

Pràctica 3. Anàlisi descriptiva univariable: variables quantitatives o numèriques

Aquesta tercera pràctica complementa l'anterior en l'objectiu d'introduir l'ús del programari SPSS per a l'anàlisi descriptiva d'una única variable, però ara referida a una variable que està mesurada a nivell numèric (variables quantitatives, ja siguin contínues o discretes, i que a l'SPSS s'identifiquen com d'"escala"). Per tal de realitzar aquesta anàlisi descriptiva procedirem a obtenir taules de distribucions de freqüències, gràfics per a representar la informació de les taules (histogrames, polígons de freqüències, diagrames de caixa i gràfics de tija i fulla), així com els estadístics de resum adients per a aquest tipus de variables: de tendència central (moda, mediana, mitjana), de posició o tendència no central (màxim, mínim i percentils), de dispersió (variància, desviació típica, coeficient de variació), i de forma de la distribució (asimetria i curtosi).


► 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 a l'apartat de “Etiquetado de tablas pivote” de la pestanya de “Etiquetas de resultados”.

► En aquesta pràctica treballarem de nou amb l'arxiu **GSS93 reducido.sav** del programari, ubicat a la carpeta de l'SPSS.

1. Taula de distribució de freqüències, histograma i estadístics

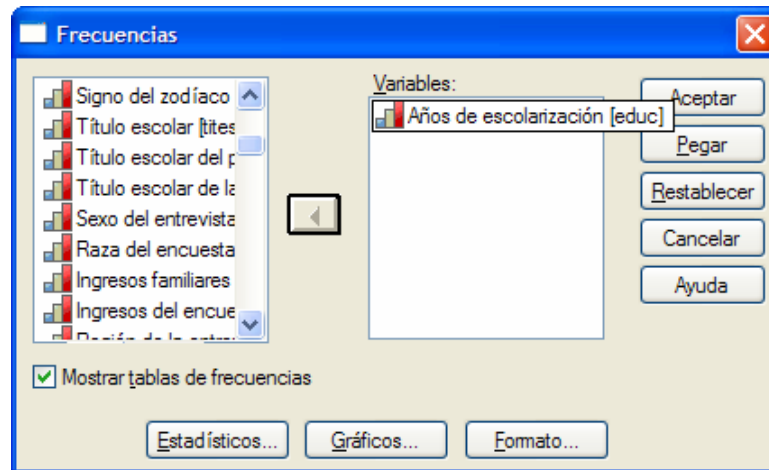
Considerarem la variable quantitativa discreta **EDUC**.

Codificada de tipus “**numèrica**” amb valors del 0 fins el 20 que indiquen el nombre d'anys d'escolarització de la persona entrevistada. A més la variable conté el valor 98, declarat valor perdut, que correspon als “no sap”. Si bé la variable és d'escala, com veurem, té el símbol  d'ordinal, està definit a la matriu aquest nivell de mesura, però no obsta el seu tractament quantitatiu.

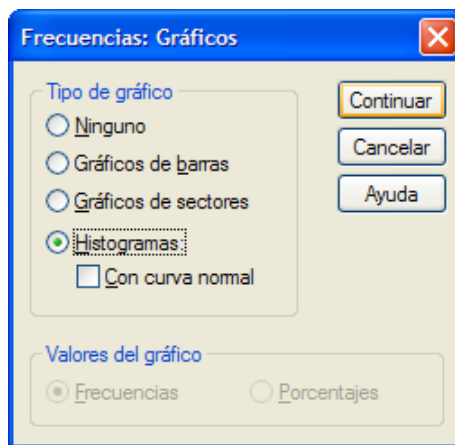
► A través del procediment “Frecuencias” demanaren: (1) la taula de distribució de freqüències, (2) l'histograma i (3) els estadístics. A través del menú:

Analizar / Estadísticos descriptivos / Frecuencias...

Ens apareix el quadre de diàleg del procediment “Frecuencias”:

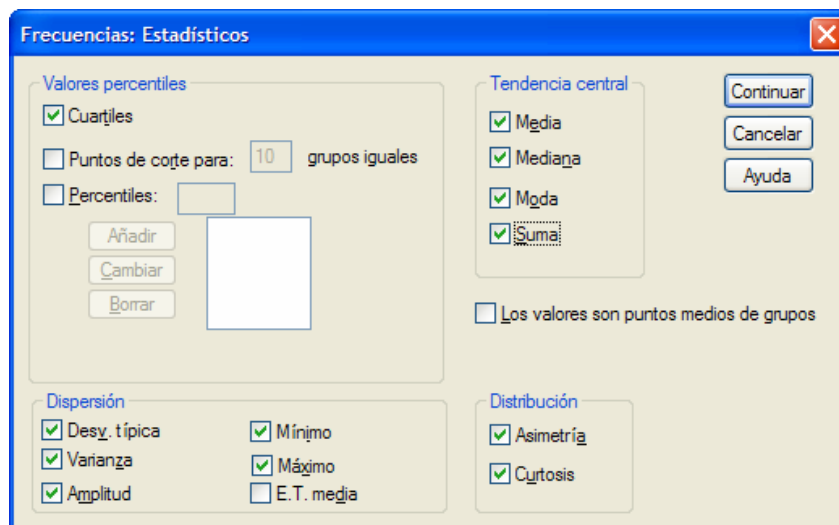


- (1) Seleccionem la variable **educ** i la col·loquem al requadre de “Variables”. D’aquesta manera obtenim la taula de distribucions de freqüències.
- (2) Per a demanar l’histograma cliquem sobre el botó “Gráficos...”. Ens apareix el quadre de diàleg on seleccionarem “**Histogramas**”:



I cliquem sobre “Continuar”.

- (3) Cliquem sobre el botó “Estadísticos...” i demanarem tots els estadístics de tendència central, els quartils, les mesures de dispersió i les de distribució:



I cliquem sobre “Continuar”. Finalment cliquem “Acceptar” al quadre de diàleg principal i observem els resultats següents:

Frecuencias

Estadísticos		
educ Años de escolarización		
N	Válidos	1496
	Perdidos	4
Media		13,04
Mediana		12,00
Moda		12
Desv. típ.		3,074
Varianza		9,450
Asimetría		-,309
Error típ. de asimetría		,063
Curtosis		,708
Error típ. de curtosis		,126
Rango		20
Mínimo		0
Máximo		20
Suma		19504
Percentiles	25	12,00
	50	12,00
	75	15,75

educ Años de escolarización					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	2	,1	,1	,1
	2	4	,3	,3	,4
	4	7	,5	,5	,9
	5	7	,5	,5	1,3
	6	20	1,3	1,3	2,7
	7	26	1,7	1,7	4,4
	8	59	3,9	3,9	8,4
	9	45	3,0	3,0	11,4
	10	55	3,7	3,7	15,0
	11	81	5,4	5,4	20,5
	12	445	29,7	29,7	50,2
	13	135	9,0	9,0	59,2
	14	166	11,1	11,1	70,3
	15	70	4,7	4,7	75,0
	16	208	13,9	13,9	88,9
	17	46	3,1	3,1	92,0
	18	71	4,7	4,7	96,7
	19	24	1,6	1,6	98,3
	20	25	1,7	1,7	100,0
	Total	1496	99,7	100,0	
Perdidos	98 No sabe	4	,3		
Total		1500	100,0		

- Les freqüències fan referència a 1496 casos dels 1500 totals, els 4 casos que corresponen als valors 98 de “No sabe” estan declarats com a valors perduts, i no es tindran en compte en els càlculs.
- La distribució de freqüències s’ha resumit i expressat a través de diversos estadístics que ens informen de les seves característiques:
 - Veiem en primer lloc que el conjunt de valors varien entre el **0** (el valor **mínim**) i el **20** (valor **màxim**). Per tant, el recorregut o el **rang** és **20**.
 - La **moda**, el valor més freqüent, és **12** anys d’escolarització, i correspon a un total de 445 casos, és a dir, el 29,7% de la mostra ha estat escolaritzat 12 anys. Amb variables quantitatives aquest estadístic tan sols és informatiu quan tenim un nombre reduït de valors, com és el cas.
 - La **mediana**, valor tal que el 50% d’observacions són inferiors a ell i el 50% superiors, és també **12**. El valor que correspon al percentatge acumulat del 50,2% és 12 anys d’escolarització. És a dir, el 50,2% dels individus tenen una escolarització igual o inferior a 12 anys, i, en particular, el 50% acumulat de la mostra correspon també al 12.
 - El mínim, el màxim, la moda i la mediana es poden observar directament sobre la taula de freqüències sense necessitat de consultar la taula d’estadístics.
 - Els **quartils** primer i segon (P_{25} i P_{50}) coincideixen amb el valor 12, el que revela una concentració d’efectius en els 12 anys d’escolarització. El tercer quartil (P_{75}) és 15,75, i s’ha calculat interpolant un valor aproximat entre el 15 i el 16:

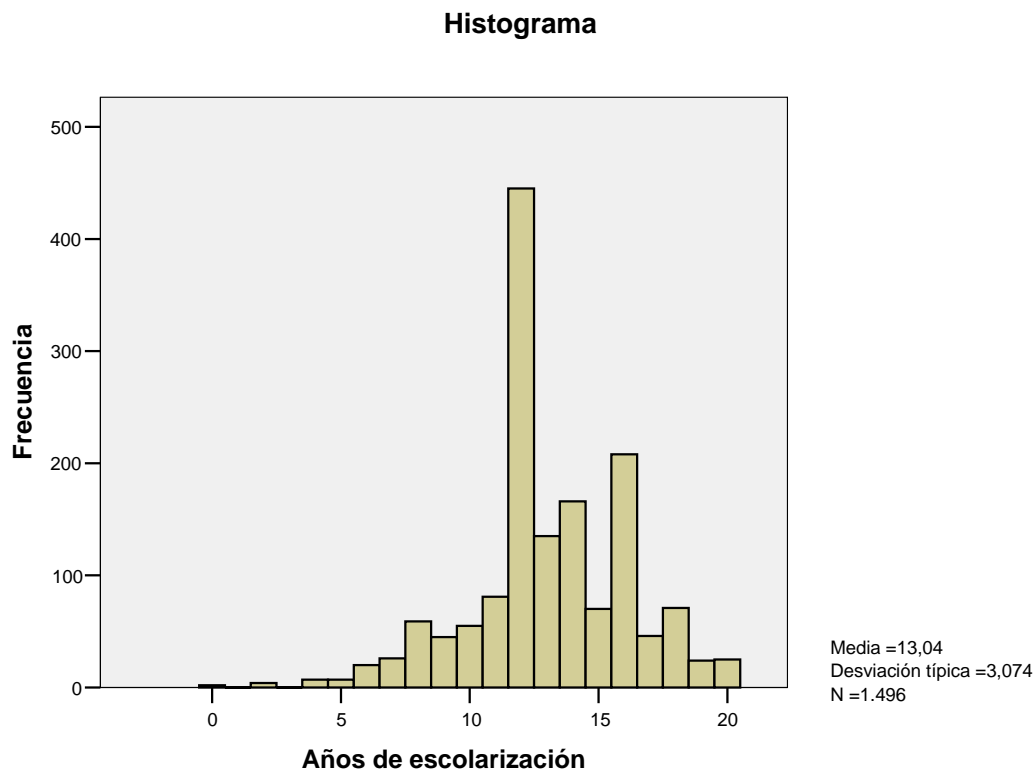
$$P_{75} = (0,25 \times 15) + (0,75 \times 16) = 15,75$$

- La **mitjana**, la suma de totes les observacions dividida per nombre total d’observacions, és de **13,04**, un valor superior als de la moda i la mediana. Aquest valor superior ve donat per la influència d’alguns individus amb molts anys d’escolarització.
- Amb la dada de l’estadístic de la **suma (19504)** podem comprovar el càlcul de la mitjana:

$$\text{Mitjana } \bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\sum_{i=1}^{1496} x_i}{1496} = \frac{19504}{1496} = 13,04$$

- La **variància** és **9,450**, i la **desviació típica 3,074**. És a dir, considerant les desviacions dels valors respecte del valor mitjà (13,04), la mitjana de totes aquestes desviacions és de 3,074 anys d’escolarització. Podeu comprovar amb la calculadora que la desviació típica és l’arrel quadrada de la variància.
- La mesura d’**asimetria** ens dona el valor **-0,309**, que ens indica, pel seu valor negatiu, l’existència d’un cert biaix cap a l’esquerra, és a dir, la presència de valors menys freqüents a l’esquerra i una major concentració de freqüències en els valors mitjans o alts.
- La mesura de **curtosi** de **0,708**, en ser positiva, ens informa de que es tracta d’una distribució apuntada en relació a la distribució normal (aquest aspecte es veurà més endavant en el curs).

Aquesta descripció de la distribució de la variable **educ** es pot observar igualment a través del seu histograma:



► La representació gràfica de l'histograma es pot editar per a modificar la seva aparença a través de la configuració d'algunes opcions o propietats del mateix.

- Per a editar-lo fem doble-clic sobre el gràfic a l'editor de resultats i entrem a l'"Editor de gráficos".
- Si no disposem de la "Ventana de propiedades" l'obrim a través del menú contextual o amb *Ctrl+T*.
- Podem realitzar, entre altres, els canvis següents:
 - Si seleccionem l'eix *x* (l'horitzontal, dels anys d'escolarització), a la pestanya de "Opciones del histograma" canviar "Tamaños de clase" per a considerar un nombre d'interval·ls o una amplada d'interval·ls personalitzats.
 - També a les propietats de l'eix *x* podem fer servir la pestanya "Escala" per a canviar el mínim, el màxim, l'increment i l'origen de l'escala.
 - Si el seleccionem, podem també operar aquests mateixos canvis sobre l'eix *y* (el vertical, les freqüències de cada interval·l).
 - Si sobre el gràfic despleguem el menú contextual podem seleccionar "Mostrar etiquetas de datos" per a veure impressionades les freqüències absolutes de cada interval·l.

- A través del menú “**Gráficos**” es pot obtenir igualment l’**histograma**:

Gráficos / Histograma...

O bé a través de **Gráficos / Interactivos / Histograma...**, opció que permet a més obtenir histogrames de freqüències acumulades.

- Ara repetirem la mateixa anàlisi descriptiva amb la variable **edad**.

Estadísticos

edad Edad del encuestado		
N	Válidos	1495
	Perdidos	5
Media		46,23
Mediana		43,00
Moda		28 ^a
Desv. típ.		17,418
Varianza		303,386
Asimetría		,500
Error típ. de asimetría		,063
Curtosis		-,700
Error típ. de curtosis		,126
Rango		71
Mínimo		18
Máximo		89
Suma		69109
Percentiles	25	32,00
	50	43,00
	75	59,00

a. Existen varias modas. Se mostrará el menor de los valores.

- Les freqüències fan referència a 1495 casos dels 1500 totals, els 5 casos que corresponen als valors 99 de “No contesta” estan declarats com a valors perduts, i no es tindran en compte en els càlculs.
- Completeu la informació de les afirmacions següents:
 - Veiem en primer lloc que el conjunt de valors varien entre el ____ (el valor **mínim**) i el ____ (valor **màxim**). Per tant, el recorregut o el **rang** és ____.
 - La **moda** (o modes) , són els valors més freqüents, ____ i ____ anys, i correspon a un total de ____ casos (el ____ %). Què implica la nota a peu de taula **a**? Recordeu que amb variables amb molts valors hi ha una tendència a obtenir freqüències baixes i, per tant, a trobar-nos més d’una moda. En aquests casos l’estadístic deixa de ser representatiu de la distribució de la variable.
 - La **mediana**, valor tal que el 50% d’observacions són inferiors a ell i el 50% superiors, és ____ . Aquest valor correspon al ____% acumulat i, per tant, fins a aquest valor hi ha acumulat el ____ % dels individus.

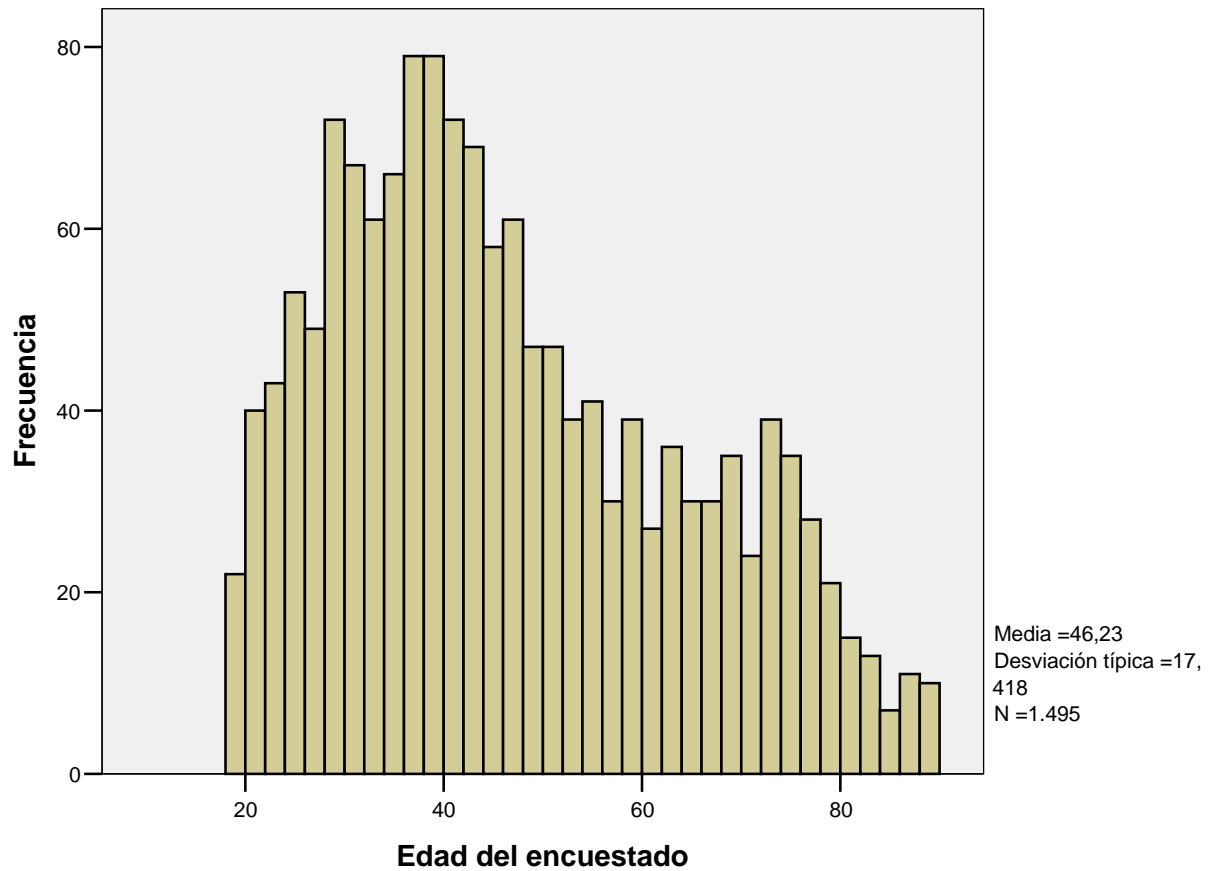
- Els **primer quartil** és el valor ____ i el **tercer quartil** és ____.
- Quin és el **percentil 10** (P_{10}) ? ____ . És a dir, el ____ % de la mostra té fins a ____ anys.
- Quin és el **percentil 90** (P_{90}) ? ____ . És a dir, el ____ % de la mostra té fins a ____ anys.
- La **mitjana**, la suma de totes les observacions dividida per nombre total d'observacions, és de ____, un valor més ____ que la mediana, per la influència d'alguns individus amb ____ anys. Podeu comprovar que la mitjana és la suma ____ dividida pel total d'individus ____.
- La **variància** és ____, i la **desviació típica** ____ . En quina unitat s'expressa la desviació? ____ I la variància? ____.
- La mesura d'**asimetria** ens dóna el valor ____, que ens indica l'existència d'un biaix cap a ____.

edad Edad del encuestado					edad Edad del encuestado						
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado			Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	18	5	,3	,3	,3	Válidos	56	12	,8	,8	72,0
	19	17	1,1	1,1	1,5		57	18	1,2	1,2	73,2
	20	18	1,2	1,2	2,7		58	25	1,7	1,7	74,9
	21	22	1,5	1,5	4,1		59	14	,9	,9	75,9
	22	15	1,0	1,0	5,2		60	16	1,1	1,1	76,9
	23	28	1,9	1,9	7,0		61	11	,7	,7	77,7
	24	23	1,5	1,5	8,6		62	17	1,1	1,1	78,8
	25	30	2,0	2,0	10,6		63	19	1,3	1,3	80,1
	26	27	1,8	1,8	12,4		64	13	,9	,9	80,9
	27	22	1,5	1,5	13,8		65	17	1,1	1,1	82,1
	28	42	2,8	2,8	16,7		66	19	1,3	1,3	83,3
	29	30	2,0	2,0	18,7		67	11	,7	,7	84,1
	30	36	2,4	2,4	21,1		68	16	1,1	1,1	85,2
	31	31	2,1	2,1	23,1		69	19	1,3	1,3	86,4
	32	28	1,9	1,9	25,0		70	9	,6	,6	87,0
	33	33	2,2	2,2	27,2		71	15	1,0	1,0	88,0
	34	25	1,7	1,7	28,9		72	19	1,3	1,3	89,3
	35	41	2,7	2,7	31,6		73	20	1,3	1,3	90,6
	36	42	2,8	2,8	34,4		74	18	1,2	1,2	91,8
	37	37	2,5	2,5	36,9		75	17	1,1	1,1	93,0
	38	41	2,7	2,7	39,7		76	13	,9	,9	93,8
	39	38	2,5	2,5	42,2		77	15	1,0	1,0	94,8
	40	36	2,4	2,4	44,6		78	14	,9	,9	95,8
	41	36	2,4	2,4	47,0		79	7	,5	,5	96,3
	42	30	2,0	2,0	49,0		80	6	,4	,4	96,7
	43	39	2,6	2,6	51,6		81	9	,6	,6	97,3
44	28	1,9	1,9	53,5	82	10	,7	,7	97,9		
45	30	2,0	2,0	55,5	83	3	,2	,2	98,1		
46	29	1,9	1,9	57,5	84	3	,2	,2	98,3		
47	32	2,1	2,1	59,6	85	4	,3	,3	98,6		
48	20	1,3	1,3	60,9	86	5	,3	,3	98,9		
49	27	1,8	1,8	62,7	87	6	,4	,4	99,3		
50	21	1,4	1,4	64,1	88	3	,2	,2	99,5		
51	26	1,7	1,7	65,9	89	7	,5	,5	100,0		
52	21	1,4	1,4	67,3	Total	1495	99,7	100,0			
53	18	1,2	1,2	68,5	Perdidos	99 No contesta	5	,3			
54	19	1,3	1,3	69,8	Total	1500	100,0				
55	22	1,5	1,5	71,2							

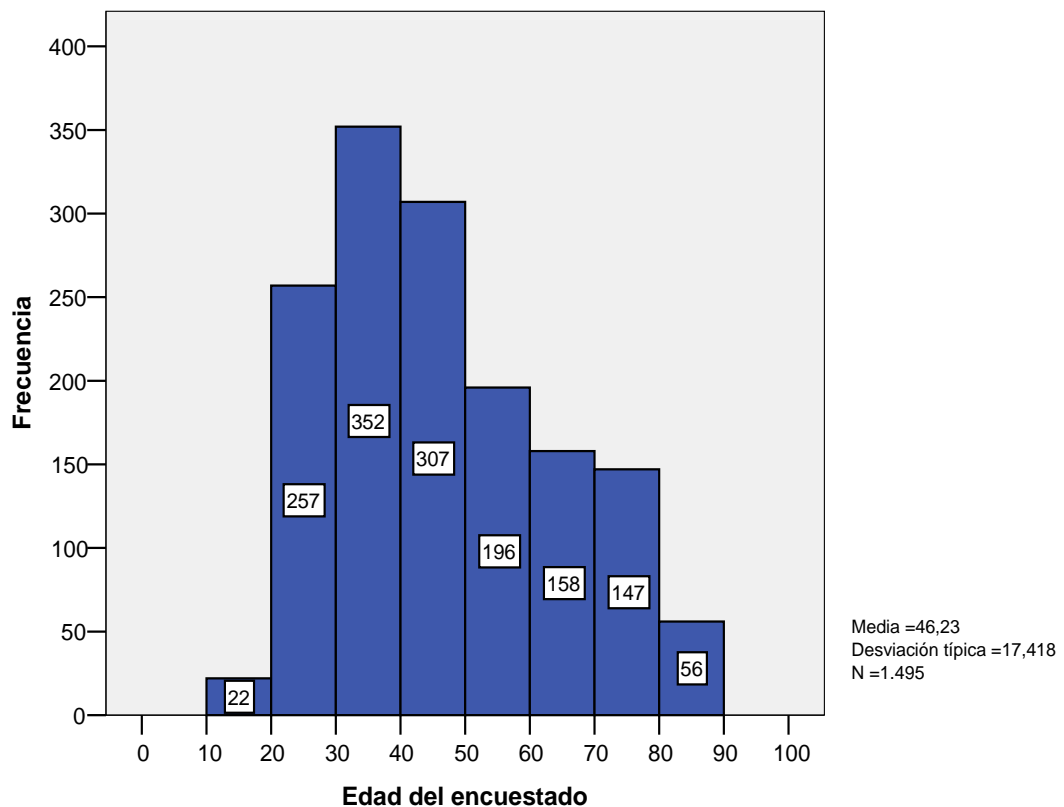
- A partir de l'**histograma** que genera el programari, editeu-lo i canvieu les opcions següents:
 - L'amplada dels intervals a través de la pestanya de "Opciones del histograma" una vegada seleccionat l'eix x. Considereu intervals d'una amplada de 10.
 - A la pestanya "Escala" canvieu el mínim a 0, el màxim a 100, i l'increment a 10.

- Seleccioneu l'eix *y* per a definir un increment de 50.
- Al menú contextual seleccioneu "Mostrar etiquetas de datos".

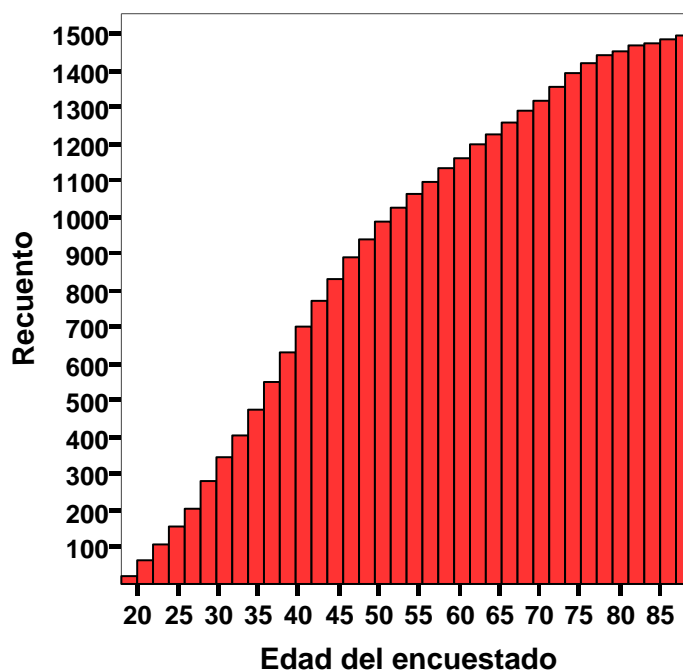
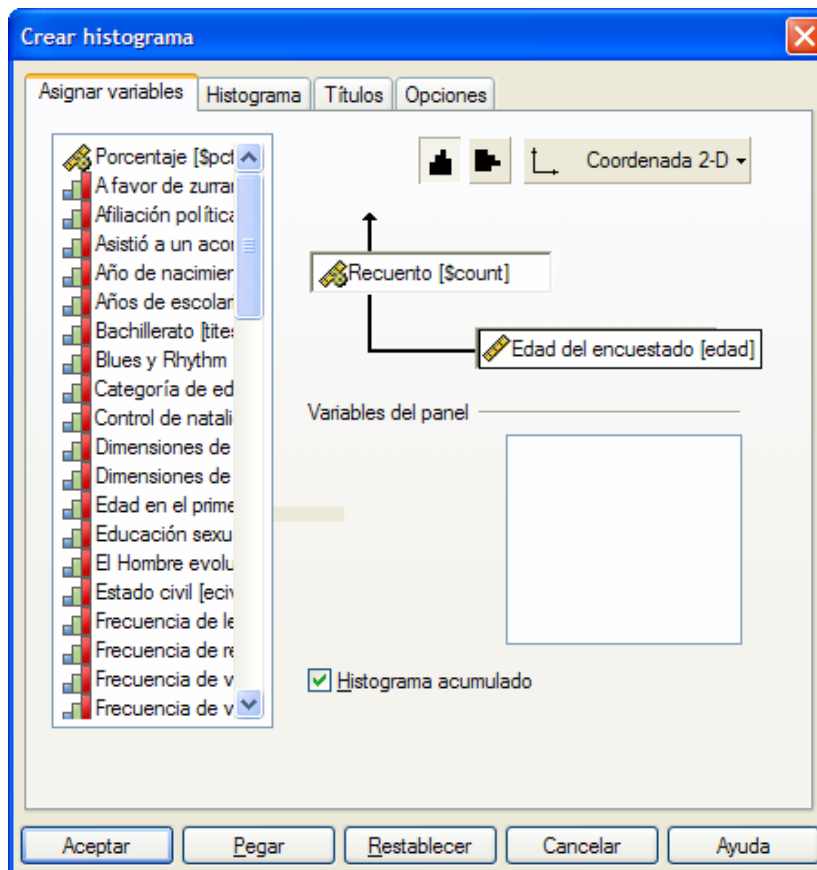
Histograma



Histograma modificat

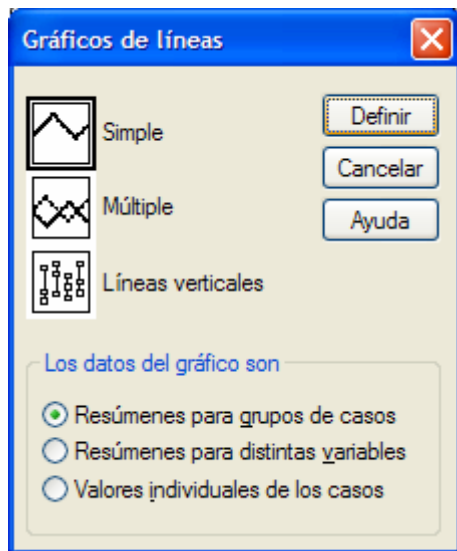


► Podem obtenir un histograma de freqüències acumulades si el demanem des dels gràfics “**Interactivos**”. Al quadre de diàleg seleccionem la variable **edad**, canviem el seu nivell de mesura a “escala” (menú contextual), l’arrosseguem fins a l’eix horitzontal i marquem “Histograma acumulado”.



► Una altre tipus de representació gràfica d'una variable quantitativa és el **polígon de freqüències**, una gràfic de línies que es construeix a partir de l'histograma unint els punts mitjans de les bases superiors dels rectangles. Aquesta representació **no** es pot obtenir amb l'SPSS, però sí podem obtenir un gràfic de línies. Per a obtenir-lo anirem al menú: **Gràfics / Línies...**

Ens apareix el quadre següent:



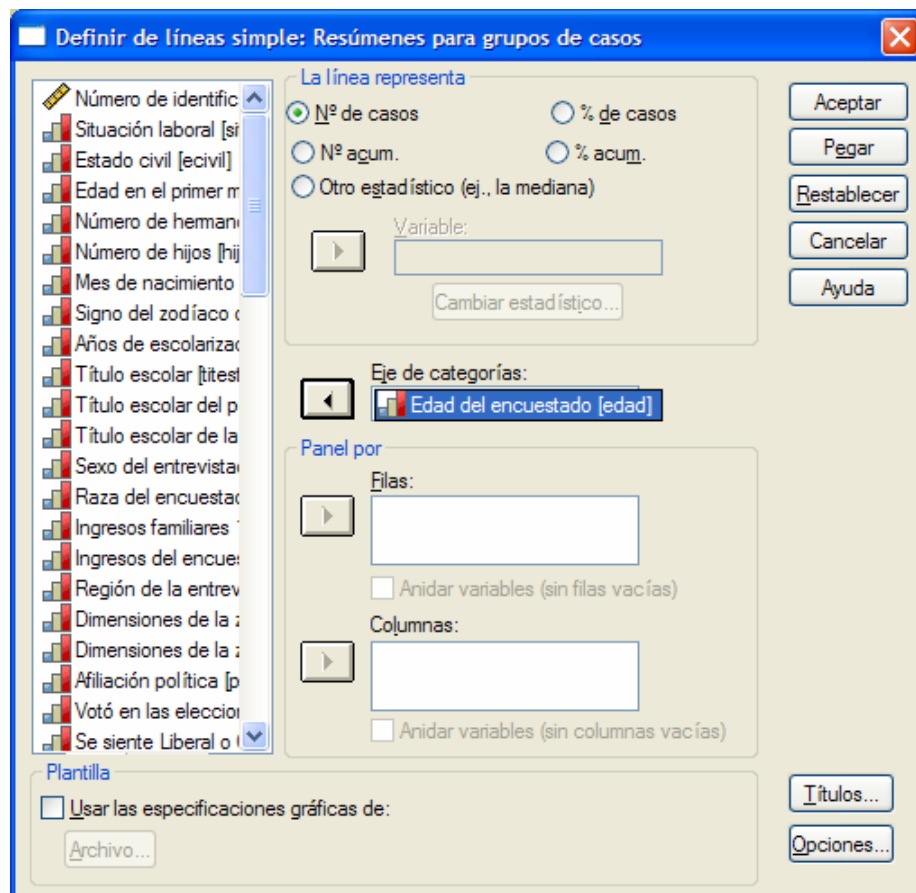
Optem per:

- “Simple”. Gràfic que mostra una sola línia que uneix un punt de cada categoria, cas o variable, de l'eix de categories.
- “Resúmenes para grupos de casos”: el gràfic resumeix una única variable dins dels subgrups definits per una variable categòrica.

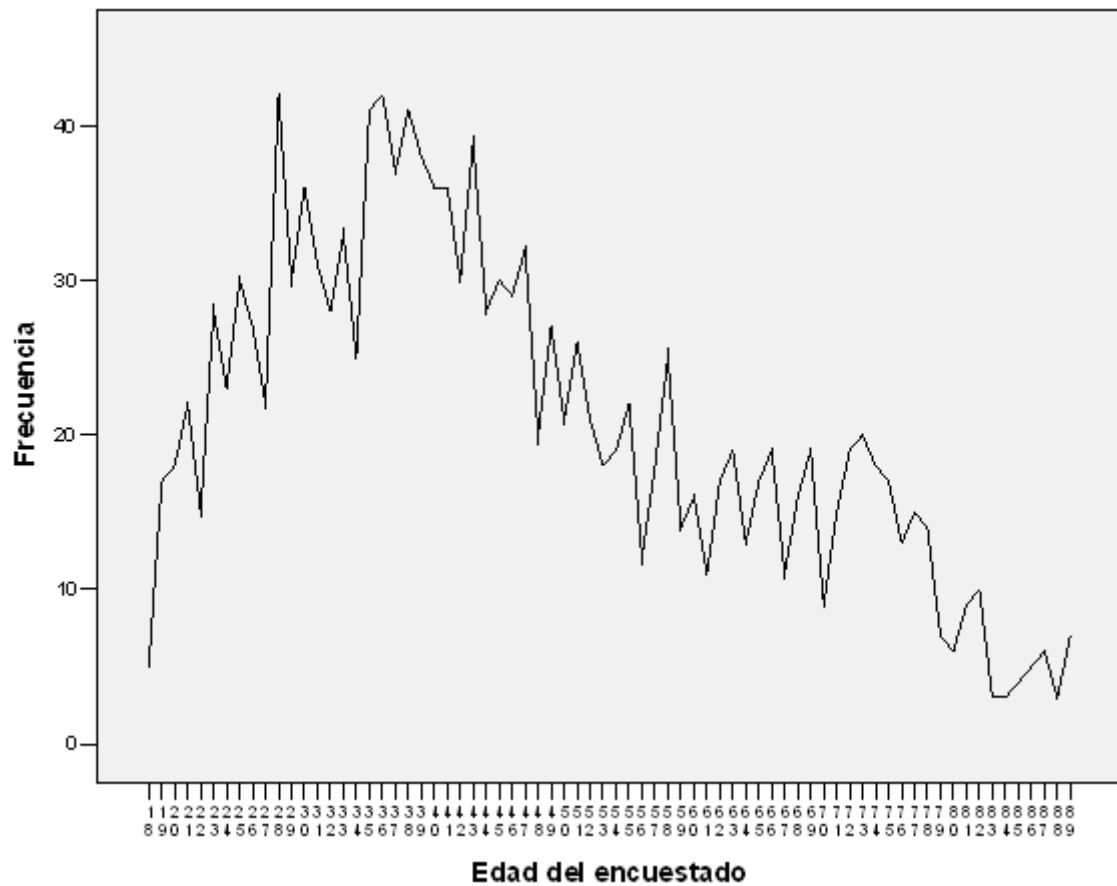
Són les opcions per defecte.

I cliquem sobre “Definir”.

Al quadre de diàleg col·locarem la variable **edad** a “**Eje de categorías:**”:



En clicar sobre “Aceptar” obtenim el resultat següent:



► Quan disposem de la informació de la dispersió (variància i desviació típica) de dues variables que es calculen a partir mitjanes diferents, no té sentit comparar-les directament amb aquests estadístics. L'índex adient és el **coeficient de variació**, que és una mesura de dispersió relativa que es defineix com el quocient entre la desviació típica i la mitjana, multiplicat per 100:

$$CV = \frac{s}{\bar{x}} \times 100$$

Considerem la comparació de les variables **educ** i **edad**. L'SPSS no ens proporciona directament el càlcul del coeficient de variació, però podem demanar els estadístics necessaris d'ambdues variables:

Estadísticos

		educ Años de escolarización	edad Edad del encuestado
N	Válidos	1496	1495
	Perdidos	4	5
Media		13,04	46,23
Desv. típ.		3,074	17,418
Varianza		9,450	303,386

Els coeficients de variació seran:

$$CV(educ) = \frac{s}{\bar{x}} \times 100 = \frac{3,074}{13,04} \times 100 = 23,54\%$$

$$CV(edad) = \frac{s}{\bar{x}} \times 100 = \frac{17,418}{46,23} \times 100 = 37,68\%$$

Amb aquests resultats podem afirmar que la variable *edad* té una més gran dispersió relativa.

► Exercici 1

Podeu repetir les anàlisis anteriors amb altres variables quantitatives de la matriu: **hijos**, **edadboda**, **indsocec**, **horastv**.

2. Anàlisi exploratòria d'una variable quantitativa

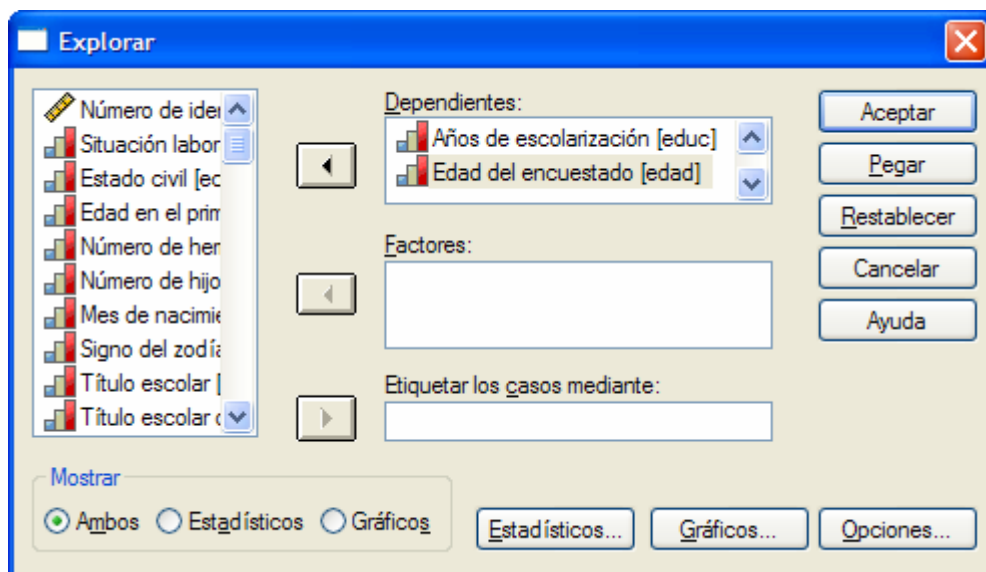
El procediment “**Explorar**” de l'SPSS ens permet realitzar una anàlisi exploratòria destinada a inspeccionar les dades, identificar valors atípics, obtenir descripcions, comprovar supòsits de les variables i caracteritzar diferències entre subpoblacions (grups de casos).

En aquesta pràctica utilitzarem el procediment amb una finalitat de descripció de les característiques de la distribució d'una variable quantitativa, i per a obtenir dos tipus de representacions gràfiques: el **diagrama de caixa** i el **gràfic de tija i fulla**.

Per a explorar les dades, a través del menú farem:

Analizar / Estadísticos descriptivos / Explorar...

Ens apareix el quadre de diàleg del procediment on col·locarem, al quadre de “Dependientes”, les dues variables que venim analitzant, *educ* i *edad*.



A continuació cliquem sobre “Aceptar” i observem els resultats.

Descriptivos

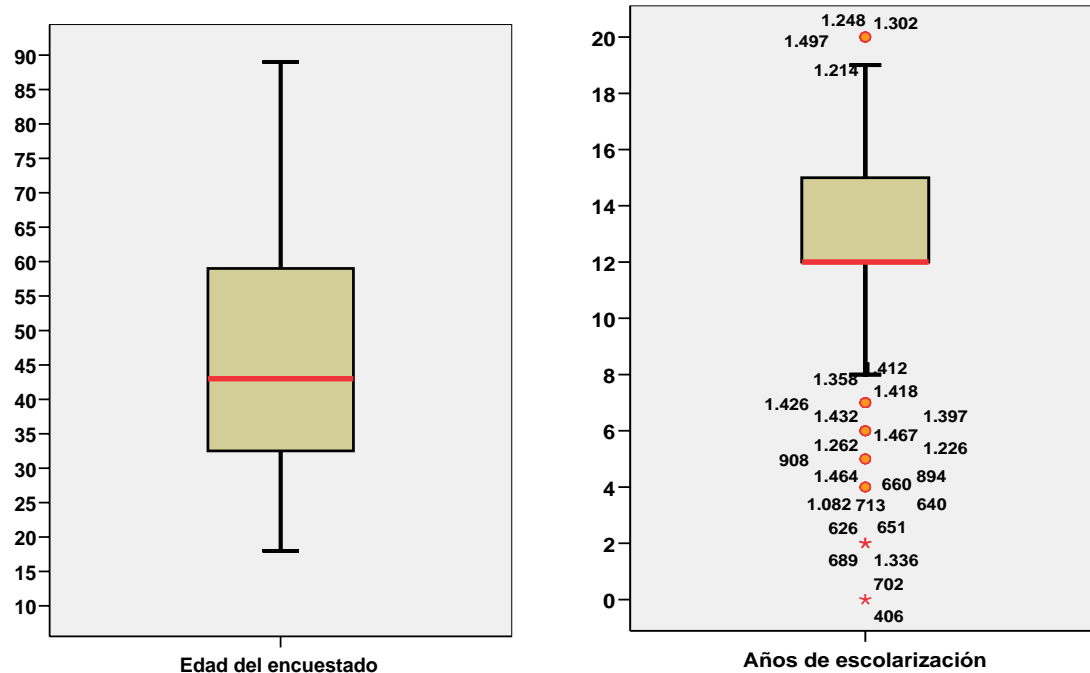
			Estadístico	Error típ.
educ Años de escolarización	Media		13,03	,079
	Intervalo de confianza para la media al 95%	Límite inferior	12,87	
		Límite superior	13,19	
	Media recortada al 5%		13,10	
	Mediana		12,00	
	Varianza		9,420	
	Desv. típ.		3,069	
	Mínimo		0	
	Máximo		20	
	Rango		20	
	Amplitud intercuartil		3	
	Asimetría		-,315	,063
	Curtosis		,718	,127
edad Edad del encuestado	Media		46,25	,451
	Intervalo de confianza para la media al 95%	Límite inferior	45,36	
		Límite superior	47,13	
	Media recortada al 5%		45,67	
	Mediana		43,00	
	Varianza		303,822	
	Desv. típ.		17,430	
	Mínimo		18	
	Máximo		89	
	Rango		71	
	Amplitud intercuartil		27	
	Asimetría		,498	,063
	Curtosis		-,704	,127

- Obtenim en primer lloc una taula de **descriptius** on podem comprovar que obtenim la major part dels estadístics que hem vist amb el procediment “Frecuencias”. Però tenim algun de nou:
- La **mitjana retallada al 5%** que és la mitjana aritmètica calculada després d’haver eliminat el 5% dels casos més grans i el 5% dels menors (els casos extrems), el que proporciona una millor estimació de la tendència central. Podem comprovar com la mitjana de la variable *educ* (13,03) canvia a 13,10, un valor superior, quan es retalla. En el cas de la variable *edad* baixa de 46,25 a 45,67.
- L’amplitud o el **rang interquartil (RI)** que és una mesura de la dispersió de les dades. És la distància entre el tercer quartil (el percentil 75) y primer quartil (el percentil 25). S’obtenen els resultats següents:

$$RI(educ) = P_{75} - P_{25} = 15 - 12 = 3$$

$$RI(edad) = P_{75} - P_{25} = 59 - 32 = 27$$

- El **diagrama de caixa** (*box plot*) és un altre resultat d'interès d'aquest procediment que ens proporciona els trets característics de la distribució de les dades, de posició, dispersió i simetria, en aquest cas referits a les variables *educ* i *edad*.



- Al primer diagrama podem observar la manca de valors extrems o atípics (*outliers*) que sí hi apareixen en el segon. Els valors extrems que s'identifiquen amb el símbol ● corresponen a aquells casos que es situen a una distància dels límits de la caixa superior a 1,5 vegades la desviació interquartil (diferència entre el tercer i el primer quartil, la distància que determina la caixa). Els *outliers* severes es reconeixen per l'asterisc (*) i corresponen als casos situats a una distància per sobre de 3 vegades la desviació interquartil. En tots aquests casos a més s'inclou el número d'identificació del cas.
- La mediana de l'edat és el valor 43, i s'identifica per la línia gruixuda de la caixa. En el cas dels anys d'escolarització, la mediana 12 es situa en l'extrem inferior de la caixa, que correspon també al primer quartil; això ens mostra la concentració d'efectius en aquest valor.
- Es pot comprovar per tant la major simetria de la variable edat i el comportament més atípic d'alguns valors baixos i alts de la variable d'escolarització.
- Un diagrama de caixa es pot obtenir directament a través del menú de gràfics:

Gráficos / Diagramas de caja...

- El procediment d'explorar també ens proporciona el **gràfic de tija i fulla**. Es tracta d'una forma particular d'histograma tombat que reitera la lectura del comportament de les distribucions.

Edad del encuestado Stem-and-Leaf Plot

Frequency	Stem &	Leaf
22,00	1 .	8999
106,00	2 .	000011112223333344444
150,00	2 .	55555566666777788888899999
152,00	3 .	000000011111222223333344444
199,00	3 .	5555555666666667777778888889999999
168,00	4 .	000000011111122222333333444444
138,00	4 .	55555566666677777888899999
104,00	5 .	0000111122223334444
91,00	5 .	555566777788888999
76,00	6 .	000112223333444
82,00	6 .	5556666778889999
81,00	7 .	00111222233334444
66,00	7 .	5556667778889
31,00	8 .	0112234
25,00	8 .	56789

```
Stem width: 10
Each leaf: 5 case(s)
```

Años de escolarización Stem-and-Leaf Plot

[illegible]

```
Stem width: 1
Each leaf: 9 case(s)
```

- El procediment d'explorar també proporciona els histogrames que aquí no hem adjuntat.

► Exercici 2

Podeu repetir l'anàlisi exploratòria amb altres variables quantitatives de la matriu: **hijos**, **edadboda**, **indsocec**, **horastv**.