Estadística Descriptiva y Análisis de Datos

Diplomatura en Estadística. Curso 2007/2008.

Formulario Tema 4.

Sea (X, Y) una variable bidimensional que puede tomar $k \times r$ pares de valores diferentes (x_i, y_j), $i = 1, \ldots, k$, $j = 1, \ldots, r$, sobre los n individuos de una muestra. Supondremos que tenemos los datos ordenados de manera que $x_1 < x_2 < \ldots < x_k$, $y_1 < y_2 < \ldots < y_r$. Es práctico presentar los datos en forma de tabla:

Table 1: **Tabla de doble entrada y tabla de contingencias**

$X \setminus Y$	y_1	y_2	$\ldots$	y_j	$\ldots$	y_r	n_X
x_1	n_{11}	n_{12}	$\ldots$	n_{1j}	$\ldots$	n_{1r}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\ldots$	n_{2j}	$\ldots$	n_{2r}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\ldots$	n_{ij}	$\ldots$	n_{ir}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	$\ldots$	n_{kj}	$\ldots$	n_{kr}	$n_{k\bullet}$
n_Y	$n_{\bullet 1}$	$n_{\bullet 2}$	$\ldots$	$n_{\bullet j}$	$\ldots$	$n_{\bullet r}$	n

- La **frecuencia absoluta** n_{ij} es el número de veces que aparece el par (x_i, y_j) en los n individuos de la muestra.
  Se cumple que $\sum_{i=1}^{k} \sum_{j=1}^{r} n_{ij} = n$.
- La **frecuencia relativa** es la proporción de individuos para los cuales se ha observado el par (x_i, y_j), es decir, $f_{ij} = n_{ij}/n$.
- La **frecuencia absoluta marginal** del valor x_i de la variable X es $n_{i\bullet} = \sum_{j=1}^{r} n_{ij}$.
- La **frecuencia relativa marginal** del valor x_i de la variable X es $f_{i\bullet} = n_{i\bullet}/n$.
- La **frecuencia absoluta marginal** del valor y_j de la variable Y es $n_{\bullet j} = \sum_{i=1}^{k} n_{ij}$.
- La **frecuencia relativa marginal** del valor y_j de la variable Y es $f_{\bullet j} = n_{\bullet j}/n$.
- La **distribución** de X **condicionada** a que Y tome el valor y_j, $X_{|Y=y_j}$, se obtiene de la columna j-ésima de la tabla 1.
- La **distribución** de Y **condicionada** a que X tome el valor x_i, $Y_{|X=x_i}$, se obtiene de la fila i-ésima de la tabla 1.

Características numéricas marginales.

- Medias marginales:

$$\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{k} x_i\, n_{i\bullet}, \quad \overline{\mathbf{y}} = \frac{1}{n}\sum_{j=1}^{r} y_j\, n_{\bullet j}$$

- Varianzas marginales:

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{\mathbf{x}})^2\, n_{i\bullet} = \overline{\mathbf{x}^2} - \overline{\mathbf{x}}^2, \quad s_Y^2 = \frac{1}{n}\sum_{j=1}^{r}(y_j - \overline{\mathbf{y}})^2\, n_{\bullet j} = \overline{\mathbf{y}^2} - \overline{\mathbf{y}}^2$$

- Desviaciones típicas marginales: $s_X = \sqrt{s_X^2}$, $s_Y = \sqrt{s_Y^2}$.

Características numéricas conjuntas para tablas de doble entrada.

- Covarianza:

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{r}(x_i - \overline{\mathbf{x}})\,(y_j - \overline{\mathbf{y}})\, n_{ij} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{r} x_i\, y_j\, n_{ij} - \overline{\mathbf{x}}\,\overline{\mathbf{y}} = \overline{x\,y} - \overline{\mathbf{x}}\,\overline{\mathbf{y}}$$

- Coeficiente de correlación lineal de Pearson:

$$r_{XY} = \frac{s_{XY}}{s_X\, s_Y}$$

Medidas de associación para tablas de contingencias.

- Estadístico χ^2:

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{(n_{ij} - n_{i\bullet}n_{\bullet j}/n)^2}{n_{i\bullet}n_{\bullet j}/n},$$

- Estadístico χ^2 con la correción de Yates:

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{r}\frac{(|n_{ij} - n_{i\bullet}n_{\bullet j}/n| - 0.5)^2}{n_{i\bullet}n_{\bullet j}/n}$$

- Coeficiente de contingencia de Pearson:

$$C = \sqrt{\frac{Q/n}{1 + Q/n}}$$

- V de Cramer:

$$V = \sqrt{\frac{Q/n}{\min\{k-1, r-1\}}}$$

- λ de Goodman-Kruskal:

$$\lambda = \frac{P(\text{error sin información de } X) - P(\text{error, dado } X)}{P(\text{error sin información de } X)}$$

Sean C el número de pares concordantes en la tabla 1 y D el número de pares discordantes en la tabla 1.

- γ de Goodman-Kruskal:

$$\gamma = \frac{C - D}{C + D}$$

- $\mathcal{D}$ de Sommer:

$$\mathcal{D} = \frac{C - D}{n(n-1)/2 - T_X}, \quad \text{donde} \quad T_X = \sum_{i=1}^{k} \frac{n_{i\bullet}(n_{i\bullet} - 1)}{2}.$$

- τ_B y τ_C de Kendall:

$$\tau_B = \frac{C - D}{\sqrt{(n(n-1)/2 - T_X)\,(n(n-1)/2 - T_Y)}}, \quad \tau_C = \frac{\min\{k, r\}(C - D)}{\min\{k-1, r-1\}\, n^2},$$

donde $T_X = \sum_{i=1}^{k} \frac{n_{i\bullet}(n_{i\bullet}-1)}{2}$. $T_Y = \sum_{j=1}^{r} \frac{n_{\bullet j}(n_{\bullet j}-1)}{2}$.

Medidas especiales.

- Coeficiente de correlación de Spearman:

$$r_S = 1 - \frac{6 \sum_{i=1}^{k} d_i^2}{k(k^2 - 1)},$$

donde $d_i = x_i - y_i$ representa la diferencia entre los rangos asignados al i-ésimo elemento de un colectivo de k elementos.