§2 方差

设有两批灯泡寿命（小时）为：

A：服从[950,1050]均匀分布 →平均寿命为1000小时；

B：服从[700,1300] 均匀分布 →平均寿命为1000小时；

问题：哪批灯泡的质量更好？

单从平均寿命这一指标无法判断，进一步考察灯泡寿命X与均值1000小时的偏离程度。

在平均的意义下，如何表示X与EX的偏离程度或集中程度或误差大小？

直观上看，$|X-EX|$表示了 X与EX的偏差，因此利用

$E\{|X-EX|\}$可以表示 X与EX的平均偏离程度。

但为方便起见，利用 $E\{(X-EX)^2\}$来度量

X与EX的偏离程度。

定义：

设X是一个随机变量，若$E\left\{[X-E(X)]^2\right\}$存在，则称其为$X$的方差，记为$D(X)$或$Var(X)$，即 $\boxed{D(X)=Var(X)=E\left\{[X-E(X)]^2\right\}}$

将$\sqrt{D(X)}$记为$\sigma(X)$,称为X的标准差或均方差，它是与随机变量X具有相同量纲的量。

方差意义：D(X)数值的大小表示了X在E(X)附近的集中程度。

D(X)越大，X就越分散，

D(X)越小，X就越集中。

D(X)是刻画X取值分散程度的一个量。

- 对于**离散型**随机变量X，其分布律为：$P(X=x_k)=p_k \quad k=1,2,\cdots$

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

- 对于**连续型**随机变量X,其概率密度为$f(x)$,

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

此外，利用数学期望的性质，可得方差计算公式：

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

证明：$D(X)=E\left\{[X-E(X)]^2\right\}=E\left\{X^2-2XE(X)+[E(X)]^2\right\}$

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

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

例1 设随机变量X具有特征值：

均值 $EX = \mu$

方差 $DX = \sigma^2 \neq 0,$

记 $X^* = \dfrac{X - \mu}{\sigma}$

则：

$E(X^*) = 0$

$D(X^*) = 1$

$\boxed{X^* = \dfrac{X - \mu}{\sigma}}$ 称为X的**标准化变量**

证：$E(X^*)=\frac{1}{\sigma}E(X-\mu)=\frac{1}{\sigma}[E(X)-\mu]=0$

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

$$=E[(\frac{X-\mu}{\sigma})^2]=\frac{1}{\sigma^2}E[(X-\mu)^2]$$

$$=\frac{\sigma^2}{\sigma^2}=1$$

方差的性质：

1. 设C是常数，则$D(C)=0$

2. 设X是随机变量，C是常数，则有$D(CX)=C^2D(X)$

3. 设X,Y是两个随机变量，

则有$D(X+Y)=D(X)+D(Y)+2E\{[X-E(X)][Y-E(Y)]\}$

特别，若X,Y相互独立，则有$D(X+Y)=D(X)+D(Y)$

综合上述三项，设X,Y相互独立，a,b,c是常数，

则$D(aX+bY+c)=a^2D(X)+b^2D(Y)$

4、$D(X)=0 \underset{\text{必要}}{\overset{\text{充分}}{\Leftrightarrow}} P\{X=E(X)\}=1$

证明：

1. $D(C)=E\left\{[C-E(C)]^2\right\}=0$

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

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

3. $D(X+Y)=E\left\{[(X+Y)-E(X+Y)]^2\right\}=E\left\{[(X-E(X))+(Y-E(Y))]^2\right\}$

$$=E\left\{[X-E(X)]^2\right\}+E\left\{[Y-E(Y)]^2\right\}+2E\left\{[X-E(X)][Y-E(Y)]\right\}$$

$$=D(X)+D(Y)+2E\left\{[X-E(X)][Y-E(Y)]\right\}$$

当X,Y相互独立时，$X-E(X)$与$Y-E(Y)$相互独立

故$E\left\{[X-E(X)][Y-E(Y)]\right\}=E[X-E(X)]E[Y-E(Y)]=0$

所以$D(X+Y)=D(X)+D(Y)$

例2：设随机变量X具有0-1分布，其分布律为：

$$P(X=0)=1-p,\ P(X=1)=p,\ 求D(X)。$$

解：

$$E(X)=0\cdot(1-p)+1\cdot p=p$$

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

所以 $D(X)=E(X^2)-[E(X)]^2$

$$=p-p^2=p(1-p)$$

例3：设$X \sim b(n,p)$，求$E(X), D(X)$。

解：随机变量X是n重伯努利试验中事件A发生的次数，

设P(A)=p。引入随机变量：

$$X_k = \begin{cases} 1 & A\text{在第}k\text{次试验发生} \\ 0 & A\text{在第}k\text{次试验不发生} \end{cases} \quad k = 1, 2, \cdots n$$

于是$X_1, X_2, \cdots, X_n$相互独立，服从同一$(0-1)$分布：

X_k	0	1
p_k	1−p	p

易知：$X = X_1 + X_2 + \cdots X_n$

故知：$E(X) = E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i) = np$

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

即$E(X) = np$，$D(X) = np(1-p)$

以n, p为参数的二项分布变量，可分解为n个相互独立且都服从以p为参数的$(0-1)$分布的随机变量之和。

例4：设$X \sim \pi(\lambda)$，求 $D(X)$。

解：X的分布律为：$P(X=k)=\dfrac{\lambda^k e^{-\lambda}}{k!}$ $k=0,1,2,\cdots$ $\lambda>0$

由上节例5已算得$E(X)=\lambda$

$$E(X)=\sum_{k=0}^{\infty} kP\{X=k\}=\sum_{k=0}^{\infty} k\frac{\lambda^k e^{-\lambda}}{k!}=\lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}$$

$$=\lambda e^{-\lambda}\sum_{m=0}^{\infty}\frac{\lambda^m}{m!}=\lambda e^{-\lambda}\cdot e^{\lambda}=\lambda$$

而 $E(X^2)=E\left[X(X-1)+X\right]=E[X(X-1)]+E(X)$

$$=\sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k e^{-\lambda}}{k!}+\lambda \quad =\lambda^2 e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}+\lambda$$

$$=\lambda^2 e^{-\lambda}e^{\lambda}+\lambda \quad =\lambda^2+\lambda$$

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

例5： 设$X \sim U(a,b)$，求$D(X)$。

解：X的概率密度为： $f(x)=\begin{cases}\dfrac{1}{b-a} & a<x<b \\ 0 & 其他\end{cases}$

上节例6已算得：$E(X)=\dfrac{a+b}{2}$

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

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

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

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

例6：设随机变量X服从指数分布，其概率密度为：

$$f(x)=\begin{cases}\dfrac{1}{\theta}e^{-\frac{x}{\theta}} & x>0\\ 0 & x\le 0\end{cases}\quad \theta>0，求E(X),D(X)。$$

解：$E(X)=\int_{-\infty}^{+\infty}xf(x)dx=\int_{0}^{+\infty}x\dfrac{1}{\theta}e^{-\frac{x}{\theta}}dx$

$$=-xe^{-\frac{x}{\theta}}\Big|_0^{+\infty}+\int_0^{+\infty}e^{-\frac{x}{\theta}}dx=\theta$$

$$E(X^2)=\int_{-\infty}^{+\infty}x^2f(x)dx=\int_{0}^{+\infty}x^2\frac{1}{\theta}e^{-\frac{x}{\theta}}dx$$

$$=-x^2e^{-\frac{x}{\theta}}\Big|_0^{+\infty}+\int_0^{+\infty}2xe^{-\frac{x}{\theta}}dx=2\theta^2$$

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

即对指数分布而言，方差是均值的平方，而均值恰为参数 θ

例7：设$X \sim N(\mu,\sigma^2)$，求$E(X),D(X)$。

解：先求标准正态变量$Z=\dfrac{X-\mu}{\sigma}$的数学期望和方差

Z的概率密度为：$\varphi(t)=\dfrac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$

于是 $E(Z)=\dfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}te^{-\frac{t^2}{2}}dt=0$

$$D(Z)=E(Z^2)=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}t^2e^{-\frac{t^2}{2}}dt$$

$$=-\frac{1}{\sqrt{2\pi}}te^{-\frac{t^2}{2}}\Big|_{-\infty}^{+\infty}+\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt=1$$

因为$X=\mu+\sigma Z$，故$E(X)=E(\mu+\sigma Z)=\mu$,

$D(X)=D(\mu+\sigma Z)=\sigma^2D(Z)=\sigma^2$

即正态分布的两个参数μ,σ^2分别是该分布的数学期望和方差。

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 $EP(\lambda)$	$f(x)=\begin{cases}\frac{1}{\theta}e^{-\frac{1}{\theta}x}, & x>0 \\ 0, & others\end{cases}$	θ	θ^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

▲ 一个重要结果

若$X_i \sim N(\mu_i, \sigma_i^2), i = 1,2,\cdots,n,$且它们相互独立。则它们的线性组合

$$C_1X_1 + C_2X_2 + \cdots + C_nX_n \sim N\left(\sum_{i=1}^{n} C_i\mu_i, \sum_{i=1}^{n} C_i^2\sigma_i^2\right)$$

有关数字特征（均值、方差）的使用两个不等式

马尔可夫（Markov）不等式

- 设X是只取非负值的连续型随机变量，且具有数学期望E(X),则对于任意正数 ε ，有 $P\{X \geq \varepsilon\} \leq \dfrac{E(X)}{\varepsilon}$

$$
\begin{aligned}
E(X) &= \int_0^{\infty} xf(x)dx \\
&= \int_0^{\varepsilon} xf(x)dx + \int_{\varepsilon}^{\infty} xf(x)dx \\
&\geq \int_{\varepsilon}^{\infty} xf(x)dx
\end{aligned}
$$

在$[\varepsilon,\infty)$内总有$X \geq \varepsilon$，且$\int_{\varepsilon}^{\infty} f(x)dx$就是$P\{X \geq \varepsilon\}$

$$E(X) \geq \int_{\varepsilon}^{\infty} xf(x)dx \geq \int_{\varepsilon}^{\infty} \varepsilon f(x)dx$$

$$= \varepsilon \int_{\varepsilon}^{\infty} f(x)dx = \varepsilon P\{X \geq \varepsilon\}$$

切比雪夫(Chebyshev)不等式：

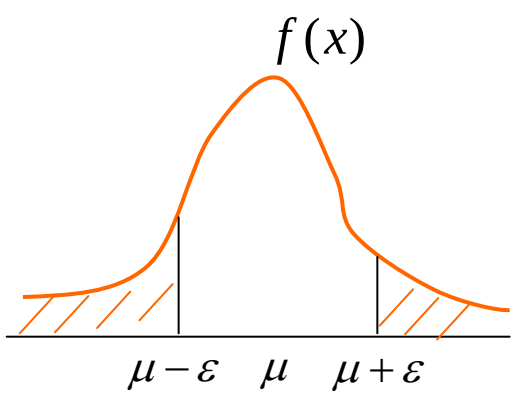
设随机变量X具有数学期望$E(X)=\mu$,方差$D(X)=\sigma^2$

则对于任意$\varepsilon>0$,都有：$P\{|X-E(X)|\geq\varepsilon\}\leq\dfrac{\sigma^2}{\varepsilon^2}$

定理的**等价形式**为：$P\{|X-E(X)|<\varepsilon\}\geq 1-\dfrac{\sigma^2}{\varepsilon^2}$

证明:(仅就X为连续型时证之) 设X的概率密度为$f(x)$,

则 $P\{|X-\mu|\geq\varepsilon\}=\int\limits_{|x-\mu|\geq\varepsilon} f(x)dx \leq \int\limits_{|x-\mu|\geq\varepsilon}\dfrac{(x-\mu)^2}{\varepsilon^2}f(x)dx$



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

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

例：设一个工厂某产品一周的产量X是随机变量

（1） 设$E(X)=500$, 估算$P\{X>1000\}$范围

则由马尔可夫不等式知：

$$P\{X>1000\}\le\frac{E(X)}{1000}=\frac{1}{2}$$

（2） 设$E(X)=500$, $D(X)=100$, 估算$P\{400<X<600\}$范围

则由切比雪夫不等式知：

$$P\{400<X<600\}$$

$$=P\{-100<X-500<100\}$$

$$\ge 1-\frac{100}{100^2}=0.99$$

切比雪夫(Chebyshev)不等式：

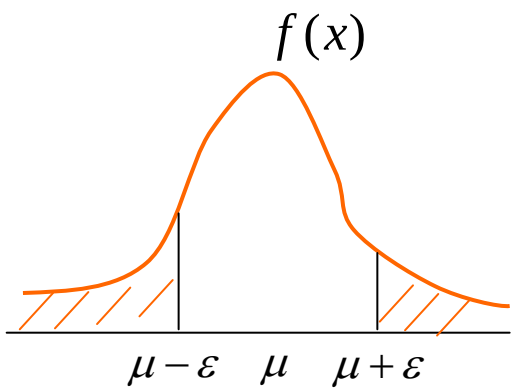
设随机变量X具有数学期望$E(X)=\mu$,方差$D(X)=\sigma^2$

则对于任意$\varepsilon>0$,都有：$P\{|X-\mu|\geq\varepsilon\}\leq\dfrac{\sigma^2}{\varepsilon^2}$

定理的**等价形式**为：$P\{|X-\mu|<\varepsilon\}\geq 1-\dfrac{\sigma^2}{\varepsilon^2}$

证明:(仅就X为连续型时证之) 设X的概率密度为$f(x)$,

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



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

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

例：在n重贝努里试验中，若已知每次试验事件A出现的概率为0.75，试利用契比雪夫不等式估计n，使A出现的频率在0.74至0.76之间的概率不小于0.90。

解：设在n重贝努里试验中，事件A出现的次数为X，

则$X \sim b(n, 0.75)$,

$$E(X) = np = 0.75n, D(X) = npq = 0.1875n,$$

又 $f_n(A) = \dfrac{X}{n}$

而$P\left\{0.74 < \dfrac{X}{n} < 0.76\right\} = P\left\{\left|X - 0.75n\right| < 0.01n\right\}$

$$\geq 1 - \frac{0.1875n}{(0.01n)^2} = 1 - \frac{1875}{n} \geq 0.90$$

$$\Rightarrow n \geq 18750$$