大数据的统计学基础——第6周

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

关注炼数成金企业微信

■提供全面的数据价值资讯，涵盖商业智能与数据分析、大数据、企业信息化、数字化技术等，各种高性价比课程信息，赶紧掏出您的手机关注吧！

赌金分配

- 在17世纪，有一个赌徒向法国著名数学家帕斯卡挑战，给他出了一道题目：甲乙两个人赌博，他们两人获胜的机率相等，比赛规则是先胜三局者为赢家，赢家可以获得100法郎的奖励。当比赛进行到第三局的时候，甲胜了两局，乙胜了一局，这时由于某些原因中止了比赛，那么如何分配这100法郎才比较公平？
- 分析：假设继续再赌下次，则有如下结果

第四局	甲胜	乙胜	
第五局		甲胜	乙胜

- 前三局中，甲已胜了两局，乙胜了一局

赌金分配

- 所以甲最终获胜的概率是3/4，乙最终获胜的概率是1/4
- 根据甲乙两人的获胜概率分配赌金
- 甲的期望所得值为100*（3/4）=75法郎；乙的期望所得值100*（1/4）=25法郎

- 若设X为甲最终获得的赌金，则

X	100	0
P	3/4	1/4

- 从而X的期望值，也就是甲最终获得的赌金的期望值为

$$100 \times \frac{3}{4} + 0 \times \frac{1}{4} = 75(\text{法郎})$$

- 这个故事里出现了“期望”这个词，数学期望由此而来

离散型随机变量的数学期望

- 设离散型随机变量X的分布律为$P\{X = x_k\} = p_k$，k=1,2,……。若级数$\sum_{k=1}^{\infty} x_k p_k$绝对收敛，则称$\sum_{k=1}^{\infty} x_k p_k$为随机变量X的数学期望，记为E(X)。即$E(X)=\sum_{k=1}^{\infty} x_k p_k$

- 例：随机变量X的分布律如下

X	0	1	2	3
P	0.2	0.1	0.5	0.2

- 则$E(X)= \sum_{k=1}^{4} x_k p_k$=0*0.2+1*0.1+2*0.5+3*0.2=1.61

就是求随机变量的取值乘以相应的概率的和

运动员选拔

- 设某教练员有甲、乙两名射击运动员, 现需要选拔其中的一名参加运动会, 根据过去的记录显示, 二人的技术水平如下:

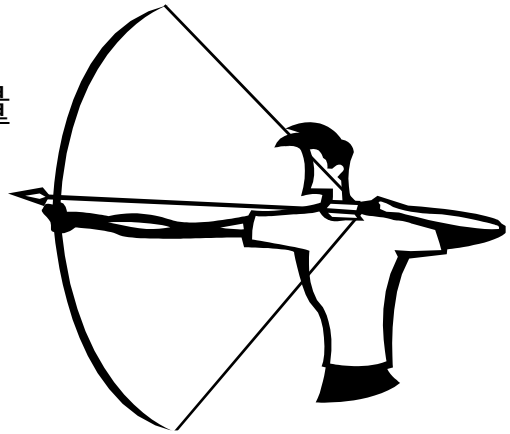
甲射手

击中环数	8	9	10
概率	0.3	0.1	0.6

乙射手

击中环数	8	9	10
概率	0.2	0.5	0.3

- 试问哪个射手技术较好?
- 一个射击运动员的射击水平，可以通过他的平均得分来衡量



运动员选拔

- 如何计算平均得分？
- 假设甲乙两人每人射击了10次，那么理论上，甲乙的得分是：

甲	8	8	8	9	10	10	10	10	10	10
乙	8	8	9	9	9	9	9	10	10	10

- 那么理论上，甲的平均得分为：
- （8+8+8+9+10+10+10+10+10+10）/10=（8*3+9*1+10*6）/10=8*0.3+9*0.1+10*0.6=9.3（环）
- 乙的平均得分为：
- （8+8+9+9+9+9+9+10+10+10）/10=（8*2+9*5+10*3）/10=8*0.2+9*0.5+10*0.3=9.1（环）
- 所以甲比乙的射击技术好

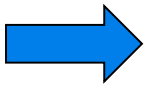
运动员选拔

- 若用X记录甲射击一次击中的环数，用Y记录乙射击一次击中的环数，则

X	8	9	10
P	0.3	0.1	0.6

Y	8	9	10
P	0.2	0.5	0.3

- X的期望值：8*0.3+9*0.1+10*0.6=9.3
- Y的期望值：8*0.2+9*0.5+10*0.3=9.1
- 随机变量的期望值=均值



与理论上的平均得分相等

新生婴儿得分

例 1 某医院当新生儿诞生时，医生要根据婴儿的皮肤颜色、肌肉弹性、反应的敏感性、心脏的搏动等方面的情况进行评分，新生儿的得分 X 是一个随机变量. 据以往的资料表明 X 的分布律为

X	0	1	2	3	4	5	6	7	8	9	10
p_k	0.002	0.001	0.002	0.005	0.02	0.04	0.18	0.37	0.25	0.12	0.01

试求 X 的数学期望 $E(X)$.

解 $$\begin{aligned} E(X) &= 0\times 0.002+1\times 0.001+2\times 0.002+3\times 0.005+4\times 0.02 \\ &\quad +5\times 0.04+6\times 0.18+7\times 0.37+8\times 0.25+9\times 0.12+10\times 0.01 \\ &= 7.15(\text{分}) \end{aligned}$$

这意味着，若考察医院出生的很多新生儿，例如 1000 个，那么一个新生儿的平均得分约 7.15 分，1000 个新生儿共得分约 7150 分. □

候车时间

例 3 按规定，某车站每天 8:00～9:00，9:00～10:00 都恰有一辆客车到站，但到站的时刻是随机的，且两者到站的时间相互独立. 其规律为

到站时刻	8:10	8:30	8:50
	9:10	9:30	9:50
概率	$\frac{1}{6}$	$\frac{3}{6}$	$\frac{2}{6}$

一旅客 8:20 到车站，求他候车时间的数学期望.

解 设旅客的候车时间为 X（以分计）. X 的分布律为

X	10	30	50	70	90
p_k	$\frac{3}{6}$	$\frac{2}{6}$	$\frac{1}{6}\times\frac{1}{6}$	$\frac{1}{6}\times\frac{3}{6}$	$\frac{1}{6}\times\frac{2}{6}$

在上表中，例如

$$P\{X=70\}=P(AB)=P(A)P(B)=\frac{1}{6}\times\frac{3}{6},$$

其中 A 为事件“第一班车在 8:10 到站”，B 为“第二班车在 9:30 到站”. 候车时间的数学期望为

$$E(X)=10\times\frac{3}{6}+30\times\frac{2}{6}+50\times\frac{1}{36}+70\times\frac{3}{36}+90\times\frac{2}{36}$$
$$=27.22(\text{分}).$$

□

(0-1)分布的数学期望

- 若X服从（0-1）分布，参数p=0.5，求E(X)。

X	0	1
P	0.5	0.5

- 则E(X)=0*0.5+1*0.5=0.5=p
- 更一般情况：

X	0	1
P	1-p	p

- E(X)=0*(1-p)+1*p=p
- 结论：若X服从参数为p的（0-1）分布，则E(X)=p

二项分布的数学期望

- 体育课中小明进行投篮练习，若小明每次投中的概率是0.6，记X为3次投篮投中的次数。求E(X)。
- 解：X~B(3,0.6)，则
- P(X=0)=0.4*0.4*0.4=0.064
- P(X=1)=3*0.6*0.4*0.4=0.288
- P(X=2)=3*0.6*0.6*0.4=0.432
- P(X=3)=0.6*0.6*0.6=0.216

X	0	1	2	3
P	0.064	0.288	0.432	0.216

- 故E(X)=0*0.064+1*0.288+2*0.432+3*0.216=1.8=3*0.6

二项分布的数学期望

- 更一般地情况，若X~B(n,p),则X的分布律为$P\{X=x_k\}=\binom{n}{k}p^k(1-p)^{n-k}$

$$E(X)=\sum_{k=0}^{n}k\cdot P\{X=k\}=\sum_{k=0}^{n}k\cdot C_n^k p^k(1-p)^{n-k}=\sum_{k=0}^{n}\frac{kn!}{k!(n-k)!}p^k(1-p)^{n-k}$$

$$=\sum_{k=1}^{n}\frac{np(n-1)!}{(k-1)![(n-1)-(k-1)]!}p^{k-1}(1-p)^{(n-1)-(k-1)}$$

$$=np\sum_{k=1}^{n}\frac{(n-1)!}{(k-1)![(n-1)-(k-1)]!}p^{k-1}(1-p)^{(n-1)-(k-1)}$$

令i=k-1

$$=np\sum_{i=0}^{n-1}\frac{(n-1)!}{i![(n-1)-i]!}p^{i}(1-p)^{(n-1)-i}$$

$$=np\sum_{i=0}^{n-1}C_{n-1}^{i}p^{i}(1-p)^{(n-1)-i}$$

二项式定理:$(x+y)^n=\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}$

$$=np[p+(1-p)]^{n-1}=np$$

连续型随机变量的数学期望

- 将离散型随机变量的定义类比到连续型随机变量上
- 设连续型随机变量X的概率密度为f(x)，若积分$\int_{-\infty}^{\infty} xf(x)\,dx$绝对收敛，则称积分$\int_{-\infty}^{\infty} xf(x)\,dx$的值为随机变量X的数学期望。记为E(X)，即

$$\mathrm{E}(X)=\int_{-\infty}^{\infty} xf(x)\,dx$$

$$\mathrm{E}(X)=\sum_{k=1}^{\infty} x_k p_k$$

例子

5. 设在某一规定的时间间隔里，某电气设备用于最大负荷的时间 X（以 min 计）是一个随机变量，其概率密度为

$$f(x)=\begin{cases}\dfrac{1}{1\ 500^2}x, & 0\leqslant x\leqslant 1\ 500,\\ \dfrac{-1}{1\ 500^2}(x-3\ 000), & 1\ 500<x\leqslant 3\ 000,\\ 0, & \text{其他.}\end{cases}$$

求 $E(X)$.

均匀分布的数学期望

- 设X~U（2，4），求E(X)。

- X的概率密度为$f(x)=\begin{cases}\frac{1}{4-2}=\frac{1}{2}, & 2<x<4\\ 0, & \text{其他}\end{cases}$

- 根据定义，X的数学期望为

$$E(X)=\int_{-\infty}^{\infty}xf(x)\,dx=\int_{2}^{4}\frac{x}{2}dx=\frac{x^2}{4}|_2^4=\frac{16}{4}-\frac{4}{4}=3 \text{——>恰好是2与4的中点}$$

- 更一般地情况，若X~U(a,b)，则

$$E(X)=\int_{-\infty}^{\infty}xf(x)\,dx=\int_{a}^{b}\frac{x}{b-a}dx=\frac{x^2}{2(b-a)}|_a^b=\frac{b^2}{2(b-a)}-\frac{a^2}{2(b-a)}=\frac{(b-a)(b+a)}{2(b-a)}$$

$$=\frac{b+a}{2}$$

⟹ 服从均匀分布的随机变量的期望值位于区间（a,b）中点

正态分布的数学期望

- 设 $X \sim N(\mu,\sigma^2)$，其概率密度为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\sigma>0,-\infty<x<\infty.$$

- 则有

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

令 $\frac{x-\mu}{\sigma}=t\Rightarrow x=\mu+\sigma t$

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

$$=\mu\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{t^2}{2}}dt+\frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}te^{-\frac{t^2}{2}}dt=0$$

$=\mu$ 标准正态分布的概率密度积分=1

分布的数学期望

- 见附表1 几种常见的概率分布表

分布	参数	分布律或概率密度	数学期望
(0－1)分布	$0<p<1$	$P\{X=k\}=p^k(1-p)^{1-k}, k=0,1$	p
二项分布	$n\geqslant 1$ $0<p<1$	$P\{X=k\}=\binom{n}{k}p^k(1-p)^{n-k}$ $k=0,1,\cdots,n$	np
均匀分布	$a<b$	$f(x)=\begin{cases}\dfrac{1}{b-a}, & a<x<b\\ 0, & \text{其他}\end{cases}$	$\dfrac{a+b}{2}$
正态分布	μ $\sigma>0$	$f(x)=\dfrac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/(2\sigma^2)}$	μ

赌博的公平性

- 美国的轮盘中常用的轮盘上有38个数字，每一个数字被选中的概率都是相等的。赌注一般押在其中某一个数字上，如果轮盘的输出值和这个数字相等，那么下赌者可以将相当于赌注35倍的奖金(原注包含在内)，若输出值和下压数字不同，则赌注就输掉了。因此，考虑到38种所有的可能结果，以1美元赌注押一个数字上获利的期望值为：
- -1*（37/38）+35*（1/38）≈-0.0526
- 结果约等于-0.0526美元。也就是说，平均起来每赌1美元就会输掉5美分，即美式轮盘以**1美元作赌注的期望值为0.9474美元**。
- 在赌博中，一场每位**参与者获利期望值为0（没有净利或净亏）的游戏通常会被叫做“公平竞赛”。**
- 所以这样的赔率设计是不公平的。

赌博的公平性

- 某个赌博游戏规则如下：每个参加者每次先付赌金1元，然后将三个骰子一起掷出，他可以赌某个点数，譬如赌“1”点。如果三枚骰子中出现一个“1”点，庄家除把赌金1元还外，再奖1元；如果出现两个“1”点，除发还赌金外，再奖2元；如果全是“1”点，那么，除发还赌金外，再奖3元。试问这样的游戏规则对下注者是否公平？
- 用X记参加者最终的获利。
- X的可能取值：-1，1，2，3
- $P\{X=-1\}=\frac{5}{6}*\frac{5}{6}*\frac{5}{6}=\frac{125}{216}$
- $P\{X=1\}=\binom{3}{1}\left(\frac{5}{6}\right)^2\frac{1}{6}=\frac{75}{216}$
- $P\{X=2\}=\binom{3}{2}\left(\frac{1}{6}\right)^2\frac{5}{6}=\frac{15}{216}$
- $P\{X=3\}=\frac{1}{6}*\frac{1}{6}*\frac{1}{6}=\frac{1}{216}$

赌博的公平性

X	-1	1	2	3
P	125/216	75/216	15/216	1/216

- 所以E(X)=-1*(125/216)+1*(75/216)+2*(15/216)+3*(1*216)=-17/216
- 所以，平均每参与216次，会输17元。对于庄家来说，只要长期有人参与这个游戏，肯定是会赢钱。这个赌博的设计不公平。

数学期望的性质

- 1.设C是常数，则有E(C)=C
- 2.设X是一个随机变量，C是常数，则有E(CX)=CE(X)
- 3.设X，Y是两个随机变量，则有E(X+Y)=E(X)+E(Y) ——>可以推广到任意有限个随机变量之和的情况
- 4.设X，Y是相互独立的两个随机变量，则E(XY)=E(X)E(Y)
- 5.设Y是随机变量X的函数：Y=g(X)（g是连续函数），则

  1）X是离散型， $\mathrm{E}(\mathrm{Y})=\sum_{k=1}^{\infty} g(x_k)p_k$

  2）X是连续型，$\mathrm{E}(Y)=\int_{-\infty}^{\infty} g(x)f(x)\,dx$

例子

例 12 一民航送客车载有 20 位旅客自机场开出，旅客有 10 个车站可以下车. 如到达一个车站没有旅客下车就不停车. 以 X 表示停车的次数，求 $E(X)$（设每位旅客在各个车站下车是等可能的，并设各位旅客是否下车相互独立）.

解 引入随机变量

$$X_i=\begin{cases}0, & \text{在第 } i \text{ 站没有人下车},\\ 1, & \text{在第 } i \text{ 站有人下车},\end{cases} \quad i=1,2,\cdots,10.$$

易知 $$X=X_1+X_2+\cdots+X_{10}.$$

现在来求 $E(X)$.

按题意，任一旅客在第 i 站不下车的概率为 $\frac{9}{10}$，因此 20 位旅客都不在第 i 站下车的概率为 $\left(\frac{9}{10}\right)^{20}$，在第 i 站有人下车的概率为 $1-\left(\frac{9}{10}\right)^{20}$，也就是

$$P\{X_i=0\}=\left(\frac{9}{10}\right)^{20}, P\{X_i=1\}=1-\left(\frac{9}{10}\right)^{20}, i=1,2,\cdots,10.$$

由此

$$E(X_i)=1-\left(\frac{9}{10}\right)^{20}, i=1,2,\cdots,10.$$

进而

$$\begin{aligned}E(X)&=E(X_1+X_2+\cdots+X_{10})\\&=E(X_1)+E(X_2)+\cdots+E(X_{10})\\&=10\left[1-\left(\frac{9}{10}\right)^{20}\right]=8.784(\text{次}).\end{aligned}$$

方差——数据的离散程度

- 运动员选拔
- 设某教练员有甲、乙两名射击运动员, 现需要选拔其中的一名参加运动会, 根据过去的记录显示, 二人的技术水平如下:

击中环数		8	9	10
概率	甲	0.3	0.2	0.5
	乙	0.1	0.6	0.3

- 试问哪个射手技术较好?
- X记甲击中环数，Y记乙击中环数
- E(X)=8*0.3+9*0.2+10*0.5=9.2
- E(Y)=8*0.1+9*0.6+10*0.3=9.2

平均水平相等，考察发挥的稳定性——方差

方差——数据的离散程度

- 假设甲乙两人每人各射击了10次，理论上击中的环数

甲	8	8	8	9	9	10	10	10	10	10
乙	8	9	9	9	9	9	9	10	10	10

- 则甲的方差为: $\frac{1}{10}[(8-9.2)^2+(8-9.2)^2+\cdots\ldots+(10-9.2)^2]=\frac{1}{10}[3\times(8-9.2)^2+2\times(9-9.2)^2+5\times(10-9.2)^2]=0.3\times(8-9.2)^2+0.2\times(9-9.2)^2+0.5\times 9.2)^2=0.844$

- 乙的方差为：$\frac{1}{10}[(8-9.2)^2+(9-9.2)^2+\cdots\ldots+(10-9.2)^2]=\frac{1}{10}[1\times(8-9.2)^2+6\times(9-9.2)^2+3\times(10-9.2)^2]=0.1\times(8-9.2)^2+0.6\times(9-9.2)^2+0.3\times 9.2)^2=0.4$ ⟹ 乙的稳定性更好

随机变量的方差

- 设X施一个随机变量，若$E\{[X-E(X)]^2\}$存在，则称E{[X-E(X)]}为X的方差，记为D(X)或Var(X)，即D(X)=Var(x)= $E\{[X-E(X)]^2\}$
- $\sqrt{D(X)}$称为X的标准差。

- 若X是离散型随机变量,则$D(X)=\sum_{k=1}^{\infty}[x_k-E(X)]^2 p_k$
- 若X是连续型随机变量，则$D(X)=\int_{-\infty}^{\infty}[x-E(X)]^2 f(x)\,dx$
- $D(X)=E\{[X-E(X)]^2\}=E\{X^2-2XE(X)+[E(X)]^2\}=E(X^2)-2E(X)E(X)+[E(X)]^2=E(X^2)-[E(X)]^2$ $E(X^2)=\mathrm{D(X)}+[E(X)]^2$
- 与第一周方差的化简式$\sigma^2=\frac{1}{N}\sum_{i=1}^{N}{X_i}^2-\mu^2$是一样的。

标准化

例 1 设随机变量 X 具有数学期望 $E(X)=\mu$，方差 $D(X)=\sigma^2\neq 0$. 记

$$X^*=\frac{X-\mu}{\sigma},$$

则

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

$$D(X^*)=E(X^{*2})-[E(X^*)]^2=E\left[\left(\frac{X-\mu}{\sigma}\right)^2\right]$$

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

即 $X^*=\frac{X-\mu}{\sigma}$ 的数学期望为 0，方差为 1. X^* 称为 X 的**标准化变量**.

（0-1）分布的方差

例 2 设随机变量 X 具有(0－1)分布，其分布律为

$$P\{X=0\}=1-p, \quad P\{X=1\}=p.$$

求 $D(X)$.

解

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

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

由(2.4)式

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

均匀分布的方差

例 4 设随机变量 $X\sim U(a,b)$，求 $D(X)$.

解 X 的概率密度为

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

上节例 7 已算得 $E(X)=\dfrac{a+b}{2}$. 方差为

$$\begin{aligned}D(X) &= E(X^2)-[E(X)]^2\\ &=\int_a^b x^2\,\frac{1}{b-a}\mathrm{d}x-\left(\frac{a+b}{2}\right)^2=\frac{(b-a)^2}{12}.\end{aligned}$$

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$\mathrm{D(CX)} = C^2 D(X)$，$D(X+C) = D(X)$
- 3.设X，Y是两个随机变量，则有 $D(X+Y) = D(X) + D(Y) - 2E\{(X - E(X))(Y -$

二项分布的方差

- 设随机变量X~B(n,p)，求D(X)
- 由二项分布的定义知道，随机变量X是n重伯努利试验中试验成功的次数，且每次试验成功的概率为p。
- 引入随机变量$X_k=\begin{cases}1,\text{第}k\text{次试验成功}\\0,\text{第}k\text{次试验失败}\end{cases}$ k=1,2.....n
- 则有$X=X_1+X_2+\cdots+X_n$。$X_1,X_2,\cdots,X_n$相互独立且都服从参数为p的（0-1）分布
- 故$D(X_1)=D(X_2)=\cdots=D(X_n)=p(1-p)$
- D(X)=D$(X_1+X_2+\cdots+X_n)$=np(1-p)

正态分布的方差

例 7 设随机变量 $X\sim N(\mu,\sigma^2)$，求 $E(X),D(X)$.

解 先求标准正态变量

$$Z=\frac{X-\mu}{\sigma}$$ → 服从标准正态分布

的数学期望和方差. Z 的概率密度为

$$\varphi(t)=\frac{1}{\sqrt{2\pi}}e^{-t^2/2},$$

于是 $$E(Z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}te^{-t^2/2}dt=\frac{-1}{\sqrt{2\pi}}e^{-t^2/2}\Big|_{-\infty}^{\infty}=0,$$

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

↓ 分部积分，公式 $\int u\,dv=uv-\int v\,du.$

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

因 $X=\mu+\sigma Z$，即得

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

正态分布

- 对于n个随机变量X_i，若$X_i \sim N(\mu_i, \sigma_i^2)$，$i = 1,2 \ldots\ldots n$,且相互独立（独立同分布），则

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

- 例：若X~N(0,1)，Y~N(2,4)，且X与Y相互独立，则X+Y~N(2,5)
- 例：

例 8 设活塞的直径(以 cm 计) $X \sim N(22.40, 0.03^2)$，气缸的直径 $Y \sim N(22.50, 0.04^2)$，X, Y 相互独立. 任取一只活塞，任取一只气缸，求活塞能装入气缸的概率.

解 按题意需求 $P\{X<Y\}=P\{X-Y<0\}$. 由于

$$X-Y \sim N(-0.10, 0.002\,5),$$

故有

$$P\{X<Y\}=P\{X-Y<0\}$$

$$=P\left\{\frac{(X-Y)-(-0.10)}{\sqrt{0.002\,5}}<\frac{0-(-0.10)}{\sqrt{0.002\,5}}\right\}$$

$$=\Phi\left(\frac{0.10}{0.05}\right)=\Phi(2)=0.977\,2. \qquad \square$$

协方差与相关系数

- 称$E\{(X-E(X))(Y-E(Y))\}$为随机变量X与Y的协方差，记为Cov(X,Y),即

$$\text{Cov(X,Y)}= E\{(X-E(X))(Y-E(Y))\}=\text{E(XY)-E(X)E(Y)}$$

- 称$\rho_{XY}=\frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$为随机变量X与Y的相关系数
- 相关系数用于衡量两个随机变量之间的线性相关性
- 当X与Y相互独立时，
- $\text{Cov(X,Y)}= E\{(X-E(X))(Y-E(Y))\}=\text{E\{XY-XE(Y)-YE(X)+E(X)E(Y)\}}$

$$=\text{E(XY)-E(X)E(Y)-E(X)E(Y)+E(X)E(Y)}$$

$$=\text{E(X)E(Y)-E(X)E(Y)}$$

$$=0$$

- 此时，X与Y的相关系数也为0,称X与Y不相关

协方差的性质

- 1. Cov(aX,bY)=abCov(X,Y)，其中，a，b为常数
- 2. Cov(X1+X2,Y)=Cov(X1,Y)+Cov(X2,Y)
- Cov(X1+X2,Y1+Y2)=Cov(X1,Y1)+Cov(X1,Y2)+Cov(X2,Y1)+Cov(X2,Y2)
- 3. 方差是特殊的协方差：Cov(X,X)=D(X)
- 证明：
- 1. Cov(aX,bY)=E(abXY)-E(aX)E(bY)=abE(XY)-abE(X)E(Y)=abCov(X,Y)
- 2. Cov(X1+X2,Y)=E[(X1+X2)Y]-E(X1+X2)E(Y)=E(X1*Y)+E(X2*Y)-[E(X1)E(Y)-E(X2)E(Y)]=Cov(X1,Y)+Cov(X2,Y)
- 3. Cov(X,X)=E(X^2)-[E(X)]^2=D(X)

相关系数的性质

- 1. $|\rho_{XY}| \leq 1$
- 当$-1 < \rho_{XY} < 0$时，称X与Y成负相关；当$0 < \rho_{XY} < 1$时，称X与Y成正相关
- 2. $|\rho_{XY}| = 1$当且仅当存在常数a,b使$P\{Y = a + bX\} = 1$成立。此时称X与Y完全线性相关

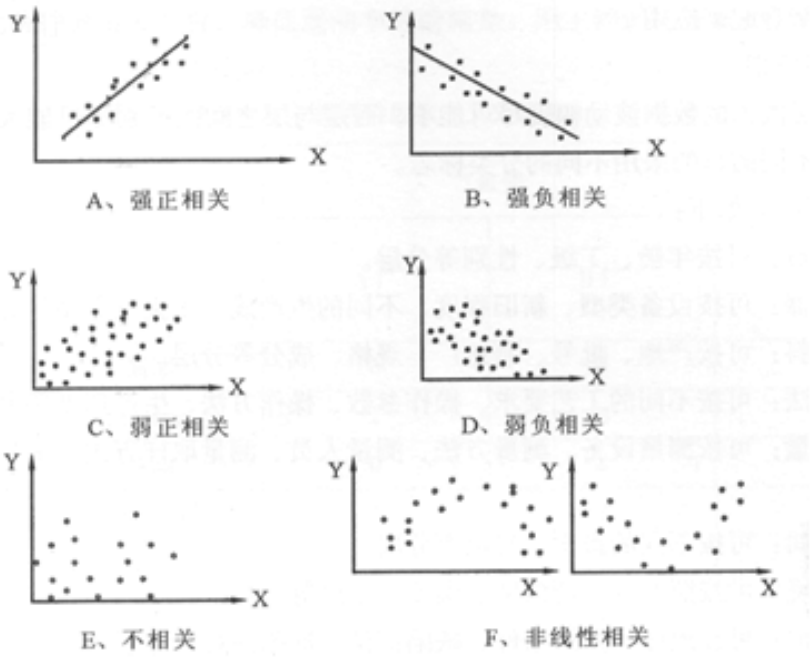


图 8-21　不同类型散布图

不相关与相互独立

- 当$\rho_{XY}=0$时（即Cov(X,Y)=0)我们称X与Y不相关

设（X，Y）的分布为：

Y \ X	-1	0	1
-1	1/8	1/8	1/8
0	1/8	0	1/8
1	1/8	1/8	1/8

容易验证 $\mathrm{cov}(X,Y)=E(XY)-E(X)E(Y)=0 \Rightarrow X,Y$不相关

$P_{ij}\neq P_iP_j, i,j=1,2,3 \Rightarrow X,Y$不独立.

例子

例 1 设(X,Y)的分布律为

Y \ X	-2	-1	1	2	$P\{Y=i\}$
1	0	1/4	1/4	0	1/2
4	1/4	0	0	1/4	1/2
$P\{X=i\}$	1/4	1/4	1/4	1/4	1

易知 $E(X)=0, E(Y)=5/2, E(XY)=0$，于是 $\rho_{XY}=0$，X,Y 不相关. 这表示 X,Y 不存在线性关系. 但，$P\{X=-2,Y=1\}=0 \neq P\{X=-2\}P\{Y=1\}$，知 X,Y 不是相互独立的. 事实上，X 和 Y 具有关系：$Y=X^2$，Y 的值完全可由 X 的值所确定. □

矩

定义 设 X 和 Y 是随机变量，若

$$E(X^k), \quad k=1,2,\cdots$$

存在，称它为 X 的 k **阶原点矩**，简称 k **阶矩**.

若

$$E\{[X-E(X)]^k\}, k=2,3,\cdots$$

存在，称它为 X 的 k **阶中心矩**.

若

$$E(X^kY^l), \quad k,l=1,2,\cdots$$

存在，称它为 X 和 Y 的 $k+l$ **阶混合矩**.

若

$$E\{[X-E(X)]^k[Y-E(Y)]^l\}, \quad k,l=1,2,\cdots$$

存在，称它为 X 和 Y 的 $k+l$ **阶混合中心矩**.

显然，X 的数学期望 $E(X)$ 是 X 的一阶原点矩，方差 $D(X)$ 是 X 的二阶中心矩，协方差 $\mathrm{Cov}(X,Y)$ 是 X 和 Y 的二阶混合中心矩.

协方差矩阵

- 对于n维随机变量X=(X1,X2,……,Xn),记cij=Cov(Xi,Xj),称矩阵

$$C = \begin{pmatrix} c11 & \cdots & c1n \\ \vdots & \ddots & \vdots \\ cn1 & \cdots & cnn \end{pmatrix}$$

- 为X的协方差矩阵。
- 对于二维随机变量(X,Y),(X,Y)的协方差矩阵为

$$\begin{pmatrix} D(X) & Cov(X,Y) \\ Cov(X,Y) & D(Y) \end{pmatrix}$$

- 协方差矩阵都是对称矩阵

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间