# 概率统计B

## Probability and Statistics

张思容

zhangsirong@buaa.edu.cn

数学与系统科学学院，北京航空航天大学
School of Mathematics and System Sciences, BUAA

October 14, 2014

# Week 4：随机变量与离散分布

1. 随机变量
   - 随机变量
   - 分布函数
2. 离散随机变量
   - 二项分布
   - 泊松分布
   - 复合分布及分布的特征

# Review: 回顾

RECALL:

- 概率模型=样本空间+概率律
- 古典模型: 一个因素影响下的概率模型。
- 复杂系统(或重复试验): 很多个(独立)因素影响下的概率模型?

TODAY

- 随机变量与分布函数;
- 二项分布。

Galton box: (bean machine)

# 随机变量的定义

什么是随机变量? random variable

- 直接解释: 对每一个样本结果给出一个(实)数值。
- 数学解释; 定义在一个样本空间上的函数。
  对比: 函数的自变量与因变量。

## EXAMPLE (投硬币)

投硬币三次, 样本空间有8个元素。记每个结果中正面(H)的个数为$X$.
$X$是个随机变量。

$X$取值$0, 1, 2, 3 \in \mathcal{R}$. 同理可定义随机变量$Y$为每个结果中反面的个数。

## Definition (定义:随机变量)

随机变量$X$是定义在样本空间$\Omega$上的实值函数:即每一个样本点$\omega$,$X(\omega)$是个实数。 通常用$X, Y, Z$或$\xi, \eta$等表示。

例子: 人体正常体温$T$ =多少摄氏度。华氏度$= 18/10c + 32$?

# 概率律的转移

## Remark (push-back)

随机变量引导概率律转移到实数上。
设$x \in \mathcal{R}$, $X^{-1}(x) = \{\omega_1, \omega_2, \dots\}$.
则$P(X = x) = \sum_{\omega_i} P(\omega_i), \omega_i \in \Omega$.

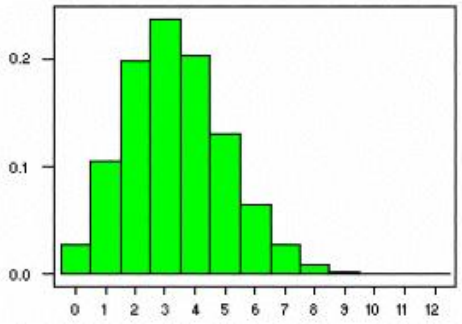
- $X$取值为离散的,称为离散随机变量, 概率律变成一个概率序列, 称为分布列;
- $X$取值包含连续区间, 概率律呢?
  注意:$P(X = x_0) = 0$!, 概率律变成一个(密度)函数。

## EXAMPLE (人体温度)

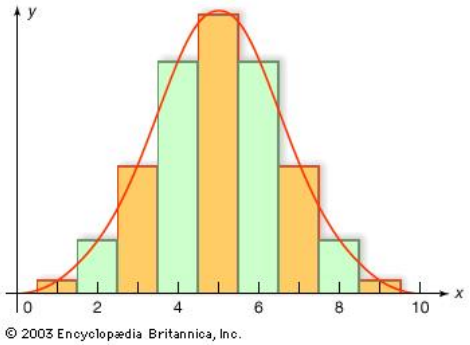
正常体温在$\mathcal{R}$上的有意义表示$36.8 \le X \le 37.2$(事件), 概率计算:
$P(36.8 \le X \le 37.2) = P(\{\omega : 36.8 \le X(\omega) \le 37.2\})$。
区间足够小? 得到一个密度函数。

# 分布函数的定义

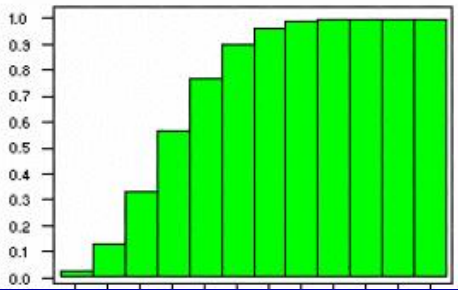
离散分布




连续分布




累积分布




**Definition (概率分布函数)**

给定随机变量$X$，定义函数

$$F(x) = P(X \leq x), -\infty \leq x \leq +\infty$$

称$F$为$X$的概率分布函数，简称分布函数或累积分布函数$CDF$。

分布函数的充分必要条件：

- $0 \leq F(x) \leq 1$.
- $F(x)$单调非减
- $F(x)$右连续；***

通常随机变量对应一些常见分布函数，直接称随机变量为某个分布。

# 随机变量的注解***

- 随机变量不仅仅是一个函数$X:\Omega\to\mathcal{R}$;
  更重要的是得到分布函数。可以记为一个变换$(\Omega,P)\to(\mathcal{R},F)$。
  事实上分布函数更重要。
- 随机变量的函数还是随机变量(复合函数)。$\Omega\to\mathcal{R}\to\mathcal{R}$
- 区分分布函数可以用函数的数字特征：期望，方差等等。
- 随机变量分类：离散随机变量，非离散随机变量(我们一般仅仅考虑其中的连续随机变量);

# 分布函数的判断

分布函数的性质

- $F(-\infty)=0, F(+\infty)=1$
- $P(a \le X \le b)=F(b)-F(a)$
- $P(X=x)=F(x)-F(x^-)$ ***

**EXAMPLE**

设$F(x)=a+be^{-x}, x \ge 0$, $F(x)$是否可以是某个随机变量的分布函数? 特别求概率$P(X \le \ln 2)$.

$a=1, b=-1, P=1/2$

说明： 有很多分布函数，但常见只有几种。

# 随机变量与决策

## EXAMPLE (轮盘赌 Roulette)

美式轮盘赌有*38*个数值，最简单的玩法押一个数值一元，可赢得*35*元。 问一次押注的期望赢钱多少?

解答: $EX = 35\frac{1}{38} + (-1)\frac{37}{38} = -1/19 \approx -0.053$

# 定义

**Definition (离散随机变量)**

随机变量取值为有限或可数个时，称为离散随机变量。
记其值为$x_1.x_2,\ldots,x_k,\ldots$。
特别记$P(X=x_k)=p_k, k=1,2,\ldots$，$P$是$X$对应的概率分布，称$\{p_k\}$为概率分布列。

- 常记为

| $X$ | $x_1$ | $x_2$ | $x_3$ | $\ldots$ |
|---|---|---|---|---|
| $P$ | $p_1$ | $p_2$ | $p_3$ | $\ldots$ |

- $p_k \geq 0, \sum p_k = 1.$
  注：$\Omega = \bigcup_i \{\omega | X(\omega) = x_i\}$
- 分布函数$F(x)$是阶跃函数。

# 二项分布

## EXAMPLE (两点分布，Bernouli贝努利分布)

$X$取值为0或1时的概率分布是
$P(x=1)=p, P(X=0)=1-p$, 记为$B(1,p)$.
模型：最简单随机变量=一次试验是否成功。

## Definition (二项分布或二项随机变量)

随机变量的取值为$0,1,2,\ldots,N$,且满足
$P(X=k)=C_N^k p^k q^{N-k}, k=1,2,\ldots,N, p+q=1$.
称其满足二项分布，记为$B(N,p)$.

- 模型：$N$次独立试验中的成功次数。
- 称为二项分布因为其值为满足二项式定理的各项。
  $(p+q)^N=\sum C_N^k p^k q^{N-k}$,可递推计算,$N$充分大时,需近似计算;
- 应用：投票权大小(委员会组成); 比赛规则(三局两胜,五局三胜?)...

# 应用例子

## EXAMPLE (基因遗传)

设某个生物特征(眼睛颜色，或左撇子)由一对基因决定。$D$为显性基因，$r$为隐性基因。后代从父母中各得到一个基因。只要有一个显性基因，就必然出现该生物特征。问一对混合型父母(基因为$Dr$)的四个后代中有三个有该生物特征的概率是多少?

解答: 服从二项$B(4,3/4)$，$P(X=3)=C_4^3(\frac{3}{4})^3\frac{1}{4}=\frac{27}{64}$.

## EXAMPLE (安全)

设飞机上每个引擎的失效概率相同为$p$且互相独立.如果一个飞机上的一半引擎正常，则飞机可以正常运行。问$p$为多少时，用四个引擎的飞机比用两个引擎安全?

解答: 四个引擎正常个数服从$B(4,p)$，飞机正常飞行即$X\geq 2$，$P(X\geq 2)=1-(1-p)^4-4p(1-p)^3$;类似两个引擎飞机正常飞行$Y\geq 1, P(Y\geq 1)=1-(1-p)^2$。计算可得$p\geq 2/3$时，四个引擎的飞机更安全.

# N次独立试验中的分布

- 模型： $N$次独立试验 或$N$次有放回抽签。
- 二项分布: $X = k$,试验中的成功次数, $B(N, p)$
  $X = X_1 + X_2 + \cdots + X_N$,其中$X_i$是第$i$次试验是否成功的两点分布。
- 几何分布： $Y = k$,首次成功时的试验次数。
  [几何分布]: 随机变量取值为$1, 2, \ldots, \infty$,且满足
  $P(Y = k) = p(1-p)^{k-1}, k = 1, 2, \ldots, \infty.$
- ***负二项分布： $Y = k$,第$r$次成功时的试验次数。$NB(r, p)$
  $Y = Y_1 + Y_2 + \cdots + Y_r$, $Y_i$一次成功时的试验次数。
  [负二项分布] 随机变量取值为$1, 2, \ldots, \infty$,且满足
  $P(Y = k) = C_{k-1}^{r-1} p^r (1-p)^{k-r}, k = 1, 2, \ldots, \infty.$

Banach火柴盒问题：有两个火柴盒(各有$N$个)，随机选取一个,直到发现一个盒子空了为止。问另外一个盒子有$k$个火柴的概率。
答案： 固定一个盒子： $Y \sim NB(N+1, 1/2)$,试验次数$Y = 2N + 1 - k$.
$P(Y = 2N + 1 - k) = C_{2N-k}^{N} (1/2)^{2N-k+1}$.
两个盒子的概率为上面两倍。

# 无放回抽签的分布***

**Definition (超几何分布)**

随机变量取值为$0, 1, 2, \ldots, n$,且满足

$P(X = k) = \frac{C_M^k C_{N-M}^{n-k}}{C_N^n}, k = 1, 2, \ldots, n;$

称其满足超几何分布，记为$H(n, M, N)$.

- 模型：从$(N, M)$大样本中抽取小样本$n$,$M$是$N$中特别一类(如次品)，则$n$中次品个数$k$服从超几何分布。特别$n = 1, k = 1$有抽签原理!.
- ***较大$N, M$时,可用二项分布逼近。$B(n, p), p = M/N$,即有放回与无放回差别不大。
- 实用例子：估计某地区的某动物总数$N$。
  捕捉一批动物$M$，做标记放回，过一段时间，再捕捉一批动物$n$,其中有标记的为$k$, 利用超几何分布，估计$N$。
  概率为$P(X = k) = \frac{C_M^k C_{N-M}^{n-k}}{C_N^n}$是关于$N$的一个序列，
  我们求最大值对应的$N$,计算有$N = nM/k$. 称为最大似然估计。

# 作业

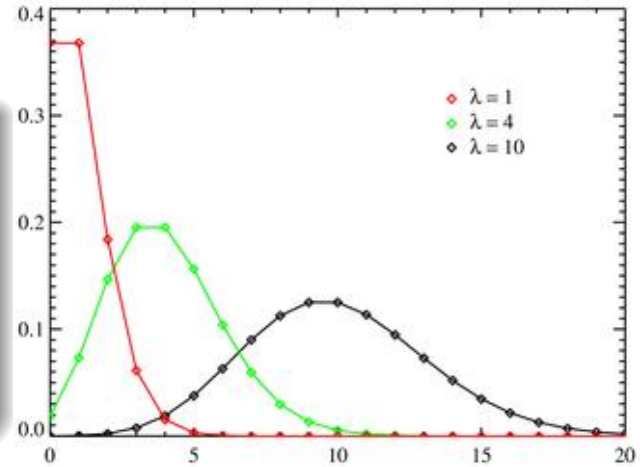
北航教材:

P44 习题二. 1,4,6,8,12,13

# 定义

**Definition (泊松分布)**

随机变量取值为$0, 1, 2, \ldots, k, \infty$,满足
$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, k = 0, 1, 2, \ldots$; $\lambda$为正常数.
称其满足泊松分布，记为$\mathcal{P}(\lambda)$.




- 模型：大量试验中的小概率事件发生次数。
  每年发生战争(或地震)的次数,某一小时进入某邮局的顾客数； 寿命超过100岁的人数;
- 逼近二项分布：当$n$大,$p$小，$np$合适大小时，可用$\lambda = np$的泊松分布近似计算。
- *** 泊松范例：$n$个小概率事件发生的概率$p_i$,所有事件互相独立或弱相依，则事件发生次数服从 $\lambda = \sum p_i$的泊松分布。
  配对问题：$N$个人得到自己作业的次数近似服从$\lambda = N * 1/N = 1$的泊松分布.

# 应用例子

- (事故发生) 设某高速公路段每天发生事故的次数服从$\lambda = 3$的泊松分布，求今天不发生事故的概率。
  解答： $P(X = 0) = e^{-3} \approx 0.05$.
- (印刷错误)设某本书任一页有印刷错误的次数服从$\lambda = 1$的泊松分布， 求某一页出现至少一个错误的概率
  解答： $P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-1} \approx 0.633$

**EXAMPLE (生日问题***)**

任意两个人生日相同的事件是小概率，弱相依的； 设发生事件的次数服从$\lambda = C_n^2 \frac{1}{365}$的泊松分布，求$n$个人里至少两个人同生日的概率。

解答： 一次试验为任选两个人为同一天生日； $p = 1/365$, 试验次数(任选两个人)$C_n^2$. 出现两个人生日相同次数$X$服从泊松分布。
有$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-\lambda}$
如果要求$P(X \geq 1) > 0.5$,即$e^{-n(n-1)/730} \geq 0.5$; 计算有$n \geq 23$.
多少人中有三个人生日相同? $\lambda = C_n^3(\frac{1}{365})^2$.

# 复合函数的分布

构造新的分布： $g(X)$

- 设$X$是随机变量,$g$是$R \to R$的函数, $Y = g(X) : \Omega \to R \to R$， $Y$是随机变量;
- 离散分布: $p_0, p_1, p_2 \cdots$
  例子: $Y = X^2$, $Y = 4 \to X = \pm 2$,
  $P(Y = 4) = P(g^{-1}(4)) = P(X = 2) + P(X = -2)$;
  参见第四章例1.

# 分布的特征: 期望 expectation

## Definition (离散随机变量期望)

设离散随机变量$X$的概率分布列为$p_i$,且$\sum_{i=0}^{+\infty}|x_i|p_i < +\infty$,
称$EX = \sum_{i=0}^{+\infty} x_i p_i$为$X$的期望(值).

- 随机变量的平均特征：又称均值 (概率加权平均);
- 期望可能无穷，有限取值随机变量有有限期望;
- 掷骰子的期望值是$EX = (1+2+3+4+5+6)/6 = 3.5$

常见离散分布的期望;

- 两点分布(贝努利分布) $EX = p$
- 二项分布$B(N,p)$， $EX = np$
- 泊松分布$\mathcal{P}(\lambda)$,$EX = \lambda$
- 几何分布$EX = 1/p$;
- 超几何分布$H(n,M,N)$, $EX = nM/N$.

# 作业

北航教材: