

Descripción conjunta de dos variables

Tema 2

Descripción breve del tema

1. Introducción
2. Distribuciones de frecuencias
 - Distribución conjunta
 - Distribuciones marginales
 - Distribuciones condicionadas
 - Independencia entre variables
3. Representaciones gráficas
 - Diagrama de dispersión
 - Otras representaciones
4. Medidas de dependencia lineal
 - Covarianza
 - Correlación

Objetivos

- Reconocer la importancia del estudio conjunto de varias variables.
- Calcular distribuciones marginales y condicionadas a partir de la conjunta.
- Comprender el concepto de independencia entre variables.
- Conocer medidas de dependencia lineal e interpretar su signo.

Descripción breve del tema

1. Introducción
2. Distribuciones de frecuencias
 - Distribución conjunta
 - Distribuciones marginales
 - Distribuciones condicionadas
 - Independencia entre variables
3. Representaciones gráficas
 - Diagrama de dispersión
 - Otras representaciones
4. Medidas de dependencia lineal
 - Covarianza
 - Correlación

Introducción

- El estudio conjunto de dos variables, nos permite analizar las relaciones entre ellas.
- La ausencia total de relación lleva el nombre de independencia.
- Puede predecirse una variable a partir de otra.
- Las representaciones gráficas son de gran ayuda.

Descripción breve del tema

1. Introducción
2. **Distribuciones de frecuencias**
 - Distribución conjunta
 - Distribuciones marginales
 - Distribuciones condicionadas
 - Independencia entre variables
3. Representaciones gráficas
 - Diagrama de dispersión
 - Otras representaciones
4. Medidas de dependencia lineal
 - Covarianza
 - Correlación

Distribuciones de frecuencias

Podemos tener cualquier combinación de variables cuantitativas / cualitativas.

Partimos de dos variables X e Y tales que

- X toma k valores distintos, $x_1, \dots, x_k$
(si cuantitativa $x_1 < \dots < x_k$)
- Y toma r valores distintos, $y_1, \dots, y_r$
(si cuantitativa $y_1 < \dots < y_r$)
- **Frecuencia absoluta conjunta del par** (x_i, y_j)
número de veces que el dato (x_i, y_j) aparece en la muestra, n_{ij}
- **Frecuencia relativa conjunta del par** (x_i, y_j)
cociente de frecuencia absoluta entre tamaño, $f_{ij} = n_{ij}/n$

Distribución conjunta de frecuencias

Podemos construir una tabla (de contingencia o de doble entrada) para la distribución conjunta de frecuencias de X e Y (con frec. relativas o absolutas).

$X \backslash Y$	y_1	y_2	$\dots$	y_r
x_1	n_{11}	n_{12}	$\dots$	n_{1r}
x_2	n_{21}	n_{22}	$\dots$	n_{2r}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_k	n_{k1}	n_{k2}	$\dots$	n_{kr}

Distribución conjunta de frecuencias PAU 2005, Madrid

UNI. \ RES.	Aprobados	Suspensos	UNI. \ RES.	Aprobados	Suspensos
Alcalá de Henares	2780	617	Alcalá de Henares	0'09	0'02
Autónoma de Madrid	7203	1664	Autónoma de Madrid	0'233	0'054
Carlos III	2734	636	Carlos III	0'088	0'021
Complutense de Madrid	9686	2608	Complutense de Madrid	0'313	0'084
Rey Juan Carlos	2189	796	Rey Juan Carlos	0'071	0'026

Ignacio Cascos

Depto. Estadística, Universidad Carlos III

9

Distribuciones marginales

Se obtienen al estudiar cada variable por separado.

Nos interesa sólo el comportamiento de X (o de Y).

□ Frecuencias absolutas marginales

$n_{i.}$ número de veces que la variable X toma el valor x_i en la muestra. Es la suma de n_{ij} con j de 1 a r .

$n_{.j}$ número de veces que la variable Y toma el valor y_j en la muestra. Es la suma de n_{ij} con i de 1 a k .

□ Frecuencias relativas marginales

cociente de frecuencia absoluta entre tamaño, $f(x_i)=f_{i.}=n_{i.}/n$

cociente de frecuencia absoluta entre tamaño, $f(y_j)=f_{.j}=n_{.j}/n$

Ignacio Cascos

Depto. Estadística, Universidad Carlos III

10

Distribuciones marginales PAU 2005

UNI. \ RES.	Aprobados	Suspensos	Total	UNI.	$n_{i.}$	$f_{i.}$
Alcalá de Henares	2780	617	3397	ual	3397	0'11
Autónoma de Madrid	7203	1664	8867	uam	8867	0'287
Carlos III	2734	636	3370	uc3m	3370	0'109
Complutense de Madrid	9686	2608	12294	ucm	12294	0'398
Rey Juan Carlos	2189	796	2985	urjc	2985	0'096
Total	24592	6321	30913	RES.	$n_{.j}$	$f_{.j}$
				Aprobado	24592	0'796
				Suspensos	6321	0'204

Ignacio Cascos

Depto. Estadística, Universidad Carlos III

11

Distribuciones condicionadas

Nos interesan sólo los individuos que toman el valor x_i en la variable X .

La variable Y definida en ese conjunto se denomina **variable condicionada** y se representa $Y|X=x_i$

□ Frecuencias relativas condicionadas

la frecuencia relativa de y_j condicionada a $X=x_i$ representa la proporción de individuos que presentan $Y=y_j$ de entre los que tienen $X=x_i$; $f(y_j|x_i)=n_{ij}/n_{i.}=f(x_i, y_j)/f(x_i)$

la frecuencia relativa de x_i condicionada a $Y=y_j$ representa la proporción de individuos que presentan $X=x_i$ de entre los que tienen $Y=y_j$; $f(x_i|y_j)=n_{ij}/n_{.j}=f(x_i, y_j)/f(y_j)$

Ignacio Cascos

Depto. Estadística, Universidad Carlos III

12

Distribuciones condicionadas

RESULTADO UNIV=ual	n_{1j}	$f(y_j x_1)$	RESULTADO UNIV=uam	n_{2j}	$f(y_j x_2)$
Aprobados	2780	0'818	Aprobados	7203	0'812
Suspensos	617	0'182	Suspensos	1664	0'188

RESULTADO UNIV=ucm	n_{4j}	$f(y_j x_4)$	RESULTADO UNIV=urjc	n_{5j}	$f(y_j x_5)$
Aprobados	9686	0'788	Aprobados	2189	0'733
Suspensos	2608	0'212	Suspensos	796	0'267

Ignacio Cascos

Depto. Estadística, Universidad Carlos III

13

Distribuciones condicionadas

RESULTADO UNIV=uc3m	n_{3j}	$f(y_j x_3)$
Aprobados	2734	0'811
Suspensos	636	0'189

RESULTADO	$n_{.j}$	$f_{.j}$
Aprobados	24592	0'796
Suspensos	6321	0'204

En la tabla superior se estudia la variable Resultado condicionada a que la variable Universidad tome el valor Carlos III.

En la tabla inferior se estudia la variable marginal Resultado.

Ignacio Cascos

Depto. Estadística, Universidad Carlos III

14

Independencia entre variables

Dos variables son independientes si no existe ninguna relación entre ellas.

A partir del valor que toma una variable no obtenemos ninguna información sobre el valor que tomará la otra.

La distribución de frecuencias relativas condicionadas de $Y|X=x_i$ es la misma, independientemente del valor que tome i

$$f(y_j|x_i) = f(y_j) \text{ para todo } i, j$$

X e Y independientes si $f(x_i, y_j) = f(y_j|x_i)f(x_i) = f(x_i)f(y_j)$ para todo i, j

Ignacio Cascos

Depto. Estadística, Universidad Carlos III

15

Independencia entre variables

Ejemplo: Realizamos una encuesta a 280 personas para estudiar si el hábito de fumar está relacionado con el sexo del individuo (si es más frecuente en hombres o en mujeres).

H\F	0	1	$n_{i.}$	H\F	0	1	$f_{i.}$
0	72	48	120	0	9/35	6/35	3/7
1	96	64	160	1	12/35	8/35	4/7
$n_{.j}$	168	112	280	$f_{.j}$	3/5	2/5	1

Ignacio Cascos

Depto. Estadística, Universidad Carlos III

16

Descripción breve del tema

1. Introducción
2. Distribuciones de frecuencias
 - Distribución conjunta
 - Distribuciones marginales
 - Distribuciones condicionadas
 - Independencia entre variables
3. Representaciones gráficas
 - Diagrama de dispersión
 - Otras representaciones
4. Medidas de dependencia lineal
 - Covarianza
 - Correlación

Diagrama de dispersión

En una nube de puntos o diagrama de dispersión, cada dato (x_i, y_j) de la muestra, se representa por un único punto de abscisa x_i y ordenada y_j

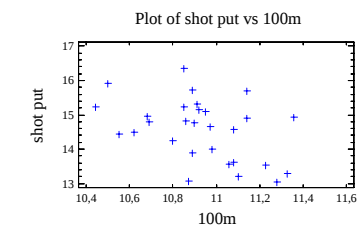
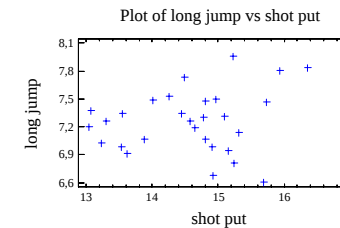
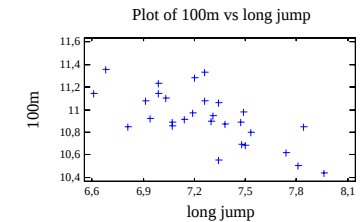
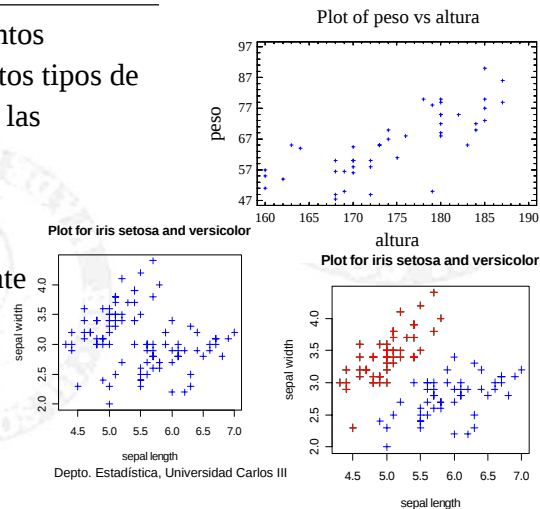


Diagrama de dispersión

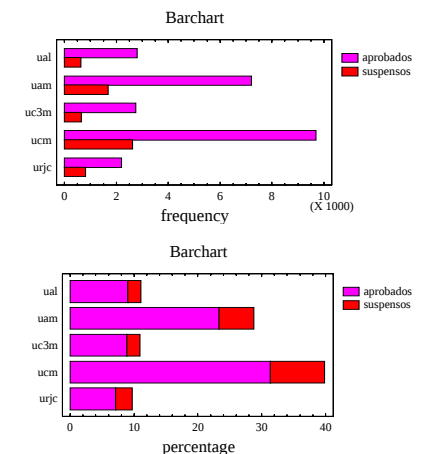
En una nube de puntos podemos ver distintos tipos de dependencias entre las variables.

También podemos observar visualmente la presencia de mezcla de poblaciones



Diagramas de barras agrupados

Para variables cualitativas o cuantitativas discretas con pocos valores, la representación más habitual son los diagramas de barras agrupados



Descripción breve del tema

1. Introducción
2. Distribuciones de frecuencias
 - Distribución conjunta
 - Distribuciones marginales
 - Distribuciones condicionadas
 - Independencia entre variables
3. Representaciones gráficas
 - Diagrama de dispersión
 - Otras representaciones
4. **Medidas de dependencia lineal**
 - Covarianza
 - Correlación

Medidas de dependencia lineal

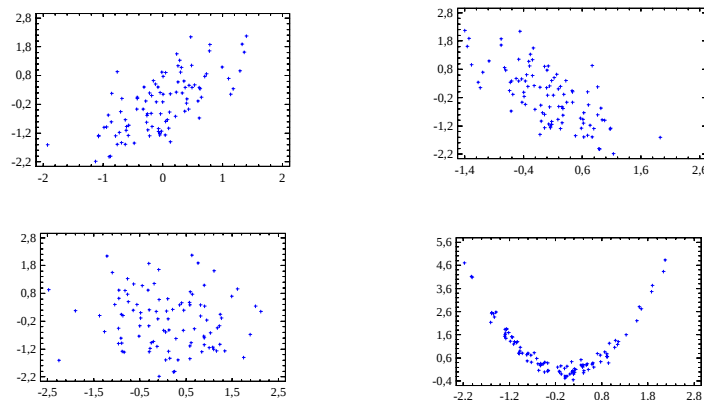
- **Covarianza:** mide la asociación lineal entre dos variables

$$s_{XY} = \sum_{i=1}^k \sum_{j=1}^r f_{ij} (x_i - \bar{x})(y_j - \bar{y}) \quad ; \quad s_{XY} = \sum_{i=1}^k \sum_{j=1}^r f_{ij} x_i y_j - \bar{x} \cdot \bar{y}$$

Si X e Y son independientes, entonces $s_{XY} = 0$, pero el recíproco no es cierto.

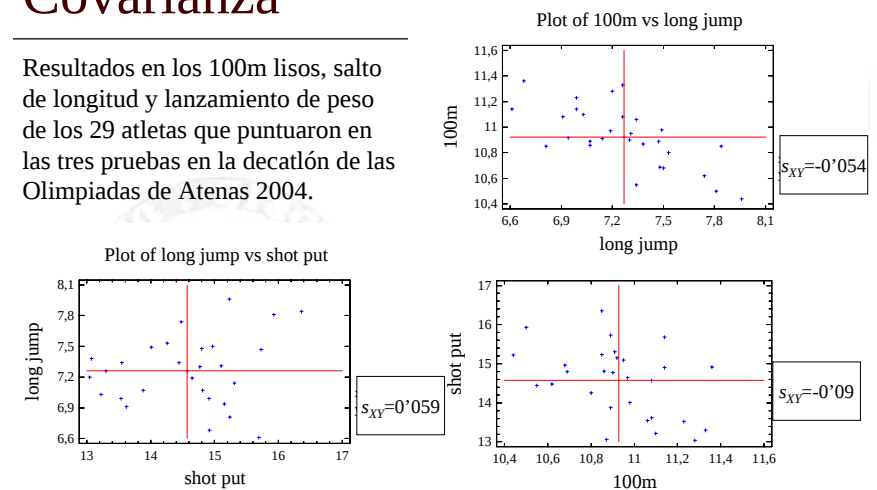
La covarianza viene dada en la unidad producto de las de X e Y y su signo resulta muy interesante.

Signo de la covarianza



Covarianza

Resultados en los 100m lisos, salto de longitud y lanzamiento de peso de los 29 atletas que puntuaron en las tres pruebas en la decatlón de las Olimpiadas de Atenas 2004.



Correlación

- **Correlación:** mide la asociación lineal entre dos variables y es **adimensional**.

Tiene el mismo signo que la covarianza.

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

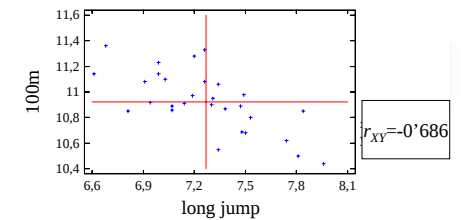
$$-1 \leq r_{XY} \leq 1$$

- Si $r_{XY} = 0$ X e Y se dicen incorreladas.
- Si $r_{XY} = 1$ ó $r_{XY} = -1$, la relación lineal entre X e Y es perfecta.

Correlación

Resultados en los 100m lisos, salto de longitud y lanzamiento de peso de los 29 atletas que puntuaron en las tres pruebas en la decatlón de las Olimpiadas de Atenas 2004.

Plot of 100m vs long jump



Plot of long jump vs shot put

