Introducció a l'estadística

PID_00247914

Ester Bernadó

Temps mínim de dedicació recomanat: 2 hores

Universitat Oberta de Catalunya

Índex

Introducció

Són molts els experts que etiqueten els nostres temps amb el terme *era de la informació*, també anomenada *era de la informació i de les comunicacions* o *era digital*. I probablement hi estareu molt d'acord, ja que l'impacte de la informació i les tecnologies de la informació a les nostres vides no passa desapercebut. Són múltiples les activitats diàries que estan influïdes per les tecnologies (per prudència, no farem servir el terme *determinades*). Per exemple, el mòbil i les seves nombroses aplicacions, les xarxes socials, el correu electrònic o les reunions virtuals. Ja no podem concebre maneres de treballar, de desplaçar-nos, de comunicar-nos, d'establir relacions... sense l'ús d'aquestes tecnologies. Algú es va preguntar un cop: «On es buscava la informació abans de Google?».

Amb l'era de la informació s'està produint una altra revolució, subjacent i sovint amagada als nostres quefers diaris. Aquesta és l'era de les dades (Mayer-Schönberger i Cukier, 2013). I és que tot allò que realitzem amb les tecnologies, com enviar un correu, publicar un tema en una xarxa social, fer servir un navegador, fer una transacció bancària o una compra al supermercat queda registrat. Dit d'aquesta manera, potser l'estudiant ha d'establir una analogia amb la distopia de George Orwell narrada a la seva novel·la *1984*, on es descriu una societat permanentment controlada pel Gran Germà, el qual vigila la població mitjançant diferents pantalles. No és la intenció d'aquest mòdul endinsar-nos en un debat com aquest, sinó despertar l'interès per les dades i per les múltiples oportunitats que la seva anàlisi ofereix per millorar la qualitat de vida de les persones, el benefici econòmic d'un negoci, els diagnòstics mèdics o la comprensió de nosaltres mateixos, per citar tan sols alguns exemples. Expliquen Mayer-Schönberger i Cukier (2013) que es pot predir en temps real l'avanç de la grip als Estats Units per les recerques a Google, a diferència de les prediccions amb dades epidemiològiques que només són capaces de detectar un brot de grip amb dues setmanes de retard.

Més enllà de les dades personals registrades, hi ha un gran volum de dades que es capturen a partir de diferents dispositius. El concepte de l'*internet de les coses* inclou qualsevol dispositiu físic o virtual que sigui capaç de recol·lectar dades i compartir-les en xarxa, com edificis, vehicles i, en general, qualsevol dispositiu electrònic, sensors i programari. Amb això s'obre la porta a un món d'aplicacions infinites, com els edificis intel·ligents, l'optimització del transport públic, la gestió intel·ligent de la xarxa elèctrica, el control i la gestió preventiva de trànsit o les ciutats intel·ligents.

Són tres els elements bàsics d'aquesta era de les dades:

- Les dades en si, o bancs de dades, on s'emmagatzema la informació.

- Els algoritmes o programari computacional, capaç d'extreure informació d'interès d'aquestes dades.

- Les aplicacions, o allò que es fa a partir d'aquestes anàlisis.

El gran avanç tecnològic ha propiciat l'emmagatzematge d'enormes volums de dades i una gran capacitat computacional per realitzar anàlisis automàtiques i eficients. Això ha generat un gran interès en les infinites possibilitats d'aplicació.

L'interès per les dades no neix amb l'era de la informació. S'origina amb la curiositat humana per comprendre el món en què vivim, predir-lo i dominar-lo. Matemàtics, astrònoms, físics, antropòlegs, biòlegs i historiadors han recol·lectat dades, realitzat anàlisis i extret conclusions que han permès el progrés de la ciència i la societat. La història de Johannes Kepler i Tycho Brahe n'és un bon exemple. Tycho Brahe (1546-1601) va ser un noble danès que va dedicar gran part de la seva vida a fer observacions precises de l'univers, abans de la invenció del telescopi. Va observar i anotar les posicions relatives dels planetes d'una manera molt més precisa que les realitzades en l'època. Johannes Kepler (1571-1630) va ser un matemàtic i astrònom que volia demostrar l'estructura geomètrica perfecta i divina de les òrbites dels planetes coneguts fins llavors. Però no disposava de dades suficients per demostrar l'òrbita circular i concèntrica dels planetes de la seva teoria. La trobada entre Brahe i Kepler va permetre la unió d'unes observacions sistemàtiques i precises realitzades durant trenta-cinc anys, amb la passió per extreure un model d'aquestes dades. A la mort de Brahe el 1601, Kepler va poder disposar de les seves observacions. Kepler es va esforçar per trobar un model matemàtic que encaixés en la teoria, però no aconseguia explicar les observacions de Brahe. Obstinat i perseverant, no va voler menysprear un error de vuit minuts d'arc en una òrbita ni donar les observacions per errònies. Així que, finalment, després de diverses iteracions, va renunciar a l'òrbita circular i va provar amb una òrbita el·líptica, amb la qual va arribar a enunciar les conegudes tres lleis de Kepler.

Al cor de les anàlisis de dades se situen les matemàtiques i, en especial, l'estadística. El terme actual *estadística*, introduït originàriament per l'alemany Gottfried Achenwall com *statistik* el 1749, prové del llatí *statisticum collegium* (consell d'estat) i del seu derivat italià *statista* (home d'estat, polític). Es referia a l'anàlisi de dades d'estat, especialment orientada a la recol·lecció sistemàtica de dades demogràfiques i econòmiques per part dels estats. De fet, aquesta orientació es va originar ja en les civilitzacions antigues com l'egípcia (3050 aC), la xinesa (2200 aC) i l'antiga Roma, on es recol·lectaven censos demogràfics per a la planificació de l'agricultura i l'economia (com la captació d'impostos). L'estadística és una disciplina en constant evolució que s'ha enriquit amb diferents aproximacions, tècniques i procediments.

Un dels censos més famosos de la història romana i potser de tota la història occidental precisament es va dur a terme l'any 0. La Bíblia narra que Maria i Josep viatjaven cap a Betlem per formar part de les dades censals romanes quan, de camí, Maria va donar a llum a Jesús. Així és com el cens romà i la incipient estadística van ser crucials per al naixement de la cristiandat.

Actualment, l'estadística estableix les bases de l'anàlisi de dades, tal com veurem al llarg d'aquest mòdul. En primer lloc, es descriu l'estadística com la ciència de les dades, definició que es matisa en els apartats següents mitjançant les principals dimensions d'aplicació de l'estadística i la distinció entre estadística descriptiva i inferencial. Ja que l'estadística és un enfocament necessari per a l'anàlisi de dades, però no exclusiu, s'introdueix el concepte de mineria de dades i *machine learning* (aprenentatge automàtic) per la seva rellevància i actualitat dins de l'anàlisi de dades. Cal definir i relacionar l'estadística i la mineria de dades, ja que són disciplines amb un elevat grau de relació, tant en les seves aplicacions com en l'ús d'alguns mètodes, alhora que provenen d'enfocaments relativament diferents. L'objectiu d'aquest capítol és centrar les bases per a la comprensió general de l'estadística en el marc de l'anàlisi de dades i preparar l'estudiant per a l'aprofundiment en les seves teories i mètodes.

1. Què és l'estadística?

Històricament, l'estadística ha estat la ciència de recol·lecció, anàlisi, interpretació, presentació i organització de les dades. Alguns experts simplement defineixen el terme *estadística* com la ciència de les dades. Però què són les dades? Fent una simplificació, les dades són nombres en context. I és que l'estadística va més enllà de realitzar càlculs sobre les dades. Consisteix a interpretar aquests càlculs en el context en què es produeixen amb l'objectiu de descriure la informació inherent, predir comportaments futurs o prendre decisions. Moore, McCabe i Craig (2012) ho exemplifiquen de la manera següent:

> «Calcular la media y desviación del peso al nacer de 1.000 niños es simple aritmética, interpretar además su significado es estadística. La estadística se fundamenta en la teoría de probabilidades, que describe el modelo matemático inherente a los fenómenos aleatorios. La estadística también involucra el juicio, que es necesario para aplicar el procedimiento adecuado en función de las condiciones del problema.»
>
> Moore, McCabe i Craig (2012)

Aprofundim una mica més en les dades. Sovint es confon el terme *dades* amb el d'*informació*. En efecte, les dades són una font d'informació, però aquesta informació no és òbvia, ni immediata. Imaginem, per exemple, un fitxer de dades sobre els clients d'un banc als quals se'ls concedeix un crèdit, tal com mostra la taula 1. Suposem que disposem de dos mil casos (registres) com els descrits. Quin tipus d'informació proporciona la taula? A simple vista, és un llistat de dades que necessita ser analitzat, interpretat i presentat en una forma que aporti informació útil.

Taula 1. Dades de sol·licitud de préstecs d'una entitat bancària

ID	Nom	Edat	Sexe	Estat civil	Nombre de fills	Nivell salarial	Crèdit sol·licitat	Préstec hipotecari	Motiu de préstec
567	José Pérez	46	H	C	2	3	20.000	65.000	Vehicle
765	María Sol	34	M	S	0	2	5.000	0	Estudis
965	Simón Martín	52	H	C	2	4	350.000	0	Reformes

L'estadística permet extreure informació de les dades. Per exemple, algunes de les preguntes que es poden plantejar són:

- Quants fills tenen de mitjana les persones que demanen crèdits?
- Quin és l'estat civil més freqüent entre les persones que demanen crèdits?

- L'edat mitjana de les persones que sol·liciten un crèdit és de trenta-cinc anys?
- Les dones demanen crèdits d'un valor superior al dels homes?
- Hi ha relació entre el nivell d'ingressos i la quantitat sol·licitada del crèdit?
- L'import del crèdit sol·licitat és diferent segons el motiu de préstec?

En definitiva, les dades en si no aporten informació. És mitjançant l'ús de tècniques d'anàlisi de dades que podem plantejar preguntes sobre les dades i saber-ne les corresponents respostes. Com veurem més endavant, aquestes preguntes es formulen en forma d'hipòtesis que se sotmeten a prova mitjançant l'ús de les tècniques estadístiques adequades, que porten al rebuig o a l'acceptació de les hipòtesis.

Per exemple, es pot formular la hipòtesi següent: les dones demanen crèdits de valor superior al dels homes.

A partir de la mostra disponible, es posa a prova la hipòtesi, i el resultat serà el rebuig o l'acceptació de la hipòtesi amb un determinat nivell de confiança (que s'expressa com un percentatge, com, per exemple, amb un 95% de nivell de confiança).

Així és com, a grans trets i d'una manera molt simplificada, l'estadística realitza una extracció d'informació de les dades. De fet, hi ha un conjunt concret de preguntes (o hipòtesis) que es poden fer sobre les dades. Sabent aquests tipus de preguntes, en quins casos s'han d'aplicar (les condicions d'aplicació) i la interpretació de les respostes, tenim un espectre amplíssim d'extracció d'informació sobre les dades.

Dèiem en la definició que l'estadística també intervé en la recol·lecció de les dades. Sovint, les dades de què es disposa per a l'anàlisi estan prèviament guardades en bases de dades, i són les dades que es fan servir per la seva disponibilitat. Altres vegades, és possible dissenyar la recollida de dades, i és en aquests casos que l'estadística pot orientar en la recollida d'una mostra representativa de la població d'interès. Així mateix, l'estadística també aporta eines addicionals d'interpretació de la informació mitjançant l'ús de gràfics i visualitzacions.

L'objectiu d'un curs d'estadística seria el de dotar l'estudiant de les eines d'anàlisi estadística, amb les quals serà capaç de fer les preguntes adequades, recollir les dades necessàries, aplicar les tècniques adequades a cada situació i extreure'n les interpretacions correctes, utilitzant visualitzacions si és necessari i sabent jutjar, a més, els errors o marges de confiança de les conclusions obtingudes.

2. Aplicacions de l'estadística

L'estadística és present en multitud de dominis: les ciències socials, els negocis, el diagnòstic i el tractament mèdic, el control de qualitat dels productes, la predicció meteorològica, l'educació, la investigació, i un llarg etcètera. Allà on hi ha dades, allà tenim l'estadística. Newbold (1997) agrupa en sis tipus les aplicacions de l'estadística. Comentàvem abans que el tipus de preguntes o anàlisis estadístiques que es poden fer sobre les dades era finit. De forma anàloga, Newbold agrupa aquests tipus de preguntes o anàlisis en sis:

- extracció de conclusions de variables numèriques,
- gestió de la incertesa,
- anàlisi de relacions,
- mostreig,
- predicció,
- presa de decisions amb incertesa.

En els apartats següents es pot veure una breu descripció de cada tipus juntament amb diversos exemples.

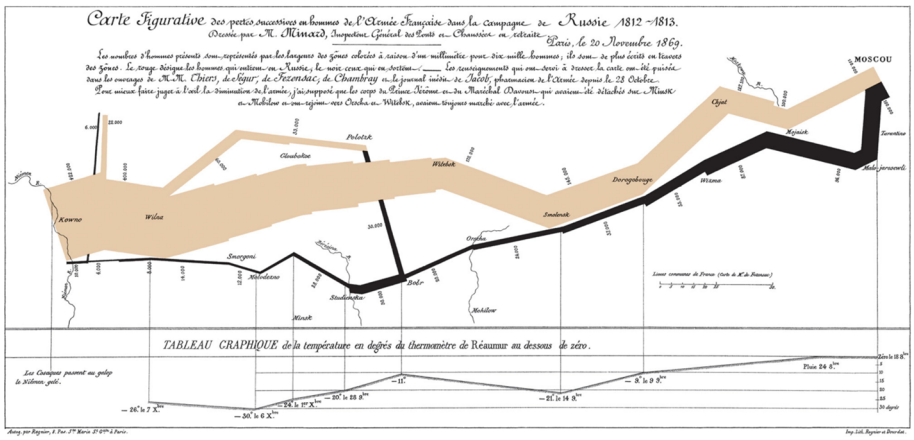
2.1. Extracció de conclusions de variables numèriques

Una part important de l'estadística gestiona dades numèriques recol·lectades en forma de llista de dades i que sovint sol tenir un gran volum. La funció de l'estadística és extreure i sintetitzar les característiques fonamentals d'aquesta llista de dades numèriques.

L'interès per les dades numèriques i estadístiques bàsiques és molt antic. A l'antiguitat, els babilonis, els xinesos, els egipcis, els grecs i els romans van fer censos de població, amb els quals pretenien estimar quants diners podien recollir-se amb els impostos, calcular quants soldats es podien reclutar per a un exèrcit o quant de menjar caldria (Rooney, 2009).

L'any 1662, John Graunt, considerat el pare de la demografia i de l'estadística, va publicar unes estadístiques sobre la mortalitat a Londres. Va mostrar el nombre de morts per malaltia com un intent d'alleujar l'ansietat de la població amb relació a determinades malalties. També va ser el primer que va documentar que naixien més nens que nenes. Va proporcionar una estimació de la població de Londres, amb la qual cosa va demostrar el seu ràpid creixement. Va mostrar que la disminució de la població en una plaga es compensava amb un ràpid creixement a causa de l'augment dels naixements (Rothman, 1996). Una altra contribució històrica notable és el gràfic de Charles Minard sobre la campanya de Napoleó a Rússia l'any 1812. El gràfic següent mostra visual-

ment la mida de l'exèrcit en l'anada i en la tornada al llarg dels quilòmetres recorreguts i la temperatura. Només quatre de cada cent soldats van tornar de la incursió napoleònica (Rooney, 2009).



Les mesures més habituals per resumir la informació numèrica són les de mitjana, moda o valor més freqüent, les mesures de dispersió i els rangs interquartílics. S'acompanyen sovint de gràfics típics com els histogrames o diagrames de caixa.

2.2. Gestió de la incertesa

Tomeo i Uña (2003) defineixen l'estadística com:

> «La ciencia que utiliza los números para el estudio de las leyes que dependen del azar, tratando de descubrir mediante el razonamiento inductivo la causa general a que obedece el modelo particular analizado.»
>
> Tomeo i Uña (2003)

Si parem atenció a la primera part de la definició, l'estudi de l'atzar o els fenòmens aleatoris es refereix a l'estudi d'aquells fenòmens el resultat dels quals és diferent, encara que es produeixi en les mateixes condicions, en contraposició als fenòmens *deterministes*, que produeixen el mateix resultat davant les mateixes condicions.

Estudi de l'atzar

El llançament d'una moneda o un dau no dona sempre el mateix resultat.

En aquests casos, en lloc de certesa, parlem de probabilitats: la probabilitat que surti cara en llançar una moneda, o un sis en llançar un dau. Newbold es refereix a l'estadística com la ciència de la incertesa.

Galileu afirmava que la llei de l'atzar era una mostra de la incapacitat humana. Segons Galileu, el que fa que el llançament d'una moneda generi un resultat impredictible és que no s'ha llançat la moneda en idèntiques condicions (Tomeo i Uña, 2003).

2.3. Anàlisi de relacions

L'anàlisi de relacions entre les variables d'un problema és una de les qüestions més freqüents que aborda l'estadística.

Un exemple il·lustratiu es va presentar arran de la pregunta sobre si existia relació entre el càncer de mama i la presa d'anticonceptius orals (Armitage, Berry i Matthews, 2002, pàg. 4). Per donar-hi resposta, el 1990 l'Organització Mundial de la Salut va realitzar un estudi en dotze centres de deu països diferents amb el títol *Collaborative Study of Neoplasia and Steroid Contraceptives*. A cada hospital es van escollir casos de dones amb càncer de mama que compliren determinats requisits d'edat i residencials. Una altra mostra de control es va prendre a partir de dones que van visitar el mateix hospital, amb criteris d'edat i residencials similars, i que no prenien anticonceptius orals. L'estudi va incloure 2.116 casos i 13.072 casos de control. Es van estudiar altres variables com l'edat, l'edat del primer fill, l'índex socioeconòmic i l'historial familiar de càncer de mama. El risc de càncer de mama per usuàries d'anticonceptius orals va ser identificat com 1,15 més gran que les que no prenien anticonceptius, la qual cosa va demostrar ser una associació molt feble en comparació amb altres factors més influents.

L'anàlisi de relacions permet estudiar si les variables identificades d'un determinat àmbit o domini varien conjuntament. Així s'identifiquen variables que estan associades entre elles.

Anàlisi de relacions

El consum elèctric de les llars és superior si el dia és molt calorós o fred; la contaminació atmosfèrica d'una ciutat està relacionada amb el nombre de cotxes que circulen i amb les condicions atmosfèriques.

Habitualment, s'usa el terme específic *anàlisi de correlacions* per denominar aquest tipus d'estudis. Un cas particular d'aquest tipus és l'anàlisi de correlacions lineals, que mesura el grau de dependència lineal entre un conjunt de variables. Així mateix, convé fer una distinció entre correlació i causalitat. Si dues variables es demostren correlacionades, no vol dir que hi hagi una relació de causalitat, és a dir, que una d'elles sigui la causa de la variació de l'altra. Tendim a atribuir massa fàcilment efectes de causalitat entre variables correlacionades. El cas és tan freqüent que la frase «correlation does not imply causation» (la correlació no implica causalitat) té fins i tot una entrada a Wikipedia

i s'escriuen llibres amb curioses correlacions, com que el nombre de suïcidis per ofegament o estrangulament està correlacionat amb la despesa del Govern dels Estats Units en ciència, espai i tecnologia (Vigen, 2015).

A més de l'anàlisi de correlacions, una altra manera d'estudiar les relacions entre variables és amb la construcció de **models de regressió**.

Els models de regressió miren d'extreure una funció matemàtica que relacioni les variables entre elles. El model més simple és el model de regressió lineal. L'anàlisi de relacions s'acompanya, així mateix, de gràfics il·lustratius com els gràfics de dispersió.

Quan es fan aquest tipus d'anàlisi amb finalitat predictiva és habitual anomenar aquestes relacions de «causa-efecte».

2.4. Mostreig

En l'estudi esmentat sobre la possible relació entre el càncer de mama i l'ús d'anticonceptius, es volia demostrar la relació entre aquestes variables en la població. No obstant això, com que no era possible incloure-hi tota la població, es va realitzar l'estudi sobre una mostra per tal de generalitzar els resultats a tota la població.

Un exemple típic de mostreig és el dels sondejos d'opinió en dies previs a unes eleccions. Els diaris solen mostrar els resultats de la intenció de vot de la població, resultats que evidentment s'han generalitzat a partir d'una mostra de la població. En aquests casos, quan es tria una mostra de la població, l'objectiu no és descriure la mostra per ella mateixa, sinó generalitzar les conclusions extretes en la mostra a tota la població.

Els mètodes de mostreig que fan servir aleatorització planificada s'anomenen *mostrejos probabilístics*. La variant més bàsica és el mostreig aleatori simple. En aquesta variant, tots els individus de la població han de tenir la mateixa probabilitat de pertànyer a la mostra. Imaginem que tota la població d'interès apareix en una guia telefònica. Cada persona tindria assignat un nombre entre 1 i N, on N és la grandària de la població. El mostreig aleatori simple consistiria a seleccionar k individus a l'atzar, escollint cada vegada un dels individus entre 1 i N de forma aleatòria. D'aquesta manera, cada individu té la mateixa probabilitat de formar part de la mostra.

Quan aquest tipus de procediments no són possibles (habitualment no es disposa d'una llista de la població d'interès), hi ha altres esquemes de mostreig com el mostreig sistemàtic, el mostreig estratificat i el mostreig per conglomerats.

Per exemplificar-ne un, el mostreig estratificat (Scheaffer, 1999) consisteix a separar la població en grups disjunts, denominats *estrats*, i seleccionar una mostra aleatòria simple de cada estrat. La grandària de la mostra també és rellevant, i, tot i que sol ser més costós, lent o difícilment accessible, és preferible treballar amb mostres grans. El mètode de mostreig determina en gran mesura la representativitat de la mostra i, per tant, les generalitzacions que se n'extreguin poden ser més o menys precises.

El 1934 l'estadístic i matemàtic Jerzy Neyman va dissenyar el primer mètode de mostreig estratificat. Dos anys més tard, el sondeig electoral Gallup va predir la victòria de Roosevelt a les eleccions de presidència dels Estats Units fent servir el mostreig estratificat. El més curiós del cas és que la predicció del mostreig estratificat va predir correctament la victòria, a diferència d'un altre sondeig realitzat amb una mostra molt més àmplia de la població que predia la victòria de l'oponent. Des de llavors el sondeig electoral de George Gallup fa servir aquest tipus de mostrejos (Scheaffer, 1999).

2.5. Predicció

Sovint les dades amb què treballa l'estadística es troben prèviament guardades en bases de dades i representen una fotografia estàtica del domini que es vol estudiar. L'exemple sobre la concessió de crèdits a clients seria un d'aquests casos. Altres exemples podrien ser el rendiment dels estudiants d'un curs, el diagnòstic i tractament dels pacients d'una consulta mèdica, els productes amb més vendes d'un determinat negoci, etcètera. Les anàlisis típiques aplicables poden ser de tipus descriptiu, com extreure valors numèrics de la mostra i anàlisis de relacions, tal com s'ha exemplificat anteriorment.

L'estadística també s'ocupa d'un altre tipus de dades que estan associades a un component temporal. En aquests casos, les dades segueixen una seqüència temporal, que s'ha recol·lectat al llarg d'un període de temps i que té l'objectiu de predir l'evolució de les dades en el futur. Per exemple, la temperatura màxima i mínima de cada dia a Barcelona al llarg dels últims deu anys. Un altre exemple clàssic és l'evolució de la borsa. L'estudi del comportament de les dades en el futur es bassa en els models de sèries temporals. D'una banda, el modelatge d'una sèrie temporal permet ajustar un model a les dades passades i descompondre'l en els seus components principals (tendència, estacionalitat, aleatorietat). A partir del modelatge de la sèrie temporal es pot predir l'evolució que tindrà en el futur.

El 1909, un estadístic de Bell Telephone va predir el nombre d'operadores requerides en centrals telefòniques per atendre la creixent demanda de trucades telefòniques. La predicció de la demanda futura va mostrar que totes les dones americanes entre disset i seixanta anys haurien de treballar d'operadores cap a l'any 1930 per satisfer el volum esperat de trucades. La predicció va accelerar la invenció del primer commutador automàtic, que va ser dissenyat i posat en servei per Bell dos anys més tard de la predicció (Drucker, 2006).

2.6. Presa de decisions amb incertesa

En multitud de situacions, es necessita prendre decisions entre un conjunt d'opcions alternatives, sense conèixer per endavant les conseqüències d'aquestes alternatives. Hi ha un elevat grau d'incertesa sobre com serà el comportament futur, a causa de l'efecte incert de determinats factors.

Per exemple, un empresari ha de decidir si invertir el seu pressupost en la línia de producció del producte A o del producte B, sabent que la inversió repercutirà en un augment de producció. No obstant això, desconeix amb exactitud la demanda futura d'aquests productes, amb la qual cosa no pot assegurar quina inversió resultarà més rendible. L'empresari pot imaginar diferents escenaris possibles i, a partir d'això, aplicar un criteri que ajudi a minimitzar-ne el risc o a maximitzar-ne els guanys. Per a fer-ho, hi ha diferents mètodes, des dels basats en càlculs de maximització dels guanys o minimització de pèrdues fins a mètodes que incorporen càlculs probabilístics.

3. Estadística descriptiva i estadística inferencial

3.1. Estadística descriptiva

L'estadística descriptiva es refereix al conjunt de tècniques per recol·lectar i presentar dades numèriques (Armitage, Berry i Matthews, 2002, pàg. 2). Segons Tomeo i Uña (2003), l'estadística descriptiva tracta de la descripció numèrica de conjunts, i és particularment útil quan aquests són de molts elements: valorant matemàticament i analitzant el col·lectiu representat pel conjunt sense pretendre obtenir-ne conclusions més generals, cosa que és objecte de l'estadística inferencial. En definitiva, l'estadística descriptiva tracta de resumir les dades d'una mostra en lloc d'extreure conclusions sobre la població en general.

Bàsicament, l'estadística descriptiva realitza tres tipus de funcions:

- En primer lloc, estudia la distribució de la variable o variables d'interès (quins valors utilitza i com es distribueixen aquests valors).

- En segon lloc, calcula resums numèrics d'aquests valors com les mesures de tendència central (la mediana i la mitjana) i la dispersió (desviació estàndard).

- Finalment, s'encarrega de visualitzar gràficament aquesta informació, de manera que es pot comprendre ràpidament com és la distribució de les dades en relació amb la seva dispersió i simetria.

Estadística descriptiva

Es disposen de dades del pes en néixer de cinc-cents nens nascuts l'any 1970 en un hospital de Barcelona. La tendència central d'aquestes dades conclouria que la mitjana de pes en néixer és de 3,54 kg, i la mediana, de 3,30 kg. L'estudi de la dispersió podria resultar en una desviació estàndard de 1,30 kg, que correspondria a la mitjana de les diferències al quadrat respecte el pes mitjà. L'extracció d'aquests valors es realitza per a la mostra específicament, sense intenció de generalitzar les conclusions a tota la població de nens nascuts el 1970 a Barcelona. Altres anàlisis descriptives es poden fer mitjançant visualitzacions gràfiques, on s'aportaria informació addicional de com es distribueixen aquests pesos.

3.2. Estadística inferencial

L'estadística inferencial va més enllà de l'estadística descriptiva i consisteix en l'obtenció de resultats que generalitzen els comportaments observats en les dades i que permeten extreure conclusions de caràcter més general sobre la població (Alea i altres, 1999). Segons la definició de Gibergans, Gil i Rovira

(2009), l'estadística inferencial es basa en l'extracció de conclusions sobre una població a partir d'una mostra (un subconjunt dels individus de la població) i precisa amb quins marges de confiança són vàlides aquestes afirmacions.

Els mètodes d'inferència estadística es divideixen principalment en dos:

- Mètodes d'estimació de paràmetres
- Mètodes de contrastos d'hipòtesis

L'estimació de paràmetres consisteix a fixar valors concrets als paràmetres que caracteritzen la distribució de probabilitat de la població. El contrast d'hipòtesis permet validar hipòtesis estadístiques que fan referència al valor d'un paràmetre poblacional (el valor esperat, la variància, la proporció, etc.) o la relació que hi ha entre paràmetres anàlegs de dues poblacions.

En termes generals, el contrast d'hipòtesis permet decidir si l'evidència empírica aportada per la mostra és o no compatible amb la hipòtesi referida a la població sobre la qual s'intenta generalitzar (Alea i altres, 1999). Aquest tipus d'anàlisi es presenta amb freqüència en múltiples dominis: en medicina, per comparar l'eficàcia de dos tractaments diferents; en agricultura, per conèixer el millor fertilitzant; en educació, per comparar el millor mètode d'estudi; en màrqueting, per conèixer la campanya amb més adquisició de clients; etc.

En l'exemple citat anteriorment, la mostra d'estudi composta per cinc-cents nens nascuts en un hospital de Barcelona el 1970 podria servir de base per a inferir paràmetres sobre la població. Per exemple, es podria deduir que el pes dels nens i nenes en néixer a la Barcelona del 1970 està comprès entre 3,43 kg i 3,66 kg amb un nivell de confiança del 95%. Mitjançant contrastos d'hipòtesis, es pot validar si la mitjana dels nens en néixer és igual o superior a un determinat valor (per exemple, 3,5 kg), o bé si la mitjana del pes en néixer dels nens és superior al de les nenes. També es podrien comparar els paràmetres de dues poblacions d'interès (recollides el 1970 i el 2010 a Barcelona) i preguntar-nos si els nens i nenes nascuts l'any 2010 tenen un pes superior als nens i nenes nascuts l'any 1970. Com ja hem esmentat, la representativitat de les mostres és crucial perquè les conclusions extretes siguin precises.

4. Estadística davant de mineria de dades

A partir de la «moda de les dades», s'ha despertat un interès creixent en l'estadística i altres disciplines, com el *machine learning* (traduït com 'aprenentatge automàtic') i el *data mining* (mineria de dades); alhora que s'ha reetiquetat l'estadística amb noms com analítica o ciència de les dades i han emergit d'altres com *big data* (Mayer-Schönberger i Cukier, 2013). En aquest nou panorama, és difícil distingir quin domini pertany a cada disciplina. La línia que separa estadística i mineria de dades és molt fina.

Tradicionalment, l'estadística prové de l'interès molt incipient de les civilitzacions antigues per fer censos de població fonamentant-se en les matemàtiques i la teoria de les probabilitats. La mineria de dades sorgeix d'un enfocament computacional, a partir de l'ús d'ordinadors per emmagatzemar bases de dades i del desenvolupament d'algoritmes computacionals que permetin incorporar intel·ligència artificial o aprenentatge artificial per extreure informació d'aquestes dades, aprofitant la gran potència de càlcul, i alimentar processos de decisió.

Estadística i mineria de dades tenen molts punts de trobada, i per això és difícil marcar una línia divisòria. No obstant això, pot ser pedagògic (encara que probablement reduccionista) contrastar les dues disciplines, perquè l'estudiant pugui fer-se una imatge preliminar que podrà anar enriquint a mesura que s'endinsi en els detalls de cada disciplina. Vegem, doncs, què és la mineria de dades en comparació amb l'estadística.

Segons Han i Kamber (2001), el terme *mineria de dades* es refereix a l'extracció de coneixement de grans quantitats de dades. Principalment, la disciplina s'ha centrat en l'extracció de patrons de bases de dades. Un terme similar ja una mica en desús és el de KDD (*knowledge discovery in databases*), que engloba tot el procés d'anàlisi des de la preparació de dades fins a la seva interpretació (Han i Kamber, 2001):

1) Neteja de dades
2) Integració de dades
3) Selecció de dades
4) Transformació de dades
5) Extracció de coneixement (mineria de dades)
6) Avaluació de patrons
7) Presentació del coneixement

Pyle (1999) diu que històricament l'anàlisi estadística s'ha orientat a la verificació i a la validació d'hipòtesis; una aproximació que s'ha vist influïda per la ciència. S'estableix una hipòtesi, es recullen les evidències i es confronten amb la hipòtesi per veure si pot ser rebutjada o no. Segons Pyle, la mineria de dades «gira la truita». En lloc d'establir les hipòtesis que s'han de testejar, l'analista pren un conjunt de dades i pregunta: «Quines són les hipòtesis que aquest conjunt de dades suporta?». Pyle afegeix un altre element que hem de tenir en compte: els grans volums de dades. L'estadística fa servir mètodes en què cal imaginar potencials relacions entre les dades, que són posteriorment validades amb els mètodes d'anàlisi apropiats. Però els grans volums de dades actuals fan difícil que aquestes hipòtesis prèvies puguin ser visualitzades prèviament i, tot i això, podrien haver-hi altres relacions en les dades no observades o no imaginades. Tot això requereix una automatització que la mineria de dades pot proveir.

Un estudi de mineria de dades sobre les compres realitzades en un supermercat dels Estats Units va llançar la correlació entre la compra de bolquers i la compra de cerveses. La història il·lustra una associació imprevista entre dos productes que el sentit comú no hauria associat. Tot i que la veracitat de la història és qüestionable, és un exemple il·lustratiu que ens permet aclarir els dominis de l'estadística i de la mineria de dades. En aquest cas, els algoritmes computacionals buscaven qualsevol tipus de relació entre dades i es van trobar amb aquesta associació sorprenent. Per contra, des de l'estadística s'hauria formulat prèviament una hipòtesi com: «Fins a quin punt la compra de bolquers està correlacionada amb la compra de fruita i verdura». Difícilment una persona hauria formulat la hipòtesi «Fins a quin punt la compra de bolquers està correlacionada amb la compra de cervesa», ja que manca de sentit comú.

Pyle ens ofereix un altre exemple il·lustratiu de les diferències entre l'estadística i la mineria de dades:

> «Una companyia de targetes de crèdit tenia un equip d'estadístics que treballaven per descobrir "interaccions interessants en les dades". La seva aproximació era escollir mostres representatives de les dades i buscar interaccions interessants entre elles. Algunes de les interaccions van ser estadísticament significatives i van ser propostes per a accions de màrqueting, mentre que d'altres no van ser significatives i es van descartar.
>
> La companyia va contractar un equip d'experts en mineria de dades per analitzar les mateixes dades. No van començar amb una mostra de les dades i unes hipòtesis per testejar. Van escollir un conjunt de dades i van fer servir algoritmes per extreure'n les possibles interaccions d'interès que podien suportar les dades. Amb la llista potencial d'interaccions generades, van refinar el model. Els miners de dades van buscar els factors entre els grups que fossin responsables de major benefici per a l'empresa (modelatge inferencial) i van dissenyar models que podien predir quin tipus de persones podrien ser (modelatge predictiu).
>
> Els estadístics van construir models de regressió lineals i no lineals, van analitzar els residuals i els intervals de confiança. Els miners van usar diverses tècniques, entre elles regles d'inducció, arbres de decisió i xarxes neuronals. Els arbres de decisió van ser els que van presentar millor rendiment i es van utilitzar per extreure el coneixement de les

dades. Una de les informacions que van extreure és que el 0,1% dels clients presentaven un patró alt de benefici per a l'empresa. Entre ells, el 30% de les persones que compraven equipament d'esquí es gastaven més de tres mil dòlars en un període de trenta dies. El resultat es va fer servir per oferir campanyes de màrqueting directament enfocades a aquest sector amb un retorn de la inversió molt elevat per a l'empresa.»

Pyle (1999, pàg. 488)

Pyle raona que una fluctuació del 0,1% podria haver estat insignificant per a un estadístic, però va ser identificada per algoritmes de mineria de dades amb un resultat significatiu des del punt de vista comercial. És així que les aproximacions de l'estadística i la mineria de dades presenten enfocaments complementaris.

Com ja hem apuntat anteriorment, voler marcar una línia divisòria entre l'estadística i la mineria de dades és un enfocament reduccionista. L'exemple il·lustra de forma particular una visió general de l'estadística com la verificació d'hipòtesis i la mineria de dades com una versió algorítmica de la recerca d'aquestes hipòtesis. Si aquest enfocament ens ajuda a comprendre què és l'estadística i què és la mineria de dades, ens serveix com a marc inicial preliminar. Però no ens agradaria quedar-nos limitats per aquest marc. Vegem què en diuen Witten, Frank i Hall (2011):

«¿Cuál es la diferencia entre aprendizaje automático y estadística? [...] En realidad, no deberíamos buscar una línea divisoria entre aprendizaje automático y estadística, puesto que hay un contínuum –a la vez que multidimensional– de técnicas de análisis de datos. Algunos derivan de las habilidades adquiridas en cursos de estadística y otros están más asociados con los tipos de algoritmos de aprendizaje automático que surgieron con la ciencia computacional. Históricamente, las dos disciplinas han tenido diferentes tradiciones. Si lo forzamos a un único punto diferencial, este sería que la estadística se ha preocupado por testear hipótesis, mientras que el aprendizaje automático se ha preocupado del proceso de generalización formulado como una búsqueda en un espacio de posibles hipótesis. Pero esto es una gran simplificación: la estadística es mucho más que contrastes de hipótesis y muchas técnicas de aprendizaje automático no involucran ningún tipo de búsqueda.

»En el pasado, se han desarrollado varios esquemas a la par en aprendizaje automático y estadística. Uno de ellos son los árboles de inducción. Cuatro estadísticos (Breiman *et al*., 1984) publicaron el libro *Classification and regression trees* a mediados de los años ochenta, y a lo largo de las décadas de los setenta y los ochenta un investigador de aprendizaje automático prominente, J.Ross Quinlan, desarrolló un sistema para inferir árboles de decisión a partir de ejemplos. Los dos proyectos, independientes entre sí, produjeron resultados similares para generar árboles de inducción a partir de ejemplos, y los autores solo fueron conscientes de los trabajos ajenos mucho más tarde.»

Witten, Frank i Hall (2011, pàg. 28)

Els autors del llibre mostren un bon exemple en què l'estadística i la mineria de dades treballen de la mà. Molts mètodes de mineria de dades estan dissenyats en forma d'algoritmes computacionals però contenen al seu interior mètodes estadístics per extreure patrons de coneixement. Això es dona en totes les fases del procés d'extracció de coneixement. Les fases de preprocés i preparació de dades contenen tècniques estadístiques com la detecció d'*outliers* (o valors extrems), la normalització de les dades o la reducció del nombre d'atributs amb mètodes com l'anàlisi de components principals, per citar alguns exemples. En la fase de modelatge es fan servir algoritmes basats en la teoria bayesiana, algoritmes d'inducció d'arbres usant mètriques basades en entropia o mètodes

estadístics per evitar l'anomenat *overfitting* (sobreaprenentatge). Les tècniques de mostreig estadístiques es fan servir per obtenir estimadors precisos de la qualitat dels algoritmes i els contrastos d'hipòtesis ajuden a identificar els algoritmes que són capaços d'extreure els millors patrons de les dades.

En definitiva, tot i que una versió reduccionista de l'estadística i la mineria de dades estableix orígens diferents i marca diferències en la formulació i la recerca de les hipòtesis, hi ha múltiples escenaris en què no solament representen vies d'actuació complementàries, sinó que, a més, comparteixen els mateixos enfocaments.

5. Aplicacions de la mineria de dades

Han i Kamber (2001) també distingeixen entre l'anàlisi de dades de tipus descriptiu i de tipus predictiu en l'àmbit de la mineria de dades. En l'anàlisi descriptiva es caracteritzen les propietats de les dades. Un exemple és categoritzar els clients habituals d'un supermercat segons els seus hàbits de compra. En l'anàlisi predictiva es realitzen inferències sobre les dades actuals amb l'objectiu de realitzar prediccions sobre el futur. Per exemple, com es comportarà el client del supermercat si es fa una campanya de descomptes.

Hi ha un conjunt finit de tipologies de preguntes que es poden realitzar sobre les dades des de la mineria de dades. Entre les més habituals hi ha:

- caracterització de les dades o discriminació en categories (classificació),
- regressió,
- anàlisi d'associació,
- *clustering* o agrupació,
- anàlisi d'*outliers*,
- anàlisi de tendències o sèries temporals.

Sense intenció d'endinsar-nos en detall en aquestes preguntes, a continuació realitzem una breu descripció d'aquest tipus d'aplicacions:

1) Classificació. La classificació consisteix en la categorització d'unes dades en un conjunt de classes predefinit prèviament. Habitualment, es disposen de dades registrades en un fitxer o base de dades, caracteritzats per un conjunt d'atributs i una classe associada. En termes d'estadística, els atributs correspondrien a les variables independents i la classe associada correspondria a la variable dependent. Els algoritmes de classificació construeixen un model que explica les relacions inherents entre els atributs i la seva classe. Aquest procés s'anomena també *generalització*, ja que sorgeix de casos particulars (la mostra) i dona com a resultat un model general que explica els patrons inherents a aquestes dades.

Si prenem l'exemple de la taula 1, els atributs o variables independents serien els mostrats en la taula (edat, sexe, estat civil, nivell d'ingressos...) i la classe podria ser si el client torna el crèdit en el temps i les condicions establerts. La classificació construiria un model que discriminaria de manera general quin tipus de clients tornen el crèdit concedit i quins clients no ho fan. El model podria tenir capacitat explicativa i, per tant, utilitzar-se per comprendre quins paràmetres i valors influeixen en la devolució o no devolució dels crèdits concedits. A més, té capacitat predictiva, ja que pot fer-se servir per predir el comportament de futurs clients.

2) Regressió. La regressió, també anomenada *predicció numèrica*, consisteix a construir un model que relacioni un conjunt d'atributs amb una variable numèrica. És com la classificació, on es relacionen atributs amb la classe, en què la classe és un valor numèric en lloc de nominal. Els algoritmes de regressió comparteixen el mateix objectiu que la regressió estadística, que és el de construir models que expliquin les relacions inherents entre les variables independents i les variables dependents (numèriques). Varien en el tipus de mètodes fets servir i, per tant, en el tipus de models construïts; tot i que també hi ha algoritmes de regressió que es basen en els mateixos models matemàtics que els utilitzats per l'estadística.

Tornant a l'exemple de la taula 1, parlaríem de regressió si prenem com a variable independent el valor del crèdit sol·licitat. Amb això, els algoritmes de regressió intentarien construir un model que explicaria l'import del crèdit a partir de la resta de característiques. Evidentment, la qualitat del model depèn de si hi ha tals relacions en les dades i, per tant, caldria avaluar l'error comès en la construcció d'aquests models.

3) Anàlisi d'associació. L'anàlisi d'associacions consisteix en la recerca i extracció de models que mostren les associacions entre atributs d'un conjunt de dades. Una manera habitual de representar aquestes associacions és mitjançant les *regles d'associació*. En aquest tipus d'enfocament, no hi ha una classe o variable dependent identificada a priori, com succeeix en la classificació i regressió. Aquest enfocament tracta totes les variables «la mateixa manera» i investiga si hi ha qualsevol tipus de relació entre elles.

Si tornem a fer servir l'exemple de la taula 1, veurem que les regles d'associació podrien ser del tipus: si el nombre de fills és dos o tres i l'estat civil és casat, està relacionat amb un crèdit per a vehicle o reformes i l'import oscil·la entre vint mil i quaranta mil. Com es veu, el model conté un valor descriptiu en forma de regles. Igual que en els casos anteriors, la qualitat d'aquests models depèn tant de l'existència d'aquestes relacions inherents a les dades com de la capacitat del model d'identificar aquest tipus de relacions.

Com es comentava anteriorment, l'enfocament de l'estadística per trobar aquest tipus d'associacions s'inicia en hipòtesis prèviament establertes. Per exemple, es podria formular si el nombre de fills està relacionat amb l'import del crèdit. Però aquestes associacions han de ser imaginades a priori. Mitjançant la mineria de dades, l'algoritme busca «a cegues» quin tipus de relacions es poden extreure. L'algorisme realitza una recerca de les hipòtesis més prometedores dins l'espai de totes les possibles hipòtesis. L'estudiant es pot imaginar que aquesta cerca a cegues necessita una gran capacitat de còmput per poder dur-se a terme.

4) Agrupació (*clustering*). L'agrupació consisteix a separar les dades en diferents grups o categories. Cal distingir-la de la classificació, on les categories de la mostra estan assignades a priori. En l'agrupació no hi ha categories preesta-

blertes i és el mateix algoritme el que agrupa les dades segons les seves característiques, de manera que els casos d'una mateixa categoria presentin un grau de similitud elevat i alhora es diferenciïn de la resta de casos dels altres grups.

L'agrupació s'anomena també *segmentació* o, en anglès, *clustering*. En el cas de la taula 1, l'agrupació revelaria quins tipus de clients formen part del conjunt de dades. Per exemple, se'n podria extreure que un segment de dades està format per persones solteres, d'entre vint i trenta anys, que sol·liciten un préstec per llogar un habitatge. Un altre grup de dades podria ser el de persones amb família, casats, amb una hipoteca i que demanen un préstec d'entre vint mil i quaranta mil euros per comprar un vehicle o fer reformes a casa.

5) Anàlisi d'*outliers*. En els enfocaments descrits prèviament, els algoritmes de classificació, regressió o agrupació intenten trobar les regularitats en les dades, és a dir, models que són capaços de descriure el comportament general de les dades. En l'anàlisi d'*outliers* (també anomenats valors extrems) es tracta precisament del contrari: trobar els casos anòmals, les rareses, dins dels patrons generals. Tant els algoritmes descrits com l'estadística solen eliminar les dades anòmales perquè poden introduir desviacions en les anàlisis que alteren les conclusions. En canvi, l'anàlisi de valors extrems pretén precisament descobrir aquests casos anòmals perquè són l'objecte d'interès de l'estudi. Parlem, per exemple, de la detecció de frau en transaccions de targetes de crèdit o de la detecció de petroli abocat en oceans.

6) Anàlisi de sèries temporals. En l'anàlisi de sèries temporals, les dades de l'estudi tenen un component temporal. Ens referim, per exemple, a l'evolució de la cotització de l'IBEX-35 en els últims deu anys, o la temperatura màxima i mínima diària de Madrid al llarg dels tres últims anys. L'objectiu de l'estudi de les sèries temporals és modelar el comportament d'aquestes dades al llarg del temps per predir el comportament futur. Com podrà intuir l'estudiant, la predicció del comportament futur no és una tasca gens fàcil, ja que no és possible identificar totes les dades que influeixen en aquest comportament (imaginem totes les variables que poden influir en la cotització de l'IBEX 35). En definitiva, el modelatge de fenòmens complexos de la realitat no és una tasca senzilla.

Resum

En aquest mòdul hem navegat per la superfície de l'anàlisi de dades, i especialment de l'estadística. Diem «navegar» perquè hem fet un recorregut, encara que no exhaustiu, pels seus orígens i ens hem detingut en algunes de les seves anècdotes, amb l'ànim de despertar l'interès i la curiositat de l'estudiant. El recorregut ha visitat les principals aplicacions de l'estadística sobre l'extracció de conclusions de les dades, la gestió de la incertesa, l'anàlisi de les relacions entre variables, el mostreig, la predicció i la presa de decisions. Així mateix, la distinció entre estadística descriptiva i inferencial és clau per comprendre si s'extreuen conclusions sobre la mostra d'estudi o inferències sobre la població.

L'anàlisi de dades es configura avui dia com un panorama multidisciplinari. Per aquest motiu és convenient contextualitzar l'estadística en aquest marc i, alhora, contrastar-la amb la jove disciplina de la mineria de dades. Hem descrit breument els principals enfocaments de la mineria de dades, com la classificació, la regressió, l'associació, l'agrupació, l'anàlisi d'*outliers* i les sèries temporals. Estadística i mineria de dades provenen d'enfocaments diferents, de caràcter matemàtic i científic, la primera, i computacional, la segona. Tot i això, comparteixen l'objectiu d'extreure informació útil de les dades que ajudin a comprendre els fenòmens que ens envolten en un ampli espectre de dominis, com la ciència, la medicina, la meteorologia, l'economia i l'empresa, per citar-ne només alguns.

Després d'aquest breu recorregut animem l'estudiant a aprofundir en les tècniques i els mètodes de l'estadística, ja que així trobarà els fonaments de l'anàlisi de dades.

Bibliografia

Alea, V.; Guillén, M.; Muñoz, M. C.; Torrelles, E.; Viladomiu, N. (1999). *Estadística aplicada a les ciències econòmiques i socials*. Barcelona: Edicions Universitat de Barcelona, McGraw-Hill.

Armitage, P.; Berry, G.; Matthews, J. N. S. (2002). *Statistical Methods in Medical Research (Fourth Edition)*. Malden, Massachusetts: Blackwell Science.

Dodge, Y. (2006). *The Oxford Dictionary of Statistical Terms*. Oxford: Oxford University Press.

Drucker, P. F. (2006). *Innovation and Entrepreneurship. Practice and Principles*. Nova York: Harper Collins.

Gibergans, J.; Gil, A. J.; Rovira, C. (2009). *Estadística*. Barcelona: FUOC.

Han, J.; Kamber, M. (2001). *Data Mining.Concepts and Techniques*. San Diego: Academic Press.

Mayer-Schönberger, V.; Cukier, K. (2013). *Big Data. A Revolution That Will Transform How We Live, Work and Think*. Regne Unit: John Murray.

Moore, D. S.; McCabe, G. P.; Craig, B. A. (2012). *Introduction to the Practice of Statistics* (7a. ed.). Nova York: W. H. Freeman and Company.

Newbold, P. (1997). *Estadística para los negocios y la Economía* (4a. ed.). Nova Jersey: Pearson, Prentice Hall.

Pyle, D. (1999). *Data Preparation for Data Mining*. San Francisco: Morgan Kaufmann Publishers, Inc.

Rooney, A. (2009). *Historia de las matemáticas. De la construcción de las pirámides hasta la explotación del infinito*. Barcelona: Oniro.

Rothman, K. J. (1996). *Lessons from John Graunt. The Lancet*. Amsterdam: Elsevier.

Scheaffer, R. L. (1999). *Sampling Methods and Practice*. Florida: Universitat de Florida, NCSSM Statistics Leadership Institute.

Thomas, D. B. (1991, núm 6). «The WHO collaborative study of neoplasia and steroid contraceptives: The influence of combined oral contraceptives on risk of neoplasms in developing and developed countries» [article en línia]. *Contraception* (vol. 43, pàg. 695-710). Disponible a: <https://doi.org/10.1016/0010-7824(91)90010-D>

Tomeo, V.; Uña, I. (2003). *Lecciones de Estadística descriptiva. Curso teórico-práctico*. Madrid: Thomson.

Vigen, T. (2015). *Spurious correlations*. Nova York: Hachette Books.

Witten, I.; Frank, E.; Hall, M. A. (2011). *Data Mining. Practical Machine Learning Tools and Techniques*. (3a. ed.). Burlington: Morgan Kaufmann.