

Introducción al Análisis Estadístico en Ciencias Sociales con



Adrián Arias Díaz-Faes - INGENIO (CSIC-UPV)



INGENIO [CSIC-UPV] Ciudad
Politécnica de la
Innovación | Edif 8E 4º
Camino de Vera s/n
46022 Valencia

tel +34 963 877 048
fax +34 963 877 991



ugr

Universidad
de Granada

ÍNDICE

Parte I

Estadística Descriptiva

Medidas de Tendencia Central

Estadística Inferencial

Parte II

Análisis de Correlación y Regresión lineal simple

Análisis de regresión múltiple

Análisis Estadístico - Parte I

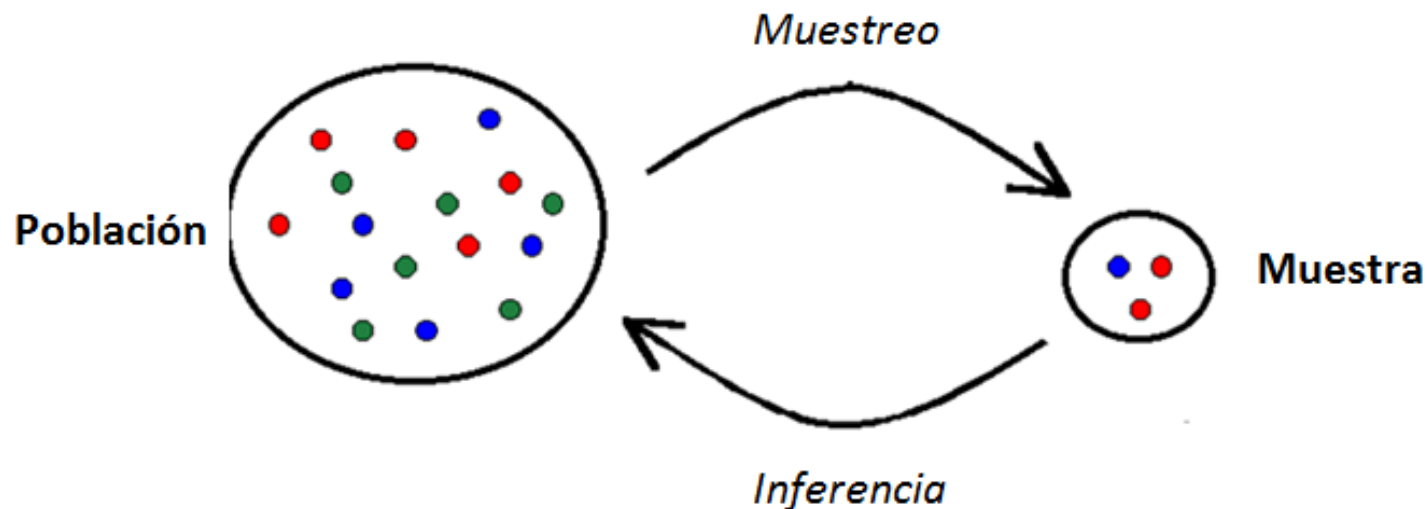
- Muestreo y Estadística Descriptiva
- Medidas de Tendencia Central
- Estadística Inferencial
 - T de Student y relacionados
 - ANOVA
 - Tablas de contingencia
- Tamaños del efecto

1. Población y muestra (i)

Población: Es el conjunto de individuos (sujetos, empresas, etc...), sobre los que se quiere estudiar una o más características. Habitualmente, no se tiene acceso a todos los sujetos, objetos, sucesos, etc., disponibles, sino con una parte de ellos

Aquella parte de elementos que realmente vamos a estudiar u observar, es lo que denominamos **Muestra** → debe representar a la población en sus características y en el comportamiento del fenómeno estudiado

Muestra: → subgrupo de la población de la que se recolectan los datos



1. Población y muestra (ii)

- ✓ El **tamaño muestral** es el nº de individuos que componen la muestra → ENCUESTAS

DEPENDE DE

- ✓ **Variabilidad de los datos**
- ✓ **Precisión de la estimación → Nivel de error:** la máxima diferencia admitida entre el verdadero valor del parámetro y el valor estimado a partir de la muestra (*suele estar entre un 3% y un 5%*)
- ✓ **Nivel de confianza (1-α):** nivel de certeza (probabilidad) de que nuestra muestra recoja el verdadero valor poblacional

Cuanto menor queremos que sea el error y mayor el nivel de confianza, mayor tendrá que ser el tamaño muestral

$$n = \frac{Z_{\alpha}^2 N p q}{e^2 (N - 1) + Z_{\alpha}^2 p q}$$

N = tamaño de la población

e = nivel de error

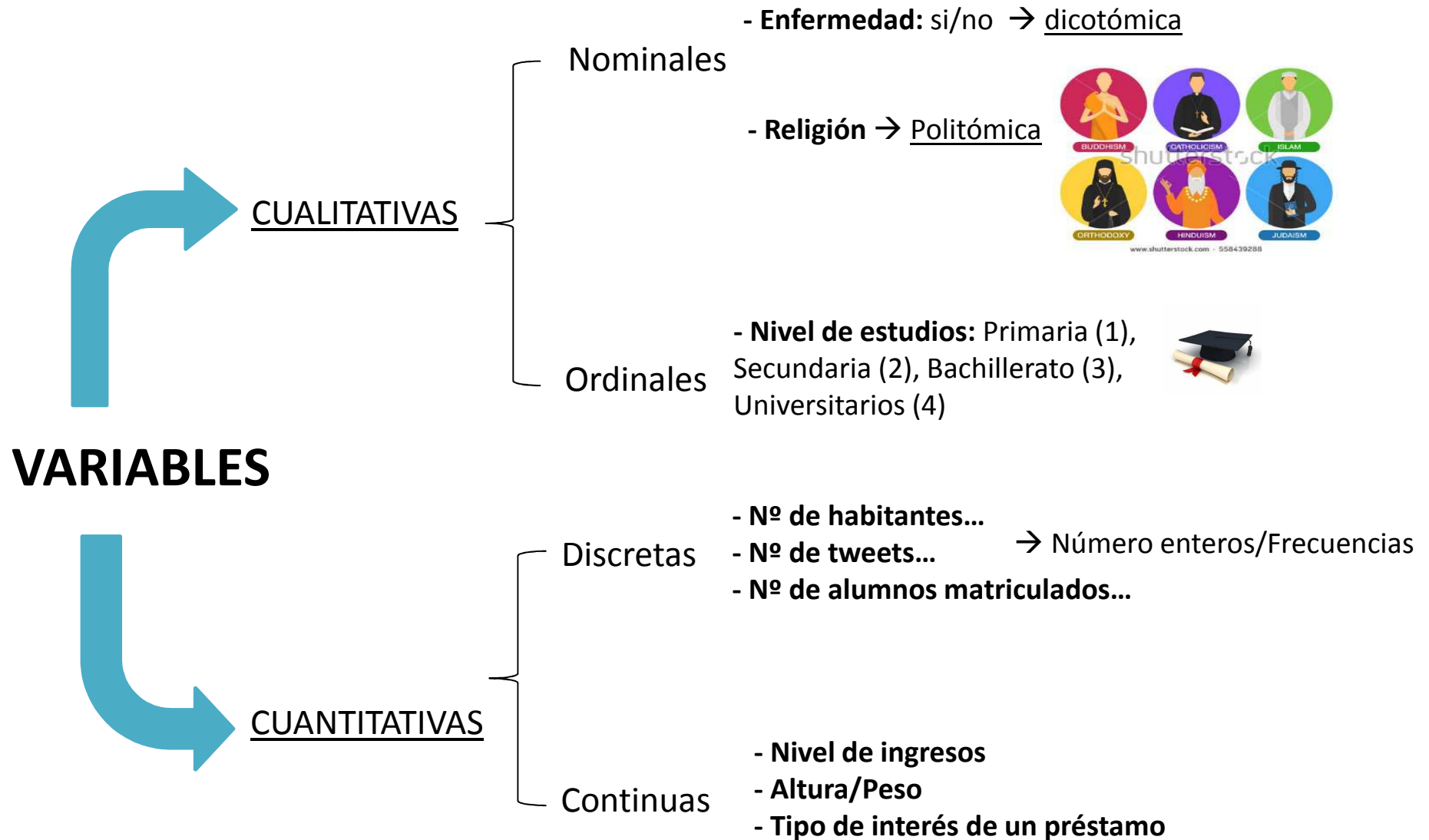
Z_{α} = constante que depende del nivel de confianza

$p=q= 0,5$



Nivel de confianza deseado	Puntuación z
80 %	1.28
85 %	1.44
90 %	1.65
95 %	1.96
99 %	2.58

2. Tipos de variables



Nº de hijos	Frecuencia absoluta (f_i)	Frecuencia absoluta acumulada (F_i)	Frecuencia relativa (h_i)	Frecuencia relativa acumulada (H_i)
0	65	65	0,325	0,325
1	21	86	0,105	0,430
2	33	119	0,165	0,595
3	35	154	0,175	0,770
4	21	175	0,105	0,875
5	17	192	0,085	0,960
6	8	200	0,040	1
Total	200		1,00	

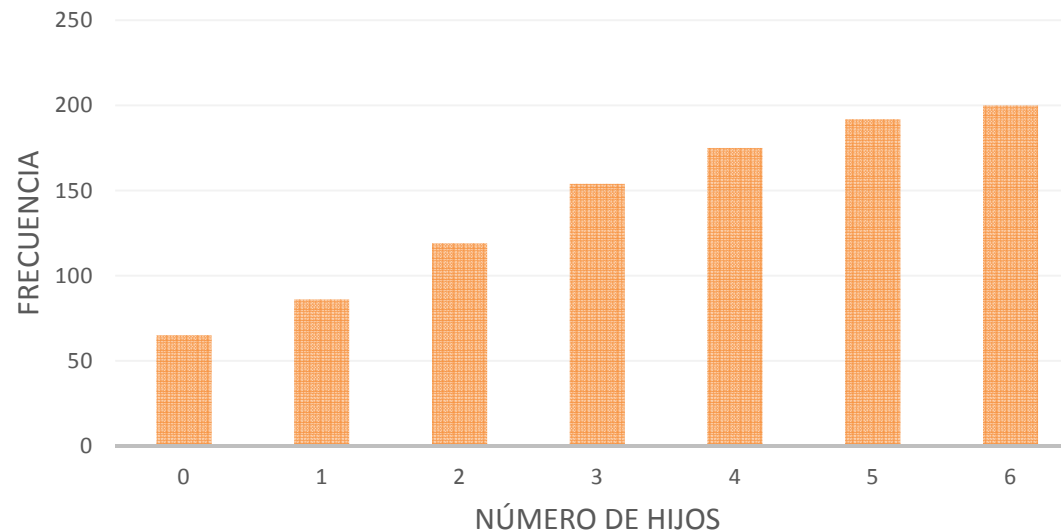
Frecuencia absoluta (f_i) → número de veces que se repite cada valor de la variable en el conjunto de todas las observaciones de la misma

Frecuencia relativa ($h_i = f_i/n$) → es el cociente entre la frecuencia absoluta y el número total de datos u observaciones

Frecuencia absoluta acumulada ($\sum_{j=1}^i f_j$) → es la suma de las frecuencias absolutas de los valores inferiores o iguales al considerado

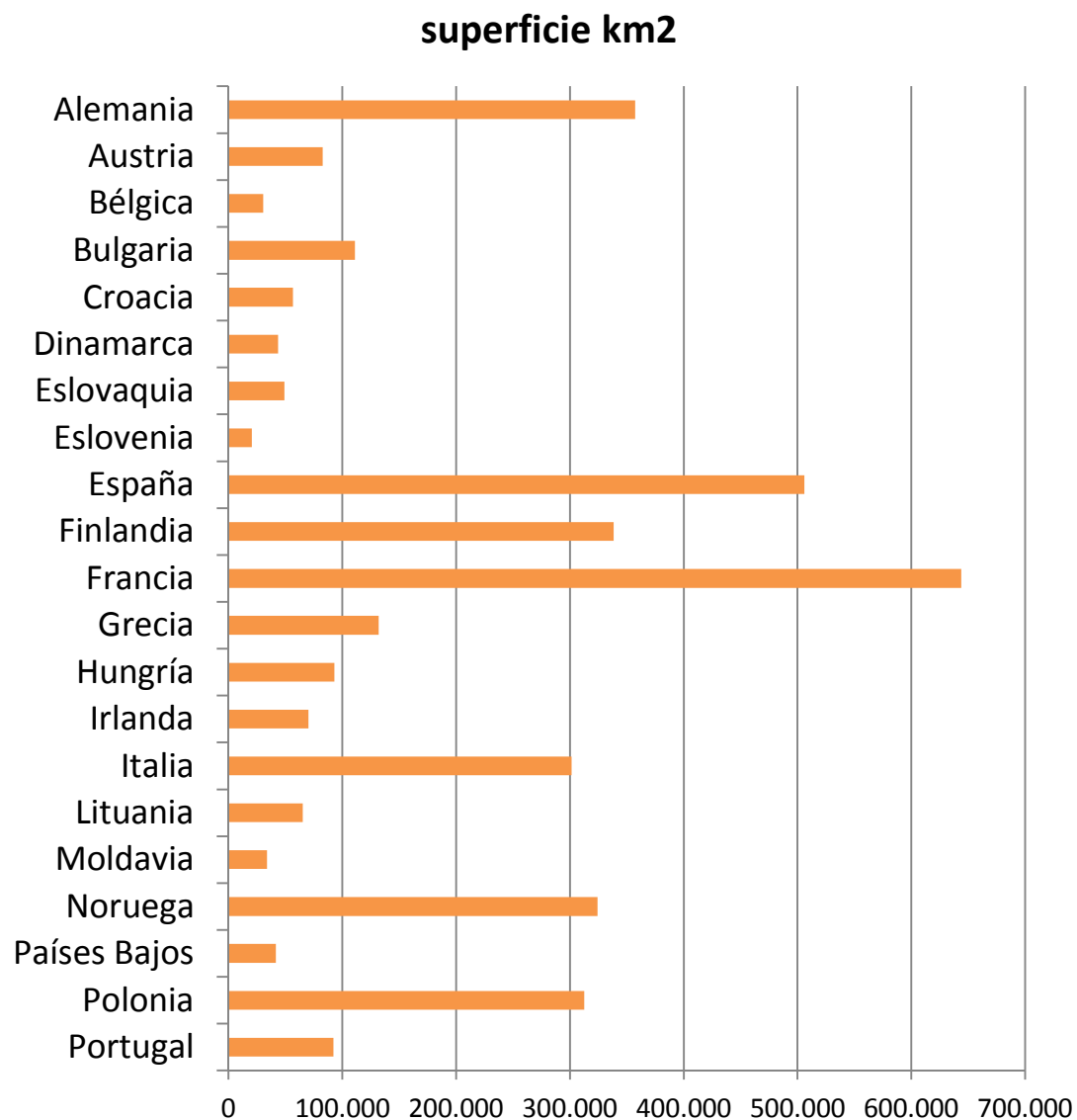
Discreta

Frecuencia acumulada



Continua

país	superficie km2
Alemania	357.376
Austria	82.700
Bélgica	30.513
Bulgaria	111.000
Croacia	56.594
Dinamarca	43.560
Eslovaquia	49.035
Eslovenia	20.521
España	505.990
Finlandia	338.524
Francia	643.801
Grecia	131.957
Países Bajos	41.543
Hungría	93.03
Irlanda	70.273
Italia	301.338
Lituania	65.300
Moldavia	33.700
Noruega	324.219
Polonia	312.677
Portugal	92.082




3. Medidas de tendencia central (i)

En cualquier análisis de datos el primer paso es hacer una **descripción de la muestra**

MEDIA: $\bar{x} = \frac{\sum x_i}{n}$ Si la variable es cuantitativa
Muy afectada por valores extremos

Medidas de dispersión: evaluar la representatividad de la media

✓ **VARIANZA** $s^2 = \frac{\sum (x_i - \bar{x})^2}{n}$ Como de dispersos están los datos respecto a la media
Expresado en el cuadrado de las unidades de la variable

✓ **DESVIACIÓN TÍPICA** $s = + \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$  Valor positivo de la raíz cuadrada de la varianza

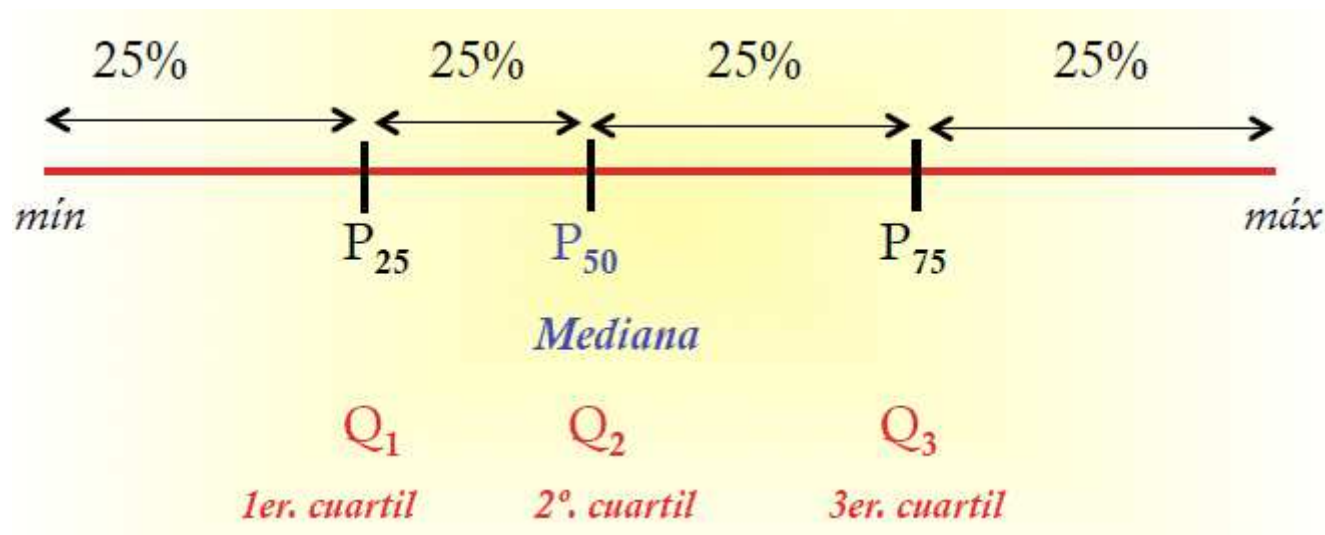
✓ **ERROR ESTÁNDAR** $ESM = s / \sqrt{n}$ Variabilidad de la media en el muestreo

3. Medidas de tendencia central (ii)

MEDIANA: valor de la distribución que, una vez ordenados los datos, ocupa el lugar central.

Nº impar de términos: $\{1, 2, 3, 4, 5\} \rightarrow M_e = 3$

Nº par de términos: $\{1, 2, 6, 7, 9, 10, 13, 14\} \rightarrow M_e = 7+9/2 = 8$



Variables cualitativas \rightarrow Frecuencias, datos ordinales

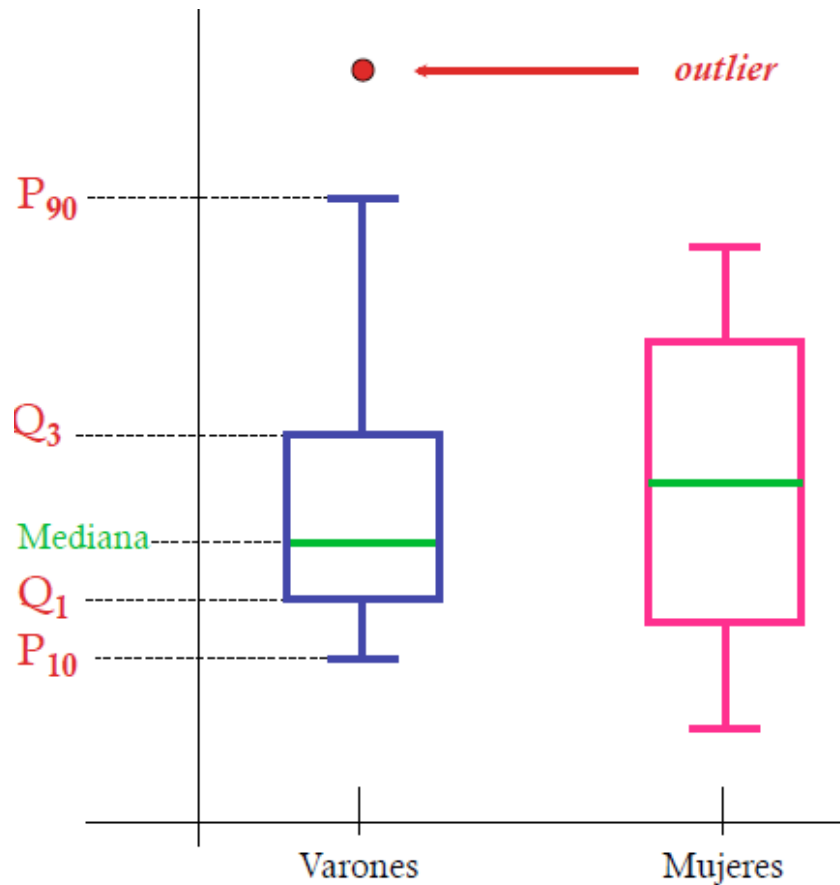
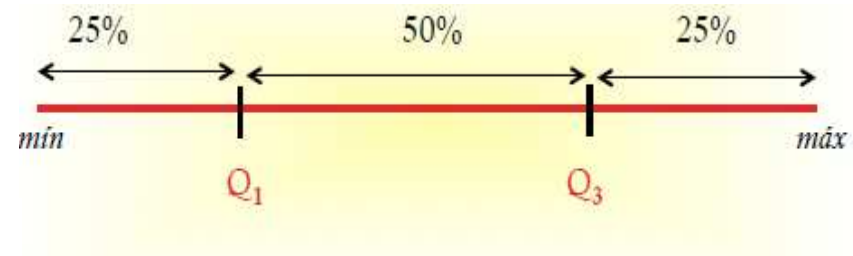
No influenciada por valores extremos \rightarrow Variables cuantitativas

Difícil de determinar en el caso de variables agrupadas en intervalos

3. Medidas de tendencia central (iii)

Medida de dispersión MEDIANA:

Rango intercuartílico = $Q_3 - Q_1$



Medidas de posición no centrales:

- ✓ Cuartiles (Q_i) → 25%
- ✓ Deciles (D_i) → 10%
- ✓ Percentiles (P_i) → 1%

4. Contrastes de hipótesis (i)

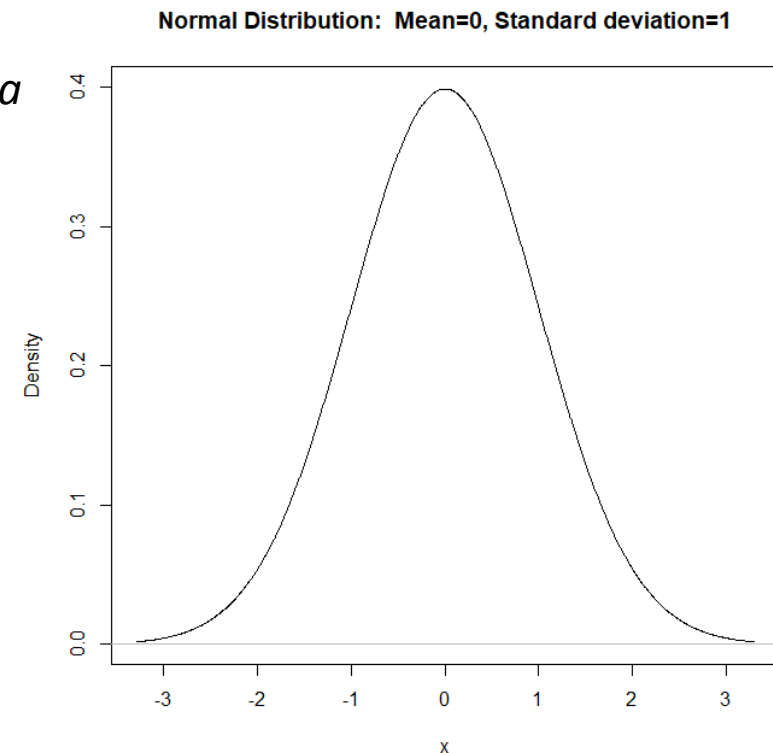
Se busca aceptar o rechazar una hipótesis estadística acerca de un parámetro o característica de la población, pero que se contrasta a partir de los resultados de una muestra de la población

Salario entre el personal técnico de la empresa X que es una multinacional

- ✓ *El salario de un técnico es de 28.000€*
- ✓ *Si medimos el salario de otros técnicos de la empresa ¿será exactamente el mismo?*
- ✓ *El salario, en este caso, es una cantidad que cambia (ligeramente) entre el personal técnico*
- ✓ *¿Cómo podemos modelar su comportamiento?*



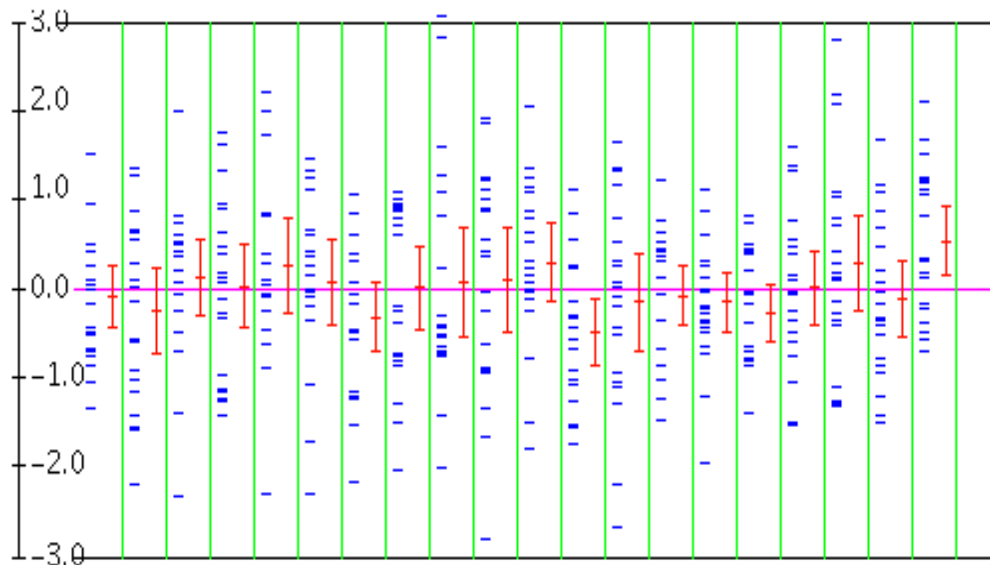
**Suponemos que la variable sigue una
DISTRIBUCIÓN NORMAL**



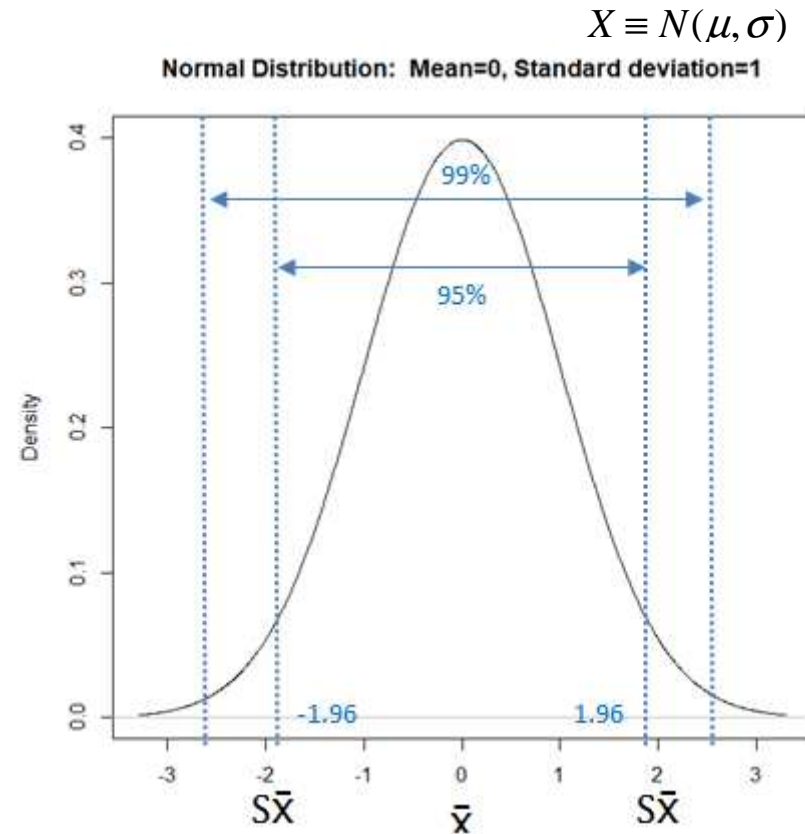
4. Contrastes de hipótesis (ii)

Intervalo de confianza

- ✓ Distintas muestras tienen distintos intervalos



- ✓ Si fijamos el intervalo de confianza en el 95% → la media de la población está entre los límites especificados con una probabilidad del 95%. El 5% no lo contienen



Nivel de confianza deseado	Puntuación z
80 %	1.28
85 %	1.44
90 %	1.65
95 %	1.96
99 %	2.58

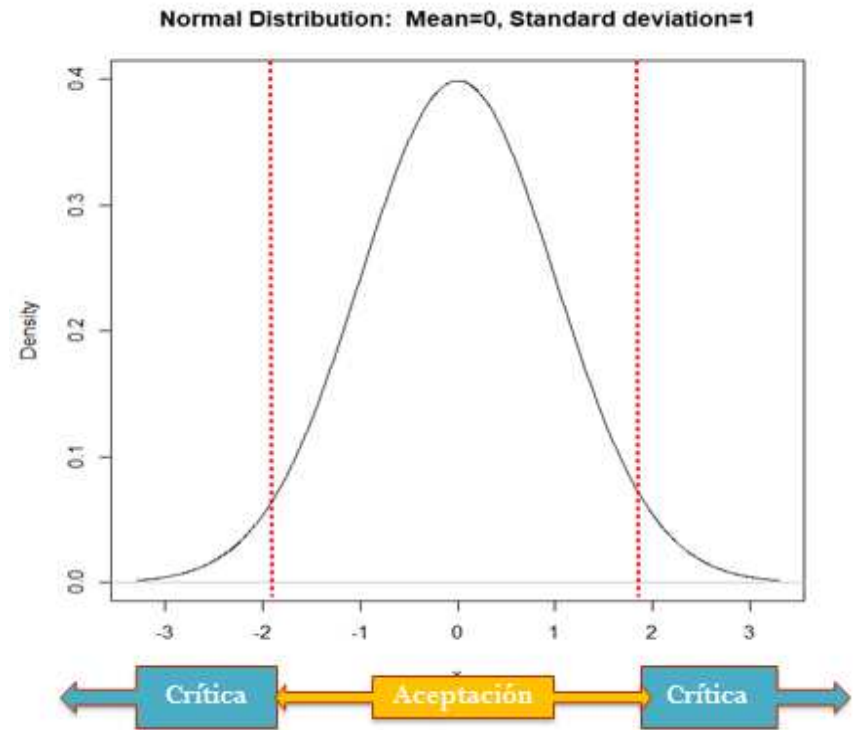
4. Contrastes de hipótesis: aceptando o rechazando H_0

Hipótesis nula (H_0): la se quiere contrastar. Será la que se acepte o rechace como consecuencia del contraste



Hipótesis alternativa (H_a): cualquier otra hipótesis que difiera de la formulada. Se acepta cuando se rechaza H_0

- ✓ Rechazaremos la H_0 (aceptando la alternativa) cuando la discrepancia entre la media observada y la teórica sea grande
- ✓ Aceptaremos la H_0 si la media muestral está dentro del intervalo seleccionado y la rechazaremos en caso contrario
- ✓ Estamos asumiendo un riesgo del 5% (*nivel de significación*) de equivocarnos y rechazar indebidamente H_0 (riesgo tipo I)

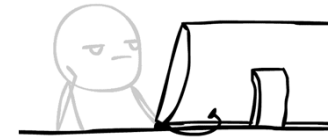


4. Contrastes de hipótesis en la práctica

La situación más habitual en la investigación aplicada es aquella en la que quieren **comparar las distribuciones de dos muestras en una variable**

Ejemplo, Igualdad de Medias: supongamos ahora que en la empresa X, además de personal técnico también hay trabajadores pertenecientes a otras categorías profesionales tales como directivo/as y personal de seguridad, y nos gustaría saber si hay diferencias en los salarios medios entre, por ejemplo, el personal técnico y el directivo

MY CODE ISN'T WORKING ...



¿Existen diferencias en el salario entre ambas categorías profesionales?

Hipótesis nula (H_0): no hay diferencias medias en el salario entre directivos/as y personal técnico



Hipótesis alternativa (H_a): existen diferencias medias en el salario entre directivos/as y personal de seguridad

✓ **Nivel de significación:** Generalmente el 0.05 (5%) o el 0.01 (1%)

➡ **P-valor**

4. Contrastes de hipótesis: p-valor y selección de test

- ✓ **P-valor:** la probabilidad de error en que incurriríamos en caso de rechazar la hipótesis nula con los datos de que disponemos

¿Mayor o menor que el nivel de significación?

Si $p\text{-valor} \geq 0.05$ acepto la H_0

Si $p\text{-valor} < 0.05$ rechazo H_0

Si acepto H_0 al 5 y al 1% se dice que el contraste es “no significativo”

Si rechazo H_0 se dice que el contraste es “significativo” al 1%: “altamente significativo”

tanto al 5% como al 1%: “probablemente significativo”

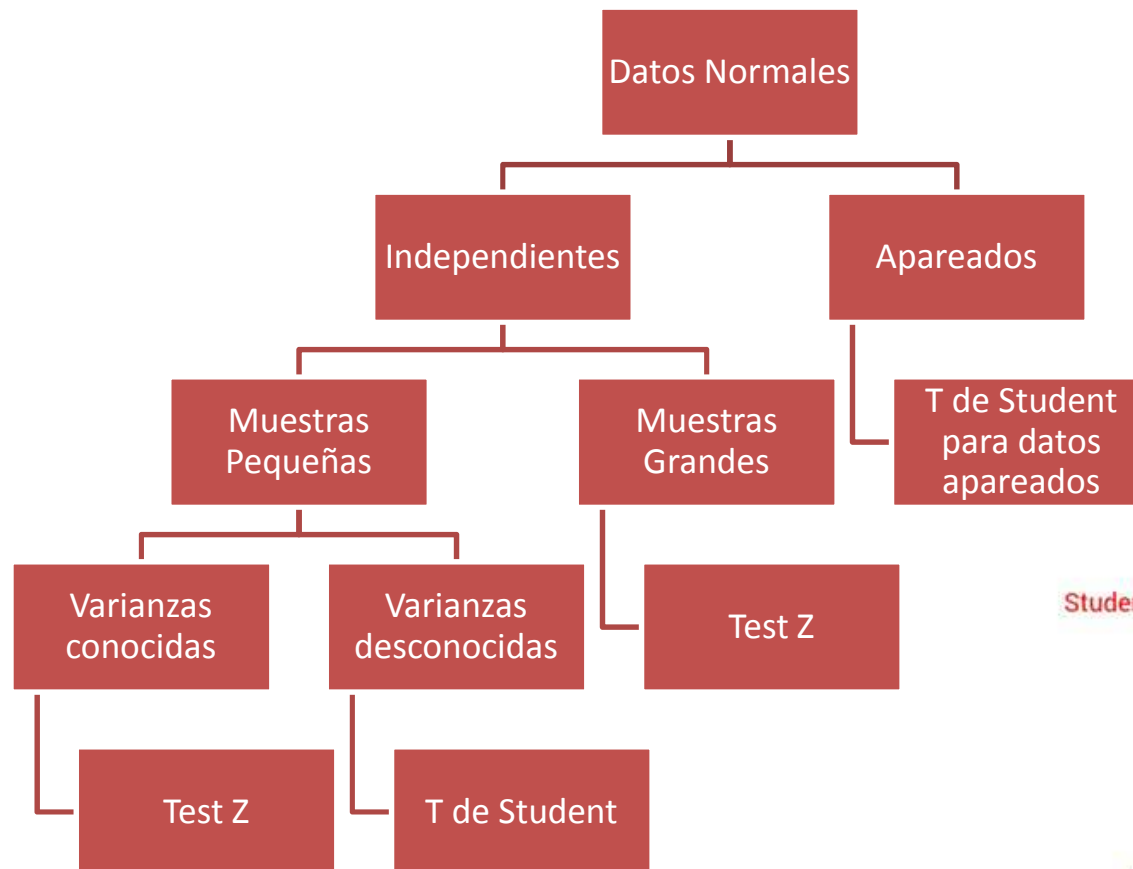
Selección del test de hipótesis a aplicar

- ✓ **Variable normal:** Medida de tendencia central (media aritmética)
- ✓ **Variable no normal:** Medida de tendencia central (mediana)
- ✓ **Datos independientes:** aquéllos que se obtienen al realizar el contraste con dos muestras distintas
- ✓ **Datos apareados:** aquéllos que se obtienen al realizar dos contrastes sobre una misma muestra

Ej.: Diferencia en el salario medio entre directivo/as y personal técnico de la empresa X

Ej.: Diferencia entre el salario inicial y el actual entre el personal técnico de la empresa X

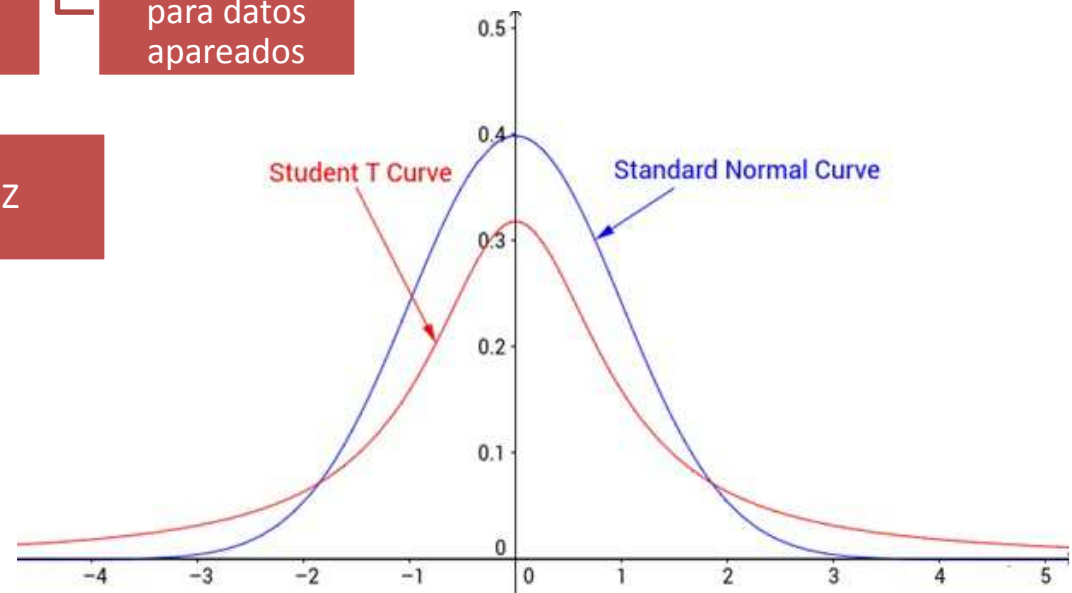
4. Contrastes de hipótesis: selección del test



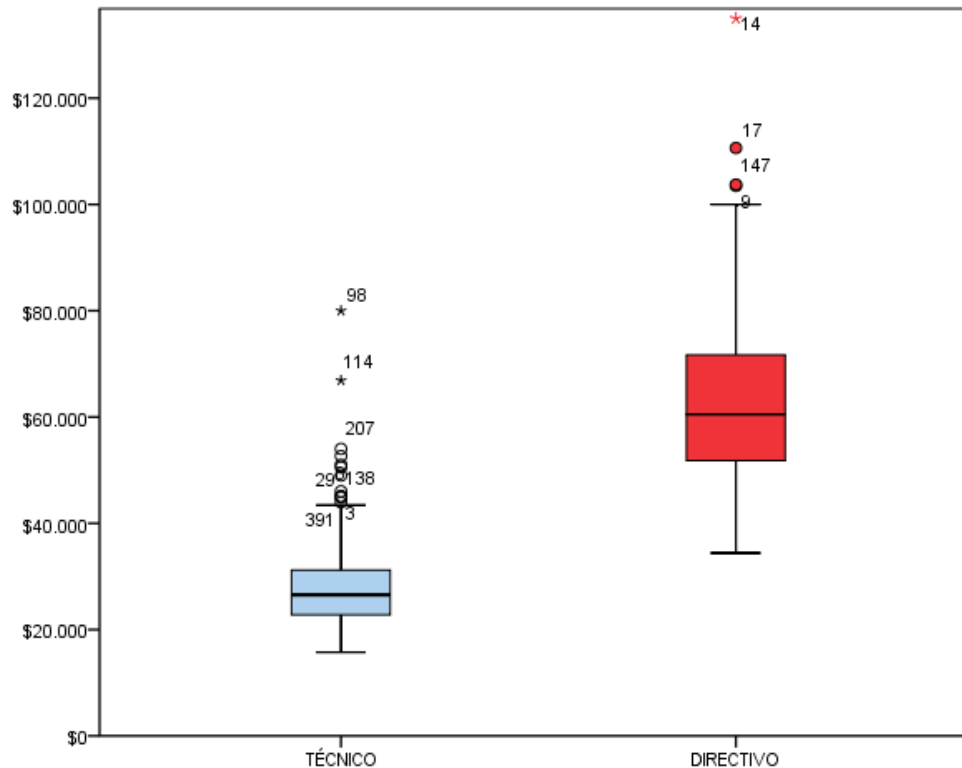
Distribución Normal



T de Student



4. Contrastes de hipótesis: ejemplo salarios



Descriptivos

	N	Mean	SD
TÉCNICO	363	\$27.838,54	\$7.567,96
DIRECTIVO	84	\$63.977,80	\$18.244,78

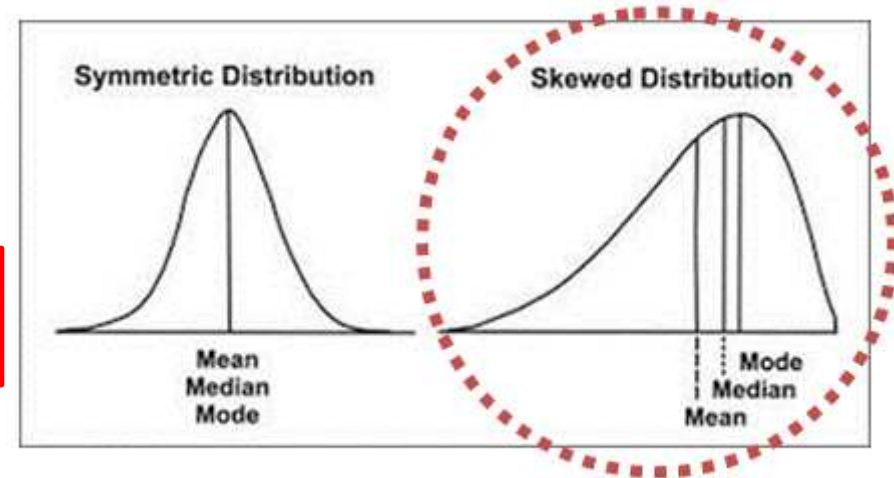
Rechazo $H_0 \rightarrow$ Diferencias altamente significativas entre los salarios medios de técnicos y directivos

Student's t-test	df	P-value	Mean Difference	Std. Error Difference
-17,803	89,708	0,000	-\$36.139,26	\$2.029,91

4. Contrastes de hipótesis: test no paramétricos

- ✓ En ocasiones en **Ciencias Sociales** es complicado que los datos se ajusten a una distribución normal → **distribuciones muy asimétricas**
- ✓ La dispersión es muy grande o la media está muy afectada por valores extremos
- ✓ Para abordar estos problemas, se emplean contrastes que utilizan la **mediana** y que no emplean parámetros de una distribución concreta
 - ☐ Comparan MEDIANAS
 - ☐ Trabajan con rangos de orden en lugar de que con los datos originales

U de Mann Whitney: muestras independientes
Test de Wilcoxon: datos apareados

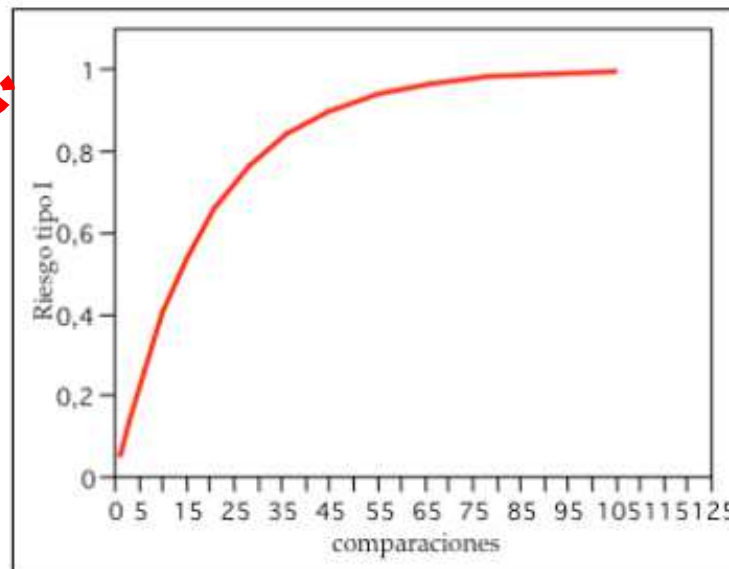


- ✓ ¿Normalidad?
 - ☐ **Métodos gráficos:** Histogramas, Box-Plot, Gráficos de normalidad
 - ☐ **Test de Normalidad:** Kolmogorov-Smirnov, Shapiro-Wilk

5. Análisis de la Varianza (ANOVA)

- ✓ Una variable cuantitativa (2 grupos) y una cualitativa → **T de Student**
- ✓ ¿Y si tengo **más de 2 grupos**? → ~~T de Student~~ **ANOVA** ✓
- ✓ **Error tipo I**: rechazo indebido de H_0
- ✓ **Nivel de significación (α)** = 5% → Intervalo de confianza 0.95

grupos	comparaciones	riesgo tipo I
2	1	0,05
3	3	0,1426
4	6	0,2649
5	10	0,4013
6	15	0,5367
7	21	0,6594
8	28	0,7622
9	36	0,8422
10	45	0,9006
11	55	0,9405
12	66	0,9661
13	78	0,9817
14	91	0,9906
15	105	0,9954



Si hiciésemos los contrastes con parejas la probabilidad de cometer el Error Tipo I = $1^3 = 0.1426 \rightarrow 14\%$

5. ANOVA: supuestos y ejemplo

- ✓ Se basa en la **comparación de la variabilidad entre y la variabilidad dentro de los grupos**, rechazaremos la H_0 siempre que la variabilidad “entre” sea grande, pero utilizando como patrón de comparación la variabilidad “dentro”

Supuestos de partida:

- ✓ **Normalidad**
- ✓ **Homocedasticidad** (igual de varianzas) → Test de Levene
- ✓ **Independencia**

Más robusto frente a la normalidad, pero más sensible a la desigualdad de varianzas, especialmente si el tamaño de muestra de los grupos difiere mucho entre sí

- ✓ Alternativa no paramétrica → **Kruskal-Wallis**

Ejemplo: la empresa X cuenta en su plantilla con 3 tipos de personal: técnico, de seguridad y directivo. Nos gustaría saber si hay diferencias medias entre los grupos y también, entre qué pares de grupos

H_0 : las medias de los grupos son iguales
 H_a : al menos un par es diferente



Aceptaremos un efecto de la categoría profesional siempre que produzca mayores diferencias en que las que habría sin ella

5. ANOVA: post-hoc test

- ✓ El ANOVA es una prueba de significación a nivel global, nos dice si hay diferencias, pero no entre qué pares de medidas
- ✓ Es necesario realizar los contrastes tras el ANOVA para encontrarse esas diferencias

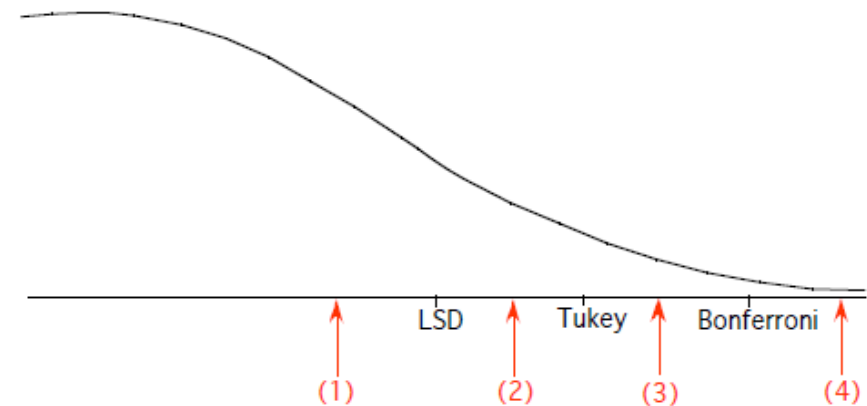
ANOVA	
F de Snedecor	371,106
P-valor	0.000

Test tras el ANOVA

- ✓ **Test LSD**: diferencia mínima significativa → sin control del riesgo tipo I
- ✓ **Test de Tukey**: tamaños de grupos iguales $\alpha_{Tukey} = \alpha / r$ donde $r = n^o$ grupos
- ✓ **Test de Bonferroni**: tamaños de grupos desiguales → el más conservador

$$\alpha_{Bonferroni} = \alpha / [r(r-1)/2]$$

Bonferroni		P-valor
Técnico	Seguridad	0,951
	Directivo	0,000
Seguridad	Directivo	0,000



6. Tablas de contingencia: dos cualitativas

- ✓ Las **tablas de contingencia** son tablas que recogen información sobre de **variables** aleatorias **cualitativas** → **test chi-cuadrado**

- ✓ Los datos aparecen como tablas de **frecuencias**

	Técnico	Seguridad	Directivo
Hombre	157	27	74
Mujer	206	0	10

¿La categoría profesional puede considerarse independiente del sexo?

H_0 → las dos variables en estudio son independientes

H_a → las dos variables en estudio están relacionadas

Chi-cuadrado → evalúa la discrepancia entre las frecuencias observadas y las esperadas

- ✓ **Frecuencias observadas:** número de individuos de nuestra muestra que pertenecen a cada combinación de categorías
- ✓ **Frecuencias esperadas:** la que cabría esperar si ambas variables fuesen independientes
 - Total marginal fila * total marginal columna / total global

6. Tablas de contingencia: test chi-cuadrado

Ejemplo: ¿La categoría profesional puede considerarse independiente del género?

Frecuencias observadas					Frecuencias esperadas				
	Técnico	Seguridad	Directivo	Total		Técnico	Seguridad	Directivo	Total
Hombre	157	27	74	258	Hombre	197,6	14,7	45,7	258
Mujer	206	0	10	216	Mujer	165,4	12,3	38,3	216
Total	363	27	84	474	Total	363	27	84	474

$$f_{o11} = 157$$

$$f_{e11} = (258 \times 363) / 474 = 197,6$$

Se miden las discrepancias calculando las diferencias entre ambas magnitudes para todas las casillas de la tabla

$$\chi^2_{\text{exp}} = \sum_i \sum_j \frac{(f_{o_{ij}} - f_{e_{ij}})^2}{f_{e_{ij}}}$$

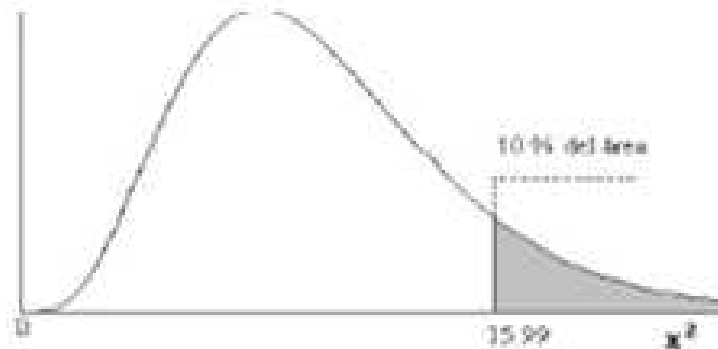
$$\chi^2_{\text{experimental}} = (157 - 197,6)^2 / 197,6 \dots + (216 \times 84) / 474 = 79,27$$

χ^2 crítico: tabla de la chi-cuadrado, se busca el valor en base al nivel de significación y los grados de libertad. Al 5% = 5,99, al 1% = 9,21

Rechazaremos H_0 cuando $\chi^2_{\text{experimental}} > \chi^2_{\text{crítico}}$

$79,27 > 5,99 \rightarrow$ Rechazamos $H_0 \rightarrow$ **P-valor < 0.01, las 2 variables están altamente relacionadas**

6. Tablas de contingencia: valor crítico



Valor crítico de la chi-cuadrado

Nivel de confianza 95%

2 grados de libertad

$\frac{\chi^2}{df}$	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005	$\frac{\chi^2}{df}$
1	3.841	3.841	3.841	3.841	3.841	3.841	3.841	3.841	3.841	3.841	3.841	3.841	3.841	1
2	5.991	5.991	5.991	5.991	5.991	5.991	5.991	5.991	5.991	5.991	5.991	5.991	5.991	2
3	7.879	7.879	7.879	7.879	7.879	7.879	7.879	7.879	7.879	7.879	7.879	7.879	7.879	3
4	9.488	9.488	9.488	9.488	9.488	9.488	9.488	9.488	9.488	9.488	9.488	9.488	9.488	4
5	10.597	10.597	10.597	10.597	10.597	10.597	10.597	10.597	10.597	10.597	10.597	10.597	10.597	5
6	11.578	11.578	11.578	11.578	11.578	11.578	11.578	11.578	11.578	11.578	11.578	11.578	11.578	6
7	12.592	12.592	12.592	12.592	12.592	12.592	12.592	12.592	12.592	12.592	12.592	12.592	12.592	7
8	13.601	13.601	13.601	13.601	13.601	13.601	13.601	13.601	13.601	13.601	13.601	13.601	13.601	8
9	14.617	14.617	14.617	14.617	14.617	14.617	14.617	14.617	14.617	14.617	14.617	14.617	14.617	9
10	15.636	15.636	15.636	15.636	15.636	15.636	15.636	15.636	15.636	15.636	15.636	15.636	15.636	10
11	16.678	16.678	16.678	16.678	16.678	16.678	16.678	16.678	16.678	16.678	16.678	16.678	16.678	11
12	17.730	17.730	17.730	17.730	17.730	17.730	17.730	17.730	17.730	17.730	17.730	17.730	17.730	12
13	18.759	18.759	18.759	18.759	18.759	18.759	18.759	18.759	18.759	18.759	18.759	18.759	18.759	13
14	19.753	19.753	19.753	19.753	19.753	19.753	19.753	19.753	19.753	19.753	19.753	19.753	19.753	14
15	20.707	20.707	20.707	20.707	20.707	20.707	20.707	20.707	20.707	20.707	20.707	20.707	20.707	15

✓ **Grados de libertad:** $(N^{\circ} \text{ filas} - 1) * (N^{\circ} \text{ columnas} - 1)$

RESUMIENDO LOS CONTRASTES...



Análisis de la relación entre dos variables

Parte I. Contrastes de hipótesis

- ✓ **Una variable cualitativa (2 grupos) y otra cuantitativa** → T de Student y relacionados
- ✓ **Una variable cualitativa (+ 2 de grupos) y otra cuantitativa** → ANOVA, test tras el ANOVA para las diferencias entre qué pares de grupos
- ✓ **Variables cualitativas** → Tablas de contingencia, chi-cuadrado

Parte II.

- ✓ **Dos variables cuantitativas**
 - ☐ Análisis de correlación
 - ☐ Análisis de regresión lineal simple
- ✓ **Más de 2 de variables (una dependiente, varias independientes)**
 - ☐ Análisis de regresión lineal múltiple

Más allá del P-valor, los tamaños del efecto

Una diferencia estadísticamente significativa no es necesariamente una diferencia grande y tampoco es necesariamente una diferencia importante

Significación estadística y práctica:

- ✓ El hecho de que exista una significación estadística no significa que el resultado sea relevante desde el punto de vista práctico
- ✓ Es más **difícil** encontrar diferencias con **muestras pequeñas**
- ✓ Si la **muestra es muy grande** puedo encontrar **diferencias** significativas **sin importancia** práctica real
- ✓ Adoptar un nivel de confianza fijo, convierte en resultado en dicotómico



¿La magnitud de la diferencia es grande o pequeña?

TAMAÑOS DEL EFECTO: Permite analizar y cuantificar la intensidad de la relación entre dos variables o la diferencia entre dos grupos

D de Cohen (Cohen, 1988)

G de Hedges

Más allá del P-valor, los tamaños del efecto



Díaz-Faes, A. A., Costas, R., Galindo, M. P., & Bordons, M. (2015). Unravelling the performance of individual scholars: Use of Canonical Biplot analysis to explore the performance of scientists by academic rank and scientific field. *Journal of Informetrics*, 9(4), 722-733.

Hedge's g by field and academic rank

Academic rank	First Author	Research Level	MNCS
Post-doc (CHEM) vs. Researcher (CHEM)	1.35	0.16	0.11
Post-doc (CHEM) vs. Professor (CHEM)	1.41	0.18	0.14
Tenure (CHEM) vs. Professor (MAT)	0.34	0.07	0.01
Researcher (CHEM) vs. Tenure (MAT)	0.40	0.58	0.07
Post-doc (MAT) vs. Professor (MAT)	1.60	0.17	0.07
Tenure (MAT) vs. Professor (CHEM)	0.50	0.61	0.05
Post-doc (CHEM) vs. Professor (MAT)	1.83	0.12	0.16

Effect size: $g < 0.20$ trivial, $g \geq 0.20$ small effect, $g \geq 0.50$ medium effect, $g \geq 0.80$ large effect and $g \geq 1.30$ very large effect.

Análisis Estadístico - Parte II

- Análisis de Correlación
- Análisis de Regresión Lineal Simple
- Análisis de Regresión Lineal Múltiple
- ¿Y si no hay variable dependiente y explicativa/s y todas están relacionadas?
Análisis Multivariante

Análisis de la relación entre variables

1 VARIABLE CUANTITATIVA Y 1 VARIABLE CUALITATIVA

- ❑ **T DE STUDENT Y RELACIONADOS** → Datos independientes o apareados, ¿siguen una distribución normal?
- ❑ **ANOVA (+ 2 DE GRUPOS)** → Detecta diferencia globales, test post-hoc para las diferencias entre qué pares de grupos

2 VARIABLES CUALITATIVAS → TABLAS DE CONTIGENCIA Y CHI-CUADRADO

VARIABLES CUANTITATIVAS

ANÁLISIS DE CORRELACIÓN

ANÁLISIS DE REGRESIÓN SIMPLE

ANÁLISIS DE REGRESIÓN MÚLTIPLE → Una variable dependiente, varias variables explicativas

} 2 variables
cuantitativas

Dependencia entre variables cuantitativas

En la práctica nos encontramos con frecuencia con situaciones en las que **en base a nuestro conocimiento previo del problema en estudio** pensamos que hay una relación de dependencia entre las variables

PAÍS	ESPERANZA	RENTA
Alemania	77	25.855
Austria	77	26.850
Bélgica	77	25.380
Bulgaria	71	1.320
Croacia	72	4.520
Dinamarca	75	33.260
Eslovaquia	73	3.700
Eslovenia	75	9.760
España	78	14.080
Finlandia	77	24.110
Francia	78	24.940
Grecia	78	11.650
Holanda	78	24.760
Hungría	71	4.510
Irlanda	76	18.340
Italia	78	20.250
Lituania	71	2.440
Moldavia	67	410
Noruega	78	34.330
Polonia	73	3.900
Portugal	75	10.690

La tabla proporciona información acerca de la esperanza de vida media y el nivel de renta de varios países europeos para una muestra de 21 países europeos

¿Cómo cuantificar la relación entre estas 2 variables?

Diagramas de dispersión (i)

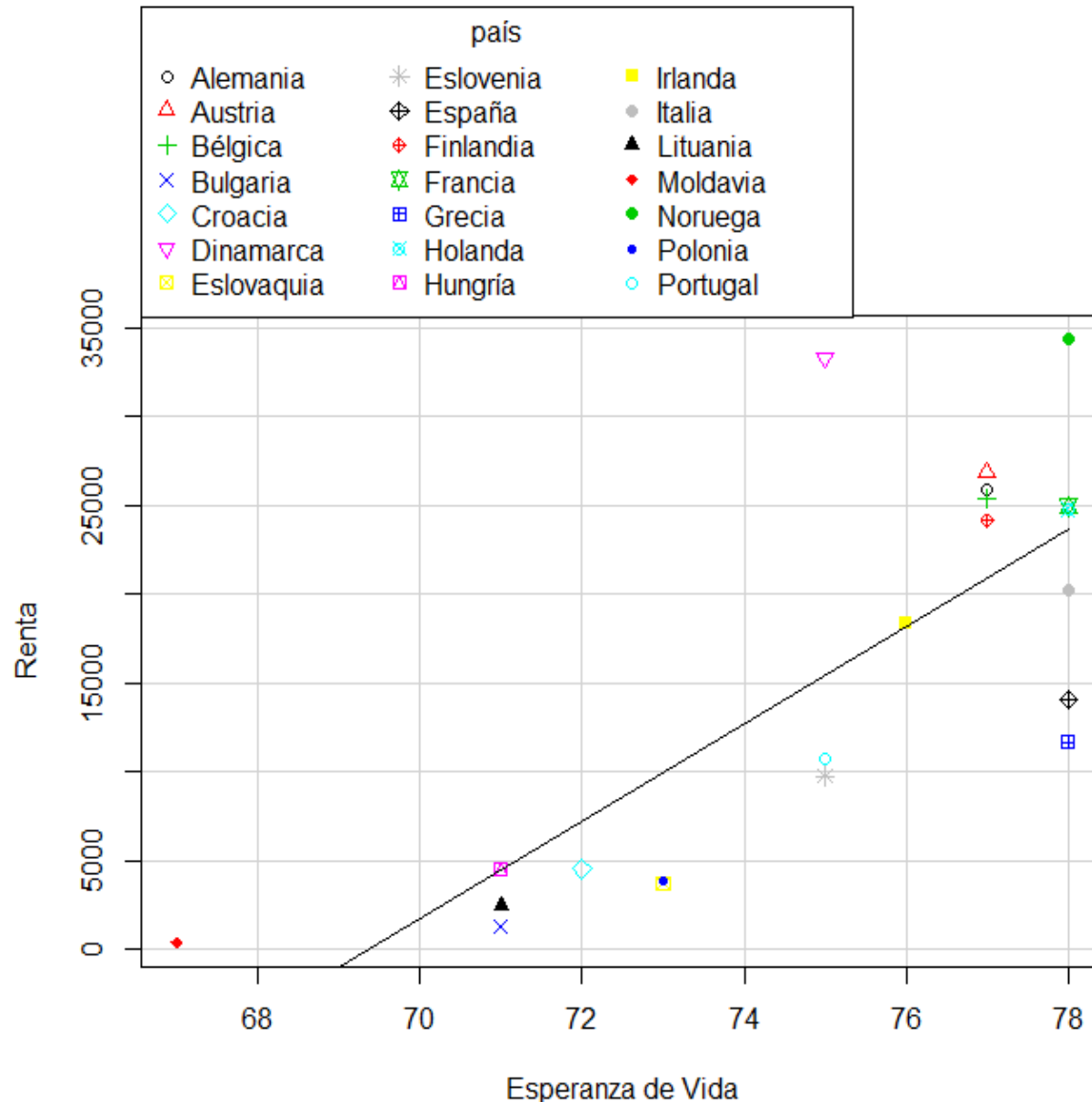
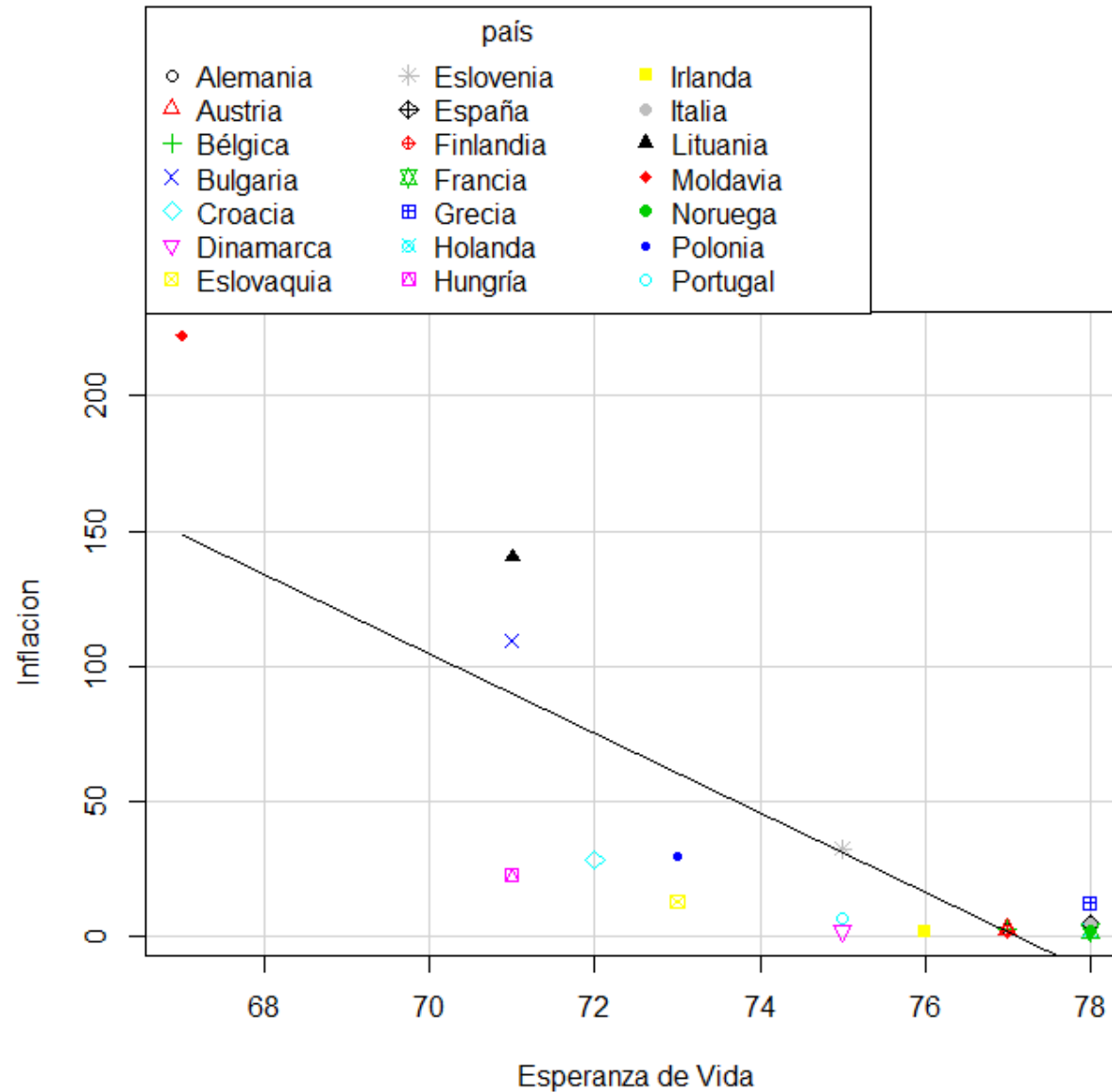


Gráfico en el que una de las variables (X) se coloca en el eje de abscisas, la otra (Y) en el de ordenadas y los (x_i, y_i) se representan como una nube de puntos

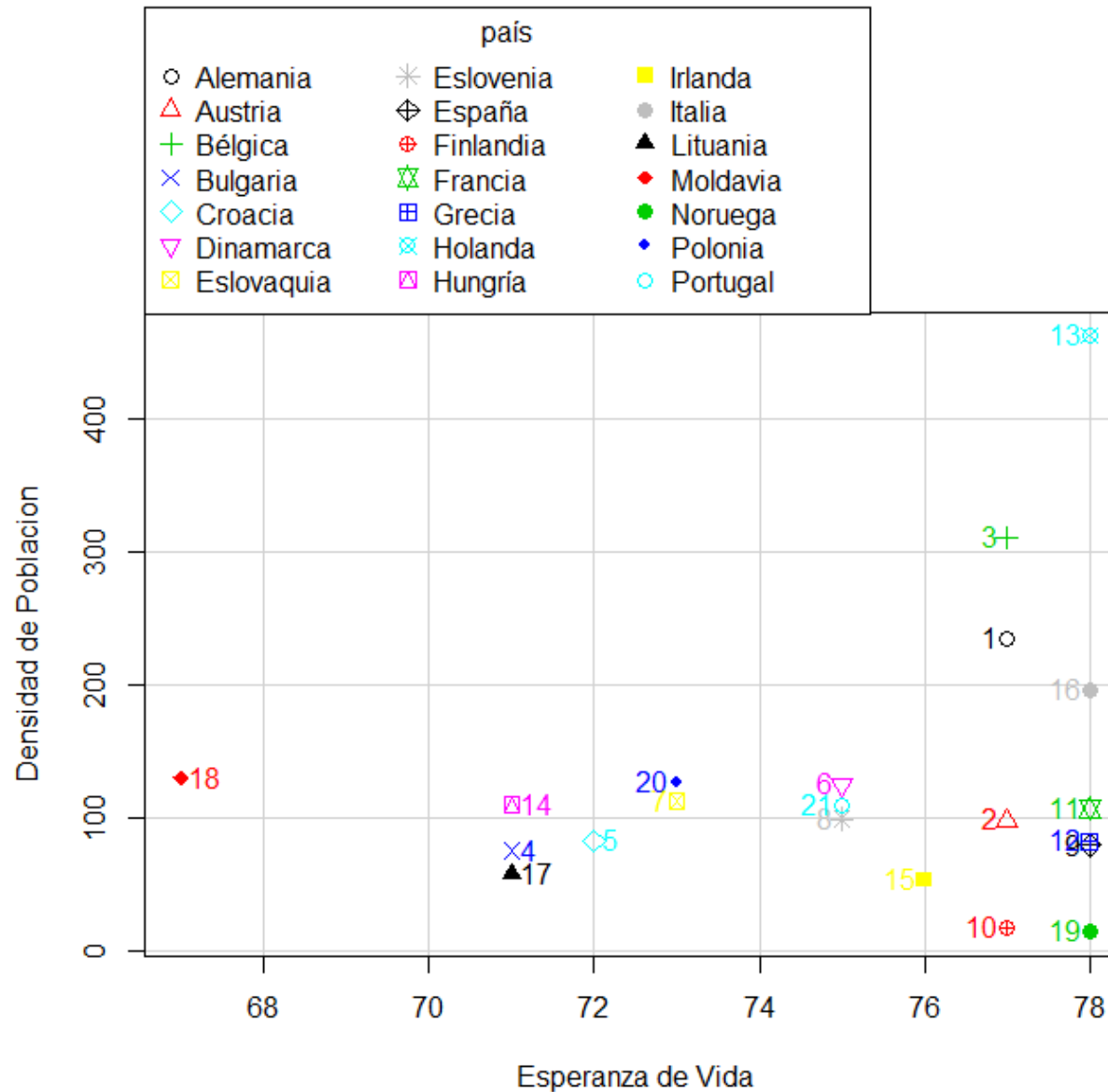
Cada dato es un punto cuyas coordenadas son los valores de las variables

La **forma de la nube de puntos** nos informa sobre el **tipo de relación** existente entre las variables

Diagramas de dispersión (ii)



Diagramas de dispersión (iii)



¿Cómo cuantificamos la relación?



Covarianza



Correlación
Análisis de Regresión

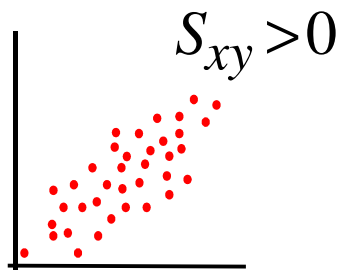
Covarianza

Es una medida de lo que se dispersan/asocian los valores de dos variables aleatorias respecto de las medidas de ambas

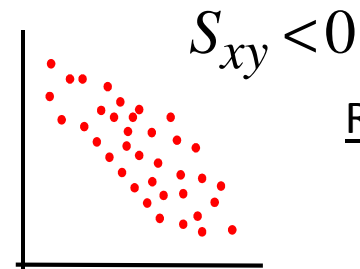
Propiedades:

$$S_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

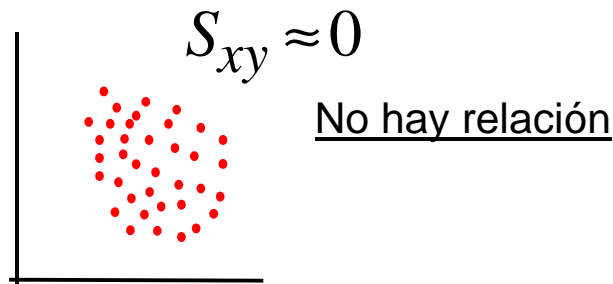
- ✓ Depende de la dimensión de las variables
- ✓ El signo refleja el tipo de relación: **+ relación directa / - inversa**
- ✓ La **magnitud** muestra el grado de relación



Relación directa



Relación inversa



No hay relación

Coeficientes de correlación

Versión normalizada de la covarianza

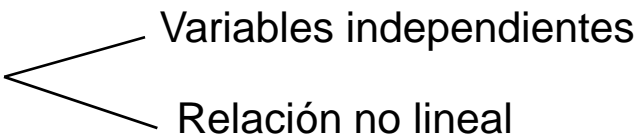
$$\hat{\rho} \equiv r = \frac{S_{xy}}{S_x S_y}$$

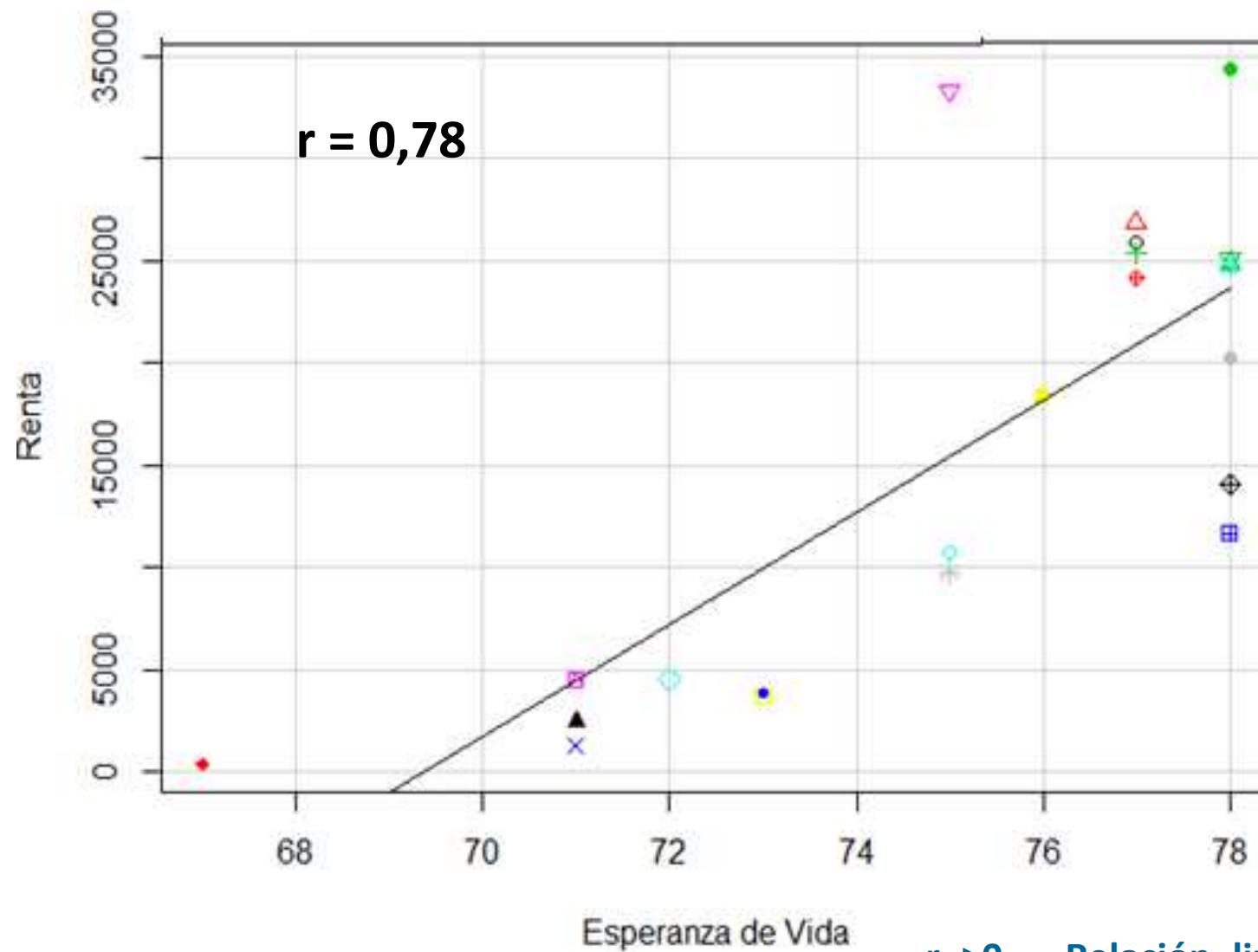
Coeficiente de correlación lineal de Pearson

- ✓ Se define como el cociente entre la covarianza y el producto de las desviaciones típicas
- ✓ Sus valores oscilan entre -1 y +1
- ✓ Mantiene el signo de la covarianza que refleja el tipo de relación
- ✓ Cuanto más se aproxime $|r|$ a 1, mayor es el grado de relación entre las variables

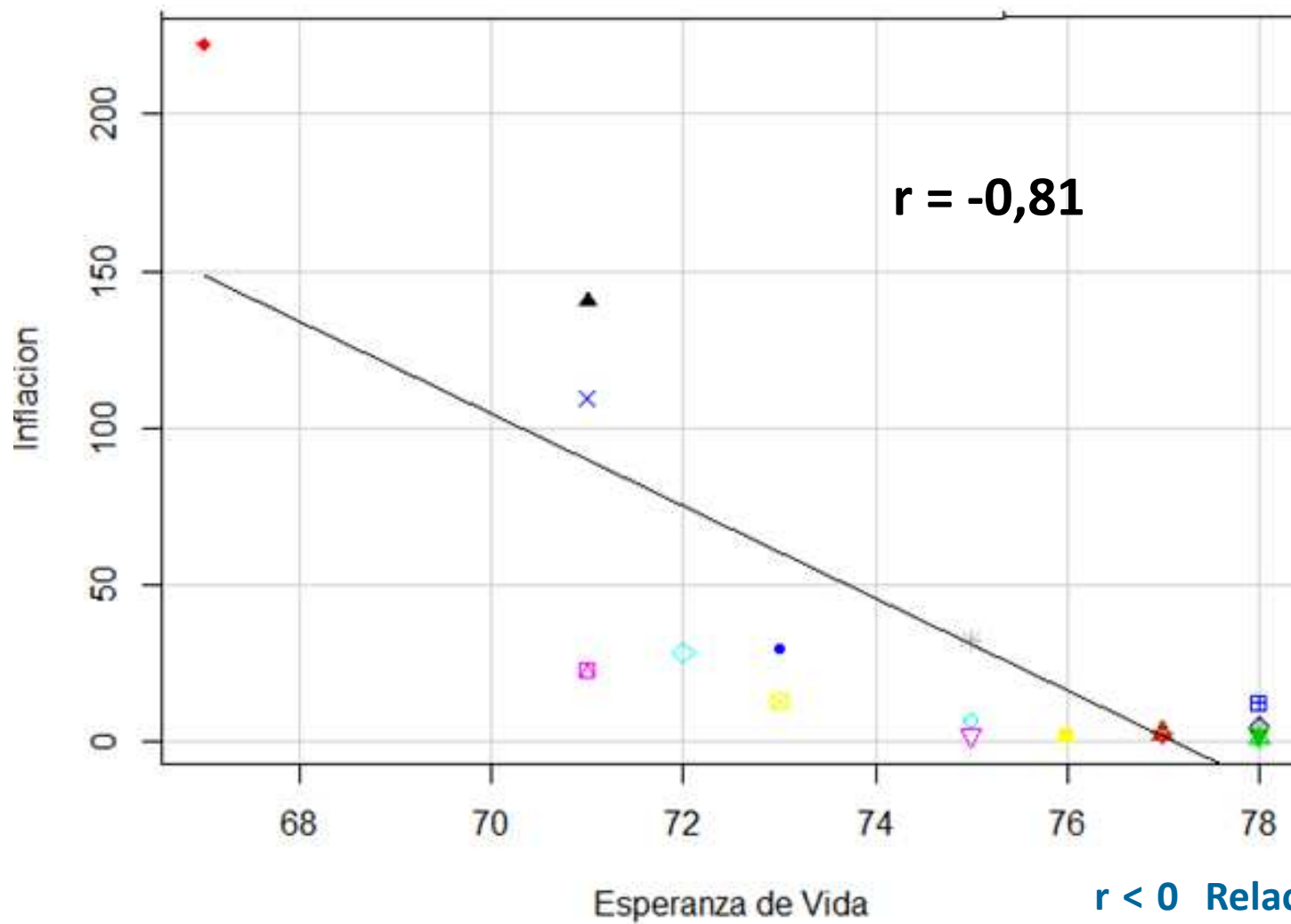
$r > 0$ Relación lineal directa: A medida que aumentan los valores de una variable aumentan los valores de la otra

$r < 0$ Relación lineal inversa: A medida que aumentan los valores de una variable disminuyen los valores de la otra

$r = 0$ 
Variables independientes
Relación no lineal

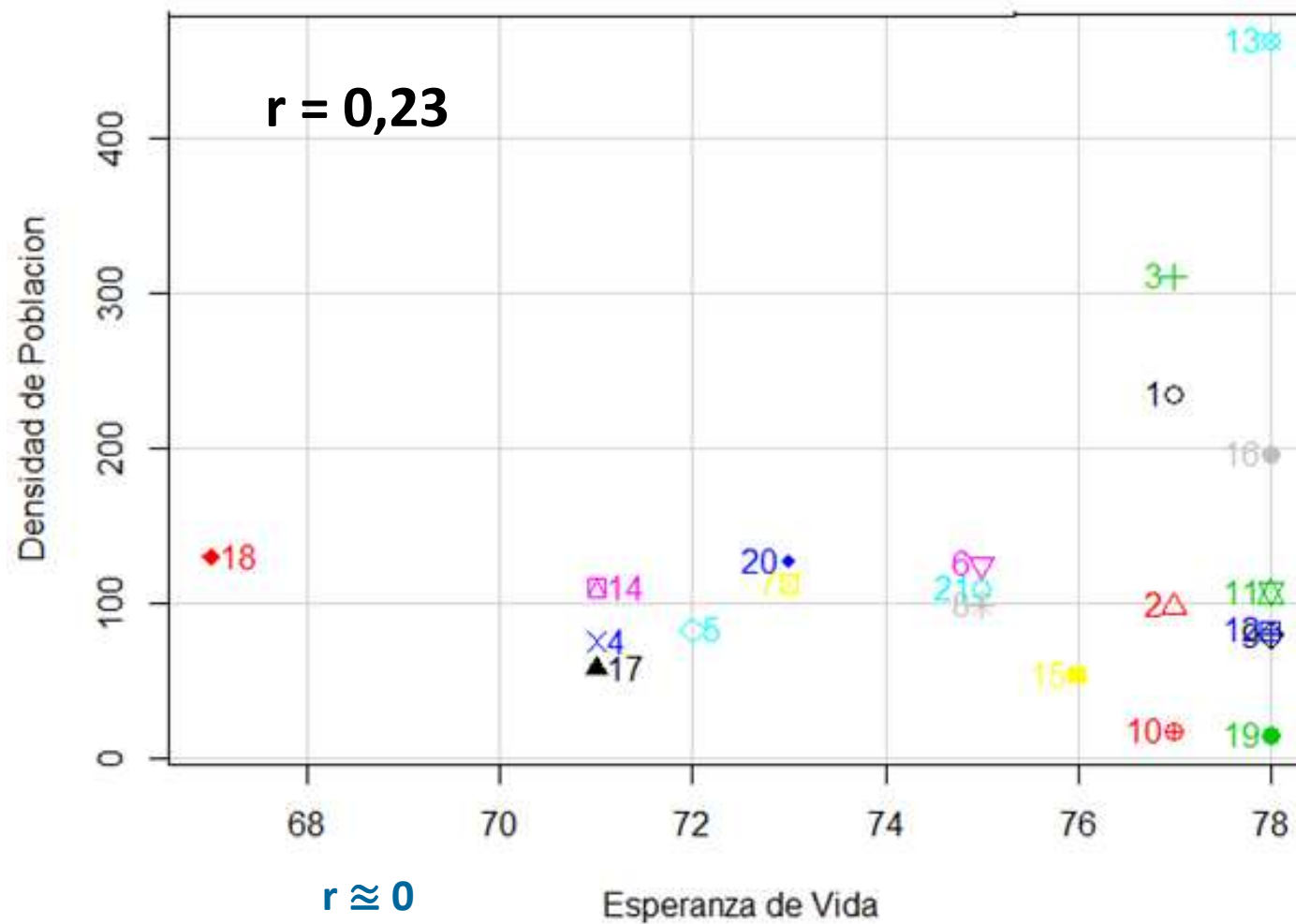


$r > 0$ **Relación lineal directa.** A medida que aumentan los valores de una variable aumentan los valores de la otra



$r < 0$ Relación lineal inversa:

A medida que aumentan los valores de una variable disminuyen los valores de la otra

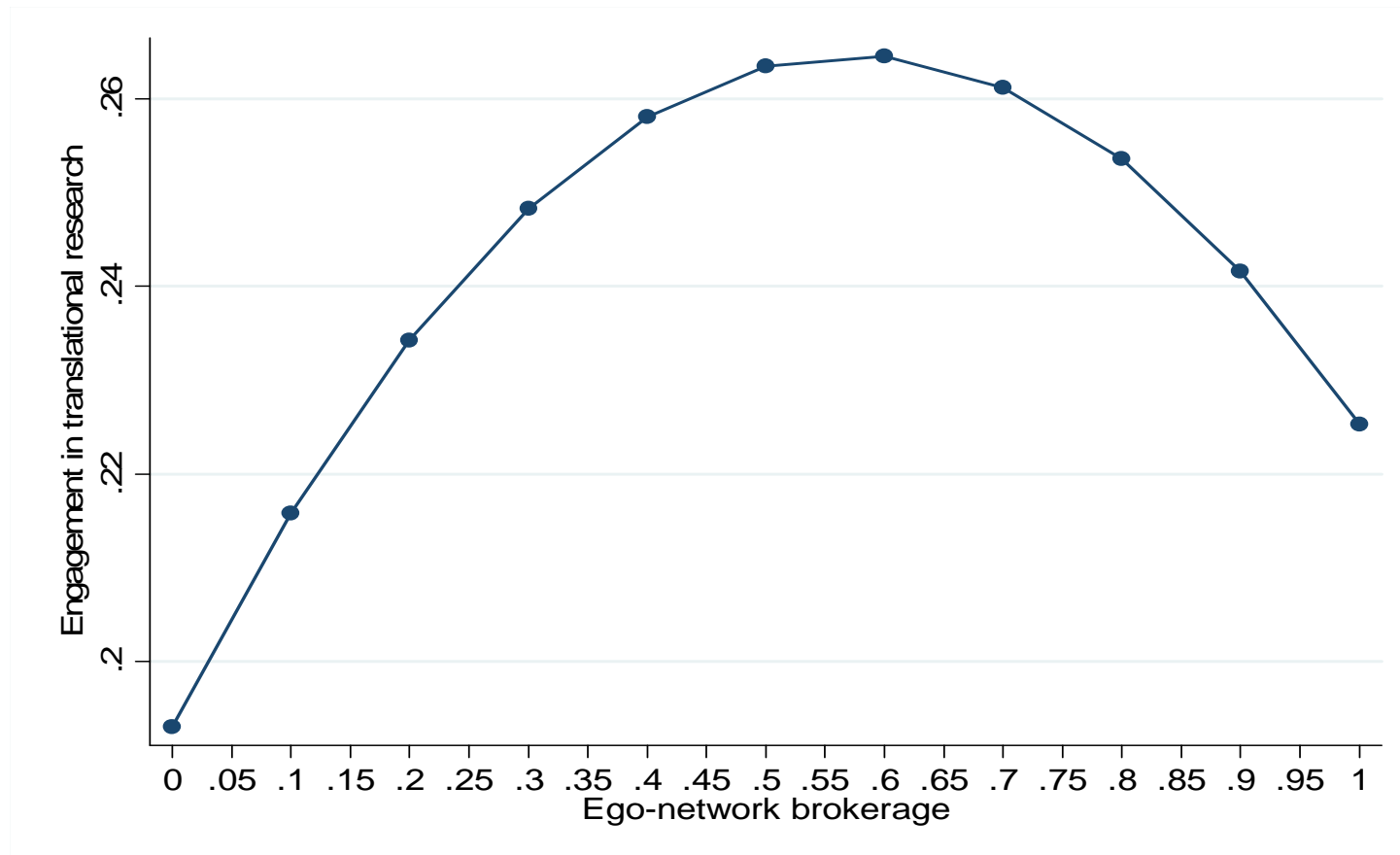


Si el coeficiente de correlación es próximo a cero las variables no tienen porqué ser independientes

La relación puede ser *no lineal*

Relación curvilínea

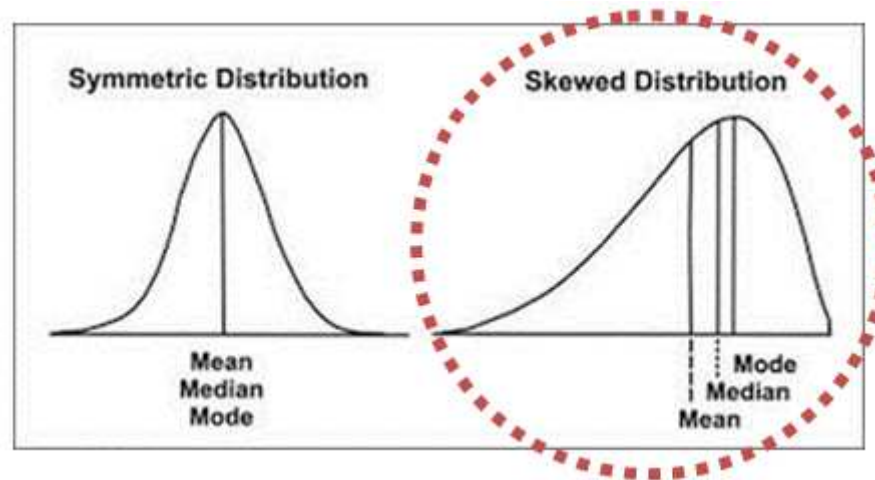
Fuente: Llopis & D'Este (2014)



Relación entre dispersión de red de los investigadores y su participación en actividades relacionadas con la investigación traslacional

Otros coeficientes de correlación

Cuando las variables numéricas presentan distribuciones muy sesgadas o también en el caso de variables ordinales



- ✓ **Rho de Spearman:** Versión no paramétrica del coeficiente de correlación Pearson. Se basa en los rangos de los datos en vez de en los valores reales y es apropiado para datos ordinales
- ✓ **Tau de Kendall:** Medida no paramétrica de asociación para variables ordinales o de rangos que tiene en consideración los empates

Relación de dependencia entre variables cuantitativas

✓ **Análisis de correlación**

- ☐ ¿Cómo están relacionadas las dos variables?
- ☐ ¿La relación es fuerte o débil?

✓ **Análisis de regresión**

- ☐ ¿Cuál es el tipo de dependencia existente?
- ☐ ¿Podemos predecir el comportamiento de la variable dependiente en base a la independiente?

Variable X

- ✓ Variable explicativa
- ✓ Variable regresora
- ✓ Covariable
- ✓ Variable independiente

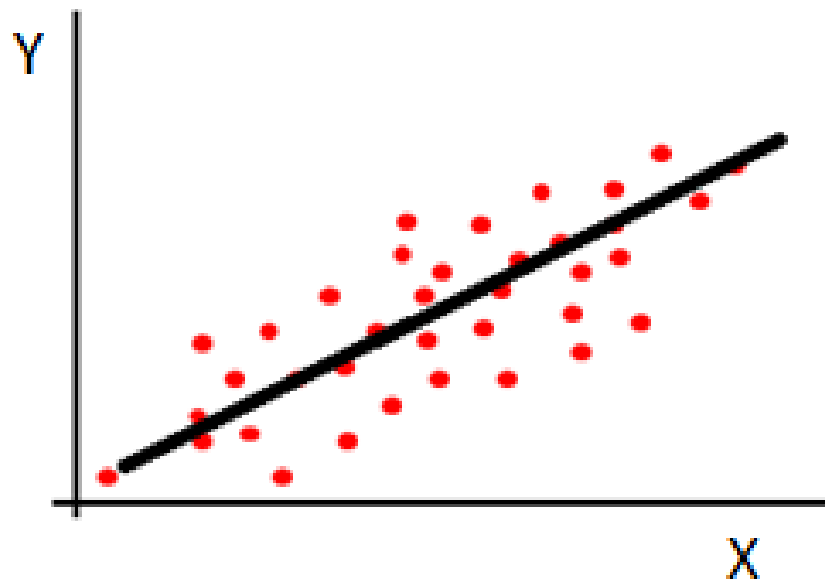
Variable Y

- ✓ Variable dependiente
- ✓ Variable respuesta
- ✓ Variable explicada

Análisis de regresión lineal simple (i)

El análisis de regresión trata de **determinar o explicar el comportamiento** de una variable (dependiente) **basándose en el conocimiento de otra** (independiente)

- ✓ **Variable dependiente (Y):** sobre la que queremos poner el foco y saber si está afectada y en qué medida por la/s variable/s independiente
- ✓ **Variable independiente/explicativa (X):** es la variable/s que puede o no afectar a la variable dependiente



Se intenta aproximar una recta que recoja a la mayor parte de los valores de la nube de puntos (línea media)

A la ecuación que describe la relación entre las variables se le denomina **ecuación de regresión**

Análisis de regresión lineal simple (ii)

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \dots, n$$

Donde:

Y_i es la i ésima observación de la **variable dependiente**

X_i es la i ésima observación de la **variable independiente**

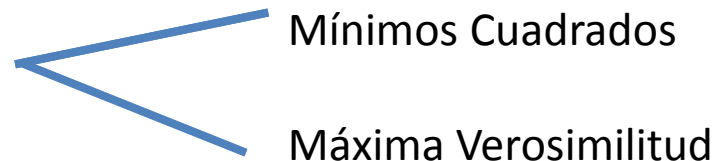
β_0 es el **término independiente** \rightarrow la altura a la que la recta corta el eje de ordenadas.

Valor que se le asigna a la variable dependiente en caso de que la independiente fuera 0

β_1 es la pendiente \rightarrow **coeficiente de regresión**, es decir, **el incremento que sea produce en la variable dependiente(Y) cuando la variable independiente (X) aumenta en una unidad**

ε_i es el **error** aleatorio no observable asociado con Y_i

Existen diversos métodos para estimar los parámetros del modelo



Poder explicativo del modelo

La recta de regresión tiene carácter de línea media, por ello debe ir acompañada siempre de una medida de su representatividad; es decir, de una medida de dispersión

- Poca dispersión → Alta representatividad del modelo
- Muchas dispersión → Baja representatividad del modelo

La forma más habitual de medir la bondad de ajuste del modelo es el Coeficiente de Determinación (R^2) → Indica el porcentaje de variaciones controladas o explicadas por el modelo

- ✓ Cuanto más se aproxime R^2 a la 1, mayor **poder explicativo o mayor bondad de ajuste** del modelo.

$R^2 = 1$ Dependencia funcional

$R^2 = 0$ Variables independientes / Modelo inadecuado

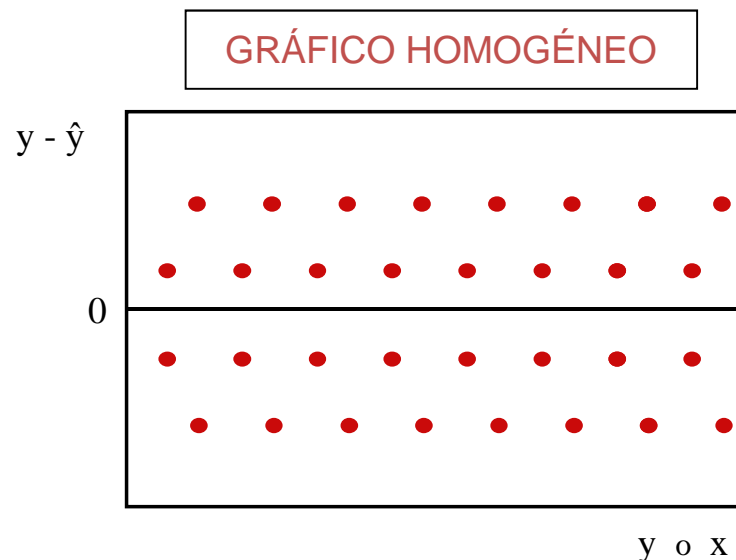
$$R^2 = 1 - \frac{S_e^2}{S_y^2}$$

$R^2 \times 100$ = Porcentaje de variaciones explicadas por el modelo

Supuestos básicos del modelo de regresión

Una vez ajustado y obtenido el modelo, si estuviésemos interesados en realizar predicciones hay que asegurarse de que no se violan las hipótesis sobre las que se sustenta → **Esta información la aportan los residuos del modelo**

- ✓ **Normalidad:** de los valores de la variable dependiente → Gráficos de Probabilidad, Test de Kolmogorov-Smirnov
- ✓ **Homocedasticidad:** igual de varianzas, es decir que la variabilidad de Y es la misma para todos los valores de X → Gráficos de residuos, Test de Levene
- ✓ **Independencia:** las observaciones muestrales son independientes. Esto no suele ser un problema salvo que se trabaje con series temporales → Gráficos de residuos, Test de Durbin-Watson

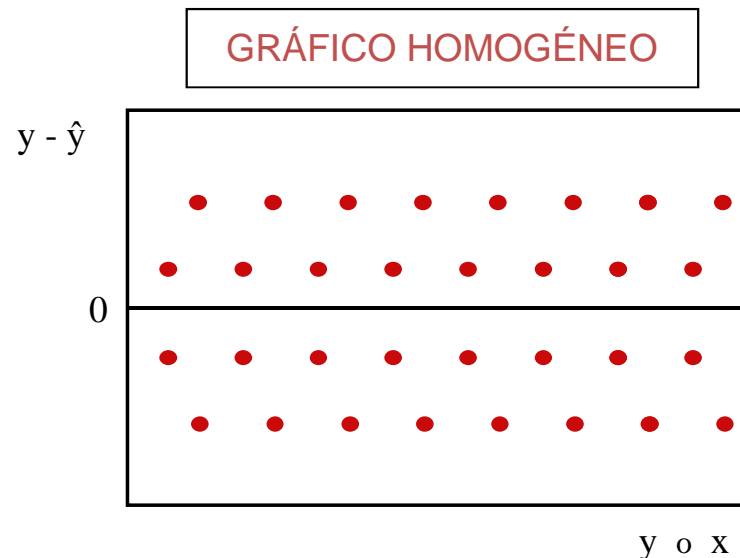


Poder explicativo vs. Poder predictivo

Gráficos de residuales son diagramas de dispersión que permiten evaluar el poder predictivo del modelo.

- ✓ Colocamos en el eje X (abscisas) la variable dependiente o la independiente
- ✓ En el eje y (ordenadas) los residuos

Interpretación: si los residuos son homogéneos y pequeños, entonces el modelo presenta un alto poder ALTO PODER PREDICTIVO



En la investigación en **Ciencias Sociales** suelen ser complicado hacer predicciones debido a las características de los propios datos

Outliers

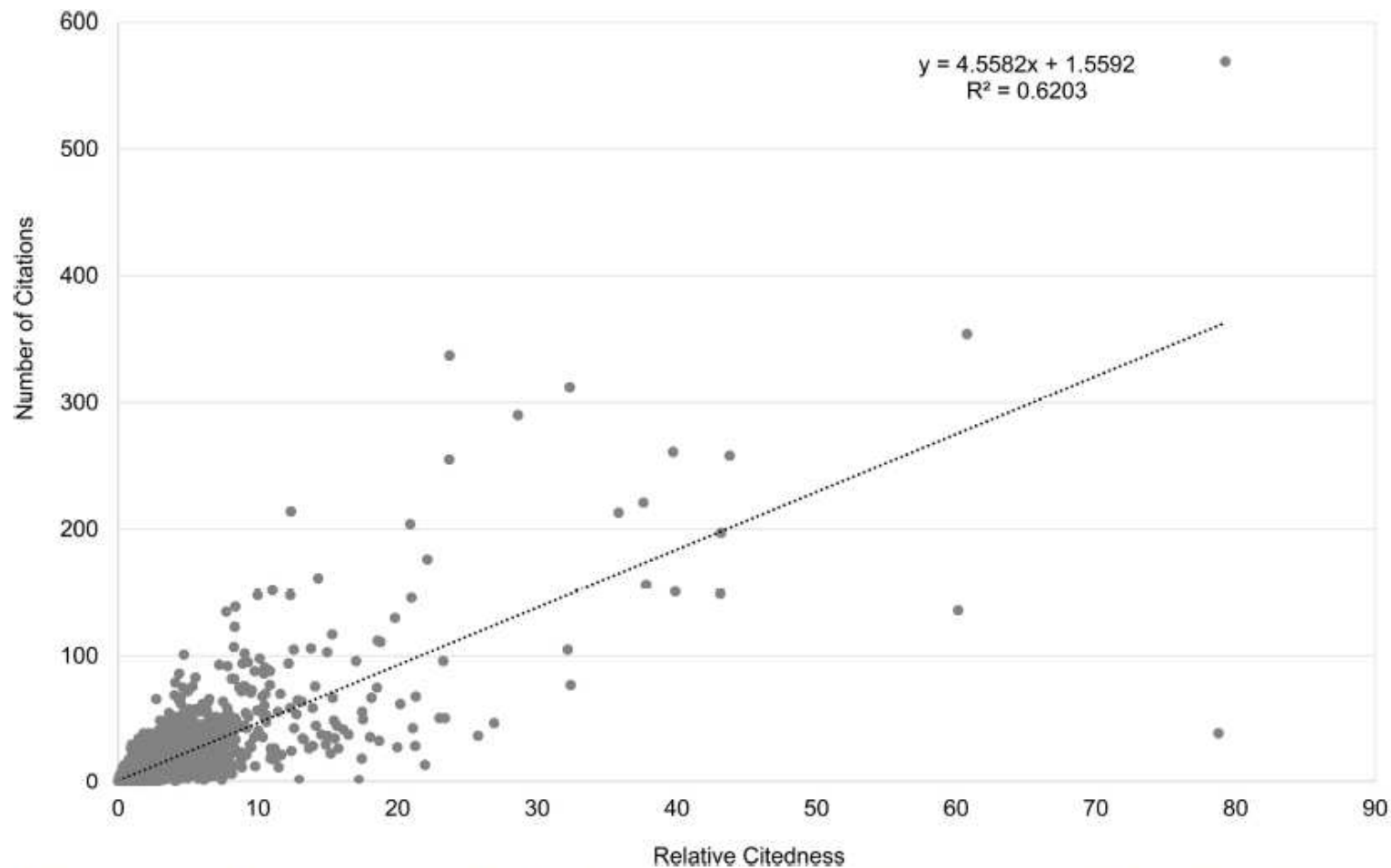


Fig 3. Relation between *Number of Citations* and *Relative Citedness* for articles published in 1992–2006. Data from Clarivate Analytics' National Citation Report for Norway.

Fuente: Zhang, Rousseau & Sivertsen (2017)

Ejemplo regresión lineal simple (i)

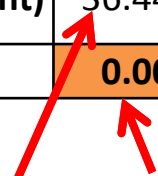
Nos gustaría saber si el número de ordenadores por 1000 habitantes de los países europeos depende del nivel de renta medio

$R^2 = 0,89$



El 89% de las variación producida en el nº d ordenadores viene explicado por el nivel de renta medio

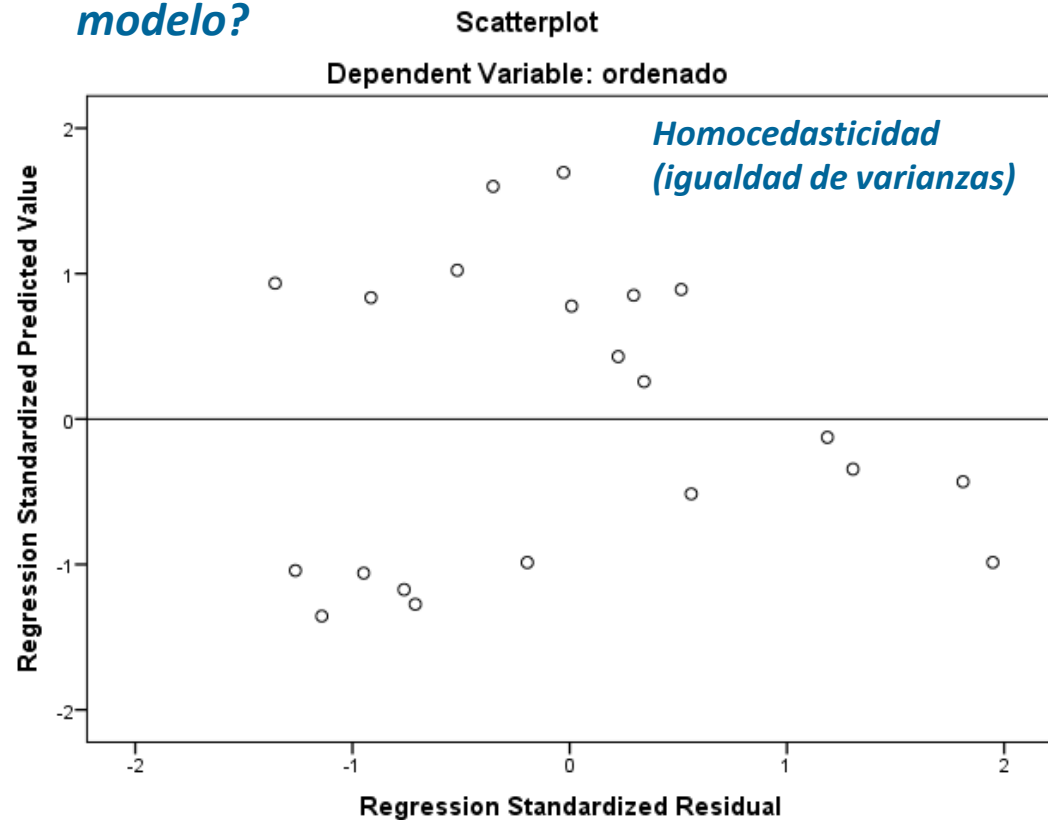
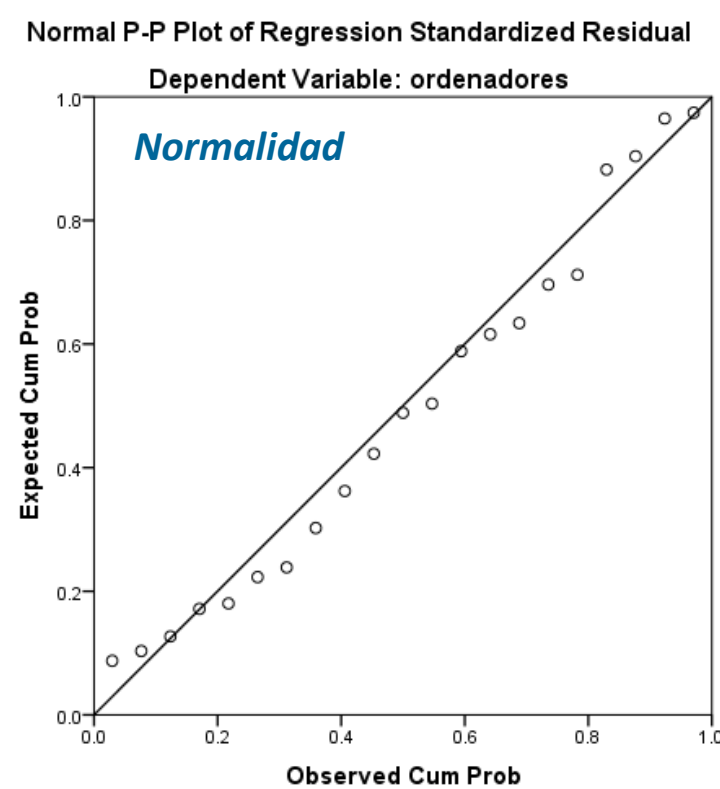
	Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval	
	B	S.E.	Beta			Lower Bound	Upper Bound
Model (Constant)	36.441	11.519		3.164	0.005	12.332	60.551
renta	0.008	0.001	0.945	12.543	0.000	0.006	0.009


$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \dots, n$$

El nº medio de ordenadores por cada 1000 habitantes es 36,44. Por cada aumento en una unidad de la renta, el nº de ordenadores aumenta en 0.008

Ejemplo regresión lineal simple (ii)

¿Se cumple las hipótesis de normalidad, homocedasticidad e independencia del modelo?



¿Cuál sería el nº esperado de ordenadores por cada 1000 habitantes para España que tiene una renta media de 14.080?

$$Y = 36,44 + 0,08 * 14.080 = 149.1$$

Análisis de regresión lineal múltiple

Existen varias variables independientes ($X_1, X_2, X_3 \dots X_n$) que explican el comportamiento de la variable dependiente Y

- ✓ Pensamos que hay diferentes factores que pueden afectar a nuestra variable de interés y nos gustaría determinar:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

- 1) El efecto de cada variable explicativa
- 2) Explorar o confirmar un modelo teórico previo
- 3) Explicar/predecir el comportamiento de la variable dependiente a partir de las explicativas



Coeficiente de regresión

Se interpretan como el efecto que una de las variables independientes tiene sobre la variable dependiente Y cuando el resto de variables se mantienen constantes

Colinealidad

Cuando las variables independientes están altamente correlacionadas entre ella se habla de colinealidad. Esto puede afectar muy negativamente a los coeficientes de regresión al hacerlo inestable

Síntomas

- 1) Altas correlaciones entre al menos un par de variables
- 2) Una variable que se cree importante, deja de ser significativa al introducirse en un modelo múltiple
- 3) El signo de algún coeficiente es contrario al esperado
- 4) Errores estándar de los estimadores muy grandes

Supongamos ahora que creemos que además de la renta, hay otras variables que puede influenciar el nº de ordenadores por habitante, por ejemplo, la inflación del país o la densidad de población

Parece que los coeficientes no están afectados ¿?

r de Pearson	renta
Inflación	-.617**
Densidad	0.241

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	S.E.	Beta		
(Constant)	67.138	16.136		4.161	0.001
Renta	0.007	0.001	0.851	9.655	0.000
Densidad	-0.078	0.062	-0.089	-1.263	0.224
Inflación	-0.294	0.136	-0.187	-2.156	0.046

Diagnóstico de colinealidad

Es posible obtener con el modelo de regresión una serie de indicadores que estiman como de afectado puede estar nuestro modelo por la relación entre las variables explicativas

¿Qué coeficientes están afectados?

Se pueden emplear medidas relacionadas con el R^2 (bondad de ajuste)

- ✓ **Factor de Inflación de la Varianza ($1/1-R^2$) → VIF > 10**
- ✓ **Tolerancia ($1-R^2$) → Valor mínimo .10**

	Collinearity Statistics	
	Tolerance	VIF
(Constant)		
Renta	0.598	1.673
Densidad	0.942	1.062
Inflación	0.620	1.614

¿Hasta donde alcanza?

Condition Number < 30

	Eigenvalue	Condition Index	Variance Proportions			
			(Constant)	Renta	Densidad	Inflacion
1	2.708	1.000	0.02	0.02	0.04	0.01
2	0.945	1.693	0.00	0.04	0.01	0.41
3	0.264	3.202	0.05	0.17	0.92	0.01
4	0.083	5.711	0.94	0.78	0.03	0.56

Posibles soluciones:

- ✓ Eliminar variable del modelo
- ✓ Cambiar escala de medida: transformar variables (centrar, estandarizar)

Variables dummy

Los modelos de regresión están pensados para trabajar con variables cuantitativas, pero en la práctica solemos estar interesados en introducir también variables categóricas

La categoría a la que se le asigna el valor 0 es la que se usa como referencia y con la que se compara el resto

Variables dicotómicas → SI = 1; NO = 0

- ✓ Presenta o no una característica: Enfermedad (S/N), Sexo (H/M)
- ✓ Se está o no por encima de cierto umbral...



¿Y con variables categóricas?

- ✓ Nivel de estudios: Primaria, Secundaria, Bachillerato, Universitarios
- ✓ Religión: Cristianismo, Protestantismo, Budismo, Hinduismo...

→ La solución es crear tantas variables como categorías menos 1. Nos aseguramos que nos mutuamente excluyentes

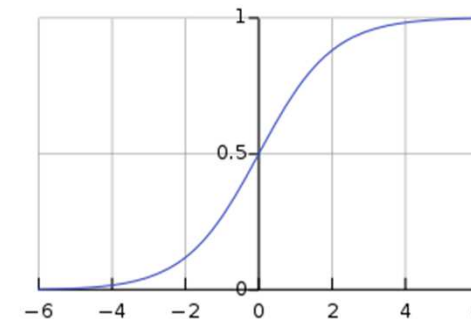
Otros modelos de regresión

La elección del modelo regresión va a depender de tipo y características de la variable dependiente

✓ Regresión logística

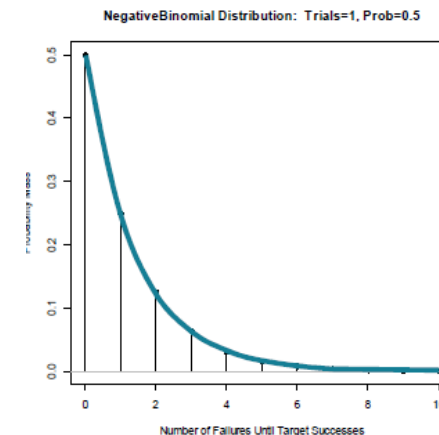
Variable dependiente dicotómica $\rightarrow 0 - 1$

Como el valor estimado debe ser una probabilidad entre 0 y 1 no podemos usar regresión lineal



✓ Regresión de Poisson / Binomial Negativa

Para situaciones donde tenemos frecuencias: nº de citas que reciben los artículos, nº de personas con una determinada característica



Análisis multivariante

¿Y si no hay variable dependiente y explicativas y todas están relacionadas?

Se dan cita 2 aspectos importantes

- 1) Descripción de una variable como una función de un conjunto de otras variables, es decir, **involucra múltiples variables que están correlacionadas entre sí**. Por tanto, el carácter multivariante de los datos no descansa únicamente en el número de variables sino en las múltiples combinaciones posibles entre las mismas
- 2) Hace referencia a la **reducción de la dimensión**. La idea subyacente es que aprovechando la estructura de relaciones entre las variables es posible simplificar el problema en estudio a unas pocas dimensiones con mínima pérdida de información.

- ✓ **Variable muy correlacionadas**
- ✓ **Distribuciones muy asimétricas**
- ✓ **Menor potencia de contraste de los test estadísticos**

Problemas desde una
perspectiva estadística clásica



Buenas noticias desde una
perspectiva multivariante!!!



Análisis factorial

Método de Análisis Multivariante que **intenta explicar las relaciones entre un conjunto de variables observables mediante un número reducido de variables hipotéticas** que se obtiene a partir de las **correlaciones** entre las variables observables

*Se estudia el desgaste profesional (burnout)
Este concepto no se puede medir directamente se emplea una serie de variables observables que representan factores o variables latentes que no son directamente observables*

Subescalas	Ítems
Agotamiento Emocional	1,2,3,6,8,13,14,16,20
Autoestima Profesional	4,7,9,12,17,18,19,21
Despersonalización	5,10,11,15,22

1. Me siento emocionalmente agotado por mi trabajo
2. Me siento cansado al final de la jornada de trabajo
3. Me siento fatigado cuando me levanto por la mañana y tengo que enfrentarme con otro día de trabajo

Factorial Exploratorio M.B.I.		Factor 1	Factor 2	Factor 3
Agotamiento Emocional	ítem-1	,843		
	ítem-2	,782		
	ítem-3	,751		
	ítem-8	,825		
	ítem-13	,696		
	ítem-14	,590		
	ítem-20	,717		
Autoestima Profesional	ítem-4		,510	
	ítem-7		,622	
	ítem-9		,713	
	ítem-12		,640	
	ítem-17		,536	
	ítem-18		,680	
	ítem-19		,694	
Despersonalización	ítem-21		,114	
	ítem-5			,565
	ítem-6	,240		,622
	ítem-10			,633
	ítem-11			,583
	ítem-15			,483
	ítem-16	,182		,745
	ítem-22			,522