Ch 4 连续型随机变量

回顾前一次课

- 0/1分布：X $\sim Ber(p)$，$E(X) = p$ $\quad \mathrm{Var}(X) = p(1-p)$
- 二项分布：$X \sim B(n,p)$，$E(X) = np$ $\quad \mathrm{Var}(X) = np(1-p)$
- 泊松分布：$X \sim P(\lambda)$, $E(X) = \lambda$ $\quad \mathrm{Var}\,(X) = \lambda$

  - 泊松定理：$\binom{n}{k} p^k (1-p)^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda}$

- 几何分布：$X \sim G(p)$，$\mathrm{E}(X) = \frac{1}{p}$ $\quad \mathrm{Var}(X) = \frac{1-p}{p^2}$

  - 无记忆性：$P(X > m + n \mid X > m) = P(X > n)$

应用案例：德国二战坦克问题、集卡问题

连续随机变量

离散型随机变量用分布列将随机变量的取值和概率全部罗列出来

一些随机现象的试验结果可能不能一一列举出来，如候车的等待时间、一个地区的降雨量、一盏电灯的寿命等

特别对于连续性随机变量，它在任意一个特定值的概率为0，此时用分布列来描述这一类型的随机变量则根本行不通

对于可能随机变量，我们可能更关心在某个区间内的概率，而不是它在某个特定点值的概率．例如，对于一盏电灯而言，更关心寿命介于1000~2000个小时之间的概率，而不是恰好1005个小时的概率．即关注于随机变量X在一个区间$[x_1, x_2]$上的概率

$$P(x_1 \leq X \leq x_2)$$

分布函数

给定任意随机变量X和实数x, 函数

$$F(x) = P(X \leq x)$$

称为随机变量X的**分布函数**, 分布函数的本质是概率

- 分布函数$F(x)$定义在$(-\infty, +\infty)$的普通函数，将概率与普通函数联系起来, 有利于利用数学分析的知识来研究随机变量
- 分布函数不限制随机变量的类型，无论时离散型随机变量还是非离散型随机变量, 都有各自的分布函数

对任意实数$x_1 < x_2$, 有

$$P(x_1 < X \leq x_2) = P(X \leq x_2) - P(X \leq x_1) = F(x_2) - F(x_1)$$

分布函数的性质

定理 分布函数$F(x)$具有如下性质:

- 单调性: 若$x_1 < x_2$, 则$F(x_1) \le F(x_2)$
- 规范性: $F(x) \in [0,1]$且$F(-\infty) = 0, F(+\infty) = 1$
- 右连续性: $F(x+0) = F(x)$

任何分布函数满足上述三性质

满足上述三性质的函数必是某随机变量的分布函数

分布函数可由上述三性质完全刻画

分布函数表示概率

有了分布函数$F(x)$, 就很容易计算随机变量X的概率, 如

$$P(X > a) = 1 - F(a)$$

$$P(X < a) = F(a-0) = \lim_{x \to a^-} F(x)$$

$$P(X = a) = F(a) - F(a-0)$$

$$P(X \geq a) = 1 - F(a-0)$$

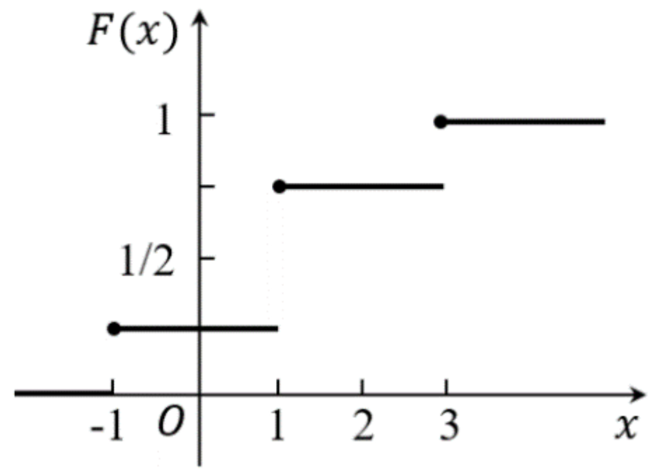
$$P(a \leq X \leq b) = F(b) - F(a-0)$$

对离散型随机变量X，设其分布列为$p_k = P(X = x_k)$ $(k = 1,2,\cdots)$
可得X的分布函数为

$$F(x) = P(X \leq x) = \sum_{k: x_k \leq x} p_k$$

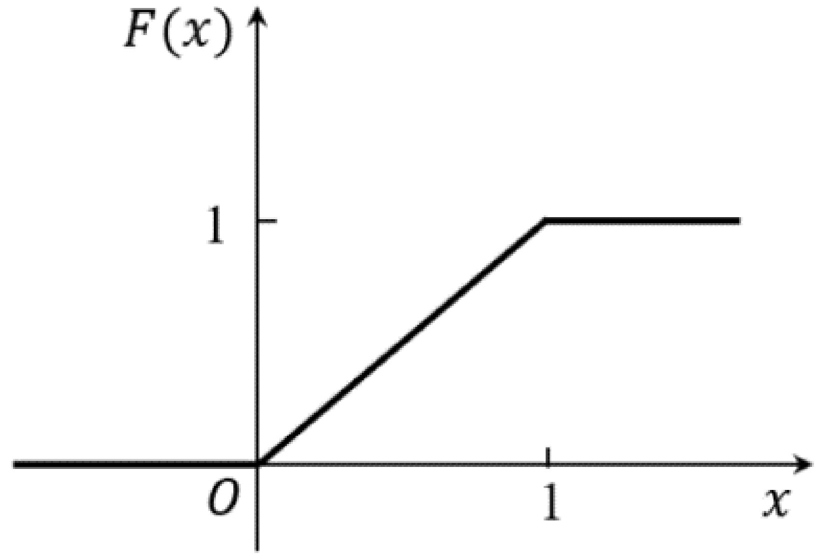
例题

随机变量X的分布列为$P(X=-1)=P(X=3)=1/4$和$P(X=2)=1/2$, 求X的分布函数



例题

在[0,1]区间随机抛一个点，用X表示落点的坐标，假设X落入[0,1]区间内任一子区间的概率与区间长度成正比, 求X的分布函数



例题

随机变量X的分布函数

$$F(x) = A + B \cdot \arctan(x) \qquad x \in (-\infty, +\infty)$$

求$P(X \leq 1)$

Ch 4.2 密度函数

概率密度函数

连续型随机变量: 随机变量的取值充满整个区间$[a,b]$或$(a,+\infty)$, 例如火车的到站时间、或一盏灯泡的寿命等

设随机变量X的分布函数为$F(x)$, 如果存在可积函数$f(x)$, 使得对任意实数x有

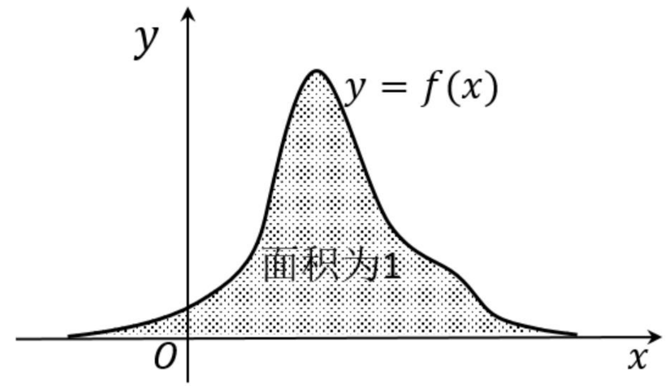
$$F(x)=\int_{-\infty}^{x} f(t)dt$$

成立, 则称X为**连续型随机变量**, 函数$f(x)$为随机变量X的**概率密度函数**, 简称**密度函数**

引理: 密度函数$f(x)$满足非负性$f(x)\geq 0$和规范性$\int_{-\infty}^{+\infty} f(t)dt=1$

密度函数的几何解释

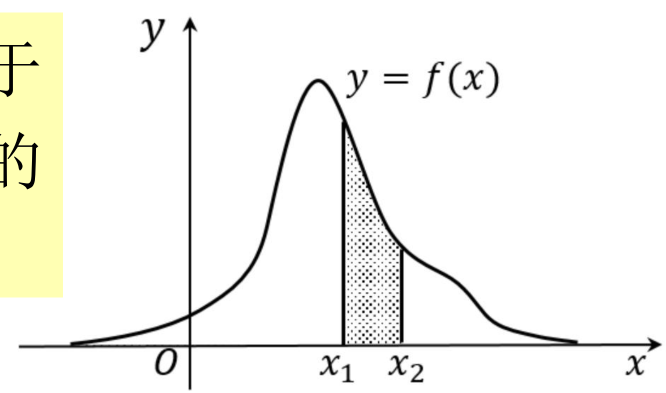
根据规范性可知曲线$y = f(x)$与x轴所围成的面积为1



对任意$x_1 \le x_2$, 有

$$P(x_1 < X \le x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(t)dt$$

几何解释: X落入区间$(x_1, x_2]$的概率等于x轴, $x = x_1$, $x = x_2$和$y = f(x)$所围成的曲边梯形的面积



定理

定理 对连续随机变量X, 其分布函数$F(x)$在整个实数域上连续; 若$f(x)$在x点连续, 则$F(x)$在x点可导, 且$F'(x) = f(x)$

定理 对连续型随机变量X和常数x, 有$P(X = x) = 0$

- 一个事件的概率为0, 不能推出该事件是不可能事件; 一个事件的概率为1, 也不能推出该事件是必然事件
- $P(a \le X \le b) = P(a < X < b) = P(a \le X < b) = P(a < X \le b)$
- 概率密度函数不是概率：$P(X = x) = 0 \ne f(x)$

概率与密度函数关系

若$f(x)$在点x连续, 根据连续性有

$$\lim_{\Delta x\to 0}\frac{P(x-\Delta x\le X\le x+\Delta x)}{\Delta x}=\lim_{\Delta x\to 0}\frac{\int_{x-\Delta x}^{x+\Delta x}f(t)dt}{\Delta x}$$

$$=\lim_{\Delta x\to 0}\frac{2\Delta x f(\xi)}{\Delta x}=2f(x)$$

其中$\xi\in(x-\Delta x,x+\Delta x)$, 由此可得

$$P(x-\Delta x\le X\le x+\Delta x)\approx 2f(x)\Delta x$$

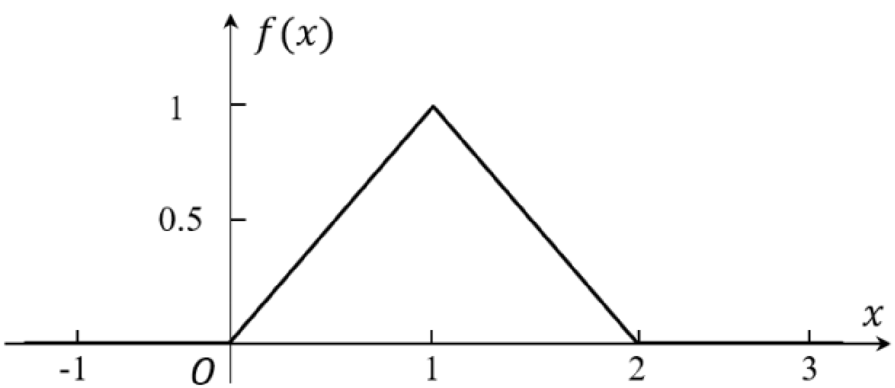
概率密度$f(x)$越大, 则X在x附近取值的概率越大

例题

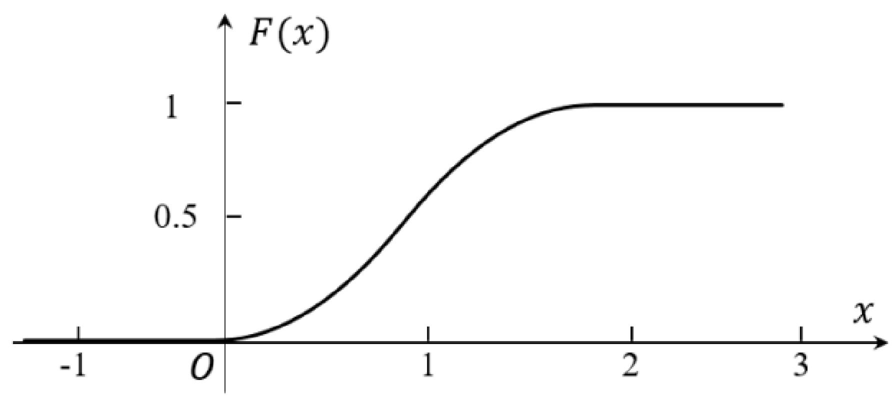
设连续随机变量X的密度函数

$$f(x)=\begin{cases} x & 0<x\le 1 \\ a-x & 1<x\le 2 \\ 0 & \text{其它} \end{cases}$$

求分布函数F(x)



(a) 概率密度函数



(b) 分布函数

例题

对连续随机变量X, 当$x \in (0,3)$时密度函数$f(x) = cx^2$, 在其它点的密度函数$f(x) = 0$. 设随机变量

$$Y = \begin{cases} 2, & X \leq 1 \\ X, & X \in (1,2) \\ 1, & X \geq 2 \end{cases}$$

求概率$P(Y \geq X)$

案例分析：集卡活动

小朋友喜欢参加各种集卡活动, 如奥特曼卡和叶罗丽卡等. 事实上很多成年人也对集卡游戏并不陌生，例如80年代的葫芦娃洋画、或90年代的小虎队旋风卡等

问题：市场上有n种不同类型的卡片，假设一个小朋友每次都能以等可能概率、独立地收集一张卡片，问一个小朋友在平均情况下至少要收集多少张卡才能收集齐n种不同类型的卡片

引理: 对任意的随机变量$X_1, X_2, \cdots, X_n$有

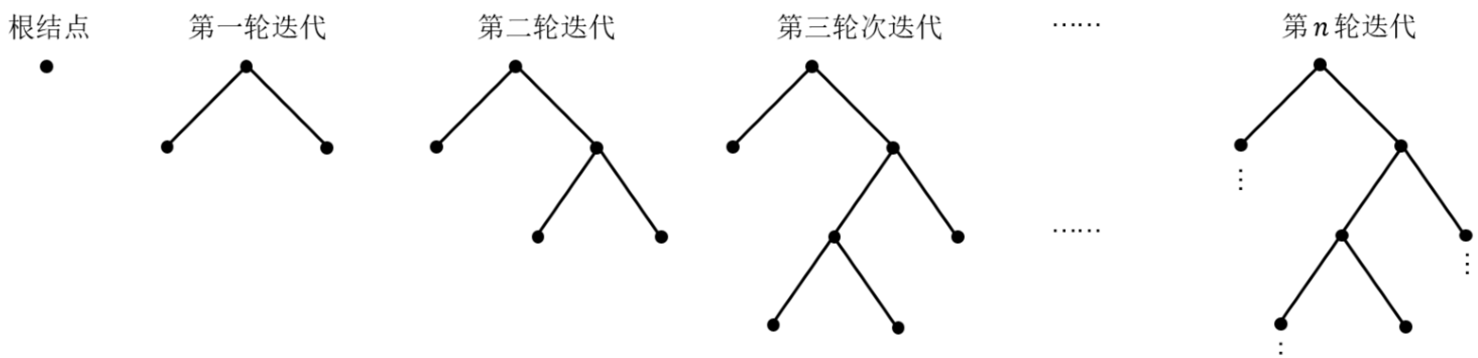
$$\boldsymbol{E(X_1+X_2+\cdots+X_n)=E(X_1)+E(X_2)+\cdots+E(X_n)}$$

案例分析：随机二叉树叶子结点的高度

随机二叉树的构造: 首先给定二叉树的根结点, 然后在每一轮的迭代过程中执行以下两步操作:

- 在当前所有的叶子结点中随机选择一个叶子结点;
- 被选中的叶子结点生长出左、右两个叶子结点，此时被选中的叶子结点成为一个内部结点;

重复上述过程n步, 则该随机树最后得到n个叶子结点.



一个叶子结点的高度是从根节点到该叶子结点的路径中边的条数

问题: 在最后生成的随机树中, 任意一个叶子结点的平均高度