《多元统计分析》第一次上机作业

3180103000 许乐乐

习题一

利用R软件生成随机数，分析如下几个参数对中心极限定理(CLT)的影响：

1. **样本容量 n：是否n的取值越大，CLT的拟合效果越好?**
2. **模拟次数 m: 分析不同模拟次数对于CLT拟合效果的影响；**
3. **随机数的分布：在相同样本容量和模拟次数下，取自不同分布的随机数对CLT有什么影响?**

- 中心极限定理（Central Limit Theorem，CLT）

  设随机变量$X_1, X_2, \ldots, X_n$独立同分布，并且具有有限的数学期望和方差，
  $EX_i = \mu,\ DX_i = \sigma^2\ (i = 1, 2, \ldots, n)$，则

$$S_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \xrightarrow{D} N(0,1)$$

- 思路：
  - 利用R语言中的函数产生服从某分布的随机数n个
  - 运用公式求值

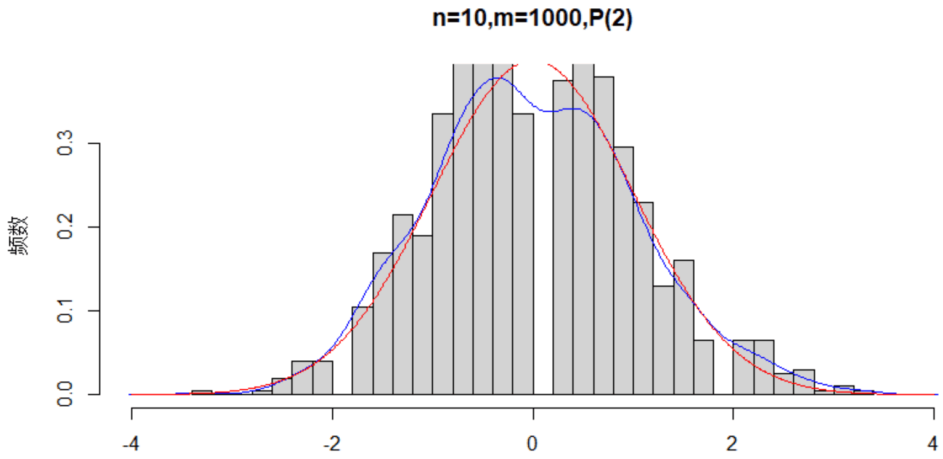
$$S_n = \frac{\sum_{i=1}^{n} X_i - nEX_i}{\sqrt{nDX_i}}$$

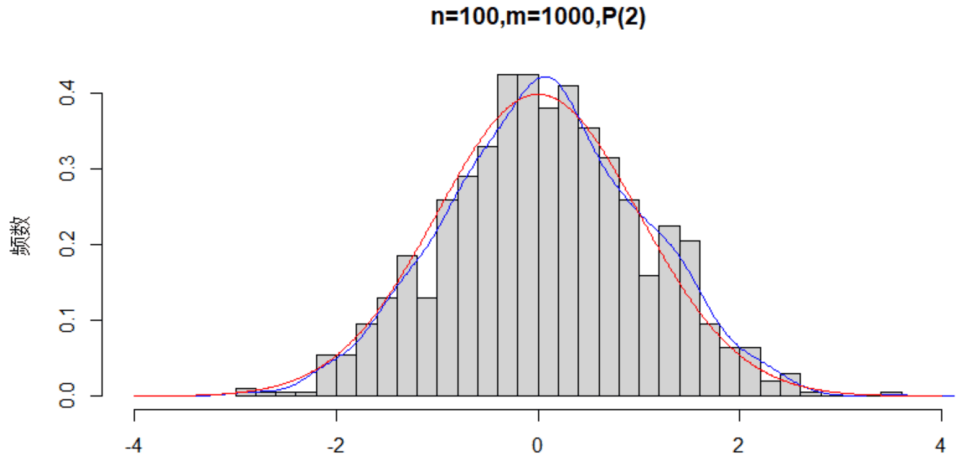
  - 重复上述步骤m次，获得m个样本点
  - 求出样本点的核密度函数，画出图像，并与标准正态分布的密度函数图像进行比较
- 代码：

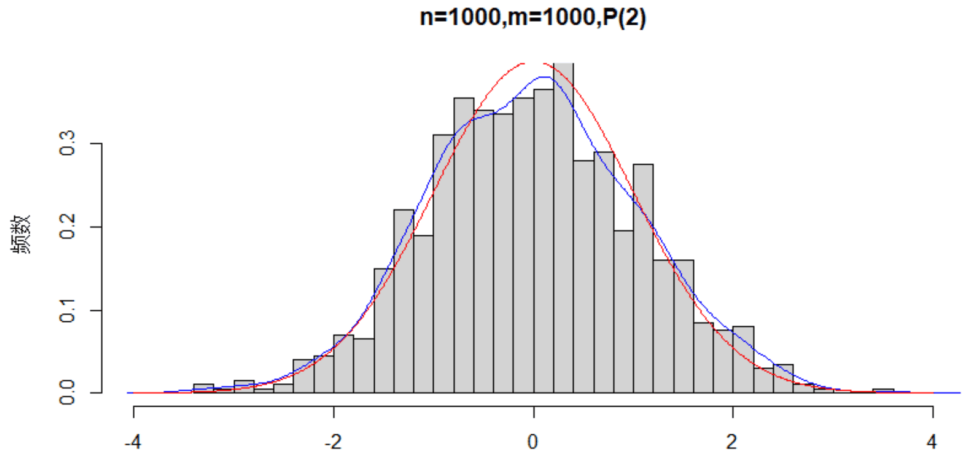
```
#以 n=1000,m=1000,Possion分布P(2)为例
library(ggplot2)
n=1000
m=1000
lambda=2
s<-c(1:m)
for(i in 1:m){
  x=rpois(n,lambda)
  s[i]=(sum(x)-n*lambda)/(sqrt(n)*lambda)
}
tmp.dens=density(s)
hist(s,breaks=30,freq=FALSE,xlim=c(-4,4),ylim=c(0,max(tmp.dens$y)),
     main='n=1000,m=1000,P(2)',xlab='', ylab='频数') # 样本点直方图
lines(tmp.dens, col='blue') # 样本点核密度
y<-seq(-4,4,0.01)
lines(y,dnorm(y,mean=0,sd=1),type="l",col="red",lty=1) #标准正态密度函数
```

- 图表与结论

  1. 样本容量 n对CLT的拟合效果的影响：
     - 样本容量 n为10，100，1000（m=1000，泊松分布P(2)随机数）：

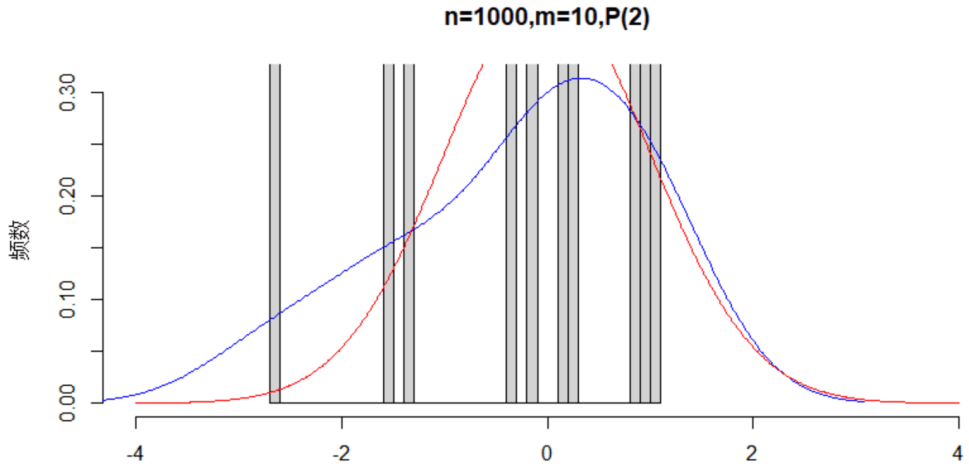


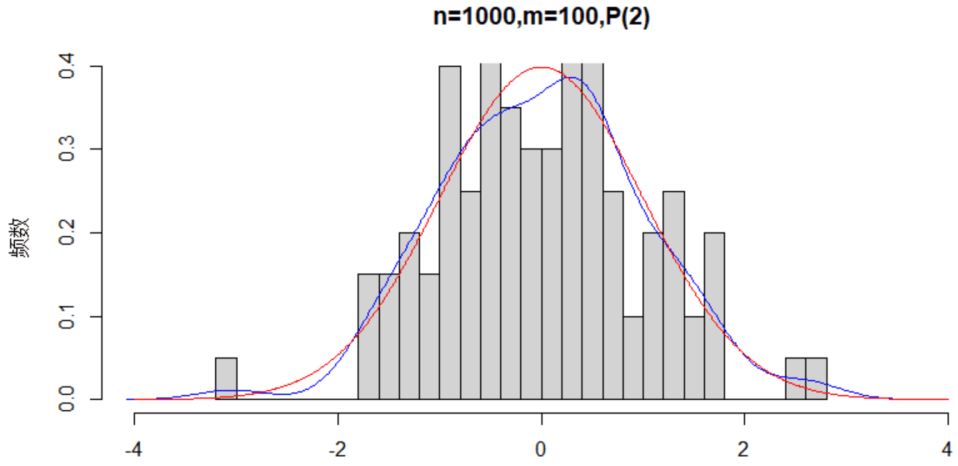


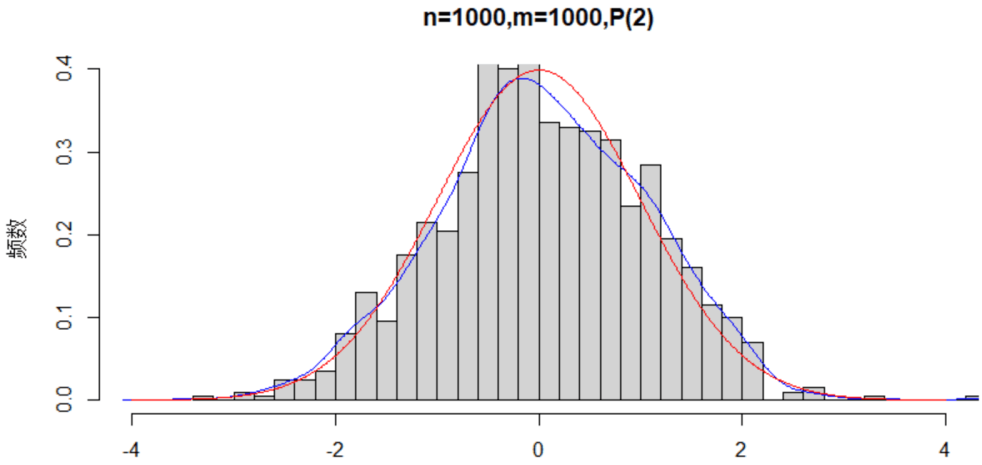


由直方图和核密度可以看出，n的取值越大，直方图中某样本区间空缺的情况越少，核密度曲线与标准正态密度曲线越贴合，样本点的分布越趋近于标准正态分布。因此，n的取值越大，CLT的拟合效果越好。

2. 模拟次数m对CLT的拟合效果的影响：
    - 模拟次数 m为10，100，1000（n=1000，泊松分布P(2)随机数）：



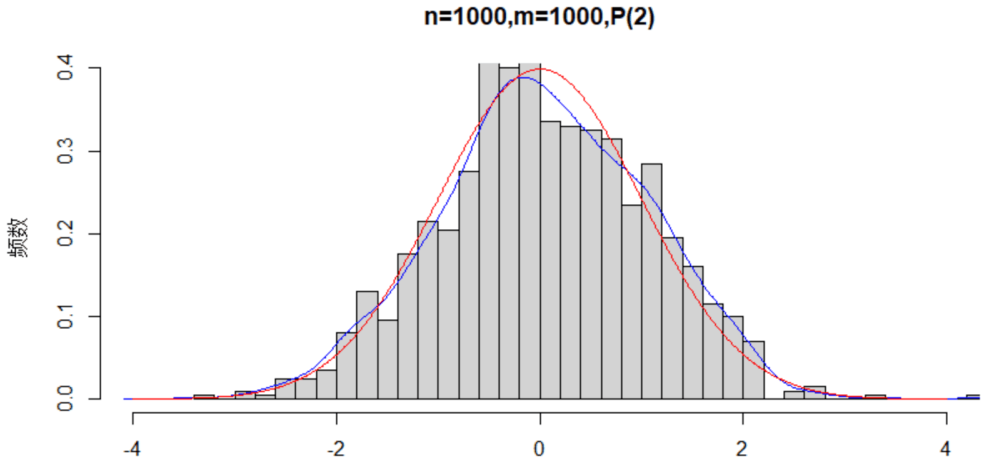


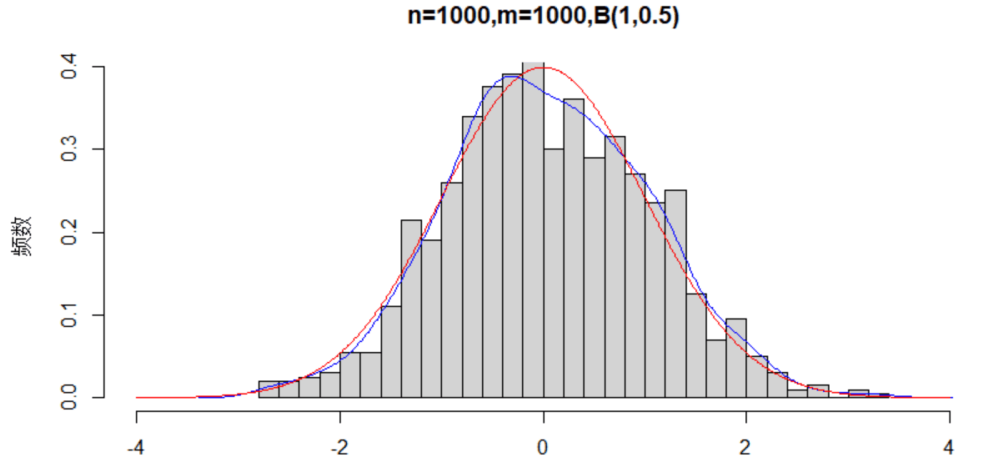


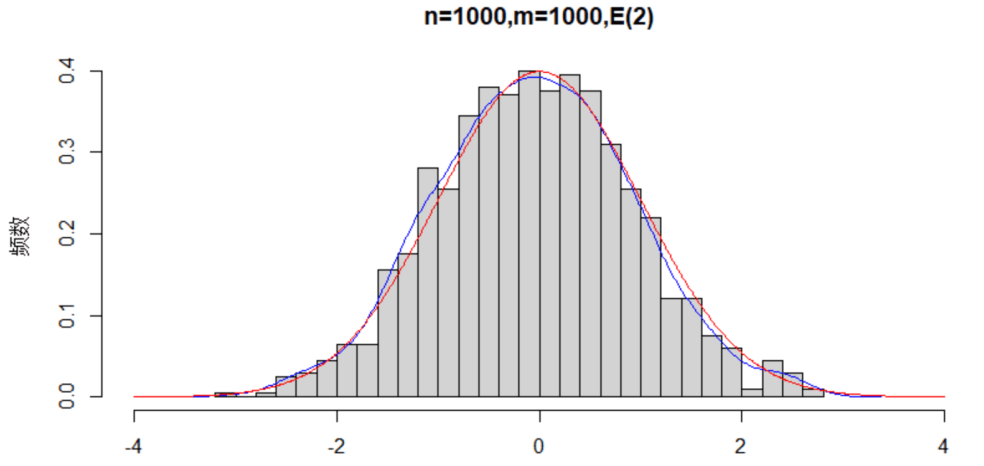
由直方图和核密度可以看出，m的取值越大，直方图中某样本区间空缺的情况越少，核密度曲线与标准正态密度曲线越贴合，样本点的分布越趋近于标准正态分布。直方图和核密度随着m的取值变化时的波动比随着n的取值变化时的波动更大，因为样本点的容量是m。因此，m的取值越大，CLT的拟合效果越好。

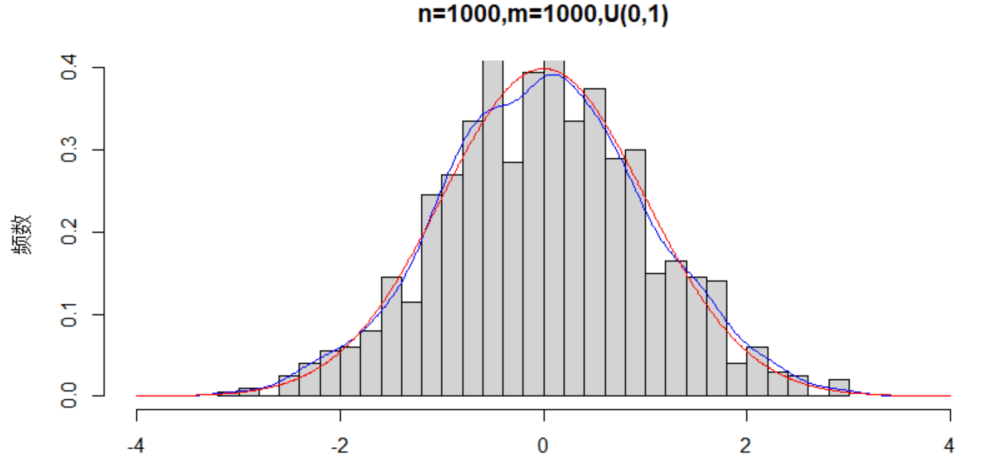
3. 随机数的分布对CLT的拟合效果的影响：

   - 随机数的分布为泊松分布P(2)，二项分布B(1,0.5)，指数分布E(2)，均匀分布U(0,1)（n=1000，m=1000）：


n=1000,m=1000,P(2)
频数
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4

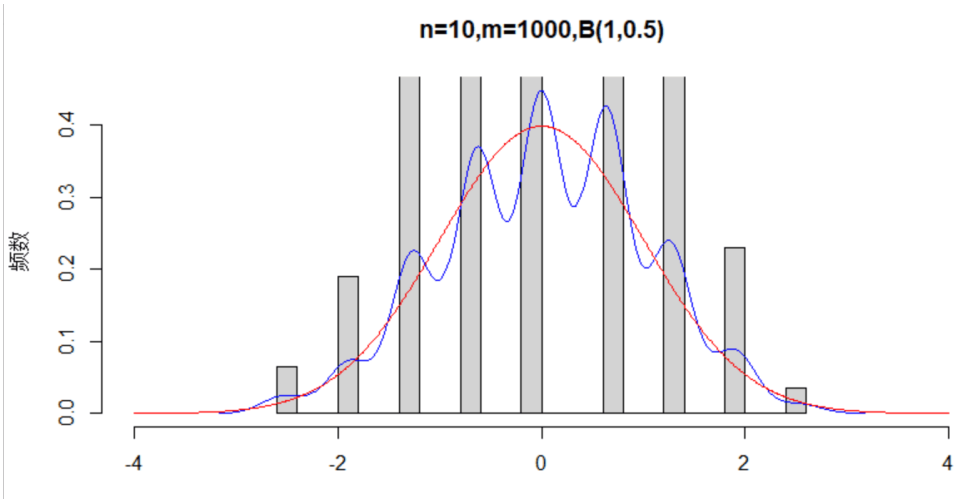

n=1000,m=1000,B(1,0.5)
频数
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4

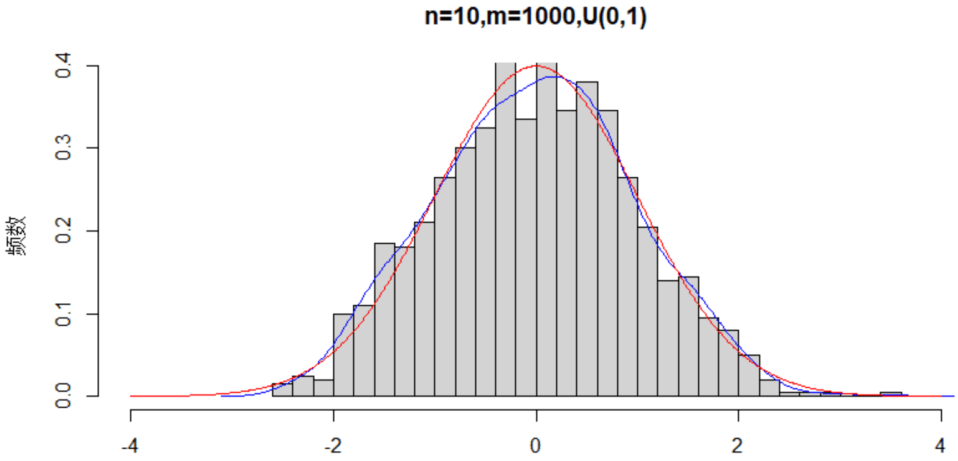

n=1000,m=1000,E(2)
频数
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4


n=1000,m=1000,U(0,1)
频数
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4

由直方图和核密度可以看出，在n和m的取值较大时，不论是服从什么分布的随机数，样本点的分布都趋近于标准正态分布。因此，在n和m的取值较大时，取自不同分布的随机数对CLT影响不大。

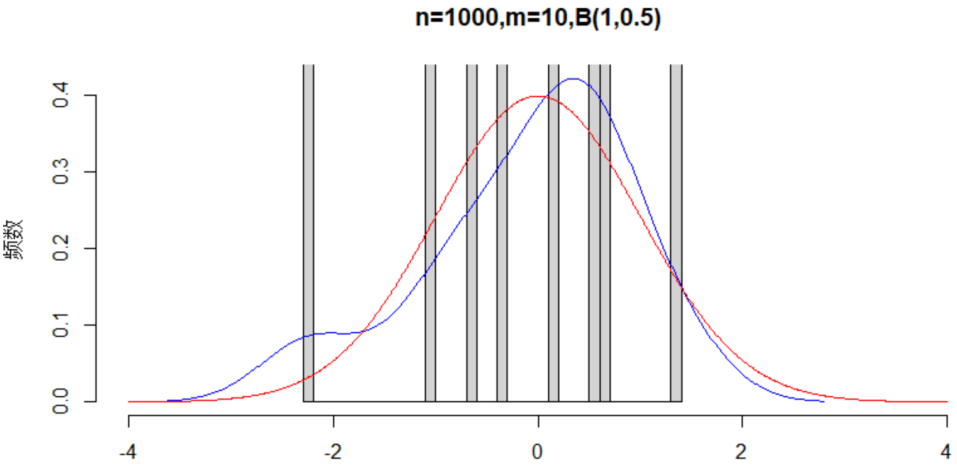
- 随机数的分布为二项分布B(1,0.5)，均匀分布U(0,1)（n=10，m=1000）：

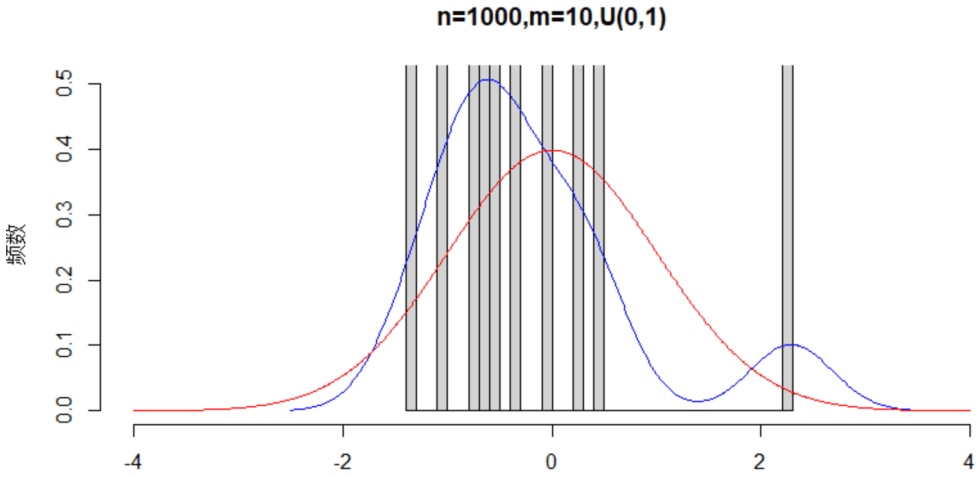




由直方图和核密度可以看出，当n取值较小时，服从二项分布的随机数生成的直方图的样本区间空缺较多，而均匀分布的则不会，这是由于随机数服从离散或均匀分布的缘故。因此，在n取值较小时，取自不同分布的随机数对CLT影响很大。

- 随机数的分布为二项分布B(1,0.5)，均匀分布U(0,1)（n=1000，m=10）：





由直方图和核密度可以看出，当m取值较小时，二项分布和均匀分布进行CLT拟合的效果都很差。因此，在m取值较小时，无法看出取自不同分布的随机数对CLT的影响。