Pràctica 6. Anàlisi exploratòria de variables numèriques: una variable i diverses variables

Aquesta pràctica comença repassant l'anàlisi exploratòria que havíem vist en la pràctica 3, afegint la possibilitat de dividir en grups les observacions. No utilitzarem la segmentació de la pràctica 4, però si que farem servir la selecció de casos.

També utilitzarem l'anàlisi exploratòria per comparar diverses variables corresponents als mateixos casos. En aquesta situació, farem servir el mètode de les transformacions de dades de la pràctica 5 per calcular increments o variacions de certes variables. Finalment introduirem el gràfic de màxims i mínims.

▶ Recordeu activar les opcions:

- "Mostrar comandos en anotaciones" a la pestanya de "Visor".
- "Nombre y etiquetas" per a les variables i "Valores i etiquetas" per als valors l'apartat de "Etiquetado de tablas pivot" de la pestanya de "Etiquetas de resultados".

▶ En aquesta pràctica treballarem amb l'arxiu *pibhabeuropa.sav*. És un fitxer obtingut de la base de dades Eurostat i conté el producte interior brut per habitant de diverses regions d'Europa.

0. Etapa preliminar: Selecció de casos

Els tres primers casos (les tres primeres files del fitxer) no corresponen a cap regió d'Europa: són resums (són les mitjanes de l'Europa de 25 estats (eu25), l'Europa dels 15 prèvia a l'ampliació de 2004 (eu15), i els nous estat membres (nms10)). En conseqüència, filtrarem els tres primers casos utilitzant la variable *"incorporacio" (incorporació)*. Aquesta variable codifica amb "0" les dades de resum i amb "1", "2" o "3" les regions segons la seva incorporació a la Unió Europea. Per tant ens quedarem amb les dades per a les quals la variable *"incorporacio"* prengui valors iguals o superiors a 1.

Per fer la selecció de casos, anem a la barra del menú i seleccionem:

Datos $-->$ **Selecionar casos**

Activem **Si se satisface la condición** i cliquem a **Si...**

Ens apareix la finestra Seleccionar casos: Si. Seleccionem la variable *incorporacio* i cliquem la fletxa de manera que entri en el quadrat blanc. Allà escrivim

incorporacio $>= 1$

o bé

incorporacio $\sim= 0$

i cliquem **Continuar** i **Aceptar**. Veiem que els tres primers casos estan filtrats (i recordeu que es crea una nova variable de filtre).

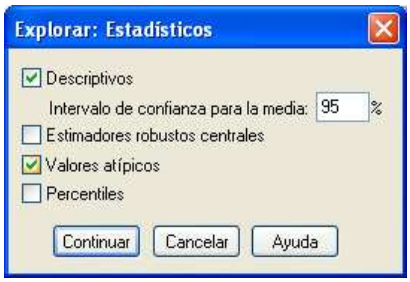
1. Anàlisi exploratòria d'una variable

1.1. **Exploració (sense subgrups).** Un cop filtrats els casos passem a l'anàlisi exploratòria de la variable *pibhab2002*. En el menú activem:

Analizar − − > Estadisticos descriptivos − − > Explorar

A la casella Dependientes hi posem la variable *PIB per habitant 2002* i a Etiquetar los casos mediante la variable *Codi Regió*.

Deixem per defecte de la casella Ambos (gràfics i estadístics). Cliquem a Estadísticos i activem l'opció Valores atípicos, per poder veure els valors extrems.



Obtenim

Explorar

Resumen del procesamiento de los casos

	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
PIB per habitant 2002 (euros/hab)	249	100,0%	0	,0%	249	100,0%

Descriptivos

			Estadístico	Error típ.
PIB per habitant 2002 (euros/hab)	Media		20169,553	608,0830
	Intervalo de confianza para la media al 95%	Límite inferior	18971,888	
		Límite superior	21367,219	
	Media recortada al 5%		19779,653	
	Mediana		21326,200	
	Varianza		92071463	
	Desv. típ.		9595,3876	
	Mínimo		3605,4	
	Máximo		75025,2	
	Rango		71419,8	
	Amplitud intercuartil		10655,2	
	Asimetría		,836	,154
	Curtosis		4,003	,307

Valores extremos

			Número del caso	CodiRegió	Valor
PIB per habitant 2002 (euros/hab)	Mayores	1	159	uki1	75025,2
		2	107	lu	51110,5
		3	4	be10	50771,0
		4	31	de60	44150,5
		5	131	se01	39651,5
	Menores	1	252	sk04	3605,4
		2	236	pl31	3707,5
		3	237	pl32	3777,1
		4	251	sk03	3953,8
		5	246	pl62	3955,9

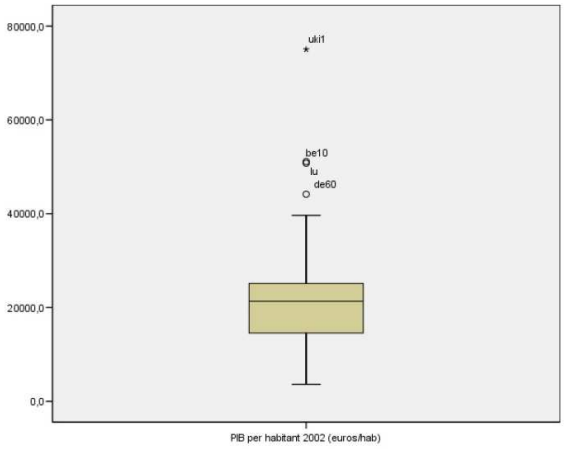
PIB per habitant 2002 (euros/hab)

PIB per habitant 2002 (euros/hab) Stem-and-Leaf Plot

```
 Frequency    Stem &  Leaf

     5,00        0 .  33333
    23,00        0 .  44444444444444455555555
     7,00        0 .  6666667
     3,00        0 .  899
     9,00        1 .  000011111
    13,00        1 .  2222233333333
    14,00        1 .  44444444455555
    22,00        1 .  66666666667777777777
    12,00        1 .  888888899999
    28,00        2 .  0000000000011111111111111111
    30,00        2 .  222222222222222223333333333333
    35,00        2 .  44444444444444444455555555555555555
    16,00        2 .  6666666677777777
     7,00        2 .  8888899
     5,00        3 .  00111
     4,00        3 .  2222
     4,00        3 .  4455
     4,00        3 .  6667
     4,00        3 .  8889
     4,00 Extremes    (>=44151)

 Stem width:    10000,0
 Each leaf:       1 case(s)
```



Les etiquetes de les regions corresponents als valors extrems són:

- "uki1" (Londres)
- "be10" (regió de Brusel·les)
- "de60" (Hamburg)
- Segons la taula, tenim un altre valor extrem (tapat per "be10" en el diagrama) que és "lu" (Luxemburg).

► QÜESTIONS:

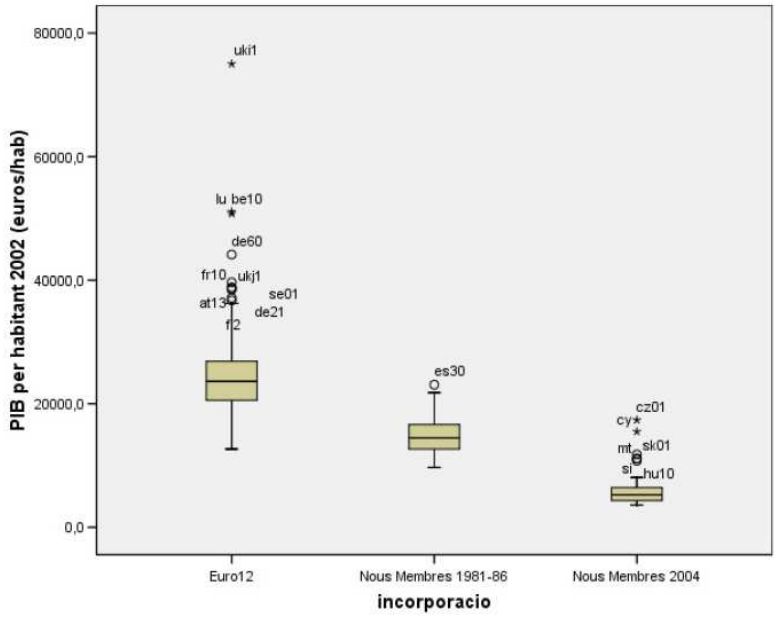
- Compareu els valors de la mitjana, la mitjana retallada i la mediana.
- Comenteu-ne també l'asimetria
- Calculeu-ne el coeficient de variació: $CV = \ldots\ldots\ldots$
- Comenteu la forma del diagrama de tija i fulles. Creieu que val la pena considerar subgrups?

1.2. **Exploració d'una variable considerant subgrups.** Analitzem la mateixa variable d'abans però ara per als diferents grups segons la data d'incorporació a la Unió Europea. La variable *incorporacio* codifica els tres grups:

1. Euro 12: països membres abans de 1980.
2. Incorporacions durant 1981-1996 (Espanya, Grècia i Portugal).
3. Nous Membres 2004.

Escollim de nou explorar, però ara, en la finestra "Explorar", en la casella "Factores" hi posem la variable "incorporació". En les altres caselles mantenim les variables que teniem (a la casella "Dependientes" la variable "PIB per habitant 2002" i a "Etiquetar los casos mediante" la variable "Codi Regió").

I obtenim els estadístics i diagrames corresponents. Només reproduïm els diagrames de caixa:



► QÜESTIONS:

- Compareu els tres grups de regions. Hi ha una diferència clara?
- Quines són les dispersions relatives de cada grup?
  - Euro 12: $CV = \ldots\ldots\ldots$
  - Incorporacions durant 1981-1996: $CV = \ldots\ldots\ldots$
  - Nous Membres 2004: $CV = \ldots\ldots\ldots$

2. Anàlisi exploratòria de diverses variables

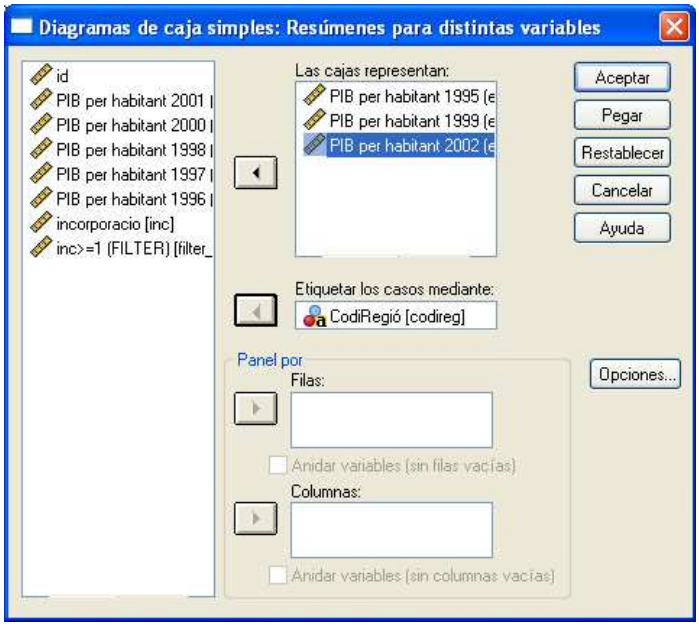
2.1. **Diagrames de caixa de diverses variables.** Comparem el PIB de diversos anys, és a dir de diverses variables, mitjançant els diagrames de caixa. Volem comparar les dades dels anys 1995, 1999 i 2002.

Des del menú activem:

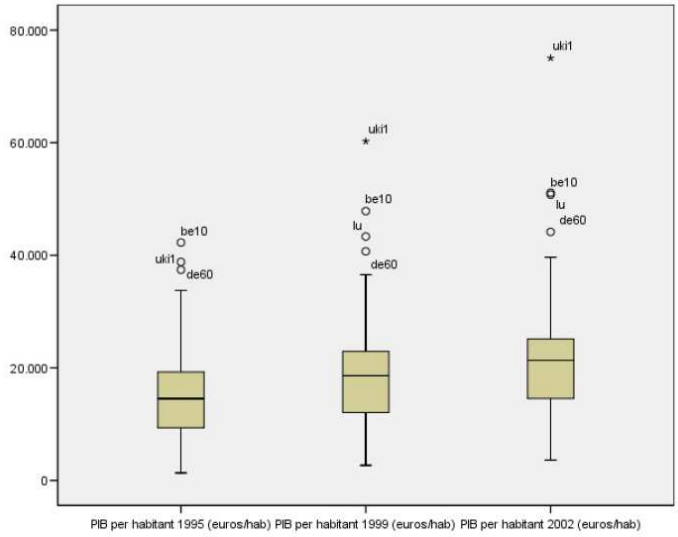
Gráficos – – > Diagramas de caja

i escollim les opcions:

Incorporem les variables per als anys 1995, 1999 i 2002 (a la casella Las cajas representan). Les entrem en aquest ordre, que és el que sortirà després als diagrames de caixa. Etiquetem les regions mitjançant el seu codi (a la casella Etiquetar los casos mediante posem la variable *Codi Regió*).



Obtenim:



Observem un augment dels tres quartils en el temps. El mínim augmenta molt més a poc a poc, mentre que el valor extrem "uki1" (Londres) augmenta espectacularment. Aquest augment no és aliè a la cotització lliura esterlina/euro.

2.2. **Definició de noves variables per comparar.** A vegades és útil calcular els increments o diferències per comparar variables. Això es fa per exemple quan comparem una variable mesurada en dos moments diferents (doncs tenim dues variables per als mateixos casos, és una situació de mostres aparellades).

En el nostre cas creem una nova variable, que sigui l'*increment percentual del PIB per habitant entre 1995 i 2002*. Fixem-nos que és la *taxa* de 2002 respecte a 1995, i l'anomenarem "taxapib". La fórmula per aquesta taxa és:

$$\frac{\textbf{pibhab2002} - \textbf{pibhab1995}}{\textbf{pibhab1995}} \times 100.$$

Activem el menú:

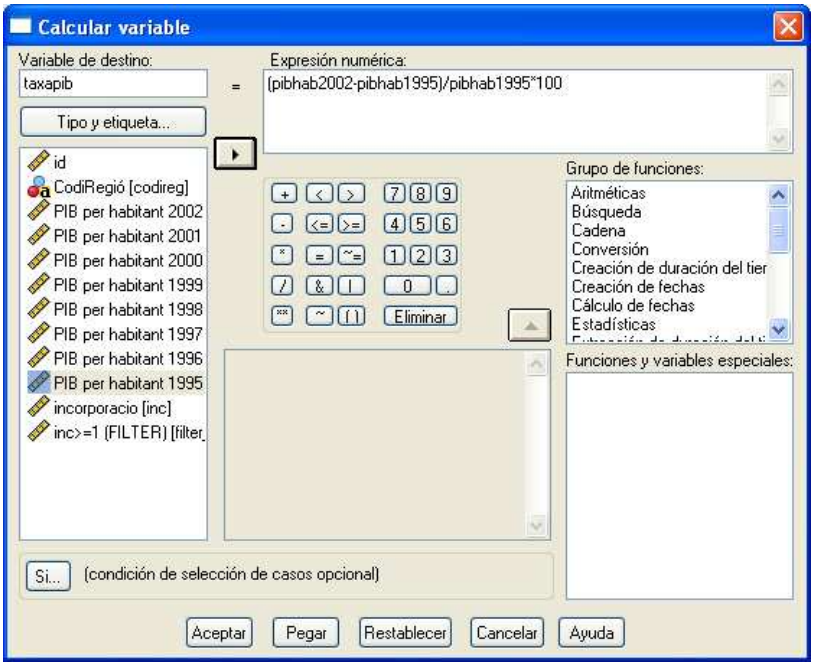
Transformar − − > Calcular

I s'obre la finestra "Calcular Variable". En la casella Variable de destino hi escrivim el nom de la nova variable *taxapib* i clicant al botó de sota Tipo y Etiqueta li posem l'etiqueta

"Taxa del PIB per hab. (1995-2002)"

Ara tornem a la finestra "Calcular Variable". A la casella "Expresión numérica:" hi posarem la fórmula:

(pibhab2002-pibhab1995)/pibhab1995*100

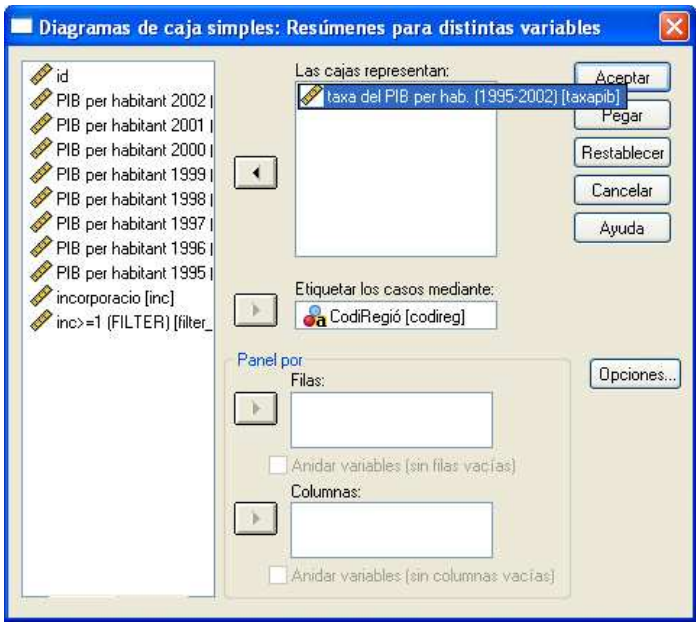


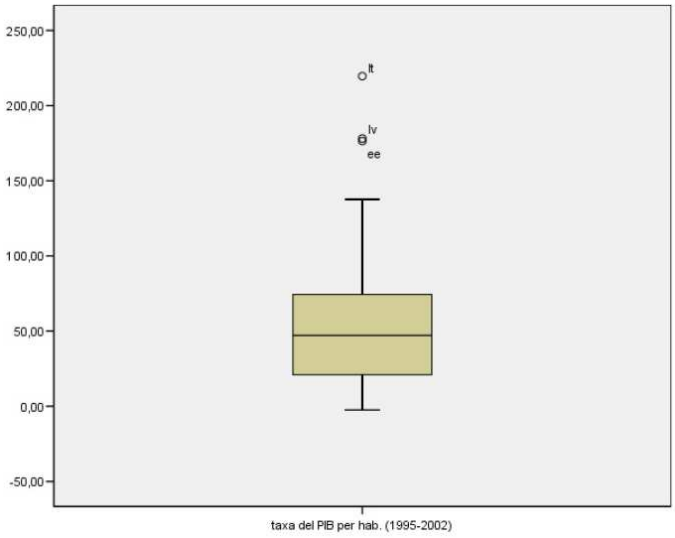
Observeu que no cal escriure els noms de les variables, només cal seleccionar-les i clicar la fletxa. Després cliquem a "aceptar".

Ja tenim la nova variable creada. En fem un diagrama de caixa amb el mètode habitual: activem

Gráficos − − > Diagramas de caja

i escollim les opcions **Simple** i **Resúmenes para distintas variables**. Posem la nova variable a Las cajas representan.





Els codis dels valors extrems que es veuen són:

- "lt" Lituania
- "ee" Estònia

▶ QÜESTIONS:

- Quin creieu que pot ser el tercer outlier que té l'etiqueta tapada?
- Observeu que el mínim és negatiu, a prop de zero. Què vol dir?

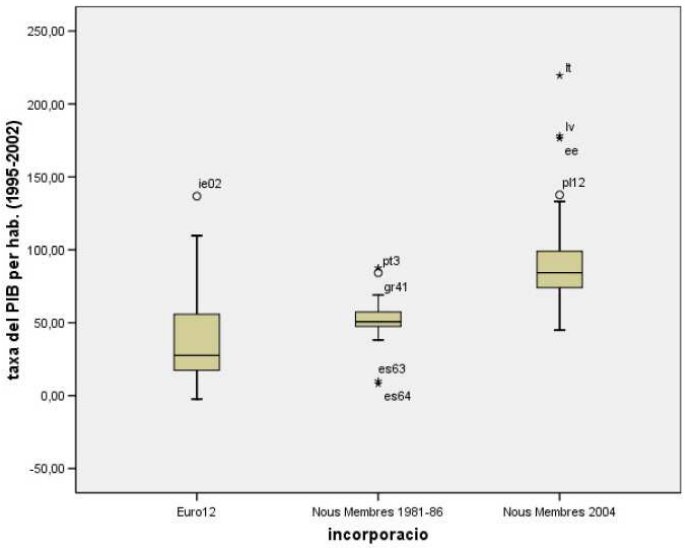
▶ El fet que els valors extrems siguin del nous estat membres, ens indueix a analitzar la variable per grups segons la data d'incorporació com en la secció 1.2. Aquest cop no només voldrem els diagrames de caixa, sinó que demanarem els estadístics de l'opció **Explorar**. Activem:

Analizar − − > Estadisticos descriptivos − − > Explorar

I omplim les caselles de la manera següent:

- Dependientes: *increment percentual entre 1995 i 2002 [taxapib]*
- Factores: *incorporacio [incorporacio]*
- Etiquetar los casos mediante: *Codi regió [codireg]*

A més dels descriptors i dels diagrames de tija i fulles, obtenim els diagrames de caixa:



Els codis dels valors extrems són:

- "ie02" sud i est d'Irlanda
- "es64" Ciutat autònoma de Melilla
- "es63" és el codi tapat per "es64", i es pot deduir fàcilment a qui correspon.
- "gr41" Voreio Aigaio (província egea, propera a Turquia).
- "pt3" Madeira.
- "lt" Lituània
- "ee" Estònia
- "pl12" Mazowieckie (província que conté Varsòvia).

► QÜESTIÓ:

- Compareu les dades dels tres grups, tant dels estadístics com dels diagrames de caixa.

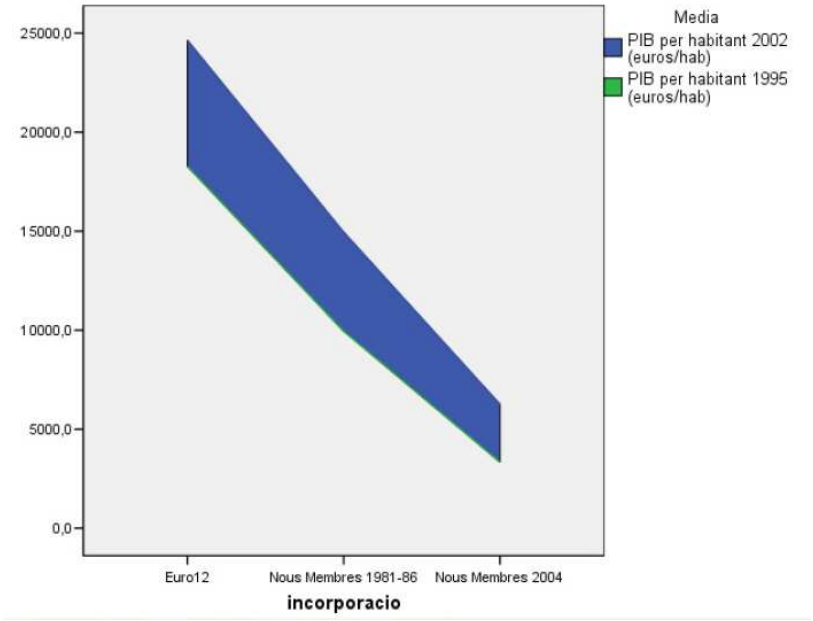
2.3. **Gràfic de màxims i mínims.** Observem que l'increment percentual augmenta en els països que s'incorporen més tard. Cal dir que aquests partien de més avall i volem comparar aquest increment en termes absoluts. Per a això ens serà útil fer un gràfic de màxims i mínims, que ens representa les diferents mitjanes de cada grup en els anys 2002 i 1995. Activem el menú:

Gráficos − − > Máximos y mínimos

S'obre una finestra "Gráficos/Máximos y mínimos", escollim les opcions:

- Área de diferencia.
- Resúmenes para distintas variables.

S'obre una finestra "Definir áreas de diferencias: Resúmenes para distintas variables". Els resums per defecte són les mitjanes i no les canviarem. A l'opció 1a posarem la variable *PIB per habitant 2002* i a la 2a, *PIB per habitant 1995*. A l'opció Eje de categorias, *incorporacio*.



▶ QÜESTIÓ:

- Quina conclusió en traieu?

2.4. **Exercici.** Obriu el fitxer *Mundo95.sav*

- Compareu els diagrames de caixa de les variables esperança de vida femenina i esperança de vida masculina.
- Definiu una variable que sigui la diferència entre l'esperança de vida femenina i la masculina. Feu-ne un diagrama de caixa un diagrama de màxims i mínims per a les diferens regions del món (variable "region"). Interpreteu els resultats.