TEMA 4

ALGUNOS MODELOS DE DISTRIBUCIONES DE PROBABILIDAD MULTIDIMENSIONALES

Distribución multinomial

La distribución multinomial es una distribución discreta multivariante y, como su nombre indica, es una generalización de la distribución binomial cuando el experimento aleatorio considerado no tiene sólo dos resultados posibles, éxito o fracaso, sino tres o más.

Modelo probabilístico

Consideremos un experimento aleatorio con $k+1$ posibles resultados o sucesos $A_1, A_2, \cdots A_{k+1}$ exhaustivos y mutuamente excluyentes, es decir, constituyen una partición del espacio muestral

$$\bigcup_{i=1}^{k+1} A_i = \Omega, \qquad A_i \cap A_j = \emptyset,\ \forall i \neq j.$$

Si denotamos por $p_1, p_2, \cdots, p_{k+1}$ las probabilidades de los diferentes sucesos, es decir

$$p_i = \mathrm{P}(A_i), \quad i = 1, 2, \cdots, k+1$$

se verifica que

$$\sum_{i=1}^{k+1} p_i = \sum_{i=1}^{k+1} \mathrm{P}(A_i) = \mathrm{P}\left(\bigcup_{i=1}^{k+1} A_i\right) = \mathrm{P}(\Omega) = 1$$

y, de aquí

$$p_{k+1} = 1 - \sum_{i=1}^{k} p_i.$$

Supongamos que realizamos n repeticiones independientes del experimento en las mismas condiciones y, por tanto, las probabilidades $p_i = \mathrm{P}(A_i),\ i = 1, 2, \cdots, k+1$ se mantienen constantes en las n repeticiones.

Si consideramos $x_1, x_2, \cdots, x_k$ enteros no negativos tales que $x_1 + x_2 + \cdots x_k \leq n$, entonces la probabilidad de que en n repeticiones del experimento ocurra exactamente x_i veces el suceso $A_i,\ i = 1, 2, \cdots, k$ y, por tanto $x_{k+1} = n - (x_1 + x_2 + \cdots x_k)$ veces el suceso A_{k+1} es

$$\frac{n!}{x_1!x_2!\ldots x_k!(n - \sum_{i=1}^k x_i)!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k} (1 - \sum_{i=1}^{k} p_i)^{n - \sum_{i=1}^k x_i}.$$

Si consideramos ahora $k+1$ variables aleatorias que representan X_i= Número de veces que ocurre el suceso A_i en las n repeticiones del experimento, $i = 1, 2, \cdots, k+1$ (claramente X_{k+1} está completamente determinada a partir de las anteriores ya que $X_{k+1} = n - \sum_{i=1}^{k} X_i$) entonces

$$P[X_1 = x_1, X_2 = x_2, \cdots, X_k = x_k]$$

$$= \begin{cases} \dfrac{n!}{x_1!x_2!\dots x_k!\left(n - \sum_{i=1}^{k} x_i\right)!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k} (1 - \sum_{i=1}^{k} p_i)^{n - \sum_{i=1}^{k} x_i} & x_i = 0, 1, \cdots, n, \\ & \sum_{i=1}^{k} x_i \leq n \\ 0 & \text{en otro caso} \end{cases}$$

Definición

Una variable aleatoria k-dimensional $X = (X_1, X_2, \cdots, X_k)$ se dice que sigue una distribución multinomial de parámetros n y $p_1, p_2, \cdots, p_k$ si su función masa de probabilidad viene dada por la expresión anterior y se denotará por $X \sim M(n; p_1, p_2, \cdots, p_k)$.

Se puede probar fácilmente que es una función masa de probabilidad. En efecto son cantidades positivas y su suma en los posibles valores de $x_1, x_2, \cdots, x_k$ es

$$\left(p_1 + p_2 + \cdots + p_k + \left(1 - \sum_{i=1}^{k} p_i\right)\right)^n = 1.$$

Función generatriz de momentos

$$M_X(t_1, t_2, \cdots, t_k) = E\left[e^{t_1 X_1 + t_2 X_2 + \cdots + t_k X_k}\right]$$

$$= \sum_{\substack{x_1, x_2, \cdots, x_k = 0 \\ x_1 + x_2 + \cdots + x_k \leq n}}^{n} e^{t_1 x_1 + t_2 x_2 + \cdots t_k x_k} \frac{n!}{x_1!x_2!\dots x_k!\left(n - \sum_{i=1}^{k} x_i\right)!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k} (1 - \sum_{i=1}^{k} p_i)^{n - \sum_{i=1}^{k} x_i}$$

$$= \sum_{\substack{x_1, x_2, \cdots, x_k = 0 \\ x_1 + x_2 + \cdots + x_k \leq n}}^{n} \frac{n!}{x_1!x_2!\dots x_k!\left(n - \sum_{i=1}^{k} x_i\right)!} (p_1 e^{t_1})^{x_1} (p_2 e^{t_2})^{x_2} \dots (p_k e^{t_k})^{x_k} (1 - \sum_{i=1}^{k} p_i)^{n - \sum_{i=1}^{k} x_i}$$

$$= \left(p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_k e^{t_k} + \left(1 - \sum_{i=1}^{k} p_i\right)\right)^n \qquad \forall t_1, t_2, \cdots, t_k \in \mathbb{R}.$$

Distribuciones marginales

La función generatriz de momentos de cada componente X_i se puede obtener como

$$\mathrm{M}_{X_i}(t_i) = \mathrm{M}_X(0, \cdots, 0, t_i, 0, \cdots, 0) = (p_i e^{t_i} + (1 - p_i))^n$$

de donde se deduce que $X_i \sim B(n, p_i),\ i = 1, 2, \cdots, k$ y, por tanto $\mathrm{E}[X_i] = np_i$ y $\mathrm{Var}[X_i] = np_i(1 - p_i)$.

De igual forma se prueba que la distribución de cualquier subvector $(X_{n_1}, X_{n_2}, \cdots, X_{n_j}),\ j < k$ es $M(n; p_{n_1}, p_{n_2}, \cdots, p_{n_j})$.

Vector de medias y matriz de covarianzas

Calculemos $\mathrm{Cov}(X_i, X_j) = \mathrm{E}X_iX_j - \mathrm{E}X_i\mathrm{E}X_j$

$$\mathrm{E}X_iX_j = \left.\frac{\partial^2 M_X(t_1, \cdots, t_k)}{\partial t_i \partial t_j}\right|_{(t_1,\cdots,t_k)=(0,\cdots,0)}$$

$$= n(n-1)\left(p_1\mathrm{e}^{t_1} + \cdots + p_k\mathrm{e}^{t_k} + \left(1 - \sum_{i=1}^{k} p_i\right)\right)^{n-2} p_ip_j\ \mathrm{e}^{t_i}\mathrm{e}^{t_j}\big|_{(t_1,\cdots,t_k)=(0,\cdots,0)} = n(n-1)p_ip_j$$

y de aquí

$$\mathrm{Cov}(X_i, X_j) = n(n-1)p_ip_j - n^2p_ip_j = -np_ip_j$$

Por tanto el vector de medias y la matriz de covarianzas vienen dados por

$$\mathrm{E}X = (np_1, np_2, \cdots, np_k)$$

$$\Sigma_X = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_k \\ -np_1p_2 & np_2(1-p_2) & \cdots & -np_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -np_1p_k & -np_2p_k & \cdots & np_k(1-p_k) \end{pmatrix}$$

Distribuciones condicionadas

$$X_1, X_2, \cdots, X_h|(X_{h+1}, \cdots X_k) = (x_{h+1}, \cdots, x_k)$$

$$\sim M\left(n - \sum_{i=h+1}^{k} x_i; \frac{p_1}{1 - \sum_{i=h+1}^{k} p_i}, \frac{p_2}{1 - \sum_{i=h+1}^{k} p_i}, \cdots, \frac{p_h}{1 - \sum_{i=h+1}^{k} p_i}\right)$$

y, en particular

$$X_i|X_j = x_j \sim M\left(n - x_j; \frac{p_i}{1 - p_j}\right) \equiv B\left(n - x_j; \frac{p_i}{1 - p_j}\right)$$

Regresión y correlación

Limitándonos al estudio de la regresión bidimensional, la curva de regresión de X_i sobre $X_j,\ i,j=1,2,\cdots,k,\ i\neq j$ es

$$x_i = \mathrm{E}[X_i|X_j = x_j] = (n-x_j)\frac{p_i}{1-p_j}$$

que coincide con la correspondiente recta de regresión. El coeficiente de correlación entre dos componentes cualesquiera $X_i,\ X_j,\ \ i,j=1,2,\cdots,k,\ \ i\neq j$ es

$$\rho_{X_iX_j} = -\frac{np_ip_j}{\sqrt{np_i(1-p_i)np_j(1-p_j)}} = -\left[\frac{p_ip_j}{(1-p_i)(1-p_j)}\right]^{1/2}.$$

Propiedad reproductiva

La distribución multinomial es reproductiva respecto al parámetro n. En efecto, sean X e Y dos vectores aleatorios independientes distribuidos según dos distribuciones multinomiales

$$X = (X_1, X_2, \cdots, X_k)' \sim M(n, p_1, p_2, \cdots, p_k)$$

$$Y = (Y_1, Y_2, \cdots, Y_k)' \sim M(m, p_1, p_2, \cdots, p_k)$$

entonces el vector aleatorio $Z = X+Y$ se distribuye según una $M(n+m, p_1, p_2, \cdots, p_{k)}$.

Demostración.- Calculemos la función generatriz de momentos de $Z = X+Y$, que, por la independencia será el producto de las funciones generatrices de momentos de X e Y

$$\mathrm{M}_Z(t_1,t_2,\cdots,t_k) = \mathrm{M}_X(t_1,t_2,\cdots,t_k)\mathrm{M}_Y(t_1,t_2,\cdots,t_k)$$

$$= \left(p_1\mathrm{e}^{t_1} + p_2\mathrm{e}^{t_2} + \cdots + p_k\mathrm{e}^{t_k} + \left(1 - \sum_{i=1}^{k} p_i\right)\right)^{n+m}$$

y, por la unicidad de la función generatriz de momentos, se deduce el resultado deseado.

Distribución normal bidimensional

Es la generalización bidimensional de la normal unidimensional

$$X \sim \mathcal{N}(\mu, \sigma^2) \;\Leftrightarrow\; f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-(x-\mu)^2/2\sigma^2} \quad x \in \mathbb{R}\ (\mu \in \mathbb{R}, \sigma > 0)$$

Función de densidad

Un vector aleatorio (X, Y) de tipo continuo tiene distribución normal bidimensional si su función de densidad es

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

con $x, y \in \mathbb{R}$ y $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 > 0$, $-1 < \rho < 1$

Probemos que es función de densidad para lo cual

1. $f(x,y) > 0$ (evidente)
2. $\int\int f(x,y)dxdy = 1$

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

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

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

De aquí es inmediato que $\int\int f(x,y)dxdy = 1$.

Marginales

$$X \sim \mathcal{N}(\mu_1, \sigma_1^2) \qquad Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

Condidionadas

$$X|Y = y \sim \mathcal{N}(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y-\mu_2), \sigma_1^2(1-\rho^2))$$

$$Y|X = y \sim \mathcal{N}(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), \sigma_2^2(1 - \rho^2))$$

Regresión y correlación

Curvas de regresión

$$x = \mathrm{E}[X|Y = y] = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2)$$

$$y = \mathrm{E}[Y|X = x] = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)$$

que coinciden con las rectas de regresión y los ECM asociados son

$$\mathrm{ECM}(X/Y) = \mathrm{E}[\mathrm{Var}[X|Y]] = \sigma_1^2(1 - \rho^2)$$

$$\mathrm{ECM}(Y/X) = \mathrm{E}[\mathrm{Var}[Y|X]] = \sigma_2^2(1 - \rho^2)$$

Además

$$\eta^2_{Y/X} = \eta^2_{X/Y} = \rho^2_{X,Y} = \rho^2$$

Identificación de coeficientes. Vector de medias y matriz de covarianzas

A partir de las distribuciones marginales se deduce que

$$\mathrm{E}X = \mu_1, \quad \mathrm{E}Y = \mu_2, \quad \mathrm{Var}X = \sigma_1^2, \quad \mathrm{Var}Y = \sigma_2^2$$

Por otra parte

$$\gamma_{X/Y} = \rho\frac{\sigma_1}{\sigma_2}, \quad \gamma_{Y/X} = \rho\frac{\sigma_2}{\sigma_1}$$

y de aquí

$$\rho^2_{X,Y} = \rho^2, \quad \rho_{X,Y} = \rho$$

Por tanto, el vector de medias y la matriz de covarianzas son

$$\mu = (\mu_1, \mu_2)$$
$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

Observemos que la matriz de covarianzas es no singular ($\rho^2 \neq 1$) y semidefinida positiva.

Se tiene que

Independencia $\Leftrightarrow$ Incorrelación $\Leftrightarrow$ Σ Diagonal

La segunda equivalencia es evidente y la primera es debido a que

$$\rho = 0 \;\Leftrightarrow\; f(x, y) = f_1(x)f_2(y)$$

Nota: Independencia $\Leftrightarrow$ Incorrelación se verifica si (X,Y) tiene distribución normal; no es válido si cada variable tiene distribución normal unidimensional pero la conjunta no.

Expresión matricial de la función de densidad

Se puede probar que la función de densidad de la distribución normal bidimensional se puede escribir como

$$f(x) = \frac{1}{2\pi|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)^T\right\}$$

con $x=(x_1,x_2)$, $\mu=(\mu_1,\mu_2)$ y

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

y la notación es $X \sim \mathcal{N}_2(\mu,\Sigma)$

Extensión: Distribución normal multidimensional

Un vector aleatorio $X=(X_1,X_2,\ldots,X_n)$ se dice que tiene distribución normal n-dimensional, y se nota $X \sim \mathcal{N}_n(\mu,\Sigma)$ si su función de densidad es

$$f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)^T\right\}$$

Función generatriz de momentos

$$M_{(X,Y)}(t_1,t_2) = \mathrm{E}\left[\mathrm{e}^{t_1X+t_2Y}\right] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\mathrm{e}^{t_1x+t_2y}f(x,y)dxdy =$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\mathrm{e}^{t_1x}\mathrm{e}^{t_2y}f_2(y)f(x/y)dxdy = \int_{-\infty}^{+\infty}\mathrm{e}^{t_2y}f_2(y)\int_{-\infty}^{+\infty}\mathrm{e}^{t_1x}f(x/y)dxdy =$$

$$= \int_{-\infty}^{+\infty}\mathrm{e}^{t_2y}f_2(y)M_{X/Y=y}(t_1)dy = \int_{-\infty}^{+\infty}\mathrm{e}^{t_2y}f_2(y)\exp\left\{t_1\left(\mu_1+\rho\frac{\sigma_1}{\sigma_2}(y-\mu_2)\right)+\frac{t_1^2\sigma_1^2(1-\rho^2)}{2}\right\}dy =$$

$$\exp\left\{t_1\mu_1 - t_1\rho\frac{\sigma_1}{\sigma_2}\mu_2+\frac{t_1^2\sigma_1^2(1-\rho^2)}{2}\right\}\int_{-\infty}^{+\infty}\exp\left\{y\left(t_2+t_1\rho\frac{\sigma_1}{\sigma_2}\right)\right\}f_2(y)dy =$$

$$\exp\left\{t_1\mu_1 - t_1\rho\frac{\sigma_1}{\sigma_2}\mu_2+\frac{t_1^2\sigma_1^2(1-\rho^2)}{2}\right\}M_Y\left(t_2+t_1\rho\frac{\sigma_1}{\sigma_2}\right) =$$

$$= \exp\left\{t_1\mu_1 - t_1\rho\frac{\sigma_1}{\sigma_2}\mu_2+\frac{t_1^2\sigma_1^2(1-\rho^2)}{2}\right\}\exp\left\{\mu_2\left(t_2+t_1\rho\frac{\sigma_1}{\sigma_2}\right)+\frac{(t_2+t_1\rho\frac{\sigma_1}{\sigma_2})^2\sigma_2^2}{2}\right\} =$$

$$= \exp\left\{t_1\mu_1 - \rho\frac{\sigma_1}{\sigma_2}t_1\mu_2 + \frac{t_1^2\sigma_1^2(1-\rho^2)}{2} + \mu_2 t_2 + t_1\mu_2\rho\frac{\sigma_1}{\sigma_2} + \frac{t_2^2\sigma_2^2}{2} + \frac{t_1^2\rho^2\sigma_1^2}{2} + 2\frac{t_1t_2\rho\sigma_1\sigma_2}{2}\right\} =$$

$$= \exp\left\{t_1\mu_1 + t_2\mu_2 + \frac{t_1^2\sigma_1^2 + t_2^2\sigma_2^2 + 2\rho t_1t_2\sigma_1\sigma_2}{2}\right\} = \exp\left\{t\mu^T + \frac{t\Sigma t^T}{2}\right\}, \qquad t = (t_1, t_2)$$

Combinaciones lineales de las componentes de un vector normal bidimensional

$$X = (X_1, X_2) \sim \mathcal{N}_2\left(\mu = (\mu_1, \mu_2), \Sigma_{X_1,X_2} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

y sea $A_{2\times q}$ ($q \leq 2$) de rango máximo q. Entonces

$$Y = XA \;\; (q\text{-dimensional}) \;\sim\; \mathcal{N}_q(\mu A, A^T\Sigma A)$$

Dem.- Sea $t \in \mathbb{R}^q$

$$M_Y(t) = \mathrm{E}\left[\mathrm{e}^{tY^T}\right] = \mathrm{E}\left[\mathrm{e}^{(tA^T)X^T}\right] = M_X(tA^T) =$$

$$= \exp\left\{(tA^T)\mu^T + \frac{(tA^T)\Sigma(tA^T)^T}{2}\right\} = \exp\left\{t(\mu A)^T + \frac{t(A^T\Sigma A)t^T}{2}\right\}$$

y dado que A es de rango máximo, $|A^T\Sigma A| \neq 0$ se deduce

$$Y = XA \sim \mathcal{N}_q(\mu A, A^T\Sigma A)$$

NOTA.- $XA_{2\times 2} = (a_{11}X_1 + a_{12}X_2, a_{21}X_1 + a_{22}X_2)$ con lo que cualquier vector (bidimensional) cuyas componentes sean combinaciones lineales de X_1 y X_2, de forma que la matriz A que define tales combinaciones sea no singular, tiene distribución normal bidimensional.

Corolario.- Cualquier combinación lineal de las componentes de un vector normal bidimensional tiene distribución normal unidimensional.

$$X = (X_1, X_2) \sim \mathcal{N}_2\left(\mu = (\mu_1, \mu_2), \Sigma_{X_1,X_2} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

$$a_1X_1 + a_2X_2 = (X_1, X_2)\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}_{2\times 1} \sim \mathcal{N}_1\left((\mu_1, \mu_2)\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, (a_1, a_2)\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}\right)$$

$$\equiv \mathcal{N}_1(a_1\mu_1 + a_2\mu_2, a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + 2\rho\sigma_1\sigma_2a_1a_2)$$

siempre que $\mathrm{rg}(A) = 1$ (a_1 ò $a_2 \neq 0$).