TEMA 4: DISTRIBUCIONES BIDIMENSIONALES

4.1.- DISTRIBUCIONES BIDIMENSIONALES DE FRECUENCIAS.

4.2.- REPRESENTACIONES GRÁFICAS.

4.3.- MOMENTOS DE DISTRIBUCIONES BIDIMENSIONALES.

4.1- DISTRIBUCIONES BIDIMENSIONALES DE FRECUENCIAS.

Podemos estudiar dos o más caracteres cuantitativos diferentes de forma simultánea. Ejemplo, en un curso de estudiantes el peso y la altura, las notas de matemáticas y estadística, y así llamamos distribución bidimensional a un conjunto ordenado de pares de valores de dos caracteres (Xi, Yj) asociado a las frecuencias absolutas n_{ij} o relativas f_{ij} de dichos pares.

Siendo n_{ij} el número de veces que se presenta conjuntamente el par de valores (Xi, Yj), es decir la frecuencia absoluta bidimensional, igual que en el caso unidimensional. La frecuencia relativa bidimensional del par (Xi,Yj) será $f_{ij} = n_{ij}/N$, es decir el cociente entre la frecuencia absoluta correspondiente (n_{ij}) y la suma de las frecuencias absolutas bidimensionales (N), y se representa por (Xi Yj n_{ij}).

Ante esta situación, nos podemos preguntar ¿Por qué se realiza este estudio simultáneo? ¿Cuál es su fin?. Se puede evidentemente, estudiar cada variable X e Y, por separado, y calcular sus medidas características, pero el interés reside en el fin de poder estudiar las posibles relaciones entre X e Y. ¿Existe relación entre las notas de matemáticas y estadística?.

O sea, lo que implícitamente se busca es si existe una relación causal entre X e Y. No existe instrumento estadístico que permita afirmar relaciones de causalidad, pero si existe instrumento estadístico que permita revelar la existencia de coincidencias entre los valores de 2 variables, y así

INDEPENDENCIA Y RELACIÓN FUNCIONAL DE DOS VARIABLES

Decimos que si existen coincidencias, va a existir relación, pero ¿cuál es la intensidad de esa relación? ¿Qué graduación tiene? ¿Cuáles serán los extremos de una relación? Se presentan las siguientes situaciones:

a) No existe relación entre X e Y, y decimos que las variables son independientes.
b) Existe relación perfecta entre X e Y: las variables tienen dependencia funcional, es decir su relación es expresable de la forma y = f(x), son determinables perfectamente los elementos de Y conocido X y viceversa.
c) Situaciones intermedias: existen otros caracteres como estatura y peso que no cabe duda que existe interrelación pero no son definibles funcionalmente en sentido matemático. En estas relaciones = coincidencias, existe dependencia estadística y ésta puede ser más o menos fuerte.

   La dependencia estadística admite grados = intensidad relación, pero la dependencia funcional no admite grados.

TABLA DE CORRELACIÓN: presentación de una variable bidimensional. Es una tabla de doble entrada. Sea una población con 2 caracteres X e Y, se representa por (X_i Y_j, n_{ij}) siendo X_i, Y_j: 2 valores cualesquiera y n_{ij} la frecuencia absoluta conjunta del valor i der X con el j de Y (número de veces que se repite). El número total de individuos observados es N.

TABLAS DE CORRELACIÓN (datos cuantitativos)
TABLAS DE CONTINGENCIA (datos cualitativos)

Distribución de frecuencias absolutas conjuntas n_{ij} (con las que se representa el par simultáneo (x_i, y_j) y frecuencias marginales

$x_i \backslash y_j$	y_1	y_2	…	Y_k	$n_{i\cdot}$
x_1	n_{11}	n_{12}	…	n_{1k}	$n_{1.}$
x_2	n_{21}	n_{22}	…	n_{2k}	$n_{2.}$
….	…	…	…	…	….
X_h	n_{h1}	n_{h2}	…	n_{hk}	$n_{h.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	….	$n_{.k}$	N

Ejemplo: Sobre una muestra de 100 alumnos se miden las siguientes características:

X= asignación mensual

Y= gasto mensual en actividades culturales, expresadas en 10 €.

$x_i \backslash y_j$	5	10	15	$n_{i\cdot}$
15	5	15	10	30
25	5	20	5	30
50	12	8	20	40
$n_{.j}$	22	43	35	N=100

pueda tomar la otra), correspondiendo las frecuencias marginales a la última fila y la última columna de la tabla de correlación.

Frecuencia marginal de X_1:

$$n_{1.} = n_{11} + n_{12} + \cdots + n_{1k} = \sum_{j=1}^{k} n_{1j} = n_{1.}$$

Frecuencia marginal de Xi:

$$n_{i.} = n_{i1} + n_{i2} + \cdots + n_{ik} = \sum_{j=1}^{k} n_{ij} = n_{i.}$$

Análogamente la frecuencia marginal de Yj:

$$n_{.j} = n_{1j} + n_{2j} + \cdots + n_{hk} = \sum_{i=1}^{h} n_{ij} = n_{.j}$$

Definimos la frecuencia marginal de Xi como $f_{i.} = n_{i.}/N$ y la frecuencia relativa marginal de Yj como $f_{.j} = n_{.j}/N$.

$\{(x_i;n_{i.})$: $i=1,2,\ldots,h\}$, distribución marginal de X

VARIABLE X		
Xi	ni.	fi.=ni./N
X1	n1.	f1.
X2	n2.	f2.
Xi	ni.	fi.
Xh	nh.	fh.
	N	1

VARIABLE Y		
Yj	n.j	f.j=n.j/N
Y1	n.1	f.1
Y2	n.2	f.2
Yj	n.j	f.j
Yk	n.k	f.k
	N	1

Evidentemente se cumple:

$$\sum_{i=1}^{h} n_{ij} = \sum_{j=1}^{k} n_{.j} = N = \sum_{i=1}^{h} \sum_{i=1}^{k} n_{ij}$$

DISTRIBUCIONES CONDICIONADAS: Se trata de otro tipo de distribución unidimensional extraíble de una variable bidimensional, al definir una condición para algún valor de X o Y. En este caso tendremos las frecuencias relativas en una distribución condicionada de X a un valor de Y = Yj como $f_{i/j} = n_{ij}/n_{.j}$ y la frecuencia relativa condicionada de Y a un valor de X = Xi como $f_{j/i} = n_{ij}/n_{i.}$

VARIABLE X condicionada a Y=Y2		
Xi/Y2	ni/2	fi/2=ni2/n.2
X1	n12	f1/2
X2	n22	f2/2
Xi	ni2	fi/2
Xh	nh2	fh/2
	n.2	1

Generalizando:

X condicionada a Y=Yj (X/Y=Yj)		
Xi/Yj	ni/j	fi/j=nij/n.j
X1	n1j	f1/j=n1j/n.j
X2	n2j	f2/j=n2j/n.j
Xi	nij	fi/j=nij/n.j
Xh	nhj	fh/j=nhj/n.j
	n.j	1

Y condicionada a X=Xi (Y/X=Xi)

Yj/Xi	nj/i	fj/i=nij/ni.
Y1	ni1	f1/i=ni1/ni.
Y2	ni2	f2/i=ni2/ni.
Yj	nij	fj/i=nij/ni.
Yk	nik	fk/i=nik/ni.
	ni.	1

Ejemplo anterior, distribución de X condicionada a Y1=5.

X condicionada a Y1=5 (Xi/y=5)

Xi/Y=5	ni/Y1=5	fi/1
15	5	0,23
25	5	0,23
50	12	0,55
	n.j=22	1,00

INDEPENDENCIA ESTADÍSTICA: los valores que tome una variable no vendrán afectados por los que tome la otra. Decimos que dos variables Xe Y son independientes estadísticamente si la frecuencia relativa conjunta es el producto de las frecuencias relativas marginales, siendo ésta la condición necesaria y suficiente de independencia.

También decimos que dos variables son independientes si las frecuencias relativas condicionadas son iguales a las frecuencias relativas marginales, indicándonos que el condicionamiento como tal no tiene efecto: las variables son independientes.

$$f_{i/j} = \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}n_{.j}/N}{n_{.j}} = \frac{n_{i.}}{N}$$

$$f_{j/i} = \frac{n_{ij}}{n_{i.}} = \frac{n_{i.}n_{.j}/N}{n_{i.}} = \frac{n_{.j}}{N}$$

Ejemplo para X1 Y2 del ejemplo inicial:

$$\frac{15}{100} \neq \frac{30}{100} * \frac{43}{100}$$

$$f_{1/2} = \frac{15}{43} \neq \frac{30}{100}$$

4.2.- REPRESENTACIONES GRÁFICAS.

La representación gráfica más utilizada consiste en representar cada pareja

También se puede representar en 3 dimensiones, con un eje para X, otro eje para Y y el tercer eje para las frecuencias.

4.3.- MOMENTOS DE DISTRIBUCIONES BIDIMENSIONALES

Momentos respecto al origen (a_{rs}). Momento de orden r,s respecto al origen, dada la distribución bidimensional (Xi,Yj,nij) se define como:

$$a_{rs} = \sum_{i=1}^{h} \sum_{j=1}^{k} x_i^r \, y_j^s \frac{n_{ij}}{N}$$

Momentos de primer orden: a_{10}, a_{01} (r+s=1)

r=1 y s=0

$$a_{10} = \sum_{i=1}^{h} \sum_{j=1}^{k} x_i^1 \, y_j^0 \frac{n_{ij}}{N} = \sum_{i=1}^{h} x_i \sum_{j=1}^{k} \frac{n_{ij}}{N} = \sum_{i=1}^{h} x_i \frac{n_{i.}}{N} = \bar{x}$$

r=0 y s=1

$$a_{01} = \sum_{i=1}^{h} \sum_{j=1}^{k} x_i^0 \, y_j^1 \frac{n_{ij}}{N} = \sum_{j=1}^{k} y_j \sum_{i=1}^{h} \frac{n_{ij}}{N} = \sum_{j=1}^{k} y_j \frac{n_{.j}}{N} = \bar{y}$$

Momentos de segundo orden: a_{20}, a_{02}, a_{11} (el más interesante de una distribución bidimensional) (r+s=2) r=0 y s=2

$$a_{02} = \sum_{i=1}^{h}\sum_{j=1}^{k} x_i^0\, y_j^2 \frac{n_{ij}}{N} = \sum_{j=1}^{k} {y_j}^2 \sum_{i=1}^{h} \frac{n_{ij}}{N} = \sum_{j=1}^{k} {y_j}^2 \frac{n_{.j}}{N}$$

r=1 y s=1

$$a_{11} = \sum_{i=1}^{h}\sum_{j=1}^{k} x_i \;\; y_j \;\; \frac{n_{ij}}{N}$$

Momentos respecto a la media (m_{rs})

$$\boldsymbol{m_{rs}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{h}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{k}} (\boldsymbol{x_i} - \boldsymbol{\bar{x}})^{\boldsymbol{r}} \left(\boldsymbol{y_j} - \boldsymbol{\bar{y}}\right)^{\boldsymbol{s}} \frac{n_{ij}}{N}$$

Momentos de primer orden: m_{10}, m_{01}($r+s=1$)

r=1 y s=0

$$\boldsymbol{m_{10}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{h}}\sum_{\boldsymbol{j=1}}^{\boldsymbol{k}} (\boldsymbol{x_i} - \boldsymbol{\bar{x}})^{\boldsymbol{1}} \left(\boldsymbol{y_j} - \boldsymbol{\bar{y}}\right)^{\boldsymbol{0}} \frac{n_{ij}}{N} = \sum_{i=1}^{h} (\boldsymbol{x_i} - \boldsymbol{\bar{x}}) \frac{n_{i.}}{N} = 0$$

r=0 y s=1

$$\boldsymbol{m_{01}} = \sum^{\boldsymbol{h}}\sum^{\boldsymbol{k}} (\boldsymbol{x_i} - \boldsymbol{\bar{x}})^{\boldsymbol{0}} \left(\boldsymbol{y_j} - \boldsymbol{\bar{y}}\right)^{\boldsymbol{1}} \frac{n_{ij}}{\;} = \sum^{k} \left(\boldsymbol{y_j} - \boldsymbol{\bar{y}}\right) \frac{n_{.j}}{\;} = 0$$

Momentos de segundo orden: m_{20}, m_{02}, m_{11} (covarianza=Sxy)

(r+s=2)

r=2 y s=0

$$\boldsymbol{m_{20}} = \sum_{i=1}^{h}\sum_{j=1}^{k}(\boldsymbol{x_i}-\bar{\boldsymbol{x}})^2\,(\boldsymbol{y_j}-\bar{\boldsymbol{y}})^0\frac{n_{ij}}{N} = \sum_{i=1}^{h}(\boldsymbol{x_i}-\bar{\boldsymbol{x}})^2\frac{n_{i.}}{N} = S_x^2$$

r=0 y s=2

$$\boldsymbol{m_{02}} = \sum_{i=1}^{h}\sum_{j=1}^{k}(\boldsymbol{x_i}-\bar{\boldsymbol{x}})^0\,(\boldsymbol{y_j}-\bar{\boldsymbol{y}})^2\frac{n_{ij}}{N} = \sum_{j=1}^{k}(\boldsymbol{y_j}-\bar{\boldsymbol{y}})^2\frac{n_{.j}}{N} = S_y^2$$

r=1 y s=1

$$\boldsymbol{m_{11}} = \sum_{i=1}^{h}\sum_{j=1}^{k}(\boldsymbol{x_i}-\bar{\boldsymbol{x}})\quad(\boldsymbol{y_j}-\bar{\boldsymbol{y}})\quad\frac{n_{ij}}{N} = S_{xy} = Cov(x,y)$$

Cálculo de los momentos centrales en función de los momentos respecto al origen. De igual forma que en las distribuciones unidimensionales, la varianza y covarianza se pueden calcular en función de los momentos respecto al origen.

$$m_{20} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 n_{i.}}{N} = \frac{\sum_{i=1}^{n}\left(x_i^2+\bar{x}^2-2\bar{x}\ x_i\right)n_{i.}}{N} =$$

$$= \frac{\sum_{i=1}^{n}x_i^2\,n_{i.}}{N} + \bar{x}^2\frac{\sum n_{i.}}{N} - 2\bar{x}\frac{\sum_{i=1}^{n}x_i\quad n_{i.}}{N}$$

$$m_{11} = \sum_{i=1}^{h}\sum_{j=1}^{k}(x_i - \bar{x}) \quad (y_j - \bar{y}) \quad \frac{n_{ij}}{N} =$$

$$= \sum_{i=1}^{h}\sum_{j=1}^{k}(x_i y_j - x_i\bar{y} - y_j\bar{x} + \bar{x}\bar{y})\frac{n_{ij}}{N} =$$

$$= \sum_{i=1}^{h}\sum_{j=1}^{k} x_i \quad y_j \quad \frac{n_{ij}}{N} - \bar{y}\sum_{i=1}^{h} x_i \frac{n_{i.}}{N} - \bar{x}\sum_{j=1}^{k} y_j \frac{n_{.j}}{N} + \bar{x}\bar{y}$$

$$= a_{11} - \bar{y}\bar{x} - \bar{x}\bar{y} + \bar{x}\bar{y} = a_{11} - \bar{x}\bar{y} = a_{11} - a_{10}a_{01}$$

$$= S_{xy}$$

Valor de la covarianza en caso de independencia estadística:

Condición de independencia estadística

$$\boxed{\frac{n_{ij}}{N} = \frac{n_{i.}}{N} * \frac{n_{.j}}{N} \quad (\forall i = 1,2,..,h) \; y \; (\forall j = 1,2,..,k)}$$

Cálculo de a_{11} con independencia estadística

$$a_{11} = \sum_{i=1}^{h}\sum_{j=1}^{k} x_i \quad y_j \quad \frac{n_{ij}}{N} = \sum_{i=1}^{h}\sum_{j=1}^{k} x_i \quad y_j \quad \frac{n_{i.}}{N}\frac{n_{.j}}{N} = \sum_{i=1}^{h} x_i \frac{n_{i.}}{N}\sum_{j=1}^{k} y_j \frac{n_{.j}}{N} = a_{10}a_{01}$$

Y por lo tanto:

$$m_{11} = a_{11} - a_{10}a_{01} = a_{10}a_{01} - a_{10}a_{01} = 0$$

Si las variables son independientes , su covarianza es cero, pero el recíproco no siempre es cierto, es decir, covarianza nula no implica independencia.