# Pràctica 7. Relació de dues variables categòriques: Taules de contingència

En aquesta pràctica s'estudien les taules de contingència, que són l'eina bàsica per estudiar la relació o la independència entre variables categòriques (variables que classifiquen individus en grups).

En particular es veu com obtenir els percentatges (globals, per files i per columnes) i les gràfiques de barres (per files o per columnes i com agrupar-les), i s'interpreten. Finalment, veurem com introduir taules de contingència mitjançant la ponderació.

▶ Recordeu activar les opcions d'edició (edición $-->$ opciones):

- "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 *enquestaser.sav*. Són dades d'una enquesta de serveis.

## 1. Variables categòriques: taula de contingència de les freqüències observades i els % del total

Suposem dues variables categòriques, per exemple *edat* (grups d'edat) i *pc* (Ordinador a casa). Vegeu les etiquetes dels valors de les variables.

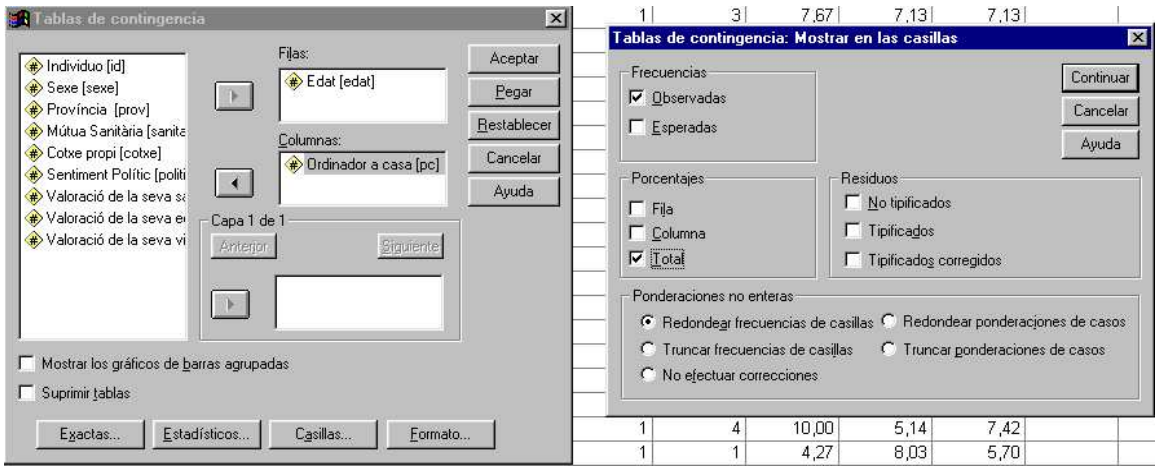
Una taula de contingència és una taula de doble entrada (*crosstab*) per *fer el recompte conjunt* de les freqüències de dues variables categòriques. Els valors d'una de les variables apareixen a les files, i els de l'altra, a les columnes; a les caselles hi ha les freqüències de la distribució conjunta.

El procediment és

**Analizar $-->$ Estadísticos descriptivos $-->$ Tablas de contingencia**

A la finestra de diàleg, seleccionem la variable *edat*, per a les files i *ordinador a casa*, per a les columnes.

A Caselles, activem **Frecuencias / Observadas** i **Porcentajes / Total**



Es produeix el resultat següent:

**Tabla de contingencia Edat * Ordinador a casa**

| | | | Ordinador a casa | | |
|---|---|---|---|---|---|
| | | | SÍ | NO | Total |
| Edat | 16-24 | Recuento | 46 | 21 | 67 |
| | | % del total | 3,8% | 1,7% | 5,5% |
| | 25-34 | Recuento | 159 | 117 | 276 |
| | | % del total | 13,0% | 9,6% | 22,6% |
| | 35-49 | Recuento | 438 | 135 | 573 |
| | | % del total | 35,8% | 11,0% | 46,9% |
| | 50-64 | Recuento | 208 | 92 | 300 |
| | | % del total | 17,0% | 7,5% | 24,5% |
| | 65 i + | Recuento | 0 | 6 | 6 |
| | | % del total | ,0% | ,5% | ,5% |
| Total | | Recuento | 851 | 371 | 1222 |
| | | % del total | 69,6% | 30,4% | 100,0% |

- A cadascuna de les caselles, hi tenim el nombre de casos (recompte o freqüència absoluta) i el percentatge del total d'enquestats (1222) que corresponen a cada combinació de l'encreuament de les dues variables *edat* x *ordinador a casa*. Per exemple, **159** enquestats (**un 13.0%** del total) són del grup d'edat **25-34 anys** i **tenen** ordinador a casa. La distribució del conjunt de percentatges del total s'anomena **distribució conjunta de les variables** *edat* x *ordinador a casa*.
- Als totals de fila, hi tenim la distribució dels grups d'*edat*: per exemple, del grup d'edat **50-64 anys** hi ha **300** enquestats (**un 24.5%** del total). La distribució dels grups d'edat amb els respectius totals s'anomena **distribució marginal de la variable fila** (*edat*).
- Als totals de columna, hi tenim la distribució del fet de tenir o no *ordinador a casa*: per exemple, **371** enquestats (**un 30.4%** del total) **no tenen** ordinador a casa. La distribució dels grups de *sentiment polític* amb els respectius totals s'anomena **distribució marginal de la variable columna** (*ordinador a casa*).

## 2. Percentatges condicionats per files (i/o per columnes)

Repetim un cop més el procediment i activem altres caselles (de l'opció casillas):

**Frecuencias / Observadas** i **Porcentajes / Filas**

**Tabla de contingencia Edat * Ordinador a casa**

| | | | Ordinador a casa | | |
|---|---|---|---|---|---|
| | | | SÍ | NO | Total |
| Edat | 16-24 | Recuento | 46 | 21 | 67 |
| | | % de Edat | 68,7% | 31,3% | 100,0% |
| | 25-34 | Recuento | 159 | 117 | 276 |
| | | % de Edat | 57,6% | 42,4% | 100,0% |
| | 35-49 | Recuento | 438 | 135 | 573 |
| | | % de Edat | 76,4% | 23,6% | 100,0% |
| | 50-64 | Recuento | 208 | 92 | 300 |
| | | % de Edat | 69,3% | 30,7% | 100,0% |
| | 65 i + | Recuento | 0 | 6 | 6 |
| | | % de Edat | ,0% | 100,0% | 100,0% |
| Total | | Recuento | 851 | 371 | 1222 |
| | | % de Edat | 69,6% | 30,4% | 100,0% |

La taula ens dóna la **distribució de percentatges condicionats per cada categoria fila**; és a dir:

- Si ens situem a la primera fila, condicionem per (equivalentment, ens restringim a) persones d'entre **16 i 24 anys**, de les quals un **68.7% tenen** ordinador a casa, per exemple.
- Anàlogament, d'entre les persones d'entre **25 i 34 anys**, un **57.6% tenen** ordinador a casa i un **42.4% no en tenen**.
- També veiem que, d'entre els de **65 i + anys**, ningú **no té** ordinador a casa.

El condicionament per les categories de la variable *edat (grups d'edat)* són els rellevants, atès que expliquen les **diferències del fet tenir o no ordinador a casa en funció de l'edat**.

També podem fer, si ho creiem interessant, el condicionament per les categories de la variable columna *ordinador a casa*; activariem dins de casillas:

**Frecuencias / Observadas** i
**Porcentajes / Columna**

i es produiria el resultat següent:

**Tabla de contingencia Edat * Ordinador a casa**

| | | | Ordinador a casa | | |
|---|---|---|---|---|---|
| | | | SÍ | NO | Total |
| Edat | 16-24 | Recuento | 46 | 21 | 67 |
| | | % de Ordinador a casa | 5,4% | 5,7% | 5,5% |
| | 25-34 | Recuento | 159 | 117 | 276 |
| | | % de Ordinador a casa | 18,7% | 31,5% | 22,6% |
| | 35-49 | Recuento | 438 | 135 | 573 |
| | | % de Ordinador a casa | 51,5% | 36,4% | 46,9% |
| | 50-64 | Recuento | 208 | 92 | 300 |
| | | % de Ordinador a casa | 24,4% | 24,8% | 24,5% |
| | 65 i + | Recuento | 0 | 6 | 6 |
| | | % de Ordinador a casa | ,0% | 1,6% | ,5% |
| Total | | Recuento | 851 | 371 | 1222 |
| | | % de Ordinador a casa | 100,0% | 100,0% | 100,0% |

**Observació.** En aquest context, però, el condicionament d'interès és per files. Es pretén analitzar l'edat com a factor explicatiu de tenir o no ordinador a casa, i no a l'inrevés.

Fixeu-vos que els percentatges per fila sumen 100% a cada fila i els percentatges per columna sumen 100% a cada columna.

## 3. Gràfiques. Interpretació de la relació

Les gràfiques són l'alternativa visual de les taules i permeten analitzar els resultats de manera més ràpida i clara. Concretament, ens preguntem

**Tenir o no ordinador a casa depèn de l'edat?**

Equivalentment,

**Els grups d'edat: presenten diferències en quant a tenir o no ordinador a casa?**

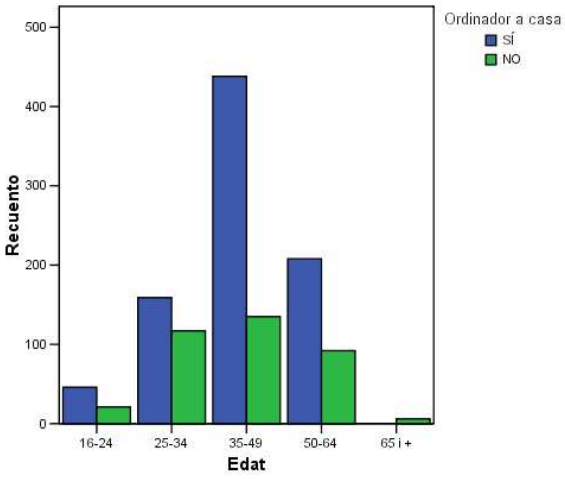
Veurem gràfiques de barres agrupades i apilades de la distribució conjunta i de la distribució condicionada per files.

Per obtenir <u>només</u> els diagrames de barres, dins del procediment de taules de contingència: activem

**Mostrar gráficos de barras agrupadas** i
**Suprimir tablas**

tal i com veieu a la finestra de diàleg:

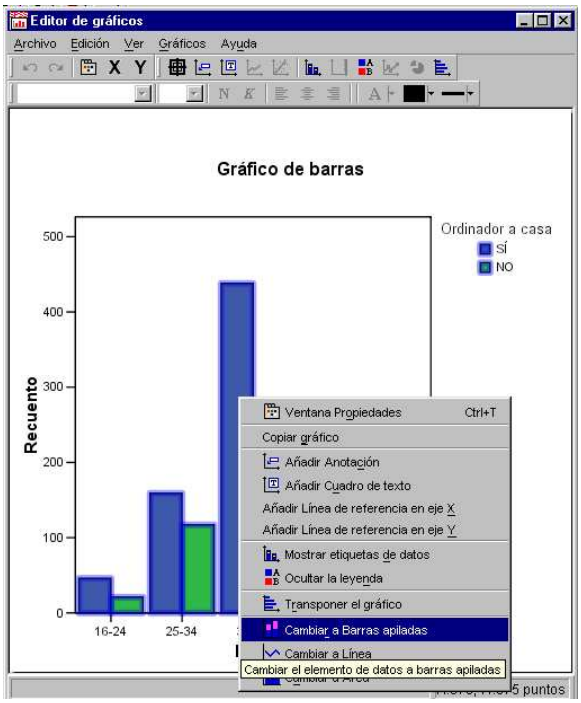

La gràfica resultant és:

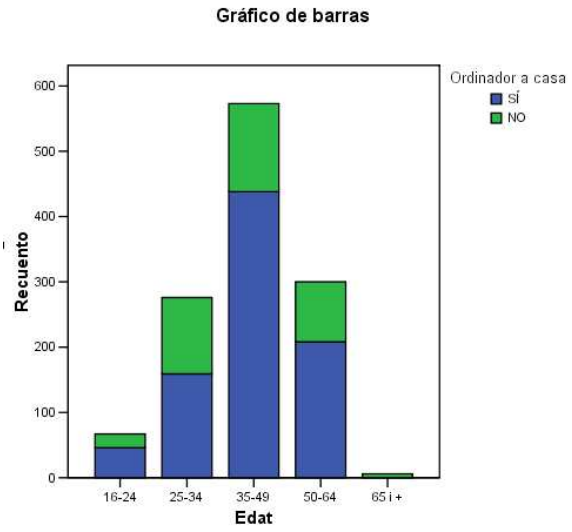



Les barres apilades corresponen a la taula de contingència de la distribució conjunta. Tenen altures diferents, per a cada grup d'edat, atès que la grandària de la mostra és diferent en cada grup, la qual cosa en dificulta la comparació i l'anàlisi de la situació.

Seguidament, editem la gràfica (recordeu: doble click) per tal de convertir-la a barres apilades. Situant el cursor sobre les barres i prement el botó dret, s'obre una finestra i escollim l'opció **Cambiar a Barras apiladas**:

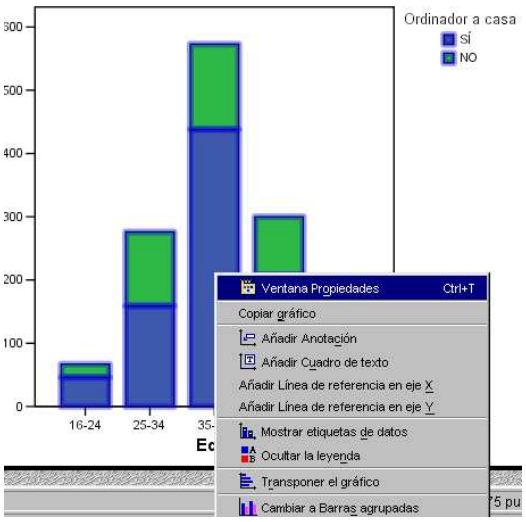
La gràfica canviada és:



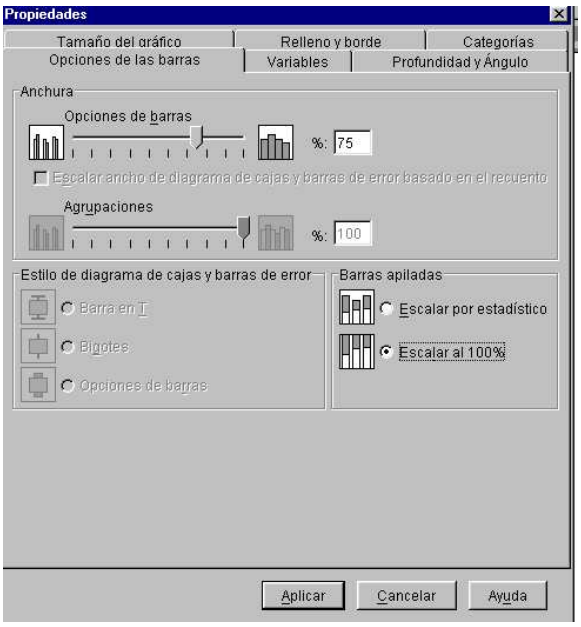



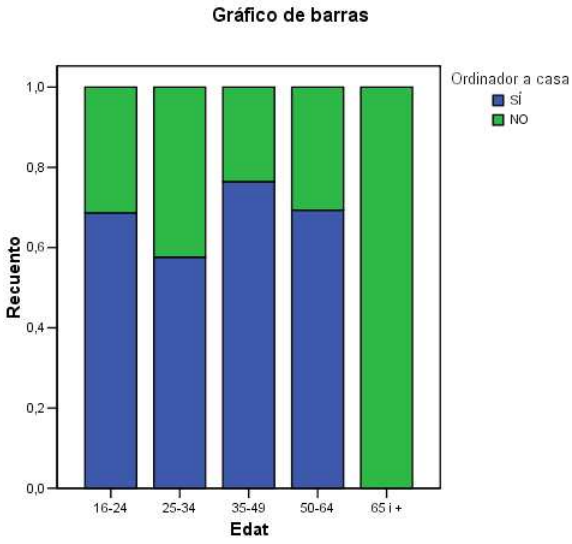

Les barres segueixen tenint diferents alçades en funció de la mida de cada grup. Per acabar, i per facilitar la interpretació, si volem que totes les barres tinguin igual

altura, situem de nou el cursor sobre les barres, premem el botó dret i seleccionem la **Ventana Propiedades**:




I s'obre una nova finestra, en la qual seleccionem la pestanya "Opciones de las barras": **Opciones de las barras - Escalar al 100%**:




Hem obtingut la gràfica definitiva (excepte mida, colors, text, etc.), amb els grups d'edat reduïts a la mateixa mida (mida 1 o 100%).

**Interpretació**: Observem que les barres canvien en funció del grup d'edat. Per un cantó el grup de 65 anys o més té un comportament completament diferent dels altres i no hi ha barra del *si*. També observem que el grup de 25-34 anys té un valor inferior als altres menors de 65. Per tant, *tenir ordinador a casa* presenta una dependència aparent de la edat.

**Observació.** Cal fer atenció en el fet que el grup de més de 65 només té 6 individus.

Si totes les barres tinguessin (aproximadament) les mateixes piles, no hi hauria relació entre les dues variables.

▶ Exercici:

Repetiu tots els procediments anteriors per analitzar la relació entre les variables *polític (sentiment polític)* i *sanitat (mútua sanitària)*. En particular:

- Quin condicionament entre les variables té sentit? És a dir, té sentit demanar si tenir o no mútua sanitària depèn del sentiment polític? I té sentit demanar si el sentiment polític pot dependre de tenir o no mútua sanitària?
- Construïu totes les taules (freqüències i percentatges totals, per files i per columnes).
- Quin percentage dels que tenen mútua se sent d'esquerres?
- Quin percentatge dels que se senten de centre no tenen mútua sanitària?
- Construïu les gràfiques de barres agrupades, apilades i apilades amb percentatges per files.
- Creieu que hi ha relació entre les dues variables?

## 4. Introducció d'una taula de contingència donada a una base de dades

De vegades la taula de contingència entre dues variables apareix publicada i la volem analitzar, per obtenir els percentages i les gràfiques. Si ho volem fer amb l'SPSS haurem d'introduir les dades de manera que el paquet sigui capaç de processar la informació.

Per exemple, disposem de la següent taula que mostra la distribució dels desocupats a Catalunya conjuntament en les variables *edat (grup d'edat)* i *sexe*:

**Població ocupada assalariada de Catalunya**

Tercer trimestre de 2005 en milers d'habitats.

| Grup d'edat | **Homes** | **Dones** | **Total** |
|---|---|---|---|
| 16–24 anys | 184,7 | 144,8 | 329,5 |
| 25–34 anys | 484,4 | 408,0 | 892,0 |
| 35–44 anys | 413,7 | 322,2 | 736,0 |
| 45–54 anys | 302,1 | 227,3 | 529,4 |
| 55 anys i més | 171,6 | 93,2 | 264,9 |
| **Total** | **1.556,6** | **1.195,6** | 2.752,1 |

Unitat:. Milers de persones

Font: IDESCAT, a partir de les dades de l'enquesta de població activa de l'INE

L'introduïm a l'SPSS, en una nova base de dades que anomenarem **epa.sav**, de la següent forma:

| | grupedat | sexe | frequencia |
|---|---|---|---|
| 1 | 1 | 1 | 184,7 |
| 2 | 1 | 2 | 144,8 |
| 3 | 2 | 1 | 484,4 |
| 4 | 2 | 2 | 408,0 |
| 5 | 3 | 1 | 413,7 |
| 6 | 3 | 2 | 322,2 |
| 7 | 4 | 1 | 302,1 |
| 8 | 4 | 2 | 227,3 |
| 9 | 5 | 1 | 171,6 |
| 10 | 5 | 2 | 93,2 |
| 11 | | | |

Hem creat tantes files, **10**, com caselles té la distribució conjunta; és a dir, 5 categories d'edat $\times$ 2 categories de sexe = 10 categories de la distribució conjunta.

Recordem que hem d'assignar etiquetes als valors numèrics amb els quals hem codificat cada variable.

Les freqüències tenen decimals, atès que la unitat són milers de desocupats; cal especificar-ho bé en l'etiqueta de la variable freqüència.

| | |
|---|---|
| **Important:** | En primer lloc, cal **ponderar** els casos per freqüència.<br><br>Quan estem creant la taula de contigència, dins de la finestra que s'obre quan premem **Casillas** cal triar l'opció **No efectuar correciones** perque treballem amb ponderacions que tenen decimals. (Vegeu la figura de la pàgina 1.) |

▶ Exercici:

Feu una anàlisi completa de les dades d'aquesta taula.

1) Construïu les taula de contingència de freqüències observades i % del total, per files i per columnes.
2) Entre les dones assalariades i ocupades, quin percentatge té més de 55 anys?
3) Entre els homes assalariats i ocupats, quin percentatge té menys de 35 anys?
4) Quin percentatge de dones hi ha entre els assalariats i ocupats de menys de 25 anys. I entre els que tenen més de 55 anys?
5) Feu les grafiques barres agrupades, apilades i apilades amb percentatges.
6) Interpreteu la possible relació entre les dues variables. Creieu que en la distribució de l'ocupació assalariada per sexes hi ha diferències notòries en funció del grup d'edat?