Grau interuniversitari (UB-UPC) d'Estadística
Software Estadístic: Pràctica final amb R
Data límit d'entrega: 18 de desembre de 2016

Instruccions

Cada grup ha d'entregar un arxiu comprimit amb nom `Cognom1_Cognom2.zip` (o `Cognom.zip` si es realitza la pràctica individualment) que consti de la següent informació:

- Una àrea de treball que contingui les dades utilitzades en l'Exercici 1.
- Una memòria PDF que reculli la resolució de cadascun dels apartats de l'Exercici 1.
- Els dos *scripts* d'R utilitzats per resoldre cadascun dels dos exercicis de la pràctica, amb els comentaris necessaris dins del codi afegits darrera d'un coixinet (`#`). Igualment, amb la sintaxi `library`, s'indicarà la utilització d'una determinada llibreria.
- A la programació dels dos *scripts* d'R seguiu les recomanacions d'estil a `https://google.github.io/styleguide/Rguide.xml`.

Tant en els *scripts* com a la memòria PDF hi figuraran els noms dels integrants del grup, i l'entrega (via el Campus Virtual) de l'arxiu `Cognom1_Cognom2.zip` es podrà efectuar fins a la data límit indicada a la capçalera d'aquest document.

Exercici 1 (5.5 punts)

Heu d'aconseguir una base de dades (ASCII, EXCEL, SPSS o SAS) en què es mesurin una sèrie de variables sobre **un mínim de 50 individus**. En concret, cada observació constarà d'una variable identificadora `id` (pot ser un nom de persona, una marca, un codi, etc), i estarà acompanyada de set variables més: Una variable categòrica binària, una variable categòrica politòmica (amb més de 2 categories) i cinc variables de tipus numèric. Es valorarà la dificultat i l'originalitat de les dades triades, i en cap cas s'admetran els següents conjunts de dades:

- Les utilitzades a pràctiques d'R d'altres anys o en altres assignatures (els alumnes repetidors de l'assignatura que van fer el treball sí podran fer servir les seves dades de l'any passat en el cas que realitzin la pràctica de forma individual).
- Les ja utilitzades per realitzar qualsevol tipus d'informe estadístic que es pugui trobar.
- Les que estiguin incorporades dins d'alguna llibreria d'R o bé incloses en qualsevol altre programa estadístic.

La utilització de dades no permeses suposarà la invalidació de l'exercici.

La memòria en PDF per a aquest exercici contindrà la resolució dels següents apartats:

a) Redacteu un apartat d'introducció que contingui els següents punts:
  - El tipus de dades que utilitzeu en el treball.
  - La font d'obtenció de les dades (incloent l'enllaç a internet en cas d'haver-ne).
  - El significat de les diverses variables segons la seva nomenclatura dins la base de dades.
  - L'objectiu de l'anàlisi que es realitza.

b) Importeu les dades a R com un *data frame* anomenat `df` i substituïu el nom de cada registre de `df` pel corresponent nom de la variable `id` (la qual desapareix al fer la substitució). Independentment de que `df` contingui valors perduts, afegiu-hi aleatòriament 1 *missing* a la 1a variable numèrica i 2 *missings* a la 2a variable numèrica.

c) Realitzeu una anàlisi descriptiva univariant de totes les vostres variables (tant categòriques com numèriques), posant especial atenció sobre aquelles variables que continguin algun valor anòmal (*outlier*). Comenteu tots els resultats numèrics i gràfics.

d) Estimeu la matriu de correlacions `matR` amb tota la informació disponible de les vostres variables numèriques, obtenint el nivell de significació d'aquelles correlacions que considereu més rellevants. A continuació realitzeu una anàlisi descriptiva multivariant entre aquelles variables (del tipus que siguin) que més poden contribuir a comprendre la informació. Comenteu tots els resultats numèrics i gràfics.

e) Redacteu un apartat de conclusions que inclogui els principals resultats obtinguts a la vostra anàlisi descriptiva. Realitzeu un judici crític d'aquests resultats, tot esmentant les causes que poden explicar en cada cas el comportament de les variables.

Exercici 2 (4.5 punts)

Programeu una funció que faci el següent donat un *data frame* (`dades`) i el nom d'una de les seves variables categòriques (`cvar`) com a arguments principals:

1. Comprovar que `dades` sigui un *data frame*. En cas contrari, la funció ha d'avortar la seva execució tornant un missatge d'error en català o castellà.

2. Comprovar que `dades` contingui una variable categòrica amb nom `cvar`. En cas contrari, la funció ha d'avortar la seva execució tornant un missatge d'error en català o castellà.

3. Comprovar que `dades` contingui almenys una variable numèrica. En cas contrari, la funció ha d'avortar la seva execució tornant un missatge d'error en català o castellà.

4. Tornar la següent informació sobre el *data frame* `dades`:

   (a) Nombre de files i columnes,

   (b) Taula de freqüències dels tipus de variables,

   (c) Nombre de *missings* per variable,

   (d) Fila amb més *missings*.

5. Calcular diferents indicadors numèrics (mitjana, mediana, desviació estàndard, etc.) de totes les variables numèriques per a les diferents categories de `cvar`.

6. De forma opcional, dibuixar gràfics de mosaics per a la resta de variables categòriques per representar les distribucions condicionals d'aquestes variables en funció de `cvar`.

7. De forma opcional, guardar tots aquests gràfics en un sol document pdf.

Apliqueu la funció a vostres dades de l'Exercici 1 i comenteu la sortida de la funció.