Lecture 2

一元随机变量与分布

Probability and Statistics
Beihang University

2.1 随机变量的定义

定义 2.1.1 **随机变量**是定义在样本空间 S 上的实数值函数, 通常用大写字母表示，例如 X

定义 2.1.2 取值于有限集合或由互不相同的实数组成的无穷序列 (即可数多个实数值) 的随机变量称为**离散随机变量**, 可取值与某区间中所有实数的随机变量成为连续随机变量.

定义 2.1.3 假设 X 是一个随机变量. 那么 X 的**分布**定义为所有如下形式的概率的集合: $P(X \in C)$ 对任意 $C \subset \mathbb{R}$. 其中 $\{X \in C\}$ 是 $\{\omega \in S : X(\omega) \in C\}$ 的简写 . ω 通常表示一个样本, 或是在随机过程中也成为一个轨道.

例题 2.1.1 假设一个硬币正面出现的概率为 p. 令 N 为投掷直到正面首次出现所需的次数. 假设投掷是相互独立的, 那么 N 是一个取值于 1,2,3,...... 的随机变量

- 每一次连续投掷是一次实验，这些实验组成了样本空间. 对

应每一次实验 N 有唯一的取值, 即第一次正面在一次投掷序列中的位置, 因此它是一个随机变量. N 的分布为

$$P(N=1)=p$$

$$P(N=2)=(1-p)p$$

$$\vdots$$

$$P(N=n)=(1-p)^{n-1}p$$

$$\vdots$$

并且 $P(N\in C)=0$ 对任意 $C\backslash\{1,2,3,......\}$.

离散分布

定理 2.1.1 令 X 为一个随机变量，其取值为 $\{x_i\}_{i=1}^{\infty}$, $x_i \neq x_j$, $\forall i \neq j$. 那么其分布可以由如下式子来计算，对任意 $C \subset \mathbb{R}$,

$$P(X \in C) = \sum_{x_i \in C} P(X = x_i),$$

并且

$$\sum_i P(X = x_i) = 1.$$

- 注意到 $\{X = x_i\} \cap \{X = x_j\} = \varnothing, \forall i \neq j$ 并运用概率公理. ///

对离散分布, $f(x) = P(X = x)$ 通常称为 X 的概率函数. 上面的定理说明对离散随机变量，概率函数即可确定其分布.

例题 2.1.2 连续投掷一个均匀硬币 10 次. 各次投掷相互独立. 令

X 为正面出现次数. 试确定 X 的概率函数.

■ 每一次实验包含 10 次投掷, 所有实验组成一个样本数目为 2^{10} 的样本空间. 根据题设, 每一个样本具有概率 $\frac{1}{2^{10}}$. 容易看出 X 是定义在这些样本上的随机变量. 其取值范围是 $R = \{0, 1, ..., 10\}$. 对 $\forall x \in R$,

$$f(x) = P(X = x) = C_{10}^{x} \cdot \frac{1}{2^{10}}.$$

Bernoulli 分布

定义 2.1.4 如果随机变量 Z 取值为 $\{0, 1\}$, 并且 $P(Z = 1) = p$, 那么我们称 Z 为具有参数 p 的 Bernoulli 随机变量. 其分布则称为具有参数 p 的 Bernoulli 分布.

二项分布

假设一台机器生产某产品, 所生产的产品以概率 p 出现次品. 假设机器生产各个产品的过程是相互独立的. 现对该机器生产的产品进行抽查, 每次检验抽取 n 个产品, 令 X 是在抽检的 n 个产品中次品的个数. 显然 X 是一个随机变量. 它的分布由以下概率函数确定:

$$P(X=x)=C_n^x p^x(1-p)^{n-x}.$$

定义 2.1.5 上面定义的随机变量 X 称为具有参数 n 和 p 的二项随机变量, 其分布则称为具有参数 n 和 p 的二项分布.

连续分布

定义 2.1.6 随机变量 X 称为**连续分布**, 如果存在定义在 $\mathbb{R}$ 上的非负函数 $f(x)$ 使得下面的等式成立,

$$P(X\in E)=\int_E f(x)\,dx$$

其中 E 是任意有界, 无界，开或者闭的实数区间. **非负**函数 $f(x)$ 则称为该连续分布的**密度函数**.

需要注意的是在这个定义中我们并不假定密度函数 $f(x)$ 的连续性, 或是有界性.

定理 2.1.2 由概率公理可知, 密度函数应满足

$$\int_R f(x)\, dx = 1.$$

定理 2.1.3 假设 X 是具有密度函数 $f(x)$. 那么对任意实数 $a \in \mathbb{R}$,

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

因此, $\forall b > a$ 总有 $P(a < X < b) = P(a < X \leqslant b)$.

从该定理可知, 对一个连续型随机变量, 其分布的密度函数

并不是唯一的, 例如我们可以修改密度函数在有限个点上的值而不改变原来的分布函数. 对应于同一个分布函数的密度函数都是几乎处处相等的.

定义 2.1.7 给定两个函数 f, g 以及事件 $D = \{f \neq g\}$. 如果 $P(D) = 0$, 那么我们称 f 和 g 是**几乎处处**相等的, 写作 $f = g, a.e.$

例题 2.1.3 如下函数满足密度函数的条件

$$f(x) = \begin{cases} 0, & x \leqslant 0 \\ \frac{1}{(1+x)^2}, & x > 0 \end{cases}$$

例题 2.1.4 如下无界函数满足密度函数的条件

$$f(x) = \begin{cases} \frac{2}{3}x^{-1/3}, & 0 < x < 1 \\ 0, & \text{其他} \end{cases}$$

// 一致分布（均匀分布）

定义 2.1.8 如果随机变量 X 的取值在区间 $[a,b]$, 并且 X 取值于 $[a,b]$ 的任何子区间的概率与子区间长度成比例. 那么随机变量 X 称为在 $[a,b]$ 上具有一致分布.

定理 2.1.4 如果随机变量 X 称为在上具有一致分布, 那么 X 的密度函数是

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leqslant x \leqslant b \\ 0, & \text{其他} \end{cases}$$

- 由于 X 取值在区间 $[a,b]$ 上, 其密度函数在此区间之外都应为 0. 任取 $x \in (a,b)$, $\epsilon > 0$ 使得 $[x-\epsilon, x+\epsilon] \subset [a,b]$. 根据一致分布的定义, 等式

$$P(x-\epsilon \leqslant X \leqslant x+\epsilon) = \int_{x-\epsilon}^{x+\epsilon} f(t)\,dt$$

不依赖于 x 的位置, 对 x 求导数得到

$$0 = \frac{d}{dx} \int_{x-\epsilon}^{x+\epsilon} f(t)\, dt = f(x+\epsilon) - f(x-\epsilon)\,,\ a.e.$$

由 x 和 ϵ 的任意性, 可知 f 在区间 $[a, b]$ 上为常数（$a.e.$）. 运用

$$\int_a^b f(x)\, dx = \int_{-\infty}^{\infty} f(x)\, dx = 1$$

即可得出定理结论.

2.2　随机变量的分布函数

定义 2.2.1 给定概率空间以及概率 P. 该空间上的随机变量 X 的 **(累积) 分布函数**定义为

$$F(x) = P(X \leqslant x), \forall x.$$

由概率的基本性质可以得出分布函数的以下性质

定理 2.2.1 对于分布函数成立

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

(ii) $x_1 \leqslant x_2 \Longrightarrow F(x_1) \leqslant F(x_2)$, 即 F 是单调不减函数.;

(iii) $F(-\infty) = 0$, $F(\infty) = 1$;

(iv) $F(x) = F(x+)$, 即 F 是右连续函数.

(v) $P(X < x) = F(x-)$;

(vi) $P(X = x) = F(x) - F(x-)$;

(i) 只需注意 $\{X \leqslant x_2\}$ 可写成不相交集合的并集:

$$\{x_1 < X \leqslant x_2\} \cup \{X \leqslant x_1\}.$$

(ii) $\{X \leqslant x_1\} \subset \{X \leqslant x_2\}$

(iii) $F(-\infty)$ 的含义是 $F(-\infty) = \lim_{x \to -\infty} P(X \leqslant x)$, 然而 $\{X \leqslant x\}$ 是关于 x 的不减少事件, 当 $x \to -\infty$ 时 $\{X \leqslant x\}$ 收敛到空集, 因此由概率的上连续性 (参见上一课笔记) 可知 $F(-\infty) = 0$. 同理可证 $F(\infty) = 0$.

(iv) 根据上一性质, $F(x)$ 存在左右极限. 由于 $\{X \leqslant y\}$ 是关于 y 的不减少事件, 当 $y \downarrow x$ 时, 该集合下降收敛到 $\{X \leqslant x\}$, 由概率的上连续性得到 $\lim_{y \downarrow x} F(y) = F(x)$, 正是右连续性.

(v) 当 $y \uparrow x$ 时, $\{X \leqslant y\}$ 递增收敛到 $\{X < x\}$, 由概率的下连续性,

$$F(x-) = \lim_{y \uparrow x} F(y) = P(X < x).$$

(vi) 注意到 $\{X \leqslant x\} = \{X = x\} \cup \{X < x\}$.

我们会看到如果一个函数 $H(x)$ 满足以下条件:(1) $H(x) \geqslant 0$, $H(x)$ 单调不减; (2) $H(x)$ 右连续; (3) $H(-\infty) = 0$, $H(\infty) = 1$; 那么存在随机变量 X, 使得 X 的分布函数正好是 $H(x)$, 因此我们通常把满足上面三个条件的函数直接地称为**分布函数**, 而不指明与其对应的随机变量.

定理 2.2.2 假设连续型随机变量 X 的密度函数是 $f(x)$, 那么它的分布函数

$$F(x) = \int_{-\infty}^{x} f(z)\, dz$$

是连续的, 并且 $F(x)$ 在 $f(x)$ 的连续点上导数存在, 即, 如果 $f(x)$ 在 x 连续那么

$$F'(x) = f(x)$$

- 连续随机变量在任何单点处概率为 0, 因此 $F(x) - F(x-) = 0$, 因此是左连续的, 从而连续. 后一结论则是微积分基本定理

的运用.

例题 2.2.1 假设 X 是一个参数 p 的 Bernoulli 随机变量, 确定其分布函数.

■ 根据定义 $P(X=1)=p, P(X=0)=1-p$. 因此其分布函数为

$$F(x)=\begin{cases} 0, & x<0; \\ 1-p, & 0\leqslant x<1; \\ 1 & x\geqslant 1. \end{cases}$$

例题 2.2.2 假设随机变量 X 在 $[a,b]$ 上具有一致分布, 确定其分布函数.

■ 其分布函数为

$$F(x)=\begin{cases} 0, & x\leqslant a; \\ \int_a^x \frac{1}{b-a}du=\frac{x-a}{b-a}, & a<x\leqslant b; \\ 1 & x>b. \end{cases}$$

例题 2.2.3 验证下面的函数是一个分布函数, 并确定其密度函数,

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

- 容易验证 $F(x)$ 是一个分布函数. 除了 $x=0$ 以外, $F(x)$ 的导数处处存在, 因此其密度函数可以写作

$$f(x)=\begin{cases}0, & x\leqslant 0\\ \frac{1}{(1+x)^2}, & x>0\end{cases}.$$

分位函数

由于累积分布函数是单调的, 其广义逆函数存在, 该函数称为分位函数. 具体地

定义 2.2.2 假设随机变量 X 的分布函数是 $F(x)$, 那么我们定义

$$F^{-1}(p) = \inf\{x : F(x) \geqslant p\}, \forall p \in (0,1).$$

该函数称为 X 的**分位函数**.

例题 2.2.4 假设随机变量 X 在 $[a,b]$ 上具有一致分布, 确定其分位函数.

■ 当 $x \in (a,b)$, X 的分布函数是严格递增的, 因此存在逆函数. 求解

$$p = \frac{x-a}{b-a}, \forall p \in (0,1).$$

得到其分位函数 $F^{-1}(p) = x = (b-a)p + a$.