Tema 3: Análisis de datos bivariantes

1. Representaciones y gráficos:
   - Tabla de frecuencia absoluta / tabla de frecuencia relativa;
   - Frecuencias marginales y condicionales
   - Diagrama de dispersion

2. Resumen numérico:
   - Covarianza
   - Coeficiente de correlación
   - Recta de regresión lineal

Lecturas recomendadas:

- Capítulos 7 a 9 del libro de Peña y Romo (1997)
- Capítulo 9 del libro de Portilla (2004)

Motivación

En el tema 2, estudiemos las características de una variable. No obstante, en muchas situaciones medimos dos o más variables conjuntamente:

Número de idiomas habladas y Provincia de Nacimiento
Población y Escaños de una comunidad

Además de analizar las variables de manera individual, queremos ver si hay relación entre ellas.

Datos (x_1,y_1), (x_2,y_2), …, (x_N,y_N)

3.1: Representaciones y gráficos

X = Número de idiomas habladas (1,2,3)
Y = Provincia de Nacimiento (Cataluña, Galicia, Pais Vasco, Otro)

Resultados de 40 personas:

(1,O) (2,C) (2,G) (1,G) (2,P) (2,C) (1,O) (2,O) (2,C) (3,P)
(2,C) (2,G) (1,G) (1,O) (2,O) (1,P) (2,C) (2,P) (2,O) (2,P)
(3,C) (2,G) (1,O) (1,O) (2,O) (2,C) (2,P) (3,C) (2,G) (2,P)
(1,O) (1,G) (1,O) (2,C) (3,C) (2,P) (2,G) (1,G) (2,C) (1,O)

La tabla de doble entrada

X / Y	C	G	P	O	
1	0	4	1	8	
2	8	5	6	4	
3	3	0	1	0	
					40

Hay 40 personas en la muestra

Hay tres Catalanes que hablan tres idiomas.

La tabla con frecuencias relativas

X / Y	C	G	P	O	
1	0	0,1	0,025	0,2	
2	0,2	0,125	0,15	0,1	
3	0,075	0	0,025	0	
					1

¿Que hacemos si no nos interesa la comunidad de origen?

Las frecuencias marginales

X / Y	C	G	P	O	Total
1	0	4	1	8	13
2	8	5	6	4	23
3	3	0	1	0	4
Total	11	9	8	12	40

¿Cuál es el número medio de idiomas habladas?

¿Y si sólo nos interesa el número de idiomas que hablan los gallegos?

X / Y	C	G	P	O	Total
1	0	0,1	0,025	0,2	0,325
2	0,2	0,125	0,15	0,1	0,575
3	0,075	0	0,025	0	0,1
Total	0,275	0,225	0,2	0,3	1

Las frecuencias condicionadas

X dado Y=G	Frecuencia	Frec. Rel.
1	4	0,44444444
2	5	0,55555556
3	0	0
Total	9	1

= 0,125 / 0,225 es la proporción de gallegos que hablan dos idiomas

¿Cuál es el número medio de idiomas habladas por los gallegos?

¿Hay diferencia con el resultado anterior?

¡Ojo!

Muchas (la mayoría) de las tablas que salen en la prensa, encuestas, … son tablas de frecuencias condicionadas.

Pregunta 1

En los últimos seis meses, ¿ha adquirido Ud. o algún miembro de su hogar alguno de los siguientes bienes?

	TOTAL	RECUERDO DE VOTO EN ELECCIONES GENERALES DE 2011													
		PP	PSOE	IU	UPyD	CiU	Otros	No tenía edad	En blanco	Voto nulo	No votó	No recuerda	N.C.	No tiene la nacionalidad	No especifica la nacionalidad
Automóvil/moto															
Sí, él/ella ha adquirido	3,4	3,1	3,9	4,8	0,0	0,0	6,6	3,7	3,8	11,1	4,0	1,4	2,4	4,7	0,0
Sí, lo ha adquirido otra persona de su hogar	2,6	4,6	2,5	1,6	0,0	5,3	4,9	5,6	3,8	11,1	0,9	0,0	1,7	1,6	0,0
No	94,0	92,3	93,6	93,7	100,0	94,7	88,5	90,7	92,3	77,8	95,2	98,6	95,9	93,8	100,0
N.S	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
N.C	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
(N)	(1404)	(260)	(204)	(63)	(24)	(19)	(61)	(54)	(52)	(9)	(227)	(69)	(296)	(64)	(2)

TOTAL RECUERDO DE VOTO EN ELECCIONES GENERALES DE 2011

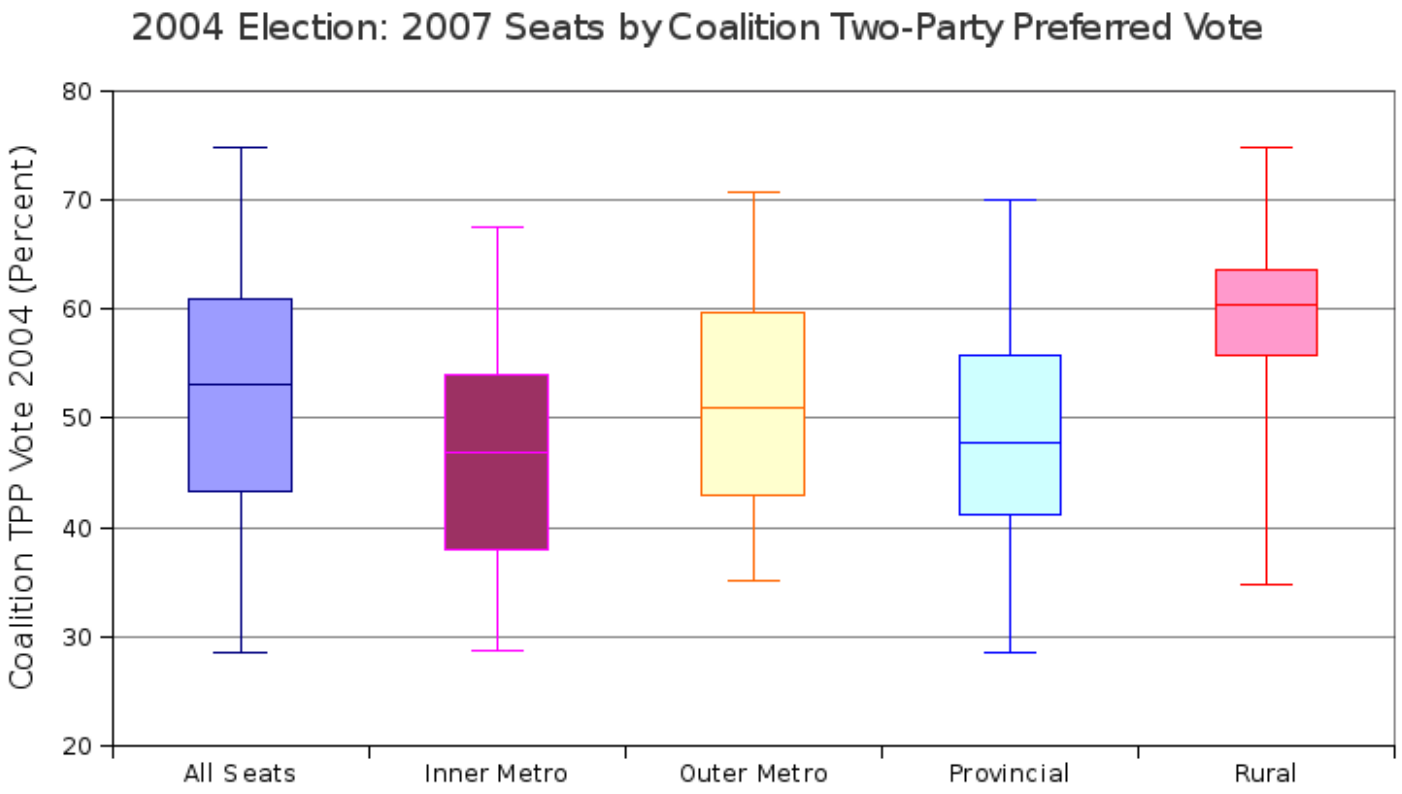
¿Cuántas personas han adquirido un coche?
¿Cuántas votantes del PSOE han comprado un coche?

CIS septiembre 2013

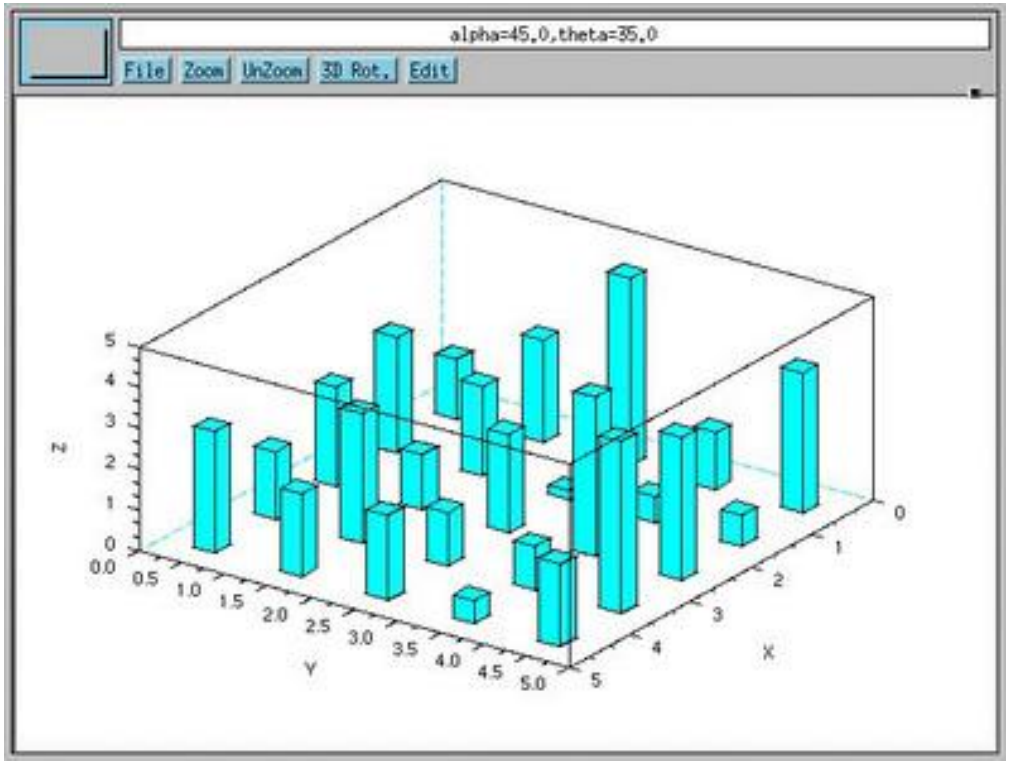
Resúmenes gráficos

Múltiples diagramas de caja

Sirven para medir una variable cuantitativa y una cualitativa



Histogramas tridimensionales



Diagramas de dispersión

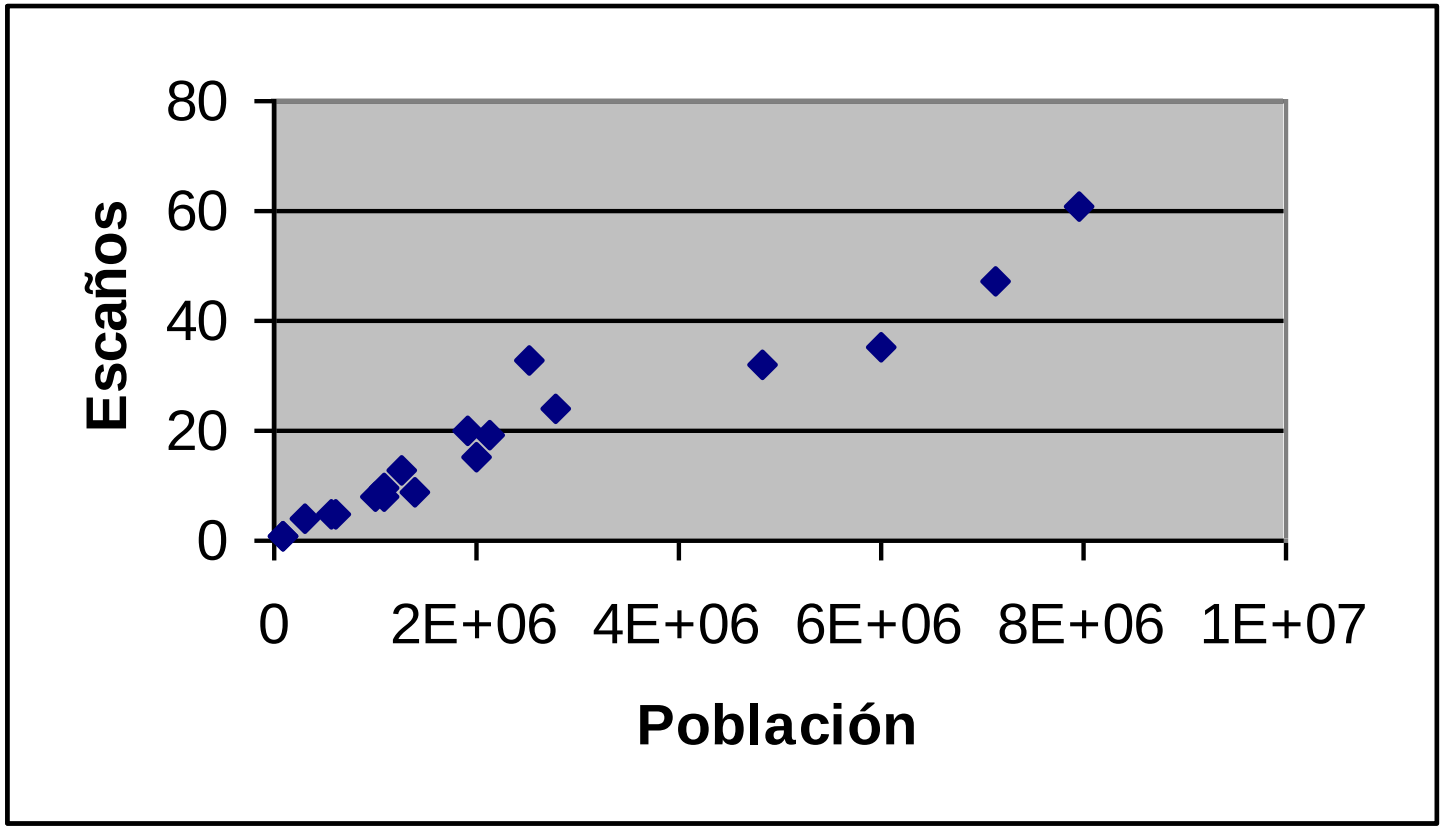
¿El número de escaños está relacionado con la población?

Comunidad	Población *(enero 2006)*	% población española *(enero 2006)*	Escaños en el Congreso	% de escaños en el Congreso
Andalucía	7.975.672	17,84%	61	17,43%
Cataluña	7.134.697	15,96%	47	13,43%
Madrid	6.008.183	13,44%	35	10%
Valencia	4.806.908	10,75%	32	9,14%
Galicia	2.767.524	6,19%	24	6,86%
Castilla y León	2.523.020	5,64%	33	9,43%
País Vasco	2.133.684	4,77%	19	5,43%
Canarias	1.995.833	4,46%	15	4,28%
Castilla-La Mancha	1.932.261	4,32%	20	5,71%
Murcia	1.370.306	3,06%	9	2,57%
Aragón	1.277.471	2,86%	13	3,71%
Extremadura	1.086.373	2,43%	10	2,85%
Asturias	1.076.896	2,41%	8	2,29%
Baleares	1.001.062	2,24%	8	2,29%
Navarra	601.874	1,35%	5	1,43%
Cantabria	568.091	1,27%	5	1,43%
La Rioja	306.377	0,69%	4	1,14%
Ceuta	75.861	0,17%	1	0,29%
Melilla	66.871	0,15%	1	0,29%

Comunidades cuya representación parlamentaria es **inferior** al porcentaje de población dentro de la sociedad española.

Comunidades cuya representación parlamentaria es **superior** al porcentaje de población dentro de la sociedad española.

El diagrama de dispersión



Hay una relación aproximadamente lineal. ¿Cómo medirla?

Ejercicio

	Tabaquismo Y_j				
Género X_i		Fumador	No fumador	Ex fumador	Total
	Varón	30	50	20	100
	Mujer	30	10	10	50
	Total	60	60	30	150

Hallar la distribución de frecuencias relativas y las distribuciones marginales.

Parece existir alguna relación entre sexo y hábitos de tabaquismo?

Ejercicio

Completar el siguiente cuadro, sabiendo que de un total de 200 personas, 92 viven en zonas urbanas, mientras 20 hombres lo hacen en zonas rurales, y que el total de hombres es 98.

	ZONA URBANA	ZONA RURAL	TOTAL
HOMBRES			
MUJERES			
TOTAL			

a) Expresa el cuadro de forma tal que te permita estudiar la distribución de hombres y mujeres al interior de la zona urbana y la rural, o sea considerando el "perfil columna". ¿Qué te sugiere este estudio?
b) Expresa el cuadro de forma tal que te permita estudiar la distribución de los residentes en la zona urbana y rural, al interior del grupo de hombres y del de mujeres, o sea considerando el "perfil fila". ¿Qué te sugiere este estudio?
c) Se va a realizar un estudio específico con las personas de la zona rural; ¿qué porcentaje de éstas son hombres?
d) Se va a realizar un estudio específico con las personas de la zona urbana; ¿qué porcentaje de éstas son hombres?
e) Compara los resultados del puntos c y d. ¿Qué conclusiones sacas a partir de ellos?
f) ¿Cuántas mujeres viven en la zona urbana?; ¿qué porcentaje del total de personas constituyen?; ¿qué porcentaje del total de su zona?

Ejercicio (Pregunta de Test)

La Encuesta de Pobreza y Desigualdades Sociales realizada por el Gobierno Vasco viene investigando en Euskadi el fenómeno de la pobreza. El objetivo central de la EPDS es el conocimiento, estudio y evaluación de las distintas líneas de pobreza, y de su incidencia en Euskadi, así como la obtención de indicadores asociados de desigualdad social. Se ha realizado la encuesta a 1000 personas y están interesados en analizar la relación entre dos de ellas:

X: ¿Con el ingreso mínimo para llegar a final de mes usted como diría que viviría su hogar?

1. Muy pobre, 2. Pobre, 3. Apañándoselas, por debajo de la media, 4. Por encima de la media, confortable, 5. Rico

Y: ¿Se les agotan los alimentos que compran y no tienen dinero para conseguir más?

1. A menudo, 2. Algunas veces, 3. Nunca

Se ha obtenido la siguiente tabla conjunta de frecuencias relativas:

Su hogar vive ...	**Se agotan los alimentos ...**		
	A menudo	Algunas veces	Nunca
Muy pobre	0.05	0.03	0
Pobre	0.15	0.1	0.01
Apañándoselas, por debajo de la media	0.1	0.15	0.2
Por encima de la media, confortable	0.03	0.05	0.05
Rico	0	0.01	0.07

Señala cual de las siguientes afirmaciones es cierta:

a) 33 de los entrevistados afirmaron que nunca se les agotan los alimentos
b) 800 de los entrevistados afirmaron que su hogar viviría pobre
c) 330 de los entrevistados afirmaron que se les agotan los alimentos a menudo
d) 13 de los entrevistados afirmaron que su hogar viviría por encima de la media, confortable

Ejercicio (Pregunta de Test)

Un instituto estadístico ha hecho una encuesta de gente de entre 18 y 20 años para ver sus intenciones de voto en las siguientes elecciones generales. El número de personas en la muestra fue de 3000. Se quiere estudiar la relación entre intención de voto y edad. Sea X = Partido político e Y = edad:

	18	19	20
Conservative	450	300	210
Labour	250	270	330
Liberal	145	170	200
Nationalist groups	95	150	180
Independents	50	115	85

¿Cuál de las siguientes afirmaciones es la correcta?

a) Un 33.5% de la gente en la muestra tienen menos de 20 años.
b) Un 5.67% de la gente en la muestra son *Liberals* de 19 años de edad.
c) Un 10% de la gente en la muestra son *Independents*.
d) Un 10% de la gente de 19 años son *Conservatives*.

Ejercicio (Pregunta de Test)

Siguiendo de la pregunta anterior, marca la respuesta correcta.

a) Un 31.76% de la gente que pretenden votar *Labour* tienen 20 años.
b) Un 29.41% de la gente que pretenden votar *Labour* tienen 19 años.
c) Un 29.41% de la gente que pretenden votar *Labour* tienen 18 años.
d) Un 38.82% de la gente que pretenden votar *Labour* tienen 19 años.

	18	19	20
Conservative	450	300	210
Labour	250	270	330
Liberal	145	170	200
Nationalist groups	95	150	180
Independents	50	115	85

Ejercicio (Pregunta de Examen)

Considera la variable X que expresa el nivel de educación de una persona y la variable Y que refleja la cantidad de euros que gasta por sus vacaciones por año.

La siguiente tabla muestra la distribución conjunta de frecuencias absolutas de X e Y:

	Y=[0,200)	Y=[200,400)	Y=[400,600)	Y=[600,800)	Y=[800,1000]
X=”sin finalización de estudios”	8	13	8	9	5
X=”con finalización de estudios”	8	9	3	6	3
X=”con finalización de estudios y de la formación universitaria”	3	5	5	7	8

Total: 100

b) Considerando las personas que tienen la finalización de estudios (pero no la formación universitaria), calcular la porcentaje de ellos que hayan pagado entre 800 y 1000 euros por sus vacaciones por año. (**0,5 puntos**)
c) Calcula la distribución marginal de frecuencias de Y. (**0,5 puntos**)
d) Estimar la media de Y. **(0,5 puntos)**

Ejercicio (Pregunta de Examen)

En el siguiente recuadro se presentan los resultados (porcentajes) a la pregunta sobre el uso de Internet como fuente de información acerca de política o de sociedad teniendo en cuenta el género de la persona encuestada.

	TOTAL	SEXO	
		HOMBRE	MUJER
Usa internet para obtener información acerca de la política o la sociedad			
Todos los días	13,8	17,0	10,8
3-4 días por semana	7,5	7,4	7,5
1-2 días por semana	5,7	5,5	6,0
Con menor frecuencia	5,6	5,7	5,6
Nunca	66,3	63,3	68,9
N.S.	0,7	0,7	0,7
N.C.	0,5	0,5	0,5
(N)	(2479)	(1219)	(1260)

¿Cuál de las siguientes afirmaciones es correcta?

(a) Aproximadamente 807 hombres encuestados nunca utilizan Internet para obtener información acerca de la política o la sociedad.

(b) El 27,8% de las personas encuestadas utilizan, todos los días, Internet para obtener información acerca de la política o la sociedad.

(c) Aproximadamente 63 mujeres encuestadas no contestaron (N.C.) a esta pregunta.

(d) Ninguna de las anteriores.