

Curso Doctorado: Metodología y Técnicas de Investigación

Metodología y Técnicas de Investigación
Carmen Antón Martín
Juan Antonio Rodríguez Sanz
M^a Valle Santos Álvarez

PROCEDIMIENTOS DE ANÁLISIS DE DATOS

- Número de variables: 1, 2, más de 2
- Descripción vs Inferencia (muestra vs población)
- Nivel de medición
 - Var. Nominales
 - Var. Ordinales
 - Var. de intervalos o de Relación

ESTADÍSTICA DESCRIPTIVA

- Objeto: Proporcionar medidas de resumen de los datos contenidos en una muestra: Medidas de tendencia central y Medidas de dispersión.
- Medidas de tendencia central: Media, mediana, moda
- Media: variables de intervalo

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

- Mediana: variables en escala ordinal o de intervalos. Valor central cuando los datos se agrupan ordenados por la variable considerada

X1 10

X2 20

X3 70 (Mediana) Media=148

X4 140

X5 500

- Moda: Variables en escala nominal o superior. Categoría de variable nominal que ocurre con mayor frecuencia

- Medidas de dispersión. Complementan a las medidas de tendencia central. Media de la amplitud de la distribución de la variable.
- Desviación estándar. Datos de intervalos.

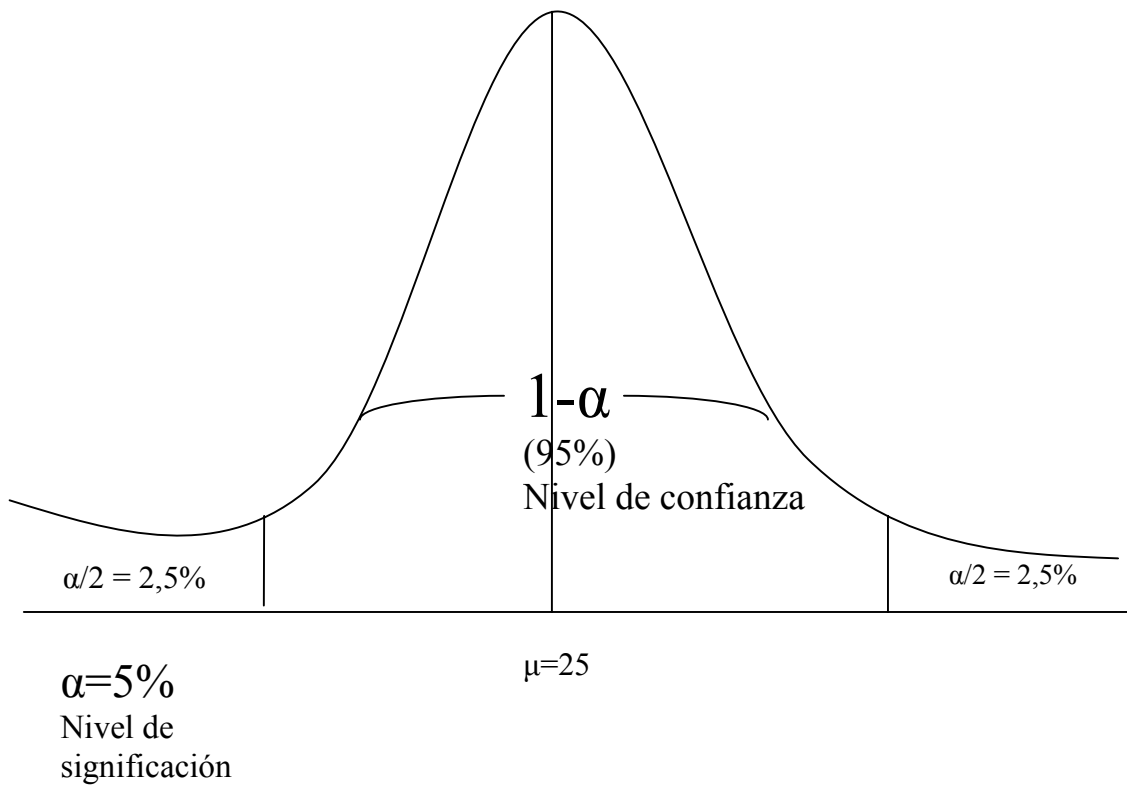
$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

- Coeficiente de variación

$$CV = \frac{s}{\bar{x}}$$

- Frecuencias relativas y absolutas. Variables en escala nominal y superiores. Gráficos de frecuencias.
- Rango intercuartílico. Usual para variables ordinales. Diferencia entre el percentil 75 y el 25.
- Medidas de oblicuidad: Asimetría, curtosis.

Resumen de errores. Prueba de hipótesis		
Condición verdadera		
	H_0 es verdadera	H_0 es falsa
No rechaza H_0 (»aceptar H_0)	<u>Dec. Correcta</u> Nivel de confianza Prob = $1-\alpha$	<u>Error tipo II</u> Prob = β
Rechaza H_0	<u>Error tipo I</u> Nivel de significación Prob = α	<u>Dec. Correcta</u> Potencia de la prueba Prob = $1-\beta$



Paso 1: Formular una hipótesis nula y una alternativa

La hipótesis nula supone que un parámetro de la población adquiere determinado valor, por ejemplo la edad media de la población de la comunidad universitaria es 25 años.

$$H_0: \mu=25$$

$$H_1: \mu \neq 25$$

Prueba bilateral: Se trata de comprobar si la hipótesis nula es rechazada, por lo tanto, si la alternativa puede ser aceptada. Se dice que no tenemos evidencia para rechazar una hipótesis nula, nunca diremos que es aceptada, siempre se pueden encontrar nuevos datos que contradicen H_0 .

Al examinar la distribución muestral de un estadístico (p.e. valor medio), podemos determinar si el valor de la muestra es suficientemente diferente o no del valor pronosticado en H_0 . Si es tan diferente como para caer en las colas podemos suponer que no es debido a un error muestral y diremos que H_0 es rechazada con un cierto nivel de significación (α) o error de tipo I o con un cierto nivel de confianza ($1-\alpha$).

Prueba unilateral:

$$H_0: \mu = 25$$

$$H_1: \mu \neq 25$$

Error tipo I o Error tipo α

Rechazar H_0 cuando verdaderamente es cierta.

Datos de la muestra \rightarrow estadístico o valor a comprobar

Cuando + lejos esté de H_0 más probable es que sea falsa \rightarrow se rechaza.

Se puede cometer un error: rechazar algo que sea cierto: error a minimizar (α). Motivo: elección muestral inadecuada: estadístico puede estar realmente en la cola de la distribución muestral: caso aislado. Rechazo de hipótesis y error.

Se fija α : proporción de la distribución muestral alejada de H_0 que hace que esta sea rechazada, aunque sea cierta: % de error asumido: 5%

También podemos hablar de la probabilidad de no rechazar H_0 (\gg aceptar) cuando es verdadero: Nivel de confianza: 95%. Proporción de la distribución muestral de un estadístico que está dentro de una cierta distancia del verdadero valor.

PASOS EN LA PRUEBA DE HIPÓTESIS

1) Formular hipótesis nula y alternativa

Prueba z o t para análisis univariante: permite comparar la media generada en una muestra con una media teóricamente existente en una población y concluir si la media hipotética de la población es cierto o no.

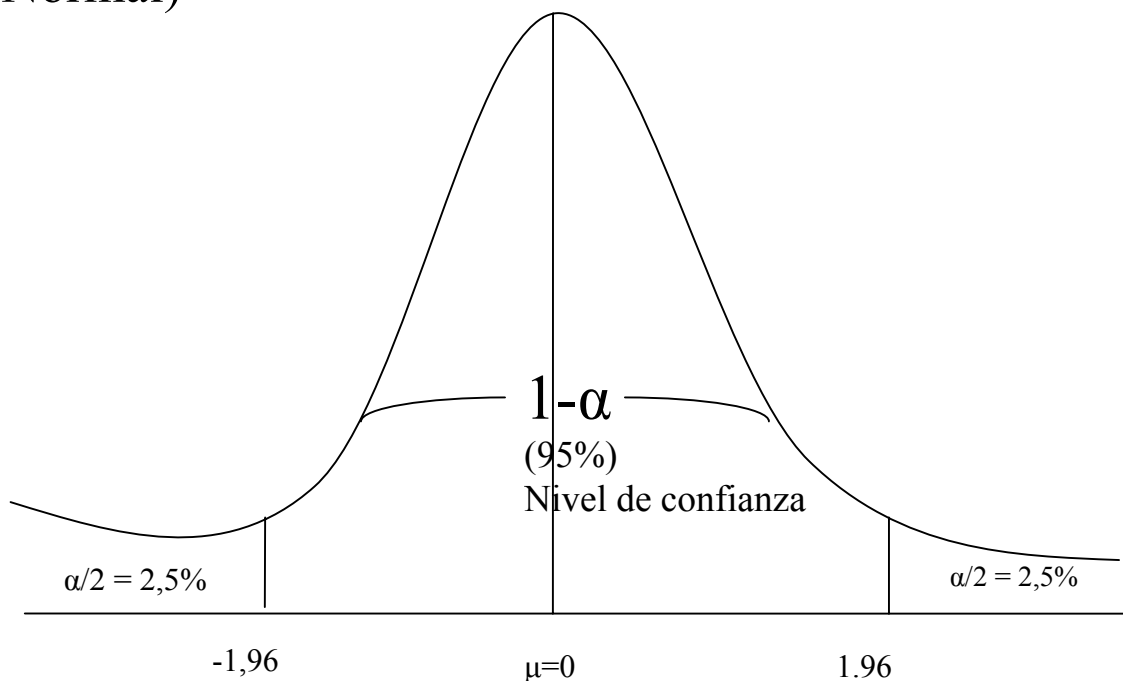
z (cualquier tamaño muestral y σ de la población conocida o tamaño > 30 y no se conoce σ).

$x=24$; $s=5$; $n=100$

$$H_0: \mu=23$$

$$H_1: \mu \neq 23$$

2) Selección de la prueba estadística apropiada: z (TCL: la distribución muestra de la media sigue una Normal)



3) Especificar un nivel de significación: p.e. $\alpha=0,05 = 0,01$

4) Buscar el estadístico de prueba en las tablas para un α dado

$\alpha=0,1$: $z=1,64$

$\alpha=0,05$; $z=1,96$

5) Realización de la prueba estadística (la del paso 2) sobre los datos de la muestra disponibles.

Si se conoce la σ de la población:

$$z = \frac{\bar{x} - \mu}{\sigma_x} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Si no se conoce la σ de la población:

$$z = \frac{\bar{x} - \mu}{s_x} = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

μ : media teórica

\bar{x} media observada

s/\sqrt{n} : error estandar

$$z = \frac{24 - 23}{\frac{s}{\sqrt{100}}} = \frac{1}{0,5} = 2$$

La diferencia de medias es 2 veces el error estandar.

6) Comparación del valor del estadístico obtenido en el paso 5 con el valor del paso 4. Si es $>$ se rechaza H_0 .

$2 > 1,96$: Rechazo de la hipótesis nula. No se puede concluir que la edad media de la población sea 23.

DATOS NOMINALES: JI CUADRADO

Para variables de tipo nominal: el interés suele radicar en realizar inferencias sobre la forma en la que se distribuyen los encuestados o individuos de la muestra a través de las categorías de una variable nominal.

$$X^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

k número de categorías

O_i n° observaciones en la categoría i

E_i n° hipotético de encuestados en la categoría i

X^2 tabla para diversos grados de libertad: k-1 y para distintos valores de α

Ejemplo: Distribución ocupacional de usuarios de envases retornables

ESTADISTICA DESCRIPTIVA PARA 2 VARIABLES DE INTERVALOS

COEFICIENTE DE CORRELACIÓN LINEAL

Es una medida del grado hasta el cuál están asociadas dos variables de intervalos: r_{xy} .

$$x = x_i - \bar{x}$$

$$y = y_i - \bar{y}$$

$$xy = (x_i - \bar{x})(y_i - \bar{y})$$

xy : Nos da idea con su signo de la dirección total de la relación. Tiene dos sesgos: tamaño de la muestra y escala.

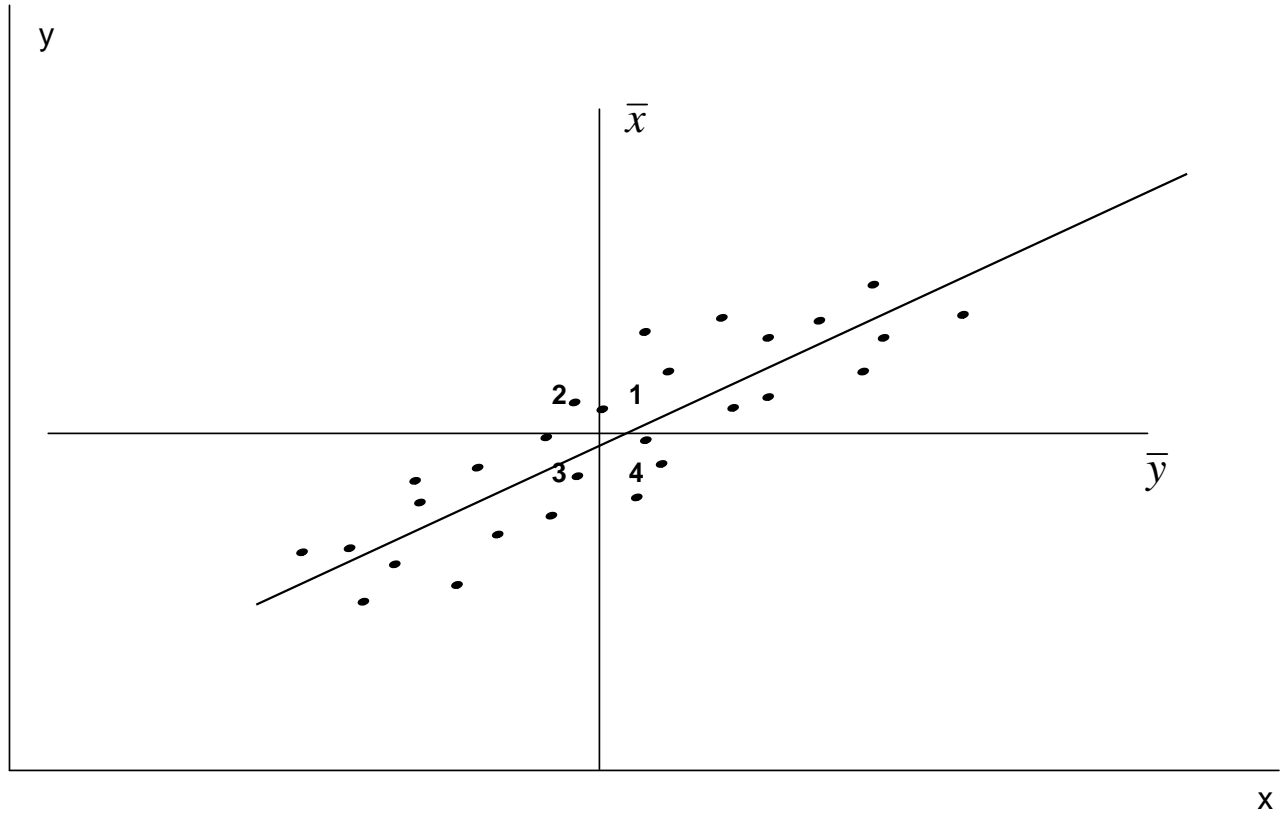
$$Cov(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Resuelve el defecto del tamaño

Para el efecto de la escala se divide por el producto de las estimaciones

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

Medida estandarizada de la covariación y comprendida entre 0 y 1



PRUEBA z o t SOBRE LA DIFERENCIA DE MEDIAS

Se trata de analizar si una diferencia observa entre 2 medias generadas por una muestra es lo suficientemente grande o no para ser considerada significativa. (Similar a la univariante).

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

Si no se conocen σ_1 y σ_2 y se supone que no son iguales:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_{x_1}^2 + s_{x_2}^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Si no se conocen σ_1 y σ_2 y se supone que son iguales:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} + \frac{n_1 + n_2}{n_1 \cdot n_2}}$$

REGRESIÓN SIMPLE

Var. dependiente en escala de intervalos y un cierto número de independientes tb. en escala de intervalos.

Se trata de adaptar una relación lineal a los datos arrojados por X e Y: ha de ser el mejor ajuste

y_i observación i-ésima de la var. Dependiente

\bar{y} media de la variable dependiente

\hat{y}_i valor pronosticado de la observación i-ésima de la variable dependiente

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

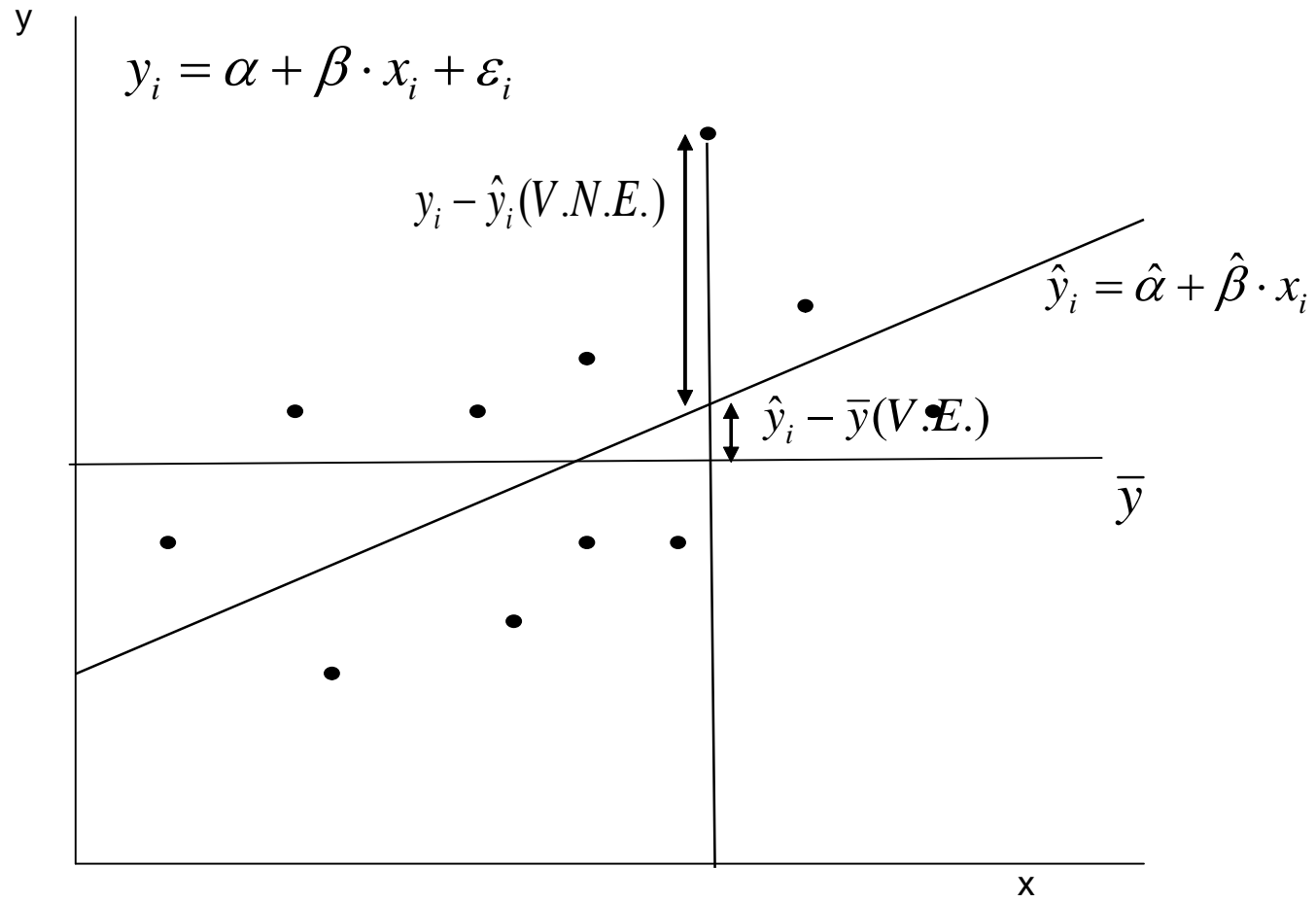
Variación Total = Var. Explicada por la regresión + Var. No explicada por la regresión

Se suman las desviaciones para todas las observaciones y se elevan al cuadrado

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

SC Totales = SC explicados por la regres. + SC no explicados por la regresión

Variación total = Variación explicada + Var. no explicada



Procedimiento para obtener una línea de regresión bien ajustada a los datos. Minimizar la suma de los cuadrados de los errores (MCO): minimizar los errores producidos en la estimación de la línea de la regresión.

$$\hat{y}_t = \hat{\alpha} + \hat{\beta} \cdot x_t$$

$$e_t = y_t - \hat{y}_t$$

$$\text{Min} \sum_{t=1}^T e_t^2$$

$$\hat{\beta} = \frac{\sum x_t - y_t}{\sum x_t^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x}$$

significado de los coeficientes

$$\hat{y}_t = \hat{\alpha} + \hat{\beta} \cdot x_t$$

$$\hat{y}_t = 6.2 + 0.93 \cdot x_t$$

Y: Calificación del estudiante en un curso de gestión

X: calificación del estudiante en un curso en marketing

$$\sum_{t=1}^T e_t^2 = \sum (y_t - \hat{y}_t)^2 = SCE = 334.17$$

Variación explicada: Coeficiente de determinación: R^2

$$R^2_{xy} = \frac{\text{Variación explicada}}{\text{Variación total}} = \frac{SCR}{SCT} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 0.77$$

ESTADÍSTICA INFERENCIAL

ESTADÍSTICO t SOBRE EL COEFICIENTE DE REGRESIÓN

$$H_0: \beta=0$$

$$H_1: \beta \neq 0$$

Error estandar del estimativo

$$s_{yx} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

Error estandar del coeficiente de regresión:

$$s_b = \sqrt{\frac{s_{yx}}{\sum x_i^2}}$$

$$t = \frac{b}{s_b}$$

Se compara con el valor de la distribución de una t de student de n-2 grados de libertad. $t=5.17$ $t(5\%)= 1.96$: H_0

ESTADÍSTICO F SOBRE SCR

SCR y SCE se convierte a varianzas dividiendo entre los grados de libertad. Varianza de la regresión y varianza del error: se comparan.

$$F = \frac{\text{Varianza explicada}}{\text{Varianza no explicada}} = \frac{SCR / 1}{SCE / (n - 2)}$$

$F=26.71$; $F^1_{n-2} (\alpha=0.05) =5.32$ Se acepta H_0

Datos de panel. Stata 8.0, 9.0.

Instrucciones

1. Convertir datos .xls a datos formato Stata (.dta)
2. Entrar en Stata: Comandos
3. set memory 20000
4. set matsize 80
5. tsset iden year, yearly (define el panel de datos)
6. xtreg vd vi1 vi2 vi3, fe i(iden)
7. xtreg vd vi1 vi2 vi3, re i(iden)
8. xttest0 (para contrastar la existencia de efectos fijos)
9. xthaus (para contrastar su correlación con el resto de v.i.: permite decantarse por fixed effects o random effects)