**Exercici 1.- Eliminar els identificadors del conjunt de dades (0.5p)**

1. Quins són els atributs identificadors en el conjunt de dades?

DNI i número de la seguretat social “SS”.

1. Descriu les comandes que has emprat.

*dades\_sense\_identificador<-dades[3:5]*

Bàsicament dades\_sense\_identificar rep la variable dades, però traient la primera y segona columna.

**Exercici 2.- Re-identificació (0.5p)**

1. Quins d’aquests atributs o combinacions poden conduir a la identificació única d’un registre del conjunt de dades?

Es poden donar diferents combinacions per a identificar, a mes a mes al ser unes dades amb tant poques persones es molt més fàcil. Identificar a una persona nomes el CP, perquè existeix una persona amb aquets codi únicament. També la identificació mitjançant la combinació CP i edat, com que les dades que tenim, i per últim, la combinació CP i salari.

1. Indica quines comandes has fet servir per obtenir aquesta informació, explicitant el significat de les comandes i dels seus paràmetres.

*x<- dades\_sense\_identificador [1]*

*y<- dades\_sense\_identificador [2]*

*c<- dades\_sense\_identificador [3]*

*which(x==25560)* #Resultat posició 18, identificat nomes amb CP.

*which(x==17800 & y==32)* #Resultat posició 1, identificat amb CP i edat.

*which(x==43870 & c==78*) #Resultat posició 94, identificat amb CP i salari.

Les primeres instruccions emmagatzemen les dades de CP, edat i salari. Which es una instrucció que troba una/es comparacions, retornant la posició.

**Exercici 3.- Aplicació de soroll additiu (0.5p)**

1. Indica la comanda que has fet servir, junt amb els paràmetres emprats i el significat de cadascun d’ells?

*dades.an=addNoise(dades\_sense\_identificador,variables=’Edat’,noise=20,method=’additive’)*

1. Executa el següent codi adjunt i explica els resultats obtinguts, justifica la resposta.

Es mostra un gràfic, aquest ens mostra molt valors en forma de punts, cada punt esta format per dues dades combinades (x,y). Es a dir, al Original tenim el valor de *dades\_sense\_identificador* columna edat i al Masked el valor de dades.an columna edat però amb el soroll ja afegit.

També tenim una línia vermella que aporta poca informació, per a referenciar bé el valor dels punts.

**Exercici 4.- Aplicació del mètode d’intercanvi de rang additiu (rang swap) (0.5p)**

1. Indica la comanda que has fet servir, junt amb els paràmetres emprats i el significat de cadascun d’ells.

*dades.rs=rankSwap(dades\_sense\_identificador,variables=’Edat’,TopPercent=10, BottomPercent=10, K0=NULL,R0=NULL, missing=NA, seed=NULL)*

Guardem el resultat en la variable dades.rs, la comanda rankSwap s’encarrega de intercanviar els valors dins d’un rang especificat. Aquesta comanda rep un paràmetres, aquells que estan en NULL no fens interessen, el paràmetre *dades\_sense\_identificador* es d’on extraiem les dades, el paràmetre variables ens dius que nomes mirem el la columna edat. Per aplicar el mètode pertorbatiu necessitem un percentatge del 10%, aquest especificat als dos paràmetres TopPercent i BottomPercent,

1. Genera una gràfica similar a la que hem emprat en l’exercici anterior apartat b), on es pugui veure de forma visual i ràpid la dispersió o alteració de l’atribut Edat després d’aplicar el mètode d’intercanvi del rang.

*plot(cbind(dades$Edat,dades.rs$Edat),ylim=c(min(dades$Edat),max(dades$Edat)),xlim=c(min(dades$Edat),max(dades$Edat)),xlab="Original", ylab="Masked", main="Metode Pertorbatiu")*

*abline(a=0, b=1, col=”red”)*

**Exercici 5.- Càlcul de la utilitat (0.5p)**

1. Utilitza la comanda dUtility de la llibreria sdcMicro per a estimar la pèrdua de informació que s’ha produït en generar la versió pertorbada de l’atribut Edat emprant els mètodes de soroll additiu i intercanvi de rang (generats en els dos exercicis anteriors). Indica la comanda que has fet servir, junt amb els paràmetres emprats i el significat de cadascun d’ells.

La comanda utilitzada per estimar la pèrdua d’informació que s’ha produït en generar la versió pertorbada de l’atribut Edat emprant el mètode soroll additiu és:

*dUtility(obj= dades\_sense\_identificador[‘Edat’],xm=dades.an$xm)*

La comanda utilitzada per estimar la pèrdua d’informació que s’ha produït en generar la versió pertorbada de l’atribut Edat emprant el mètode intercanvi de rang és:

*dUtility(obj= dades\_sense\_identificador[‘Edat’],xm=dades.rs[‘Edat’])*

El paràmetre *obj* indica la dada o variable amb els valors original i el paràmetre *xm* indica la dada o variable amb els valors pertorbats, és a dir, els paràmetres ens indiquen la versió original i la versió pertorbada.

1. Comenta els resultats obtinguts i descriu el significat dels valors obtingut i com es calculen.

En el cas de la versió pertorbada emprant el soroll additiu, el valor obtingut ha

sigut 70.65247 que ens indica que hi ha bastant diferència entre els valors perturbats i els valors originals, per tant la pèrdua d’informació i utilitat és bastant elevada.

En el cas de la versió pertorbada emprant intercanvi de rang, el valor obtingut ha sigut 9.444746, que a diferencia del cas anterior, la diferencia entre els valors perturbats i originals es menor i per tant no es perd tanta informació i té més utilitat.

Aquest valors son calculats mesurant les distàncies absolutes entre els valors originals i els perturbats, escalant aquestes distàncies per l’arrel quadrada de 2 de la desviació estándar dels valors originals (el mètode utilitzat si no s’especifica que és IL1s).

**Exercici 6.- Càlcul del nivell de privacitat (0.5p)**

1. Utilitza la comanda dRisk de la llibreria sdcMicro per a estimar el risc o nivell de privacitat en la versió pertorbada de l’atribut Edat, emprant els mètodes de soroll additiu i intercanvi de rang (generats en els dos exercicis anteriors). Indica la comanda que has fet servir, junt amb els paràmetres emprats i el significat de cadascun d’ells.

La comanda utilitzada per estimar el risc o nivell de privacitat que s’ha produït en generar la versió pertorbada de l’atribut Edat emprant el mètode soroll additiu és:

*dRisk(obj=* *dades\_sense\_identificador[‘Edat’],xm=dades.an$xm)*

La comanda utilitzada per estimar per estimar el risc o nivell de privacitat que s’ha produït en generar la versió pertorbada de l’atribut Edat emprant el mètode intercanvi de rang és:

*dRisk(obj=* *dades\_sense\_identificador[‘Edat’],xm=dades.rs[‘Edat’])*

El paràmetre *obj* indica la dada o variable amb els valors original i el paràmetre *xm* indica la dada o variable amb els valors perturbats, és a dir, els paramètres ens indiquen la versió original i la versió pertorbada.

1. Comenta els resultats obtinguts i descriu el significat dels valors obtingut i com es calculen.

En el cas de la versió pertorbada emprant el soroll additiu, el valor obtingut ha sigut 0.19 que ens indica que el nivell de privacitat és baix, per tant el risc és elevat.

En el cas de la versió pertorbada emprant intercanvi de rang, el valor obtingut ha sigut 0.17, que a diferencia del cas anterior, el nivell de privacitat és lleugerament més baix i per tant té més risc, tot i que la diferència és mínima..

Es calcula construint un interval (basat en la desviació estàndard) al volant de cada valor de la variable pertorbada. Aleshores es comprova si els valors originals es troben o no en aquests intervals. Amb el paràmetre k es pot ampliar o reduir el interval, però en el nostre cas no hem utilitzat aquesta opció.

**Exercici 7.- Micro-agregació univariant i multivariant (1p)**

1. Aplica el mètode de micro-agregació univariant amb un nivell d’agregació igual a 3 als atributs Edat i Salari, de forma independent, sobre conjunt de dades original (sense identificadors). Indica les comandes i paràmetres que has fet servir i la comenta la seva funció.

Comandes per atribut edat i salari:

m1\_Edat<-microaggregation(obj=dades\_sense\_identificador,variables = 'Edat', method = “onedims”,aggr=3)

m1\_Salari<-microaggregation(obj=dades\_sense\_identificador,variables='Salari', method = “onedims”,aggr=3)

Els paràmetres utilitzats son, el paràmetre *obj* indica el conjunt de dades original, el paràmetre *variables* indica sobre quin atribut apliquem el mètode i el paràmetre *aggr* indica el nivell d’agregació.

1. Els histogrames són gràfiques que presenten la freqüència d’aparició segons els valors. Per tant, són molt interessants per veure quins són els valors més freqüents i quins són únics. Crea quatre gràfiques que presentin la informació de l’atribut Edat i Salari abans i després del procés de micro-agregació, i que permetin comparar de forma visual i senzilla, com s’han modificat el valors d’aquests dos atributs. Indica el codi que has emprat, detallant la funció dels diferents Nota: la funció par(mfrow=c(2,2)) us pot ajudar a posar les 4 gràfiques en una sola imatge, de manera que simplifica la visualització.

Codi utilitzat:

*par(mfrow=c(2,2))*

*hist(dades\_sense\_identificador$Edat)*

*hist(dades\_sense\_identificador$Salari)*

*hist(m1\_Edat$mx$Edat)*

*hist(m1\_Salari$mx$Salari)*

Lla primera línia ens ajudar a posar les 4 gràfiques en una sola imatge, com podem veure en la figura 1.

La funció de les pròximes 4 línies es generar 4 histogrames diferents. Línia 2 histograma de l’atribut Edat abans, línia 3 histograma de l’atribut Salari abans, línia 4 histograma de l’atribut Edat després i línia 5 histograma de l’atribut Salari després.

Gráfico, Histograma, Gráfico de cajas y bigotes

Descripción generada automáticamente

Figura 1. Visualització dels 4 histogrames diferents. A la primera columna podem comparar l’atribut Edat i a la segona columna podem comparar l’atribut Salari.

1. Aplica el mètode de micro-agregació multivariant MDAV amb un nivell d’agregació igual a 3 als atributs Edat i Salari sobre conjunt de dades original (sense identificadors). Indica les comandes i paràmetres que has fet servir i la comenta la seva funció.

Comanda utilitzada:

*m1\_Edat\_Salari<-microaggregation(obj= dades\_sense\_identificador,variables=c("Edat","Salari"),method = “mdav”,aggr=3)*

Els paràmetres utilitzats son, el paràmetre *obj* indica el conjunt de dades original, el paràmetre *variables* indica sobre quins atribut apliquem el mètode, el paràmetre *aggr* indica el nivell d’agregació i el paràmetre *method* indica quin mètode de micro-agregació multivariant utilitzem.

1. Crea les gràfiques que presentin la informació dels atributs Edat i Salari anonimitzats en la seva versió multivariant, i compara els resultats obtinguts amb els obtinguts en el cas univariant.

Gráfico, Histograma, Gráfico de cajas y bigotes

Descripción generada automáticamente

Figura 2. Visualització dels 4 histogrames diferents. A la primera columna podem comparar l’atribut Edat en el cas univariant (primera fila) amb el cas multivariant (segona fila) i a la segona columna podem comparar l’atribut Salari en el cas univariant (primera fila) amb el cas multivariant (segona fila).

1. Realitza una estimació del risc i de la utilitat en els dos casos (univariant i multivariant). Comenta i justifica els resultats obtinguts.

Estimació de la utilitat i del risc:

-Cas univariant:

-Atribut Edat:

-Comanda Utilitat i Risc:

*dUtility(obj=dades\_sense\_identificador['Edat'],xm=m1\_Edat$mx['Edat'])*

*dRisk(obj= dades\_sense\_identificador ['Edat'],xm=m1\_Edat$mx['Edat'])*

-Resulta Utilitat: 1.760311. No molta diferència entre els valors perturbats i els valors originals, per tant la pèrdua d’informació es bastant lleu i té una bona utilitat .

-Resultat Risc: 0.89. El nivell de privacitat és elevat, per tant el risc és poc..

-Atribut Salari

-Comanda Utilitat i Risc:

*dUtility(obj= dades\_sense\_identificador ['Salari'],xm=m1\_Salari$mx['Salari'])*

*dRisk(obj= dades\_sense\_identificador ['Salari'],xm=m1\_Salari$mx['Salari'])*

-Resultat Utilitat: 2.03656. No molta diferència entre els valors perturbats i els valors originals, per tant la pèrdua d’informació es bastant lleu i té una bona utilitat.

-Resultat Risc: 0.84. El nivell de privacitat és elevat, per tant el risc és poc..

-Cas multivariant:

-Atribut Edat:

-Comanda Utilitat i Risk:

*dUtility(obj= dades\_sense\_identificador ['Edat'],xm=m1\_Edat\_Salari$mx['Edat'])*

*dRisk(obj= dades\_sense\_identificador ['Edat'],xm=m1\_Edat\_Salari$mx['Edat'])*

-Resultat Utilitat: 10.35875. Bastant diferència entre els valors perturbats i els valors originals, més que en el cas univariant, per tant la pèrdua d’informació és major i perd utilitat.

-Resultat Risc: 0.29. El nivell de privacitat és menor que en el cas univariant, per tant, té més risc.

-Atribut Salari:

-Comanda Utilitat:

*dUtility(obj=dades\_sense\_identificador ['Salari'],xm=m1\_Edat\_Salari$mx['Salari'])*

*dRisk(obj=dades\_sense\_identificador ['Salari'],xm=m1\_Edat\_Salari$mx['Salari'])*

-Resultat Utilitat: 9.754602. Bastant diferència entre els valors perturbats i els valors originals, més que en el cas univariant, per tant la pèrdua d’informació és major i per utilitat.

-Resultat Risc: 0.21. El nivell de privacitat és menor que en el cas univariant, per tant, té més risc.

**Exercici 8.- Generalització d’atributs (1p)**

1. Crea una funció pròpia que permeti generalitzar els codis postals de població, convertint-los en codis de província. Detalle el codi de la funció, així com alguns exemples del seu funcionament i valors de retorn.

Funció:

*generalitzacio\_CP <-function(llista\_CP){*

*for(i in llista\_CP){*

*aux = substr(i,1,2)}*

*nova\_llista = paste(aux,"\*\*\*",sep="")*

*return(nova\_llista)*

*}*

En la funció, el procés de generalitzar es troba en la línia 3. El que fem es que extraiem només els dos primers números del codi postal (aquest són el codi de província). Realitzem un bucle for per aplicar aquest procés a tots els codis de la llista, i finalment, a la llista resultant del bucle l’hi afegim els caràcters “\*\*\*” després de cada codi postal. La funció retorna aquesta nova llista creada.

Exemples:

*c <- data.frame("CP"=c(17820,08500,17100,43870))*

*c1 <- data.frame("CP"=c(08240,08500,08018))*

*generalitzacio\_CP(c)*

*generalitzacio\_CP(c1)*

Valors de retorn:

[1]"17\*\*\*" "08\*\*\*" "17\*\*\*" "43\*\*\*"

[1]"08\*\*\*" "08\*\*\*" "08\*\*\*"

1. Aplica la funció que has creat a l’atribut CP del conjunt de dades, per a tots els individus (registres) existents. Comenta el resultat obtingut.

Comanda:

*generalitzacio\_CP(dades\_sense\_identificador['CP'])*

Resultat:

[1] "17\*\*\*" "25\*\*\*" "17\*\*\*" "08\*\*\*" "43\*\*\*" "25\*\*\*" "17\*\*\*" "43\*\*\*"

[9] "17\*\*\*" "43\*\*\*" "43\*\*\*" "08\*\*\*" "08\*\*\*" "17\*\*\*" "17\*\*\*" "08\*\*\*"

[17] "17\*\*\*" "25\*\*\*" "25\*\*\*" "25\*\*\*" "17\*\*\*" "08\*\*\*" "25\*\*\*" "25\*\*\*"

[25] "43\*\*\*" "43\*\*\*" "08\*\*\*" "17\*\*\*" "17\*\*\*" "25\*\*\*" "17\*\*\*" "08\*\*\*"

[33] "08\*\*\*" "08\*\*\*" "25\*\*\*" "25\*\*\*" "08\*\*\*" "08\*\*\*" "43\*\*\*" "17\*\*\*"

[41] "17\*\*\*" "25\*\*\*" "08\*\*\*" "17\*\*\*" "25\*\*\*" "43\*\*\*" "17\*\*\*" "17\*\*\*"

[49] "43\*\*\*" "43\*\*\*" "43\*\*\*" "25\*\*\*" "43\*\*\*" "17\*\*\*" "08\*\*\*" "25\*\*\*"

[57] "17\*\*\*" "08\*\*\*" "08\*\*\*" "25\*\*\*" "43\*\*\*" "17\*\*\*" "17\*\*\*" "08\*\*\*"

[65] "43\*\*\*" "17\*\*\*" "08\*\*\*" "43\*\*\*" "17\*\*\*" "43\*\*\*" "08\*\*\*" "25\*\*\*"

[73] "25\*\*\*" "43\*\*\*" "08\*\*\*" "08\*\*\*" "17\*\*\*" "43\*\*\*" "08\*\*\*" "25\*\*\*"

[81] "08\*\*\*" "08\*\*\*" "17\*\*\*" "43\*\*\*" "25\*\*\*" "08\*\*\*" "43\*\*\*" "25\*\*\*"

[89] "25\*\*\*" "17\*\*\*" "17\*\*\*" "43\*\*\*" "43\*\*\*" "43\*\*\*" "08\*\*\*" "43\*\*\*"

[97] "43\*\*\*" "25\*\*\*" "08\*\*\*" "17\*\*\*"

El resultat obtingut és que només tenim 4 tipus diferents de codis postals (17\*\*\*, 25\*\*\*,08\*\*\*,43\*\*\*) que corresponen a les províncies de Girona, LLeida, Barcelona i Tarragona. Dades completament correctes però més generals, fent que les individualitats de cada usuari del conjunt de dades originals quedin difuminades entre els altres usuaris en el conjunt de dades generalitzat.

Tenim 27 individus de Girona, 22 de Lleida, 26 de Barcelona i 25 de Tarragona.

Aproximadament tenim el mateix número d’individus en cada província, per tant, identificar un individu per l’atribut CP és molt més complicat si utilitzem la generalització d’aquest.

1. Aplica la funció que has creat a l’atribut CP del conjunt de dades, però només a aquells registres que estiguin en perill de ser re-identificats, és a dir, que la seva freqüència d’aparició sigui igual o inferior a 3 (≤ 3) en tot el conjunt de dades. Comenta el resultat obtingut i compara-ho amb els resultats de l’apartat anterior.

Comandes:

*table(dades\_sense\_identificador['CP'])*

Amb aquesta comanda podem observar que els registres amb freqüència d’aparició igual o inferior a 3 són el 08401, 25600 i 25610.

*registres\_perill = dades\_sense\_identificador[dades\_sense\_identificador $CP %in% c("08401", "25600","25610"), ]*

Amb aquesta comanda extraiem els registres del conjunt de dades amb freqüència d’aparició igual o inferior a 3.

*generalitzacio\_CP(registres\_perill['CP'])*

Amb aquesta comanda apliquem la funció a l’atribut CP.

Resultat:

[1] "08\*\*\*" "08\*\*\*" "25\*\*\*" "25\*\*\*" "08\*\*\*"

Tenim 2 tipus diferents de codis postals (08\*\*\* i 25\*\*\*), a diferència de l’apartat anterior, que corresponen a les províncies de Barcelona i Lleida. Només tenim 5 individus amb el registre CP amb perill de ser re-identificat, 3 individus de Barcelona i 2 de Lleida. El resultat que aconseguim és que aquests individus ja no estiguin en perill segons l'atribut CP gràcies a la generalització, com en l'apartat anterior, amb la diferència de que estaran en més perill degut a que només tenim 2 províncies en comptes de 4.

**Exercici 9.- Alternatives per mesurar la pèrdua de informació (1p)**

a. Mostra la gràfica obtinguda.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

b. Comenta i justifica els resultats obtinguts.

Podem veure una comparativa dels 4 mètodes d’emmascarament i la diferencia d’aquest amb les dades originals que s’han aplicat.

El mètode de soroll additiu, en general té més errors, ja que introdueix soroll. Podem veure que el seu error va creixent.

El mètode d’intercanvi de rang, no creix uniformement, però també introdueix molt error.

El millor mètode amb aquestes dades es la micro-agregació univariant, podem veure que el seu error creix lentament, i es diferencia bastant de les altres mostres.

El mètode de micro-aggregació multivariant, presenta uns valors molts semblant amb el de soroll additiu, però en alguns registres te menys error que aquest.