数据科学基础 I (Matlab)

东北大学

数据科学中的数学基础与运算基础

连续随机变量vs离散随机变量

- 若随机变量X的取值为有限个或可数个，则称X为离散型随机变量。
  - 掷硬币的朝上面：正面、反面
  - 掷骰子的朝上面：1、2、3、4、5、6
  - 某班级的人数
  - ……

连续随机变量vs离散随机变量

- 若随机变量X的取值为不可数个，则称X为连续型随机变量。
  - 某地区人的身高
  - 在车站的候车时长
  - 某化学品中某一成分的占比
  - ……

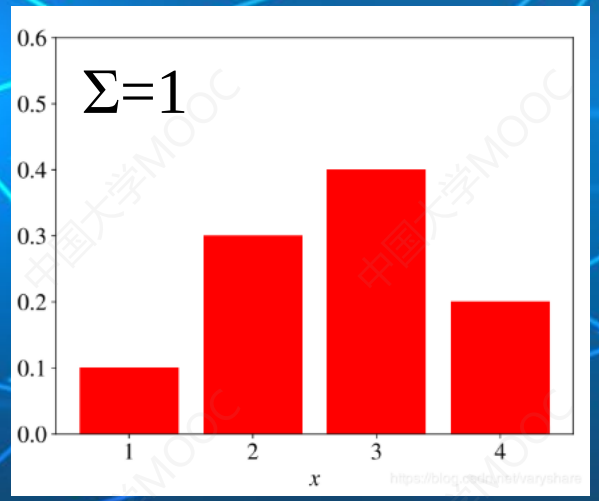
对随机变量取值的描述

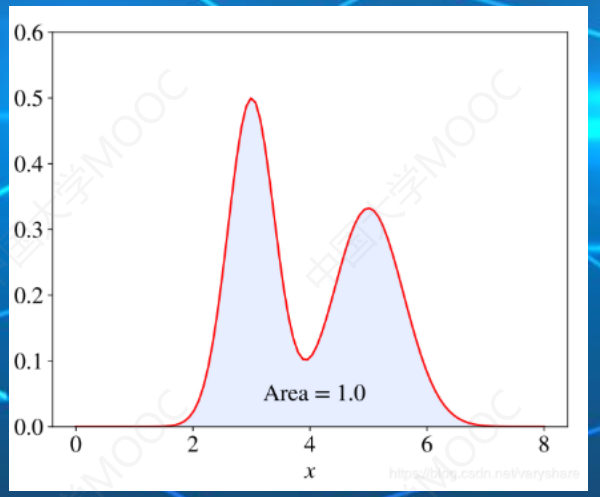
- 期望E
  - 刻画了随机变量取值的平均水平
- 方差D
  - 刻画了随机变量取值与期望的平均偏离程度
- 标准差
  - 方差的平方根

对随机变量取值的分布描述

- 离散随机变量——概率质量函数
  Probability Mass Function
  - 指各个分类的概率
- 连续随机变量——概率密度函数
  Probability Density Function
  - 指数据落在某一段连续的区间的概率

对随机变量取值的分布描述





常见离散随机变量的概率分布

- 伯努利分布
- 二项分布
- 几何分布
- 泊松分布

常见离散随机变量的概率分布

伯努利分布

- 描述只有两种可能结果的单次随机试验
- 分布函数：$P(X=k)=p^k\ (1-p)^{1-k}$
- 期望：$E=p$
- 方差：$D=p(1-p)$

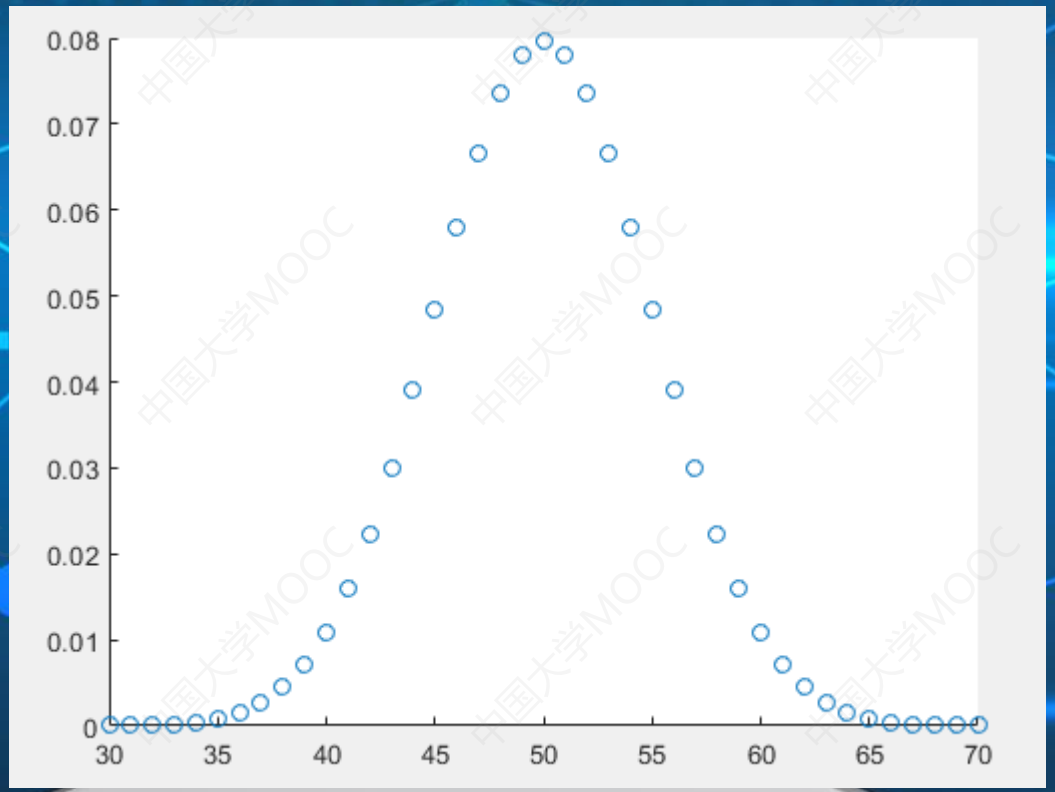
常见离散随机变量的概率分布

二项分布

- 二项分布是n重伯努利试验（即独立的进行n次伯努利试验）成功次数的离散概率分布，记为$X \sim B(n,p)$。
- 分布函数：$P(X=k)=\left(\frac{n!}{k!(n-k)!}\right)p^k(1-p)^{1-k}$
- 期望：$E=np$
- 方差：$D=np(1-p)$

常见离散随机变量的概率分布

二项分布



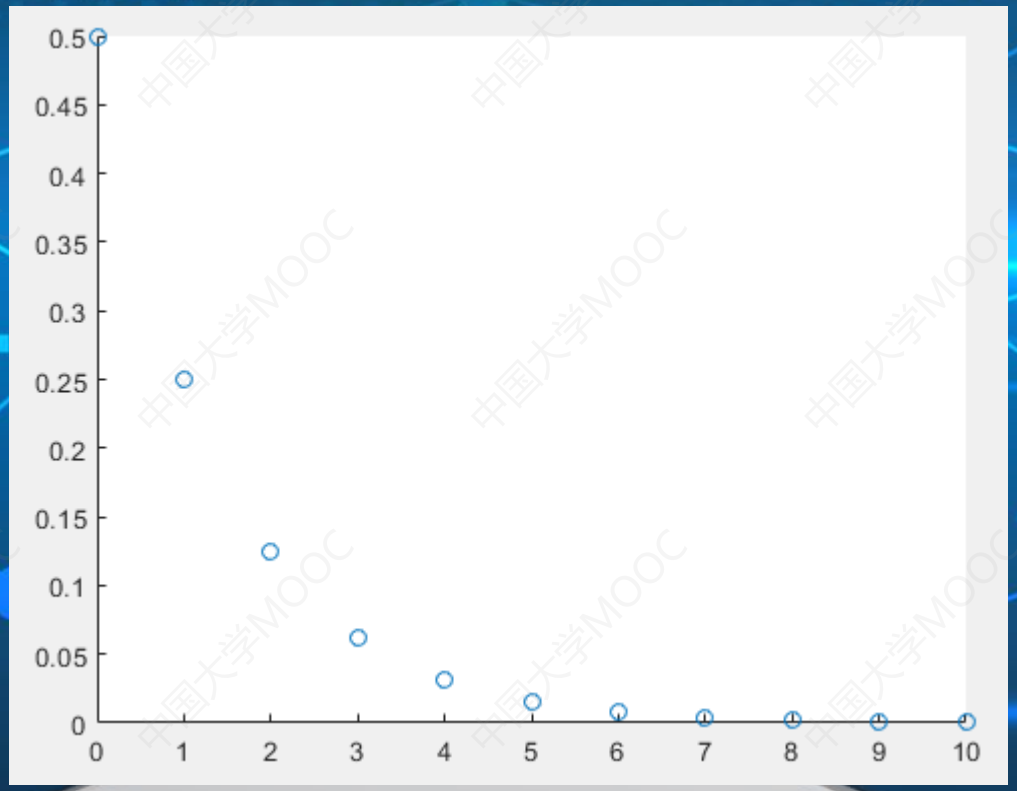
常见离散随机变量的概率分布

几何分布

- 在n次伯努利试验中，第k次试验才得到第一次成功的概率分布称为几何分布。
- 分布函数：$P(X=k)=(1-p)^{k-1}p$
- 期望：$E=1/p$
- 方差：$D=\frac{1-p}{p^2}$

常见离散随机变量的概率分布

几何分布



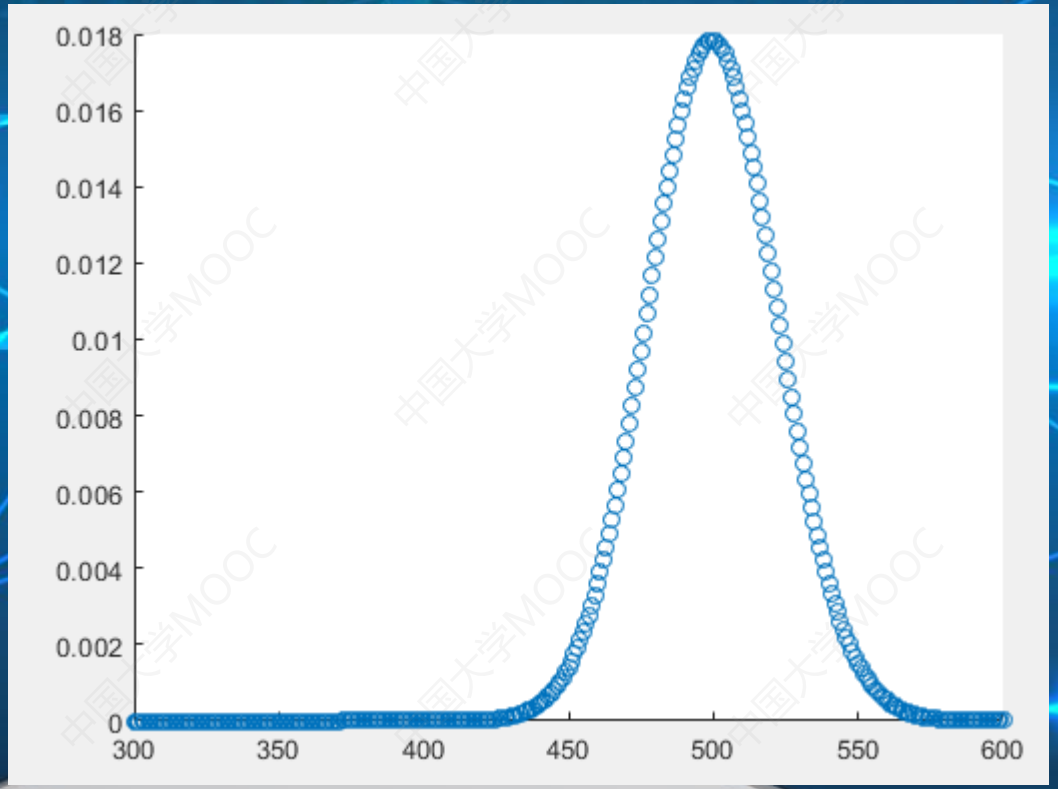
常见离散随机变量的概率分布

泊松分布

- 描述单位时间/面积内，随机事件发生的次数。
- 分布函数：$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda}$
- 期望：$E=\lambda$
- 方差：$D=\lambda$

常见离散随机变量的概率分布

泊松分布



常见连续随机变量的概率分布

- 正态分布（高斯分布）
- 均匀分布
- 指数分布

常见连续随机变量的概率分布

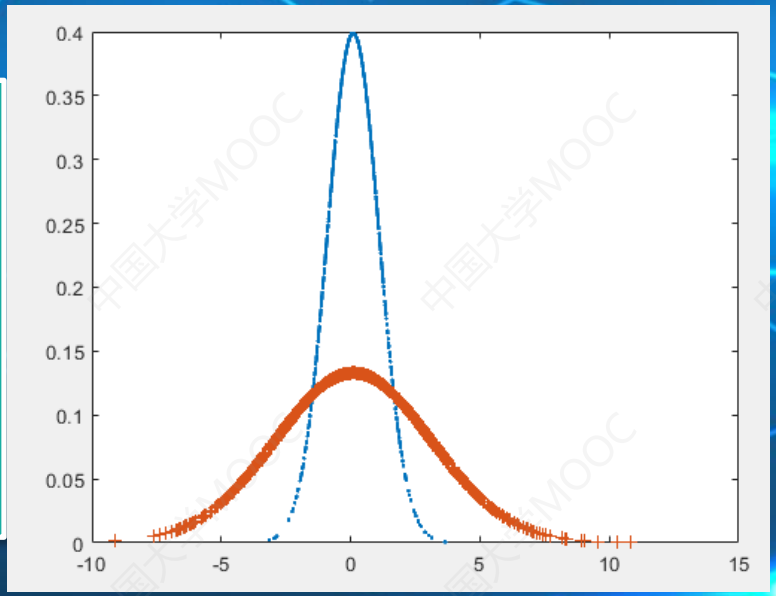
正态分布（高斯分布）

- 表现为两边对称，是一种钟形的概率分布。
- 概率密度函数：$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{1(\frac{1}{2})\left[\frac{x-\mu}{\sigma}\right]^2}$
- 期望：$E=\mu$
- 方差：$D=\sigma^2$

常见连续随机变量的概率分布

正态分布（高斯分布）

```
>>x=normrnd(mu,sigma,m,n
 %产生m*n阶均值为mu，方差为sigma正态分布的随机数矩阵
>> d=pdf('norm',x,mu,sigma)
>> plot(x,d,'.')
```



常见连续随机变量的概率分布

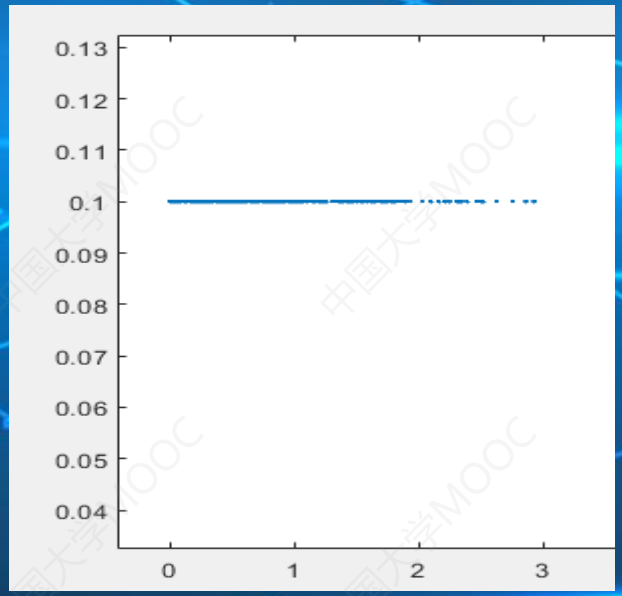
均匀分布

- 均匀分布是指连续型随机变量所有可能出现值的出现概率都相同
- 概率密度函数：$f(x)=\frac{1}{b-a}(a\le x\le b)$
- 期望：$E=\frac{b-a}{2}$
- 方差：$D=\frac{(b-a)^2}{12}$

常见连续随机变量的概率分布

均匀分布

```
>>x=unifrnd (a,b,m, n)
%产生m*n阶［a，b］均匀分布U（a，b）的随机数矩阵
>> d=pdf('unif',x,A,B)
>> plot(x,d,'.')
```



常见连续随机变量的概率分布

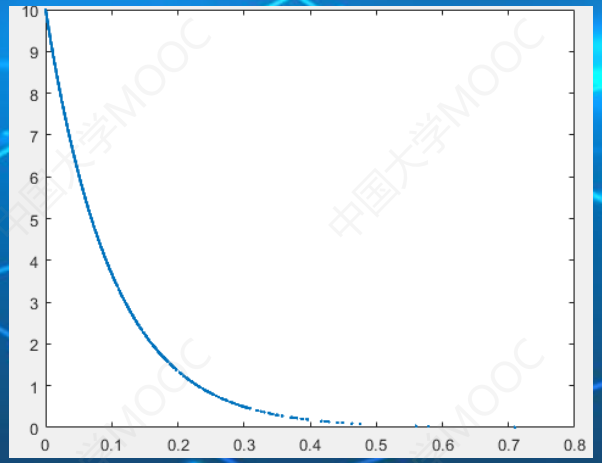
指数分布

- 指数分布通常用来表示随机事件发生的时间间隔
- 概率密度函数： $f(x)=\frac{1}{\theta}e^{-x/\theta}\ (x>0)$
- 期望： $E=\theta$
- 方差： $D=\theta$

常见连续随机变量的概率分布

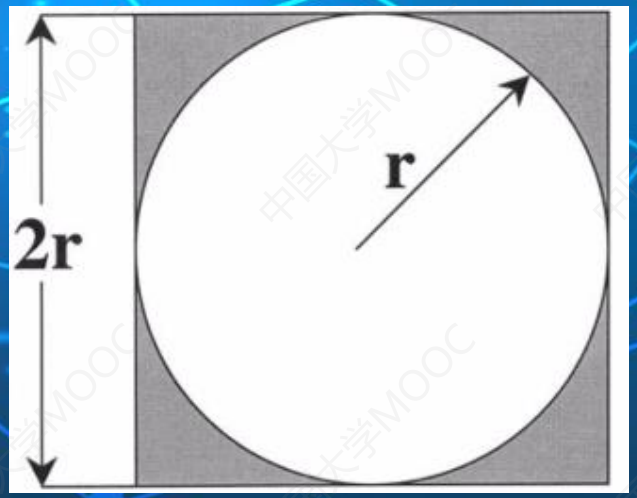
指数分布

```
>>x=exprnd (mu,m, n)
 %产生m*n阶均值为mu的指数分布的随机数矩阵
>>d=pdf('exp',x,mu);
>> plot(x,d,'.')
```



概率的计算应用——蒙特卡洛Monte Carlo

- 是一种近似的模拟计算过程



$$p = \frac{\pi r^2}{4r^2} = \frac{\pi}{4}$$