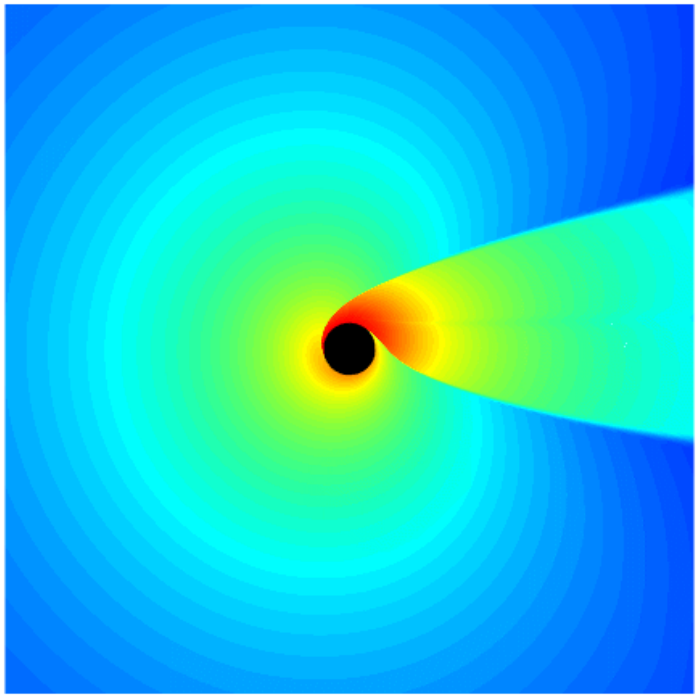
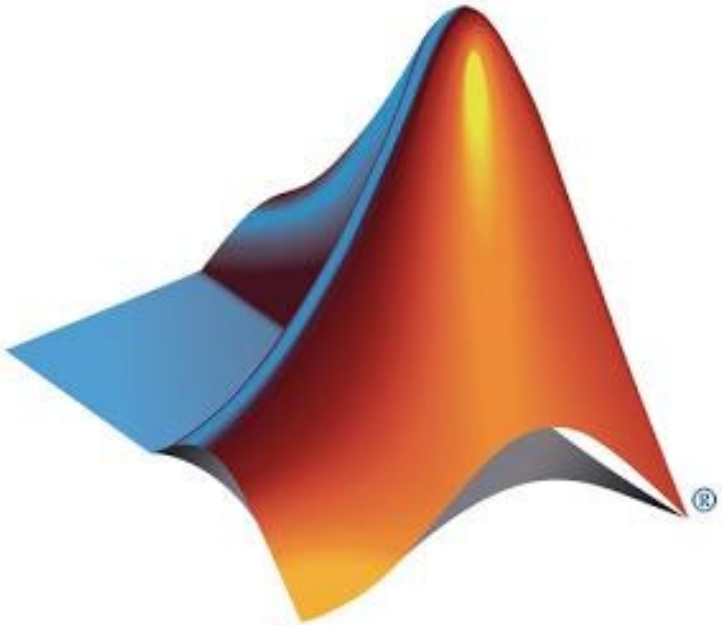
数学实验
Mathematical Experiments





实验十二：
概率统计实验
Probability and Statistics

◆实验的背景

所有现象的"因"和"果",即"条件"和"结果"之间在客观上都存在着一定的规律.这种规律通常可以分成两类:一类是确定性的规律,另一类是非确定性规律.对于确定性的系统,当已知条件是充分的时,那么实验的结果也是确定的,即在每一次试验以前,可以预见试验产生的后果.但若条件不充分,那么就无法预测试验的"结果",这时就产生了“因果律的破缺”的随机现象.

◆实验的背景

随机现象在实践中是大量遇到的,虽然无法由"因"预测"果",但是当进行大量重复试验时，因果之间仍会呈现一种统计规律.当然,整个解题思路和我们已经习惯的确定性方法有很大区别.概率方法建立在"重复试验"的基础之上,统计规律只有在大量重复后才会呈现出来,诸如随机变量、分布、均值方差等概念无一不体现了重复的思想.用Matlab软件进行统计实验,可以方便地重现这一思想,更好地理解和掌握概率统计的内容.

◆实验的背景

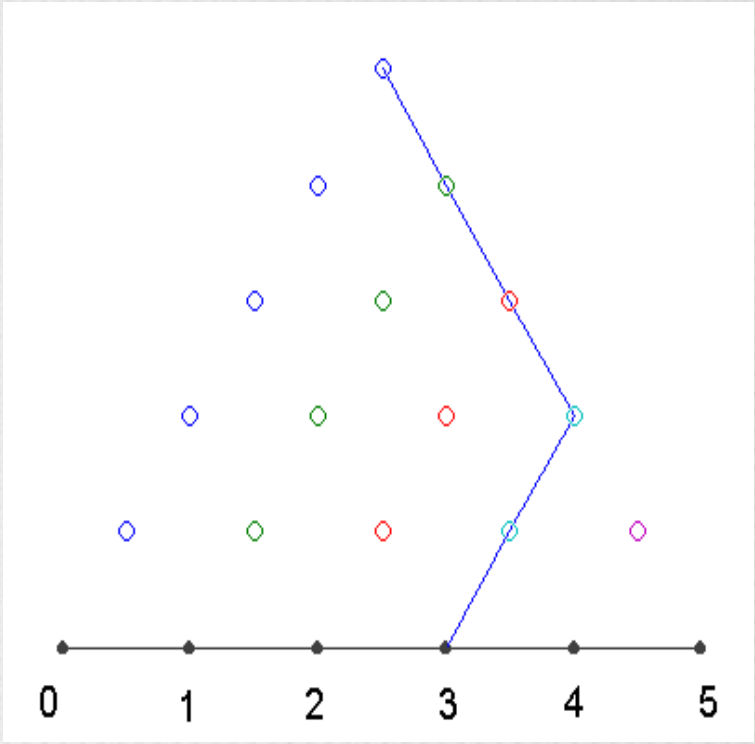
本次实验，我们希望通过若干有趣的例子来理解统计学的一些基本思想以及如何用Matlab解决相关问题。

实验1：Galton钉板实验

Galton钉板实验：实验与观察

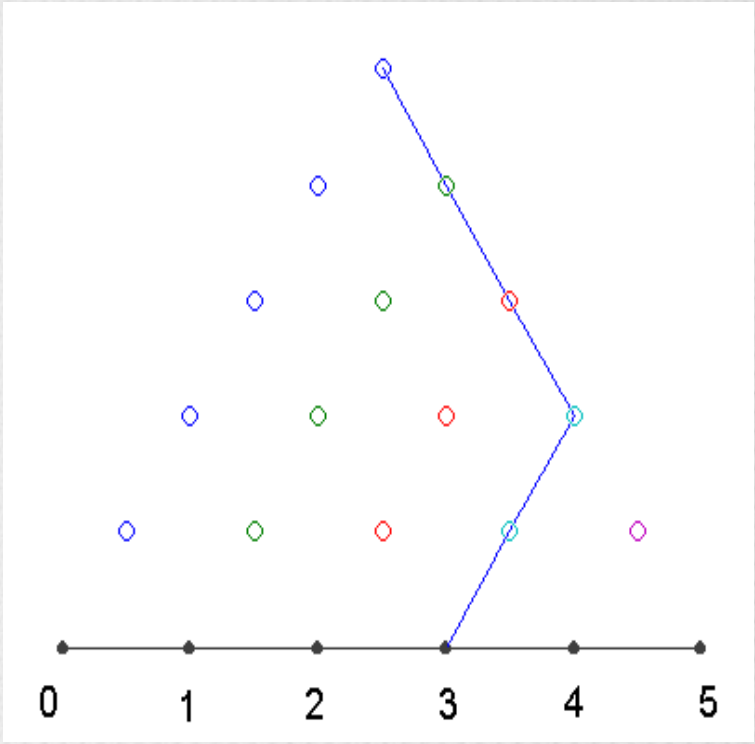
Galton钉板试验是由英国生物统计学家Galton设计的.在一板上钉有n排钉子，如图所示，其中n=5，即有5排钉子的情况.图中15个圆点表示15颗钉子，在钉子的下方有6个格子，分别编号为0,1,2,3,4,5.

自Galton钉板的上方扔进一小球,任其自由下落，在下落的过程中当小球碰到钉子时，从左边落下与从右边落下的机会相等.碰到下一排钉子时又是如此.



Galton钉板实验：实验与观察

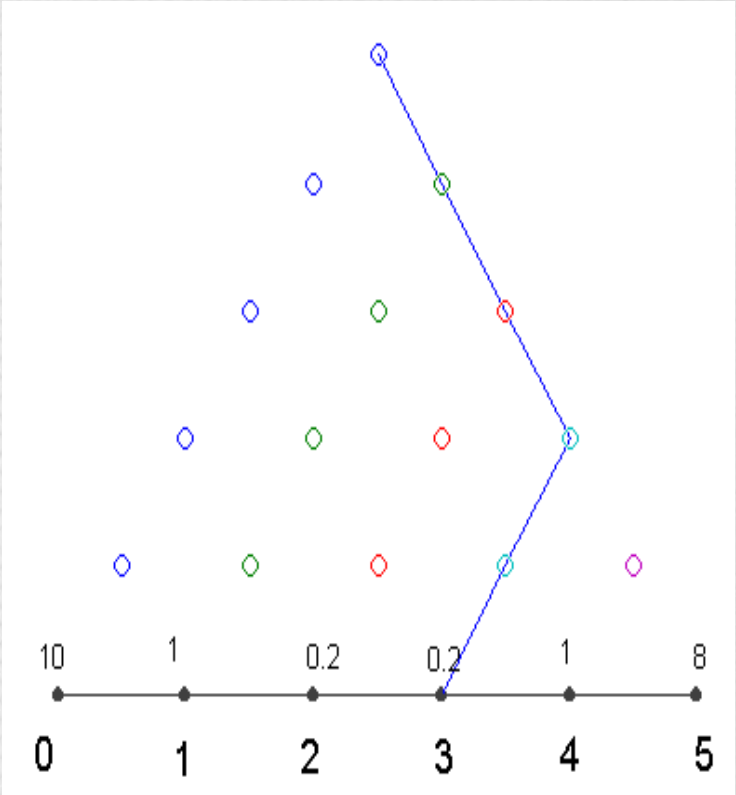
- 小球最后落入底板中的某一个格子，图中用一条折线显示小球下落的一条轨迹.向Galton钉板扔进一个小球，显然不能预测小球会落到哪一个格子。
- 如果不断地重复扔球过程，将会发现什么结果呢？
- 这个问题让你联想到生活中什么例子？



Galton钉板实验：实验与观察

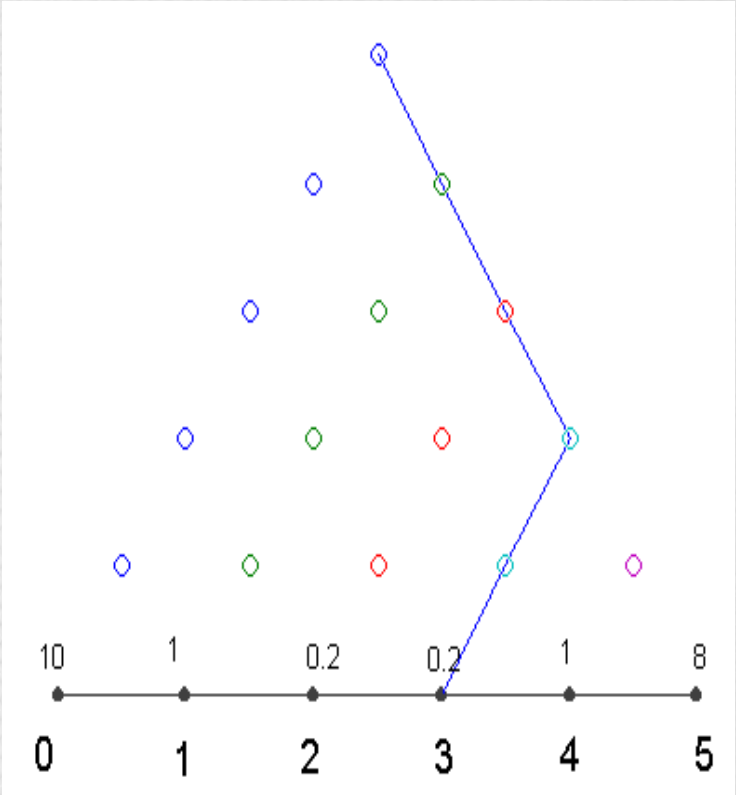
- 博彩问题（仅举例，本课程**不提倡**）
- 在每一格子中放上适当价值的奖品；
- 例如，奖品价值依次为 10　1

  0.2　0.2　　1　　8　（元）
- 扔一次小球你要付1元给庄主；
- 如果小球落入某个格子；
- 你将获得相应价值的奖品；
- 你合算吗？庄主会赚吗？



Galton钉板实验：实验与观察

- 小球落入哪一个格子是不确定的；
- 所以要计算落入每一个格子的可能性；
- 试想向Galton板中扔10000个小球；
- 这些小球将堆积起来；
- 小球的堆积形状告诉了我们什么呢？



- 下面让我们通过数学实验来探索其中的规律吧。

Galton钉板实验：实验与观察

- 实验内容:模拟Galton钉板试验,观察和体会概率分布的意义.
- 模拟Gatlon钉板试验的步骤很简单.

◆ Galton钉板实验：实验与观察

- **Step 1**：确定钉子的位置：将钉子的横、纵坐标存储在两个矩阵X和Y之中.
- **Step 2**：在Galton钉板试验中，小球每碰到钉子下落时都具有两种可能性.设向右的概率为p,向左的概率为q=1-p，这里p=0.5,表示向左向右的机会是相同的.

回顾：首先产生一均匀随机数U,这只需调用随机数发生器指令rand(m,n).

rand(m,n)指令：用来产生m×n个(0,1)区间中的随机数,并将这些随机数存于一个m×n矩阵中，每次调用rand(m,n)的结果都会不同.如果想保持结果一致，可与rand('seed,s)配合使用，这里s是一个正整数.

◆ Galton钉板实验：实验与观察

- **Step 1**：确定钉子的位置：将钉子的横、纵坐标存储在两个矩阵X和Y之中.
- **Step 2**：在Galton钉板试验中，小球每碰到钉子下落时都具有两种可能性.设向右的概率为p,向左的概率为q=1-p，这里p=0.5,表示向左向右的机会是相同的.

回顾：首先产生一均匀随机数U,这只需调用随机数发生器指令rand(m,n).

例如【rand('seed',1),u=rand(1,6)】

u=0.5129 0.4605 0.3504 0.0950 0.4337 0.7092

而且再次运行该指令时结果保持不变.除非重新设置种子seed的值，如

【rand('seed',2),u=rand(1,6)】

u=0.0258 0.9210 0.7008 0.1901 0.8673 0.4185

这样结果才会产生变化.

Galton钉板实验：实验与观察

- **Step 1：** 确定钉子的位置：将钉子的横、纵坐标存储在两个矩阵X和Y之中.
- **Step 2：** 在Galton钉板试验中，小球每碰到钉子下落时都具有两种可能性.设向右的概率为p,向左的概率为q=1-p，这里p=0.5,表示向左向右的机会是相同的.

具体模拟过程如下：

将[0,1]区间分成两段，区间[0,p)和[p,1].如果随机数u∈[0,p)，让小球向右落下；若u ∈[p,1]，让小球向左落下.将这一过程重复n次，并用直线连接小球落下时所经过的点，这样就模拟了小球从顶端随机地落入某一格子的过程.

◆ Galton钉板实验：实验与观察

- **Step 3**：用直线连接小球落下时所经过的点，这样就模拟了小球从顶端随机地落入某一格子的过程。
- **Step 4**：模拟小球堆积的形状.

输入扔球次数m(例如m=50、100、500等等)，计算落在第i个格子的小球数在总球数m中所占的比例，这样当模拟结束时，就得到了频率$f_i = \frac{m_i}{m}, i = 0, \cdots, n$，用频率反映小球的堆积形状。

Galton钉板实验：实验与观察

用如下动画指令结构制作动画：

moviein(n)：创建动画矩阵;制作动画矩阵数据;

getframe:拷贝动画矩阵;

movie(mat,m)：播放动画矩阵m次.

关于这一点可参见本课程实验一的内容和指令说明.

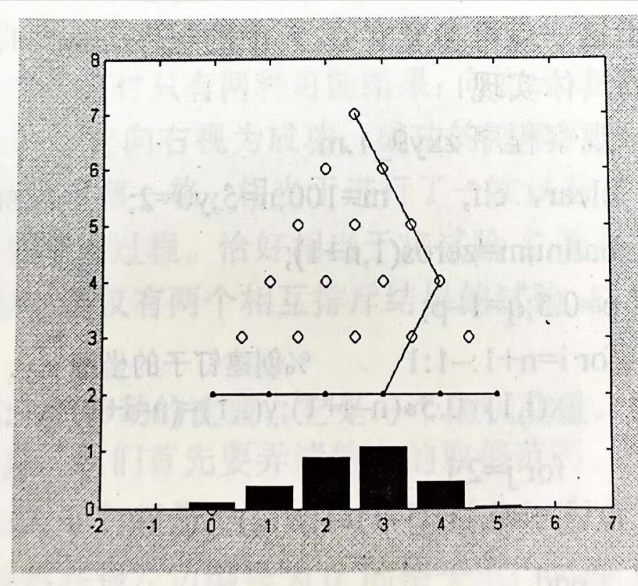
下面的观测程序就是上面步骤的具体实现.

观察程序Exp12_1.m

Galton钉板实验：实验与观察

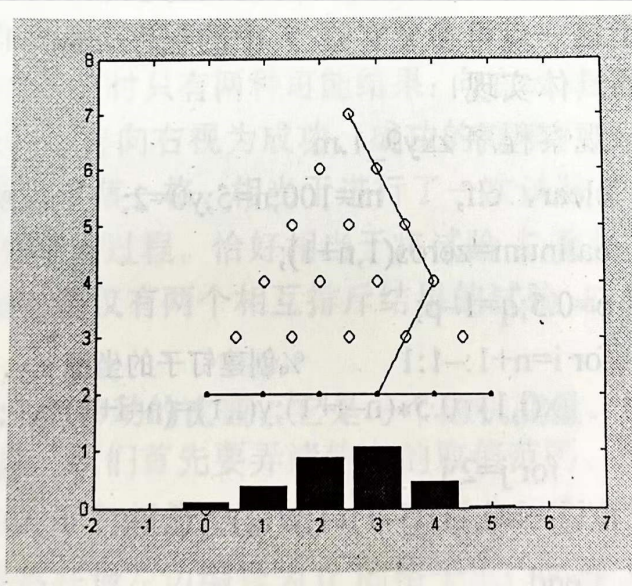
实验：

- 扔100个小球
- 向右概率p=0.5



用二项分布描述 Galton 钉板模型

Galton钉板模型可以看成一个有趣的游戏，有人也拿它作为一种赌博盈利的工具.但在概率统计学家的眼中，它却是一个非常有用的概率模型.



当你扔小球时，你和庄家关心什么？

对，是小球落入格子的编号数 X

（有些绕口，但很重要）

在投球前，你不能说你的小球会落在第0个格子（虽然你很希望）。但你可以说小球将落在第X个格子。X是一个随机数是概率论中重要的讨论对象-----随机变量！！！

用二项分布描述 Galton 钉板模型

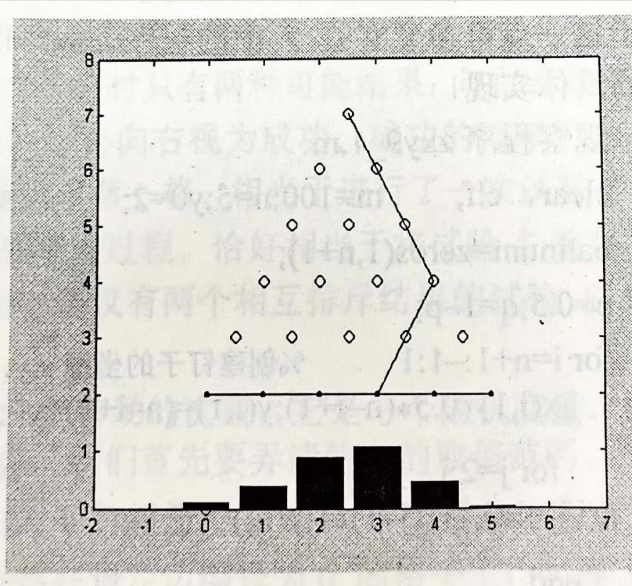
实际上，更应该关心的是 X 的分布列

分布列是小球落在各格子里的概率：

P(X=0)，P(X=1)，P(X=2)，P(X=3)，

P(X=4)，P(X=5)

想一想，它是不是表现了大量投球后小球堆积的极限形状呢？（比较频率和概率）

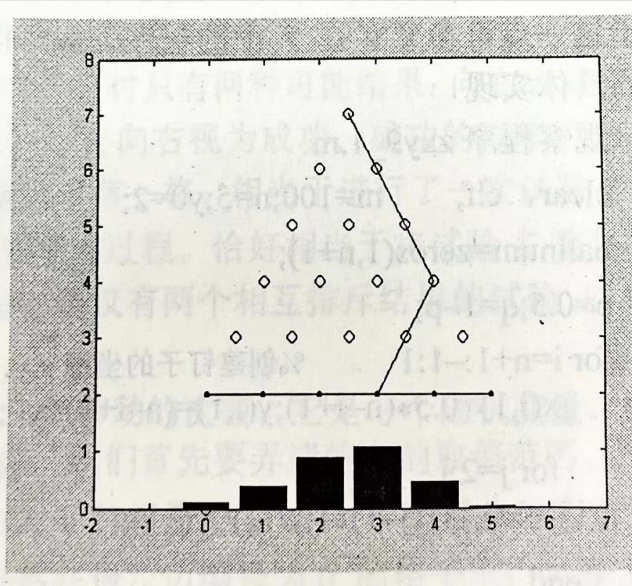


不要把Galton钉板简单地当作消遣

- 它是一个有用的概率模型

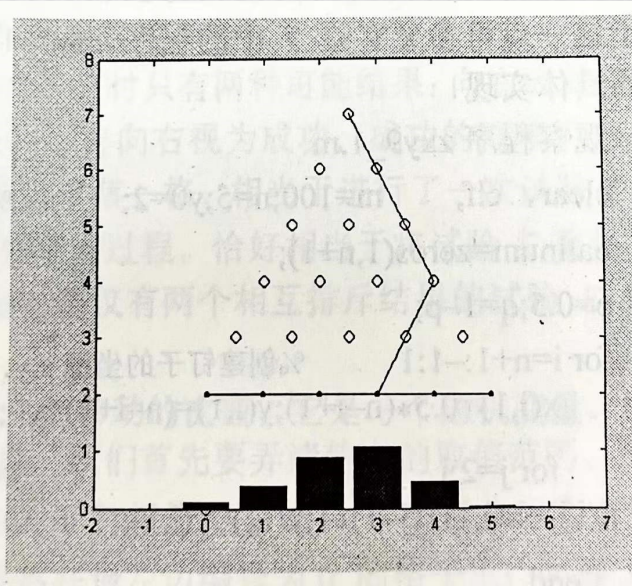
用二项分布描述 Galton 钉板模型

小球自上方落下，经过n个钉子. 每经过一个钉子时只有两种可能结果：向右或向左. 这是一个具有两个结果（成功和失败）的随机试验E，将向右视为成功，成功的概率为p, 向左为失败，失败率为q=1−p. 小球碰到一个钉子下落一格，相当于进行了一次试验E.



小球自顶端落下，碰到n个钉子，最终落在某个格子的过程，恰好相当于将试验E重复了n次，因此一次投球过程就是一个n重贝努利试验（将仅有两个相互排斥结果的试验E独立重复n次，构成了n重Bernoulli试验E^n）.

用二项分布描述 Galton 钉板模型



n重Bernoulli试验的成功次数X正好是小球向右移动的次数，它是一个随机变量. 根据概率论，n重Bernoulli试验的成功次数X 服从二项分布B(n, p). 上面模拟对应于n=5, p=0.5的情形。对于一个随机变量，我们首先要弄清楚它的取值范围，X的取值范围为0,1,2,…, n, 这是什么意思呢？

在Galton钉板模型中X=0表示小球向右移动零次，即小球一直向左移动，所以它恰好要落在最左边编号为0的格子里；同理X=1表示小球恰好要落在编号为1的格子里，依此类推，这就是说，X是小球最终落进的格子编号数，也对应为小球向右移动的次数.

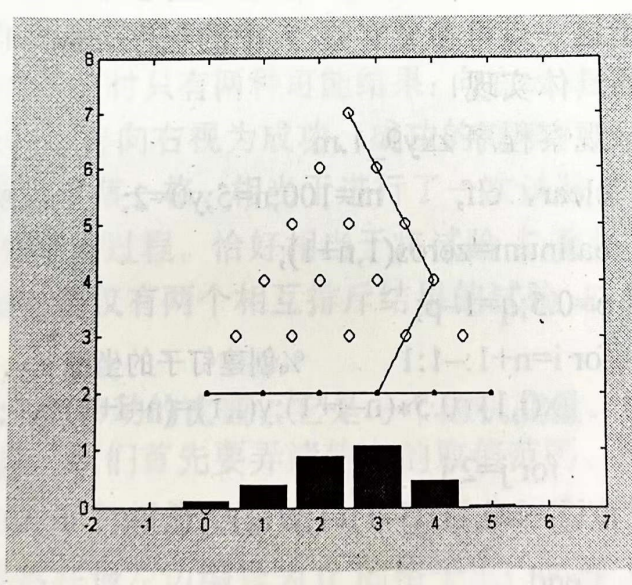
◆ 用二项分布描述 Galton 钉板模型

二项分布随机变量X的分布列为

$$p_i = P(X = i) = C_n^i p^i q^{n-i}, i = 0,1,\cdots,n$$

(这里取n=5, p=0.5, q=1-p).

用**Matlab的统计工具箱**可以方便的计算这一分布列.

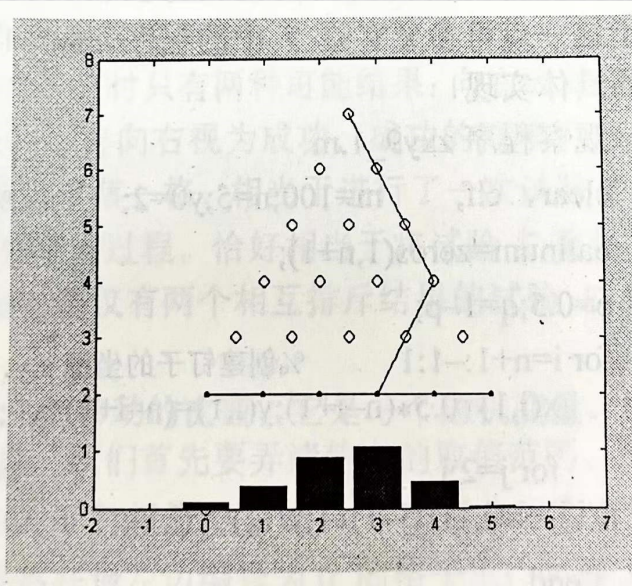


- binopdf指令：binopdf(x,n,p)用来计算二项分布列,参数n和p分别为试验次数和成功概率.只要给定x,就可以计算x处相应的概率、x可以是向量或矩阵.
- binornd指令：用Matlab模拟服从二项分布的随机变量也是方便的,只需调用的二项分布发生器R=binornd(n,p,s,m),参数n、p和binopdf(x,n,p)中是一样的,运行该指令后得到一个s×m矩阵R

用二项分布描述 Galton 钉板模型

模拟向图中的n=5层Galton钉板模型中投一个小球，只需模拟该小球将落入哪一个格子.设小球落入了第X个格子，则X是服从二项分布B(5,0.5)的随机变量.

所以，如果不追求趣味性，对X的模拟就是对Galton钉板试验投球过程的模拟.



观察:用二项分布随机数指令binornd模拟Galton钉板模型的10次投球过程.【X=binornd(5,0.5,1,10)】

{X=1 3 2 2 3 3 0 2 1 2}

思考： 这相当于向Galton钉板投球 ？？次，第一个球落入编号为 ？？的格子里，第二个球落在编号为？？的格子里

用二项分布描述 Galton 钉板模型

实验:用binornd模拟5000次投球过程，观察小球堆积的情况.

```
n=5;p=0.5;
m=5000;
rand('seed',3);
R=binornd(n,p,1,m);%模拟二项分布的随机数，相当于模拟投球m次
```

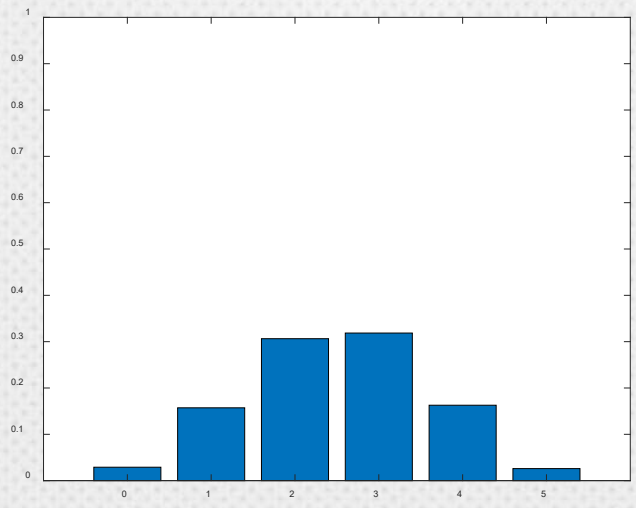
思考接下来如何编程统计小球堆积?

讲解程序Exp12_2.m

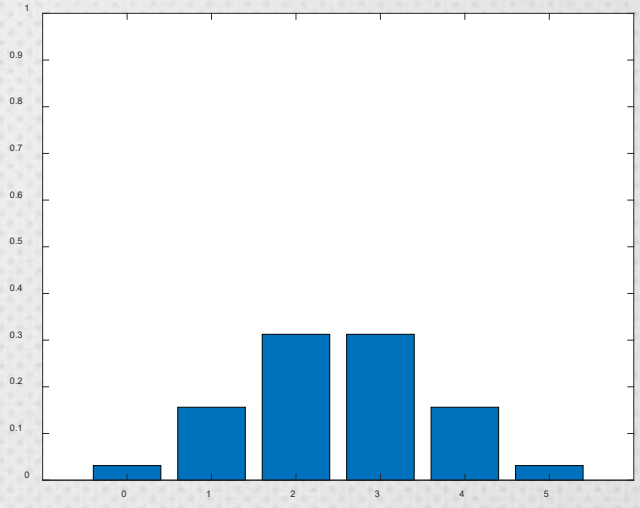
用二项分布描述 Galton 钉板模型

观察:对前面的Galton钉板模型，选择n=5,p=0.5，画出二项分布列的分布图.

【n=5;p=0.5;x=[0:1:n];f=binopdf(x,n,p),
bar(x,f),axis([-1 6 0 1])】



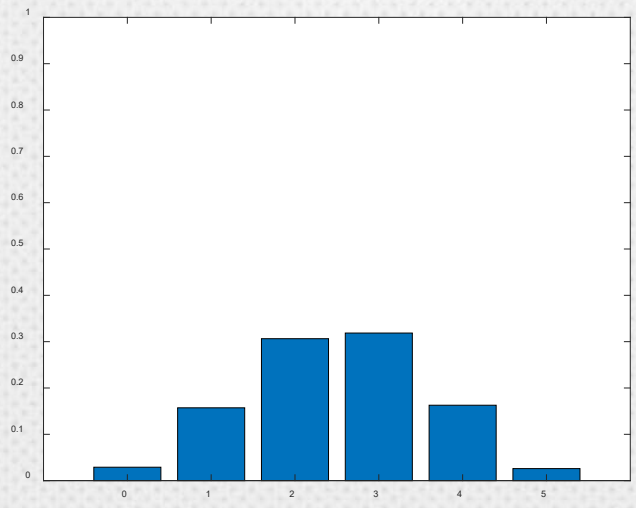
5000次实验得到的频率图



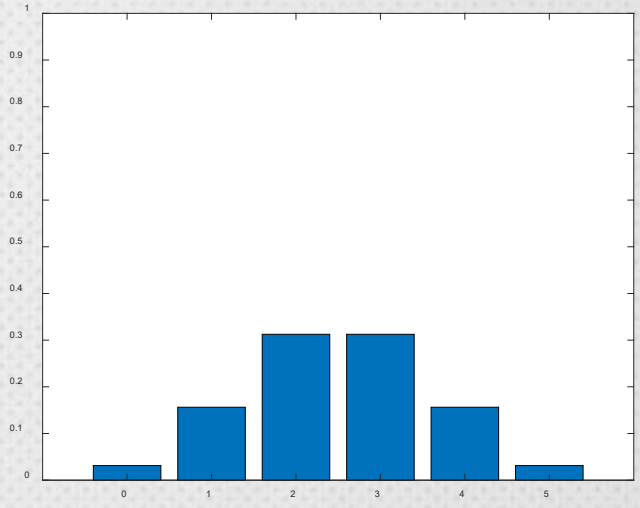
理论分布B(5,0.5)的PDF

用二项分布描述 Galton 钉板模型

对比图中的结果，分布列用来描述大量投球后，小球的堆积的“极限”形状，这个极限反映了小球沿实轴(格子)的分布规律，它只有在大量重复投球后才会呈现出来，一旦获得了这一规律，就获得了所研究的随机模型的所有的统计信息.



5000次实验得到的频率图



理论分布B(5, 0.5)的PDF

数学期望与平均收益

分布列起着什么作用呢？当你准备向Galton钉板投入一个小球时，你能期望你投出的小球落在编号为0或5的格子中吗？如果这是一个抽奖游戏，扔一次小球需要付出1元代价，同时在不同的格子中设置了不同价值的奖品，下表给出了一种奖品设置的情况，抽奖者一般的希望是奖品回报大于所付出的代价，这一点能够实现吗？

奖品的设置

格子编号	0	1	2	3	4	5
奖品的价值/元	5	1	0.2	0.2	1	5

数学期望与平均收益

当然，小球有可能落入5元的格子中，但是见好就收吧.

如果你继续扔m次，将每次获得的奖品价值相加并除以m，就得到了每次抽奖的平均回报.

如果这个平均回报的价值超过抽一次奖所付出的代价1元钱，那么你当然是高兴的了.但是会是这样的吗？

奖品的设置

格子编号	0	1	2	3	4	5
奖品的价值/元	5	1	0.2	0.2	1	5

数学期望与平均收益

观察：模拟5000次抽奖过程，抽奖一次支付1元，按下表获得回报. 计算总收益和一次抽奖所得的平均收益.

运行程序Exp12_3.m 通过随机模拟，估算了1次抽奖的平均收益.

- 学过概率论的同学，可以计算理论的期望值，与实验结果进行比较：

奖品的设置

格子编号	0	1	2	3	4	5
奖品的价值/元	5	1	0.2	0.2	1	5

数学期望与平均收益

可见抽奖5000次，每次抽奖的平均回报为0.7506元<抽奖的代价1元.因此作为抽奖者总体上要亏，而抽奖的主办者总体上是要赚钱的(思考：5000次抽奖主办者将获取利润多少？)

奖品的设置

格子编号	0	1	2	3	4	5
奖品的价值/元	5	1	0.2	0.2	1	5

数学期望与平均收益

- 这是因为由统计概率的定义，频率 $\lim_{m\to\infty}\frac{m_i}{m}=p_i$，于是实际模拟计算的平均值趋向于理论平均值(数学期望)$Ef(X)=\sum_{i=0}^{n}f_ip_i$.
- 数学期望可以理解为由于随机变量X以p_i的概率取到值i(即小球落入第i格的概率为p_i)，
- 这意味着抽奖者以p_i的概率获取价值f_i，所以若以概率p_i对函数值f_i做折扣：即计算折扣值f_ip_i，并把所有折扣值加总，就得到了理论均值或数学期望$Ef(X)$.
- 运行下面的程序，获得理论期望值

```
f=[5,1,0.2,0.2,1,5];x=0:5;pu=binopdf(x,5,p);EF=sum(f.*pu)
```

实验2：电力供应问题

◆思考

电力供应问题. 某车间有200台车床互相独立的工作，由于经常需要检修、测量、调换刀具等种种原因需要停车，这使每台车床的开工率只有60%. 而每台车床在开动时需耗电1kW。

显然向该车间供电200kW可以保证有足够电力供这些车床使用，但是在电力比较紧张的情况下，给这个车间供给电力太多将造成浪费，太少又影响生产.如何解决这一矛盾？

◆思考

一种解决方案是保证有基本足够的电力供应该车间，比如要求在8小时的生产过程中允许有半分钟的电力不足，半分钟约占8小时的0.1%，用概率论的语言就是：应供给多少电力才能以99.9%的概率保证不会因为电力不足而影响生产？

◆思考

在任意时刻,把某台车床在工作看作成功,则成功概率为p=0.6;否则是失败,失败的概率为0.4.

这样逐台考察200台车床的工作情况,就相当于进行了一个试验次数为n=200的贝努利试验(也就相当于Galton钉板模型的小球从顶端落在钉板下方的某一格子之中,只不过这时小球向右的概率为0.6),而200重贝努利试验的成功次数X恰好是在这一时刻工作着的车床数,因此X～B(200,0.6).

◆一种思路

假设需供电m(kW)确保m个车床可以供电,则以99.9%的概率保证不会因为电力不足而影响生产的含义是确定m,使得

$P(X \leq m) = \sum_{k=0}^{m} C_{200}{}^{k}(0.6)^{k}(0.4)^{200-k} \geq 0.999$,

上式左边是X的分布函数$F(x) = k = P(X \leq x)$在整数点m处的取值.

◆思考与练习

- binocdf指令. 可以用指令binocdf(x,n,p)直接计算这一概率,binocdf用来计算二项分布的分布函数(cdf),其参数的意义和前面已经用过的计算二项分布列(pdf)的指令binopdf是相同的,不再重述.

◆思考与练习

练习.

(1)计算分布函数在某些点的取值

F(m), m=0, 1, 2, …, 200, 并将它绘于图上, 辅助某些必要的计算, 求出问题中所需要的供电功率数。

(2)将8小时按半分钟分成若干时间段, 共有

8×60×2=960个时段, 用二项分布模拟8小时车床运行的情况, 观察已算得的供电功率数是否能够基本保证车间正常工作. (思考, 如何实验?)

◆思考与练习

(3) 用德莫佛-拉普拉斯中心极限定理，当n充分大时有 $\frac{X-np}{\sqrt{npq}} \sim N(0,1)$，这意味着 $X \sim N(np, npq)$，分两种情形观察这种近似：（实验课实验题）计算二项分布的理论分布列并将其与正态分布密度函数进行比较；模拟车床运行的情况，计算出频率后与正态密度函数进行比较.

用正态分布计算供电功率数，与已算得的功率数比较差异. 如果车床较少（例如只有10台车床），两者的差别大吗？试比较之. （5分）

实验3：如何制定胖和瘦的标准?

实验3

描述某一类人群的身高，可以使用正态随机变量.

- 这是因为大多数人的身高在人群的平均身高附近波动，呈现着一种对称分布：特别高或特别矮的人是很少的，与正态分布体现的中间大两头小的特点比较吻合.
- 同理，也可用正态随机变量来描述人群的体重.

实验3

随着生活水平的日益提高，人们越来越关注自己的健康. 如果摄取营养过多而运动不足，将导致人的发胖，这既影响健美、也对健康不利. 针对人的希望健美和健康的心理，各种减肥药品和健身器械应运而生，其广告铺天盖地.

但是，评定一个人过胖或过瘦并不是一件很容易的事情.

实验3

制定一个正常体重标准，是人们所关心的.

这一标准是和具体的人群对象有关.

例如男性、女性应该有不同的标准，不同的人种标准也可能不一样等等.

实验3

从数学建模的观点来看，我们需要根据一些必要的统计数据来建立模型，这一模型是随机的，必须根据实测数据，用概率统计的思想和方法来处理.

下面所用的方法只是提供了一种解决问题的思路，这种思路在实践中经常被采用，真正要确定正常体重标准可能要复杂得多.

◆ 模拟数据

(1)观察：模拟身高和体重的数据. 当然你可以自己收集数据(比如收集同学的身高和体重数据)来进行分析. 这里我们没有实际数据，因此采用模拟方法. 在模拟中假设了某一人群的身高是正态随机变量$X \sim N(\mu, \sigma^2)$，比如$\mu = 170cm, \sigma = 4.5cm$. 又假定体重Y和身高有如下关系

$$Y = rX + s + u, \qquad (1)$$

这里r、s是适当的常数，$u \sim N(0, \sigma^2{}_u)$为零均值的正态随机变量，反映了一种随机扰动. 设定具体的参数值，模拟一组身高和体重的数据.

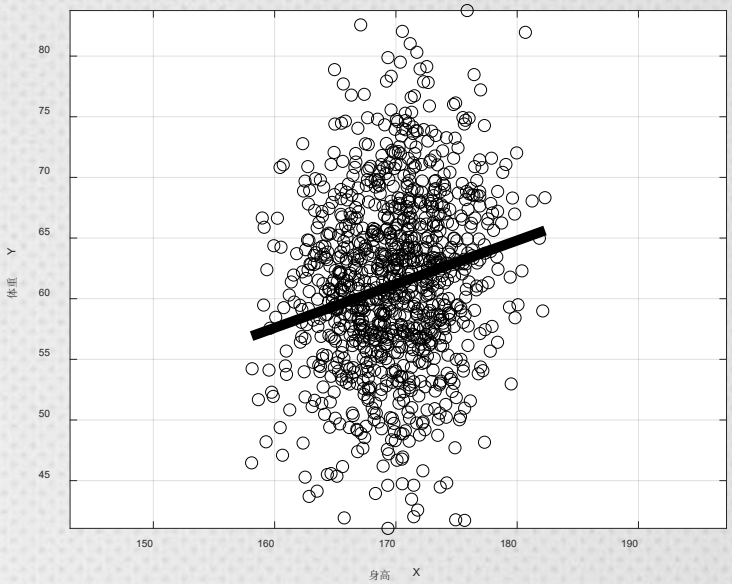
◆ 实验与思考

步骤1：产生n个正态随机数X，代表n个人的身高；再产生n个零均值的正态随机扰动U，给定适当的r值和s值，由(9.1)式计算n个随机数Y，它们代表着这n个人的体重(实际中可以通过对n个人测得这两组数据，用这些数据来建立模型，当然这时数据不一定有关系式(1)，所以更复杂).现在我们就有了n个人身高和体重的数据.

◆ 实验与思考

思考： 用指令hist画出身高数据的直方图，也可以用normpdf和plot指令绘制描述身高的随机变量X的概率密度函数$f_X(x)$；同理也可画出体重数据的直方图和描述体重的随机变量Y的概率密度函数$f_Y(y)$. 想一想，概率密度函数$f_X(x)$

$f_Y(y)$提供了身高和体重之间的

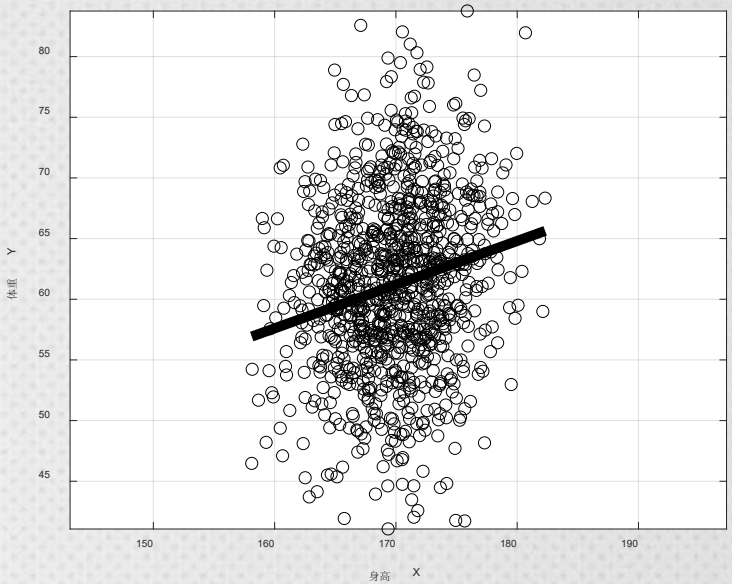


模拟身高体重数据的散点图

实验与思考

思考：

- 为什么要讨论多元随机变量？
- 图中的直线有什么特别的意义吗？



模拟身高体重数据的散点图

◆ 实验与思考

思考：如何绘制二维直方图和二元正态分布密度函数图象？

可以作二维直方图进一步在量方面反映图中散点的分布情况，这需要编制相应的程序：

将x-y平面分成若干网格，计算落入每个网格的散点个数(频数)，再将每个频数除以样本点的总数n，便得到频率. 将频率绘在三维图形中(用mesh或surf指令，当然也可以自己编程绘出直方的效果)，同时可绘制相应的等值线图(用contour指令)，这样便可更清楚观测散点的分布情况.

◆ 实验与思考

思考：如何绘制二维直方图和二元正态分布密度函数图象？

由正态随机变量在线性变换下的不变性，(X, Y)仍然服从二维正态分布，且

$(X,Y)\sim N(\mu, r\mu+s, \sigma^2, r^2\sigma^2+\sigma^2{}_u, \rho_{XY})$.

这是因为$EX=\mu, EY=E(rX+s+u)=r\mu+s+E(u)=r\mu+s, DX=\sigma^2,$

$D(Y)=D(rX+s+u)=D(rX+s)+D(u)$

$=D(rX)+D(u)=r^2D(X)+\sigma^2{}_u=r^2\sigma^2+\sigma^2{}_u,$

$cov(X,Y)=cov(X,rX+s+u)=rcov(X,X)+cov(X,u)=rD(X)=r\sigma^2,$

$\rho_{XY}=\frac{cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}=\frac{r\sigma^2}{\sigma\sqrt{r^2\sigma^2+\sigma^2{}_u}}=\frac{r\sigma}{\sqrt{r^2\sigma^2+\sigma^2{}_u}}=\frac{r}{|r|}\frac{1}{\sqrt{1+\sigma^2{}_u/r^2\sigma^2}}.$

由此可画出二元正态密度函数的图形（三维图和等值线图），并同二维直方图进行对比.

◆ 实验与思考

运行参考程序Exp12_7.m改变参数，观察二维直方图和理论分布的图形和身高和体重的概率关系.

二维直方图和二维正态密度函数图的对比（左上图：二维直方图；右上图：二维直方图的等值线图；左下图：密度函数图；右下图：密度函数的等值线图）

◆ 实验与思考

在实际应用中，条件正态分布是很重要的，以二维的情况说明之.

由概率论的结论，设$(X,Y)\sim N(\mu_1,\mu_2,{\sigma_1}^2,{\sigma_2}^2,\rho)$，则给定X=x下Y的条件分布密度函数为

$$f_{Y|X}(y|x)=\frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2){\sigma_2}^2}[y-(\mu_2+\frac{\sigma_2}{\sigma_1}\rho(x-\mu_1))]\right\}^2.$$

即$f_{Y|X}(y|x)$为正态分布$N(\mu_2+\frac{\sigma_2}{\sigma_1}\rho(x-\mu_1),(1-\rho^2){\sigma_2}^2)$的概率密度。称$E(Y|X=x)=\int_{-\infty}^{\infty}yf_{Y|X}(Y|X)dy$为给定X=x下Y的**条件均值**. 称$\mathrm{Var}(Y|X=x)=\int_{-\infty}^{\infty}\big(y-E(Y|X=x)\big)^2f_{Y|X}(Y|X)dy$为给定X=x下Y的条件方差，这两个条件数字特征是非常有用的.

◆ 实验与思考

在正态的情形，显然有

$$E(Y|X=x)=\mu_2+\frac{\sigma_2}{\sigma_1}\rho(x-\mu_1),$$

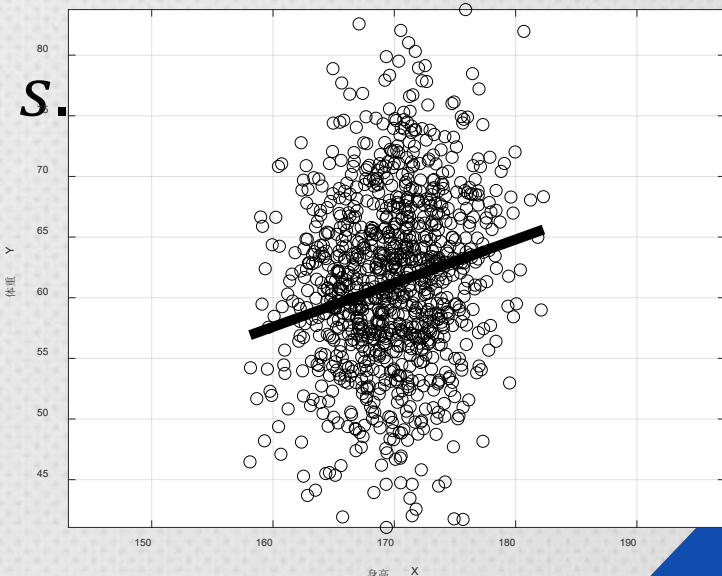
$$\mathrm{Var}(Y|X=x)=\left(1-\rho^2\right){\sigma_2}^2.$$

利用上面的结论，对于所讨论的身高体重问题，

将相应的参数代入上式中计算得

$$E(Y|X=x)=rx+s.$$

思考图中直线的意义？



◆ 实验与思考

- 然而，在实际中我们并不知道所需的参数，即体重和身高的均值、方差、相关系数等等都是未知的(甚至我们并不知道代表身高和体重随机变量X、Y是否服从二维正态分布)，只有一些观测到的数据，这时候如何制定正常体重的标准呢？
- 身高和体重之间的关系，与确定性函数不同；体重不能表成身高的函数，即我们不能根据某人的身高确定出他的体重. 身高和体重之间的关系可以用两者的联合分布密度来刻画.

◆ 实验与思考

- 如果假定身高体重服从二元正态分布，就可以用条件数学期望$E(Y|X=x)$由身高对体重进行合理预测（这是一种条件平均），并且用条件概率密度$f_{Y|X}(y|x)$给出相应的区间估计。

◆ 实验与思考

另一种方法是，假定两者之间满足某种模型，例如，假定体重Y和身高X有如下关系

$$Y = rX + s + u,$$

其中r、s是适当的常数，$u \sim N(0, \sigma_u{}^2)$为零均值正态随机变量，反应了一种随机扰动.

上面的模型称为线性回归模型。

下面我们讨论更一般的多元线性回归。（思考：与拟合的关系？）

◆ 实验与思考

- 用多元线性回归指令regress做体重预测.

多元线性回归是最常用的统计方法之一，一般模型为

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m + \varepsilon, \varepsilon \sim N(0, \sigma^2),$$

其中σ未知. 若得到n组独立观测数据

$\{y_i, x_{i1}, x_{i2}, \ldots, x_{im}\}, i = 1,2, \ldots, n, n > m$，将它们代入回归模型中得到

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2),$$

$$i = 1,2, \ldots, n.$$

◆ 实验与思考

可写成下面的矩阵形式

$$Y = X\beta + \varepsilon,$$

其中

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{m1} \\ 1 & x_{12} & \cdots & x_{m2} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{1n} & \cdots & x_{mn} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

回归分析任务是根据数据估计未知参数β和σ^2，并作统计分析给出模型的可信程度的定量评价.

根据最小二乘法可得到未知参数的估计.

◆ 实验与思考

即通过解

$$\min_{\beta_0,\beta_1,\cdots,\beta_m}\sum_{i=1}^{n}\varepsilon_i=\min_{\beta_0,\beta_1,\cdots,\beta_m}\sum_{i=1}^{n}(y_i-(\hat{\beta}_0+\hat{\beta}_1x_{i1}+\cdots+\hat{\beta}_mx_{im}))^2$$

确定$\beta_0,\beta_1,\cdots,\beta_m$的估计值$\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_m$，这实际上是解一个超定方程组.Matlab指令为

[b,bint,r,rint,stats]=regress(Y,X,alpha)

◆ 实验与思考

Matlab指令为

[b,bint,r,rint,stats]=regress(Y,X,alpha)

- 其中Y、X是的数据矩阵；alpha是显著性水平，缺省时为0.05；b是参数的估计值；bint是的置信区间；r是残差，令Y的回归估计值为$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_m X_m$，残差定义为$r_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_m x_{im}), i = 1,2,\cdots,n$;又令$Q = \sum_{i=1}^{n} {r_i}^2$为**残差平方和**，可用**剩余标准差**$s^2 = Q/(n-m-1)$作未知参数$\sigma^2$的无偏估计；rint是残差r置信度为1-alpha的置信区间，使用rcoplot(r,rint)指令可显示残差图.

◆ 实验与思考

Matlab指令为[b,bint,r,rint,stats]=regress(Y,X,alpha)

- stats中含有几个检验回归模型的统计量. 可将平方和$S=\sum_{i=1}^{n}(y_i-\bar{y})^2$分解为残差平方和Q与回归平方和U两个部分S=Q+U, 回归平方和定义为$U=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$. stats(1)为$R^2=U/Q$的值, R也称为相关系数, R越大表示Y与$X_1,\cdots,X_m$关系越密切, 通常R大于0.8或0.9才认为相关关系成立; stats(2)为$F=\frac{U/m}{Q/(n-m-1)}$的值; 可以证明$F\sim F(m,n-m-1)$, 这是用来检验Y是否与$X_1,\cdots,X_m$存在线性关系的统计量(原假设为$H_0:\beta_j=0,j=0,1,\cdots,m$); stats(3)是与F对应的概率值p, 若p<alpha拒绝H_0, 可认为回归模型成立.

◆ 实验与思考

Matlab指令为

[b,bint,r,rint,stats]=regress(Y,X,alpha)

- 对模拟的100对身高体重数据，运行程序Exp12_8.m

了解指令regress的用法,理解相关结果.

- 数据产生的真实模型是$Y = 0.36X + u, u \sim N(0, \sigma^2{}_u)$，而上面的估计值与此相差较大，这是什么原因呢？试减小扰动均方差σ_u，再模拟数据并进行回归分析.你观察到什么结果？还可以自己设置模型并模拟数据，观察回归的效果。

实验4：极大似然估计

◆从一个应用例子谈起

1. 如何决定废品率？

产品的质量总是生产者、管理者高度注重的问题. 废品存在是在所难免的，质检员的任务之一是监控产品的废品率. 设某工厂生产了一大批产品，为测算产品废品率，质检员随机抽取了n=50件产品进行质量检验.

- 根据以往的经验，他将产品质量分为6个不同档次，对应的废品率分别为0. 01、0. 02、0. 03、0. 04、0. 05、0. 06；现在质检员要根据对50件产品检查的结果，决定该批产品档次，请为他提供一种合理的方案.

◆从一个应用例子谈起

观察：

(1)模拟抽样数据. 设想有一批产品，你可以设定它们的档次，例如设定废品率为p=0.04. 请模拟质检员随机抽取n=50个样品进行检验.

用统计术语，对任一件产品，它要么是正品，要么是废品. 以随机变量X表示这一事实，则$X=\begin{cases}1,\text{产品为废品},\\0,\text{产品为正品}.\end{cases}$X服从两点分布，即$P(X=1)=p, P(X=0)=q=1-p$. 称X为总体，它的分布称为总体分布. 总体分布决定了产品的档次，p的取值范围是0.01、0.02、0.03、0.04、0.05、0.06，取不同值决定了产品不同的档次.

◆从一个应用例子谈起

观察：

(1)模拟抽样数据. 设想有一批产品，你可以设定它们的档次，例如设定废品率为p=0. 04. 请模拟质检员随机抽取n=50个样品进行检验.

质检员对产品n次抽样相当于对总体X复制了n次，得到了n个独立同分布的随机变量$X_1, X_2, \ldots, X_n$，

$X_1, X_2, \ldots, X_n$称为容量为n的简单样本.

这样模拟质检员的抽样行为就相当于对总体X作n次模拟，模拟的结果相当于得到了简单样本的一组样本观察值$x_1, x_2, \ldots, x_n$.

◆从一个应用例子谈起

观察：

(1)模拟抽样数据. 设想有一批产品，你可以设定它们的档次，例如设定废品率为p=0. 04. 请模拟质检员随机抽取n=50个样品进行检验.

程序Exp12_9. m模拟质检员抽取10个产品的检查过程.

运行一次程序得到

i　1 2 3 4 5 6 7 8 9 10

xi 0 0 0 0 1 1 0 0 0　0

故质检员抽到的10件产品中，有两件是废品，8件是正品.

设想质检员再抽取10个产品进行检查，得到的数据会是相同的吗？

◆从一个应用例子谈起

- **思考：**

(1) 以上三次模拟，每次都得到不同结果，还可以继续做多次模拟看看结果，它们是相同的吗？

- 是否能够理解“简单样本X_1, X_2, …, X_10是独立同分布的随机变量”的含义？
- 对质检员来说，该批产品的档次也就是实际废品率p=0.04是已知的吗？

◆从一个应用例子谈起

- **思考：**

(1)数据意味着信息

质检员通过对产品的检验获得容量为n的简单样本观察值$x_1, x_2, \ldots, x_n$，这些x_i取值或为1(表示检验的产品为废品)或为0(表示为正品).

他必须根据这些数据估计产品的档次(你可以想一想，根据上面的样本观察值，如何估计产品的废品率呢？)

◆从一个应用例子谈起

- 定义似然函数L(p)的定义为

$$L(p) = P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$$

$$= P(X_1 = x_1)P(X_2 = x_2) \cdots P(X_n = x_n)$$

$$= p^{x_1}q^{1-x_1}p^{x_2}q^{1-x_2} \cdots p^{x_n}q^{1-x_n} = p^{\sum_{i=n}^{1} x_i} q^{n-\sum_{i=n}^{1} x_i}.$$

注：因x_i仅取0或1值，故$P(X_i = x_i) = p^{x_i}q^{1-x_i}$是两点分布的另一种表示.

◆从一个应用例子谈起

“似然”是“像”的意思. 当质检员获得一批样本，需根据样本提供的信息推断p的哪一个可能值更“像”真实总体，是p=0.02？还是p=0.04?等等.

很自然，应该有一个度量指标来衡量未知参数和总体的相似性，似然函数正是这样的相似指标.

注意到L(p)是简单样本$X_1, X_2, \ldots, X_n$取值于样本某个特定观察值$(x_1, x_2, \ldots, x_n)$的(联合)概率，而$(x_1, x_2, \ldots, x_n)$反映了真实总体X的某些特征.

◆从一个应用例子谈起

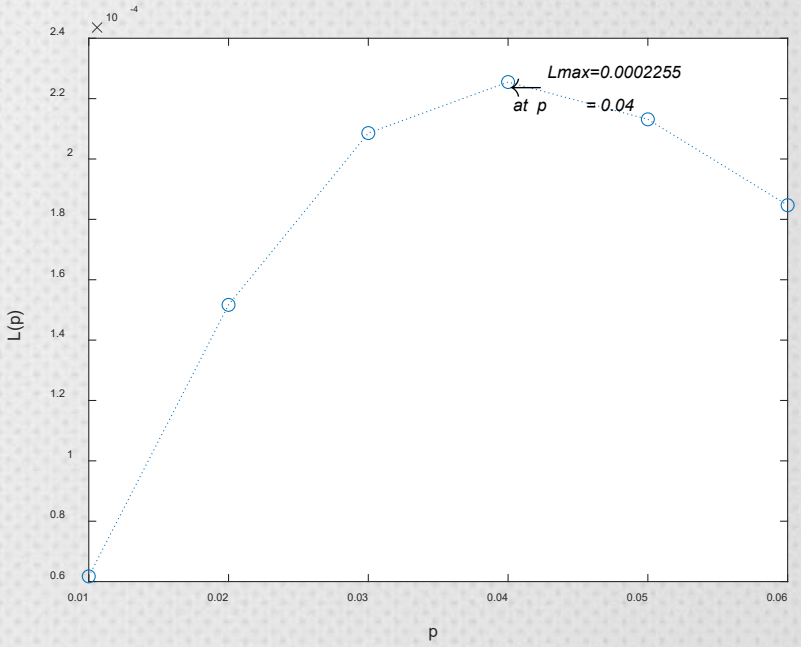
“似然”是“像”的意思. 当质检员获得一批样本，需根据样本提供的信息推断p的哪一个可能值更“像”真实总体，是p=0. 02？还是p=0. 04?等等. 很自然，应该有一个度量指标来衡量未知参数和总体的相似性，似然函数正是这样的相似指标.

对于p的两个可能的取值p_1和p_2，若$L(p_1) < L(p_2)$，则p_2比p_1更像总体. 因此，如果某个p_0达到了似然函数$L(p)$的最大值，则和其它p值相比，p_0最像真实总体，我们把这个p_0作为真实废品率的估计，这就是极大似然估计.

◆从一个应用例子谈起

注意产品分为6个档次，即未知参数p有6个可能的取值：0.01、0.02、0.03、0.04、0.05、0.06.假设该批产品的实际的废品率为$p_0 = 0.04$，而质检员得到了n=50个样本观察值$(x_1, x_2, \ldots, x_{50})$

计算6个档次的似然函数值，作出曲线图，比较它们的大小运行观察程序Exp12_10.m，该程序对p=0.01、0.02、0.03、0.04、0.05、0.06分别计算似然函数L(p)的值，并将其绘制成曲线，观察最大值所对应的情况.



◆从一个应用例子谈起

6个似然函数值的最大者为$L(0.04) = \max_{p} L(p) = 0.0002255$, 即似然函数$L(p)$的最大值恰好在该产品的真实废品率p=0. 04达到，这是偶然的吗？

- **思考：**如果质检员事先并不知道产品的可分为6个档次，他又应该如何估计产品的废品率呢？