

BLOC 2. VAD - Dades dels germans

1 Introducció

Seguidament, anem a treballar amb **Variables Aleatòries Discretes** (VAD). En lloc d'inventar-nos les probabilitats de les variables emprades, la font d'informació en la que ens basarem serà un recull de dades introduïdes per altres alumnes. Això no és tan estrany, sobretot si es disposa de moltes dades. Però de cap manera estem dient que qualsevol mostra pot servir per a caracteritzar una VAD. Simplement, com a exercici didàctic, no explotarem la informació amb eines d'estadística descriptiva, sinó que suposarem que ens permeten conèixer les distribucions de probabilitat subjacents. Els objectius d'aquesta sessió són:

- (1) Veure conceptes de VAD
- (2) Aprendre a manipular vectors i `data.frames` amb R.

Les dades que farem servir es troben al document online:

<https://docs.google.com/spreadsheets/ccc?key=0ArpNEQeErsDJdDIhSHp6STNYOTIBV2gyWjZVZm9XTnc>


A la nostra base de dades tindrem la següent informació:

Nom de la variable	Explicació de la variable
grup	Grup de l'assignatura de PE
edat	Edat
gen	Gènere
nger	Nombre germans (tu inclòs)
genger	Ordre de germans segons gènere
gerpa	Nombre de germans del pare (el pare inclòs)
germa	Nombre de germans de la mare (la mare inclosa)

Importar les dades a R

Per carregar les dades al R, feu els següents passos:

1. Creeu un nou script amb la comanda (no la executeu):

```
DF <- read.table("clipboard", header=TRUE, sep='\t',
stringsAsFactors = TRUE)
```
2. Seleccionem les cel·les B2 a H170 i les copiem al portapapers (Ctrl+C a Windows)
3. Executeu la comanda del pas 1 amb Ctrl + Enter o amb el símbol 

Inspeccionar les dades

```
# Comprovació de la lectura de dades
dim(DF)           # Nombre d'observacions i de variables
names(DF)         # Noms de les variables
head(DF)          # Primeres files de les dades
summary(DF)       # Descriptiva de les variables a la base
```

2 Una única VAD

2.1 Exemples

1.- Definiu i dibuixeu la funció de probabilitat i de distribució de la variable discreta *nger*.

Crearem una funció que faci el que demana l'enunciat per qualsevol variable

```
f <- function(x) {

##-- Càlcul funcions
  t <- table(x)
  k <- as.numeric(names(t))      # Valors que pren la VAD
  p <- as.numeric(prop.table(t)) # Función de probabilidad
  F <- cumsum(p)                 # Función de distribució
  df <- data.frame(k,p,F)        # Taula amb dues funcions
  print(k);print(p);print(F)

  ##-- Gràfic funcions
  par(mfrow=c(1,2),las=1)
  plot(k,p,type="h",main="Probabilitat")
  plot(k,F,type="S",main="Distribució")

  return(df)
}

# DF <- read.table("clipboard", header=TRUE, sep='\t')
Df.nger <- f(x = DF$nger)
Df.nger
```

2.- Calculeu amb R la esperança i la variància del nombre de germans.

```
##-- Esperança (E) i variància (V)
E <- with(Df.nger, sum(k*p))
V <- with(Df.nger, sum((k-E)^2*p))
E;V
```

2.2 Exercicis

1. Estudieu les variables 'Edat', 'Número de germans del pare' i 'Número de germans de la mare' com a variables aleatòries: determineu les funcions de probabilitat, representeu-les gràficament i trobeu els principals indicadors de tendència central i dispersió.
2. A nivell poblacional, te sentit pensar que la mateixa variable pot representar la distribució del número de germans del pare o de la mare. Obteniu les funcions de probabilitat i distribució de la variable: "número mitjà de germans de pare i mare".

3 Dues VADs

3.1 Exemples

3.- A la primera part es va estudiar la distribució d'una sola VAD, i ara veurem l'estudi de dues VADs que han estat mesurades a la mateixa unitat d'observació (l'alumne). L'objectiu es determinar si existeix una relació entre algun parell de variables concret. Genereu la funció de probabilitat conjunta per les variables nombre de germans (nger) i nombre de germans del pare (gerpa)

```
##-- Esperança (E) i variància (V)
tab.conj <- with(DF, table(nger, gerpa))
round(prop.table(tab.conj), 3)
```

		gerpa									
		1	2	3	4	5	6	7	8	9	10
nger	1	0.048	0.065	0.054	0.018	0.006	0	0	0	0	0
	2	0.06	0.196	0.161	0.113	0.048	0	0.018	0.024	0.006	0.006
	3	0.018	0.042	0.018	0.036	0.012	0.006	0.006	0	0	0.006
	4	0	0.012	0.006	0.006	0	0.006	0	0	0	0
	5	0	0	0	0	0.006	0	0	0	0	0

4.- Quina és la probabilitat que el nombre de fills sigui el mateix que el nombre de germans a la família del pare?

```
sum(diag(prop.table(tab.conj)))
```

5.- Representeu la funció de probabilitat conjunta amb el paquet "scatterplot3d".

```
par(mfrow=c(1,1), las=1)
conj <- prop.table(tab.conj)
X <- rep(rownames(conj), 10)
Y <- rep(colnames(conj), each=5)
scatterplot3d(X, Y, conj, type='h', lwd=3, lab=c(5, 5, 1), box=FALSE,
color=X, xlab='Germans fill', ylab='Germans pare', zlab='p')
```

6.- Trobeu les funcions de probabilitat de les variables 'nger' i 'gerpa'. Trobeu les seves esperances i les seves desviacions tipus. Empreu aquesta informació per calcular la correlació entre les dues variables.

```
##-- Funcions de probabilitat
px <- apply(conj,1,sum)
py <- apply(conj,2,sum)
px;py

##-- Esperances
kx <- as.numeric(names(px))
Ex <- sum(kx*px)
ky <- as.numeric(names(py))
Ey <- sum(ky*py)
Ex;Ey

##-- Desviacions tipus
Vx <- sum((kx-Ex)^2*px)
Vy <- sum((ky-Ey)^2*py)
Sx <- sqrt(Vx)
Sy <- sqrt(Vy)

##-- Càlcul de la correlació
Cov <- sum(((kx-Ex) %*% t((ky-Ey))) * conj)
Corr <- Cov / (Sx*Sy)
Corr
```

7.- Per obtenir la probabilitat condicionada l'únic que s'ha de fer és dividir la probabilitat conjunta per la probabilitat marginal. Obteniu la distribució condicionada del número de germans del pare condicionada al número de germans del fill i viceversa

```
round(conj/apply(conj,1,sum),3) # Condicionat als germans del fill
round(conj/apply(conj,2,sum),3) # Condicional als germans del pare
```

3.2 Exercicis

- Podem repassar una coneguda propietat de les VA: $E(XY) = E(X) E(Y) + \text{Cov}(X,Y)$? Per fer això, hauràs de crear una variable XY multiplicant els valors de X i Y (germans de l'alumne i germans del pare, per exemple), trobar el seu valor esperat i comprovar que coincideix amb el producte d'esperances més la covariància. Recorda que $\text{Cov}(X, Y) = \rho \sigma_1 \sigma_2$.
- Una vegada teniu les funcions de probabilitat condicionada, podeu trobar indicadors com esperança i desviació tipus per cada cas. Calculeu el número esperat de germans (per a un fill) si a la família del pare eren 2. I si fossin 4?