Técnicas Estadísticas de Análisis de Datos

- Descripción de datos. Estadísticos de una variable
- Distribuciones de probabilidad e intervalos de confianza
  - Contrastes de hipótesis. Tipos
- Relaciones entre atributos
  - **Nominales- Numéricos**: Tests de comparación de medias (muestras dependientes e independientes) y análisis de varianza.
  - **Numéricos - Numéricos**: Análisis de Regresión
  - **Nominales-Nominales**: Tablas de Contingencia. Tests de independencia y comparación de proporciones.
- Aplicación de técnicas estadísticas a la clasificación
  - Clasificación mediante regresión numérica
  - Clasificador bayesiano

Análisis de una variable (muestra de datos)

- **Estadísticos:** resumen (describen) toda la información contenida en una muestra de datos :
  - Variables continuas
    - medidas centrales (media, moda, mediana)
    - medidas de dispersión (rango, varianza, desviación estándar, percentiles)
    - medidas de forma (histograma)
  - Variables nominales
    - frecuencias relativas (probabilidades), moda
    - media y varianza de probabilidad estimada
- **Muestra**: y_i; i =1…n; toma valores en un rango continuo/discreto

Estadísticos centrales

- **Media** (esperanza) muestral: promedio de todos los valores

$$\mathrm{media}(y) = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- **Moda**: valor que aparece más veces
- **Mediana**: valor que deja el mismo número de casos a ambos lados

$$\mathrm{mediana}(y) = y_i \mid \mathrm{N^o\ casos}\left(y_j \le y_i\right) = \mathrm{N^o\ casos}\left(y_k \ge y_i\right)$$

  - equivale a ordenar el vector de datos y tomar el valor central
  - menos sensible frente a valores extremos poco probables

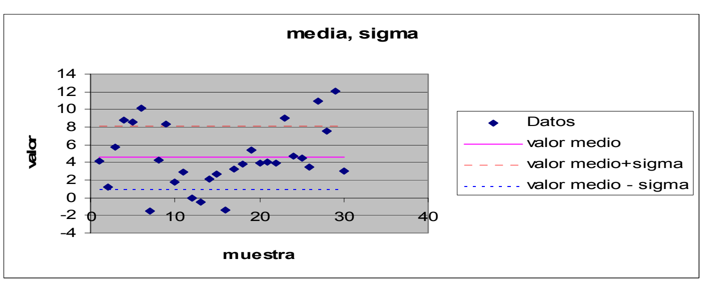
Estadísticos de dispersión

- **Recorrido (intervalo, o rango):**
  max(yi)-min(yi)
- **Varianza:** promedio de desviaciones con respecto a valor medio

$$\mathrm{Var}(y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i{}^2 - n\bar{y}^2\right]$$

- **Desviación estándar** (típica): raíz cuadrada de la varianza

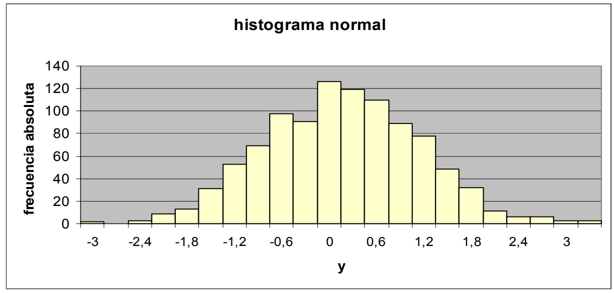
$$\mathrm{desv}(y) = \sigma_y = \sqrt{\mathrm{Var}(y)}$$



Histograma

Estimación de la distribución de densidad de probabilidad:
frecuencia absoluta o relativa de valores de y_i por unidad de intervalo

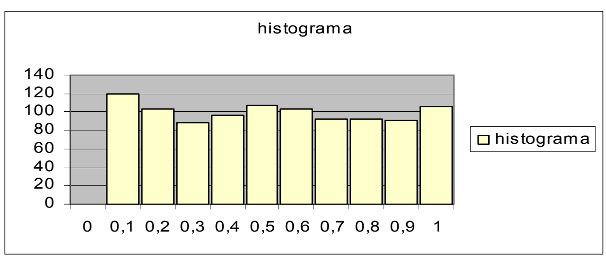
Nº de casos en intervalo

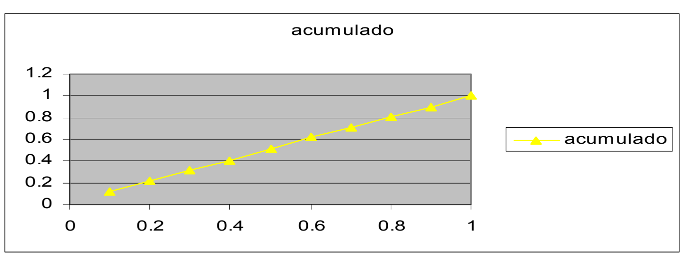


intervalos de clase

La suma total de frecuencias absolutas es el número de datos
La suma de frecuencias relativas es 1

Ejemplo: histograma de variable uniforme



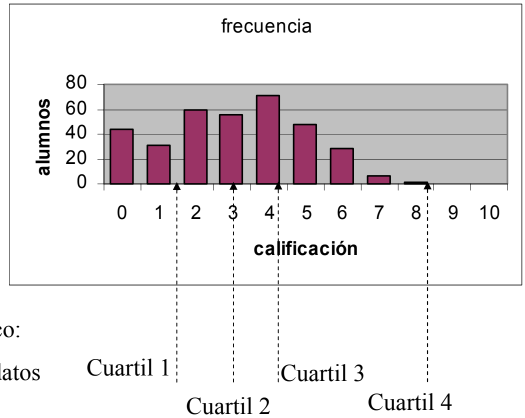


Cuantiles del histograma

- **Cuantil:** valores que dividen el recorrido de datos en k partes de la misma frecuencia (percentiles: 100 partes, cuartiles: 4 partes, etc.)
- Ejemplo: cuartiles

Calificación
2,8
0,6
5
3,1
3,9
4,9
1
0
6,55
...

porcentaje	cuartiles
0,25	1,4
0,5	2,725
0,75	4
1	7,7



Recorrido inter-cuartílico:

[1.4, 4]: contiene 50% datos

Estadísticos de variable nominal

- y_i nominal: toma valores de un conjunto discreto (categorías): $\{v_{i1}, \ldots, v_{iki}\}$
- **Distribución de frecuencias** de cada valor

$$p_1 = 100(n_1 / n)\%$$
$$p_2 = 100(n_2 / n)\%$$
$$\vdots$$
$$p_{ki} = 100(n_{ki} / n)\%$$
$$n = \sum_{j=1}^{k_i} n_j$$

- **Moda**: valor que aparece más veces

$$\max_{j} (n_j)$$

Media y varianza de frecuencias estimadas

- Cálculo de cada frecuencia
  - para una categoría dada: m casos de n

    $p = m/n$
  - puede verse como asignar: $v_i=1$ cada ejemplo en la categoría

    $v_i=0$ en el resto

$$p = \frac{1}{n}\sum_{i=1}^{n} v_i$$

- Varianza de p:

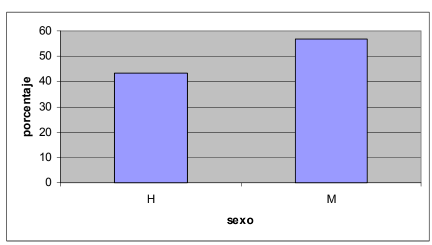
$$\mathrm{Var}(p) = \frac{1}{n}\sum_{i=1}^{n}(v_i - p)^2 = p(1-p)$$

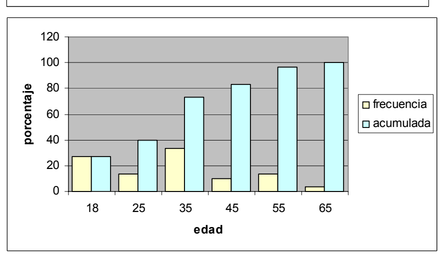
$$\sigma_p = \sqrt{p(1-p)}$$

  - caso máxima varianza: $p=0.5$

Ejemplo variable nominal y numérica

Edad	Sexo
23	M
25	M
18	H
37	M
45	H
62	H
43	M
40	H
60	M
54	H
28	H
18	H
54	M
29	H
42	M
26	M
32	M
41	M
37	M
36	H
53	H
21	M
24	H
21	H
45	M
64	H
22	M
61	M
37	M
66	M





Distribución Normal

- Curva de gran interés por explicar datos en muchas situaciones
  - Aplicada por primera vez como distribución por A. Quetelet (1830)

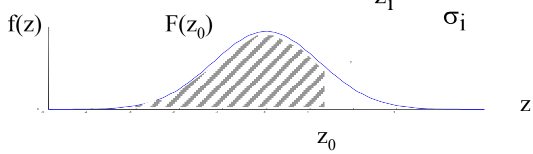
$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} z^2\right]$$

- distribución simétrica: coincide media y mediana en 0
- se dispone del valor de la distribución de probabilidad: área bajo la curva de $f_Z(z)$ para cualquier valor:

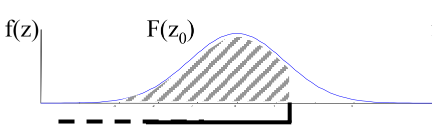
z	$F_Z(z)$
-3	0.001349967
-2.5	0.00620968
-2	0.022750062
-1.5	0.066807229
-1	0.15865526
-0.5	0.308537533
0	0.5
0.5	0.691462467
1	0.84134474
1.5	0.933192771
2	0.977249938
2.5	0.99379032
3	0.998650033

Tipificar o estandarizar variables: Se mide el desplazamiento respecto a la media en unidades de desviación típica:

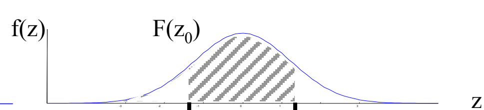
$$z_i = \frac{y_i - \bar{y}}{\sigma_i}$$



Distribución Normal e Intervalos de Confianza



Una cola (unilateral)



Simétrico dos colas (bilateral)

- Ej.: se conocen parámetros de una población con distribución normal: media: μ= 115; desviación típica:σ= 20
  - ¿casos inferiores a 70? z=(70-115)/20, F(z)=0,012
  - ¿casos superiores a 150? z=(150-115)/20, 1-F(z)=0,04
  - ¿en intervalo 90-130? F((130-115)/20)-F((90-115)/20)=0,667
  - ¿qué intervalos simétrico tienen el 80%, 95% de los casos (intervalos de confianza)? $z=F^{-1}(\alpha/2)$; $y=\mu \pm z\sigma$
    - 80%: $z_{0.1}$=1,28; 115± $z_{0.1}$*20=[89.3, 140.6]
    - 95%: $z_{0.025}$=1,96; 115± $z_{0.025}$*20=[75.8, 154.2]

RELACIONES DE VARIABLES. TEST DE HIPOTESIS

ANÁLISIS DE VARIAS VARIABLES

- Objetivo: analizar la interrelación (dependencia) entre los valores de distintas variables, haciendo uso de los datos disponibles
  - Numéricas (retardo, carga, distancia,…)
  - Nominales (tipo de avión, condición visibilidad, …)
- Herramienta de análisis: tests de hipótesis
  - **Numéricas-numéricas**: análisis de regresión y covarianza
  - **Nominales-nominales**: tablas de contingencia
  - **Nominales-numéricas**: comparación de medias, análisis de varianza

ANÁLISIS ESTADÍSTICO DE DATOS

ANÁLISIS DE VARIAS VARIABLES - NUMÉRICA-NUMÉRICA

- Permite identificar relaciones entre variables numéricas y construir modelos de regresión
- Se consideran relaciones de una variable de salida (dependiente) con múltiples variables de entrada (independientes)
- Estimación de una función (**Regresión Lineal**) que mejor “explique” los datos

$$\{(\vec{X}_1, y_1), (\vec{X}_2, y_2), \ldots, (\vec{X}_n, y_n)\}$$

$$\vec{X}: \quad \text{vectores con M dimensiones}$$

$$g(.): \quad R^M \longrightarrow R$$

$$\vec{X} \longrightarrow \hat{y} = g(\vec{X})$$

Mínimos Cuadrados

- Estima vector de coeficientes que minimiza error

$$\hat{y}_i = g_i(\vec{X}) = a_0 + \sum_{p=1}^{I} a_p x_p = (\vec{A}^t) * \vec{X}$$

$$(\vec{A}) = [a_0 \quad a_1 \quad \cdots \quad a_I]^t; \quad \vec{X} = [1 \quad x_1 \quad \cdots \quad x_I]^t$$

- Objetivo: dadas N muestras, determinar coeficientes que minimicen el error de predicción global

$$\varepsilon = \sum_{j=1}^{n} [g(\vec{X}_j) - y_j]^2$$

- El método de mínimos cuadrados selecciona, como estimación de la recta de regresión poblacional, aquella para la cual esta suma de cuadrados es menor.
- Problema clásico de minimización de función cuadrática: solución única

Mínimos Cuadrados

- Solución genérica matricial

$$\vec{y} = \begin{bmatrix} y^1 \\ \vdots \\ y^N \end{bmatrix}; \quad \hat{g} = \begin{bmatrix} \hat{y}^1 \\ \vdots \\ \hat{y}^N \end{bmatrix} = \begin{bmatrix} g(\vec{X}^1) \\ \vdots \\ g(\vec{X}^N) \end{bmatrix} = \begin{bmatrix} 1 & x_1^1 & \cdots & x_I^1 \\ 1 & x_1^2 & \cdots & x_I^2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^N & \cdots & x_I^N \end{bmatrix} \vec{A} = H * \vec{A}$$

- Solución MC:

$$\vec{A} = [H^t H]^{-1} H^t \vec{y}$$

[(1+F)x1] = [(1+F)xN] [Nx(1+F)] [(1+F)xN] [Nx1]

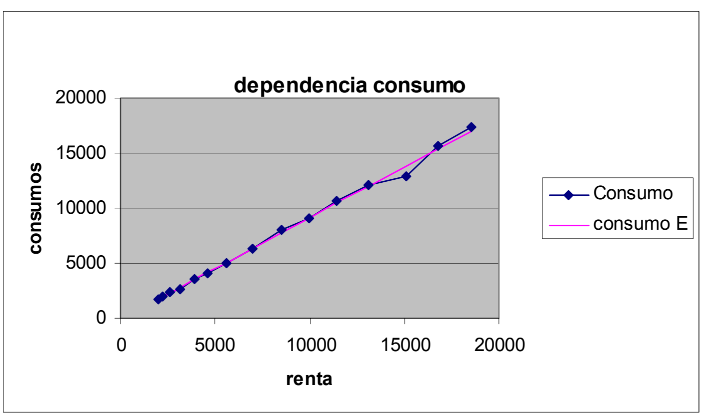
Ejemplo: regresión lineal de 1 variable

Año	Renta	Consumo	consumo E
1970	1959,75	1751,87	1683,473374
1971	2239,09	1986,35	1942,43325
1972	2623,84	2327,9	2299,11261
1973	3176,06	2600,1	2811,043671
1974	3921,6	3550,7	3502,190468
1975	4624,7	4101,7	4153,993607
1976	5566,02	5012,6	5026,63666
1977	6977,84	6360,2	6335,452914
1978	8542,51	7990,13	7785,967518
1979	9949,9	9053,5	9090,676976
1980	11447,5	10695,4	10479,01488
1981	13123,04	12093,8	12032,31062
1982	15069,5	12906,27	13836,76054
1983	16801,6	15720,1	15442,48976
1984	18523,5	17309,7	17038,76316

Estimación Lineal	
a1	a0
0.927041871	-133.296932

$$\text{ConsumoE} = a0 + a1 * \text{Renta}$$

Ejemplo: regresión lineal de 1 variable



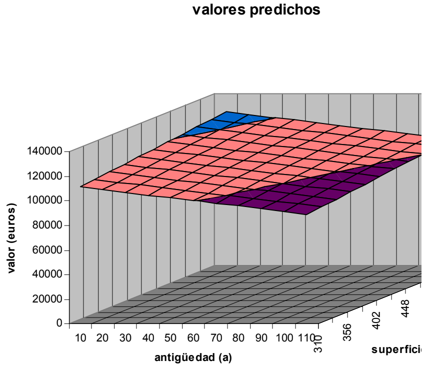
Ejemplo: regresión lineal de 2 variables

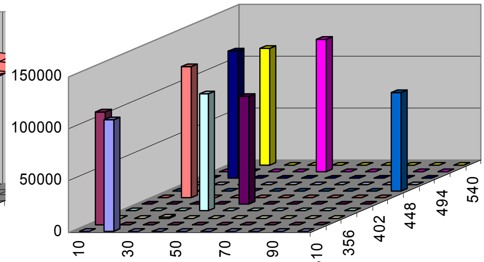
x1 Superficie	x2 Antigüedad	y Valor	Valor predicho
310	20	106,287 Euros	109,180 Euros
333	12	107,784 Euros	112,283 Euros
356	33	113,024 Euros	108,993 Euros
379	43	112,275 Euros	108,128 Euros
402	53	104,042 Euros	107,262 Euros
425	23	126,497 Euros	115,215 Euros
448	99	94,311 Euros	99,800 Euros
471	34	106,961 Euros	115,469 Euros
494	23	122,006 Euros	119,233 Euros
517	55	126,497 Euros	113,518 Euros
540	22	111,527 Euros	122,132 Euros

Estimación Lineal		
a2	a1	a0
-220.444829	58.2271936	95538.7217

$$\text{Valor} = a0 + a1 * \text{Superficie} + a2 * \text{Antigüedad}$$

Ejemplo: regresión lineal de 2 variables

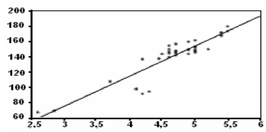


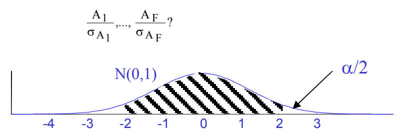


Evaluación del modelo de regresión

Análisis de validez del modelo asumido:

- Medidas de "parecido" entre variable de salida estimada y real, influencia de variables de entrada
  - Factor de Correlación
  - Error de predicción
- Análisis de "calidad" del modelo
  - Error en coeficientes
  - Hipótesis de significatividad de parámetros: t-Student





Factor de correlación

- Factor de correlación entre datos y predicciones:

$$\mathrm{Corr}(\hat{y}, y) = \frac{1}{\sqrt{S_{\hat{y}} S_y}} \sum_{j=1}^{n} (\hat{y}_j - \bar{\hat{y}})(y_j - \bar{y}) = \frac{\mathrm{Cov}(\hat{y}, y)}{\sqrt{\mathrm{Var}(\hat{y})\mathrm{Var}(y)}}$$

  - El factor de correlación varía entre -1 y 1.
- En general, se puede hacer factores de correlación entre cualquier par de variables numéricas: indica el grado de relación lineal existente.
  - -1: existe asociación lineal negativa perfecta.
  - 1 positiva perfecta.
  - 0 no hay asociación lineal.

Matrices de covarianza y correlación

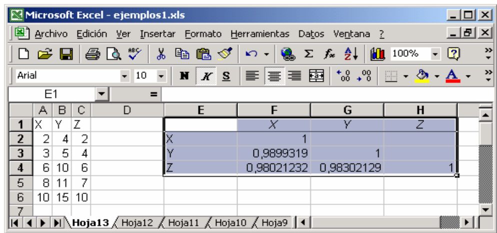
Muestra de vectores aleatorios: $\{\vec{X}_1, \vec{X}_2, ..., \vec{X}_n\}$

- Matriz de covarianzas:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\vec{X}_i$$

$$\hat{C}_{\vec{X}} = \frac{1}{n}\sum_{i=1}^{n}(\vec{X}_i - \hat{\mu})(\vec{X}_i - \hat{\mu})^t = \begin{bmatrix} var(x_1) & cov(x_1, x_2) & \cdots & cov(x_1, x_2) \\ cov(x_1, x_2) & var(x_1) & & \\ \vdots & & \ddots & \vdots \\ cov(x_1, x_I) & \cdots & & var(x_I) \end{bmatrix}$$

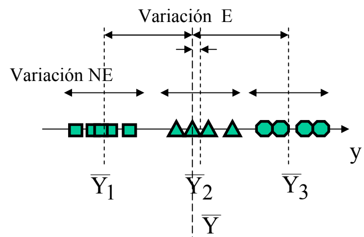
- La matriz de correlaciones es similar, normalizada



ANÁLISIS ESTADÍSTICO DE DATOS

ANÁLISIS DE VARIAS VARIABLES - NUMÉRICA-NOMINAL

- Mide la relación entre variables numéricas y nominales, o nominales y nominales (proporciones)
- Analiza las diferencias de medias condicionadas a variable nominal: impacto de la variable nominal sobre la continua
- Dos tipos de análisis:
  - Con dos medias o proporciones: significatividad de la diferencia **t-student**
  - Más de dos valores distintos: **Análisis de Varianza**

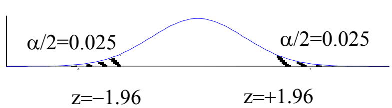


1. Comparación de dos medias

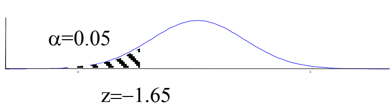
- Se plantea como un test de hipótesis, dividiendo los datos en dos grupos, cada uno con su media y varianza.
- Hipótesis sobre diferencia de medias: $D = \bar{y}_1 - \bar{y}_2$
  - H_0: la diferencia de medias en la población es nula D=0.
  - Hipótesis alternativa A: las medias son distintas: D!=0.
  - Hipótesis alternativa B: la media de 1 es mayor que 2: $\bar{y}_1 > \bar{y}_2$
  - Hipótesis alternativa C: la media de 1 es menor que 2: $\bar{y}_1 < \bar{y}_2$
- Situaciones posibles:
  - Muestras independientes: conjuntos distintos.
  - Muestras dependientes: mismo conjunto, con dos variables a comparar en cada ejemplo.

Contrastes de dos medias

- Hipótesis alternativa A

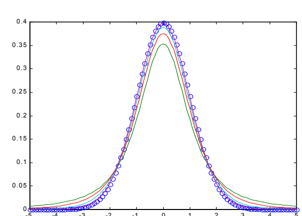


- Hipótesis alternativa B:



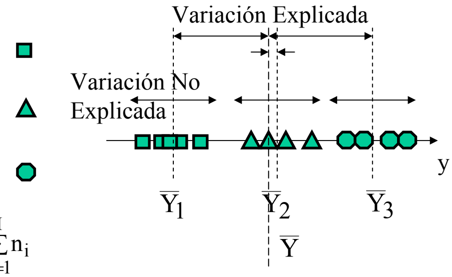
- Cuando las muestras son pequeñas no es válida la hipótesis de normalidad de los estadísticos de medias

$$\bar{y} \pm t_{\alpha/2,GL}\,\sigma$$



2. Análisis de varianza (ANOVA)

Niveles	Observaciones
1	$Y_{11}, Y_{12}, \ldots Y_{1j}, \ldots Y_{1n1}$
...	...
i	$Y_{i1}, Y_{i2}, \ldots Y_{ij}, \ldots Y_{ini}$
...	...
I	$Y_{I1}, Y_{I2}, \ldots Y_{Ij}, \ldots Y_{InI}$



- Número total de elementos: $n = \sum_{i=1}^{I} n_i$
- Media por nivel: $\overline{Y}_i = \frac{1}{n_i} \sum_{i=1}^{I} Y_{ij}$
- Media total: $\overline{Y} = \frac{1}{n} \sum_{i=i}^{I} \sum_{j=1}^{n_i} Y_{ij}$

variación explicada: variabilidad entre grupos

variación no explicada (residual): variabilidad dentro de los grupos

- Relación entre "cuadrados":

$$\sum_{i=i}^{M} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y})^2 = \sum_{i=i}^{M} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2 + \sum_{i=1}^{M} n_i (\overline{Y}_i - \overline{Y})^2$$

ANÁLISIS ESTADÍSTICO DE DATOS

ANÁLISIS DE VARIAS VARIABLES - NOMINAL-NOMINAL

- Analiza la interrelación entre los valores de variables nominales según distribución de casos
- Herramienta para dos variables: **tabla de contingencia**
  - distribución de casos (frecuencias) para las distintas combinaciones de valores de las dos variables

	variable 2				totales 1
variable 1	valor 1	valor 2	…	valor p2	
valor 1	n_{11}	n_{12}	…	n_{1p2}	t1
valor 2	n_{21}	n_{22}	…	n_{2p2}	t2
…	…	…	…	…	…
valor p1	n_{p11}	n_{p12}	…	n_{p1p2}	tp1
totales 2	t'1	t'2	…	t'p2	t

Probabilidades marginales:

$P_i = t_i/t$

Casos "esperados"

$E_{ij} = t(t_i/t)(t'_j/t) = t_i t'_j/t$

Probabilidades marginales: $P_j = t'_j/t$

Relación entre variables nominales-nominales

- Objetivo: analizar la interrelación (dependencia) entre los valores de variables nominales
- Herramienta para dos variables: **tabla de contingencia**
  - distribución de casos (frecuencias) para las distintas combinaciones de valores de las dos variables

	variable 2				totales 1
variable 1	valor 1	valor 2	…	valor p2	
valor 1	n_{11}	n_{12}	…	n_{1p2}	t1
valor 2	n_{21}	n_{22}	…	n_{2p2}	t2
…	…	…	…	…	…
valor p1	n_{p11}	n_{p12}	…	n_{p1p2}	tp1
totales 2	t'1	t'2	…	t'p2	t

Probabilidades marginales:

$P_i=t_i/t$

Probabilidades marginales:
$P_j=t'_j/t$

Estimación del nº esperado de observaciones

$E_{ij}=t(t_i/t)(t'_j/t)= t_i t'_j/t$

Contraste Chi-2 de variables nominales

- Es aplicable en análisis bi-variable (normalmente clase vs atributo)
- Determina si es rechazable la hipótesis de que dos variables son independientes
  - Bajo hipótesis H_0 se determinan los casos en el supuesto de variables independientes. Los valores esperados se determinan con probabilidades marginales de las categorías: $E_{ij}=tP_i\ P_j$ (valores esperados).
  - Nuestro contraste de hipótesis nula de no asociación estará basado en las magnitudes de las diferencias entre los valores observados y los esperados bajo la hipótesis nula.
  - El estadístico Chi-cuadrado mide la diferencia entre los valores observados y los valores esperados.

$$\chi^2 = \sum_{i=1}^{p1}\sum_{j=1}^{p2}(O_{ij}-E_{ij})^2 / E_{ij}$$

Ejemplo

Microsoft Excel - Libro1

Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ?

J20 =

reales	SEXO		
VOTO	varón	mujer	total
PP	115	165	280
CDS	20	21	41
PSOE	367	364	731
IU	104	76	180
total	606	626	1232

diferencias	SEXO	
VOTO	varón	mujer
PP	3.75037504	3.63055475
CDS	0.00138633	0.00134204
PSOE	0.15367406	0.14876435
IU	2.69987046	2.61361262

chi2	prob (3 gl)
12.9995797	0.00463751

esperados	SEXO	
VOTO	varón	mujer
PP	137.727273	142.272727
CDS	20.1672078	20.8327922
PSOE	359.566558	371.433442
IU	88.538961	91.461039

Hoja1 Hoja2 Hoja3

Listo NUM

EJEMPLOS VALIDACIÓN HIPÓTESIS

ANÁLISIS DE VARIAS VARIABLES - NOMINAL-NUMÉRICA

VUELO	COMP.	FECHA	entrada A	salida A	…
X-1322	IB	30/07/2005	11:33	11:52	…
C-1144	KLM	30/07/2005	12:01	12:15	
…	…	…			

– Hay relación entre tiempo en retardo y: franja horaria (mañana-tarde-noche), tipo de día (diario-finsemana), compañía …

• Mayor grado de relación?

tiempo sector	mañana	tarde	noche
	11,3	10	12,4
	8,5	6,4	9,4
	15,6	8,5	8,2
	…	…	…
media	9	11,3	8,5
desv std	1,4	0.8	1.8

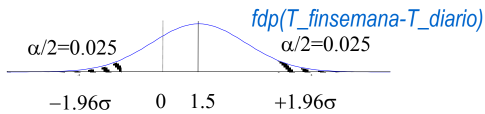
tiempo sector	diario	finsemana
	11,3	12,4
	6,4	8,5
	15,6	10
	…	…
media	8,5	7,5
desv std	1.1	1.2

EJEMPLOS VALIDACIÓN HIPÓTESIS

ANÁLISIS DE VARIAS VARIABLES - NOMINAL-NUMÉRICA

Hipótesis (análogo a comparación de prestaciones!)

- **Hipótesis nula H_0**: la diferencia de medias según tipo día es nula D=0
- Hipótesis alternativa: las medias son distintas: D!=0



tiempo sector	diario	finsemana
	11,3	12,4
	6,4	8,5
	15,6	10
	...	...
media	8,5	7,5
desv std	1.1	1.2

- Mayor grado de relación? Más evidencia estadística para rechazar la hipótesis de independencia

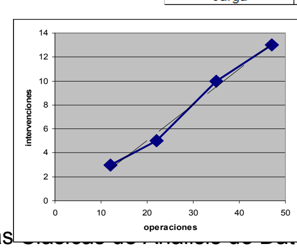
EJEMPLOS VALIDACIÓN HIPÓTESIS

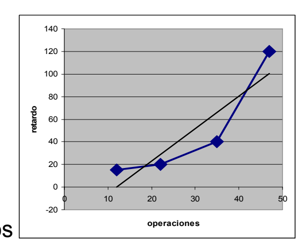
ANÁLISIS DE VARIAS VARIABLES - NUMÉRICA-NUMÉRICA

operaciones	retardo	intervenciones	carga
12	15	3	0,13
22	20	5	0,22
35	40	10	0,43
47	120	13	0,56
...	...	...	...

– Qué variables están "más linealmente" relacionadas …

CORRELACIÓN	*operaciones*	*retardo*	*intervenciones*	*carga*
operaciones	1,000			
retardo	0,899	1,000		
intervenciones	0,994	0,888	1,000	
carga	0,994	0,888	1,000	1,000





EJEMPLOS VALIDACIÓN HIPÓTESIS

ANÁLISIS DE VARIAS VARIABLES – NOMINAL-NOMINAL

– Dependencia entre grado de retardo y tipo de avión, visibilidad,…

VUELO	…	PESO	VISIB	T PLAN	T LLEG	RETARDO
X-1322	IB	mediano	IFR	15:45	16:10	medio
C-1144	KLM	ligero	VFR	12:15	12:31	medio
…	…	…	…		…	…

	Tipo avión		
Grado retardo	Ligero	Mediano	Pesado
	alto	nulo	alto
	nulo	alto	alto
	nulo	medio	muy alto
	muy alto	nulo	nulo
	…	…	…

tipo avión/ retardo	Ligero	Mediano	Pesado
nulo	75	323	15
medio	32	405	6
alto	8	45	7
muy alto	2	15	1

EJEMPLOS VALIDACIÓN HIPÓTESIS

ANÁLISIS DE VARIAS VARIABLES – NOMINAL-NOMINAL

– **Hipótesis nula H_0**: las variables retardo y categoría son independientes: $E_{ij}=t(t_i/t)(t'_j/t)$

tipo avión/ retardo	Ligero	Mediano	Pesado	**total**
nulo	75	323	15	**413**
medio	32	405	6	**443**
alto	8	45	7	**60**
muy alto	2	15	1	**18**
total	**117**	**788**	**29**	**934**

tipo avión/ retardo	Ligero	Mediano	Pesado	total
nulo	**51,74**	**348,44**	**12,82**	413
medio	**55,49**	**373,75**	**13,75**	443
alto	**7,52**	**50,62**	**1,86**	60
muy alto	**2,25**	**15,19**	**0,56**	18
total	117	788	29	934

$$\chi^2 = \sum_{i=1}^{p1}\sum_{j=1}^{p2} (E_{ij}-O_{ij})^2 / E_{ij}$$