实验五　概率论与应用数理统计实验

自然界中的许多现象，在一定条件下有多种可能的结果发生，事前人们无法预言将出现哪种结果，即事件发生的结果同时具有不唯一性和不确定性，这类现象称为随机现象．尽管对每一次试验来讲，其结果的出现是无规律的，但经过大量重复试验后，总体上却能呈现出一定的规律性．概率论与应用数理统计就是研究和揭示这种随机现象统计规律性的一门数学科学．通过本实验，我们可以了解随机现象及其发生的概率，一、多维随机变量的分布规律，参数估计与假设检验，方差分析与回归分析等概率统计的基本方法．

第一节　随机事件及其概率

一、古典概型及其模拟

由古典概型的定义知，古典概型基于这样两个原则：一是所有可能发生的结果只有有限个；二是每一种可能出现的结果机会是相同的．例如掷硬币，我们可以用计算机模拟掷硬币这一过程．在 MATLAB 中提供了一个在[0,1]区间上均匀分布的随机函数，其用法为

rand(N)	返回一个 $N\times N$ 的随机矩阵
rand(N,M)	返回一个 $N\times M$ 的随机矩阵
rand(P_1,P_2,$\cdots$,P_n)	返回一个 $P_1\times P_2\times\cdots\times P_n$ 随机数组

为了模拟掷硬币出现正面或者反面，规定随机数小于 0.5 时为反面，否则为正面，可用 round(　)函数将其变成 0－1 阵，然后将整个矩阵的各元素值加起来再除以总的元素个数即为出现正面的概率．以连续掷 10 000 次硬币为例，重复做 100 次试验模拟出现正面的概率．程序如下

```
for i = 1:100
  a(i) = sum(sum(round(rand(100))))/10000;
end
mx = max(a);
mn = min(a);
ma = mean(a);
```

```
a
mx,mn,ma
```

该程序输出的四项中，a 为试验 100 次中每次出现正面的频率(重复 10 000 次)，mx 和 mn 分别为 100 次试验中出现正面频率的最大值和最小值，ma 为 100 次试验出现正面的平均频率.连续执行该程序 5 次，结果如下(为节省篇幅，后 4 次的 a 略去)

```
a = 0.5056   0.4964   0.4960   0.5019   0.5013   0.4984   0.4965   0.5041
    0.5017   0.4993   0.5005   0.4961   0.4992   0.4927   0.5076   0.5090
    0.4919   0.5033   0.5046   0.4919   0.4970   0.5096   0.5121   0.4990
    0.5002   0.4972   0.4899   0.4975   0.5014   0.4994   0.4958   0.4885
    0.4988   0.5103   0.4970   0.4990   0.5004   0.4957   0.4964   0.5016
    0.5074   0.4978   0.5053   0.4957   0.4983   0.4974   0.4980   0.5021
    0.4883   0.4942   0.4990   0.4995   0.4921   0.4957   0.4999   0.5003
    0.4960   0.4958   0.4959   0.5028   0.4949   0.5061   0.5053   0.4978
    0.4989   0.5043   0.4997   0.5053   0.5019   0.5044   0.5063   0.4991
    0.5001   0.4966   0.5047   0.4992   0.4987   0.4995   0.4994   0.4998
    0.5021   0.5086   0.4977   0.4914   0.4984   0.4944   0.5001   0.4935
    0.4955   0.5109   0.5116   0.4986   0.4963   0.4929   0.5008   0.4974
    0.5022   0.4913   0.4935   0.5046
  mx = 0.5121      mn = 0.4883      ma = 0.4995
  mx = 0.5130      mn = 0.4850      ma = 0.5004
  mx = 0.5121      mn = 0.4894      ma = 0.5004
  mx = 0.5114      mn = 0.4867      ma = 0.5007
  mx = 0.5130      mn = 0.4821      ma = 0.4996
```

有兴趣的同学可以改变程序中的试验次数或重复次数，进一步观察不同次的试验中，出现正面的频率变化情况(包括频率变化区间和平均频率)有何不同.想一想由上述结果能得出什么结论?

再如掷骰子，用计算机模拟这一过程，并观察出现 1 点、2 点、…、6 点的情况有何规律? 程序及执行结果如下:

```
for i = 1:100
  A = ceil(6 * rand(100));
  for j = 1:6
    a(i,j) = sum(sum(~((A - j)&ones(100))))/10000;
  end
end
```

```
mx = max(a);
mn = min(a);
ma = mean(a);
a
mx,mn,ma
```

执行结果为

```
mx =   0.1745   0.1764   0.1755   0.1752   0.1766   0.1745
mn =   0.1577   0.1572   0.1578   0.1577   0.1569   0.1597
ma =   0.1662   0.1669   0.1664   0.1664   0.1669   0.1671
mx =   0.1758   0.1774   0.1759   0.1776   0.1752   0.1740
mn =   0.1557   0.1565   0.1561   0.1560   0.1554   0.1559
ma =   0.1669   0.1669   0.1666   0.1671   0.1658   0.1667
mx =   0.1744   0.1754   0.1779   0.1755   0.1763   0.1741
mn =   0.1534   0.1581   0.1588   0.1559   0.1549   0.1561
ma =   0.1671   0.1667   0.1667   0.1672   0.1662   0.1661
mx =   0.1769   0.1738   0.1769   0.1753   0.1747   0.1770
mn =   0.1572   0.1568   0.1569   0.1598   0.1567   0.1591
ma =   0.1662   0.1664   0.1670   0.1670   0.1667   0.1667
mx =   0.1737   0.1757   0.1769   0.1778   0.1758   0.1748
mn =   0.1584   0.1558   0.1558   0.1574   0.1586   0.1552
ma =   0.1666   0.1665   0.1665   0.1667   0.1672   0.1665
```

上述程序仍然试验 100 次，每次重复 10 000 次，并执行 5 次程序，为节省篇幅，将执行结果中的 a 略去. 其中 100 次模拟试验频率 a，最大频率 mx，最小频率 mn，平均频率 ma 均为 6 列，分别对应于掷 1 点、2 点、…、6 点的频率. 出现每个点的频率与 $1/6\approx0.166\ 7$ 很接近，说明掷骰子出现各种点的概率是相同的.

二、统计概率及其模拟

由于古典概型是建立在事件发生的等可能基础上的概率，而现实生活中许多现象的出现并不是等可能的. 例如某品种的小麦，当种下一粒后，其发芽与不发芽的机会并不相同. 那么，这类概型就不能建立在等可能基础上，即不能使用古典概型的定义. 而统计概型的定义是建立在频率基础之上的，就是说某事件出现的频率如果稳定在某数值 a 附近，则称数值 a 为该事件出现的概率. 由于统计概型中的概率 a 是一个理论上的数值，实际问题中根本无法直接得到该数值，因而通常在试验次数充分多时，利用频率值近似代替概率值. 由前面古典概

型中的掷硬币模拟和掷骰子模拟可以看出，在试验次数充分多的情况下，掷硬币出现正面和反面的频率均在0.5左右，故出现正面和反面的概率均为0.5；而在掷骰子的模拟中，出现各种点的频率均在1/6左右，因而其概率均为1/6.我们可用程序模拟频率随样本容量的增加而变化的趋势与概率值之间的关系.仍以掷硬币为例，观察当样本容量分别为n=10,100,1 000,10 000,100 000,1 000 000时频率的变化.输入下述程序，并观察比较反复五次执行结果的异同.

```
for i = 1:6
  a(i) = sum(round(rand(1,10^i)))/10^i;
end
a
```

```
n =       10    100    1000    10000  100000  1000000
a =   0.5000  0.5400  0.4990  0.5131  0.5012  0.5000
a =   0.5000  0.5700  0.4940  0.4985  0.5006  0.4996
a =   0.9000  0.5100  0.5110  0.5077  0.4972  0.4996
a =   0.7000  0.4400  0.4980  0.5021  0.4996  0.5008
a =   0.6000  0.4900  0.4870  0.5018  0.4987  0.5006
```

从上述执行结果可以看出，当样本容量不够大时，其频率的波动范围很大，即频率不够稳定，即使有时达到0.5，但最大已达0.9.然而随着样本容量的增加，频率的波动范围越来越小，相差仅有10^{-3}左右.类似地，可以模拟掷骰子试验，请读者自己模拟.

进一步模拟：任给一数$0<\lambda<1$，从[0,1]区间随机取一数a，求$P\{a<\lambda\}=?$可以这样模拟：在[0,1]区间任取一数λ（如取$\lambda=0.7$或0.6），再产生1 000 000个随机数，分别取前10,100,1 000,10 000,100 000,1 000 000个统计其小于0.7或0.6的频率，然后将程序分别执行5次，观察执行结果中频率随样本容量大小变化的范围，程序如下

```
lamda = 0.7;
a = round(rand(1,1000000) - lamda + 0.5);
for i = 1:6
  b = a(1:10^i);
  c(i) = 1 - sum(b)/(10^i);
end
c
```

执行结果为

```
n =     10     100      1000   10000  100000  1000000
c = 0.8000  0.7200  0.6960  0.7012  0.6986  0.7000
```

```
c = 0.6000   0.6900   0.7120   0.7081   0.6975   0.6997
c = 0.8000   0.7800   0.7620   0.7019   0.7007   0.7004
c = 0.5000   0.7100   0.6890   0.7020   0.7026   0.7009
c = 0.7000   0.7200   0.6960   0.7057   0.6975   0.6999
```

再将 $\lambda=0.7$ 改为 $\lambda=0.6$,程序执行 5 次,结果为

```
n =     10       100      1000     10000    100000   1000000
c =  0.6000   0.6700   0.6080   0.5939   0.6016   0.6002
c =  0.7000   0.5300   0.5800   0.5991   0.6003   0.5997
c =  0.2000   0.5400   0.5890   0.5925   0.5975   0.5998
c =  0.8000   0.7000   0.6000   0.5981   0.5975   0.5996
c =  0.7000   0.6500   0.6060   0.5992   0.5986   0.6001
```

由上述执行结果可以看出,随着样本容量的增大,随机数小于 0.7 的频率逐渐稳定在 0.7 附近,可见,$P\{\lambda<0.7\}=0.7$.同样有 $P\{\lambda<0.6\}=0.6$.

三、条件概率、全概公式与伯努利概型

若事件 B 的发生会影响到事件 A 的发生,则事件 B 发生的条件下,事件 A 发生的概率称为条件概率.计算条件概率可用公式

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

若事件 A 的发生与事件 B 的发生与否没有关系,即事件 B 发生与否不会影响到事件 A 的发生,反之亦然,则称事件 A 与事件 B 是相互独立的,这时有:$P(AB)=P(A)P(B)$.如袋中有 10 只球,其中白球 7 只,黑球 3 只.分有放回和无放回两种情况,分三次取球,每次取一个,分别求① 第三次摸到了黑球的概率,② 第三次才摸到黑球的概率,③ 三次都摸到了黑球的概率.并用计算机模拟这一过程.

当有放回地摸球时,由于三次摸球互不影响,因此三次摸球相互独立,从理论上可求得:① 第三次摸到黑球的概率为 $\frac{3}{10}=0.3$;②第三次才摸到黑球的概率为 $\frac{7}{10}\cdot\frac{7}{10}\cdot\frac{3}{10}=0.147$;③ 三次都摸到黑球的概率为 $\frac{3}{10}\cdot\frac{3}{10}\cdot\frac{3}{10}=0.027$.用计算机模拟这一过程时,可在[0,1]区间上产生三次随机数来模拟三次摸球.当随机数小于 0.7 时可认为摸到了白球,否则认为摸到了黑球.重复 10^6 次分别求上述三种情况出现的频率,并与理论值进行比较.反复五次进行相互比较,程序如下

```
a = round(rand(1000000,3) - 0.2);
for i = 1:6
  b = a(1:10^i,3);
```

```
  c(i) = sum(b)/(10^i);
end
c
for i = 1:6
  b = (~a(1:10^i,1))&(~a(1:10^i,2))&a(1:10^i,3);
  d(i) = sum(b)/(10^i);
end
d
for i = 1:6
  b = a(1:10^i,1)&a(1:10^i,2)&a(1:10^i,3);
  e(i) = sum(b)/(10^i);
end
e
```

执行结果为

```
n =     10       100     1000   10 000   100 000   1 000 000
c =  0.5000   0.2800   0.2740   0.2981   0.3007   0.3005
d =  0.1000   0.1400   0.1270   0.1469   0.1473   0.1472
e =  0.1000   0.0300   0.0300   0.0262   0.0269   0.0271
c =  0.4000   0.2600   0.3000   0.3021   0.2990   0.3004
d =  0.1000   0.1200   0.1470   0.1432   0.1462   0.1470
e =  0.1000   0.0200   0.0270   0.0275   0.0272   0.0271
c =  0.1000   0.2700   0.2920   0.3011   0.3017   0.3002
d =       0   0.1000   0.1460   0.1461   0.1468   0.1473
e =       0   0.0400   0.0260   0.0250   0.0276   0.0271
c =  0.4000   0.3200   0.2970   0.3067   0.3004   0.2999
d =  0.3000   0.1700   0.1520   0.1522   0.1494   0.1469
e =       0   0.0200   0.0280   0.0267   0.0267   0.0270
c =  0.4000   0.3000   0.2870   0.2933   0.2968   0.3004
d =  0.1000   0.1300   0.1350   0.1434   0.1451   0.1471
e =  0.1000   0.0200   0.0270   0.0238   0.0269   0.0271
```

执行结果中 c 为第三次摸到黑球的概率，d 为第三次才摸到黑球的概率，e 为三次都摸到黑球的概率. 可以看到，随着试验次数的增加，其频率都会逐渐稳定在理论值附近.

当无放回地摸球时，由于第二次摸球会受到第一次的影响，第三次摸球又会受到前两次的影响，因而三次摸球相互影响，并不独立. 从理论上可求得：① 第

三次摸到黑球的概率为$\frac{7}{10}\cdot\frac{6}{9}\cdot\frac{3}{8}+\frac{7}{10}\cdot\frac{3}{9}\cdot\frac{2}{8}+\frac{3}{10}\cdot\frac{7}{9}\cdot\frac{2}{8}+\frac{3}{10}\cdot\frac{2}{9}\cdot\frac{1}{8}=0.3$,② 第三次才摸到黑球的概率为$\frac{7}{10}\cdot\frac{6}{9}\cdot\frac{3}{8}=0.175$,③ 三次都摸到了黑球的概率为$\frac{3}{10}\cdot\frac{2}{9}\cdot\frac{1}{8}=0.008\ 3$.用计算机模拟这一过程时,在[0,1]区间模拟第一次摸球,当值小于0.7时认为摸到了白球,否则认为摸到了黑球;第二次摸球时由于少了一个球,故可在长度为0.9的区间上模拟,若第一次摸到白球,可将区间设为[0.1,1],否则区间设为[0,0.9];第三次摸球可依此类推,其模拟程序如下

```
a = rand(1000000,3);
a(:,1) = round(a(:,1) - 0.2);
a(:,2) = round(a(:,2) * 0.9 - 0.2 - 0.1 * (a(:,1) - 1));
a(:,3) = round(a(:,3) * 0.8 - 0.2 - 0.1 * (a(:,1) - 1) - 0.1 * (a(:,2) -
1));
for i = 1:6
  b = a(1:10^i,3);
  c(i) = sum(b)/(10^i);
end
c
for i = 1:6
  b = (~a(1:10^i,1))&(~a(1:10^i,2))&a(1:10^i,3);
  d(i) = sum(b)/(10^i);
end
d
for i = 1:6
  b = a(1:10^i,1)&a(1:10^i,2)&a(1:10^i,3);
  e(i) = sum(b)/(10^i);
end
e
```

执行结果为

```
n =     10       100     1000   10 000   100 000   1 000 000
c =  0.5000   0.3000   0.2960   0.2985   0.2985   0.2993
d =  0.1000   0.1700   0.1740   0.1768   0.1742   0.1746
e =       0        0   0.0060   0.0095   0.0087   0.0084
c =  0.4000   0.2200   0.2640   0.3007   0.3013   0.3004
d =  0.1000   0.1400   0.1450   0.1756   0.1758   0.1752
```

```
e =       0         0    0.0030    0.0080    0.0079    0.0083
c =  0.4000    0.3200    0.3040    0.3034    0.2994    0.3004
d =  0.1000    0.1900    0.1740    0.1726    0.1743    0.1751
e =       0    0.0100    0.0080    0.0088    0.0079    0.0084
c =  0.1000    0.2700    0.3060    0.3016    0.3004    0.3004
d =       0    0.1000    0.1850    0.1733    0.1742    0.1753
e =       0         0    0.0050    0.0082    0.0085    0.0084
c =  0.4000    0.2900    0.2800    0.3039    0.3007    0.3002
d =  0.2000    0.1600    0.1660    0.1758    0.1778    0.1752
e =       0    0.0100    0.0110    0.0083    0.0082    0.0083
```

可见,模拟结果与理论计算是吻合的.这里在理论上计算第三次摸到黑球的概率时,我们用到了一个很重要的公式——全概率公式,即若 $A_1, A_2, \cdots, A_n$ 构成一个完备事件组,且事件 B 的发生总是伴随着事件 $A_i (i = 1,2,\cdots,n)$ 中的某一个的发生而发生,则有

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

这里第三次摸到黑球的四种情况分别是:{白,白,黑},{白,黑,黑},{黑,白,黑},{黑,黑,黑}.这四种情况构成了完备事件组.现若进一步提出如下问题:① 当不放回时,已知第三次摸到了黑球,问前两次是黑球的概率为多少? ② 若有放回地连续摸 10 次,则恰有 3 次摸到黑球的概率为多少? 第一问显然是一逆概问题,由逆概公式即贝叶斯公式得到其概率应为 $3/10 \times 2/9 \times 1/8/0.3 = 1/36 \approx 0.0278$.第二问则属伯努利概型,所谓伯努利概型是指相同条件下,进行 n 次独立重复试验,每次试验只有事件 A 发生或不发生两种结果,且 $P(A) = p$, $P(\bar{A}) = 1 - p = q$.这里 A 为{摸到的是黑球},故 $P(A) = 0.3, P(\bar{A}) = 0.7$.于是由二项概率公式有,10 次有放回摸球中,恰有 3 次摸到黑球的概率为 $C_{10}^3 \times 0.3^3 \times 0.7^7 \approx 0.266\ 8$.同样我们可以在计算机上模拟这两个过程.程序及模拟结果如下

```
a = rand(1000000, 3);
a(:, 1) = round(a(:, 1) - 0.2);
a(:, 2) = round(a(:, 2) * 0.9 - 0.2 - 0.1 * (a(:, 1) - 1));
a(:, 3) = round(a(:, 3) * 0.8 - 0.2 - 0.1 * (a(:, 1) - 1) - 0.1 * (a(:, 2) -
1));
for i = 1:6
  b = a(1:10^i, 3);
  c(i) = sum(b);
```

```
    b = a(1:10^i, 1)&a(1:10^i, 2)&a(1:10^i, 3);
    d(i) = sum(b);
    e(i) = d(i)./c(i);
end
e
e =       0       0  0.0114  0.0266  0.0263  0.0277
e =       0  0.0313  0.0263  0.0290  0.0264  0.0278
e =       0       0  0.0163  0.0272  0.0282  0.0278
e =       0  0.0345  0.0393  0.0273  0.0273  0.0275
e =  0.3333  0.0345  0.0340  0.0278  0.0276  0.0278
a = round(rand(1000000,10) - 0.2);
for i = 1:6
    b = sum(a(1:10^i,:), 2) - 3;
    c(i) = sum(~b)/(10^i);
end
c
c =  0.3000  0.2600  0.3000  0.2712  0.2665  0.2666
c =  0.4000  0.2200  0.2660  0.2713  0.2686  0.2674
c =  0.2000  0.1900  0.2470  0.2613  0.2662  0.2665
c =  0.4000  0.2600  0.2840  0.2727  0.2670  0.2666
c =  0.4000  0.2900  0.2580  0.2674  0.2668  0.2669
```

从上述模拟结果可以看出，模拟的结果与理论计算是吻合的．请读者读懂上述两段程序，并写出模拟的思路．

第二节　随机变量的分布及其数字特征

随机变量的统计行为完全决定于其概率分布．按随机变量的取值不同，通常可将其分为离散型、连续型和奇异型三大类．由于奇异型在实际应用中很难碰到，因而这里我们不作讨论，仅讨论离散型和连续型两类随机变量的概率分布及其数字特征．

一、离散型随机变量的分布及其数字特征

如果随机变量 X 的所有可能取值为有限个或无穷可列个，则称 X 为离散型随机变量．设 X 的所有可能值为 $X_1, X_2, \cdots$，并且 X 取这些值的概率为

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