Introducció al disseny i anàlisi d'enquestes

Aplicacions estadístiques a la selecció de mostres i a l'anàlisi de qüestionaris

Ángel A. Juan, Alicia Vila i Patricia Carracedo

PID_00242447

Temps de lectura i comprensió: **3 hores**

UOC
Universitat Oberta de Catalunya

Índex

Introducció

Les enquestes i qüestionaris s'han convertit en una eina d'investigació d'ús quotidià en l'anomenada *societat de la informació*. La idea d'usar dades provinents d'una mostra –composta per un nombre relativament petit d'elements– per a obtenir informació sobre tota una població és utilitzada a diari pels mitjans de comunicació, ja sigui premsa escrita, televisió, ràdio o fins i tot Internet.

En efecte, les enquestes i els qüestionaris s'usen per a sondejar l'estat d'opinió dels potencials votants d'unes eleccions, per a conèixer el potencial interès de nous béns o serveis al mercat, per a predir l'acceptació que tindran determinades decisions governamentals o estratègiques, per a conèixer millor els membres d'una comunitat o d'una organització, per a detectar demandes potencials dels consumidors que no estan essent satisfetes, etc. En investigació, a més, les tècniques basades en l'ús d'enquestes i qüestionaris representen probablement l'eina d'investigació social més comuna en articles i publicacions científiques.

Tanmateix, el pas de dades mostrals a informació sobre la població no és trivial, ja que requereix de tot un procés metòdic que inclou el disseny de les preguntes (per a evitar d'introduir-hi biaixos innecessaris), el disseny de la mostra (per a minimitzar tant com sigui possible l'error mostral), la realització de l'enquesta i l'anàlisi dels resultats. Molt sovint aquest procés es realitza massa a corre-cuita i de manera poc rigorosa, amb la qual cosa els resultats que s'obtenen són poc fiables i gens creïbles des d'un punt de vista científic. En aquest mòdul es presenten i discuteixen els conceptes bàsics d'aquestes tècniques, des de les claus d'un bon qüestionari i d'un bon disseny mostral fins a exemples de com es poden aplicar les tècniques estadístiques treballades durant el curs per a representar numèricament i gràficament la informació obtinguda sobre la població.

Objectius

Els objectius docents que es pretenen assolir amb aquest mòdul són els següents:

1. Entendre la importància de les enquestes i els qüestionaris en la societat de la informació.

2. Conèixer els aspectes clau a considerar quan s'elaboren les preguntes d'un qüestionari.

3. Conèixer els tipus d'escales més habituals en els qüestionaris, així com el tipus de dades que cada una produeix.

4. Introduir-se en els tipus de mostreig més habituals en els estudis d'enquestes, en particular: el mostreig aleatori simple, el mostreig sistemàtic, el mostreig per estrats i el mostreig per conglomerats.

5. Saber calcular estimacions puntuals i per intervals per a diversos paràmetres poblacionals segons el tipus de mostreig usat.

6. Aprendre a utilitzar les tècniques estadístiques treballades durant el curs per a analitzar qüestionaris.

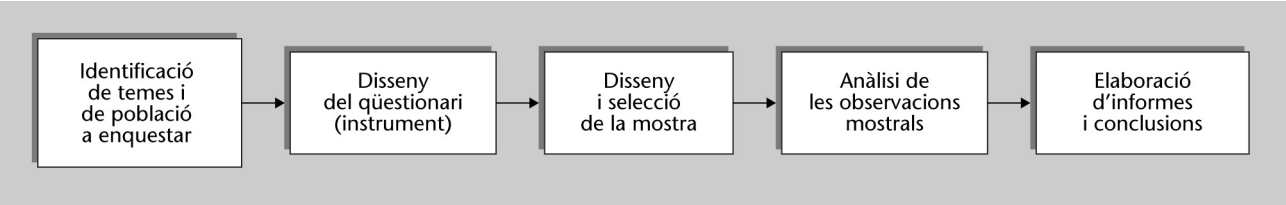
7. Aprendre a usar programari estadístic i d'anàlisi de dades com a instrument bàsic en l'aplicació pràctica dels conceptes i tècniques estadístiques.

1. Disseny de qüestionaris

Les tècniques d'investigació basades en l'ús d'enquestes s'apliquen a multitud d'àmbits diferents: en els negocis, en l'administració pública, en les ciències socials i del comportament, en les ciències de la informació i la comunicació, en les ciències de la salut, en les ciències polítiques, i en qualsevol altre àmbit en el qual les dades que puguin aportar els usuaris d'un servei o els consumidors d'un producte tinguin un paper fonamental. En la societat de la informació, les organitzacions i institucions fan un ús intensiu de les dades que expliquen com es comporten els individus, quins són els seus gustos i les seves necessitats, quina opinió tenen sobre determinats temes, etc. En aquest context, les tècniques d'investigació basades en l'ús d'enquestes permeten obtenir les esmentades dades que, després de la seva posterior anàlisi estadística, proporcionen una valuosa informació tant als investigadors teòrics d'una determinada disciplina com als responsables de prendre decisions sobre el funcionament de les organitzacions.

En general, es poden distingir sis fases seqüencials en el desenvolupament de qualsevol estudi basat en l'ús d'enquestes (figura 1): (a) identificació dels temes concrets sobre els quals es vol obtenir informació, així com de la població a enquestar, (b) disseny del qüestionari com a instrument per a obtenir les dades que es necessiten, (c) disseny i selecció d'una mostra representativa de la població, (d) obtenció de les dades mitjançant la tramesa del qüestionari als individus que componen la mostra, (e) anàlisi estadística de les observacions mostrals a fi d'inferir informació sobre la població, i (f) elaboració d'informes i conclusions.

Figura 1. Fases en el desenvolupament d'una enquesta



En aquest apartat farem especial èmfasi en la fase de disseny del qüestionari, i deixarem per a apartats posteriors altres fases clau en què les tècniques estadístiques tenen una aportació decisiva, com ara la fase de disseny i selecció de la mostra i la fase d'anàlisi de les observacions mostrals.

1.1. Elaboració de les preguntes d'un qüestionari

Les preguntes que es formulen en un qüestionari constitueixen l'aspecte més rellevant de qualsevol enquesta. Perquè aquestes compleixin el seu pa-

per de forma eficient, les preguntes d'un qüestionari s'han de centrar en els aspectes essencials sobre els quals es vol obtenir informació. Així mateix, les esmentades preguntes han de ser tan breus i clares com sigui possible a fi de facilitar la tasca de les persones enquestades i maximitzar la fiabilitat i validesa del qüestionari. Es tracta d'evitar possibles problemes com ara interpretacions errònies de les preguntes, esgotament de l'enquestat o, fins i tot, rebuig a contestar una part o la totalitat del qüestionari per la longitud d'aquest o l'esforç necessari per a entendre les preguntes i contestar-les. Aquestes problemàtiques podrien introduir biaixos i errors mostrals en les dades, la qual cosa minvaria la fiabilitat i validesa de l'enquesta i dels resultats.

És important ser curós en l'elaboració de les preguntes a fi d'evitar introduir al qüestionari problemes **d'error mostral** –a causa de l'ús d'una mostra per a estimar paràmetres poblacionals– o de **biaix** (qualsevol altre tipus d'error al qüestionari diferent de l'error mostral): si en la mateixa formulació de les pregunta s'està induint l'enquestat a respondre en un sentit concret, llavors s'està introduint un biaix en el qüestionari; si la formulació de les preguntes és ambigua i dóna peu a diferents interpretacions, llavors s'està afavorint una excessiva dispersió de les respostes, la qual cosa incrementa l'error mostral. Per tant, la manera com les preguntes són formulades en un qüestionari és determinant a l'hora d'evitar introduir-hi patrons de biaix i error mostral. Així, podem establir les recomanacions generals següents a tenir presents quan elaborem les preguntes d'un qüestionari:

- Criteris d'interpretació i resposta clars: els criteris en què l'enquestat s'ha de basar per a interpretar i contestar una pregunta han d'estar clarament especificats en el qüestionari.

- Preguntes apropiades per al conjunt d'individus que configuren la mostra: les preguntes han de poder ser respostes per tots els enquestats partint de la seva experiència o condició personal.

- Ús adequat d'expressions, exemples o alternatives de resposta: s'ha d'evitar incloure en la pregunta expressions que indueixin a una determinada resposta.

- Nivell d'actualitat de les preguntes: no s'hauria de pressuposar que l'enquestat serà capaç de recordar amb precisió quin va ser el seu comportament en el passat o la seva opinió sobre un tema esdevingut fa ja bastant temps.

- Preguntes amb un nivell de generalització / concreció adequat: s'hauria d'evitar formular preguntes massa genèriques o ambigües que es puguin interpretar de formes molt diferents i la resposta de les quals no aporti massa

informació, així com preguntes massa específiques que l'enquestat no sigui capaç de contestar amb el nivell de detall requerit.

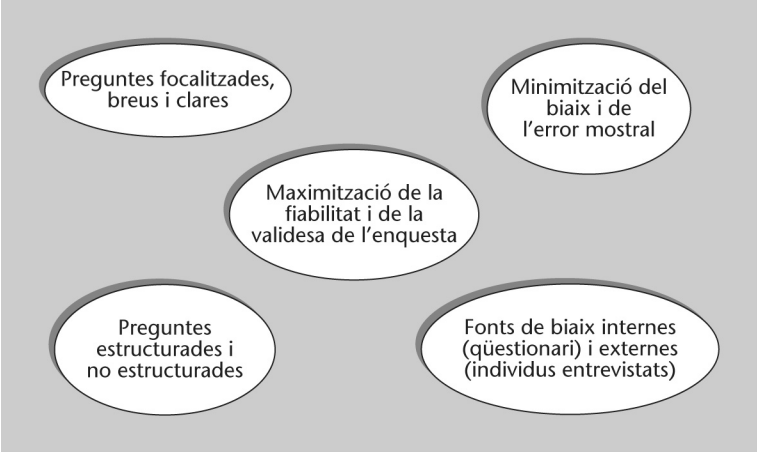
A més d'aquestes fonts internes de biaix causades pel mateix instrument de l'enquesta, hi ha també altres potencials fonts de biaix que no són originades per com s'han elaborat les preguntes sinó per les condicions en les quals s'ha respost el qüestionari. Convé conèixer i tenir presents aquestes altres fonts potencials de biaix per a evitar-les tant com sigui possible amb una correcta elecció de les condicions de l'enquesta i, en particular, de la mostra. Així, algunes d'aquestes fonts externes de biaix són les següents: respostes que busquen estar en coherència amb el que és "socialment desitjable" o amb el que l'entrevistador espera obtenir, respostes orientades a donar una bona imatge de l'enquestat, respostes amb excessiva tendència a la dicotomia (sí/no, positiu/negatiu...) o feia les opcions extremes, respostes hostils excessivament condicionades per experiències negatives recents, etc.

Hi ha dos formats bàsics per a elaborar preguntes d'un qüestionari: les **preguntes obertes** o no estructurades són aquelles que permeten a l'enquestat respondre lliurement sense estar condicionat per un conjunt de possibles alternatives de resposta. Per contra, les **preguntes estructurades** o tancades són les que que contenen en la mateixa pregunta un conjunt de possibles respostes o categories que pot escollir l'enquestat. Les preguntes estructurades són les que habitualment més s'usen en els qüestionaris, ja que a més de delimitar més clarament el context de la informació que s'espera obtenir, solen ser més fàcils i ràpides en cas de contestar, permeten comparar millor diferents grups d'enquestats i, sobretot, faciliten enormement el processament i anàlisi posterior de les dades.

Quan s'usen preguntes estructurades és important elegir bé les categories o possibles respostes alternatives de manera que aquestes constitueixin una llista completa d'opcions (incloent-hi opcions com "altres" o "no ho sap / no contesta" quan sigui necessari) i siguin mútuament excloents (llevat que siguin d'opció múltiple). Pel que fa al nombre de categories o respostes alternatives, el més recomanable és que aquest se situï entre un mínim de dos per a preguntes dicotòmiques i un màxim de sis. Afegir més categories sol dificultar en excés la tasca de l'enquestat. Cal tenir present, tanmateix, que en cas de dubte sobre el nivell de detall que es vulgui oferir en les categories, sol ser preferible optar per l'opció amb més categories, ja que sempre és possible combinar o agregar categories *a posteriori* –durant la fase d'anàlisi–, mentre que l'operació de desagregar respostes ja obtingudes en noves categories no sol ser possible sense la conseqüent pèrdua de precisió i informació.

La figura 2 sintetitza els conceptes clau que s'han de tenir en compte en l'elaboració de les preguntes de qualsevol qüestionari.

Figura 2. Conceptes clau en l'elaboració de les preguntes d'un qüestionari



1.2. Ús d'escales en preguntes estructurades

Les respostes a preguntes estructurades consisteixen, en general, a elegir una opció concreta en una llista de categories possibles. Aquestes categories segueixen una escala o graduació que pot ser simplement nominal o bé pot implicar algun tipus de relació ordinal o numèrica entre les diferents categories implicades:

- **Escales nominals:** són aquelles en què les categories no estan associades a una relació d'ordre o de magnitud. Un exemple seria una escala en què les categories fossin diferents codis postals, prefixos telefònics o identificadors del sexe ("home", "dona"). Aquest tipus d'escala proporciona dades de tipus nominal que simplement identifiquen categories, per la qual cosa és el més limitat des del punt de vista de les tècniques estadístiques que podem aplicar en les observacions obtingudes.

- **Escales ordinals:** són aquelles en les quals les categories segueixen una relació d'ordre o preferència, encara que no de magnitud, que permet classificar-les. Un exemple seria una escala de tasques seqüencials a realitzar en un procés, en què la pregunta podria ser elegir la tasca que es consideri més crítica. Aquest tipus d'escales possibilita l'ús de les anomenades tècniques estadístiques no paramètriques per a analitzar les dades obtingudes.

- **Escales d'intervals equidistants:** són les que associen una magnitud a cada categoria i en les quals el zero no significa absència de magnitud. Un exemple seria una escala graduada de l'1 al 7 per a representar nivells d'importància. Aquesta escala permet l'ús de tècniques d'inferència estadística, per la qual cosa resulta altament recomanable.

- **Escales de ràtio:** són les que associen una magnitud a cada categoria i en les quals el zero representa absència de magnitud. Un exemple seria una es-

cala graduada del 0 al 50 per a indicar la distància en quilòmetres recorreguda per l'enquestat per a acudir al seu lloc de treball. Igual com ocorria amb les escales d'intervals equidistants, les de proporció també permeten l'ús de tècniques d'inferència estadística.

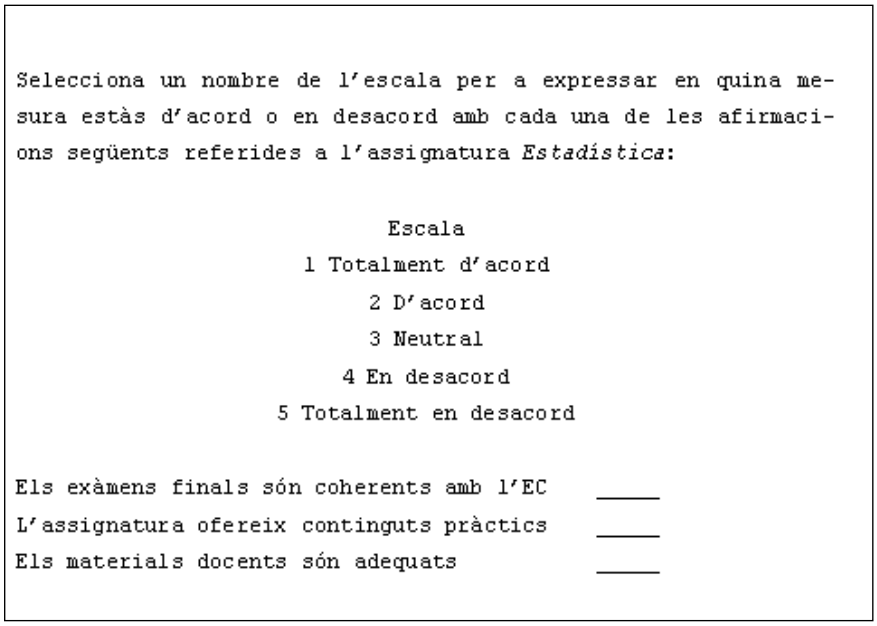
A continuació, descrivim alguns exemples d'escales particulars que s'utilitzen habitualment en els qüestionaris:

Nota

Els exemples només cobreixen algunes de les tipologies d'escales més usades. A Internet és fàcil trobar exemples de qüestionaris complets i altres tipus d'escales només buscant per termes com "survey examples", "questionnaire examples", etc.

- **L'escala de Likert**: aquesta escala sol usar-se per a obtenir el grau d'acord o desacord de l'enquestat amb una determinada afirmació (figura 3). Ja que totes les categories en una escala de Likert solen estar etiquetades (i les etiquetes o identificadors de cada categoria no han de representar pas magnituds equidistants), hi ha certa discrepància entre els experts sobre si aquesta escala ha de considerar-se simplement com una escala ordinal o bé pot considerar-se fins i tot com una escala d'intervals. Una possible solució a aquest problema seria mantenir únicament els identificadors o etiquetes dels extrems (per exemple, (1) "Molt en desacord" i (5) "Molt d'acord"), i deixar la resta d'ítems numerats però sense etiquetar, de manera que els nombres defineixin intervals equidistants. En tot cas, és aquest un tema bastant discutible sobre el qual no sembla haver-hi un consens total. Òbviament, resulta molt avantatjós poder considerar una escala de Likert com d'intervals per a poder així aplicar tècniques d'inferència estadística de forma lícita.

Figura 3. Exemple de preguntes usant una escala de Likert



- **L'escala de freqüència verbal**: aquesta escala és molt similar a la de Likert, amb la diferència que els ítems de l'escala indiquen amb quina freqüència s'ha dut a terme una determinada acció (figura 4).

Figura 4. Exemple de preguntes usant una escala de freqüència verbal

```
Selecciona un nombre de l'escala per a expressar la freqüència
amb què s'esdevenen cada un dels següents esdeveniments refe-
rits a les assignatures de la titulació que curses:

                          Escala
                        1 Sempre
                        2 Sovint
                    3 Algunes vegades
                      4 Gairebé mai
                          5 Mai

Els exàmens finals són coherents amb l'EC       _____
Les assignatures ofereixen continguts pràctics  _____
Els materials docents són adequats              _____
```

- **L'escala comparativa**: a diferència de les anteriors, els ítems d'aquesta escala indiquen com es comparen dos elements entre ells a criteri de l'enquestat (figura 5). Aquesta escala es considera com una escala d'intervals, per la qual cosa és lícit aplicar les tècniques d'inferència a les dades que s'hi obtenen.

Figura 5. Exemple d'ús d'una escala comparativa

```
Selecciona un nombre de l'escala per a expressar la teva opi-
nió sobre cada un dels temes següents:

                          Escala
                    1 Molt superior
                       2 Superior
                       3 Similar
                       4 Inferior
                    5 Molt inferior

Comparat amb el pla d'estudis anterior,
el nou pla d'estudis et sembla                  _____

Comparat amb el sistema d'avaluació anterior,
el nou sistema d'avaluació et sembla            _____
```

- **L'escala lineal numèrica**: aquesta escala és també similar a la de Likert encara que els ítems extrems solen fer referència al grau d'importància que assigna l'enquestat a un tema i els ítems intermedis no solen estar etiquetats (figura 6). Per això últim, es considera una escala d'intervals.

Figura 6. Exemple d'ús d'una escala lineal numèrica

```
Selecciona un nombre de l'escala per a expressar la teva opi-
nió sobre el nivell de rellevància de cada un dels següents
temes referits a les assignatures que curses:

                          Escala
  Màxima rellevància 1   2   3   4   5   6 Mínima rellevància

L'ús de recursos d'Internet                     _____
L'ús de materials actualitzats                  _____
L'ús dels fòrums i debats                       _____
```

- **L'escala de diferències semàntiques**: aquesta escala consisteix a definir dos extrems caracteritzats per adjectius contraposats i, posteriorment, definir una graduació d'ítems no etiquetats entre ambdós (figura 7). També es considera com una escala d'intervals.

Figura 7. Exemple d'ús d'una escala de diferències semàntiques

En relació amb la formació que reps en aquesta universitat, selecciona un valor numèric segons la proximitat respecte a cada adjectiu:

Teòrica	1	2	3	4	5	6	7	Pràctica
Econòmica	1	2	3	4	5	6	7	Cara
Actualitzada	1	2	3	4	5	6	7	Desfasada

2. Disseny i selecció de la mostra

Com ja hem comentat en l'apartat anterior, en tota enquesta hi ha dos tipus d'errors que convé tenir presents: (a) l'error mostral, que és la diferència entre l'estimador obtingut a partir de les observacions (com ara la mitjana mostral $\overline{x}$) i el verdader valor del paràmetre poblacional (com ara la mitjana poblacional μ), i (b) el biaix o error no mostral, que engloba tots els restants tipus d'errors que poden ocórrer durant el desenvolupament i anàlisi d'una enquesta, com ara errors en el disseny de les preguntes, errors causats per la no-resposta (*missing data*), errors en la selecció dels individus a enquestar, errors en el registre i processament de les dades, etc.

Exemple

Recordeu que el terme *estadístic* fa referència a una mostra, mentre que el terme *paràmetre* fa referència a tota la població. Així, per exemple, l'estadístic mitjana mostral és un estimador del paràmetre mitjana poblacional.

Les enquestes poden classificar-se segons el mètode de mostreig utilitzat. Així, es parla de **mostreig probabilístic** quan cada un dels individus que componen el marc del mostreig (elements de la població susceptibles de ser elegits) té una probabilitat coneguda de ser seleccionat. Per contra, es parla de **mostreig no probabilístic** quan no és possible conèixer quina és la probabilitat que té cada element de ser seleccionat. Els mostreigs no probabilístics poden ser de gran utilitat com a eina exploratòria, però no permeten conèixer la precisió de les estimacions que s'obtenen per als paràmetres poblacionals, per la qual cosa no donen informació sobre l'error mostral que s'està cometent. Exemples de mostreigs no probabilístics serien els següents:

- A fi de conèixer l'opinió dels estudiants d'una universitat presencial sobre el seu nou campus virtual, s'enquesten els matriculats d'una assignatura concreta.

- A fi de conèixer l'opinió dels clients d'un nou centre comercial, es demanen voluntaris per a respondre un qüestionari.

- A fi de conèixer l'opinió dels usuaris d'una base de dades documental, un directiu selecciona una mostra d'usuaris que, segons el seu criteri, són representatius del conjunt d'usuaris.

Els mostrejos probabilístics, per la seva part, sí que permeten calcular intervals de confiança per als paràmetres poblacionals a partir dels les observacions de la mostra. Això és, els mostrejos probabilístics permeten conèixer la magnitud de l'error mostral que s'està cometent. En aquest apartat es descriuran quatre dels mètodes probabilístics més populars: el mostreig aleatori simple, el mostreig sistemàtic, el mostreig estratificat, i el mostreig per conglomerats.

2.1. Mostreig aleatori simple

En un **mostreig aleatori simple**, tots els elements del marc mostral (elements de la població que són candidats a ser seleccionats) tenen la mateixa probabilitat de ser elegits. Per a seleccionar, mitjançant mostreig aleatori simple, n elements d'entre els N que componen la llista de candidats a ser elegits, se sol assignar un nombre natural (1, 2, 3..., N) a cada un dels elements de la llista i a continuació, es generen a l'atzar n nombres aleatoris diferents, que identifiquen els elements seleccionats.

D'acord amb la teoria de l'estadística inferencial, si se selecciona una mostra aleatòria prou gran (a la pràctica $n \geq 30$ sol ser suficient), el teorema central del límit permet obtenir intervals de confiança per a la mitjana poblacional μ. En particular:

Per a un nivell de confiança del 95%, un interval de confiança per a la mitjana poblacional, μ, és donat per:

$$\bar{x} \pm 1,96 \cdot \sqrt{\frac{N-n}{N}} \left(\frac{s}{\sqrt{n}} \right)$$

on s representa la desviació estàndard de les observacions mostrals.

Exemple: un diari d'economia té actualment N = 8.000 lectors subscrits. Una mostra aleatòria simple de n = 484 lectors és elegida per a realitzar una enquesta. Després d'analitzar les dades de l'esmentada enquesta, se sap que la mitjana dels ingressos mensuals dels lectors seleccionats a la mostra és de $\bar{x}$ = 30.500 euros, i que la corresponent desviació estàndard són de s = 7.040 euros.

La mitjana mostral, $\bar{x}$, és un bon estimador de la mitjana poblacional, μ. A més, un interval de confiança al 95% per a l'esmentada mitjana poblacional serà: $30.500 \pm 1,96 \cdot \sqrt{\frac{8.000-484}{8.000}} \left(\frac{7.040}{\sqrt{484}} \right)$ = (29.892,07, 31.107,93). En altres paraules, per a un nivell de confiança del 95%, els ingressos mitjans del conjunt dels 8.000 lectors subscrits al diari oscil□laran entre 29.892 i 31.108 euros.

De forma similar, és possible calcular intervals de confiança per a altres paràmetres de la població, com el total acumulat d'una població, com per exemple la demanda total de la població, la riquesa total d'una població, etc.

L'estadístic $N \cdot \bar{x}$ és un bon estimador del total acumulat d'una població, $N \cdot \mu$. A més, si (a, b) és un interval de confiança al 95% per a la mitjana poblacional, μ, un interval de confiança al 95% per a $N \cdot \mu$ és donat per ($N \cdot a$, $N \cdot b$).

Exemple: volem estimar el nombre total de visites anuals que reben els portals web de les universitats pertanyents a un rànquing que inclou les 500 millors del món. Per a això, hem seleccionat una mostra aleatòria de 50 universitats pertanyents a l'esmentat rànquing i hem obtingut els estadístics mostrals següents: el nombre mitjà de visites anuals és de 22.000, en què la desviació estàndard és de 4.000.

En primer lloc, notem que $N \cdot \bar{x}$ =11.000.000 serà un bon estimador per al nombre total de visites anuals que reben els portals de les top-500 universitats. Un interval de confiança al 95% per al nombre total de visites anuals serà: $500 \cdot 22.000 \pm 1{,}96 \cdot 500 \cdot \sqrt{\frac{500-50}{500}}\left(\frac{4.000}{\sqrt{50}}\right)$ = (10.474.077, 11.525.923). En altres paraules, per a un nivell de confiança del 95%, el nombre total de visites anuals que rebran els 500 portals web serà entre 10,47 milions i 11,53 milions.

Indicació

A l'hora de realitzar els càlculs, es recomana fer servir almenys quatre decimals per a no perdre massa precisió en l'arrodoniment, especialment quan *N* és un nombre molt gran.

Finalment, també és possible obtenir intervals de confiança per a la proporció d'elements d'una població que satisfan unes determinades condicions, com, per exemple, proporció d'individus que usen un servei, proporció d'individus amb estudis superiors, etc.

Per a un nivell de confiança del 95%, un interval de confiança per a la proporció p d'elements d'una població que compleix una determinada condició és donada per:

$$p' \pm 1{,}96 \cdot \sqrt{\left(\frac{N-n}{N}\right) \cdot \left(\frac{p'(1-p')}{n-1}\right)}$$

on p' és la proporció d'elements de la mostra que la compleixen.

Exemple: continuant amb l'exemple anterior dels portals web de les universitats pertanyents al rànquing de les top-500, es vol estimar el percentatge de portals que disposen d'un programa institucional –a l'estil del MIT OpenCourseWare– per a oferir continguts formatius en obert. De les 50 universitats que constitueixen la mostra, un total de 35 disposen del programa esmentat.

La proporció mostral p' = 35/50 = 0,70 = 70%, és un bon estimador del percentatge d'universitats en el top-500 que tindran un programa com aquest. A més, és possible obtenir un interval de confiança al 95% per a l'esmentada proporció poblacional: $0{,}7 \pm 1{,}96 \cdot \sqrt{\left(\frac{500-50}{500}\right) \cdot \left(\frac{0{,}7(1-0{,}7)}{50-1}\right)}$ = (0,5783, 0,8217).

En altres paraules, amb un nivell de confiança del 95% podem afirmar que entre el 58% i el 82% d'universitats en el top-500 disposen d'un programa de continguts formatius en obert. Observem que, en aquest cas, l'interval de con-

fiança és poc precís (hi ha uns 24 punts percentuals de diferència entre els extrems de l'interval), la qual cosa es deu al fet que la mida de la mostra és relativament petita.

2.2. Mostreig sistemàtic

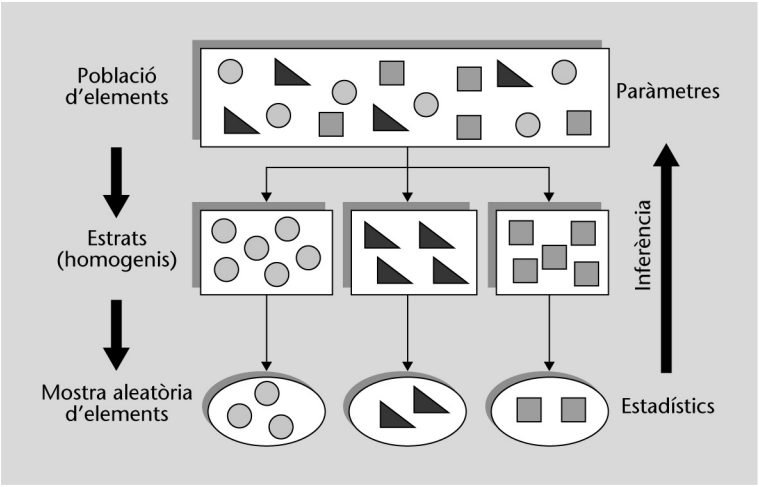
El **mostreig sistemàtic** consisteix a utilitzar una regla per a seleccionar de forma sistemàtica els elements d'una mostra. Aquest mostreig se sol usar en poblacions grans i homogènies com a alternativa al mostreig aleatori simple, especialment en aquelles situacions en les quals el procés d'assignar un nombre enter a cada element d'una llarga llista pot resultar complicat o costós en temps (com, per exemple, assignar un nombre enter a cada un dels números d'una guia telefònica, assignar un nombre enter a cada un dels clients que accedeix a un centre comercial un dia determinat, etc.). Així, per exemple, si es vol seleccionar una mostra de 30 telèfons de la guia telefònica d'una ciutat, una forma sistemàtica de fer-ho seria escollir a l'atzar el primer i, posteriorment, escollir un telèfon qualsevol de cada una de les 29 pàgines següents. Un altre exemple: si es vol entrevistar 40 clients d'un gran centre comercial, una forma sistemàtica de seleccionar la mostra seria començar per un a l'atzar i, a continuació, escollir cada 5 minuts el nou client que accedeixi al centre en aquell precís instant. Sovint, aquest tipus de mostreig es pot considerar com a equivalent a un mostreig aleatori simple, especialment quan la llista o marc mostral segueix un ordre aleatori. Per exemple, fer una selecció sistemàtica d'elements en una llista que segueix un ordre aleatori és tècnicament equivalent a realitzar directament una selecció aleatòria d'elements que no segueixin un ordre aleatori.

2.3. Mostreig aleatori estratificat (grups homogenis)

El **mostreig aleatori estratificat** se sol usar en els casos en els quals resulta fàcil agrupar els elements de la població considerada en subgrups de composició homogènia anomenats **estrats**. Per exemple, treballadors d'una organització agrupats per departament, estudiants d'una universitat agrupats per titulació, habitants d'un país agrupats per nivell de renda o edat, revistes científiques agrupades per àmbit temàtic, etc. Quan la variabilitat dins de cada estrat és menor que la variabilitat entre estrats, aquest tipus de mostreig tendeix a proporcionar més precisió que un mostreig aleatori simple a l'hora d'estimar els paràmetres poblacionals.

Així, el mostreig aleatori per estrats consisteix a: (a) classificar els N elements d'una població en H grups o estrats (de manera que els elements de cada estrat siguin similars entre ells), i (b) seleccionar a continuació una mostra aleatòria simple per a cada un dels estrats (figura 8). Els estadístics obtinguts per a cada estrat són combinats posteriorment per a obtenir estimacions d'alguns paràmetres com la mitjana, el total acumulat o la proporció de la població.

Figura 8. Mostreig aleatori estratificat



En un mostreig per estrats, és possible obtenir un bon estimador de la mitjana poblacional fent una mitjana ponderada de les mitjanes mostrals obtingudes a cada estrat. En concret, $\bar{x}_E = \frac{1}{N}\sum_{i=1}^{H} N_i \cdot \bar{x}_i$ és un bon estimador de μ, on n_i representa el nombre total d'elements de l'estrat *i*-èsim i $\bar{x}_i$ representa la mitjana de la mostra associada a l'esmentat estrat.

Per a un nivell de confiança del 95%, un interval de confiança per a la mitjana poblacional, μ, és donat per:

$$\bar{x}_E \pm 1{,}96 \cdot \sqrt{\frac{1}{N^2}\sum_{i=1}^{H} N_i (N_i - n_i)\frac{s_i^2}{n_i}}$$

on n_i i s_i representen, respectivament, la mida i la desviació estàndard de la mostra associada a l'estrat *i*-èsim.

D'altra banda, l'estadístic $N \cdot \bar{x}_E$ és un bon estimador del total acumulat d'una població, $N \square \mu$. A més, si (*a*, *b*) és un interval de confiança al 95% per a la mitjana poblacional, μ, un interval de confiança al 95% per a $N \square \mu$, és donat per $(N \cdot a, N \cdot b)$.

Finalment, un interval de confiança per a la proporció *p* d'elements d'una població que compleix una determinada condició és donada per:

$$p'_E \pm 1{,}96 \cdot \sqrt{\frac{1}{N^2}\sum_{i=1}^{H} N_i (N_i - n_i) \cdot \left(\frac{p'_i (1 - p'_i)}{n_i - 1}\right)}$$

on $p'_E = \frac{1}{N}\sum_{i=1}^{H} N_i \cdot p'_i$ és una mitjana ponderada de les proporcions p'_i d'elements de la mostra que la compleixen per a l'estrat *i*-èsim.

Exemple: fa dos anys, es van graduar en una universitat un total de 1.500 estudiants. Per a conèixer el salari mitjà dels esmentats estudiants, tant a nivell global com per titulació, es van agrupar els estudiants per titulacions (estrats) i es va enquestar un total de 180 exestudiants. La taula 1 inclou, per ordre de columnes, el nombre de graduats en cada estrat, la mida de cada mostra, la mitjana mostral, la desviació estàndard mostral i la proporció d'estudiants amb un sou superior als 36.000 euros anuals.

Taula 1. Estadístics obtinguts per a cada estrat

Titulació (estrat)	N_i	n_i	$\bar{x}_i$	s_i	p'_i
Administració i Direcció d'Empreses	500	45	30.000	2.000	4/45
Informació i Documentació	350	40	28.500	1.700	2/40
Eng. Informàtica	200	30	31.500	2.300	7/30
Psicologia	300	35	27.000	1.600	1/35
Eng. Telecomunicació	150	30	31.000	2.250	6/30
Total	1.500	180			

Un bon estimador del salari mitjà per al conjunt de 1.500 graduats és donat per la mitjana ponderada de les diferents mitjanes mostrals:

$$\bar{x}_E = \frac{1}{1.500}(500 \cdot 30.000 + 350 \cdot 28.500 + 200 \cdot 31.500 + 300 \cdot 27.000 + 150 \cdot 31.000)$$

= 29.350 euros.

A més, es pot obtenir el corresponent interval de confiança, per a un nivell de confiança del 95%, per a la mitjana poblacional:

$$29.350 \pm 1,96 \cdot \sqrt{\frac{1}{1.500^2}\left(500 \cdot (500-45)\frac{2.000^2}{45} + \ldots + 150 \cdot (150-30)\frac{2.250^2}{30}\right)} =$$

= (29.079,33, 29.620,67), amb la qual cosa es pot afirmar, amb un nivell de confiança del 95%, que el salari mitjà del total de 1.500 graduats d'aquesta universitat està entre 29.079 i 29.621 euros per any. Per a fer aquest tipus de càlculs, és convenient usar un full de càlcul (figura 9).

Figura 9. Ús d'Excel per a fer càlculs en mostreig estratificat

	A	B	C	D	E	F	G	H
1	**Titulació (estrat)**	**N(i)**	**n(i)**	**x-bar(i)**	**s(i)**	**p(i)**	**N(i) * x-bar(i)**	**N(i) * (N(i) - n(i)) * (s(i)^2 / n(i))**
2	Administració i Direcció d'Empreses	500	45	30.000	2.000	16.528	15.000.000	20.222.222.222
3	Informació i Documentació	350	40	28.500	1.700	14.642	9.975.000	7.839.125.000
4	Eng. Informàtica	200	30	31.500	2.300	40.024	6.300.000	5.995.333.333
5	Psicologia	300	35	27.000	1.600	12.785	8.100.000	5.814.857.143
6	Eng. Telecomunicació	150	30	31.000	2.250	39.994	4.650.000	3.037.500.000
7	*Totals*	*1.500*	*180*				*44.025.000*	*42.909.037.698*
8								
9			z =	1,96				
10			x(E) =	29.350				
11			s(E) =	138,10				
12		x(E) - z*s(E) =		29.079,3306				
13		x(E) + z*s(E) =		29.620,6694				
14								

$\sum_{i=1}^{5} N_i \cdot \bar{x}_i$

$\sum_{i=1}^{5} N_i \left(N_i - n_i\right) \frac{s_i^2}{n_i}$

En segon lloc, es poden estimar els ingressos anuals totals del conjunt dels 1.500 graduats, $N \cdot \mu$, per a saber quin serà el seu potencial impacte sobre l'economia local. En aquest cas, ja que l'estimador de μ era $\bar{x}_E$ = 29.350 euros, l'estimador puntual de $N \cdot \mu$ serà 1.500 $\bar{x}_E$ = 44.025.000 i un interval de confiança al 95% serà donat per: (1.500 □ 29.079,3306, 1.500 □ 29.620,6694) = (43.618.995,86, 44.431.004,14). En altres paraules, es pot afirmar amb un nivell de confiança del 95% que seran necessaris entre 43,6 i 44,4 milions d'euros per a cobrir els salaris anuals dels 1.500 graduats.

En tercer lloc, un bon estimador del percentatge d'estudiants de la població els ingressos de la qual superen els 36.000 euros serà donat per la mitjana ponderada dels percentatges a cada estrat: $p'_E = \frac{1}{1.500}\left(500\frac{4}{45} + \ldots + 150\frac{6}{30}\right)$ = 0,0981, i així, aproximadament, només un 9,8% dels salaris dels 1.500 graduats serà superior als 36.000 euros anuals. Finalment, es pot obtenir un interval de confiança al 95% per al percentatge poblacional anterior:

$$0,0981 \pm 1,96 \cdot \sqrt{\frac{1}{1.500^2}\left(500(500-45)\frac{(4/45)(41/45)}{45-1} + \ldots + 150(150-30)\frac{(6/30)(24/30)}{30-1}\right)}$$

= (0,0584, 0,1379), és a dir, es pot afirmar amb un 95% de confiança que el percentatge de graduats en la promoció de fa dos anys els ingressos de la qual superen els 36.000 euros anuals oscil□la entre un 5,8% i un 13,8%.

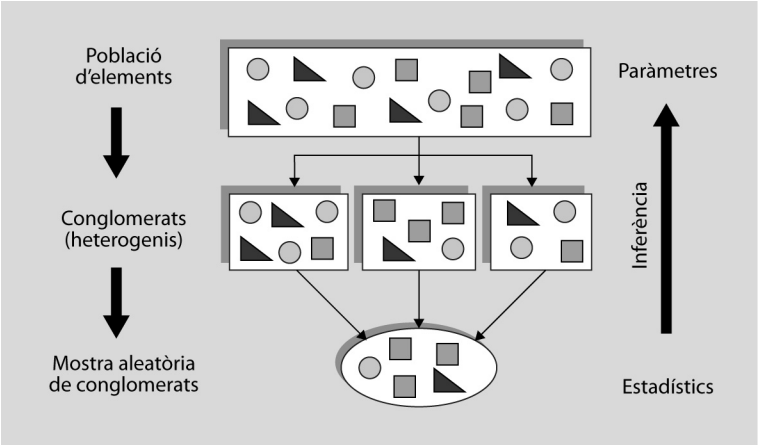
2.4. Mostreig per conglomerats (clústers o grups heterogenis)

El **mostreig per conglomerats** se sol usar en els casos en els quals resulta fàcil agrupar els elements de la població considerada en subgrups de composició

heterogènia anomenats conglomerats, cada uns dels quals és una representació a petita escala de la població total (és a dir, es pressuposa una gran variabilitat entre els elements d'un mateix conglomerat). Per exemple: els habitants d'una gran ciutat poden ser agrupats per barris, els usuaris d'un servei web poden ser agrupats per països de procedència, les revistes científiques poden ser agrupades per editorial, etc. De fet, una de les principals aplicacions del mostreig per conglomerats està relacionada amb el mostreig per àrees o regions geogràfiques, on els conglomerats solen ser països, regions, ciutats o barris. El mostreig per conglomerats permet reduir els costos de desplaçaments entre zones geogràficament disperses i, alhora, evita haver de generar llistats exhaustius de tota la població, ja que només són necessaris llistats exhaustius de cada conglomerat seleccionat.

Així, el mostreig per conglomerats consisteix a: (a) classificar els N elements d'una població en H grups o conglomerats (de manera que els elements de cada conglomerat presentin molta variabilitat entre ells), (b) seleccionar a continuació una mostra aleatòria simple de h conglomerats, i (c) per a cada conglomerat de la mostra seleccionada, o bé enquestar a cada un dels elements que el componen –mostreig per conglomerats en una etapa– o bé seleccionar una nova mostra aleatòria d'elements per a enquestar –mostreig en dues etapes– (figura 10). Si bé tant en un cas com en un altre és possible obtenir estimadors puntuals i per intervals per a diversos paràmetres poblacionals, es tractarà només el mostreig per conglomerats en una etapa (és a dir, se suposarà que, una vegada seleccionada la mostra de conglomerats, s'enquesta tots els elements de cada conglomerat seleccionat).

Figura 10. Mostreig per conglomerats



En un mostreig per conglomerats, és possible obtenir un bon estimador de la mitjana poblacional μ mitjançant l'expressió $\bar{x}_C = \frac{\sum_{i=1}^{h} y_i}{\sum_{i=1}^{h} N_i}$, on n_i representa el nombre total d'elements del conglomerat i-èsim i y_i representa el valor total de les observacions de l'esmentat conglomerat.

Per a un nivell de confiança del 95%, un interval de confiança per a la mitjana poblacional, μ, és donat per:

$$\bar{x}_C \pm 1{,}96 \cdot \sqrt{\frac{H-h}{H \cdot h\left(\frac{N}{H}\right)^2}\left(\frac{\sum_{i=1}^{h}\left(y_i - \bar{x}_C \cdot N_i\right)^2}{h-1}\right)}$$

D'altra banda, l'estadístic $N \cdot \bar{x}_C$ és un bon estimador del total acumulat d'una població, $N \square \mu$. A més, si (a, b) és un interval de confiança al 95% per a la mitjana poblacional, μ, un interval de confiança al 95% per a $N \square \mu$, és donat per $(N \cdot a, N \cdot b)$.

Finalment, un interval de confiança per a la proporció p d'elements d'una població que compleix una determinada condició és donada per:

$$p'_C \pm 1{,}96 \cdot \sqrt{\frac{H-h}{H \cdot h\left(\frac{N}{H}\right)^2}\left(\frac{\sum_{i=1}^{h}\left(m_i - p'_C \cdot N_i\right)^2}{h-1}\right)}$$

on m_{eu} és el nombre d'elements del conglomerat i-èsim que compleix una determinada característica i $p'_C = \frac{\sum_{i=1}^{h} m_i}{\sum_{i=1}^{h} N_i}$ és bon estimador de la mitjana d'elements de la població que compleixen l'esmentada característica.

Exemple: el sistema sanitari d'atenció primària d'un país és compost per un total de 12.000 metges distribuïts en 1.000 centres d'atenció primària (conglomerats). A fi d'obtenir certa informació sobre la població de metges considerada, i davant de la dificultat de realitzar enquestes a metges de tots els centres, es du a terme un mostreig per conglomerats en el qual se seleccionen de manera aleatòria un total de 10 centres d'atenció primària. A continuació, es passa una enquesta als metges de cada un dels centres escollits. La taula 2 inclou, per ordre de columnes, l'identificador del centre, el nombre de metges que hi treballen, el nombre total de visites associades amb una certa malaltia que rep el centre en una setmana normal, i el nombre de metges que són dones.

Taula 2. Estadístics obtinguts per a cada conglomerat de la mostra

Centre (conglomerat)	Nombre de metges N_i	Total de visites y_i	Nombre de dones m_i
CAP-01	8	320	2
CAP-02	25	1.125	8
CAP-03	4	115	0
CAP-04	17	714	6
CAP-05	7	247	1
CAP-06	3	94	2
CAP-07	15	634	2
CAP-08	4	147	0
CAP-09	12	481	5
CAP-10	33	1.567	9
Totals	**128**	**5.444**	**35**

En primer lloc, un bon estimador per al nombre mitjà de visites setmanals que rep cada metge és donat per: $\bar{x}_C = \frac{5.444}{128}$ = 42,5313, és a dir, de mitjana cada metge del sistema sanitari rebrà unes 43 visites setmanals.

És possible obtenir un interval de confiança al 95% per a l'esmentada mitjana poblacional:

$$42,5313 \pm 1,96 \cdot \sqrt{\frac{1.000-10}{1.000 \cdot 10\left(\frac{12.000}{1.000}\right)^2}\left(\frac{(320-42,5313\cdot 8)^2+\ldots+(1.567-42,5313\cdot 33)^2}{10-1}\right)}$$

= 42,5313 ± 1,96 □ 1,7299 = (39,14, 45,92). En altres paraules, es pot afirmar amb un nivell de confiança del 95% que la mitjana de visites setmanals per metge en el sistema sanitari del país és entre 39 i 46 (figura 11).

Figura 11. Ús d'Excel per a realitzar càlculs en mostreig per conglomerats

	A	B	C	D	E
1	**Centre**	**Nombre de metges**	**Total de visites**	**Nombre de dones**	
2	**(conglomerat)**	N_i	Y_i	m_i	**[y(i) - x(C)*N(i)]^2**
3	CAP-01	8	320	2	410,06
4	CAP-02	25	1125	8	3809,20
5	CAP-03	4	115	0	3038,77
6	CAP-04	17	714	6	81,56
7	CAP-05	7	247	1	2572,39
8	CAP-06	3	94	2	1128,54
9	CAP-07	15	634	2	15,75
10	CAP-08	4	147	0	534,77
11	CAP-09	12	481	5	862,89
12	CAP-10	33	1567	9	26722,03
13	*Totals*	*128*	*5444*	*35*	*39175,97*
14					
15		z =	1,96		
16		x(C) =	42,53		
17		s(C) =	1,73		
18		x(C) - z*s(C) =	39,14		
19		x(C) + z*s(C) =	45,92		

$$\sum_{i=1}^{10}\left(y_i - \bar{x}_C \cdot N_i\right)^2$$

En segon lloc, es poden estimar les visites setmanals totals del conjunt dels 12.000 metges, N □ μ, per a saber quin serà el seu potencial impacte sobre el sistema sanitari. En aquest cas, ja que l'estimador de μ era $\bar{x}_C$ = 42,5313, l'estimador puntual de N □ μ serà 12.000 $\bar{x}_C$ = 510.375 i un interval de confiança al 95% serà donat per: (12.000 □ 39,1.406, 12.000 □ 45,9219) = (469.687,38, 551.062,62). En altres paraules, es pot afirmar amb un nivell de confiança del 95% que el sistema d'atenció primària del país rebrà entre 469.687 i 551.063 visites en una setmana normal.

En tercer lloc, un bon estimador del percentatge de metges que són dones serà donat per: $p'_C = \frac{35}{128}$ = 0,2734, és a dir, aproximadament el 27,3% dels metges del sistema d'atenció primària són dones. Finalment, es pot obtenir un interval de confiança al 95% per al percentatge poblacional anterior:

$$0,2734 \pm 1,96 \cdot \sqrt{\frac{1.000-10}{1.000 \cdot 10\left(\frac{12.000}{1.000}\right)^2}\left(\frac{(2-0,2734 \cdot 8)^2 + \ldots + (9-0,2734 \cdot 33)^2}{10-1}\right)}$$

= (0,2066, 0,3402). Així, doncs, es pot afirmar amb un 95% de confiança que el percentatge de dones en la població de metges d'assistència primària oscil·la entre un 20,7% i un 34,0%.

3. Anàlisi de qüestionaris: estudi parcial d'un cas

En aquest apartat presentem un cas d'estudi en el qual es mostren exemples de l'ús de tècniques estadístiques per a analitzar diferents tipus de preguntes pertanyents a una enquesta. L'objectiu de l'enquesta era obtenir informació concreta sobre la visió (i l'actitud) de les grans empreses d'una determinada comunitat autònoma respecte al fenomen de l'externalització dels serveis, sistemes i tecnologies de la informació. Per a això, vam dissenyar una enquesta formada per diverses preguntes, algunes de les quals basades en escales nominals i altres en escales d'intervals equidistants. La població objectiu de l'enquesta eren els directius de serveis, sistemes i tecnologies de la informació de les empreses, amb seu social en l'esmentada comunitat autònoma, el volum de facturació o d'empleats de les quals superaven unes determinades quantitats establertes *a priori* pels investigadors. Del llistat complet d'empreses que complien els esmentats requisits, vam seleccionar una mostra aleatòria de 100 empreses i vam enviar el qüestionari als corresponents directius, del qual vam obtenir una taxa de resposta superior al 80%. L'aleatorietat de la mostra i l'alta taxa de resposta obtinguda són dos factors imprescindibles a l'hora de generalitzar, amb certes garanties, els resultats de l'enquesta al conjunt de la població d'empreses que satisfan les característiques anteriorment descrites.

Aclariment

L'objectiu últim d'aquesta secció no és explicar amb detall un cas complet d'anàlisi d'una enquesta (ja que per a això caldria un mòdul sencer), sinó proporcionar exemples concrets de com es poden utilitzar molts dels conceptes i tècniques vistos en mòduls anteriors per a analitzar enquestes. Així doncs, aquesta secció mostra com es poden combinar moltes de les tècniques estadístiques vistes anteriorment per a extreure informació a partir de les dades d'una enquesta.

3.1. Exemple d'ús d'estadístics descriptius i intervals de confiança

Una de les preguntes de l'enquesta demanava especificar el nombre de treballadors de plantilla del departament de tecnologies de la informació i la comunicació (TIC). L'esmentada pregunta està associada a una variable aleatòria discreta, per la qual cosa es poden considerar els estadístics descriptius d'aquesta tal com mostra la figura 12.

Nota

Tant les sortides (*outputs*) com els gràfics d'aquesta secció han estat generats amb el R Commander, utilitzant els menús i les opcions ja explicats en mòduls anteriors.

Figura 12. Estadístics descriptius de la variable *Empleados*

```
   Empleados
 Min.   :  1.00
 1st Qu.: 57.00
 Median : 58.00
 Mean   : 56.18
 3rd Qu.: 60.00
 Max.   :450.00
 NA's   :5


    mean       sd IQR 0% 25% 50% 75% 100%  n NA
56.18182 49.78208   3  1  57  58  60  450 77  5


        One Sample t-test

data:  Empleados
t = 9.903, df = 76, p-value = 2.534e-15
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 44.88267 67.48097
sample estimates:
mean of x
 56.18182
```

Recordatori R Commander

Per a obtenir els estadístics descriptius, useu *Estadísticos > Resúmenes > Conjunto de datos activo* o *Estadísticos > Resúmenes > Resúmenes numéricos*. Per a obtenir l'interval de confiança, utilitzeu *Estadísticos > Medias > Test t para una media.*

Observem que aquesta pregunta va ser contestada correctament per un total de 77 dels 82 directius que van respondre l'enquesta (5 directius van deixar sense contestar aquesta pregunta). La mitjana de treballadors del departament TIC és de 56,18 per a les empreses que van contestar la pregunta. Observem també que el nombre de treballadors en l'esmentat departament és molt variable, i oscil□la entre un mínim d'1 treballador i un màxim de 450, la qual cosa fa pensar en diferents nivells d'externalització dels serveis i sistemes TIC. Atès que la mostra és aleatòria, s'ha pogut obtenir un interval de confiança per a la mitjana de treballadors en departaments TIC de totes les empreses de la població considerada. En aquest cas, usant un nivell de confiança del 95%, hem obtingut l'interval (44,88267, 67,48097). Així, doncs, amb un 95% de confiança, podem afirmar que en mitjana aquests departaments tenen entre 45 i 67 empleats. Així mateix, resulta possible agrupar els valors obtinguts per a la variable anterior en categories d'empreses segons el nombre d'empleats en el departament TIC, la qual cosa permet obtenir taules i gràfics circulars per a representar les freqüències associades a cada tipus d'empresa participant en l'enquesta (figures 13 i 14).

Figura 13. Taula de freqüències per a cada categoria

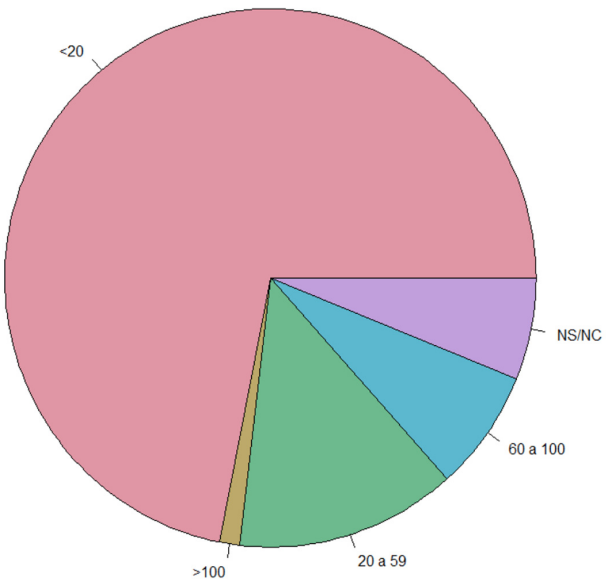
```
counts:
Categoría
     <20      >100  20 a 59 60 a 100     NS/NC
      59         1        11         6         5

percentages:
Categoría
     <20      >100  20 a 59 60 a 100     NS/NC
   71.95      1.22    13.41      7.32      6.10
```

Recordatori R Commander

Per a obtenir una taula de freqüències, useu *Estadísticos > Resúmenes > Distribución de frecuencias*.

Figura 14. Gràfic circular representant els percentatges de cada categoria



Recordatori R Commander

Per a obtenir un diagrama circular, useu *Gráfica de sectores*.

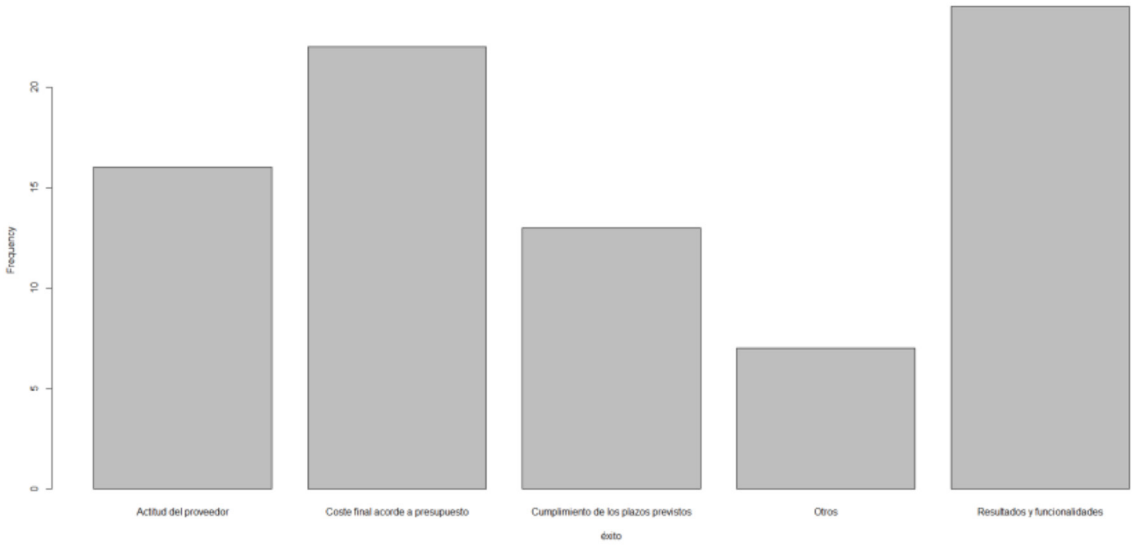
En aquest cas s'aprecia que aproximadament un terç (71,95%) de les empreses participants tenen departaments TIC relativament petits (menys de 20 empleats), la qual cosa indueix a pensar que tindran pocs serveis i sistemes d'informació externalitzats.

Una altra de les preguntes de l'enquesta demanava seleccionar, d'entre una llista de factors, els (un o més) que es tenien en compte a l'hora de valorar el nivell d'èxit d'un projecte TIC finalitzat. Atès que es tracta d'una pregunta amb resposta múltiple (es poden seleccionar diversos factors alhora), en aquest cas es pot emprar un diagrama de barres, com es mostra a la figura 15, per a representar el percentatge de citacions de cada factor i caracteritzar així els factors més freqüentment citats.

Recordatori R Commander

Per a obtenir un diagrama de barres, useu *Gráficas* > *Gráfica de barras*. Noteu que és possible personalitzar els gràfics (per exemple, fent que les barres siguin horitzontals) mitjançant la pestanya *Opciones* a dins de *Gráfica de barras*.

Figura 15. Gràfic de barres amb freqüència de citacions de factors d'èxit



En aquest cas, queda clar que a l'hora de valorar l'èxit d'un projecte hi ha tres factors que s'usen gairebé sempre ("resultats i funcionalitat", "cost final d'acord amb el pressupost" i "compliment dels terminis previstos"). Observeu que el factor *otros* ha estat seleccionat en menys casos, la qual cosa indica que potser hi ha un factor no considerat entre els anteriors que també tingui la seva importància relativa.

3.2. Exemple d'ús de contrastos d'hipòtesis per a comparar 2 grups

En una altra de les preguntes del qüestionari se li proposava a l'enquestat una llista de cinc ítems o motius pels quals una empresa podia optar per l'externalització dels seus serveis i sistemes TIC (per exemple: "superar les limitacions de les qualificacions professionals i tècniques de l'equip intern", "promoure canvis organitzatius, estructurals o culturals interns", "aconseguir millors nivells de qualitat del servei o sistema final", "reduir

els costos totals", etc.). A continuació se li demanava valorar, usant una escala lineal numèrica, la importància de cada un dels esmentats ítems o motius d'externalització, tant des d'un punt de vista teòric com des d'un punt de vista pràctic (per exemple, l'enquestat havia d'emetre dues avaluacions per a cada ítem: d'una banda la corresponent a la importància teòrica o hipotètica del motiu d'externalització i, de l'altra, la corresponent a la importància real manifestada a la pràctica quotidiana). L'escala lineal numèrica oscil□lava entre 1 (molt poc important) i 5 (molt important). Un dels objectius d'aquesta pregunta era determinar si per a cada un dels ítems hi havia diferències significatives entre la importància hipotètica o teòrica i la importància real a la pràctica del dia a dia (aquestes diferències posarien de manifest l'existència d'altres factors associats amb la pràctica diària que alteraven significativament el nivell d'importància teòric de cada motiu). En aquest cas es va optar per realitzar un contrast d'hipòtesi per a comparar les dues mitjanes que s'obtenien per a cada un dels ítems (és a dir, per a cada motiu es realitza un contrast d'hipòtesi sobre la igualtat de la puntuació mitjana teòrica i la puntuació mitjana pràctica). La figura 16 mostra la sortida de R Commander per als dos primers tests corresponents als dos primers motius de la llista (ítems A1 i A2). S'observa que en el cas del primer motiu d'externalització considerat, no sembla haver-hi diferències significatives, per a un nivell de significació $\alpha = 0{,}05$, entre les mitjanes respectives de les puntuacions teòriques (A1T) i les pràctiques (A1P). Per contra, en el cas del segon motiu, el *p*-valor obtingut és molt baix (*p-value* = 0,000), cosa que evidencia l'existència de diferències significatives entre la importància hipotètica del motiu i la importància a la pràctica.

Figura 16. Test d'hipòtesi per a comparar mitjanes de motius

```
> with(Dataset, (t.test(A1P, A1T, alternative='two.sided', conf.level=.95, paired=TRUE)))

        Paired t-test

data:  A1P and A1T
t = 1, df = 81, p-value = 0.3203
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.02413869  0.07291918
sample estimates:
mean of the differences
             0.02439024


> with(Dataset, (t.test(A2P, A2T, alternative='two.sided', conf.level=.95, paired=TRUE)))

        Paired t-test

data:  A2P and A2T
t = -3.7016, df = 81, p-value = 0.0003894
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5812567 -0.1748409
sample estimates:
mean of the differences
             -0.3780488
```

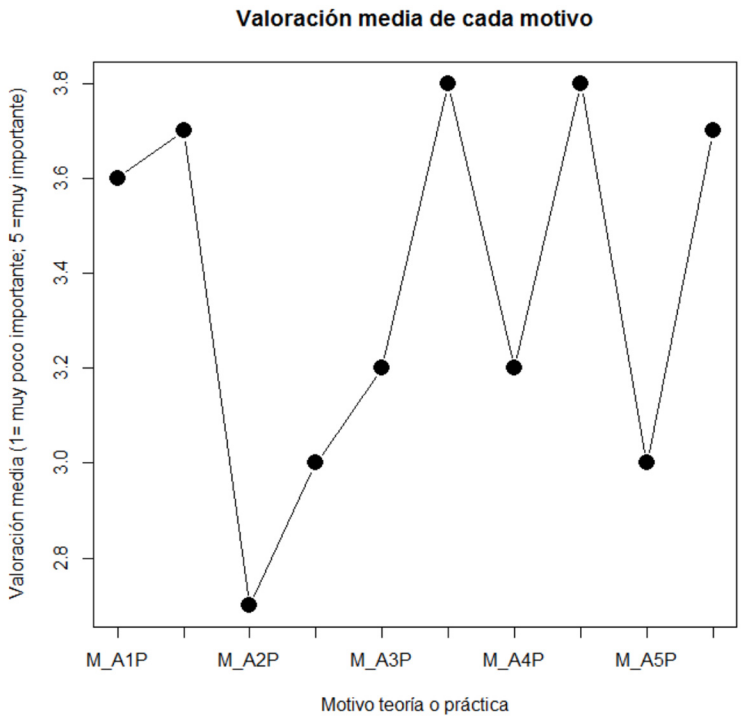
Recordatori R Commander

Per a fer un contrast d'hipòtesis per a dues mostres dependents, useu *Estadísticos > Medias > Test t para datos relacionados.*

La figura 17 mostra el valor d'importància mitjà obtingut per a cada un dels cinc motius d'externalització considerats, tant des d'un punt de vista teòric com des d'un punt de vista pràctic. S'observa que, per a tots els parells teoria - pràctica, el valor teòric sempre és superior al valor pràctic. Això fa sos-

pitar que si bé alguns motius d'externalització haurien de ser considerats molt importants, a la pràctica això no sempre és possible a causa de la influència d'altres factors (condicions laborals, recursos disponibles, etc.). Precisament els contrastos d'hipòtesi permeten detectar aquells casos en els quals les diferències entre teoria i pràctica són significatives. S'observa també en aquesta figura quina és la importància relativa de cada motiu a l'hora de decidir sobre externalitzar o no els serveis i sistemes TIC.

Figura 17. Comparació visual de la importància relativa dels ítems



Recordatori R Commander

La figura mostra un núvol de punts obtingut amb *Gráfica de las medias*.

3.3. Exemple d'ús d'ANOVA per a comparar més de 2 grups

A fi de disposar d'informació sobre el percentatge de serveis i sistemes TIC que les empreses externalitzaven, en una de les preguntes se li va demanar a l'enquestat estimar l'esmentat valor percentual. En particular, es pretenia analitzar si aquest percentatge era el mateix per a totes les empreses amb independència de la mida del seu departament TIC o si, per contra, aquest percentatge depenia de manera significativa del nombre de treballadors en nòmina que tingués l'esmentat departament.

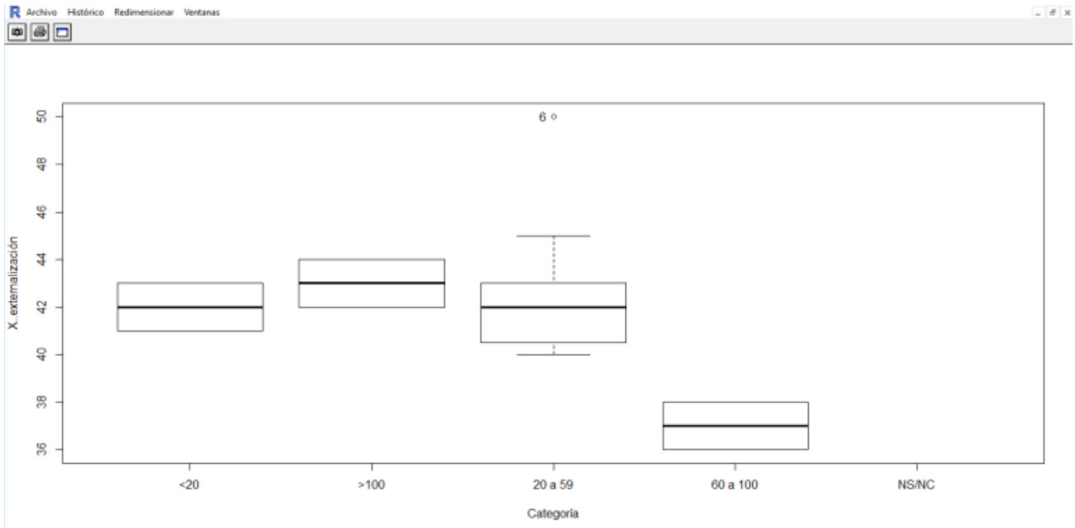
Com que s'havien predefinit quatre categories o nivells diferents d'empreses segons la dimensió del departament TIC (vegeu figura 14), resulta necessari aplicar un test ANOVA per a donar resposta al dubte formulat. La figura 18 mostra una comparativa dels diferents diagrames de caixes i bigotis (*boxplots*) per categoria o nivell. Visualment no s'observen grans diferències entre els diferents grups, llevat de potser una certa diferència amb el grup d'empreses amb departaments entre 60 i 100 empleats, els percentat-

Recordatori R Commander

Per a obtenir un diagrama de caixa (*box-plot*) múltiple, utilitzeu l'opció *Gráficos > Diagrama de cajas*. Recordeu que R Commander representa la mediana, no la mitjana als *box plots*. Aquesta és la línia horitzontal de dins les caixes.

ges d'externalització dels quals semblen una mica inferiors a la resta (fins i tot a les d'una mida més gran). En tot cas, aquestes possibles diferències visuals no semblen gaire clares.

Figura 18. *Box plots* de percentatge d'externalització per nivell



La figura 19 mostra la sortida ANOVA, que ajuda a aclarir els dubtes: un *p*-valor de 0 indica que s'han trobat indicis suficients per a rebutjar la hipòtesi nul□la que el percentatge mitjà d'externalització és el mateix per a tots els grups. Així, doncs, sembla que la mida del departament TIC tingui una influència decisiva en el percentatge de serveis i sistemes TIC que acaben essent externalitzats.

Figura 19. Contrast ANOVA per a comparar les mitjanes de percentatges

```
> summary(AnovaModel.2)
             Df Sum Sq Mean Sq F value   Pr(>F)
Categoría     3  146.8   48.94   27.12 6.81e-12 ***
Residuals    73  131.7    1.80
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
5 observations deleted due to missingness

> with(Dataset, numSummary(X..externalización, groups=Categoría, statistics=c("mean", "sd")
              mean        sd data:n data:NA
<20       41.98276 0.8269873     58       0
>100      43.00000 1.4142136      2       0
20 a 59   42.45455 2.9449495     11       0
60 a 100  37.00000 0.8944272      6       0
NS/NC          NaN        NA      0       5
```

3.4. Exemple d'ús de correlació i regressió lineal

En una de les últimes preguntes del qüestionari, es demanava als enquestats estimar les quantitats (en euros) que tenien previst invertir durant el pròxim any en adquisició de programari i nous sistemes informàtics. Sembla lògic pensar que aquestes quantitats poden estar inversament relacionades amb els percentatges d'externalització de cada empresa, això és, es podria esperar que a major percentatge d'externalització de serveis i sistemes TIC, menor inversió prevista en adquisició de programari i nous sistemes informàtics. Per a tractar de corroborar aquesta impressió i detectar una possible correlació lineal en-

tre ambdues variables es va calcular el coeficient de correlació lineal entre ambdues. La figura 20 mostra que, en efecte, hi ha una forta correlació lineal negativa entre ambdues variables, ja que el coeficient de correlació és de –0,9760152.

Figura 20. Coeficient de correlació lineal entre externalització i inversió

```
        Pearson's product-moment correlation

data:  Inversión and X..externalización
t = -38.826, df = 75, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9847262 -0.9624301
sample estimates:
       cor
-0.9760152
```

Recordatori R Commander

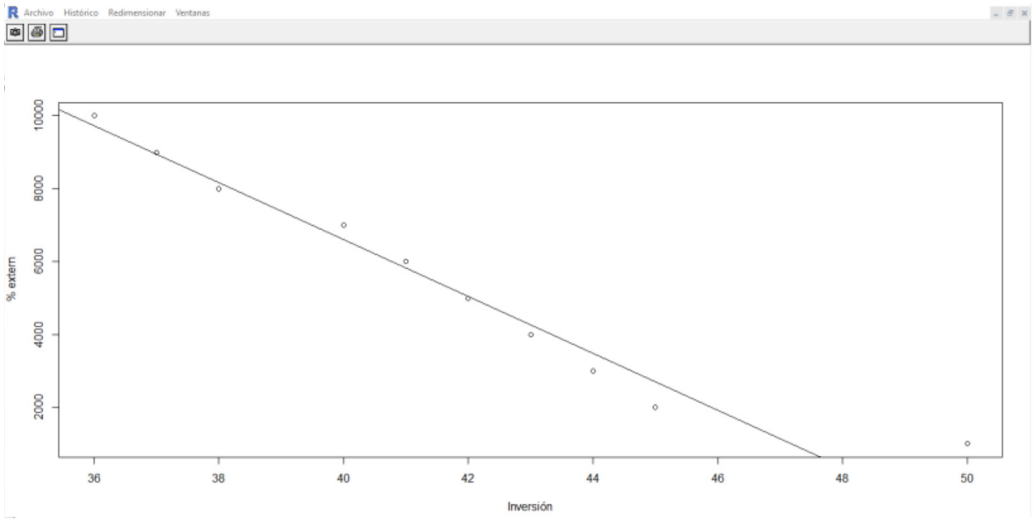
Per a calcular el coeficient de correlació, feu servir l'opció Estadísticos > Resúmenes > Test de correlación.

Té sentit, doncs, representar la recta de regressió de la inversió sobre el nivell d'externalització. L'esmentada recta es mostra a la figura 21. Com que el coeficient de determinació associat és molt alt (*R–sq* = 95,26%), es pot usar fins i tot a l'equació de l'esmentada recta per a fer estimacions sobre la inversió futura de les empreses en nous equips informàtics a partir del seu nivell d'externalització de serveis i sistemes TIC.

Recordatori R Commander

Per a obtenir el model de regressió lineal, useu l'opció *Estadísticos > Ajuste de modelos > Regresión lineal.*

Figura 21. Recta de regressió de la inversió sobre el nivell d'externalització




```
> summary(RegModel.5)

Call:
lm(formula = Inversión ~ X..externalización, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-703.81 -263.08  -42.71  177.66 2194.35

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     37787.28     837.98   45.09   <2e-16 ***
X..externalización  -779.63      20.08  -38.83   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 335.1 on 75 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.9526, Adjusted R-squared:  0.952
F-statistic:  1507 on 1 and 75 DF,  p-value: < 2.2e-16
```

Resum

Les tècniques d'investigació social basades en l'ús d'enquestes i qüestionaris són cada vegada més esteses en tots els àmbits. Tanmateix, dissenyar un bon qüestionari no és una tasca fàcil, i convé tenir presents aspectes clau com la brevetat i claredat de les preguntes, el tipus d'escala utilitzada o l'anàlisi posterior que es pretén aplicar a les dades de la mostra.

En el disseny del qüestionari i del mostreig cal tractar de minimitzar tant l'error mostral com l'error no mostral o biaix. Per a això resulta necessari conèixer bé les diferents tècniques bàsiques de mostreig que s'usen en cada cas (mostreig aleatori simple, mostreig sistemàtic, mostreig estratificat i mostreig per conglomerats).

Finalment, una vegada obtingudes les dades de l'enquesta, convé saber quines tècniques estadístiques es poden aplicar en cada cas i quin tipus d'informació poden proporcionar, tant de manera numèrica com gràfica. Precisament, l'anàlisi dels resultats obtinguts mitjançant l'ús d'aquestes tècniques comporta sovint un procés de reflexió important, és a dir, el programa estadístic sempre serà capaç de calcular nombres i generar resultats, però no sempre aquests resultats tindran sentit ni seran vàlids. És tasca de l'investigador comprovar si se satisfan les hipòtesis necessàries per a aplicar cada tècnica estadística, i interpretar i validar, si s'escau, els resultats generats pels ordinadors.

Exercicis d'autoavaluació

1) Seleccioneu un tema i dissenyeu un qüestionari per a obtenir-ne informació. El qüestionari ha de contenir una pregunta per cada tipus d'escala (nominal, ordinal, d'intervals equidistants i de proporció). Argumenteu la validesa del qüestionari i especifiqueu quin tipus de tècniques estadístiques es poden fer servir per a analitzar cada pregunta.

2) Entre els investigadors d'una universitat s'ha realitzat un estudi per a conèixer els seus hàbits de treball. Entre altres coses, l'estudi pretenia obtenir informació sobre el nombre mitjà d'articles que un investigador llegeix anualment, així com sobre quin percentatge d'aquests estan en anglès. Com que la universitat té tres grans àmbits d'investigació (*E-learning*, Computació i Societat de la Informació), es va dissenyar un mostreig per estrats en el qual es va classificar cada investigador a l'estrat corresponent al seu àmbit d'investigació. La taula següent resumeix les dades de l'enquesta:

Àmbit d'investigació (estrat)	N_i	n_i	$\bar{x}_i$	s_i	p'_i
E-learning	200	20	138	30	0,50
Computació	250	30	103	25	0,78
Societat de la Informació	100	25	210	50	0,21

Amb l'ajuda d'un model de full de càlcul (MS Excel o Open Office Calc), es demana:

a) Obtenir un interval de confiança al 95% per a la mitjana d'articles llegits anualment per la població d'investigadors de la universitat.
b) Obtenir un interval de confiança al 95% per al total d'articles llegits anualment pel conjunt d'investigadors de la universitat.
c) Obtenir un interval de confiança al 95% per al percentatge d'articles llegits que estan en anglès.

3) Les 25 biblioteques universitàries d'un país empren un total de 300 professionals en el seu servei d'obtenció de documents (SOD). A fi d'obtenir informació sobre el nombre mitjà de documents difícils d'obtenir que se sol□liciten anualment, se selecciona una mostra aleatòria de 4 biblioteques universitàries i s'enquesta cada un dels professionals del SOD respectiu. Es vol, a més, obtenir informació sobre el nombre d'experts en Tecnologies de la Informació i Comunicació de cada servei SOD analitzat. La taula inferior mostra la informació obtinguda:

Biblioteca (conglomerat)	Nombre de professionals N_i	Total de documents "difícils" y_i	Nombre d'experts en TIC m_i
SOD-01	7	95	1
SOD-02	18	325	6
SOD-03	15	190	6
SOD-04	10	140	2

Amb l'ajuda d'un model de full de càlcul (MS Excel o Open Office Calc), es demana:

a) Obtenir un interval de confiança al 95% per a la mitjana de documents "difícils d'obtenir" processats anualment per un SOD.
b) Obtenir un interval de confiança al 95% per al total de documents "difícils d'obtenir" que són processats anualment pel global dels SOD.
c) Obtenir un interval de confiança al 95% per al percentatge d'especialistes en TIC que treballen en els SOD del sistema universitari.

4) En un estudi es van entrevistar 8 individus elegits a l'atzar per a avaluar el potencial de venda d'un producte abans i després de llançar una forta campanya publicitària per televisió. L'interès per comprar el producte va ser determinat per cada individu, abans i després de la campanya, usant una escala entre 0 i 10, on els valors més grans representaven un interès més gran en adquirir el producte. La taula següent mostra els resultats obtinguts:

Individu	Després	Abans
1	6	5
2	6	4
3	7	7
4	4	3
5	3	5
6	9	8
7	7	5
8	6	6

Contrasteu la hipòtesi nul□la que, en mitjana, l'interès per adquirir el producte no ha variat després de la campanya. Useu un nivell de confiança del 95%.

5) En un estudi es van visitar cinc ciutats d'una província per a preguntar als residents sobre els seus hàbits a l'hora de fer la compra. Una de les preguntes versava sobre el nombre de dies per mes que realitzaven la compra fora de la seva província. Un total de 30 persones van participar en l'enquesta, que van proporcionar les observacions que s'inclouen en la taula següent:

Ciutat 1	Ciutat 2	Ciutat 3	Ciutat 4	Ciutat 5
1	3	1	2	5
3	3	6	5	3
2	4	2	7	2
1	3	5	4	9
1	9	6	8	8
0	7	3	1	6

Es demana:

a) Determinar si hi havia o no diferències significatives entre els hàbits de compra dels residents en funció de la seva ciutat (usar un nivell de confiança del 95%).
b) Obtenir el coeficient de correlació entre la ciutat de residència i el nombre de vegades per mes que es compra fora de la província.

Solucionari

1) Pregunta oberta, consulteu el primer apartat d'aquest material per a comprovar la validesa del qüestionari proposat.

2) La figura següent mostra els resultats obtinguts amb el model Excel. Així, amb un nivell de confiança del 95% es pot afirmar que:
a) El nombre mitjà d'articles llegits per any i investigador oscil□la entre 129 i 142.
b) El total d'articles llegits per any oscil□la entre 70.675 i 78.025.
c) El percentatge dels articles llegits que està en anglès oscil□la entre el 47% i el 68%.

	A	B	C	D	E	F	G	H	I	J
1	**Àmbit de recerca (estrat)**	**N(i)**	**n(i)**	**x-bar(i)**	**s(i)**	**p'(i)**	**N(i) * x-bar(i)**	**N(i) * (N(i) - n(i)) * (s(i)^2 / n(i))**	**N(i) * p'(i)**	**N(i) * (N(i) - n(i)) * [p'(i) * (1 - p'(i)) / (n(i) - 1)]**
2	*E-learning*	200	20	138	30	0,50	27.600	1.620.000	100	473,68
3	Computació	250	30	103	25	0,78	25.750	1.145.833	195	325,45
4	Societat de la informació	100	25	210	50	0,21	21.000	750.000	21	51,84
5	*Totals*	*550*	*75*				*74.350*	*3.515.833*	*316*	*850,98*
6										
7				z =	1,96					
8				x(E) =	135,18		p'(E) =	0,57		
9				s(E) =	3,41		sp(E) =	0,05		
10			x(E) - z*s(E) =		**128,50**		p'(E) - z*sp(E) =	**0,47**		
11			x(E) + z*s(E) =		**141,86**		p'(E) + z*sp(E) =	**0,68**		
12				N * a =	**70.674,89**					
13				N * b =	**78.025,11**					
14										

3) La figura següent mostra els resultats obtinguts amb el model Excel. Així, amb un nivell de confiança del 95% es pot afirmar que:
a) El nombre mitjà de documents "difícils" sol□licitats per any en cada SOD oscil□la entre 129 i 142.
b) El total de documents "difícils" sol□licitats per any en el conjunt dels SOD oscil□la entre 3.635 i 5.365.
c) El percentatge d'especialistes en TIC d'entre els empleats en el conjunt dels SOD oscil□la entre 0,21 i 0,39.

F3 =(D3-E10*B3)^2

	A	B	C	D	E	F
1	**SOD**	**Nombre de professionals**	**Total de documents "difícils"**	**Nombre d'especialistes TIC**		
2	**(conglomerat)**	**N(i)**	**y(i)**	**m(i)**	**[y(i) - x(C)*N(i)]^2**	**[m(i) - p'(c)*N(i)]^2**
3	SOD-01	7	95	1	100,00	1,21
4	SOD-02	18	325	6	3025,00	0,36
5	SOD-03	15	190	6	1225,00	2,25
6	SOD-04	10	140	2	100,00	1,00
7	*Totals*	*50*	*750*	*15*	*4450,00*	*4,82*
8						
9		z =	1,96			
10		x(C) =	15,00	p'(C) =	0,3	
11		s(C) =	1,47	sp(C) =	0,05	
12		x(C) - z*s(C) =	**12,12**	p'(C) - z*sp(C) =	**0,21**	
13		x(C) + z*s(C) =	**17,88**	p'(C) + z*sp(C) =	**0,39**	
14		N*a =	**3635,18**			
15		N*b =	**5364,82**			

4) En aquest cas, cal usar un contrast d'hipòtesi per a dues poblacions dependents (ja que són els mateixos individus els que contesten al test en dos moments diferents). La sortida de R Commander mostra un p-valor = 0,2168 > 0,05 = α, amb la qual cosa no es pot rebutjar la hipòtesi nul□la que ambdues mitjanes són iguals. En altres paraules, no s'han trobat evidències suficients per a afirmar, a un nivell de confiança del 95%, que la campanya publicitària ha tingut efecte en la intenció de compra del producte per part dels consumidors.

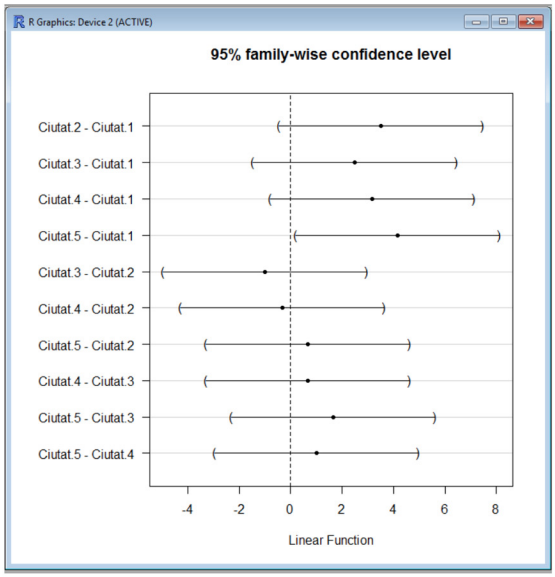
```
        Paired t-test

data:  Abans and Després
t = -1.3572, df = 7, p-value = 0.2168
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.7138923  0.4638923
sample estimates:
mean of the differences
                -0.625
```

5) En la sortida següent de R Commander es mostra un estadístic $F = 2{,}849$, el qual té un *p*-valor associat $p = 0{,}0449 > 0{,}05 = \alpha$ (ja que en aquest cas el nivell de consideràvem que era del 95%). Així, doncs, no hi ha evidències suficients per a rebutjar la hipòtesi nul□la que totes les mitjanes són iguals, ab la qual cosa, no sembla haver-hi diferències significatives entre els hàbits de compra dels residents de les diferents ciutats. S'observa que, en efecte, els intervals de confiança dos a dos toquen el 0.

```
> summary(AnovaModel.6)
            Df Sum Sq Mean Sq F value Pr(>F)
factor       4     62   15.50   2.849 0.0449 *
Residuals   25    136    5.44
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> with(StackedData, numSummary(variable, groups=factor, statistics=c("mean", "sd")))
              mean       sd data:n
Ciutat.1 1.333333 1.032796      6
Ciutat.2 4.833333 2.562551      6
Ciutat.3 3.833333 2.136976      6
Ciutat.4 4.500000 2.738613      6
Ciutat.5 5.500000 2.738613      6
```



La sortida següent mostra que la correlació entre la variable *ciutat* i la variable *dies* (ambdues generades a partir de les dades inicials) és de 0,440, valor que no sembla correspondre amb una correlació forta. En efecte, el *p*-valor de 0,015 fa pensar que, per a un nivell de confiança del 99%, ambdues variables no estan fortament correlacionades. Observeu que aquesta conclusió és bastant coherent amb la que hem obtingut anteriorment per al test ANOVA.

```
        Pearson's product-moment correlation

data:  Ciutat and Dies
t = 2.5955, df = 28, p-value = 0.01487
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.09522432 0.69101942
sample estimates:
      cor
0.4403855
```