Sessió 8

Dades multivariants 2: relació entre dos factors

Veurem com s'explora numèricament i gràficament la relació entre dues variables categòriques (dos factors).

8.1 Taules de contingència: distribucions conjunta i condicionades per files i per columnes

Treballarem amb el fitxer `gss93_reducido.csv`. El carreguem i l'anomenem `gss`

Estudiarem la relació entre els factors estat civil (`ecivil`) i sexe (`sexo`).

- La funció `table` ens permet fer la **taula de freqüències absolutes o de percentatges de la distribució conjunta**:

```
> attach(gss)
> table(sexo,ecivil)
        ecivil
sexo     Casado Divorciado Separado Soltero Viudo
  Hombre    383         75        9     142    31
  Mujer     412        138       31     144   134

#la desem en la variable taula i comprovem que és de la classe table
> taula<-table(sexo,ecivil)
> class(taula)

# per tenir percentatges de la distribució conjunta
> round(100*taula/sum(taula),1)
        ecivil
sexo     Casado Divorciado Separado Soltero Viudo
  Hombre   25.6        5.0      0.6     9.5   2.1
  Mujer    27.5        9.2      2.1     9.6   8.9

# el mateix s'aconsegueix amb la funció prop.table
> round(prop.table(taula)*100,1)
```

- Podem afegir els **totals de files i de columnes** a la taula amb la funció `addmargins`

```
addmargins(taula)
        ecivil
sexo     Casado Divorciado Separado Soltero Viudo  Sum
  Hombre    383         75        9     142    31  640
  Mujer     412        138       31     144   134  859
  Sum       795        213       40     286   165 1499
```

- Les **distribucions marginals** associades a la taula de la distribució conjunta es poden obtenir amb `margin.table`

```
> margin.table(taula,1)  # distribució marginal de la variable
sexo
Hombre  Mujer
   640    859

> margin.table(taula,2)  # distribució marginal de la variable
                       # que hi ha per columnes
ecivil
    Casado Divorciado   Separado    Soltero      Viudo
       795        213         40        286        165
```

- La **distribució condicionada per files** s'obté amb `prop.table(taula,1)`

```
> cond.files<-prop.table(taula,1)
> round(cond.files*100,2)
        ecivil
sexo     Casado Divorciado Separado Soltero Viudo
  Hombre  59.84      11.72     1.41   22.19  4.84
  Mujer   47.96      16.07     3.61   16.76 15.60
```

En la taula anterior les files sumen 100. La primera fila és la distribució en percentatges de la variable `ecivil` condicionada a `sexo` igual a `Hombre`, i la segona fila és la distribució de la mateixa variable però condicionada a sexe femení.

- Podem comprovar que les files de la taula `cond.files` sumen 100, amb `addmargins(...,2)`:

```
> round(addmargins(cond.files*100,2),3)  #arrodonim a 3 decimals
        ecivil
sexo      Casado Divorciado Separado Soltero   Viudo     Sum
  Hombre  59.844     11.719    1.406  22.188   4.844 100.000
  Mujer   47.963     16.065    3.609  16.764  15.600 100.000
```

- De manera anàloga podem trobar la **distribució condicionada per columnes**, amb `prop.table(taula,2)` i afegir-hi les sumes de columnes

```
> cond.cols<-prop.table(taula,2)
> round(addmargins(cond.cols*100,1),3) # arrodonim a 3 decimals
        ecivil
sexo      Casado Divorciado Separado Soltero   Viudo
  Hombre  48.176     35.211   22.500  49.650  18.788
  Mujer   51.824     64.789   77.500  50.350  81.212
  Sum    100.000    100.000  100.000 100.000 100.000
```

Taules amb la funció apply directament

- **Taula de distribució conjunta, en percentatges**

```
taula<-table(sexo,ecivil) # taula distribució conjunta freqs absolutes
taula/sum(taula)*100      # distribució conjunta en percentatges
```

- **Afegir marges**

```
taula2<-cbind(taula,apply(taula,1,sum))
rbind(taula2,apply(taula2,2,sum)) # conjunta i sumes de files i cols
```

- **Distribucions marginals**

```
apply(taula,1,sum)  # marginals de files
apply(taula,2,sum)  # marginals de columnes
```

- **Distribucions condicionals**

```
# condicional per files
 t(apply(taula,1,function(x) x/sum(x)*100))
# condicional per columnes
 apply(taula,2,function(x) x/sum(x)*100)
```

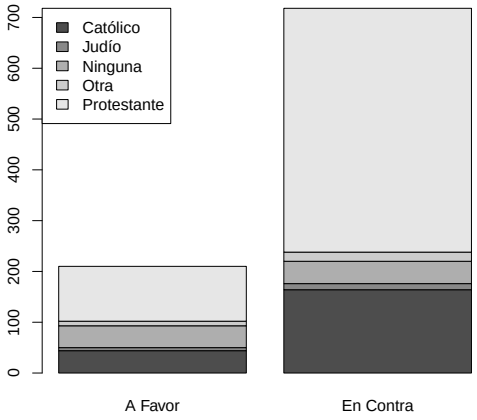
Pràctica:
• Amb les taules anteriors contesta les preguntes següents:
Quin percentatge de la mostra són dones divorciades?
Quin percentatge de vidus són homes?
Quin percentatge de dones són solteres?
• Fes les taules de la distribució conjunta amb freqüències absolutes i amb percentatges (dues taules) i les taules de les distribucions condicionals de les variables `museos` i `telenov`. Contesta les preguntes següents amb les taules que has creat:
- Quin percentatge de la mostra va a museus i veu telenoveles rarament?
- Quin percentatge dels que no van a museus veuen diariament telenoveles?
-Quin percentatge dels que mai veuen telenoveles van a museus?
- Et semblen associades (relacionades) les 2 variables? Raona la resposta.

8.2 Gràfiques bivariants de dades categòriques

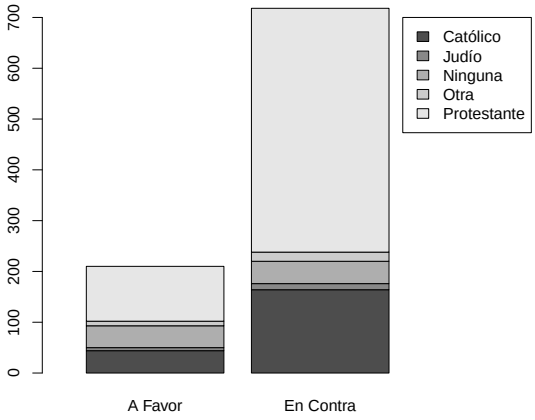
Treiem dues noves variables categòriques: `legdroga` (a favor o en contra de la legalització de la droga) i `relig` (religió)

- Anem a representar primer la distribució conjunta amb un diagrama de barres apilades.
  En primer lloc hem de calcular la taula, i després fem un `barplot`

```
> reldroga<-table(relig,legdroga)
> barplot(reldroga)
# perquè surti la llegenda hem de demanar-ho
> barplot(reldroga,col=gray.colors(5))    # gràfic 1
legend("topleft",legend=levels(relig),fill=gray.colors(5))

# una altra possibilitat és fer lloc fora del gràfic i
#               especificar posició de la llegenda
> barplot(reldroga,col=gray.colors(5),xlim=c(0,3))   # gràfic 2
> legend(2.5,700,legend=levels(relig),fill=gray.colors(5))
```



gràfic 1



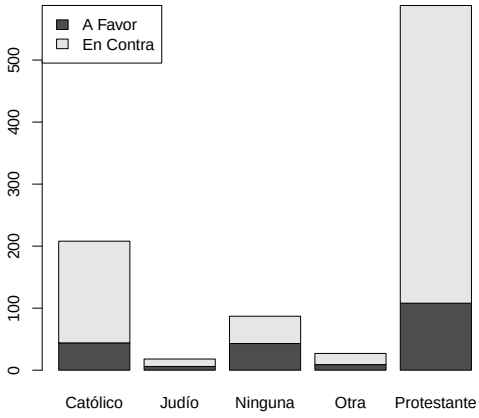
gràfic 2

- Si el que volem és **intercanviar els papers** de les dues variables transposem la taula (gràfic 3)

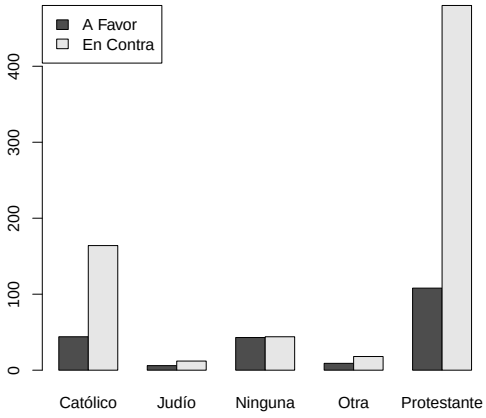
```
> barplot(t(reldroga),col=gray.colors(2))    # gràfic 3
legend("topleft",legend=levels(legdroga),fill=gray.colors(2))
```

- Per a fer gràfiques **no apilades** s'utlitza l'opció `beside=TRUE` (gràfic 4)

```
barplot(t(reldroga),col=gray.colors(2),beside = T)    # barres al costat
legend("topleft",legend=levels(legdroga),fill=gray.colors(2))
```



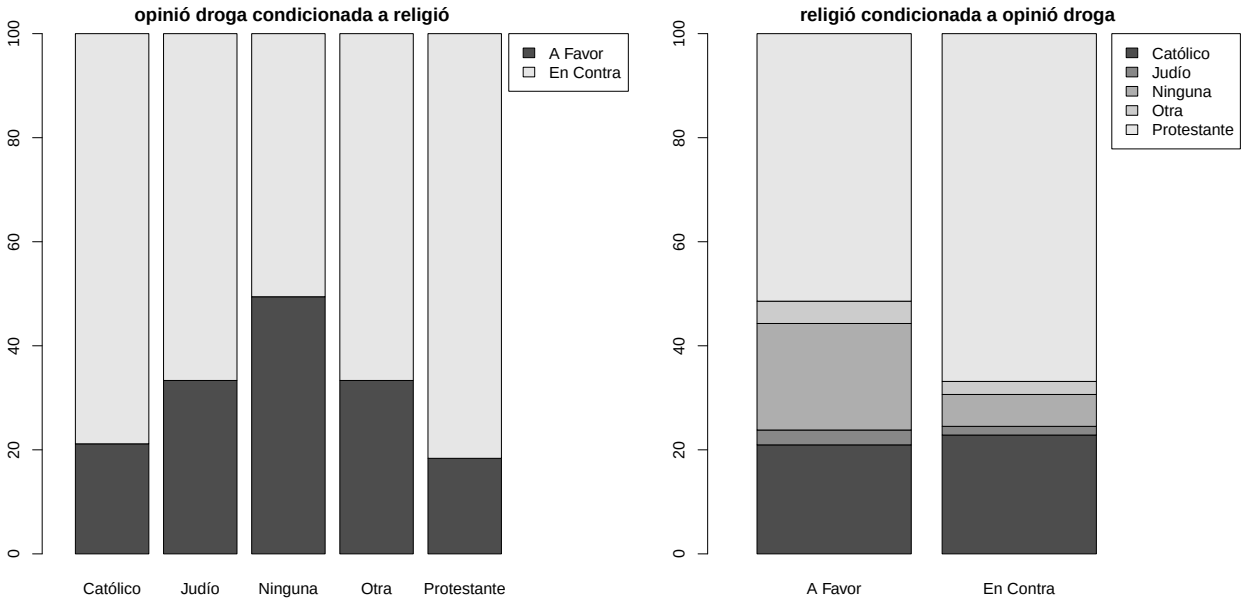
gràfic 3



gràfic 4

- Si el que volem és representar la **distribució condicional** amb barres apilades, primer hem de calcular la taula com hem fet abans, amb `prop.table`. Comencem calculant les dues distribucions condicionades

```
reldrog.condf<-prop.table(reldroga,1)*100  # condicionem per files
reldrog.condc<-prop.table(reldroga,2)*100  # condicionem per columnes

# si volem fer el gràfic de barres apilades condicionant per files,
# hem de transposar primer abans de fer el barplot
barplot(t(reldrog.condf),xlim=c(0,6.3),
                  main="opinió droga condicionada a religió")
legend(6.1,100,legend=levels(legdroga),fill=gray.colors(2))

barplot(reldrog.condc,xlim=c(0,3),col=gray.colors(5),
                  main="religió condicionada a opinió droga")
legend(2.5,100,legend=levels(relig),fill=gray.colors(5))
```

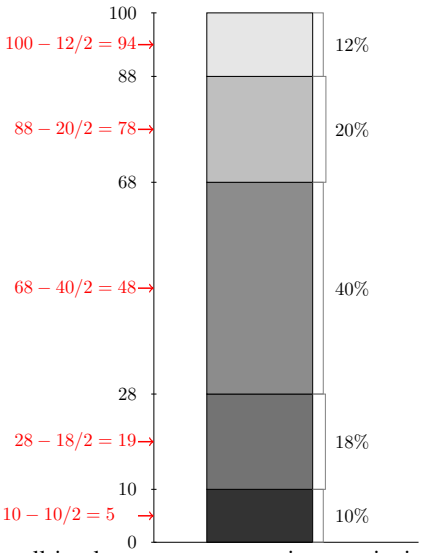


- Sovint convé afegir els percentatges de cada categoria dins de cada rectangle del gràfic de barres apilades. S'hi posa amb la instrucció `text`. Recordem que la funció `text` té com a arguments les coordenades del punt on col·locar el text i el text en sí. També podem utilitzar `text` per posar més d'una etiqueta de cop, donant com a arguments un vector de longitud n amb les coordenades x dels punts, un altre vector de longitud n amb les coordenades y dels punts i un vector de longitud n amb els textos que volem col·locar en els punts.

  A continuació veurem com calcular les altures on posar el text amb els percentatges dels diferents rectangles d'una barra apilada.

  Fixem-nos en el dibuix, que correspondria a una barra amb una distribució donada per la taula

x_i	p_i
x_1	10
x_2	18
x_3	40
x_4	20
x_5	12

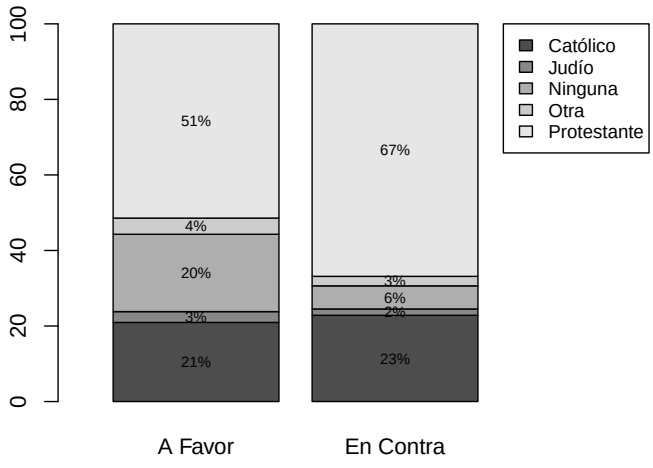


  Volem col·locar les etiquetes a les altures (les coordenades y) en vermell i volem que aquestes etiquetes siguin els percentatges de la taula. Per fer-ho en R tenim la funció `cumsum` que ens permet calcular els números on acaben els rectangles en que queda dividida la barra. Si volem el punt mitjà en vertical del rectangle, hem de restar la meitat de l'altura del rectangle

```
> (alturesrect<-c(10,18,40,20,12))
[1] 10 18 40 20 12
> (finalrect<-cumsum(alturesrect))
[1]  10  28  68  88 100
> (puntsmitjans<-finalrect-alturesrect/2)
[1]  5 19 48 78 94
```

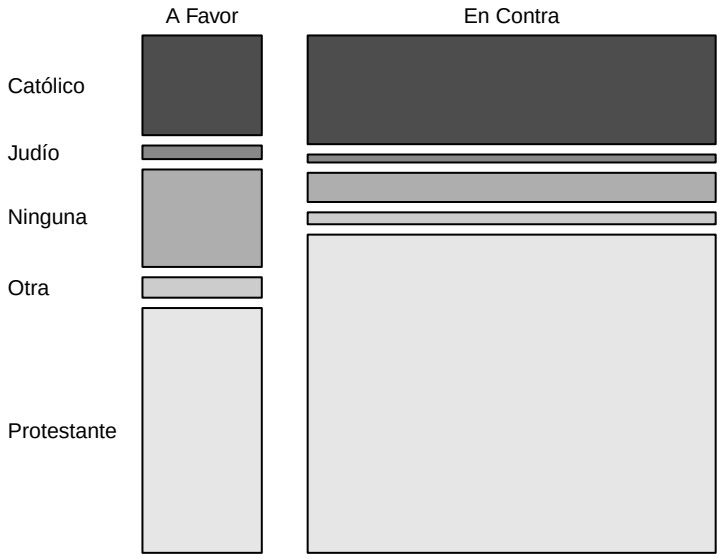
Anem a fer ara el mateix amb el gràfic de barres apilades de religió condicionada a opinió sobre la droga.

```
x<-barplot(reldrog.condc,xlim=c(0,3.3))
# l'anterior instrucció fa el gràfic
# i guarda a x les posicions de les barres en l'eix x
legend("topright",legend=levels(relig),fill=gray.colors(5),cex=0.8)

# preparem les etiquetes i les seves posicions:

# com que a cada barra hi posarem 5 etiquetes,
# repetim les components de x 5 cops
etiquetesx<-rep(x,each=5)

# com hem fet abans calculem el punt mitjà de cada rectangle amb apply
etiquetesy<-as.vector(apply(reldrog.condc,2,function(x) cumsum(x)-x/2))
# apply amb marge 2 serveix per processar les dues barres de cop.

# el vector etiquetes conté el text que posarem, és un vector de
# longitud 10, perquè la matriu reldrog.condc es converteix en vector
# al aplicar paste
etiquetes<-paste(round(reldrog.condc),"%",sep="")

# ja podem afegir el text,
# etiquetesx, etiquetesy, etiquetes tenen longitud 10
text(etiquetesx,etiquetesy,etiquetes,cex=0.7)
```



- El gràfic de mosaic és com una mena de gràfic de barres apilades però amb les barres més juntes i d'amplades proporcionals a la mida del grup que representa cada barra.

```
mosaicplot(t(reldroga),col=2:6,las=1,off=8,main="")
```



8.2.1 Colors: les "palettes"

Per fer gràfics de barres o d'altres, com el de mosaic en què són importants els colors, podem utilitzar les predefinides, com ara `gray.colors` o bé `rainbow`, especificant quants colors volem i quins de la paleta

```
barplot(reldrog.condc,xlim=c(0,3.3),col=rainbow(5))
barplot(reldrog.condc,xlim=c(0,3.3),col=rainbow(8)[2:6])
barplot(reldrog.condc,xlim=c(0,3.3),col=gray.colors(6)[2:6])
barplot(reldrog.condc,xlim=c(0,3.3),col=terrain.colors(5)
```

Però també hi ha d'altres paletes, Podeu consultar per exemple
https://developer.r-project.org/Blog/public/2019/04/01/hcl-based-color-palettes-in-grdevices/

```
# En aquest cas es canvia la paleta que hi ha per defecte
# que són els colors que surten quan posem col=1:5
# proveu què passa quan executem
mosaicplot(t(reldroga),col=1:5,las=1,off=8,main="")
palette(hcl.colors(5, "viridis"))
mosaicplot(t(reldroga),col=1:5,las=1,off=8,main="")
palette(hcl.colors(5,"Purples"))
mosaicplot(t(reldroga),col=1:5,las=1,off=8,main="")
palette(hcl.colors(5,"Temps"))
mosaicplot(t(reldroga),col=1:5,las=1,off=8,main="")

# per tornar a la paleta per defecte:
palette("default")
```

Pràctica: Analitza amb taules i gràfiques les relacions entre una parella de variables categòriques de l'arxiu `gss93_reducido`. No oblidis d'interpretar els resultats numèrics i gràfics.

8.3 Avaluació de la dependència-independència: freqüències esperades

Observant a les taules de la secció 1 les diferències entre les distribucions condicionades per files o bé entre les distribucions condicionades per files i la distribució marginal corresponent (o bé treballant per columnes) podem deduir si hi ha una forta dependència o no de les dues variables. Però per avaluar la independència de dues variables categòriques amb més precisió es comparen les **freqüències observades** n_{ij} que apareixen a la taula de la distribució conjunta amb les **freqüències esperades**, que es calcuen mitjançant la fórmula

$$e_{ij} = \frac{n_{i,\cdot} \times n_{\cdot,j}}{n}$$

- Podem calcular les **freqüències esperades** utilitzant el producte de matrius. Per exemple, anem a fer el càlcul per a les variables `sexo` i `ecivil`. Si escrivim la distribució marginal de la variable `sexo` com la matriu

$$A = \begin{pmatrix} 859 \\ 640 \end{pmatrix}$$

i la distribució marginal de la variable `ecivil` com la matriu

$$B = \begin{pmatrix} 213 & 795 & 286 & 40 & 165 \end{pmatrix}$$

El producte de les dues matrius dividit pel total d'observacions $n = 1499$ donarà les freqüències esperades

$$\frac{1}{1499} AB = \frac{1}{1499} \begin{pmatrix} 859 \\ 640 \end{pmatrix} \begin{pmatrix} 213 & 795 & 286 & 40 & 165 \end{pmatrix} = \begin{pmatrix} \frac{859 \cdot 213}{1499} & \frac{859 \cdot 792}{1499} & \cdots & \frac{859 \cdot 165}{1499} \\ \frac{640 \cdot 213}{1499} & \frac{640 \cdot 792}{1499} & \cdots & \frac{640 \cdot 165}{1499} \end{pmatrix}$$

Fem els càlculs amb R (recorda que el producte de dues matrius A i B s'escriu com `A%*%B`):

```
> (a<-as.matrix(margin.table(taula,1)))
       [,1]
Hombre  640
Mujer   859

> (b<-as.matrix(margin.table(taula,2)))
           [,1]
Casado      795
Divorciado  213
Separado     40
Soltero     286
Viudo       165
# transposem b
> (b<-t(b))
     Casado Divorciado Separado Soltero Viudo
[1,]    795        213       40     286   165
# fem el càlcul:
> round(a%*%b/sum(taula),1)
       Casado Divorciado Separado Soltero Viudo
Hombre  339.4       90.9     17.1   122.1  70.4
Mujer   455.6      122.1     22.9   163.9  94.6
```

Ara tocaria comparar la taula de la distribució conjunta `taula` amb la que acabem d'obtenir. Si s'assemben els dos factors són independents. Veiem que no ho són.

```
        ecivil
sexo     Casado Divorciado Separado Soltero Viudo
  Hombre    383         75        9     142    31
  Mujer     412        138       31     144   134
```

- R pot fer el **càlcul automàtic de les freqüències esperades** amb la funció `chisq.test`

```
> contingencia<-chisq.test(sexo,ecivil)
> contingencia$expected
        ecivil
sexo       Casado Divorciado Separado  Soltero    Viudo
  Hombre 339.4263   90.94063 17.07805 122.1081 70.44696
  Mujer  455.5737  122.05937 22.92195 163.8919 94.55304
> contingencia$observed
        ecivil
sexo     Casado Divorciado Separado Soltero Viudo
  Hombre    383         75        9     142    31
  Mujer     412        138       31     144   134
```

El p valor associat a una taula de contingència

El p valor associat a la taula de contingència serveix per quantificar el grau d'independència, com veureu a Estadística Inferencial. Es pot caluclar amb la funció `chisq.text`. Un p valor petit ($p < 0.05$) indica que les variables són dependents.

```
contingencia$p.value
[1]  2.013619e-13
```

Veiem que les variables `sexo` i `ecivil` són dependents.

Pràctica: Calcula la distribució conjunta i la taula de freqüències esperades de les variables `eutan` i `relig`. Són independents les dues variables?