Resumen-Tema-2.pdf

ferluque

ESTADÍSTICA DESCRIPTIVA E INTROD A LA PROBABILIDAD

1º Doble Grado en Ingeniería Informática y Matemáticas

Facultad de Ciencias
Universidad de Granada

Tema 2

Distribución conjunta de dos carácteres estadísticos

En una población de tamaño n, se han observado dos variable estadísticas X e Y, de las cuales, la variable X ha presentado k modalidades distintas (x_1, x_2, ..., x_k) y la variable Y ha presentado p modalidades distintas (y_1,y_2,...,y_p) con **distribución de frecuencias conjuntas** $\{(x_i, y_j); n_{ij}\}_{i=\{1,...,k\}/j=\{1,...,p\}}$

- **Frecuencia absoluta (n_{ij})**: Cantidad de individuos que han presentado **simultáneamente** la modalidad x_i del carácter X y la modalidad y_j del carácter Y
- **Frecuencia relativa (f_{ij})**: Proporción de la población que ha presentado **simultáneamente** la modalidad x_i del carácter X y la modalidad y_j del carácter Y
- **Frecuencia absoluta (relativa) condicionada ($n_{i.}$ ($f_{i.}$))**: Número total de individuos (proporción de individuos) que han presentado la modalidad x_i del carácter X, sin tener en cuenta las modalidades del carácter Y
- **Frecuencia absoluta (relativa) condicionada ($n_{.j}$($f_{.j}$))**: Número total de individuos (proporción) que han presentado la modalidad y_j del carácter Y, sin tener en cuenta las modalidades del carácter X

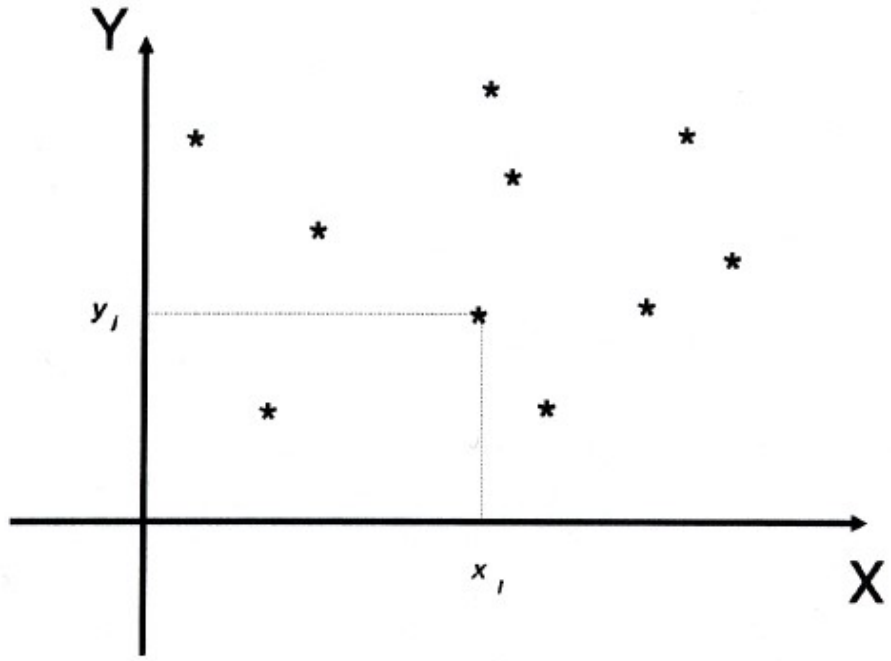
  $n = \sum_{i=1}^{k} n_{i.} = \sum_{j=1}^{p} n_{.j}$

Tablas bidimensionales

$X \quad Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_p	$n_{i.}$
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1p}	$n_{1.}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2p}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ip}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\cdots$	n_{kj}	$\cdots$	n_{kp}	$n_{k.}$
$n_{.j}$	$n_{,1}$	$n_{,2}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.p}$	n

Representaciones gráficas

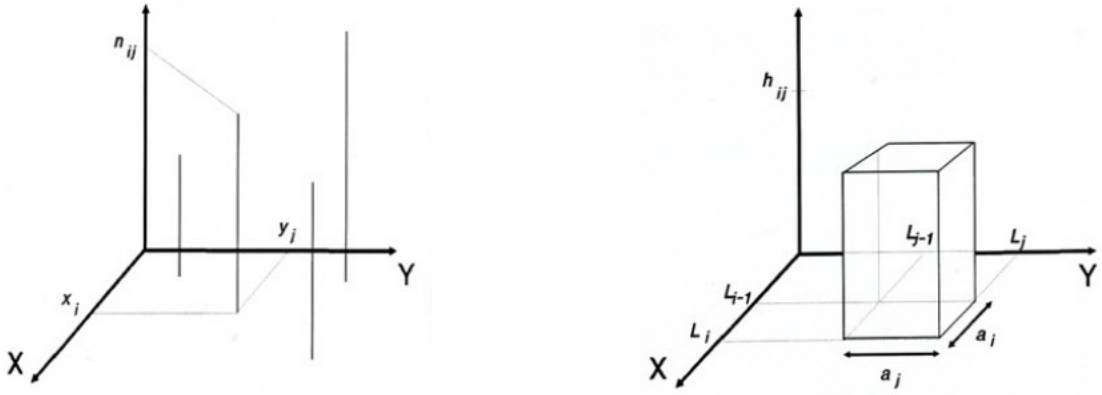
Diagrama de dispersión



No se tiene en cuenta los valores n_{ij}, sólo qué pares (x_i, y_j) tienen $n_{ij} \neq 0$

Estereograma

Izquierda: Variables cualitativas y discretas // Derecha: Variables continuas



En este caso, la altura va relacionada con cada n_{ij} de la siguiente manera:

$h_{ij} = \frac{n_i j}{(L_i - L_{i-1})(L_j - L_{j-1})}$

En las variables discretas $n_{ij} = h_{ij}$

Distribuciones marginales

- **Distribución marginal de X**: Las modalidades x_i junto con las frecuencias $n_{i.}$. Se realizan sin tener en cuenta las modalidades de Y
- **Distribución marginal de Y**: Las modalidades y_j junto con las frecuencias $n_{.j}$. Se realizan sin tener en cuenta las modalidades de X

Serie Original
8 de junio solo en

Disney+

Distribuciones condicionadas

$X_{/Y=y_j}$:

En esta distribución condicionada, $n_{ij};\ i = 1, 2, \ldots, k$ individuos en $n_{\cdot j}$ presentan la modalidad x_i de X. Por ello, la **frecuencia relativa de la modalidad x_i de X en los individuos que presentan la modalidad y_j de Y es:

$$f_{i/j} = f_i^j = \frac{n_{ij}}{n_{\cdot j}}$$

Análogamente se define la distribución condicionada $Y_{/X=x_i}$

Existen p distribuciones condicionadas de X para una sola modalidad de Y y k distribuciones condicionadas de Y para una sola modalidad de X

Relaciones

$$f_{ij} = \frac{n_{ij}}{n} = \frac{n_{i\cdot}}{n} \cdot \frac{n_{ij}}{n_{i\cdot}} = \frac{n_{\cdot j}}{n} \cdot \frac{n_{ij}}{n_{\cdot j}}$$

$$f_{ij} = f_{i\cdot} \cdot f_j^i = f_{\cdot j} \cdot f_i^j$$

Dependencia e independencia estadística

Dos carácteres X e Y serán **estadísticamente dependientes** cuando la variación en uno de ellos influya en la distribución del otro

Serán **estadísticamente independientes** si las distribuciones de X condicionadas a cada valor y_j de Y son todas idénticas para cualquier valor de j. Es decir f_i^j no depende de j. O bien, las frecuencias absolutas son proporcionales

Además, tenemos también que $f_i^j = f_{i\cdot}$

De forma análoga se define la **independencia de Y de X**

Proposición

$$Si\ el\ carácter\ X\ es\ independiente\ del\ carácter\ Y,\ entonces\ Y\ es\ independiente\ de\ X$$

$$X\ es\ independiente\ de\ Y \Leftrightarrow Y\ es\ independiente\ de\ X$$

Dependencia funcional

X es **funcionalmente dependiente** de Y si a cada modalidad y_j le corresponde una única modalidad x_i con frecuencia no nula.

Y es **funcionalmente dependiente** de X si a cada modalidad x_i le corresponde una única modalidad y_j con frecuencia no nula.

La dependencia funcional **no siempre es recíproca**

Momentos bidimensionales

No centrales

$$m_{rs} = \sum_i \sum_j n_{ij} x_i^r y_j^s$$

Centrales

$$\mu_{rs} = \sum_i \sum_j n_{ij} (x_i - \bar{x})^r (y_j - \bar{y})^s$$

- Cabe destacar que si $r = 0$, los momentos coinciden con los de la variable Y y si $s = 0$, coinciden con los de la variable X
- El momento μ_{11} se denomina covarianza de las variables X e Y, y también se suele representar por σ_{xy}
  - $\sigma_{xy} = \mu_{11} = m_{11} - m_{10}m_{01}$
  - X e Y independientes $\Rightarrow m_{rs} = m_{r0}m_{0s}$
  - X e Y independientes $\Rightarrow \sigma_{xy} = 0$

Regresión

Con el objetivo de encontrar relaciones entre dos variables que no son ni **funcionalmente dependientes** ni **estadísticamente independientes**, la estadística trata de buscar una expresión que las relacione de la siguiente forma: $Y = f(X_1, X_2, \ldots, X_n)$

Para ello, denominaremos a unas de las variables como **variable independiente** y a la otra como **variable dependiente**

Por tanto, hacer regresión consiste en ajustar lo mejor posible una función a una serie de valores observados encontrando una curva que, aunque no pase por todos los puntos de la nube, al menos esté lo más próxima posible a ellos.

Regresión de tipo I

Sin duda, el valor más representativo del comportamiento de una variable es la media aritmética. Es por eso que se define la *curva de regresión de tipo I de* Y/X como la curva que pasa por los puntos $(x_i, \bar{y}_i); i = 1, \ldots, k$. Siendo $\bar{y}_i$ la media de la distribución $Y_{/X=x_i}$. Análogamente se define la de X/Y

Esta curva es la que mejor se ajusta a los datos observados según el criterio de mínimos cuadrados. Sin embargo, esta curva no tiene gran utilidad práctica pues no nos permite hacer predicciones en la mayoría de los casos, ya que la conocemos sólo en puntos aislados.

Ajuste de funciones por mínimos cuadrados

Consiste en encontrar una función $f(x_i, a_0, a_1, \ldots, a_n)$ que minimice la media de los cuadrados de los residuos:

$$\psi(a_0, a_1, \ldots, a_n) = \sum_i \sum_j f_{ij}(y_j - f(x_i, a_0, a_1, \ldots, a_n))^2$$

Para hayar $a_0, \ldots, a_n$ debemos de hacer las respectivas derivadas parciales e igualarlas a 0:

$$\frac{\partial \psi}{\partial a_r} = 0$$

De esta forma, obtenemos los valores de a_r y hayamos la función f

De forma análoga se haría para la función $f(y_j, b_0, b_1, \ldots, b_m)$

Nota: Cambios de variables:

- **Hipérbola equilátera**:

  $y = a + \frac{b}{x} \Rightarrow x' = \frac{1}{x} \Rightarrow y = a + bx'$
- **Potencial**:

  $y = ax^b \Rightarrow y' = \log y // a' = \log a // x' = \log x \Rightarrow \log y = \log a + b \log x \Rightarrow y' = a' + bx'$
- **Exponencial**:

  $y = ab^x \Rightarrow \log y = \log a + x \log b \Rightarrow y' = a' + b'x \Rightarrow y' = \log y // a' = \log a // b' = \log b$

Varianza residual: Como siempre, en las funciones lineales en los parámetros, la media de los residuos es cero, usamos la **varianza residual**, que se define como la media de los residuos al cuadrado:

$\sigma_{ry}^2 = \sum_i \sum_j f_{ij}(y_j - \hat{y}_i)^2 = \sum_i \sum_j f_{ij}(y_j - f(x_i))^2$

En funciones no lineales también se aplica la varianza residual.

Para explicar el grado de ajuste, tomaremos como medida la proporción de la varianza total de la variable Y, lo que hace posible que la varianza de la variable se pueda dividir en suma de varianza residual y varianza explicada por la regresión, de forma que:

$\sigma_y^2 = \sigma_{ry}^2 + \sigma_{ey}^2$, donde

$\sigma_{ey}^2 = \sum_i \sum_j f_{ij}(\hat{y}_j - \bar{y})^2$

denominada, **varianza explicada por la regresión**. A mayor σ_{ey}^2, mejor ajuste.

Con el objetivo de poder hacer comparaciones, se define ahora el **coeficiente de determinación** o en general **razón de correlación**:

$\eta_{Y/X}^2 = \frac{\sigma_{ey}^2}{\sigma_y^2} = 1 - \frac{\sigma_{ry}^2}{\sigma_y^2}$

De esta forma $0 \leq \eta_{Y/X}^2 \leq 1$, lo que nos permite hacer comparaciones entre diferentes distribuciones.

Mientras más se acerque $\eta_{Y/X}^2$ a 1, mejor será el ajuste

Caso lineal

$\eta_{Y/X}^2 = \eta_{X/Y}^2 = r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2}$

Además, r^2 coincide con el producto de las pendientes de las rectas.

La raíz cuadrada del coeficiente de determinación lineal anterior (con el signo de la covarianza) recibe el nombre de **coeficiente de correlación lineal**

$r = \pm\sqrt{r^2} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$