Relació ente dues variables categòriques

Variables categòriques són aquelles que, o bé prenen un nombre finit de valors o bé són variables numèriques que hem agrupat en un nombre finit de classes o intervals. És a dir, són variables de tipus qualitatiu (nominal o ordinal), o bé variables numèriques que hem convertit a categòriques agrupant els seus valors en classes o intervals.

Si tenim dues variables categòriques (per exemple, color dels ulls i color dels cabells) avaluades sobre una mateixa població o mostra d'individus, fins ara hem estudiat com analitzar cada variable per separat.

El nostre objectiu és analitzar si hi ha relació o no entre dues variables categòriques avaluades sobre una mateixa mostra.

Distribució conjunta

Suposem que tenim dues variables categòriques: color dels ulls i color dels cabells avaluades en una mostra de 6800 persones. Obtenim les següents freqüències:

ULLS	n_i	f_i	p_i
Blau	2811	0.413	41.3%
Verd	3132	0.461	46.1%
Castany	857	0.126	12.6%
	6800	1	100%

CABELLS	n_i	f_i	p_i
Ros	2829	0.416	41.6%
Castany	2632	0.387	38.7%
Negre	1223	0.18	18%
Pelroig	116	0.017	1.7%
	6800	1	100%

Si volem estudiar la relació entre les dues variables, ens hem de fer preguntes del següent tipus:

1. Entre les persones que tenen color de cabell negre, quantes tenen els ulls blaus? I entre les de color de cabell ros? Representen la mateixa proporció en ambdos casos?

2. O bé, entre les persones de color d'ulls castany quantes són de color de cabell negre? és la mateixa proporció que entre les de colot d'ull castany?

Taula de contingència: és una taula que ens permet resumir les freqüències observades en dues variables categòriques avaluades en una mateixa mostra. Els valors d'una variables es col.loquen en files, els valors de l'altra variable en columnes i a cada casella posem la freqüencia o nombre d'observacions corresponents als valors fila i columna.

Ulls/Cab	Ros	Castany	Negre	Pelroig	Total
Blau	1768	807	189	47	2811
Verd	946	1387	746	53	3132
Castany	115	438	288	16	857
Total	2829	2632	1223	116	6800

A la última columna tenim resumida la variable color d'ulls i a la última fila tenim resumida la variable color de cabell.

Hi ha 189 persones amb cabell negre i ulls blaus. I 1768 amb cabell ros i ulls blaus.

Hi ha 288 persones d'ulls castanys i cabell negre.

En la mateixa taula podem estudiar els percentatges respecte el total de la mostra

$$\frac{freq.casella}{totalmostra} \times 100\%$$

Ulls/Cab	Ros	Castany	Negre	Pelroig	Total
Blau	26%	11.9%	2.8%	0.7%	41.3%
Verd	13.9%	20.4%	11%	0.8%	46.1%
Castany	1.7%	6.4%	4.2%	0.2%	12.6%
Total	41%	38.7%	18%	1.7%	100%

Un 2.8% de la mostra té cabell negre i ulls blaus.

Un 26% de la mostra té cabell ros i ulls blaus.

Un 4.2% de la mostra té cabell negre i ulls castanys.

Ara bé, si volem analitzar la influència d'una variable en l'altra el que hem de fer és estudiar la distribució de valors un cop fixada una variable. Per exemple, si només ens fixem en les persones de cabell negre, quins són els percentatges pels diferents color d'ulls? Són els mateixos si només ens fixem en les persones de cabell ros? O pelroig?

Aleshores, fem una taula amb els percentatges calculats per columna $\frac{freq.casella}{totalcolumna} \times 100\%$

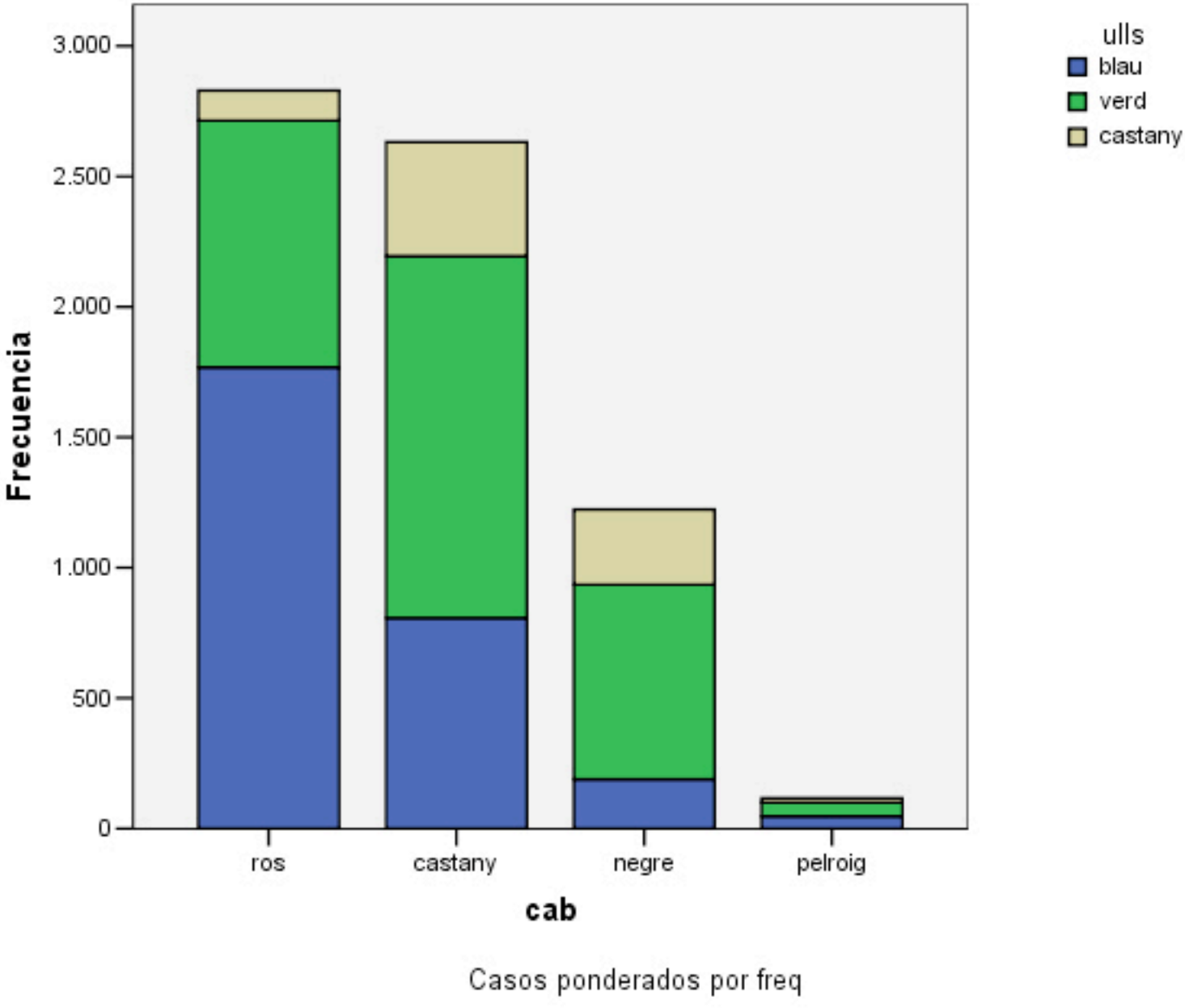
Ulls/Cab	Ros	Castany	Negre	Pelroig	Total
Blau	62.5%	30.7%	15.5%	40.5%	41.3%
Verd	33.4%	52.7%	61%	45.7%	46.1%
Castany	4.1%	16.6%	23.5%	13.8%	12.6%
Total	100%	100%	100%	100%	100%

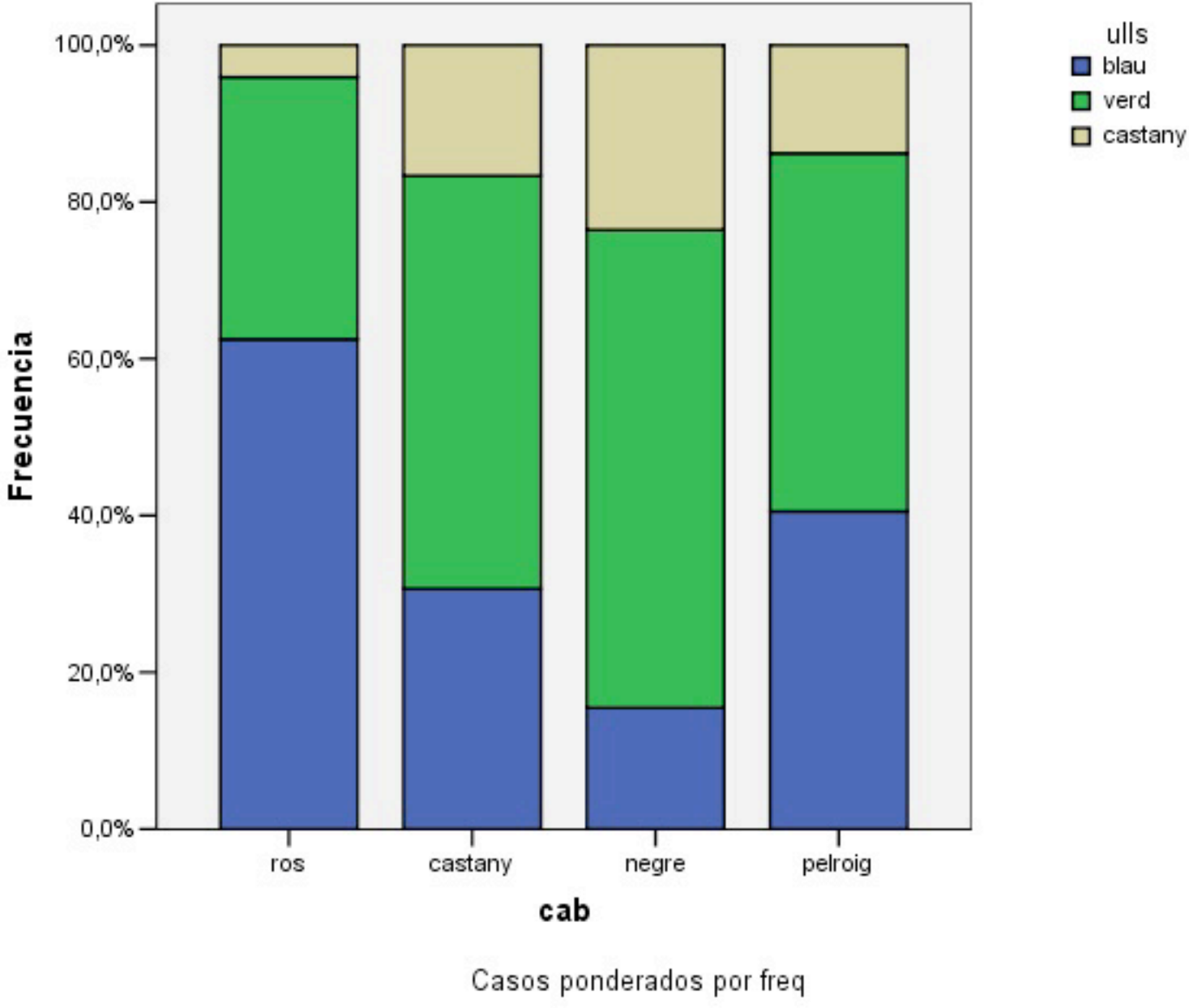
La última columna continua tenint el resum de la variables color d'ulls.

Entre les persones de color de cabell negre, un 15.5% té ulls blaus, un 61% verds i un 23.5% castanys.

Entre les persones de color de cabell ros, un 62.5% té ulls blaus, un 33.4% verds i un 4.1% castanys.

Amb aquestes dades, podem sospitar que les dues variables estan relacionades? És a dir, podem sospitar que el color de cabell influeix en el color d'ulls de les persones?


ulls
blau
verd
castany
Frecuencia
3.000
2.500
2.000
1.500
1.000
500
0
ros
castany
negre
pelroig
cab
Casos ponderados por freq


ulls
blau
verd
castany
100,0%
80,0%
60,0%
40,0%
20,0%
0,0%
Frecuencia
ros
castany
negre
pelroig
cab
Casos ponderados por freq

També podem fer una taula amb els percentatges calculats per files $\frac{freq.casella}{totalfila} \times 100\%$

Ulls/Cab	Ros	Castany	Negre	Pelroig	Total
Blau	62.9%	28.7%	6.7%	1.7%	100%
Verd	30.2%	44.3%	23.8%	1.7%	100%
Castany	13.4%	51.1%	33.6%	1.9%	100%
Total	41.6%	38.7%	18%	1.7%	100%

La última fila continua tenint el resum de la variables color de cabells.

Entre les persones de color d'ulls blau, un 62.9 = 1768/2811·100% tenen cabells rossos, un 28.7% castanys, un 6.7% negres i un 1.7% són pelroges.

Entre les persones de color de cabell ros, un 62.5% té ulls blaus, un 33.4% verds i un 4.1% castanys.

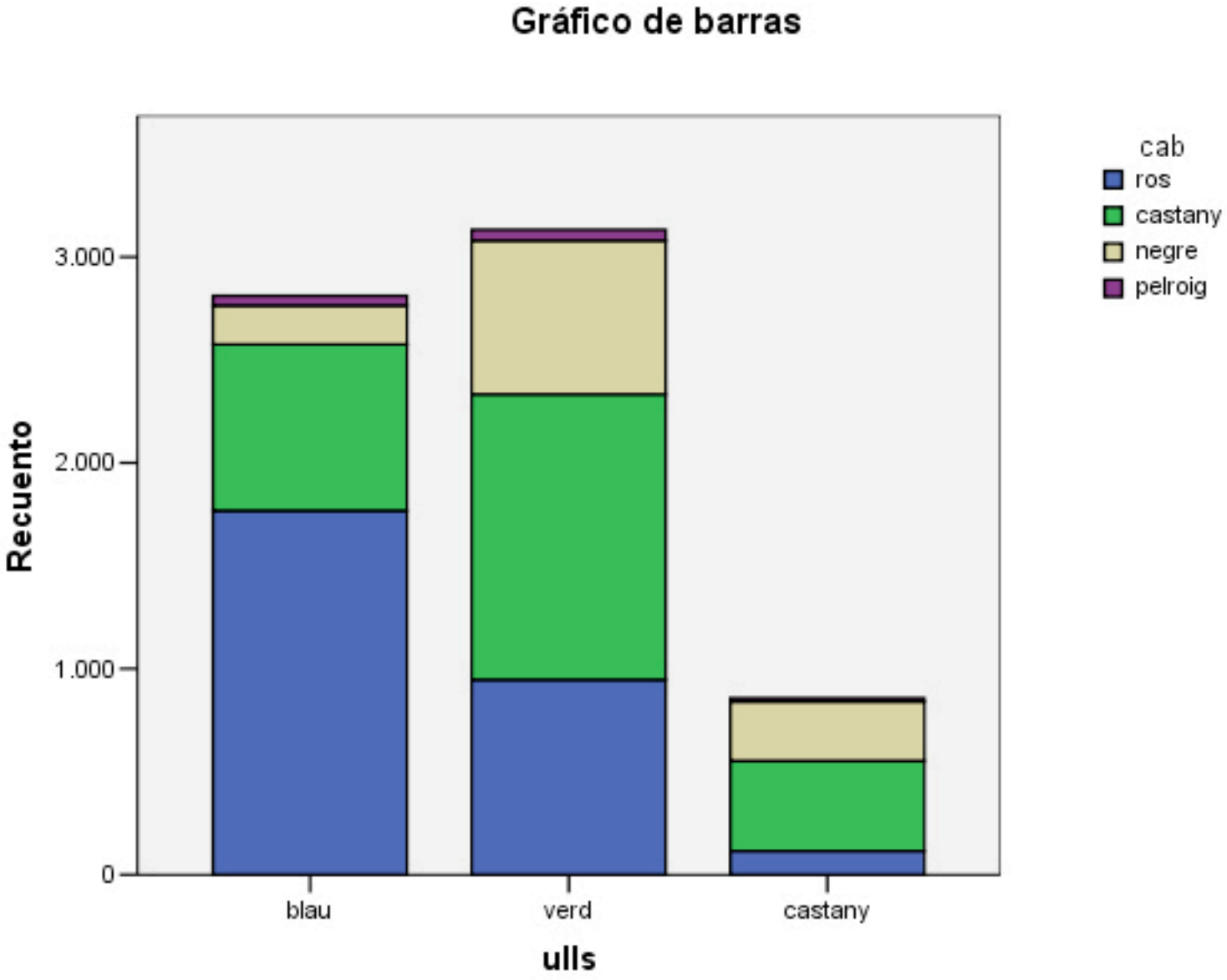


Gráfico de barras
cab
ros
castany
negre
pelroig
Recuento
3.000
2.000
1.000
0
blau
verd
castany
ulls

Amb aquestes dades, podem sospitar que les dues variables estan relacionades? És a dir, podem sospitar que el color de cabell influeix en el color d'ulls de les persones?

Si les variables fossin independents entre elles i no s'influissin aleshores els diferents colors d'ulls es repartirien en les mateixes proporcions independentment del color del cabell, és a dir, si hi ha un 41.6% de persones d'ulls de color blau, també esperaríem que si només ens fixem en les de cabell negre també surti un mateix percentatge, és a dir, que hi hagi $1223 \cdot 41.6/100 \simeq 506$ persones de cabell negre i ulls blaus. Però n'hi ha 189.

Les freqüències esperades són les que surtirien si les variables fossin independents i totes les files (o columnes) es repartissin en les mateixes proporcions.

$$\frac{totalfila \cdot totalcolumna}{total}$$

U/C	Ros	Castany	Negre	Pelroig	Tota
B	1768(1169.5)	807(1088)	189(505.6)	4748(48)	281
V	946(11303)	1387(1212.3)	746(563.3)	53(53.4)	313
C	115(356.5)	438(331.7)	288(154.1)	16(14.6)	85
Tot	2829	2632	1223	116	680

Les freqüències esperades es troben entre parèntesis. Observeu que en algunes caselles hi ha molta diferència.