3-1：随机变量的数字特征

金百锁

第三章随机变量的数字特征

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。**
   - 度量” 中心”: **期望, 中位数**
   - 度量散布程度: **方差, 绝对偏差, 极差**
   - 分布形状: **偏度系数, 峰度系数**
   - 相关程度: **相关系数**

3.1 数学期望 (均值) 及中位数

3.1.1 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**. 我们先看如下的一个例子

↑Example

一甲乙两人赌技相同, 各出赌金 100 元, 约定先胜三局者为胜, 取得全部 200 元. 现在甲胜 2 局乙胜 1 局的情况下中止, 问赌本该如何分?

↓Example

解: 如果继续赌下去而不中止, 则甲有 3/4 的概率取胜, 而乙胜的概率为 1/4. 所以, 在甲胜 2 局乙胜 1 局的这个情况下, 甲能期望 "得到" 的数目, 应当确定为

$$200 \times \frac{3}{4} + 0 \times \frac{1}{4} = 150(\text{元}),$$

而乙能 "期望" 得到的数目, 则为

$$200 \times \frac{1}{4} + 0 \times \frac{3}{4} = 50(\text{元}).$$

如果引进一个随机变量 X, X 等于在上述局面 (甲值 2 胜乙 1 胜) 之下, 继续赌下去甲的最终所得, 则 X 有两个可能的值: 200 和 0, 其概率分别为 3/4 和 1/4. 而甲的期望所得, 即 X 的 "期望" 值, 即等于

$$X\text{的可能值与其概率之积的累加}$$

这就是 "数学期望" 这个名称的由来. 另一个名称 "均值" 形象易懂, 也很常用.

↑Example

甲乙两人射击水平如下所示

甲:

击中环数	8	9	10
概率	0.3	0.1	0.6

乙:

击中环数	8	9	10
概率	0.2	0.5	0.3

试问两人谁的水平高?

↓Example

假设两人分别射击 N 次, 则他们各自射击的总环数大概为

$$\text{甲}: 8 * 0.3N + 9 * 0.1N + 10 * 0.6N = 9.3N$$

$$\text{乙}: 8 * 0.2N + 9 * 0.5N + 10 * 0.3N = 9.1N$$

因此, 在 N 次射击后, 两人的平均击中环数分别为 9.3 和 9.1, 因此甲的水平稍高一些.

↑Example

通常每条选择题有五个答案，只有一个是正确的。在某次考试中，李老师共拟 20 题，每题 5 分，满分是 100 分。他决定对每一个错误答案倒扣若干分，但应该倒扣多少分才合理呢？

↓Example

使乱撞一通的学生一无所获。换句话说，如果学生完全靠运气的话，他的总分的数学期望值应该是 0。假定对一个错误答案倒扣 x 分，而正确答案得 5 分。随意选一个答案，选着错误答案的概率是 4/5，选着正确答案的概率是 1/5，所以总分的数学期望值是 $(5\times 1/5 - x\times 4/5)\times 20$。要它是 0，即是要 $5 - 4x = 0$，由此 $x = 5/4 = 1.25$，即是对每一个错误答案应该倒扣 1.25 分。

下面我们就给出数学期望 (均值) 的定义:
对一般的离散型分布, 我们有

Definition

设 X 为一离散型随机变量，其分布律为

$$P(X = x_i) = p_i, \quad i = 1, 2, \cdots$$

如果 $\sum_{i=1}^{\infty} |x_i| p_i < +\infty$，则称

$$\sum_{i=1}^{\infty} x_i p_i$$

为随机变量 X 的数学期望（均值），用符号 EX 表示．若 $\sum_{i=1}^{\infty} |x_i| p_i = +\infty$，则称 X 的数学期望（均值）不存在．

对连续型随机变量, 其数学期望的定义如下

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 $\{t_i\}$, $x_{i-1} < t_i \le x_i$ 且有分布律

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

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

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

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

如果连续型随机变量 X 的概率密度函数为 $f(x)$，则当

$$\int_{-\infty}^{\infty} |x| f(x) dx < \infty$$

时，我们将积分

$$\int_{-\infty}^{\infty} x f(x) dx$$

Definition

的值称为 X 的数学期望，记作 EX. 如果

$$\int_{-\infty}^{\infty} |x| f(x) dx = \infty,$$

则称 X 的数学期望不存在.

下面求解几种常见分布的数学期望.

1. 二项分布 $X \sim B(n,p)$:

$$\begin{aligned} EX &= \sum_{k=0}^{n} k.\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \\ &= np \cdot \sum_{i=0}^{n-1} \frac{(n-1)!}{i!(n-1-i)!}p^i(1-p)^{n-1-i} \\ &= np. \end{aligned}$$

2. Poisson 分布 $X \sim P(\lambda)$:

$$EX = \lambda.$$

3. 均匀分布 $X \sim U[a,b]$:

$$EX = \int_a^b x\frac{1}{b-a}dx = \frac{a+b}{2}.$$

4. 正态分布 $X \sim N(\mu, \sigma^2)$:

$$\begin{aligned} EX &= \int_{-\infty}^{+\infty} \frac{x}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \int_{-\infty}^{+\infty} (\sigma y + \mu). \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\ &= \mu. \end{aligned}$$

5. 指数分布 $X \sim Exp(\lambda)$:

$$EX = \int_0^{\infty} x\lambda e^{-\lambda x} dx = 1/\lambda.$$

6. 卡方分布 $X \sim \chi_n^2$:

$$EX = n.$$

7. t 分布 $X \sim t_n$:

$$EX = 0.$$

↑Example

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。

↓Example

解: 由于

$$\sum_{k=1}^{\infty}|(-1)^k\frac{2^k}{k}|\frac{1}{2^k}=\sum_{k=1}^{\infty}\frac{1}{k}=+\infty$$

因此 X 的数学期望不存在。而尽管

$$\sum_{k=1}^{\infty}(-1)^k\frac{2^k}{k}\frac{1}{2^k}=\sum_{k=1}^{\infty}(-1)^k\frac{1}{k}=-ln2.$$

↑Example

(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.

↓Example

解: 容易看出,$p(x)$ 非负, 并且

$$\int_{-\infty}^{\infty} p(x)dx = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{1}{1+x^2}dx = \frac{1}{\pi}\arctan x \,|_{-\infty}^{\infty} = 1,$$

所以 $p(x)$ 是一个密度函数 (称为 Cauchy 分布), 但是

$$\int_{-\infty}^{\infty} |x|p(x)dx = \frac{2}{\pi}\int_{0}^{\infty} \frac{x}{1+x^2}dx = \infty,$$

所以 Cauchy 分布的期望不存在. #

3.1.2 数学期望的性质

1. 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 $c_1, c_2, \ldots, c_n$ 为常数, 则有

$$E(c_1X_1 + c_2X_2 + \cdots + c_nX_n) = c_1EX_1 + c_2EX_2 + \cdots + c_nEX_n,$$

这里假定各变量的期望都存在.

假设随机变量 $X \sim B(n, p)$, 求 EX.

↑Example
↓Example

解: 令 $I_i \sim B(1, p)$, $i = 1, 2, \ldots, n$, 则 $X = \sum_{i=1}^n I_i$ 且 $EI_i = p$. 所以, $EX = \sum_{i=1}^n EI_i = np$.

2. 若干个独立随机变量之积的期望, 等于各变量的期望之积, 即

$$E(X_1X_2\cdots X_n) = EX_1EX_2\cdots EX_n,$$

这里假定各变量相互独立且期望都存在.

3. (随机变量函数的期望) 设随机变量 X 为离散型, 有分布 $P(X = a_i) = p_i$, $i = 1, 2, \ldots$, 或者为连续型, 有概率密度函数 $f(x)$. 则

$$Eg(X) = \begin{cases} \sum_i g(a_i)p_i, & \sum_i |g(a_i)|p_i < \infty; \\ \int_{-\infty}^{+\infty} g(x)f(x)dx, & \int_{-\infty}^{+\infty} |g(x)|f(x)dx < \infty. \end{cases}$$

假设 c 为常数, 则 $EcX = cEX$.

↑Example ↓Example

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

↑Example ↓Example

解: 由 $X \sim N(0,1)$, 我们有

$$\begin{aligned} EX^2 &= \int_{-\infty}^{+\infty} x^2 . \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\ &= 1. \end{aligned}$$

所以, $EY = EX^2 + 1 = 2$.

设随机变量 $X \sim B(n,p)$, 求 $Y = X(n-X)$ 的数学期望.

↑Example
↓Example

解: 由 $X \sim B(n,p)$, 则 $EX = np$, $EX^2 = Var(X) + (EX)^2 = np(1-p) + n^2p^2$. 我们有

$$\begin{aligned} EY = EX(n-X) &= \sum_{k=0}^{n} k(n-k)\binom{n}{k} p^k (1-p)^{n-k} \\ &= nEX - EX^2 \\ &= n(n-1)p(1-p). \end{aligned}$$

设 $X \sim U(-1,1)$, 求 $E(2|X|)$.

↑Example
↓Example

解: 由性质 (3),

$$E(2|X|) = 2\frac{1}{2}\int_{-1}^{1}|x|dx = 2\int_{0}^{1}xdx = 1$$

从福利彩票 33 个红色球中随机选取 6 个中奖号码，且满足条件 $X_1 < X_2 < \cdots < X_6$, 求 $EX_i, i = 1, \cdots 6$.

↑Example ↓Example

解: 令 $Y_1 = X_1 - 1, Y_i = X_{i+1} - X_i - 1, i = 2,3,4,5,6$, 以及 $Y_7 = 33 - X_6$, Y_i 表示两个相邻开奖号码的间隔。$Y_i, i = 1, \cdots, 7$ 是同分布的, 因为 $P(Y_i = k)$ 没有理由比 $P(Y_j = k)$ 大或小。所以它们的期望相同，注意到

$$E\left(\sum_{i=1}^{7} Y_i\right) = 33 - 6 = 27$$

两边取期望, 利用同分布随机变量期望相同得, $E(Y_1) = 27/7$ 故

$$E(X_1) = 1 + 27/7 \approx 4.86$$

$$E(X_i) = E(X_1)\, i = (34/7)i, \quad i = 1, 2, \cdots, 6.$$

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.

↑Example

↓Example

解: 设

$$Y_i = \begin{cases} 1, & \text{第 } i \text{ 个车站有人下车} \\ 0, & \text{第 } i \text{ 个车站无人下车} \end{cases} \qquad i = 1, \cdots, 20.$$

则显然 $X = \sum\limits_{i=1}^{20} Y_i$, 所以

$$\begin{aligned} EX &= \sum_{i=1}^{20} EY_i = \sum_{i=1}^{20} P(\text{第 } i \text{ 个车站有人下车}) \\ &= \sum_{i=1}^{20} [1 - 0.9^{20}] = 8.784. \end{aligned}$$

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.

↑Example
↓Example

解: 设

$$Y_i = \begin{cases} 1, & \text{第 } i \text{ 个车站有人下车} \\ 0, & \text{第 } i \text{ 个车站无人下车} \end{cases} \qquad i = 1, \cdots, 20.$$

则显然 $X = \sum\limits_{i=1}^{20} Y_i$, 所以

$$\begin{aligned} EX &= \sum_{i=1}^{20} EY_i = \sum_{i=1}^{20} P(\text{第 } i \text{ 个车站有人下车}) \\ &= \sum_{i=1}^{20} [1 - 0.9^{20}] = 8.784. \end{aligned}$$

3.1.3 条件期望 (Conditional Mean)

黄胖警官的射击术还不算太差，发第一枪已经有八成机会中靶。而且如果第一枪不中，再瞄准发第二枪，命中率有九成。如果仍然不中，他发第三枪必定中的。现在他要射 10 具靶，你估计他要发多少枪呢？

↑Example

↓Example

先看第一枪便中的，概率是多少？答案是 0.8。再看两枪才中的，概率是多少？答案是（ 1 −0.8 ）× 0.9 = 0.18（为甚么？）最后看三枪才中的，概率是多少？答案是 1 −0.8 −0.18 = 0.02（为甚么？）如果 X 是发枪次数，它的数学期望值 E(X) 就是 1 × 0.8 + 2 × 0.18 + 3 × 0.02 = 1.22，即是说射中一具靶黄警官大概要发 1.22 枪,10 具靶 12.2 枪.

我们知道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 $E(Y|x)$.

Definition

设 X 和 Y 为随机变量, 若 (X,Y) 为离散型, 且在给定 $X=x$ 之下, Y 有分布 $P(Y=a_i|X=x)=p_i$, $i=1,2,\ldots$, 或者 (X,Y) 为连续型, 且在给定 $X=x$ 之下, Y 的条件密度函数为 $f(y|x)$. 则

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

期望所具有的性质条件期望同样满足.

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

↑Example
↓Example

解: 由于 $Y|X=x\sim N(b+\rho\frac{\sigma_2}{\sigma_1}(x-a),(1-\rho^2)\sigma_2^2)$, 所以由二维正态分布的性质知 $E(Y|X=x)=b+\rho\frac{\sigma_2}{\sigma_1}(x-a)$.

[**注**]: 条件期望 $E(Y|X=x)$ 是 x 的函数, 当我们将 x 换为 X 时, $E(Y|X)$ 就是一个随机变量.

我们有如下的公式成立:

定理 1 (Law of total expectation). 设 X, Y 为两个随机变量. 则有

$$EX = E\{E[X|Y]\} \qquad [\text{全期望公式}]$$

证: 我们仅在连续型随机变量的情形下证明此定理. 设 X 的 p.d.f 为 $f(x)$, Y 的 p.d.f 为 $p(y)$, $X|Y=y$ 的 p.d.f 为 $q(x|y)$. 则

$$\begin{aligned} EX &= \int_{-\infty}^{\infty} x f(x)dx = \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} q(x|y)p(y)dydx \\ &= \int \int_{-\infty}^{\infty} xq(x|y)dxp(y)dy = \int_{-\infty}^{\infty} E[X|Y=y]p(y)dy \\ &= E\{E[X|Y]\} \end{aligned}$$

[**推广**]: 当 $g(X)$ 为可积随机变量时, 有 $Eg(X) = E\{E[g(X)|Y]\}$.

由此得到求解期望的第二种方法: 先求解 $h(x) = E(Y|X=x)$, 再求解 $Eh(X)$, 即可求得 EY.

(几何分布的期望) 连续抛掷一枚出现正面的概率为 p 的硬币直到出现正面为止。问需要抛掷的次数的期望多少?

↑Example

↓Example

解: 设 N 为需要抛掷的次数，令 Y 为第一次抛出的结果，$Y=1$ 为正面，$Y=0$ 为反面。

$$E(N)=E(E(N|Y))=p+(1-p)(1+E(N)), E(N)=1/p.$$

一窃贼被关在有 3 个门的地牢里, 其中第一个门通向自由. 出这门走 3 个小时便可以回到地面; 第 2 个门通向另一个地道, 走 5 个小时将返回到地牢; 第 3 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 个门的可能性总相同, 求他为获得自由而奔走的平均时间.

↑Example

↓Example

解: 设这个窃贼需要走 X 小时才能到达地面, 并设 Y 代表他每次对 3 个门的选择情况, Y 各以 1/3 的概率取值 1, 2, 3. 则

$$EX = E[E(X|Y)] = \sum_{i=1}^{3} E(X|Y=i)P(Y=i)$$

注意到 $E(X|Y=1) = 3, E(X|Y=2) = 5 + EX,$

$E(X|Y=3)=7+EX$, 所以

$$EX=\frac{1}{3}[3+5+EX+7+EX]$$

即得到 $EX=15$.

设 $(X,Y)\sim N(a,b,\sigma_1^2,\sigma_2^2,\rho)$, 试计算 EXY.

↑Example
↓Example

解: 先算得

$$E(XY|X=x)=xE(Y|X=x)=x(b+\rho\frac{\sigma_2}{\sigma_1}(x-a));$$

所以

$$\begin{aligned} EXY &= E(bX+\rho\frac{\sigma_2}{\sigma_1}X^2-\rho\frac{\sigma_2}{\sigma_1}aX) \\ &= ab+\rho\frac{\sigma_2}{\sigma_1}(a^2+\sigma_1^2)-\rho\frac{\sigma_2}{\sigma_1}a^2 \\ &= ab+\rho\sigma_1\sigma_2. \end{aligned}$$

3.1.4 中位数 (Median)

我们已经知道, 随机变量 X 的数学期望就是它的平均值, 因此从一定意义上, 数学期望刻画了随机变量所取之值的 "中心位置". 但是, 我们也可以用别的数字特征来刻画随机变量的 "中心位置". 中位数就是这样一种数字特征.

称 m 为连续型随机变量 X 的中位数，如果

$$P(X \le m) \ge \frac{1}{2}, \quad P(X \ge m) \ge \frac{1}{2}.$$

Definition

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.

- 中位数总是存在的.
- 和期望值相比中位数的一个优点是它受个别特别大或特别小的值的影响很小, 而期望则不然: 收入差距非常大时, 中位数比均值更加有效

虽则中位数有这些优点, 但在概率统计中, 无论理论和应用上, 数学期望的重要性都超过中位数, 其原因有一下两个方面:

1. 均值有很多优良的性质, 这些性质时使得在数学处理上很方便. 例如, $E(X_1+X_2)=EX_1+EX_2$, 而 X_1+X_2 的中位数与 X_1, X_2 各自的中位数之间, 不存在简单的联系, 这使中位数在数学上的处理很复杂且不方便;
2. 中位数本身固有的某些缺点: 中位数可以不唯一, 且对于离散型随机变量不易定义.

设随机变量 $X \sim B(1, \frac{1}{2})$, 求 X 的中位数.

↑Example
↓Example

解: 由于 X 的分布函数为

$$F(x) = \begin{cases} 0, & x \leq 0 \\ \frac{1}{2}, & 0 < x < 1 \\ 1, & x \geq 1 \end{cases}$$

由中位数的定义知区间 (0,1) 内的每一个数都是 X 的中位数, 所以此例说明中位数可以不唯一.

中位数的定义是 p 分位数定义的特例:

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

$$P(X \leq \mu_p) \geq p, \quad P(X \geq \mu_p) \geq 1 - p.$$

Definition