# *Anàlisi Multivariant de Dades*

## *Grau Interuniversitari en Estadística*

## *Prova de Laboratori*

***17 juny 2016***

**NOM: _____**

Llegiu atentament les instruccions:

Connecteu-vos a la pàgina habitual de l'assignatura i baixeu-vos les dades Crimes (a l'apartat Reference DataSets)

Les dades fan referència a homicidis ocorreguts a l'estat de Califòrnia entre 1950 i 2000

Seguiu les instruccions de l'enunciat i contesteu a les preguntes, SOBRE AQUEST MATEIX FULL D'EXAMEN (no s'admeten fulls de respostes en llàpis; Assegureu-vos de posar el nom e l'examen en l'espai habilitat a tal efecte sobre aquestes instruccions). Al darrera, podeu afegir tots els grafics que es demana adjuntar EN FULLS FIRMATS I DEGUDAMENT MARCATS AMB EL NÚMERO DE LA PREGUNTA .

Cal escriure amb lletra del propi puny les respostes a cada pregunta

**Pregunta 1)** Executeu la següent instrucció de lectura

```
df<-read.csv("Crimes.csv", sep=';', row.names=1)
```

I observeu que R crea un data.frame de 1316 observacions i 0 variables. Expliqueu per què passa això, corregiu l'error .

Escriviu aquí dessota la instrucció R corregida que genera el dataframe corecte

df<-read.csv ("Crimes.cv", _____)

**Pregunta 2)** Amb quina instrucció de R podeu saber la classe de la variable COUNTY?

Instrucció R: _____

De quin tipus diu R que és aquesta variable?

Utilitzeu aquesta instrucció adequadament per esbrinar quines variables reconeix R com a factors. Quines són (poseu una llista amb el nom i posició a la BD entre parèntesi):

Reproduïu aquí el codi R que heu utilitzat per a respondre aquesta pregunta:

**Pregunta 3)** Efectueu l'anàlisi descriptiva gràfica i numèrica de la variable Month

Adjunteu en full marcat AMB EL VOSTRE NOM i etiqueta "Pregunta3" el resultat

Codi R utilitzat:

**Pregunta 4)** Creeu una variable nova que es digui "MonthLit" i, per a cada mes, contingui el literal del mes de l'homicidi segons la següent equivalència

| Month | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MonthLit | Jan | Feb | Mar | Ap | May | Jun | Jul | Aug | Sep | Oct | Nov | Dic |

Codi R:

Adjunteu en full apart els resultats de la descriptiva gràfica i numèrica de la variable MonthLit

Comenteu la distribució de la variable

**Pregunta 5)** Afegiu la variable MonthLit al dataframe i poseu aquí sota quin codi R heu executat per fer-ho. Assegureu-vos que el dataframe actualitza adequadament la metainformació corresponent

Codi R:

Com mireu si la transformació de MONTH a MonthLit s'ha fet correctament?

Instrucció R: _____

Què surt?

**Pregunta 6)** Feu el summary(df) i reviseu si cal algun altre tractament previ per poder iniciar l'anàlisi multivariant. Contesteu a les següents preguntes

a) Hi ha dades faltants? Si/No
Cas que n'hi hagi ompliu la següent taula per ordre d'aparició a la base de dades:

| Posició variable | Nom variable | Tipus (num, cat, ord) | Representació de la dada mancant | Quantes dades mancants conté la var |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

b) Hi ha alguna altra variable que sigui absolutament necessari transformar explícitament com a factor? Quina/es? Per què?

c) Hi ha algun factor que sigui ordinal i calgui fixar-ne l'ordenació de les modalitats? Quin? Com ho faríeu?

**Pregunta 7)** Construïu un vector amb els índexos de les variables actives anomenat Actives on:

a) el MONTH no hi sigui, i es mantingui el MonthLit.
b) No hi hagi cap variable numèrica amb dades mancants

Instruccions R utilitzades:

c) Quantes variables numèriques han quedat actives? ___ Quantes qualitatives?___ Com ho heu calculat? Instruccions R:

**Pregunta 8)** Segons la configuació en tipus de les variables Actives, quina funció faríeu servir per calcular les distàncies entre els individus:

```
a) D  <- dist(df)
b) D  <- daisy(df[], metric = "gower", stand=TRUE)
c) D  <- daisy(df[,Actives],  stand=TRUE)
d) S  <- daisy(df[,Actives], metric = "gower", stand=TRUE); D<-S^2
e) D  <- daisy(df[,Actives], metric = "manhattan")
f) D  <- dist(df[,Actives])
```

Executeu l'opció seleccionada i construïu D. Feu summary(D) i ompliu la següent taula (5 decimals)

| | Min | Q1 | Median | Mean | Q3 | Max |
|---|---|---|---|---|---|---|
| D | | | | | | |

**Pregunta 9)** Amb quines instruccions R feu cluster jeràrquic de "Actives" pel mètode de Ward?

Codi R:

Aporteu la imatge del dendrograma resultant (annexeu-la a l'examen)

A la vista del dendrograma preferiu un tall en 3, 7 o 15 classes? Per què?

Executeu el tall seleccionat i construïu una variable d'R anomenada P que el contingui

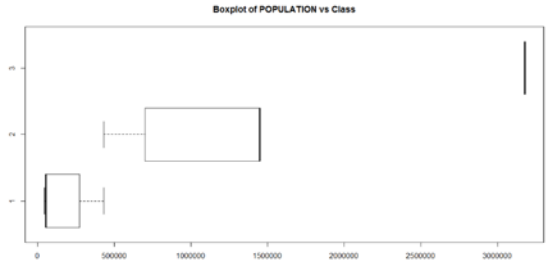
Quina grandària tenen les classes?

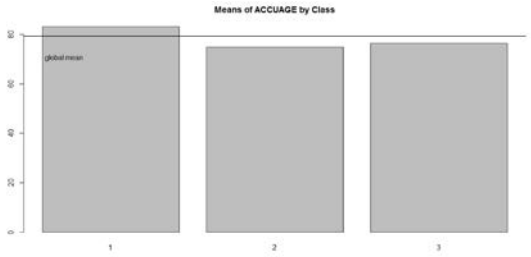
| P | Classe1 | Classe2 | … |
|---|---|---|---|
| Size | | | |

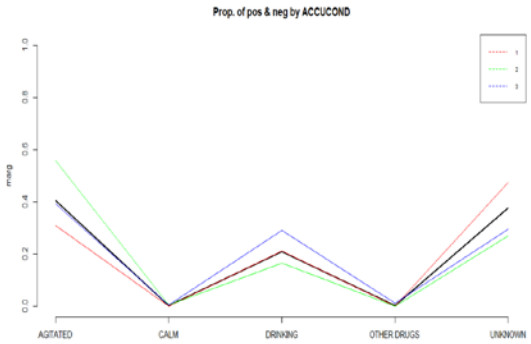
Amb quina instrucció R ho heu esbrinat? _____

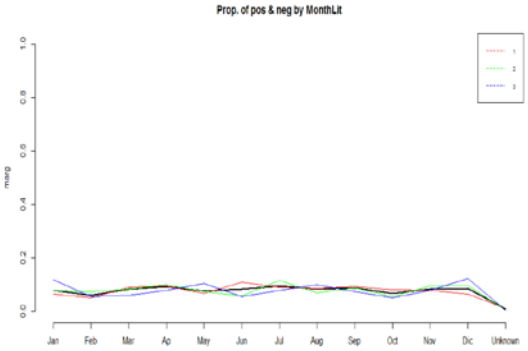
Afegiu la variable P a la columna 32 de la base de dades (després us farà falta)

**Pregunta 10)** Marqueu amb una V aquells dels següents gràfics de profiling corresponents a variables que discriminen les classes i amb una X les que no




Marqueu amb FLETXA CAP AMUNT aquells gràfics que indiquen que la variable es més gran a la classe 1, si n'hi ha alguna

Describiu el perfil de cada classe en funció d'aquestes variables:

Classe1

Classe2

Classe3

**Pregunta 11)** Calculeu les mitjanes de les observacions útils de les variables VicAge i Accuage per cada classe i ompliu la taula següent (utilitzeu tantes classes com considereu oportunes d'acord amb el que heu decidit a la pregunta 9.

Instrucció R per fer-ho:

| **Posició variable** | **Nom variable** | **Representació de la dada mancant** | **Mitjana per Classe1** | **Mitjana per Classe2** | **Mitjana per Classe3** | **…..** |
|---|---|---|---|---|---|---|
| | Viccage | | | | | |
| | AccuAge | | | | | |

Substituïu a la base de dades els valors mancants de les variables Viccage i Accuage per les mitjanes locals a la classe on pertanyi cada observació i ompliu la següent taula:

| D | Min | Q1 | Median | Mean | Q3 | Max |
|---|---|---|---|---|---|---|
| Vicage original | | | | | | |
| Vicage without missings | | | | | | |
| VicageClean | | | | | | |
| AccuAge Original | | | | | | |
| AccuAge without missings | | | | | | |
| AccuAgeClean | | | | | | |

**Pregunta 12)** Construïu un dataframe de R anomenat ActivesACP que contingui les variables numèriques, excepte MONTH, DAY, YEAR, NUMWOMEN i NUMMEN i efectueu una Anàlisi en components principals amb les dades centrades i normalitzades. Poseu aquí dessota el vector d'indexos amb les posicions de les variables que agafeu de df per intervenir a l'ACP

indexosActivesACP: c(_____)

Comanda per l'ACP?

Quantes components principals obteniu?____

Annexeu diagrama de barres amb la inercia acumulada en cada subespai de dimensió creixent

Quants eixos calen per conservar un mínim del 80% de la informació?_____

Adjunteu la projecció dels individus del primer pla factorial, distingint el color dels punts segons la variable county. Comenteu el gràfic aquí dessota

Adjunteu la projecció de les  variables numeriques actives en “darkgray”.

Quin county es mes gran?_____

Quin county té menys graduats de High School?_____

Codifiqueu els missings de YEAR en NA:

Intrucció R: _____

Creeu un dataframe illustrativesNum que contingui les variables numèriques illustratives YEAR i NUM.WOMEN. Adjunteu la projecció conjunta de les variables numeriques actives en “darkgray” i les numèriques illustratives “darkgreen”. Les coordenades de les i.lustratives sobre el pla factorial les heu de calcular amb la instrucció

```
PhiIllustratives  = cor(illustrativesNum,Psi, use="pairwise.complete.obs")
```

Què podeu dir de les illustratives?

Adjunteu un altre mapa que afegeixi a l’anterior la projecció de les seguents variables qualitatives:

COUNTY, WEAPON, ACCURACE, VICRACE, ACCUSEX, RELATION

Codi R per fer aquesta projecció:

Quines armes prefereixen les dones?_____

A quins counties actuen  preferenment assassins dona?_____

I les armes dels mes cultes preferentment, son:_____

I les dels que assassinen al conjuge?_____

Perfil de crim que involucra víctimes de raça americana-nativa?

On i com assassinen els African American?