概率论与数理统计

第四章

杨青
yangq@ustc.edu.cn

第四章：　随机变量的数字特征和极限定理

大纲：

- 数字特征
  - 数学期望
  - 中位数、众数
  - 方差、矩、协方差、相关系数
  - 熵
- 大数定律
- 中心极限定理

随机变量的性质描述

1. 随机变量的**分布函数是对随机变量的概率性质最完整的刻画.**

2. 有些时候我们更关心随机变量的某方面 “特征”(完全由分布函数决定的):

    - 某行业工人的平均工资 (这里工资的分布情况不是最关心的), 或者某行业工人的工资散布程度

3. **能够刻画随机变量某些方面的性质特征的量称为随机变量的数字特征。**

    - 描述随机变量分布的“中心位置”：期望，中位数，众数等
    - 描述随机变量分布的“散布程度”：方差，标准差，极差等
    - 描述随机变量之间的“关系”: 相关系数、协方差等
    - 描述随机变量分布的“形状”: 偏度系数, 峰度系数等
    - 描述随机变量的“不确定性程度”：熵

1.1. 数学期望(Expectation)

数学期望也称均值 (Mean), 是随机变量的一个最基本的数字特征.

定义 4.1 离散型随机变量的期望

设随机变量 X 为离散型随机变量, 其分布律为 $\mathbb{P}(X = x_k) = p_k, \quad k = 1, 2, \cdots$. 如果

$$\sum_{k\geq 1} |x_k| p_k < \infty, \tag{4.1}$$

则称

$$\mathbb{E}X = \sum_{k\geq 1} x_k p_k \tag{4.2}$$

取值的加权平均

为随机变量 X 的数学期望 (Expectation), 简称期望. ♣

绝对收敛条件 (4.1) 是保证期望 (4.2) **收敛到唯一的数**, 与x_i的排列次序无关.

若$\sum_{k=1}^{\infty} |x_k| p_k = +\infty$, 则称$X$的**数学期望不存在**.

- 从"概率的统计定义"出发, 假设把试验独立重复N次, 每次把X的有限个取值 $x_1, \ldots, x_n$记录下来, 设取到 $x_k, k = 1, \ldots, n$的次数分别为 $N_1, \ldots, N_n$, 则X取值的平均$\bar{X}$为 $\bar{X} = \sum_{k=1}^{n} x_k \frac{N_k}{N}$.
- 由概率的统计定义, 当N充分大时, 频率N_k/N很接近概率p_k, 这就是说, 随机变量X的数学期望正是在大量独立重复试验下, X在试验中取值的平均.

例子

- 某场考试中, 有一大题是选择题. 每一小题的 4 个答案中只有一个是正确的, 若规定选对得 2 分, 不选 0 分, 选错扣 1 分, 当有一题你没有把握选择时, 你应该不选还是任选一个答案?

定义 4.2 连续型随机变量的期望

设 $X \sim f(x)$, 如果

$$\int |x| f(x) dx < \infty \tag{4.3}$$

(常表示为 $\mathbb{E}|X| < \infty$), 则称

加权积分

$$\mathbb{E}X = \int x f(x) dx \tag{4.4}$$

为连续型随机变量的数学期望, 简称期望. 否则称为不存在数学期望. 这里绝对可积条件(4.3)是保证期望有确定的值即存在的条件. ♣

- 如果随机变量X, Y有相同的分布，由期望的定义可知：$EX = EY$.
- 如果随机变量$X_1, X_2, \dots, X_n$的分布都相同，我们称它们为**同分布的(identical)**, 此时$EX_1 = EX_2 = \cdots = EX_n$.
- 如果$X_1, X_2, \dots, X_n$不仅有相同的分布, 而且相互独立, 则称$X_1, X_2, \dots, X_n$是**相互独立有相同分布的**(independent and identically distributed，简写**i.i.d.**)随机变量.

补充(离散→连续)

不妨设连续型随机变量 X 的密度 $f(x)$ 的非零取值范围为 $(a,b), a<b$ 可以为 $\mp\infty$, 则可以通过将 X 离散化来考虑 X 的期望:

1. 取点集 $\{x_i\}$, 使得 $a = x_0 < x_1 < \ldots < x_n = b$, 区间长为 $\Delta x_i = x_i - x_{i-1}$.
2. 定义一个新的离散型随机变量 X', 其所有可能取值点为 $\{t_i\}$, $x_{i-1} < t_i \leq x_i$ 且有分布律

$$P(X' = t_i) = p_i = P(x_{i-1} < X \leq x_i) \approx f(t_i)\Delta x_i$$

3. 从而有离散型随机变量期望的定义有:$(\Delta x_i \to 0)$

$$EX' = \sum t_i p_i \approx \sum t_i f(t_i)\Delta x_i \to \int_R x f(x)dx := EX < \infty,$$

$$\text{如果} \sum |x_i| p_i \approx \sum |t_i| f(t_i)\Delta x_i \to \int_R |x| f(x)dx < \infty.$$

例：一些常见分布的期望

- 二项分布 $X \sim B(n,p)$
- Poisson分布 $X \sim P(\lambda)$
- 正态分布 $X \sim N(\mu,\sigma^2)$
- 均匀分布 $X \sim U(a,b)$
- 指数分布 $X \sim Exp(\lambda)$

例：期望不存在

设 $r.v.$ X 的分布律为

$$P\left(X = (-1)^k \frac{2^k}{k}\right) = \frac{1}{2^k}, \quad k = 1, 2, \cdots$$

则 X 的数学期望不存在。

(Cauchy 分布) 设

$$p(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathcal{R},$$

则：该分布的期望不存在.

- **性质一：** 若干个随机变量线性组合的期望, 等于各变量期望的线性组合. 假设 $c_1, c_2, \dots, c_n$ 为常数, 则有

$$E(c_1X_1 + c_2X_2 + \cdots + c_nX_n) = c_1E(X_1) + c_2E(X_2) + \cdots + c_nE(X_n)$$

这里假设各随机变量X_i的期望都存在.

- 例：假设随机变量$X \sim B(n, p)$，求 $E(X)$.

- 例：假设随机变量X_r服从参数为(r, p)的负二项分布，求$E(X_r)$.

◆ **性质二：** 若干个**独立**随机变量的积的期望等于各变量的期望之积，即：

$$E(X_1X_2\cdots X_n)=E(X_1)E(X_2)\cdots E(X_n)$$

这里假设各随机变量X_i的期望都存在.

◆ **性质三：** 若X $\geq$ Y, 则 $EX\geq EY$.

◆ **性质四：** **(随机变量函数的期望)** 设$\mathbf{X}$为一个n 维随机变量, 有分布函数 $F_X(\boldsymbol{x})$, $\mathbf{Y}=g(\mathbf{X})$ 为 m 维随机变量且分布函数为 $F_Y(\boldsymbol{y})$, 若 $\mathbf{Y}$ 的各分量存在期望, 则

$$E\mathbf{Y}=\int_{\mathbb{R}^m}\boldsymbol{y}dF_Y(\boldsymbol{y})\text{ (求}\boldsymbol{y}\text{的分布再求期望)}$$

$$=Eg(\mathbf{X})=\int_{\mathbb{R}^n}g(\boldsymbol{x})dF_X(\boldsymbol{x})=\begin{cases}\sum\limits_{\boldsymbol{x}}g(\boldsymbol{x})P(X=\boldsymbol{x}), & \mathbf{X}\text{为离散型；}\\ \int_{\mathbb{R}^n}g(\boldsymbol{x})f_X(\boldsymbol{x})d\boldsymbol{x}, & \mathbf{X}\text{为连续型.}\end{cases}$$

这里随机向量取期望指每个分量都取期望.

设随机变量 $X \sim N(0,1)$, 求 $Y = X^2 + 1$ 的数学期望.

飞机场载客汽车上有 20 位乘客, 离开机场后共有 10 个车站可以下车, 若某个车站没有人下车则该车站不停车. 设乘客在每个车站下车的可能性相等, 以 X 表示停车的次数, 求 EX.

1.2. 条件数学期望(条件期望)

- 条件分布也是一个概率分布，因此类似数学期望的定义，我们可以给出条件期望的定义. 在给定了随机变量 X 取值 x 的条件之下，Y的条件期望，我们记为 $E(Y|X=x)$，也可简记为 $E(Y|x)$.
- 设X和Y为随机变量，

  若(X,Y)为离散型，且在给定$X=x$下，Y有条件分布率$P(Y=y_i|X=x)$，$i=1,2,\dots$，

  或(X,Y)为连续型，且在给定$X=x$下，Y的条件密度函数为$f(y|x)$. 则

$$E(Y|X=x)=\begin{cases}\sum_i y_i P(Y=y_i|X=x), & (X,Y)\text{为离散型;}\\ \int_{-\infty}^{+\infty} y f(y|x)dy, & (X,Y)\text{ 为连续型.}\end{cases}$$

- 期望所具有的性质条件期望同样满足.

例子

- 设$(X,Y) \sim N\left(a,b,\sigma_1^2,\sigma_2^2,\rho\right)$，试计算$E(Y|X=x)$.

- 设X,Y为两个随机变量，则有

$$EY = E[E(Y|X)]$$

$$Eg(Y) = E[E(g(Y)|X)]$$

- 设$(X,Y) \sim N\left(a,b,\sigma_1^2,\sigma_2^2,\rho\right)$，试计算$E(XY)$.

例子

- 一窃贼被关在有3个门的地牢里，其中第一个门通向自由，出这个门走3小时便可以回到地面。第二个门通向另一个地道，走5个小时将返回地牢。第三个门通向更长的地道，走7个小时也回到地牢。若窃贼每次选择3个门的可能性总是相同，求他为获得自由而奔走的平均时间。

定义 4.4 中位数 (median)

设随机变量 $X \sim F(x)$, 若存在常数 m, 满足

$$\mathbb{P}(X \geq m) = 1 - F(m-0) \geq 1/2, \quad \mathbb{P}(X \leq m) = F(m) \geq 1/2,$$

其中 $F(m-0) = \mathbb{P}(X < m)$, 则常数 m 称为随机变量 X 的中位数(median). ♣

- 从定义上可以看出，m 这个点把 X 的分布从概率上一分两半：在 m 左边占一半，m 右边也占一半，从概率上说，m 这个点正好居于中央，这就是“中位数”得名的由来.

- 例：若随机变量$X \sim B\left(1, \frac{1}{2}\right)$，求X的中位数.

- 连续型随机变量$X \sim f(x)$, 当$f(m) > 0$ 时，中位数唯一且满足$F(m) = 1/2$.

定义 4.5 众数 (mode)

- 若 X 为离散型随机变量，则其概率质量函数最大值对应的随机变量的取值称为众数，记为 m_d.
- 若 X 为连续型随机变量，$X \sim f(x)$, 则使 $f(x)$ 达到最大值的 x 称为众数, 记为 m_d.

♣

- 设随机变量 X 服从对数正态分布, 即$\ln X \sim N\left(\mu, \sigma^2\right)$, 求$X$的期望、中位数和众数.

定义 4.6 p 分位数

设 $0 < p < 1$, 称 Q_p 是随机变量 X 的 p 分位数, 如果

$$\mathbb{P}(X \le Q_p) \ge p, \quad \mathbb{P}(X \ge Q_p) \ge 1 - p.$$

♣

- 中位数的定义是 p 分位数定义的特例（$p = 0.5$）.
- 连续型随机变量的分位数唯一，因此上述定义中两个不等式化为一个等式 $P(X \le Q_p) = p$.

当 $p = 0.25, 0.5, 0.75$ 时, 称分位数$Q_{0.25}, Q_{0.5}, Q_{0.75}$ 为四分位数, 它们把X 的取值分为概率相同的四段.

称 $IQR(X) = Q_{0.75} - Q_{0.25}$为内四分位距 (Interquartile range).

当 p 恰好取百分比例时得到的分位数称为百分位数.

- 设$X \sim N\left(\mu, \sigma^2\right)$，求$X$ 的内四分位矩$IQR(X)$.