Lliurament: **Divendres 03/04/2020**.

1 Introducció

Aquesta pràctica consta de dues parts. A la primera part analitzarem variables categòriques mitjançant l'anàlisi de correpondències (CA) i aprendrem a interpretar els seus resultats. A la segona part estudiem l'ús de la distribució normal multivariant per dades multivariants, i aplicarem alguns mètodes de inferència multivariant com el T^2 de Hotelling. Tal com s'ha fet a les pràctiques anteriors, cada grup d'estudiants ha de realitzar el guió de la pràctica a continuació, realitzant els càlculs i gràfics necessaris en l'entorn R. Els resultats obtinguts i les respostes a les preguntes s'han de recollir en un document amb la vostre solució, fet per exemple en R markdown, o bé en Microsoft Word, o latex. Procureu posar els vostres noms i cognoms i número de grup a l'inici del document. De matrius grans nomes cal incloure les primeres 5 files a l'informe. Empleneu la mateixa numeració dels ítems de l'enunciat al document amb la vostre solució. Lliureu la solució del qüestionari en format PDF, pujant-la a l'entorn Atenea (atenea.upc.edu), a l'apartat de la tasca corresponent, abans de la data final de la tasca.

2 Exercicis

1. (20p) **Anticonceptius a Indonèsia**.

   Les dades que fem servir en aquesta pràctica provenen d'un estudi sobre ús de mètodes anticonceptius a Indonèsia a l'any 1987, i són extretes d'un arxiu de dades per machine learning (archive.ics.uci.edu). Les dades corresponen als resultats d'una enquesta en la qual es van consultar 1473 dones, i estan enregistrades al fitxer `cmc.dat`. Es disposa de les variables amb la seva codificació corresponent, com descrit a continuació:

Age:	Wife's age
WifeEduc:	Wife's education; 1=low, 2, 3, 4=high
HusbEduc:	Husband's education; 1=low, 2, 3, 4=high
Childs:	Number of children ever born
Rel:	Wife's religion; 0=Non-Islam, 1=Islam
Work:	Wife's now working?; 0=Yes, 1=No
Occup:	Husband's occupation; 1, 2, 3, 4
Standard:	Standard-of-living index; 1=low, 2, 3, 4=high
Media:	Media exposure; 0=Good, 1=Not good
Method:	Contraceptive method used; 1=No-use, 2=Long-term, 3=Short-term

   (a) Carregueu el fitxer cmc.dat a l'entorn R. Verifiqueu si les variables prenen valors dins el rang esperat i comenteu, si procedeix, el que feu amb possibles dades mancants.

   (b) Re-codifiqueu la variable *Method*, canviant 1 per "none", 2 per "long-term" i 3 per "short-term". Feu que la variable sigui reconeguda com a categòrica dins R amb `factor`. Re-codifiqueu també convenientment la variable *WifeEduc*, passant-la també a categòrica amb `factor`.

   (c) (1p) Feu una taula de contingència entre les variables *Method* i *WifeEduc* utilitzant la funció `table`.

(d) (2p) Contrasteu mitjançant una prova de chi-quadrat (funció `chisq.test`) la hipòtesi d'independència. Doneu el valor de l'estadístic de prova i el valor p i feu constar la vostre conclusió.

(e) (1p) Calculeu els vectors de pesos columna i pesos fila. Quin és el mètode anticonceptiu més utilitzat?

(f) (1p) Calculeu la matriu de perfils fila i la matriu de perfils columna.

(g) (1p) Quin és el perfil *marginal* de la taula de perfils fila?

(h) (1p) Instal.leu el paquet `ca` amb `install.packages("ca")`. Feu un anàlisi de correspondències simples de la taula que acabeu de crear, amb les instruccions `resultats <- ca(taula)` i `summary(resultats)`. Quantes dimensions existeixen en la solució de l'anàlisi? Quantes dimensions es necessiten per representar adequadament la taula de contingència?

(i) (1p) Feu un biplot dels perfils amb la instrucció `plot(resultats)`, posant la opció `map="rowprincipal"`. Interpreteu la gràfica obtinguda.

(j) (1p) Quin dels perfils fila s'assembla mes al perfil marginal?

(k) (1p) Categoritzeu la variable *Age* en quatre franges (min,Q1), (Q1,Me), (Me,Q3) i (Q3,max). Podeu fer servir la funció `cut` amb l'opció `breaks` per fer-ho. Feu una taula amb el número d'observacions a cadascun dels intervals d'edat.

(l) (1p) Feu la codificació interactiva de la variable *Age* categoritzada i la religió (*Rel*), creant una nova variable categòria, anomenada *AgeRel* amb 8 categories. Feu una taula amb les freqüencies de les 8 categories.

(m) (1p) Feu una taula de contigència de la nova variable categòrica *AgeRel* amb *Method*.

(n) (2p) Apliqueu l'anàlisi de correspondències simple a la nova taula. Visualitzeu els resultats amb `plot`. Podeu millorar la visualització amb la opció `map="rowgreen"`. Interpreteu els resultats, descrivint el perfil de les dones que utilitzen els diferents mètodes.

(o) (2p) Feu un anàlisi de correspondències múltiple, utilitzant les tres variables *WifeEduc, Method* i *AgeRel*, amb la funció `mjca` i utilitzant la matriu d'indicadors (opció `lambda="indicator"`). Quantes dimensions té l'anàlisi? Quina és la inèrcia de la matriu d'indicadors? Quina és la bondat de representació de la matriu d'indicadores en dues dimensions?

(p) (1p) Utilitzeu la funció `plot` per fer una gràfica de la solució en dues dimensions. Interpreteu els resultats.

(q) (1p) Podeu superposar els individus al plot amb la instrucció `points` aplicat al camp `rowpcoord` dels resultats del MCA. Estan representades totes les 1473 dones? Argumenteu la resposta.

(r) (2p) Repetiu l'anàlisi de correspondències fent servir la matriu de Burt. Doneu la descomposició de la inèrcia i feu el biplot. Hi han diferències respecte a l'anàlisi utilitzant la matriu d'indicadores?

2. (15p) **Distribució de característiques esquelètiques d'homes i dones**.

Considerem la distribució de les característiques esquelètiques d'un conjunt de 507 persones, principalment entre 20 i 40 anys. Les 9 variables esquelètiques de la base de dades, en terminologia anglesa, són:

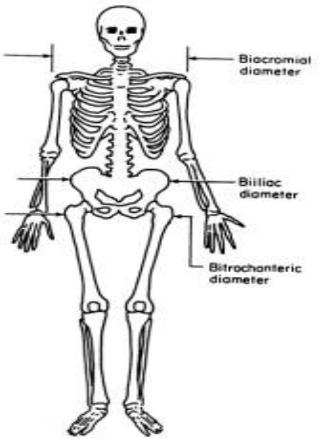

Figure 1: Mesures esquelètiques

- Biacromial diameter (mireu la figura)
- Biiliac diameter, or "pelvic breadth" (mireu la figura)
- Bitrochanteric diameter (mireu la figura)
- Chest depth between spine and sternum at nipple level, mid-expiration
- Chest diameter at nipple level, mid-expiration
- Elbow diameter, sum of two elbows
- Wrist diameter, sum of two wrists
- Knee diameter, sum of two knees
- Ankle diameter, sum of two ankles

La base de dades també conté 12 mesures de circumferència que no farem servir. També hi ha algunes variables addicionals d'un altre naturalesa, i són:

- Age (years)
- Weight (kg)
- Height (cm)
- Gender (1 - male, 0 - female)

Les dades provenen d'un estudi publicat per Heinz et.al. (2003). Per a més informació sobre les dades podeu consultar l'article:

Heinz. G., Peterson, L.J., Johnson, R.W. and Kerk, C.J. (2003) Exploring Relationships in Body Dimensions. *Journal of Statistics Education* 11, 2.

Les dades es troben en el fitxer body.dat i la descripció completa de les variables en body.txt

(a) Extreieu del fitxer `body.dat` les 9 primeres columnes, que contenen les 9 característiques esquelètiques. Totes les variables estan expressades en centímetres (cm). El sentit d'algunes variables esta indicat a la figura 1. Etiqueteu les 9 variables adequadament, segons l'ordre especificat a dalt. Extreieu també la última columna del fitxer `body.dat`, que conté el sexe de la persona.

(b) (2p) Mireu la normalitat bivariant de `Wrist diameter` and `Knee diameter`. Feu un diagrama bivariant d'aquestes variables. Calculeu el vector de mitjanes i la matriu de covariancies de les dues variables. Afegiu un contour de 95% amb la funció `ellipse` del package `ellipse`.

(c) (2p) Quantes observacions cauen fora de l'ellipse? Quantes esperaries que caiguéssin fora suposant certa la normalitat bivariant?

(d) (1p) Com podries saber si una observació esta fora de l'ellipse sense mirar la gràfica? Refeu un diagrama bivariant d'aquestes variables, marcant segons un criteri numèric, les observacions fora de l'ellipse amb un color diferent.

(e) Sel.leccioneu pels càlculs a continuació, nomes les dades de les dones (`gender=0`).

(f) (1p) Utilizeu `pairs` per fer les diagrames bivariants de cada parella de variables. Quin patró s'espera si la normalitat bivariant es compleix? Existeixen parelles de variables que no tenen el patró esperat?

(g) (1p) Programem el procediment per fer un gràfica chi-quadrat ("chisquare plot") per investigar normalitat multivariant. Calculem la matriu de dades centrades, la matriu de covariancies i la seva inversa (utilitzant `solve`). Calculem les distàncies Mahalanobis al quadrat respecte al vector de mitjanes. Ordeneu aquestes distàncies de mes petit a mes gran. Mostreu el vector ordenat a l'informe.

(h) (1p) Determineu el rang d'aquestes distàncies, mitjançant el càlcul de $(i-\frac{1}{2})/n$. Utilitzeu el rang per calcular els quantils corresponents d'una distribució de chi-quadrat. Quants graus de llibertat cal aplicar? Mostreu el vector de quantils.

(i) (1p) Grafiqueu les distàncies al quadrat versus els quantils, afegint una recta amb pendent 1 i constant zero a la gràfica (utilitzant `abline`). Donen els resultats suport a la idea que les dades són normal multivariant?

(j) (1p) Feu una gràfica panel (3×3, amb `par(mfrow=c(3,3))` i feu el "normal probability plot" de cadascuna de les set variables que estem estudiant. És creible la normalitat *marginal* de les variables?

(k) (1p) Probeu el chisquare-plot per les dades dels homes. Segueix aquest grup la distribució normal multivariant?

(l) (1p) Instal.leu el paquet `ICSNP` amb `install.packages("ICSNP")`. Compareu els vectors de mitjanes dels dos grups amb una prova de T^2 de Hotelling, assumint igualtat de matrius de covariancies. Existeixen diferencies significatives entre els dos grups? Doneu el valor de l'estadístic T^2i el valor p.

(m) (1p) Per les dades en qüestió, és rellevant sí les matrius de variancies i covariancies són iguals o no? Perque si o no?

(n) (1p) Contrasteu igualtat de mitjanes per cadascuna de les variables per separat, fent una prova de t de student amb `t.test`, sense asumir igualtat de variancies. Per quines variables trobeu diferencies significatives? (utilitzeu $\alpha = 0.001$).

(o) (1p) Feu boxplots de les set variables estratificades per la variable `gender`. Que observeu?