第十一周 正态分布专题

11.1 正态分布的相关与独立

二元正态分布的两个重要性质：

（1） 二元正态分布的边缘分布为一元正态分布。但是逆命题不成立，即边缘密度均为正态，联合分布未必是二元正态。

（2） 如果$(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，则

X,Y 相互独立 $\Leftrightarrow$ X,Y 不相关， 即 $Cov(X,Y)=0$ 或 $\rho=0$。

注：对一般随机变量 X,Y， X,Y 相互独立可以推出 X,Y 不相关。但是 X,Y 不相关则不能推出 X,Y 一定相互独立。

二元正态分布从不相关推出独立的性质是很容易验证：

$$\frac{1}{2\pi\sigma_1\sigma_2}e^{\frac{-1}{2}\left[\frac{(x-\mu_1)^2}{\sigma_1^2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right]}=\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}\cdot\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}$$

****

定理：设随机变量$(X,Y)\sim N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$，则$(a_1X+b_1Y,a_2X+b_2Y)$也服从二元正态分布。

计算$aX+bY$的分布参数

$$E(aX+bY)=aE(X)+bE(Y)=a\mu_1+b\mu_2$$

$$Var(aX+bY)=Var(aX)+Var(bY)+2Cov(aX,bY)=a^2\sigma_1^2+b^2\sigma_2^2+2ab\rho\sigma_1\sigma_2$$

$$aX+bY\sim N(a\mu_1+b\mu_2,a^2\sigma_1^2+b^2\sigma_2^2+2ab\rho\sigma_1\sigma_2)。$$

****

例 11.1.1 (X,Y)服从二维正态分布，X,Y 都服从$N(0,\sigma^2)$，X,Y 的相关系数为0.6。如果 $aX-Y$ 和 $X+Y$ 相互独立，试求常数a的值。

解： $(aX-Y,X+Y)$ 服从二维正态分布，所以 $aX-Y,X+Y$ 独立当且仅当它们不相关，即 $Cov(aX-Y,X+Y)=0$ 。

$$Cov(aX-Y,X+Y)=aVar(X)-Var(Y)+(a-1)Cov(X,Y)$$

$$=(a-1)\sigma^2+0.6(a-1)\sigma^2=0$$

故 $a=1$ 。

特别注意

因为两个正态分布的随机变量的联合分布未必是二元正态分布，所以两个正态分布随机变量 X,Y 不相关不能推出 X,Y 一定相互独立。只有 X,Y 的联合分布为二元正态分布时，才有 X,Y 不相关与独立的等价关系。

下面这段话摘自维基百科：

It is sometimes mistakenly thought that one context in which uncorrelatedness implies independence is when the random variables involved are normally distributed. However, this is incorrect if the variables are merely marginally normally distributed but not jointly normally distributed.

****

11.2 边缘密度均为正态，联合分布不是二元正态的例子

本节我们给出两个边缘密度均为正态，联合分布不是二元正态的例子。选取这两个例子，一方面的考虑是它们能够增进同学们对正态分布相关与独立性质的理解，另一方面，这两个例子也是全概率公式和全期望公式这两个重要概率计算工具的很好的应用实例。

边缘密度均为正态，联合分布不是二元正态的例子（1）

例 11.2.1 设随机变量 $X\sim N(0,1)$，$W\sim\begin{pmatrix}1 & -1\\ 1/2 & 1/2\end{pmatrix}$，且 X 与 W 相互独立，令 $Y=WX$，

验证：（1） $Y\sim N(0,1)$，（2） $Cov(X,Y)=0$，（3） X,Y 不独立。

解：（1） 计算 $Y=WX$ 的分布函数，利用全概率公式

$$F_Y(y)=P(Y\le y)=P(W=1)\cdot P(Y\le y|W=1)+P(W=-1)\cdot P(Y\le y|W=-1)$$

$$=\frac{1}{2}\cdot P(X\le y)+\frac{1}{2}\cdot P(X\ge -y)\qquad [\text{注：}P(X\ge -y)=P(X\le y)]$$

$$=\frac{1}{2}\cdot P(X\le y)+\frac{1}{2}\cdot P(X\le y)=P(X\le y)=\Phi(y)\Rightarrow Y\sim N(0,1)$$

（2）计算$Cov(X,Y)$，先计算$E(XY)$，利用全期望公式

$$E(XY)=E\left(E(XY|W)\right)$$

$$=P(W=1)\cdot E(XY|W=1)+P(W=-1)\cdot E(XY|W=-1)$$

$$=\frac{1}{2}\cdot E(X^2)+\frac{1}{2}\cdot E(-X^2)=0$$

$$Cov(X,Y)=E(XY)-E(X)E(Y)=0$$

（3） 举一个反例， 考虑 $f_{X,Y}(1,0)=0$， 而 $f_X(1)>0$， $f_Y(0)>0$

$f_{X,Y}(1,0)\neq f_X(1)\cdot f_Y(0)$ $\Rightarrow$ X,Y 不独立 。

****

边缘密度均为正态，联合分布不是二元正态的例子（2）

例 11.2.2 设随机变量 $X\sim N(0,1)$， 令$Y=\begin{cases}X, & if \quad |X|\ge c\\ -X, & if \quad |X|<c\end{cases}$， 验证

(1) $Y\sim N(0,1)$，(2) 存在常数c，使X,Y不相关，(3) 对任意$c>0$，X,Y不独立。

解：（1）计算Y的分布函数，利用全概率公式

$$F_Y(y)=P(Y\le y)=P(|X|\ge c)\cdot P(Y\le y||X|\ge c)+P(|X|<c)\cdot P(Y\le y||X|<c)$$

$$=P(|X|\ge c)\cdot\frac{P(Y\le y,|X|\ge c)}{P(|X|\ge c)}+P(|X|<c)\cdot\frac{P(Y\le y,|X|<c)}{P(|X|<c)}$$

$$=P(X\le y,|X|\ge c)+P(-X\le y,|X|<c)$$

利用X分布的对称性，以及$|X|<c$的对称性，有

$$P(-X\le y,|X|<c)=P(X\ge -y,|X|<c)=P(X\le y,|X|<c)$$

$$=P(X\le y,|X|\ge c)+P(X\le y,|X|<c)$$

$$= P(X \le y) = \Phi(y) \Rightarrow Y \sim N(0,1)$$

(2)存在常数c，使X,Y不相关，计算$Cov(X,Y)$，先计算$E(XY)$，利用全期望公式

$$E(XY) = E\left(E(XY|Y)\right) = P(|X| \ge c)\cdot E(XY||X| \ge c) + P(|X| < c)\cdot E(XY||X| < c)$$

$$= P(|X| \ge c)\cdot E(X^2||X| \ge c) + P(|X| < c)\cdot E(-X^2||X| < c)$$

$$= P(|X| \ge c)\cdot E(X^2||X| \ge c) + P(|X| < c)\cdot E(X^2||X| < c)$$

$$- 2P(|X| < c)\cdot E(X^2||X| < c)$$

$$= E(X^2) - 2\cdot P(|X| < c)\cdot E(X^2||X| < c)$$

$$= 1 - 2\cdot P(|X| < c)\cdot \int_{-c}^{c} x^2 \cdot \frac{\varphi(x)}{P(|X| < c)} dx = 1 - 4\cdot \int_0^c x^2 \cdot \varphi(x) dx$$

$$E(XY) = E\left(E(XY|Y)\right) = 1 - 4\cdot \int_0^c x^2 \cdot \varphi(x) dx$$

$$Cov(X,Y) = E(XY) - E(X)\cdot E(Y) = E(XY) = 0 \Rightarrow 1 - 4\cdot \int_0^c x^2 \cdot \varphi(x) dx = 0$$

$c = 0$时，$\int_0^c x^2 \cdot \varphi(x) dx = 0$，$E(XY) = 1 - 4\cdot \int_0^c x^2 \cdot \varphi(x) dx = 1 > 0$

$c \to \infty$时，$\int_0^c x^2 \cdot \varphi(x) dx = \frac{1}{2} E(X^2) = \frac{1}{2}$，$E(XY) = 1 - 4\cdot \int_0^c x^2 \cdot \varphi(x) dx = -1 < 0$

所以必然存在某一个$c > 0$，使得 $E(XY) = 1 - 4\cdot \int_0^c x^2 \cdot \varphi(x) dx = 0$，此时$c \approx 1.5383$。

注：随机变量 $X \sim N(0,1)$，$Y = \begin{cases} X, & if \quad |X| \ge c \\ -X, & if \quad |X| < c \end{cases}$

(3)对任意$c > 0$，X,Y不独立。

举一个反例，因为或者$Y = X$，或者$Y = -X$，

所以 $f_{X,Y}(1,0) = 0$，

而 $f_X(1) > 0$，$f_Y(0) > 0$，

$f_{X,Y}(1,0) \ne f_X(1)\cdot f_Y(0) \Rightarrow X,Y$不独立 。

****

11.3 二项分布的正态近似

参数 p 固定，n 很大时的二项分布随机变量的概率计算

先考虑一个简单的例子，假设随机地抛掷一枚均匀地硬币 100 次，求正面恰好出现 50 次的概率是多少？正面出现次数在 40 至 60 次之间的概率是多少？正面出现次数超过 80 次的概率是多少？

设随机地抛掷硬币 100 次，得到正面的次数为随机变量 X，$X \sim B\left(100, \frac{1}{2}\right)$。

$$P\left(X=50\right)=C_{100}^{50}\left(\frac{1}{2}\right)^{100}=\frac{100!}{50!50!}\cdot\frac{1}{2^{100}}$$

$$P\left(40\le X\le 60\right)=\sum_{k=40}^{60}C_{100}^{k}\left(\frac{1}{2}\right)^{100}=\sum_{k=40}^{60}\frac{100!}{k!\left(100-k\right)!}\left(\frac{1}{2}\right)^{100}$$

$$P\left(X\ge 80\right)=\sum_{k=80}^{100}C_{100}^{k}\left(\frac{1}{2}\right)^{100}=\sum_{k=80}^{100}\frac{100!}{k!\left(100-k\right)!}\left(\frac{1}{2}\right)^{100}$$

上述这些表达式形式复杂，它们的具体取值难以计算。其中主要的计算困难来自于阶乘项，早在十八世纪，数学家斯特林就给出了一个很方便的估计阶乘的公式。

****

阶乘的估计：斯特林（Stirling，1692-1770）公式

$$n!\sim\sqrt{2\pi n}\left(\frac{n}{e}\right)^{n},\quad \text{即}\quad \lim_{n\to\infty}\frac{n!}{\sqrt{2\pi\, n}\left(\frac{n}{e}\right)^{n}}=1$$

n	$n!$	$\sqrt{2\pi\, n}\left(\frac{n}{e}\right)^{n}$	相对误差
5	120	118.2	1.65%
10	3628800	3598695.6	0.83%
50	3.0414×10^{64}	3.0363×10^{64}	0.17%
100	9.3326×10^{157}	9.3248×10^{157}	0.08%

****

二项分布的正态近似（当 n 充分大时）

我们只考虑 $p=\frac{1}{2}$， n 为偶数的特殊情形。

设 $X\sim B\left(n,\frac{1}{2}\right)$， 令 $n=2m$，

$$a_k=P\left(X=m+k\right)=\frac{n!}{(m+k)!(m-k)!}\left(\frac{1}{2}\right)^n=\frac{n!}{m!m!}\left(\frac{1}{2}\right)^n\frac{m!m!}{(m+k)!(m-k)!}$$

由斯特林公式 $\frac{n!}{m!m!}\left(\frac{1}{2}\right)^n\sim\frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\sqrt{2\pi m}\left(\frac{m}{e}\right)^m\cdot\sqrt{2\pi m}\left(\frac{m}{e}\right)^m}\left(\frac{1}{2}\right)^n=\frac{1}{\sqrt{\pi m}}$

$X\sim B\left(100,\frac{1}{2}\right)$， $P\left(X=50\right)=C_{100}^{50}\left(\frac{1}{2}\right)^{100}=\frac{100!}{50!50!}\cdot\frac{1}{2^{100}}\approx\frac{1}{\sqrt{50\pi}}=0.0798$。

$$\frac{m!m!}{(m+k)!(m-k)!}=\frac{m!}{(m+k)!}\cdot\frac{m!}{(m-k)!}=\frac{m\cdot(m-1)\cdots(m-k+1)}{(m+1)(m+2)\cdots(m+k)}=\sum_{j=1}^{k}\frac{m-j+1}{m+j}$$

$\forall\ j=1,2,\cdots,k$， $\frac{m-j+1}{m+j}=\frac{m+\frac{1}{2}-\left(j-\frac{1}{2}\right)}{m+\frac{1}{2}+\left(j-\frac{1}{2}\right)}=\frac{1-\frac{j-1/2}{m+1/2}}{1+\frac{j-1/2}{m+1/2}}$

由于 $\frac{1}{2}\ln\frac{1+t}{1-t}=\frac{1}{2}\left(\ln(1+t)-\ln(1-t)\right)=t+\frac{t^3}{3}+\frac{t^5}{5}+\cdots$

$$\Rightarrow\frac{m\cdot(m-1)\cdots(m-k+1)}{(m+1)(m+2)\cdots(m+k)}=\prod_{j=1}^{k}\frac{1-\frac{j-1/2}{m+1/2}}{1+\frac{j-1/2}{m+1/2}}=e^{\sum_{j=1}^{k}\ln\frac{1-\frac{j-1/2}{m+1/2}}{1+\frac{j-1/2}{m+1/2}}}\sim e^{-2\sum_{j=1}^{k}\frac{j-1/2}{m+1/2}}\sim e^{\frac{-k^2}{m}}$$

由 $\frac{n!}{m!m!}\left(\frac{1}{2}\right)^n\sim\frac{1}{\sqrt{\pi m}}$， $\frac{m!m!}{(m+k)!(m-k)!}\sim e^{\frac{-k^2}{m}}$

可得： $P\left(X=m+k\right)=\frac{n!}{m!m!}\left(\frac{1}{2}\right)^n\frac{m!m!}{(m+k)!(m-k)!}\approx\frac{1}{\sqrt{\pi m}}\cdot e^{\frac{-k^2}{m}}$

$$\Rightarrow P\left(m-k\le X\le m+k\right)\approx\sum_{i=-k}^{k}\frac{1}{\sqrt{\pi m}}e^{\frac{-i^2}{m}}=\sum_{i=-k}^{k}\frac{1}{\sqrt{2\pi}}e^{\frac{-1}{2}\left(\frac{i}{\sqrt{m/2}}\right)^2}\cdot\frac{1}{\sqrt{m/2}}\approx\int_{\frac{-k}{\sqrt{m/2}}}^{\frac{k}{\sqrt{m/2}}}\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}dx$$

$$\Rightarrow P\left(m-k\le X\le m+k\right)\approx\int_{\frac{-k}{\sqrt{m/2}}}^{\frac{k}{\sqrt{m/2}}}\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}dx=\Phi\left(\frac{2k}{\sqrt{n}}\right)-\Phi\left(\frac{-2k}{\sqrt{n}}\right)$$

$$P\left(m-k\le X\le m+k\right)=P\left(\frac{-k}{\sqrt{n}/2}\le\frac{X-m}{\sqrt{n}/2}\le\frac{k}{\sqrt{n}/2}\right)\approx\Phi\left(\frac{2k}{\sqrt{n}}\right)-\Phi\left(\frac{-2k}{\sqrt{n}}\right)$$

$$\Rightarrow\frac{X-m}{\sqrt{n}/2}\sim N\left(0,1\right)\Rightarrow\frac{X-m}{\sqrt{n}/2}X\sim N\left(\frac{n}{2},\frac{n}{4}\right)$$

****

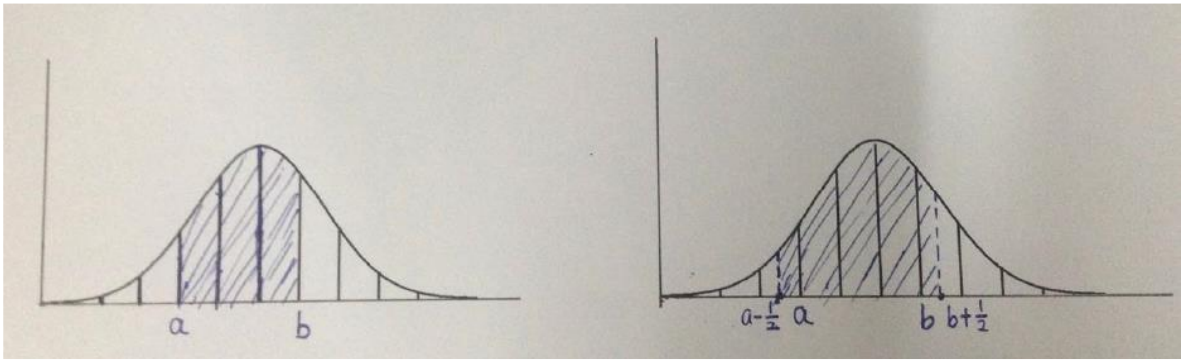
18 世纪法国数学家棣莫弗和拉普拉斯定理分别发现了二项分布和正态分布的近似关系，定理以两个人的名字命名。

棣莫弗—拉普拉斯定理 若 $X\sim B\left(n,p\right)$，则对任何两个常数 a 和 b，$-\infty<a<b<+\infty$，

$\lim\limits_{n\to\infty}P\left(a<\dfrac{X-np}{\sqrt{np\left(1-p\right)}}<b\right)=\int_a^b\dfrac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$；即 $n\to\infty$ 时，$X\sim N\left(np,np\left(1-p\right)\right)$。

实际的近似计算

$X\sim B\left(n,p\right)$，$Y\sim N\left(np,np\left(1-p\right)\right)$，$P\left(a\le X\le b\right)\approx P\left(a-\dfrac{1}{2}\le Y\le b+\dfrac{1}{2}\right)$



图中显示了二项分布的分布列和相应的正态分布的密度函数。X 大于等于 a 小于等于 b 的概率可以由正态分布密度函数从 a 到 b 进行积分计算，即左图的阴影部分。但使用这种方法，当估计 X=a 的概率时，近似值就会为 0。弥补这个缺陷的方法是对任何正整 k，X=k 的概率用相应正态分布在[k-1/2, k+1/2]内的概率来近似。这样 X 大于等

于 a 小于等于 b 的概率可以由正态分布密度函数从 a-1/2 到 b+1/2 进行积分计算，即右图阴影部分的面积，这样的近似，精度较好。

****

11.4 正态近似计算实例

例 11.4.1 随机地抛掷一枚均匀地硬币 100 次，求正面恰好出现 50 次的概率，正面出现次数在 40 至 60 次之间的概率，以及正面出现次数超过 80 次的概率。

解：设随机地抛掷硬币 100 次，得到正面的次数为随机变量 X，$X \sim B(100,1/2)$。

则 $X \sim N(50,5^2)$，令 $Y \sim N(50,5^2)$，则 $\dfrac{Y-50}{5} \sim N(0,1)$

$$P(X=50) \approx P(49.5 \le Y \le 50.5) = P\left(\frac{49.5-50}{5} \le \frac{Y-50}{5} \le \frac{50.5-50}{5}\right)$$

$$= \Phi(0.1) - \Phi(-0.1) = 0.0797$$

$$P(40 \le X \le 60) \approx P(39.5 \le Y \le 60.5) = P\left(\frac{39.5-50}{5} \le \frac{Y-50}{5} \le \frac{60.5-50}{5}\right) = 0.9643$$

$$P(X \ge 80) \approx P(Y \ge 79.5) = P\left(\frac{Y-50}{5} \ge \frac{79.5-50}{5}\right) = 1-\Phi(5.9) = 1.82\times 10^{-9}$$

****

例 11.4.2 设系统由 100 个相互独立的部件组成，运行时间每个部件损坏的概率为 0.1，至少有 85 个部件完好是系统才能正常工作，求系统正常工作的概率。

解：设正常工作部件的数目为随机变量 X，则 $X \sim B(100,0.9)$，

$P(\text{正常工作}) = P(X \ge 85)$；

X 近似服从正态分布 $N(90,9)$，令 $Y \sim N(90,3^2)$，则 $\dfrac{Y-90}{3} \sim N(0,1)$

$$P(\text{正常工作}) = P(X \ge 85) \approx P(Y \ge 84.5) = P\left(\frac{Y-90}{3} \ge \frac{84.5-90}{3}\right)$$

$$=1-\Phi\left(-\frac{5.5}{3}\right)=\Phi\left(\frac{5.5}{3}\right)=0.967$$

****

例 11.4.3 设系统由一些相互独立的部件组成，运行时间每个部件损坏的概率为 0.1，至少有 80%个部件完好是系统才能正常工作，问部件数 n 至少为多少才能使系统正常工作的概率不小于 0.95。$\Phi(1.645)=0.95$ 。

解：设正常工作部件的数目为随机变量 X，则 $X\sim B(n,0.9)$，

$X\sim N(0.9n,0.09n)$，　$\dfrac{X-0.9n}{0.3\sqrt{n}}\sim N(0,1)$。

$$P(X\geq 0.8n)=P\left(\frac{X-0.9n}{0.3\sqrt{n}}\geq\frac{0.8n-0.9n}{0.3\sqrt{n}}\right)\approx 1-\Phi\left(-\frac{0.1n}{0.3\sqrt{n}}\right)=\Phi\left(\frac{\sqrt{n}}{3}\right)\geq 0.95$$

$$\Rightarrow\quad \frac{\sqrt{n}}{3}\geq 1.645\quad\Rightarrow\quad n\geq 25\text{。}$$

****