Chpt.2 Random Variables & Probability Distributions

第二章 随机变量及其分布

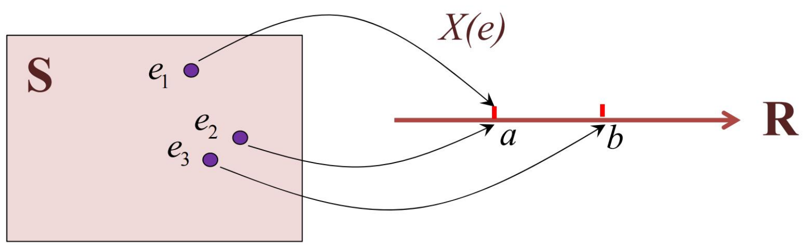
上节回顾

- 概率空间

  $\{\Omega, \mathcal{F}, P\}$

- 随机变量

  $X: \Omega \to \mathbb{R}$



Ω => X

A =>$\{X \in L\}$

P(A) =>P($X \in L$)

- 离散型随机变量及分布律
- 常见分布
  - 单点分布
  - 0-1分布
  - 二项分布 $b(k,n,p) = b(n-k,n,1-p)$

    最有可能成功的次数为(n＋1)p

思考题：　食堂开设多少窗口合理

某学校有500人，拟建一个食堂，开设m个窗口。窗口数m太小，则经常派长队；窗口数m太大，则不经济。

假定在每一个指定时刻，500个人中每人是否去食堂是独立的，每人去食堂的概率都是0.1。

问题：”　在营业中任意时刻，保证每个窗口的排队人数（包括正在打菜的那个人）不超过10”这个事件的概率不小于0.9, 则至少需开设多少个窗口？

Example 2.8 邮局开设多少服务窗口合理

假设X代表在食堂的人数，则**X~B(500,0.1)**。“在营业中任意时刻每个窗口的排队人数（包括正在被服务的那个人）不超过10” 这件事相当于m个窗口总的人数X：{ X≤10m } 这个事件。其发生的概率为

$$P\{X \le 10m\} = \sum_{k=0}^{10m} P\{X = k\} = \sum_{k=0}^{10m} C_{500}^{k}(0.1)^{k}(0.9)^{500-k}$$

题目要求此概率大于0.9，即

$$\sum_{k=0}^{10m} C_{500}^{k}(0.1)^{k}(0.9)^{500-k} \ge 0.9$$

找一个自然数m，使得上述不等式成立，此m即为问题的答案。

[3] $n\to\infty$ 时的渐近性质

假定p与n有关，记作p_n。 考虑$n\to\infty$的情况，有下面的定理：

[泊松(Poisson)定理] 如果存在正常数λ，当 $n\to\infty$ 时有 $np_n\to\lambda$，则

$$\lim_{n\to\infty} b(k,n,p)=\frac{\lambda^k}{k!}e^{-\lambda} \qquad k=0,1,2,\cdots$$

[3] $n \to \infty$ 时的渐近性质

Proof: 记$\lambda_n = np_n$

$$b(k,n,p) = C_n^k (p_n)^k (1-p_n)^{n-k}$$

$$= \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!}\left(\frac{\lambda_n}{n}\right)^k \left(1-\frac{\lambda_n}{n}\right)^{(n-k)}$$

$$= \frac{\lambda_n^k}{k!} \frac{n(n-1)(n-2)\cdots(n-k+1)}{n^k} \left(1-\frac{\lambda_n}{n}\right)^n \Big/ \left(1-\frac{\lambda_n}{n}\right)^k$$

$$\downarrow \qquad \downarrow \qquad \downarrow \qquad \downarrow$$

$$\boxed{\frac{\lambda^k}{k!}} \qquad \boxed{1} \qquad \boxed{e^{-\lambda}} \qquad \boxed{1}$$

$$\to \frac{\lambda^k}{k!} e^{-\lambda} \qquad (n \to \infty)$$

[3] n →∞ 时的渐近性质

$$\lim_{n\to\infty} b(k,n,p)=\frac{\lambda^k}{k!}e^{-\lambda} \qquad k=0,1,2,\cdots$$

定理的条件 $np_n \to \lambda$ 意味着：当 n很大时，p_n 必定很小.

因此，利用泊松定理，对于二项分布，当 n 很大，p 很小时有以下近似式：

$$C_n^k p^k (1-p)^{n-k} \approx \frac{\lambda^k e^{-\lambda}}{k!}$$

其中 $\boldsymbol{\lambda = np}$

实际计算中，$n \geq 20$, $p \leq 0.05$ 时近似效果就很好.

请同学们利用编程，绘图观察两者之间差距的变化（绘图是以n作为变量，而不是b(k,n,p)中的k）

Example 2.9 为保证设备正常工作，要配备适量的维修人员 .

设共有300台设备，每台的工作相互独立，发生故障的概率都是0.01. 在通常情况下，一台设备的故障可由一人来处理 .

问：至少应配备多少维修人员，才能保证当设备发生故障时不能及时维修的概率小于0.01?

分析：

设X为300台设备同时发生故障的台数，300台设备独立工作，每台出故障概率p=0.01 .

可见， **X~B(n,p)，n＝300, p＝0.01**

设需配备N个维修人员，所求问题就是满足$P(X>N) < 0.01$ 或 $P(X\le N)\ge 0.99$ 的最小的N.

$$P\{X>N\}=\sum_{k=N+1}^{300} C_{300}^{k}(0.01)^{k}(0.99)^{300-k}$$

$$\approx \sum_{k=N+1}^{300} \frac{3^{k}e^{-3}}{k!}$$

$$\approx \sum_{k=N+1}^{\infty} \frac{3^{k}e^{-3}}{k!}$$

$$\sum_{k=x}^{\infty} \frac{\lambda^{k}}{k!}e^{-\lambda}$$

x	$\lambda=2.0$	$\lambda=2.5$	$\lambda=3.0$	$\lambda=3.5$	$\lambda=4.0$	$\lambda=4.5$	$\lambda=5.0$
0	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
1	0.864665	0.917915	0.950213	0.969803	0.981684	0.988891	0.993262
2	0.593994	0.712703	0.800852	0.864112	0.908422	0.938901	0.959572
3	0.323324	0.456187	0.576810	0.679153	0.761897	0.826422	0.875348
4	0.142877	0.242424	0.352768	0.463367	0.566530	0.657704	0.734974
5	0.052653	0.108822	0.184737	0.274555	0.371163	0.467896	0.559507
6	0.016564	0.042021	0.083918	0.142386	0.214870	0.297070	0.384039
7	0.004534	0.014187	0.033509	0.065288	0.110674	0.168949	0.237817
8	0.001097	0.004247	0.011905	0.026739	0.051134	0.086586	0.133372
9	0.000237	0.001140	0.003803	0.009874	0.021363	0.040257	0.068094
10	0.000046	0.000277	0.001102	0.003315	0.008132	0.017093	0.031828
11	0.000008	0.000062	0.000292	0.001019	0.002840	0.006669	0.013695
12	0.000001	0.000013	0.000071	0.000289	0.000915	0.002404	0.005453
13	0.000000	0.000002	0.000016	0.000076	0.000274	0.000805	0.002019
14	0.000000	0.000000	0.000003	0.000019	0.000076	0.000252	0.000698
15	0.000000	0.000000	0.000001	0.000004	0.000020	0.000074	0.000226
16	0.000000	0.000000	0.000000	0.000001	0.000005	0.000020	0.000069
17	0.000000	0.000000	0.000000	0.000000	0.000001	0.000005	0.000020
18	0.000000	0.000000	0.000000	0.000000	0.000000	0.000001	0.000005

我们求满足 $\sum_{k=N+1}^{\infty} \frac{3^{k}e^{-3}}{k!} < 0.01$

查泊松分布表得

$$\sum_{k=9}^{\infty} \frac{e^{-3}3^{k}}{k!} \approx 0.0038, \qquad \sum_{k=8}^{\infty} \frac{e^{-3}3^{k}}{k!} \approx 0.012$$

$K=N+1\geq 9$即可，即$N\geq 8$.　即至少需配备8个维修人员.

（四）泊松分布(Poisson)

从上面的泊松定理可引入另一类重要的分布。

设随机变量X可取一切非负整数值，取这些值的概率为：

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda} \qquad k=0,1,2,\cdots\cdots$$

其中λ>0是它的参数，称X服从参数为λ的泊松分布，简记 $X\sim\pi(\lambda)$

泊松分布适合于描述单位时间（空间）内随机事件发生次数的概率。如某一服务设施在一定时间内受到的服务请求的次数，汽车站台的候车人数、机器出现的故障数、自然灾害发生的次数、DNA序列的变异数、放射性原子核的衰变数、激光的光子数分布等等。

Q：泊松分布为什么可以表示单位时间内事件发生次数的概率？

假设单位时间内某一事件发生的平均数目为λ

为方便记，设所观察的这段时间为[0,1),取一个很大的自然数n，把时间段[0,1)分为等长的n段：

$$l_1=[0,\frac{1}{n}],l_2=[\frac{1}{n},\frac{2}{n}],\ldots,l_i=[\frac{i-1}{n},\frac{i}{n}],\ldots,l_n=[\frac{n-1}{n},1]$$

我们做如下两个假定：

1. 在每段 l_i 内，恰发生一个事故的概率，近似的与这段时间的长 $\frac{1}{n}$ 成正比，可设为 $\frac{\lambda}{n}$ 。当n很大时，$\frac{1}{n}$ 很小时，在 l_i 这么短暂的一段时间内，要发生两次或者更多次事故是不可能的。因此在 l_i 这段时间内不发生事故的概率为 $1-\frac{\lambda}{n}$ 。

2. $l_i,\ldots,l_n$ 各段是否发生事故是独立的

把在[0,1)时段内发生的事故数X视作在n个划分之后的小时段 $l_i,\ldots,l_n$ 内有事故的时段数，则按照上述两个假定，X应服从二项分布 $B(n,\frac{\lambda}{n})$ 。于是，我们有

$$P(X=i)=\binom{n}{i}(\frac{\lambda}{n})^i(1-\frac{\lambda}{n})^{n-i}$$

注意到当 $n\to\infty$ 取极限时，我们有

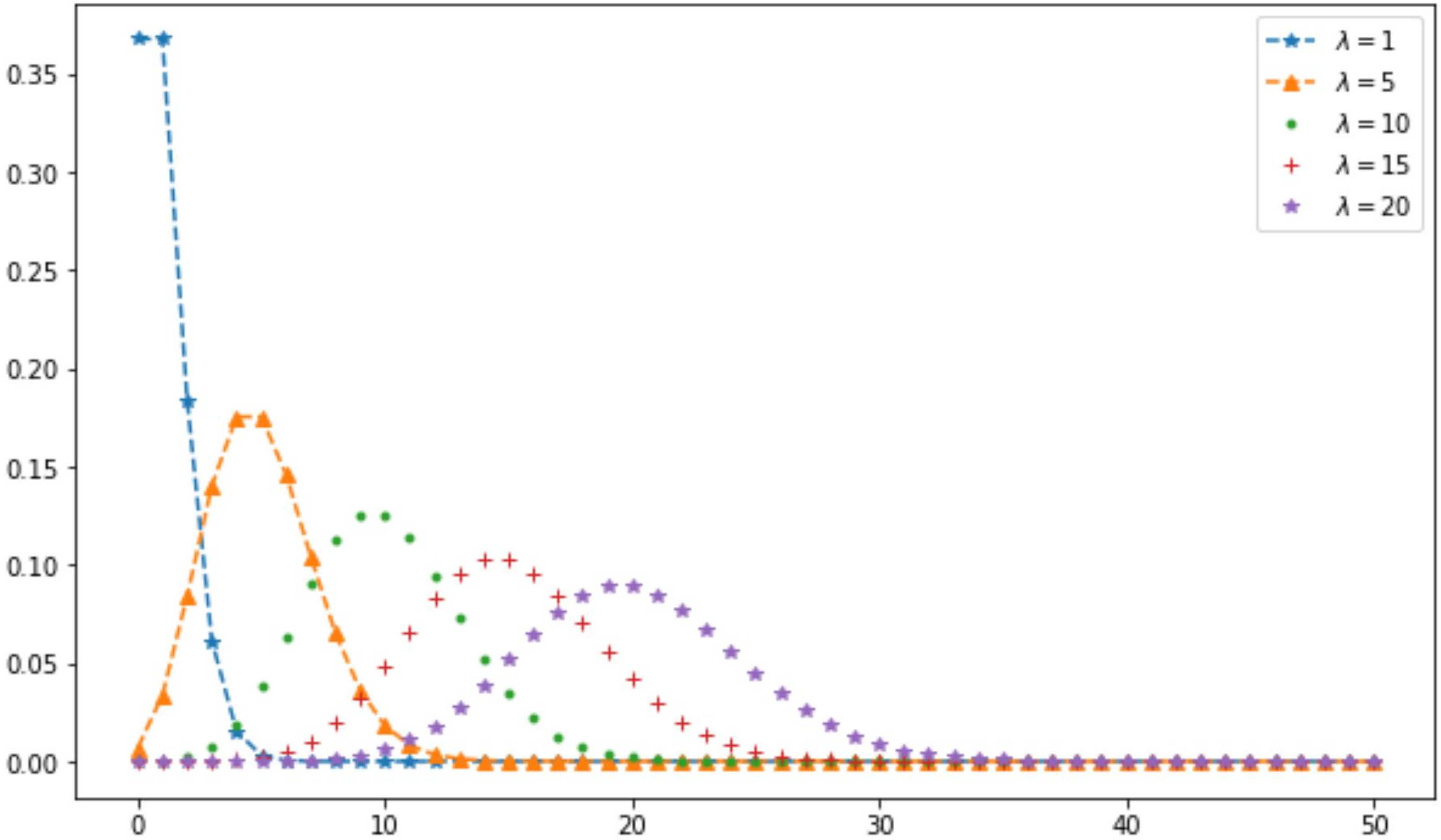
$$\frac{\binom{n}{i}}{n^i}\to\frac{1}{i!},(1-\frac{\lambda}{n})^n\to e^{-\lambda}$$

二项分布的极限就是泊松分布；
泊松分布的连续形式就是正态分布

因此

$$P(X=i)=\binom{n}{i}(\frac{\lambda}{n})^i(1-\frac{\lambda}{n})^{n-i}$$
$$=\frac{e^{-\lambda}\lambda^i}{i!}$$

泊松分布 vs 正态分布



0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
10
20
30
40
50
$\lambda = 1$
$\lambda = 5$
$\lambda = 10$
$\lambda = 15$
$\lambda = 20$

（五）几何分布

单次试验中事件A 发生概率为p，现重复多次，直到A出现为止。定义X为事件A首次发生时所进行的试验次数，则X为一随机变量，可能的取值为

$$P(X=k)=(1-p)^{k-1}p \qquad (k=1,2,\cdots,)$$

称X服从参数为 p 的几何分布

2.3　随机变量的分布函数

2.3.1 WHY and HOW

WHY：分布列不能表示取值不可数的随机变量的分布，比如机器的寿命,无法一一罗列这些值及其概率，有必要引入新的概率分布表示法。

HOW：

- 事实上，我们经常关心的是一个随机变量落入一个区间的概率，如机器寿命大于 T 的概率，产品中次品数少于 k 件的概率。
- 注意到随机变量取值为实数，在实数轴上的任意的区间都可以用 $(-\infty, x]$ 来表示，进一步的由于任意波雷尔集B*是左开右闭区间的（有限或可列）并、（有限或可列）交、逆产生的集合，故此可以定义在区间$(-\infty, x]$上概率*。

2.3 随机变量的分布函数

2.3.2 定义

设随机变量X，对任意的实数x，令随机变量X取值落入区间 $(-\infty, x]$ 内的概率为F(x)

$$F(x) = P(X \le x) \quad -\infty < x < +\infty$$

称为随机变量X的分布函数 (distribution function)。

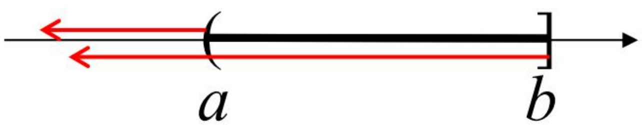
2.3 随机变量的分布函数

分布函数的用途

可以表示随机变量落入任何一个区间的概率大小

X的分布函数为$F(x)$，a,b为两个实数，$a<b$，则

$$P(a<X\le b)=P(X\le b)-P(X\le a)$$
$$=F(b)-F(a);$$



$$P(a<X<b)=P(a<X\le b-0)=F(b-0)-F(a);$$

$$P(X=b)=P(X\le b)-P(X<b)=F(b)-F(b-0).$$

注：这里的0表示任意小的一个数，或者说是一个无穷小量。

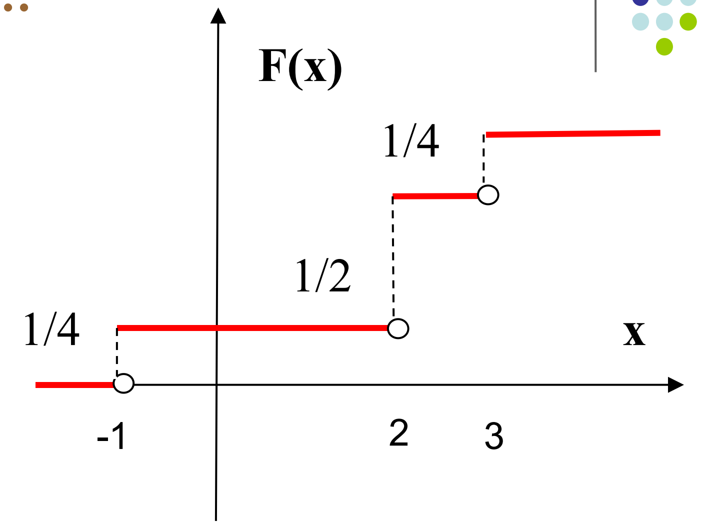
例：随机变量X的分布律为

X	-1	2	3
p	1/4	1/2	1/4

求X的分布函数F(x)。

例：随机变量X的分布律为

X	-1	2	3
p	1/4	1/2	1/4



求X的分布函数F(x)

$$F(x)=\begin{cases}0 & x<-1\\ \dfrac{1}{4} & -1\le x<2\\ \dfrac{3}{4} & 2\le x<3\\ 1 & x\ge 3\end{cases}$$

$$P\left(X\le\frac{1}{2}\right)=F\left(\frac{1}{2}\right)=\frac{1}{4}$$

$$P\left(\frac{3}{2}<X\le\frac{5}{2}\right)=F\left(\frac{5}{2}\right)-F\left(\frac{3}{2}\right)=\frac{1}{2}$$

$$P(2\le X\le 3)=F(3)-F(2)+P(X=2)=\frac{3}{4}$$

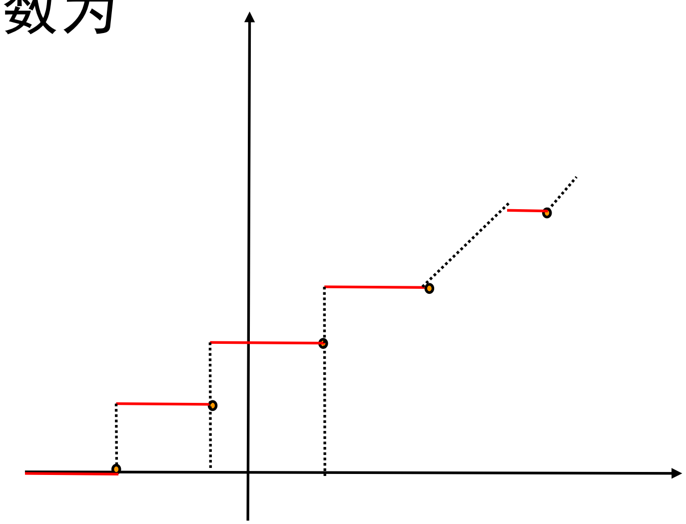
2.3 随机变量的分布函数

推广 离散随机变量的分布函数，随机变量X的分布列为

$$\begin{pmatrix} x_1 & x_2 & \cdots & x_k & \cdots \\ p(x_1) & p(x_2) & \cdots & p(x_k) & \cdots \end{pmatrix}$$

$x_1 < x_2 < \cdots\cdots < x_k < \cdots\cdots$ ，则X的分布函数为

$$F(x) = \begin{cases} 0, & x < x_1, \\ p(x_1), & x_1 \le x < x_2, \\ \cdots & \cdots \\ \sum\limits_{i \le k} p(x_i), & x_k \le x < x_{k+1}, \\ \cdots & \cdots \end{cases}$$



离散型随机变量的分布函数F(x)是分段函数，在 x_k(k=1，2， …)各有一跳跃，跃度为 $p(x_k)$ ，在每一段 $[x_k, x_{k+1})$ 中都是常数，呈阶梯形。

2.3.3 分布函数的特性

分布函数 $F(x)=P(X\le x)$ －∞<x<+∞

已经是一个普通函数，就可以用数学分析、实变函数来研究。

概率事件 ⇒ 随机变量(统一描述&数字化) ⇒ 分布函数(函数工具)

(1) $F(x)$是一个不减的函数， $\forall x_1\le x_2, F(x_1)\le F(x_2)$

(2) $0\le F(x)\le 1$

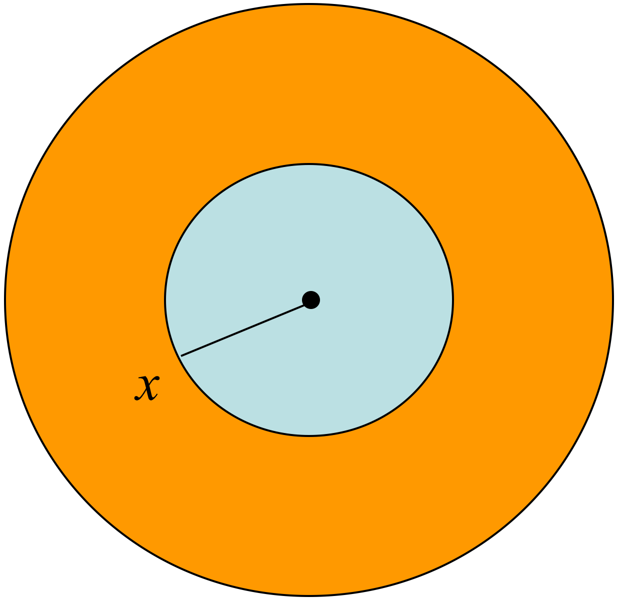
(3) $F(-\infty)=\lim_{x\to-\infty}F(x)=0$ $F(\infty)=\lim_{x\to\infty}F(x)=1$

Example （pp.50 例2）

一个半径为2米的圆盘形靶子，设击中的点落在靶上任何一同心圆内的概率与该同心圆的面积成正比，并设射击都能中靶，以X表示弹着点与圆心的距离。

求：随机变量X的分布函数。

解：如图，X是随机变量，x是每次讨论时定下的阈值，是一个常量。



[1] 若$x<0$，则$\{X\le x\}$是不可能事件

$$F(x)=P\{X\le x\}=0$$

[2] 若$0\le x\le 2$, 由假设知道$\{0\le X\le x\}$的概率正比于圆的面积πx^2 $\longrightarrow$ $P\{0\le X\le x\}=\frac{\pi x^2}{4\pi}=\frac{x^2}{4}$

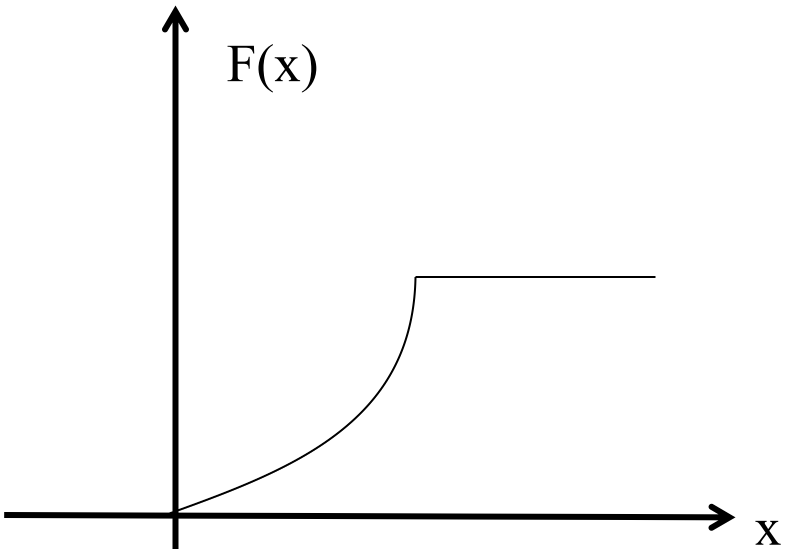
$$F(x)=P\{X\le x\}=P(X<0)+P\{0\le X\le x\}=x^2/4$$

[3] 若$x\ge 2$，由题意$\{X \le x\}$是必然事件

$$F(x) = P\{X \le x\} = 1$$

综合以上知道：

$$F(x)=\begin{cases}0 & x<0\\ \dfrac{x^2}{4} & 0\le x<2\\ 1 & x\ge 2\end{cases}$$



同时分布函数可以写为积分形式

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

$$f(t)=\begin{cases}\dfrac{t}{2} & 0<t<2\\ 0 & \text{其他}\end{cases}$$

此分布函数是个连续函数，因此X不是离散型随机变量，而是连续型随机变量。

2.4 连续型随机变量的分布

2.4.1 连续型随机变量及密度函数

猜测连续型随机变量的分布函数（依据）：

"对于离散随机变量，分布函数为落在 $(-\infty, x]$ 样本点的概率之和"

离散型 $F(x) = P(X \le x) = \sum_{i:x_i \le x} P(X = x_i)$

↓

连续型 $F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt$

Definition 1 如果对于随机变量 X 的分布函数$F(x)$，存在非负函数 $f(x)$，使对于任意实数 x 满足

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

则称 X 为连续型(continuous)随机变量，称 $f(x)$ 为 X 的概率密度函数，简称为**密度函数(density function)**。

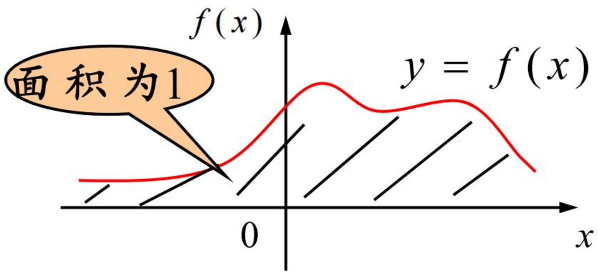
2.4.2 连续型随机变量的特性

连续型随机变量的分布函数$F(x)$ 具有下列数学性质：

(1) $F(x)$ 必定处处连续；

在$f(x)$ 的连续点，$F(x)$可导，且 $F'(x)=f(x)$.

(2) $\int_{-\infty}^{+\infty} f(x)dx = 1.$



(3) 随机变量X落在区间$(a，b]$的概率:

$$P\{a<x\le b\}=F(b)-F(a)$$

$$=\int_a^b f(x)dx.$$

对于连续型的随机变量X,有

$P(X\in D)=\int_D f(x)dx$, 任意$D\subset R$.

(4) 特别，对任一常数c，$P\{X=c\}=0.$

2.4 连续型随机变量的分布

Remark 1：对连续型随机变量，计算在一点的概率是没有意义的，这也是不能用分布列描写连续型随机变量的理由之一。

Remark 2：对连续型随机变量而言，$\{X=C\}$概率为0，但是一个可能发生的事件（如在一个区间上方抛点，点可以落到任意一点 x，但是落到任意一点的概率为0）；

因此，$P(A)=0$并不表明$A=\varnothing$

但$A=\varnothing$必然有$P(A)=0$

同样，$P(A)=1$也并不表明$A=S$（样本空间）

但$A=S$必然有$P(A)=1$

主观题 10分

设置

例1：设X的概率密度为 $f(x)=\begin{cases} cx+1/6, & 0<x<2; \\ 0, & 其他. \end{cases}$

求：(1)常数c的值；(2) X的概率分布函数$F(x)$；(3) $P(-1<X<1)$的值.

作答

思考题

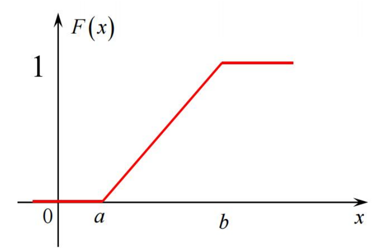
- 除了离散型和连续型，还有其他类型的随机变量吗？

2.4.3 常见的连续型随机变量

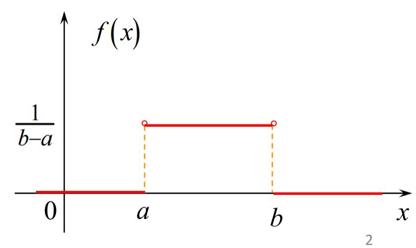
[一] 均匀(Uniform)分布

考虑在一个区间$[a,b]$上方抛点，假设点落入$[a,b]$中任意一个等长度区间内的概率（可能性）是相同的（或者说点落入某个区间的可能性大小只与区间长度有关，与区间端点无关），用随机变量X表示抛点得到的值，则X的分布函数F(x)为

$$F(x)=P\{X\le x\}=\begin{cases}0, & x\le a\\ \dfrac{x-a}{b-a}, & a<x<b\\ 1, & x\ge b\end{cases}$$



$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a<x<b,\\ 0, & else\end{cases}$$



称具有上述概率密度$f(x)$的连续型随机变量X为在区间$[a, b]$上服从**均匀分布**，简记作$X\sim U(a, b)$。

2.4.3 常见的连续型随机变量

[一] 均匀(Uniform)分布

Example 考察一个数据，它在小数点n位后四舍五入，则其真值x与其近似值$\hat{x}$之间的误差$\varepsilon=\hat{x}-x$，一般假定服从$\left[-0.5\times10^{-n},0.5\times10^{-n}\right]$上的均匀分布。

利用均匀分布理论可以对大量运算后的数据进行误差分析。这对于使用计算机解题时是很重要的，因为计算机的字长总是有限的。

Example 列车从车站每40分钟开出一次，乘客到站时刻是机会均等的，问等候超过10分钟的概率是多少？

以T表示等候时间，它是一个随机变量。T取值范围为 [0, 40]，其服从均匀分布 U[0, 40]。

$$P\{T\geq10\}=\int_{10}^{40}\frac{1}{40}dt=\frac{3}{4}$$

[二] 指数(Exponential)分布

若X的概率密度函数为 $f(x)=\begin{cases}\lambda e^{-\lambda x}, & x>0;\\ 0, & x\le 0,\end{cases}$

其中$\lambda>0$,就称X服从参数为λ的**指数分布**(*Exponential*),

记为$X\sim E(\lambda)$ 或 $X\sim Exp(\lambda)$.

分布函数为 $F(x)=\begin{cases}1-e^{-\lambda x}, & x>0;\\ 0, & x\le 0.\end{cases}$

指数分布常用来描述产品的寿命。

[二] 指数分布

性质：指数分布具有无记忆性

$$F(x)=\begin{cases}1-e^{-\lambda x}, & x>0;\\ 0, & x\le 0.\end{cases}$$

对于 $t_0>0, t>0,$

$$P(X>t_0+t\mid X>t_0)=\frac{P(X>t_0+t, X>t_0)}{P(X>t_0)}$$

$$=\frac{P(X>t_0+t)}{P(X>t_0)}=\frac{1-F(t_0+t)}{1-F(t_0)}$$

$$=\frac{e^{-\lambda(t_0+t)}}{e^{-\lambda t_0}}=e^{-\lambda t}=P(X>t).$$

指数分布具有“**无记忆性**”：使用s小时后元件的寿命不低于t的概率与初始时使用寿命为t的概率相等。

指数分布的用途:

- 指数分布可以用来表示独立随机事件发生的时间间隔，比如旅客进机场的时间间隔、中文维基百科新条目出现的时间间隔等等;
- 在排队论中，一个顾客接受服务的时间长短也可用指数分布来近似;
- 无记忆性的现象(连续时).

主观题 10分

设置

设某人电话通话时间X(分钟)服从指数分布，概率密度为

$$f(x)=\begin{cases}\frac{1}{15}e^{-\frac{x}{15}}, & x>0;\\ 0, & x\le 0.\end{cases}$$

求：(1) 她的通话时间在10～20分钟之间的概率;

(2) 若她已打了10分钟，求她继续通话超过15分钟的概率.

(即，若她已打了10分钟，求她总共通话超过25分钟的概率.)

作答