代入期望 $E(X)$ 可得

$$E(X)=k\binom{n}{k}^{-1}\sum_{i=k}^{n}\binom{i-1}{k-1}\frac{i}{k}=k\binom{n+1}{k+1}\Big/\binom{n}{k}=\frac{k(n+1)}{k+1}\ .$$

由于仅做了一次观察, 将观察中 k 个数的最大值近似期望 $E[X]$, 即 $E(X)\approx\max(x_1,x_2,\cdots,x_n)$, 由此估计

$$n\approx\max(x_1,x_2,\cdots,x_n)\left(1+\frac{1}{k}\right)-1,$$

从而完成 n 的估计.

例如, 如果观察到被击毁坦克编号分别为 $17,68,94,127,135,212$, 根据上面的推到可估计出

$$n\approx 212\times(1+1/6)-1=246.$$

针对德国坦克数量的实际估计情况见下表, 可以发现利用上述所提的统计估计方法接近德国的实际产量, 比英国的情报估计准确得多.

时间	统计估计	英国情报估计	德国实际产量
1940-06	169	1000	122
1941-06	244	1550	271
1942-08	327	1550	342

3.5.2 集卡活动

很多小朋友喜欢各种集卡活动, 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. 问题可以描述为: 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.

这里先补充一个需要用到的引理, 后面将给出详细的证明:

引理 3.5 对任意的随机变量 $X_1,X_2,\cdots,X_n$ 有

$$E(X_1+X_2+\cdots+X_n)=E(X_1)+E(X_2)+\cdots+E(X_n)\ .$$

用 X 表示收集齐 n 种不同类型的卡片所需要的收集次数, 用 X_k 表示收集齐第 $k-1$ 种和第 k 种不同类型卡片之间所需要的收集次数 ($k\in[n]$), 于是有 $X=X_1+X_2+\cdots+X_n$. 我们的问题是计算期望 $E(X)$.

很容易发现随机变量 X_k 服从参数为 p_k 的几何分布. 当已经收集到 $k-1$ 种不同类型的卡片时, 再获得一张新卡的概率

$$p_k=1-(k-1)/n\ .$$

根据几何分布的性质有 $E[X_k] = 1/p_k = n/(n-k+1)$. 利用引理 3.5 有

$$E(X) = E\left(\sum_{k=1} X_k\right) = \sum_{k=1} E(X_k) = \sum_{k=1}^{n} \frac{n}{n-k+1} = n\sum_{k=1}^{n} \frac{1}{k} = nH(n) \ ,$$

这里 $H(n)$ 表示参数为 n 的调和数, 即 $H(n) = \sum_{k=1}^{n} 1/k$. 关于调和数有

引理 3.6 调和数 $H(n) \in [\ln(n+1), 1+\ln(n)]$.

证明 因为函数 $1/x$ 在 $x \in (0, +\infty)$ 单调递减, 有

$$\ln(n+1) = \int_{x=1}^{n+1} \frac{1}{x} dx \leqslant \sum_{k=1}^{n} \frac{1}{k} = 1 + \sum_{k=2}^{n} \frac{1}{k} \leqslant 1 + \int_{x=1}^{n} \frac{1}{x} dx = 1 + \ln(n).$$

最后得到 $n\ln(n+1) \leqslant E(X) \leqslant n + n\ln n$.

3.5.3 随机二叉树叶子结点的高度

在机器学习中, 随机树和随机森林是一类经典的分类或回归算法, 随机树叶子结点的高度估计对学习算法性能的分析具有重要作用. 本节考虑完全随机的二叉树中一个叶子结点的平均高度. 随机二叉树的构造过程非常简单: 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作为划分结点;
- 被选中的叶子结点变成一个内部结点, 生长出左、右两个叶子结点.

重复上述过程 n 步, 最后得到具有 n 个叶子结点的随机二叉树. 在这一构造过程中, 最关键的一步是随机选择的叶子结点作为划分结点. 随机二叉树构造的示意图如下所示:

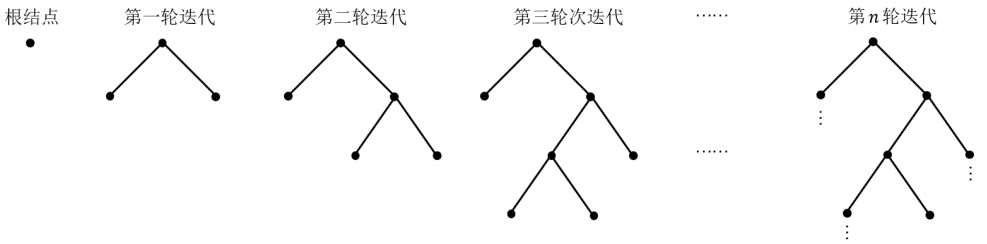


图 3.3 随机二叉树构造的示意图

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. 求解的问题: 在最后生成的随机二叉树中, 求任意一个叶结点的平均高度.

用随机变量 X 表示任意给定的一个叶结点的高度, 并用随机变量 X_i 表示在第 i 轮迭代过程中该叶子的祖先结点是否恰好被选中作为划分结点, 而在第 i 轮迭代过程中恰好有 i 个叶结点, 则有

$$X_i = \mathrm{Ber}(1/i) \qquad 且 \qquad X = X_1 + X_2 + \cdots + X_n.$$

根据期望的性质和引理 3.6 有

$$E[X] = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} 1/i = H(n) \in [\ln(n+1), 1+\ln(n)] \ .$$

由此可知一个叶子结点的平均高度为 $\Theta(\ln n)$.

第 4 章　连续型随机变量

4.1 分布函数

离散型随机变量利用概率分布列将随机变量的取值和对应的概率全部罗列出来. 然而一些随机现象的试验结果可能不止可列个取值, 此时不能一一列举出来, 例如候车的等待时间、一个地区的降雨量、一盏电灯的寿命等. 特别地, 对于连续性随机变量, 它在任意一个特定值的概率为 0 (将在 4.2 节介绍), 此时用分布列来描述这一类型的随机变量就根本行不通.

对于一些非离散型随机变量, 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, 我们关心其寿命大于 1000 个小时的概率, 而不是恰好 1005 个小时的概率. 针对这些随机现象, 我们关注于随机变量 X 在一个区间 $[x_1, x_2]$ 上的概率 $P(x_1 \leqslant X \leqslant x_2)$. 为此引入分布函数的概念:

定义 4.1　给定随机变量 X, 对任意实数 $x \in (-\infty, +\infty)$, 函数

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

称为随机变量 X 的 **分布函数** (cumulative distribution function).

分布函数 $F(x)$ 是定义在 $(-\infty, +\infty)$ 的普通函数, 将普通函数与随机事件的概率关联起来, 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.

分布函数的本质是概率, 考虑随机事件 $\{X \in (-\infty, x]\}$ 的概率. 对任意实数 $x_1 < x_2$ 有

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

若已知随机变量 X 的分布函数 $F(x)$, 则可以知道 X 落入任意区间 $(x_1, x_2]$ 上的概率, 因此分布函数完整地了刻画随机变量的统计规律性. 分布函数具有良好的分析性质:

定理 4.1　分布函数 $F(x)$ 具有如下性质:

- 单调性: 若 $x_1 < x_2$, 则 $F(x_1) \leqslant F(x_2)$;
- 规范性: $F(x) \in [0, 1]$, 且 $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$, $F(+\infty) = \lim_{x \to +\infty} F(x) = 1$;
- 右连续性: $F(x + 0) = \lim_{\Delta x \to 0^+} F(x + \Delta x) = F(x)$.

证明　根据概率的非负性, 对任意 $x_1 < x_2$ 有

$$F(x_2) - F(x_1) = P(x_1 < X \leqslant x_2) \geqslant 0 \ .$$

根据规范性有

$$\begin{aligned} 1 = P(-\infty < X < +\infty) &= \sum_{n=-\infty}^{+\infty} P(n < X \leqslant n+1) = \sum_{n=-\infty}^{+\infty} F(n+1) - F(n) \\ &= \lim_{n\to-\infty} F(n) - \lim_{m\to+\infty} F(m) \ . \end{aligned}$$

根据 $F(x)$ 的单调性有 $F(-\infty) = \lim_{x\to-\infty} F(x) = \lim_{n\to-\infty} F(n)$ 和 $F(+\infty) = \lim_{x\to+\infty} F(x) = \lim_{n\to+\infty} F(n)$, 以及结合 $F(-\infty), F(+\infty) \in [0,1]$ 和 $F(+\infty) - F(-\infty) = 1$ 可得

$$F(-\infty) = 0 \quad \text{和} \quad F(+\infty) = 1 \ .$$

针对右连续性, 设 $\{x_n\}_{n=1}^{\infty}$ 是一个单调下降的数列且 $x_n \to x$, 则有

$$F(x_1) - F(x) = P(x < X \leqslant x_1) = \sum_{n=1}^{+\infty} F(x_n) - F(x_{n+1}) = F(x_1) - \lim_{n\to+\infty} F(x_n) \ .$$

于是得到 $\lim_{n\to+\infty} F(x_n) = F(x)$, 再结合函数 $F(x)$ 的单调性有

$$F(x+0) = \lim_{n\to+\infty} F(x_n) = F(x) \ ,$$

由此完成证明.

通过上面的证明发现, 分布函数的三条基本性质, 分别对应于概率的三条公理. 因此, 任何分布函数都满足三条基本性质, 而满足上面三条基本性质的函数必是某随机变量的分布函数.

有了分布函数, 就很容易计算随机变量 X 在很多区间上的概率, 例如

$$\begin{aligned} P(X > a) &= 1 - F(a) \\ P(X < a) &= F(a-0) = \lim_{x\to a^-} F(x) \\ P(X = a) &= F(a) - F(a-0) \\ P(X \geqslant a) &= 1 - F(a-0) \\ P(a \leqslant X \leqslant b) &= F(b) - F(a-0). \end{aligned}$$

针对离散型的随机变量 X, 设其分布列为 $p_k = P(X = x_k)$ $(k = 1, 2, \cdots)$, 根据概率的可列可加性可得 X 的分布函数为

$$F(x) = P(X \leqslant x) = \sum_{k:\, x_k \leqslant x} p_k \ . \tag{4.1}$$

例 4.1 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.

解 当 $x<-1$ 时, 根据 (4.1) 有

$$F(x)=P(X\leqslant x)=P(\emptyset)=0;$$

当 $-1\leqslant x<2$ 时, 根据 (4.1) 有

$$F(x)=P(X\leqslant x)=P(X=-1)=\frac{1}{4};$$

当 $2\leqslant x<3$ 时, 根据 (4.1) 有

$$F(x)=P(X\leqslant x)=P(X=-1)+P(X=2)=\frac{3}{4};$$

当 $x\geqslant 3$ 时有 $F(x)=1$. 如图 4.1(a) 所示, 分布函数 $F(x)$ 是一条阶梯形的曲线, 在 $x=-1,2,3$ 处有跳跃点.

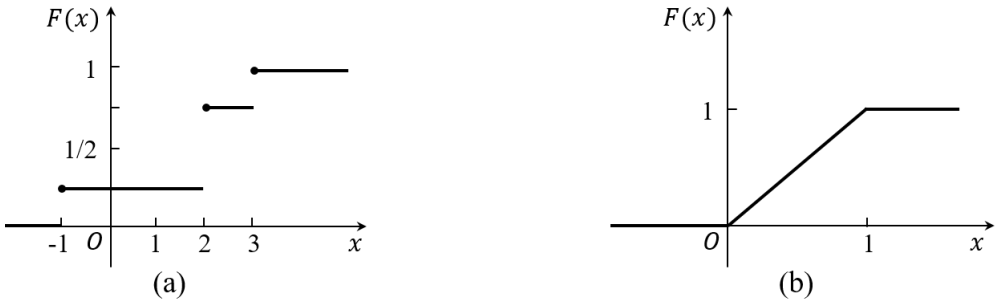


图 4.1 图 (a) 和 (b) 分别给出了例 4.1 和 4.2 的分布函数

例 4.2 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.

解 设随机变量 X 的分布函数为 $F(x)$, 其中 $x\in[0,1]$, 当 $x<0$ 时有 $F(x)=0$; 当 $x>1$ 时有 $F(x)=1$. 当 $x\in[0,1]$ 时有

$$F(x)=P(X\leqslant x)=kx.$$

根据 $F(1)=1$ 求解可得 $k=1$. 从而得到 X 的分布函数为

$$F(x)=\begin{cases}0 & x<0,\\ x & 0\leqslant x\leqslant 1,\\ 1 & x>1.\end{cases}$$

如图 4.1(b) 所示, 分布函数 $F(x)$ 是一条连续的折线.

例 4.3 随机变量 X 的分布函数 $F(x)=A+B\arctan x$, $x\in(-\infty,+\infty)$, 求 $P(X\leqslant 1)$.

解 由分布函数的性质有

$$0 = F(-\infty) = \lim_{x\to-\infty} A + B\arctan x = A - \pi B/2,$$
$$1 = F(+\infty) = \lim_{x\to+\infty} A + B\arctan x = A + \pi B/2,$$

求解可得 $A = 1/2$ 和 $B = 1/\pi$, 从而得到 $P(X \leqslant 1) = 3/4$.

4.2 概率密度函数

离散型随机变量的取值是有限个或可列个离散的单点, 本节研究连续型随机变量, 即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. 离散型和连续型随机变量是实际应用中常遇到的两种随机变量.

定义 4.2 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

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

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** (probability density function), 简称 **密度函数**.

下面给出概率密度函数的一系列性质:

引理 4.1 概率密度函数 $f(x)$ 满足非负性 $f(x) \geqslant 0$ 和规范性 $\int_{-\infty}^{+\infty} f(t)dt = 1$.

任意概率密度函数必然满足非负性和规范性; 而对满足非负性和规范性的任意函数 $f(x)$, 其必为某个随机变量的密度函数, 并有分布函数为 $F(x) = \int_{-\infty}^{x} f(t)dt$, 密度函数完整地刻画了随机变量的统计规律. 分布函数和密度函数都能刻画连续随机变量的统计规律, 但密度函数在图形上对各种分布特征的显示要优越得多, 比分布函数更常用.

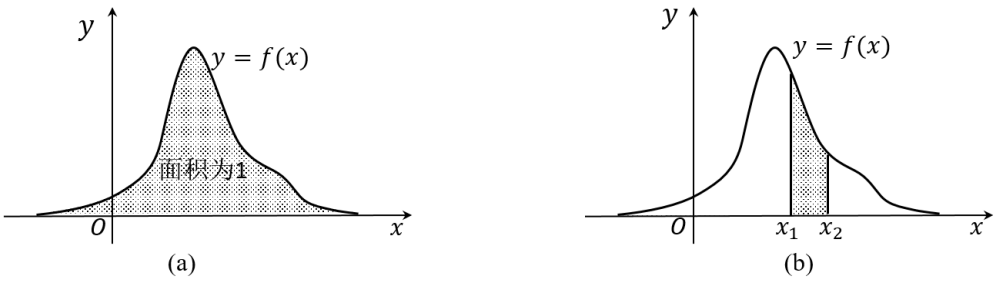


图 4.2 概率密度函数的几何解释

根据规范性可知曲线 $y = f(x)$ 与 x 轴所围成的面积为 1 (如图 4.2(a) 所示). 对任意 $x_1 < x_2$, 有

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

由此给出概率密度的几何解释: 随机变量 X 落入区间 $(x_1, x_2]$ 的概率等于由 x 轴, $x = x_1$, $x = x_2$ 和 $y = f(x)$ 所围成的曲边梯形的面积, 如图 4.2(b) 所示.

引理 4.2 对连续随机变量 X, 分布函数 $F(x)$ 在整个实数域上连续; 若密度函数 $f(x)$ 在 x 点连续, 则分布函数 $F(x)$ 在 x 点可导, 且有 $F^{'}(x) = f(x)$.

证明 该引理根据函数的积分性质直接可得: 若函数 $f(x)$ 在实数域上可积, 则积分函数

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

在实数域上连续; 若函数 $f(x)$ 在实数域上连续, 则 $F(x) = \int_{-\infty}^{x} f(t)dt$ 在实数域上可导, 且有 $F'(x) = f(x)$ 成立.

引理 4.3 对任意常数 c 和连续型随机变量 X, 有 $P(X = c) = 0$.

证明 对任意 $\Delta x > 0$ 有事件 $\{X = x\} \subset \{X \in (x - \Delta x, x]\}$, 根据积分中值定理有

$$P(X = x) \leqslant \lim_{\Delta x \to 0} P(x - \Delta x \leqslant X \leqslant x) = \lim_{\Delta x \to 0} \int_{x-\Delta x}^{x} f(t)dt \leqslant \lim_{\Delta x \to 0} f(\xi)\Delta x = 0,$$

其中 $\xi = \arg\max_{x \in (x-\Delta x, x]} f(x)$, 根据概率的非负性完成证明.

根据上面的引理, 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. 此外, 连续随机变量的概率无需强调端点, 因为

$$P(a \leqslant X \leqslant b) = P(a < X < b) = P(a \leqslant X < b) = P(a < X \leqslant b).$$

因为 $f(x) \neq 0 = P(X = x)$, 由此说明概率密度函数不是概率.

若 $f(x)$ 在点 x 连续, 由连续性定义有

$$\lim_{\Delta x \to 0} \frac{P(x - \Delta x \leqslant X \leqslant x + \Delta x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{\int_{x-\Delta x}^{x+\Delta x} f(t)dt}{\Delta x} = \lim_{\Delta x \to 0} \frac{2\Delta x \cdot f(\xi)}{\Delta x} = 2f(x),$$

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

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

若概率密度 $f(x)$ 越大, 则 X 在 x 附近取值的概率越大.

例 4.4 设随机变量 X 的密度函数

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

求其分布函数 $F(x)$.

解 根据概率密度的规范性有

$$1=\int_{-\infty}^{+\infty}f(t)dt=\int_0^1 tdt+\int_1^2(a-t)dt=a-1,$$

从而求解出 $a=2$, 于是得到具体的密度函数 $f(x)$. 当 $x\leqslant 0$ 时有 $F(x)=0$; 当 $0<x\leqslant 1$ 时, 有

$$F(x)=\int_0^x f(t)dt=x^2/2;$$

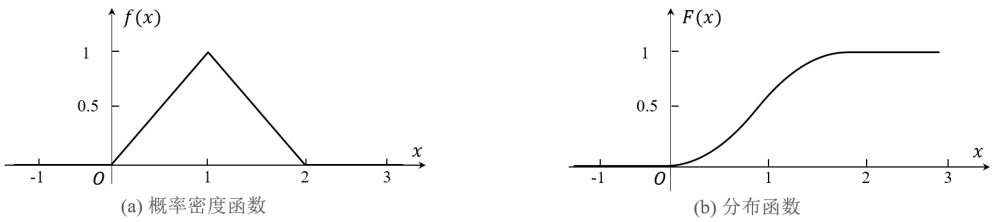
当 $1<x\leqslant 2$ 时, 有

$$F(x)=\int_0^1 f(t)dt+\int_1^x f(t)dt=1/2+\int_1^x(2-t)dt=-x^2/2+2x-1;$$

当 $x\geqslant 2$ 时有 $F(x)=1$. 综合可得

$$F(x)=\begin{cases}0 & x\leqslant 0,\\ x^2/2 & 0<x\leqslant 1,\\ -x^2/2+2x-1 & 1<x\leqslant 2,\\ 1 & x\geqslant 2.\end{cases}$$

随机变量 X 的密度函数和分布函数如图 4.3 所示.



(a) 概率密度函数 (b) 分布函数

图 4.3 例 4.4 中随机变量 X 的密度函数和分布函数图

例 4.5 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

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

求随机变量 Y 的分布函数, 以及计算概率 $P(Y \geqslant X)$.

解 根据概率密度函数的规范性有 $1 = \int_{-\infty}^{+\infty} f(t)dt = 9c$, 由此可得 $c = 1/9$.

用 $F_Y(y)$ 表示随机变量 Y 的分布函数. 当 $y < 1$ 时, 有 $F_Y(y) = P(Y \leqslant y) = 0$; 当 $y \geqslant 2$ 时, 有 $F_Y(y) = P(Y \leqslant y) = 1$; 当 $1 \leqslant y < 2$ 时有

$$\begin{aligned} F_Y(y) &= P(Y \leqslant y) = P(Y = 1) + P(1 < Y \leqslant y) \\ &= P(X \geqslant 2) + P(1 < X \leqslant y) = \int_2^3 t^2/9dt + \int_1^y t^2/9dt = (18 + y^3)/27 \ . \end{aligned}$$

由此可得随机变量 Y 的分布函数为

$$F_Y(y) = \begin{cases} 0, & y < 1, \\ (18 + y^3)/27, & y \in [1,2), \\ 1, & y \geqslant 2. \end{cases}$$

可以观察发现随机变量 Y 不是连续型随机变量, 也不是离散型随机变量. 最后计算概率

$$P(X \leqslant Y) = P(X < 2) = \int_0^2 t^2/9dt = 8/27.$$

例 4.6 已知一个靶半径为 2 米的圆盘, 击中靶上任一同心圆盘上的点的概率与该圆盘的面积成正比. 假设射击都能击中靶, 用 X 表示击中点与圆心的距离, 求 X 的概率密度函数.

解 根据题意分析随机变量 X 的分布函数 $F(x)$. 当 $x < 0$ 时有 $F(x) = 0$; 当 $0 \leqslant x \leqslant 2$ 时有

$$F(x) = P(X \leqslant x) = P(0 \leqslant X \leqslant x) = kx^2.$$

根据分布函数的性质有 $F(2) = 1 = 4k$, 求解可得 $k = 1/4$, 进一步得到 X 的概率密度

$$f(x) = \begin{cases} x/2 & 0 \leqslant x \leqslant 2 \\ 0 & \text{其它.} \end{cases}$$