2.4 随机变量的函数的概率分布

在理论和应用上,经常碰到这种情况:已知某个或某些随机变

量 $X_1,\cdots,X_n$ 的分布，现另有一些随机变量 $Y_1,\cdots,Y_m$，它们都是 $X_1,\cdots,X_m$ 的函数：

$$Y_i = g_i(X_1,\cdots,X_n),i = 1,\cdots,m \tag{4.1}$$

要求$(Y_1,\cdots,Y_m)$的概率分布．事实上我们已经考虑过这样的一个例子，即例 3.6.

在数理统计学中常碰到这个问题．在那里，$X_1,\cdots,X_n$ 是原始的观察或试验数据，$Y_1,\cdots,Y_m$ 则是为某种目的将这些数据“加工”而得的量，称为“统计量”．例如，$X_1,\cdots,X_n$ 可能是对某个未知量 a 作 n 次量测的结果，量测有误差，我们决定用 $X_1,\cdots,X_n$ 的算术平均值 $\overline{X}=(X_1+\cdots+X_n)/n$ 去估计未知量 a．$\overline{X}$ 就是 $X_1,\cdots,X_n$ 的函数．

2.4.1 离散型分布的情况

这种情况比较简单，故只须稍加解释．例如，变量 X 取 6 个值 $-2,-1,0,1,2,3$，其概率分别为 $1/12,3/12,3/12,2/12,1/12$ 和 $2/12$，而 $Y=X^3$．则 Y 取 $-8,-1,0,1,8,27$ 这 6 个值，它们没有相重的，故取这些值的概率，就仍如上述．

但若考虑 $Y=X^2$，则情况有所不同．相应于 X 的 6 个值的 Y 值分别为 $4,1,0,1,4,9$，其中有相重的．相重值的概率需要合并起来：

$$P(Y = 0) = P(X = 0) = 3/12$$
$$P(Y = 1) = P(X = 1) + P(X = -1) = 2/12 + 3/12 = 5/12$$
$$P(Y = 4) = P(X = 2) + P(X = -2) = 1/12 + 1/12 = 2/12$$
$$P(Y = 9) = P(X = 3) = 2/12$$

一般情况在原则上也一样：把 $Y=g(X_1,\cdots,X_n)$可以取的不同值找出来，把与某个值相应的全部$(X_1,\cdots,X_n)$值的概率加起来，即得 Y 取这个值的概率．当然，在实际做的时候，涉及的计算也可能并不简单．

例 4.1 设$(X_1,X_2,\cdots,X_n)$服从多项分布 $M(N;p_1,\cdots,$

$p_n)$, $n \geqslant 3$. 试求 $Y = X_1 + X_2$ 的分布.

Y 取值为 $0,1,\cdots,N$. 指定 k,有

$$P(Y = k) = \sum{}' \frac{N!}{k_1!k_2!k_3!\cdots k_n!} p_1^{k_1} p_2^{k_2} p_3^{k_3} \cdots p_n^{k_n}$$

这里 $\sum'$ 表示求和的范围为

$$k_i \text{ 为非负整数}, k_1 + k_2 = k, k_1 + \cdots + k_n = N$$

记 $p_i' = p/(1 - p_1 - p_2)$, $i = 3,\cdots,n$,则 $p_3' + \cdots + p_n' = 1$. 将上式写为

$$P(Y = k) = \frac{N!}{k!(N-k)!}(1 - p_1 - p_2)^{N-k} \sum{}'' \frac{k!}{k_1!k_2!} p_1^{k_1} p_2^{k_2}$$

$$\cdot \sum{}''' \frac{(N-k)!}{k_3!\cdots k_n!} p_3'^{k_3} \cdots p_n'^{k_n}$$

这里 $\sum''$ 求和的范围为:k_1, k_2 为非负整数,$k_1 + k_2 = k$. $\sum'''$ 求和的范围为;$k_3,\cdots,k_n$ 为非负整数,$k_3 + \cdots + k_n = N - k$. 由于 $p_3' + \cdots + p_n' = 1$. 由(2.4)式知 $\sum'''$ 这个和的值是1. $\sum''$ 这个和的值为 $(p_1 + p_2)^k$. 于是得到

$$P(Y = k) = \frac{N!}{k!(N-k)!}(p_1 + p_2)^k [1 - (p_1 + p_2)]^{N-k}$$

$$= b(k; N, p_1 + p_2)$$

即 Y 服从二项分布 $B(N, p_1 + p_2)$.

如果从概率意义的角度去考虑,这个结果不用计算就可以知道:在定义多项分布时有 n 个事件 $A_1, A_2, A_3, \cdots, A_n$. $X_1, X_2, X_3, \cdots, X_n$ 分别是它们在 N 次试验中发生的次数. 现若记 $A = A_1 + A_2$,则事件 $A, A_3, \cdots, A_n$ 仍构成一个完备事件群,其概率分别为 $p_1 + p_2, p_3, \cdots, p_n$. 记 $Y = X_1 + X_2$,则 $(Y, X_3, \cdots, X_n)$ 构成多项分布 $M(N; p_1 + p_2, p_3, \cdots, p_n)$,而 Y 成为这个多项分布的一个边缘分布. 于是按例 2.7 即得出上述结论.

这就是我们前面几个地方曾提及的概率思维. 概率论中有不少结果可以用纯分析方法证明,但如利用概率思维,有时证明可以简化,学习概率论的一个要素在于锻炼这种概率思维.

例 4.2 设 X_1 和 X_2 独立,分别服从二项分布 $B(n_1,p)$和 $B(n_2,p)$(注意 p 是公共的),求 $Y=X_1+X_2$ 的分布.

Y 之可能值为 $0,1,\cdots,n_1+n_2$. 固定 k 于上述范围内,由独立性假定,有

$$\begin{aligned}P(Y=k) &= \sum{}' P(X_1=k_1,X_2=k_2)\\ &= \sum{}' \binom{n_1}{k_1} p^{k_1}(1-p)^{n_1-k_1}\binom{n_2}{k_2} p^{k_2}(1-p)^{n_2-k_2}\\ &= \sum{}' \binom{n_1}{k_1}\binom{n_2}{k_2} p^{k}(1-p)^{n_1+n_2-k}\end{aligned}$$

此处 $\sum{}'$ 求和的范围为:k_1,k_2 为非负整数,$k_1+k_2=k$. 按第一章公式(2.5),得 $\sum{}'\binom{n_1}{k_1}\binom{n_2}{k_2}=\binom{n_1+n_2}{k}$,于是

$$P(Y=k)=\binom{n_1+n_2}{k}p^k(1-p)^{n_1+n_2-k}=b(k;n_1+n_2,p)$$

即 Y 服从二项分布 $B(n_1+n_2,p)$. 这个结果很容易推广到多个的情形:若 $X_i\sim B(n_i,p)$, $i=1,\cdots,m$,而 $X_1,\cdots,X_m$ 独立,则 $X_1+\cdots+X_m\sim B(n_1+\cdots+n_m,p)$. 证明不难用归纳法作出,细节留给读者.

上述结论如用"概率思维",则不证自明:按二项分布的定义,若 $X\sim B(n,p)$,则 X 是在 n 次独立试验中事件 A 出现的次数,而在每次试验中 A 的概率保持为 p. 令 X_i 是在 n_i 次试验中 A 出现的次数,每次试验 A 出现的概率为 p. 故 $Y=X_1+\cdots+X_m$ 是在 $n_1+\cdots+n_m$ 次独立试验中,A 出现的次数,而在每次试验中 A 出现的概率保持为 p. 故按定义即得 $Y\sim B(n_1+\cdots+n_m,p)$.

例 4.3 设 X_1,X_2 独立,分别服从波哇松分布 $P(\lambda_1)$和 $P(\lambda_2)$(见例 1.2). 证明:$Y=X_1+X_2$ 服从波哇松分布 $P(\lambda_1+\lambda_2)$.

Y 的可能值仍为一切非负整数. 固定这样一个 k,则由独立性假定及波哇松分布的形式(1.7),有

$$
\begin{aligned}
P(Y=k) &= \sum{}' P(X_1=k_1, X_2=k_2) \\
&= \sum{}' P(X_1=k_1)P(X_2=k_2) \\
&= \sum{}' \mathrm{e}^{-\lambda_1}\lambda_1^{k_1}/k_1! \cdot \mathrm{e}^{-\lambda_2}\lambda_2^{k_2}/k_2! \\
&= \mathrm{e}^{-(\lambda_1+\lambda_2)}/k! \sum{}' \frac{k!}{k_1!k_2!}\lambda_1^{k_1}\lambda_2^{k_2}
\end{aligned}
$$

这里 $\sum{}'$ 的求和范围与上例相同,因而这个和等于$(\lambda_1+\lambda_2)^k$.故

$$P(Y=k)=\mathrm{e}^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^k/k!$$

因而证明了所要的结果.这结果也自然地可推广到多个的情形.

在例 1.2 后面我们对波哇松分布通过二项分布而产生的过程作了一个解释,利用这个解释的架构,不须计算即可容易看出这个结论.我们留给读者自己去完成.这样解释的目的,倒不在于为了避免计算(就本例而言,计算很简单,可能比通过上述解释还简便些),而是它使人了解为什么会有这个结果(前面几个例子也如此).形式的计算使人相信结果是对的,但不能提供直观上的启发性.

2.4.2 连续型分布的情况:一般讨论

本节的其余部分将讨论更有兴趣的连续型情况.这一段对处理这种问题的一般方法作些介绍,然后在 2.4.3,2.4.4 两段中,分别对两个在数理统计学上重要的情况专门进行讨论,并由此引出在数理统计学上几个重要的概率分布.

先考虑一个变量的情况.设 X 有密度函数 $f(x)$.设 $Y=g(x)$,g是一个严格上升的函数,即当 $x_1<x_2$ 时,必有 $g(x_1)<g(x_2)$.又设 g 的导数g'存在.由于 g 的严格上升性,其反函数 $X=h(Y)$存在且 h 的导数h'也存在.

任取实数 y.因 g 严格上升,有

$$P(Y\leqslant y)=P(g(X))\leqslant y)=P(X\leqslant h(y))=\int_{-\infty}^{h(y)}f(t)\mathrm{d}t$$

Y 的密度函数$l(y)$,即是这个表达式对 y 求导数(见定义 1.3).

有

$$l(y)=f(h(y))h'(y) \tag{4.2}$$

如果 $Y=g(X)$ 而 g 是严格下降，则 $\{g(X)\leqslant y\}$ 相当于 $\{X\geqslant h(Y)\}$. 于是

$$P(Y\leqslant y)=P(g(X)\leqslant y)=P(X\geqslant h(y))=\int_{h(y)}^{\infty}f(t)\mathrm{d}t$$

对 y 求导数，得 Y 的密度函数

$$l(y)=-f(h(y))h'(y) \tag{4.3}$$

因为当 g 严格下降时，其反函数 h 也严格下降，故 $h'(y)<0$. 这样 $l(y)$ 仍为非负的. 总结(4.2),(4.3)两式，得知在 g 严格单调(上升下降都可以)的情况下，总有 $g(X)$ 的密度函数 $l(y)$ 为

$$l(y)=f(h(y))\,|h'(y)| \tag{4.4}$$

例 4.4 $Y=aX+b$, $a\neq0$. 反函数为 $X=(Y-b)/a$. 由(4.4)得出：$aX+b$ 的密度函数为

$$l(y)=f((y-b)/a)/|a| \tag{4.5}$$

若 X 有正态分布 $N(\mu,\sigma^2)$，则据正态密度函数的表达式(1.14)和公式(4.5)，易算出 $aX+b$ 服从正态分布 $N(a\mu+b,a^2\sigma^2)$. 特别，当 $Y=(X-\mu)/\sigma$ 时，有 $Y\sim N(0,1)$. 这一点在例 1.6 中已指出过了.

当 $Y=g(X)$ 而 g 不为严格单调时，情况复杂一些，但并无原则困难. 我们不去考虑一般情况，而只注意一个特例 $Y=X^2$. 仍以 f 记 X 的概率密度. 因 Y 非负，有 $P(Y\leqslant y)=0$ 当 $y\leqslant0$. 若 $y>0$，则有

$$\begin{aligned}P(Y\leqslant y)&=P(X^2\leqslant y)=P(-\sqrt{y}\leqslant X\leqslant\sqrt{y})\\&=\int_{-\sqrt{y}}^{\sqrt{y}}f(t)\mathrm{d}t\end{aligned}$$

对 y 求导数，得 Y 的密度函数 $l(y)$ 为

$$l(y)=\frac{1}{2}y^{-1/2}[f(\sqrt{y})+f(-\sqrt{y})],\text{当 } y>0$$

而当 $y\leqslant0$ 时 $l(y)=0$. 下面的特例很重要.

例 4.5 若 $X \sim N(0,1)$，试求 $Y=X^2$ 的密度函数.

以 $f(x)=(\sqrt{2\pi})^{-1}e^{-x^2/2}$代入上式，得

$$l(y)=(\sqrt{2\pi y})^{-1}e^{-y/2}, \text{当 } y>0. \ l(y)=0 \text{ 当 } y \leqslant 0 \tag{4.6}$$

现在考虑多个变量的函数的情况，以两个为例. 设(X_1,X_2)的密度函数为 $f(x_1,x_2)$，Y_1,Y_2 都是(X_1,X_2)的函数：

$$Y_1=g_1(X_1,X_2), Y_2=g_2(X_1,X_2) \tag{4.7}$$

要求(Y_1,Y_2)的概率密度函数 $l(y_1,y_2)$. 在此，我们要假定(4.7)是(X_1,X_2)到(Y_1,Y_2)的一一对应变换，因而有逆变换

$$X_1=h_1(Y_1,Y_2), X_2=h_2(Y_1,Y_2) \tag{4.8}$$

又假定 g_1,g_2 都有一阶连续偏导数. 这时，逆变换(4.8)的函数 h_1,h_2 也有一阶连续偏导数，且在一一对应变换的假定下，贾可比行列式

$$J(y_1,y_2)=\begin{vmatrix} \partial h_1/\partial y_1 & \partial h_1/\partial y_2 \\ \partial h_2/\partial y_1 & \partial h_2/\partial y_2 \end{vmatrix} \tag{4.9}$$

不为 0.

现在我们在(Y_1,Y_2)的平面上任取一个区域 A. 在变换(4.8)之下，这区域变到(X_1,X_2)平面上的区域 B. 就是说，事件$\{(Y_1,Y_2)\in A\}$等于事件$\{(X_1,X_2)\in B\}$. 考虑到 f 是(X_1,X_2)的密度函数，有

$$P((Y_1,Y_2)\in A)=P((X_1,X_2)\in B)=\iint_B f(x_1,x_2)dx_1dx_2$$

使用重积分变数代换的公式，在变换(4.8)之下，上式最右端一项的重积分变换为

$$P((Y_1,Y_2)\in A)=\iint_A f(h_1(y_1,y_2),h_2(y_1,y_2)) \cdot |J(y_1,y_2)|dy_1dy_2 \tag{4.10}$$

此式对(Y_1,Y_2)平面上任何区域 A 都成立. 于是，按定义 2.2(见(2.5)式))，即得(Y_1,Y_2)的密度函数为

$$l(y_1,y_2)=f(h_1(y_1,y_2),h_2(y_1,y_2))\,|J(y_1,y_2)| \tag{4.11}$$

一个重要的特例是线性变换

$$Y_1=a_{11}X_1+a_{12}X_2,Y_2=a_{21}Z_1+a_{22}X_2 \tag{4.12}$$

假定变换的行列式 $a_{11}a_{22}-a_{12}a_{21}\neq 0$,则逆变换(4.8)存在且仍为线性变换:

$$X_1=b_{11}Y_1+b_{12}Y_2,X_2=b_{21}Y_1+b_{22}Y_2 \tag{4.13}$$

此变换的贾可比行列式为常数:

$$J(y_1,y_2)=J=b_{11}b_{22}-b_{12}b_{21}=(a_{11}a_{22}-a_{12}a_{21})^{-1}$$

按(4.11)式,得出(Y_1,Y_2)的密度函数为

$$l(y_1,y_2)=f(b_{11}y_1+b_{12}y_2,b_{21}y_1+b_{22}y_2)\,|b_{11}b_{22}-b_{12}b_{21}| \tag{4.14}$$

例 4.6 再回过头来考虑例 3.6.为与此处记号一致,把该例中的 R 和 Θ 分别记为 Y_1,Y_2,这时逆变换(4.8)为

$$X_1=Y_1\cos Y_2,X_2=Y_1\sin Y_2$$

贾可比行列式为

$$J(y_1,y_2)=\begin{vmatrix}\cos y_2 & -y_1\sin y_2\\ \sin y_2 & y_1\cos y_2\end{vmatrix}=y_1$$

因为(X_1,X_2)的密度函数为

$$f(x_1,x_2)=\frac{1}{2\pi}\exp\left[-\frac{1}{2}(x_1^2+x_2^2)\right]$$

而 $x_1^2+x_2^2=y_1^2\cos^2 y_2+y_1^2\sin^2 y_2=y_1^2$,由公式(4.11),得$(Y_1,Y_2)$的概率密度函数为$\frac{1}{2\pi}e^{-y_1^2/2}y_1$.变量范围为 $0\leqslant y_1<\infty,0\leqslant y_2<2\pi$.在这个范围之外为 0.这与例 3.6 中求出的一致.

本例还提醒了我们一点:必须注意变换以后变量的范围.光从公式(4.11)上有时并不能看清这一点.在本例中,因为(Y_1,Y_2)是点的极坐标,其范围易于判定,在有些例子中,则需经过一定的判断.看下面的例子.

例 4.7 设 X_1,X_2 独立,都服从指数分布(1.20),其中 $\lambda=1$.

而设 $Y_1=X_1+X_2, Y_2=X_1-X_2$,求(Y_1,Y_2)的密度函数.

用公式(4.11)不难算出密度函数为 $l(y_1,y_2)=\frac{1}{2}e^{-y_1}$.问题在于:这个表达式只在一定范围 B 内有效,在 B 外为0.B 是什么?这就要考虑到(X_1,X_2)只在第一象限 A 内大于0.A 的两条边,即两轴的正半部,分别相应于(Y_1,Y_2)平面上的直线 $Y_1=Y_2$ 和 $Y_1=-Y_2$(见图2.10).另外,$Y_1=X_1+X_2$ 必大于0,Y_1 必大于 Y_2.故(Y_1,Y_2)只能落在上述两条直线所夹出的包含 Y_1 正半轴的那部分,即图2.10中标示的 B.

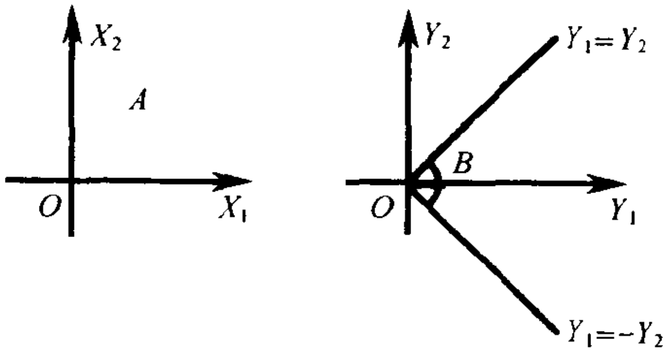

图2.10

有时,我们所要求的只是一个函数

$$Y_1=g_1(X_1,X_2)$$

的分布.一个办法是对任何 y 找出$\{Y_1\leqslant y\}$在(X_1,X_2)平面上对应的区域$\{g_1(X_1,X_2)\leqslant y\}$,记为 A_y.然后由 $P(Y_1\leqslant y)=\iint_{A_y} f(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2$ 找出 Y_1 的分布.另一个办法是配上另一个函数 $Y_2=g_2(X_1,X_2)$,使(X_1,X_2)到(Y_1,Y_2)成一一对应变换.然后按(4.11)找出(Y_1,Y_2)的联合密度函数 $l(y_1,y_2)$.最后,Y_1 的密度函数由公式 $\int_{-\infty}^{\infty} l(y_1,y_2)\mathrm{d}y_2$ 给出(见(2.9)).后面将给出使用这个方法的重要例子.

以上所说可完全平行地推广到 n 个变量的情形:设$(X_1,\cdots,X_n)$有密度函数 $f(x_1,\cdots,x_n)$,而

$$Y_i = g_i(X_1, \cdots, X_n), i = 1, \cdots, n$$

构成$(X_1, \cdots, X_n)$到$(Y_1, \cdots, Y_n)$的一一对应变换,其逆变换为

$$X_i = h_i(Y_1, \cdots, Y_n), i = 1, \cdots, n$$

此变换的贾可比行列式为

$$J(y_1, \cdots, y_n) = \begin{vmatrix} \partial h_1/\partial y_1 & \cdots & \partial h_1/\partial y_n \\ \cdots & \cdots & \cdots \\ \partial h_n/\partial y_1 & \cdots & \partial h_n/\partial y_n \end{vmatrix}$$

则$(Y_1, \cdots, Y_n)$的密度函数为

$$l(y_1, \cdots, y_n) = f(h_1(y_1, \cdots, y_n), \cdots, h_n(y_1, \cdots, y_n)) \cdot |J(y_1, \cdots, y_n)| \qquad (4.15)$$

2.4.3 随机变量和的密度函数

设(X_1, X_2)的联合密度函数为 $f(x_1, x_2)$,要求

$$Y = X_1 + X_2$$

的密度函数.

一个办法是考虑事件

$$\{Y \leqslant y\} = \{X_1 + X_2 \leqslant y\}$$

它所对应的(X_1, X_2)坐标平面上的集合 B,就是图 2.11 中所示的直线 $x_1 + x_2 = y$ 的下方那部分.按密度函数的定义有

$$P(Y \leqslant y) = P(X_1 + X_2 \leqslant y) = \iint_B f(x_1, x_2)\mathrm{d}x_1\mathrm{d}x_2$$

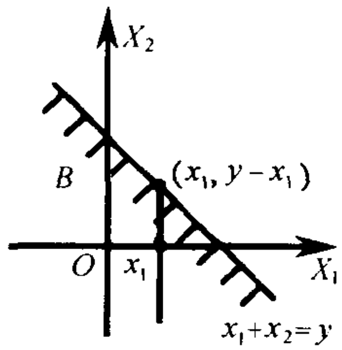

图 2.11

将重积分化为累积分,先固定 x_1 对 x_2 积分,积分范围为 $-\infty$ 到 $y - x_1$,如图所示.然后再对 x_1 从 $-\infty$ 到 ∞ 积分.结果得

$$P(Y \leqslant y) = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{y-x_1} f(x_1, x_2)\mathrm{d}x_2\right)\mathrm{d}x_1$$

对 y 求导数,即得 Y 的密度函数为

$$l(y) = \int_{-\infty}^{\infty} f(x_1, y - x_1)\mathrm{d}x_1$$

$$=\int_{-\infty}^{\infty} f(x, y-x)\mathrm{d}x \tag{4.16}$$

作变数代换 $t=y-x$(注意 y 是固定的),再把积分变量 t 换回到 x,也得到

$$l(y)=\int_{-\infty}^{\infty} f(y-x, x)\mathrm{d}x \tag{4.17}$$

如果 X_1, X_2 独立,则 $f(x_1, x_2)=f_1(x_1)f_2(x_2)$. 这时(4.16)和(4.17)有形式

$$\begin{aligned} l(y) &= \int_{-\infty}^{\infty} f_1(x)f_2(y-x)\mathrm{d}x \\ &= \int_{-\infty}^{\infty} f_1(y-x)f_2(x)\mathrm{d}x \end{aligned} \tag{4.18}$$

这个方法在数学上一点不足的地方是要通过在积分号下求导数. 这在理论上是有条件的. 另一个做法是配上另一个函数,例如 $Z=X_1$. 则

$$Y=X_1+X_2, Z=X_1$$

构成 (X_1, X_2) 到 (Y, Z) 的一一对应变换. 逆变换为

$$X_1=Z, X_2=Y-Z$$

贾可比行列式为 -1,绝对值为 1. 按公式(4.11),得 (Y, Z) 的联合密度函数为 $f(z, y-z)$. 再依公式(2.9),求得 Y 的密度函数 $l(y)$ 仍为(4.16)式.

例 4.8 设 X_1, X_2 独立,分别服从正态分布 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$. 求 $Y=X_1+X_2$ 的密度函数.

由假定,利用(4.18)的第一式,有

$$l(y)=\frac{1}{2\pi\sigma_1\sigma_2}\int_{-\infty}^{\infty}\exp\left[-\frac{1}{2}\left(\frac{(x-\mu_1)^2}{\sigma_1^2}+\frac{(y-x-\mu_2)^2}{\sigma_2^2}\right)\right]\mathrm{d}x \tag{4.19}$$

经过一些初等代数的运算,不难得到

$$\begin{aligned} &(x-\mu_1)^2/\sigma_1^2+(y-x-\mu_2)^2/\sigma_2^2 \\ &\quad=(\sigma_1^2+\sigma_2^2)^{-1}(y-\mu_1-\mu_2)^2+(ax-b)^2 \end{aligned}$$

其中

$$a=\sigma_1\sigma_2/\sqrt{\sigma_1^2+\sigma_2^2}$$

$$b=\frac{\sigma_1\sigma_2}{\sqrt{\sigma_1^2+\sigma_2^2}}(\mu_1\sigma_1^{-2}+(y-\mu_2)\sigma_2^{-2})$$

代入(4.19),得

$$l(y)=(2\pi\sigma_1\sigma_2)^{-1}\exp\left[-\frac{(y-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right]\int_{-\infty}^{\infty}e^{-\frac{1}{2}(ax-b)^2}dx$$

注意 a,b 都与 x 无关,作变数代换 $t=ax+b$,并利用 $\int_{-\infty}^{\infty}e^{-t^2/2}dt=\sqrt{2\pi}$(见(1.15)式),即得

$$l(y)=\left(\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}\right)^{-1}\times$$

$$\exp\left[-\frac{1}{2}(y-\mu_1-\mu_2)^2/(\sigma_1^2+\sigma_2^2)\right] \quad (4.20)$$

这正是正态分布 $N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$的密度函数.由此可见,两个独立的正态变量的和仍服从正态分布,且有关的参数相加.

有趣的是,这个事实的逆命题也成立:如果 Y 服从正态分布,而 Y 表成两个独立随机变量 X_1,X_2 之和,则 X_1,X_2 必都服从正态分布.这个事实称为正态分布的“再生性”:一条蚯蚓砍成两段,仍各成一条蚯蚓,这称为蚯蚓的再生性;此处亦然:一个正态变量 Y 砍成独立的两段 $X_1,X_2(Y=X_1+X_2)$,各段 X_1,X_2 仍不失其正态性.这个深刻命题的证明超出了本书的范围.

不难证明:即使 X_1,X_2 不独立,只要其联合分布为二维正态 $N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$,则 $Y=X_1+X_2$ 仍为正态:$Y\sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2+2\rho\sigma_1\sigma_2)$.证明与本例相仿,细节留给读者.

本例直接推广到 n 个变量的情形:若 $X_1,\cdots,X_n$ 相互独立,分别服从正态分布 $N(\mu_1,\sigma_1^2),\cdots,N(\mu_n,\sigma_n^2)$,则 $X_1+\cdots+X_n$ 服从正态分布$N(\mu_1+\cdots+\mu_n,\sigma_1^2+\cdots+\sigma_n^2)$.

证明很容易.以三个变量的情形为例.记

$$Y=X_1+X_2+X_3=Z+X_3,Z=X_1+X_2$$

按本例结果有 $Z\sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$. 又按定下 3.3,知 Z 与 X_3 独立. 对 Z 和 X_3 应用本例,即得

$$Y = Z + X_3 \sim N(\mu_1+\mu_2+\mu_3,\sigma_1^2+\sigma_2^2+\sigma_3^2)$$

在介绍下面这个重要例子之前,我们先要引进两个重要的特殊函数:

Γ 函数(读作 Gamma 函数)$\Gamma(x)$:通过积分

$$\Gamma(x) = \int_0^{\infty} \mathrm{e}^{-t} t^{x-1} \mathrm{d}t, x > 0 \tag{4.21}$$

来定义. 此积分在 $x>0$ 时有意义.

β 函数(读作 Beta 函数)$\beta(x,y)$:通过积分

$$\beta(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1} \mathrm{d}t, x>0, y>0 \tag{4.22}$$

来定义. 此积分在 $x>0, y>0$ 时有意义.

直接算出 $\Gamma(1) = \int_0^{\infty} \mathrm{e}^{-t} \mathrm{d}t = 1$, 而在作变数代换 $t=u^2$ 后, 算出

$$\Gamma(1/2) = \int_0^{\infty} \mathrm{e}^{-t} t^{-1/2} \mathrm{d}t = \int_0^{\infty} \mathrm{e}^{-u^2} u^{-1}(2u\mathrm{d}u)$$
$$= 2\int_0^{\infty} \mathrm{e}^{-u^2} \mathrm{d}u = \int_{-\infty}^{\infty} \mathrm{e}^{-u^2} \mathrm{d}u$$

令 $u=v/\sqrt{2}$,并利用(1.15)式,得

$$\Gamma(1/2) = \frac{1}{\sqrt{2}} \int_{-\infty}^{\infty} \mathrm{e}^{-v^2/2} \mathrm{d}v = \frac{1}{\sqrt{2}} \sqrt{2\pi} = \sqrt{\pi}$$

Γ 函数有重要的递推公式:

$$\Gamma(x+1) = x\Gamma(x) \tag{4.23}$$

事实上, $\Gamma(x+1) = \int_0^{\infty} \mathrm{e}^{-t} t^x \mathrm{d}t$. 作分部积分,有

$$\int_0^{\infty} \mathrm{e}^{-t} t^x \mathrm{d}t = -\int_0^{\infty} t^x \mathrm{d}(\mathrm{e}^{-t}) = -t^x \mathrm{e}^{-t}\Big|_0^{\infty} + x\int_0^{\infty} \mathrm{e}^{-t} t^{x-1} \mathrm{d}t$$
$$= x\Gamma(x)$$

由算出的 $\Gamma(1)$ 和 $\Gamma(1/2)$, 可得出当 n 为正整数时, $\Gamma(n)$ 和

$\Gamma(n/2)$之值(后者当 n 为奇数时,否则 $n/2$ 为整数):

$$\Gamma(n)=(n-1)!,\Gamma(n/2)=1\cdot 3\cdot 5\cdots(n-2)2^{-(n-1)/2}\sqrt{\pi} \tag{4.24}$$

例如

$$\begin{aligned}\Gamma(4)&=\Gamma(3+1)=3\Gamma(3)=3\cdot 2\Gamma(2)=3\cdot 2\cdot 1\Gamma(1)\\&=3\cdot 2\cdot 1=3!\\ \Gamma(7/2)&=\Gamma(5/2+1)=(5/2)\Gamma(5/2)\\&=(5/2)(3/2)\Gamma(3/2)\\&=(5/2)(3/2)(1/2)\Gamma(1/2)=1\cdot 3\cdot 5\cdot 2^{-3}\sqrt{\pi}\end{aligned}$$

Γ 函数与 β 函数之间有重要的关系式:

$$\beta(x,y)=\Gamma(x)\Gamma(y)/\Gamma(x+y) \tag{4.25}$$

这个公式的证明见本章附录 A.

由 Γ 函数的定义易知:若 $n>0$,则函数

$$k_n(x)=\begin{cases}\dfrac{1}{\Gamma\left(\dfrac{n}{2}\right)2^{n/2}}e^{-x/2}x^{(n-2)/2}, & 当\ x>0\\ 0, & 当\ x\leqslant 0\end{cases} \tag{4.26}$$

是概率密度函数.实际上,由 $k_n(x)$的定义知它非负.又(作变数代换 $x=2t$)

$$\int_0^{\infty}e^{-x/2}x^{(n-2)/2}dx=2^{n/2}\int_0^{\infty}e^{-t}t^{(n-2)/2}dt=2^{n/2}\Gamma\left(\frac{n}{2}\right)$$

故知 $\int_{-\infty}^{\infty}k_n(x)dx=\int_0^{\infty}k_n(x)dx=1$. 因而证明了它是密度函数.这个密度函数在统计学上很重要且很有名,它称为"自由度 n 的皮尔逊卡方密度"(相应的分布则称为卡方分布),常记为 χ_n^2. K.皮尔逊是英国统计学家,现代统计学的奠基人之一.在本书第五章中将涉及他的工作.

例 4.9 若 $X_1,\cdots,X_n$ 相互独立,都服从正态分布 $N(0,1)$*,则 $Y=X_1^2+\cdots+X_n^2$ 服从自由度 n 的卡方分布 χ_n^2.

从例 4.5,并注意 $\Gamma(1/2)=\sqrt{\pi}$,看出本例的结果当 $n=1$ 时成立.于是可用归纳法,设此结果当 n 改为 $n-1$ 时成立.表 Y 为 $Z+X_n^2$,其中 $Z=X_1^2+\cdots+X_{n-1}^2$,则由归纳假设,知 Z 有密度函数 $k_{n-1}(x)$.由例 4.5 知 X_n^2 有密度函数 $k_1(x)$.再由定理 3.3,知 Z 与 X_n^2 独立.于是按公式(4.18)(用前一式),知 Y 的密度函数为

$$l(y)=\int_{-\infty}^{\infty}k_{n-1}(x)k_1(y-x)\mathrm{d}x=\int_0^y k_{n-1}(x)k_1(y-x)\mathrm{d}x$$

后一式是因为,$k_{n-1}(t)$ 和 $k_1(t)$ 都只在 $t>0$ 时才不为 0,故有效的积分区间为 $0\leqslant x\leqslant y$.以(4.26)中的表达式(n 分别改为 $n-1$ 和 1)代入上式,得

$$l(y)=\left(\Gamma\left(\frac{n-1}{2}\right)2^{\frac{n-1}{2}}\Gamma\left(\frac{1}{2}\right)2^{\frac{1}{2}}\right)^{-1}\mathrm{e}^{-y/2}\cdot\int_0^y x^{(n-3)/2}(y-x)^{-1/2}\mathrm{d}x \qquad (4.27)$$

在积分中作变数代换 $x=yt$,得

$$\begin{aligned}&\int_0^y x^{(n-3)/2}(y-x)^{-1/2}\mathrm{d}x\\&=y^{(n-2)/2}\int_0^1 t^{(n-3)/2}(1-t)^{-1/2}\mathrm{d}t\\&=y^{(n-2)/3}\beta\left(\frac{n-1}{2},\frac{1}{2}\right)\\&=y^{(n-2)/2}\Gamma\left(\frac{n-1}{2}\right)\Gamma\left(\frac{1}{2}\right)\Big/\Gamma\left(\frac{n}{2}\right)\end{aligned}$$

以此代入(4.27),即得 $l(y)=k_n(y)$.从而证明了本例结果对 n 也成立,这完成了归纳证明.

---

* 常把这说成 $X_1,\cdots,X_n$ 独立同分布并缩记为 iid.(independently identically distributed),并说 $X_1,\cdots,X_n$ 有公共分布 $N(0,1)$.注意不要混淆"公共"分布和"联合"分布.整个这假定可简记为:$X_1,\cdots,X_n$ iid,$\sim N(0,1)$.

本例也解释了在定义卡方分布时提到的“自由度 n”这个名词.因为 Y 表为 n 个独立变量 $X_1,\cdots,X_n$ 的平方和,每个变量 X_i 都能随意变化,可以说它有一个自由度,共有 n 个变量,因此有 n 个自由度.当然这个解释只在 n 为正整数时才有效(注意 $k_n(x)$ 的定义中并不必须限制 n 为正整数,只要 $n>0$ 就行).实际上,自由度这个名词通常也只用在 n 为整数时.

卡方分布有如下的重要性质:

1. 设 X_1,X_2 独立,$X_1\sim\chi_m^2,X_2\sim\chi_n^2$,则 $X_1+X_2\sim\chi_{m+n}^2$.

证明可以直接利用和的密度公式(4.18)得到.更简便的是从卡方变量的表达式出发,设 $Y_1,\cdots,Y_{m+n}$ 独立且都有分布 $N(0,1)$.令 $X_1=Y_1^2+\cdots+Y_m^2,X_2=Y_{m+1}^2+\cdots+Y_{m+n}^2$.按本例,有

$$X_1\sim\chi_m^2,X_2\sim\chi_n^2$$

而

$$X_1+X_2=Y_1^2+\cdots+Y_{m+n}^2$$

为 $m+n$ 个标准正态变量的平方和.按本例其分布为 χ_{m+n}^2,明所欲证.

2. 若 $X_1,\cdots,X_n$ 独立,且都服从指数分布(1.20),则

$$X=2\lambda(X_1+\cdots+X_n)\sim\chi_{2n}^2$$

首先,由 X_i 的密度函数为(1.20),知 $2\lambda X_i$ 的密度函数为 $\dfrac{1}{2}\mathrm{e}^{-x/2}$(当 $x>0.x\leqslant0$ 时为 0).但在(4.26)中令 $n=2$,可知这正好是 χ_2^2 的密度函数,因此 $2\lambda X_i\sim\chi_2^2$.再因 $X_1,\cdots,X_n$ 独立,利用刚才证明的性质,即得所要的结果.

2.4.4 随机变量商的密度函数

设(X_1,X_2)有密度函数 $f(x_1,x_2)$,$Y=X_2/X_1$.要求 Y 的密度函数.为简单计,限制 X_1 只取正值的情况.

事件$\{Y\leqslant y\}=\{X_2/X_1\leqslant y\}$可写为$\{X_2\leqslant X_1y\}$,因为 $X_1>0$.这相应于图 2.12 中所标出的区域 B.通过化重积分为累积分,

得到

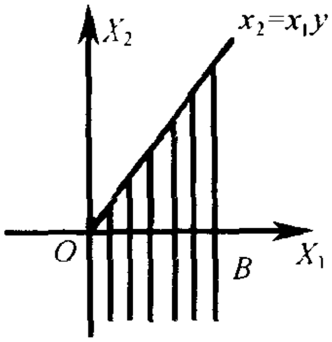

图 2.12

$$P(Y\leqslant y)=\iint_B f(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2$$

$$=\int_0^{\infty}\left[\int_{-\infty}^{x_1y}f(x_1,x_2)\mathrm{d}x_2\right]\mathrm{d}x_1$$

对 y 求导,得 Y 的密度函数为

$$l(y)=\int_0^{\infty}x_1f(x_1,x_1y)\mathrm{d}x_1 \tag{4.28}$$

若 X_1,X_2 独立,则 $f(x_1,x_2)=f_1(x_1)\cdot f_2(x_2)$,而上式成为

$$l(y)=\int_0^{\infty}x_1f_1(x_1)f_2(x_1y)\mathrm{d}x_1 \tag{4.29}$$

(4.28)式也可以通过添加一个变换 $Z=X_1$,再运用公式(4.11)和(2.9)得到,建议读者自己去完成.这个做法不须在积分号下求导数.

下面考察两个在统计学上十分重要的例子.

例 4.10 设 X_1,X_2 独立,$X_1\sim\chi_n^2$ 独立,$X_2\sim N(0,1)$,而 $Y=X_2/\sqrt{X_1/n}$.求 Y 的密度函数.

记 $Z=\sqrt{X_1/n}$.先要求出 Z 的密度函数 $g(z)$.有

$$P(Z\leqslant z)=P(\sqrt{X_1/n}\leqslant z)=P(X_1\leqslant nz^2)$$

$$=\int_0^{nz^2}k_n(x)\mathrm{d}x$$

两边对 Z 求导,得 Z 的密度函数为

$$g(z)=2nzk_n(nz^2)$$

其次,以 $f_1(x_1)=2nx_1k_n(nx_1^2)$ 和 $f_2(x_2)=\sqrt{2\pi}^{-1}\mathrm{e}^{-x_2^2/2}$ 应用公式(4.29),得 Y 的密度函数,记之为 $t_n(y)$,等于

$$t_n(y)=\sqrt{2\pi}^{-1}(2^{n/2}\Gamma(n/2))^{-1}\int_0^{\infty}2nx_1^2\mathrm{e}^{-nx_1^2/2}(nx_1^2)^{(n-2)/2}$$

$$
\begin{aligned}
&\cdot \mathrm{e}^{-(x_1 y)^2/2}\mathrm{d}x_1\\
&=\sqrt{2\pi}^{-1}(2^{n/2}\Gamma(n/2))^{-1}2n^{n/2}\\
&\cdot\int_0^\infty x_1^n \exp\left[-\frac{1}{2}(nx_1^2+x_1^2y^2)\right]\mathrm{d}x_1
\end{aligned}
\tag{4.30}
$$

作变数代换 $x_1=\sqrt{2/(n+y^2)}\sqrt{t}$，上面的积分变为

$$
\begin{aligned}
&\frac{1}{2}\left(\frac{2}{n+y^2}\right)^{(n+1)/2}\int_0^\infty \mathrm{e}^{-t}t^{(n-1)/2}\mathrm{d}t\\
&=\frac{1}{2}\left(\frac{2}{n+y^2}\right)^{(n+1)/2}\Gamma\left(\frac{n+1}{2}\right)
\end{aligned}
$$

以此代入(4.30)，并略加整理，即得 $Y=X_2/\sqrt{X_1/n}$ 的密度函数为

$$
t_n(y)=\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}\left(1+\frac{y^2}{n}\right)^{-\frac{n+1}{2}} \tag{4.31}
$$

这个密度函数称为“自由度 n 的 t 分布”的密度函数，常简记为 $Y\sim t_n$. 这个分布是英国统计学家 W. 哥色特在 1908 年以“student”的笔名首次发表的. 它是数理统计学中最重要的分布之一，今后我们将见到这个分布在统计学上的许多应用.

这个密度函数关于原点对称，其图形与正态 $N(0,1)$ 的密度函数的图形相似，以后我们将见到(见第三章 3.4 节)，当自由度 n 很大时，t 分布确实接近于标准正态分布.

例 4.11 设 X_1,X_2 独立，$X_1\sim\chi_n^2$，$X_2\sim\chi_m^2$，而 $Y=m^{-1}X_2/n^{-1}X_1$. 求 Y 的密度函数.

因为 X_1,X_2 独立，故 $n^{-1}X_1$ 和 $m^{-1}X_2$ 也独立. 由 $X_1\sim\chi_n^2$ 和 $X_2\sim\chi_m^2$ 易求出 $n^{-1}X_1$ 和 $m^{-1}X_2$ 的密度函数分别为 $nk_n(nx_1)$ 和 $mk_m(mx_2)$. 以此代入(4.29)，得 Y 的密度函数，记之为 $f_{mn}(y)$(注意 m 在前，m 是分子 X_2 的自由度)，等于

$$
\begin{aligned}
f_{mn}(y)&=mn\int_0^\infty x_1k_n(nx_1)k_m(mx_1y)\mathrm{d}x_1\\
&=mn\left[2^{m/2}\Gamma\left(\frac{m}{2}\right)2^{n/2}\Gamma\left(\frac{n}{2}\right)\right]^{-1}
\end{aligned}
$$

$$\cdot \int_0^\infty x_1 e^{-nx_1/2}(nx_1)^{n/2-1}e^{-mx_1y/2}(mx_1y)^{m/2-1}dx_1$$

$$= \left[2^{(m+n)/2}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\right]^{-1} m^{m/2}n^{n/2}y^{m/2-1}$$

$$\cdot \int_0^\infty e^{(my+n)x_1/2}x_1{}^{(m+n)/2-1}dx_1$$

作变数代换 $t=(my+n)x_1/2$,上式的积分化为

$$2^{(m+n)/2}(my+n)^{-(m+n)/2}\int_0^\infty e^{-t}t^{(m+n)/2-1}dt$$

$$=2^{(m+n)/2}(my+n)^{-(m+n)/2}\Gamma\left(\frac{m+n}{2}\right)$$

以此代入上式,得

$$f_{mn}(y) = m^{m/2}n^{n/2}\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}y^{m/2-1}(my+n)^{-(m+n)/2},$$

$$y>0 \tag{4.32}$$

当 $y\leqslant 0$ 时 $f_{mn}(y)=0$,因为 Y 只取正值.

这个分布称为“自由度 m,n 的 F 分布”(注意分子的自由度在前).它也是数理统计学上的一个重要分布,有很多应用,常记为 F_{mn}:$Y\sim F_{mn}$.

人们有时把 χ^2,t 和 F 这三个分布合称为“统计上的三大分布”,就是因为它们在统计学中有广泛的应用.这些应用的相当大一部分根由,在于以下的几条重要性质.它们的证明可参见本章附录 B.

1° 设 $X_1,\cdots,X_n$ 独立同分布,有公共的正态分布 $N(\mu,\sigma^2)$.记 $\overline{X}=(X_1+\cdots+X_n)/n$,$S^2=\sum_{i=1}^n(X_i-\overline{X})^2/(n-1)$.则

$$(n-1)S^2/\sigma^2=\sum_{i=1}^n(X_i-\overline{X})^2/\sigma^2\sim\chi_{n-1}^2 \tag{4.33}$$

2° 设 $X_1,\cdots,X_n$ 的假定同 1°,则

$$\sqrt{n}(\overline{X}-\mu)/S\sim t_{n-1} \tag{4.34}$$

3° 设 $X_1,\cdots,X_n,Y_1,\cdots,Y_m$ 独立，X_i 各有分布 $N(\mu_1,\sigma_1^2)$，Y_j 各有分布 $N(\mu_2,\sigma_2^2)$，则

a. $$\left[\sum_{j=1}^{m}(Y_j-\overline{Y})^2/(\sigma_2^2(m-1))\right]\Big/\left[\sum_{i=1}^{n}(X_i-\overline{X})^2/(\sigma_1^2(n-1))\right] \sim F_{m-1,n-1} \tag{4.35}$$

b. 若 $\sigma_1^2=\sigma_2^2$，则

$$\sqrt{\frac{nm(n+m-2)}{n+m}}\left[(\overline{X}-\overline{Y})-(\mu_1-\mu_2)\right] \Big/\left[\sum_{i=1}^{n}(X_i-\overline{X})^2+\sum_{j=1}^{m}(Y_j-\overline{Y})^2\right]^{1/2} \sim t_{n+m-2} \tag{4.36}$$