Introducció a la neteja i anàlisi de dades

PID_00265701

Mireia Calvo González
Diego Oswaldo Pérez Trenard
Laia Subirats Maté

Temps mínim de dedicació recomanat: 6 hores

Universitat Oberta de Catalunya

Mireia Calvo González

Enginyera de telecomunicacions per la Universitat Politècnica de Catalunya (2011), Màster en Enginyeria Biomèdica per la Universitat de Barcelona i per la Universitat Politècnica de Catalunya (2014) i Doctora en Processament de senyals i telecomunicacions per la Universitat de Rennes 1 i en Enginyeria Biomèdica per la Universitat Politècnica de Catalunya (2017). Des del 2012 ha treballat com a investigadora en diferents entorns acadèmics, clínics i industrials, aplicant el processament de dades a l'estudi de diferents malalties cardíaques i respiratòries. Des del 2017 col·labora amb la UOC com a docent en el Màster de Data Science.

Diego Oswaldo Pérez Trenard

Enginyer electrònic per la Universitat Simón Bolívar (2015), especialització en High Tech Imaging (HTI) per la Universitat Télécom SudParis (2014) i doctor en Senyals, Imatges i Visió per la Universitat de Rennes 1 (2018). Des del 2014, ha treballat com a enginyer de recerca i desenvolupament a l'Institut Nacional de Salut i Investigació Mèdica (INSERM) i al Laboratori de Processament de Senyals i Imatges (LTSI), aplicant coneixements en electrònica i en processament de dades a l'estudi de diferents malalties neurològiques, cardíaques i respiratòries. Des del 2018, col·labora com a docent en el màster de Data Science de la UOC.

Laia Subirats Maté

Enginyera de Telecomunicacions per la Universitat Pompeu Fabra (2008), màster en Telemàtica per la Universitat Politècnica de Catalunya (2009) i doctora en Informàtica per la Universitat Autònoma de Barcelona (2015). Des de 2009, treballa com a investigadora a Eurecat (Centre Tecnològic de Catalunya) aplicant la ciència de dades a diferents camps com ara la salut, el medi ambient o l'educació. Des de 2016, col·labora amb la UOC com a docent en el màster de Data Science i en el grau d'Informàtica. És especialista en intel·ligència artificial, ciència de dades, salut digital i representació del coneixement.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats per la professora: Isabel Guitart Hormigo (2019)

Primera edició: setembre 2019

Av. Tibidabo, 39-43, 08035 Barcelona
Realització editorial: FUOC

Índex

Introducció

En l'actualitat, grans quantitats de dades són emmagatzemades diàriament, de manera que l'aplicació de mètodes robustos que permetin analitzar i extreure informació, i posteriorment coneixement d'aquestes dades és de summa importància. Cada vegada que comprem per Internet, per exemple, generem una sèrie de dades relacionades amb el procés de compra que, després de ser netejades i analitzades, es converteixen en informació útil per als propietaris dels comerços electrònics. Gràcies al coneixement extret d'observar tendències i patrons de comportament en els diferents tipus de clients, es poden identificar els seus gustos per així millorar la seva experiència de comprar i augmentar, en última instància, el nombre de vendes.

Cal destacar que, per dotar de robustesa les anàlisis aplicades amb l'objectiu de generar coneixement, és clau la qualitat de les dades analitzades.

La **neteja de dades**, o *data cleaning* en anglès, és el conjunt de processos que permeten corregir o eliminar aquelles mostres errònies d'una base de dades. Aquests processos permeten identificar dades incompletes, incorrectes, inexactes o no pertinents, a fi d'eliminar-les o corregir-les, i així, obtenir bases de dades de més qualitat.

Així mateix, verificar que les dades compleixen les suposicions requerides per les proves estadístiques aplicades és fonamental. Les anàlisis estadístiques més comunament utilitzades, com la correlació de Pearson o la prova *t* de Student, assumeixen certes característiques o suposicions sobre les dades que s'han de complir perquè aquestes tècniques, i per tant les conclusions extretes, siguin vàlides. Algunes de les suposicions més habituals són el fet que les dades es trobin distribuïdes normalment, així com que els grups de dades presentin variàncies semblants.

Tot i que la robustesa dels mètodes d'anàlisi es mesuren generalment amb la taxa d'error de Tipus I, el *data cleaning* també pot afectar de manera significativa la potència estadística, la mida de l'efecte i l'exactitud a l'hora d'aplicar aquests mètodes, i per tant la seva replicabilitat, així com la minimització de la taxa d'error de Tipus II. Les dades errònies no només poden conduir a la violació de suposicions en les dades, com la normalitat o l'homogeneïtat de la variància (homoscedasticitat), sinó que també poden dur a l'estimació errònia de paràmetres i efectes sense causar una desviació significativa d'aquestes suposicions. Tractar amb eficàcia els valors extrems (*outliers*) d'una mostra ge-

neralment millorarà la potència estadística i la mida de l'efecte i disminuirà els errors de Tipus I i de Tipus II, per la qual cosa tendirà a millorar el resultat de les anàlisis i estimacions.

No obstant això, aquesta important etapa de preprocessat no sempre rep l'atenció necessària. En un treball de revisió bibliogràfica realitzat sobre els articles publicats durant el 2009 a les revistes científiques de l'American Psychological Association (APA, per les seves sigles en anglès), es va reportar que només entre el 22 % i el 38 % dels treballs feien referència a algun procés de *data cleaning*. Entre el 16 % i el 18 % dels estudis van reportar anàlisis de valors extrems (*outliers*), entre el 10 % i el 32 % van verificar la distribució de les dades, i entre el 32 % i el 45 % van reportar l'aplicació d'algun mètode per gestionar les dades perdudes. Això no s'ha d'interpretar com que menys de la meitat dels treballs van realitzar algun tipus de neteja en les dades analitzades, però atès el seu impacte en els resultats, resulta sorprenent que aquests processos de neteja no es descriguin sempre en detall, tal com es fa amb els mètodes d'anàlisi aplicats. Per exemple, davant d'una base de dades que contingui un nombre significatiu de dades perdudes, serà rellevant descriure si aquestes dades es van eliminar o es van imputar i, si fos així, mitjançant quin mètode, ja que això modificarà els resultats i, per tant, les conclusions extretes d'aquests resultats.

Font bibliogràfica

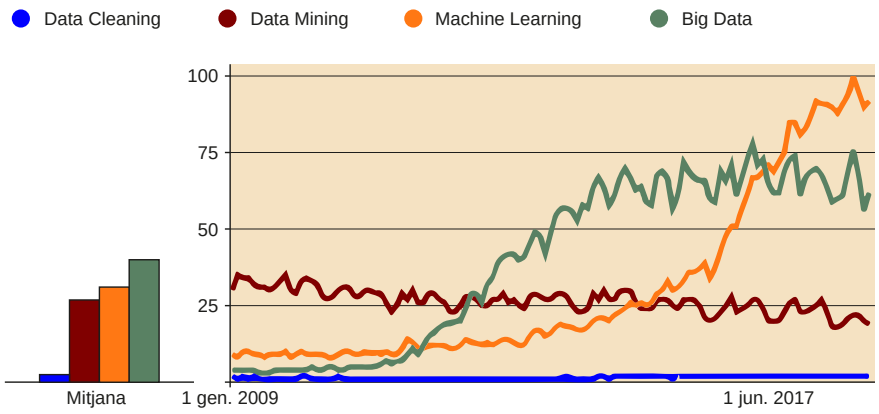
Osborne, Jason W.; Kocher, Brady, Tillman, David (2012). «Sweating the small stuff: do authors in APA journals clean data or test assumptions (and should anyone care if they do)?» [conferència]. En: *Annual meeting of the Eastern Education Research Association* (2012: Hilton Head, SC).

De fet, s'estima que el 80 % de la feina d'un científic de dades és invertida en processos de neteja. Això no obstant, la figura 1, extreta de Google Trends, mostra com el *data cleaning* no ha aconseguit capturar la mateixa atenció que el *big data*, el *data mining* o el *machine learning* durant l'última dècada.

Bibliografia recomanada

Squire, Megan (2015). *Clean Data*. Birmingham: Packt Publishing.

Figura 1. Popularitat de termes relacionats amb les dades, entre gener de 2009 i 2019



Font: Google Trends.

La **mineria** o **exploració de les dades** (*data mining*, en anglès) és l'anàlisi automàtica o semiautomàtica de les dades, amb l'objectiu de descobrir informació útil de manera eficaç, escalable i flexible.

En funció del context, aquesta anàlisi pot perseguir diferents objectius: identificar grups de registres de dades (*clustering*), detectar registres poc usuals (anomalies) o identificar dependències (mineria per regles d'associació), entre d'altres. Tot i que tradicionalment es realitzava mitjançant mètodes estadístics, actualment aquests mètodes es combinen amb altres tècniques procedents de la intel·ligència artificial, l'aprenentatge automàtic i els sistemes de bases de dades, que cada vegada gaudeixen de més popularitat.

Tot procés de descobriment de coneixement a partir d'una sèrie de dades ha d'incloure'n, per tant, la neteja, així com la seva anàlisi. Per això, després de tractar la captura i l'emmagatzematge de les dades (veure mòdul de *web scraping*), aquest mòdul didàctic se centra en els processos de neteja i anàlisi de les bases de dades emmagatzemades.

L'apartat "Neteja de dades" detalla les etapes de neteja més habituals; passant per la integració, selecció i reducció de les dades, possibles transformacions com la discretització, així com la gestió de dades perdudes (*missing data*) i de valors extrems (*outliers*).

A continuació, a l'apartat "Anàlisi de dades", s'enumeren les principals tècniques d'anàlisi que permeten explorar les dades, amb l'objectiu d'identificar tendències i patrons d'interès. Des de l'anàlisi estadística descriptiva i inferencial, passant per la regressió, la correlació, les anàlisis de supervivència i enumerant alguns models supervisats i no supervisats. Així mateix, es comenta breument la representació visual dels resultats com a mètode addicional d'anàlisi.

Finalment, després d'un breu resum dels continguts més rellevants del mòdul, es proposen alguns exercicis d'autoavaluació, així com les seves solucions, amb els quals poder comprovar l'assimilació dels conceptes principals que es presenten aquí.

Aquest mòdul didàctic s'acompanya d'un repositori GitHub on s'inclou el codi descarregable d'alguns dels exemples presentats. Tot i que aquests exemples es proporcionen en R, els processos de neteja i anàlisi descrits en aquest material poden implementar-se en altres llenguatges de programació com Python.

Tots els exemples estan basats en conjunts de dades disponibles a R. No obstant això, molts dels mètodes utilitzats es basen en tècniques *randomitzades*, per la qual cosa, com que no fixen llavors específiques en cada un dels exemples, els resultats mostrats en aquest material no tenen per què coincidir sempre amb els que poden trobar.

Així mateix, els exemples proposats il·lustren algunes de les funcionalitats dels paquets presentats, però es recomana consultar l'ajuda i documentació de cadascuna de les funcions proporcionades, per aprofundir en les seves funcionalitats i poder treure així el màxim partit.

Objectius

En aquest material didàctic es proporcionen les eines fonamentals que permetran assimilar els objectius següents:

1. Comprendre el significat i els beneficis potencials de la neteja de dades.

2. Conèixer la dificultat de netejar una base de dades determinada.

3. Conèixer els principals mètodes per a la neteja de dades.

4. Conèixer les principals tècniques d'exploració de les dades.

5. Saber aplicar processos de neteja, validació i anàlisi utilitzant R.

6. Poder extreure informació útil de les bases de dades disponibles.

1. Neteja de dades

En el cicle de vida de les dades, l'etapa de neteja (o preprocessat) es compon del conjunt de processos que permeten identificar aquells registres incomplets, incorrectes, inexactes o no pertinents d'un conjunt de dades, per tal d'eliminar-les o corregir-les. Aquesta etapa permet millorar la qualitat de les dades, per la qual cosa serà extremadament important a l'hora de treure el màxim rendiment a la seva anàlisi posterior.

Els apartats següents descriuen alguns dels mètodes de neteja més utilitzats. Tanmateix, és important destacar que aquests s'aplicaran en funció del context, és a dir, del tipus de dades que es volen tractar i de l'anàlisi que es vulgui realitzar posteriorment. Així, tot i que s'enumeren diversos mètodes, no sempre serà necessari aplicar-los tots, de la mateixa manera que podran afegir-se altres mètodes al procés de neteja.

1.1. Integració

La **integració** o **fusió de les dades** consisteix en la combinació de dades procedents de múltiples fonts, per tal de crear una estructura de dades coherent i única que contingui més quantitat d'informació.

Bibliografia recomanada

Han, Jiawei; Kamber, Micheline; Pei, Jian (2011). *Data mining: Concepts and Techniques*. Waltham: Elsevier.

Aquesta fusió es pot realitzar de manera horitzontal, és a dir, afegint nous atributs a la base de dades original. Atès que les diferents fonts no sempre tindran el mateix nombre de registres i aquests no estaran ordenats seguint el mateix criteri, abans de la integració serà fonamental identificar un atribut que serveixi d'«identificador únic» i, per tant, que relacioni adequadament els nous atributs amb els registres existents. Amb l'objectiu d'evitar inconsistències i redundàncies, és important destacar que aquest identificador podrà tenir diferents noms i formats en cadascuna de les fonts. Si utilitzéssim, per exemple, el nom dels diferents clients com a identificador, aquest es podria escriure de diferents maneres en cadascuna de les bases de dades (només amb el primer cognom, amb o sense el nom, especificant el títol, etc.). També podria passar que diferents clients tinguessin el mateix nom. Per això, per evitar duplicitats, és habitual utilitzar identificadors numèrics sintètics que identifiquin unívocament cada un dels registres; en l'exemple, cadascun dels clients.

En l'exemple següent es fusiona la informació d'un grup d'autors amb la dels llibres que han escrit, agafant els atributs «name» i «surname» com a identificadors. En aquest cas, no podem distingir si les dues entrades de Ripley fan referència al mateix autor.

Exemple

```
>authors <- data.frame( surname = c("Tukey", "Venables", "Tierney", "Ripley",
"McNeil"), nationality = c("US", "Australia", "US", "UK", "Australia"),
retired = c("yes", rep("no", 4)))

>books <- data.frame( name = c("Tukey", "Venables", "Tierney", "Ripley",
"Ripley", "McNeil"), title = c("Exploratory Data Analysis", "Modern Applied
Statistics ...", "LISP-STAT", "Spatial Statistics", "Stochastic Simulation",
"Interactive Data Analysis"), other.author = c(NA, "Ripley", NA, NA, NA, NA))

>example1<-merge(authors, books, by.x="surname", by.y="name")
>example1

  name                          title  other.author nationality   retired
1   McNeil  Interactive Data Analysis          <NA>   Australia        no
2   Ripley         Spatial Statistics          <NA>          UK        no
3   Ripley      Stochastic Simulation          <NA>          UK        no
4  Tierney                  LISP-STAT          <NA>          US        no
5    Tukey  Exploratory Data Analysis          <NA>          US       yes
6 Venables Modern Applied Statistics ...    Ripley    Australia        no
```

D'altra banda, es poden realitzar fusions verticals amb l'objectiu d'incloure nous registres a una base de dades original. Per exemple, quan l'inventari de les diferents botigues d'una cadena es reculli en bases de dades separades, aquestes poden voler-se integrar per analitzar-les conjuntament. El següent exemple mostra la fusió vertical de dues bases de dades, mitjançant la funció `rbind`():

Exemple

```
>data1 <- data.frame(CustomerId = c(1:6), Product = c(rep("Oven", 3),
rep("Television", 3)))

>data2 <- data.frame(CustomerId = c(4:7), Product = c(rep("Television", 2),
rep("Air conditioner", 2)))

>example2<-rbind(data1,data2)
example2

   CustomerId     Product
1         1           Oven
2         2           Oven
3         3           Oven
4         4     Television
5         5     Television
6         6     Television
7         4     Television
8         5     Television
9         6 Air conditioner
10        7 Air conditioner
```

En aquest tipus de fusió serà molt important que el format de les bases de dades a integrar sigui el mateix; en cas contrari, apareixeran inconsistències i errors en l'estructura de dades final. Així, es pot donar el cas que en una de les botigues no sigui procedent recollir un dels atributs perquè es tracti d'un servei que només s'ofereixi en altres botigues de la cadena. Si aquest atribut només apareix en alguna de les bases de dades d'origen, caldrà afegir-lo també

en aquelles botigues on no sigui procedent recollir-lo, per després deixar el camp buit. D'aquesta manera, es mantindrà el mateix format en totes les bases que es volen fusionar.

Així mateix, la informació de les diferents botigues es pot recollir en diferents formats. Per exemple, si aquestes botigues es troben a Europa i els Estats Units, els atributs que indiquin mesures o preus es recolliran molt probablement en unitats diferents, per la qual cosa caldrà aplicar un procés previ de conversió de manera que tots els registres es representin en un mateix sistema de mesura i en una mateixa moneda.

Finalment, un cop integrades les dades, sempre serà necessari verificar que tant la fusió com les conversions prèvies s'han realitzat correctament, així com que no hi ha elements duplicats a la base de dades nova. R proporciona les funcions `duplicated()` i `unique()` per identificar els registres duplicats. Així mateix, la funció `distinct()` del paquet `dplyr` permet analitzar aquesta duplicitat acotant la recerca només a aquells atributs especificats. Aquesta eina és particularment útil quan algun dels atributs ha de mostrar forçosament valors únics per a cada registre.

Exemple

```
>unique(example1)

  name                      title   other.author nationality    retired
1   McNeil  Interactive Data Analysis      <NA>    Australia         no
2   Ripley         Spatial Statistics      <NA>           UK         no
3   Ripley      Stochastic Simulation      <NA>           UK         no
4  Tierney                  LISP-STAT      <NA>           US         no
5    Tukey  Exploratory Data Analysis      <NA>           US        yes
6 Venables Modern Applied Statistics ...  Ripley   Australia         no

>example1 %>% distinct(surname, .keep_all=TRUE)

  name                      title   other.author nationality   retired
1   McNeil  Interactive Data Analysis      <NA>    Australia        no
2   Ripley         Spatial Statistics      <NA>           UK        no
3  Tierney                  LISP-STAT      <NA>           US        no
4    Tukey  Exploratory Data Analysis      <NA>           US       yes
5 Venables Modern Applied Statistics ...  Ripley   Australia        no
```

1.2. Selecció

Una de les primeres etapes en el preprocessat de les dades és el **filtratge** o **selecció de dades** d'interès. Per a un estudi en particular ens pot interessar analitzar només aquelles persones majors de 50 anys, o només la mostra procedent d'un municipi concret.

En aquesta fase també és habitual realitzar una exploració de les dades (*screening*, en anglès), amb l'objectiu d'analitzar globalment les seves característiques i identificar fortes correlacions entre atributs, de manera que es pugui

prescindir d'aquella informació més redundant. Així mateix, se sol aprofitar aquest primer *screening* per analitzar les suposicions en les dades requerides per les proves estadístiques que s'aplicaran posteriorment.

D'altra banda, i contràriament a la selecció, el preprocessat de les dades també pot incloure la **creació de noves variables** a partir de l'extracció de característiques de les dades originals. Per exemple, una variable comunament utilitzada en medicina és l'índex de massa corporal (IMC), que relaciona l'alçada i el pes d'un individu segons la fórmula següent:

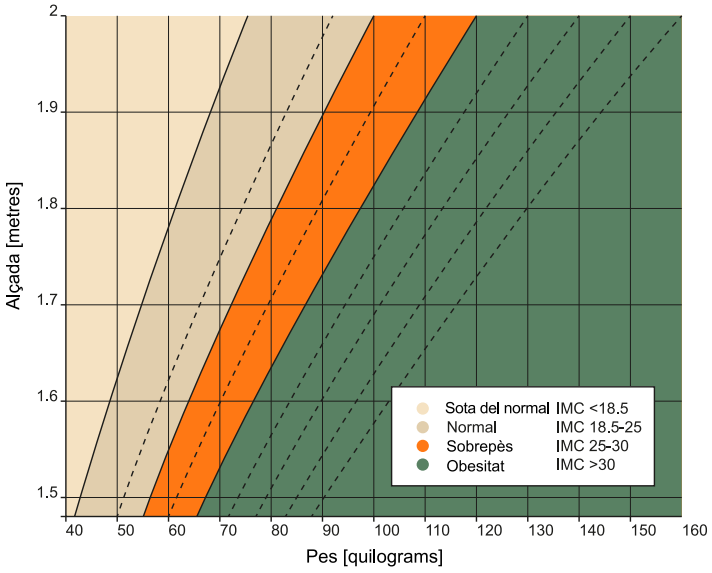
Bibliografia recomanada

Han, Jiawei; Kamber, Micheline; Pei, Jian (2011). *Data Mining: Concepts and Techniques*. Waltham: Elsevier.

$$IMC = \frac{massa}{alçada^2} \qquad (1)$$

Aquest paràmetre s'utilitza com a indicador de greix corporal i per tant d'obesitat, factor de risc en nombroses malalties com la diabetis, la hipertensió, el càncer o l'apnea del son, entre d'altres. La figura 2 mostra la relació entre l'IMC i els diferents graus d'obesitat.

Figura 2. Índex de massa corporal (IMC)



Font: *Viquipèdia*.

1.3. Reducció

Treballar amb grans quantitats de dades pot convertir la tasca d'anàlisi en un procés molt complex i fins i tot impracticable. Les tècniques de **reducció de dades** permeten obtenir una representació reduïda d'aquestes dades, mantenint la integritat de la mostra original. Així, les anàlisis aplicades sobre la mostra de dades reduïda produiran els mateixos resultats (o molt semblants) que si s'apliquessin sobre la mostra total.

Els apartats següents descriuen detalladament algunes de les tècniques de reducció més comunament utilitzades, agrupant-les en dos grups principals: reducció de la dimensionalitat i reducció de la quantitat.

Alguns autors, com Jiawei *et al.* (2011), a més, afegeixen la compressió de les dades com un grup de mètodes dissenyats per reduir-los. Com el seu nom indica, aquests mètodes obtenen una representació comprimida de les dades originals que, quan són capaços de reconstruir la mostra original completament, es denominen mètodes sense pèrdues (*lossless*). En tractament d'imatges, hi ha diversos formats d'emmagatzematge que poden comprimir aquestes imatges, amb o sense pèrdues. El format PNG, per exemple, permet comprimir una imatge de manera totalment reversible, per la qual cosa la imatge recuperada és idèntica a l'original. D'altra banda, JPG és el mètode de compressió més utilitzat en fotografia, ja que després d'eliminar la informació menys apreciable, permet obtenir importants compressions, mantenint qualitats d'imatge molt elevades.

1.3.1. Reducció de la dimensionalitat

Aquest conjunt de mètodes té per objectiu reduir el nombre d'atributs sota consideració. Es poden dividir en mètodes paramètrics i no paramètrics. Els primers estimen les dades mitjançant un model, de manera que només cal emmagatzemar els paràmetres d'aquest model en lloc de la base de dades original. Alguns exemples d'aquest tipus són els models de regressió, que s'estudiaran a l'apartat 2.2.3. Regressió.

Dins dels no paramètrics, entre els més utilitzats hi ha l'anàlisi de components principals i les transformacions *wavelet* que projecten les dades originals en un espai de dimensions més reduït. D'altra banda, en la selecció de subconjunts d'atributs (*Attribute subset selection*), es detecten i s'eliminen aquells atributs més irrellevants o redundants del conjunt de dades. A Jiawei *et al.* (2011), es detalla més en profunditat cadascun d'aquests mètodes.

Sense entrar en detalls en l'explicació teòrica de l'**anàlisi de components principals** (ACP), aquesta tècnica permet descriure un conjunt de dades de n atributs, en termes de m noves variables no correlacionades, o components principals, on $m < n$. Aquests components s'ordenen segons la quantitat de variància de les dades originals que descriuen. Tècnicament, l'ACP és una transformació lineal que busca la projecció segons la qual les dades quedin més ben representades en termes de mínims quadrats, convertint un conjunt d'observacions d'atributs possiblement correlacionats en un conjunt de noves variables sense correlació lineal, anomenades *components principals*.

L'exemple següent es basa en el conjunt de dades `mtcars`, disponible a R, que comprèn les dades d'11 atributs, per 32 models de cotxe. Atès que l'ACP funciona principalment amb dades numèriques, s'exclouen les variables categòriques de la mostra, deixant 9 atributs per als 32 models. A continuació, s'aplica

l'ACP mitjançant la funció `prcomp()`, centrant i escalant les dades per a què el fet de combinar valors a diferents escales no afecti significativament els resultats.

Exemple

```
>mtcars.pca <- prcomp(mtcars[,c(1:7,10,11)], center = TRUE, scale = TRUE)
>summary(mtcars.pca)

Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7    PC8     PC9
Standard deviation     2.3782 1.4429 0.71008 0.51481 0.42797 0.35184 0.32413 0.2419 0.14896
Proportion of Variance 0.6284 0.2313 0.05602 0.02945 0.02035 0.01375 0.01167 0.0065 0.00247
Cumulative Proportion  0.6284 0.8598 0.91581 0.94525 0.96560 0.97936 0.99103 0.9975 1.00000
```

El resultat són 9 components principals (PC1-PC9), cadascun dels quals explica un percentatge de variància del *dataset* original. Així, la primera component principal explica pràcticament 2/3 de la variància total i les dues primeres components descriuen el 86 % de la variància. Atès que les quatre primeres components ja expliquen el 95 % de la variància, es podria treballar només amb aquest subconjunt (PC1-PC4), que conté pràcticament la totalitat de la informació continguda en el conjunt de dades original. A més de la solució obtinguda per a cada registre en el nou espai de components principals (`mtcars.pca$x`), entre altres informacions, el resultat d'aplicar aquesta funció permet analitzar els pesos associats a cada atribut en la transformació lineal aplicada per l'ACP resultant.

Exemple

```
> mtcars.pca$rotation
            PC1         PC2         PC3          PC4         PC5         PC6         PC7         PC8          PC9
mpg  -0.3931477  0.02753861 -0.22119309 -0.006126378 -0.3207620  0.72015586 -0.38138068 -0.12465987  0.11492862
cyl   0.4025537  0.01570975 -0.25231615  0.040700251  0.1171397  0.22432550 -0.15893251  0.81032177  0.16266295
disp  0.3973528 -0.08888469 -0.07825139  0.339493732 -0.4867849 -0.01967516 -0.18233095 -0.06416707 -0.66190812
Hp    0.3670814  0.26941371 -0.01721159  0.068300993 -0.2947317  0.35394225  0.69620751 -0.16573993  0.25177306
drat -0.3118165  0.34165268  0.14995507  0.845658485  0.1619259 -0.01536794  0.04767957  0.13505066  0.03809096
Wt    0.3734771 -0.17194306  0.45373418  0.191260029 -0.1874822 -0.08377237 -0.42777608 -0.19839375  0.56918844
qsec -0.2243508 -0.48404435  0.62812782 -0.030329127 -0.1482495  0.25752940  0.27622581  0.35613350 -0.16873731
gear -0.2094749  0.55078264  0.20658376 -0.282381831 -0.5624860 -0.32298239 -0.08555707  0.31636479  0.04719694
carb  0.2445807  0.48431310  0.46412069 -0.214492216  0.3997820  0.35706914 -0.20604210 -0.10832772 -0.32045892
```

1.3.2. Reducció de la quantitat

Aquest grup de mètodes substitueix el volum de dades original per una representació alternativa amb un volum de mostres o registres menor. Alguns exemples d'aquest tipus són els histogrames, el *clustering* o el *sampling*.

Bibliografia recomanada

Han, Jiawei; Kamber, Micheline; Pei, Jian (2011). *Data Mining: Concepts and Techniques*. Waltham: Elsevier.

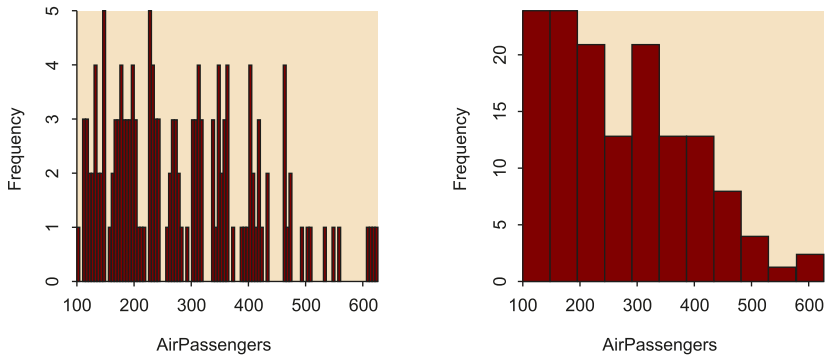
Els **histogrames** divideixen la distribució de dades d'un atribut en subconjunts dissociats, denominats contenidors (*bins*). Tot i que aquests *bins* poden representar un únic parell d'atribut-valor/freqüència quan representen rangs continus per a un atribut donat, també permeten reduir la quantitat de dades.

L'exemple següent mostra l'histograma d'un mateix atribut (`AirPassengers`, disponible a R), quan es redueix la quantitat de *bins* representats, de 100 a 10, per reduir així la mida de les dades.

Exemple

```
> hist(AirPassengers,breaks=100)
> hist(AirPassengers,breaks=10)
```

Figura 3. Histograma d'`AirPassengers`, quan es representen 100 i 10 *bins*



D'altra banda, les tècniques de *clustering* divideixen els registres de dades en grups, o clústers, de manera que els registres dins d'un mateix clúster siguin semblants entre ells i diferents als registres d'altres clústers. La similitud es defineix generalment en termes de quina proximitat tenen els registres a l'espai, basant-se en una funció de distància.

Tot i que el *clustering* s'explica amb més detall a l'apartat 2.5. Models no supervisats, ja que es tracta d'un mètode d'anàlisi, aquestes representacions també poden ser utilitzades per substituir les dades originals, amb l'objectiu de reduir-ne la mida. No obstant això, l'eficàcia d'aquesta tècnica dependrà de la naturalesa de les dades i de la seva capacitat per representar-les en grups.

Finalment, el mostreig (*sampling*, en anglès) es pot utilitzar com una tècnica de reducció de la quantitat, ja que permet que un gran conjunt de dades sigui representat per una mostra (o subconjunt) de dades aleatòries molt més petita. Aquesta reducció es pot realitzar mitjançant diferents mètodes, representats a la figura 4:

- **Mostra aleatòria simple sense substitució** (SRSWOR, per les sigles en anglès *Simple Random Sample WithOut Replacement*): dels N registres que formen el conjunt D, s'escullen s ($s < N$), on la probabilitat de cada registre de ser seleccionat és la mateixa ($1/N$).

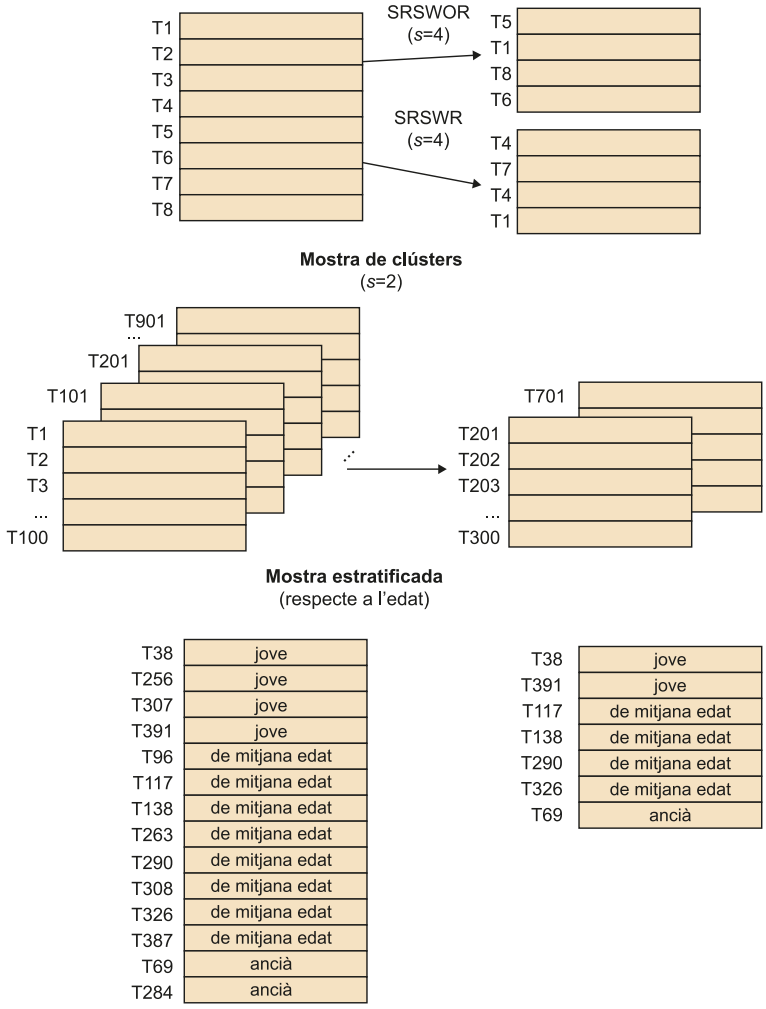
- **Mostra aleatòria simple amb substitució** (SRSWR, per les sigles en anglès *Simple Random Sample With Replacement*): en aquest cas, cada vegada que

se selecciona un registre, aquest es torna a tenir en compte en la selecció següent, de manera que cada registre pot escollir-se diverses vegades.

- **Mostra de clústers**: la selecció es realitza per clústers, és a dir, si els registres en el conjunt D s'agrupen en M clústers diferents, se seleccionen s clústers de manera aleatòria, on $s < M$.

- **Mostra estratificada**: si el conjunt D es divideix en parts perfectament dissociades, anomenades estrats, aquest mètode selecciona una mostra aleatòria per a cada estrat del conjunt. Això permet generar una mostra representativa de les dades quan aquestes són molt esbiaixades.

Figura 4. El mostreig (*sampling*) com a mètode de reducció de les dades



Adaptació de Jiawei *et al.* (2011)

El paquet `sampling` permet treballar amb aquest conjunt de mètodes a R. Algunes funcions interessants són `srswr`(), `srswor`(), `cluster`() o `strata`(). L'exemple següent mostra el codi utilitzat per seleccionar 3 clústers del conjunt de dades `swissmunicipalities`, disponible a R, prenent la variable `REG` com la que separa els $M = 7$ clústers o grups de dades i utilitzant el mètode de mostreig aleatori simple sense substitució (SRSWOR).

Exemple

```
>data(swissmunicipalities)
>cl=cluster(swissmunicipalities,clustername=c("REG"),size=3,method="srswor")
>getdata(swissmunicipalities, cl)
```

1.4. Conversió

En l'etapa de **conversió**, les dades són transformades amb l'objectiu que l'anàlisi posterior sigui més eficient i els resultats obtinguts siguin més fàcilment interpretables. Algunes tècniques de conversió habituals són la normalització, la transformació de Box-Cox o la discretització.

1.4.1. Normalització

La **normalització**, o estandarditzación, permet reduir el biaix causat per la combinació de valors mesurats a diferents escales a l'hora d'ajustar-los a una escala comuna, típicament entre (-1,1) o entre (0,1).

Depenent del context, aquesta normalització es pot aplicar mitjançant diferents mètodes, essent la normalització min-max i la normalització z-score els més comuns.

Bibliografia recomanada

Han, Jiawei; Kamber, Micheline; Pei, Jian (2011). *Data Mining: Concepts and Techniques*. Waltham: Elsevier.

La normalització min-max realitza una transformació lineal de les dades originals preservant la relació entre els valors del conjunt de dades original. Suposant A un atribut numèric amb n valors observats (v_1,..., v_n), on $minA$ i $maxA$ són els valors mínim i màxim d'aquest atribut, aquest mètode ajusta el valor v_i d'A a v_i' al rang ($minA'$, $maxA'$), per exemple (0,1), mitjançant la fórmula:

$$v_i' = \frac{v_i - minA}{maxA - minA}(maxA' - minA') + minA' \qquad (2)$$

D'altra banda, la normalització z-score transforma l'atribut original basant-se en la seva mitjana (μ_A) i desviació estàndard (σ_A) de la següent manera:

$$v_i' = \frac{v_i - \mu_A}{\sigma_A} \qquad (3)$$

La funció `scale()` aplica aquest tipus de normalització z-score a R. Atès que aquesta tècnica només tindrà sentit en variables numèriques, el següent exemple normalitza les 9 variables numèriques del conjunt de dades `mtcars`.

Exemple

```
>mtcars.scaled <- scale(mtcars[,c(1:7,10,11)])
```

1.4.2. Transformació de Box-Cox

Com s'esmenta en la introducció, algunes proves estadístiques assumeixen certes suposicions sobre les dades que s'han de complir per tal que aquestes proves, i per tant les conclusions extretes, siguin vàlides. Així, per poder aplicar proves per contrast d'hipòtesis paramètriques, com la prova *t* de Student:

1) Les variables de les dades analitzades han d'estar normalment distribuïdes.

2) Les variàncies d'aquestes variables han de romandre constants al llarg del rang observat d'alguna altra variable.

Quan no sigui així, es pot optar per utilitzar una alternativa no paramètrica, com les proves de Wilcoxon o Mann-Whitney (apartat 2.2.1), però això normalment implicarà la pèrdua de potència estadística. Per això, hi ha una altra alternativa prèvia a l'ús de proves no paramètriques que consisteix a convertir les dades per intentar millorar la seva normalitat i homoscedasticitat: la **transformació** (o família de transformacions) **de Box-Cox**.

Bibliografia recomanada

Osborne, Jason W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Thousand Oaks, CA: Sage Publications.

Després d'estimar un coeficient de transformació λ òptim, els diferents valors en el nou conjunt de dades queden definits a partir de l'expressió següent:

$$y_i^{\lambda} = \left(y_i^{\lambda} - 1\right) / \lambda \quad \lambda \neq 0;$$

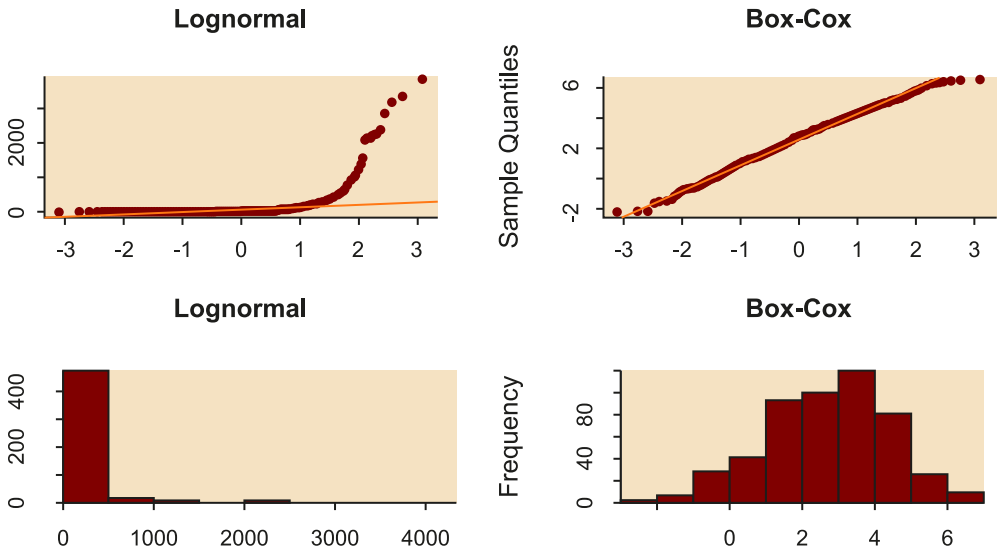
$$y_i^{\lambda} = ln\left(y_i\right) \quad \lambda = 0.$$

El paquet `DescTools` d'R permet estimar el valor òptim de λ mitjançant la funció `BoxCoxLambda()`, per després aplicar aquesta transformació amb la funció `BoxCox()`. A continuació, es mostra un exemple on la transformació de Box-Cox permet normalitzar una sèrie de dades sintètiques de tipus *lognormal* (figura 5).

Exemple

```
>x <- rlnorm(500, 3, 2)
>x.norm <- BoxCox(x, lambda = BoxCoxLambda(x))

>par(mfrow=c(2,2))
>qqnorm(x, main="Lognormal")
>qqline(x,col=2)

>qqnorm(bx, main="Box-Cox")
>qqline(bx,col=2)

>hist(x,main="Lognormal")
>hist(bx, main="Box-Cox")
```

Figura 5. Gràfic Q-Q i histograma del conjunt de dades *lognormal*, abans i després d'aplicar la transformació de Box-Cox



1.4.3. Discretització

La **discretització** consisteix a substituir els valors numèrics d'un atribut per etiquetes, categories o nivells, que poden ser conceptuals (nen/adult) o rangs de valors (0-18/19-100). La **dicotomització** és un cas particular de discretització en dues etiquetes o categories.

Tot i que aquesta fase de preprocessat resulta de vegades molt convenient a l'hora d'interpretar i comparar els resultats de diferents grups de dades, s'ha d'utilitzar amb moderació perquè a l'hora de discretitzar les dades s'estarà perdent informació que pot resultar d'interès.

Bibliografia recomanada

Osborne, Jason W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Thousand Oaks, CA: Sage Publications.

En aquest context, els mètodes de *clustering* tornen a ser una eina per discretitzar un atribut numèric, dividint els seus valors en grups o clústers. Aquesta tècnica té en compte la distribució de l'atribut, així com la proximitat de les dades a cada grup, de manera que permet una discretització de les dades d'alta qualitat.

Un altre grup de mètodes àmpliament utilitzats discretitza les dades aplicant criteris d'*equal-width* o *equal-frequency binning*, és a dir, dividint les dades de manera que l'amplada dels rangs o les freqüències de cada categoria siguin iguals, i substituint les dades per les mitjanes o medianes de cada *bin*. Tot i que són mètodes menys sofisticats per no tenir en compte la informació de cada categoria, la seva simplicitat és el que els fa especialment interessants.

Bibliografia recomanada

Han, Jiawei; Kamber, Micheline; Pei, Jian (2011). *Data Mining: Concepts and Techniques*. Waltham: Elsevier.

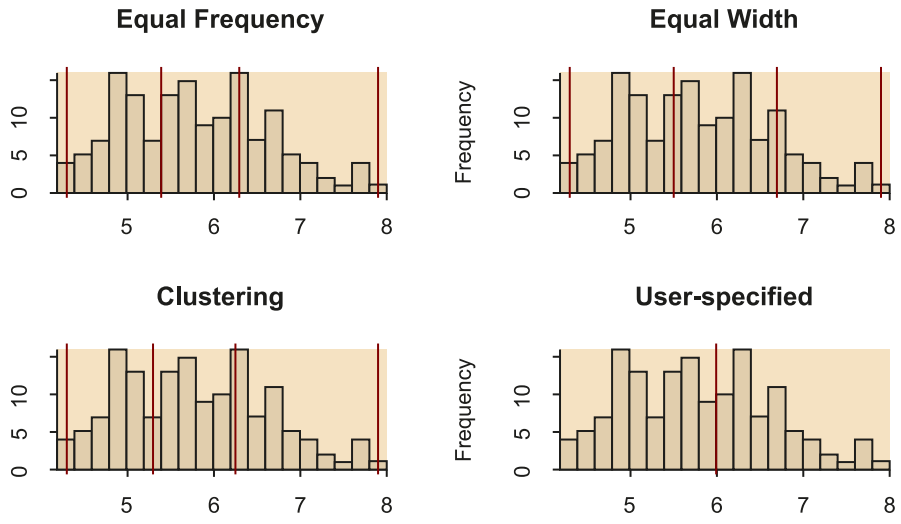
A R, el paquet `arules` permet discretitzar les dades utilitzant els mètodes esmentats, mitjançant la funció `discretize()`. El següent codi mostra un exemple de discretització en 3 nivells (la base de dades `iris` es compon de 3 tipus

de flors), tenint en compte els mètodes *equal-frequency*, *equal-width*, *clustering* i fixant els intervals manualment (*user-specified*). Es mostra la divisió en 3 categories de manera gràfica (figura 6), sobre l'histograma original.

Exemple

```
>data(iris)
>x <- iris[,1]

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

>#equal frequency
>hist(x, breaks = 20, main = "Equal Frequency")
>abline(v = discretize(x, breaks = 3, onlycuts = TRUE), col = "red")

>#equal interval width
>hist(x, breaks = 20, main = "Equal Width")
>abline(v = discretize(x, method = "interval", breaks = 3, onlycuts = TRUE),
col = "red")

>#  k-means clustering
>hist(x, breaks = 20, main = "Clustering")
>abline(v = discretize(x, method = "cluster", breaks = 3, onlycuts = TRUE),
col = "red")

# user-specified
hist(x, breaks = 20, main = "User-specified")
abline(v = discretize(x, method = "fixed", breaks = c(-Inf, 6, Inf), onlycuts
= TRUE), col = "red")
```

Figura 6. Resultat de discretitzar un conjunt de dades mitjançant els mètodes d'*equal-frequency binning*, *equal-width binning*, *clustering* i escollint manualment els intervals (*user-specified*)



1.5. Dades perdudes

Un dels riscos associats a la conversió de dades és la pèrdua d'informació. Aquesta pot ser parcial, quan es redueix la mida o el nivell de precisió pel que fa a les dades originals, però també pot ser completa (**dades perdudes** o *missing data*).

Bibliografia recomanada

Squire, Megan (2015). *Clean Data*. Birmingham: Packt Publishing.

Les dades buides o no definides es poden presentar en diferents formats, típicament "","" " o NA (*not available,* en anglès), però en alguns contextos poden, fins i tot, tenir valors numèrics com 0 o 999. Per tant, serà fonamental identificar en cada cas els valors que indiquin la pèrdua de dades.

En el cas de valors numèrics, aquests generalment es consideraran dades perdudes quan no formin part del domini de l'atribut. Per exemple, en el cas de l'índex de massa corporal o IMC (apartat 1.2), aquest mai no podrà ser 0 de manera que, si en un conjunt de dades trobem una sèrie de registres amb el valor de l'IMC a 0, aquesta dada molt probablement estarà indicant la pèrdua d'informació.

Així mateix, a vegades es pot tractar de valors buits legítims, i no de dades perdudes, quan aquest camp no sigui procedent per a un registre determinat. Per exemple, quan es pretengui recollir informació sobre els resultats d'una prova clínica i aquesta no s'hagi realitzat per a un pacient concret.

Segons Osborne (2013), depenent del context, hi ha múltiples possibles solucions a la pèrdua de dades. D'una banda, si la informació és coneguda i suposa una inversió de temps acceptable, la millor solució consisteix a completar manualment els registres que falten.

Més informació a:

Osborne, Jason W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Thousand Oaks, CA: Sage Publications.

Així mateix, es pot substituir el conjunt de valors perduts per una mateixa constant o etiqueta, com per exemple «Desconegut». Aquesta tècnica també pot ser útil quan els valors perduts tinguin un significat comú, com «No és procedent».

Altres aproximacions substitueixen els registres perduts per una mateixa mesura de tendència central, és a dir, per la mitjana o la mediana d'aquest atribut, depenent de la distribució de les dades. Aquesta mitjana es pot calcular per a tota la mostra o per a cadascuna de les classes o categories que la descriguin. Per exemple, es podrien calcular per separat les mitjanes d'homes i dones.

Finalment, altres aproximacions es basen en la implementació de mètodes probabilistes per predir (o imputar) els valors perduts. Alguns d'aquests mètodes són les regressions, les inferències basades en models bayesians o els arbres de decisió.

Tot i que les aproximacions que tracten d'imputar valors perduts es presenten com una alternativa particularment interessant amb l'objectiu de no perdre informació útil, s'han d'aplicar amb precaució i utilitzant mètodes adequats, per no introduir error i falsejar els resultats (Osborne, 2013). L'últim grup d'aproximacions, però, és el que gaudeix de més popularitat, ja que tracta de capturar la quantitat més gran d'informació de les dades per així predir els valors perduts.

Un dels mètodes més populars és el kNN (per les sigles en anglès, *k-Nearest Neighbours*), ja que permet predir valors en conjunts de dades multidimensionals formats per dades mixtes (contínues, discretes, ordinals o nominals). No obstant això, aquest mètode és molt sensible a l'elecció del valor *k*. Quan és massa elevat, s'inclouen valors significativament diferents del resultat esperat, mentre que quan *k* és massa baix implica la pèrdua de valors significatius. Per això, un altre mètode més robust que està guanyant popularitat en els últims anys ja que també permet treballar amb conjunts de dades mixtos multidimensionals és *missForest*.

Bibliografia recomanada

Osborne, Jason W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Thousand Oaks, CA: Sage Publications.

Stekhoven, Daniel J.; Bühlmann, Peter (2011, gener). «MissForest: Non-parametric missing value imputation for mixed-type data». *Bioinformatics* (vol. 28, núm. 1, pàg. 112-118).

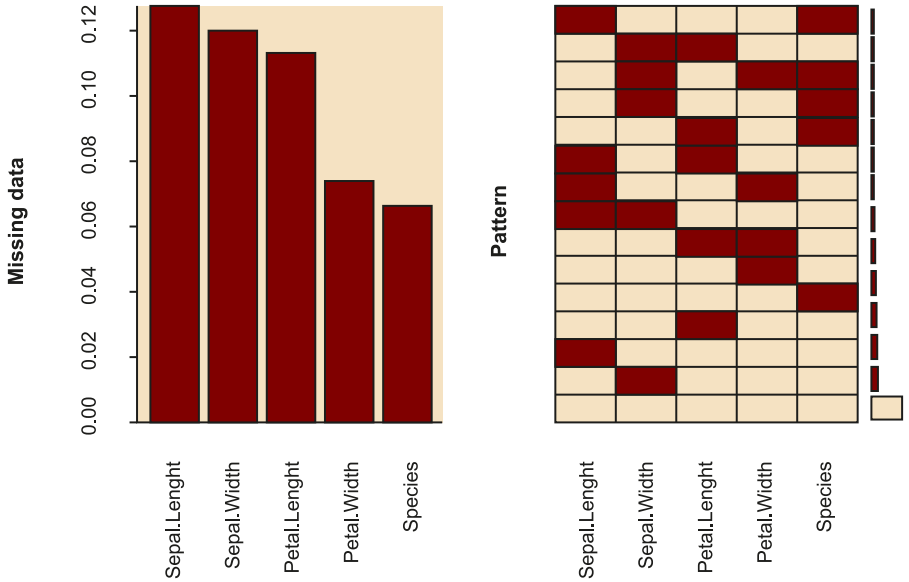
A R, hi ha diversos paquets que permeten aplicar la imputació de dades mitjançant kNN, com `DMwR` mitjançant la funció `knnImputation()`, o `VIM` mitjançant `kNN()`. Aquest últim, a més, permet representar gràficament els valors perduts d'un conjunt de dades mitjançant la funció `aggr()`. Així mateix, el paquet missForest permet aplicar aquest mètode d'imputació de dades perdudes.

En el següent exemple es mostren els resultats d'aplicar aquestes funcions (figura 7), després d'utilitzar la funció `missForest::prodNa()` per introduir valors perduts, de manera sintètica, en el conjunt de dades `iris`.

Exemple

```
>data(iris)
>iris.mis <- prodNA(iris, noNA = 0.1)

>aggr(iris.mis, numbers=TRUE, sortVars=TRUE, labels=names(iris.mis),
cex.axis=.7, gap=3, ylab=c("Missing data","Pattern"))
```

Figura 7. Percentatge de valors perduts en cada variable del conjunt de dades (esquerra). Representació de patrons en el conjunt de dades, tenint en compte –en vermell– els valors perduts (dreta)




```
>##kNN 1 (VIM package)
>kNN1.imp<-kNN(iris.mis, k=3)
```

```
>##kNN 2 (DMwR package)
>kNN2.imp <- knnImputation(iris.mis, 3)

>##missForest
>missForest.imp<-missForest(iris.mis, variablewise = TRUE)
```

1.6. Valors extrems

Els **valors extrems** (*extreme scores* o *outliers*) són aquelles dades que es troben molt allunyades de la distribució normal d'una variable o població. Són observacions que es desvien tant de la resta que aixequen sospites sobre si van ser generades mitjançant el mateix mecanisme. Així mateix, aquests valors poden afectar de manera adversa els resultats de les anàlisis posteriors, a l'hora d'incrementar l'error en la variància de les dades i esbiaixar significativament els càlculs i estimacions.

Bibliografia recomanada

Osborne, Jason W. (2010, març). «Data cleaning basics: Best practices in dealing with extreme scores». *Newborn and Infant Nursing Reviews* (vol. 10, núm. 1, pàg. 37-43).

Aquests valors poden aparèixer per diferents raons, per la qual cosa s'apliquen diferents solucions en funció del context. En alguns casos, aquestes dades seran legítimes i formaran part de la mostra, de manera que no s'haurà de modificar el conjunt de dades i es contemplaran els *outliers* a l'anàlisi.

En altres casos, es pot detectar una desviació sistemàtica en el grup de valors extrems, que se solucionarà amb una operació matemàtica senzilla. Per exemple, a l'hora d'integrar els conjunts de dades procedents de les diferents botigues d'una cadena, ens podem trobar que, encara que la majoria de les botigues reporten el preu dels seus productes en euros, la botiga que es troba a l'aeroport de Ciutat de Mèxic ha reportat els preus en pesos mexicans. Tot i que aquests preus es veuran com *outliers* inicialment, només cal aplicar la conversió de pesos a euros en aquest grup de dades per a què el conjunt sigui coherent.

En els casos en què els *outliers* siguin errors en les dades, complicats de corregir, generalment es tractaran com a valors perduts, de manera que s'optarà per eliminar o corregir el registre mitjançant els mètodes d'imputació de dades esmentades a l'apartat anterior (apartat 1.5).

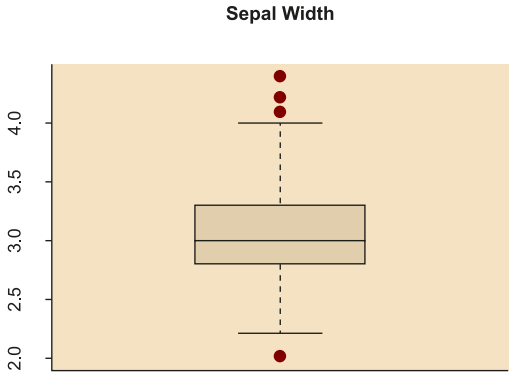
Tot i que la decisió sobre què es considera un valor extrem ha estat un tema controvertit durant dècades, generalment es considera que quan un valor es troba allunyat 3 desviacions estàndard respecte a la mitjana del conjunt és un *outlier*. Per això, en molts treballs s'utilitza la representació de les dades mitjançant gràfics de caixes (*boxplots*), amb l'objectiu de detectar aquests *outliers*.

El següent exemple mostra la detecció de valors extrems mitjançant aquesta tècnica a R. A la gràfica resultant de la funció boxplot() s'identifiquen 4 *outliers*, representats en forma de cercles i el valor dels quals es pot recuperar del resultat `out` (figura 8).

Exemple

```
>iris.bp<-boxplot(iris$Sepal.Width,main="Sepal Width")
>iris.bp$out
[1] 4.4 4.1 4.2 2.0
```

Figura 8. *Boxplot* de `Sepal.Width`, del conjunt de dades `iris`, on es representen 4 *outliers*



Altres mètodes que permeten identificar valors extrems en dades multidimensionals es basen en la distància de Mahalanobis o la distància de Cook. A partir d'aquestes distàncies, els valors extrems es poden identificar a l'hora d'aplicar un llindar o d'especificar un nombre fix d'*outliers*.

La distància de Mahalanobis determina la similitud entre variables aleatòries multidimensionals, basant-se en la correlació entre aquestes variables. Per altra banda, la distància de Cook estima el grau d'influència de cadascun dels punts que formen el conjunt quan es realitza una anàlisi de regressió per mínims quadrats.

En l'exemple següent es busca identificar els valors extrems d'un conjunt de dades incloent el pes i l'alçada de 16 individus. Quan se seleccionen com a *outliers* aquells punts que es troben, unidimensionalment, a més de 2 desviacions estàndard de la mitjana, una de les mostres que s'identifica com a normal (per no presentar un valor de més de 2 desviacions estàndard en cap dels atributs analitzats) sembla ser un *outlier* quan s'analitza visualment amb la resta de dades (figura 9). Això es confirma a l'hora de calcular les distàncies de Mahalanobis de cada mostra i seleccionar els 2 primers candidats a *outliers*.

Bibliografia recomanada

Newton, Rae R.; Rudestam, Kjell E. (1999). *Your Statistical Consultant: Answers to Your Data Analysis Questions*. Thousand Oaks, CA: Sage Publications.

Mahalanobis, Prasanta C. (1936, gener). «On the generalized distance in statistics». *Proceedings of the National Institute of Science of India* (vol. II, núm. 1).

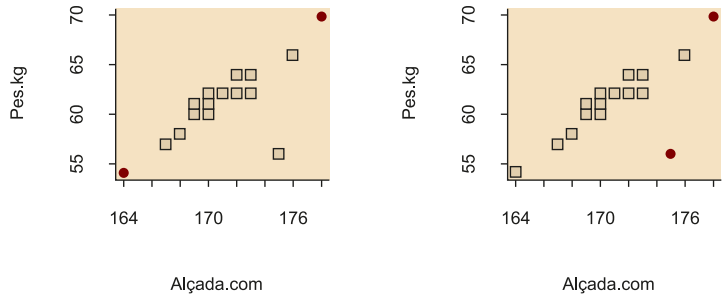
Cook, R. Dennis. (1977, febrer). «Detection of Influential Observations in Linear Regression». *Technometrics*, American Statistical Association (vol. 19, núm. 1, pàg. 15-18).

Exemple

```
>ap <- data.frame(Altura.cm=c(164, 167, 168, 169, 169, 170, 170, 170, 171,
172, 172, 173, 173, 175, 176, 178), Peso.kg=c( 54,  57,  58,  60,  61,  60,
61,  62,  62,  64,  62,  62,  64,  56,  66,  70))

>#Criterio +/-2SD
>Altura.outlier <- abs(scale(ap$Altura.cm)) > 2
>Peso.outlier <- abs(scale(ap$Peso.kg)) > 2
>pch <- (Altura.outlier | Peso.outlier) * 16

>par(mfrow=c(1,2))
>plot(ap, pch=pch)
```

```
>#Criterio distancia Mahalanobis (los dos outliers más extremos)
>n.outliers   <- 2
>m.dist.order <- order(mahalanobis(ap, colMeans(ap), cov(ap)), decreasing=TRUE)
>is.outlier   <- rep(FALSE, nrow(ap))
>is.outlier[m.dist.order[1:n.outliers]] <- TRUE
>pch <- is.outlier * 16

>plot(ap, pch=pch)
```

Figura 9. Representació del conjunt de dades `ap` i *outliers* detectats (cercles vermells) segons diferents criteris. Esquerra: dues desviacions estàndard. Dreta: distància de Mahalanobis



Finalment, hi ha altres mètodes més sofisticats per a la detecció d'*outliers* que es basen en models estadístics, supervisats o no supervisats, del conjunt de dades per intentar detectar les anomalies o errors. Per exemple, mitjançant tècniques de *clustering* és possible identificar conjunts de dades que s'allunyin significativament dels valors esperats de la mostra.

Bibliografia recomanada

Han, Jiawei; Kamber, Micheline; Pei, Jian (2011). *Data Mining: Concepts and Techniques*. Waltham: Elsevier.

2. Anàlisi de dades

L'**anàlisi** o **exploració de les dades** té com a objectiu explicar les principals característiques d'aquests dades, per així intentar respondre les preguntes plantejades en el marc d'un projecte de dades.

Depenent de la naturalesa d'aquestes dades, així com dels objectius del projecte, es poden aplicar diferents tipus d'anàlisi. Els següents apartats descriuen algunes de les més utilitzades.

2.1. Anàlisi estadística descriptiva

Les **estadístiques descriptives** són estimacions, valors calculats a partir d'una mostra de dades, que descriuen o resumeixen les característiques intrínseques d'aquesta mostra. La mitjana, la mediana o la desviació estàndard són estimacions comunament utilitzades.

Bibliografia recomanada

Jarman, Kristin H. (2013). *The art of data analysis: how to answer almost any question using basic statistics*. Hoboken, NJ: Wiley.

Moltes vegades, aquesta etapa es realitza, fins i tot, prèviament al procés de neteja de les dades, ja que proporciona una visió general de les dades molt valuosa a l'hora d'identificar els processos de neteja i anàlisi més adequats per al tipus de dades que es volen estudiar.

L'anàlisi estadística descriptiva es pot dividir principalment en els dos tipus següents:

- Les mesures de tendència central: s'inclouen les mesures que representen el centre de la distribució de dades, com la mitjana, la mediana, la moda o el rang mitjà.

- Les mesures de dispersió: és habitual calcular el rang, els quartils, el rang interquartílic, la variància o la desviació estàndard.

Algunes funcions interessants a R que permeten calcular algunes d'aquestes mesures descriptives són `mean()`, `median()`, `quantiles()`, `var()` o `sd()`. Així mateix, la funció `summary` () proporciona un resum de cada un dels atributs del conjunt, incloent-hi el mínim, el màxim, la mitjana, la mediana, el primer i tercer quartils, així com el nombre de NA (sigles de *not available*). A continuació, es mostra un exemple amb les dades d'`iris` modificades per contenir NA:

Exemple

```
>summary(iris.mis)
```

```
Sepal.Length    Sepal.Width        Petal.Length    Petal.Width      Species
 Min.   :4.400   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa  :45
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:48
 Median :5.800   Median :3.000   Median :4.200   Median :1.300   virginica :47
 Mean   :5.873   Mean   :3.058   Mean   :3.706   Mean   :1.209   NA's    :10
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
 NA's   :19      NA's   :18       NA's   :17    NA's   :11
```

Una altra funció interessant a R que permet mostrar de manera compacta l'estructura d'un conjunt de dades és `str()`, ja que mostra el nombre d'observacions i atributs o variables del conjunt, així com els seus tipus. L'exemple següent descriu les dimensions del conjunt de dades `iris.mis`, així com els tipus de variables que conté (quatre numèrics i un factor amb 3 nivells o categories).

Exemple

```
>str(iris.mis)

'data.frame':    150 obs. of  5 variables:
$ Sepal.Length: num  5.1 4.9 NA 4.6 5 NA 4.6 5 4.4 4.9 ...
$ Sepal.Width : num  NA 3 NA 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
$ Petal.Length: num  1.4 NA 1.3 1.5 NA 1.7 1.4 1.5 1.4 1.5 ...
$ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 NA 0.2 0.1 ...
$ Species  : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 NA 1 1 1 1 1 NA …
```

2.2. Anàlisi estadística inferencial

Aquest tipus d'anàlisi té per objectiu modelar les dades a través d'una distribució coneguda. Partint de la premissa que el conjunt de dades estudiat representa una fracció de la totalitat d'una població, el seu objectiu és inferir com és aquesta població, assumint un grau d'error en les estimacions pel fet de disposar d'una mostra reduïda de les dades.

Els següents apartats descriuen alguns exemples d'anàlisi d'aquest tipus, com són la comparació de grups mitjançant els contrastos d'hipòtesis, les regressions o les correlacions.

2.2.1. Comparació d'un o dos grups

A l'apartat sobre neteja de les dades es feia referència a la importància del *data screening* per verificar característiques importants en les dades a l'hora d'identificar els mètodes d'anàlisi més adequats (veure apartat 1.2). Per exemple, saber si les dades segueixen una distribució normal, així com si presenten homoscedasticitat, serà fonamental per poder aplicar proves per contrast d'hipòtesis de tipus paramètric.

Els apartats següents presenten diversos mètodes basats en l'anàlisi estadística de les dades, dissenyats per comprovar la normalitat i l'homoscedasticitat, així com per comparar parells de grups de dades.

Comprovació de la normalitat

Amb l'objectiu de verificar la suposició de la normalitat, algunes de les proves més habituals són els tests de Kolmogorov-Smirnov i de Shapiro-Wilk. Tot i que tots dos comparen la distribució de les dades amb una distribució normal, el test de Shapiro-Wilk es considera un dels mètodes més potents per contrastar la normalitat. Assumint com a hipòtesi nul·la que la població està distribuïda normalment, si el p-valor és més petit que el nivell de significació, generalment α=0,05, llavors la hipòtesi nul·la és rebutjada i es conclou que les dades no compten amb una distribució normal. Si, per contra, el p-valor és major a α, es conclou que no es pot rebutjar aquesta hipòtesi i s'assumeix que les dades segueixen una distribució normal.

El següent fragment de codi a R mostra com es poden aplicar aquestes proves, mitjançant les funcions `ks.test()` i `shapiro.test()`, respectivament.

Exemple

```
>ks.test(iris$Sepal.Length, pnorm, mean(iris$Sepal.Length), sd(iris$Sepal.Length))

 One-sample Kolmogorov-Smirnov test

data:  iris$Sepal.Length
D = 0.088654, p-value = 0.1891
alternative hypothesis: two-sided

>shapiro.test(iris$Sepal.Length)

 Shapiro-Wilk normality test

data:  iris$Sepal.Length
W = 0.97609, p-value = 0.01018
```

En l'exemple es presenta un cas controvertit en el qual s'obtenen resultats diferents per a cadascuna de les proves. Mentre que segons Kolmogorov-Smirnov les dades segueixen una distribució normal, el test de Shapiro-Wilk rebutja la hipòtesi nul·la i considera que no és així.

El **teorema central del límit** s'aplica a la distribució de la mitjana de la mostra d'un conjunt de dades. La mitjana d'una mostra de qualsevol conjunt de dades és cada vegada més normal a mesura que augmenta la quantitat d'observacions. Així, a mesura que augmenta la mida de la mostra N, la distribució de la mitjana de la mostra s'assembla cada vegada més a una distribució normal amb una (vertadera) mitjana de població μ i variància σ^2/N.

Bibliografia recomanada

Jarman, Kristin H. (2013). *The art of data analysis: how to answer almost any question using basic statistics*. Hoboken, NJ: Wiley.

Atès que la prova de Shapiro-Wilk es considera més robusta, una posició més conservadora conclouria que les dades no segueixen una distribució normal. Això no obstant, si el conjunt de dades es compon d'un nombre de registres prou gran, pel teorema central del límit, es podria considerar que les dades segueixen una distribució normal.

Comprovació de l'homoscedasticitat

Així mateix, algunes proves estadístiques requereixen la comprovació prèvia de l'homoscedasticitat en les dades, és a dir, de la igualtat de variàncies entre els grups que s'han de comparar. Entre les proves més habituals hi ha el test de Levene, que s'aplica quan les dades segueixen una distribució normal, així com el test de Fligner-Killeen, que es tracta de l'alternativa no paramètrica, utilitzada quan les dades no compleixen amb la condició de normalitat. En ambdues proves, la hipòtesi nul·la assumeix igualtat de variàncies en els diferents grups de dades, de manera que p-valors inferiors al nivell de significació indicaran heteroscedasticitat.

El codi següent a R mostra un exemple del funcionament d'aquestes proves a la base de dades `InsectSprays`, mitjançant les funcions `leveneTest`() del paquet `car` i `fligner.test`(), respectivament.

Exemple

```
> leveneTest(count ~ spray, data = InsectSprays)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value    Pr(>F)
group  5  3.8214 0.004223 **
      66
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> fligner.test(count ~ spray, data = InsectSprays)

    Fligner-Killeen test of homogeneity of variances

data:  count by spray
Fligner-Killeen:med chi-squared = 14.483, df = 5, p-value = 0.01282
```

Atès que les dues proves resulten en un p-valor inferior al nivell de significació (< 0,05), es rebutja la hipòtesi nul·la d'homoscedasticitat i es conclou que la variable `count` presenta variàncies estadísticament diferents per als diferents grups de `spray`.

Comparació entre dos grups de dades

Quan la normalitat i l'homoscedasticitat es compleixin (p-valors majors que el nivell de significació), es podran aplicar proves per contrast d'hipòtesis de tipus paramètric, com la prova *t* de Student. En els casos en què no es com-

pleixin, s'hauran d'aplicar proves no paramètriques com Wilcoxon (quan es comparin dades dependents) o Mann-Whitney (quan els grups de dades siguin independents).

En la prova *t* de Student, la hipòtesi nul·la assumeix que les mitjanes dels grups de dades són les mateixes, mentre que en les proves no paramètriques s'assumeix que les distribucions dels grups de dades són les mateixes. Per tant, només si el p-valor resultant de la prova és menor al nivell de significació es rebutjarà la hipòtesi nul·la i es conclourà que hi ha diferències estadísticament significatives entre els grups de dades analitzades.

A R, la prova *t* de Student s'aplica mitjançant la funció `t.test()`. El codi següent comprova la normalitat i homoscedasticitat de les dades `sleep` a R, per aplicar posteriorment la prova paramètrica *t* de Student.

Exemple

```
>shapiro.test(sleep$extra)
 Shapiro-Wilk normality test

data:  sleep$extra
W = 0.94607, p-value = 0.3114

>leveneTest(extra ~ group, data = sleep)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.2482 0.6244
      18

>t.test(extra ~ group, data = sleep)
 Welch Two Sample t-test

data:  extra by group
t = -1.8608, df = 17.776, p-value = 0.07939
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3654832  0.2054832
sample estimates:
mean in group 1 mean in group 2
           0.75            2.33
```

El p-valor resultant de la prova *t* de Student és més gran que el nivell de significació, això vol dir que no s'observen diferències estadísticament significatives entre els grups de dades de `sleep` per a la variable `extra`.

D'altra banda, les proves de Wilcoxon i Mann-Whitney s'apliquen indistintament mitjançant la funció `wilcox.test()`. L'exemple següent compara les distribucions d'`airquality` mitjançant aquesta prova, després de comprovar que no compleix les suposicions requerides pels tests paramètrics.

Exemple

```
> shapiro.test(airquality$Ozone)

    Shapiro-Wilk normality test
```

```
data:  airquality$Ozone
W = 0.87867, p-value = 2.79e-08

> fligner.test(Ozone ~ Month, data = airquality)

    Fligner-Killeen test of homogeneity of variances

data:  Ozone by Month
Fligner-Killeen:med chi-squared = 19.341, df = 4, p-value = 0.0006736

> wilcox.test(Ozone ~ Month, data = airquality, subset = Month %in% c(5, 8))

    Wilcoxon rank sum test with continuity correction

data:  Ozone by Month
W = 127.5, p-value = 0.0001208
alternative hypothesis: true location shift is not equal to 0
```

En aquest cas, sí que s'observen diferències estadísticament significatives en la qualitat de l'aire en termes de l'ozó (`airquality$Ozone`), entre els mesos de maig i agost.

Finalment, a vegades es voldrà comparar si hi ha diferències significatives en una variable categòrica entre els grups definits per una altra variable categòrica. En aquest cas, es pot aplicar el test de χ^2 a R, mitjançant la funció `chisq.test()`, com mostra l'exemple següent. A partir de les freqüències de cada gust de gelat per a cada un dels grups, s'observa que homes i dones mostren diferències significatives en els seus gustos.

Exemple

```
> men = c(100, 120, 60)
> women = c(350, 200, 90)
> ice.cream.survey = as.data.frame(rbind(men, women))
> names(ice.cream.survey) = c('chocolate', 'vanilla', 'strawberry')
> ice.cream.survey

   chocolate vanilla strawberry
men      100     120          60
women    350     200          90
> chisq.test(ice.cream.survey)

    Pearson's Chi-squared test

data:  ice.cream.survey
X-squared = 28.362, df = 2, p-value = 6.938e-07
```

2.2.2. Comparació entre més de dos grups

L'anàlisi de variància unidireccional, també conegut com a ANOVA (en anglès, *analysis of variance*) d'un únic factor, és una extensió de la prova *t* de Student, amb l'objectiu de comparar les mitjanes entre més de dos grups de dades.

El següent exemple mostra el resultat d'aplicar un test d'ANOVA a R mitjançant la funció `aov()`, a l'hora de comparar el `Sepal.Width` del conjunt de dades `iris`, entre les diferents espècies de flors (`Species`). La funció `summary.aov()` resumeix aquest resultat, permetent concloure que les diferents espècies mostren amplades del sèpal estadísticament diferents.

Exemple

```
> shapiro.test(iris$Sepal.Width)

    Shapiro-Wilk normality test

data:  iris$Sepal.Width
W = 0.98492, p-value = 0.1012

> leveneTest(Sepal.Width ~ Species, data = iris)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group   2  0.5902 0.5555
      147

> res.aov <- aov(Sepal.Width ~ Species, data = iris)
> summary(res.aov)
          Df Sum Sq Mean Sq F value Pr(>F)
Species    2  11.35   5.672   49.16 <2e-16 ***
Residuals 147  16.96   0.115
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

D'altra banda, l'alternativa no paramètrica als contrastos d'hipòtesis de més de 2 grups, és el test de Kruskal-Wallis. A R, s'aplica mitjançant la funció `kruskal.test()`, com mostra l'exemple següent.

Exemple

```
> shapiro.test(airquality$Ozone)

    Shapiro-Wilk normality test

data:  airquality$Ozone
W = 0.87867, p-value = 2.79e-08

> fligner.test(Ozone ~ Month, data = airquality)

    Fligner-Killeen test of homogeneity of variances

data:  Ozone by Month
Fligner-Killeen:med chi-squared = 19.341, df = 4, p-value = 0.0006736

> kruskal.test(Ozone ~ Month, data = airquality)

    Kruskal-Wallis rank sum test

data:  Ozone by Month
Kruskal-Wallis chi-squared = 29.267, df = 4, p-value = 6.901e-06
```

Atès que el p-valor obtingut és menor que el nivell de significació, es pot concloure que el nivell d'ozó mostra diferències significatives per als diferents mesos de l'any.

2.2.3. Regressió

La **regressió lineal** és un model matemàtic que té com a objectiu aproximar la relació de dependència lineal entre una variable dependent i una (o una sèrie) de variables independents.

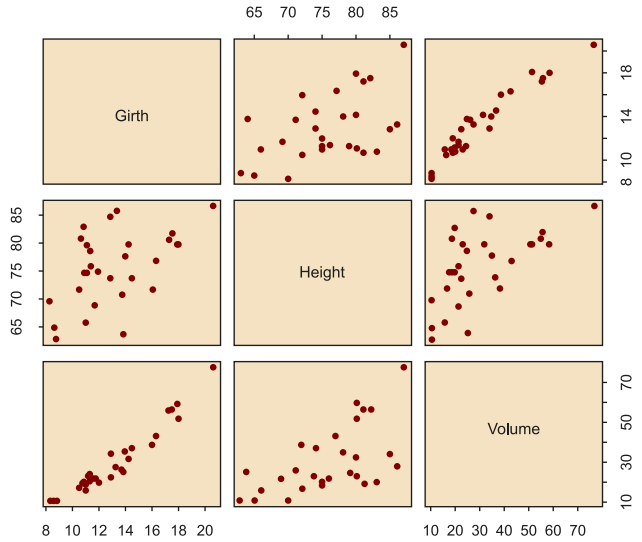
A R, la regressió lineal s'aplica mitjançant la funció `lm()`. Aquesta funció pot ser simple o múltiple segons les variables independents que s'incloguin en la fórmula que s'introdueix com a argument.

El següent codi a R mostra un exemple de cada tipus, per al conjunt de dades `trees`. En primer lloc, s'estima un model simple del volum, a partir del perímetre (`Girth`), després d'intuir visualment (figura 10) certa relació lineal entre aquestes dues variables. Posteriorment, s'implementa un model múltiple del volum, a partir del perímetre i l'alçada dels arbres (`Girth i Height`). Gràcies a la funció `summary()` s'analitzen detalladament els resultats de cada un dels models.

Exemple

```
> plot(trees)
```

Figura 10. Representació de cada parell de variables del conjunt de dades `trees`




```
> m1 = lm(Volume~Girth,data=trees)
> summary(m1)

Call:
lm(formula = Volume ~ Girth, data = trees)

Residuals:
   Min      1Q Median    3Q    Max
-8.065  -3.107  0.152  3.495  9.587

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.9435  3.3651  -10.98 7.62e-12 ***
```

```
Girth       5.0659  0.2474   20.48  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.252 on 29 degrees of freedom
Multiple R-squared:  0.9353,    Adjusted R-squared:  0.9331
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16


> m2 = lm(Volume~Girth+Height,data=trees)
> summary(m2)

Call:
lm(formula = Volume ~ Girth + Height, data = trees)

Residuals:
     Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877  8.6382  -6.713 2.75e-07 ***
Girth         4.7082  0.2643  17.816  < 2e-16 ***
Height        0.3393  0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,    Adjusted R-squared:  0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

Essent el coeficient de determinació (R^2 o *R-squared*) una mesura de qualitat del model que pren valors entre 0 i 1, es comprova com el volum i el perímetre es correlacionen amb força, donant lloc a un *R-squared* de 0,9353. A l'hora d'introduir l'alçada, aquest *R-squared* millora fins a 0,948 perquè també es correlaciona amb el volum de manera significativa, tot i que menys.

Així mateix, la funció `lm()` permet implementar models polinòmics més complexos, com en l'exemple següent:

Exemple

```
> m3 = lm(Volume~Girth+I(Girth^2),data=trees)
> summary(m3)

Call:
lm(formula = Volume ~ Girth + I(Girth^2), data = trees)

Residuals:
    Min     1Q     Median     3Q    Max
-5.4889 -2.4293 -0.3718  2.0764  7.6447

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.78627   11.22282   0.961 0.344728
Girth       -2.09214    1.64734    -1.270 0.214534
I(Girth^2)    0.25454    0.05817   4.376 0.000152 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.335 on 28 degrees of freedom
Multiple R-squared:  0.9616,    Adjusted R-squared:  0.9588
F-statistic: 350.5 on 2 and 28 DF,  p-value: < 2.2e-16
```

Es pot observar que el terme que relaciona el volum amb el perímetre de manera quadràtica resulta ser el més significatiu, millorant l'*R-squared* fins a 0,9616.

Finalment, si un cop estimat el model de regressió volguéssim utilitzar-lo per predir el resultat en noves mostres de dades, s'utilitzaria la funció `predict()` de la manera següent:

Exemple

```
> pred.frame<-data.frame(Girth=seq(10,16,2))
> predict(m3,newdata=pred.frame)
    1     2     3     4
15.31863 22.33400 31.38568 42.47365
```

D'altra banda, la **regressió logística** és un tipus d'anàlisi de regressió utilitzat per predir el resultat d'una variable dicotòmica dependent, en funció d'una sèrie de variables independents o predictores. Atès que aquest model estima les probabilitats d'ocurrència, en lloc d'utilitzar un model additiu que podria predir valors fora del rang (0,1) utilitza una escala transformada, basada en una funció logística.

Així, el model lineal per a probabilitats transformades es defineix com:

$$logit(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_k x_k \qquad (4)$$

on $logit(p) = log\left(\frac{p}{1-p}\right)$ i p representa la probabilitat d'ocurrència d'una de les categories.

A R, aquest tipus de models s'estimen mitjançant la funció `glm()`, especificant la família com a binomial. El següent codi mostra un exemple d'aplicació sobre el conjunt de dades `BreastCancer` del paquet `mlbench`, on es combinen les variables `Cell.size` i `Cell.shape` per estimar si un tumor és maligne o benigne.

Exemple

```
> data(BreastCancer, package="mlbench")
> bc <- BreastCancer[complete.cases(BreastCancer),]
> m4<- glm(Class ~ Cell.size+Cell.shape,data=bc, family="binomial")
> summary(m4)

Call:
glm(formula = Class ~ Cell.size + Cell.shape, family = "binomial",
 data = bc)

Deviance Residuals:
  Min        1Q     Median    3Q    Max
-2.6080  -0.0868  -0.0868   0.0000   3.3416

Coefficients:
              Estimate Std. Error z value   Pr(>|z|)
(Intercept)     9.57037  819.07884   0.012     0.991
Cell.size.L    11.57767  971.41961   0.012     0.990
Cell.size.Q    -6.17166 1905.62757  -0.003     0.997
```

```
Cell.size.C     8.20862 1050.06309   0.008     0.994
Cell.size^4    18.04156 1699.37615   0.011     0.992
Cell.size^5     4.55711 1084.68606   0.004     0.997
Cell.size^6    -8.65383 1474.48372  -0.006     0.995
Cell.size^7     1.55134 1538.21302   0.001     0.999
Cell.size^8    10.18306 1226.09621   0.008     0.993
Cell.size^9     0.08582 2675.24882   0.000     1.000
Cell.shape.L   17.78829 2558.23603   0.007     0.994
Cell.shape.Q    8.88783 1476.66754   0.006     0.995
Cell.shape.C    5.64269 1282.10652   0.004     0.996
Cell.shape^4   -1.99598 2612.89511  -0.001     0.999
Cell.shape^5   -5.73970 3110.33104  -0.002     0.999
Cell.shape^6   -5.75511 2642.37488  -0.002     0.998
Cell.shape^7   -3.90172 1695.00477  -0.002     0.998
Cell.shape^8   -1.51847  805.59705  -0.002     0.998
Cell.shape^9   -0.82841  251.11006  -0.003     0.997


(Dispersion parameter for binomial family taken to be 1)

 Null deviance: 884.35  on 682  degrees of freedom
Residual deviance: 180.13  on 664  degrees of freedom
AIC: 218.13

Number of Fisher Scoring iterations: 19
```

En aquest cas, la bondat del model s'avaluarà mitjançant la mesura AIC (criteri d'informació d'Akaike, per les seves sigles en anglès *Akaike Information Criterion*). Atès que aquesta mesura té en compte tant la bondat de l'ajustament (l'error) com la complexitat del model, quan es comparin diversos models candidats se seleccionarà aquell que resulti en el menor AIC.

2.2.4. Correlació

El **coeficient de correlació** és una mesura de l'associació entre dues variables. Aquest coeficient pot prendre valors entre -1 i 1, on els extrems indiquen una correlació perfecta i el 0 indica l'absència de correlació. El signe és negatiu quan valors elevats d'una variable s'associen amb valors petits de l'altra, i el signe és positiu quan ambdues variables tendeixen a incrementar o disminuir simultàniament.

El coeficient de correlació de **Pearson** és el més utilitzat entre variables relacionades linealment. No obstant això, per poder aplicar-se, requereix que la distribució de les dues variables sigui normal, així com que es compleixi el criteri d'homoscedasticitat.

Així, la correlació de **Spearman** apareix com una alternativa no paramètrica que mesura el grau de dependència entre dues variables. Aquest mètode no comporta cap suposició sobre la distribució de les dades, tot i que les variables que es volen comparar s'han de mesurar almenys en una escala ordinal.

A R, la funció `cor()` permet calcular la correlació entre les variables que componen un conjunt de dades. Per exemple:

Exemple

```
> cor(trees)

           Girth Height   Volume
Girth  1.0000000 0.5192801 0.9671194
Height 0.5192801 1.0000000 0.5982497
Volume 0.9671194 0.5982497 1.0000000
```

No obstant això, el resultat anterior no dona cap indicació sobre si la correlació és significativament diferent de zero. Per a això, cal fer servir la funció `cor.test()` i especificar les variables que es volen comparar.

Exemple

```
> cor.test(trees$Volume,trees$Height)

    Pearson's product-moment correlation

data:  trees$Volume and trees$Height
t = 4.0205, df = 29, p-value = 0.0003784
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3095235 0.7859756
sample estimates:
   cor
0.5982497
```

Com s'observa del resultat anterior, `cor.test()` analitza per defecte la correlació de Pearson, tot i que també permet analitzar altres correlacions com la de Spearman, que serà més indicada quan les dades no segueixin una distribució normal.

Exemple

```
> cor.test(trees$Volume,trees$Height, method="spearman")

    Spearman's rank correlation rho

data:  trees$Volume and trees$Height
S = 2089.6, p-value = 0.0006484
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.5787101
```

En tots dos casos el p-valor és significatiu i el coeficient de correlació és més gran que 0,57. Tanmateix, podem comprovar com la condició de normalitat no es compleix per a la variable `Volume`, de manera que el test més adequat en aquest cas serà el de Spearman. S'observa, per tant, una correlació de 0,579 entre `Volume` i `Height`. Seria erroni afirmar que aquesta correlació és de 0,598, ja que ens estaríem basant en el resultat del test de Pearson, que suposa normalitat en les dades.

2.3. Anàlisi de supervivència

L'**anàlisi de la supervivència** tracta generalment amb dades que no estan distribuïdes amb normalitat. Així mateix, aquestes dades sovint són censurades, és a dir, no es coneix la seva supervivència exacta, ja que va més enllà del període d'estudi. Aquesta tècnica s'utilitza àmpliament en camps com la biologia i la medicina, però també en enginyeria, per a l'estudi de la fiabilitat de certes aplicacions (Dalgaard, 2008).

L'**estimador de Kaplan-Meier** és un dels mètodes més utilitzats perquè es tracta d'un estimador no paramètric de la funció de supervivència que té en compte la censura.

A R, el paquet `survival` permet realitzar aquest tipus d'anàlisi. Mitjançant la funció `Surv()` es poden indicar les dades censurades per, posteriorment, implementar un estimador de Kaplan-Meier mitjançant la funció `survfit()`.

El següent exemple utilitza el conjunt de dades `melanoma`, del paquet `ISwR`. En aquest cas, la variable `status` és un indicador de l'estat del pacient al final de l'estudi:

1) mort per melanoma maligne;

2) viu l'1 de gener del 1978; i

3) mort per altres causes.

A més a més, `days` és el temps d'observació en dies, `ulc` indica si el tumor va ser ulcerat, `thick` és el gruix del tumor i `sex` conté la informació sobre el sexe del pacient (1 per a les dones i 2 per als homes). Atès que els estats 2 i 3 es consideren censurats, la funció de supervivència es modela a partir d'aquesta informació (els valors censurats es representen amb el signe +).

Exemple

```
> data(melanom)
> attach(melanom)

> Surv(days, status==1)
[1]   10+   30+   35+   99+  185   204   210   232   232+  279   295   355+  386   426   469
493+  529   621   629
[20]  659   667   718   752   779   793   817   826+  833   858   869   872   967   977   982
1041  1055  1062  1075
[39] 1156  1228  1252  1271  1312  1427+ 1435  1499+ 1506  1508+ 1510+ 1512+ 1516  1525+ 1542+
1548  1557+ 1560  1563+
[58] 1584  1605+ 1621  1627+ 1634+ 1641+ 1641+ 1648+ 1652+ 1654+ 1654+ 1667  1678+ 1685+ 1690
1710+ 1710+ 1726  1745+
[77] 1762+ 1779+ 1787+ 1787+ 1793+ 1804+ 1812+ 1836+ 1839+ 1839+ 1854+ 1856+ 1860+ 1864+ 1899+
1914+ 1919+ 1920+ 1927+
[96] 1933  1942+ 1955+ 1956+ 1958+ 1963+ 1970+ 2005+ 2007+ 2011+ 2024+ 2028+ 2038+ 2056+ 2059+
2061  2062  2075+ 2085+
[115] 2102+ 2103  2104+ 2108  2112+ 2150+ 2156+ 2165+ 2209+ 2227+ 2227+ 2256  2264+ 2339+ 2361+
2387+ 2388  2403+ 2426+
```

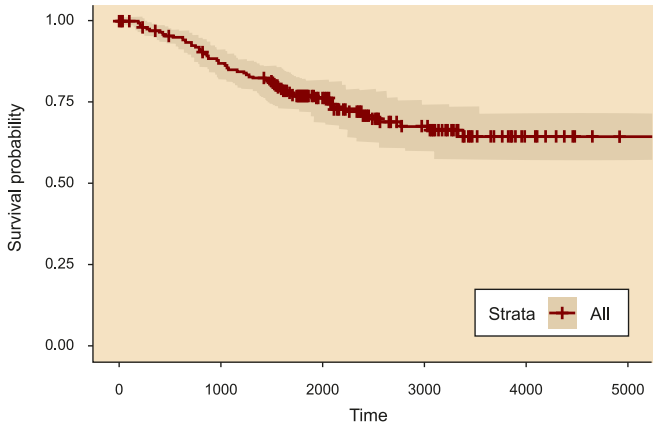
```
[134] 2426+ 2431+ 2460+ 2467  2492+ 2493+ 2521+ 2542+ 2559+ 2565  2570+ 2660+ 2666+ 2676+ 2738+
2782  2787+ 2984+ 3032+
[153] 3040+ 3042  3067+ 3079+ 3101+ 3144+ 3152+ 3154+ 3180+ 3182+ 3185+ 3199+ 3228+ 3229+ 3278+
3297+ 3328+ 3330+ 3338
[172] 3383+ 3384+ 3385+ 3388+ 3402+ 3441+ 3458+ 3459+ 3459+ 3476+ 3523+ 3667+ 3695+ 3695+ 3776+
3776+ 3830+ 3856+ 3872+
[191] 3909+ 3968+ 4001+ 4103+ 4119+ 4124+ 4207+ 4310+ 4390+ 4479+ 4492+ 4668+ 4688+ 4926+ 5565+
```

A més, l'estimador de Kaplan-Meier es pot aplicar per al conjunt de dades complet o dividint l'anàlisi en grups de dades. El següent exemple mostra el resultat d'aplicar les dues aproximacions, prenent el conjunt de dades complet i posteriorment separant l'estimació per a homes i dones. La funció `ggsurvplot()` del paquet `survminer` permet graficar el resultat, on les línies verticals representen les dades censurades.

Exemple

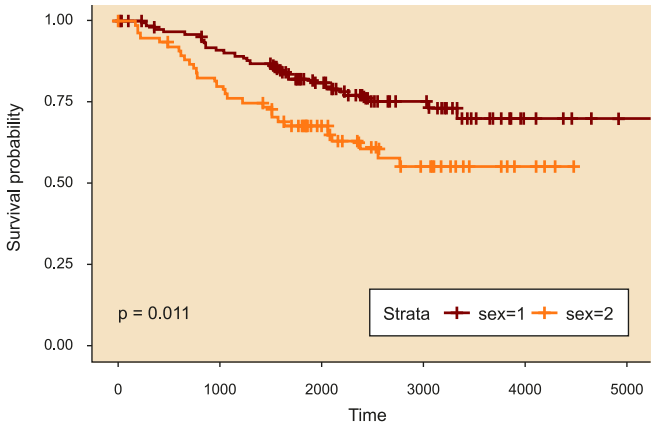
```
> surv.all <- survfit(Surv(days,status==1)~1)
ggsurvplot(surv.all,Surv(days,status==1),pval=TRUE)
```

Figura 11. Kaplan-Meier (amb interval de confiança) per a les dades de `melanoma`




```
> surv.bysex <- survfit(Surv(days,status==1)~sex)
ggsurvplot(surv.bysex,Surv(days,status==1),pval=TRUE)
```

Figura 12. Kaplan-Meier per a les dades de `melanoma`, quan se separa l'estimació per a homes (taronja) i dones (vermell)



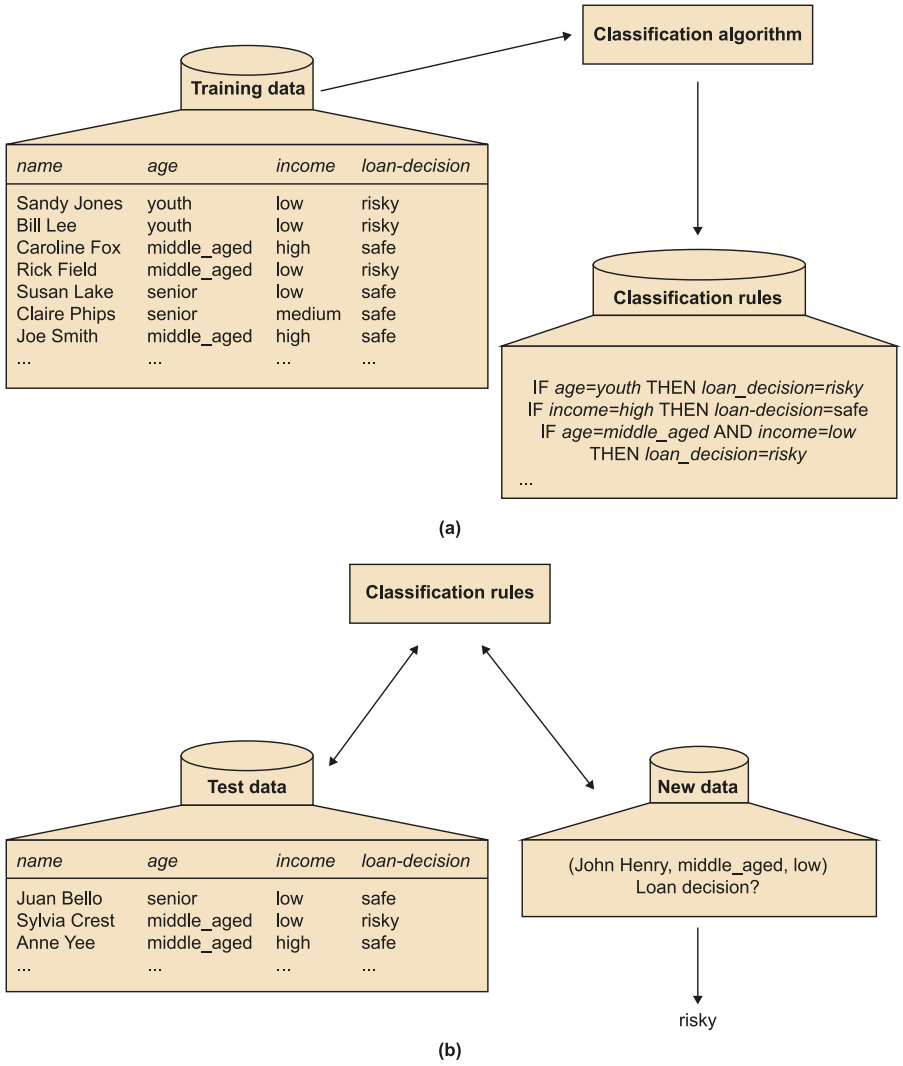
A la figura 12 s'observa com la supervivència al melanoma tendeix a ser menor en el grup d'homes (`sex = 2`), essent aquesta diferència estadísticament significativa (p-valor = 0,011).

2.4. Models supervisats

L'**aprenentatge supervisat** estima una funció o model a partir d'una sèrie de dades d'entrenament, amb l'objectiu de predir posteriorment el resultat de noves dades desconegudes. Els conjunts de dades d'entrenament estan formats per parells d'objectes que representen les dades d'entrada i els resultats desitjats. Aquests resultats poden ser un valor numèric, com en els problemes de regressió tractats a l'apartat 2.2.3, o una etiqueta de classe, com en els de classificació, que seran l'objecte d'aquest apartat.

En qualsevol problema de **classificació**, el conjunt de dades es dividirà, per tant, en els subconjunts d'entrenament (*training*) i de prova o test (*testing*). Gràcies als primers, s'entrenarà un model de classificació de manera que es defineixin una sèrie de regles de classificació. A continuació, gràcies a les dades de test, s'estimarà l'exactitud (*accuracy*) del model, de manera que, si és acceptable, les regles de classificació definides podran ser utilitzades en noves dades d'entrada amb les mateixes característiques, amb l'objectiu de predir-ne el resultat. La figura 13 representa esquemàticament aquest procés.

Figura 13. Etapes implicades en els processos de classificació: a) etapa d'entrenament i b) etapa de classificació (Jiawei *et al.*, 2011)



Els apartats següents descriuen els principals mètodes de partició de les dades que permeten definir els grups d'entrenament i test, així com els models de classificació i les mesures d'anàlisi del rendiment d'aquests models més utilitzats.

2.4.1. Partició de les dades

Hi ha diversos mètodes per classificar les dades originals en entrenament i test: el mètode d'exclusió (*holdout*), el mètode de submostreig aleatori (*random subsampling*) i la validació creuada (*cross-validation*).

Bibliografia recomanada

Han, Jiawei; Kamber Micheline; Pei, Jian (2011). *Data Mining: Concepts and Techniques*. Waltham: Elsevier.

En el mètode d'**exclusió**, les dades es divideixen aleatòriament en dos conjunts independents, el d'entrenament i el de test. Típicament, dos terços de les dades s'assignen al conjunt d'entrenament, i el terç restant es reserva per testejar el model. En general, aquesta estimació és pessimista ja que només utilitza una part de les dades originals per dissenyar el model.

El mètode de **submostreig aleatori** és una variació del mètode anterior, ja que s'aplica la mateixa tècnica *k* vegades per estimar posteriorment la precisió global del model com la mitjana de les precisions obtingudes de cada iteració.

En el mètode de la **validació creuada** de tipus *k-fold*,les dades originals es divideixen aleatòriament en *k* subconjunts (*folds*) mútuament exclusius i de mides semblants. L'entrenament i testeig es realitzen *k* vegades, a partir de totes les combinacions possibles de *k-1* subconjunts per a entrenament i deixant el subconjunt restant per testejar el model. En aquest cas, a diferència dels mètodes anteriors, cada mostra es fa servir el mateix nombre de vegades per entrenar i només una vegada per testejar. L'exactitud es calcula com el nombre total de classificacions correctes en les *k* iteracions, dividit pel nombre total de mostres en el conjunt de dades original.

El *leave-one-out* és un cas especial de validació creuada de tipus *k-fold* on *k* s'ajusta al nombre de mostres del conjunt de dades original. A cada iteració, només omet una de les mostres en la fase d'entrenament, per posteriorment utilitzar-la en el testeig del model. La validació creuada de tipus *k-fold* és una estimació pessimista i esbiaixada del rendiment del model ja que, generalment, millorarà quan s'ampliï el conjunt d'entrenament. Tot i que la validació creuada de tipus *leave-one-out* redueix significativament aquest biaix a l'hora d'utilitzar pràcticament la totalitat del conjunt de dades en l'etapa d'entrenament, tendeix a presentar una alta variància (es poden obtenir estimacions molt diferents a l'hora de variar la mostra seleccionada per a l'etapa de prova). Així, com l'error de l'estimador dependrà finalment tant del biaix com de la variància, l'ús d'un mètode o un altre dependrà del context.

En la validació creuada estratificada, els *folds* s'estratifiquen de manera que la distribució de classes a cada *fold* sigui aproximadament el mateix que en el conjunt de dades original. Generalment es recomana utilitzar una validació creuada estratificada amb 10 *folds*, encara que la potència de càlcul permeti utilitzar més *folds*, a causa del seu biaix i variància relativament baixos.

A R, el mètode d'exclusió o *holdout* es pot aplicar mitjançant la funció `holdout()` del paquet `rminer`. Reprenent l'exemple del càncer de mama (`BreastCancer`), el codi següent divideix les dades originals en 2/3 per a entrenament i 1/3 per testeig.

Exemple

```
> h<-holdout(bc$Class,ratio=2/3,mode="stratified")
> data_train<-bc[h$tr,]
> data_test<-bc[h$ts,]
> print(table(data_train$Class))

benign malignant
 296    159

> print(table(data_test$Class))
```

```
benign malignant
 148     80
```

Atès que l'exemple aplica una partició de les dades estratificada, les proporcions entre tumors benignes i malignes es mantenen en els dos conjunts de dades, deixant 2/3 per a l'entrenament (296 tumors benignes i 159 de malignes), i la resta per a l'etapa de test (148 tumors benignes i 89 de benignes).

Un altre paquet interessant a R és `caret`, ja que permet utilitzar una gran quantitat de mètodes d'entrenament i classificació. La funció `trainControl()` permet especificar les característiques del procés d'entrenament i testeig. Els següents exemples de codi mostren algunes possibilitats d'aquesta eina. En el primer cas s'aplica un *leave-one-out*; el segon fa referència a una validació creuada de tipus *10-fold* i el tercer indica la repetició 10 vegades d'una validació creuada de tipus *4-fold*.

Exemple

```
> train_control1<- trainControl(method="LOOCV")
> train_control2<- trainControl(method="cv", number=10)
> train_control3<- trainControl(method="repeatedcv", number=4, repeats=10)
```

2.4.2. Mesures del rendiment

A l'hora d'avaluar la bondat o el rendiment del model de classificació, hi ha diverses mesures que permeten interpretar el resultat d'aplicar aquest model al subconjunt de testeig.

Si s'entenen els valors positius com aquells registres que pertanyen a la classe d'interès (per exemple, pacients que no responen a un tractament) i negatius la resta (pacients que responen al tractament), es defineixen els termes següents, representats a la matriu de confusió de la figura 14:

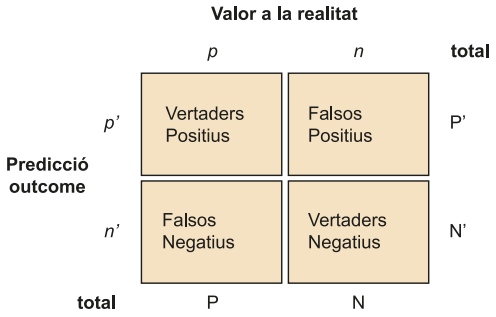
- **Veritables positius (*VP*)**: són els registres positius correctament classificats. Per exemple, els pacients que no responen al tractament que han estat classificats correctament com a «no responedors».

- **Falsos positius (*FP*)**: fan referència als registres negatius que van ser incorrectament classificats com a positius, és a dir, els pacients que van respondre al tractament, però van ser classificats com a «no responedors».

- **Veritables negatius (*VN*)**: són els registres negatius correctament classificats. Són els pacients que van respondre al tractament i van ser classificats com a «responedors».

- **Falsos negatius (*FN*)**: fan referència als registres positius classificats com a negatius. En l'exemple, aquells pacients que no responen al tractament, però que són classificats com a «responedors».

Figura 14. Matriu de confusió



Font: *Viquipèdia*.

L'**exactitud** (*accuracy*) d'un classificador fa referència al conjunt de registres correctament classificats i es calcula com a:

$$Exactitud = \frac{VP + VN}{P + N} \qquad (5)$$

Això no obstant, aquesta mesura no informa de la capacitat del model per classificar els registres positius i negatius per separat. Per a això, la **sensibilitat** (o taxa de veritables positius) quantifica la proporció de registres positius correctament identificats, mentre que l'**especificitat** (o taxa de veritables negatius) mesura la proporció de registres negatius que s'identifiquen correctament. Aquestes mesures es defineixen com a:

$$Sensibilitat = \frac{VP}{P} \qquad (6)$$

$$Especificitat = \frac{VN}{N} \qquad (7)$$

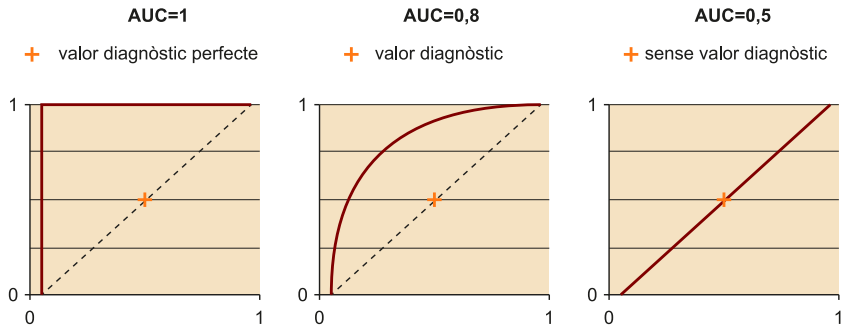
Una altra mesura àmpliament utilitzada és la precisió, que expressa la proporció de registres classificats com a positius que efectivament ho són, i es calcula:

$$Precisió = \frac{VP}{VP + FP} \qquad (8)$$

Les corbes ROC (*receiver operating characteristic*, per les sigles en anglès) són una altra eina visual per comparar models de classificació. A R, el paquet `pROC` permet visualitzar i treballar amb aquestes corbes. Les ROC mostren el compromís entre la taxa de veritables positius (TVP), equivalent a la sensibilitat, i la taxa de falsos positius (TFP), equivalent a 1-Especificitat. Donat un model i un subconjunt de dades per al testeig, la TVP és la proporció de registres positius que han estat correctament etiquetats pel model, mentre que la TFP és la proporció de registres negatius que han estat erròniament etiquetats com a positius.

A partir de la representació de la corba ROC, és habitual mesurar la seva àrea o AUC (*area under the curve,* per les sigles en anglès). Aquest paràmetre proporciona informació sobre la qualitat del model, essent menys precís a mesura que l'AUC s'acosta a 0,5 i mostrant una exactitud perfecta quan és 1. La figura 15 mostra tres exemples de corbes ROC juntament amb el valor de la seva AUC.

Figura 15. AUC obtinguda per a diferents corbes ROC



Font: *Viquipèdia.*

2.4.3. Mètodes de classificació

Hi ha gran quantitat de mètodes de classificació, dissenyats per adaptar-se a diferents tipus de dades. Per això, dependrà del context la idoneïtat de cada mètode per a un problema concret. Alguns exemples són els arbres de decisió, els models bayesians o els basats en regles (*rule-based models*). Així mateix, tècniques més sofisticades de classificació inclouen les màquines de suport vectorial, els *random forests* o les xarxes neuronals. A Jiawei *et al.*, 2011, es pot trobar més informació sobre cadascun d'aquests mètodes.

Tot i que molts d'aquests mètodes tenen el seu propi paquet a R, `caret` inclou més de 200 possibilitats, especificades en aquest enllaç:

Enllaç d'interès

Models de classificació disponibles en el paquet `caret` d'R: http://topepo.github.io/caret/available-models.html

El següent fragment de codi mostra un exemple d'aplicació d'un *random forest* al conjunt de dades sobre càncer de mama (`BreastCancer`), mitjançant una validació creuada amb 4 *folds*. Amb l'objectiu d'analitzar la bondat del model en un conjunt de dades completament desconegut, es divideixen les dades en entrenament i test per només aplicar l'entrenament mitjançant validació creuada al primer subconjunt. Posteriorment, mitjançant la funció `predict()` es prediu el resultat de les dades del subconjunt de test i es representen les diferents mesures de bondat del model, mitjançant la funció `confusionMatrix()`, especificant com a positius els casos de càncer maligne.

Exemple

```
> data(BreastCancer, package="mlbench")
> bc <- BreastCancer[complete.cases(BreastCancer),-1]
```

```
> h<-holdout(bc$Class,ratio=2/3,mode="stratified")
> data_train<-bc[h$tr,]
> data_test<-bc[h$ts,]

> train_control<- trainControl(method="cv", number=4)
> mod<-train(Class~., data=data_train, method="rf", trControl = train_control)

> pred <- predict(mod, newdata=data_test)
> confusionMatrix(pred,data_test$Class,positive="malignant")

Confusion Matrix and Statistics

          Reference
Prediction   benign malignant
  benign        142         6
  malignant       6        74

               Accuracy : 0.9474
                 95% CI : (0.9099, 0.9725)
 No Information Rate : 0.6491
 P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8845
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9250
            Specificity : 0.9595
         Pos Pred Value : 0.9250
         Neg Pred Value : 0.9595
             Prevalence : 0.3509
         Detection Rate : 0.3246
   Detection Prevalence : 0.3509
   Balanced Accuracy : 0.9422

    'Positive' Class : malignant
```

2.5. Models no supervisats

L'**aprenentatge no supervisat** consisteix a adaptar un model a les observacions donades, ja que, a diferència de l'aprenentatge supervisat, no es té un coneixement *a priori* de les dades. Tot i que existeixen diferents mètodes d'aprenentatge no supervisat, els més utilitzats són els basats en l'agrupament o *clustering*.

L'agrupament o *clustering* és el procés mitjançant el qual s'agrupa un conjunt de dades en múltiples subgrups o clústers, de manera que els registres dins d'un mateix clúster tinguin una alta similitud i siguin molt diferents dels registres d'altres clústers. Aquestes diferències i similituds s'avaluen d'acord amb els valors dels atributs que descriuen cada registre i sovint impliquen mesures de distància. Aquest tipus de mètodes també es pot dividir en les categories següents: mètodes de partició, mètodes jeràrquics i mètodes basats en la densitat.

Bibliografia recomanada

Han, Jiawei; Kamber, Micheline; Pei, Jian (2011). *Data Mining: Concepts and Techniques*. Waltham: Elsevier.

Els **mètodes de partició**, donat un conjunt de dades amb n objectes o registres, construeixen k particions en les dades, on cada partició representa un clúster que ha de contenir com a mínim un objecte. Els mètodes *k-means* i *k-medoids* són exemples populars d'aquest tipus. A *k-means*, cada objecte pertany al clúster el valor mitjà del qual és més proper. Així, mentre que a *k-means* cada

clúster queda representat pel valor mitjà, a *k-medoids* és l'objecte més proper al centre que representa cada clúster, raó per la qual aquesta segona aproximació és més robusta davant del soroll i els *outliers*.

A R, el mètode *k-means* es pot aplicar mitjançant la funció `kmeans()`. L'exemple següent mostra el resultat d'aplicar aquesta funció sobre la base de dades `iris`, per a $k = 3$. L'última taula compara el resultat esperat amb l'obtingut pel mètode aplicat, on es pot observar que la classe «setosa» se separa adequadament, mentre que «versicolor» i «virginica» mostren cert encavalcament.

Exemple

```
> iris.cl<-iris
> iris.cl$Species<-NULL
> kmeans.res<-kmeans(iris.cl,3)
> table(iris$Species,kmeans.res$cluster)

             1  2  3
  setosa     0  0 50
  versicolor 48  2  0
  virginica  14 36  0
```

El mètode *k-medoids* se sol aplicar mitjançant l'algoritme PAM (*partitioning around medoids*, per les sigles en anglès), disponible en el paquet `cluster` a través de la funció `pam()`. Així mateix, el paquet `fpc` proporciona la funció `pamk()` on no cal fixar un valor per a *k*. L'exemple següent mostra el funcionament de les dues funcions.

Exemple

```
> kmedoids.res1<-pam(iris.cl,3)
> table(iris$Species,kmedoids.res1$cluster)

             1  2  3
  setosa     50  0  0
  versicolor  0 48  2
  virginica   0 14 36

> kmedoids.res2<-pamk(iris.cl)
> table(iris$Species,kmedoids.res2$pamobject$clustering)

             1  2
  setosa     50  0
  versicolor  1 49
  virginica   0 50
```

Del segon resultat, es pot confirmar l'encavalcament entre les classes «versicolor» i «virginica», ja que, com que no es fixa el nombre de clústers a 3, l'algoritme ha considerat que les dades es poden descriure només amb 2 grups.

D'altra banda, els **mètodes jeràrquics** apliquen una descomposició jeràrquica del conjunt de dades d'origen. Aquests mètodes poden ser de tipus ascendent (*bottom-up*) o descendent (*top-down)*, en funció de com es produeixi aquesta descomposició jeràrquica. En el primer cas, tots els objectes formen grups separats a l'inici del procés. Posteriorment, els grups semblants entre ells es van

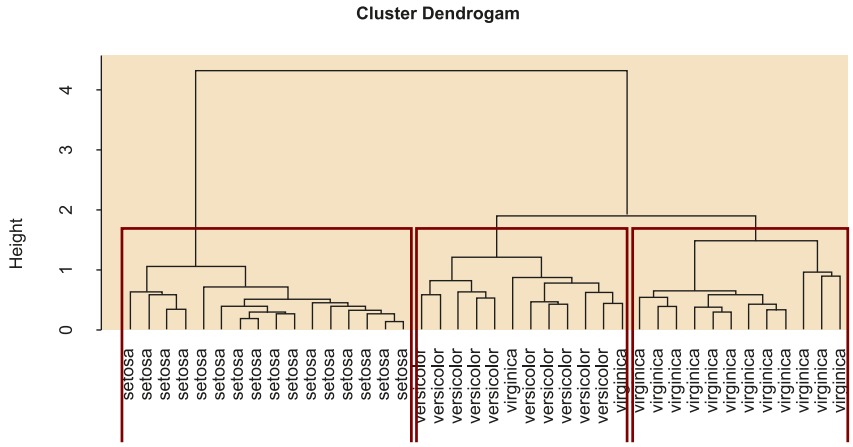
fusionant fins a formar un únic clúster (nivell superior de la jerarquia) o fins a arribar a una condició que acabi el procés. Contràriament, en els mètodes de tipus *top-down*, tots els objectes es troben en un mateix clúster i, a mesura que avança el procés, es van dividint en clústers més petits, fins que cada objecte forma un clúster o s'arriba a una condició de terminació.

A R, aquest tipus de *clustering* s'aplica mitjançant la funció `hclust()`, tal com es mostra en l'exemple següent, on s'analitza una submostra del conjunt de dades `iris`, amb l'objectiu de facilitar la representació i interpretació del resultat (figura 16).

Exemple

```
> id<-sample(1:dim(iris.cl)[1],40)
> irisSample<-iris.cl[id,]
> hc<-hclust(dist(irisSample),method="ave")
> hc

Call:
hclust(d = dist(iris.cl), method = "ave")

Cluster method   : average
Distance         : euclidean
Number of objects: 150

> plot(hc, hang = -1, labels = iris$Species[id])
> rect.hclust(hc, k = 3)
```

Figura 16. Resultat del *clustering* jeràrquic



Finalment, els **mètodes basats en la densitat** segueixen estenent un clúster donat mentre la densitat (nombre d'objectes) al «veïnat» superi algun llindar. Per exemple, per a cada registre contingut en un clúster donat, el veïnatge d'un radi donat ha de contenir com a mínim un nombre mínim de punts. Atès que aproximacions com a *k-means* solen definir clústers amb forma esfèrica i de mides semblants, aquests mètodes s'utilitzen generalment per filtrar soroll o valors extrems.

A R, es poden aplicar mitjançant la funció `dbscan()` del paquet `fpc`, especificant els valors d'`eps` (mida del veïnat) i `MinPts` (nombre mínim de punts).

Exemple

```
> ds <- dbscan(iris.cl, eps = 0.42, MinPts = 5)
table(ds$cluster, iris$Species)

      setosa versicolor virginica
  0   2          10         17
  1   48         0          0
  2   0          37         0
  3   0          3         33
```

En l'exemple, després de fixar una mida del veïnat de 0,42 i un mínim de 5 punts, es classifiquen correctament 48 plantes com a «setosa», 37 com a «versicolor» i 33 com a «virginica». La resta de plantes van ser classificades erròniament en altres clústers.

Tot i que no s'inclou un exemple al respecte, com en els mètodes de regressió i classificació, es podria predir el clúster al qual pertanyeria un nou conjunt de dades mitjançant la funció `predict()`.

3. Visualització de les dades

La **visualització de les dades** té com a objectiu comunicar la informació continguda en les dades de manera clara i eficaç mitjançant la representació gràfica. Aquesta etapa aprofita la capacitat del sistema visual humà per a la detecció de patrons i tendències.

Tot i que alguns mètodes de representació s'han anat citant al llarg del mòdul didàctic, aquest apartat pretén ser un resum d'aquelles tècniques de representació més comunes en les fases de neteja i anàlisi de les dades.

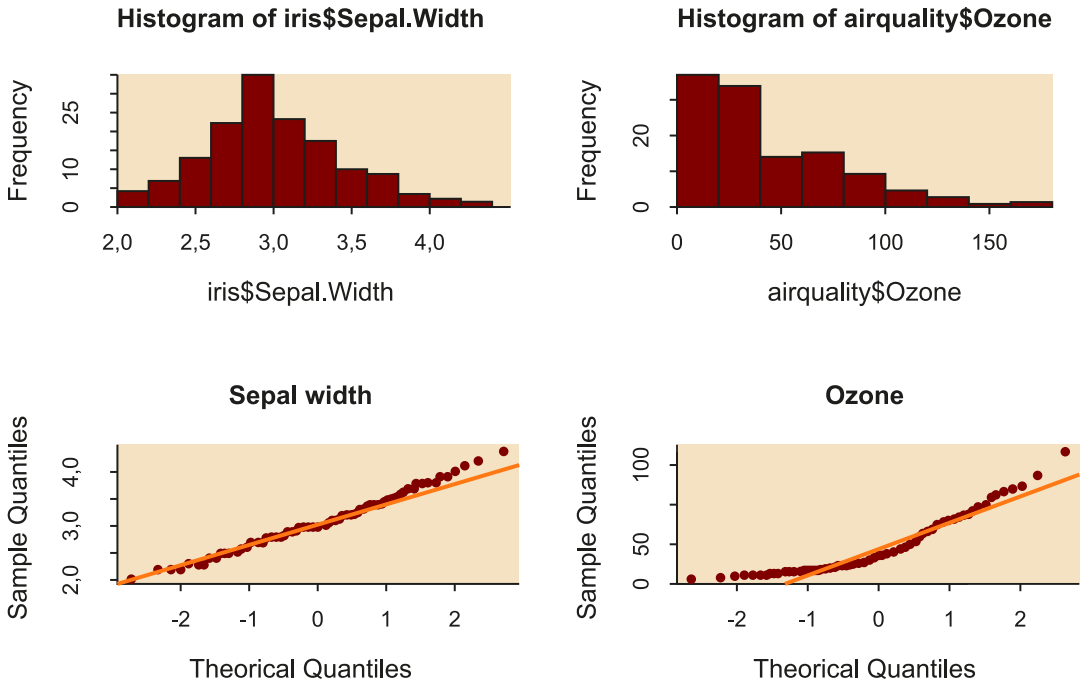
D'una banda, per a l'anàlisi de la normalitat, és molt habitual l'ús d'**histogrames** amb l'objectiu d'observar visualment si les dades semblen seguir una distribució normal. Un altre mètode àmpliament utilitzat en l'anàlisi de la normalitat són els **gràfics Q-Q** o gràfics de quantils teòrics. Aquests comparen els quantils de la distribució observada amb els quantils teòrics d'una distribució normal, de manera que com més s'aproximen les dades a una de normal, més alineats es mostren els seus punts a la recta.

L'exemple següent mostra l'histograma i el gràfic Q-Q de la variable `Sepal.Width` de les dades `iris` i de la variable `Ozone` de les dades `airquality`, mitjançant l'ús de les funcions `hist()`, `qqnorm()` i `qqline()`.

Exemple

```
> par(mfrow=c(2,2))
>
> hist(iris$Sepal.Width)
> hist(airquality$Ozone)
>
> qqnorm(iris$Sepal.Width, main="Iris")
> qqline(iris$Sepal.Width,col=2)
> qqnorm(airquality$Ozone, main="AirQuality")
> qqline(airquality$Ozone,col=2)
```

Figura 17. Histograma i gràfic Q-Q per a les variables `Sepal.Width` i `Ozone`



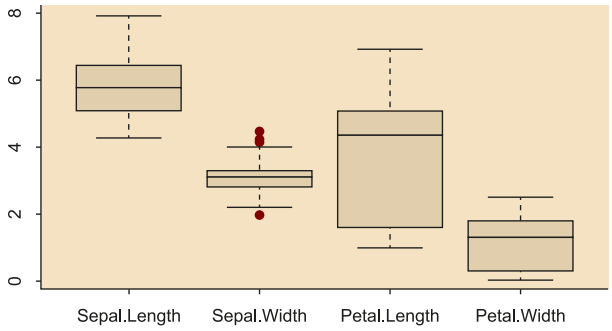
Tot i que la normalitat s'haurà de comprovar mitjançant proves estadístiques, es pot intuir que l'amplada del sèpal (`Sepal.Width`) sembla seguir una distribució normal, mentre que no serà així per a l'ozó (`Ozone`).

En l'anàlisi de valors extrems, també és habitual l'ús de *boxplots* per identificar *outliers* de manera visual. Tot i que la figura 8 ja mostra un exemple, a continuació es mostra el *boxplot* per a totes les variables del conjunt de dades `iris`, on només s'identifiquen valors extrems per a la variable `Sepal.Width`, representats amb cercles (figura 18).

Exemple

```
> boxplot(iris[,-5])
```

Figura 18. *Boxplots* per a les variables del conjunt de dades `iris`

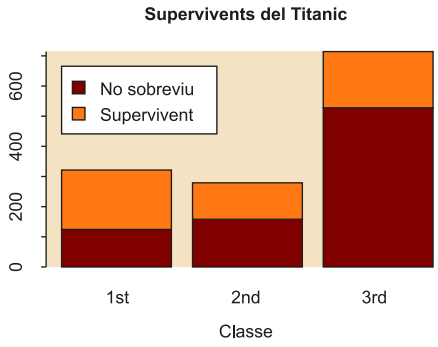


Els **gràfics de barres** (*barplots*) són especialment útils quan es treballa amb dades qualitatives (Jarman, 2013). El següent exemple representa la quantitat de supervivents i no supervivents del Titanic (conjunt de dades `TitanicSurvival`) en cadascuna de les categories de `Class`, mitjançant l'ús de la funció `barplot()`.

Exemple

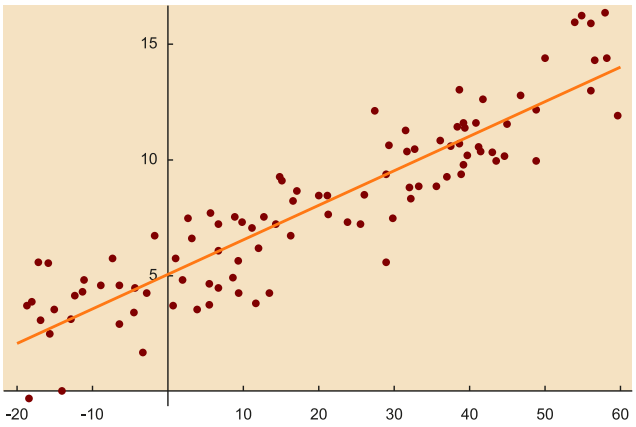
```
> titanic.data<-table(TitanicSurvival[,c(1,4)])
> barplot(titanic.data, main = "Supervivientes del Titanic", xlab = "Clase",col
= c("cadetblue4","aquamarine"))
> legend("topleft", c("No sobrevive","Superviviente"), fill =
c("cadetblue4","aquamarine"))
```

Figura 19. Gràfics de barres per representar la quantitat de supervivents i no supervivents en funció de la classe, en el conjunt de dades `TitanicSurvival`



Altres gràfics poden ajudar a interpretar els resultats de diferents models com les regressions (figura 20), les anàlisis de supervivència (veure apartat 2.3), els arbres de decisió o la importància de les variables en un model de classificació.

Figura 20. Exemple de regressió lineal



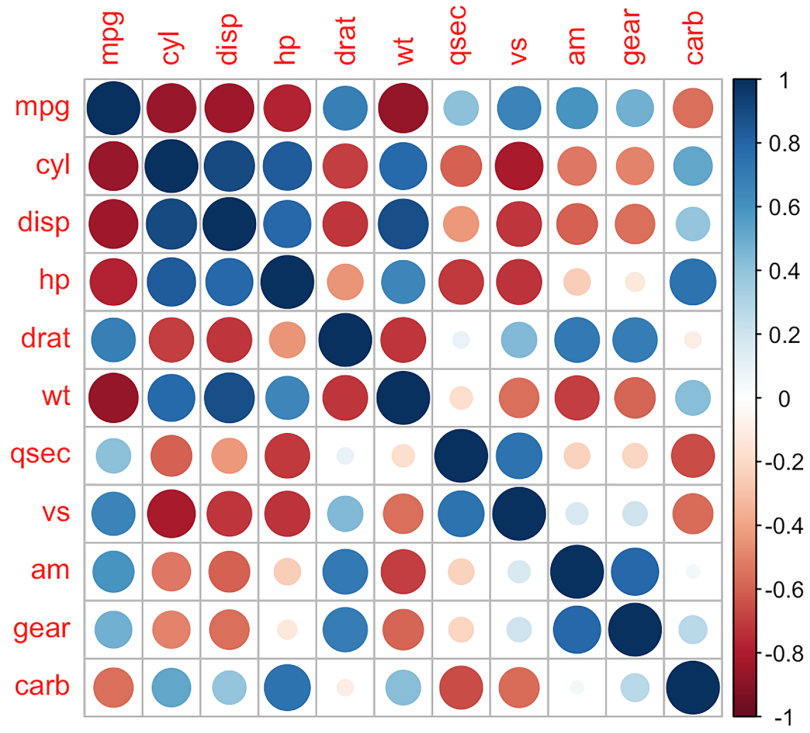
Font: *Viquipèdia.*

La funció `corrplot()`, del paquet `corrplot`, per exemple, permet representar gràficament les correlacions entre parells de variables en un conjunt de dades. Tot i que aquesta funció ofereix gran quantitat de possibilitats per a la representació d'aquestes correlacions, en l'exemple s'utilitza el mètode dels cercles, per al conjunt de dades `mtcars`.

Exemple

```
> corr.res<-cor(mtcars)
orrplot(corr.res,method="circle")
```

Figura 21. Gràfic de correlacions per al conjunt de dades `mtcars`



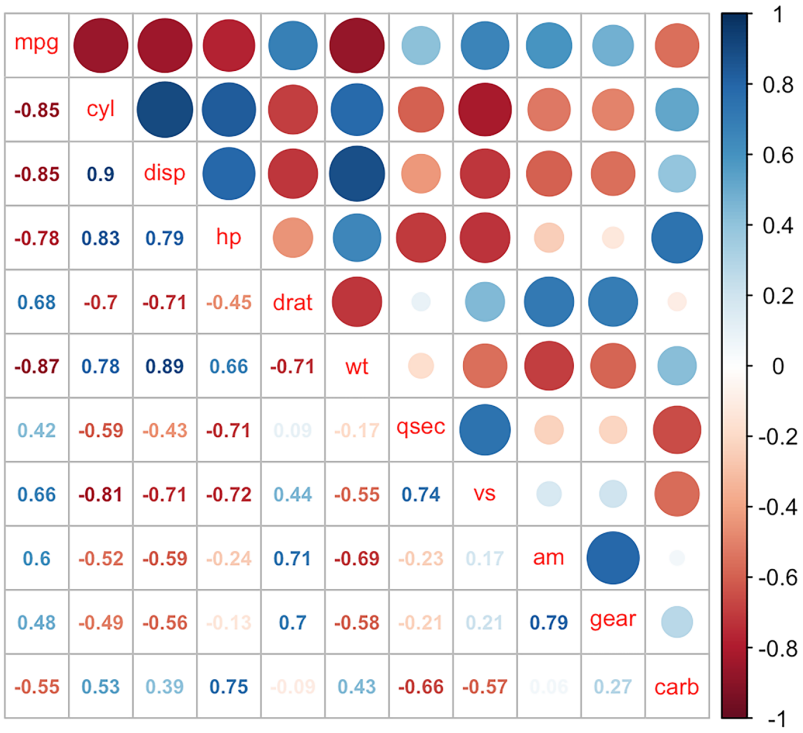
Com indica la barra lateral, el gràfic representa en vermell les correlacions negatives i en blau les positives. Així mateix, la mida i intensitat de cada un dels cercles indica el grau de correlació entre els parells de variables.

D'altra banda, la funció `corrplot.mixed()` permet barrejar dos mètodes de representació, com són els cercles i els valors numèrics. El següent exemple mostra el resultat per al conjunt de dades `mtcars`.

Exemple

```
> corrplot.mixed(corr.res,upper="circle",number.cex=.7,tl.cex=.8)
```

Figura 22. Gràfic mixt de correlacions per al conjunt de dades `mtcars`



Resum

En aquest mòdul didàctic s'han revisat els aspectes fonamentals relacionats amb la neteja i l'anàlisi de les dades. En primer lloc, s'ha presentat la utilitat i potencial de la neteja de dades o *data cleaning*, especialment útil quan es pretenen realitzar anàlisis posteriors amb l'objectiu de detectar tendències i patrons en les dades.

Així, en l'apartat "Neteja de dades", s'han revisat les principals etapes de neteja com la integració, la selecció i la reducció de les dades, tant en termes de dimensionalitat com de quantitat, destacant el mètode ACP com a tècnica comunament utilitzada per reduir la dimensionalitat de les dades. Posteriorment, s'han presentat alguns mètodes de conversió àmpliament utilitzats, com els diferents tipus de normalització per redimensionar les dades, les transformacions de Box-Cox per millorar la seva normalitat i homoscedasticitat, així com la discretització. Per acabar aquest apartat, s'han introduït diverses tècniques de detecció i correcció de dades perdudes, així com de valors extrems.

A l'apartat "Anàlisi de dades" s'han revisat els principals mètodes d'anàlisi de les dades. S'ha revisat l'estadística descriptiva i inferencial, així com les proves per contrast d'hipòtesis paramètriques i no paramètriques. Després de presentar les proves de Kolmogorov-Smirnov i Shapiro per comprovar la normalitat de les dades, i les proves de Levene i Fligner-Killeen per comprovar l'homoscedasticitat, s'han introduït les proves *t* de Student i Wilcoxon per comparar dos grups de dades i, finalment, ANOVA i Kruskal-Wallis per a les comparacions entre més de dos grups. Posteriorment, s'han presentat diversos models de regressió i correlacions, així com les anàlisis de supervivència mitjançant l'estimador de Kaplan-Meier, alguns models supervisats com el mètode de classificació i models no supervisats com els mètodes d'agrupament o *clustering*.

Finalment, a l'apartat "Visualització de les dades" s'han resumit algunes funcions útils per a la visualització de les dades, entesa com una etapa addicional i complementària a l'anàlisi de les dades.

Exercicis d'autoavaluació

1. Enumera i descriu breument les 6 etapes principals de la neteja de dades.

2. Explica la diferència entre la reducció de la dimensionalitat i de la quantitat. Posa un exemple de cada tipus i descriu breument en què consisteix.

3. Explica la diferència entre els conceptes de normalització i normalitat.

4. En què consisteix la regressió logística i en quins casos resulta més adequada que una regressió lineal?

5. Explica les similituds i diferències entre les correlacions de Pearson i Spearman.

6. Què són les dades censurades en l'anàlisi de supervivència? Posa'n un exemple.

7. Enumera i explica breument els diferents tipus de particions de les dades per a entrenament i test en classificació.

8. A partir de les dades `ToothGrowth`, disponibles a R, analitza:

a) Si la variable `len` segueix una distribució normal.

b) Si la variància de la variable `len` es manté constant per als grups definits per `supp`.

c) Compara els valors de la variable `len` entre els grups de dades definits per `supp` mitjançant la prova per contrast d'hipòtesis més adequada. Interpreta el resultat.

9. A partir de les dades `iris`, disponibles a R:

a) Analitza si les variables `Sepal.Length`, `Sepal.Width`, `Petal.Length` i `Petal.Width` segueixen una distribució normal.

b) Calcula la correlació entre parells de variables.

c) Representa gràficament aquestes correlacions, mostrant els valors numèrics en el triangle inferior de la figura, així com la representació mitjançant el·lipses en el triangle superior. Interpreta el resultat.

Solucionari

1. Les etapes principals de la neteja de dades són:

a) Integració: aquesta etapa consisteix en la combinació de dades procedents de diferents fonts, per tal de crear una estructura coherent i única, que contingui més quantitat d'informació.

b) Selecció: es basa en el filtratge de les dades que es troben dins d'un fitxer o base de dades amb l'objectiu de seleccionar només les dades d'interès. Per a això, s'apliquen criteris de cerca que ens permetin discernir, en el conjunt de dades, aquells que són realment necessaris per a l'anàlisi posterior.

c) Reducció: consisteix a aplicar certs algoritmes de tractament amb l'objectiu d'obtenir una representació reduïda de les dades, mantenint la integritat de la mostra original. Així, les anàlisis aplicades sobre la mostra de dades reduïda produiran els mateixos resultats (o molt semblants) que si s'apliquessin sobre la mostra total.

d) Conversió: són una sèrie de tècniques que tenen com a objectiu transformar o modificar el format original de les dades a un format més pla i comprensible, amb l'objectiu que l'anàlisi posterior sigui més eficient i els resultats obtinguts siguin més fàcilment interpretables.

e) Gestió de dades perdudes: representa un dels problemes més comuns trobats en bases de dades. Poden sorgir per diferents causes, per la qual cosa, depenent de la seva naturalesa, hi ha diferents solucions per resoldre aquest inconvenient. Per exemple, completar manualment els registres que falten, substituir el conjunt de valors perduts per una mateixa constant o etiqueta, substituir per una mateixa mesura de tendència central, o la implementació de mètodes probabilistes per predir els valors perduts.

f) Anàlisi de valors extrems: consisteix a identificar aquelles dades que es troben molt allunyades de la distribució normal d'una variable o població. Aquestes observacions es desvien tant de la resta que aixequen sospites sobre si van ser generades mitjançant el mateix mecanisme. Com les dades perdudes, poden aparèixer per diferents motius, de manera que, segons la seva naturalesa, s'apliquen diferents criteris per evitar que afectin negativament en el procés d'anàlisi.

2. Tot i que tots dos mètodes tenen com a objectiu reduir la quantitat de dades d'un conjunt dades de manera que es mantingui la integritat de la mostra original, la **reducció de la dimensionalitat** permet reduir el nombre d'atributs, mentre que la **reducció de la quantitat** redueix el nombre de mostres sota consideració.

Un exemple de mètode per reduir la dimensionalitat és l'**anàlisi de components principals** (ACP). Aquesta tècnica permet descriure un conjunt de dades de n atributs, en termes de m noves variables no correlacionades, o components principals, on $m < n$. Aquests components s'ordenen segons la quantitat de variància de les dades originals que descriuen.

D'altra banda, un exemple de mètode per reduir la quantitat de les dades és el *sampling*, ja que permet que un gran conjunt de dades sigui representat per un subconjunt de dades molt més petit, seleccionat de manera **aleatòria.**

3. La **normalització** és un tipus de conversió de les dades que permet reduir el biaix causat per la combinació de valors mesurats a diferents escales a l'hora d'ajustar-los a una escala comuna, típicament entre (-1,1) o entre (0,1). Depenent del context, aquesta normalització es pot aplicar mitjançant diferents mètodes, essent la normalització min-max i la normalització z-score els mètodes més comuns.

La **normalitat** fa referència a la propietat d'un conjunt de dades de seguir una distribució normal. És important destacar que, quan unes dades es normalitzen, s'està canviant la seva escala de representació, però això no té per què millorar la seva normalitat.

4. La **regressió logística** és un tipus d'anàlisi de regressió utilitzat per predir el resultat d'una variable dicotòmica dependent, en funció d'una sèrie de variables independents o predictores. Atès que aquest model estima les probabilitats d'ocurrència, en lloc d'utilitzar un model additiu que podria predir valors fora del rang (0,1), utilitza una escala transformada basada en una funció logística.

Per tant, quan la variable resultat (dependent) només pugui prendre dos valors, la regressió logística serà més adequada que la **regressió lineal**.

5. Tots dos coeficients de correlació són una mesura de l'associació entre dues variables. Aquests poden prendre valors entre -1 i 1, on els extrems indiquen una correlació perfecta

i el 0 indica l'absència de correlació. El seu signe serà negatiu quan valors elevats d'una variable s'associïn a valors petits de l'altra, i serà positiu quan ambdues variables tendeixin a incrementar o disminuir simultàniament.

El coeficient de **correlació de Pearson** és el més utilitzat entre variables relacionades linealment. No obstant això, per poder-se aplicar, requereix que la distribució de les dues variables sigui normal, així com que es compleixi el criteri d'homoscedasticitat.

D'altra banda, la **correlació de Spearman** apareix com una alternativa no paramètrica que mesura el grau de dependència entre dues variables. Aquest mètode no comporta cap suposició sobre la distribució de les dades, tot i que les variables que es volen comparar s'han de mesurar almenys en una escala ordinal.

6. En l'anàlisi de supervivència, es diu que les dades d'un conjunt estan **censurades** quan no se'n coneix la supervivència exacta, ja que va més enllà del període d'estudi.

Per exemple, en un estudi sobre la supervivència d'una sèrie de pacients després d'aplicar un tractament en particular, si les dades es recullen tan sols un any després de l'aplicació d'aquest tractament, aquells pacients que segueixin vius després de l'any seran dades censurades. Com que no es recullen dades més enllà d'aquest període no en podrem saber la supervivència exacta.

7. En el **mètode d'exclusió** (*holdout*), les dades es divideixen aleatòriament en dos conjunts independents, el d'entrenament i el de test. Típicament, dos terços de les dades s'assignen al conjunt d'entrenament i el terç restant es reserva per testejar el model.

El **submostratge aleatori** (*random subsampling*) és una variació del mètode anterior, ja que s'aplica la mateixa tècnica *k* vegades per posteriorment estimar la precisió global del model com la mitjana de les precisions obtingudes de cada iteració.

En la **validació creuada** (*cross-validation*) de tipus *k-fold* les dades originals es divideixen aleatòriament en *k* subconjunts (*folds*) mútuament exclusius i de mides semblants. L'entrenament i testeig es realitzen *k* vegades, a partir de totes les combinacions possibles de *k-1* subconjunts per a entrenament i deixant el subconjunt restant per testejar el model. En aquest cas, l'exactitud es calcula com el nombre total de classificacions correctes en les *k* iteracions, dividit pel nombre total de mostres en el conjunt de dades original.

El *leave-one-out* és un cas especial de validació creuada de tipus *k-fold* on *k* s'ajusta al nombre de mostres del conjunt de dades original. A cada iteració, només omet una de les mostres en la fase d'entrenament, per a posteriorment utilitzar-la en el testeig del model.

Finalment, en la **validació creuadaestratificada**, els *folds* s'estratifiquen de manera que la distribució de classes a cada *fold* sigui aproximadament la mateixa que en el conjunt de dades original.

8. a) S'aplica el test de Shapiro i es comprova que les dades segueixen una distribució normal (p-valor > 0,05).

```
> shapiro.test(ToothGrowth$len)

    Shapiro-Wilk normality test

data:  ToothGrowth$len
W = 0.96743, p-value = 0.1091
```

b) Atès que les dades segueixen una distribució normal, s'aplica el test de Levene i es comprova que també presenten homoscedasticitat (p-valor > 0,05).

```
> leveneTest(ToothGrowth$len ~ ToothGrowth$supp)

Levene's Test for Homogeneity of Variance (center = median)
```

```
       Df F value Pr(>F)
group    1  1.2136 0.2752
       58
```

c) Atès que es compleixen les condicions de normalitat i homoscedasticitat, s'aplica la prova *t* de Student, comprovant que no hi ha diferències significatives entre els grups de dades (p-valor > 0,05). Per tant, el mètode de subministrament de la vitamina C sembla no afectar en el creixement dental.

```
> t.test(ToothGrowth$len ~ ToothGrowth$supp)

    Welch Two Sample t-test

data:  ToothGrowth$len by ToothGrowth$supp
t = 1.9153, df = 55.309, p-value = 0.06063
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1710156  7.5710156
sample estimates:
mean in group OJ mean in group VC
     20.66333      16.96333
```

9. a) S'aplica el test de Shapiro i es comprova que només Sepal.Width segueix una distribució normal (p-valor > 0,05).

```
> shapiro.test(iris$Sepal.Length)

    Shapiro-Wilk normality test

data:  iris$Sepal.Length
W = 0.97609, p-value = 0.01018

> shapiro.test(iris$Sepal.Width)

    Shapiro-Wilk normality test

data:  iris$Sepal.Width
W = 0.98492, p-value = 0.1012

> shapiro.test(iris$Petal.Length)

    Shapiro-Wilk normality test

data:  iris$Petal.Length
```

```
W = 0.87627, p-value = 7.412e-10


> shapiro.test(iris$Petal.Width)


    Shapiro-Wilk normality test


data:  iris$Petal.Width
W = 0.90183, p-value = 1.68e-08
```

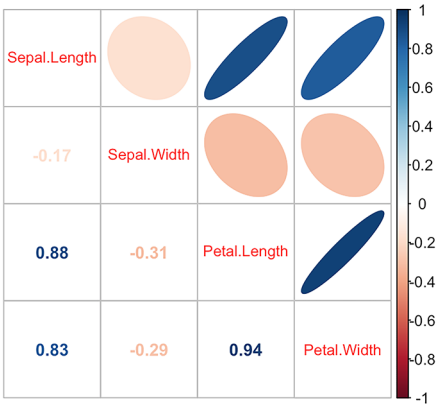
b) Atès que les dades en general no segueixen una distribució normal, s'aplica el test de Spearman per calcular les correlacions entre parells de variables.

```
> corr.res<-cor(iris[,-5], method="spearman")
> corr.res
             Sepal.Length  Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000   -0.1667777 0.8818981   0.8342888
Sepal.Width    -0.1667777   1.0000000  -0.3096351  -0.2890317
Petal.Length    0.8818981   -0.3096351 1.0000000   0.9376668
Petal.Width     0.8342888   -0.2890317 0.9376668   1.0000000
```

c) El següent gràfic mixt representa les correlacions entre parells de variables.

```
> corrplot.mixed(corr.res,upper="ellipse",number.cex=.9,tl.cex=.8)
```

Figura 23. Gràfic mixt de correlacions per al conjunt de dades `iris`



Com indica la barra lateral, el gràfic representa en vermell les correlacions negatives i en blau les positives. Així mateix, l'amplada i intensitat de cadascuna de les el·lipses indica el grau de correlació entre els parells de variables, essent les el·lipses més estretes i intenses les que mostren correlacions més elevades.

Així, la correlació més important s'observa entre `Petal.Lengthvs` *vs* `Petal.Width` (0,94), seguida de `Petal.Length` *vs* `Septal.Length` (0,88) i `Petal.Width` *vs* `Sepal.Length` (0,83). La resta de parells de variables presenten correlacions negatives poc significatives.

Glossari

ACP *m.* Vegeu ***principal component analysis***.

AIC Vegeu **criteri d'informació d'Akaike**.

analysis of variance *m.* Extensió de la prova *t* amb l'objectiu de comparar les mitjanes entre més de dos grups de dades. sigla **ANOVA**.

ANOVA *m.* Vegeu ***analysis of variance***.

area under the curve *f.*Àrea sota la corba ROC.
sigla **AUC**.

AUC *f.* Vegeu ***area under the curve***.

criteri d'informació d'Akaike *m.*Criteri per mesurar la bondat d'un model.
sigla **AIC**.

clustering És un tipus de tècnica d'aprenentatge no supervisat que divideix els registres de dades en grups, o clústers, de manera que els registres dins d'un mateix clúster siguin semblants entre ells i diferents dels registres d'altres clústers. La similitud es defineix generalment en termes de com de prop es troben els registres a l'espai, basant-se en una funció de distància.

coeficient de determinació *m.* (R^2 o *R-squared*) Mesura de qualitat del model que pren valors entre 0 i 1, i la proporció de variació dels resultats que pot explicar-se pel model.

corba ROC *f.* Vegeu ***receiver operating characteristic***.

exactitud *f.* Es refereix a com de prop del valor real es troba el valor mesurat. En classificació, fa referència a la proporció de mostres correctament classificades per a un classificador concret.

heteroscedasticitat *f.* Un model de regressió lineal presenta heteroscedasticitat quan la variància dels errors no és constant en totes les observacions realitzades.

histograma **m.** És una representació gràfica d'una variable en forma de barres, on la superfície de cada barra és proporcional a la freqüència dels valors representats.

homoscedasticitat *f.* Un model predictiu presenta homoscedasticitat quan la variància de l'error condicional a les variables explicatives és constant al llarg de les observacions.

IMC *m.* Vegeu **índex de massa corporal**.

imputar Substituir valors no informats en una observació.

índex de massa corporal *m.* Concepte que relaciona l'altura i el pes d'un individu.
sigla **IMC**.

kNN Vegeu ***k-Nearest neighbours***.

k-Nearest neighbours Permet predir valors en conjunts de dades multidimensionals formats per dades mixtes (continus, discrets, ordinals o nominals).
sigla **kNN**.

NA Vegeu ***not available***.

not available És un indicador d'una dada buida. sigla **NA**.

potència estadística *f.* La potència d'una prova estadística o el poder estadístic és la probabilitat que la hipòtesi nul·la sigui rebutjada quan la hipòtesi alternativa és verdadera (és a dir, la probabilitat de no cometre un error de Tipus II).

principal component analysis *m.* Procediment estadístic que utilitza una transformació ortogonal per convertir un conjunt d'observacions o variables possiblement correlacionades en un conjunt de valors de variables linealment no correlacionades anomenades *components principals*.
sigla **ACP**.

receiver operating characteristic És una representació gràfica de la sensibilitat enfront de l'especificitat per a un sistema classificador binari segons es varia el llindar de discriminació. Una altra interpretació d'aquest gràfic és la representació de la raó o ràtio de vertaders positius enfront de la raó o ràtio de falsos positius també segons es varia el llindar de discriminació (valor a partir del qual vam decidir que un cas és un positiu).
sigla **ROC**.

SRSWOR *f.* Sigles en anglès de mostra aleatòria simple sense substitució.

SRSWR *f.* Sigles en anglès de mostra aleatòria simple amb substitució.

mida de l'efecte *m.* En estadística, la mida de l'efecte és una mesura de la força d'un fenomen (per exemple, el canvi en el resultat després d'una intervenció experimental). La mida de l'efecte calculat a partir de dades és una estadística descriptiva que transmet la magnitud estimada d'una relació sense fer cap declaració sobre si la relació aparent en les dades reflecteix una veritable relació a la població.

taxa d'error de Tipus I *f.* L'error de Tipus I, també anomenat error de tipus alfa (α) o fals positiu, és l'error que es comet quan l'investigador rebutja la hipòtesi nul·la essent aquesta veritable a la població.

taxa d'error de Tipus II *f.* L'error de Tipus II, també anomenat error de tipus beta (β) (β és la probabilitat que existeixi aquest error) o fals negatiu, es comet quan l'investigador no rebutja la hipòtesi nul·la essent aquesta falsa a la població.

Bibliografia

Cook, R. Dennis (1977, febrer). *«Detection of Influential Observations in Linear Regression»*. Technometrics, American Statistical Association (vol. 19, núm. 1, pàg. 15-18).

Dale, Kyran (2016). *Data Visualization with Python and JavaScript: Scrape, Clean, Explore & Transform Your Data*. Sebastopol, CA: O'Reilly Media.

Dalgaard, Peter (2008). *Introductory statistics with R*. Berlín: Springer.

Han, Jiawei; Kamber, Micheline; Pei, Jian (2011). *Data Mining: Concepts and Techniques*. Waltham: Elsevier.

Jarman, Kristin H. (2013). *The art of data analysis: how to answer almost any question using basic statistics*. Hoboken, NJ: Wiley.

Mahalanobis, Prasanta C. (1936, gener). *«On the generalized distance in statistics»*. Proceedings of the National Institute of Science of India (vol. II, núm. 1).

McKinney, Wes (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and Ipython*. Sebastopol, CA: O'Reilly Media.

Newton, Rae R.; Rudestam, Kjell E. (1999). *Your Statistical Consultant: Answers to Your Data Analysis Questions*. Thousand Oaks, CA: Sage Publications.

Osborne, Jason W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Thousand Oaks, CA: Sage Publications.

Osborne, Jason W. (2010, març). *«Data cleaning basics: Best practices in dealing with extreme scores»*. Newborn and Infant Nursing Reviews (vol. 10, núm. 1, pàg. 37-43).

Osborne, Jason W.; Kocher, Brady, Tillman, David (2012). *«Sweating the small stuff: do authors in APA journals clean data or test assumptions (and should anyone care if they do)?»* [conferencia]. En: Annual meeting of the Eastern Education Research Association (2012: Hilton Head, SC).

Squire, Megan (2015). *Clean Data*. Birmingham: Packt Publishing.

Stekhoven, Daniel J.; Peter Bühlmann (2011, gener). *«MissForest: Non-parametric missing value imputation for mixed-type data»*. Bioinformatics (vol. 28, núm. 1, pàg. 112-118).