PAC2 Primavera 2021

UOC

Les PACs es basaran en una base de dades obtinguda a partir del repositori de microdades del "Banc Mundial" a [https://microdata.worldbank.org/index.php/catalog/424/get-microdata].

Conté indicacions, entre d'altres de

1. *City* = Nom de la ciutat
2. *Country* = País
3. *Population2000* = Població de la ciutat a l'any 2000.
4. *PM10Concentration1999* = PM10 concentrations (micro grams per cubic meter) in residential areas of cities larger than 100,000, l'any 1999
5. *Region* = Classificació en regió geogràfica
6. *IncomeGroup* = Classificació segons nivell d'ingressos del país.

Per importar les dades podem usar la següent instrucció i per comprovar que tot funcioni mostrem els 3 primers registres:

```
dadesPM10<-read.table("AirPollution2000WB_UOC2.csv", header=TRUE,
          sep=";",na.strings="NA",
          fileEncoding = "UTF-8", quote = "\"",
          colClasses=c(rep("character",4),rep("numeric",2),
                       rep("character",2)))
head(dadesPM10,3)
```

```
##   Cod     Country Citycode                 City Population2000
## 1 AFG Afghanistan    40003               Herat         323741
## 2 AFG Afghanistan    40001               KABUL        2457496
## 3 AFG Afghanistan    40002 Kandahar (Quandahar)         411752
##   PM10Concentration1999     Region IncomeGroup
## 1                    46 South Asia  Low income
## 2                    46 South Asia  Low income
## 3                    51 South Asia  Low income
```

Us pot ser útil consultar el següent material:

1. Mòdul: Probabilitat i variables aleatòries.
2. Anàlisi de dades i estadística descriptiva amb R.
3. Activitats Resoltes de Probabilitat i variables aleatòries.

NOM:

PAC2

Un cop importades les dades,

Pregunta 1. (30%)

a) Feu una taula amb el nombre de ciutats que hi ha en cadascuna de les regions. Mostreu el resultat. (10%)

A partir de les dades obtingudes en la taula anterior, calculeu les següents probabilitats:

b) Probabilitat que una ciutat triada a l'atzar estigui en la regió `Latin American & Caribbean`. (10%)

c) Probabilitat que una ciutat triada a l'atzar `pugui estar a Àsia o que estigui a alguna regió que contingui ciutats d'Àsia`. (10%)

Solució:

a) Fem:

```
table(dadesPM10$Region)
```

```
##
##        East Asia & Pacific      Europe & Central Asia
##                        839                        871
##  Latin America & Caribbean Middle East & North Africa
##                        467                        191
##               North America                 South Asia
##                        256                        402
##         Sub-Saharan Africa
##                        192
```

b) Probabilitat que una ciutat triada a l'atzar estigui en la regió `Latin American & Caribbean`:

$$\frac{467}{839+871+467+191+256+402+192} = \frac{467}{3218} = 0.1451$$

c) Probabilitat que una ciutat triada a l'atzar `pugui estar a Àsia o que estigui a alguna regió que contingui ciutats d'Àsia`: serà la probabilitat que estigui en `South Asia` més la probabilitat que estigui en `Europe & Central Asia` més la probabilitat que estigui en `East Asia & Pacific`, és a dir,

$$\frac{402}{3218} + \frac{871}{3218} + \frac{839}{3218} = \frac{2112}{3218} = 0.6563$$

Pregunta 2. (70%)

En aquest segon exercici, en primer lloc codificarem la variable `PM10Concentration1999` en 5 categories:

- `Concentració Molt Baixa (MB)`, per valors de `PM10Concentration1999` iguals o inferiors al primer quartil, és a dir, `PM10` $\leq Q_1$.
- `Concentració Baixa (B)`, per valors de `PM10Concentration1999` superiors al primer quartil i iguals o inferiors al segon quartil, és a dir, $Q_1 <$ `PM10` $\leq Q_2$.
- `Concentració Moderada (M)`, per valors de `PM10Concentration1999` superiors al segon quartil i iguals o inferiors a la mitjana aritmètica, és a dir, $Q_2 < PM10 \leq \overline{X}$.
- `Concentració Alta (A)`, per valors de `PM10Concentration1999` superiors a la mitjana aritmètica i iguals o inferiors al tercer quartil, és a dir, $\overline{X} <$ `PM10` $\leq Q_3$.
- `Concentració Molt Alta (MA)`, per valors de `PM10Concentration1999` superiors al tercer quartil, és a dir, `PM10` $> Q_3$.

a) Crear la variable `Tipus_Concentracio` amb les especificacions anteriors de manera que podrà ser: `Molt Baixa (MB), Baixa (B), Moderada (M), Alta (A)` i `Molt Alta (MA)`. (10%)

b) Trobar la taula de contingència d'aquesta nova variable, `Tipus_concentracio` i la variable `Nivell d'ingressos (IncomeGroup)`. (10 %)

Indicant les fórmules i calculant les probabilitats manualment a partir de la taula de contingència que ens ha donat R en l'apartat anterior, es demana:

c) Probabilitat que una ciutat estigui en el grup de nivell d'ingressos alts. (10%)
d) Probabilitat que una ciutat estigui en el grup de concentració molt alta. (10%)
e) Probabilitat que una ciutat estigui en el grup d'ingressos baixos i estigui també en el grup de concentració de `PM10Concentration1999` molt alta.(10%)
f) Probabilitat que una ciudad estigui en el grup de concentració de `PM10Concentration1999` moderada sabent que està en el grup de nivell d'ingressos alts. (10%)
g) Els successos `ser una ciutat d'ingressos alts` i `estar en el grup de contaminació de PM10Concentration1999 baixa`, són independents? Per què? (10%)

Solució:

a) En primer lloc calculem el resum estadístic corresponent a la concentració de PM10Concentration1999:

```
summary(dadesPM10$PM10Concentration1999)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     6.0    24.0    38.0    51.1    71.0   359.0
```

A continuació codifiquem la variable concentració de `PM10Concentration1999` creant la nova variable `Tipus_Concentracio` i llistem els primers registres per comprovar que obtenim el que volem: (10%)

```
Tipus_concentracio<-cut(dadesPM10$PM10Concentration1999,
  breaks=c(0, 24.0, 38.0, 51.1, 71.0, 359.0),labels=c("Molt Baixa","Baixa",
     "Moderada","Alta","Molt Alta"),include.lowest=TRUE)
table(Tipus_concentracio)
```

```
## Tipus_concentracio
## Molt Baixa      Baixa   Moderada       Alta  Molt Alta
##        878        772        411        366        791
```

b) Taula de contingència de la variable, `Tipus_concentracio` i la variable `Nivell d'ingressos (IncomeGroup)`: (10%)

```
conting<-table(dadesPM10$IncomeGroup,Tipus_concentracio)
rownames(conting)<-c("Ing. Alts", "Ing. Baixos", "Ing. Mig Baixos", "Ing. Mig Alts")
conting
```

```
##                  Tipus_concentracio
##                   Molt Baixa Baixa Moderada Alta Molt Alta
##   Ing. Alts              513   391      138   32        21
##   Ing. Baixos              6    15       20   19        39
##   Ing. Mig Baixos         17   130      114  126       369
##   Ing. Mig Alts          342   236      139  189       362
```

c) Probabilitat que una ciutat estigui en el grup de nivell d'ingressos alts (`IA`): (10%)

$$P(\texttt{IA}) = \frac{513 + 391 + 138 + 32 + 21}{3218} = \frac{1095}{3218} = 0.3403$$

d) Probabilitat que una ciutat estigui en el grup de concentració de `PM10Concentration1999` molt alta (`MA`): (10%)

$$P(\texttt{MA}) = \frac{21 + 39 + 369 + 362}{3218} = \frac{791}{3218} = 0.2458$$

e) Probabilitat que una ciutat estigui en el grup d'ingressos baixos (`IB`) i estigui també en el grup de concentració de `PM10Concentration1999` molt alta (`MA`):(10%)

$$P(\texttt{IB} \cap \texttt{MA}) = \frac{39}{3218} = 0.01212$$

f) Probabilitat que una ciudad estigui en el grup de concentració de `PM10Concentration1999` moderada (`M`) sabent que està en el grup de nivell d'ingressos alts (`IA`): (10%)

$$P(\texttt{M}|\texttt{IA}) = \frac{P(\texttt{M} \cap \texttt{IA})}{P(\texttt{IA})} = \frac{138}{513 + 391 + 138 + 32 + 21} = \frac{138}{1095} = 0.1260$$

g) Els successos `ser una ciutat d'ingressos alts (IA)` i `estar en el grup de contaminació de PM10Concentration1999 baixa (B)`, són independents? Per què?

(10%)

Per què siguin independents s'ha de verificar que:

$$P(\texttt{IA} \cap \texttt{B}) = P(\texttt{IA}) \cdot P(\texttt{B})$$

Aquests successos no són independents ja que:

$$P(\texttt{IA} \cap \texttt{B}) = \frac{391}{3218} = 0.1215$$

$$P(\texttt{IA}) \cdot P(\texttt{B}) = 0.3403 \cdot \frac{391 + 15 + 130 + 236}{3218} = 0.3403 \cdot 0.2399 = 0.08164.$$