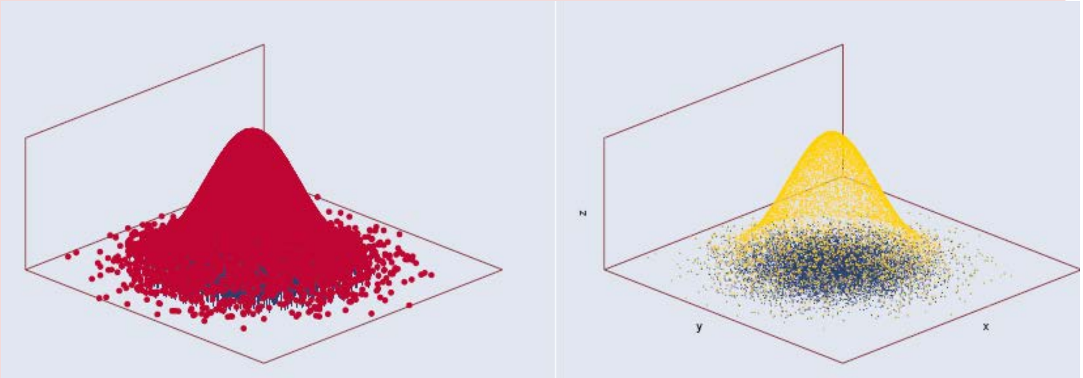
$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}e^*$$

$$*=\frac{-1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right]$$

显然，二元正态分布由 5 个常参数决定。只要这几个参数一定，则随机有序数组(X,Y)在（x,y）的无穷小邻域内出现的概率就是确定的 f(x,y)。

上机 11：二元正态分布图

```
ssc install scat3                        //在运行 sct3 之前需要先下载该命令
drawnorm x y, n(10000) clear             //产生 10000 个标准二维正态随机向量
g   f=0.5/_pi*exp(-0.5*(x^2+y^2))        //计算其概率密度
*绘出概率密度的三维图
scat3  x  y  f,  mcolor(gold)  shadow(msize(0))
```

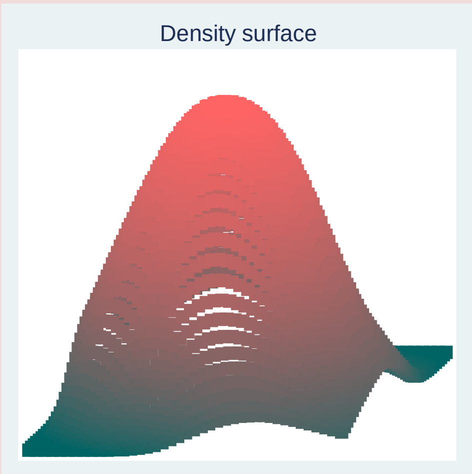


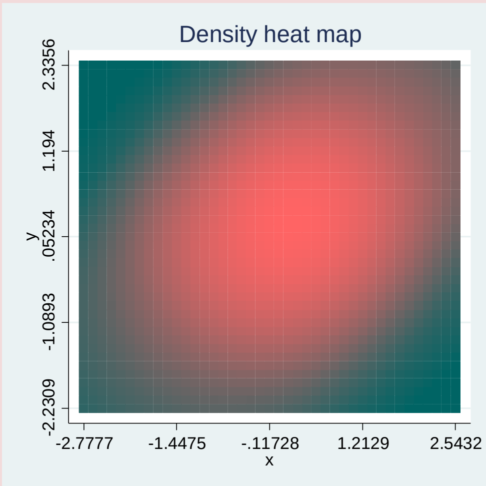
ssc install tddens //三维密度函数图

matrix C = (1, .9 \ .9, 1)

drawnorm x y, n(100) corr(C) clear

tddens x y,s





（二）多元正态分布

多元正态分布的密度公式为

$$f(y)=\frac{1}{(2\pi)^{\frac{n}{2}}(\det\Sigma)^{\frac{1}{2}}}e^{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)}$$

多元正态分布密度由均值和协方差阵所决定。因此，已知随机向量服从正态分布，又知道了均值和方差阵，即可决定其分布。

标准多元正态向量的期望为 0,方差为 I,因此分布密度公式

$$f(y)=\frac{1}{(2\pi)^{\frac{n}{2}}}e^{-\frac{1}{2}y'y}=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y_i^2}=\prod_{i=1}^{n}f(y_i)$$

定理 3-1：多元正态分布的性质

（1）若 y 服从多元正态分布，则 y 中的每个元素 y_i 都是正态分布的。任何 k 个元素也服从多元正态分布（再生性：蚯蚓被断成几截，各截都能成为新的蚯蚓）

（2）y 中任意两个元素 y_i 和 y_j 相互独立的充分必要条件是它们不相关

（3）正态随机向量的线性组合仍然服从正态分布.标准正态分布随机向量可以经线性组合得到任意的多维正态向量,反之,**任意的多维正态向量可以线性组合为标准的多维正态分布向量**

$$\begin{aligned}
&\boldsymbol{x} \ \rightarrow \ N(0, I) \\
&\boldsymbol{y} \ \rightarrow \ N(\boldsymbol{\mu}, \Sigma) \\
&\boldsymbol{y} \ = \ \Sigma^{\frac{1}{2}}\boldsymbol{x} + \boldsymbol{\mu} \\
&\boldsymbol{z} \ = \ A\boldsymbol{y} \ \rightarrow \ N(A\boldsymbol{\mu}, A\Sigma A')
\end{aligned}$$

（三）卡方分布

自由度为 k 的卡方分布为独立的 k 个服从标准正态分布随机变量的平方和，用向量形式来表达，k 维标准正态分布随机向量 $\boldsymbol{x}$

$$\boldsymbol{x} \ \rightarrow \ N(0, I_k)$$

$\boldsymbol{x}$ 的内积即为 k 个平方和，因此根据定义，它服从卡方分布，自由度为 k

$$\boldsymbol{x}'\boldsymbol{x} \ \rightarrow \ \chi_k^2$$

因为

$$\boldsymbol{x}'\boldsymbol{x} \ = \ x_1^2 + x_2^2 + \cdots + x_k^2$$

对于 k 维非标准多元正态分布，通过标准化后，其内积亦服从卡方分布

$$\boldsymbol{x} \ \rightarrow \ N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \to \chi_k^2$$

因为

$$\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu}) \to N(0, I_k)$$

另外一个常用的重要定理是，由多元标准正态随机变量和幂等对称阵构成的二次型服从卡方分布，其自由度为该幂等对称阵的秩（迹）。

$$\boldsymbol{x}'A\boldsymbol{x} \to \chi_r^2 \qquad (if \qquad \boldsymbol{x} \to N(0, I), rankA = r, A = A', AA = A)$$

证明：

$$\because\ rankA = r, A = A', AA = A$$
$$\therefore\ A = H\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}H'$$
$$\boldsymbol{x}'A\boldsymbol{x} = \boldsymbol{x}'H\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}H'\boldsymbol{x}$$
$$\boldsymbol{y} = H'\boldsymbol{x}$$
$$\boldsymbol{x} \to N(0, I_k)$$
$$E\boldsymbol{y} = E(H'\boldsymbol{x}) = 0$$
$$Var(\boldsymbol{y}) = Var(H'\boldsymbol{x}) = H'Var(\boldsymbol{x})H = H'IH' = I$$
$$\boldsymbol{y} \to \mathbf{N(0, I)}$$
$$\boldsymbol{x}'A\boldsymbol{x} = \boldsymbol{y}'\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}\boldsymbol{y} = y_1^2 + y_2^2 + \cdots + y_r^2 \to \chi_r^2$$

上机 12：多元正态分布

```
clear
mat  m=(3,4,5,0)                                    //均值阵
mat  v=(9,9.6,0.01,0\9.6,16,12,0\0.01,12,25,0\0,0,0,1)    //方差阵
drawnorm  x1  x2  x3  x4,n(10000)  means(m)  cov(v)  clear
mata
x=st_data(.,.)                     //导入服从多元正态分布的随机矩阵
```

```
n=rows(x)                         //观察值数
m=st_matrix("m")                  //均值阵
v=st_matrix("v")                  //方差阵
z=J(n,1,.)
for (i=1;i<=n;i++) {
z[i]=(x[i,.]-m)*invsym(v)*(x[i,.]-m)'        //(x − μ)'Σ^{-1}(x − μ) → χ²_k
}
st_store(.,st_addvar("double","z"),z)       //导出生成的服从卡方随机变量
end
g  chi=rchi2(4)                              //用命令直接生成卡方随机变量
tw (kdensity  z) (kdensity chi)              //绘图
```

（四）F 分布

根据定义，F 分布为两个经自由度调整后的标准正态分布的平方和之比，而多元标准正态分布可表达成随机向量的内积形式，因此

$$\boldsymbol{x} \to \boldsymbol{N(0, I)}$$

$$\frac{\boldsymbol{x}_m'\boldsymbol{x}_m / m}{\boldsymbol{x}_n'\boldsymbol{x}_n / n} \to F_{m,n}$$

与上一节的卡方分布推导类似，由两个幂等对称阵构成的二次型服从 F 分布

$$\frac{\boldsymbol{x}'A\boldsymbol{x} / p}{\boldsymbol{x}'B\boldsymbol{x} / q} \to F_{p,q} \qquad (rankA = p, rankB = q, A = A', AA = A, B = B', BB = B)$$

八、主成分分析

给定 k 阶随机向量 **x**，设 k 阶随机向量 **y** 等于 **x** 与其特征向量的积，于是向量 **x** 和 **y** 的方差阵满足

$$\mathbf{y} = H'\mathbf{x}, HH' = I$$
$$Var(\mathbf{y}) = Var(H'\mathbf{x}) = H'Var(\mathbf{x})H$$
$$Var(\mathbf{x}) = \Sigma = H\Lambda H'$$
$$Var(\mathbf{y}) = H'H\Lambda H'H = \Lambda$$

经过主成分变换后的 **y**，若按特征值大小重新排序，前面的成分方差大（即变异度很大），后面的方差越来越小，即变异度很少，甚至越来越接近于零，即几乎没有变异。由于没有变异，就无法利用其信息来区别不同的观察对象。

比如我们要区别不同的人，只能根据人的种种特征来区别，人有多个特征相当于随机变量的维数，有些维度的变异度不大，好比大家都穿校服就不能靠衣服来区别是张三李四，如果头发有长有短可以靠头发来区别，但如果都是光头就没办法根据头发长短来识别人了，而且这些不同的维度特征具有内在的相关性（比如身高和体重正相关），主成分可以将较多的特征综合到较少的几个显著特征上，并且将不同维度的相关性消除，新生成的随机向量 **y** 的各个维度是不相关的（其方差阵的非对角线元素为零）。

下面的程序先求出 n 个变量 **x** 的样本方差阵，将样本方差阵分解为特征值和特征向量（系数），用特征向量对原变量加权求和得到各主成分 **y**。

上机 13：主成分分析

```
clear
set more off
set obs 100
gen x=uniform()
g y=x
pca y x,cov     //若有 cov 选项则采用原始数据而非标化数据
predict z                  //主成分得分也用原始数据而非标化数据
putmata X=(y x),replace  //将 STATA 数据导入 mata 矩阵 X
mata
C=variance(X )                              //求方差阵而非相关系数阵
symeigensystem(C, V=., l=.)                 //X 的方差阵分解
```

```
l                                            //特征值
round(V,0.0001)                              //特征向量组成的矩阵
p=X*V        //特征向量与原始数据相乘
end
getmata (p*)=p
g    np=sqrt(2)*(x+y)/2
list   y   x   z*   p*   np   in   1/10
tw   (sc   y   x)   (sc   p1 p2)
```

用特征向量旋转空间，使得原来第一象限中的 45 度角射线旋转到 y 轴上，x 轴全变为 0，不再起作用，被降维了。

```
*用 auto.dta 数据做主成分分析
sysuse auto,clear
pca weight mpg length,cov      //若有cov选项则采用原始数据而非标化数据
predict y1-y3                  //主成分得分也用原始数据而非标化数据
putmata X=(weight mpg length),replace   //将STATA数据导入mata矩阵X
mata
C=variance(X )                           //求方差阵而非相关系数阵
symeigensystem(C, V=., l=.)              //X的方差阵分解
l                                        //特征值
round(V,0.0001)                          //特征向量组成的矩阵
p=X*V        //特征向量与原始数据相乘
end
```

```
getmata (p*)=p

list   y*   p*    in   1/10
```

采用原始数据方差阵进行分解，方差将由绝对值较大的哪个维度所决定，如一个变量是人均 GDP，另一个变量是总量 GDP，则人均 GDP 几乎对总方差没有什么贡献。有时，不同维度单位都不相同，此时绝对值大小并不能说明什么，因为相互不可比，因此，多数主成分的分析不用原始数据，而是先进行标准化变换或采用相关系数矩阵进行分解时，这种主成分分析实际上相当于假设 **x** 的各分量等权。

将 **x** 的每个变量标准化有两种方法，一种是取其均值和标准差，然后减均值再除以标准差，另外一种是减最小值，然后再除以极差（即最大值减最小值的差）。用第一种方法标准化后的数据求主成分，相当于直接对其相关系数矩阵 R 求主成分，即对 R 的分解。这正是 STATA 软件中默认的处理方式。此时命令 pca 后不带任何参数，实质上是默认为 corr,即用相关系数矩阵进行分解。但在进行 predict 时，需要将标准化后的变量与系数矩阵相乘，而不可以用原始变量与之相乘。

```
clear

set more off

sysuse auto,clear

pca weig mpg leng                 //主成分分析，用相关系数矩阵求主成分

predict y1-y3                     //用的是X标准化后的值与特征向量相乘

egen sw=std(weight)               //将变量标准化，即减均值后除以标准差

egen sm=std(mpg)

egen sl=std(length)

putmata X=(sw sm sl),replace

mata

C=correlation(X )                 //求X的相关系数阵

symeigensystem(C, V=., l=.)       //相关系数阵的分解
```

```
l                                   //特征值，pca的第一张表
round(V,0.0001)                     //特征向量，pca的第二张表
p=X*V                               //主成分得分p，必须与标准化后数据相乘
end
getmata (p*)=p
list  y*  p*  in  1/10
```

比较是否标准化原始数据后得到的两种主成分的结果，可以发现特征值有巨大的差异，前者最大的特征值达到604495，后者仅为2.7.