第2次课

教学内容：

1. 方差
2. 协方差,相关系数

教学目的及目标：

理解方差、协方差的定义和性质，并能正确计算随机变量及其函数的方差，以及两个随机变量的协方差和相关系数。

教学重点：

方差、协方差的定义、性质和计算,相关系数的定义

教学难点：

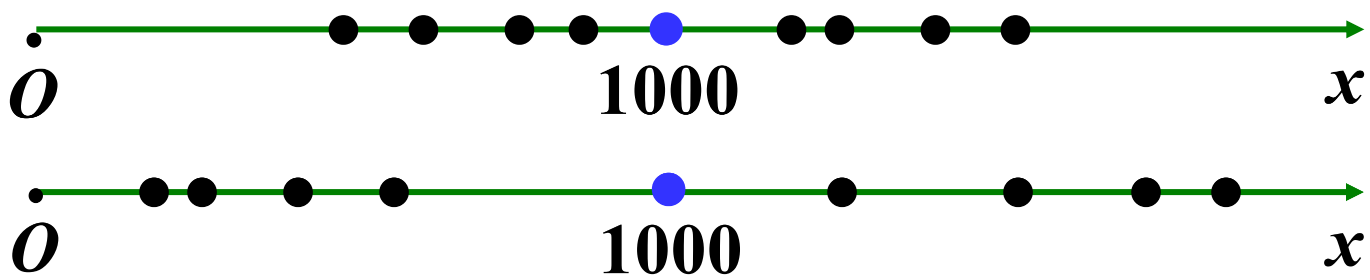
协方差的计算

§4.2　随机变量的方差

一、　概念的引入

数学期望反映随机变量的平均取值，但仅有着一个数字特征往往是不够的. 如：

例　有两批灯泡, 其平均寿命都是 $E(X)$=1000时.



问：那批灯泡质量更好一些？为什么？

研究随机变量与其均值的平均偏离程度是十分必要的。

如何衡量X与EX的平均偏离程度？

$X-E(X)$：X的离差（deviation）

易见恒有：$E[X-E(X)]=0$

原因： X关于E(X)的正负偏差相互抵消

考虑 $E|X-E(X)|$？

绝对值运算不方便。

考虑 $E[X-E(X)]^2$。

——这就是X的方差(Variance) 。

二、 定义

X的函数$(X\text{-}EX)^2$的数学期望

设 X 是一个随机变量，若 $E\{[X-E(X)]^2\}$ 存在，则称 $E\{[X-E(X)]^2\}$ 为 X 的方差，记为 $D(X)$ 或 $\mathrm{Var}(X)$，即

数

$$D(X)=\mathrm{Var}(X)=\underline{E\{[X-E(X)]^2\}}.$$

称 $\sqrt{D(X)}$ 为$r.v.X$的标准差或均方差，记为$\sigma(X)$.

两者量纲相同

说明 方差是来体现随机变量 X 取值分散程度的量。如果$D(X)$地值大，则意味着X 取值的分散程度大，$E(X)$ 的代表性差；而如果 $D(X)$ 的值小，则表示X 的取值比较集中，以 $E(X)$ 作为随机变量的取值代表性好.

三、 计算

随机变量的方差是该随机变量的一个特定函数的数学期望。

1、 利用随机变量函数的数学期望的计算公式

离散型随机变量的方差

$$D(X)=\sum_{k=1}^{+\infty}[x_k-E(X)]^2 p_k,$$

其中 $P\{X=x_k\}=p_k, k=1,2,\cdots$ 是 X 的分布律.

连续型随机变量的方差

$$D(X)=\int_{-\infty}^{+\infty}[x-E(X)]^2 f(x)\,\mathrm{d}x,$$

其中 $f(x)$ 为X的概率密度.

2、 利用简算公式

$$D(X)=E(X^2)-[E(X)]^2.$$

证明

$$D(X)=E\{[X-E(X)]^2\}$$

$$=E\{X^2-2XE(X)+[E(X)]^2\}$$

$$=E(X^2)-2E(X)E(X)+[E(X)]^2$$

$$=E(X^2)-[E(X)]^2$$

$$=E(X^2)-E^2(X).$$

四、性质

$$\left.\begin{aligned}&1.\ D(C)=0\\&2.\ D(CX)=C^2D(X)\\&3.\ D(X\pm Y)=D(X)+D(Y)\\&\qquad\qquad\pm 2E\big((X-E(X))(Y-E(Y))\big)\end{aligned}\right\}D(aX+b)=a^2D(X)$$

特别地，若X,Y相互独立，则

$$D(X\pm Y)=D(X)+D(Y)$$

性质 1 的证明： $D(C)=E\big(C-E(C)\big)^2=0$

性质 2 的证明：

$$\begin{aligned}D(aX+b)&=E\big((aX+b)-E(aX+b)\big)^2\\&=E\big(a(X-E(X))+(b-E(b))\big)^2\\&=E\left(a^2(X-E(X))^2\right)=a^2D(X)\end{aligned}$$

性质 3 的证明：

X 与 Y 的协方差 cov(X,Y)

$$D(X\pm Y)=E\left((X\pm Y)-E(X\pm Y)\right)^2$$

$$=E(X-E(X))^2+E(Y-E(Y))^2\pm 2E\left[(X-EX)(Y-EY)\right]$$

$$=D(X)+D(Y)\pm 2E\left[(X-EX)(Y-EY)\right]$$

注意到， $E\left[(X-EX)(Y-EY)\right]=E(XY)-E(X)E(Y)$

当 X,Y 相互独立时， $D(X\pm Y)=D(X)+D(Y)$

若 X,Y 相互独立 $\quad \Rightarrow \quad \cancel{\Leftarrow} \quad$ $D(X\pm Y)=D(X)+D(Y)$

$$\Updownarrow$$

$$E(XY)=E(X)E(Y)$$

推广：

若 $X_1,\cdots,X_n$ 相互独立， $a_1,a_2,\cdots,a_n,b$ 为常数

则 $$D\left(\sum_{i=1}^{n}a_iX_i+b\right)=\sum_{i=1}^{n}a_i^2D(X_i)$$

4. 对任意常数C, $D(X) \le E(X-C)^2$,当且仅当$C=E(X)$时等号成立。

证明：
$$E(X-C)^2 = E[(X-EX)-(C-EX)]^2$$
$$= E(X-EX)^2 + E(C-EX)^2 = D(X) + (C-EX)^2$$

当$C=E(X)$时，显然等号成立；

当$C \neq E(X)$时， $(C-E(X))^2 > 0$
$$E(X-C)^2 > D(X)$$

P184:12即可用此性质证明。

5. $D(X)=0 \quad \Longleftrightarrow \quad P(X=EX)=1$

此时称X依概率 1 等于常数 $E(X)$。

五、 几种重要分布的方差

1、 0-1分布

设随机变量X具有(0-1)分布，其分布律为

$P\{X=0\}=1-p$，$P\{X=1\}=p$，则$D(X)=p(1-p)$。

证：$E(X)=0\cdot(1-p)+1\cdot p=p$，$E(X^2)=0^2\cdot(1-p)+1^2\cdot p=p$，

$$D(X)=E(X^2)-[E(X)]^2=p-p^2=p(1-p)。$$

2、二项分布

设$X \sim B(n, p)$

引入随机变量 $X_1, X_2, \cdots, X_n$

$$X_i = \begin{cases} 1, & \text{第} i \text{次试验事件} A \text{发生} \\ 0, & \text{第} i \text{次试验事件} \bar{A} \text{发生} \end{cases}$$

$D(X_i) = p(1-p) \qquad i = 1, 2, \cdots, n$

$X_1, X_2, \cdots, X_n$ **相互独立，** $X = \sum_{i=1}^{n} X_i$

故 $D(X) = \sum_{i=1}^{n} D(X_i) = np(1-p)$

3、泊松分布

设$X \sim P(\lambda)$, 则

$$E(X)=\sum_{k=0}^{+\infty} k \cdot \frac{\lambda^{k} e^{-\lambda}}{k!}=\lambda$$

$$\left.\begin{aligned}
&E(X^{2})=E(X(X-1))+E(X) \\
&E(X(X-1))=\sum_{k=0}^{+\infty} k(k-1) \cdot \frac{\lambda^{k} e^{-\lambda}}{k!} \\
&\qquad =\lambda^{2} e^{-\lambda} \sum_{k=2}^{+\infty} \frac{\lambda^{k-2}}{(k-2)!}=\lambda^{2}
\end{aligned}\right\} \Rightarrow E(X^{2})=\lambda^{2}+\lambda$$

$$D(X)=E(X^{2})-E^{2}(X)=\lambda$$

所以 $D(X)=E(X^{2})-[E(X)]^{2}=\lambda$

即泊松分布的均值与方差相等，都等于参数λ

4、均匀分布

设X在区间$(a，b)$上服从均匀分布，

则$E(X)=(a+b)/2$，

$$E(X^2)=\int_{-\infty}^{+\infty} x^2 f(x)dx = \int_a^b x^2 \frac{1}{b-a}dx = \frac{b^3-a^3}{3(b-a)}$$

$$=\frac{a^2+b^2+ab}{3}$$

$$D(X)=E(X^2)-[E(X)]^2$$

$$=\frac{a^2+b^2+ab}{3}-\frac{a^2+b^2+2ab}{4} = \frac{(b-a)^2}{12}$$

5、正态分布

设 $X \sim N(\mu, \sigma^2)$, 则 $EX = \mu$

$$D(X) = E(X-EX)^2 = E(X-\mu)^2$$

$$= \int_{-\infty}^{+\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$\overset{令\frac{x-\mu}{\sigma}=t}{=} \int_{-\infty}^{+\infty} \sigma^2 t^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

$$= \sigma^2$$

6、指数分布

若X服从指数分布 $f(x)=\begin{cases}\lambda e^{-\lambda x} & x>0 \\ 0 & x\le 0\end{cases}$ 则$E(X)=1/\lambda$

$$E(X^2)=\int_{-\infty}^{+\infty}x^2 f(x)\,\mathrm{d}x=\int_0^{+\infty}x^2\cdot\lambda\mathrm{e}^{-\lambda x}\,\mathrm{d}x$$

$$=\int_0^{+\infty}x^2\cdot d\left(-\mathrm{e}^{-\lambda x}\right)=-x^2e^{-\lambda x}\Big|_0^{+\infty}+2\int_0^{+\infty}x\mathrm{e}^{-\lambda x}dx$$

$$=\frac{2}{\lambda}\int_0^{+\infty}x\lambda\mathrm{e}^{-\lambda x}dx=\frac{2}{\lambda}E\left(X\right)=\frac{2}{\lambda^2}$$

于是 $DX=E\left(X^2\right)-\left(EX\right)^2=1/\lambda^2$

表1 几种常见分布的均值与方差

分布	分布率或 密度函数	数学期望	方差
0—1分布	$P(X=k)=p^k(1-p)^{1-k}$ $k=0,1$	p	p(1-p)
二项分布b(n,p)	$P(X=k)=C_n^k p^k(1-p)^{1-k}$ $k=0,1,\ldots,n$	np	np(1-p)
泊松分布 $\pi(\lambda)$	$P(X=k)=\lambda^k e^{-\lambda}/k!$ $k=0,1,\ldots,$	λ	λ
均匀分布U(a,b)	$f(x)=\begin{cases}1/(b-a), & a<x<b\\ 0, & \text{其它}\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
指数分布 $EP(\lambda)$	$f(x)=\begin{cases}\lambda e^{-\lambda x}, & x>0\\ 0, & \text{其它}\end{cases}$	$1/\lambda$	$1/\lambda^2$
正态分布 $N(\mu,\sigma^2)$	$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $-\infty<x<\infty$	μ	σ^2

例1 已知X, Y独立，且都服从$N(0,0.5)$, 求 $E(|X-Y|)$.

解 $X \sim N(0,0.5), Y \sim N(0,0.5)$

$$E(X-Y)=0,\ D(X-Y)=1$$

故 $X-Y \sim N(0,1)$

$$E(|X-Y|)=\int_{-\infty}^{+\infty}|z|\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz$$

$$=\frac{2}{\sqrt{2\pi}}\int_{0}^{+\infty}ze^{-\frac{z^2}{2}}dz=\sqrt{\frac{2}{\pi}}$$

问题：若X, Y独立，且X~$N(1,3)$, Y~$N(2,4)$,则2X-3Y~?

2X-3Y~N(-4,48)

一般地，

独立的n个正态变量的线性组合仍服从正态分布：

若$X_i \sim N(\mu_i, \sigma_i^2)\quad i=1,2,\cdots n$ 且它们相互独立

则它们的线性组合：

$$\boxed{\begin{aligned}&C_0+C_1X_1+C_2X_2+\cdots+C_nX_n\\&\sim N(C_0+C_1\mu_1+\cdots+C_n\mu_n, C_1^2\sigma_1^2+C_2^2\sigma_2^2+\cdots+C_n^2\sigma_n^2)\end{aligned}}$$

$C_0, C_1, C_2 \cdots C_n$是不全为0的常数

例6 设X表示独立射击直到击中目标 n 次为止所需射击的次数，已知每次射击中靶的概率为 p，求$E(X), D(X)$.

解 令 X_i 表示击中目标 $i-1$ 次后到第 i 次击中目标所需射击的次数，$i=1,2,\dots,n$

$X_1, X_2, \cdots, X_n$ 相互独立，且 $X=\sum_{i=1}^{n} X_i$

$$P(X_i=k)=pq^{k-1},\quad k=1,2,\cdots$$

$$p+q=1$$

$$E(X_i)=\sum_{k=1}^{+\infty} kpq^{k-1}=p\sum_{k=1}^{+\infty} kq^{k-1}=p\frac{1}{(1-q)^2}=\frac{1}{p}$$

$$E(X_i^2)=\sum_{k=1}^{+\infty}k(k-1)pq^{k-1}+\sum_{k=1}^{+\infty}kpq^{k-1}$$

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

$$=pq\frac{d^2}{dx^2}\left(\sum_{k=0}^{+\infty}x^k\right)\Bigg|_{x=q}+\frac{1}{p}$$

$$=pq\frac{2}{(1-x)^3}\Bigg|_{x=q}+\frac{1}{p}=\frac{2-p}{p^2}$$

$$D(X_i)=\frac{2-p}{p^2}-\frac{1}{p^2}=\frac{1-p}{p^2}$$

故 $$E(X)=\sum_{i=1}^{n}E(X_i)=\frac{n}{p}$$

$$D(X)=\sum_{i=1}^{n}D(X_i)=\frac{n(1-p)}{p^2}$$

本例给出了几何分布与帕斯卡分布的期望与方差

例7 将编号分别为 $1 \sim n$ 的 n 个球随机地放入编号分别为 $1 \sim n$ 的n 只盒子中，每盒一球. 若球的号码与盒子的号码一致，则称为一个配对. 求配对个数 X 的期望与方差.

解 $$X_i = \begin{cases} 1, & i \text{ 号球放入} i \text{ 号盒} \\ 0, & \text{其它} \end{cases} \qquad i = 1, 2, \cdots, n$$

则 $$X = \sum_{i=1}^{n} X_i$$

但 $X_1, X_2, \cdots, X_n$ 不相互独立,

X_i	1	0
P	$\frac{1}{n}$	$1-\frac{1}{n}$

$i=1,2,\cdots,n$

$$E(X)=\sum_{i=1}^{n}E(X_i)=n\cdot\frac{1}{n}=1$$

$$E(X^2)=E\left(\sum_{i=1}^{n}X_i\right)^2=E\left(\sum_{i=1}^{n}X_i^2+2\sum_{1\le i<j\le n}^{n}X_iX_j\right)$$

$$=\sum_{i=1}^{n}E(X_i^2)+2\sum_{1\le i<j\le n}^{n}E(X_iX_j)$$

X_i^2	1	0
P	$\frac{1}{n}$	$1-\frac{1}{n}$

$$E(X_i^2)=\frac{1}{n} \qquad i=1,2,\cdots,n$$

X_iX_j	1	0
P	$\frac{1}{n(n-1)}$	$1-\frac{1}{n(n-1)}$

$$E(X_iX_j)=\frac{1}{n(n-1)} \qquad i,j=1,2,\cdots,n$$

$$E(X^2)=\sum_{i=1}^{n}E(X_i^2)+2\sum_{1\le i<j\le n}^{n}E(X_iX_j)$$

$$=\sum_{i=1}^{n}\frac{1}{n}+2\sum_{1\le i<j\le n}^{n}\frac{1}{n(n-1)}$$

$$=n\cdot\frac{1}{n}+2\cdot C_n^2\cdot\frac{1}{n(n-1)}$$

$$=2$$

$$D(X)=E(X^2)-E^2(X)=1$$

例： 设X，Y是两个随机变量，EX^2和EY^2存在. 证明：

$$[E(XY)]^2 \le EX^2 EY^2$$

等号成立充要条件：存在常数a使得$P(Y=aX)=1$

解： 考虑实变量t的函数

$$g(t) = E[Y+tX]^2 = E(X^2)t^2 + 2E[XY]t + E(Y^2)$$

$$\because \forall t, g(t) \ge 0 \therefore$$

$$[2E(XY)]^2 \le 4EX^2 EY^2,$$

即有$[E(XY)]^2 \le EX^2 EY^2$.

等号成立充要条件：存在 t_0 使得 $g(t_0)=0$

$$\therefore g(t_0)=E[Y+t_0X]^2=0$$

$$\because E[Y+t_0X]^2=D[Y+t_0X]\ +(E[Y+t_0X])^2=0$$

$$\therefore D[Y+t_0X]\ =0, E[Y+t_0X]=0$$

$$\therefore P[Y+t_0X=0]\ =1 \quad \text{令} a=-t_0$$

$$\therefore P[Y=aX]\ =1$$

小结： 此不等式为柯西-施瓦茨不等式的一种概率表现形式，应熟练掌握.

标准化随机变量

设随机变量 X 的期望$E(X)$、方差$D(X)$都存在, 且 $D(X) \neq 0$, 则称

$$X^* = \frac{X - E(X)}{\sqrt{D(X)}}$$

为 X 的标准化随机变量. 显然,

$$E(X^*) = 0, \quad D(X^*) = 1$$

r.v.的期望与方差并不能唯一确定其分布。

例如

X	-1	0	1
P	0.1	0.8	0.1

$$E(X)=0, D(X)=0.2$$

与

Y	-2	0	2
P	0.025	0.95	0.025

$$E(Y)=0, D(Y)=0.2$$

有相同的期望方差但是分布却不相同

六、切比雪夫不等式

定理4.2.1 **设随机变量X有期望$E(X)$和方差σ^2，则对于任给 $\varepsilon>0$,恒有**

$$P\{|X-E(X)|\geq\varepsilon\}\leq\frac{\sigma^2}{\varepsilon^2}$$

证明 这里仅就连续型随机变量的情形给出证明。设X是一个连续型随机变量，其概率密度为f(x)

则对于任意常数$\varepsilon>0$，有

$$P\{|X-\mu|\geq\varepsilon\}=\int_{|x-\mu|\geq\varepsilon}f(x)dx\leq\int_{|x-\mu|\geq\varepsilon}\frac{(x-\mu)^2}{\varepsilon^2}f(x)dx$$

$$\leq\int_{-\infty}^{+\infty}\frac{(x-\mu)^2}{\varepsilon^2}f(x)dx=\frac{1}{\varepsilon^2}\int_{-\infty}^{+\infty}(x-\mu)^2f(x)dx$$

$$=\frac{D(X)}{\varepsilon^2}=\frac{\sigma^2}{\varepsilon^2}$$

$$P\{|X-E(X)|\geq\varepsilon\}\leq\frac{\sigma^2}{\varepsilon^2} \quad \text{或} \quad P\{|X-E(X)|<\varepsilon\}\geq 1-\frac{\sigma^2}{\varepsilon^2}$$

由切比雪夫不等式可以看出，若 σ^2越小，则事件$\{|X-E(X)|<\varepsilon\}$的概率越大，即随机变量X集中在期望附近的可能性越大.

当方差已知时，切比雪夫不等式给出了*r.v* X与它的期望的偏差不小于ε 的概率的估计式. 如取$\varepsilon=3\sigma$

$$P\{|X-E(X)|\geq 3\sigma\}\leq\frac{\sigma^2}{9\sigma^2}\approx 0.111$$

例3 已知正常男性成人血液中，每一毫升白细胞数平均是7300，均方差是700．利用切比雪夫不等式估计每毫升白细胞数在5200~9400之间的概率．

解：设每毫升白细胞数为X

依题意，$E(X)=7300, D(X)=700^2$

所求为 $P(5200 \le X \le 9400)$

$P(5200 \le X \le 9400)$

$=P(5200-7300 \le X-7300 \le 9400-7300)$

$= P(-2100 \le X-E(X) \le 2100) = P\{|X-E(X)| \le 2100\}$

由切比雪夫不等式

$$P\{|X-E(X)| \le 2100\} \ge 1-\frac{D(X)}{(2100)^2} = 1-\left(\frac{700}{2100}\right)^2 = 1-\frac{1}{9} = \frac{8}{9}$$

即每毫升白细胞数在5200~9400之间的概率不小于8/9．

例4 在每次试验中，事件A发生的概率为 0.75, 利用切比雪夫不等式求：n需要多么大时，才能使得在n次独立重复试验中, 事件A出现的频率在0.74~0.76之间的概率至少为0.90?

解：设X为n 次试验中，事件A出现的次数，则

$X \sim B(n, 0.75)$, $E(X)=0.75n$, $D(X)=0.75*0.25n=0.1875n$

所求为满足 $P(0.74 < \frac{X}{n} < 0.76) \geq 0.90$ 的最小的n .

$P(0.74 < \frac{X}{n} < 0.76)$ 可改写为

$$P(0.74n< X<0.76n) =P(-0.01n<X-0.75n< 0.01n)$$

$$= P\{ |X-E(X)| <0.01n\}$$

$P(0.74n< X<0.76n\ ) = P(-0.01n<X-0.75n< 0.01n)$

$= P\{\ |X-E(X)|\ <0.01n\}$

在切比雪夫不等式中取 $\varepsilon = 0.01\ n$，则

$$P(0.74 < \frac{X}{n} < 0.76) = P\{\ |X-E(X)|\ <0.01n\}$$

$$\geq 1-\frac{D(X)}{(0.01n)^2} = 1-\frac{0.1875n}{0.0001n^2} = 1-\frac{1875}{n}$$

依题意，取 $1-\frac{1875}{n} \geq 0.9$

解得

$$n \geq \frac{1875}{1-0.9} = 18750$$

即n 取**18750**时，可以使得在n次独立重复试验中, 事件A出现的频率在**0.74~0.76**之间的概率至少为**0.90**.