SOCIOESTADÍSTICA

GRAU EN SOCIOLOGIA
CODI 34412

Prof. Jordi Giner Monfort
Dept. Sociologia i Antropologia Social
Universitat de València

Índex

Índex de taules

Índex de gràfiques

1. Introducció

En aquest text es desenvolupen els continguts de l'assignatura *Socioestadística* des del punt de vista de la seua fonamentació teòrica i metodològica, és a dir, no només des d'un plantejament purament teòric, sinó també pràctic. Segons la guia docent, l'assignatura s'organitza entorn de cinc grans àrees: organització i bases de la socioestadística; socioestadística descriptiva unidimensional; socioestadística descriptiva bidimensional; probabilitat i introducció a la inferència; i, finalment, el mostreig.

El primer apartat tracta de posar les bases epistemològiques i metodològiques de la recerca quantitativa, tant des de les teories sobre les quals se sustenta, com també l'organització del procés de la investigació, sense entrar en el desenvolupament de les tècniques en concret. No deixa de ser important la inclusió en aquest apartat de la necessitat de l'estadística, ja que la predisposició envers l'assignatura no sol ser positiva. En un segon subapartat del primer tema es tracten els conceptes bàsics d'estadística, cosa que té a veure amb la manera com es mesura en sociologia, la construcció de les variables i els tipus de variables.

El segon apartat desenvolupa la socioestadística descriptiva unidimensional, cosa que implica l'anàlisi variable a variable, sense tenir en compte la relació que hi puga haver entre elles. Per fer-ho, s'endinsa en l'anàlisi de freqüències, mesures de posició i de dispersió i representacions gràfiques.

En el tercer apartat s'aborda la socioestadística descriptiva bidimensional. Tracta d'introduir l'anàlisi de parells de variables, novament a partir de l'anàlisi de freqüències i de les seues representacions gràfiques, però també a partir de l'anàlisi de l'associació i la covariació, amb la introducció de proves estadístiques com khi quadrat, la correlació i la regressió.

El quart apartat introdueix la probabilitat com a base necessària per tal de dur a terme procediments inferencials, tot introduint de manera genèrica els conceptes de mostra i població i la dificultat d'arribar a resultats exactes. S'introdueixen les principals distribucions de probabilitat, amb les seues característiques pròpies i les aplicacions que tenen en sociologia. Això implica tractar de les distribucions normal tipificada, khi quadrat, t de Student, F de Fisher-Snedecor i les seues respectives taules de distribució de probabilitats.

Per acabar, l'últim apartat se centra en el mostreig. Una vegada vistos els conceptes de mostra i població en el tema 3, i la importància de la probabilitat per a la inferència en el

tema 4, es tracta d'aprofundir en allò que representa el mostreig, com calcular la dimensió de la mostra en funció de diferents criteris i per què és rellevant en sociologia. En un primer apartat se centra l'atenció en el teorema del límit central i en la seua aplicació pràctica sobre els comportaments socials, entre altres. En un segon apartat s'introdueixen les estimacions i els contrastos a partir d'exemples pràctics.

No s'ha d'oblidar que el curs de socioestadística serveix també d'introducció a l'assignatura Tècniques quantitatives d'investigació social, que s'imparteix en segon. També volem fer constar que la presentació de l'assignatura que fem tot seguit, i que coincideix amb la que presentem a l'aula, és bàsicament freqüentista en el seu vessant inferencial i, per tant, s'allunya dels plantejaments bayesians. Tampoc no entra a fons en la diferenciació entre l'estadística paramètrica i la no paramètrica, ja que no està entre els objectius d'aquesta assignatura. Tot i amb això, tant una qüestió com l'altra s'introduiran, encara que siga de manera puntual, durant el curs.

2. Fonamentació de la socioestadística

2.1. Organització i bases de la socioestadística

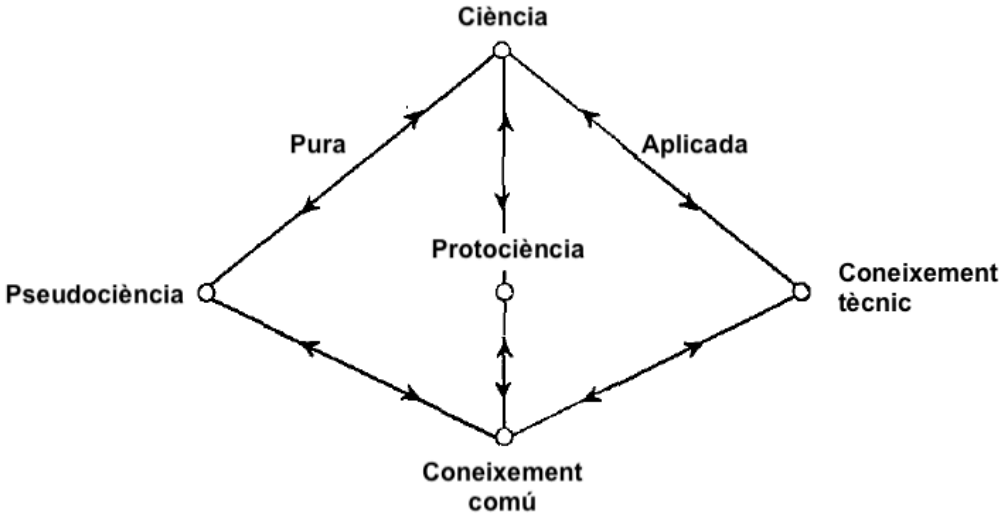
L'ànsia de l'ésser humà de conèixer el seu entorn, un coneixement ordinari d'allò que l'envolta, és possiblement tan antiga com l'ésser humà mateix. Si bé el primer pensament racional, i per tant allunyat encara que siga mínimament del coneixement ordinari, no s'originaria fins als primers pensadors hel·lènics, caldrà esperar almenys fins al segle XVI per trobar un plantejament modern del coneixement científic, el més acurat dels coneixements a què podem aspirar. Tanmateix, el coneixement del sentit comú, acrític, precientífic, es pot considerar com un lloc enmig del coneixement ordinari i el científic (Asensi-Artiga i Parra-Pujante, 2002), o fins i tot un moment previ al de l'adquisició del coneixement científic. Si alguna cosa diferencia el coneixement del sentit comú i el de la ciència, això no és l'objecte. Com explica Mario Bunge, l'objecte pot ser analitzat des del punt dels dos punts de vista, evidentment amb conclusions molt diferents en ambdós casos, però no és aquest el punt en què es fa més notable la distinció. Més aviat, allò que distingeix els dos modes de coneixement és el procediment amb què s'adrecen a l'objecte d'estudi, això és, el mètode (Bunge, 2004).

Si bé en les ciències naturals la distància entre el discurs d'allò quotidià i el discurs científic és evident, de tal manera que una persona llega tindrà problemes per interpretar-lo, això no és tant palpable en el cas de les ciències socials. És més, com assenyalen Bourdieu, Chamboredon i Passeron (2002), les ciències socials, i la sociologia en particular, s'enfronten a la familiaritat de l'univers social que estudien, cosa que s'erigeix en un obstacle epistemològic de primer ordre. Contràriament a la física, per exemple, on hi ha l'oposició laboratori-vida quotidiana, en sociologia no hi ha aquesta contraposició, com tampoc hi ha una distància equiparable entre els llenguatges que fan servir cadascuna de les dues facetes. Així, les ruptures epistemològiques en sociologia aconsegueixen distanciar el coneixement científic en l'àmbit social del coneixement del sentit comú. El coneixement del sentit comú només pot aspirar a una objectivitat limitada, molt sovint paralitzadora de la imaginació científica, que veu com allò de què s'ocupa el sentit comú trau necessitat a l'ànsia de coneixement objectiu.

Bourdieu (2001), tot parafrasejant Kant, recorda que la teoria sense investigació empírica és buida, mentre que la investigació empírica, sense recolzament teòric, resta també buida. L'afirmació no és fútil perquè suposa una crítica a la ciència feta només des de la teoria o des de les dades sense més aprofundiment, no només teòric o empíric,

sinó també ontològic i axiològic, és a dir: quin és el per què de la investigació (el requeriment explícit o la demanda implícita en paraules d'Ibáñez) i també quin és el per a què de la investigació social que es du a terme. En definitiva, com apunta Bourdieu, com més proper estiga l'estudi als pressupòsits quotidians de la ciència hegemònica, més probable serà que siga acceptat per la comunitat científica, però més difícil serà que trenque amb el coneixement del sentit comú.

El coneixement ordinari a què ens hem referit anteriorment té tres dimensions si seguim Bunge (2004). Aquestes dimensions el poden acostar més o menys al coneixement científic, però en qualsevol cas queden a mig camí perquè hi falte algun o alguns dels elements intrínsecs del coneixement científic. Així, el coneixement ordinari té tres dimensions, que el poden acostar més o menys al coneixement científic, però que queden a mig camí: el coneixement tècnic, especialitzat però no científic, propi de les arts i les habilitats professionals; la protociència, que tot i ser laboriosa i continga observació empírica o fins i tot experimentació no preveu el treball teòric; i la pseudociència, que encara que tinga l'aspecte de ciència, no coincideix amb qüestions tan importants com el plantejament, les tècniques o el cos de coneixements, és a dir, ni en l'àmbit axiològic, ni en el teòric, ni en l'empíric. Tot i això, continua Bunge, la relació de la ciència amb el coneixement ordinari és significativa. Per exemple, utilitza les habilitats artesanals pròpies del coneixement tècnic, a les quals a més pot beneficiar gràcies a la seua sistematització, és a dir, a l'aportació de l'aparell empíric. També utilitza dades en brut aconseguides per la protociència, malgrat que en absència de control n'hi ha que poden ser inútils. I fins i tot es pot establir una relació amb la pseudociència, atès que l'origen de pràcticament totes les disciplines està en enfocaments pseudocientífics. A més, la ciència mateixa, en algunes ocasions, defensa teories que cristal·litzen en dogmes i, per la incapacitat pròpia de refutar o perfilar-les, acaben esdevenint corpus de coneixements o creences de caràcter pseudocientífic. Per tant, en la pseudociència falta, entre altres coses, l'experimentació i la refutació (Bunge, 2004: p. 32 i s.).



Gràfica 1. Esquema del coneixement comú i el coneixement científic.

Font: Adaptació de Bunge (2004: p. 33).

Des del punt de vista metodològic, es poden diferenciar diferents perspectives, que tenen relació amb la ruptura epistemològica de què parlen Bourdieu, Chamboredon i Passeron relacionada alhora amb la superació del que hi ha preestablert, del coneixement del sentit comú (2002). Jesús Ibáñez en diferencia tres (Ibáñez, 2002):

- La perspectiva distributiva, que aplica la dimensió referencial del component simbòlic en connexió amb el llenguatge, que es concreta metodològicament en l'enquesta estadística, que utilitza com a dispositiu de captura de dades sobre una matriu.
- La perspectiva estructural, que se centra en la dimensió estructural del component simbòlic i es concreta metodològicament en els grups de discussió, on la xarxa de captura de dades s'expandeix fins a captar dues dimensions: d'individu a grup i de pregunta/resposta a la conversa.
- I finalment, la perspectiva dialèctica, que se centra en el component semiòtic i en el pla metodològic i que es concreta metodològicament en la socioanàlisi, on l'expansió en termes de captura és estratègica i implica una major llibertat que multiplica les possibilitats de la xarxa.

Aquests exercicis classificatoris han estat reproduïts per uns altres autors del nostre entorn, com ara Alfonso Ortí, per a qui els nivells d'aprehensió de la realitat social també són tres i tenen diferents conseqüències en l'àmbit del que és constituent, conscient, analític, epistemològic i metodològic (Ortí, 1995). Així, a un primer nivell dels fets, allò que és manifest en termes freudians, en segueix un segon, el dels discursos, allò que

és latent o preconscient, i encara un tercer, el de les motivacions, allò que Ortí classifica com a propi de les pulsions. Cadascun d'aquests tres nivells d'aprehensió desenvolupa unes unitats bàsiques dels processos d'anàlisi social. Així, el nivell del que és manifest s'associa amb el registre de dades, les explicacions causals i, en definitiva, el model estadístic. Enfront d'aquest acostament a la realitat, els nivells latent i profund s'encarreguen dels textos, símptomes i símbols, i per tant es relaciona amb models lingüístics i heurístics. Amb aquesta aproximació a la interpretació de la realitat, i el popular esquema a què dona lloc l'obra d'Ortí (1995: p. 93), s'enfoquen de manera genèrica les tres aproximacions empíriques, metodològiques i ontològiques en investigació social. En aquest cas, deixem de banda les aproximacions de caire més qualitatiu i ens centrem en la que s'encarrega d'allò manifest que és, en definitiva, la quantitativa.

Les ruptures epistemològiques, o obstacles si seguim la nomenclatura de Bachelard (2002/1938), són també les que fan avançar el coneixement científic. Per exemple, el primer obstacle epistemològic a vèncer és el de l'experiència directa, allò que constituiria l'estadi precientífic en el desenvolupament de la ciència segons el seu punt de vista i que arribaria pràcticament fins al segle XVIII. És aproximadament en aquest moment que comença l'estadi científic, que arriba fins al segle XX, que és quan comença l'època de la nova mentalitat científica que, segons l'autor francès, queda inaugurada amb l'obra d'Einstein el 1905, que va complir el paper de desconstruir conceptes que fins aleshores es consideraven fixos (Bachelard, 2002/1938: p. 18). Així, si apliquem aquest esquema a la sociologia, es podria dir que la primera pedra per a l'estadi científic arriba tard, aproximadament a la darreria del segle XIX, fins i tot es podria dir que té en el moment en què Durkheim publica les normes del mètode sociològic (Durkheim, 1895). El segon moment, el de la nova mentalitat científica o maduresa en el camp, es podria situar a mitjan segle XX, quan, en paraules de Gigerenzer i Murray, s'arriba a la revolució inferencial i tot el que aquesta suposa per a l'avanç de les tècniques quantitatives (Gigerenzer i Murray, 1987: p. 6).

Tot i amb això, la interpretació del positivisme i per extensió dels mètodes quantitatius com un paral·lel de les ciències exactes en l'àmbit de les ciències socials no deixa de ser una comparació falsa per a alguns autors. Allò que al segle XIX pretenia ser la física social, diuen Bourdieu, Chamboredon i Passeron, no era sinó un espill -objectivisme provisori- en què pretenien imitar mecànicament la lògica objectivista de les ciències naturals, sense tenir en compte que les ciències socials -o les ciències humanes, com diuen- necessiten una autonomia pròpia (Bourdieu, Chamboredon i Passeron, 2002: p. 19).

2.1.1. Mètode i fases d'una investigació empírica quantitativa

Sembla que qualsevol material docent sobre tècniques d'investigació social haja de començar per tractar els diferents paradigmes. I potser tinguen raó aquelles persones que històricament ho han organitzat d'aquesta manera. El concepte de paradigma, com l'entenem avui dia, és originari de la interpretació que en va fer Thomas Kuhn en la seua obra sobre les revolucions científiques (1962). Per a Kuhn, el paradigma és una estructura conceptual de caràcter complex que guia la comunitat científica a l'hora d'acostar-se a la realitat que volen estudiar, tant pel que fa a l'elecció del problema a estudiar, les passes a seguir i els instruments que hauran d'utilitzar. És, per tant, una mena de guia per a la ciència. Per a Kuhn, hi ha una ciència normal que és la predominant en una determinada disciplina científica i que, al mateix temps, és acceptada per tota la comunitat científica. Durant el desenvolupament d'aquesta fase *normal* hi ha avanços significatius en el coneixement científic, com caldria esperar, fins que en un moment donat hi ha una revolució per la qual se substitueix el paradigma dominant per un altre de nou. Mentre que això ha sigut històricament visible en les ciències naturals, en les ciències socials no hi ha hagut un paradigma que haja sigut amplament compartit per la comunitat científica, cosa que fa que segons Kuhn encara avui dia es puguen considerar en un estat preparadigmàtic o multiparadigmàtic (Corbetta, 2007: p. 6).

No obstant això, sí que hi ha un consens elevat sobre els paradigmes o perspectives fonamentals en la teoria sociològica: el positivisme i l'interpretativisme[1]. Cadascun d'aquests desenvolupa maneres diferents d'interpretar la realitat social i també mètodes diferenciats en la manera de conèixer-la, cosa que té efectes també en les tècniques d'investigació i en les diferents anàlisis que despleguen cadascun. En definitiva, es tracta de respondre a les qüestions ontològiques (què és allò que volem conèixer i de quina manera es manifesta), epistemològiques (quina relació tenim amb allò que volem conèixer) i metodològiques (com podem observar allò que volem conèixer). De la relació entre aquestes tres qüestions naixen les dues perspectives a què hem fet referència anteriorment i que guien els subapartats següents.

[1] No tota la comunitat científica, però, està d'acord en la divisió binària. De fet, autors com Manuel García Ferrando (1974) aposten per una divisió en tres: a més dels paradigmes positivista (Durkheim) i interpretativista (Weber), caldria tenir en compte un tercer paradigma, l'estructuralista (Marx).

El positivisme

El positivisme com a tal és, en la pràctica, el primer paradigma de què disposa la sociologia en el moment del seu naixement. Apareix al segle XIX i és producte de la fe de les primeres persones que van tractar la sociologia en el paradigma imperant en les ciències naturals, és a dir, en l'estudi de la realitat social mitjançant el marc conceptual, les tècniques d'investigació i els instruments d'anàlisi que li són pròpies (Corbetta, 2007). Les aportacions, primerament, de Henry de Saint-Simon (1803) i després del seu deixeble Auguste Comte (1848) anaven en el sentit de crear una física social a imatge de l'evolució de les ciències naturals. Per bé que abans de les formulacions de Comte ja hi havia hagut intents d'utilitzar eines metodològiques pròpies del positivisme, desproveïdes de qualsevol fonamentació teòrica[2]. Aquesta perspectiva és la pròpia dels garants de l'ordre i de la llei, que en definitiva era la guia del positivisme, com revelen alguns fets, com ara la inscripció de la bandera de Brasil, inspirada en una cita de Comte: "*L'amour pour principe et l'ordre pour base; le progrès pour but"*, que no era sinó el seu lema que apareixia de manera sistemàtica, entre altres, en la portada del seu *Système de Politique Positive* (Comte, 1851-1854).

Al marge d'aquells experiments purament empírics, el primer que va aplicar la perspectiva teòrica a la investigació empírica fou Émile Durkheim, qui en la pràctica va traslladar la perspectiva teòrica positivista a la praxi empírica en el seu manual fundacional sobre el mètode sociològic (1895). Ja en el prefaci de l'edició original llança la que serà la base del seu mètode: considerar els fets socials com si foren coses, de la mateixa manera que ocorre amb les coses en l'àmbit de les ciències naturals. Els fets socials, aleshores, no estarien subjectes a la voluntat de l'ésser humà i tindrien les seues pròpies normes, que són les que cal descobrir de manera objectiva mitjançant la investigació científica. És important incorporar, arribats en aquest punt, la perspectiva de gènere que en el positivisme clàssic estaria representada, com a mínim, per les figures de Harriet Martineau i Florence Nightingale, especialment la darrera i la seua aportació a l'estadística moderna (McDonald, 1994).

El positivisme actua, seguint Corbetta (2007), mitjançant procediments inductius, és a dir, passant d'allò particular a allò universal. El procediment inductiu assumeix la presència d'un ordre i unes normes organitzadores de caràcter universal, cosa que ha de

[2] Vegeu, en aquest sentit, les protoenquestes de Jean-Baptiste Colbet al segle XVII sobre qüestions d'ordenació administrativa o de les formes de vida de la població francesa; o les protoenquestes d'Anglaterra comandades per Gregory King durant la mateixa època (Echegaray, 2018).

servir de guia a la comunitat científica. El plantejament original del positivisme és tan ingenu que al segle XX es desenvolupa una evolució que se sol conèixer com neopositivisme (aquell corrent sorgit entre els anys 30 i 60 del segle XX) i postpositivisme (el que imperarà a partir dels anys 60). Paul Lazarsfeld, un dels principals representants del neopositivisme, de qui parlarem més endavant, era partidari d'analitzar la realitat social mitjançant el que ell anomenava el llenguatge de les variables: tot era reduïble a variables i aquestes, com a elements neutrals, objectius i operatius en el nivell matemàtic, esdevenien el centre de l'anàlisi social. Es podria dir que l'objectivació de l'objecte d'estudi arribava així al seu màxim nivell.

També és fruit del positivisme modern la incorporació de la categoria de refutabilitat, no des del punt de vista de la comprovació de les hipòtesis mitjançant l'observació de les dades empíriques, sinó per la via de la comprovació de la no invalidació de la hipòtesi inicial. Això també posarà la base de la provisionalitat de les hipòtesis, sotmeses ara a processos constants de comprovació, i deixarà de banda l'ideal de la ciència com a certesa, amb el que això significa d'abandó de la ingenuïtat primigènia d'aquest paradigma.

En la pràctica, el positivisme es tradueix majoritàriament en l'aplicació de tècniques quantitatives. Entre les característiques d'aquest enfocament, en podem trobar les següents (Francés et al., 2014: p. 56):

- Dona privilegi a la lògica hipoteticodeductiva per a la construcció del coneixement.
- El seu objectiu és la mesura objectiva de variables que, una vegada creuades, seran claus per al funcionament de la realitat social.
- L'articulació del procés investigador gira entorn del maneig de conceptes operatius i unívocs.
- L'estructura del procés investigador està altament formalitzada, amb una seqüència lògica i lineal de fases de la investigació.
- El paper de la persona investigadora és extern a la realitat, amb una voluntat de neutralitat i asèpsia que controle l'entrada de biaixos analítics.

L'interpretativisme

Si el positivisme es basa en la mirada ingènuament cientifista de Comte i Durkheim, l'interpretativisme beu de la sociologia comprensiva de Max Weber, el plantejament bàsic de la qual és que la realitat no pot ser observada directament, sinó que ha de ser interpretada. De fet, l'origen a la crítica cientifista s'ha de buscar en Wilhelm Dilthey (1883), que al segle XIX ja diferenciava entre les ciències de la natura i les ciències de l'esperit. Si en les ciències de la natura l'objecte és exterior a la persona que l'investiga, en les ciències de l'esperit no hi ha separació entre observador i la realitat estudiada, per la qual cosa és necessari un procés diferent al de l'explicació, com és ara la comprensió. Weber bevia del plantejament de Wilthey, però alhora pretenia allunyar la ciència social de plantejaments valoratius, és a dir, que no renunciava a l'objectivitat, ni tampoc a la generalització. De manera molt resumida es pot dir que l'orientació comprensiva que proposa es dirigeix a entendre l'objectiu de l'acció humana (Corbetta, 2007: p. 21). Weber resol l'aparent incompatibilitat d'una anàlisi individualista amb una orientació objectivista mitjançant els tipus ideals, les formes d'acció social que es troben de manera recurrent en la conducta dels individus i en els quals, en certa manera, es pot reconèixer una certa uniformitat. Com a models teòrics, els tipus ideals ajuden a interpretar la realitat, sense el rang de lleis amb què opera el positivisme, sinó més avait a partir de connexions causals.

Els plantejaments interpretativistes posteriors, representats bàsicament per l'interaccionisme simbòlic, la fenomenologia i l'etnometodologia, no són sinó replantejaments, en termes metodològics, de la teoria weberiana. Això sí, són replantejaments que es fan des d'òptiques microestructurals, enfront del pla macrosociològic que guiava la teoria weberiana, i ja no s'interessen per qüestions com l'economia, el poder o la religió, sinó pel món de la vida quotidiana, que les ciències socials havien deixat de banda fins aquell moment.

L'interpretativisme, a diferència del positivisme, desenvolupa majoritàriament tècniques qualitatives, que solen presentar les característiques següents (Francés et al., 2014: p. 57):

- Predomina la lògica de coneixement inductiva sobre la deductiva, de manera que no hi ha una formulació prèvia d'hipòtesis a contrastar. Més aviat es podria dir que es troben hipòtesis generades en el mateix procés d'investigació.
- Assumeix una perspectiva de l'anàlisi social holística, en què el fenomen s'interpreta com un tot que no és possible descompondre en variables mesurables, sinó que s'ha d'entendre en la seua complexitat.

- Resulta un enfocament flexible pel que fa a l'ordre intern del procés d'investigació, com també en les normes de procediment.
- Els conceptes que utilitza són orientatius i interpretables, amb la qual cosa no cal una definició operativa que separe de manera estricta uns conceptes dels altres.
- Es dona privilegi a la interpretació de la realitat sobre el seu mesurament. Amb l'anàlisi no es busca la generalització, sinó la troballa d'elements significatius i explicatius.
- La persona investigadora assumeix un paper interactiu amb els subjectes investigats, tant per les tècniques utilitzades, com també per la pròpia posició davant el procés de construcció del coneixement.
- La construcció del coneixement s'orienta a partir de la intersubjectivitat compartida dels subjectes, i no tant mitjançant la recerca d'elements objectius.

El pluralisme metodològic

Resulta evident que, com han sigut presentats, positivisme i interpretativisme tenen associades lògiques d'investigació oposades. El positivisme desenvolupa un corpus de tècniques al voltant de l'anàlisi estadística de dades, mentre que l'interpretativisme se centra en l'anàlisi hermenèutica dels discursos. És, en definitiva, l'històric debat dels nombres contra les lletres. De fet, la defensa fèrria dels plantejaments propis de cadascun dels paradigmes ha estat una constant que encara avui dia s'escolta a les aules i als passadissos de les facultats, no només de ciències socials. No obstant això, cada vegada és més habitual trobar investigacions empíriques en les quals s'utilitzen dues o més tècniques, tant del mateix mètode (qualitatiu o quantitatiu) com de diferents mètodes (és a dir, quantitatiu i qualitatiu). A aquesta utilització simultània de diferents mètodes se l'ha anomenat pluralisme metodològic, triangulació metodològica, articulació metodològica o mètodes mixtos. En definitiva, com argumenta Byrman, més que tractar-se de dues maneres d'investigar allunyades per plantejaments epistemològics, es tracta de tècniques diferents, tot i que aquestes tècniques comporten maneres d'entendre la realitat social oposades (Byrman, 2004: p. 32).

Deixant de banda l'estudi de Charles Booth sobre la vida i el treball de la població londinenca a la fi del segle XIX i els intents de l'escola de Chicago d'utilitzar mètodes his-

tòrics i estadístics, el punt inicial de la triangulació metodològica es considera la publicació de l'article de Campbell i Fiske en el *Psychological Bulletin* (1959). Per a Norman Denzin, la triangulació en investigació social es pot classificar en quatre tipus diferents (1970):

- Triangulació de dades: la modalitat més habitual, ja que fa servir diferents fonts de dades sobre un mateix objecte de coneixement.
- Triangulació d'investigadors: equivalent a considerar equips interdisciplinaris.
- Triangulació teòrica: que fa ús de diferents perspectives teòriques sobre un mateix problema d'investigació.
- Triangulació metodològica: inclou la triangulació intramètode (dins el mateix cos de mètode) i intermètodes (combinació de diferents mètodes). L'últim model, l'intermètodes, s'estima el més útil perquè és la base de la validació creuada, és a dir, arribar als mateixos resultats amb mètodes diferents.

En la investigació empírica quotidiana, la utilització de diferents mètodes per tal de dur a terme una investigació planteja algunes problemàtiques. En primer lloc, es pot donar el cas que en el plantejament metodològic hi haja un mètode que siga central, mentre que l'altre hi és secundari o té un ús instrumental (per exemple, una enquesta per detectar discursos que explorar en grups de discussió, o a l'inrevés, grups de discussió a fi de treballar conceptes a mesurar en una enquesta). En segon lloc, s'ha de decidir ordre en què s'aplicaran les tècniques, cosa que pot estar relacionada amb el punt anterior. En definitiva, a les qüestions sobre què fa cada tècnica i quina importància té en la recerca, s'hauria d'afegir també en quin moment s'apliquen dins la triangulació.

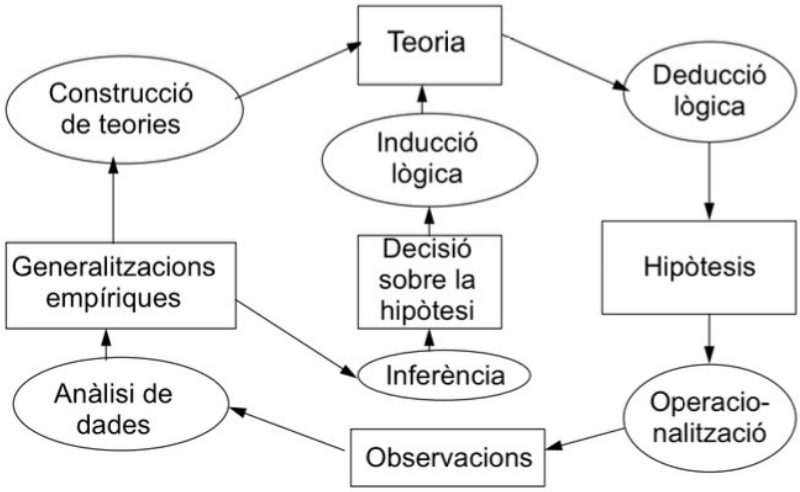
D'altra banda, hi ha una branca de la metodologia actual que tendeix a considerar que mètodes quantitatius i mètodes qualitatius no s'han d'entendre com a opcions oposades dicotòmiques, sinó més aviat com els extrems d'un *contínuum* en què es pot situar qualsevol estudi (Creswell, 2014).

2.1.2. El procés d'investigació

De la informació anterior es pot deduir que el paradigma imperant en tècniques quantitatives és el positivista o, dit altrament, que quan s'apliquen tècniques quantitatives al coneixement d'un fenomen social se sol fer des d'una òptica positivista. Aquesta òptica

positivista implica seguir l'organització de la recerca segons el marc científic comunament acceptat. Per a Merton (1965), això significa que la ciència siga pública i no privada, cosa que implica dues condicions, de les quals la primera és el control. Si es fan públics els procediments utilitzats, la comunitat científica pot comprovar com ha sigut el procés de recerca i, en últim terme, validar-ne els resultats. Però, que la ciència siga pública implica acumulació, ja que no s'han d'ignorar els coneixements previs a què ha arribat la comunitat científica[3]. Això posa les bases de dos dels apartats més importants de qualsevol investigació, juntament amb el resultats: l'apartat metodològic i les referències bibliogràfiques.

El procés d'investigació, però, comença molt abans. L'esquema de Walter Wallace[4] a partir de quatre processos cognitius en forma circular potser siga un dels que més èxit ha tingut a l'hora de representar el procés d'investigació (Wallace, 1971). Així, la investigació científica enceta un recorregut circular que en la seua versió més senzilla comença amb la revisió teòrica, passa per la via de la deducció a les hipòtesis, mitjançant l'operacionalització arriba a l'observació empírica, i finalment acaba amb l'anàlisi i interpretació i la generalització empírica, que mitjançant la inducció genera una nova teoria. Tot seguit analitzem cada pas de manera un poc més pausada.



[3] Així naix el concepte que utilitza, entre altres, Google Scholar: "a coll de gegants". És una citació recurrent des de l'antiguitat, però Merton la va adoptar per al títol del seu llibre, on glossa aquestes dues característiques fonamentals de la ciència (1965).

[4] De vegades és convenient parar-se a explicar la vida d'algunes persones que considerem clàssics de la sociologia. Així, si casos com el de Merton serien l'exemple d'una vida atzarosa i rocambolesca (vegeu l'obituari en Kaufman, 2003), Wallace ha de ser l'exemple de la condició de negritud en l'acadèmia nord-americana. De fet, va ser el director de la *senior thesis* (ací en diguem TFG) de Michelle Robinson -després Obama- a la *University of Princeton*, allà pel 1989, que precisament versava sobre la qüestió de l'estudiantat negre de la universitat de Princeton (vegeu Hotchkiss, 2005).

Gràfica 2. El cicle continu de la ciència, o cercle de Wallace.

Font: Adaptació de Wallace (1971).

La teoria

La primera fase en el cercle de Wallace és la de la teoria. Qualsevol recerca, com hem vist que argumentava Merton, ha d'estar basada en els coneixements previs de la comunitat científica. Per això, cal iniciar una fase prospectiva sobre el tema en qüestió. Aquesta recerca prèvia ha de servir també per contrastar les afirmacions, en forma d'hipòtesi, que caldrà verificar o refutar, a les quals hem vist que s'arriba mitjançant un procés deductiu. La teoria es pot interpretar com un grup de proposicions lògicament interconnectades de les quals es poden deduir uniformitats empíriques (Merton, 1949). La fase teòrica s'ha enriquit en els darrers anys per la possibilitat d'accedir a milers de documents en línia indexats electrònicament, cosa que l'allunya de la lentitud en què vivia quan s'havia de fer de manera analògica. Centenars de fonts electròniques de caràcter acadèmic, científic i social serveixen ara de prestatgeries en línia, moltes de les quals en accés obert, que faciliten la feina a les persones investigadores, i certifiquen la sentència de Merton sobre l'avanç científic a col de gegants. Per a Cea (2012), la teoria pot ajudar a decidir els esquemes classificatoris a utilitzar; plantejar els conceptes teòrics que orienten l'anàlisi; formular problemes d'investigació de rellevància social; concretar idees generals sobre la producció del canvi social; i també formular hipòtesis.

D'altra banda, cal diferenciar entre els tipus de teoria que podem trobar: en primer lloc, hi ha les grans teories o teories globals, que ofereixen explicacions abstractes, molt sovint sense suport empíric, sobre la realitat social. Alguns aspectes d'aquestes teories poden ser útils a l'hora d'interpretar els resultats del procés d'investigació. Cea cita alguns exemples de teories generals, com ara la teoria marxista del desenvolupament social i la teoria del sistema social de Parsons (Cea, 2012: p. 39). En segon lloc, hi ha les teories de rang mitjà, aquelles que se centren en aspectes concrets de la realitat social i en variables que poden ser mesurades empíricament. Són molt més fàcils de ser analitzades mitjançant la verificació o refutació d'hipòtesis, i seguint Cea, n'és un exemple la teoria del suïcidi de Durkheim.

La hipòtesi

La hipòtesi o hipòtesis d'investigació sorgeixen del procés deductiu a què sotmetem la teoria. Una hipòtesi és una formulació parcial de la teoria, i se situa en un nivell inferior en termes de generalitat, àmbit geogràfic o lapse de temps. La hipòtesi s'ha de plantejar de manera positiva, com a formulació de relació entre dues o més variables, i ha de ser

comprovable empíricament. Normalment treballem amb només una hipòtesi, perquè en funció de l'operacionalització dels conceptes implicats poden desencadenar un procés d'investigació suficientment llarg com per adoptar-ne més.

Mitjançant el procés d'observació empírica -o específica en el cas de Wallace-, es procedeix a comprovar la hipòtesi. Per exemple, en el cas de la teoria del suïcidi de Durkheim, es pot comprovar que un major grau d'individualisme comporta un major nivell de suïcidis. O que els individus que són casat i tenen fills se suïciden menys que les persones que no compleixen aquestes condicions. Seguint Corbetta, el criteri de la verificació empírica és el criteri del caràcter científic, i si una hipòtesi pot ser comprovada, passarà al corpus teòric de la sociologia. Tot i això, cal tenir en compte, com apunta el sociòleg italià, que en ciències socials es corre un risc especialment elevat de formular teories vagues o confuses que siguen de difícil operacionalització (2007: p. 73).

L'observació empírica

Després de formular la hipòtesi es posa en marxa el procés d'operacionalització, és a dir, la transformació de les proposicions implicades en la hipòtesi en conceptes mesurables a partir dels quals llançar les proves per verificar-la o refutar-la. La fase d'observació és també la fase en què es despleguen les tècniques d'investigació, i per tant és la fase més tècnica. També sol ser una de les parts de la investigació que més temps, energia i recursos consumeix, atès que, especialment en les tècniques quantitatives, sol demanar una inversió gran en termes de personal, si es tracta d'una enquesta.

L'operacionalització de la hipòtesi es desenvolupa, una vegada més, amb l'ajuda del corpus teòric que ha servit de base a la recerca. No es poden ignorar, tot seguint Merton, les aportacions acadèmiques i científiques anteriors, i això se sol traduir a l'adopció de mesures i escales que ja han sigut provades amb èxit, tot i que també a la millora d'eines existents o a l'avaluació de noves maneres d'acostar-se al concepte. Val a dir que el procés d'operacionalització, en funció de l'amplitud del concepte a mesurar, pot tornar-se molt complex perquè tractarà d'incorporar diferents vessants a la mesura. Aquest és el cas, per exemple, de l'operacionalització del concepte *qualitat de vida* que exposa Cea d'Ancona i que acaba representant 11 àrees diferents de diferents àmbits de les necessitats humanes, que es descomponen en 39 dimensions i, finalment, en 251 indicadors (2012: p. 77).

La definició dels indicadors a mesurar està en funció de la tècnica que s'utilitze, i també dels recursos disponibles, cosa sobre la qual tractarem quan parlem del projecte d'investigació. Així, és habitual que si els recursos són limitats, com és el cas de moltes recerques actualment, els indicadors que s'utilitzen siguen de fonts secundàries. Crear una eina per mesurar una sèrie d'indicadors com els que argumentava anteriorment Cea d'Ancona, i aplicar-la sobre una mostra prou representativa de la població, exigeix a l'equip d'investigadors una inversió en equips de recerca, treball de camp i codificació considerable de què, ara per ara, no es disposa a la universitat. Certament, els processos d'enquestació s'han abaratit molt els darrers anys, ja que s'hi han introduït tècniques que així ho han fet possible: la generació automàtica de números de telèfon, per exemple, ha fet que no hàgem de dependre de marcs mostrals que s'havien de comprar o aconseguir de la manera més fiable possible. El mateix cost de les telefonades no és el mateix ara que fa vint o trenta anys. Per no parlar de les enquestes autoadministrades amb l'ús de plantilles d'internet. Igualment, el procés de tabulació i codificació és ara molt més ràpid gràcies a l'ús de tauletes electròniques o les plantilles d'internet, on el procés de codificació es fa a priori i cada persona que emplena l'enquesta en realitat fa l'esforç de codificació que anteriorment havia de realitzar l'equip d'investigació.

La generalització empírica

El pas previ a les generalitzacions empíriques és l'anàlisi. Qualsevol procés de recerca ha de passar per una fase d'anàlisi. Abans hem diferenciat entre l'anàlisi hermenèutica, la pròpia de les tècniques qualitatives, i l'anàlisi estadística, d'ús en les tècniques quantitatives. Això es tradueix també en diferents tipus d'anàlisi: d'una banda, l'anàlisi per casos o per temes, propi de les tècniques quantitatives; d'altra banda, l'anàlisi per variables, com és el cas de les tècniques quantitatives. En el cas de les tècniques quantitatives, hem de diferenciar entre l'anàlisi descriptiva (univariant o bivariant), que és de caràcter deductiu, i l'anàlisi inferencial, de caràcter inductiu. Ambdós tipus d'anàlisi poden conduir a la verificació o refutació de la hipòtesi inicial, però només l'anàlisi inferencial ofereix una mesura en termes del que adés s'ha expressat com a refutabilitat, expressada en significativitat a partir de diferents distribucions (normal tipificada, khi quadrat, t de Student i F de Fisher-Snedecor). En tot cas, l'anàlisi també està relacionada amb el tipus de disseny que es presente, els recursos disponibles i la tècnica d'investigació aplicada.

La generalització empírica és el procés que segueix l'observació i que, mitjançant l'anàlisi, tanca el cercle de la investigació tal com el planteja Wallace. Val a dir que la generalització només seria aplicable, *stricto sensu*, en aquelles recerques en què s'ha dut a terme un procés inferencial, i per tant representatiu de la població, o en aquelles que s'ha treballat sobre les dades de l'univers. Les recerques exploratòries es quedarien en un àmbit pseudogeneralitzable que hauria de ser contrastat mitjançant noves recerques en altres espais, moments, grups socials, etcètera. També cal apuntar que el procés de generalització empírica ha de partir de la hipòtesi o hipòtesis inicials, per tal de verificar-les o refutar-les i, d'aquesta manera, fixar el coneixement científic sobre el tema en qüestió. Això només es pot dur a terme, com apuntava anteriorment Merton, mitjançant un plantejament públic de la ciència, és a dir, fent difusió dels resultats pels mitjans que siguen necessaris perquè la comunitat acadèmica i científica en siga sabedora. No entrarem a debatre el negoci que s'ha creat els darrers anys al voltant del mercat de les publicacions, les grans corporacions editorials que ho fan possible i les institucions acadèmiques i administratives que ho afavoreixen. Tanmateix, sí que hi ha dos factors que han fet que cada vegada siga més fàcil accedir de primera mà als treballs dels equips d'investigació, siga quina siga la seua especialitat. En primer lloc, com hem vist anteriorment, l'accés als documents en format electrònic, cosa que facilita la consulta i agilita la citació de treballs, encara que estiguen allunyats geogràficament o escrits en altres llengües diferents de la de l'equip d'investigació. En segon lloc, el compromís d'algunes administracions, com ara les del Regne Unit, de publicar en obert totes aquelles recerques que han estat finançades amb fons públics[5]. Aquest és un pas que, encara avui dia, no han pres totes les administracions, i això continua afavorint un mercat de les publicacions privat i tancat a una part de la comunitat científica que hi pot no tenir accés (per exemple, a països de l'hemisferi sud, on els convenis amb els grans grups editorials poden ser econòmicament inassolibles).

La generalització empírica comporta la creació d'una nova teoria, l'aprofundiment en una que ja hi és o la falsació d'una teoria prèvia. Com a tal, i tancant ara sí el cercle de Wallace, es podrà constituir en la base d'una nova recerca que en faça ús, també en aquest triple sentit de creació-aprofundiment-falsació.

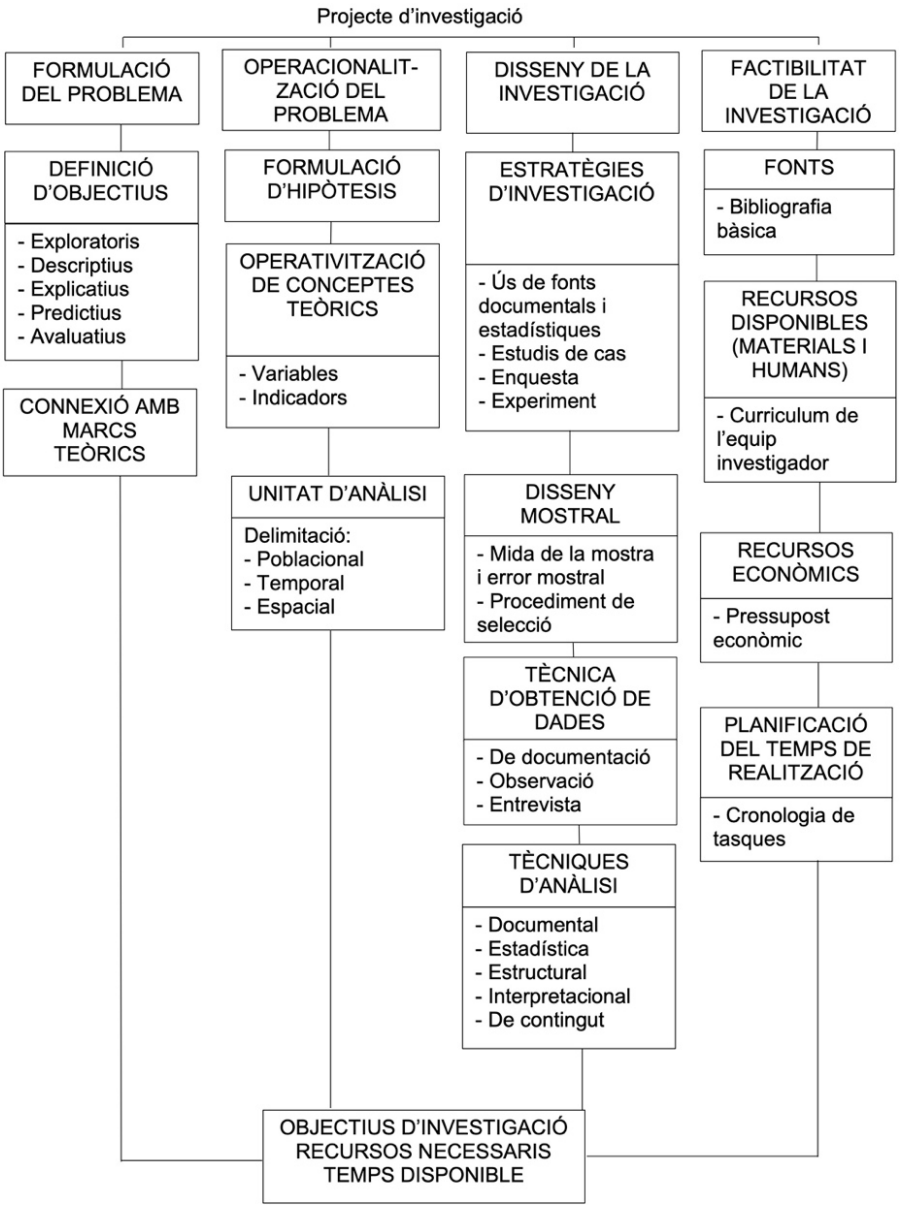
2.1.3. El projecte d'investigació

[5] Vegeu la iniciativa del govern britànic de 2012 Public access to publicly-funded research (https://www.gov.uk/government/speeches/public-access-to-publicly-funded-research) o també la recentment creada *Coalition S* en la qual, per cert, no participa cap agència espanyola.

Tot el procés de recerca exposat en els punts anteriors no seria possible si no hi haguera una motivació per dur-lo endavant. Aquesta motivació pot tenir diferents orígens: un treball acadèmic (d'alguna assignatura, de final de grau, màster o tesi doctoral); un concurs, subvenció o premi; la convocatòria d'una beca o plaça amb perfil d'investigació; el fet de pertànyer a un organisme acadèmic o d'investigació o a una empresa dedicada a la investigació social; o simplement una curiositat personal que ens empeny a dur-lo a terme. Totes les motivacions tenen en comú que han de passar un filtre que és el que n'assegura la qualitat, en termes de concurrència competitiva la major part de les vegades, i hi assigna els recursos econòmics o n'aprova la realització. Evidentment, cadascun dels orígens de la motivació donarà lloc a diferents tipus de recerca, en funció entre altres coses dels recursos disponibles. Les recerques, diguérem, canòniques es caracteritzen per tenir una sèrie de punts en comú, dins d'allò que hom anomena projecte d'investigació.

El projecte d'investigació, com avisa Cea d'Ancona, és el pas del procés d'investigació ideal a la pràctica habitual (2012: p. 43). És on hi ha compresos els objectius, el marc teòric i també els recursos econòmics, humans i la seua temporalització. El projecte d'investigació és també on ha de figurar el disseny de la investigació: des del pla de la investigació, les tècniques que es faran servir per recollir les dades, les operacions d'anàlisi previstes, els objectius que cal acomplir i fins i tot el pla de publicacions previst. Segons Cea d'Ancona, a més del disseny de la investigació, tot projecte ha de incloure tres components essencials: la formulació del problema d'investigació, l'operacionalització del problema d'investigació i la factibilitat de la investigació en termes de recursos humans, econòmics i temporals. L'estructura es pot consultar, en tot cas, en la gràfica que presentem a continuació.

Pel que fa a la formulació del problema d'investigació, cal esmentar els objectius generals i específics de la investigació, les raons que expliquen l'elecció del problema central de l'estudi i justificar la rellevància del tema. En funció de quina siga la transcendència del tema escollit, serà més o menys fàcil obtenir el finançament que en moltes ocasions cal per poder dur a terme la recerca. D'altra banda, s'ha d'aclarir quin plantejament es dona a la investigació, que pot anar d'allò més vague i exploratori, a allò més precís, com ara les recerques de tipus predictiu o avaluatiu. La formulació del problema, en tot cas, ha d'estar basada en la literatura científica existent fins al moment, de la qual hauria d'eixir un marc teòric encara que siga preliminar.



Gràfica 3. El projecte d'investigació

Font: Adaptat de Cea d'Ancona (2012: p. 44)

En segon lloc, i una vegada delimitats els objectius de l'estudi, cal acotar-lo des del punt de vista conceptual, poblacional i temporal. En aquest sentit, la hipòtesi o hipòtesis plantejades són clau perquè se situen en el marc determinat i en relació amb els objectius formulats que, de manera habitual, s'han treballat, ni que siga de manera exploratòria, anteriorment. Dins el projecte d'investigació cal donar a conèixer l'operacionalització

d'allò que es vol analitzar. També és important donar a conèixer la unitat d'anàlisi i com es contactarà amb cadascun dels individus de la mostra o d'on s'espera obtenir els resultats.

En tercer lloc, s'ha de fer esment al disseny de la investigació, és a dir, l'estratègia que se seguirà per assolir els objectius de la investigació. En funció de l'estratègia que triem, s'han de desenvolupar els punts següents: el disseny mostral, que inclou tant la mida de la mostra com el procediment de selecció d'aspectes que, en tot cas, desenvolupem posteriorment. Un altre aspecte que cal destacar són les tècniques de recollida de la informació que es faran servir, és a dir, les tècniques que s'aplicaran sobre la mostra escollida (anàlisi de fonts secundàries, enquesta a partir de qüestionari, panel d'experts, etcètera). D'altra banda, també cal fer menció de les tècniques d'anàlisi previstes i de l'estratègia en termes de temporalitat (si és en un moment donat, si s'estudia l'evolució del fenomen, etcètera).

En quart lloc, cal esmentar la factibilitat de la investigació, apartat que inclou les condicions mínimes amb què ha de comptar la recerca per tal que siga viable. Això inclou aspectes immaterials com ara les característiques de l'equip d'investigació: quantitat de persones, temps de dedicació, currículum de l'equip investigador i de la persona que n'estarà al càrrec, possibilitat de contractar més personal o externalitzar serveis, etcètera. També incorpora aspectes materials, com ara l'accés a la bibliografia especialitzada, la disposició de material informàtic i programes adequats per a l'anàlisi i la redacció de resultats, però també allò que en l'àmbit administratiu es coneix com a material fungible. D'altra banda, cal incorporar informació sobre la temporalització del projecte d'investigació, on es puga observar en quin moment es treballarà en cada fase del procés d'investigació. Per dur-ho a terme és molt útil l'adopció d'un diagrama de Gantt en què les columnes són setmanes o mesos -en funció de com siga de llarg el projecte- i les files les tasques o microtasques a desenvolupar.

D'altra banda, cal tenir en compte que, en funció de la convocatòria i l'organisme responsable, hi poden haver altres requisits, documents o proves que aportar. Per exemple, al Regne Unit qualsevol investigació, siga del departament que siga, ha de disposar d'un informe favorable d'un comitè ètic que, en el cas de la major part de les universitats espanyoles, només és necessari quan es tracta d'investigació biomèdica o de tipus psicològic.

Amb tota aquesta informació també s'ha de valorar el cost econòmic del projecte d'investigació, una informació que no és fàcil de calcular i que, en el cas de l'administració planteja alguns reptes, com ara que les subvencions i ajudes solen retallar el càlcul inicial però no les tasques o els objectius plantejats; i també que les empreses solen guardar el zel professional a l'hora de fer públics els preus de les tècniques que duen a terme[6].

2.1.4. El disseny de la investigació

En funció dels objectius de la investigació, del marc temporal i del context en què es preveu fer la investigació. Atesos els objectius, el disseny de la investigació pot ser de tipus exploratori, descriptiu, explicatiu, predictiu o avaluatiu.

Les recerques de tipus exploratori són aquelles que es duen a terme amb la finalitat de familiaritzar-se amb el problema d'investigació. No hi ha un interès per aprofundir en la qüestió investigada, sinó només de conèixer-la millor. Pot ser un pas previ a una investigació més profunda, per exemple per verificar la factibilitat o per posar en pràctica els mitjans que poden utilitzar-se per dur-la a terme. Com argumenta Cea d'Ancona, els estudis exploratoris no solen constituir-se com a finalistes, sinó com a pas previ per a altres estudis posteriors. De fet, l'escassa representativitat de les mostres que s'utilitzen, siga per la mida o pels procediments de selecció de la mostra, contribueixen al seu ús limitat i en cap cas inferencial.

Les recerques de tipus descriptiu poden ser finalistes, o també servir com a pas previ d'un procés d'investigació més ambiciós. En aquest cas, es pot aplicar qualsevol de les tècniques d'investigació quantitatives (anàlisi de dades secundàries, enquestes, estudis de cas o experiments). L'única diferència amb les recerques d'altres tipus és la profunditat de l'anàlisi, que en el cas de les descriptives se sol quedar en l'àmbit dels encreuaments univariables i bivariables de tipus bàsic.

[6] En aquest sentit, és útil el paper dels col·legis oficials de Sociologia. En el cas del COLPIS de Catalunya, va tenir un document actiu d'orientació de preus de les diferents tècniques quantitatives i qualitatives i en alguna ocasió també ha fet algun curs dirigit al càlcul de pressupostos per a projectes. Al marge d'això, també són útils els preus públics de concursos d'investigació, com ara els que fa el CEO de la Generalitat de Catalunya, que sol incloure el pressupost en les fitxes tècniques, i més recentment també els que fa RTVE amb motiu dels sondejos electorals a peu d'urna. Pensem que és molt important que l'alumnat conega quin és el cost real d'una investigació social com a part del procés formatiu en tècniques d'investigació, no només quantitatives, sinó de qualsevol tipus.

Les recerques de tipus explicatiu afegeixen a les de tipus descriptiu la capacitat de buscar causes o raons a partir de les mateixes tècniques d'investigació. A més, utilitza mètodes d'anàlisi més complexos, entre els quals podem trobar correlacions, proves d'independència, regressions, etcètera. La diferència de les recerques explicatives amb les de tipus predictius es pot dir que és mínima, atès que les últimes incorporen als factors que ja coneixem l'objectiu de predir l'evolució del fenomen estudiat.

Per acabar, el disseny de tipus avaluatiu té com a objectiu principal l'estudi de l'efectivitat d'un programa, política o iniciativa. L'avaluació es pot classificar al seu torn en impacte, procés, valoració de necessitats, anàlisi de sistemes, anàlisi cost benefici i de conjunt. Tot seguit, analitzem els tipus de disseny d'avaluació un per un.

- L'avaluació d'impacte tracta d'analitzar els efectes del programa o política sobre la població objectiu. Es pot dir que és l'anàlisi més popular, perquè tracta de mesurar l'èxit o el fracàs d'una iniciativa a partir dels resultats del programa.
- L'avaluació de procés tracta d'esbrinar el funcionament real del programa analitzat, en el pla teòric, però també en el pràctic, per poder destriar les causes del seu èxit o fracàs.
- L'avaluació de necessitats tracta de detectar les necessitats que hauria de cobrir el programa o iniciativa en qüestió.
- L'avaluació mitjançant l'anàlisi de sistemes tracta d'estudiar les relacions entre els subsistemes implicats en la posada en pràctica del programa o iniciativa i els sistemes que hi estan relacionats.
- L'avaluació de cost-benefici estudia els costos del programa amb els seus resultats, cosa que se sol fer des del punt de vista exclusivament econòmic.
- L'avaluació de conjunt és la més complexa de les formes d'avaluació i inclou l'objectiu del programa o iniciativa, les alternatives de què es pot disposar i els costos de l'actuació.

Deixant de banda l'objectiu, el disseny de les recerques, atès el rang temporal que comprenen, poden ser de tipus seccional o longitudinal. Els dissenys seccionals són aquells en què la recollida de la informació s'efectua en un únic moment. Val a dir que aquest tipus de disseny d'investigació és compatible amb els dissenys anteriors, de manera que podem tenir una investigació predictiva i seccional. Els dissenys longitudinals, per contra, són aquells en què la recollida de la informació es fa en diferents moments, amb l'objectiu específic d'analitzar-ne la variació en el temps. La variació es pot observar almenys de tres maneres diferents: de tendències, de cohort i de panel.

- L'anàlisi de tendències s'encarrega d'observar l'evolució de la població objecte d'estudi en diferents moments de temps, amb la particularitat que les mostres per a cadascuna de les mesures són diferents.
- L'anàlisi de cohort tracta d'estudiar l'evolució, però en aquest cas no de tota la població sinó d'una cohort o subpoblació que comparteix alguna característica, com ara l'edat, per observar-ne l'evolució en relació amb alguna variable central de l'estudi.
- L'anàlisi de panel incorpora a l'estudi longitudinal que les persones que formen part de la mostra són exactament les mateixes, amb la qual cosa es pot comprovar quina n'ha sigut l'evolució durant el període estudiat. Per a Cea d'Ancona, la bondat d'aquest model en termes d'anàlisi de les causes del canvi està acompanyada de disfuncions com ara el desgast de la mostra en ser interrogada diferents vegades i la introducció del biaix per l'efecte aprenentatge, per tal com s'utilitza el mateix qüestionari de manera repetitiva (Cea, 2012: p. 57).

A banda dels tipus de disseny per objectiu i per temporalitat, també hi ha la classificació del disseny per experimentalitat. Així, en funció de cinc criteris de classificació, es pot dir que ens trobem davant d'un estudi preexperimental, experimental o quasiexperimental (Cea, 2012: p. 58). Aquests cinc criteris classificadors són els següents: la manera com se seleccionen les unitats d'observació, el nombre d'observacions dutes a terme, el grau d'intervenció de la persona que condueix la investigació, el control de variables explicatives alternatives a les centrals de l'estudi (factor que explica la validesa interna) i la possibilitat de generalització dels resultats (factor que interpel·la la validesa externa).

- Dissenys preexperimentals: atesos els criteris d'adscripció de l'experimentalitat, els dissenys preexperimentals es caracteritzen per l'absència de manipulació de les variables que intervenen en el procés d'investigació, és a dir, que s'observa el fenomen sense introduir cap modificació o alteració. En segon lloc, només s'efectua una mesura del fenomen, tot i que en pot incloure diferents variables. I a l'últim, no hi ha control de fonts d'invalidesa de la investigació. Això, juntament amb l'absència de grups de control i aleatorització en la formació de la mostra. Qualsevol enquesta tradicional cabria dins de la categoria preexperimental.
- Disseny experimental: inclou la manipulació experimental, per la qual l'investigador manipula a priori les variables que concorreran en el fenomen estudiat. També té en compte la formació de grups de control equivalents en les seues característiques al grup experimental, a excepció de les variables a mesurar. A

l'últim, preveu l'aleatorització de la formació de la mostra, de manera que els individus s'assignen de manera aleatòria als grups experimental i de control. Aquestes mesures incrementen la validesa interna, és a dir, la capacitat d'establir explicacions alternatives a l'observada per la influència de variables externes al model. Tanmateix, el model experimental no té un valor elevat pel que fa a la validesa externa, ja que la manipulació prèvia i l'escàs nombre de persones que conformen la mostra fan que els resultats no siguen extrapolables a l'univers.

- Disseny quasiexperimental: se situa entre les dues opcions anteriors, tret de la circumstància que normalment no té lloc en un laboratori, sinó en contextos de la vida real. Això implica que la distribució de les unitats no és equivalent entre el grup de control i l'experimental. I a diferència del disseny preexperimental, la persona encarregada de la investigació no es limita a observar, sinó que pot intervenir en la situació central de la recerca.

Per acabar, els dissenys d'investigació poden incloure tres tipus d'orígens de dades en funció del grau de control que té la persona investigadora. Aquesta circumstància, com les anteriors, s'encreua amb cadascuna de les classificacions en termes de disseny. D'aquesta manera, podem diferenciar entre les dades primàries, secundàries i terciàries. Les dades primàries són aquelles que ha pogut reunir l'equip d'investigació mateix. Les de tipus secundari han sigut recollides per altres equips d'investigació i la seua anàlisi ens permet realitzar agregacions ulteriors o, en el millor dels casos, treballar a partir de les microdades originals. A l'últim, l'origen terciari de les dades només permet a l'equip d'investigador acostar-se a la forma amb què altres persones han presentat el resultat de la investigació, la qual cosa se sol traduir en el fet que només es poden consultar les taules o gràfiques de manera agregada sense cap més possibilitat d'aprofundiment que la que donen les intencions primeres de qui elabora l'informe de resultats o la seua plasmació en forma d'article, capítol o llibre.

2.1.5. Conceptes bàsics d'estadística

La mesura en sociologia

La mesura, tal com explica González, implica comparar una magnitud amb una altra de la seua mateixa espècie considerada com a unitat, o amb una altra magnitud equivalent, cosa que dificulta la seua aplicació en ciències socials (González, 2000). La qüestió és com es pot mesurar una actitud o un posicionament a partir d'una escala preexistent i quina és la seua exactitud. Això implica que, en ciències socials, cal considerar la mesura com l'assignació de símbols als elements d'un conjunt de magnituds, propietats, objectes o esdeveniments o, dit d'una altra manera, l'assignació de nombres a objectes o idees com una forma de representació de les seues propietats. En paraules de Stevens, la mesura només és possible sempre que hi ha una espècie d'isomorfisme entre les relacions empíriques dels objectes o fenòmens i les propietats del que s'observa (Stevens, 1951). L'assignació d'una xifra a una propietat implica una certa semblança entre les estructures d'ambdós sistemes, tot i que això desemboque, finalment, en una certa ambigüitat en qualsevol tipus de mesura que utilitzem en ciències socials en general, i en sociologia en particular. Assumir aquesta feblesa significa també tenir clar que les ciències socials no seran tan exactes com les ciències naturals, però no per això s'han de deixar de contrastar empíricament aquelles hipòtesis que es plantege en el pla teòric.

La mesura, en tot cas, només és possible si prèviament hi ha un model matemàtic que la fa possible. Entre altres coses, mesurar implica una cosa tan bàsica com l'existència de numerals, la presència d'unes normes d'addició, el comptatge i, prèviament, l'existència d'una escala de mesura de la quantitat o intensitat de l'objecte o fenomen considerat (Stevens, 1951).

El problema amb la mesura en sociologia apareix en el moment en què es pretén captar de manera precisa una realitat, i transformar les observacions en dades. En aquest sentit, Paul Lazarsfeld explica en el seu conegut article *Evidence and Inference in Social Research* (1958) que no hi ha disciplina científica que s'enfronte al seu objecte d'estudi amb plena concreció, sinó que més aviat se solen centrar en les seues propietats i intenten abordar les relacions que es poden establir entre si, de manera que l'objectiu final sol ser trobar les lleis que expliquen aquestes relacions. Per tant, una de les primeres

coses a tenir en compte és que mesurar vol dir també restringir allò que s'observa. Seguint Lazarsfeld (1958), el procés pel qual un concepte es trasllada a un índex empíric passa per quatre estadis:

1. Imaginació o constructe del problema teòric. Pot començar com un acte creatiu o aparèixer del fet que la persona al càrrec de la investigació ha observat certes regularitats. En qualsevol cas, hi ha una impressió general, una noció aproximada d'allò que volem mesurar que és la que mobilitza la intenció de resoldre el problema de mesura.
2. Especificació del concepte. Una vegada tenim el primer constructe, cal especificar-ne les dimensions, unes vegades de manera lògica, d'altres de manera empírica. Així, s'arriben a cobrir totes les possibles dimensions d'allò que es vol mesurar, en un procés que pot ser realment ampli. A mesura que s'especifiquen les dimensions es guanya en el detall amb què coneixem en fenomen investigat, però alhora es perd en profunditat del concepte en si. Com apunta González, no hi ha una norma sobre com fer el pas de concepte a dimensions, ni de quantes dimensions són les ideals, sinó que solen ser la intuïció i l'experiència prèvia de la persona investigadora les que ho solen delimitar (González, 2000: p. 213).
3. Selecció d'indicadors. Una vegada seleccionades les dimensions, cal trobar indicadors per a cadascuna de les dimensions, tot tenint en compte que no sol ser suficient un únic indicador per dimensió, sinó que en fan falta més d'un. En alguns casos, els indicadors estan donats per l'existència d'estudis previs, mentre que en unes altres ocasions cal fer una tria de quins indicadors formen part d'allò que volem mesurar i quins altres en són externs.
4. Creació d'índexs. En paraules de Lazarsfeld, es tracta de tornar a ajuntar *Humpty Dumpty*, és a dir, combinar els indicadors per tal de tenir una idea global d'allò que s'ha mesurat i poder extraure'n conclusions que tinguen a veure amb el primer constructe teòric. Els índex poden cristal·litzar en la mesura ideal d'una determinada realitat, de manera que a partir d'aleshores cada vegada que algú vulga determinar la quantitat o la intensitat d'un fenomen, s'hi referisca a partir de l'índex que la mesura.

Pel que fa als indicadors, González en recull de dos tipus principals: els descriptius i els analítics. Els descriptius són aquells que expliciten la regularitat que hi ha en un conjunt de dades, mentre que els analítics transmeten el valor de les dades (2000: p. 218). Al mateix temps, i tenint en compte la seua capacitat de recollir les variacions en les dimensions que mesuren, en distingeix tres subtipus:

a. Els normatius: són aquells que es refereixen a aspectes sobre els quals hi ha un alt grau de consens pel que fa a la seua mesura i, per tant, no hi té lloc la discussió sobre quin podria ser el sentit que es dona a la variació.
b. Els objectius: són aquells que utilitzen dades físiques referents als individus o col·lectius i que es considera que no estan sotmesos a una interpretació subjectiva[7].
c. Els subjectius: es tracta dels indicadors que fan servir interpretacions subjectives de les dades o la realitat que mesuren.

Tal com argumenta González, l'amplitud d'una variable en sociologia pot ser diferent en funció del que s'intente representar (2000: p. 216). En tots els casos, les variables representen propietats de les unitats d'anàlisi. Per exemple, com veurem posteriorment, les nominals dicotòmiques classifiquen al voltant de dues propietats excloents, les ordinals classifiquen de manera ordenada les unitats d'anàlisi en funció de les seues propietats i les d'interval afegeixen a l'ordenació la mesura de la distància exacta entre cadascuna de les propietats. Pel que fa a les variables, Lazarsfeld les considera mesures empíriques individuals o grupals que incorporen algunes operacions en la seua conformació, de manera que s'inclou informació sobre individus i col·lectius (1958).

1. Algunes característiques sobre proposicions generalitzadores: el significat de les variables pot ser ambigu si no s'examinen en el context en què s'utilitzen.
    a. Tracten sobre un conjunt d'elements, els casos o unitats d'investigació.
    b. Aquests elements són considerats comparables.
    c. Cada element té un cert valor per a cada propietat, siga aquesta qualitativa o quantitativa.
    d. Les proposicions afirmen relacions entre les propietats.

2. Significat especial del que és *col·lectiu* i d'allò que implica ser *membre*: Un col·lectiu és qualsevol element d'una proposició format per membres. Però, els membres d'un col·lectiu no tenen per què ser individus, sinó que poden ser un altre tipus d'agrupacions. Per tant, quan es parla d'elements col·lectius en una proposició, cal especificar quins són els membres del col·lectiu analitzat.

[7] En aquest punt, González (2000: p. 219) cita el sexe com un dels indicadors objectius sobre el qual hi ha acord i no hi ha lloc a interpretació subjectiva. No obstant això, actualment el sexe és actualment un indicador que es posa en dubte, precisament per la diferent interpretació que se'n pot fer des del vessant biològic i social i, també, per les diferents conseqüències que pot tenir per a la vida quotidiana.

3. Propietats dels col·lectius: se'n poden distingir de tres tipus.
    a. Propietats analítiques: derivades de la implementació d'alguna operació matemàtica per a alguna propietat de cada membre individual.
    b. Propietats estructurals: propietats dels col·lectius que són obtingudes a partir d'operacions sobre les relacions de cada membre respecte d'alguns o de la resta dels membres.
    c. Propietats globals: les característiques dels col·lectius tenen a veure amb propietats que no es basen en informació sobre les propietats dels membres individuals.

4. Distinció entre les propietats analítiques dels col·lectius: les propietats analítiques ofereixen en ocasions dades que, a primera vista, resulten similars, com ara una mitjana comparada entre dues comunitats, però que en observar-ne les desviacions s'arriba a la conclusió que no són realment iguals. De la mateixa manera, es pot constatar que diferències en proporcions entre grups poden tenir a veure amb la seua conformació prèvia. En tots dos casos cal tenir en compte la configuració interna dels grups analitzats per tal d'entendre millor el que s'està mesurant.

5. Propietats dels membres individuals dels col·lectius: es pot dir que les recerques en què les unitats d'anàlisi són individus formen el gros del treball empíric. En el cas que els individus siguen considerats com a membres d'un col·lectiu -ara diríem, els classifiquem mitjançant una variable independent-, aleshores les seues propietats es poden classificar també en funció de si la resta del col·lectiu entra en la caracterització dels seus membres o no, tot seguint la classificació següent:
    a. Propietats absolutes: són les característiques d'individus que han estat obtingudes sense fer us d'informació sobre les característiques del col·lectiu o d'informació sobre les relacions de l'individu en comparació amb la resta dels membres. Un exemple de propietats absolutes és el sexe.
    b. Propietats relacionals: les propietats dels individus són obtingudes a partir d'informació sobre les relacions entre el membre i els altres membres. Un exemple que esmenta Lazarsfeld és la introducció d'una mesura sociomètrica en què cada individu és puntuat per la resta de membres del grup en una escala d'un a cinc, i així l'individu es podria caracteritzar pel resultat que dona la resta del grup, pel total que avalua l'individu a la resta

del grup, per la desviació mitjana dels valors respecte a la resta dels membres del grup, etc. (Lazarsfeld, 1958: p. 114).

c. Propietats comparatives: caracteritzen un membre per la comparació entre el seu valor, en termes absoluts o relacionals, amb la distribució d'aquesta propietat en el col·lectiu del qual és membre.
d. Propietats contextuals: descriuen un membre per una propietat del seu col·lectiu.

Unitats d'anàlisi

Com s'ha vist en l'apartat anterior, la unitat d'anàlisi més freqüent en ciències socials és l'individu. Tanmateix, molt sovint no disposem de dades desagregades o microdades, sinó de dades agregades que fan referència als individus o fins i tot a agregacions d'individus (barris, municipis, províncies, comunitats autònomes, països, regions, etc.). No s'ha de confondre això amb el fet que de vegades la unitat d'anàlisi puga ser un conjunt d'individus o col·lectiu. Per exemple, a nivell d'anàlisi, es poden trobar diferents agrupacions de persones, que també es coneixen com a conglomerats, com ara les seccions censals, els municipis, les comarques, les províncies o fins i tot els països. Cadascun d'aquests conglomerats pot esdevenir una unitat d'anàlisi constituïda a partir de variables registrades de manera individual com a casos. Seguint Corbetta (2007: p. 80), en aquest cas es pot diferenciar entre la unitat d'anàlisi i la unitat de registre, que se situa en un nivell inferior. També cal tenir en compte que el nivell d'agregació que suposa una institució és també superior al de l'individu i al del grup d'individus, per exemple una escola o una empresa. El quart nivell d'unitats d'anàlisi és l'esdeveniment, per exemple les eleccions d'un país. L'últim tipus d'unitat d'anàlisi és el que Corbetta anomena representació simbòlica o producte cultural, i està representat per documents secundaris (articles de premsa, notícies radiofòniques, fotografies), que també es poden interpretar com a unitats d'anàlisi.

En ocasions es pot diferenciar entre la unitat d'anàlisi i la unitat de mostreig, perquè no sempre coincideixen. Sí que han de coincidir, però, els objectius de la investigació amb la unitat d'anàlisi. Vegeu, per exemple, un cas pràctic al voltant de la discriminació de les dones en els anuncis de televisió: si el que volem esbrinar és si homes i dones apareixen representats de manera diferent en els mitjans de comunicació, aleshores la nostra unitat d'anàlisi no han de ser únicament les dones que apareixen en els mitjans, sinó també els homes per tal de comprovar si homes i dones són representats amb la

mateixa freqüència i amb els mateixos papers o rols (Hernández, Fernández i Baptista, 2014: p. 172).

En el seu manual ja clàssic sobre investigació social, Earl Babbie (2013: p. 103) proposa diferents unitats d'anàlisi entre:

- Individus: la unitat més habitual en recerca social, tot i que en ocasions són agafats com a part de grups més amplis.
- Grups socials: sense que hi haja d'haver un compromís o document de pertinença, sempre que siguen agafats com una entitat singular.
- Organitzacions: com, per exemple, associacions, empreses, corporacions, etc. La pertinença està més marcada i el fet que la institucionalització siga major fa que el tipus de variables que poden ser considerades siguen diferents a les dels grups socials.
- Interaccions socials: com a unitat d'anàlisi ens poden interessar cridades telefòniques, besos, baralles, correus electrònics, etc. Tot i que els individus en siguen els protagonistes, no es tracta de mesurar-los com a tals, sinó el producte de la seua interacció.
- Artefactes socials: qualsevol producte humà o del seu comportament, com ara llibres, cotxes, cançons, etc. poden ser analitzats com a unitats rellevants en un estudi empíric.

Altres autors, com Morris Rosenberg, encara apunten més unitats d'anàlisi, en aquest cas entre individus, grups, organitzacions, institucions, espais, cultures i societats (Rosenberg, 1968: p. 239 i s.).

Tipus de variables

La variable és un concepte crucial dins la investigació quantitativa, ja que no solament és la base de la recollida de dades, sinó també la perspectiva bàsica a partir de la qual s'estructura l'anàlisi (Cea, 2012). Una variable és, per definició, allò que varia. Aplicat al camp de les ciències socials, la variable és la característica d'un objecte, fenomen o persona que té almenys dos atributs o categories (si només en tinguera un, no es podria dir que varia). Per a Corbetta, en canvi, la variable és un concepte operacionalitzat (2007). Luis Camarero i altres consideren la variable en ciències socials com un conjunt de valors que classifiquen la població objecte d'anàlisi en distints grups, a partir de diferents categories classificatòries (Camarero et al., 2013: p. 29). A cadascun dels atributs

de la variable els anomenem categories, i en funció de quants en tinga i de quin tipus siguen, parlarem de diferents classificacions de variables amb diferents conseqüències tant pel que fa a la mesura com a l'anàlisi.

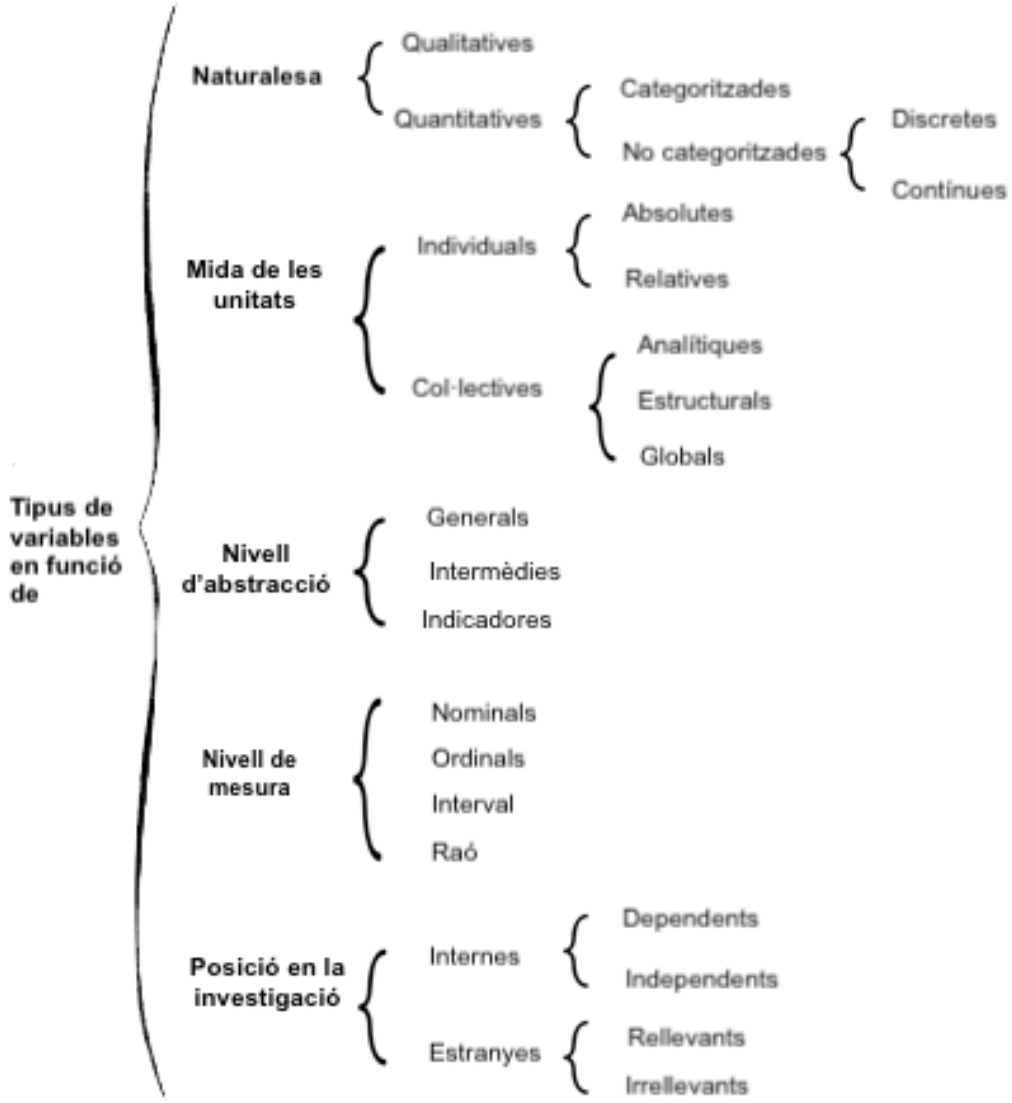
Perquè una variable i les seues categories estiguen plantejades correctament, i per tant mesuren allò que diuen mesurar i siguen analitzables com a tals, han de complir almenys tres premisses: exhaustivitat, exclusivitat i precisió, que tot seguit desenvolupem.

Les variables han de ser exhaustives perquè han d'incloure tota la variació de categories que garantisca que tots els subjectes estiguen representats en les opcions de resposta, definides a priori. Això és relativament senzill en variables com l'edat, però quan es tracta de variables categòriques la cosa es pot complicar. Per això, sempre és convenient tenir en compte l'opció *altres*, encara que això eixample el nombre de categories a analitzar, i també el *no sap* i el *no contesta*. Una recerca prèvia al moment de l'operacionalització per tal d'avaluar els possibles tancaments de les variables sol ser garantia d'una mesura exhaustiva, com també ho és un petit *pretest* abans de posar en pràctica el qüestionari a fi d'avaluar la bondat de les mesures proposades.

A més, les categories de les variables han de ser excloents, cosa que implica que una persona només es pot classificar en una única opció. Aquesta qualitat només afecta les variables d'una única resposta, ja que en ocasions poden haver-hi variables en què es poden utilitzar tancaments multiresposta. En el cas de les variables numèriques és molt important deixar els intervals tancats, de manera que no es repetisquen el final d'un i el començament de l'altre, perquè en aquests casos la classificació es torna impossible.

Per acabar, les categories han de ser precises. Sempre cal apostar pel tancament de pregunta més precís enfront del més genèric o agregat. Si s'ha de fer alguna agrupació, sempre es pot fer després del moment en què s'haja fet la recollida de les dades. En canvi, fer el camí invers no és possible. Així, en el qüestionari és molt millor preguntar l'edat de manera oberta que no agrupar-la per intervals. Només s'aconsella el contrari en aquelles variables en què ja hi ha una certa tradició en la mesura o en què els agregats són aconsellables pel perill de no resposta (cas del salari mensual en les enquestes del *Centro de Investigaciones Sociológicas*).

Les variables es poden classificar, com a mínim, d'acord amb el nivell de mesura, l'escala de mesura[8], el nivell d'abstracció i la funció en la investigació. Totes les variables compleixen un tipus, de manera excloent, de cadascuna de les classificacions. No obstant això, en la gràfica següent podem observar que les classificacions poden anar més enllà. Sierra en proposa una més, la mida de les unitats, que veurem tot seguit (Sierra, 2001).



Gràfica 4. Tipus de variables en funció de diferents criteris.

Font: Adaptat de Sierra, (2001: p. 106).

[8] Com en tantes altres coses, sol haver-hi discrepàncies entre diferents autors a l'hora de denominar les mateixes coses. Per exemple, en el cas del nivell de mesura, Francesc La-Roca s'hi refereix com a escala de la mesura, tot i que fa referència a la mateixa classificació que Stevens en la seua classificació de variables històrica (La-Roca, 2006: p. 23; Stevens, 1951: p. 25).

Tenint en compte l'escala de mesura, les variables es poden classificar en qualitatives i quantitatives. Entre les quantitatives, en podem distingir les categoritzades i les no categoritzades, és a dir, si s'han agrupat o no, i discretes, en funció de si l'existència de valors és ininterrompuda o no. En el cas que tinguem valors de manera ininterrompuda des del valor mínim fins al màxim, ens trobem davant una variable contínua, com és el cas de l'edat: des del valor mínim fins al màxim podem trobar valors enmig, que a més podrien anar tornant-se acurats a mesura que n'especifiquem més dades, per exemple sobre la data de naixement, hora, minut, segon, etcètera. Per contra, les variables discretes no presenten valors entre categories o atributs de la variable, sinó que hi ha una discontinuïtat que fa que desconeixem què hi ha enmig, és a dir, que només adopta determinats valors dins del rang (Sierra, 2001: p. 105). Totes les variables nominals i també les ordinals són discretes, i també les quantitatives que no es puguen dividir, com ara el nombre de fills.

En relació amb la mida de les unitats, les variables es distribueixen entre individuals i col·lectives: les individuals es refereixen a observacions relatives a individus, mentre que les col·lectives es refereixen a unitats d'observació que impliquen conjunts, grups o col·lectius. Pel que fa a les variables de tipus individuals, es poden expressar en nombres absoluts, és a dir, sense cap referència a altres característiques de l'individu, o en termes relatius, en què la variable inclou la comparació amb una característica d'origen. Alhora, les variables relatives poden ser de tipus comparatiu, relacional o contextual. Les comparatives fan referència a valors comparats amb el de la resta del grup a què pertany l'individu, per exemple ser el primer o l'últim en alguna característica. Les relacionals fan referència al rol de l'individu o la posició que ocupa en el grup més ampli del qual forma part. Per acabar, les variables relatives contextuals estableixen una relació entre l'individu i l'entorn social a què pertany (Sierra, 2001: p. 107). Entre les variables col·lectives, en destaquen les de tipus analític, les estructurals i les globals. Les analítiques són aquelles en què el sistema es fonamenta en els individus; les globals, la fonamentació és col·lectiva; i les estructurals, les variables s'originen de les dades individuals, basades en les relacions socials i interaccions dins del col·lectiu.

Segons el nivell d'abstracció podem considerar tres tipus de variable. En les de tipus general, que són les que no es poden observar directament per la seua amplitud, cal un procés d'operacionalització que les transforme en ítems mesurables. Aquest és el cas, seguint l'exemple de Cea d'Ancona (2012: p. 91), de variables com la classe social, que no aporta cap informació si l'única cosa que fa un qüestionari és demanar a la persona

interessada que s'autodefinisca en termes de classe social. En segon lloc, hi ha les variables de tipus intermedi, que expressen alguna dimensió compresa en una variable genèrica. No són encara una variable observable empíricament, sinó que més aviat necessiten un procés d'operacionalització per transformar-les en indicadors mesurables. Aquest és el cas dels àmbits educatiu, laboral i econòmic, sense entrar en la mesura de cadascun d'ells (Cea, 2012: p. 91). A l'últim, hi ha les variables empíriques, que són directament mesurables, de tal manera que són equivalents als indicadors. L'exemple que posa Cea d'Ancona és el dels cursos acadèmics acabats, ocupació principal o ingressos, per a cadascun dels vessants educatiu, laboral i econòmic esmentats anteriorment.

Tenint en compte la posició de les variables en la investigació, es pot distingir entre les que són internes al plantejament empíric i les que en resulten estranyes, és a dir, que no estan incloses en el model inicial, i que poden resultar rellevants per a la seua interpretació o no. Pel que fa a les variables internes, es pot distingir entre les dependents i les independents. Així, les variables dependents són aquelles els valors de les quals, en una investigació, depenen de les variables independents. Per tant, el que es vol analitzar és la seua variació per tal d'establir les relacions de dependència, covariació o causalitat que corresponguen. S'expressen mitjançant la lletra y en l'anàlisi estadística. Un exemple d'aquest tipus de variable és, segons Cea (2012: p. 91), una investigació en què la variable central a estudiar és la qualificació acadèmica. En segon lloc, hi ha les variables independents, que són aquelles que prediuen el valor que adoptarà la variable dependent, raó per la qual també reben el nom de variables explicatives o predictores. En aquest cas, se simbolitzen mitjançant la lletra x en l'anàlisi estadística. En el cas anterior, seguint l'exemple de Cea d'Ancona, les hores d'estudi (x_1), però també l'assistència a classe (x_2), el cocient d'intel·ligència (x_3), la motivació (x_4) o el sexe (x_5). D'altra banda, i tenint en compte la seua funció, també podem trobar les variables pertorbadores, que intermedien en la relació entre variable independent i dependent. A més, les variables pertorbadores poden ser de control, si hi ha un control abans o després de la recollida de dades, de manera que es pot observar l'efecte de les diferents variables, en aquest cas $x_{\{1;4\}}$ en la variable dependent. El cas contrari, és l'aparició d'una variable aleatòria, de difícil control, que influeix sobre la variable dependent sense que càpiga la possibilitat de controlar-la. Aquest és el cas, per exemple, de variables com ara posar-se malalt o rebre una mala notícia instants abans de fer la prova qualificatòria de l'assignatura.

Per acabar, d'acord amb el nivell de mesura, les variables es poden classificar en quatre grups: nominals, ordinals, d'interval i de raó. Aquesta classificació és la que va presentar

el psicòleg Stanley Stevens durant els anys quaranta i que va popularitzar posteriorment en el seu recull sobre psicologia experimental (1951).

El nivell de mesura nominal és el més qualitatiu -o no mètric- entre tots els tipus de variables. Expressa una qualitat de l'objecte o individu i no hi ha un ordre determinat entre les seues categories que, tanmateix, sí que han de ser excloents. Això fa que les opcions d'anàlisi estadística siguen menors que amb altres tipus de variables. També cal dir que, tot i ser qualitativa, la seua representació en qualsevol programari d'anàlisi estadística es fa a partir d'una representació codificada numèrica. Tot i amb això, aquest fet no implica ordre, sinó que s'assignen els nombres de manera arbitrària. El sexe és una de les variables nominals més fàcils de reconèixer, perquè a més se sol representar de manera categòrica d'acord amb una classificació biològica. La seua codificació, és a dir, la representació numèrica de les diferents categories de la variable, que en aquest cas no presenta isomorfisme amb les seues qualitats, se sol fer utilitzant l'1 per als homes i el 2 (o en ocasions el 6, per una qüestió d'evitar errors en la digitació durant el procés de tabulació) per a les dones. Per a Frankfort-Nachmias i Leon-Guerrero, la dicotomització mereix un comentari a banda tant pel que fa a les variables nominals com a les ordinals i fins i tot les d'interval, ja que és un procés molt comú que simplifica les tècniques d'anàlisi que es poden aplicar, el tipus d'estadístics que se'n poden extraure i les tècniques que s'hi poden aplicar, com ara el recompte i la correlació de contingència (Frankfort-Nachmias i Leon-Guerrero, 2000: p. 17; Stevens, 1951).

En segon lloc, hi ha les variables de tipus ordinal que, com les variables nominals, expressen una qualitat. A diferència d'aquestes, però, impliquen l'existència d'un ordre. Així, podem saber quina és l'ordenació de les categories, cosa que amplia les possibilitats d'anàlisi. Entre les variables que es poden caracteritzar com a ordinals hi ha les escales d'acord o satisfacció o diferents nivells, com ara el d'estudis. En aquest cas, l'ordre del codi sí que és convenient que tinga una correspondència amb allò que mesura, de manera que vaja de menys a més. No cal que hi haja un altre tipus de correspondència, perquè per les pròpies característiques de les variables ordinals no hi pot haver un altra ordenació. Aquest és el cas de variables amb un nivell d'acord que va del *totalment en desacord* al *totalment d'acord*. En la seua construcció s'ha d'estar a l'aguait de fer que tinga forma d'espill, ja que moltes vegades l'expressió lingüística que s'utilitza no és exactament l'equivalent en forma negativa. En el cas que ens ocupa, la codificació d'aquesta variable va del *totalment en desacord*, que tindrà el codi 1, al *totalment d'acord*, que va precedit per *en desacord*, *indiferent* i *d'acord*, que tindrà el codi *5*. Stevens atribueix a les variables ordinals operacions matemàtiques més complexes, com

és ara la mitjana (1951), i estableix que amb una interpretació ampla de les ordinals numèriques o discretes podem estendre el de l'interval a les anàlisis pròpies del següent nivell.

Finalment, hi ha les variables de tipus numèric, que se solen classificar en interval i raó (cardinals en paraules de Corbetta). La diferència entre ambdues és mínima: les dues són quantitatives, les dues suposen un ordre, en les dues coneixem la distància exacta entre les categories, però només en la de raó sabem que hi ha un punt zero vertader, és a dir, allà on es puguen mesurar zero unitats. Per a Stevens, el punt zero és simplement una convenció (1951), ja que les operacions que es poden plantejar amb variables d'interval i variables de raó són equivalents. També és la variable més versàtil pel que fa als tipus d'anàlisi que es poden dur a terme, no només perquè accepta tècniques d'anàlisi més complexes, sinó perquè per recodificació es poden fer també anàlisis més senzilles. La variable d'exemple per a les d'interval és l'altura mesurada en centímetres o l'edat, mentre que per a les de raó podem esmentar els ingressos mensuals mesurats en euros.

Models d'escales

La mesura de les actituds i les motivacions es pot estructurar a partir d'una sèrie de tancaments preestablerts de preguntes que faciliten la seua traducció en variables empíriques. Es tracta de les escales de mesura, mitjançant les quals es pot construir un cos d'indicadors capaç de representar situacions més complexes. Seguint Corbetta, una escala és un conjunt coherent d'elements que es consideren indicadors d'un concepte més general (2007). En l'àmbit de les ciències socials, les escales s'utilitzen per a la mesura d'actituds, sobre les quals es poden distingir els components cognoscitiu (informació que té el subjecte sobre l'objecte mesurat); afectiu (sentiment associat a l'objecte mesurat); i conductual (reacció davant l'objecte mesurat). Així, ens podem aproximar a la mesura de l'objecte per la via de la direcció de l'actitud (sentit positiu o negatiu per a l'individu), de la magnitud de l'actitud (grau de favorabilitat o desfavorabilitat del subjecte) o la intensitat de l'actitud (sentiment associat a l'actitud per part del subjecte). En qualsevol cas, les escales més utilitzades en la construcció de qüestionaris són l'escala Thurstone, l'escala Likert, l'escalograma de Guttman i el diferencial semàntic d'Osgood, que analitzarem tot seguit.

L’escala Thurstone deu el nom a Louis Thurstone, un especialista en psicometria, que la va desenvolupar el 1928 per mesurar les actituds envers la religió. Aquesta escala es caracteritza per fer mesuraments indirectes de l’actitud a partir de les opinions dels subjectes enquestats. El primer pas en la conformació de l'escala consisteix a construir un conjunt d’ítems vinculats a l’actitud estudiada en forma de proposicions categòriques, sovint sorgides d’una revisió de la bibliografia, entrevistes prèvies o revisió d’escales semblants. El segon pas és avaluar els ítems amb l’ajuda d’un grup de persones expertes, que són els que han de valorar cada ítem en funció de com són favorables envers l’objecte d’estudi, des del valor 0, molt desfavorable, al 10, molt favorable. Acabada aquesta ordenació, cal seleccionar aquells indicadors que formaran part de l’escala definitiva, cosa que es fa en desestimar totes aquelles mesures esmentades per les persones expertes que s'allunyen més de dues desviacions típiques del punt mitjà de totes les valoracions. Sobre la bateria proposada, es demana a les persones investigades que diguen si estan d’acord o en desacord amb les afirmacions. El grau d'acord o desacord originalment s'organitzava al voltant d'una resposta dicotòmica, però es pot combinar amb respostes més complexes, per exemple de tres, quatre o cinc graus d'acord. Agafant com a referència la mitjana dels valors proposada per les persones expertes, es pregunta als individus per la seua postura a favor o en contra, de manera que finalment s’obté un valor a considerar, resultat de la suma ponderada de les postures a favor i la resta de les postures en contra. Aquest indicador es pot obtenir tant individualment com, i això és el que ens interessa, a nivell de mostra i/o submostra. En la taula següent es pot apreciar el valor de la ponderació atorgat per les persones expertes a cadascuna de les formulacions, que és en definitiva el que constituirà l'índex actitudinal de posició a favor o en contra de la guerra, tal com el presenten Francés *et al*. (2014: p. 139).

Valor escalar	Ítems	D'acord	En desacord
(5,6)	1. La defensa és l'única justificació per a la guerra.		
(4,0)	2. És quasi impossible disposar d'una gran força armada i no sentir-se temptat a utilitzar-la.		
(5,6)	3. La pau i la guerra són essencials per al progrés.		
(0,4)	4. Tots els països haurien de desarmar-se immediatament.		
(2,3)	5. La guerra condueix a la misèria milions de persones que no tenen veu en la seua declaració.		
(9,0)	6. El militarisme és necessari per a la defensa adequada i la protecció dels individus d'un país.		
(6,6)	7. Hauríem de tenir un poc d'entrenament militar en les nostres escoles.		
(1,9)	8. En una guerra violenta, els ciutadans haurien de poder rebutjar la crida a les armes.		
(3,4)	9. Les diferències civils i nacionals poden resoldre's sense guerres.		
(10,0)	10. La guerra ennobleix i estimula les més altes i millors qualitats de la humanitat.		

Taula 1. Exemple de proposta d'escala Thurstone sobre actituds envers la guerra

Font: p. Adaptat de Francés *et al.* (2014: p. 139)

L'escala Likert deu el nom a Rensis Likert, també especialista en psicometria, que la va presentar el 1932. La seua proposta compartia amb l'escala Thurstone que disposava d'una sèrie d'enunciats que en la pràctica funcionen com a reactius de les opinions, però en el cas de Likert no es fixava en una escala contínua prefixada sinó en graus ordinals de favorabilitat o desfavorabilitat envers l'objecte d'estudi, que habitualment es classifica en graus de 5 a 7 categories que van del *totalment d'acord* al *totalment en desacord*, passant pel *d'acord, indiferent i en desacord*. De la mateixa manera que l'escala Thurstone, els components de l'escala Likert s'han de sotmetre a una mostra representativa, que són els que responen, i en funció del resultat en termes de quartils se seleccionen aquelles mesures situades entre els quartils Q_1 i Q_3. Pel que fa a l'anàlisi, cal tenir en compte que les categories resultants són de tipus ordinal, amb la qual cosa l'anàlisi no permet fer comparacions individuals, però sí amb la variable d'interval resultant.

	Totalment d'acord	D'acord	Ni d'acord ni en desacord	En desa-cord	Totalment d'acord
Vivim en una societat violenta.	1	2	3	4	5
La situació social de la dona ha millorat molt els darrers anys.	1	2	3	4	5
La violència contra les dones és molt greu.	1	2	3	4	5
Les dones pateixen discriminació social, econòmica i laboral.	1	2	3	4	5
Quan una dona és agredida per la seua parella, alguna cosa deu haver fet per provocar-ho.	1	2	3	4	5
Pel bé dels fills, la dona que pateix violència no ho hauria de denunciar.	1	2	3	4	5
En cas que un dels pares haguera de deixar de treballar, hauria de ser la dona.	1	2	3	4	5
La violència domèstica és un assumpte familiar, i no hauria d'eixir d'allí.	1	2	3	4	5

Taula 2. Exemple d'escala Likert sobre violència masclista

Font: Adaptat de Francés *et al.* (2014: p. 141)

L'escala o escalograma de Guttman deu el nom al matemàtic i sociòleg Louis Guttman. Més que mesurar les actituds a partir d'operacions sobre un conjunt de valors amb el qual obtenir un valor interval, la variable resultant és el resultat d'un conjunt d'enunciats als quals l'individu entrevistat va donant la seua opinió. Aquests enunciats s'escalonen de manera que estar d'acord amb un determinat ítem implica estar d'acord amb tots els ítems precedents. Aleshores, el procés de construcció passa pel disseny d'una sèrie d'indicadors relativament nombrosos que, a més, cal ordenar jeràrquicament. Amb les respostes ordenades, cal eliminar-ne aquelles que, aplicat el quocient de reproductibilitat, donen un valor per sobre de 0,90. Aquest coeficient es calcula a partir de la fórmula:

$$R = 1 - \frac{E}{(Q * N)}$$

On E representa el nombre total d'errors a les puntuacions atorgades; Q representa el nombre d'ítems en l'escalograma i N el nombre de subjectes que responen. Una vegada seleccionats els ítems que han de formar part de l'escalograma, s'han de col·locar en

forma dicotòmica (d'acord/en desacord) per tal d'avaluar individualment la quantitat d'ítems amb què s'ha mostrat d'acord. Resulta un model de fàcil anàlisi, tot i que pot tornar-se limitat per tal com perd importància la multidimensionalitat de les actituds.

	En desacord	D'acord
Que visquen persones d'una altra ètnia o nacionalitat al meu país.	0	1
Que visquen persones d'una altra ètnia o nacionalitat a la meua ciutat.	0	1
Mantenir una conversa amb persones d'una altra ètnia o nacionalitat.	0	1
Que una persona d'una altra ètnia o nacionalitat siga veïna meua.	0	1
Que un familiar meu es case amb una persona d'una altra ètnia o nacionalitat.	0	1

Taula 3. Exemple d'escalograma de Guttman d'actituds enfront del racisme

Font: Adaptat de Francés *et al.* (2014: p. 143).

L'última escala que cal tenir en consideració és el diferencial semàntic d'Osgood. L'escala porta el nom del seu creador, Charles Osgood, un psicòleg nord-americà que va desenvolupar l'escala del diferencial semàntic per tal de mesurar els significats connotatius. La idea principal d'Osgood és semblant a les anteriors, però afegeix complexitat al model de Guttman. L'actitud ja no és dicotòmica, sinó que se situa en un espai entre dos paràmetres, entre els quals s'ha de situar l'individu, cosa que és un indicador de l'actitud que es vol mesurar. En aquest cas, la construcció de l'escala passa també per la selecció d'una sèrie de categories en forma d'adjectius bipolars: en la mostra se seleccionen els conjunts d'adjectius que presenten una càrrega factorial més elevada per a cada dimensió, derivada de les puntuacions de la mostra de subjectes. Una vegada seleccionats, se situen en bateria, separats per un rang de set categories. A fi d'evitar la resposta automàtica, és convenient girar alguns dels ítems, sempre col·locant la puntuació final en el sentit de l'adjectiu positiu d'aquell fenomen que volem estudiar. L'anàlisi sol passar pel càlcul de les mitjanes, amb la qual cosa s'obté un valor interval que pot ser analitzat amb qualsevol tipus d'operació estadística. Tot i que és una escala que ha demostrat ser útil, algunes de les crítiques posen en dubte el seu paper per a l'anàlisi actitudinal.

Independents	7	6	5	4	3	2	1	Dependents
Intolerants	1	2	3	4	5	6	7	Tolerants
Integrats	7	6	5	4	3	2	1	Marginats
Desconfiats	1	2	3	4	5	6	7	Confiats
Vergonyosos	1	2	3	4	5	6	7	Sociables
Valorats	7	6	5	4	3	2	1	Desvalorats
Improductius	1	2	3	4	5	6	7	Productius

Taula 4. Exemple de diferencial semàntic d'Osgood sobre el sentit social de les persones grans.

Font: Adaptat de Francés *et al.* (2014: p. 144).

Del concepte a les dades

Una vegada s'ha passat dels conceptes a les mesures concretes, és moment d'aplicar la tècnica de recollida de dades que faça possible una anàlisi encaminada a descriure la nostra població objecte d'estudi o a donar una resposta a la hipòtesi que hàgem plantejat. Això se sol fer a partir de les tècniques quantitatives d'investigació social, bàsicament qüestionaris o també de l'anàlisi de dades generades de manera més o menys automatitzada, cosa que cada vegada és més habitual amb la introducció de les noves tecnologies i l'internet de les coses. Aquestes tècniques són explicades amb molt més detall en l'assignatura *Tècniques quantitatives d'investigació social* que, com s'ha pogut comprovar en el tercer apartat, s'imparteix durant el segon curs i té assignats 9 crèdits. És per això que no hi insistirem en aquest punt, més enllà de donar-ne algun apunt a l'aula.

El pas lògic següent consisteix a tabular els resultats, i això vol dir passar els resultats en paper o en dades brutes a un programa d'anàlisi que ens permeta dur a terme les operacions que siguen necessàries i pertinents. En el nostre cas, farem servir el programa SPSS, que és un dels que hi ha disponibles a les aules d'informàtica de treball obert que, com s'ha pogut comprovar en el segon apartat, en són unes quantes, tant a la Facultat de Ciències Socials com als aularis. Ja hem pogut constatar anteriorment que el pas de les dades a la matriu numèrica no és tan fàcil com sembla i que implica tota una sèrie de decisions que cal tenir en compte, especialment pel que fa a la codificació de les variables nominals, les respostes obertes, els valors perduts, etc.

Mitjançant la tabulació arribem a una matriu numèrica que, en el cas de l'SPSS, ofereix les variables en columnes i els casos en files. En ocasions, com quan es tracta de dades provinents de fonts secundàries, com ara els baròmetres del CIS, aquestes dades s'han d'interpretar mitjançant uns registres de variables que proporciona la institució mateixa. En aquests casos és com millor es veu què significa una variable, què són les categories i com es mesuren o quin és el nivell de mesura de la variable. Així, per a una variable nominal, és fàcil descobrir que els casos solen variar entre uns pocs valors; per contra, per a una variable d'interval com l'edat, el nombre de categories és molt més nombrós. Tanmateix, fins i tot així cal dur a terme una anàlisi, tal com veurem en l'apartat següent. Si no, podem caure en el parany de pensar que una variable amb un alt índex de variació és, per defecte, d'interval, quan pot tractar-se d'una variable nominal amb una gran amplitud de categories, com ara la Classificació Nacional d'Ocupacions (CNO-11), amb més de 700 categories.

Població i mostra: cens i enquesta

En els seus orígens, l'estadística va començar sent una expressió de la mesura de la població desenvolupada per l'estat a fi de controlar la població amb vista a poder preveure de quantes persones podia comptar per a l'exèrcit i també la quantitat d'ingressos de què podia disposar en un moment donat (Sánchez, 2001). Aquest és el cas dels censos que es van dur a terme durant el segle XVIII, particularment els que va dirigir Gottfried Achenwall des de la Universitat de Göttingen (1752) o els que es van fer sota el regnat de Carles III (Aranda primer, el 1768, i Floridablanca després, el 1785). No és casualitat que els censos apareguen en aquest moment històric, segle XVIII, en què ja hi havia una certa confiança en la intel·ligibilitat de la natura i el concepte d'estat modern començava a perfilar-se. En tots dos casos es tractava de comptar tota la població resident en un territori, cosa que anava associada als conceptes de població (caps de família o llars fins a temps recents) i també d'estat nació i el seu territori. D'aquella intenció primigènia de comptar la població en van sorgir almenys tres models d'estadística. El model anglès es basava en l'estudi de la mortalitat a partir de registres escrits, molt més orientat a l'àmbit numèric i fortament influït per les figures de William Petty i John Graunt, especialment el segon, al qual es considera el fundador de la demografia moderna. El model alemany s'interessava pels costums, la població o qualsevol aspecte rellevant, i destacava els aspectes literaris sobre els numèrics. A la fi, el model francès, sota l'influx

d'Abraham de Moivre, s'aproximava més a l'alemany que no a l'anglès (Sánchez, 2001; Korstanje, 2009).

Evidentment, el fet d'arreplegar informació d'una població és una tasca molt costosa, quant a esforç, temps i diners[9]. No ha d'estranyar, doncs, que aparegueren obres que no es basaven directament sobre el conjunt de la població. L'enquesta tal com l'entenem avui dia té l'origen al segle XVII, quan Sébastien Le Prestre, marqués de Vauban, va publicar el seu *Méthode génerélle et facille pour faire le denombrement des peuples* (1686), que en la pràctica eren unes normes per dur a terme els censos poblacionals, algunes de les quals encara són útils ara mateix. Per exemple, proposava comptar totes les persones de la llar, independentment de la seua edat, i incloure dades com el sexe, l'edat o l'estat civil. Fins i tot va dissenyar un qüestionari per facilitar l'arreplegada d'informació i una anàlisi relativament ràpida. Uns cent anys més tard, John Sinclair va editar *Statistical Account of Scotland*, una obra publicada entre el 1791 i el 1825 amb què interrogava la població escocesa sobre les seues característiques sociodemogràfiques, però també sobre la pràctica religiosa o l'activitat econòmica (Cea, 2012: p. 186; Leti, 2000). A partir d'aquells moments és relativament habitual trobar enquestes, com ara les de Charles Booth sobre la població de Londres (1902-1903), la de Benjamin Seebohm sobre la pobresa i les condicions laborals (1906) o l'obra de Fréderic Le Play (1855) sobre els obrers a diferents països europeus (Cea, 2012: p. 187). Una de les qüestions comunes a totes les investigacions primigènies, especialment les del segle XIX i principi del XX, és la preocupació per les condicions de vida i treball de la classe obrera, preocupació que denota la inseguretat de les classes dominants respecte d'una novetat desconeguda fins aleshores: l'aparició d'una massa obrera, que no sabien de quina manera afrontar.

Fet i fet, però, una de les figures clau per al desenvolupament de l'enquesta moderna és Max Weber. El mateix personatge a qui devem el desenvolupament futur del paradigma interpretatiu i la metodologia qualitativa, és qui dona l'impuls a dues qüestions claus en la correcta aplicació de l'enquesta: d'una banda, la selecció dels informants; de l'altra, la millora del qüestionari. Gràcies a la seua obra, compresa entre el 1892 i el 1908, amb base quantitativa i mostres en alguns casos de més de 1.000 persones,

[9] A tall d'exemple, els censos d'Aranda i Floridablanca es configuraren en uns tres anys. També cal recordar el canvi metodològic que va implantar l'Instituto Nacional de Estadística respecte del cens de 2011, el primer en què es deixa de preguntar a tota la població i es passa a preguntar només a una mostra mitjançant qüestionaris autoemplenats, amb un resultat molt pobre que va fer impossible continuar les sèries i fer anàlisis mínimament rigoroses en poblacions d'abast mitjà.

qüestionaris amb desenvolupament de 27 preguntes i anàlisis en alguns casos comparatives, es pot dir que l'enquesta entra al segle XX (Cea, 2012: p. 189; Brain, 2001).

Un dels avanços que més va fer per la popularització de l'enquesta va ser l'aplicació de la teoria de la probabilitat i, en particular, el mostreig representatiu i probabilístic. La qüestió de la representativitat va ser introduïda per Anders Kiær el 1897. El mostreig probabilístic és degut a Arthur Bowley, qui el 1915 va publicar una investigació sobre la pobresa a la ciutat de Londres en què, per primera vegada, s'utilitzava una selecció aleatòria d'informants. Jerry Neyman va adaptar la idea de la selecció aleatòria i va introduir l'estratificació de la mostra (Neyman, 1934), moment en el qual també defensa el mostreig per conglomerats, el mostreig en poblacions finites i l'error de mostreig (Cea, 2012: p. 189 i ss.; Alasuutari, Brickman i Brannen, 2008).

Amb les bases de l'enquesta fetes públiques, no ha d'estranyar que durant els anys 1920 i 1930 foren moltes i variades les mostres d'enquestes. De fet, en aquell moment sorgeixen empreses de prospecció de mercats centrades en la investigació comercial i els estudis preelectorals. Potser un dels casos de major èxit és el que va enfrontar les prediccions electorals de la revista *The Literary Digest*, que defensava que les mostres de major mida (en aquest cas els seus subscriptors) tenien també un major poder explicatiu, enfront del disseny mostral representatiu per quotes de sexe i edat de Gallup i Crossley. La mostra més petita va ser capaç de predir la victòria de Roosevelt el 1936, i això va situar l'enquesta com una tècnica fiable per predir esdeveniments (Squire, 1988). Pocs anys després, Paul Lazarsfeld va fundar el *Bureau of Applied Social Research* a Columbia, des d'on estudia aspectes electorals, però també posa en pràctica l'enquesta per panel i l'anàlisi de dades en taules encreuades, tot i que encara no aplicarà models d'inferència estadística ni estimació d'intervals de confiança (Cea, 2012: p. 193). A Samuel Stouffer i els seus col·laboradors de *The National Opinion Research Center* (NORC) devem l'ús de les enquestes autoemplenades, la formulació de diferents preguntes per mesurar un mateix fenomen i la millora de l'anàlisi (1941). Per la mateixa època comencen a aparèixer les primeres publicacions crítiques amb plantejaments com el biaix en l'entrevista o l'efecte de les expectatives de la persona entrevistada sobre els resultats (Cea, 2012: p. 194). Aquests avanços, fonamentalment centrats en l'àmbit nord-americà, coincideixen amb l'expansió del funcionalisme i el paradigma positivista, per la qual cosa seran majoritaris tant en l'àmbit acadèmic com també en les publicacions i l'ús que en fa l'administració. En el moment de màxim auge de l'enquesta com a tècnica d'investigació social, l'estat espanyol estava sotmès a una guerra iniciada pel colp d'estat feixista del 1936 o directament sota la misèria, l'ostracisme i l'autarquia

provocada pel règim franquista. No és estrany que, per bé que la primera enquesta es puga datar el 1883 (*Comisión de Reformas Sociales*, sobre el problema social), de la qual se sap que sorgiren multitud de rèpliques territorials tant a nivell social com cultural i etnogràfic (Lisón, 1968; Duque, 2003), les primeres enquestes modernes han d'esperar als anys 50 del segle XX. Una de les primeres, centrada en l'estudiantat universitari de Madrid, va ser obra del feixista Manuel Fraga i del matemàtic Joaquín Tena l'any 1949 (Cea, 2012: p. 197). Les primeres enquestes tenen una clara limitació política, d'altra banda gens estranya en el funcionalisme, i se centren sobre grups professionals (metges, estudiants, empresaris, treball domèstic), família i joventut. Pocs anys després, el 1963, va aparèixer l'Instituto de la Opinón Pública, que el 1976 es transforma en el Centro de Investigaciones Sociológicas. Al mateix temps comença a desenvolupar-se un entramat d'empreses d'investigació social, començant per l'Instituto Eco (1958), vinculat a Jesús Ibáñez, o DATA (1965), vinculat a Amando de Miguel i Juan Linz. Una valoració del mercat dels estudis d'opinió el 1979 estimava el seu impacte en 2.000 milions de pessetes. L'any 2000 se situava al voltant dels 22.882 milions de pessetes (Cea, 2012: p. 198). Una estimació de la patronal *AEDEMO* (Asociación Española de Estudios de Mercado) per a l'any 2013, en plena crisi econòmica, situava l'impacte en 438 milions d'euros, uns 73.000 milions de pessetes al canvi. L'última dada que ha fet pública *AEDEMO* és de 506 milions el 2017, uns 84.000 milions de l'antiga moneda.

Criteris de qualitat en els dissenys d'investigació: fiabilitat i validesa

Seguint María Ángeles Cea d'Ancona (2012: p. 62), la qual al seu torn s'inspira en alguns dels articles apareguts en *International Journal of Social Research Methodology*, cal fer un esment especial als criteris de qualitat dins de l'àmbit de la investigació. És significatiu pensar en els criteris de qualitat, perquè avui dia se sotmeten a avaluació centenars de projectes d'investigació en diverses convocatòries, entre les quals hi ha les assignatures de metodologia, que també han d'avaluar, en aquesta ocasió, no tant el projecte com el resultat final del nostre estudiantat. Aleshores, saber quins són els criteris, no ja acadèmics, sinó científics, a l'hora d'avaluar un disseny d'investigació és més que pertinent.

El primer determinant de qualitat ha de ser l'ajust de la investigació a l'objectiu principal declarat. També cal analitzar l'encaix dels recursos (humans, materials i econòmics) i del temps de què es disposa per fer efectiva la investigació. Qualsevol dels punts anteriors que no s'ajuste a allò que s'espera o a allò que és de sentit comú dins el camp de la investigació afectarà la qualitat de la investigació que es vol dur a terme. A més, diu

Cea, cal tenir en compte almenys quatre criteris d'avaluació que mesuren la validesa del disseny:

- Validesa interna: com s'ha abordat anteriorment, la validesa interna mesura el control d'explicacions alternatives a la que es vol analitzar. És particularment important quan s'està davant d'investigacions de tipus explicatiu perquè controla la relació de causalitat entre variables i destria quina és la variable que té un major efecte sobre la dependent. Això vol dir, per tant, controlar les variables pertorbadores. El control de la validesa interna es pot fer a priori, és a dir, abans de recollir la informació, o posteriorment a la fase de recollida. El primer cas és el propi de les investigacions de tipus experimental, en què el control de la validesa interna es fa a partir de la creació de grups de control, mentre que el segon cas és el de les investigacions de tipus correlacional, en què la validesa es controla a partir d'anàlisis multivariables i, en tot cas, tenint en compte les variables que puguen influir en el resultat.
- Validesa externa: la mesura de la validesa externa es relaciona amb la capacitat del disseny de ser generalitzable a la població d'on s'ha extret la mostra, però també a altres moments i espais. Això estarà relacionat amb el nombre de persones incloses en la mostra, però també de les seues característiques i la manera amb què hagen estat triades les persones que en formen part. En principi, els procediments de selecció aleatoris o probabilístics afavoreixen que la mostra siga representativa de la població, mentre que els no aleatoris o no probabilístics fan que la mostra no siga representativa pel tipus de selecció duta a terme.
- Validesa de constructe: es tracta d'una variant de la validesa interna que se centra en el mesurament dels conceptes centrals de la investigació. En el moment de dur a terme l'operacionalització, han de quedar cobertes totes les dimensions possibles del concepte que es vol estudiar.
- Validesa de conclusió estadística: en aquest cas és una variant de la validesa externa, relacionada amb la significativitat de l'anàlisi estadística. Com major siga la mida de la mostra, menor serà la variància de les variables (o menors errors típics), cosa que es tradueix en majors possibilitats d'inferència en les proves d'hipòtesis.

Cal afegir-hi, a aquestes mesures de la validesa en el disseny, les de validesa de la mesura. La primera ja és coneguda perquè coincideix amb una de les fonts de validesa del disseny: la de constructe. La segona, seguint Cea (2012: p. 112), és la de contingut, és a dir, el grau en què una mesura empírica dona cobertura a la varietat de significats

que inclou el concepte. És evident que, tant aquesta com les altres mesures, no tenen un equivalent numèric en termes d'avaluació, sinó que més aviat es tracta d'un nivell d'optimització de la validesa en què ha d'haver-hi un mínim de preocupació per la qualitat de la mesura. D'altra banda, hi ha la validesa de criteri, amb què es contrasta la mesura amb algun criteri preexistent, comunament acceptat, de mesura del mateix concepte. Dins la validesa de criteri podem trobar la validesa concurrent, si la mesura nova correlaciona amb un criteri adoptat per al mateix moment, cosa que es pot comprovar amb el quocient de correlació. D'altra banda, hi ha la validesa de criteri predictiva, quan la comparació no es fa amb un factor concurrent en termes temporals, sinó amb un criteri futur. En aquest cas, també s'han d'aplicar càlculs de correlació per tal d'avaluar-ne la validesa. Una de les mesures que augmenten la validesa, en tots els casos, és la decisió de dur a terme una operacionalització múltiple, és a dir, mesurar el mateix concepte de maneres diferents en la mateixa recerca -per exemple, en el mateix qüestionari-, tot intentant incloure les diferents dimensions que té.

Al concepte de validesa, a més, s'ha d'afegir el de fiabilitat, que en part es relaciona també amb el que explicava Merton sobre la ciència pública. La fiabilitat mesura la possibilitat d'obtenir els mateixos resultats, o resultats consistents, aplicant els mateixos procediments de mesura. Es pot dir que una mesura fiable no sempre és vàlida, però que una mesura vàlida normalment és fiable. En el cas de la fiabilitat, sí que hi ha mesures estadístiques per tal d'avaluar-la, que analitzem tot seguit: el mètode de test-retest, el mètode de les dues meitats i el mètode d'*alfa de Cronbach* (pel seu creador, el psicòleg Lee Cronbach).

- Test-retest: és la manera més senzilla de comprovar la fiabilitat. S'aplica el mateix instrument a la mateixa mostra en dos moments de temps diferents (com a mínim allunyats un de l'altre un mes, per mirar de fer disminuir l'efecte memòria). La fiabilitat s'estableix a partir del càlcul del quocient de correlacions en les respostes de les dues mesures, sent aquestes variables contínues, i s'estableix com a mesura de fiabilitat estàndard una $r \geq 0{,}80$, on una r propera a 1 indica una fiabilitat perfecta i una r propera a 0 una fiabilitat nul·la amb la fórmula següent, on y_1 és la primera mesura i y_2 és la segona:

$$r_{y_{i1},y_{i2}} = \frac{cov(y_{i1}, y_{i2})}{\sqrt{var(y_{i1})\ var(y_{i2})}}$$

- Mètode de les dues meitats: la comprovació de la fiabilitat es fa sobre subjectes diferents, però en un mateix moment, tot dividint els ítems que han de ser mesurats, i finalment comparant els resultats de les dues meitats, totes amb unes característiques semblants. A continuació, s'aplica la mateixa mesura de correlació sobre els resultats, amb idèntica interpretació que el mètode anterior.
- Mètode d'*alfa de Cronbach*: en aquest cas el que s'avalua és la matriu de variàncies-covariàncies d'una escala, el pas previ al càlcul de la correlació, de manera que no hi ha una estandardització. El quocient *alfa* dona una mesura de la variància comuna entre cada parell d'ítems involucrats en el test amb la fórmula següent, on *k* és el nombre d'ítems, S_i^2 és la variància de l'ítem *i* i S_t^2 és el resultat de la suma de tots els ítems, i la interpretació novament es fa agafant el valor mínim de fiabilitat 0,80:

$$\alpha = \left[\frac{k}{k-1}\right]\left[1 - \frac{\sum_{i=1}^{k} S_i^2}{S_i^2}\right]$$

2.2. Socioestadística descriptiva unidimensional

L'agregació quantitativa d'individus amb una perspectiva sumatòria i centrada en l'arreplega d'una quantitat xicoteta d'informació per a ser analitzada és un procés que apareix a partir del segle XVII i que beu de diferents factors i tradicions. Per un costat, l'aritmètica política anglesa i els recomptes parroquials de batejos, matrimonis i morts organitzats per Graunt (Desrosières, 1998: p. 18). No obstant, la denominació *Estadística* ens arriba per via de l'escola alemanya, que precedeix l'anglesa en termes temporals, de la mateixa manera que la configuració de les matrius en files i columnes es deu a la necessitat del oficials alemanys per a representar la vasta diversitat territorial de tres-cents microestats previs a l'existència d'un estat unificat i poderós. Així mateix, va resultar fonamental la tradició francesa d'instaurar les proves escrites, per damunt de l'oralitat, com una manera de demostrar la certesa dels esdeveniments enregistrats (Desrosières, 1998: p. 23). També va resultar un element decisori el major poder de l'Estat francés, en comparació amb el britànic i l'alemany, primer com a monarquia i després com a república i imperi, a l'hora d'instaurar per la via jacobina les diferents mesures en les demarcacions. Pràcticament fins al segle XIX tota aquesta informació serà utilitzada de manera privativa pels successius governs, fins que la il·lustració comence a fer públics els seus resultats, bàsicament a partir de les publicacions de Sébastien Bottin en 1799 (Desrosières, 1998: p. 34).

Com explica Juan José Sánchez Carrión respecte de l'aparició de l'estadística moderna, un requisit previ al mesurament d'una població, és que aquesta siga reconeguda com a tal i es puga formar un agregat, cosa que, per exemple, no passava en l'Europa del segle XVII, en què no tots els habitants eren considerats subjectes polítics actius i, per tant, mai haurien entrat en una hipotètica enquesta per mostreig (Sánchez, 2001: p. 37). Per tant, un primer mecanisme, previ al recompte, és la identificació com a unitats agregables, és a dir, ciutadans, i això no ocorre fins que no tenen un reconeixement com a tal. Si abans parlàvem de la importància del recompte en l'aparició de l'estadística, un avanç important el va constituir la codificació, és a dir, el tall del continu i la classificació dels individus en funció d'un criteri determinat, que en el moment iniciàtic de l'estadística moderna podia ser més o menys senzill i visible (cas del sexe) o podia estar construït a partir de classificacions més complexes (Sánchez, 2001: p. 41). D'altra banda, alguns supòsits medievals, com ara el que invocà Guillem d'Occam en el segle XIV (*Pluralitas*

non est ponenda sine necesitatite, literalment, la pluralitat no s'ha de considerar sense necessitat) són aplicables a l'estadística actual (Desrosières, 1998: p. 68).

Per a Jesús Ibáñez, el desenvolupament de la perspectiva distributiva moderna, que és la que se centra en la mesura estadística dels conceptes sociològics, implica l'aparició de dues pinces: la de la selecció de la mostra i la de la situació comunicativa de l'entrevista (Ibáñez, 2002). La primera pinça és la que captura els cossos (mostreig), mentre que la segona captura les ànimes (aplicació del qüestionari). Utilitzant l'etimologia d'*investigació* (del llatí *vestigium*, literalment, seguir les petjades) Ibáñez proposa un símil amb la caça, en què la construcció de la mostra esdevé una cacera de cossos efectuada pel predador-investigador, que en realitat exerceix d'auxiliar del vertader predador. D'altra banda, hi ha la situació comunicativa, l'entrevista, en la qual hi ha implicats uns entrevistadors, que no tenen la mateixa probabilitat d'entrevistar en funció del seu aspecte i origen social i que exerceixen de subjectes domats pel poder, que entrevisten els objectes extrets del seu context, i de qui només els interessen les respostes del qüestionari. L'aplicació d'enquestes es tractaria, per a Ibáñez, d'una situació individualista, irreversible, simplificant i homogeneïtzadora (Ibáñez, 2002, 73). Precisament, Ibáñez comparteix la perspectiva individualista en termes d'agregacions amb dos personatges als quals els separen tres-cents i set-cents anys: Adolphe Quetelet i Guillem d'Occam. Guillem d'Occam sostenia una postura nominalista, segons la qual només es podia interpretar la realitat a partir dels individus com a entitats físiques, cosa que constitueix la base de la seua filosofia i el conegut precepte de la navalla d'Occam. Aquesta llei, originada a partir de la polèmica de les possessions terrenals de la congregació franciscana i la incapacitat de prendre-la com una unitat quan no era més que la suma d'individus es transformarà, mitjançant el que Desrosières anomena transmutació màgica del coneixement estadístic (1998: p. 70), en la base sobre la qual, per oposició, Quetelet desenvoluparà, quatre segles després, la seua teoria sobre l'home mitjà[10] que veurem a continuació.

2.2.1. Anàlisi de freqüències

L'anàlisi més senzilla que es pot dur a terme, donada una variable determinada (i per tant amb més d'un valor diferent) per a un mínim de dos casos (i per tant, una mostra,

[10] En aquest punt utilitzem l'expressió *home mitjà* (*homme moyen* en l'original) conscients que es tracta d'un llenguatge masculinitzant, però també del fet que la ciència, en temps de Quetelet, tenia aquesta inclinació que encara arrossega avui dia.

encara que siga xicoteta) és la distribució de freqüències. En notació matemàtica, direm que es tracta d'una variable x per a *i* casos, és a dir, x_i. Aquest és el cas de la matriu de dades que presentem tot seguit, en la qual es presenten deu casos extrets de manera aleatòria del baròmetre 3201 (*Encuesta Social General Española*), confeccionat pel CIS l'any 2017. En la matriu de respostes apareixen un conjunt de variables de diferents característiques, per a aquesta mostra de 10 persones, en què trobem:

- x_0: variable auxiliar, de tipus ordinal, que expressa l'ordenació dels individus de la mostra,
- x_1: sexe, variable nominal dicotòmica amb els valors H (home) i D (dona),
- x_2: edat, variable d'interval, mesurada en anys,
- x_3: escala de satisfacció amb l'aparença personal, variable ordinal numèrica, del 0 (gens satisfet/a) al 10 (molt satisfet/a),
- x_4: contactes en *Instagram*, variable d'interval,
- x_5: grau d'acord amb la confiança que, en el futur, tot anirà millor amb els valors DA (d'acord) ID (indiferent) i ED (en desacord).

Així, la matriu té una estructura en la qual trobem classificats en files els casos, expressats mitjançant la notació matemàtica n_i, en què trobaríem n_1 com a primer cas i n_{10} com a últim cas; i classificades en columnes, les variables, des de x_0, la variable auxiliar, fins a x_5, la que mesura el grau de confiança en el futur. Abans de recollir les dades d'aquestes 10 persones que conformen la mostra, hem hagut de prendre algunes decisions que tenen a veure amb la part introductòria d'aquesta fonamentació teòrica: en primer lloc, dissenyar l'instrument de recollida de dades i escollir de quina manera s'incorporen els casos, el tipus de qüestionari, etcètera. En segon lloc, establir unes escales de mesura per a les variables i triar la manera en què s'introdueix la variable en termes de pregunta: per exemple, del sexe i de la valoració de la confiança en el futur, una expressada en termes binaris i l'altra amb tres possibles categories (podria haver estat en altres termes, amb més amplitud o introduir les opcions *no sap/no contesta*). En el cas de les variables quantitatives, s'ha hagut d'escollir una escala existent i comunament acceptada per la població. Per exemple, l'edat, que es mesura en anys, se sol preguntar de manera directa, però altres vegades és una pregunta més complexa (el CIS sol fer aquesta pregunta amb el format "*¿Cuántos años cumplió ud. en su último cumpleaños?"*). En el cas de la satisfacció amb l'aparença personal i els contactes en *Instagram* poden existir altres mesures que poden ser interessants, com ara una satisfacció més restringida (per exemple, de cinc graus) o una mesura d'activitat en xarxes, més que únicament els contactes. També s'hauran pres decisions sobre la manera en què

se selecciona la mostra de persones que analitzem i quina funció tindran les dades que s'extrauen, en termes de possibilitat d'inferència, cosa que veurem posteriorment[11].

1	D	22	8	40	DA
2	H	24	3	5	ED
3	D	23	7	89	DA
4	H	25	4	1	DA
5	H	22	8	204	DA
6	D	23	7	175	DA
7	H	22	6	231	ID
8	D	23	8	180	DA
9	H	23	6	108	DA
10	D	24	9	254	ED

Taula 5. Matriu de dades d'exemple.

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

Analitzar les freqüències de les variables de la matriu d'una en una significa escollir per quina es començarà l'anàlisi. En el cas que ens ocupa, ens centrarem en la variable edat, que anteriorment hem etiquetat com a x_2. Passar de la matriu de dades a un agregat significa dur a terme un recompte lògic de les freqüències de cadascuna de les categories de la variable escollida. Aquest recompte tindrà relació amb les unitats amb què es mesure la variable, en aquest cas els anys, i la precisió amb què l'hàgem sol·licitada en el moment de recollir les dades. Per exemple, l'agrupació seria diferent si haguérem demanat detall dels dies o les hores viscudes, en comptes dels anys.

Valors x_i	Recompte	Freqüència absoluta n_i
22	III	3
23	IIII	4
24	II	2
25	I	1
$\sum$		10

Taula 6. Recompte de la variable edat de la matriu d'exemple

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

[11] Tot i que en l'assignatura de Socioestadística s'introdueixen tots aquests conceptes, serà en segon, durant la impartició de Tècniques Quantitatives d'Investigació Social, quan s'hi aprofundirà.

El següent pas lògic és el d'expressar les freqüències absolutes en termes relatius al total de la població o mostra, expressat com a sumatori en l'última fila (N). El càlcul de la freqüència relativa, expressada amb la notació matemàtica fr_i es fa a partir dels valors de les freqüències absolutes respecte del total de la mostra (N), de manera que el sumatori de les freqüències absolutes donarà com a resultat la unitat. L'avantatge de treballar amb freqüències relatives és que afegeixen a la informació del recompte el pes de cada categoria sobre la unitat. De fet, encara que fem el càlcul sobre la unitat, és fàcil fer la transformació en termes percentuals, potser més intel·ligibles atès que hi ha més costum en la interpretació amb base 10 o amb base 100 que sobre la unitat, cosa que té relació amb l'isomorfisme de què parlava Stevens.

$$fr_i = \frac{n_i}{N}$$

x_i	n_i	Freqüència relativa fr_i
22	3	0,3
23	4	0,4
24	2	0,2
25	1	0,1
∑	10	1

Taula 7. Taula de freqüències amb freqüències relatives

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

De la mateixa manera, un càlcul que pot ajudar a entendre la distribució de la variable és registrar l'acumulació dels valors de cada categoria, a la qual anomenarem na_i (n acumulada d'i), seguint l'ordre lògic (sumatori de la freqüència absoluta d'x_i més l'acumulada anterior x), de manera que l'última categoria siga el sumatori dels casos acumulats, allò que hem anomenat N i que s'expressaria de la manera següent:

$$na_i = \sum_{X \leq x_i} n_i$$

Això deixaria la taula de freqüències com segueix:

x_i	n_i	fr_i	*Freqüències absolutes acumulades* na_i
22	3	0,3	3
23	4	0,4	7
24	2	0,2	9
25	1	0,1	10
$\sum$	10	1	

Taula 8. Taula de freqüències amb freqüències absolutes, relatives i absolutes acumulades

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

De la mateixa manera, es pot calcular la freqüència relativa acumulada, cosa que ens deixa observar de quina manera es distribueixen les freqüències relatives sumades de manera continuada fins al valor de la unitat, que serà el de l'última categoria. Això es pot expressar amb la fórmula següent:

$$fra_i = \frac{N_i}{N} = \frac{1}{N}\sum_{X \leq x_i} n_i = \sum_{X \leq x_i} \frac{n_i}{N} = \sum_{X \leq x_i} fr_i$$

La qual cosa deixaria finalment la taula de freqüències de la manera següent:

x_i	n_i	fr_i	na_i	Freqüències relatives acumulades fra_i
22	3	0,3	3	0,3
23	4	0,4	7	0,7
24	2	0,2	9	0,9
25	1	0,1	10	1
$\sum$	10	1		

Taula 9. Taula de freqüències amb freqüències absolutes, relatives i absolutes i relatives acumulades

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

La taula de freqüències estàndard, per tant, tindrà el format següent:

x_i	n_i	fr_i	na_i	fra_i
x_1	n_1	fr_1	na_1	fra_1
...	...	...	...	...
x_i	n_i	fr_i	na_i	fra_i
...	...	...	...	...
x_n	n_n	fr_n	na_n	fra_n
$\sum$	N	1		

Taula 10. Taula de freqüències en format estàndard

Font: Elaboració pròpia

Resulta evident que en els exemples anteriors hi ha pocs casos i també poques categories en una variable que sol ser tan ampla com l'edat. Quan una variable és contínua o quan té moltes categories i ens interessa retallar-la per tal de fer-la més intel·ligible, podem transformar-la en una nova variable composta d'intervals, de manera que obtenim una distribució de freqüències semblant, però a partir de categories agrupades. A banda de l'avantatge de facilitar-ne l'anàlisi, hi ha un inconvenient, que és la pèrdua d'informació resultat de disminuir el nombre de categories. Alguns consells per a dur a terme una agrupació són:

a. Mantindre la regularitat en els intervals
b. Evitar que hi haja massa categories buides
c. Mantindre una certa proporció amb la mida de la població
d. Evitar superposicions i definir clarament els límits, especialment en els extrems

En qualsevol cas, en agrupar les categories d'una variable quantitativa perdem informació, i algunes de les operacions que abans eren relativament fàcils de dur a terme, ara han de passar per un pas previ, que és el càlcul de la marca de classe, que a efectes operatius rep la mateixa notació matemàtica que les categories, això és, x_i, però en aquest cas es calcula a partir de la fórmula següent, en la qual l_{i-1} és el límit anterior obert de l'interval i l_i és el límit posterior tancat. A efectes de notació matemàtica, s'assenyalen amb un claudàtor obert en el límit inferior i un claudàtor tancat en el límit superior:

$$x_i = \frac{l_{i-1} + l_i}{2}$$

Així, si agafem la variable x_4, el nombre de contactes en l'aplicació d'*Instagram*, i l'agrupem en intervals de cinquanta en cinquanta contactes, obtenim el recompte següent:

$]l_{i-1}+l_i]$	Recompte	n_i
]0-50]	IIII	4
]50-100]	I	1
]100-150]	I	1
]150-200]	II	2
]200-250]	II	2
$\sum$		10

Taula 11. Recompte de freqüències de la variable x_4 agrupada en intervals

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

A partir del recompte es poden calcular les marques de classe de cada categoria (x_i), que a partir d'ara seran els valors de referència per a subsegüents operacions.

$]l_{i-1}+l_i]$	x_i	n_i	fr_i	na_i	fra_i
]0-50]	25	4	0,4	4	0,4
]50-100]	75	1	0,1	5	0,5
]100-150]	125	1	0,1	6	0,6
]150-200]	175	2	0,2	8	0,8
]200-250]	225	2	0,2	10	1
$\sum$		10	1		

Taula 12. Taula de freqüències amb freqüències absolutes, relatives i absolutes i relatives acumulades

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

Una vegada vista l'anàlisi de freqüències, caldrà diferenciar els nivells d'anàlisi que es poden dur a terme amb cadascuna de les variables implicades en l'anàlisi. Per això, caldrà que tinguem en compte quin és el nivell de mesura de les variables, tal com hem vist anteriorment (nominal, ordinal o d'interval), vist que això ens permetrà aplicar-hi un càlcul o un altre. Així, encara que físicament siga possible calcular a mà o demanar al programari estadístic que calcule la mitjana o la desviació típica de la variable sexe, nominal dicotòmica, no tindria sentit més enllà del que suposa la distribució de freqüències. Totes les mesures de posició, dispersió i forma es poden calcular amb variables quantitatives, com ara les ordinals numèriques i les variables d'interval. Per contra, per

a les variables qualitatives, les nominals i les ordinals, les mesures que podem dur a terme són molt més limitades, perquè entre altres coses allò que ens permeten en una anàlisi unidimensional és l'anàlisi de freqüències i poc més. Això mateix podem observar en la taula següent, a mode de resum, dels tipus d'operacions estadístiques que podem desenvolupar per a les variables en funció de l'escala de mesura.

		Escala de mesura		
		Qualitativa		Quantitativa
		Nominal	Ordinal	
Posició central	Moda	X	X	X
	Mediana		X	X
	Mitjana			X
Posició	Valors extrems			X
	Quantils		X	X
Dispersió	Rang			X
	Rang interquartílic			X
	Variància			X
	Desviació típica			X
	Coeficient de variació			X
Forma	Asimetria			X
	Curtosi			X

Taula 13. Mesures de resum segons l'escala de mesura

Font: Adaptat de López-Roldán i Fachelli (2015, 3.3: p. 26)

2.2.2. Mesures de posició

El camí de l'agregació fins al càlcul de la mesura de posició més important, la mitjana, només va ser possible quan va existir un determinat acord en el fet que aquests agregats eren del mateix tipus, de manera que formaven un nou ens superior, cosa que convergeix en la figura d'Adolphe Quételet. La novetat en Quetelet no és la mesura d'un indicador sobre la regularitat dels nombres, que ja anteriorment havien aplicat alguns doctors com Arbuthnot o rectors com Sussmilch, tot i que sempre des del punt de vista de la divina providència (Desrosières, 1998: p. 74). Tampoc ho era la creació d'una nova mesura mitjana per tal de fer estimacions, cosa que ja havia aplicat anteriorment Sébastien Vauban per estimar poblacions a partir de mostres de naixements, matrimonis i defuncions (Desrosières, 1998: p. 27). Per contra, Quetelet, que se centrava en l'estudi de les característiques físiques, va observar mitjançant l'aplicació d'histogrames (gràfica a la qual ens referirem a continuació) que moltes de les mesures sobre el cos seguien

una distribució que seguia la llei binomial, allò que des de 1894, gràcies a l'obra de Karl Pearson, coneixem com la corba normal. Així, distingeix tres tipus de mesures de la mitjana (objectiva, subjectiva i aritmètica), a partir de les quals dedueix que, atesa l'escassa varietat que mostren mesures com l'altura i la forma del cos humà, es podia considerar que si triem dos grups a l'atzar, les mitjanes d'altura seran molt semblants. El següent pas lògic va ser la utilització d'aquesta mesura com a element extern, comparatiu i destinat a la presa de decisions, cosa que es produirà a partir de mitjans del segle XIX. Com apunta Sánchez Carrión, el pas de l'home mitjà com a ideal a l'home mitjà com a mesura d'allò mediocre, incapaç d'adaptar-se per sobreviure, passa pel cosí segon de Charles Darwin, Francis Galton (Sánchez, 2001: p. 11), que va contribuir, entre altres coses, a la formulació de mesures de posició com ara la mediana o els quantils (específicament, els quartils), però sobretot se'l reconeix per aportacions en l'àmbit de les mesures de dispersió, covariació i correlació, com veurem posteriorment.

Moda

La primera de les mesures de posició és la moda (Mo_x), el paràmetre de posició central més elemental i intuïtiu: si seguir la moda és practicar allò més habitual, la mesura de la moda no deixa de ser el valor de la distribució amb major freqüència (La-Roca, 2006-41). Es tracta també de l'únic paràmetre de posició que es pot aplicar a les variables nominals. Pot passar que no existisca un valor modal, sinó més d'un, la qual cosa minora el valor explicatiu del paràmetre.

En el cas de tenir una variable en la qual les categories estan compostes per intervals, per exemple, una variable contínua que s'ha transformat en discreta, tal com s'ha aplicat en l'exemple anterior, el càlcul de la moda s'ha de fer sobre la marca de classe, de manera que s'aplica la següent fórmula compensatòria. L_{i-1} és l'interval inferior de la categoria modal, n_{i+1} és la freqüència absoluta de la categoria posterior a la modal, n_{i-1} és la freqüència absoluta de la categoria prèvia a la modal i a_i és l'amplitud de la categoria modal.

$$Mo_x = L_{i-1} + \left(\frac{n_{i+1}}{n_{i-1} + n_{i+1}}\right) * a_i$$

En el cas que els intervals siguen de diferents amplades, caldria calcular-ne les altures a partir de la fórmula $h_i = n_i/a_i$, de manera que en comptes d'utilitzar les freqüències, s'operaria amb les altures.

Mediana

La mesura de la mediana és un dels paràmetres que va utilitzar Galton en el seu famós experiment sobre la mesura del pes d'un bou, si bé allò que ha quedat en l'imaginari col·lectiu és la utilització de la mitjana, com veurem a continuació[12]. La mediana (Me) ofereix el valor central, és a dir, el que divideix la distribució en dues parts iguals. Per a poder-la calcular cal que la variable en qüestió siga, com a mínim, ordinal i que les seues categories s'hagen disposat de manera vertical. Si tenim la taula de freqüències relatives acumulades (fra_i) i la variable és contínua, només caldrà fixar-se en quin és el valor que ocupa la posició central, és a dir, el punt $\frac{N}{2}$ o la freqüència relativa 0,5. En el cas que tinguem un número senar d'observacions, la mediana ocuparà el lloc central de la sèrie, mentre que si el número és parell, la mediana serà el terme mitjà dels dos valors més centrats (La-Roca, 2006: p. 42). La fórmula de la mediana és relativa, ja que ofereix la posició de l'indicador en la mostra, que en tot cas, s'hauria de calcular amb la fórmula següent:

$$Me = \frac{n+1}{2}$$

En el cas de tenir una variable amb les categories organitzades per intervals, el càlcul de la mediana s'ha de fer amb el valor ponderat de l'interval en qüestió. Així, la fórmula de la mediana seria lleugerament diferent i incorporaria l'interval inferior de la categoria central (L_{i-1}), la freqüència acumulada prèvia a la de la categoria central (Na_{i-1}) i també l'amplada de l'interval de la categoria central (a_i) i la freqüència absoluta de la mateixa referència (n_i), combinades de la manera següent:

$$Me = L_{i-1} + \left(\frac{N}{2} - Na_{i-1}\right) * \frac{a_i}{n_i}$$

[12] La història sobre l'experiment del pes del bou que organitzà Galton a la fira de ramaderia de Plymouth és un poc més enrocada del que sembla a priori. La conta amb tot detall Surowiecki (2005: p. XI). Tal com es pot observar en els articles originals publicats en *Nature*, les primeres mesures que utilitza Galton són la mediana i els percentils agafats de cinc en cinc (Galton, 1907a). Tres setmanes després, un lector avesat li demanà si no seria millor calcular la mitjana. És quan obté la fascinant dada tan precisa sobre el pes del bou que li va donar popularitat (Galton, 1907b).

Mitjana

Els primers càlculs de mitjanes aritmètiques deriven, com a mínim, de l'obra de Ptolomeu, tal com és citada, segles després, per Al-Biruni (Eisenhart, 1974: p. 31). Aquells càlculs, però, fan referència al rang mitjà o valor intermedi, que encara no feien referència a allò social ni tampoc a una instància major o externa a l'origen de les dades sobre les quals s'havia calculat el paràmetre. Com s'ha vist anteriorment, haurem d'esperar primer els càlculs de Vauban, i després la formulació de la mitjana en Quetelet per tal d'acostar-nos al concepte actual de mitjana.

La mitjana aritmètica és la més representativa, però demana que la variable siga contínua, per la qual cosa no és possible utilitzar-la amb variables nominals, com s'apuntava abans. Es representa amb el símbol $\bar{x}$ (o μ quan es tracta de la mesura de l'univers[13]) i s'obté de la suma de tots els productes de cada valor per la seua freqüència, dividint el resultat pel total d'observacions o casos. La seua fórmula és la següent:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{N} = \frac{\sum_{i=1}^{n} x_i}{N}$$

Si estem treballant amb una taula de freqüències com la que hem presentat anteriorment en format de matriu, només caldria fer el sumatori del producte entre x_i i n_i dividit pel total d'observacions (N), o el que és el mateix, el sumatori del producte de les categories o marques de classe (x_i) per les seues freqüències relatives (fr_i), de manera que la fórmula quedaria de la manera següent:

$$\bar{x} = \sum_{i=1}^{I} x_i fr_i$$

[13] La lletra grega μ es pronuncia mi.

x_i	n_i	fr_i	na_i	fra_i	$x_i fr_i$
22	3	0,3	3	0,3	6,6
23	4	0,4	7	0,7	9,2
24	2	0,2	9	0,9	4,8
25	1	0,1	10	1	2,5
$\sum$	10	1			$\bar{x}_i$= 23,1

Taula 14. Taula de freqüències amb càlcul de la mitjana

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

Novament, en el cas que les dades amb què treballem estiguen agrupades en intervals, el càlcul es faria a partir de la marca de classe, que ja estaria codificada en la taula de distribució de freqüències com a x_i.

Les propietats de la mitjana són variades: en primer lloc, si multipliquem una variable x_i per una constant C, aleshores la mitjana $\bar{x}_i$ queda multiplicada per la mateixa constant. De la mateixa manera, si sumem a una variable x_i una constant C, aleshores la mitjana $\bar{x}_i$ augmenta un valor equivalent al de la constant. En tercer lloc, la mitjana de la suma de dues variables x_1 i x_2 és equivalent a la suma de les seues mitjanes $\bar{x}_1 + \bar{x}_2$. Finalment, la suma de les desviacions de totes les dades respecte a la mitjana sempre és zero.

El problema de la mitjana, com observa La-Roca, és que tot i ser un dels paràmetres més utilitzats, no se sol tenir en compte algunes de les seues limitacions. La primera és la restricció a les dades quantitatives i, per tant, la limitació a variables ordinals numèriques i d'interval. La segona és la importància de la distribució dels valors, que poden presentar una dispersió molt elevada, amb la qual cosa la mitjana deixa de ser definitòria de la posició central, i tindria més sentit fer-ne el càlcul a partir d'un paràmetre com la mediana. En paraules de La-Roca, una excursió amb bicicleta en què la mitjana de la pendent és zero no implica que no comporte cap d'esforç (2006: p. 46). Per tant, no s'hauria d'interpretar sense tenir en compte mesures de dispersió perquè, altrament, pot estar altament afectada pels valors extrems.

Entre les possibles aplicacions de la mitjana, hi destaca l'ús per a comparar subpoblacions, tant entre elles, com també amb la mesura de l'entitat major. Per tal de dur a terme aquest últim paràmetre, caldria aplicar la següent fórmula de càlcul de la mitjana

del conjunt, de manera que els pesos relatius de les diferents subpoblacions estarien inclosos en el càlcul combinat:

$$\bar{x} = \frac{\bar{x}_1 N_1 + \ldots + \bar{x}_k N_k}{N_1 + \ldots + N_k}$$

Quantils

Una de les aportacions de Francis Galton a l'estadística descriptiva, a banda de la mediana, va ser la utilització dels quantils com a mesures de posició no centrals. En el cas del cèlebre experiment de la fira de ramaderia de Plymouth, el mateix Galton utilitza els percentils en el càlcul col·lectiu del pes del bou. No obstant, la paraula quantil no apareix fins l'any 1940, mentre que la primera denominació en l'obra de Galton era *equi-postile*. La primera de les successives denominacions dels quantils va ser la dels quartils i els octils (1879), seguida dels decils (1882) i finalment els percentils (1885), d'acord amb les observacions de Jeffrey Aronson (2001).

Els quantils no deixen de ser generalitzacions de la mediana, és a dir, mesuren les posicions relatives en una distribució de freqüències, respecte de diferents maneres de considerar possibles divisions de la distribució. Es pot dir, per tant, que els quantils es manifesten en grups etiquetats segons el nombre de parts en què es divideix la distribució. Els quantils més habituals són els que divideixen una distribució en quatre parts (quartils), deu parts (decils) i cent parts (percentils). El càlcul es fa de manera semblant a la mediana, en què només canvia el denominador de la fórmula. En el cas que tinguem una distribució de freqüències contínua i les categories estiguen ordenades, el càlcul es fa localitzant en la columna de freqüències acumulades el punt corresponent a la posició relativa que es vulga calcular. Per exemple per al tercer quartil (o Q_3) es procediria amb la fórmula següent:

$$Q_3 = 3\frac{N}{4}$$

Per a l'anàlisi dels quartils en una distribució de freqüències ens pot ser de molta utilitat el diagrama de caixes que, com veurem tot seguit, incorpora no només informació dels quartils, sinó també de la mediana i també un indicador de la dispersió en la figura dels bigots. Per últim, també incorpora la presència de valors fora de la normalitat o extrems.

De manera anàloga al càlcul dels quartils, les fórmules per al decil i el percentil, per exemple el decil 6 (o D_6) i el percentil 79 (o P_{79}) serien les següents:

$$D_6 = 6\frac{N}{10}$$

$$P_{79} = 79\frac{N}{100}$$

Novament, en el cas que les variables implicades en el càlcul estiguen agrupades en intervals, el càlcul s'ha de fer de manera que el resultat estiga ponderat amb l'interval inferior de la categoria central (L_{i-1}) en combinació amb la freqüència acumulada prèvia a la de la categoria del quantil en qüestió (Na_{i-1}) i també l'amplada de l'interval de la categoria en qüestió (a_i) i la freqüència absoluta de la mateixa referència (n_i), combinades de la manera següent per als tres quantils més habituals (quartils, decils i percentils):

$$Q_p = L_{i-1} + \left(\frac{PN}{4} - Na_{i-1}\right) * \frac{a_i}{n_i}$$

$$D_p = L_{i-1} + \left(\frac{PN}{10} - Na_{i-1}\right) * \frac{a_i}{n_i}$$

$$P_p = L_{i-1} + \left(\frac{PN}{100} - Na_{i-1}\right) * \frac{a_i}{n_i}$$

De la combinació dels quartils Q_1 i Q_3, la mesura de la mediana i els valors màxim i mínim d'una distribució, naix allò que alguns autors apunten com el resum de cinc nombres (*the five numbers summary*) que, combinat amb les gràfiques de caixes, ofereixen la informació necessària per tal de caracteritzar una distribució de freqüències determinada (Moore, Notz i Fligner, 2018: p. 126).

2.2.1. Mesures de dispersió

Si les mesures de posició ens retornaven indicadors respecte de la posició del paràmetre en la distribució de freqüències, les mesures de dispersió ofereixen una informació complementària, en aquest cas, la distància de cadascun dels valors de la distribució respecte dels indicadors de posició, tant la mitjana com la mediana. El propi Quetelet ja

era conscient de la presència de desviacions respecte de la tendència central d'una mesura, que ell interpretava com a imperfeccions en la realització d'un model i que, metafòricament, identificava amb la història del rei de Prússia que encarregà diferents escultures a diversos escultors, sent que cap d'elles era un model perfecte del rei i també totes eren diferents entre sí mateixes (Desrosières, 1998: p. 75). Tot i que Quetelet ja era conscient de la mesura de la desviació entorn a la mitjana, plantejada com a errors, novament va ser Francis Galton, l'últim dels *gentelman* de la ciència en paraules de Stigler (Desrosières, 1998: p. 128) qui deduí el funcionament de la desviació en 1877. Això ho aconseguí gràcies a dues coses: la primera, una intuïció que el va fer conjuminar conceptes procedents de diferents autors i tradicions: la mitjana de Quetelet, la divisió en categories de Booth, la distribució normal de Gauss i, del seu cosí Darwin, l'herència dels atributs individuals (Desrosières, 1998: p. 115). En segon lloc, la invenció d'una màquina, el *quincunx*[14], amb la qual provà de manera empírica un dels debats que havia portat de bòlit els seus col·legues Poisson, Bienamyé i Lexis: de quina manera podia un procés aleatori, composat al seu temps de seleccions aleatòries, conduir a resultats regulars i aplicables a la probabilística. Amb el *quincunx* va aconseguir unir els conceptes del triangle de Pascal, el teorema del límit central i la distribució normal, tot demostrant que una distribució binomial, quan està formada per un elevat nombre de casos, s'aproxima a una distribució normal. Com en tantes altres coses, els avanços intuïtius de Galton van ser desenvolupats per Karl Pearson, qui en 1894 publicà les seues aportacions a la teoria de l'evolució, dins l'òptica eugenèsica compartida per ambdós, on es parlava ja obertament de desviació estàndard i del seu càlcul matemàtic (Pearson, 1894).

En termes generals, es pot dir que el càlcul de la posició central no es pot interpretar de manera correcta si no s'acompanya d'alguna mesura de dispersió que ajude a posar-la en el seu context (La-Roca, 2006: p. 54).

Rang

El rang o recorregut és un paràmetre que mesura la distància entre el valor màxim i el valor mínim en una distribució de freqüències. És un paràmetre molt senzill, però pot ser rellevant en distribucions de freqüències en què no treballem de manera directa les dades sinó que ho fem directament sobre la matriu i es tracta d'una mostra molt gran. En estos casos pot ser interessant d'observar el rang, no només per a conèixer la dispersió

[14] També conegut com la màquina de Galton o la màquina de pèsols, perquè l'element amb què Galton provà la distribució aleatòria en una forma normal es va fer amb aquesta lleguminosa.

de la resposta, sinó també per observar els valors màxims i mínims. La fórmula del rang (Re) és molt senzilla:

$$Re = x_{max} - x_{min}$$

Recorregut i desviació interquartílica

El recorregut interquartílic és una altra de les mesures incorporades per Galton a finals del segle XIX. Es tracta d'una altra mesura de dispersió senzilla, que s'obté de la resta entre el tercer quartil i el primer quartil. Això ofereix, precisament, la mesura de la dispersió entre els quartils superior i inferior a la mediana, cosa que com veurem posteriorment, ens ofereix també la gràfica de caixes i bigots. La fórmula del recorregut interquartílic, per tant, seria la següent:

$$RQ = Q_3 - Q_1$$

De manera paral·lela, es pot calcular la desviació interquartílica, que ofereix el punt intermedi del recorregut interquartílic, que pot ser útil comparar amb la mediana per tal d'observar la diferència entre ambdues. La fórmula, per tant, seria la següent:

$$DQ = \frac{Q_3 - Q_1}{2}$$

El mateix concepte es pot aplicar als percentils, de manera que obtindríem una desviació interpercentílica a efectes d'estudi de la dispersió de la variable. Una de les mesures més utilitzades en aquest sentit és la que compara les posicions del P_{10} i el P_{90}, amb la qual cosa la fórmula quedaria de la següent manera:

$$DP = \frac{P_{90} - P_{10}}{2}$$

Variància

La variància, representada pels símbols S^2 o σ^2 (*sigma quadrat)* segons es tracte de distribucions mostrals o poblacionals, mesura la dispersió respecte a la mitjana aritmètica. De la mateixa manera que la mitjana, només es pot calcular en variables quantitatives. El càlcul de la variància es fa a partir de la mitjana dels quadrats de les desviacions

dels valors de la variable respecte de la seua mitjana. La quadratura de la sostracció, com explica La-Roca, es deu al fet que cal eliminar el problema de que la suma de les desviacions respecte de la mitjana, com s'ha pogut observar en l'apartat anterior és igual a zero i, per tant, cal controlar el signe, sense que això passe per l'aplicació de valors absoluts -que ofereix un mínim quan el punt respecte de la que es calcula és la mediana- mentre que la funció quadràtica ofereix un mínim quan s'aplica sobre la mitjana (La-Roca, 2006: p. 54). Per tant, això, quan es tracta d'una població, s'aplica sobre la variància poblacional σ_x^2, que es concreta en la següent fórmula:

$$\sigma_x^2 = \frac{\sum_{i=1}^{k}(x_i - \mu_x)^2 n_i}{n} = \sum_{i=1}^{k}(x_i - \mu_x)^2 fr_i$$

En el cas del càlcul del paràmetre per a una mostra, s'aplica la correcció de la quasivariància (La-Roca, 2006: p. 55), que es diferencia de la variància perquè en el denominador apareix la mida de la mostra menys una unitat $(\mathrm{n} - 1)$. Així la fórmula quedaria de la següent manera:

$$S_x^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{n - 1}$$

Una de les propietats de la variància és, com a resultat del fet que estiga configurada com una suma de quadrats, que el seu valor sempre és positiu o nul. Igualment, la seua interpretació es dificulta pel fet que es mesura en les unitats de la variable elevades al quadrat, amb la qual cosa no es pot dir que es tracte d'un paràmetre intuïtiu.

Desviació estàndard o desviació típica

El càlcul de la desviació típica ofereix un avantatge sobre el de la variància en tant que està mesurada en les mateixes unitats que la variable sobre la qual s'ha calculat la distribució de freqüències. Això s'aconsegueix mitjançant l'aplicació d'una arrel quadrada, que és la solució a la divergència incorporada en afegir el quadrat per tal d'evitar la consabuda suma zero de les desviacions respecte de la mitjana. Per tant, la fórmula de la desviació quedaria de la següent manera en el cas de la desviació típica poblacional:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \mu_x)^2 n_i}{n}} = \sqrt{\sum_{i=1}^{k}(x_i - \mu_x)^2 fr_i}$$

En el cas de la desviació típica mostral, com s'ha avançat, caldria treballar a partir de la quasivariància, amb la qual cosa la fórmula quedaria de la següent manera:

$$S_x = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{n-1}}$$

A efectes de dur a terme el càlcul de la desviació típica a mà, és convenient organitzar successives columnes a partir de les quals calcular la resta sobre el valor de la mitjana, elevar al quadrat, fer la multiplicació per les freqüències absolutes (si és sobre una població es pot fer directament sobre les freqüències relatives), la divisió pel denominador -en el cas de la quasivariància-, dur a terme el sumatori i, finalment, aplicar l'arrel quadrada de la variància resultant.

x_i	n_i	fr_i	na_i	fra_i	$x_i fr_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \frac{n_i}{n-1}$
22	3	0,3	3	0,3	6,6	-1,10	1,21	0,4033
23	4	0,4	7	0,7	9,2	-0,10	0,01	0,0044
24	2	0,2	9	0,9	4,8	0,90	0,81	0,1800
25	1	0,1	10	1	2,5	1,90	3,61	0,4011
$\sum$	10	1			$\bar{x}_i$= 23,1			$S_x^2 = 0{,}9889$
								$S_x = 0{,}9944$

Taula 15. Taula de freqüències amb càlcul de la quasivariància i la desviació típica mostral

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

En el cas de dades agregades, operaríem amb les marques de classe. A més, arribat el moment, potser ens interessa fer alguna comparació entre les desviacions típiques de dues distribucions de freqüències, amb diferents mitjanes i variàncies. La solució per a aquest tipus de comparacions passa per l'aplicació del coeficient de variació o coeficient de Pearson (CV_x), que posa en contacte la desviació i la mitjana, expressant la dispersió expressada com a proporció de la mitjana, amb un valor màxim d'1 (o també calculable com a tants per cent), i s'aplica de la mateixa manera en poblacions i en mostres, tal i com es pot observar amb la següent fórmula:

$$CV_x = \frac{\sigma_x}{\mu_x} = \frac{S_x}{\bar{x}_x}$$

D'altra banda, cap la possibilitat de calcular la desviació respecte de la mediana (D_{me}). En aquest cas el càlcul ofereix la dispersió en torn al paràmetre de centralitat. A diferència de la desviació típica, on calia passar primer pel càlcul de la variància per aplicar la correcció quadràtica, en la desviació de la mediana s'aplica una correcció mitjançant valors absoluts, amb la qual cosa queda compensada l'entrada en joc dels valors negatius. Així, la fórmula seria la següent:

$$D_{me} = \sum_{i=1}^{I} |x_i - Me| fr_i$$

Tipificació de variables

Una de les possibles transformacions de les variables d'una distribució és la tipificació[15], també coneguda com a estandarditzación o reducció. Una variable tipificada (Z) és aquella que presenta una mitjana de valor 0 i una desviació típica de valor 1, que com veurem posteriorment, és la que representa la distribució normal tipificada o estandarditzada. Qualsevol variable pot esdevindre tipificada mitjançant una operació per a la qual es necessita la mitjana i la desviació típica. Cada valor (x_i) de la distribució de freqüències serà tipificat per la mitjana de la distribució i també dividit per la desviació típica de la mateixa, com es pot observar a la següent fórmula:

$$Z = \frac{x_i - \bar{x}}{S_x}$$

Mitjançant la tipificació, podem conèixer la localització relativa de qualsevol valor en el supòsit que la distribució de freqüències analitzada s'aproxime a una distribució normal. A més, quan es tracta de diferents variables, la tipificació ens permet comparar la posició del valor respecte de la mitjana. El resultat de la tipificació s'ha d'interpretar com la dis-

[15] En l'àmbit anglosaxó s'utilitza l'expressió *standarization* o *normalization*, i per tant també standard error. Per contra, l'estadística francesa de la qual ha begut l'espanyola utilitza *écart type* i per tant, *typification* -tot i que també *normalisation*-, des d'on deduïm la tipificació. En aquest text utilitzarem majoritàriament l'expressió
tipiticació.

tància en desviacions típiques del valor (x_i) respecte de la mitjana. Posteriorment tractarem de les propietats de les unitats Z, donat que són fonamentals per al desenvolupament de l'apartat inferencial.

La desigualtat de Txebyshev

El matemàtic rus Pafnuty Txebyshev va donar nom a l'operació per la qual es mesuren les distàncies respecte de la mitjana. Com altres operacions que ja hem vist anteriorment, ja havia estat formulat anteriorment pel matemàtic Jules Bienaymé, per la qual cosa en ocasions la desigualtat apareix com a Txebyshev-Bienaymé. El càlcul de la desigualtat es fa a partir del percentatge de casos (P) que es troben a una quantitat (k) de desviacions típiques (σ_x) respecte de la mitjana (μ_x) que queda per baix de $\frac{1}{k^2}$. Dit d'una altra manera, la desigualtat de Txebyshev ens facilita trobar el percentatge d'observacions que esperem trobar entre dues desviacions típiques determinades a partir de la mitjana d'una distribució de freqüències. cosa que ve determinada per la següent fórmula:

$$P(|x - \mu_x| > k\sigma) \leq \frac{1}{k^2}$$

Aquesta desigualtat es basa en el fet que en qualsevol distribució de freqüències quantitativa les distàncies entre la mitjana i els seus valors dispersos es pot trobar a k desviacions típiques, traduïbles en percentatges de la distribució, com podem observar a la següent taula:

k		P
2	$P(\mu - 2\sigma \leq \bar{x} \leq \mu + 2\sigma)$	75%
3	$P(\mu - 3\sigma \leq \bar{x} \leq \mu + 3\sigma)$	89%
4	$P(\mu - 4\sigma \leq \bar{x} \leq \mu + 4\sigma)$	94%
5	$P(\mu - 5\sigma \leq \bar{x} \leq \mu + 5\sigma)$	96%
10	$P(\mu - 10\sigma \leq \bar{x} \leq \mu + 10\sigma)$	99%

Taula 16. Percentatges de dispersió segons k desviacions típiques

Font: Adaptat de Camarero et al. (2013: p. 95).

El càlcul de les distribucions es fa a partir de l'adaptació de la fórmula original, d'on extraguem que els valors de la nostra distribució estaran a $\pm k$ desviacions típiques de la mitjana μ, tot aplicant la següent fórmula:

$$\mu \pm k\sigma$$

2.2.2. Mesures de forma

A banda de les mesures de posició i dispersió, una altra de les aproximacions a les qualitats de les distribucions de freqüències consisteix a analitzar l'apartat visual. En alguns casos, estes aproximacions visuals són la base per a determinades modelitzacions estadístiques. Si bé algunes mesures de forma es poden deduir fàcilment per les seues representacions gràfiques, particularment els histogrames, sempre serà més fiable el càlcul de les mesures de forma, particularment la asimetria i la curtosi o apuntament.

Asimetria

La asimetria en estadística s'ha d'interpretar a partir d'un eix central vertical que està constituït per la mitjana, i de la distribució al seu voltant en forma de gràfiques de barres o histogrames. Val a dir que en investigació empírica és pràcticament impossible trobar una distribució de freqüències totalment simètrica. En aquest cas, es pot deduir que el 50% de la població es troba a una banda de la mitjana i l'altre 50% a l'altre costat, de manera que la mediana coincidiria amb la mitjana. En canvi, allò que solem trobar és una inclinació dels valors cap a l'esquerra o cap a la dreta. A partir de la fórmula de la asimetria es pot deduir la forma de la distribució: un valor de 0 indica una corba simètrica, mentre que valors majors a 0 indiquen asimetria a la dreta i valors menors de 0, asimetria a l'esquerra, sent 0 la representació gràfica de la mitjana. Un indicador major de 0,8, tant en positiu com en negatiu, indica una asimetria elevada. Seguint Glass i Stanley (1974) el coeficient d'asimetria de Fisher es calcula a partir de la següent fórmula:

$$Asimetria = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3/n}{s_x^3}$$

Curtosi

La curtosi és també un indicador de forma, en aquest cas, de l'apuntament o aplanament de la distribució de freqüències. Si en el cas anterior el model ideal era el que es distribuïa de manera simètrica -però també improbable- a partir del valor central de la mitjana, en l'estudi de la curtosi el model ideal a complir és el de la distribució normal. Novament, un ajustament total a la corba normal és molt improbable, mentre que és més probable trobar distribucions leptocúrtiques, és a dir, més apuntades que la normal o platicúrtiques, és a dir, més aplanades que la corba normal. La curtosi, per tant, és un indicador de la concentració dels valors en torn a la mitjana. Concentracions elevades indiquen un apuntament de la gràfica, mentre que valors reduïts indiquen una elevada dispersió. Un valor igual a 3 indica normalitat de la variable, mentre que valors per damunt de 3 indiquen concentració, i per baix de 3 dispersió. Segons l'obra de Glass i Stanley (1974) el coeficient de curtosi de Fisher es calcula mitjançant la següent fórmula:

$$Curtosi = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4/n}{s_x^4}$$

La mateixa fórmula es pot plantejar amb una sostracció de tres unitats, de manera que els resultats s'interpretarien a partir de 0, sent aquest valor el que representaria una corba amb forma normal; valors majors de 0 representarien una corba leptocúrtica o concentrada; i valors menors de 0 representarien una corba platicúrtica o dispersa (La-Roca, 2006: p. 64).

Asimetria i curtosi són dos indicadors que ens poden acostar a la consideració de la normalitat d'una distribució de freqüències. No obstant, a nivell operatiu, considerarem que per a mostres superiors a 30-40 individus no cal testar la normalitat de les variables per a aplicar procediments paramètrics, com veurem posteriorment (Ghasemi i Zahediasl, 2012).

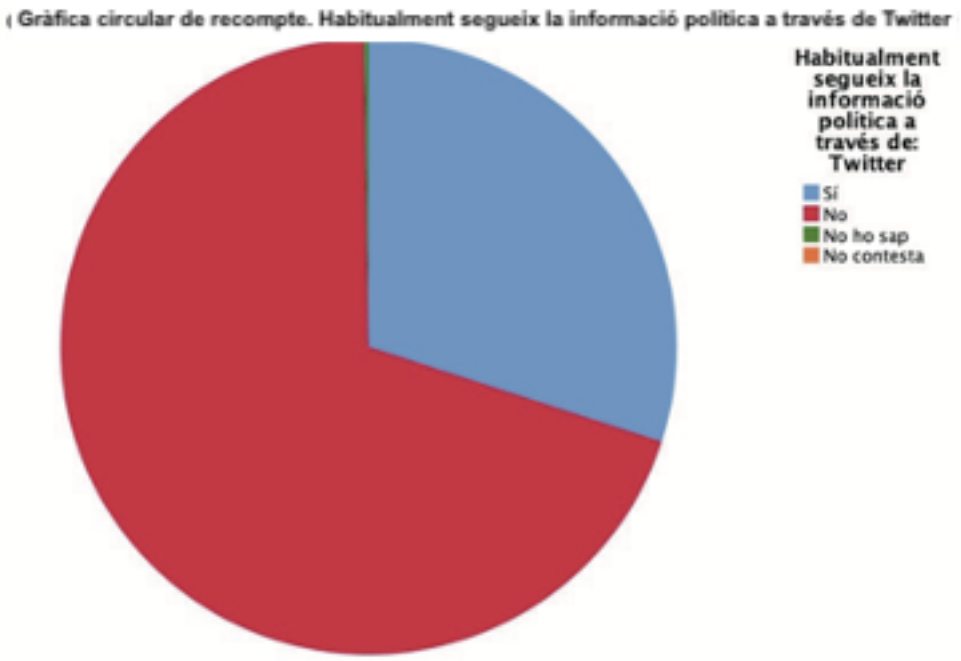
2.2.1. Gràfiques i figures

La informació que demanarem a cada variable està en funció de la profunditat que siga necessària en la nostra anàlisi. Allò més normal és demanar una taula de freqüències

d'una dimensió que, en ocasions, vindrà acompanyada d'alguna gràfica. Les que presentem ací es poden generar via el *Generador de gràfiques* de l'SPSS. L'SPSS permet extraure diferents tipus de gràfiques, tant d'una variable com de creuament de diferents variables. Entre les més habituals trobem els diagrames de barres, els histogrames -més adequats per a variables contínues-, les gràfiques de sectors o les gràfiques de caixes i bigots, molt útils per al creuament de variables contínues amb nominals.

L'SPSS permet extraure gràfiques de les variables que considerem oportú. Malgrat que alguns tipus de gràfiques estan condicionades a determinats nivells de mesura (nominal, ordinal o d'escala en la nomenclatura de l'SPSS), molt sovint comprovem que l'alumnat fa gràfiques sense cap ni peus. En aquest sentit, va molt bé recordar que en els casos de les variables nominals i ordinals es pot demanar al programari que extrega gràfiques de barres, sectors o línies, que són les més fàcils d'interpretar en termes univariants. En el cas de variables quantitatives, es poden demanar altres tipus de gràfiques que ajuden a la visualització de les dades, com ara els histogrames, les ogives, les gràfiques de caixes i bigots o *boxplot* o, en una combinació entre barres i histogrames sobre un eix doble, les piràmides de població.

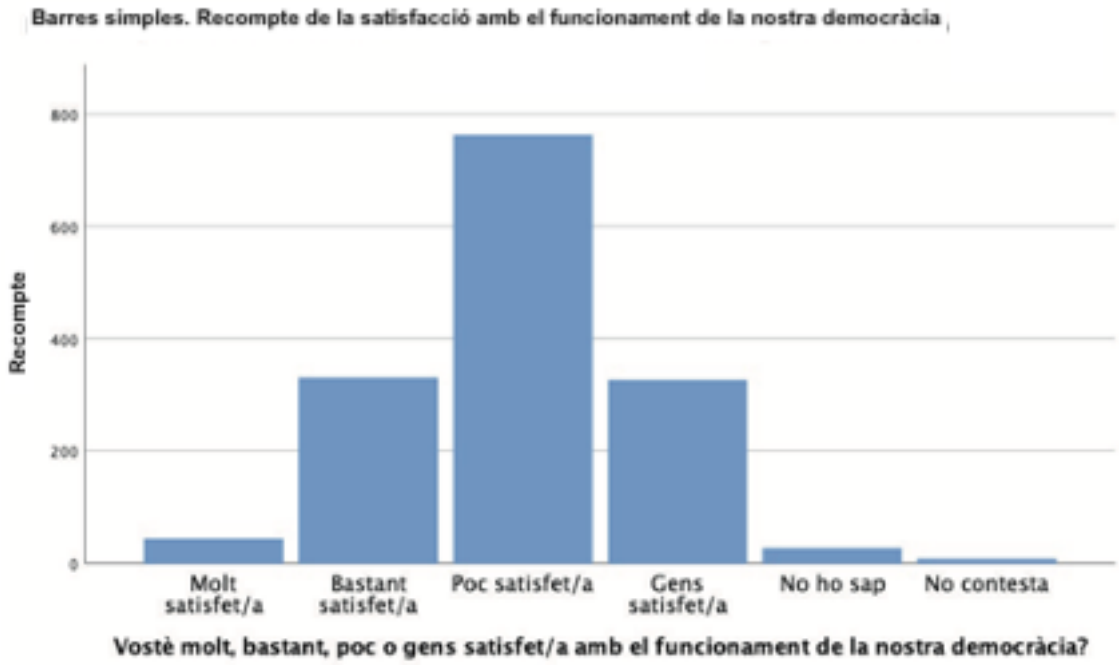
Les gràfiques que podem aplicar a les variables nominals i ordinals són també les més senzilles, donat que només es poden representar les freqüències de les categories que adopta la variable. En el cas de variables amb poques opcions de resposta, una solució bastant extesa és la dels diagrames de sectors, on amb una mirada es pot interpretar relativament ràpid els resultats d'una variable. Això sí, no és recomanable si la variable que volem analitzar té moltes opcions de resposta, perquè es torna molt complicat poder identificar cadascuna de les categories. Per a determinar l'angle que correspon a la freqüència d'una categoria ($\propto$), s'opera a partir de $fr_i * 360$.



Gràfica 5. Exemple d'un diagrama de sectors sobre el seguiment d'informació política mitjançant Twitter.

Font: Elaboració pròpia a partir del Baròmetre 942 del Centre d'Estudis d'Opinió (2019).

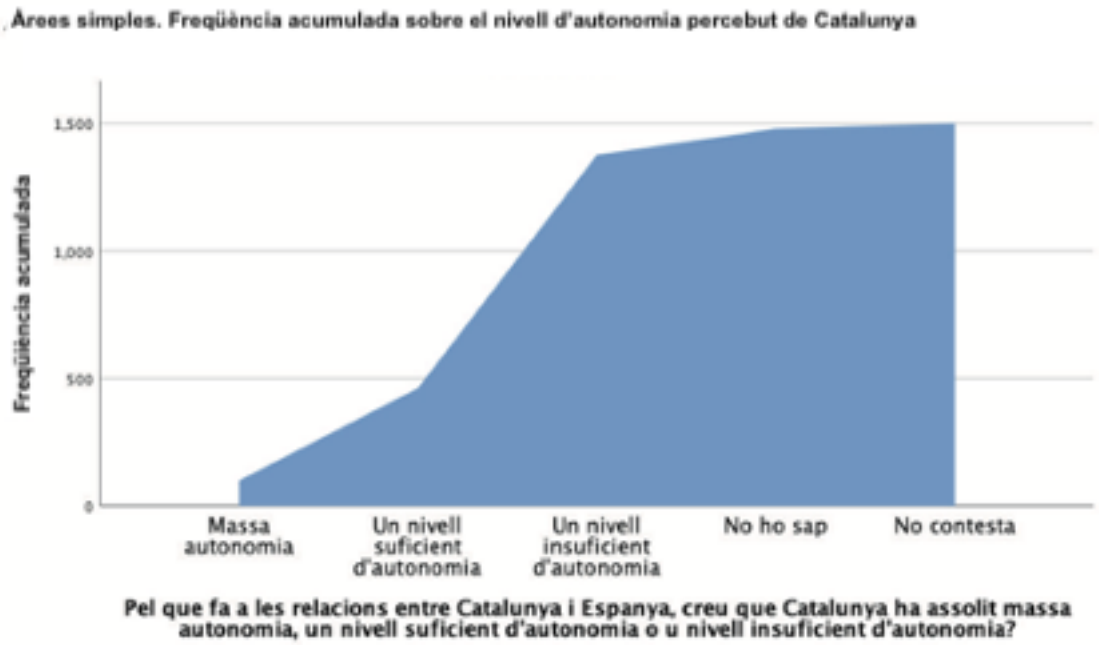
Les gràfiques de barres poden ser útils quan el número de categories de la variable és major que el que considerem per a un diagrama de sectors. Les gràfiques de barres es poden presentar en forma vertical o horitzontal, en termes absoluts o sobre una base 100, apilades, en tres dimensions, etcètera. No obstant, una presentació de tipus acadèmic sempre tendeix a ser més austera que la que podríem veure en altres suports no acadèmics, i això servei per a tot tipus de gràfiques: ni colors estridents, ni fons acolorits, ni decoració en excés. Com diu Juan Ignacio Martínez, l'ornat és pecat (2018: p. 273).



Gràfica 6. Exemple d'una gràfica de barres sobre el nivell de satisfacció amb el funcionament de la democràcia

Font: Elaboració pròpia a partir del Baròmetre 942 del Centre d'Estudis d'Opinió (2019).

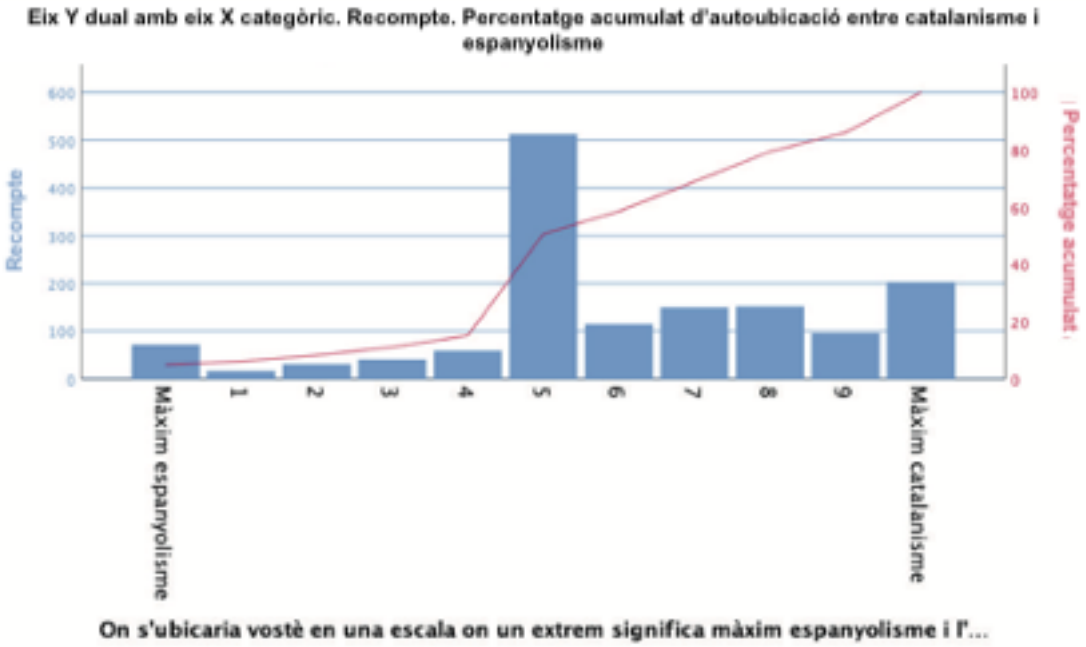
Combinant l'aspecte de les barres amb les gràfiques de línies, trobem la gràfica d'àrees, que també es poden representar amb freqüències o percentatges acumulats, de manera que permeten observar la importància de cada categoria en el resultat final.



Gràfica 7. Exemple d'una gràfica d'àrees sobre la percepció del nivell d'autonomia de Catalunya.

Font: Elaboració pròpia a partir del Baròmetre 942 del Centre d'Estudis d'Opinió (2019).

I combinant les gràfiques de barres i les de línies amb dos eixos diferents, sorgeix el diagrama de Pareto, que combina les freqüències absolutes amb els percentatges acumulats. El principi del diagrama de Pareto és el que també porta el seu nom: pocs vitals, molts trivials. O el que ve a ser el mateix, que el 80% dels problemes se sol resoldre amb el 20% de les causes. En definitiva, el diagrama de Pareto ens ajuda a distingir on està allò més important en una distribució de freqüències i percentatges acumulats.



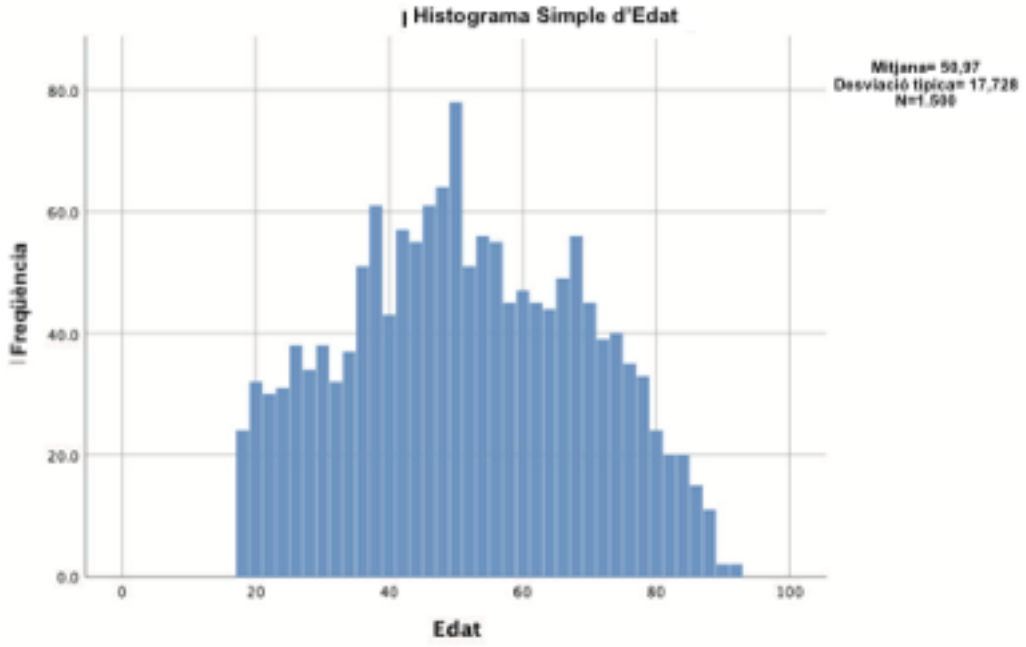
Gràfica 8. Exemple d'una gràfica de doble eix o Pareto sobre l'autoubicació entre espanyolisme i catalanisme

Font: Elaboració pròpia a partir del Baròmetre 942 del Centre d'Estudis d'Opinió (2019).

Pel que fa a les gràfiques per a variables quantitatives, cal tenir en compte la diferència entre estes variables i la resta, en tant que la distància entre categories es pot quantificar. Evidentment, cap la possibilitat d'utilitzar alguna de les gràfiques anteriors, però si, com sol ser el cas, les variables quantitatives tenen moltes categories, les opcions es limiten bastant. Una de les solucions a aquesta multitud de categories és l'agrupació prèvia a partir d'intervals de valors, que es pot fer de forma automatitzada amb l'SPSS. Per exemple, si disposem del dia, l'any i el mes de naixement d'una mostra de persones, abans de plantejar una gràfica acuradíssima i difícil de plantejar, podem transformar les

tres variables en una sola[16] que seria l'edat de la persona en el moment de contestar el qüestionari o la fitxa origen de la informació. Això simplifica bastant els càlculs.

En aquest sentit, l'histograma és un tipus de gràfica que conté la distribució d'una variable determinada a partir de la construcció de rectangles proporcionals a la freqüència de cada categoria. Els rectangles solen estar apegats i segons la seua altura i amplitud indiquen un número de casos. En el cas que els intervals siguen d'igual amplitud, la superfície del rectangle és equivalent a les freqüències, de manera que l'histograma seria equivalent a un diagrama de barres. En el cas de que no tots els intervals tinguen la mateixa mida, l'àrea del rectangle ha de representar les freqüències, de manera tal que ja no serà l'alçària allò que determine les freqüències, sinó la densitat, és a dir, la freqüència dividida per l'amplària de l'interval. Si se sumaren les freqüències de cada categoria etària, obtindríem la mida de la mostra, en absència de casos perduts. En el cas que es presenta en la següent gràfica, a més, l'SPSS ja ha extret tres indicadors bàsics que simplifiquen l'anàlisi: la mitjana, la desviació típica i la mida de la població analitzada.

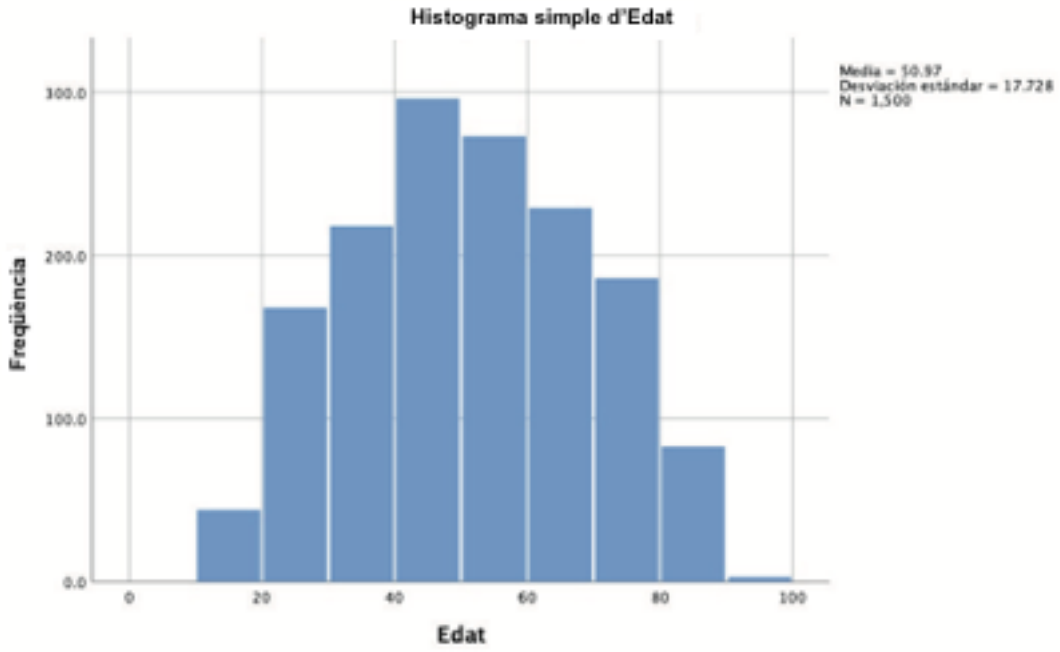


Gràfica 9. Exemple d'un histograma de la variable edat de la persona enquestada

Font: Elaboració pròpia a partir del Baròmetre 942 del Centre d'Estudis d'Opinió (2019).

[16] Això es faria, en el cas de l'SPSS, amb l'ajuda de l'assistent de data i hora, que amb unes poques instruccions ens retornaria una nova variable d'edat.

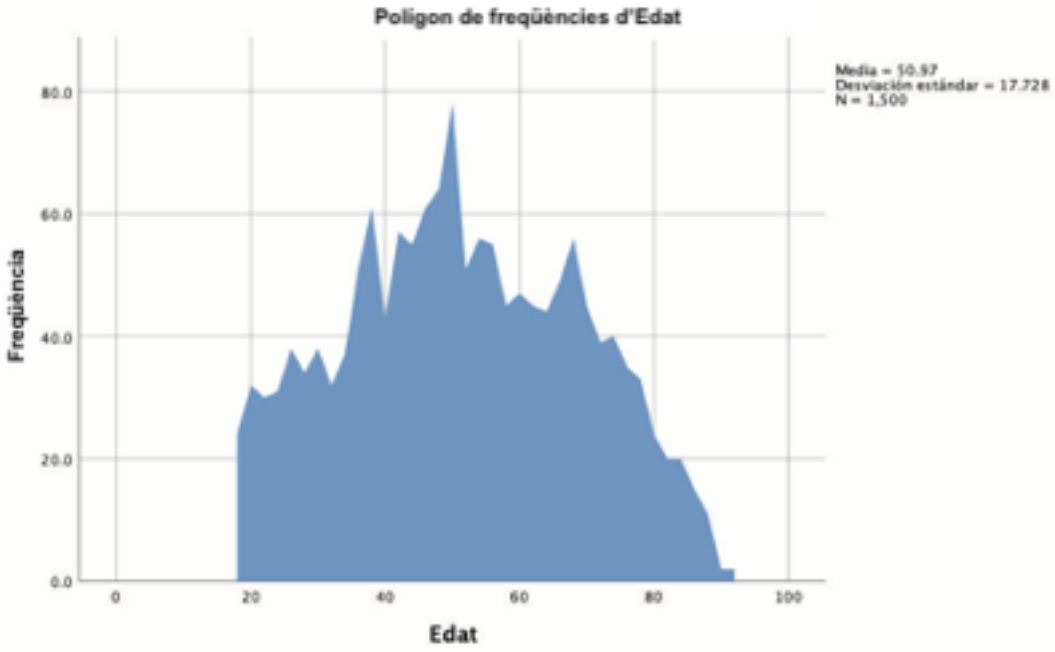
Val a dir que diferents agrupacions de les categories, per exemple, mitjançant una recodificació de la variable de cinc en cinc anys o de deu en deu, retornarien gràfiques amb una forma semblant, però amb un nivell de precisió menor, donada la pèrdua de detall que solen implicar estes transformacions., tal i com es pot observar a la següent gràfica, que representa les mateixes dades que l'anterior, però amb una agrupació etària de deu en deu anys. Cal tenir en compte que, com en l'anterior cas, de la suma de cadascuna de les categories d'edat resultaria la mida de la mostra. També s'ha de fer notar que el càlcul d'estadístics és el mateix, malgrat l'agrupació.



Gràfica 10. Exemple d'un histograma de la variable edat simplificada de la persona enquestada

Font: Elaboració pròpia a partir del Baròmetre 942 del Centre d'Estudis d'Opinió (2019).

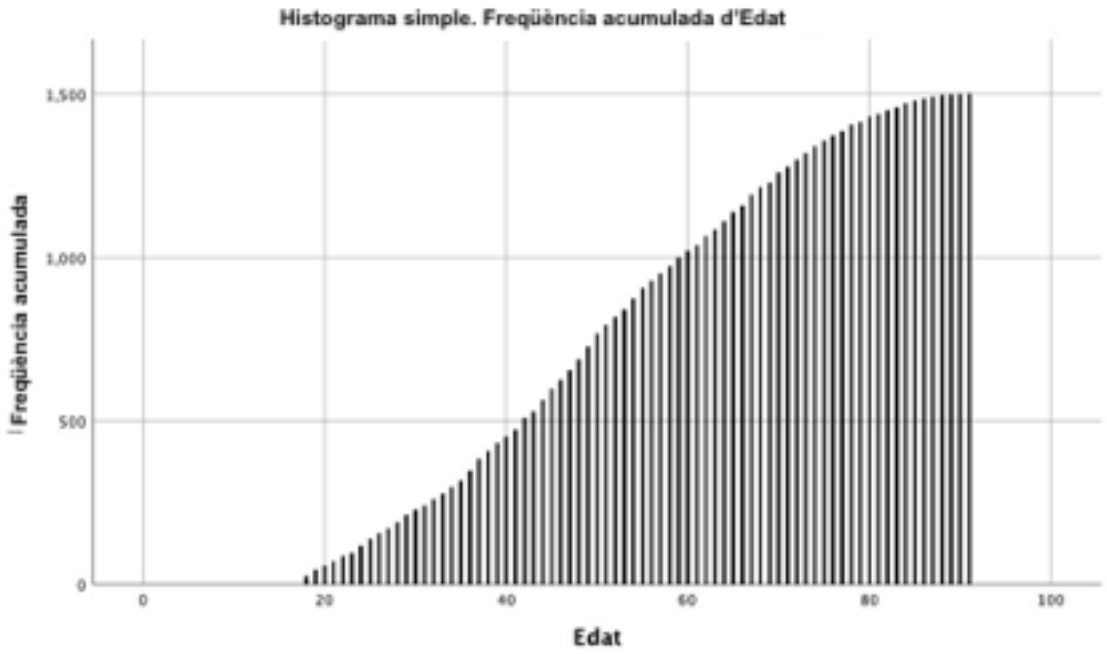
A partir de les dades d'un histograma es pot elaborar també un polígon de freqüències, que ve a ser com una gràfica d'àrees, però amb la unió de les línies dels límits superiors de cada categoria mitjançant una línia. Com en el cas anterior, es perd profunditat d'anàlisi, però la simplificació resultant pot ser positiva per a una anàlisi visual.



Gràfica 11. Exemple d'un polígon de freqüències de la variable edat de la persona enquestada

Font: Elaboració pròpia a partir del Baròmetre 942 del Centre d'Estudis d'Opinió (2019).

De la mateixa manera, un histograma de freqüències acumulades, en aquest cas substituint les barres per bigots, que resulten més fins quan ens trobem davant de tantes categories, ens oferiria una gràfica en forma d'ogiva. En aquest cas, s'arribaria al càlcul de la mida de la mostra en l'última categoria analitzada, que en la següent gràfica se situaria al voltant dels 90 anys i escaig.

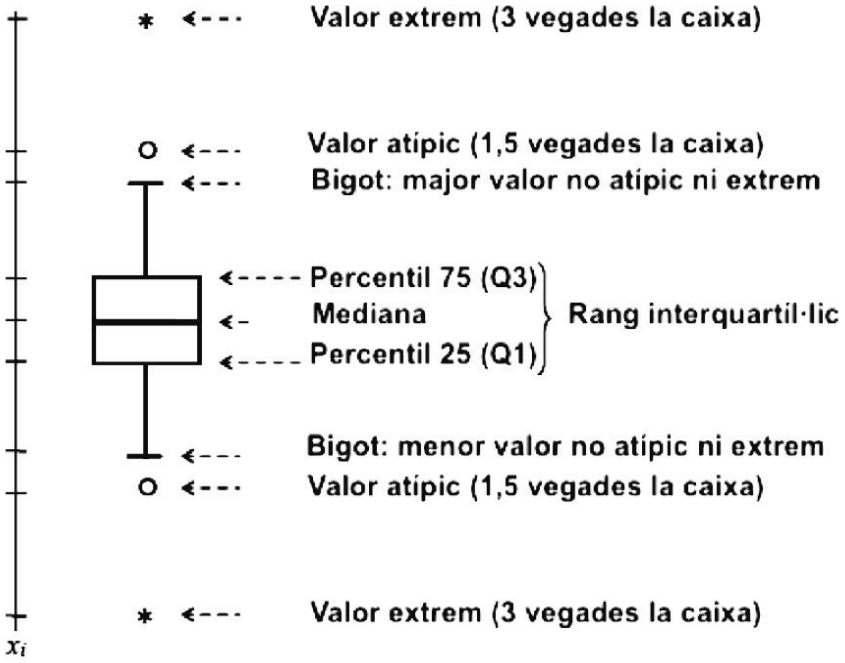


Gràfica 12. Exemple d'una piràmide de la població enquestada

Font: Elaboració pròpia a partir del Baròmetre 942 del Centre d'Estudis d'Opinió (2019).

Una de les gràfiques que més informació pot aportar sobre una variable quantitativa és el diagrama de caixes i bigots o *boxplot*. La informació, en aquest cas, és bàsicament al voltant de la dispersió i la centralitat d'una variable determinada. Una de les primeres mesures que salta a la vista és la de la mediana, més que res perquè se situa enmig de la caixa central. En funció de la curtosi i la simetria de la variable, la ratlla que representa la mediana estarà situada més cap al mig o desplaçada cap a un dels extrems, com tot seguit es podrà apreciar: si està més a prop del Q_1, és a dir, de la part de baix, indica concentració en valors baixos i a l'inrevés amb la Q_3. Els extrems de la caixa central són els indicadors dels quartils Q_1 i Q_3 o, dit d'una altra manera, dels percentils P_{25} i P_{75}. Així, caixes allargades significaran variables molt disperses, mentre que caixes d'altura reduïda implicaran variables molt concentrades en torn a la mediana (corbes més leptocúrtiques en la mesura de la curtosi) a com estan representats els límits dels quartil. Cal comprovar també la dispersió del que anomenem els bigots, que comprenen des de $Q_1 - (Q_3 - Q_1) * 1{,}5$ fins a $Q_3 + (Q_3 - Q_1) * 1{,}5$. Es pot dir, a nivell indicatiu, que entre els bigots trobem la major part de la variació de la variable en la categoria. Per últim, s'ha d'interpretar la presència de valors fora de la norma, que poden adoptar dos formats en SPSS: un cercle assenyala els *outliers*, com és el cas dels que trobem representats en la següent gràfica. En cas que aparega un asterisc, estaria indicant que el valor es troba més enllà de tres recorreguts interquartílics, és a dir, que seria el que anomenem un

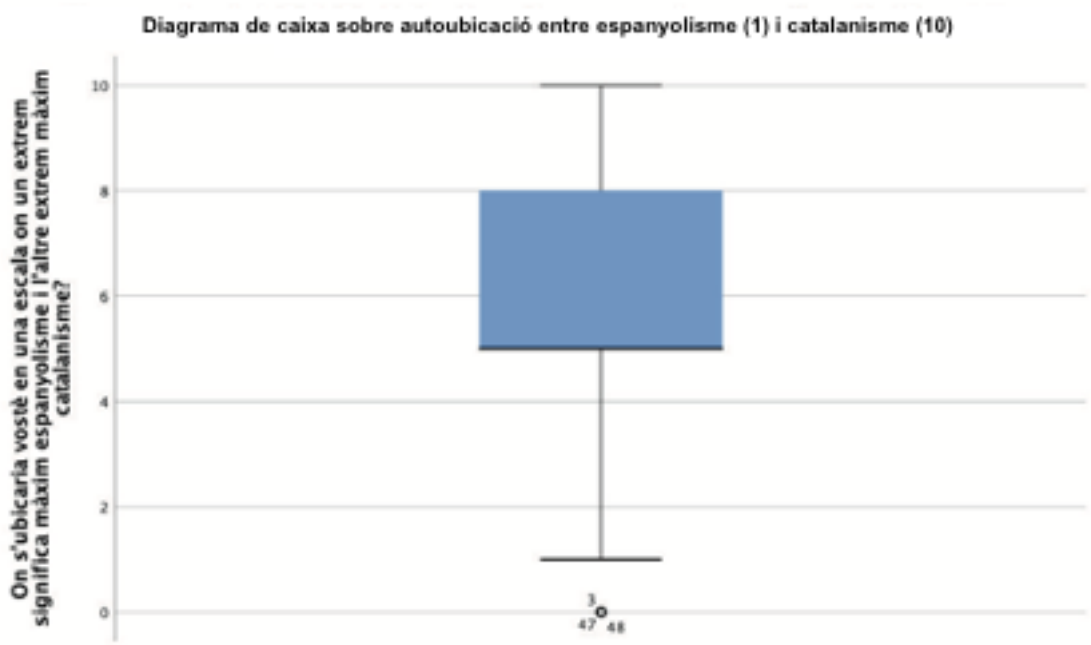
valor extrem. Els valors que apareixen al costat dels cercles o dels asteriscs assenyalen el número del cas, amb la qual cosa és fàcil tornar a la base de dades per comprovar qui són eixos casos en qüestió i avaluar si es tracta d'un error o d'un perfil determinat, que en tot cas es podria estudiar amb més deteniment.



Gràfica 13. Elements d'un diagrama de caixes i bigots.

Font: Adaptat de López-Roldan i Fachelli (2015).

En la següent gràfica podem observar el funcionament del diagrama de caixes i bigots per a una variable ordinal numèrica on el mínim, zero, és el màxim grau d'espanyolisme, mentre que el màxim, deu, és el màxim nivell de catalanisme. Es pot observar que la mediana està situada just a l'altura del 5, però a més coincideix amb el Q_1, la qual cosa vol dir que la proporció de gent que està concentrada entre Q_1 i Q_2 és molt gran respecte de la dispersió de la resta de les categories en la variable. O, dit d'una altra manera, que la major part de la mostra ha triat el punt central per a situar-se: igual d'espanyol que de català. D'altra banda, la caixa no és molt allargada, cosa que implica concentració entre els punts de Q_1 i Q_3, és a dir, entre el 5 i el 8. Això també ens dóna una idea del desplaçament dels casos cap a la dreta, o el que és el mateix, cap a posicions catalanistes. Els bigots se situen en l'1 i el 10, cosa que indica que la major part de la mostra se situa per baix d'1,5 vegades la caixa. De fet, només tres casos estan fora d'eixa mesura, i els tres es concentren en el valor 0, de màxim espanyolisme: els casos 3, 47 i 48.

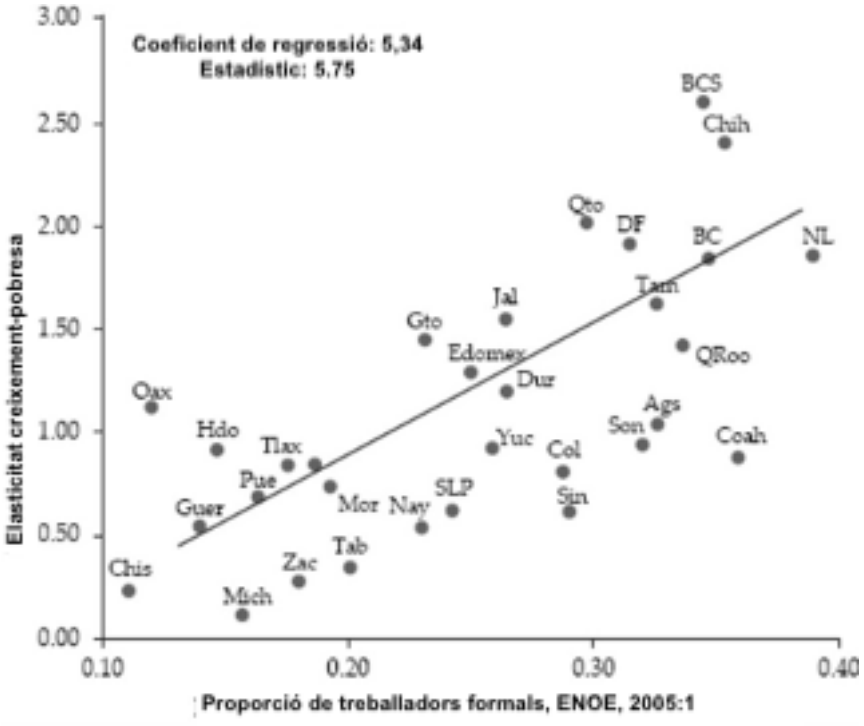


Gràfica 14. Exemple d'una piràmide de la població enquestada

Font: Elaboració pròpia a partir del Baròmetre 942 del Centre d'Estudis d'Opinió (2019).

Pel que fa a les gràfiques de creuament de variables, en destaquem de dos tipus: les de dispersió, molt útils per a variables d'interval, i les de caixes i bigots, útils per a creuar variables nominals o ordinals amb variables d'interval. La primera de les gràfiques no se sol utilitzar massa en investigació empírica, o almenys no en la major part de bases de dades de tipus socials, donada l'absència generalitzada d'indicadors de tipus interval per a creuar[17]. Més bé se sol aplicar a anàlisis conjuntes de països, a variables que han estat recalculades o a diferents tipus d'índexs agregats, purament quantitatius i numèricament continus. Aquest seria el cas de la següent gràfica, calculada a partir de dades estatals de Mèxic, i on es compara el càlcul de l'elasticitat entre creixement i pobresa, i la proporció de treballadors i treballadores formals. A partir d'aquesta gràfica, en el seu format linial, es a partir de la qual es calcula la regressió.

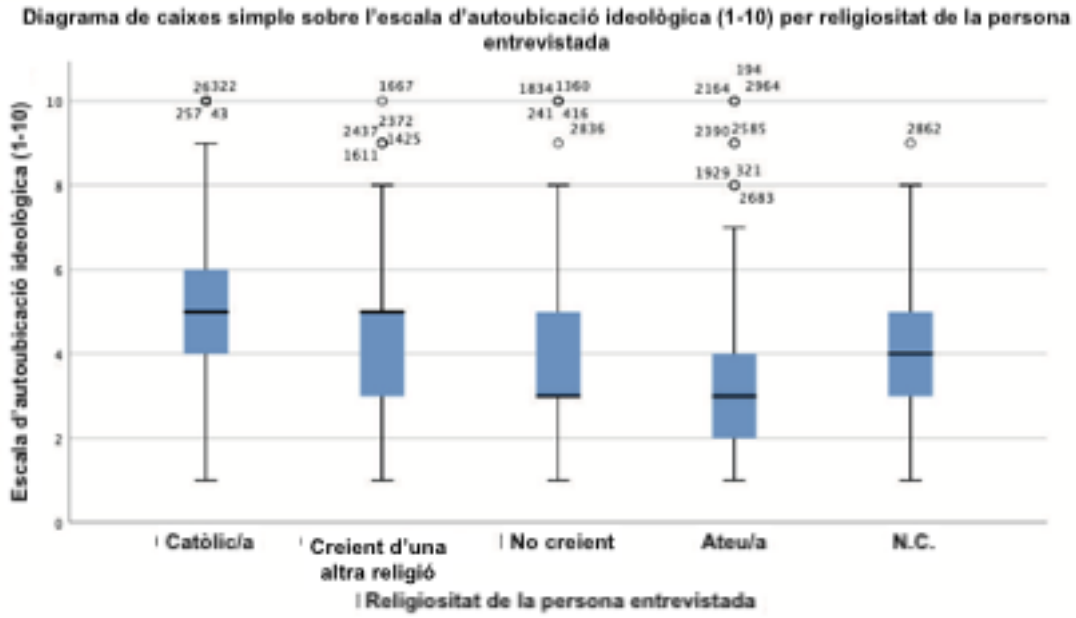
[17] Un dels indicadors que es podrien utilitzar per a creuar en anàlisi quantitativa, com és el salari, no se sol expressar en forma interval, que seria la més lògica i la que plantejaria més facilitat d'interpretació. Per contra, per a disminuir la no resposta, se sol mesurar amb una variable ordinal, amb la qual cosa es dificulta el càlcul d'indicadors i gràfiques com ara les de correlació.



Gràfica 15. Gràfica de correlació amb indicadors estadístics

Font: Adaptat de Campos i Monroy-Gómez-Franco (2016).

D'altra banda, una de les gràfiques més útils quan tractem amb variables qualitatives i quantitatives és la de caixes i bigots. En el seu format bivariable, la gràfica de caixes i bigots ens deixa contrastar, per a una variable qualitativa, quina és la distribució d'una variable quantitativa dins de cadascuna de les seues categories. Això ens ajuda a veure més clarament les relacions entre variables i també ens pot orientar en l'establiment de possibles hipòtesis a contrastar.



Gràfica 16. Gràfica de caixes i bigots de creuament entre autoubicació ideològica[18] i religiositat

Font: Elaboració pròpia a partir del baròmetre de gener de 2019 (número 3238)

D'altra banda, de la combinació de gràfiques de barres amb un doble eix es poden formular les piràmides de població, que no deixen de ser gràfiques bivariables on, en aquest cas, la variable quantitativa que és l'edat es mesura en freqüències de cadascuna de les categories. Amb una única gràfica podem conèixer com és l'estructura de la població atenent a la variable classificadora sexe i a les freqüències d'edat. En comptes de considerar edat i sexe es poden introduir altres variables que complisquen les mateixes característiques[19].

[18] Agafem l'autoubicació ideològica com a variable quantitativa, tot i que es tracta d'una ordinal numèrica, perquè una de les problemàtiques que trobem en baròmetres de tipus social és, precisament, la pràctica absència de variables d'interval. L'efecte de considerar-la com a variable d'escala en l'SPSS és el de pèrdua de profunditat, causada pel fet de no ser una variable contínua.

[19] Vegeu, per exemple, les piràmides d'edat i sexe al moment de dur a terme una migració de retorn (Giner, 2015).

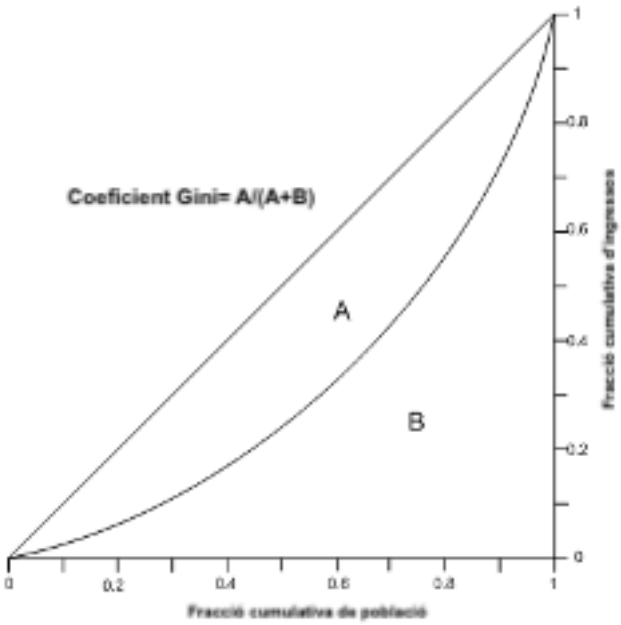
Gràfica 17. Exemple d'una piràmide de la població enquestada

Font: Elaboració pròpia a partir del Baròmetre 942 del Centre d'Estudis d'Opinió.

Corba de Lorenz i índex de Gini.

La corba de Lorenz és una representació gràfica en un diagrama cartesià on apareix l'acumulació de la freqüència ordenada d'una variable (p) respecte a la població (q), molt utilitzada per a representar la distribució de la riquesa, en concret amb l'aplicació d'índexs com el de Gini. La primera és obra d'un economista nordamericà, Max Lorenz, mentre que l'índex pren el nom del sociòleg italià Corrado Gini. Tant l'un com l'altre van nàixer a finals del segle XIX.

A la corba de Lorenz es pot llegir com un percentatge acumulat, que parteix del punt $(0,0)$ i acaba al punt $(1,1)$, que seria l'equivalent, en el cas que la gràfica tracte sobre la distribució de la riquesa, a la suma de la riquesa de totes les persones d'un determinat territori. La corba de Lorenz indica situacions de desigualtat respecte de la línia de 45º que assenyala la situació d'igualtat perfecta. Per tant, com més propera estiga la distribució a la diagonal, més igualitària serà la societat analitzada; i per contra, com més propera estiga la distribució a l'eix horitzontal i vertical dret, es detectaran més desigualtats a la societat analitzada (La-Roca, 2006: p. 139).



Gràfica 18. Model de corba de Lorenz amb el coeficient de Gini

Font: University of Denver (s.d.) Domestic Distribution of Household Income. Disponible a: https://www.du.edu/ifs/help/understand/economics/equations/sam/domdistribution.html

L'índex de Gini (I_G), aleshores, assenyala la ràtio compresa entre la corba de Lorenz i la bisectriu, i la bisectriu, a partir de la següent fórmula on P és la freqüència acumulada de la població i Q és la freqüència acumulada de la variable considerada, per exemple, els ingressos anuals.

$$I_G = \frac{\sum_{i=1}^{k-1}(P_i - Q_i)}{\sum_{i=1}^{k-1} P_i} = 1 - \frac{\sum_{i=1}^{k-1} Q_i}{\sum_{i=1}^{k-1} P_i}$$

Quan l'índex de Gini dóna com a resultat 1, assenyala desigualtat absoluta, mentre que en situacions on s'aproxima a 0 s'estima que hi ha una igualtat perfecta que, per tant, s'ajusta a la diagonal de la corba de Lorenz.

Any	Interval de renda	Freqüència	% Població	Marca ingressos	Pob*Ingr.	fri Pob*ingr.	frai pob. (P)	frai ingr*pob. (Q)	P-Q
2015	Menys de 1500	1332	3,60	750	999000	0,27	3,60	0,27	3,33
	1500-4500	4279	11,58	3000	12837000	3,51	15,19	3,78	11,40
	4500-7500	8845	23,94	6000	53070000	14,51	39,12	18,29	20,83
	7500-10000	8593	23,26	9000	77337000	21,14	62,38	39,43	22,95
	10500-13500	5605	15,17	12000	67260000	18,39	77,55	57,82	19,73
	13500-16500	3318	8,98	15000	49770000	13,61	86,53	71,42	15,10
	Més de 16500	4978	13,47	21000	104538000	28,58			
	Total	36950	100,00		365811000	100,00	284,37		93,35
								$\sum \frac{P-Q}{P}$	0,33

Gràfica 19. Exemple d'aplicació de l'índex de Gini en Excel

Font: Elaboració pròpia

De manera resumida, podríem considerar que cada tipus de gràfica es correspon de manera ideal amb un problema d'investigació en funció de quina siga la profunditat d'anàlisi. Així, les gràfiques de tipus descriptiu, exploratiu i comparatiu serien les de barres, àrees i sectors. Les gràfiques que donarien resposta a problemes d'investigació d'anàlisi i comparació de distribucions serien els histogrames, polígons de freqüències, ogives i diagrames de caixes. Els problemes d'investigació centrats en l'anàlisi de sèries temporals se centrarien en els diagrames de línies. I, per últim, la distribució conjunta de dues variables com a problema d'investigació es correspondria amb el diagrama de dispersió (Camarero et al. 2013: p. 128). Això, no obstant, no implica que un tipus de gràfica no puga ser utilitzat en un informe d'investigació si la seua finalitat última no és la que acabem d'enumerar. En tot cas, queda subjecte a la voluntat de la persona que coordine la investigació el fet d'utilitzar un tipus o un altre de gràfica.

2.3. Socioestadística descriptiva bidimensional

L'anàlisi descriptiva bidimensional afegeix a l'anàlisi unidimensional una segona dimensió o variable, de manera que, a banda de les anàlisis per separat, cap la possibilitat de dur a terme proves d'independència i causalitat. Sembla que un dels primers a detindre's en l'anàlisi de dues dimensions des del punt de vista social i no només abstracte va ser Francis Galton, que ja en 1885 avançava una primera anàlisi de correlació (reversió, si hem d'ajustar-nos al seu vocabulari) entre el pes d'uns pèsols i el de les llavors que havien produït, que més tard adaptaria, per la via de l'eugenèsia, a la descendència entre humans (Stigler, 1989). Les troballes de Galton estaven, al seu torn, inspirades en allò que anteriorment ja havien treballat el francés Auguste Bravais i l'escocés George Yule durant el segle XIX. En 1900 Karl Pearson publicà un article en què posava les bases de la correlació entre dues variables obtingudes per mostreig aleatori (Pearson, 1900). Quatre anys després, el mateix Pearson va posar nom a les taules de contingència. En aquell escrit, Pearson explica que el fonament sobre el qual basa l'expressió *contingència* és precisament que allibera l'analista de la necessitat de determinar les escales a l'hora de classificar els atributs estudiats, és a dir, que les escales són contingents (Pearson, 1904: p. 5).

Les taules de freqüència conjuntes, o taules de contingència, tenen el format estàndard següent, en el qual trobem distribuïdes en columnes les freqüències absolutes de la variable x_i i en files les de la variable y_j. Els extrems de les files són els sumatoris per columnes i per files, la qual cosa ens facilita l'observació dels valors marginals de la variable, independentment dels valors de l'altra variable. D'aquesta manera, per tant, podem tractar les variables unidimensionalment, cosa que facilita l'aplicació de tot allò que hem pogut revisar en l'apartat anterior. Per últim, la casella inferior dreta ha de sumar, tant per la via dels marginals, com de l'espai de les freqüències absolutes conjuntes, la N poblacional.

	y_1	y_2	...	y_j	...	y_q	N_x
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	$\sum n_{1j} = N_{1\cdot}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	$\sum n_{2j} = N_{2\cdot}$
...	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	$\sum n_{ij} = N_{i\cdot}$
...	...	...	...	...	...	...	...
x_p	n_{p1}	n_{p2}	...	n_{pj}	...	n_{pq}	$\sum n_{iq} = N_{q\cdot}$
N_y	$\sum n_{i1} = N_{\cdot 1}$	$\sum n_{i2} = N_{\cdot 2}$	...	$\sum n_{ij} = N_{\cdot j}$	...	$\sum n_{ip} = N_{\cdot p}$	N

Taula 17. Taula estàndard bidimensional de distribució de freqüències absolutes

Font: Elaboració pròpia

Val a dir que l'escala de mesura de les variables no és un impediment per a la seua anàlisi creuada. No obstant, treballar manualment amb variables amb molta amplitud, a més de resultar tediós, contribueix ben poc als avanços de l'alumnat. Així, agrupar les dades redueix les observacions i pot resultar més beneficiós per al càlcul, de manera que es concentre en allò que resulta important. De la matriu observada en l'apartat anterior, se'n podria deduir un encreuament entre les variables de l'escala de satisfacció amb l'aparença personal i els seguidors en Instagram. Després d'agrupar les dues variables, oferiria el resultat següent:

	]0-3]	]3-6]	]6-10]	n_x
]0-100]	1	1	2	4
]100-200]	0	1	2	3
]200-300]	0	1	2	3
n_y	1	3	6	10

Taula 18. Taula de distribució de freqüències absolutes de les variables seguidors d'Instagram (x_i) i satisfacció amb l'aparença personal (y_j)

Font: Font: Elaboració pròpia a partir del baròmetre del CIS 3201

La mateixa taula es pot expressar en freqüències relatives; en l'espai central hi hauria les freqüències relatives conjuntes, en els extrems les freqüències relatives marginals i

en l'extrem inferior dret hi hauria la suma 1 que, com a criteri de control, ha de resultar tant de les freqüències marginals d'x_i com de y_j.

	y_1	y_2	...	y_j	...	y_q	N_x
x_1	fr_{11}	fr_{12}	...	fr_{1j}	...	fr_{1q}	$\sum fr_{1j} = fr_{1\cdot}$
x_2	fr_{21}	fr_{22}	...	fr_{2j}	...	fr_{2q}	$\sum fr_{2j} = fr_{2\cdot}$
... ...		...	...	...	...	...	...
x_i	fr_{i1}	fr_{i2}	...	fr_{ij}	...	fr_{iq}	$\sum fr_{ij} = fr_{i\cdot}$
...	...	...	...	...	...	...	...
x_p	fr_{p1}	fr_{p2}	...	fr_{pj}	...	fr_{pq}	$\sum fr_{iq} = fr_{q\cdot}$
N_y	$\sum fr_{i1} = fr_{\cdot 1}$	$\sum fr_{i2} = fr_{\cdot 2}$	...	$\sum fr_{ij} = fr_{\cdot i}$	...	$\sum fr_{ip} = fr_{\cdot p}$	1

Taula 19. Taula estàndard bidimensional de distribució de freqüències relatives

Font: Elaboració pròpia

Si apliquem tot això a la taula que hem estat treballant anteriorment, tindrem la següent distribució de freqüències relatives conjuntes i marginals d'x_i i de y_j.

	]0-3]	]3-6]	]6-10]	n_x
]0-100]	0,1	0,1	0,2	0,4
]100-200]	0	0,1	0,2	0,3
]200-300]	0	0,1	0,2	0,3
n_y	0,1	0,3	0,6	1

Taula 20. Taula de distribució de freqüències relatives de les variables seguidors d'Instagram (x_i) i satisfacció amb l'aparença personal (y_j)

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

De manera semblant, podem atendre les freqüències relatives de la nostra distribució condicionades als valors d'una variable. Aquest seria el cas, atesa la distribució de la matriu sobre la qual estem treballant, de condicionar els valors dels seguidors en Instagram a l'escala de satisfacció amb l'aparença personal, o el que és el mateix, condicionar els valors de x_i als valors de y_j, amb la qual cosa obtenim les freqüències condicionades verticals, en què $fr_{i\cdot} = \frac{n_{i\cdot}}{N}$. En aquest cas, la taula estàndard seria la següent:

	y_1	y_2	...	y_j	...	y_q	N_x
x_1	fr_{11}	fr_{12}	...	fr_{1j}	...	fr_{1q}	$\sum fr_{1j} = fr_{1\cdot}$
x_2	fr_{21}	fr_{22}	...	fr_{2j}	...	fr_{2q}	$\sum fr_{2j} = fr_{2\cdot}$
... ...		...	...	...	...	...	...
x_i	fr_{i1}	fr_{i2}	...	fr_{ij}	...	fr_{iq}	$\sum fr_{ij} = fr_{i\cdot}$
...	...	...	...	...	...	...	...
x_p	fr_{p1}	fr_{p2}	...	fr_{pj}	...	fr_{pq}	$\sum fr_{iq} = fr_{q\cdot}$
N_y	1	1	...	1	...	1	1

Taula 21. Taula estàndard bidimensional de distribució de freqüències condicionades verticals

Font: Elaboració pròpia

Aplicant l'esquema a les dades que hem estat treballant, el resultat seria el següent:

	]0-3]	]3-6]	]6-10]	n_x
]0-100]	1	0,33	0,33	0,4
]100-200]	0	0,33	0,33	0,3
]200-300]	0	0,33	0,33	0,3
n_y	1	1	1	1

Taula 22. Taula de distribució de freqüències relatives condicionades verticals de les variables seguidors d'Instagram (x_i) i satisfacció amb l'aparença personal (y_j)

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

Contràriament, podem condicionar els valors de la variable y_j als de la variable x_i, amb la qual cosa tindríem la taula de valors condicionats horitzontals. La taula estàndard, aleshores, presentaria el format següent, en què $fr_{\cdot j} = \frac{n_{\cdot j}}{N}$:

		y_1	y_2	...	y_j	...	y_q	N_x
x_1		fr_{11}	fr_{12}	...	fr_{1j}	...	fr_{1q}	1
x_2		fr_{21}	fr_{22}	...	fr_{2j}	...	fr_{2q}	1
...	...		...	...	...	...	...	...
x_i		fr_{i1}	fr_{i2}	...	fr_{ij}	...	fr_{iq}	1
...		...	...	...	...	...	...	...
x_p		fr_{p1}	fr_{p2}	...	fr_{pj}	...	fr_{pq}	1
N_y		$\sum fr_{i1} = fr_{\cdot 1}$	$\sum fr_{i2} = fr_{\cdot 2}$	...	$\sum fr_{ij} = fr_{\cdot i}$	...	$\sum fr_{ip} = fr_{\cdot p}$	1

Taula 23. Taula estàndard bidimensional de distribució de freqüències condicionades horitzontals

Font: Elaboració pròpia

De la mateixa manera, la taula sobre la qual hem estat treballant presentaria l'aspecte següent, en aquest cas amb els acumulats a la banda dreta:

	]0-3]	]3-6]	]6-10]	n_x
]0-100]	0,25	0,25	0,5	1
]100-200]	0	0,33	0,66	1
]200-300]	0	0,33	0,66	1
n_y	0,1	0,3	0,6	1

Taula 24. Taula de distribució de freqüències relatives condicionades horitzontals de les variables seguidors d'Instagram (x_i) i satisfacció amb l'aparença personal (y_j)

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

També cap la possibilitat de limitar l'observació només a una part de la mostra, per exemple, aquelles persones que satisfan una condició determinada. Així, aquest nou conjunt es constituiria com una submostra del primer conjunt, que vindria a ser una mesura semblant a la condicionada, només que fixaríem la nostra atenció únicament en una part de les persones que contesten. Per exemple, si limitem les observacions a les persones que valoren positivament el seu aspecte, considerant una valoració positiva que aquesta estiga entre el 6 i el 10, obtindríem la següent distribució de freqüències de seguidors d'Instagram:

	]6-10]	fr_x
]0-100]	2	0,33
]100-200]	2	0,33
]200-300]	2	0,33
N	6	1

Taula 25. Taula de distribució de freqüències absolutes i relatives de la variable seguidors d'instagram (x_i) condicionada a una satisfacció amb l'aparença personal positiva [6-10] (y_j)

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

2.3.1. Independència i associació

A més de l'estudi de les freqüències de cadascuna de les variables implicades, una de les qüestions que cal analitzar és la relació que es dona entre aquestes. Dues variables seran independents si la freqüència relativa és igual al producte de les corresponents freqüències marginals relatives per a cada cel·la (La-Roca, 2006: p. 71), de manera que x_i i y_j seran independents si:

$$fr_{ij} = fr_{i.}fr_{.j} \forall_{i,j}$$

O també

$$n_{ij} = \frac{n_{i.}n_{.j}}{N}$$

En l'exemple següent es pot observar la qüestió de la independència a partir de dues taules on hi ha representades, en cadascuna, un grup social diferent al qual hem preguntat pel seu posicionament polític (conservador o progressista) i pel sexe de les persones que han ofert la seua resposta (home i dona).

Grup A	Cons.	Prog.	n_j
H	1	4	5
D	3	2	5
n_i	4	6	10

Grup B	Cons.	Prog.	n_j
H	2	3	5
D	2	3	5
n_i	4	6	10

Grup A	Cons.	Prog.	$f_{\cdot j}$
H	0,1	0,4	0,5
D	0,3	0,2	0,5
$f_{i\cdot}$	0,4	0,6	1

Grup B	Cons.	Prog.	$f_{\cdot j}$
H	0,2	0,3	0,5
D	0,2	0,3	0,5
$f_{i\cdot}$	0,4	0,6	1

Taula 26. Taules de freqüències absolutes i relatives de dos grups A i B sobre sexe (x_i) posicionament polític (y_j)

Font: Elaboració pròpia a partir de La-Roca (2006: p. 71)

A primera vista, si ens remetem als valors marginals, els grups A i B presenten una semblança absoluta. No obstant, hi ha diferències notables en els valors de les freqüències conjuntes: mentre que en el grup A les diferències entre homes i dones estan més pronunciades, en el grup B estan més suavitzades. Per tant, a primera vista es pot dir que en el grup A hi ha dependència, mentre que en el grup B les variables són independents.

En el grup A operaríem mitjançant la comparació dels valors conjunts i els marginals, és a dir, comparem el producte de $f_{\cdot j}$ i $f_{i\cdot}$ amb el valor conjunt f_{ij}, o el que és el mateix:

0,5*0,4= 0,2≠0,1
0,5*0,6= 0,3≠0,4

I així successivament

En canvi, en el grup B, els valors que obtindríem serien:

0,5*0,4= 0,2=0,2
0,5*0,6=0,3= 0,3

I així successivament

Fets els càlculs, però, encara no es pot afirmar quina relació hi existeix o amb quina direcció, perquè la dependència estadística és una relació conjunta i simètrica i no estableix relacions causals.

2.3.2. La relació entre variables a partir de la distribució de freqüències

A partir de les distribucions de freqüències podem conèixer quina relació hi ha entre les variables, és a dir, de quina manera els canvis en una variable, en aquest cas la variable independent, impliquen canvis en la variable dependent. La diferència entre observar i mesurar la independència o l'associació és que, en aquest estadi, es fa una mirada analítica a la distribució de freqüències.

Si agafem com a exemple la taula de freqüències relatives del grup A plantejada anteriorment, és fàcil d'observar que la variable independent, en aquest cas el sexe, deixa més homes progressistes i, per contra, més dones conservadores en la mostra estudiada. Així, es pot dir que la variable sexe determina en certa mesura el posicionament polític en el grup A.

Aquest acostament a les distribucions de freqüències és més interessant -i complicat- d'observar en presència de variables ordinals i d'interval, i moltes vegades exigeix de la persona investigadora una recodificació prèvia que agrupe les categories per tal que les relacions siguen més fàcilment observables. A l'efecte de presentació de les dades, és recomanable col·locar els valors de la variable independent en files, i la variable dependent en columnes, tot calculant-ne els percentatges en horitzontal (Almazán *et al*., 2015: p. 82).

Cal, però, detindre's bé a observar les distribucions de freqüències perquè pot donar-se allò que en estadística es coneix com la paradoxa de Simpson, anomenada així pel matemàtic que la va formular l'any 1951, Edward Simpson[20]. En l'encreuament entre dues variables categòriques hi poden estar obrant variables desconegudes i no controlades que fan que la interpretació d'una taula de freqüències siga completament la contrària de la que caldria esperar. Aquesta paradoxa ha de ser detectada abans de dur a terme qualsevol càlcul d'associació, perquè obeeix a qüestions lògiques més que no

[20] No obstant, sembla que la formulació prèvia de la paradoxa és obra de Karl Pearson, d'una banda, i Udny Yule, de l'altra, per la qual cosa en ocasions se la coneix com la paradoxa de Yule-Simpson.

matemàtiques. Un cas prototípic de la paradoxa de Simpson està explicat en el manual de Moore, Notz i Fligner a propòsit del nombre de víctimes d'accidents de trànsit que moren o sobreviuen en trasllats a l'hospital en helicòpter i en ambulància (Moore, Notz i Flinger, 2018: p. 336). La distribució original de la taula és la següent:

	Helicòpter	Ambulància
Moren	64	260
Sobreviuen	136	840

Taula 27. Taula de freqüències absolutes de la supervivència d'accidentats en transports a l'hospital amb helicòpter i ambulància

Font: Elaboració pròpia a partir de Moore, Notz i Flinger (2018: p. 336)

Quan s'analitzen les freqüències relatives per columnes, es pot comprovar que el resultat dels trasllats en helicòpter és d'una mortalitat superior, 32% dels trasllats contra el 24% dels trasllats en ambulància.

	Helicòpter	Ambulància
Moren	0,32	0,24
Sobreviuen	0,68	0,76

Taula 28. Taula de freqüències relatives de la supervivència d'accidentats en transports a l'hospital amb helicòpter i ambulància

Font: Elaboració pròpia a partir de Moore, Notz i Flinger (2018: p. 336)

No obstant, quan introduïm en aquesta relació entre variables una tercera de control, la gravetat de l'accident, es pot comprovar que les dades canvien, cosa que es perfila millor en la taula de freqüències relatives que segueix.

Greus	Helicòpter	Ambulància
Moren	48	60
Sobreviuen	52	40

Lleus	Helicòpter	Ambulància
Moren	16	200
Sobreviuen	84	800

Taula 29. Taula de freqüències absolutes de la supervivència d'accidentats en transports a l'hospital amb helicòpter i ambulància, controlades per la gravetat de l'accident

Font: Elaboració pròpia a partir de Moore, Notz i Flinger (2018: p. 337)

L'observació de les freqüències relatives mostra que l'helicòpter salva més vides que el trasllat en ambulància, concretament, el 52% dels trasllats greus i el 84% dels trasllats lleus. En canvi, en la primera taula hem vist que la diferència era notable en el nombre de morts per als trasllats en helicòpter. L'explicació rau en la quantitat de persones que

hi ha en cadascuna de les dues submostres: en el cas dels helicòpters, la meitat de la mostra prové d'accidents greus, mentre que en el cas de les ambulàncies, només suposa un 10% del total de la mostra. Per tant, poder controlar aquest tipus de variables és de suma importància per tal de poder analitzar bé la relació entre dues variables.

Greus	Helicòpter	Ambulància
Moren	0,48	0,60
Sobreviuen	0,52	0,40

Lleus	Helicòpter	Ambulància
H	0,16	0,20
D	0,84	0,80

Taula 30. Taula de freqüències relatives de la supervivència d'accidentats en transports a l'hospital amb helicòpter i ambulància, controlades per la gravetat de l'accident

Font: Elaboració pròpia a partir de Moore, Notz i Flinger (2018: p. 337)

Per tal d'aprofundir més en l'anàlisi conjunta de dues variables serà necessari diferenciar l'escala de mesura de les variables. Si es tracta de variables d'interval, la mesura de l'associació ha de passar pel càlcul de la covariància, la correlació lineal i la regressió. Si, per contra, es tracta de variables ordinals, calcularem el coeficient de correlació de Spearman o Gamma. Per últim, si les nostres dades són nominals, l'associació es mesurarà mitjançant khi quadrat o Alfa i alguna de les mesures associades. Vegem tot seguit, desglossades pel tipus de variable, les principals operacions.

2.3.3. Associació en variables d'interval: covariància, correlació lineal i regressió

Covariància

En el cas que les variables que intentem analitzar siguen d'interval, el primer pas per a fer el càlcul de la correlació lineal de Pearson és el càlcul de la covariància. Aquest paràmetre tracta de mesurar la posició relativa de les observacions bivariants respecte de les mitjanes aritmètiques de les dues variables (La-Roca, 2006: p. 73). És per això que només s'hauria de calcular quan ens trobem davant d'encreuaments de dues variables d'interval, atès que el càlcul de la mitjana és privatiu d'aquest tipus de variable. La representació gràfica és σ_{xy} en el cas de mesures poblacionals i S_{xy} en el cas de mostres, i la fórmula es pot definir com la mitjana dels productes encreuats de les desviacions respecte de les mitjanes de cada parell d'observacions:

$$S_{xy} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}(x_i - \bar{x})(x_j - \bar{y})n_{ij}}{N} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}x_i y_j n_{ij}}{N} - \left(\frac{\sum_{i=1}^{h}\sum_{j=1}^{k}x_i n_{ij}}{N}\right)\left(\frac{\sum_{i=1}^{h}\sum_{j=1}^{k}y_j n_{ij}}{N}\right)$$

O el que és el mateix, si descomponem la fórmula, la covariància S_{xy} és igual a la sostracció de la mitjana del producte i el producte de les mitjanes:

$$S_{xy} = \overline{xy} - \bar{x}\bar{y}$$

El càlcul de la primera part de l'equació es produeix del sumatori del producte de cada freqüència relativa conjunta per les seues categories o marques de classe corresponents (x_i o y_j), és a dir, de la mitjana del producte:

$$\overline{xy} = \sum_{i=1}^{h}\sum_{j=1}^{k}x_i y_i f r_{ij}$$

De l'aplicació d'aquesta fórmula i de la sostracció del producte de les mitjanes, que s'obtenen d'acord amb el que hem vist en el primer apartat, obtindrem el valor de la covariància, que cal interpretar de la manera següent: valors positius indiquen una relació directa de les variables; valors negatius indiquen una relació inversa de les variables; i per últim, valors propers a zero estarien assenyalant l'absència d'una relació clara entre les variables. Si traslladem les observacions d'ambdues variables a una gràfica de dispersió, serà fàcil observar que la recta que dibuixen els encreuaments de les categories és ascendent en el cas de la relació directa; descendent en el cas de la relació inversa i sense una forma definida de relació en el cas que el valor de la covariància s'aproxime a zero. A diferència del valor de la correlació, que com veurem a continuació sí que està definit amb un màxim i un mínim, el valor de la covariància no té un màxim i un mínim, per la qual cosa no podem deduir la intensitat de la relació entre les variables, només la direcció de la seua relació.

Considerem ara les dades provinents de la matriu que hem presentat anteriorment, i concretament les que fan referència a l'aparença personal i els seguidors en Instagram. Tot i que després de l'agrupament no constitueixen exemples de variables d'interval, serviran com a exemple i, d'altra banda, la recodificació ens serà funcional des del punt de vista de l'espai.

	]0-3]	]3-6]	]6-10]	n_x
]0-100]	1	1	2	4
]100-200]	0	1	2	3
]200-300]	0	1	2	3
n_y	1	3	6	10

	]0-3]	]3-6]	]6-10]	fr_x
]0-100]	0,1	0,1	0,2	0,4
]100-200]	0	0,1	0,2	0,3
]200-300]	0	0,1	0,2	0,3
fr_y	0,1	0,3	0,6	

x_i \ y_j	1,5	5	8,5	fr_x	$fr_x x_i$	$\frac{(x_i-\bar{x})^2 n_i}{N-1}$
50	0,1	0,1	0,2	0,4	20	3600
150	0	0,1	0,2	0,3	45	33,3333
250	0	0,1	0,2	0,3	75	4033,3333
fr_y	0,1	0,3	0,6	10	$\bar{x} = 140$	$S_x^2 = 7666{,}6666$
$fr_y y_j$	0,15	1,5	5,1	$\bar{y} = 6{,}75$		$S_y = 87{,}5595$
$\frac{(y_j-\bar{y})^2 n_j}{N-1}$	3,0625	1,0208	2,0417	$S_y^2 = 6{,}125$		
				$S_y = 2{,}4749$		

$x_i y_j fr_{ij}$	1,5	5	8,5	
50	7,5	25	85	
150	0	75	255	
250	0	125	425	

$\overline{xy} = 997{,}5$

$S_{xy} = 999{,}5 - (140 * 6{,}75)$

$S_{xy} = 999{,}5 - 945 = 54{,}5$

Taula 31. Càlcul de la covariància per a les variables seguidors d'Instagram (x_i) i satisfacció amb l'aparença personal (y_j)

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

Dels càlculs anteriors deduïm que la relació entre les variables x_i i y_j és directa, tal com mostra el signe de la covariància S_{xy}, amb la qual cosa podem dir que un major nombre de seguidors d'Instagram implica una major valoració personal, sense que puguem establir-ne una relació de causalitat. Posteriorment, caldrà valorar quina és la intensitat d'aquesta relació per tal de poder afegir més arguments a l'associació que ens ha mostrat la covariància.

Correlació lineal

Arribats a aquest punt, cal valorar la intensitat de la relació que ha mesurat la covariància, cosa que no podem fer amb l'ús exclusiu d'aquesta mesura atès que no disposem d'uns límits clars que ens ajuden a fer-ho. Com hem vist anteriorment, les bases teòriques de la correlació van ser establertes per Francis Galton que, a més a més, ja va batejar el nou paràmetre, conegut com r, tot i que anteriorment Auguste Bravais ja n'havia deduït la fórmula. Va haver de ser Karl Pearson qui ho desenvolupara posteriorment utilitzant el llenguatge matemàtic, gràcies a les contribucions del seu equip, en particular de Raphael Weldon (Desrosières, 1998: p. 131). Precisament, el mateix Pearson dona la solució per a dues de les tres mesures d'associació més habituals en l'àmbit descriptiu (coeficient de correlació i coeficient de contingència), i totes dues tenen l'origen en la mateixa publicació (Pearson, 1896).

El coeficient de correlació de Pearson es basa en el quocient entre la covariància i el producte de les desviacions típiques. Cal precisar que en la formulació original de Pearson aquest coeficient es basa en el fet que les dues variables han sigut extretes mitjançant un procés d'aleatorització i presenten una forma normal, cosa que serà important especialment en l'àmbit inferencial. A més, la correlació implica la mesura de l'ajust de les dades resultants de la correlació a una recta, per això l'adjectiu de la correlació és *lineal* i, també per això es dedueix que poden existir diferents ajustos de la correlació que no necessàriament han de ser lineals. Se simbolitzen amb la lletra r_{xy} les mostres i amb la lletra ρ_{xy} (*rho*) les poblacions. Així doncs, la fórmula simplificada del coeficient de Pearson és:

$$r = \frac{S_{xy}}{S_x S_y}$$

D'aquest coeficient sorgirà un número que tindrà el mateix signe que la covariància, ja que el valor del denominador sempre serà positiu, de la mateixa manera que ho són les desviacions típiques implicades en el quocient. Així, també es manté la interpretació de la direcció de l'associació observada en el càlcul de la covariància en el qual, com acabem de veure, valors negatius impliquen relacions inverses, valors positius impliquen relacions directes i valors propers a zero impliquen absència de relació. En el cas del coeficient de correlació, a més, el resultat se situa entre -1 i 1, de manera que la seua interpretació, a diferència de la covariància, sí que se situa entre dos límits, -1 i 1. Així,

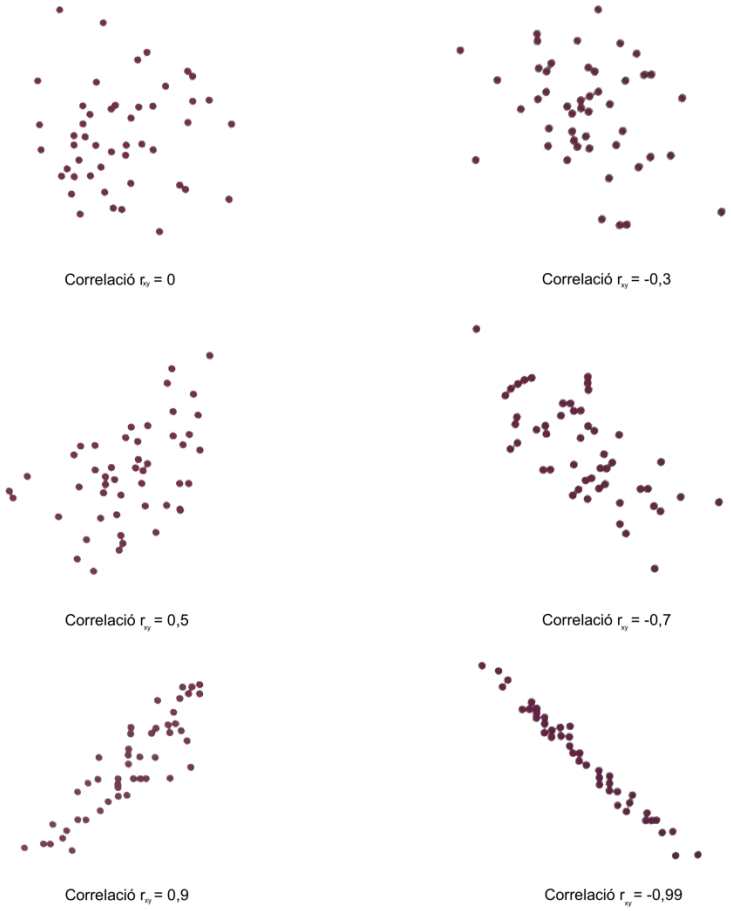
a més de la direcció de l'associació se'n pot interpretar la intensitat, com veurem tot seguit:

- $r_{xy} < 0$: relació inversa
- $r_{xy} > 0$: relació directa
- $r_{xy} = 0$: incorrelació o relació absent

El fet que el coeficient de correlació tinga un límit clar, en aquest cas en la unitat, fa possible que puguem conèixer la intensitat de l'associació entre variables. Així, valors de $|r_{xy}|$ propers a zero indiquen una incorrelació, mentre que valors de $|r_{xy}|$ propers a la unitat assenyalen una correlació perfecta. La situació més habitual, però, és que els valors de $|r_{xy}|$ se situen en una àrea intermèdia en què, com més proper siga el valor a la unitat, major intensitat de l'associació. D'aquesta manera, l'associació serà més intensa quan el valor estiga més proper a 1. Per tal de donar una aproximació a la intensitat de l'associació, es poden seguir les indicacions següents:

- $0 < |r_{xy}| \leq 0{,}2$: associació molt feble
- $0{,}2 \leq |r_{xy}| \leq 0{,}4$: associació feble
- $0{,}4 \leq |r_{xy}| \leq 0{,}6$: associació moderada
- $0{,}6 \leq |r_{xy}| \leq 0{,}8$: associació forta
- $0{,}8 \leq |r_{xy}| \leq 1$: associació molt forta

Visualment, la correlació lineal implica l'ajust dels punts de confluència de les freqüències de les variables x i y de manera que, quan estes coincidències es mostren en un plànol bivariant en forma de gràfica de dispersió, se'n pot deduir el signe i el valor atenent a dues qüestions fonamentals: la primera és la presència o no d'una inclinació en els punts que formen la gràfica. Com més evident siga la presència d'aquesta inclinació, més clar estarà el signe, tant en la correlació com en la covariància. La segona és la força de la correlació respecte de la línia, que es pot observar a simple vista per la distància de cadascuna de les confluències en el plànol respecte de la línia imaginària. És a dir, com més allunyats estiguen els punts de la línia imaginària que marca la relació lineal, més xicotet serà el valor de la correlació. En canvi, com més propers estiguen els valors de la línia imaginària de la correlació, més a prop de la unitat estarà el valor de r_{xy}. Tot això ho podem veure en la gràfica següent, en la qual apareixen exemples de correlacions properes a zero, zero, properes a la unitat i amb diferents signes.



Gràfica 20. Exemples de correlació lineal

Font: Elaboració pròpia a partir de Moore, (2007: p. 102)

Cal tenir en compte algunes consideracions prèvies a l'hora de plantejar una anàlisi de correlació (Moore, 2007: p. 103):

1. L'anàlisi d'una correlació implica que les dues variables són quantitatives[21].
2. La correlació lineal es limita a la relació lineal, i no a altres tipus de relació - curvilínia, per exemple.
3. La correlació és molt sensible als valors atípics, per la qual cosa el valor de r_{xy} pot veure's afectat quan hi ha valors atípics en la distribució de freqüències.

[21] No obstant, és possible incloure en les equacions de regressió allò que es coneix com a variables *dummy*, o falses variables quantitatives, constituïdes per variables nominals dicotòmiques que varien com a 0 i 1 i que es poden incorporar al càlcul de la regressió en SPSS. Vegeu, al respecte, el cas pràctic que explica Healey (2016: p. 351).

4. La correlació per si mateixa no explica les dues variables implicades. Cal més informació respecte de cadascuna de les variables, com per exemple les mitjanes i la desviació típica de cadascuna. Fora bo disposar també dels diagrames de caixa de cadascuna de les variables per tal de valorar la presència de valors atípics i, si s'escau, el seu control.

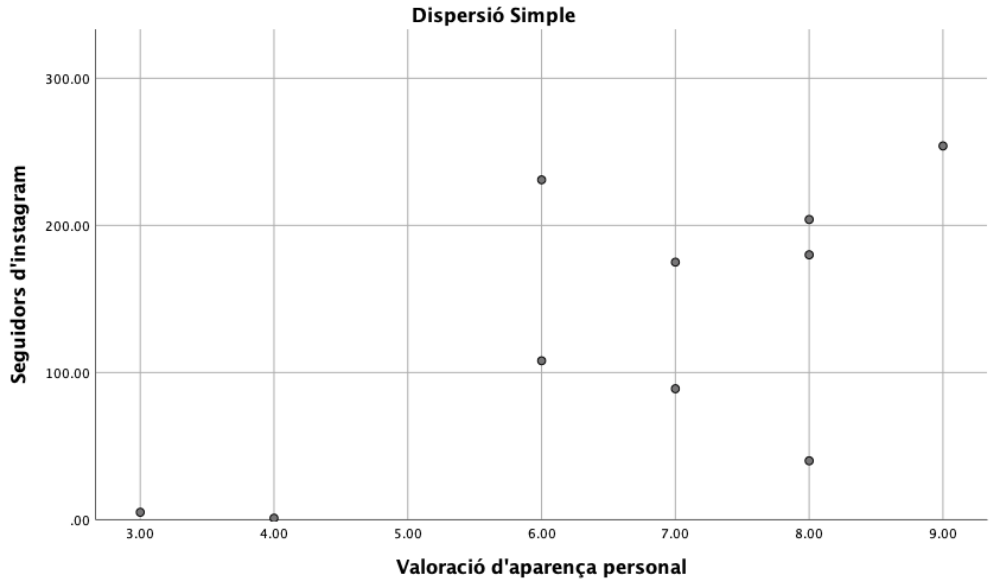
Així, per a la covariància estudiada anteriorment S_{xy} i les desviacions típiques S_x i S_y, l'indicador de correlació seria el següent. Per tant, podríem dir que la relació entre ambdues variables és feble:

$$r_{xy} = \frac{54{,}5}{87{,}5595 * 2{,}4749} = \frac{54{,}5}{216{,}7} = 0{,}251$$

Taula 32. Càlcul de la correlació per a les variables seguidors d'Instagram (x_i) i satisfacció amb l'aparença personal (y_j)

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

Una correlació r_{xy} com la que acabem de calcular implica, com en el cas de la covariància (ja hem vist que han de mantindre el mateix signe), una relació directa. A més, allò que indica la correlació és la intensitat de l'associació, que en el cas que ens ocupa seria feble, si tenim en compte que se situa entre el 0,2 i el 0,4. La gràfica de dispersió de les dades de la matriu original, tal com hem vist en l'apartat de gràfiques, seria la següent:



Gràfica 21. Gràfica de dispersió de les variables seguidors d'Instagram i valoració de l'aparença personal

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

Regressió

L'anàlisi de la correlació no té encara la sistematització d'un model que explique la relació entre variables quantitatives des del punt de vista matemàtic, en aquest cas reduït a la funció lineal. L'aplicació inferencial de la regressió, però, s'observarà posteriorment quan siga analitzada des del context probabilístic i del contrast d'hipòtesis.

L'ajust de les dades a una recta és una operació que es reprodueix almenys des que el 1805 el matemàtic francès Adrien-Marie Legendre va aplicar el mètode dels mínims quadrats a l'estimació de distàncies astronòmiques o terrestres (Desrosières, 1998: p. 61). Les aportacions de Galton primer i de Pearson després van ser necessàries per desenvolupar no només els conceptes de correlació que acabem de formular, sinó també la regressió tal com la coneixem.

La regressió per la via dels mínims quadrats és el mètode de regressió més habitual. Té l'origen en una distribució bivariant de tipus quantitatiu que es pot representar en un diagrama de dispersió i que, a priori, s'ajusta a una forma lineal, com el cas que acabem de presentar en el punt anterior. L'ajust a una recta y es basa formalment en una equació que està determinada per la fórmula següent, on a marca el punt d'intercepció quan $\hat{y} = 0$; b designa el pendent de la línia que està determinada pels valors de $\hat{y}$ per cada increment de x.

$$\hat{y} = a + bx$$

Per tant, a cada valor x_i de la variable, que designarem com a independent X, correspondran dos valors: un, el que se li assigna empíricament, és a dir, y_i. L'altre valor és el que li correspondria en el model ideal $\hat{y}$. La diferència entre valor observat y_i i valor esperat $\hat{y}$ constitueix l'error o residu e_i, d'on podem extraure que:

$$e_i = y_i - \hat{y}$$

Aleshores, la recta ideal per a una distribució és la que faça possible que les distàncies entre valors observats y_i i valors esperats $\hat{y}$ siga menor, o dit d'una altra manera, que l'error e_i siga el mínim possible. Això, després de desenvolupar el concepte d'error[22]

[22] Vegeu el desenvolupament complet del concepte d'error en La-Roca (2006: p. 104 i ss.).

desemboca en el fet que el càlcul del factor b de l'equació és el resultat de l'aplicació de la següent fórmula:

$$b = \frac{S_{xy}}{S_x^2}$$

I, per tant:

$$a = \bar{y} - b\bar{x}$$

De la matriu sobre la qual hem extret les dades anteriors, tenim les dades següents, aquesta vegada agafades sense agrupar[23]:

x_i	y_i
8	40
3	5
7	89
4	1
8	204
7	175
6	231
8	180
6	108
9	254
$\bar{x} = 6{,}6$	$\bar{y} = 128{,}7$
$S_x^2 = 3{,}6001$	$S_y^2 = 8676{,}9039$
$S_x = 1{,}8974$	$S_y = 93{,}1499$
$S_{xy} = 120{,}5333$	
$r_{xy} = 0{,}6819$	

Taula 33. Càlcul de covariància i correlació per a les variables seguidors d'Instagram (x_i) i satisfacció amb l'aparença personal (y_j) sense agrupar

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

[23] Vegeu que les dades resultants per a mitjanes canvien, i amb això tota la resta: desviacions, covariància i correlació.

Aleshores, el càlcul dels factors a i b es resoldran mitjançant les dues fórmules anteriors:

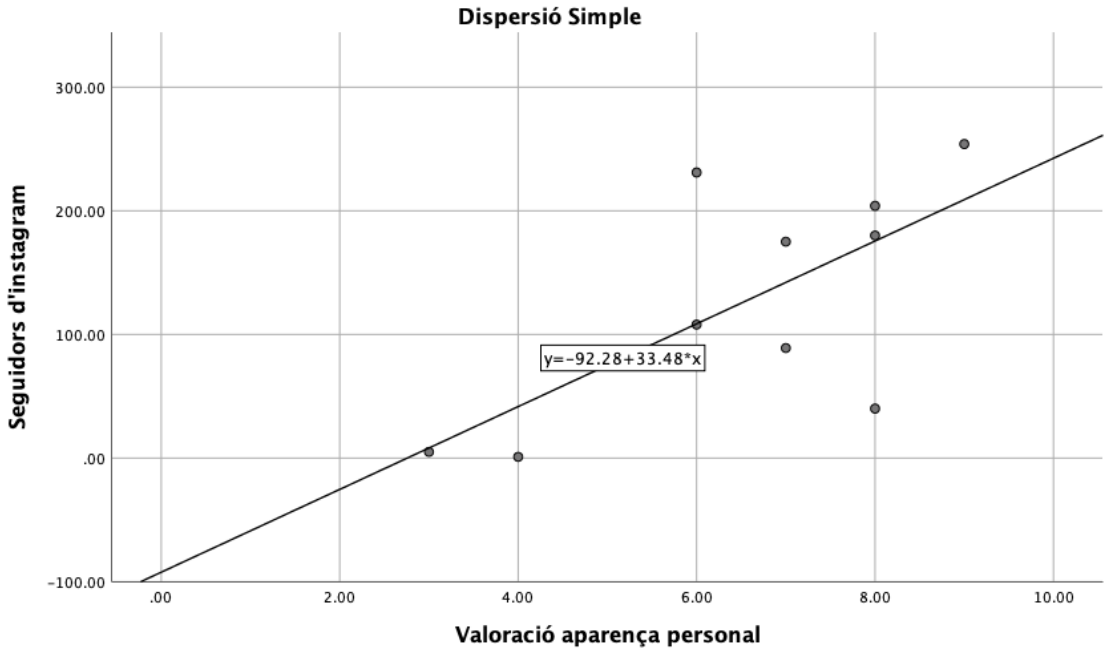
$$b = \frac{S_{xy}}{S_x^2} = \frac{120{,}5333}{3{,}6001} = 33{,}4805$$

$$a = \bar{y} - b\bar{x} = 128{,}7 - 33{,}4805 * 6{,}6 = -92{,}2713$$

Per tant, la recta queda de la manera següent:

$$y = a + bx = -92{,}2713 + 33{,}4805x$$

La transposició de la fórmula sobre el pla bivariant consisteix a localitzar primer el valor de a, que és el que talla sobre l'eix x, i després anar sumant cada valor de x amb l'elevació corresponent a b. En el cas que ens ocupa, la recta parteix de l'eix negatiu i creua el primer valor de y en -92,2713. El primer valor de la recta de regressió per sobre de $x = 0$ és en 2,756, de manera que es pot comprovar la relació directa que havíem deduït en el càlcul de la covariància i de la correlació, i també la intensitat de l'ajust de les dades sobre la recta.



Gràfica 22. Gràfica de dispersió de les variables seguidors d'Instagram i valoració de l'aparença personal amb la recta de regressió

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

Pel que fa als errors e_i , a partir de la fórmula de la regressió podem calcular la distància de cada valor de y_i respecte del nou valor de $\hat{y}$, cosa que ens serà d'utilitat posteriorment per al càlcul de les proves de significació. Es pot comprovar que, com era previsible, la suma dels errors és 0, atès que uns compensen els altres per l'efecte de centralitat de la recta extreta dels mínims quadrats:

x_i	y_i	$\hat{y}$	$e_i = y_i - \hat{y}$
8	40	175,5727	-135,5727
3	5	8,1702	-3,1702
7	89	142,0922	-53,0922
4	1	41,6507	-40,6507
8	204	175,5727	28,4273
7	175	142,0922	32,9078
6	231	108,6117	122,3883
8	180	175,5727	4,4273
6	108	108,6117	-0,6117
9	254	209,0532	44,9468
$\bar{x} = 6{,}6$	$\bar{y} = 128{,}7$	$\hat{y} = 128{,}7$	$\sum = 0$
	$S_y^2 = 8676{,}9039$	$S_{\hat{y}}^2 = 4035{,}3980$	

Taula 34. Càlcul dels errors respecte de la recta de regressió per a les variables seguidors d'Instagram (y_j) i satisfacció amb l'aparença personal (x_i)

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

A partir de la mesura de la variable observada i els errors que es produeixen sobre els valors esperats es pot deduir la proporció de la variància de la variable dependent explicada per la variable independent, és a dir, la bondat d'ajust. Aquest coeficient s'anomena coeficient de determinació i el seu símbol és R^2. La fórmula del coeficient de determinació és:

$$R^2 = \frac{S_{\hat{y}}^2}{S_y^2}$$

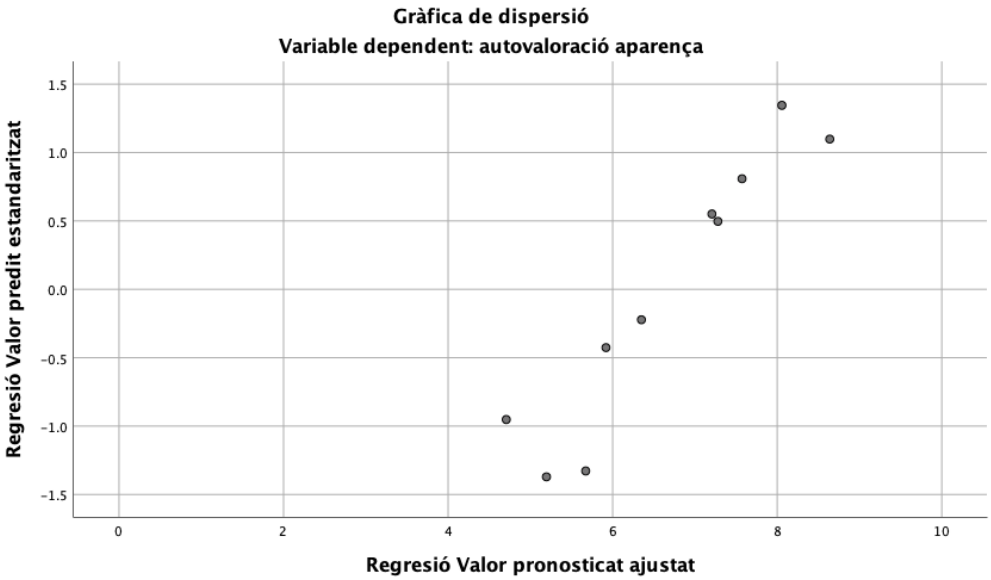
La interpretació del coeficient de determinació s'ha de fer, com en el cas del coeficient de correlació, en una escala de 0 a 1, on el valor màxim s'assoleix quan tota la variació

de y és explicada pel model, de manera que la variància explicada i la variància total són iguals i queda anul·lada la variància residual. Això suposa, per exemple, una recta empírica ajustada de manera perfecta a la recta teòrica. Per contra, quan el valor és 0, es pot deduir que la variància explicada és nul·la i la variància residual és igual a la variància total. Per tant, el model i l'ajust a la recta teòrica no expliquen res. Per a l'exemple anterior, tenim que la variància de la variable y és $S_y^2 = 8676{,}9039$ i el càlcul de la variància[24] dels errors $\hat{y}$ és $S_{\hat{y}}^2 = 4035{,}3980$. Aleshores, el coeficient de determinació és 0,4651, cosa que ens situa en un àmbit mitjà, ja que està entre el 0 i l'1. Per tant, podríem dir que un 46% de la variància de la variable contactes d'*Instagram* està determinada per l'aparença personal.

$$R^2 = \frac{4035{,}3980}{8676{,}9039} = 0{,}4651$$

D'altra banda, de la gràfica dels errors se'n poden deduir els residus individuals, a partir dels quals és possible fer una gràfica on poder constatar si hi ha observacions atípiques o algun tipus de patró de comportament dels subjectes d'anàlisi. Per exemple, per a les dades que hem estat treballant, en la gràfica apareixen les distàncies de cadascun dels punts respecte de la recta, ajustada ara a l'eix vertical entorn del valor zero. Aquesta manera de presentar les dades facilita l'observació dels valors atípics, però alhora en magnifica la desviació respecte de la recta. Si sumem tots els errors individuals, el valor resultant és zero, com s'ha pogut observar anteriorment.

[24] Variància que es calcula de la mateixa manera que la resta de les variàncies, a partir de les desviacions de cada observació sobre la mitjana de l'error.



Gràfica 23. Gràfica de dispersió dels residus de la variables valoració de l'aparença personal respecte de la recta de regressió

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

El control de casos atípics pot resultar interessant en l'anàlisi estadística, ja que en alguns casos podem trobar individus que, per la seua posició, poden alterar de manera lleugera o notable l'equació de la regressió. Per exemple, suprimir un individu pot significar des d'un moviment mínim de la recta de regressió fins al canvi de signe de la recta.

Correlació i causació

Tot i que pot haver-hi correlació entre dues variables, això no implica que hi haja una relació causal entre elles. És cert que un valor elevat en l'associació pot estar indicant un cert grau de causalitat, però ni la prova de correlació ni tampoc la regressió indiquen de manera clara quina és la variable que actua sobre l'altra. Una de les raons per les quals no és possible saber-ho és perquè falta un control de l'aparició temporal de les variables per tal de controlar-ne la causació: cal poder demostrar que la variable independent ocorre abans de la variable dependent en el temps. Una altra raó que cal tenir en compte és l'entrada en joc d'altres variables que puguen afectar la relació entre independent i dependent. En el cas que hem analitzat, hi poden estar actuant algunes variables, com ara el nivell educatiu o la classe social, que caldrà controlar per aïllar el seu efecte sobre la correlació.

2.3.4. Associació en variables ordinals: correlació ordinal de Spearman i Gamma

Rho de Spearman

L'adaptació de l'anàlisi de correlacions a variables ordinals va ser obra de Charles Spearman, psicòleg deixeble de Karl Pearson i, a més, creador de l'anàlisi factorial (Desrosières, 1998: p. 145). La correlació ordinal mesura, de manera anàloga a la correlació lineal, la direcció de la relació entre les variables i també el seu ajust a un patró preexistent. En el cas que ens ocupa, cal una ordenació de les variables, que poden provenir de variables d'interval, de manera que els valors s'alineen en nombres ordenats, per exemple, de l'1 al 5 o de l'1 al 10 o que poden provenir directament de variables ordinals, per exemple, d'escales *Likert*. En el cas que hi haja algun valor repetit, se n'ajusta el valor d'ordenació calculant la mitjana dels valors que ocuparia cada categoria i saltant un lloc l'ordre. La fórmula de Spearman tracta de deduir la diferencia d entre els ordres dels individus en les dues variables analitzades, tot incloent el seu quadrat en una fórmula molt més simple que la de la correlació, com es pot comprovar a continuació:

$$r_{xy} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}$$

La interpretació de r_{xy} s'ha de fer de la mateixa manera que el coeficient de Pearson, és a dir, que la correlació ordinal pot ser inversa o directa en funció del signe del resultat; i també pot ser una associació més intensa o més feble en funció de com de proper o llunyà estiga del valor 0.

En el cas següent observem una distribució de freqüències on trobem les notes de 10 estudiants en dues assignatures diferents. L'anàlisi de les correlacions ordinals implica una primera ordenació en què les dues variables s'ordenen de major a menor, primer pas per al càlcul de la correlació. De les diferències individuals en ambdós rangs s'obté la diferència entre rangs, d i la seua elevació al quadrat, com es pot observar en la taula següent.

Notes A	Notes B	Ordre A	Ordre B	d	d^2
5,3	6,3	2	4	-2	4

7,1	6,7	8	7	1	1
4,3	3,8	1	1	0	0
6,7	5,7	5,5	2	3,5	12,25
5,9	6,2	3	3	0	0
7,0	6,5	7	5	2	4
6,4	6,6	4	6	-2	4
8,8	8,5	10	10	0	0
8,4	7,4	9	9	0	0
6,7	6,9	5,5	8	-2,5	6,25
					$\sum = 31{,}5$

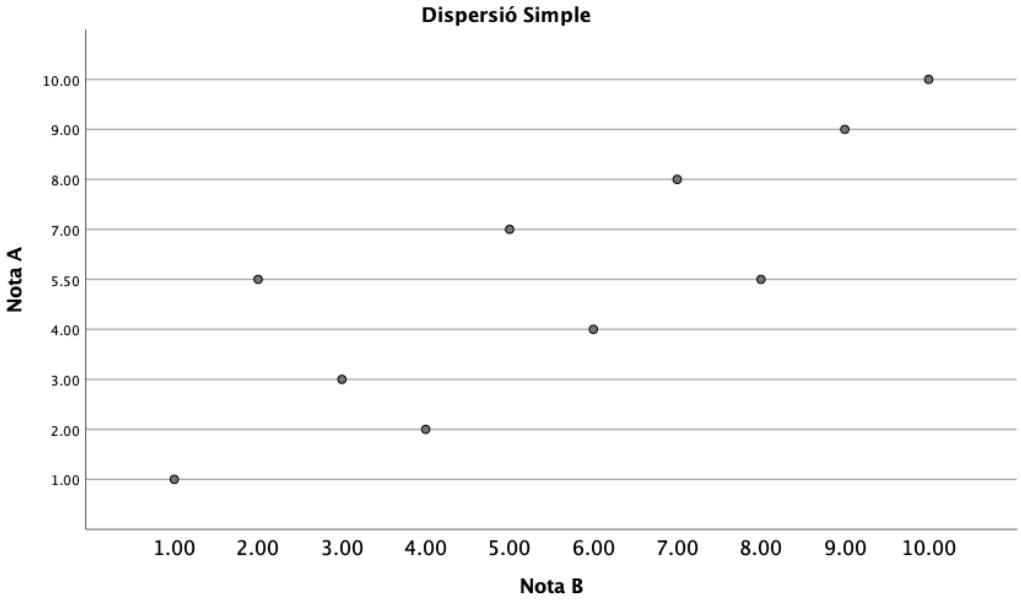
Taula 35. Taula de freqüències de les variables nota A i nota B, ordenació i càlcul de les diferències

Font: Elaboració pròpia

L'aplicació de la fórmula ens deixa el resultat següent:

$$r_{xy} = 1 - \frac{6 * 31{,}5}{10(10^2 - 1)} = 1 - \frac{189}{990} = 0{,}8091$$

El resultat de r_{xy} és de signe positiu, per la qual cosa se'n pot deduir una relació directa, i d'una intensitat bastant elevada, ja que s'aproxima bastant a la unitat, que és l'indicador de correlació perfecte. Això es pot comprovar en la gràfica de dispersió resultant de l'encreuament de les dues variables implicades en l'anàlisi, on s'aprecia tant la direcció de la relació ascendent, com també la intensitat de la relació entre les variables, cosa que es dedueix de la distància de cadascun dels punts respecte de la línia imaginària de la correlació.



Gràfica 24. Gràfica de dispersió dels residus de la variables valoració de l'aparença personal respecte de la recta de regressió

Font: Elaboració pròpia

Gamma

El coeficient gamma, també conegut com el coeficient Gamma de Goodman i Kruskal, pels seus creadors, els matemàtics Leo Goodman i William Kruskal, es basa en la lògica de la reducció proporcional de l'error. El càlcul de gamma (G) es pot interpretar com el percentatge en què una variable millora la nostra capacitat de predir l'altra variable. De la mateixa manera que la rho de Spearman, podem esbrinar la força de l'associació i també la direcció que pren. La seua fórmula posa en contacte el nombre de casos aparellats en el mateix ordre en les variables analitzades (N_s) i el nombre de casos aparellats en diferent ordre en les variables analitzades (N_d). Això implica calcular com a N_s els valors que coincideixen, per exemple, en l'ordre 1 en ambdues variables i com a N_d els valors que, en la categoria 1 d'una variable, prenen els ordres 2 i 3 en l'atra variable. Aleshores, gamma pren la fórmula següent:

$$G = \frac{N_s - N_d}{N_s + N_d}$$

Pel que fa al resultat, gamma s'ha d'interpretar com el percentatge d'errors que som capaços de reduir quan prediem l'ordre de parells en una variable a partir de l'ordre de parells en l'altra variable. Valors entre 0,00 i 0,30 s'han d'interpretar com a relacions

febles; entre 0,31 i 0,60, com a relacions moderades; i per sobre de 0,60, com a relacions fortes (Healey, 2016: p. 311). Igualment, la direcció de la relació s'ha d'interpretar a partir del signe que adopte la relació estudiada. Com que gamma és una mesura de relació simètrica, la interpretació és igual tant si agafem una mesura com a independent com si n'agafem una altra.

Agafem com a exemple la gràfica següent on es poden trobar les variables resultat en l'examen (x) i hores estudiades (y), recodificades ambdues perquè expressen tres nivells: baix, mitjà i alt.

	Baix	Mitjà	Alt	N_x
Baix	12	5	6	23
Mitjà	5	11	7	23
Alt	5	6	10	21
N_y	22	22	23	67

Taula 36. Taula de freqüències de les variables resultat en l'examen (x) i hores estudiades (y)

Font: Elaboració pròpia

En els casos de taules 2x2 N_s i N_d es calculen a partir de les diagonals. En les taules 3x3 com la que presentem, els valors de N_s s'han de calcular a partir dels valors baix i a la dreta dels valors de referència. Així, per al valor 12, el multiplicand s'obté de la suma de 11+7+6+10. Per al valor 5, a partir de la suma de 7+10. I així, successivament, de manera que els valors de l'extrem (6, 7 i 10) es multipliquen per 0 i no aporten res a l'índex N_s, com tampoc no ho fan els valors de la darrera línia, 5, 6 i 10. Així, el valor de N_s= 903.

Per al càlcul de N_d obrem de manera inversa, començant al cantó superior dret. Així, per al valor 6, el multiplicand ix de la suma 11+5+6+5. Així, tenim que N_d=419, amb la qual cosa ja podem calcular gamma.

$$G = \frac{903 - 419}{903 + 419} = \frac{484}{1322} = 0{,}3661$$

De la combinació de N_s i N_d en la fórmula anterior obtenim una $G = 0{,}3661$, que apunta a una relació directa i moderada, de manera que més hores d'estudi es relacionen amb

notes més altes en l'examen, i que a partir d'una variable podem reduir un 36,61% dels errors en predir els valors de l'altra variable.

2.3.5. Associació en variables nominals: khi quadrat i lambda

L'estudi de la independència de variables nominals suposa un abans i un després per a la sociologia, ja que gran part de les variables socials solen presentar aquesta escala de mesura. Així, resulta fonamental la seua introducció en diferents estudis preliminars de Pearson sobre la bondat de l'ajust i sobre la contingència (Pearson, 1900; 1904). No obstant això, la tècnica serà millorada posteriorment per Ronald Fisher, qui va resoldre l'apartat inferencial de la contingència quadràtica (Fisher, 1922).

La mesura de l'associació en les variables nominals no es pot referir ni a la mitjana ni tampoc a la variància, ja que per definició l'únic càlcul que es pot fer és la distribució de freqüències. Això només ens deixa la possibilitat d'estudiar la independència de les variables tal com hem vist a l'apartat inicial, cosa que implica l'estudi de les freqüències observades i esperades, i la comparació entre ambdues. Això és possible, en el cas de dues variables nominals, si considerem el càlcul de χ^2 *(khi*[25]*)* i el coeficient lambda (λ).

[25] Seguint el vocabulari d'estadística de la Universitat de Barcelona, la manera correcta de referir-nos a allò que en castellà es coneix com a *ji cuadrado* i en anglès com a *chi square* és khi quadrat (Comissió de Normalització Lingüística de la Facultat de Ciències Econòmiques i Empresarials, 1996)

Mesures basades en khi quadrat

El càlcul del coeficient χ^2 es fa, per tant, seguint la lògica de la fórmula proposada per a l'avaluació de la independència, amb què es comparen les freqüències observades amb les que caldria esperar atesa una situació d'independència entre les variables, i se n'avaluen els residus:

$$n_{ij} = \frac{n_{i.} n_{.j}}{N}$$

En el cas de χ^2 la fórmula incorpora el quadrat per tal de controlar els signes resultants, de manera que sempre tinguem un resultat positiu. Així, la prova consisteix a comparar les freqüències observades n_{ij} amb les esperades $\hat{n}_{ij}$, tot al quadrat, dividit per les freqüències esperades $\hat{n}_{ij}$.

$$\chi^2 = \sum_{i=1}^{i} \sum_{j=1}^{j} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

El resultat del coeficient s'ha de trobar entre el 0, valor que assenyala la independència entre les dues variables, i el producte del total d'observacions n pel menor dels valors de les categories $i - 1$ o $j - 1$.

A tall d'exemple, prenem com a punt de partida les dues taules de freqüències absolutes exposades anteriorment sobre posicionament polític i sexe. Cal recordar que les freqüències esperades s'obtenen del producte dels marginals i la divisió del nombre resultant per N:

Obs. A	Cons.	Prog.	n_i
H	1	4	5
D	3	2	5
n_j	4	6	10

Obs. B	Cons.	Prog.	n_i
H	2	3	5
D	2	3	5
n_j	4	6	10

Esp. A	Cons.	Prog.	n_i
H	2	3	5
D	2	3	5
n_j	4	6	10

Esp. B	Cons.	Prog.	n_i
H	2	3	5
D	2	3	5
n_j	4	6	10

Taula 37. Taula de freqüències observades i esperades de les variables sexe i posicionament polític

Font: Elaboració pròpia a partir de La-Roca (2006: p. 71).

El valor de χ^2 per als dos grups és el següent

$$\chi_A^2 = \frac{(1-2)^2}{2} + \frac{(4-3)^2}{3} + \frac{(3-2)^2}{2} + \frac{(2-3)^2}{3} = 1{,}6667$$

$$\chi_B^2 = \frac{(2-2)^2}{2} + \frac{(3-3)^2}{3} + \frac{(2-2)^2}{2} + \frac{(3-3)^2}{3} = 0$$

Tenint en compte que hi ha implicades dues variables categòriques ($\mathrm{i} = 2; \mathrm{j} = 2$), i per tant ($\mathrm{i} = 2-1; \mathrm{j} = 2-1$), amb la qual cosa la menor de les dues categories menys 1 és igual a 1, i que el nombre d'observacions (N) és 10, aleshores el valor màxim que pot arribar a assolir, tant en el grup A com en el B, és 10. Una vegada fets aquests càlculs queda clar que en el grup B les variables són independents ($\chi_B^2 = 0$), mentre que hi ha una certa associació en el grup A ($\chi_A^2 = 1{,}6667$).

El càlcul del grau d'associació, però, s'ha de fer mitjançant algun dels coeficients basats en la contingència quadràtica Tschuprow, Cramér, Yule o Pearson (Blalock, 1979: p. 304 i s.; La-Roca, 2006: p. 97). Un dels coeficients més utilitzats és, precisament, el que Pearson va presentar a principi de segle XX, i així:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

Novament, el coeficient de contingència de Pearson és 0 en situacions d'independència, i el valor màxim és la unitat, de manera que és més intuïtiu oferir la bondat d'ajust de la mesura de dependència. En el cas del grup A, l'únic en què hem mesurat una situació de dependència, el coeficient de contingència donarà el resultat següent:

$$C_A = \sqrt{\frac{1{,}6667}{1{,}6667 + 10}} = \sqrt{0{,}1429} = 0{,}378$$

La interpretació de les mesures de la força d'associació per a variables nominals s'ha de fer seguint els criteris següents: valors entre 0 i 0,10 tenen una relació feble, valors entre 0,11 i 0,30 tenen una relació moderada i valors per damunt de 0,30 tenen una relació forta (Healey, 2016: p. 303). Aleshores, el coeficient de contingència per a C_A s'ha d'interpretar com una associació forta entre les variables sexe i posicionament polític.

Un altre dels coeficients d'associació que es poden aplicar a les mesures de khi quadrat és ϕ (phi), introduït també pel matemàtic Karl Pearson. El coeficient ϕ s'aplica a taules de dues variables categòriques, és a dir, a taules 2x2. En cas d'aplicar ϕ a taules de més de dues categories per variable, el seu valor excedirà la unitat i serà més difícil d'interpretar (Healey, 2016: p. 302). La seua fórmula és molt senzilla:

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

Per a taules de més de dues categories per variable es pot aplicar el coeficient V de Cramér, obra del matemàtic Harald Cramér. El coeficient V, com el coeficient de contingència i ϕ, s'ha d'analitzar tenint en compte que el valor màxim és la unitat i que qualsevol valor que s'hi acoste indica una associació forta. En aquest cas, la fórmula de V és la següent:

$$V = \sqrt{\frac{\chi^2}{N\ (k-1)}}$$

on el denominador significa el producte del nombre d'observacions pel mínim entre els valors de les files menys 1 o el de columnes menys 1.

El coeficient lambda

Els mateixos Leo Goodman i William Kruskal, als quals anteriorment hem atribuït la paternitat del coeficient gamma per a variables ordinals, són també els creadors de lambda, una altra mesura basada en la reducció proporcional de l'error. Per dur a terme el càlcul de lambda (λ), s'han de fer dues prediccions sobre els valors de la variable dependent. La primera predicció es fa ignorant la variable independent, mentre que en la segona s'incorpora la informació de la variable independent. Així, si les variables estan associades, els errors de predicció del valor resultant són menors, de manera que si hi incloem la variable independent, les prediccions no milloraran el primer càlcul efectuat.

El primer càlcul, el de la predicció d'errors tot ignorant la variable independent (E_1), s'obté a partir dels marginals de les files, de manera que hem de triar una de les dues categories com a valor de predicció i estimar els errors generats en aplicar aquesta norma, que és el valor E_1. Per calcular E_2 cal considerar els valors en columnes, sobre la categoria amb més pes relatiu. Així, si considerem que les dues categories amb més freqüències es troben en la primera fila, hem d'operar amb $N_{.1} - N_{11}$ i $N_{.2} - N_{12}$. La suma dels errors de predicció acumulats constitueix el valor E_2. Com s'ha avançat anteriorment, variables associades tenen valors més petits en E_2 que en E_1. La fórmula del coeficient lambda és la següent:

$$\lambda = \frac{E_1 - E_2}{E_1}$$

El resultat d'aplicar la fórmula de λ pot oferir un valor entre 0 i 1, de manera que valors propers a 0 indiquen independència i valors propers a 1 indiquen associació perfecta (i per tant absència d'errors en la segona predicció). L'avantatge de λ sobre V i ϕ rau en el fet que, del resultat, en podem deduir, com en gamma, el percentatge de reducció de l'error, amb la qual cosa la simple mesura de l'associació es veu millorada.

Agafem com a exemple la taula que hem analitzant anteriorment en l'apartat d'ordinal, i deixem de considerar-hi les categories centrals, de manera que es presenten els resultats en l'examen (x) i les hores estudiades (y).

	Baix	Alt	N_x
Baix	12	6	18

Alt	5	10	15
N_y	17	16	23

Taula 38. Taula de freqüències de les variables resultat en l'examen (x) i hores estudiades (y)

Font: Elaboració pròpia

L'error predit sobre les files (E_1) s'obté d'una de les dues files. En aquest cas, pronosticarem que poques hores d'estudi donaran com a resultat resultats baixos, i per tant els errors generats seran $E_1 = 15$. El segon error predit, que posa en contacte la variable independent i la dependent, parteix de la resta $N_{\cdot 1} - N_{11}$ i $N_{\cdot 2} - N_{22}$ (atès que en la segona columna el valor amb major freqüències es troba en la segona fila). Així, $E_2 = 5 + 6 = 11$. De l'aplicació de la fórmula obtenim que $\lambda = 0{,}267$.

$$\lambda = \frac{E_1 - E_2}{E_1} = \frac{15 - 11}{15} = 0{,}267$$

La interpretació de λ en el cas que acabem de presentar s'ha de fer dient que el coneixement de les hores d'estudi millora la nostra capacitat de predir el resultat de l'examen un 26,7% respecte de no tenir-les en compte en la predicció.

Cal tenir en compte, però, que λ presenta dues febleses: la primera és que és un coeficient asimètric, i per tant, quan donem la volta al càlcul, el coeficient resultant és diferent. La segona feblesa rau en el fet que, en el cas en què les categories estiguen altament concentrades en una de les files, el resultat pot ser d'independència, fins i tot quan altres proves com ara khi quadrada donen positiu per a l'associació (Healey, 2016: p. 306).

Residus tipificats

La introducció de l'anàlisi dels residus tipificats (o estandarditzats) fa possible la comparació de valors mitjançant una relació entre les freqüències observades i les freqüències esperades. Així, els residus tipificats es calculen a partir de la fórmula següent:

$$Re = \frac{n_{ij} - \hat{n}_{ij}}{\sqrt{\hat{n}_{ij}}}$$

Dit d'una altra manera, cada residu de cada casella en el càlcul de khi quadrat ofereix el grau en què cada casella contribueix al valor final de l'estadístic χ^2.

El problema amb els residus tipificats és que, pel fet que no estan corregits, no es poden interpretar com a puntuacions tipificades (tot i que la mitjana o valor esperat és 0, la desviació típica tendeix a ser menor d'1). La correcció dels residus tipificats va ser introduïda per Haberman (1973) i consisteix a introduir com a denominador, no ja l'arrel quadrada del valor esperat, sinó l'error típic de cada casella, que de manera habitual sol representar-se com a:

$$ReC = \frac{n_{ij} - \hat{n}_{ij}}{\sqrt{\hat{n}_{ij}\left(1 - \frac{n_{\cdot j}}{n}\right)\left(1 - \frac{n_{i\cdot}}{n}\right)}}$$

D'on $n_{\cdot j}$ són els marginals de columna i $n_{i\cdot}$ són els marginals de fila. La interpretació, en aquest cas, dependrà dels valors resultants per a cada cel·la, i com que es tracta de valors normalitzats, es poden comparar a l'efecte de mesura de l'associació, amb el valor $\pm$1,96 que és el que situa la significativitat per als valors Z de la taula que, com veurem posteriorment, són els que, en termes probabilístics, marquen relacions estadísticament significatives a partir de la distribució normal tipificada. És a dir, valors superiors a 1,96 o inferiors a -1,96 s'han d'interpretar com a relacions positives o negatives i estadísticament significatives.

Tot seguit apliquem ambdues fórmules sobre l'exemple plantejat en l'exercici sobre khi quadrat, concretament sobre el grup A, que prèviament hem comprovat que ofereix un valor diferent de zero, i per tant n'hem avaluat una certa dependència. De la columna de residus tipificats, se'n desprèn la distància de cada combinació de variables respecte del valor esperat, en termes relatius. Ara bé, de l'aplicació de la fórmula dels residus tipificats corregits, se'n desprèn que la idea inicial d'associació, tot i ser vàlida, no resulta estadísticament significativa, ja que cap dels residus tipificats corregits no passa del llindar de l'1,96 crític per a un 95% de confiança[26], conceptes aquests darrers que recuperarem tot seguit.

[26] Tot i que prèviament cal haver assumit la normalitat de les variables implicades per tal de fer-ne l'aproximació a la normal, especialment tenint en compte la mida de la mostra.

Residus tipificats				Residus tipificats corregits			
	Cons.	Prog.	n_i		Cons.	Prog.	n_i
H	-0,5	0,33	5	H	-1,291	1,291	5
D	0,5	-0,33	5	D	1,291	-1,291	5
n_j	4	6	10	n_j	4	6	10

Taula 39. Taula de residus tipificats de les variables sexe i posicionament polític.

Font: Elaboració pròpia

2.4. Probabilitat i introducció a la inferència

La probabilística moderna no es desenvolupa fins al moment en què es pot deduir l'aleatorietat i separar-la, tant dels sistemes de superstició, com dels propis de la religió, cosa que ocorre al voltant del segle XVI (Kendall, 1960), particularment a partir de l'obra dels italians Gieolamo Cardano i Galileo Galilei, ambdós vinculats al protocàlcul de probabilitats relacionat principalment amb els jocs d'atzar, tant els daus com les cartes, però també amb situacions hipotètiques en les quals es pot donar una situació o l'alternativa (Desrosières, 1998: p. 47; Mateos-Aparicio, 2002: p. 3 i s.).

No obstant, les aportacions teòriques més rellevants arribarien al segle XVII amb l'obra de Blaise Pascal i Pierre de Fermat, els primers a desvincular el càlcul de probabilitats dels jocs d'atzar, tot i que aquest component segueix sent central en els seus càlculs, i serà també el que els connecte, amb Christiaan Huygens primer, i amb Jakob Bernoulli després. L'obra de Christiaan Huygens, *De Ratiociniis in Ludo Aleae*, és considerada com la primera que tracta de probabilitats encara en el segle XVII, mentre que l'aportació de Bernoulli és la primera que distingeix entre esperança matemàtica i esperança moral (Mateos-Aparicio, 2002: p. 7 i s.), i és també la que donarà lloc a la llei dels grans nombres.

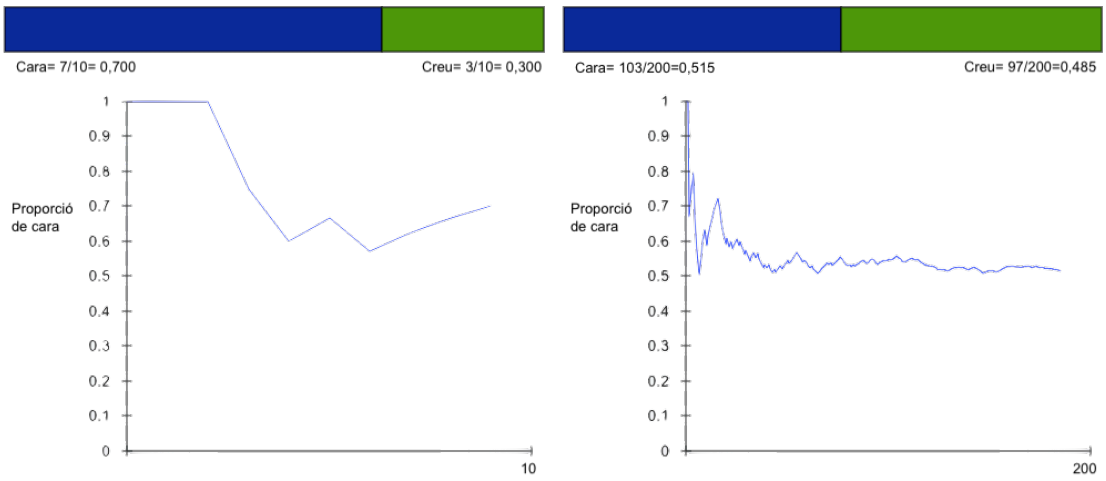
2.4.1. La llei dels grans nombres i el teorema central del límit

La llei dels grans nombres

Entre totes els teories probabilístiques que s'estaven desenvolupant al voltant del joc dels daus, les cartes o el llançament d'una moneda, Jakob Bernoulli es va centrar en esdeveniments de dos resultats: guanyar i perdre o, en el context clàssic d'extraccions d'una urna, treure una bola roja o una de blava. Avui dia ens referim a aquests esdeveniments com a situacions de tipus binomial. Fins a Bernoulli, la probabilística s'havia dedicat a estimar la possibilitat d'extraure un determinat resultat d'una distribució binomial coneguda. Allò que innova Bernoulli és l'estimació, a priori, de la composició de la distribució binomial i, per tant, el procés invers que, fins aleshores, havia estat l'habitual. Així, observa que per a una urna amb 30 boles roges i 20 boles blaves, el nombre d'observacions necessàries per tal que l'estimació de treure una bola d'un color o un altre, a

priori, ha de ser t vegades superior a la contrària (Landro i González, 2012: p. 40). El resultat és que, a mesura que hi ha més observacions, la freqüència relativa s'aproxima a un valor determinat, que serà la llei que governa tal esdeveniment. En termes matemàtics, per a una població amb mitjana μ, a mesura que les observacions augmenten, la mitjana $\bar{x}$ dels valors observats s'aproximarà a la mesura de la població μ (Moore, Notz i Fligner, 2018: p. 649).

Aquest seria el cas de la probabilitat de treure cara o creu en un experiment clàssic de llançar una moneda a l'aire. La probabilitat de treure cara (p) durant les primeres observacions s'allunyarà de l'equiprobabilitat. Si, com és el cas de la primera gràfica, limitem les observacions a 10, la sensació de l'equiprobabilitat sembla llunyana. No obstant, quan els observacions es multipliquen, la tendència de l'experiment és a l'aproximació a l'equiprobabilitat ($p = 0{,}5$)



Gràfica 25. Gràfiques de probabilitat de treure cara en un llançament de moneda amb $n = 10$ *i* $n = 200$

Font: Elaboració pròpia a partir d'https://digitalfirst.bfwpub.com/stats_applet/stats_applet_10_prob.html

El teorema central del límit

Amb l'arribada del segle XIX apareix la primera teoria clàssica sobre la probabilitat, obra de Pierre Simon Laplace, autor que introduí la inferència d'una mostra a l'univers per tal de calcular els habitants d'una població a partir d'indicadors de naixements, matrimonis i morts a París, en absència d'un cens de població (Desrosières, 1998: p. 25). La principal innovació de Laplace es va inspirar en l'obra de dos autors coetanis: per una banda, la d'Adrien-Marie Legendre, de qui ja hem parlat anteriorment quan hem tractat

els mínims quadrats sorgits del càlcul dels errors en les observacions astronòmiques, que servirien per a unir les observacions empíriques al grau de certesa d'aquestes (Desrosières, 1998: p. 62). Per una altra banda, la de Carl Gauss[27], que va comprovar que els errors sorgits dels mínims quadrats seguien una distribució normal. Amb aquestes bases teòriques es desenvolupa la teoria de la probabilitat, que tindrà en Pierre Simon Laplace un dels seus teòrics clàssics més rellevants.

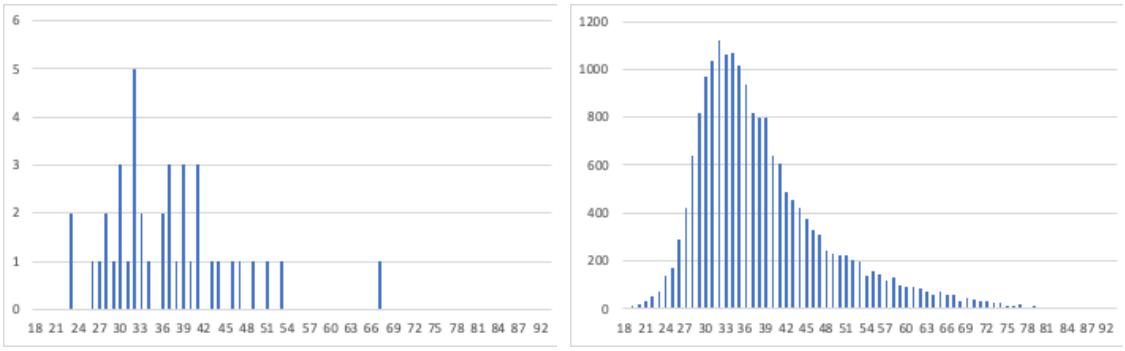
Allò que Laplace va unir fou el concepte de la normal a la interpretació amb la distribució dels errors, per a formular el teorema del límit central, segons el qual, en observar un número suficient de freqüències de variables aleatòries, fins i tot quan la seua distribució inicial no és normal, sol tendir a adoptar una forma de corba normal. Així, si prenem una mostra aleatòria simple n d'una població N amb una mitjana μ i desviació típica σ, la distribució de la mitjana $\bar{x}$ s'aproximarà a la normal:

$$\bar{x}\ és\ aproximadament\ N\left(\mu\frac{\sigma}{\sqrt{n}}\right)$$

Aquest teorema esdevindrà bàsic per tal de desenvolupar la inferència. Per exemple, basant-nos en el teorema central del límit podem utilitzar els càlculs de la probabilitat normal per a donar resposta a càlculs sobre mitjanes de mostres amb moltes observacions, fins i tot quan la distribució no continua normal.

A mode d'exemple, hem extret l'edat al matrimoni de quatre mostres aleatòries de 10 persones cadascuna de l'estadística de matrimonis de l'INE per a l'any 2018 en el territori del País Valencià. De les quatre mostres obtenim les següents mitjanes d'edat en el matrimoni: $\bar{x}_1 = 35{,}7$; $\bar{x}_2 = 36{,}7$; $\bar{x}_3 = 41{,}7$; i $\bar{x}_4 = 30{,}6$. Les quatre mostres combinades ofereixen una mitjana de $(\overline{\bar{x}} = 36{,}18)$ que és molt semblant a la que s'observa empíricament amb la mesura de tota la població ($\mu = 38{,}53$). També es pot observar en les gràfiques següents que la forma de les edats en les mostres combinades i en l'univers de referència és molt semblant.

[27] Qui, per cert, reclamava la paternitat dels mínims quadrats per davant de Legendre (Desrosières, 1998).



Gràfica 26. Gràfiques d'edat en el matrimoni al País Valencià $n = 40$ *i* $n = 18913$

Font: Elaboració pròpia a partir de l'INE, Microdades de l'Estadística de Matrimonis, 2018.

La distribució mostral

De l'anterior exercici teòric de combinació de les k diferents mostres aleatòries possibles d'un univers determinat naix la distribució mostral. Es tracta d'una distribució teòrica dels valors de l'estadística en totes les mostres possibles de la mateixa mesura que es poden extraure d'un mateix univers. En la pràctica, dur a terme més d'una mostra d'un mateix univers en què mesurem els mateixos indicadors és econòmicament poc viable i, ara com ara, una quimera. Tanmateix, si duguérem a terme l'exercici veuríem, idealment, que la distribució de mostres resultants té tres característiques:

a) Resulta ser una distribució que s'aproxima a la normalitat.
b) La seua mitjana s'aproxima molt al valor de la mitjana de l'univers.
c) La seua desviació, en canvi, se sol allunyar del valor real de l'univers.

Si ampliem l'exercici de treball amb les *k* mostres aleatòries possibles d'un univers i centrem l'atenció en la mitjana $\bar{x}$ podrem comprovar que aquesta s'aproxima molt a la mitjana observada en l'univers de referència μ. Per tant, es pot dir que l'estadístic $\bar{x}$ és un estimador no esbiaixat del paràmetre μ. L'estimació de la desviació típica, en canvi, s'ha de fer a partir de la seua versió no esbiaixada, això és, tenint en compte la mitjana de les desviacions, per la qual cosa la seua fórmula és $\frac{\sigma}{\sqrt{n}}$. Així, enfront de la distribució individual de les dades de l'univers $N(\mu, \sigma)$, la distribució mostral de totes les mitjanes possibles d'un univers és regirà a partir de la distribució $N(\bar{x}, \frac{\sigma}{\sqrt{n}})$. A més, en contacte amb el teorema del límit central que hem vist anteriorment, es pot afirmar que com més gran siga la mostra en qüestió, més petita serà la desviació típica resultant.

2.4.2. Inferència i distribucions de probabilitat

La inferència, el pas d'una mesura mostral a la seua representació poblacional, és un procés que s'instaura en l'àmbit de les ciències socials al voltant dels anys 50 del segle XX. No obstant això, l'origen s'ha de cercar en els experiments de Laplace sobre els models d'errors en observació astronòmica o, fins i tot, en la traducció que en va fer Quetelet en l'àmbit d'allò social (Gigerenzer i Murray, 1987: p. 3). La revolució inferencial, tal com va ser definida per Gerd Gigerenzer i David Murray, té lloc entre els anys 1940 i 1955, moment en el qual el pas de la mesura mostral a la poblacional esdevé el major dels objectius de la investigació empírica en l'àmbit social, als EUA primer, i a la resta del món després (Gigerenzer i Murray, 1987: p. 6). De fet, l'aplicació de la inferència és, per a alguns autors, el centre de la producció científica actual, deixant de banda qüestions com ara la minimització dels errors o la possibilitat de replicació quan, en realitat, no és més que un conjunt de ferramentes estadístiques que instauren la creença en els valors p, sobre els quals, a més, s'aprén i ensenya de maneres molt diverses i, en ocasions, contraposades (Gigerenzer i Marewski, 2015: p. 423).

El pas d'una estadística descriptiva a una estadística inferencial és fa, en primer lloc, a partir de la transformació dels probabilitats P en freqüències relatives, un pas que resulta senzill en la pràctica, però que en la teoria enfronta històricament els escoles Bayesiana i freqüentista (Gigerenzer i Marewski, 2015: p. 431). Debats a banda, hi ha quatre normes bàsiques per les quals es guia la probabilística i que fan possible aquesta transformació (Moore, 2007: p. 253):

1. La probabilitat $P(A)$ de qualsevol esdeveniment es troba entre $0 \leq P(A) \leq 1$.
2. Si prenem S com una mostra pertanyent a un model probabilístic, aleshores $P(S) = 1$.
3. Dues situacions A i B són disjuntes si no tenen concurrències en comú, de manera que mai no ocorreran de manera conjunta; aleshores $P(A\ o\ B) = P(A\ + P)(B)$.
4. Per a qualsevol situació A, $P(no\ ocòrrega\ A\ = 1 - P(A))$.

Així, si entenem la probabilitat com la proporció de repeticions que ocórrega un determinat esdeveniment a llarg termini, el pas a freqüències relatives és automàtic, atès que

el sumatori, tant d'una situació com de l'altra[28], és 1. De la mateixa manera, es pot fer el recorregut invers, és a dir, calcular la probabilitat de trobar-nos un cas aleatori dins d'una distribució de freqüències. En altres casos, les probabilitats es poden determinar a partir d'una funció teòrica que conforma, al seu torn, una distribució teòrica de probabilitat.

Dit això, podem trobar dos models probabilístics: els models discrets i els models continus. Cadascun d'aquests models analitza la probabilitat que ocórrega un determinat esdeveniment, en un cas amb valors separats i en l'altre amb valors continus.

Distribucions probabilístiques discreta i contínua

Cal entendre les funcions de distribució com una funció d'acumulació de probabilitats, de manera que si les funcions de probabilitats eren les equivalents a les freqüències relatives, les funcions de distribució ho són de les freqüències acumulades. Així, les distribucions seran sempre creixents, atès que no poden existir probabilitats negatives, i l'últim valor sempre serà la unitat.

Les distribucions probabilístiques discretes prenen un nom finit o numerable de valors possibles. Aquests valors, d'acord amb les normes que acabem d'exposar, se situen entre 0 i 1, i tots junts sumen 1. Les gràfiques de distribució probabilística discreta solen fer-se partir d'histogrames o gràfiques de barres, de manera que s'hi veu clarament la diferència entre els categories.

La distribució probabilística contínua, per contra, pren un gran nombre de valors, que poden ser infinits en tant que no hi ha límits clars entre les categories. Per tant, les representacions gràfiques ideals solen ser les de línies, en què l'àrea sota la corba entre dos punts representa la probabilitat que una variable agafe un valor que se situe en l'interval estudiat (Agresti, 2018: p. 71). De fet, es pot dir que les distribucions contínues assignen probabilitats 0 a cada resultat individual (ja que aquests poden ser infinits) i només es poden conèixer probabilitats concretes per a intervals determinats (Moore, 2007: p. 258). Una dels distribucions probabilístiques contínues més conegudes i utilitzades és precisament la distribució normal, que analitzem tot seguit. No obstant això,

[28] Una altra cosa seriosa si dins de la distribució de freqüències relatives estan incloses totes els possibles situacions que abasta la probabilitat estudiada.

en investigació social hi trobem moltes altres distribucions de probabilitats[29] que ens poden ser útils, entre les quals hi ha la uniforme, la binomial, la t de Sudent, khi quadrat o la F de Fisher-Snedecor.

La distribució uniforme

D'entre totes els distribucions probabilístiques, la uniforme és la més senzilla ja que aporta la mateixa probabilitat per a cada categoria, la qual cosa es podria traduir en una situació d'equiprobabilitat si ho entenem en termes de selecció de casos o individus d'una mostra. Es pot presentar en forma de rectangle, per a les variables discretes, o en forma de successió de punts o línia en el cas de les contínues. La fórmula de la funció de densitat de la distribució uniforme per a un interval comprés entre els valors a i b és:

$$f(x) = \frac{1}{b-a} \; per \; als \; valors \; compresos \; entre \; a \leq x \leq b$$

Així doncs, la probabilitat de trobar un cas comprés entre l'interval $[a, b]$ dependrà de la longitud de l'interval, no de la posició, donada l'equiprobabilitat constant.

A més, la distribució uniforme presenta una mitjana que ve determinada per:

$$\bar{x} = \frac{a+b}{2}$$

I una desviació típica:

$$S_x = \sqrt{\frac{(b-a)^2}{12}}$$

Aquests serien els casos clàssics del llançament d'una moneda o d'un dau de sis cares sense modificar, que a priori presentarien la mateixa probabilitat de cara/creu o de treure

[29] La major part de distribucions de probabilitat no tenen aplicació pràctica en l'àmbit de les ciències socials. De fet, la major part dels manuals d'estadística aplicada a les ciències socials solen centrar-se en la distribució normal, alguns inclouen la binomial, uns altres incorporen khi quadrada i T de Student i molt pocs tracten la distribució F de Fisher-Snedecor. En el nostre cas, les incorporem totes, incloent-hi també la distribució uniforme per tal d'assentar les bases de la interpretació de les probabilitats.

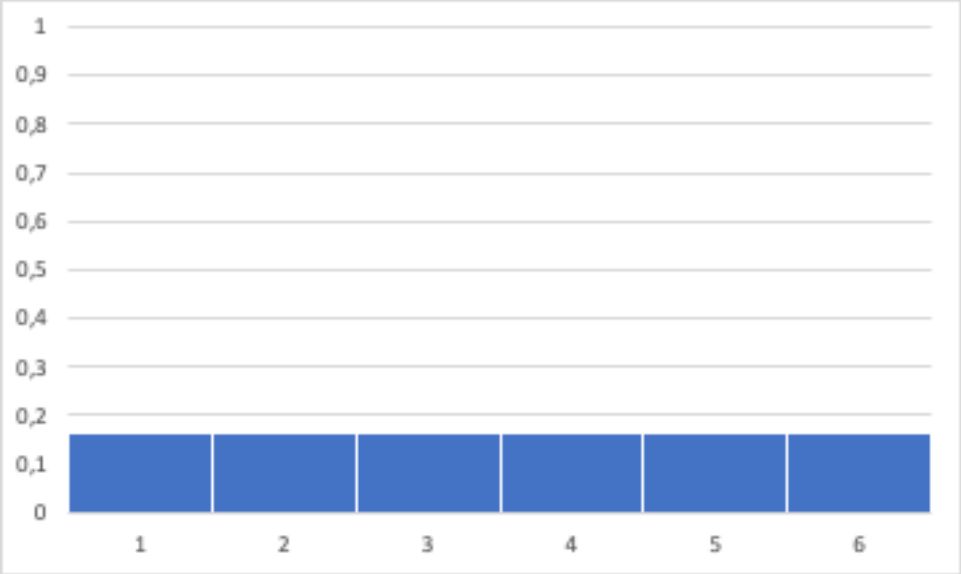
qualsevol de les sis cares. En el cas de les distribucions uniformes discretes, com és el cas de l'experiment del dau, la probabilitat de cada valor es calcularia a partir de:

$$p(x) = \frac{1}{n}$$

Així, obtindríem successivament que la probabilitat de cadascuna de les sis cares, i per tant de cada cas (o si extrapolem, de cada individu) seria la següent:

$$p(1) = \frac{1}{6} = 0{,}1667$$
$$p(2) = \frac{1}{6} = 0{,}1667$$
$$\ldots$$
$$p(6) = \frac{1}{6} = 0{,}1667$$

La qual cosa es veuria representada en termes gràfics de la manera següent:



Gràfica 27. Gràfica de probabilitat de resultats quan en tirar un dau equilibrat

Font: Elaboració pròpia

La distribució binomial

La distribució binomial és de tipus discret i és dona en aquells experiments compostos per n experiments simples i aleatoris dels quals només s'esperen dos resultats mútuament excloents, i per tant és tracta d'una distribució discreta (Cambrer et al., 2013: p. 188). Una de les diferències amb la distribució uniforme és el fet que, en les distribucions binomials, s'hi espera una mostra d'esdeveniments o casos. Les mesures poden situar-se en una de les dues categories excloents de manera que, si ho fan en la que prèviament haurem definit com a mesura d'èxit, els comptarem com a p , mentre que si recauen en la mesura contrària seran recomptades com a q o $(1-p)$, ja que com s'ha vist anteriorment, una de les propietats de les distribucions probabilístiques és que $p+q=1$.

La fórmula de la funció binomial posa en contacte el nombre d'experiments aleatoris n i la probabilitat d'obtindre èxit p.

$$f(x) \sim B(n,p)$$

A partir d'aquesta funció, podem deduir la fórmula de la distribució, que utilitza una operació de combinatòria del número de repeticions n sobre el número d'èxits k.

$$p(x=k) = \binom{n}{k} p^k q^{n-k}$$

Convé recordar que el càlcul de l'operació combinatòria es fa a partir del coeficient binomial, en què els operands són $n!$ i $k!$ expressions factorials:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

En què cada operand es resol per via de la notació factorial[30], en la qual:

$$n! = n * (n-1) * (n-2) * \ldots * (2) * (1)$$

Quan, a més, $0! = 1$.

Les distribucions binomials presenten una mitjana:

[30] La introducció del signe d'exclamació com a símbol del nom factorial es deu a Christian Kramp (o Chrétien Krempe en francès), que el formulà en 1808 tal com el coneixem (Higgins, 2008: p. 12).

$$\bar{x} = np$$

I una desviació típica (Moore, 2007: p. 329):

$$S = \sqrt{np(1-p)}$$

En el cas següent es presenta un problema a partir del qual s'ha d'aplicar allò que hem vist sobre la distribució binomial per a donar-li solució. Al País Valencià es van celebrar 565 matrimonis homosexuals d'un total de 18.914 matrimonis l'any 2018: 290 entre homes i 275 entre dones. Quina serà la probabilitat que se celebre un matrimoni entre persones del mateix sexe per cada deu que se celebren en un ajuntament valencià? Del problema, en deduïm que $n = 10$; que $k = 1$; i que $p = \frac{565}{18914}$. Aleshores:

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \binom{10}{1} 0{,}03^1 0{,}97^9 = \frac{10!}{1!\,(9)!} 0{,}03 * 0{,}76 = 10 * 0{,}03 * 0{,}76 = 0{,}228$$

La interpretació del resultat es pot fer multiplicant el número per 100, de manera que tindríem una probabilitat expressada en nombres percentuals. En aquest cas, tindríem un 22,8% de probabilitats que en deu celebracions de matrimonis d'un ajuntament valencià una siga un matrimoni entre parelles del mateix sexe.

La mateixa operació es pot resoldre, de manera aproximada, a partir de la taula de probabilitats de la distribució binomial, que s'incorpora als adjunts en la seua versió del *National Bureau of Standards* de 1952, i en la qual es poden substituir n, k i p per tal d'obtindre la probabilitat aproximada. En aquest cas, com que no existeix una p aproximada a la que hem calculat ($p = 0{,}03$), operaríem amb la més propera ($p = 0{,}05$), de manera que per a n=10 i $k = 1$ la probabilitat associada estaria al voltant de 0,3151, només unes dècimes allunyada de la que hem calculat amb l'ús de la combinatòria.

Una dels possibilitats de la distribució binomial és l'aproximació per via de la normal. Gràcies als treballs de Karl Pearson a principis de segle XX sabem que quan una distribució binomial té moltes observacions, és a dir, te una n suficientment gran[31], es poden utilitzar els càlculs de la normal per tal d'aproximar-se a les probabilitats binomials. Així,

[31] A mode indicatiu, np≥10; n(1-p)≥10 (Moore, 2007: p. 334 i s.).

per a una distribució binomial amb una n gran la distribució de x seguiria la fórmula següent:

$$N(np, \sqrt{np(1-p)})$$

Una distribució binomial ideal presentaria, per tant, els resultats següents, classificats per nombre d'èxits x i probabilitat $p(x)$:

Nombre d'èxits k	Probabilitat $p(x)$
0	$\binom{n}{0} p^0 q^n$
1	$\binom{n}{1} p^1 q^{n-1}$
2	$\binom{n}{2} p^2 q^{n-2}$
...	...
x	$\binom{n}{x} p^x q^{n-x}$
...	...
$n-1$	$\binom{n}{n-1} p^{n-1} q^1$
N	$\binom{n}{n} p^n q^0$

Taula 40. Format general d'una distribució binomial de probabilitat

Font: Cambrer et al., 2013: p. 189 i s.

Val a dir que hi ha una altra manera de fer el càlcul dels coeficients combinatoris, que és seguir el triangle de Tartaglia o de Pascal[32], en què l'eix horitzontal representa els valors de x i el vertical els valors de n (vegeu Cambrer *et al.*, 2013: 190). De la mateixa manera, el càlcul de la probabilitat es pot fer a partir d'una taula binomial de probabilitats com la que s'adjunta en els annexos, en la qual es pot consultar la probabilitat associada a un esdeveniment amb n casos, k èxits i una probabilitat preestablerta en els valors de p més habituals.

[32] D'on, tot siga dit, els noms resultants de la divisió dels valors centrals (1, 2, 6, 20, 70,...) pels valors corresponents en els laterals (1, 2, 3, 4, 5,...), és a dir, (1, 1, 2, 5, 14,...) són coneguts internacionalment com a noms *catalans* (Higgins, 2008: p. 65), no pel gentilici, sinó pel seu creador, Eugène Catalan, que va observar aquesta relació en 1855, i que té aplicacions en combinatòria i geometria.

La distribució normal

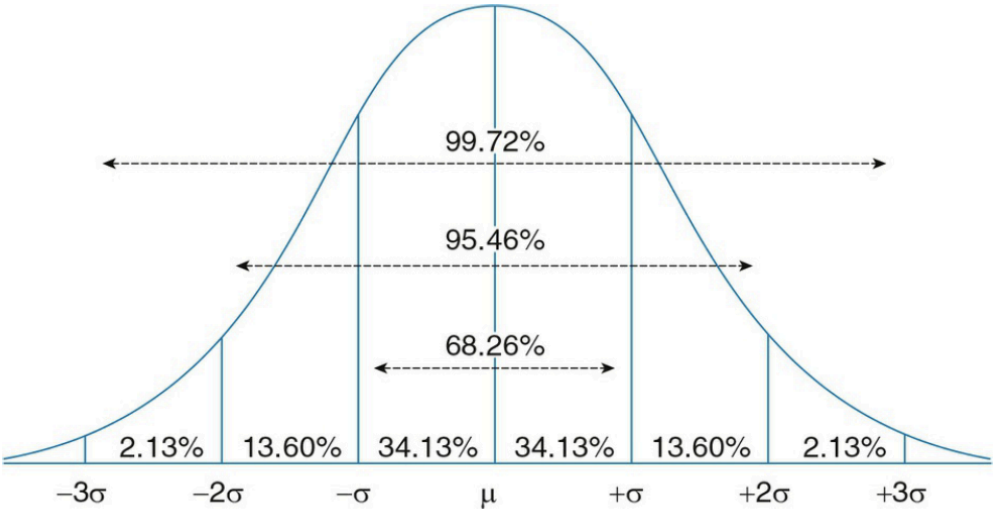
La distribució normal, també coneguda com a corba normal, distribució gaussiana o de Gauss-Laplace (Pearson, 1920) és una distribució contínua i, per tant, que representa variables d'interval. En notació matemàtica, la distribució normal es representa $N(\mu, \sigma)$.

Tal com apunten Moore, Notz i Flinger, el concepte *normal* no implica que siga comuna o fàcil de trobar, sinó tot el contrari: és una corba especial que, a més, compleix algunes característiques que la fan única (2018: p. 176):

1. Totes les corbes normals tenen una forma de campana, simètriques i amb un sol pic.
2. Qualsevol corba normal es pot definir donada la seua mitjana μ i desviació típica σ.
3. La mitjana està localitzada al centre de la corba simètrica i coincideix amb la mitjana i la moda. Si es canvia el valor de la mitjana sense canviar la desviació típica només s'aconsegueix moure la corba en l'eix horitzontal, però no en canvia la variabilitat. La seua simetria, a més, fa que puguem deduir que, de la mateixa manera que la totalitat dels casos està sota la corba, la meitat dels casos estan a cada costat del punt central definit per la mitjana.
4. La desviació típica controla la variabilitat de la corba, de manera que si la desviació típica és més gran, l'àrea sota la corba normal està més dispersa, i produeix una corba aplanada o platicúrtica; mentre que si la desviació típica disminueix, l'àrea per davall de la corba es concentra, i produeix una corba més pronunciada o leptocúrtica.

Tot i que poden existir diferents tipus de corbes normals, totes segueixen allò que Moore anomena la norma[33] 68-95-99,7 (Moore, 2007: p. 71), és a dir: el 68% de les observacions en una distribució normal es troben entre la mitjana i una desviació típica; el 95%, entre la mitjana i dues desviacions típiques; i el 99,7%, entre la mitjana i tres desviacions típiques. Això és pot observar, amb un poc més de precisió, en la gràfica de la distribució normal següent:

[33] A l'efecte de memorització, utilitzem la norma 68-95-99,7, tot i que els valors reals es corresponen amb 68,26-95,46-99,72.



Gràfica 28. Desviacions típiques per davall de la corba normal

Font: Frankfort-Nachmias i Leon-Guerrero, 2018: p. 233

Tanmateix, no totes els distribucions empíriques són normals. Abans de considerar-ne la normalitat caldrà avaluar-la de diferents maneres. La primera, i més fàcil, és fer un contrast visual a partir d'una representació visual de les dades; per exemple, un histograma, una gràfica de barres o un diagrama de caixa. Aquestes tres representacions gràfiques de les dades ofereixen una aproximació bastant fiable a la normalitat de la distribució empírica. En un estadi de major confiança, s'hi pot aplicar algun dels tests de normalitat existents, entre els quals destaquen el *Kolmogorov-Smirnov* o també el *Shapiro-Wilk*.

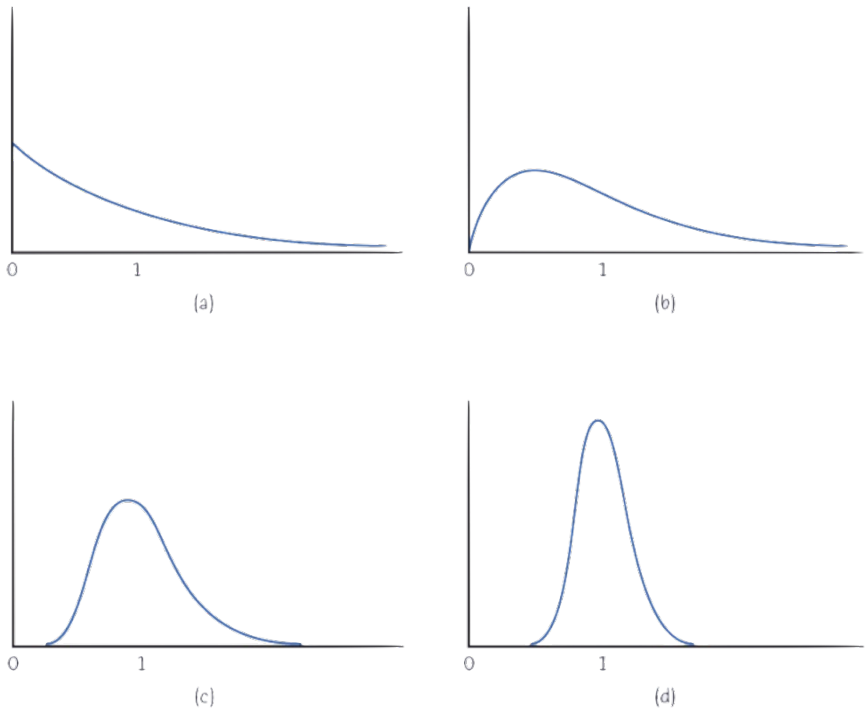
La fórmula de la funció de la distribució normal és:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2};\ \forall x \in \mathbb{R}$$

En què i i π són dues constants conegudes i només necessitem conèixer els valors de la mitjana μ i de la desviació típica σ per tal de calcular la resta de paràmetres d'una distribució normal mitjançant el càlcul d'una integral per a $f\ (x)$.

Això implica l'existència d'una corba normal per a cada distribució de freqüències, amb la qual cosa, el treball amb distribucions normals es complica. No obstant això, el teorema del límit central fa possible que, en presència de distribucions amb un gran nombre

d'observacions, siga possible utilitzar els càlculs de probabilitat de la distribució normal en problemes sobre mostres aleatòries, fins i tot quan aquestes no són normals. La pregunta, aleshores, és com de gran ha de ser una mostra aleatòria per tal que puguem igualar-ne la distribució de variables a una distribució normal. Ja hem vist anteriorment que el criteri se sol situar al voltant dels 30-40 individus (Ghasemi i Zahediasl, 2012), tot i que des de l'acadèmia no falten opinions que apunten a aquesta norma com un nombre daurat més, de la mateixa manera que ho seriosa el nivell de confiança 0,05 (Cohen, 1990; 1994). Per exemple, Healey posa el límit en els 100 individus (2016: p. 147) i Cambrer et al. en els 120 (2013: p. 278). Allò que sí que sabem, però, és que les mitjanes de mostres aleatòries són menys disperses que els que tenen el seu origen en observacions individuals; i són també més tendents a la normalitat que ho serien les observacions individuals. Vegeu, a tall d'exemple, la gràfica quàdruple que presenta Moore (2007: p. 283) respecte de les mitjanes d'una població que, a priori, se sap que no presenta una distribució normal: a mesura que les observacions sobre mostres aleatòries de la mateixa població s'incrementen, aquestes van ajustant-se més a una distribució normal, en què a) és la distribució de $\bar{x}$ per a una mostra aleatòria; b) és la distribució de $\bar{x}$ per a dues mostres aleatòries; c) és la distribució de $\bar{x}$ per a 10 mostres aleatòries; i d) és la distribució de $\bar{x}$ per a 25 mostres aleatòries.



Gràfica 29. Distribucions de la mitjana d'1, 2, 10 i 25 mostres aleatòries d'una variable no-normal

Font: Moore, 2007: p. 283.

La distribució normal estandarditzada

La distribució normal estandarditzada és la resultant del procés d'estandardització o tipificació, que com hem vist anteriorment, s'aplica a partir de la fórmula:

$$Z = \frac{x_i - \bar{x}}{S_x}$$

Aquesta fórmula ofereix una distribució normal $N(0,1)$, per tant de mitjana $\mu = 0$ i desviació típica $\sigma = 1$. Igual que hem vist en la distribució binomial, el càlcul de la probabilitat de densitat es pot fer a partir de taules estadístiques estandarditzades des d'on extraure resultats aproximats del càlcul i, de manera paral·lela, també a partir de l'ús de paquets estadístics. De la mateixa manera, doncs, es pot consultar en l'apartat d'annexos la taula de probabilitats de la distribució normal tipificada per a valors per baix de z, sent aquest valor z el resultat de la tipificació vista en apartats anteriors. En la major part de taules de la distribució normal tipificada s'ofereixen les dades per a valors per baix de z, la qual cosa condiciona la manera en què es poden dur a terme els càlculs d'estimació i càlcul d'intervals de confiança que veurem tot seguit. Els valors de z, en tot cas, es corresponen amb el càlcul de l'àrea entre la corba normal i l'eix horitzontal per la via de la resolució d'una integral que retorna un número que, multiplicat per 100, s'ajusta al percentatge de població per baix de z. En aquest cas, la fórmula per a calcular cada número z és:

$$\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx$$

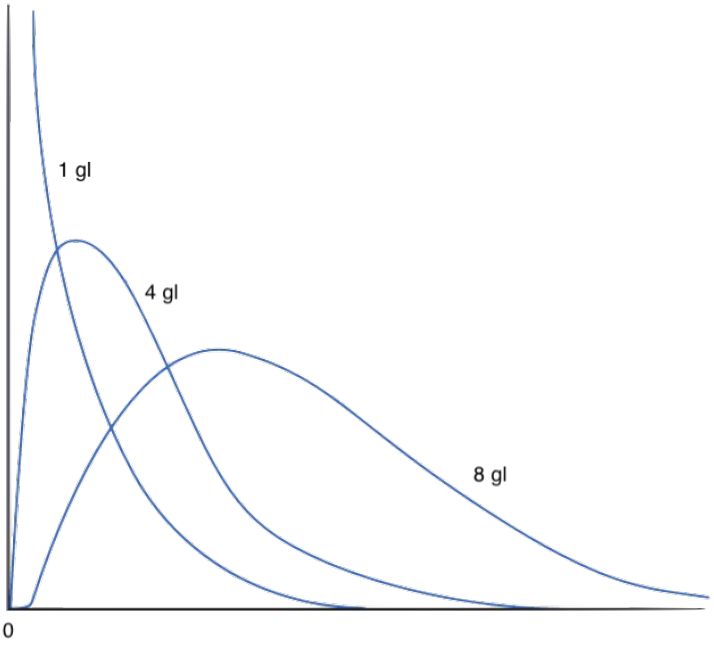
La resolució de la integral, no obstant, ofereix la mateixa distribució 69-95-99,7 que acabem de veure en el cas de la distribució normal, de manera que a l'hora de realitzar càlculs d'aproximació als percentatges d'una població per dalt o per baix d'un punt z o entre dos punts z_1i z_2 es pot intentar aproximar, primer, pel que Moore anomena el recompte empíric i després, a partir del càlcul de l'àrea coberta per la distribució normal a partir de la fórmula de z vista anteriorment i que comprovarem tot seguit en la inferència a partir de la mitjana i de l'aproximació binomial.

La distribució khi quadrat

La distribució khi quadrat és un tipus de distribucions asimètriques que presenten les següents característiques:

1. És una distribució amb una asimetria positiva, és a dir, que la cua de la dreta que dibuixa la corba de densitat és més llarga que la que es dibuixa a l'esquerra. Això diferencia aquesta distribució de la normal, que com acabem de veure, és simètrica.
2. Només pren valors positius, sent 0 el valor mínim i l'infinit el màxim. Un khi quadrat igual a zero respecte de l'associació entre dues variables implica independència absoluta i, per tant, que cada cel·la en la taula de contingència es correspon amb les seues freqüències esperades.
3. A mesura que augmenten els graus de llibertat, la distribució es torna més simètrica fins que, per damunt de 30 graus de llibertat, s'assembla a una distribució normal (Frankfort-Nachmias i Leon-Guerrero, 2018: p. 484).

Una altra diferència respecte a la distribució normal és que els seus valors venen determinats pels graus de llibertat, que es calculen, com hem vist en l'apartat d'anàlisi descriptiva, pel producte de les categories de la variable en files menys un i les categories de la variable en columnes menys un. Literalment, els graus de llibertat són el número d'elements d'un sistema que tenen llibertat per a adoptar qualsevol valor, en aquest cas, dins el creuaments entre les categories en la taula de contingència. Aquesta mesura serà especialment rellevant en el moment de dur a terme les proves de significació, donat que els valors poden ser diferents en funció dels graus de llibertat que s'adopten. En la distribució khi quadrat ocorre a més que, quants més graus de llibertat té la distribució, esdevé més simètrica, més propera a una distribució normal, i també són més probables valors elevats per a l'estadístic.



Gràfica 30. Distribucions de khi quadrat per a mostres amb 1, 4 i 8 graus de llibertat

Font: Elaboració pròpia a partir de Moore, 2007: p. 564.

La funció de probabilitat de densitat de khi quadrat és:

$$f(x|v) = \frac{x^{\frac{v-2}{2}}\, e^{\frac{-x}{2}}}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)}$$

On v són els graus de llibertat i *gamma* (Γ) és la funció Gamma.

La interpretació de la corba de densitat de khi quadrat no és tan intuïtiva com la de la corba normal. La seua asimetria i la variabilitat en funció dels graus de llibertat dificulten una aproximació empírica com la de la norma 68-95-99,7. No obstant, sí que hi ha una norma que en facilita la seua comprensió: la mitjana d'una distribució khi és igual als seus graus de llibertat (Moore, 2007: p. 565).

De la mateixa manera que hem vist en la distribució binomial i en la distribució normal, el càlcul de la probabilitat de densitat, per la seua complexitat, se sol fer a partir de taules estadístiques estandarditzades d'on l'estudiantat en pot extraure un resultat aproximat del càlcul i, evidentment, també es pot extraure el seu valor exacte a partir de l'ús del programari estadístic. En aquest cas, però, les taules estan orientades, més que cap al càlcul de l'àrea de densitat per baix d'una determinada àrea, cap als valors crítics en l'estudi de l'associació entre dues variables, com veurem a continuació.

La distribució t de Student

La distribució t, també coneguda com a t de Student[34] és, com en el cas de khi quadrat, una família de corbes, més que una sola corba amb una forma i distribució visual homogènia. En aquest cas, els valors de t venen determinats també pels graus de llibertat, que com en el cas anterior, representen el número de freqüències que tenen llibertat d'agafar x valor a l'hora de calcular un estadístic determinat. En aquest cas, els graus de llibertat estan relacionats, més que amb les categories de les variables relacionades, amb la mida de la mostra, donat que la distribució t se sol utilitzar per a caracteritzar variables quantitatives. Per exemple, la distribució t per a una mostra tindra N graus de llibertat (on N és la mida de la mostra) menys un (grau de llibertat), que és la mateixa operació que es du a terme a l'hora de calcular la desviació típica d'una mostra, expressada anteriorment com l'arrel quadrada de la quasivariància.

La distribució t es basa en l'estadístic z per a una mostra, de manera que es poden calcular els seus valors a partir de distribucions normals $N(0,1)$ tot seguint la següent fórmula, d'on es dedueix que, en absència de la desviació típica poblacional σ, se substitueix l'error estàndard ($\frac{s}{\sqrt{n}}$) per la desviació estàndard que apareix en el denominador de la fórmula original de la tipificació ($\frac{\sigma}{\sqrt{n}}$).

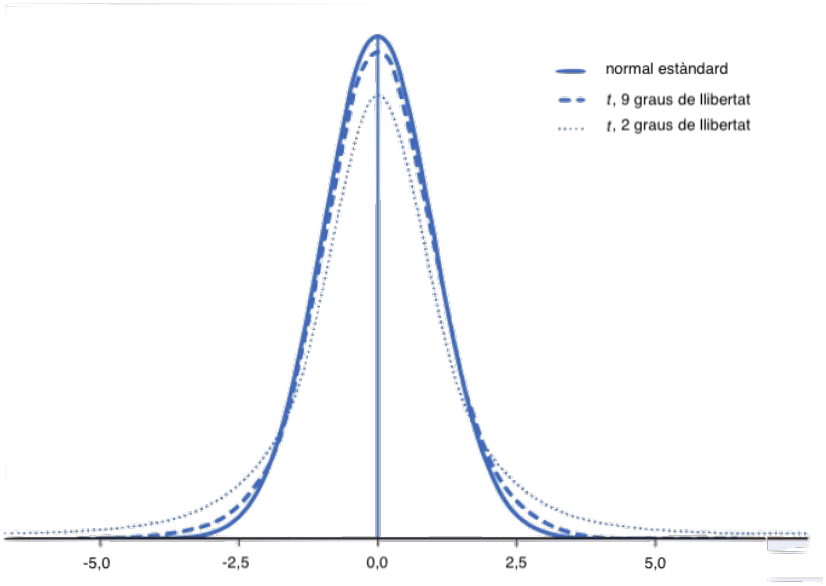
$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

La distribució t presenta les següents característiques:

1. La corba de densitat és molt semblant a la corba normal: simètrica, amb el valor 0 al centre, amb un sol màxim i forma acampanada.
2. La dispersió dels valors de la distribució t és més gran que la de la distribució normal, de manera que es pot trobar, comparativament, més probabilitat en les cues i menys en el centre.

[34] Per a saber més sobre la curiosa història de William Sealy Goset, el misteriós estudiant darrere del pseudònim *Student* en la publicació *Biometrika* (1907), vegeu la seua història glossada en l'obituari des del punt de vista estadístic pel fill de Karl Pearson (el Pearson de les proves χ^2, les correlacions i regressions) Egon Pearson (1938) i des del personal per Launce McMullen (1938).

3. S'interpreta igualment que la distribució normal estandarditzada, és a dir, que representa com de lluny està la mitjana $\bar{x}$ de la seua μ, mesurat en desviacions típiques.
4. A mesura que creixen els graus de llibertat, la forma de la distribució va acostant-se més a una corba normal, cosa que s'explica pel fet que una major mida de la mostra, que és en definitiva d'on es calculen els graus de llibertat, implica millor capacitat de predicció de σ a partir de s i també la pèrdua de les cues abultades (Moore, 2007: p. 435).



Gràfica 31. Distribucions normal estàndard i de t per 2 i 9 graus de llibertat

Font: Elaboració pròpia a partir de Moore, 2007: p. 435.

La funció de densitat de t és:

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)} 1 + \left(\frac{t^2}{v}\right)^{-\frac{v+1}{2}}$$

On Γ és la funció Gamma i v els graus de llibertat.

De la mateixa manera que hem vist en les distribucions anteriors, existeix una taula estandarditzada per tal de no haver de calcular els valors crítics de t a l'hora de fer proves de significació, en aquest cas per a demostrar igualtat o diferència de les mitjanes d'una mostra sobre una població (en el cas de mostres per baix de $n = 30$) i també per a la comparació de mitjanes de dues mostres, relacionades o no, conceptes que desenvolupem tot seguit.

La distribució F de Fisher-Snedecor

La distribució F és una família de distribucions de dos paràmetres. En aquest cas, es guia a partir dels graus de llibertat de les variàncies de n mostres, que apareixen tant al numerador com al denominador de l'estadístic. Per tant, es regeix per $F(gl_1, gl_2)$, d'on gl_1 s'obtenen a partir del número de mostres $k-1$; i gl_2 s'obté del número de mostres k pel número de persones en cada mostra ($k(n-1)$) els del denominador, tot tenint en compte que en aquest cas l'ordre sí que pot canviar el resultat de l'operació.

Les corbes de la distribució F es caracteritzen per:

1. Ser asimètriques, contràriament a la corba normal o la de la distribució t.
2. Adoptar només valors positius, sense probabilitats per baix de zero i màxim en infinit.
3. Tenir una densitat major a l'esquerre i una cua llarga a la dreta, amb el pic al voltant del valor 1.
4. En situacions d'igualtat de variàncies, F adopta un valor proper a 1, mentre que valors que s'hi allunyen assenyalen diferència de desviacions.
5. En augmentar els graus de llibertat, la corba que forma F s'assembla més a una distribució normal, distribuïda al voltant d'1.

L'estadístic F es calcula a partir de la següent fórmula, la paternitat de la qual es deu als estadístics Ronald Fisher i William Snedecor. En la fórmula es posen en contacte l'estimació interna de la variància en el denominador, que ve determinada per la següent fórmula en què es combinen les diferents variàncies de les k_i mostres:

$$S_w^2 = \frac{\sum(S_i^2)}{k}$$

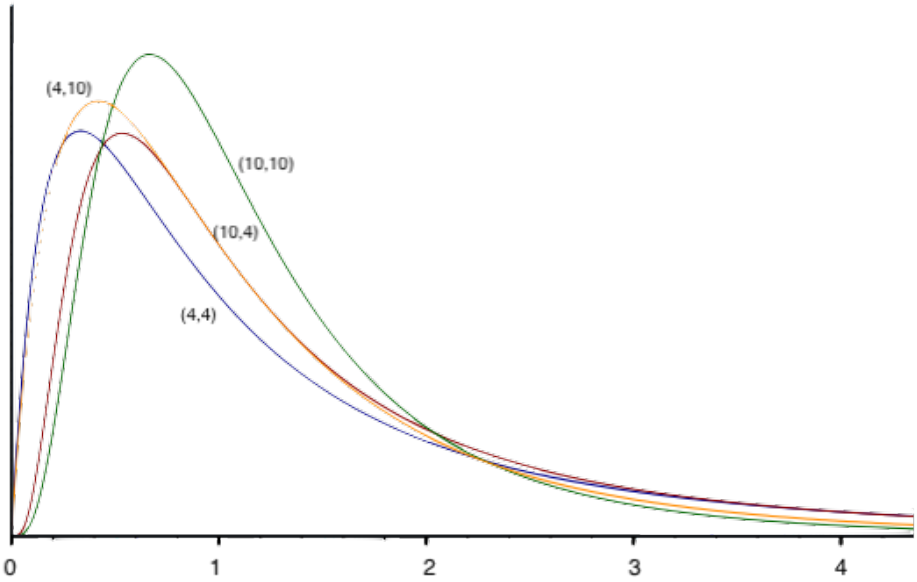
I, al denominador, l'estimació intermediant de la variància, que es calcula a partir del producte de les observacions (n) per la variància de les mitjanes ($S_{\bar{x}}^2$), mitjançant la següent fórmula:

$$S_x^2 = nS_{\bar{x}}^2$$

Aleshores, la fórmula de l'estadístic F serà la següent (Frankofrt-Nachmias i León Guerrero, 2018: p. 533 i ss.)

$$F = \frac{estimació\ intermediant\ de\ la\ variància}{estimació\ interna\ de\ la\ variància} = \frac{\sum_{i=k}^{n=1}(\bar{Y}_k - \bar{Y})^2}{\sum Y_i^2 - \sum \frac{(\sum Y_k)^2}{n_k}}$$

La distribució de densitats, com s'ha avançat anteriorment, mostra formes diverses en funció dels graus de llibertat implicats.



Gràfica 32. Distribució de F per a diferents combinacions de graus de llibertat

Font: Elaboració pròpia

La funció de probabilitat de F és:

$$f(x; gl_1, gl_2) = \frac{\sqrt{\frac{(gl_1 x)^{gl_1} {gl_2}^{gl_2}}{(gl_1 x + gl_2)^{(gl_1+gl_2)}}}}{x\mathrm{B}\left(\frac{gl_1}{2}, \frac{gl_2}{2}\right)}$$

On gl_1 i gl_2 són els graus de llibertat considerats i B és la funció beta.

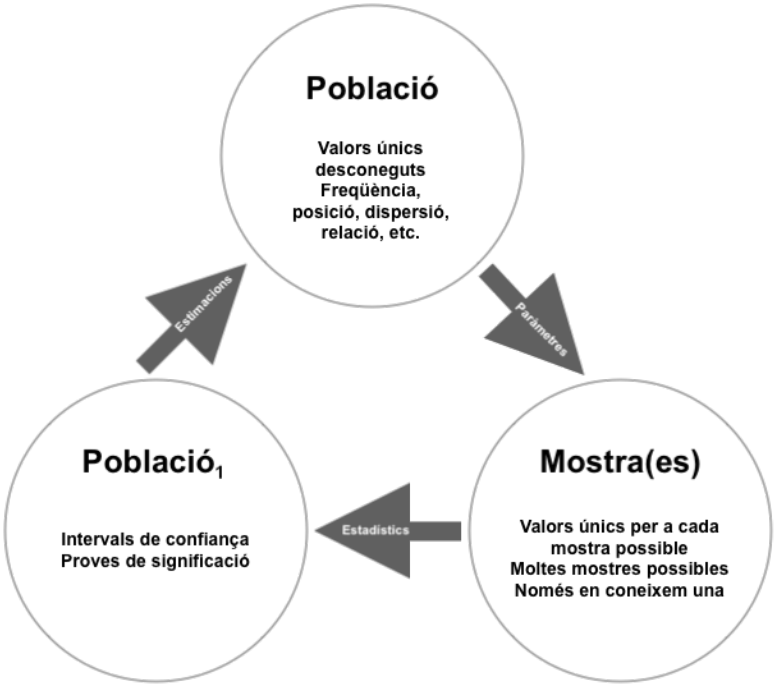
De la mateixa manera que amb la resta de funcions de densitat de probabilitat, en el cas de la distribució F es poden calcular les integrals dels principals creuaments de graus de llibertat, o es poden consultar els creuaments precalculats que figuren en annexos. També es poden consultar en annexos per a evitar fer els càlculs per a cada creuament

en què es vulga testar la igualtat de variàncies. En aquest cas però, per la gran quantitat de càlculs per a cada creuament de variables, es pot extraure una taula diferent per a cada nivell de significació, amb la qual cosa només s'exposa el més habitual, és a dir, $\propto = 0{,}05$.

2.4.3. Inferència a partir d'estimacions, intervals de confiança i proves de significació

Sobre els intervals de confiança

Els conceptes d'estimació i interval de confiança ens porten a considerar dos conceptes ja abordats anteriorment: la mitjana i la desviació típica. En concret, l'estimació ho fa a partir de posar en contacte el paràmetre poblacional amb l'estadístic mostral, amb l'objectiu d'obtindre un estimador que puga ser d'utilitat per conèixer la població objecte d'estudi. En el següent esquema reproduïm el procés inferencial amb allò que hem vist fins ara i com s'ajusta el concepte d'estimació:



Gràfica 33. Esquema del procés inferencial

Font: Elaboració pròpia

L'anàlisi amb deteniment d'una mitjana i la seua desviació típica, en el context d'una distribució mostral aleatòria i normal o aproximadament normal, ens ofereix un estimador que pot ser més o menys representatiu de la població d'on està extret. A aquesta consideració conjunta d'una estimació amb un marge d'error determinat se l'anomena interval de confiança i se sol simbolitzar com a IC. Tot i que alguns autors ja n'havien deduït la seua existència, no se sistematitza el seu ús fins a que Jerzy Neyman els introdueix en un article (Neyman, 1937), només tres anys després d'haver publicat un altre article central per a la teoria del mostreig, cosa que tractarem en el següent apartat (Neyman, 1934). El marge d'error, en aquest cas, ve determinat pel teorema central del límit, on com hem vist anteriorment, $\bar{x}$ pren els valors d'una N amb mitjana μ i desviació típica $\frac{\sigma}{\sqrt{n}}$. Aquesta desviació típica es pot interpretar en termes inferencials com a error típic $\sigma_{\bar{x}}$, un estimador que s'aproxima a la variància de la població, de tal manera que la desviació d'$\bar{x}$ es pot expressar, via quasidesviació, com a:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$$

O també, en el cas de les distribucions binomials (i per tant, per la via de les proporcions):

$$\sigma_{p} = \sqrt{\frac{p(1-p)}{n}}$$

Si tenim en compte que l'error és la distància entre el valor del paràmetre poblacional i l'estadístic mostral:

$$e = |\bar{x} - \mu|$$

Aleshores, combinant això amb la fórmula coneguda de tipificació obtenim que el valor Z estaria relacionat amb l'error mitjançant el nivell de confiança (Camarero et al., 2013: p. 235):

$$Z = \frac{x - \mu}{\sigma} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{e}{\frac{\sigma}{\sqrt{n}}}$$

Així, cada valor Z té associat un nivell de confiança que és la probabilitat de que la diferència entre estadístic i paràmetre siga menor que l'error e, que es pot calcular a partir del seu aïllament des de l'anterior fórmula:

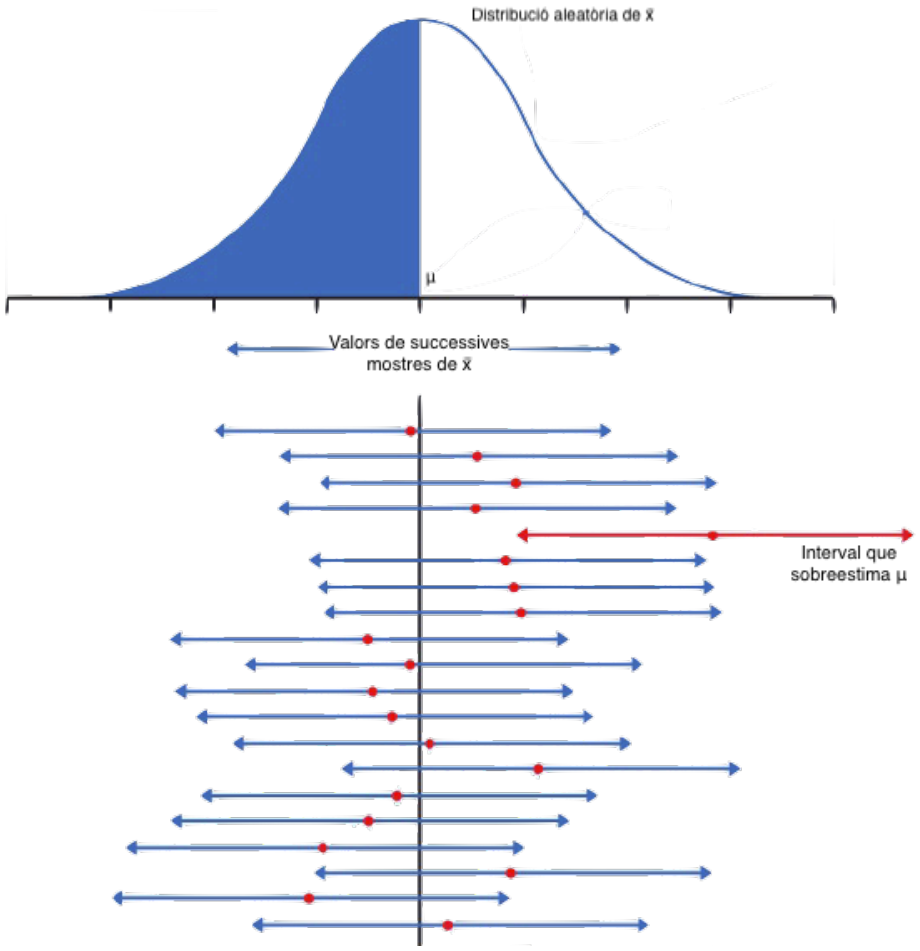
$$e = Z\frac{\sigma}{\sqrt{n}}$$

Que, en el cas de distribucions binomials, és a dir, per la via de les proporcions, quedaria de la següent manera:

$$e = Z\sqrt{\frac{p(1-p)}{n}}$$

El nivell de confiança aporta la probabilitat de que l'interval obtés en repetides mostres siga representativa del paràmetre poblacional. Com s'ha avançat anteriorment, el nivell de confiança més habitual no només en ciències socials sinó en investigació acadèmica en general, sol ser el 95%, la qual cosa es correspon amb 1,96 desviacions típiques o graus sigma, tot i que també se sol utilitzar el corresponent amb 95,5% o dues desviacions típiques o graus sigma. En combinar interval de confiança i nivell de confiança obtenim l'afirmació que sosté la investigació empírica quantitativa, en tant que dóna la seguretat estadística necessària per afirmar que un determinat estimador serà observable en la població, amb l'adició i la sostracció d'un determinat interval de, en un 95% o 95,5% de les mostres aleatòries simples possibles que s'extraguen d'aquesta població. Així es pot observar en la següent gràfica, on es pot apreciar que només una de les vint mostres extretes d'una població ofereix un valor diferent a la mitjana poblacional μ, mentre que els altres dèneu valors $\bar{x}$ estan dins de l'interval de confiança que marca la mitjana poblacional amb dues desviacions típiques[35] que és, en definitiva, l'aproximació de 95% en termes de z.

[35] En realitat, el valor de dues desviacions típiques o dos graus sigma σ es corresponen amb un 95,5% de l'àrea baix la corba normal tipificada, mentre que una àrea del 95% es correspondria amb 1,96 graus sigma.



Gràfica 34. Mitjanes i intervals de confiança respecte d'una mesura μ

Font: Elaboració pròpia a partir de Moore, 2007: p. 347.

Les proves de significació

L'estimació a partir de mitjanes i proporcions i l'establiment d'intervals de confiança són dues de les vies d'aplicació d'inferència sobre mostres per tal de calcular paràmetres poblacionals. Les proves de significació tracten d'aplicar la inferència basant-se en l'evidència de les dades mostrals sobre una afirmació al voltant d'un paràmetre poblacional. En realitat, és un raonament lògic semblant al dels intervals de confiança basat, això sí, en la concurrència de diferents mostres aleatòries simples i sota supòsits de normalitat, amb la finalitat d'avaluar dues hipòtesis: H_0 i H_1.

El procediment de formulació de proves de significació té l'origen en el primer terç del segle XX, concretament en el manual d'estadística de Ronald Fisher (1925). El propi Fisher explica en el prefaci a l'onzena edició de les catorze que s'han fet de la seua obra

en anglès que les proves de significació durant molt de temps necessitaven d'una disculpa prèvia. De fet, la proposta de situar el nivell a partir del qual s'assumeix o es declina una hipòtesi en $\alpha = 0{,}05$ es deu a la seua proposta d'una probabilitat suficientment forta en contra de la hipòtesi nul·la, en aquest cas, d'una contra vint, com hem vist anteriorment en la gràfica de $n = 20$ mostres sobre la dispersió de la mitjana. A partir de 1933, amb la publicació d'un article de Neyman i Pearson (fill), s'acaba de definir la formulació d'hipòtesis, la seua relació amb els errors I i II, especialment amb la minimització dels errors controlables de tipus II.

El primer pas en les proves de significació passa per llançar una hipòtesi. La hipòtesi, en estadística, és una afirmació sobre una població que pren forma de predicció de que un paràmetre adopta un valor numèric determinat o cau al voltant d'un rang de nombres determinats (Agresti, 2018: p. 139). La hipòtesi, en estadística, pot adoptar dos valors com hem avançat: la hipòtesi nul·la, H_0; i la hipòtesi alternativa o d'investigació[36], H_1. Potser contra el que dictaria el sentit comú, no examinem la hipòtesi d'investigació a partir de les proves de significació, sinó que ho fem contra la hipòtesi nul·la, tal com s'ha avançat a l'apartat introductori, de manera que s'intenten reunir suficients evidències en la seua contra (almenys una de cada vint mostres, ergo $p < 0{,}05$) per tal que la seua refutació siga robusta.

Seguint les indicacions d'Agresti i Healey, que proposen cadascú una aplicació de proves de significació amb cinc components -en algun cas diferents- podem esmentar les cinc passes per tal de formular una prova de significació[37] (Healey, 2016: p. 182; Agresti, 2018: p. 140 i ss.).:

a) assumpcions
b) hipòtesis
c) tria del model de distribució i valor-p[38]
d) càlcul de la prova
e) conclusió sobre la prova

[36] Com en tantes altres coses, diferents autories utilitzen diferents nomenclatures. Així, Agresti i Moore adopten H_a per a la hipòtesi alternativa (Agresti, 2018: p. 140; Moore, 2007: p. 366); Frankfort-Nachmias i Leon-Guerrero i Healey, en canvi, utilitzen H_1 per a la hipòtesi d'investigació (Frankfort-Nachmias i Leon-Guerrero, 2018: p. 363; Healey, 2016: p. 183). És significatiu que la notació moderna per la qual formulem les hipòtesis no es localitza, per exemple, en manuals com el de Fisher, on simplement s'esmenta la seua existència (Fisher, 1925).
[37] Moore, per exemple, organitza el procés al voltant de quatre passes: plantejament, planificació, resolució i conclusió (Moore, 2007: p. 372).
[38] Utilitzem ací l'expressió *valor-p* com a traducció directa de l'anglès *p-value*, on el guionet no té finalitat de sostracció, sinó de separació en termes sintàctics.

Tot seguit analitzem cadascuna de les parts per separat.

Les assumpcions prèvies inclús a la formulació de les hipòtesis estan relacionades amb els mínims per poder dur a terme la major part de les proves de significació.

1. Tipus de dades: cada tipus de dades demana un tipus de prova de significació diferent. Per tant, en funció de si tenim dades qualitatives o quantitatives, caldrà demanar una prova o una altra i, per tant, plantejar les hipòtesis en aquesta línia.
2. Aleatorització: la major part de proves d'inferència, com hem vist, es basen en l'existència d'un procés aleatori simple de selecció de la mostra.
3. Distribució de la mostra: algunes proves demanen que la variable sobre la qual es basen seguisca una distribució probabilística normal.
4. Mida de la mostra: per algunes proves cal que la mostra tinga una mida mínima, especialment en els casos en què la variable central per a la determinació de la prova s'allunya de la normalitat.

Pel que fa a les hipòtesis, un pas previ a l'aplicació de les proves de significació és expressar correctament especialment la hipòtesi nul·la, on s'argumenta l'absència d'efecte (no associació, no diferència, etcètera). Allò que pretén la investigació empírica és, a priori, el rebuig d'H_0 perquè tot el procés d'investigació està basat en la hipòtesi d'investigació H_1, cosa que implica l'existència prèvia de la formulació d'hipòtesi respecte al procés d'investigació. Com veurem a continuació, no s'argumenta igual un contrast d'hipòtesis per a una prova de significació basada en associació entre variables que una basada en diferència de mitjanes o de variàncies, tant pel que fa a H_0 com a H_1.

El tercer component per Agresti és la tria de la prova de significació, cosa que està relacionada amb el pas previ, la formulació de les hipòtesis, com també amb el pas posterior, la tria del *valor-p* (Agresti, 2018: p. 141). Healey, en canvi, proposa un tercer pas més complet, que inclou part del quart supòsit d'Agresti: la tria de la distribució probabilística i de la regió crítica α a partir de la qual fer la prova de significació. La tria del model de distribució de la mostra sobre la qual es farà la prova de significació està relacionada amb la prova estadística que se seleccionarà i també amb el tipus de dades, ambdós qüestions pertanyents a l'apartat d'assumpcions. No obstant, és significatiu que, arribats a aquest punt, s'ha de triar la regió crítica o regió de rebuig. Cada distribució

de probabilitats té una àrea per baix de la qual és possible rebutjar amb una certa consistència la presència d'una associació o diferència de mitjanes o variàncies. Aquest rebuig està relacionat, per tant, amb l'assumpció d'H_0. El punt a partir del qual es genera el rebuig d'H_0 i s'accepta H_1 ha de ser determinat en funció de la robustesa que li vulguem donar a la prova de significació i es determina pel nivell $\propto$ o pel *valor-p*[39]. Així, abans hem vist que el *valor-p* més habitual és 0,05, això és, una probabilitat de que només una de cada vint mostres aleatòries tinga un resultat a favor d'H_0. No obstant, es pot decidir situar el *valor-p* en marges de fiabilitat de la prova de significació més robustos. Per exemple, molt sovint es veuen càlculs amb un *valor-p* de 0,01 o fins i tot de 0,001. Evidentment, només els contrastos amb major evidència empírica seran capaços de passar tals nivells d'exigència, cosa que també dóna major seguretat a l'hora d'aplicar els seus resultats a qüestions que poden resultar vitals, com ara la medicina[40].

La quarta passa en el procés d'obtindre proves de significació es basa en el càlcul de l'estadístic que permetrà contrastar les hipòtesis plantejades en el primer apartat, a partir del *valor-p* de l'apartat anterior. Dins d'aquest apartat poden existir diferents possibilitats de càlcul[41], relacionades cadascuna amb la hipòtesi plantejada en el primer apartat i també de les possibilitats que plantegen les variables centrals en l'estudi. Per exemple, per a variables qualitatives, cap la possibilitat de calcular l'associació, la intensitat i la direcció, igual que per a variables ordinals; per a variables quantitatives, es pot calcular, a més de la correlació, també la igualtat o diferència de mitjanes i de variàncies. Paral·lelament, en determinades proves de significació s'ha de prendre una decisió pel que fa al plantejament de la hipòtesi, especialment en la igualtat de mitjanes. Les proves d'igualtat de mitjanes plantegen una H_0 en termes de desigualtat, per exemple, $H_0: \mu = 5$, on μ siga la mitjana de les notes dels estudiants d'un curs determinat. En aquest cas, la hipòtesi alternativa H_1 pot prendre tres formes:

- $H_1: \mu < 5$
- $H_1: \mu > 5$
- $H_1: \mu \neq 5$

[39] En realitat, nivell $\propto$ i *valor-p* representen les dues cares de la mateixa moneda, en tant que la primera expressa la proporció de l'àrea de la distribució probabilística inclosa en l'àrea de rebuig i el segon indica la probabilitat de que la prova iguale o sobrepasse el punt predit per H_1, sempre a partir de l'assumpció que H_0 és vertadera (Agresti, 2018: p. 141).

[40] Vegeu, per exemple, l'article seminal de Ioannidis sobre la falsació en el procés d'investigació (Ioannidis, 2005) o les recents demandes per les quals la biomedicina hauria de canviar el *valor-p* de 0,05 a 0,005 (Benjamin *et al*., 2018)

[41] En realitat, Healey planteja les cinc passes en el context de la tipificació (2016: p. 182), tot i que després aplica el model de les cinc passes a la resta de proves estadístiques que va plantejant.

En funció de quin dels tres plantejaments s'esculla, caldrà adoptar una forma o una altra de càlcul de l'estadístic, en aquest cas, de la t *de Student*. Però el mateix ocorre en altres plantejaments, com ara en el càlcul de les àrees per dalt o baix d'un valor z en una distribució probabilística normal. Els dos primers casos, com veurem a continuació, serien càlculs d'una cua: es vol calcular el valor per dalt o per baix d'un determinat valor de *t* en la corba de densitat, en aquest cas, 5, per a una mostra de n persones i k graus de llibertat. En canvi el tercer cas representa un càlcul de dues cues: es vol calcular el valor de t tant per dalt com per baix del punt determinat per la hipòtesi nul·la, novament per a una mostra de n persones i k graus de llibertat.

Tots estos càlculs, a més de basar-los en l'àmbit d'allò descriptiu, avancen ara cap a plantejaments inferencials en els supòsits que tot seguit veurem i, per tant, esdevenen ara mesures poblacionals per la via de l'estimació. Així, el *valor-p* marca per a cada estadístic la probabilitat que trobem el valor marcat per H_0 fora de l'àmbit d'influència de la cua que marca la significativitat estadística $\propto$ assumida amb anterioritat. Com hem vist, aquesta significativitat sol marcar-se amb $\propto < 0{,}05$, tot i que en ocasions es pot rebaixar el nombre a partir del qual rebutgem H_0 de manera que, tot i que parega un contrasentit, augmenta el nivell d'exigència: com més xicotet siga el *valor-p*, més forta és l'evidència envers H_0 (Agresti, 2018: p. 141).

Per últim l'última passa consisteix a establir les conclusions de la prova de significació a partir de la prova de significació escollida i prendre decisions al seu voltant. Així, la primera decisió que s'ha de prendre és la d'acceptar o rebutjar H_0, cosa que caldrà fer amb l'ajuda de l'estadístic escollit i del *valor-p* que tinga associat, en aquest cas a partir de la taula de probabilitats que corresponga i que hem deixat disponibles en l'annex. Només rebutgem H_0 quan el valor resultant a la taula de probabilitats siga $p < 0{,}05$. En tot cas, cal deixar palés quin és el resultat final en termes d'acceptació o rebuig d'H_0 i també quines conseqüències té això per a l'objecte d'investigació.

Inferència sobre proporcions (una i dues mostres)

A banda de la norma 68-95-99,7 cal recordar que la distribució normal estandarditzada concentra un 50% dels casos a cada banda del punt central representat per la mitjana, cosa que a més fa de manera simètrica, amb la qual cosa, recordant la gràfica sobre les

desviacions típiques sota la corba normal, podem deduir les proporcions que es troben a banda i banda o també per dalt i per baix d'una desviació típica determinada. Operar amb proporcions amb una distribució normal implica dur a terme allò que se sol conèixer com a aproximació de la binomial per la normal. A mesura que la mostra d'una distribució binomial esdevé més gran, la distribució resultant s'acosta a una distribució normal. Així, per a n mostres de x elements i una probabilitat p d'èxit, tindrem una normal de mitjana np i desviació típica $\sqrt{npq}$. Això ocorre generalment quan np és major[42] de 5. Aquesta operació és coneguda com a operació de continuïtat i té la funció de convertir una variable discreta en contínua, per tal de poder operar amb la distribució normal. En aquest cas, per tant, el valor discret x es convertirà en l'interval continu $[x - 0{,}5, x + 0{,}5]$, de manera que en comptes d'aproximar a x, per efecte de la correcció aproximem a $x + 0{,}5$ (La-Roca, 2006: p. 215 i s.).

a) Situacions amb una mostra

En els problemes d'estimacions sobre proporcions se sol donar, com a mínim, la probabilitat d'ocurrència d'un esdeveniment, de manera directa o indirecta, és a dir, amb percentatge o proporció de casos ($\hat{p}$) sobre el total de casos en la mostra. A més, també se'ns sol proporcionar la probabilitat que el mateix esdeveniment tinga lloc en la població de referència en termes binomials (p), amb la qual cosa obtenim el primer valor de l'equació per tal de calcular l'estimador: la desviació típica o l'error típic per a les proporcions. Val a dir que, en ocasions, no coneixem quina és la probabilitat que ocórrega l'esdeveniment en qüestió. En aquests casos cal agafar com a referència l'opció més conservadora, això és, que p i q, o el que és el mateix, que la probabilitat que ocórrega un esdeveniment i el seu contrari, s'igualen. Dit d'una altra manera, que $p = q = 0{,}5$. En cas que coneixem la desviació típica, hem d'obrar amb les dades que plantege el problema amb la fórmula següent:

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

Si no disposem de les dades per a la desviació típica, que sol ser el més habitual, hem d'aproximar el seu valor a partir de la fórmula de l'error típic, tenint en compte que el

[42] Camarero et al. (2013: p. 211) apunten a un np major de 5, mentre que altres autors assenyalen un límit més elevat (La-Roca, 2006: p. 215).

valor de $\hat{p}$ aproximarà a la normal si es donen les condicions necessàries (aleatorietat i mida gran de la mostra[43]):

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

En qualsevol dels dos casos, amb la desviació o l'error típic, el càlcul de l'interval de confiança s'extrau de la multiplicació amb el valor Z, i l'aplicació del valor resultant tant per dalt (interval de confiança superior) com per baix (interval de confiança inferior) de la proporció de casos amb èxit sobre el total de casos ($\hat{p}$). La fórmula, per tant, és:

$$IC = \hat{p} \pm Z\sqrt{\frac{p(1-p)}{n}}$$

La resolució dels intervals de confiança i les proves de significació pot prendre diferents camins: una de les possibles solucions és calcular l'interval de confiança per al nivell $\propto$ seleccionat i contrastar el valor de Z obtingut; una altra possible solució és calcular directament el valor de Z i contrastar-lo amb la taula normal tipificada; finalment, queda una possible solució que consisteix a calcular el valor de Z i contrastar-lo amb els valors crítics de z més habituals, que podem consultar en les darreres línies de la taula de probabilitats de la distribució normal tipificada. Així, en proves d'una cua per a un 90% de confiança el valor crític és d'1,282; per a un 95% de confiança és d'1,645; i per a un 99% de confiança és de 2,326. Aquests valors són també els que es poden consultar en la taula de la normal tipificada. També convé recordar tres de les propietats de la distribució normal estandarditzada: la primera és la norma que abans hem definit com a 68-95-99,7, la segona és la distribució simètrica que caracteritza aquestes corbes i la

[43] En qualsevol cas, es pot fer una prova de significació sobre l'aproximació normal, de manera que es comprova si un valor observat empíricament $\hat{p}$ és major, menor o diferent del valor esperat p_0, tot seguint la formulació d'hipòtesis següent:

$H_0: \hat{p} = p_0$
$H_{1a}: \hat{p} < p_0$
$H_{1b}: \hat{p} > p_0$
$H_{1c}: \hat{p} \neq p_0$

La resolució del problema passa per calcular el valor Z, que en aquest cas s'extrau de la fórmula anterior, i resulta aleshores que:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

tercera és el fet que la desviació típica total suma 1, i per tant podem aproximar-nos a una densitat de la corba desconeguda en oposició a la densitat de la corba coneguda. Aquestes indicacions ens facilitaran el càlcul d'estimacions quan els estadístics tinguen a veure amb aquestes quantitats prefixades.

D'altra banda, en proporcions amb mostres petites[44] s'introdueix l'interval de Wilson, pel qual s'ajusta el resultat del càlcul de la proporció fins i tot en mostres de menor grandària (Wilson, 1927).

$$IC = \frac{p + \frac{Z^2}{2n} + Z\sqrt{\frac{pq}{n} + \frac{Z^2}{4n^2}}}{1 + \frac{Z^2}{n}}$$

Ara veurem, tot seguit, alguns exemples relacionats amb el càlcul d'intervals de confiança. Donada una mostra aleatòria simple de 2.498 persones, s'ha trobat que 166 d'elles presentaven símptomes de coronavirus. Quina serà aquesta proporció en la població de referència per a un 99% de confiança? La resolució d'un problema com aquest passa pel càlcul, en primer lloc, per la comprovació de les condicions mínimes per al càlcul d'estimadors: que l'origen de les dades siga una mostra aleatòria i que la variable a considerar siga normal, o la grandària de la mostra permeta considerar-la com a tal. El segon pas passa pel càlcul de la proporció de la mostra sobre la qual volem establir els càlculs. En aquest cas, $p = \frac{166}{2498} = 0{,}0655$. Aquesta proporció s'observarà en la població amb un marge d'error $\sqrt{\frac{p(1-p)}{n}}$ que, en tot cas, està determinat per la confiança assumida, que en aquest cas, segons el problema formulat, és del 99%, és a dir, que deixa un 1% per cada costat, perquè en ser un interval de confiança deixa una àrea tant per dalt com per baix. Aleshores, ens hem de fixar en la z que deixa, 0,005% a banda i banda, o el que és el mateix, la que s'aproxima més a 0,995 en la taula estadística, que es troba entre 2,57 i 2,58 si l'expressem en graus z. Així, l'interval de confiança en la població sorgeix del càlcul de $0{,}0655 \pm 2{,}575\sqrt{\frac{0{,}0655*0{,}9345}{2498}}$. És a dir, que en la població trobarem una proporció equivalent a 0,0655±0,0128; [0,0527;0,0783] o el que és el mateix, en la població esperarem trobar entre 0,0527 i 0,0783 de malalts de coronavirus, que si ho expressem en termes percentuals és entre un 5,27 i un 7,83 per cent de la població.

[44] En aquest cas, el límit per a una mostra petita es considera que és aquell en què np_0 i $n(1-p_0)$ és igual a 10 o més unitats, amb la qual cosa s'assumeix que mostres de més de 10 individus o casos són suficients.

Un altre tipus de problema se'ns pot donar en situacions en què no es planteja una z coneguda, això és, en què no és cap dels valors 90, 95 ni 99, o fins i tot en cas que no recordem quina és l'àrea que conté la distribució normal per sota d'algun dels valors anteriors. La taula estadística ens pot ajudar en aquest sentit, sempre tenint en compte que els valors que conté fan referència als valors positius per sota de Z, i per això cal actuar amb cura[45]. Si agafem l'exemple anterior i canviem el 99% pel 94% de confiança, podem extraure el valor de Z anant a la taula estadística, des d'on abans haurem de recórrer a un raonament lògic: un 94% de confiança en deixa fora un 6% que, si el distribuïm en les dues àrees al voltant de la mitjana de la distribució de la mitjana de la mostra, és un 3% a cada extrem, superior i inferior, de la distribució. Aleshores, si busquem aquest 97%, o 0,97 en la taula, obtindrem els graus z corresponents al 94% de confiança, que coincideix amb el valor 1,88 de la taula. En considerar $z = 1{,}88$ estem tenint en compte un interval de confiança per dalt i per baix de la mitjana corresponent al 94% de confiança. Així, el càlcul queda de la manera següent: $0{,}0655 \pm 1{,}88\sqrt{\frac{0{,}0655*0{,}9345}{2498}} = 0{,}0655 \pm 0{,}0092;\ [0{,}0653; 0{,}0747]$.

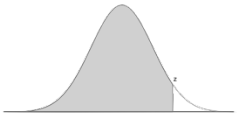
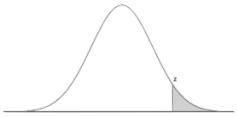
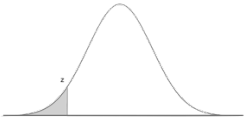
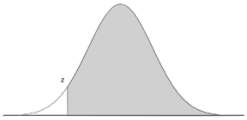
De manera semblant, els problemes poden obviar els intervals de confiança per posar damunt la taula les proves de significació i calcular les probabilitats per damunt, per sota o per damunt i per sota de Z, la qual cosa comporta considerar proves de significació d'una cua, en els dos primers casos, i de dues cues, en el darrer. Per tal d'aplicar aquests càlculs partim d'un valor de contrast, que ens sol oferir el problema, a partir del qual hem d'operar. La presentació de la taula normal tipificada mes habitual és la que dona els valors per sota de z, que és també la que s'adjunta en l'apartat d'annexos, i presenta el format següent:

$$P(Z \leq z)$$

En combinar la fórmula del càlcul d'intervals amb l'aproximació a la normal es poden extraure cinc situacions diferents, que s'extrauen de la taula normal tot just presentada. En tot cas, l'operativa amb la taula normal estarà relacionada amb les cues o regions crítiques amb què comparem els valors de Z per tal d'acceptar o rebutjar la hipòtesi. Així, càlculs per sota de z impliquen agafar les dades de la taula tal com apareixen en

[45] De la mateixa manera, el problema pot proporcionar-nos els graus z directament, a partir dels quals calcular l'interval de confiança.

l'annex, tret que Z siga un valor negatiu. Contràriament, els càlculs per damunt de z impliquen agafar el valor complementari a Z, sempre que no es tracte d'un valor negatiu. Finalment, en els casos de dues cues, s'opera sempre amb 2(1-P).

Per sota de z, quan Z és un valor positiu	$P(Z \leq z)$	
Per dalt de z, quan Z és un valor positiu	$P(Z > z) = 1 - P(Z \leq z)$	
Per sota de z, quan Z és un valor negatiu	$P(Z \leq -z) = 1 - P(Z \leq z)$	
Per dalt de z, quan Z és un valor negatiu	$P(Z > -z) = P(Z \leq z)$	
Entre dos valors de z	$P(z_1 < Z \leq z_2) = P(Z \leq z_2) - P(Z \leq z_1)$	

Gràfica 35. Fórmules i gràfiques per al càlcul de Z a partir de la z estandarditzada.

Font: Elaboració pròpia.

Així, en el problema anterior, en què havíem observat una $p = 0{,}0655$, quina és la probabilitat de trobar en l'univers de referència almenys un 5% de població amb coronavirus? A més de la p, tenim ara una $\hat{p}$, que ha de presentar la condició següent: $\hat{p} \geq 0{,}05$. Assumim que la nostra distribució p té una mitjana de 0,0655 i una desviació $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0{,}0655 * 0{,}9345}{2498}} = 0{,}0049$. Per contra, la probabilitat del valor que volem estimar presenta la forma següent $P(\hat{p} \geq 0{,}05) = P\left(\frac{\hat{p}-0{,}0655}{0{,}0049} \geq \frac{0{,}05-0{,}0655}{0{,}0049}\right) = P(Z \geq -2{,}9388)$. Tenint en compte la gràfica anterior, i atès que ens demana la probabilitat que $\hat{p}$ estiga per damunt de z i el valor que hem obtingut és negatiu, operem amb $P(Z \leq z)$, havent de substituir la part de dins del parèntesi pel valor de la taula estadística corresponent; en aquest cas, per a 2,94 tenim una probabilitat de 0,9984 que s'ha de llegir de la manera següent: si repetim la mostra diverses vegades, un 99,84% de les mostres tindran almenys un 5% de la població amb coronavirus.

El mateix exercici amb una $\hat{p} \leq 0{,}05$ té una resolució diferent perquè, en substituir la part de dins del parèntesi operem amb $1 - P(Z \leq z)$, de manera que la probabilitat serà la inversa (1-0,9984) i per tant 0,0016%, o el que és el mateix, un 0,2% de les mostres donaran valors per sota del 5% en coronavirus.

En el cas que aquest contrast es faça no introduint un criteri de major o menor que 0,05, sinó en relació amb la diferència amb el valor de referència, la comprovació ha d'anar per la via de les dues cues, que com hem vist en la gràfica anterior, suposa operar amb $2P(Z \leq z)$. En el cas proposat, el resultat per a la probabilitat de dues cues és $2(1 - 0{,}9984) = 2(0{,}0016) = 0{,}0032$, o el que és el mateix, en el 0,32% de les mostres el valor obtingut serà diferent de 0,05. Un darrer cas és el que trobem comprés entre els dos punts crítics, que hem de calcular amb la sostracció entre $\hat{p} \leq 0{,}05$ i $\hat{p} \geq 0{,}05$, que ens deixa observar els valors compresos entre els valors crítics, que en aquest cas és equivalent a $0{,}9984 - 0{,}0016 = 0{,}9968$, o el que és el mateix, en el 99,68% dels casos els valors estaran compresos dins dels valors crítics.

De manera inversa, es poden plantejar problemes en el sentit que allò que es pretén no és tant estimar el percentatge de mostres en què un valor és major, menor o diferent d'un altre valor de referència, sinó més aviat calcular la significativitat estadística d'aquesta diferència a partir d'un nivell de confiança definit per $\propto$ que habitualment se sol situar en $\propto = 0{,}05$ o també en $\propto = 0{,}01$. En definitiva, $\propto$ és qui marca el *valor-p* a partir del qual rebutgem o no H_0. Així, la hipòtesi nul·la es formula a partir de la igualtat de l'estadístic i el paràmetre, H_0: $p = p_0$, i les hipòtesis alternatives adopten alguna de les tres formes conegudes, $p < p_0$; $p > p_0$ o bé $p \neq p_0$, amb les mateixes precisions que s'han fet anteriorment. Això sí, per al contrast d'hipòtesis el que volem conèixer és ara la z, de manera que l'aïllem en la fórmula, que ara queda així:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

En aquesta ocasió, cal començar la resolució pel càlcul del valor crític de Z a partir del nivell de confiança que propose el problema o que considerem per a la seua resolució. Atès que la taula estadística de la probabilitat normal tipificada ens ofereix els valors per sota de z, el càlcul del valor crític s'ha de calcular a partir de $1 - z$. Així, per a 0,05, l'àrea que deixa el valor crític per sota és 0,95, la qual cosa es correspon amb els valors 1,64

a 1,65. Per a 0,01, el valor crític se situa entre 2,32 i 2,33. Aquests, i altres valors habituals en el càlcul de probabilitats a partir de la distribució normal tipificada, es poden trobar en les darreres files de la taula estadística de probabilitats de l'annex. Amb aquestes aproximacions resolem l'equació anterior i, en funció del plantejament de les hipòtesis i del resultat, operem amb el valor crític.

Si seguim amb l'exemple anterior, amb un nivell de confiança del 99%, es podria dir que la proporció de persones amb coronavirus en la mostra és significativament diferent d'un 5%, que és el que s'observa en altres entorns per a la incidència del mateix virus? Tal com està formulat el problema, assumim que H_0: $p = 0{,}05$ i que H_1: $p \neq 0{,}05$ i, per tant, resolem per la via de les dues cues. En aquest cas sabem que un nivell de confiança 99% $\propto = 0{,}01$, que dividit entre les dues cues suposa un 0,005 a cada costat, és a dir, una $Z = 2{,}57$, contra la qual haurem de contrastar amb la fórmula anterior que, una vegada resolta, dona 3,5227, més enllà de la zona de rebuig de la hipòtesi nul·la; per tant, hem de rebutjar la hipòtesi nul·la H_0 i acceptar la hipòtesi d'investigació o alternativa H_1.

b) Situacions amb dues mostres

En situacions en què es posen en contacte dues mostres per comparar les seues proporcions, cal tenir en compte que, per tal d'aplicar els intervals de confiança i les proves de significació, cal que es done amb caràcter previ tres condicions ja conegudes: a la qüestió de la normalitat, a més, sumem que les mostres estiguen seleccionades de manera aleatòria i que siguen independents l'una de l'altra. En aquest cas, la inferència s'avalua a partir de la diferència entre les proporcions en la població $p_1 - p_2$, per a la qual cosa s'utilitza com a estimador la diferència entre les proporcions en les mostres $\hat{p}_1 - \hat{p}_2$. Aleshores, la distribució normal de $\hat{p}_1 - \hat{p}_2$ pren ara una mitjana $(p_1 - p_2)$ i amb una desviació típica:

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

En situacions de mostres de mida suficient[46] la desviació típica s'aproxima a partir de l'error típic de $\hat{p}_1 - \hat{p}_2$, amb la fórmula següent:

[46] En aquest cas, una mida suficient és aquella en què les n mostrals combinades sumen més de 100 individus o casos (Healey, 2016: p. 219).

$$e = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

I com en l'apartat anterior, l'interval de confiança es calcula a partir de:

$$IC_{\hat{p}_1-\hat{p}_2} = Z \pm e_{\hat{p}_1-\hat{p}_2}$$

Pel que fa a les proves de significació, la fórmula resultant per al càlcul de Z és la següent:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c(1-\hat{p}_c)\frac{1}{n_1} + \frac{1}{n_2}}}$$

D'on, en aquest cas, $\hat{p}$ és el resultat de la probabilitat combinada d'ambdues mostres, que hem de calcular a partir de la fórmula següent:

$$\hat{p}_c = \frac{número\ d'èxits\ en\ les\ dues\ mostres}{número\ d'individus\ en\ les\ dues\ mostres}$$

Novament, cal plantejar les hipòtesis, que s'han d'aplicar sobre la mateixa $\hat{p}$ i, per tant, s'ha de reproduir l'operativa de la diferència de proporcions per a una mostra.

Per exemple, contraposem ara la proporció que hem estudiat $p_1 = 0{,}0655$ (166 casos positius sobre 2.498) contra la proporció d'una altra mostra -seleccionada de manera aleatòria- amb 2.249 persones i una $p_2 = 0{,}0725$ (163 casos positius sobre el total), suposant que en la segona població la proporció és significativament superior i té un nivell de confiança $\propto = 0{,}05$. Aleshores, H_0 suposa que les dues proporcions són iguals, i per contra la hipòtesi alternativa aposta per una relació basada en una única cua $H_1: p_2 \geq p_1$. Aleshores, Z s'ha de resoldre a partir de l'equació anterior, i així obtindrem que $\hat{p}_c = 0{,}0693$.

$$Z = \frac{0{,}0721 - 0{,}0655}{\sqrt{\frac{0{,}0693 * 0{,}9307}{2498} + \frac{0{,}0693 * 0{,}9307}{2249}}} = \frac{0{,}0066}{0{,}0074} = 0{,}8919$$

D'on deduïm, a través de la taula estadística per a la normal tipificada, que $P(Z \geq 0{,}8919) = 1 - 0{,}8133 = 0{,}1867$. Aquest valor no es troba inclòs dins l'àrea de rebuig,

representada per $\propto = 0{,}05$, que demana valors per sobre de 1,64, i per tant podem rebutjar que p_2 siga significativament superior a p_1.

La mateixa prova per a dues cues ofereix un resultat semblant. En aquest cas, mantindrem H_0, però no la hipòtesi alternativa, que queda $H_1: p_2 \neq p_1$. La resolució de l'equació és la mateixa, i per tant tindrem que $P(Z \neq 0{,}8919) = 0{,}1867$, que novament no està inclòs dins l'àrea de rebuig representada pel valor $\propto = 0{,}05$, i que en aquest cas se situa en 1,96.

Inferència per a mitjanes (una i dues mostres)

De la mateixa manera que hem vist en la inferència per a proporcions, hi ha la possibilitat de dur a terme càlculs inferencials per a mitjanes, tant per a una mostra, com per a dues mostres. En aquest sentit, no solament s'apliquen diferents fórmules, sinó també s'introdueix una nova distribució en el càlcul de la prova de significació, en aquest cas la *t de Student*.

a) Situacions amb una mostra

En aquests casos, cal aplicar la fórmula de tipificació per tal de convertir els valors a valors z, i a partir d'aquests obrar amb la taula de valors de la normal. Per això, però, sol ser necessari conèixer el valor de la desviació típica poblacional σ_x, que és un dels valors que sol oferir el problema. D'aquesta manera, coneguda la desviació típica, assumint que es tracta d'un mostreig aleatori simple i que la distribució s'aproxima a la normal, podem aplicar la fórmula per deduir el valor de Z.

$$Z = \frac{x_i - \bar{x}}{\sigma_x}$$

En funció del que demane el problema, ens podem trobar amb cinc situacions, que estan donades pel format de la taula amb què ens trobem. Per tal de poder utilitzar una taula estadística de la normal tipificada com la que apareix als annexos, cal tenir en compte que els problemes que es puguen plantejar tindran una resolució o una altra en funció del tipus de pregunta que generen i si la quantitat que es desitja estimar està per dalt, per sota o entre dos valors de z, tal com hem vist en l'apartat anterior.

Per acabar, cal tenir en compte que en presència de mostres de mida reduïda, que ja hem vist anteriorment que és un concepte relatiu però que podria estar al voltant de les 120 persones com a límit, s'aplica la distribució t de Student. En realitat, es pot aplicar per a qualsevol tipus de mostra, tot assumint prèviament la condició de normalitat, perquè la distribució t de Student s'adapta millor que la normal tipificada a les mostres de dimensions reduïdes (Agresti, 2018: p. 114).

De manera paral·lela a l'estimació, es pot dur a terme el càlcul dels intervals de confiança per a mitjanes poblacionals, cosa que passa novament pel càlcul de l'error típic, que en aquest cas, com hem vist anteriorment, segueix la fórmula següent:

$$e = Z\frac{\sigma}{\sqrt{n}}$$

En la major part dels problemes hi ha mostres amb una N elevada i una desviació poblacional σ coneguda. Això deixa els càlculs en la mera substitució de factors, començant pel nivell de confiança Z.

$$IC = \bar{x} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$$

Si la desviació poblacional σ és desconeguda, com en el cas d'intervals de proporcions en què cal triar el cas més conservador ($p = q = 0{,}5$), llavors s'utilitza la desviació típica S i, en comptes de l'arrel quadrada de la mostra n, escollim el denominador de la quasidesviació ($n - 1$) de la manera següent (Healey, 2016: p. 158):

$$IC = \bar{x} \pm Z\left(\frac{S}{\sqrt{n-1}}\right)$$

D'igual manera que hem vist anteriorment, si s'estableixen intervals per a mitjanes amb n de mida limitada, en comptes d'escollir Z com a estimador del nivell de confiança hem d'agafar la t de Student, típicament amb un nivell de confiança 0,05 per a $n - 1$ (Camarero *et al.*, 279) i una quasivariància que es calcula a partir de la fórmula:

$$S_{n-1} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Així, el càlcul dels intervals de confiança han de seguir la formulació següent:

$$IC = \bar{x} \pm t_{n-1}^{0,05} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N}}$$

Un exercici molt habitual en el càlcul d'inferència sobre la mitjana consisteix a calcular el percentatge de població que compleix unes determinades condicions respecte de la mitjana. Així, un primer pas és calcular els percentatges, en un univers, atesa una distribució normal. Vegem, per exemple, l'edat a la qual abandonen la llar els joves a l'estat espanyol. De mitjana, aquesta edat se situa als 29,5 anys amb una desviació típica de 4,3 anys. Quina és la probabilitat que algú s'independitze abans dels 25 anys? El primer que cal fer és calcular les unitats que hi ha entre el punt que representa la mitjana i el de l'edat que volem esbrinar a partir de la primera fórmula plantejada en l'apartat. Així, $Z = \frac{25-29,5}{4,3} = -1,0465$. Tenint en compte el que hem vist anteriorment sobre com treballar amb la taula de distribució de la normal tipificada, el que volem conèixer és l'àrea que queda per sota del valor Z, cosa que equival a calcular l'àrea que queda per damunt del valor corresponent en positiu, és a dir, $z = 1 - p(Z \leq z) = 1 - 0,8531 = 0,1469$. Això mateix es podria fer per calcular les probabilitats entre dos punts, per exemple la població que s'independitza entre els 25 i els 27 anys. Aprofitem el primer càlcul i resolem el segon $Z = \frac{27-29,5}{4,3} = -0,5814$ que, seguint l'exemple anterior, es correspon amb $= 1 - p(Z \leq z) = 1 - 0,7190 = 0,281$. Així, la distància és el resultat de la sostracció de les àrees d'ambdues edats: 0,1341, és a dir, el 13,41% dels joves.

Els càlculs, en el cas de mostres aleatòries, s'han de fer, com hem vist, a partir de l'error típic. Així, si en una mostra sobre 897 valencians i valencianes hem obtingut una mitjana de 28,7 anys, ateses les condicions de normalitat, aleatorietat i independència, podem fer els mateixos càlculs sobre la distribució normal tipificada; ara, però, en comptes d'agafar el valor de la desviació poblacional directe, l'aproximem a partir de l'error típic. Per exemple, per a la mostra que s'acaba de definir, l'interval de confiança per a $\propto = 0,05$ és $IC = \bar{x} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right) = 28,7 \pm 1,960\left(\frac{4,3}{\sqrt{897}}\right) = 28,7 \pm 0,2815; [28,4185; 28,9815]$.

Pel que fa a les proves de significació, la seua resolució és molt semblant a la que ja hem vist per a les proporcions i segueix les mateixes passes, tot i que en comptes de contrastar amb la taula de la distribució normal tipificada, ho hem de fer sobre la taula

de distribució t, ja que l'estadístic que resulta de considerar la desviació mostral s per la poblacional σ torna un resultat que no s'ajusta a la distribució normal (Moore, 2007: p. 435). Així, les fases per les quals hem de passar són: en primer lloc, plantejament de les hipòtesis nul·la i alternativa; després, càlcul de t; comparació amb el nivell de significació; i resolució del plantejament hipotètic. En aquest cas, es mesura la diferència en termes d'una cua o dues cues del paràmetre d'una mostra $\bar{x}$ respecte de l'estimador de la població μ, cosa que podem deduir de la combinació de la fórmula per al càlcul de la probabilitat de la normal tipificada i de l'error típic, de manera que:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Així, per a la mostra anterior basada en dades sobre valencians, podem dir que és diferent de la mesura poblacional considerada anteriorment, per a un nivell de confiança del 95%? Hem d'assumir, llavors, que $H_0: \bar{x} = \mu$ i $H_1: \bar{x} \neq \mu$, per a $\propto = 0{,}05$. Obtindrem que $t = \frac{28{,}7-29{,}5}{\frac{4{,}3}{\sqrt{897}}} = \frac{-0{,}8}{0{,}1436} = -5{,}5710$. Com que el contrast es resol per la via d'una distribució t, cal extraure els graus de llibertat, que en aquest cas s'extrauen de n-1, és a dir, 896 o, en la taula de referència, el valor que més s'hi aproxime. Atès que la diferència entre $n = 120$ graus de llibertat i $n = \infty$ és mínima, apostem per l'última, en la qual el valor crític a l'hora d'acceptar o rebutjar H_0 per a dues cues és 1,960. Així, ja que t excedeix el valor crític, podem assegurar que hi ha diferències significatives entre mostra i població. L'interval de confiança per a aquesta operació està constituït per $28{,}7 \pm 1{,}960 \frac{4{,}3}{\sqrt{897}} = 28{,}7 \pm 0{,}5515; [28{,}1485; 28{,}4185]$.

b) Situacions amb dues mostres

En situacions de dues mostres, com hem avançat anteriorment, la prova de significació no es fa sobre la distribució normal tipificada, sinó a partir de la distribució *t de Student*. Com en els casos anteriors, per poder utilitzar aquest tipus de prova cal assumir l'aleatorietat de les mostres que es desitja comparar, a més de la normalitat i, en aquest cas, també la independència de les mostres, o el que és el mateix, que una mostra no influeix sobre l'altra. Un cas prototípic de mostres relacionades, i per tant dependents, és aquell en què es mesura un paràmetre en una mostra, s'administra un tractament i es torna a mesurar el paràmetre per tal d'observar-ne els canvis. Amb aquestes condicions, podem fer inferència de la diferència entre $\mu_1 - \mu_2$ a partir de $\bar{x}_1 - \bar{x}_2$. Novament, cal calcular la

desviació típica, que en cas que estiguen disponibles les dades poblacionals s'extrau de la fórmula següent:

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Com sol ser habitual, no disposem de les desviacions típiques de la població de referència, per la qual cosa s'estima a partir de l'error típic, amb la fórmula següent:

$$e_{x_1 - x_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

L'interval de confiança, aleshores, s'obté de la fórmula següent[47]:

$$(x_1 - x_2) \pm t \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Posem per cas que un estudi ha plantejat el nivell de confiança de l'electorat sobre dos periòdics valencians, que anomenarem EMV i LP, a partir d'una escala de 7 valors. EMV ha obtingut $\bar{x}_1 = 4{,}77$ i LP $\bar{x}_2 = 2{,}43$, amb unes mostres de $n_1 = 76$ i $n_2 = 78$ persones respectivament, i unes desviacions típiques de $s_1 = 0{,}2$ i $s_2 = 0{,}35$. Quin és l'interval de confiança? Per a la primera part de l'equació tenim que $(x_1 - x_2) = 2{,}34$; mentre que per a la segona tenim que $\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = \sqrt{\frac{0{,}2^2}{76} + \frac{0{,}35^2}{78}} = 0{,}0458$. Comprovem el valor en la taula de distribució de t per a dues cues (ja que necessitem un interval per sobre i per sota) per a $gl = 75$. Com que en la taula de distribució de t no hi ha el valor en qüestió, ens hi aproximem tant com podren, en aquest cas a $gl = 80$, d'on extraiem que el valor de t per a un 95% de confiança i dues cues és d'1,990. Així, l'interval de confiança és $2{,}34 \pm 1{,}990 * 0{,}0458 = 2{,}34 \pm 0{,}0911$, o el que és el mateix, [2,2489; 2,4311]. Aleshores, l'estimador de la diferència serà 2,34, d'on la diferència de mitjanes se situa entre 2,2489 i 2,4311. Com que la mitjana d'EMV supera de manera ampla l'interval, es pot dir que hi ha diferències entre les dues mitjanes.

[47] Tot i que alguns manuals utilitzen una aproximació per la via dels graus z per a mostres amb una grandària suficient, ací apostem per centrar els càlculs per la via de t, ja que la diferència en el contrast és mínima.

Val a dir que l'estadístic t per a dues mostres s'aproxima a una distribució t de Student, tot i que no és exacte. Per aproximar-nos al seu valor, cal utilitzar una aproximació als graus de llibertat de la distribució, que, com hem vist anteriorment, s'extrau[48] a partir del menor dels resultats de $(n_1 - 1)$ i $(n_2 - 1)$. Després d'estandarditzar, obtenim l'estadístic *t de Student* per a dues mostres, que actua ara en substitució de Z i que té la fórmula següent:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

I així, es pot seguir l'operativa que s'ha plantejat anteriorment: o bé es calculen els intervals de confiança i posteriorment es contrasta amb el valor t obtingut; o bé després de calcular el valor de t es contrasta amb la taula de probabilitats que s'adjunta en annexos i que, en aquest cas, s'organitza a partir dels graus de llibertat. S'ha optat per adjuntar la taula de probabilitats per a una cua i dues cues, que és el format en què Fisher i Yates la presenten en la seua monografia sobre taules estadístiques (1974), i pot resultar més interessant per a l'estudiantat, no només perquè no cal calcular les dues cues, sinó també perquè es pot veure gràficament com es comporta el nivell de significació $\propto$ en ambdues situacions. En els casos de dues cues se segueix el procediment de multiplicar el valor de t per les dues cues, amb la qual cosa s'obté la probabilitat d'obtenir un valor tant per sobre de t com per sota de $-t$.

Sobre l'exemple dels periòdics EMV i LP que acabem d'analitzar, es pot dir que la valoració d'EMV és significativament superior amb un nivell de confiança del 99%? Comencem, com en les ocasions anteriors, per plantejar les hipòtesis a partir de l'enunciat, d'on es desprèn que $H_0: \bar{x}_1 = \bar{x}_2$ i $H_1: \bar{x}_1 > \bar{x}_2$. La resolució de t es fa a partir de l'equació plantejada anteriorment: $t = \frac{4{,}77 - 2{,}43}{\sqrt{\frac{0{,}2^2}{76} + \frac{0{,}35^2}{78}}} = \frac{2{,}34}{0{,}0455} = 51{,}4286$. Tot seguit, comprovem en la taula de distribució de t el valor crític per a una mostra en què els graus de llibertat estan determinats pel menor dels valors entre $(n_1 - 1)$ i $(n_2 - 1)$, que en aquest cas és 75 o

[48] En canvi, quan es calculen els graus de llibertat per a $x_1 - x_2$ amb l'ajuda de programari estadístic, sol passar que els graus de llibertat no són exactament els mateixos que s'ha calculat amb aquesta aproximació. Per tal de dur a terme un càlcul exacte sobre la distribució t, caldria calcular els graus de llibertat a partir de la fórmula següent: $gl = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}{\frac{1}{n_1 - 1}\left(\frac{S_1^2}{n_1}\right) + \frac{1}{n_2 - 1}\left(\frac{S_2^2}{n_2}\right)}$

el valor que més s'hi aproxime, que és més o menys $n = 60$. El contrast és significatiu per a una cua i per al nivell de confiança proposat ($\propto = 0{,}01$), en què la regió de rebuig està marcada per $t = 2{,}390$. Per tant, podem dir per aquesta via que la mitjana d'EMV és significativament superior a la de LP.

Val a dir que, quan utilitzem programari estadístic per calcular t, ens torna un quadre de dues files que inclou dos valors de t, a més d'un contrast F de les variàncies. Pel que fa a la primera dada, cal dir que el procediment que s'ha seguit ací ha estat el de variàncies diferents. El procediment oposat és el d'igualtat de variàncies, pel qual s'extrau la mitjana de les variàncies d'$\bar{x}_1$ i $\bar{x}_2$ i s'extrau l'estadístic t de dues mostres agrupades (Moore, 2007: p. 476). D'un altra banda, el contrast F és el resultat d'aplicar la fórmula següent:

$$F = \frac{major\ variància}{menor\ variància}$$

A partir d'aquesta fórmula s'extrauen els graus de llibertat de cadascuna de les mostres, considerant que els graus de llibertat del numerador (gl_1) apareixen primer en la distribució i, a continuació, el denominador (gl_2), ja que l'ordre d'aparició de les variables en la distribució F té efectes sobre el seu resultat final. Així, en aquest cas la presentació ha de ser $F = (gl_1, gl_2)$. La prova F suposa considerar dues hipòtesis, en què H_0 implica igualtat de les desviacions típiques σ_1 i σ_2 i H_1 implica diferència entre aquestes mateixes desviacions. Es tracta d'una prova molt sensible a la no normalitat de les variables considerades, per la qual cosa s'aconsella no usar-les en inferència (Moore, 2007: p. 477)

Inferència sobre relacions de variables (I): khi quadrat

La inferència no solament es pot utilitzar per posar a prova mostres sobre poblacions o mostres entre si, sinó també, com és el cas de les proves que presentem tot seguit, entre variables per tal d'estudiar si el que s'ha observat en l'apartat descriptiu és estadísticament significatiu, tant en els casos de mostres, com en els de poblacions. En aquest cas, agafem com a hipòtesi nul·la la igualtat entre els casos observats i casos esperats i, en contraposició, la hipòtesi d'investigació de diferència entre els casos observats i casos esperats, això és, la situació de dependència.

$$H_0: O_i = E_i$$

$$H_1: O_i \neq E_i$$

De la mateixa manera que s'ha vist anteriorment, per tal de dur a terme una prova de significació a partir de khi quadrat, cal que les mesures obtingudes estiguen extretes de manera aleatòria i siguen independents entre si. Aleshores, la hipòtesi nul·la es pot entendre com l'absència de relació entre les variables estudiades i, per contra, la hipòtesi d'investigació es planteja com l'afirmació de la relació.

Un dels primers càlculs que cal dur a terme és el dels graus de llibertat, que com s'ha vist anteriorment, és el resultat del producte de les categories de la variable en files menys un i les categories de la variable en columnes menys un. A partir d'aquest paràmetre i amb el resultat de χ^2 extret de la seua fórmula:

$$\chi^2 = \sum_{i=1}^{i} \sum_{j=1}^{j} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

només s'ha d'anar a la taula estadística per comprovar si el valor es troba dins de l'àrea crítica d'acceptació d'H_0, tot partint del mateix nivell $\propto$ plantejat en els anteriors problemes, és a dir, $p < 0{,}05$.

L'alternativa, com s'ha vist anteriorment, és calcular l'interval de confiança de khi quadrat per comprovar si el valor resultant en el càlcul del paràmetre s'hi troba dins o no. En aquest cas, la fórmula de l'interval de confiança és la següent (hi podem percebre que el que és rellevant en la formulació no és tant la mitjana com la variància):

$$\chi^2_{\frac{1-\alpha}{2}} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2}}$$

Agafem com a exemple el mateix exercici que hem desenvolupat anteriorment, en què $\chi^2_A = 1{,}6667$. Per a un nivell de confiança $\alpha = 0{,}05$, el valor crític per als graus de llibertats associats a la parella de variables $gl = (k_1 - 1)(k_2 - 1) = (2-1)(2-1) = 1$ s'extrau de la taula de distribució de khi quadrat, i així podem deduir que l'associació entre les variables no és estadísticament significativa, i per tant no podem rebutjar H_0. Perquè ho siga, hauria de tenir un valor major del valor crític, que per a $\alpha = 0{,}05$ i $gl = 1$ és de 3,841.

Inferència sobre relacions de variables (II): regressió

La inferència a partir de la regressió es basa en alguns supòsits ja vistos anteriorment: en primer lloc, l'aleatorització; en segon lloc, el fet que la mitjana d'y està relacionada amb x per la via de l'equació $y = a + b\,x$; en tercer lloc, que la desviació típica condicional σ és la mateixa per a tots els valors de x; i per últim, que la distribució de y per a cada valor de x és normal (Agresti, 2018: p. 267).

A partir del que s'ha pogut veure anteriorment al voltant de la correlació i la regressió, es pot deduir que cal estimar tres paràmetres a partir de les dades de correlació, que són els components de la recta de regressió en la població, és a dir, $\mu_y = \alpha + \beta x$, en què la inclinació de la recta b actua com a estimador de la pendent de la recta en la població β, i els punts a en la línia dels mínims quadrats actuen com a estimador dels punts en la població α.

De manera genèrica, es pot dir que una recta de regressió en la qual el valor de la pendent (β) siga 0 indica l'absència de relació linear, en la mesura que apunta a una horitzontalitat que es tradueix en el fet que no es produeixen canvis en la mitjana de y quan x té diferents resultats. Aleshores, a efectes de proves de significació, la formulació d'hipòtesis quedarà de la manera següent:

$$H_0: \beta = 0$$
$$H_{1a}: \beta > 0$$
$$H_{1b}: \beta < 0$$
$$H_{1c}: \beta \neq 0$$

Per tal de comprovar la hipòtesi de la significació de l'associació linear, partim de la fórmula següent, semblant a la del càlcul d'una prova t o z:

$$t = \frac{b}{e_b}$$

En què b és l'estimador de β i e_b és l'error estàndard dels mínims quadrats, que en el cas del valor de la pendent s'extrau a partir de la fórmula següent:

$$e_b = \frac{s}{\sqrt{\sum (x - \bar{x})^2}}$$

Es pot dir que com més petita siga s, que actua com a estimador de la desviació típica de les distribucions condicionals de y, més precisa serà l'estimació que fa b de β.

D'altra banda, la desviació típica σ s'estima a partir de l'error típic de la regressió s, que ocupa el lloc del numerador en el càlcul de l'error estàndard de l'error de b, i que s'obté mitjançant la desviació típica dels residus:

$$s = \sqrt{\frac{1}{n-2}\sum(y-\hat{y})^2}$$

En què $n-2$ actua com a graus de llibertat per a la t, que es calculen a partir de $gl(n-2)$, atès que si férem el càlcul de tots els residus, quan en quedaren $n-2$ podríem predir els dos valors que faltaria per calcular. A l'efecte de la comprovació en la taula d'estadístics, s'hi nactua de la mateixa manera que en la comparació de mitjanes, és a dir, una cua per a valors de H_1 majors o menors que 0 i dues cues quan els valors a contrastar són diferents de 0. De fet, en cas que s'accepte H_0 s'està donant per fet que no hi ha correlació entre les variables estudiades, de manera que no caldria fer els càlculs de regressió per a l'encreuament de les variables implicades. La comprovació es pot fer automàticament a partir de la taula estadística que s'adjunta als annexos, en la qual podem trobar els valors crítics a partir dels quals la correlació és significativa per a mostres de mida n i k graus de llibertat.

Per últim, el càlcul dels intervals de confiança es faria a partir de:

$$b \pm t * e_b$$

Recuperem l'exemple iniciat pàgines enrere, en el qual havíem arribat a extraure la fórmula de regressió dels seguidors d'Instagram i la valoració de l'aparença personal, en què la recta venia definida per $y = -92{,}2713 + 33{,}4805x$ i ja teníem els valors de la mitjana de x, $\bar{x} = 6{,}6$; la mitjana de y, $\bar{y} = 128{,}7$; i també els valors de la covariància $S_{xy} = 120{,}5333$ i de la correlació $r_{xy} = 0{,}6819$. Plantegem com a hipòtesis $H_0: \beta = 0$ i $H_0: \beta \neq 0$; per tant, un contrast de dues cues, i un nivell de confiança del 95%.

Així doncs, el primer càlcul hauria de ser el de l'error típic de la regressió a partir de les desviacions dels mínims quadrats.

y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
40	-88,7	7867,69
5	-123,7	15301,69
89	-39,7	1576,09
1	-127,7	16307,29
204	75,3	5670,09
175	46,3	2143,69
231	102,3	10465,29
180	51,3	2631,69
108	-20,7	428,49
254	125,3	15700,09

$$\sum = 78092{,}1$$

$$s = \frac{78092{,}1}{10 - 2} = 9761{,}5125$$

Taula 41. Càlcul de l'error de la regressió per a les variables seguidors d'Instagram (x_i) i satisfacció amb l'aparença personal (y_j)

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

Amb la dada de l'error típic de la regressió podem resoldre l'error típic dels mínims quadrats, que es calcula a partir de la fórmula en què ja s'inclou l'error típic de la regressió. Així:

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
8	1,4	1,96
3	-3,6	12,96
7	0,4	0,16
4	-2,6	6,76
8	1,4	1,96
7	0,4	0,16
6	-0,6	0,36
8	1,4	1,96
6	-0,6	0,36
9	2,4	5,76

$$\sum = \sqrt{32{,}4} = 5{,}6921$$

$$e_b = \frac{9761{,}5125}{5{,}6921} = 1714{,}9229$$

Taula 42. Càlcul de l'error típic dels mínims quadrats per a les variables seguidors d'Instagram (x_i) i satisfacció amb l'aparença personal (y_j)

Font: Elaboració pròpia a partir del baròmetre del CIS 3201

L'últim pas és el càlcul de l'estadístic t a partir de la primera fórmula d'aquest subapartat. Per al nostre exemple, $t = \frac{b}{e_b} = \frac{-92{,}2713}{1714{,}9229} = 0{,}0537$ que, per a dues cues, en la taula de probabilitats de la distribució t i $gl = 8$ té un valor crític de 2,306. Per tant, no hi ha evidència d'una relació linear entre l'aparença personal i el nombre de seguidors en Instagram.

De manera paral·lela, es pot comprovar la hipòtesi sense haver de fer els càlculs si ens fixem en la taula de probabilitats de la distribució r que figura en els annexos, en què es pot comprovar el valor a partir de la mida de la mostra n i el valor absolut de la correlació. En aquest cas, per a $n = 10$ el valor crític hauria de ser de 0,5494 en el cas d'una cua, o 0,6319 en el cas d'un supòsit de dues cues.

Inferència sobre relacions de variables (III): ANOVA d'una i dues vies

Una de les possibilitats de l'anàlisi comparativa de les mitjanes via t de Student que s'ha vist anteriorment és incrementar els grups que entren en la prova, de manera que es posen en contacte tres o més grups en relació amb una o més variables. Quan es fa una anàlisi sobre una variable ens referim a una ANOVA d'una via, mentre que quan entren en joc més variable ens referim a una ANOVA de dues vies. L'anàlisi de la variància fa possible aquesta comparació mitjançant la prova ANOVA (les inicials de l'expressió *Analysis Of Variance*), que es basa en la distribució F de Fisher-Snedecor. De fet, va ser Ronald Fisher el primer que l'introduí en el seu manual de recerca (1925: p. 211).

a) ANOVA d'una via

La prova ANOVA d'una via es basa, per tant, en la hipòtesi nul·la d'igualtat de mitjanes per a t mostres. Així, tindríem la formulació d'hipòtesis següent:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_t$$

$$H_1: almenys\ dues\ de\ les\ mitjanes\ són\ diferents$$

Per tant, el nombre de comparacions en ANOVA és més elevat que en una prova t, ja que suposa comparar cadascuna de les mitjanes amb les altres dues, i també entre les tres.

Com en ocasions anteriors, la realització d'una prova ANOVA exigeix el compliment d'algunes normes: per a cada grup, la distribució de la variable y (la quantitativa) és normal; la desviació típica ha de ser la mateixa per a cada grup; i, per últim, cal haver escollit les mostres de manera aleatoritzada i han de ser independents.

Una qüestió que introdueix Agresti és per què, si es tracta d'un mètode de comparació de mitjanes, analitza les variàncies (Agresti, 2018: p. 359). En realitat argumenta. El que fa ANOVA és comparar estimacions de la variància per a cada grup a partir de dues mesures: la primera, la variació entre cada mitjana $\bar{y}_t$ i la mitjana comuna $\bar{\bar{y}}$; la segona, l'estimació, dins de cada grup, respecte de les pròpies mitjanes. Això, com hem vist anteriorment, s'expressa a partir de la fórmula de la prova F:

$$F = \frac{estimació\ intermediant\ de\ la\ variància}{estimació\ interna\ de\ la\ variància}$$

Quan H_0 és vertadera, l'estadístic F adopta una distribució de tipus F amb un valor proper a 1. En canvi, quan H_0 és falsa, l'estadístic F adopta valors majors que 1, en la mesura que l'estimació intermediant de la variància tendeix a sobreestimar les estimacions entre els grups.

La descomposició del denominador, com s'ha vist anteriorment, passa pel càlcul de les diferències de les observacions y respecte de les seues mitjanes $\bar{\bar{y}}$ al quadrat.

$$SQ_{total} = \sum (x - \bar{\bar{x}})^2$$

La Suma de Quadrats Total es divideix entre la Suma de Quadrats Interna i la Suma de Quadrats Intermediant, de manera que:

$$SQ_{total} = SQ_{intermediant} + SQ_{interna}$$

La primera, la Suma de Quadrats Interna, s'obté a partir de la fórmula següent:

$$SQ_{interna} = \sum (n_i - 1)\, s_i^2$$

La Suma de Quadrats Intermediant s'obté a partir de la fórmula següent, en la qual es comparen les mitjanes de cada grup sobre la mitjana total elevada al quadrat. En casos

de mitjanes grupals i totals semblants, el valor de $SQ_{intermediant}$ tendirà a ser baixet, mentre que si les diferències són grans, el resultat d'aquest indicador també serà elevat.

$$SQ_{intermediant} = \sum n_i(\bar{x} - \bar{\bar{x}})^2$$

Aleshores, podem obtindre el valor de $SQ_{interna}$ coneixent els valors de SQ_{total} i de $SQ_{intermediant}$, que solen ser els càlculs més fàcils de dur a terme.

A partir dels factors anteriors, construirem l'Estimació Interna de la Variància (o Quadrat Mitjà Intern) i l'Estimació Intermediant de la Variància (o Quadrat Mitjà Intermediant), que seran denominador i numerador respectivament en el càlcul de l'estadístic F a partir del qual contrastar el valor crític.

Així, a partir de les k estimacions de variàncies obtindrem la variància estimada entre els grups, també coneguda com estimació interna de la variància o Quadrats Mitjans Interns ($QM_{interns}$), que es calcula a partir de la Suma de Quadrats Interna ($SQ_{interna}$), i hi afegirem els graus de llibertat en el denominador, de manera que quedaria la fórmula següent:

$$QM_{intern} = \frac{SQ_{intern}}{N - k}$$

El denominador coincideix amb els graus de llibertat gl_1, que són també els que apareixen en files a la taula estadística.

Pel que fa al denominador en la fórmula d'F, allò que anomenem l'estimació intermediant de la variància, sorgeix del sumatori de les diferències entre cadascuna de les mitjanes i la mitjana comuna al quadrat. Per tant, l'estimació intermediant de la variància parteix de la Suma dels Quadrats Intermediant ($SQ_{intermediant}$), a què s'afegeixen els graus de llibertat en el denominador. Així:

$$QM_{intermediant} = \frac{\sum n_i(\bar{x}_i - \bar{\bar{x}})^2}{k - 1} = \frac{n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \cdots + n_k(\bar{x}_k - \bar{\bar{x}})^2}{k - 1}$$

En aquest cas, els graus de llibertat estan determinats pel nombre de grups k menys un, i formen el segon factor en la fórmula de la funció F (gl_2).

Una vegada més, amb els graus de llibertat (gl_1, gl_2) podem contrastar el valor de F en la taula de valors crítics de la distribució F i comprovar si s'accepta o es rebutja H_0 i, per tant, si assumim o no que les mitjanes són diferents. Cal afegir que l'ANOVA d'una via assumeix que les variàncies dels grups analitzats són iguals, la qual cosa es pot observar també, com hem vist anteriorment, en la comparació de les mitjanes de dues mostres, tal com ens les ofereix l'SPSS. Per tant, es pot dir que la prova F per a dues mostres és l'equivalent a aplicar una prova t (Agresti, 2018: p. 363).

Vegem tot això a partir d'un exemple. Agafem tres grups de població que es corresponen amb y_1, y_2 i y_3, i en els quals hem mesurat el seu posicionament ideològic en una escala del 0 al 5, en què 0 és extrema esquerra i 5 és extrema dreta; n'hem obtés els resultats següents per a una mostra de 9 persones. La mesura d'ANOVA ens donarà a conèixer si hi ha diferències significatives entre els grups o, per contra, els tres grups tenen variàncies semblants, atenent a un nivell de confiança $\propto = 0{,}05$:

y_1	y_2	y_3
1	2	2
2	4	3
5	2	4

Taula 43. Resultat de mesurar el posicionament ideològic en tres grups de població

Font: Elaboració pròpia

El plantejament de les hipòtesis segueix la idea plantejada al principi de l'apartat, és a dir, H_0: $\mu_1 = \mu_2 = \mu_3$, mentre que $H_1 = almenys\ dues\ \mu_t\ són\ diferents$. Calculem primer els graus de llibertat de l'estimació interna ($gl_1 = N - k$ =6) i de l'estimació intermediant ($gl_2 = k - 1 = 2$). Coneixent els graus de llibertat podem extraure el valor crític de F a partir de la taula de distribució dels annexos, d'on podem dir que estarà al voltant de $F = 5{,}14$. Per tant, valors per damunt del valor crític indicaran desigualtat de les mitjanes, i consegüentment el rebuig de H_0.

De la distribució d'exemple obtenim que les mitjanes de cada grup són $y_1 = 2{,}6667$; $y_2 = 2{,}6667$; i $y_3 = 3$, i que la mitjana combinada de les tres mostres és $\bar{\bar{y}} = 2{,}7778$. Així doncs, la suma dels quadrats serà el resultat de combinar cada mitjana amb la mitjana comuna, d'on extraiem que:

y_1	y_2	y_3	y_1	y_2	y_3
(1-2,7778)^2	(2-2,7778)^2	(2-2,7778)^2	3,1606	0,605	0,605
(2-2,7778)^2	(4-2,7778)^2	(3-2,7778)^2	0,605	1,4938	0,0494
(5-2,7778)^2	(2-2,7778)^2	(4-2,7778)^2	4,9382	0,605	1,4938
					$\sum = 13{,}5558$

Taula 44. Resultat de la suma dels quadrats de les desviacions de la mitjana combinada $\bar{\bar{y}}$

Font: Elaboració pròpia

La suma de quadrats interna l'extraiem del sumatori del càlcul de les desviacions de cada valor respecte de la seua mitjana.

y_1	y_2	y_3	y_1	y_2	y_3
(1-2,6667)^2	(2-2,6667)^2	(2-3)^2	2,7779	0,4445	1
(2-2,6667)^2	(4-2,6667)^2	(3-3)^2	0,4445	1,7777	0
(5-2,6667)^2	(2-2,6667)^2	(4-3)^2	5,4443	0,4445	1
					$\sum = 13{,}3334$

Taula 45. Resultat del càlcul de la suma de quadrats interna

Font: Elaboració pròpia

Si coneixem la suma de quadrats total, que hem calculat anteriorment, podem extraure la suma de quadrats intermediant a partir de la sostracció de la suma de quadrats interna. Així, la suma de quadrats intermediant seria de 0,2224.

El següent pas consisteix a transformar les sumes de quadrats en estimacions de la variància, per a la qual cosa s'han de posar en contacte amb els graus de llibertat. Així, l'Estimació de la Variància Intermediant serà el resultat de la divisió de la $SQ_{intermediant}$ i els graus de llibertat que hem anomenat gl_2. Per tant:

$$Estimació\ intermediant\ de\ la\ variància = \frac{0{,}2224}{2} = 0{,}1112$$

De la mateixa manera, obtenim l'Estimació de la Variància Interna a partir del resultat de la $SQ_{interna}$ i la seua divisió pels graus de llibertat que hem anomenat gl_1. Per consegüent:

$$Estimació\ interna\ de\ la\ variància = \frac{13{,}3334}{6} = 2{,}2222$$

El resultat de F vindrà determinat per la fórmula plantejada a l'inici de l'apartat, és a dir:

$$F = \frac{estimació\ intermediant\ de\ la\ variància}{estimació\ interna\ de\ la\ variància} = \frac{0{,}1112}{2{,}2222} = 0{,}05$$

Donat que F és menor del valor crític calculat anteriorment per a la combinació de (gl_1, gl_2), podem concloure que no rebutgem la H_0, i per tant no podem dir que hi haja diferències significatives entre els tres grups analitzats.

b) ANOVA de dues vies

Si el càlcul d'una ANOVA d'una via es torna tediós quan es planteja a mà, fer els càlculs d'una ANOVA de dues vies ho pot ser encara més. De fet, només Moore el considera com un capítol complementari en el seu manual que, en tot cas, ha de ser descarregat d'Internet perquè no s'inclou en la versió impresa del llibre (Moore, 2007).

La suma de quadrats total i la suma de quadrats interna continuen calculant-se de la mateixa manera que en l'ANOVA d'una via. Ara bé, en l'ANOVA de dues vies cal considerar que tenim dos factors, R i C i també el factor conjunt RC; per tant, la suma de quadrats dels grups serà ara la dels efectes principals de cadascun dels factors i la seua interacció conjunta, de manera que:

$$SQ_{total} = SQ_R + SQ_C + SQ_{RC} + SQ_{interna}$$

De la mateixa manera ocorre amb els graus de llibertat:

$$gl_{total} = gl_R + gl_C + gl_{RC} + gl_{interna}$$
$$rcn - 1 = (r-1) + (c-1) + (r-1)(c-1) + rc(n-1)$$

Per últim, la prova de significació F es divideix ara en tres proves diferents: una per al factor R, una per al factor C i una de conjunta per a la interacció de R i C. Així:

$$F = \frac{MQ_R}{MQ_{interna}}; gl: r-1, N-I$$

$$F = \frac{MQ_C}{MQ_{interna}}; gl: c-1; N-I$$

$$F = \frac{MQ_{RC}}{MQ_{interna}}; gl: (r-1)(c-1); N-I$$

En qualsevol cas, la comprovació del valor crític es fa de la mateixa manera que amb l'ANOVA d'una via, amb la taula de distribucions de l'estadístic F que s'adjunta als annexos.

En l'exemple següent plantegem una ANOVA de dues vies, també coneguda com a anàlisi factorial, a partir de l'efecte de dos factors, edat i sexe, sobre el posicionament polític, partint de la matriu de dades que podem observar en la taula següent.

Sexe	Posicionament polític	Edat
Home	1	20
Home	2	20
Home	3	20
Dona	1	20
Dona	3	20
Dona	3	20
Home	2	21
Home	2	21
Home	3	21
Dona	2	21
Dona	3	21
Dona	4	21
Home	3	22
Home	3	22
Home	4	22
Dona	4	22
Dona	5	22
Dona	5	22

Taula 46. Matriu de dades sobre sexe, posicionament polític i edat

Font: Elaboració pròpia

D'aquesta taula, en podem extraure les mitjanes grupals següents per cadascuna de les categories dels factors, així com les mitjanes marginals corresponents a cada factor i també la mitjana total:

	20 anys	21 anys	22 anys	Mitjanes marginals
Homes	2	2,3333	3,3333	2,5555
Dones	2,333	3	4,6667	3,3333
Mitjanes marginals	2,1665	2,6667	4	2,9444

Taula 47. Matriu de mitjanes de posicionament polític per sexe i edat

Font: Elaboració pròpia

La suma de quadrats que abans hem utilitzat amb un únic factor ara s'ha de calcular per a cada categoria de cada factor i amb cada factor. Això vol dir contrastar els resultats d'homes amb la mitjana marginal d'homes i amb la mitjana marginal total i repetir els càlculs per a les dones, tal com veiem a continuació. A més, els graus de llibertat que pren aquesta distribució són $gl_{sexe} = 1$, que s'extrauen dels dos grups (homes i dones) menys un.

	y_i	$\bar{y}_{sexe}$	$\bar{\bar{y}}$	$(\bar{y}_{sexe} - \bar{\bar{y}})$	$(\bar{y}_{sexe} - y)^2$	
Home	1	2,5556	2,9444	-0,3888	0,1512	
Home	2	2,5556	2,9444	-0,3888	0,1512	
Home	3	2,5556	2,9444	-0,3888	0,1512	
Home	2	2,5556	2,9444	-0,3888	0,1512	
Home	2	2,5556	2,9444	-0,3888	0,1512	
Home	3	2,5556	2,9444	-0,3888	0,1512	
Home	3	2,5556	2,9444	-0,3888	0,1512	
Home	3	2,5556	2,9444	-0,3888	0,1512	
Home	4	2,5556	2,9444	-0,3888	0,1512	$\sum_{homes} = 1{,}3605$
Dona	1	3,3333	2,9444	0,3889	0,1512	
Dona	3	3,3333	2,9444	0,3889	0,1512	
Dona	3	3,3333	2,9444	0,3889	0,1512	
Dona	2	3,3333	2,9444	0,3889	0,1512	
Dona	3	3,3333	2,9444	0,3889	0,1512	
Dona	4	3,3333	2,9444	0,3889	0,1512	
Dona	4	3,3333	2,9444	0,3889	0,1512	
Dona	5	3,3333	2,9444	0,3889	0,1512	
Dona	5	3,3333	2,9444	0,3889	0,1512	
Dona	1	3,3333	2,9444	0,3889	0,1512	$\sum_{dones} = 1{,}3612$
						$\sum_{sexe} = 2{,}7217$

Taula 48. Suma de quadrats del posicionament polític per sexe

Font: Elaboració pròpia

Actuem de la mateixa manera amb els resultats per edats. En aquest cas, els graus de llibertat que pren aquesta distribució són $gl_{edat} = 2$, que s'extrauen de les tres agrupacions d'edat presents en la distribució menys una unitat.

	y_i	$\bar{y}_{edat}$	$\bar{\bar{y}}$	$(\bar{y}_{edat} - \bar{\bar{y}})$	$(\bar{y}_{edat} - \bar{\bar{y}})^2$	
20	1	2,1665	2,9444	-0,7779	0,6051	
20	3	2,1665	2,9444	-0,7779	0,6051	
20	3	2,1665	2,9444	-0,7779	0,6051	
20	1	2,1665	2,9444	-0,7779	0,6051	
20	2	2,1665	2,9444	-0,7779	0,6051	
20	3	2,1665	2,9444	-0,7779	0,6051	$\sum_{20} = 3{,}6308$
21	2	2,6667	2,9444	-0,2777	0,0771	
21	3	2,6667	2,9444	-0,2777	0,0771	
21	4	2,6667	2,9444	-0,2777	0,0771	
21	2	2,6667	2,9444	-0,2777	0,0771	
21	2	2,6667	2,9444	-0,2777	0,0771	
21	3	2,6667	2,9444	-0,2777	0,0771	$\sum_{21} = 0{,}4627$
22	4	4	2,9444	1,0556	1,1143	
22	5	4	2,9444	1,0556	1,1143	
22	5	4	2,9444	1,0556	1,1143	
22	3	4	2,9444	1,0556	1,1143	
22	3	4	2,9444	1,0556	1,1143	$\sum_{22} = 6{,}6857$
						$\sum_{edat} = 10{,}7792$

Taula 49. Suma de quadrats del posicionament polític per edat

Font: Elaboració pròpia

Així, la $SQ_{sexe} = 2{,}7217; gl_{sexe} = 1$. I la $SQ_{edat} = 10{,}7792; gl_{edat} = 2$.

El pas següent és calcular la suma de quadrats interns $SQ_{interns}$, que consitueixen l'aproximació a l'error, en aquest cas a partir de les mitjanes de cada categoria. En el cas del posicionament polític dels homes, per tant, triaríem la mitjana de cada grup d'edats per als homes; i per a les dones, cada mitjana de grup d'edats en els casos de les dones. Per al càlcul dels graus de llibertat agafem el nombre d'observacions de cada grup. Per a cadascuna de les tres categories d'edat de cadascun dels dos sexes tenim tres observacions, a les quals sostraurem una unitat, i sumades fan un total de $gl_{interns} = 12$.

	y_i	$\bar{y}$	$(y-\bar{y})$	$(y-\bar{y})^2$	
Home	1	2	-1	1,0000	
Home	2	2	0	0,0000	
Home	3	2	1	1,0000	
Home	2	2,3333	-0,3333	0,1111	
Home	2	2,3333	-0,3333	0,1111	
Home	3	2,3333	0,6667	0,4445	
Home	3	3,3333	-0,3333	0,1111	
Home	3	3,3333	-0,3333	0,1111	
Home	4	3,3333	0,6667	0,4445	$\sum_{homes} = 3{,}3333$
Dona	1	2,3333	-1,3333	1,7777	
Dona	3	2,3333	0,6667	0,4445	
Dona	3	2,3333	0,6667	0,4445	
Dona	2	3	-1	1,0000	
Dona	3	3	0	0,0000	
Dona	4	3	1	1,0000	
Dona	4	4,6667	-0,6667	0,4445	
Dona	5	4,6667	0,3333	0,1111	
Dona	5	4,6667	0,3333	0,1111	$\sum_{dones} = 5{,}3333$
					$\sum_{sexe} = 8{,}6667$

Taula 50. Suma de quadrats interns del posicionament polític

Font: Elaboració pròpia

Tal com hem vist anteriorment, caldria calcular la suma de quadrats total o la suma de quadrats conjunta per tal de poder calcular l'estadístic F. En aquest cas, calcularem la suma de quadrats total SQ_{total} a partir de la mitjana total $\bar{\bar{y}}$, amb el mateix mètode d'obtenció del sumatori a partir de la desviació quadrada. El càlcul dels graus de llibertat s'extrauen ara de la suma de tots els graus de llibertat, comptant també els graus de llibertat conjunts $gl_{conjunts}$ que, recordem, eren 2 i 1; per tant, $gl_{conjunts} = 2$. De la suma, per tant, obtindríem $gl_{total} = 17$.

	y_i	$\bar{\bar{y}}$	$(y-\bar{\bar{y}})$	$(y-\bar{\bar{y}})^2$	
Home	1	2,9444	-1,9444	3,7807	
Home	2	2,9444	-0,9444	0,8919	
Home	3	2,9444	0,0556	0,0031	
Home	2	2,9444	-0,9444	0,8919	
Home	2	2,9444	-0,9444	0,8919	
Home	3	2,9444	0,0556	0,0031	
Home	3	2,9444	0,0556	0,0031	
Home	3	2,9444	0,0556	0,0031	
Home	4	2,9444	1,0556	1,1143	$\sum_{homes} = 7{,}5830$
Dona	1	2,9444	-1,9444	3,7807	
Dona	3	2,9444	0,0556	0,0031	
Dona	3	2,9444	0,0556	0,0031	
Dona	2	2,9444	-0,9444	0,8919	
Dona	3	2,9444	0,0556	0,0031	
Dona	4	2,9444	1,0556	1,1143	
Dona	4	2,9444	1,0556	1,1143	
Dona	5	2,9444	2,0556	4,2255	
Dona	5	2,9444	2,0556	4,2255	$\sum_{dones} = 15{,}3614$
					$\sum_{total} = 22{,}9444$

Taula 51. Suma de quadrats totals del posicionament polític

Font: Elaboració pròpia

Dels càlculs anteriors, podem deduir-ne la suma dels quadrats d'ambdós factors, que podem obtindre, coneguts la resta de factors, aïllant-la del plantejament genèric de l'ANOVA de dues vies, a partir de la sostracció de la suma de quadrats total, de cadascuna de les sumes de quadrats de cada factor i de la suma de quadrats interna.

$$SQ_{ambdós\ factors} = SQ_{total} - SQ_{edat} - SQ_{sexe} - SQ_{interna}$$
$$= 22{,}9444 - 10{,}7792 - 2{,}7217 - 8{,}6667 = 0{,}7768$$

Així, a partir dels valors que ja tenim, calculem les estimacions de la variància o mitjanes quadrades a partir de la divisió de la suma de quadrats pels graus de llibertat, especialment de les sumes de quadrats dels factors i de la suma de quadrats interna, que representa l'error. El denominador per a la F de cada factor serà el resultat de la divisió de la suma de quadrats del factor corresponent pel resultat de la suma de quadrats interna.

S'hi comproven els valors crítics a partir dels graus de llibertat de cada relació (amb edat i amb sexe), amb compte de no equivocar-nos amb els graus de llibertat. Per exemple, la $F_{edat}(1{,}12)$ i la $F_{sexe}(2{,}12)$ donen valors crítics de F de 4,75 i de 3,88, respectivament. Per tant, podem rebutjar H_0 per a l'edat, però no per al sexe. D'altra banda, el valor F de la variància conjunta es calcularia a partir dels graus de llibertat de la $SQ_{conjunta}$, que en aquest cas és novament $F_{conjunta}\ (2{,}12)$, la qual cosa ofereix novament un valor crític de 4,75, que no supera la nostra $F_{conjunta}$, per la qual cosa no rebutjaríem la hipòtesi nul·la. Finalment, per tant, només rebutjaríem una de les tres possibilitats d'encreuament, la de l'edat.

	Resultat	gl	Estimació variància	F
SQ_{edat}	10,7792	1	$\frac{10{,}7792}{1} = 10{,}7792$	$\frac{10{,}7792}{0{,}7222} = 14{,}9255$
SQ_{sexe}	2,7217	2	$\frac{2{,}7217}{2} = 1{,}3609$	$\frac{1{,}3609}{0{,}7222} = 1{,}8844$
$SQ_{conjunta}$	0,7768	2	$\frac{0{,}7768}{2} = 0{,}3884$	$\frac{0{,}3884}{0{,}7222} = 0{,}5378$
$SQ_{interna}$	8,6667	12	$\frac{8{,}6667}{12} = 0{,}7222$	
SQ_{total}	22,9444	17		

Taula 52. Càlcul de F per a edat, sexe i l'efecte conjunt

Font: Elaboració pròpia

2.1. El mostreig

El mostreig, o la tècnica d'agafar una part pel tot, a què es refereix Desrosières (1998: p. 210) es remunta a les primeres enquestes del segle XVII, quan Sébastien Le Prestre, marqués de Vauban, publica el seu *Méthode générale et facile pour faire le dénombrement des peuples* (1686), que en la pràctica eren unes normes per a dur a terme els censos poblacionals, algunes de les quals encara són útils ara mateix. Vora cent anys més tard, John Sinclair edita l'*Statistical Account of Scotland*, una obra publicada entre 1791 i 1825 en la qual interroga la població escocesa sobre les seues característiques sociodemogràfiques, però també sobre la pràctica religiosa o l'activitat econòmica (Cea, 2012: p. 186; Leti, 2000). A partir d'aquells moments serà relativament habitual trobar enquestes, com ara les de Charles Booth sobre la població de Londres (1902-1903), la de Benjamin Seebohm sobre la pobresa i les condicions laborals (1906) o l'obra de Frédéric Le Play (1855) sobre els obrers en diferents països europeus (Cea, 2012: p. 187; Ornstein, 2013: p. 1). Fins i tot figures com Marx o Weber van utilitzar mostres en els seus estudis, entre els quals destaca l'enquesta obrera comandada per Marx (Weiss, 1979; Cea, 2012: p. 189; Brain, 2001).

Però un dels avanços que més va fer per la popularització de l'enquesta va ser l'aplicació de la teoria de la probabilitat i, en particular, el mostreig aleatori, representatiu i probabilístic. La qüestió de la representativitat va ser introduïda pel noruec Anders Kiær en l'enquesta que va planificar sobre Noruega en 1895, amb una divisió del país en àrees rurals i urbanes, primer; en carrers dels municipis, després; i finalment en edificis dels carrers (Lie, 2002: p. 391). Es podria dir que representa allò que ara coneixem com a mostreig polietàpic, tot i que no va existir-hi un procés d'aleatorització (Ornstein, 2013: p. 2). El mostreig probabilístic es deu a Arthur Bowley i Alexander Burnett-Hurst, que en 1915 publicaren una investigació sobre la pobresa a la ciutat de Londres on, per primera vegada, s'utilitzava una selecció aleatòria dels informants i fins i tot l'aplicació d'intervals de confiança (Desrosières, 1998: p. 210).

Durant els anys 1920 i 1930 és quan més s'avança en el mostreig. Potser un dels casos de major èxit és el que enfrontà les prediccions electorals de la revista *The Literary Digest*, que defenia que les mostres de major grandària (en aquest cas als seus subscriptors) tenien també un major poder explicatiu, enfront del disseny mostral representatiu per quotes de sexe i edat de Gallup i Crossley. La mostra de menor grandària va ser capaç de predir la victòria de Roosevelt en 1936, cosa que situà l'enquesta com a

una tècnica fiable per a predir esdeveniments (Squire, 1988). Pocs anys després Paul Lazarsfeld fundà el *Bureau of Applied Social Research* a Columbia, des d'on estudiarà aspectes electorals, però també posarà en pràctica l'enquesta panel i l'anàlisi de dades en taules encreuades, tot i que encara no aplicarà models d'inferència estadística ni estimació d'intervals de confiança (Cea, 2012: p. 193).

Però l'avanç definitiu en aquesta època és el que aportà Jerry Neyman, que adaptà la idea de la selecció aleatòria i introduí l'estratificació de la mostra en 1934, amb el llançament del mostreig estratègic, com diu Desrosières, a les masmorres (1998: p. 226). A Neyman també se li atribueix el mostreig per conglomerats, el mostreig en poblacions finites i l'error de mostreig, entre d'altres (Alasuutari, Brickman i Brannen, 2008; Cea, 2012: p. 189 i s.; Ornstein, 2013: p. 4).

Al llarg del text ens hem referit a la importància del mostreig, del mètode escollit per a dur-lo a terme i de com de determinant pot arribar a ser per a l'aplicació de les mesures estadístiques. Arribats a aquest punt, convé centrar el discurs en la manera amb què s'escullen els individus o casos que seran els que representaran la població, requisit bàsic per tal de poder tancar el cercle inferencial que ens porta d'univers a mostra, primer, i a l'univers una altra vegada.

Entre els diferents tipus de mostreig, aquell que resulta més interessant per a la socioestadística és l'aleatori, també conegut com a probabilístic, en la mesura que és l'únic que permet fer ús dels procediments inferencials. Per tal que una mostra siga considerada com a probabilística cal que complisca almenys la condició d'aleatorització en la selecció, cosa que deixa de banda tots aquells procediments que no seguisquen aquesta norma. Aquests procediments són els coneguts com a mostres no probabilístiques, i poden ser interessants, però en cap cas s'haurien de generalitzar sobre el grup més ampli en el qual han estat seleccionades (Healey, 2016: p. 141). Així doncs, una de les primeres decisions a l'hora de plantejar una investigació passa per com escollim les persones que formaran part de la mostra, calcular el nombre d'individus que caldria considerar i com es relacionen aquests aspectes amb paràmetres com l'error o el nivell de confiança.

2.1.1. Tipus de mostreig

La diferència entre el mostreig probabilístic i el no probabilístic es basa en la introducció de l'aleatorització en la selecció de les unitats mostrals. Així, mentre que en el mostreig probabilístic sí que es contempla la introducció de mètodes d'aleatorització, en el mostreig no probabilístic existeix una conveniència, una raó per la qual un individu i no un altre, entra a formar part de la mostra. En els procediments probabilístics cada unitat de l'univers o del marc mostral té les mateixes possibilitats de ser escollit per a la mostra (Cea, 2012: p. 298). A més, l'elecció d'una persona per tal que forme part de la mostra és independent de l'elecció de la resta de les persones que ja en formen[49] part. I existeix la possibilitat de calcular l'error mostral, i per tant de dur a terme un procés d'inferència estadística, mitjançant el càlcul de paràmetres i l'aplicació de proves de rellevància. Per contra, els individus seleccionats mitjançant procediments no probabilístics no presenten una mateixa probabilitat de formar part de la mostra final. Això genera dificultats o, directament, la impossibilitat de calcular l'error mostral. A més, la introducció de biaixos en el l'elecció de la mostra pot, fins i tot, invalidar els resultats de la investigació.

Entre els procediments no probabilístics de selecció de la mostra, per tant no adequats a procediments inferencials, hi podem distingir entre el mostreig per quotes, l'estratègic i el circumstancial. Pel que fa al mostreig per quotes, es pot dir que és el mostreig no probabilístic que proporciona millors resultats, perquè garanteix la proporcionalitat de la quota respecte de l'estructura de la població a partir de les variables seleccionades com a quotes (per exemple, sexe, edat i classe social). Val a dir que el mostreig per quotes s'incorpora, com a procediment mixt, en mostrejos polietàpics, com veurem a continuació, la qual cosa incorpora l'element d'aleatorietat en la selecció de la mostra.

Un segon tipus de mostreig no probabilístic és l'estratègic, en el qual es trien els individus de la mostra a partir de criteris d'interès particular per a la investigació que aplica la persona que la dirigeix. En principi, ha de ser un mostreig heterogeni per a facilitar diferents opinions i una mínima diversitat. L'única cosa que tindrien en comú les persones escollides és, precisament, tenir la característica, coneixement o qualitat central per a l'objecte d'estudi.

Finalment, també dins els procediments no probabilístics, hi ha el mostreig circumstancial, que comprèn el mostreig de persones voluntàries i el de *bola de neu*. El mostreig de persones voluntàries implica la voluntarietat de les persones que participaran en un estudi, cosa que eximeix l'atzar de l'elecció de la mostra i introdueix biaixos relacionats

[49] Healey ho defineix com *Equal Probability of Selection Method*, amb les sigles EPSEM (Healey, 2016: p. 142).

amb la voluntat de participar en l'estudi (per exemple, si és un estudi pel qual es paga una determinada quantitat per participar-hi, possiblement hi estaran sobrerepresentades les persones de classes socials més baixes). D'altra banda, el mostreig en forma de bola de neu suposa que, a partir de la selecció de les primeres unitats de la mostra, aquesta es va eixamplant mitjançant la incorporació de noves persones informants que són conegudes o recomanades, fins que s'arriba a la saturació teòrica pròpia de les tècniques qualitatives i es deixen d'incorporar nous membres a la mostra.

Entre els procediments de selecció mostral probabilística en trobem de diferents, tot i que els més habituals són el simple, el sistemàtic, l'estratificat i el mostreig per conglomerats, tot i que també se'n poden localitzar altres de tipus polietàpic que en combinen de diferents tipus, incloent-hi els probabilístics i no probabilístics, com veurem a continuació.

El mostreig aleatori simple és el més senzill de tots els tipus de mostreig, i també és el que serveix de referència a l'hora de calcular, entre altres coses, la grandària de la mostra, el nivell de confiança o l'error mostral[50]. El plantejament del mostreig simple és que cada unitat de l'univers o del marc mostral té la mateixa probabilitat d'entrar dins la mostra. Aquesta probabilitat està determinada pel que s'anomena fracció de mostreig, que es calcula a partir del quocient $\frac{n}{N}$, en què n és la grandària de la mostra i N el total de l'univers. La tria de la mostra es fa amb mètodes aleatoris (per exemple, la generació de x números aleatoris situats entre 1 i el valor de n, tria que normalment s'ha de sobredimensionar per tal de donar eixida a imprevistos com la no localització de les unitats mostrals o la no resposta. Aquesta tria se sol fer sense reemplaçament i s'organitza amb els mateixos mètodes de què disposen els programes electrònics de tractament de dades. En el cas de l'SPSS, per exemple, es pot optar per generar una taula de nombres aleatoris a partir dels quals escollir les persones que formen part de la mostra, o també seleccionant un nombre de casos de manera aleatòria. En altres casos, es pot disposar d'una taula de nombres aleatoris que en faciliten l'elecció[51], d'urnes amb butlletes o paperets o altres mètodes d'aleatorització en què ens assegurem de la no reposició, i per tant del fet que cada membre del marc mostral tinga les mateixes oportunitats de ser escollit. En tot cas, el mostreig aleatori simple, com el sistemàtic, impliquen la disposició

[50] En els manuals anglòfons és molt habitual trobar-se amb les sigles que el caracteritzen: *Simple Random Sample* (SRS). Vegeu, a tall d'exemple, Kalton (1983: p. 8).

[51] Encara hi ha molts manuals que, a més d'aportar les taules estadístiques de densitat, adjunten una taula de nombres aleatoris (vegeu, per exemple Blalock, 1979; o Moore, 2007). Una de les primeres mostres en aquest sentit és l'obra de Kendall i Babington, un clàssic de 60 pàgines que va millorar publicacions precedents en aportar més de 100.000 dígits (Kendall i Babington, 1938) i que als anys 50 arribaria a publicacions amb un milió de dígits aleatoris.

d'un marc mostral amb dades de contacte a partir del qual fer la selecció aleatòria, cosa que en la investigació empírica en l'àmbit acadèmic és difícil d'aconseguir atès l'alt cost per a aconseguir aquestes bases de dades i, d'altra banda, la creixent importància que tenen altres mètodes de selecció de la mostra com la utilització de mètodes de selecció automàtica de números de telèfons randomitzats. D'altra banda, el mostreig aleatori simple genera una distribució mostral de la qual, en tot cas, inferim la variable seleccionada sobre la proporció, sempre després d'estudiar, entre altres coses, la forma i els principals indicadors de centralitat i dispersió. En tot cas, es passaria d'una distribució poblacional -desconeguda- a una distribució teòrica mostral -no empírica- a una distribució mostral -en aquest cas empírica i coneguda.

El mostreig aleatori sistemàtic incorpora l'elecció aleatòria d'un primer número, que ha de ser inferior al resultant del quocient d'elevació $\frac{N}{n}$ aproximat a la unitat, a partir del qual s'aniria sumant el quocient tantes vegades com unitats tinga la mostra, fins arribar a cobrir la grandària de la mostra seleccionada. Com en el cas anterior, és convenient que el nombre de casos siga superior al de la mostra teòrica per tal de facilitar la substitució d'aquelles unitats no localitzades o que no volen contestar.

Per a aplicar el mostreig aleatori estratificat cal, a més de tenir un marc mostral amb dades de contacte, tenir també dades d'estratificació que servisquen per a classificar a la població. Aquesta selecció ha de garantir homogeneïtat dins els estrats en funció de les variables estratificadores, i també heterogeneïtat entre els estrats. A més, cal garantir la presència d'estrats amb menor pes en la mostra, la qual cosa de vegades es tradueix en la sobrerepresentació d'alguns estrats. També pot voler dir aplicar mètodes de selecció de la mostra diferents en funció dels estrats, o fins i tot infrarepresentar estrats amb escassa heterogeneïtat ja que no cal atorgar-los un pes proporcional a la seua importància relativa sobre l'univers.

Dins el mostreig aleatori estratificat, per tant, podem trobar-hi diferents combinacions en funció de quina siga la distribució del pes dels estrats en la mostra. La més senzilla és l'afixació simple, que assigna a cada estrat el mateix pes en la mostra, independentment del seu pes relatiu en l'univers, que possiblement és desconegut. Això es tradueix en estrats sobrerepresentats i estrats infrarepresentats. L'afixació proporcional distribueix la mostra en estrats de manera proporcional a la seua presència en l'univers. I per acabar, l'afixació òptima considera no només el pes relatiu sobre el total de l'univers, sinó també l'heterogeneïtat de cadascun dels estrats a partir de l'indicador de la variància,

que en tot cas s'hauria de calcular anteriorment. L'aplicació pràctica de les tres modalitats de mostreig aleatori estratificat es pot observar en la taula següent (Cea, 2012: p. 307). Sota l'encapçalament de l'afixació, s'hi pot observar com la simple és el resultat de dividir la grandària de la mostra calculada inicialment, 1.446, pel nombre d'estrats, en aquest cas, 5. El resultat, 289,2, s'arrodoneix a 289 i es perd una unitat, amb la qual cosa la mostra final seria de 1.445 persones. En el cas de l'estratificació proporcional, cada estrat es calcula en funció del seu pes en la població. Això exigeix calcular primer el pes relatiu $\left(\frac{població}{total\ univers}\right)$ per aplicar el resultat sobre el total de la mostra. El primer estrat, 431, seria el resultat de 0,298*1.446. Finalment, en el cas de la distribució òptima, inclouria el càlcul de la variància de la tercera columna, en què el resultat el donaria el càlcul $\left(\frac{percentatge\ població*variància}{\sum percentatges*variància}*n\right)$. En el cas de la primera fila, el resultat s'esdevindria del càlcul $\left(\left(\frac{29{,}8*2{,}475}{227.030{,}1}\right)*100\right)*1.446 = 469{,}95$, amb la qual cosa seleccionaríem 470 persones del primer estrat. En absència de la variància, cal substituir-la per l'error típic al quadrat a l'hora d'estimar l'afixació.

Ensenyament	Població	Variància	Afixació		
			Simple	*Proporcional*	*Òptima*
Cicles superiors	22.073	2,475	289	431	470
Graus	33.220	2,139	289	648	610
Dobles graus	6.681	2,496	289	130	143
Màsters	4.509	2,331	289	88	91
Doctorat	7.613	2,016	289	149	132
Total	74.096	2,304	1.445	1.446	1.446

Taula 53. Distribució de la mostra segons un mostreig estratificat i les seues variants

Font: Elaboració pròpia a partir de Cea (2012: p. 307)

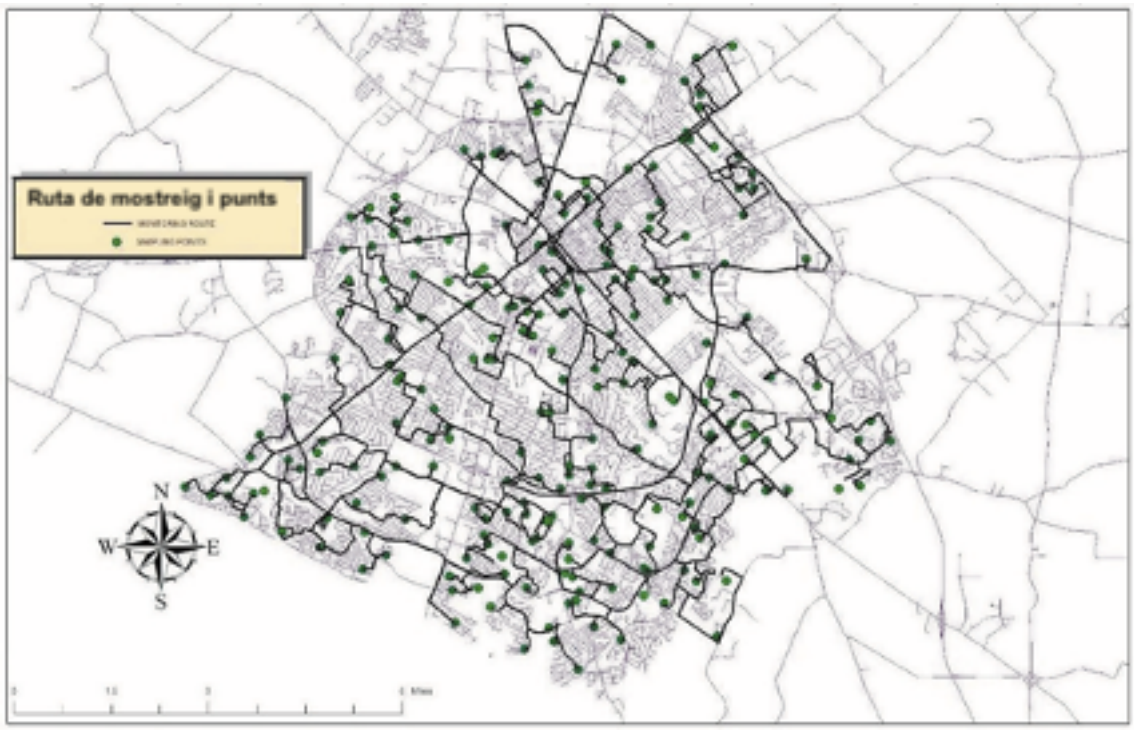
Cap la possibilitat que l'afixació estratificada no siga proporcional, perquè l'equip d'investigació ho ha decidit de manera conscient. De vegades és perquè no s'ha aconseguit arribar a cobrir tota la població dels estrats, altres vegades perquè un càlcul previ o posterior ens fa canviar la distribució dels pesos dels estrats. En aquests casos, cal aplicar-hi una ponderació de la mostra en cada estrat. Els quocients de ponderació serveixen per a tornar a la mostra la proporcionalitat que no s'ha aconseguit per la via de l'estratificació, de manera que l'indicador de ponderació multiplica cada estrat pel pes que hauria de tenir. Això aplicat, per exemple, sobre un mostreig simple com el de la tercera columna de l'exemple anterior, es traduiria en la transformació de cada estrat en

un indicador nou, producte del quocient aplicat. En aquest cas, com que el percentatge de la població és de 29,8% i el percentatge en la mostra és un 20%, calcularíem el quocient de ponderació dividint el primer entre el segon ($\frac{29,8}{20} = 1,49$). Així, els indicadors referents als graus haurien d'incrementar-se en un 1,49 per tal de ser proporcionals al seu pes en la població, i successivament amb la resta d'estrats.

En el mostreig per conglomerats, en comptes de seleccionar estrats de la població, se seleccionen conjunts poblacionals. La seua aplicació és convenient quan l'univers es troba dispers i cal desplaçar les persones de l'equip d'investigació. Com apunta Cea, els conglomerats poden ser demarcacions territorials, organitzacionals, però també institucions, centres administratius, o fins i tot conglomerats artificials com les urnes electorals (2012: p. 311). Així, el primer pas en la selecció de la mostra és escollir els conglomerats i després, dins de cadascun, s'extrauen les unitats de la mostra mitjançant algun mètode d'aleatorització. En seleccionar conglomerats, i no el total de l'univers, el treball de camp és menys costós perquè exigeix menys desplaçaments. Tanmateix, allò que convé és conèixer prèviament quina és la distribució de la mostra dins de cada conglomerat. Cal distingir entre el conglomerat monoetàpic, si totes les unitats del conglomerat componen la mostra, i el polietàpic, quan dins de cada conglomerat s'inclou la selecció aleatòria d'unitats mostrals. Normalment, els mostrejos polietàpics comprenen tres o quatre fases diferents, com ara el cas dels baròmetres del CIS: primerament se seleccionen els municipis atenent a la dimensió de la població que hi viu; després se seleccionen les seccions censals dins els municipis; i sobre aquestes seccions censals s'apliquen mètodes d'aleatorització en la selecció de les llars i finalment dels individus per quotes.

Tot i que no es tracta d'un procediment de mostreig en si mateix, el mostreig per àrees o rutes aleatòries s'aplica a la selecció última dels individus, de manera que els mapes s'utilitzen com a marcs mostrals. La primera tasca és dividir en blocs l'àrea a investigar, numerar-los i tractar de delimitar-ne l'àrea per tal que la superfície siga semblant en tots els casos. A partir d'aquest punt, cal delimitar com es fa la selecció dels blocs. L'opció més fàcil podria ser seleccionar-los tots de manera sistemàtica, però normalment s'apliquen rutes aleatòries més o menys estructurades. El mostreig per rutes aleatòries ha de començar per un punt de partida que se selecciona prèviament al treball de camp, i a partir d'aquest lloc s'hi van aplicant les instruccions per a seleccionar les unitats de llar, i dins d'aquestes, els individus. Les instruccions de la ruta comencen per seleccionar el carrer a partir del punt de partida, de manera que la persona enquestadora tindrà unes

instruccions precises per als girs en els blocs de cases. Tot seguit, se selecciona un edifici del carrer, per exemple, els acabats en 7. De l'edifici, caldrà seleccionar-ne un habitatge, per exemple, el de la primera planta parell. I finalment, de la llar caldrà entrevistar una persona que es triarà en funció de les quotes de sexe i edat normalment predefinides en el mostreig (Díaz de Rada, 2015).



Gràfica 36. Exemple de rutes aleatòries sobre un mapa

Font: Adaptat de Matei (2009)

2.1.2. Determinació de la mostra

Un dels primers passos que cal fer a l'hora de planificar el mostreig és delimitar la població a la qual s'estudia, allò que anomenem univers d'estudi, i que està compost pel conjunt d'unitats de les quals volem obtenir una informació. Aquestes unitats normalment seran individus, encara que en ocasions poden ser llars, municipis, països, etcètera. El mostreig final estarà en funció de quin siga l'univers, en quin moment se l'interroga i quines característiques té. En tot cas, l'equip d'investigació ha de ser coneixedor del nombre exacte -en la mesura del possible- d'unitats que té l'univers en qüestió. En funció d'aquests criteris, caldrà triar el marc mostral, el llistat del qual s'extrauen les unitats que finalment formaran part de la mostra. Fins fa alguns anys, aquesta era una decisió que calia prendre, especialment en el marc de les enquestes telefòniques, en

base a registres existents, com ara els directoris telefònics. Avui dia, gràcies a la generació aleatòria de números de telèfon ha desaparegut la necessitat de disposar d'un marc mostral. De la mateixa manera, les enquestes cara a cara tampoc utilitzen marc mostral, sinó l'entrada aleatòria d'unitats mostrals a partir de la selecció de llars per rutes aleatòries disperses per les poblacions. En el cas de disposar d'un marc mostral, hem d'assegurar-nos d'algunes qüestions: que és tan complet com siga possible; que està actualitzat; que les unitats no estan repetides; que no hi ha unitats que no corresponen a la població estudiada; que disposa d'informació complementària -que pot ajudar a establir quotes de mostreig- i que és fàcil d'utilitzar.

El següent pas en la determinació de la mostra és la decisió sobre la seua grandària, cosa que està en funció d'algunes variables com ara: el temps i els recursos disponibles, la modalitat de mostreig, la diversitat de l'anàlisi prevista, la variància o heterogeneïtat poblacional, el marge d'error previst, el nivell de confiança i, per descomptat, la grandària de l'univers de referència.

En funció del temps i dels recursos disponibles, triarem una grandària de mostra o una altra. Per exemple, l'enquesta telefònica, degut al seu preu i la facilitat per arreplegar grans quantitats d'informants en poc temps, pot ser una solució que s'adapte a diferents tipus d'investigació. Les enquestes autoadministrades via web, per contra, solen tenir aparellat el perill de la no resposta, amb la qual cosa la grandària de la mostra prevista haurà de ser major si volem que siga representativa.

En funció de la modalitat de mostreig ens podem trobar davant de seleccions d'informants probabilístiques que, en tot cas, exigeixen aleatorització en la selecció de les unitats mostrals i per tant capacitat de representació i inferència; o mostrejos no probabilístics on la grandària pot ser menor perquè mitjançant el mètode de selecció de la mostra no cap la possibilitat de representar l'univers.

En funció de l'anàlisi prevista, la mostra haurà de ser més o menys gran. En termes generals, com més s'haja de segmentar la mostra en l'anàlisi, major haurà de ser la mostra per tal que les estimacions obtingudes puguen assolir la rellevància estadística.

En funció de la variància o l'heterogeneïtat poblacional, la grandària de la mostra variarà, de manera que una major heterogeneïtat implica també més persones en la mostra per tal de recollir la diversitat estimada en la població. Com que normalment no tenim un

estimador de la variància en la mostra estudiada, assumim la màxima situació d'heterogeneïtat que es pot aplicar a un supòsit probabilístic dicotòmic, és a dir, que la probabilitat que ocórrega un esdeveniment ($\hat{p}$) és igual a la probabilitat que no ocórrega ($1-\hat{p}$, que també se sol representar amb $\hat{q}$), o el que és el mateix, que $\hat{p}=\hat{q}=0{,}5$. Això fa incrementar la grandària de la mostra per damunt del que suposaria valors diferents a 0,5 (per exemple, 70/30, 80/20 o 1/99), tal com es pot observar en la taula següent, en què es combinen diferents valors de $\hat{p}$ i $\hat{q}$ amb diferents errors mostrals i nivells de confiança[52].

Error mostral (%)	Nivell de confiança (%)	Valors pressuposats de p i q (%)				
		10/90	20/80	30/70	40/60	50/50
±1,0	95,5	3600	6400	8400	9600	10000
	99,7	8100	14400	18900	21600	22500
±2,0	95,5	900	1600	2100	2400	2500
	99,7	2025	3600	4725	5400	5627
±2,5	95,5	576	1024	1344	1536	1600
	99,7	1296	2304	3024	3456	3600
±3,0	95,5	400	711	933	1067	1111
	99,7	900	1600	2100	2400	2500
±4,0	95,5	225	400	525	600	625
	99,7	506	900	1181	1350	1406

Taula 54. Grandària de la mostra per a poblacions infinites en un nivell de confiança del 95,5% (2σ) i 99,7 (3σ)

Font: Adaptat de Cea (2012: p. 294)

Tanmateix, en la major part de les investigacions empíriques basades enquestes a població, solem utilitzar de manera sistemàtica $\hat{p}=\hat{q}=0{,}5$ perquè assumim que treballem amb una variància que ens és desconeguda, amb la qual cosa la grandària de la mostra estarà en funció de les dues variables que apareixen en files en la taula anterior: p. ex. error mostral i nivell de confiança.

[52] Altres manuals com el de Sierra Bravo (2001: p. 231 i s.) ofereixen taules encara més completes però valorem, per damunt que l'alumnat tinga a la seua disposició una resposta immediata al problema de la grandària de la mostra, el fet que siga capaç de calcular-la a partir de les fórmules proposades.

En funció del marge d'error admissible en l'estudi tindrem una grandària de la mostra o una altra, cosa que implica que a major error assumit, menor grandària de la mostra; i per contra, errors menors implicarien majors grandàries mostrals. Això es tradueix en el fet que mostres més grans tindran major capacitat d'aproximar-se al valor real de la variable estudiada en l'univers, mentre que les mostres més xicotetes tindran més dificultats per a predir el mateix valor. El problema que hi trobem és que reduir els marges d'error a partir de l'1,9% suposen increments de la mostra tals que allò que es guanya en termes de marge d'error suposa inversions molt grans per a ampliar la mostra més enllà de les 3.000 persones[53] que no sempre es poden justificar.

En funció del nivell de confiança de l'estimació dels paràmetres de l'univers podem trobar mostres grans, amb major probabilitat que les seues estimacions s'aproximen a les mesures de l'univers o mostres més xicotetes, que en principi tindrien menys probabilitat d'aproximar-se. Els nivells de confiança es corresponen a les àrees per davall de la corba normal acotades segons els diferents valors de desviació típica, també anomenada *sigma* (i representada amb la lletra grega σ). Entre tots els valors que pot adoptar la mostra, el més habitual que trobem en investigació empírica és el 2σ (95,5% d'àrea sota la corba normal). Augmentar el nivell de confiança d'una unitat fins als 3σ (99,7% de nivell de confiança) es traduiria en un increment de la mostra de més del doble, amb la qual cosa s'entén que no siga una acció habitual en investigació empírica[54].

Per acabar, en funció de la població que conforma l'univers, hi trobarem una grandària mostral o una altra. Això sol estar relacionat amb el teorema del límit central, pel qual, el càlcul de la mitjana d'una mostra aleatòria de qualsevol variable en poblacions grans se sol aproximar a la distribució normal. A l'efecte de càlcul, l'univers gran és aquell en què el percentatge que suposa la mostra sobre l'univers és menor del 5% (Rodríguez, Ferreras i Núñez, 1991), cosa que a efectes pràctics s'ha traduït en l'aplicació d'un factor de *Correcció de Poblacions Finites* (CPF) a partir d'universos majors de 100.000 unitats. Així, podem trobar dos tipus d'equació per a calcular la grandària de la mostra en funció de si la població és finita (universos per davall de 100.000 persones) o si és finita (per damunt d'aquesta quantitat), que són les que desenvolupà l'estadístic d'origen escocès William Cochran (1953) i que veurem a continuació.

[53] Per exemple el CIS, que és un dels organismes amb més capacitat de reunir grans mostres, sol apostar per errors que no baixen de l'1,8%, a excepció dels macrobaròmetres en què, de manera excepcional, arriba al 0,75% (vegeu, per exemple, la fitxa tècnica del recent macrobaròmetre d'abril de 2019 a http://www.cis.e s/cis/export/sites/default/-Archivos/Marginales/3240_3259/3245/FT3245.pdf)

[54] En el cas del CIS, fins i tot en els estudis més ambiciosos, com ara el macrobaròmetre a què es feia referència anteriorment, agafen com a referència un nivell de confiança de 2σ.

2.1.1. Càlcul de la mida de la mostra

La fórmula del càlcul de la mida li la devem a l'estadístic escocès William Cochran, qui la va fer pública l'any 1953 en el seu manual *Sampling Techniques* (1953: p. 50). En el desenvolupament de l'aplicació de la fórmula, Cochran utilitza un cas pràctic on un antropòleg vol conèixer les persones amb grup sanguini O d'una illa. A partir de l'estimació de l'error, que en aquest cas situa voluntàriament en ±5% i de la capacitat d'encertar l'estimació en almenys 19 de cada 20 investigacions sobre la mateixa població extrau la fórmula que avui dia utilitzem per a les poblacions infinites[55] en situacions de mostreig aleatori simple[56].

Val a dir que per a poder calcular la mida de la mostra cal conèixer abans alguns factors, cosa que es pot saber a priori o es pot aproximar. La primera dada que cal conèixer, vista la necessitat d'utilitzar-la per al càlcul de la mida de la mostra, és la mida de l'univers. En funció d'aquesta mesura, triarem una fórmula o una altra (infinita en el cas dels universos de més de 100.000 unitats, finita en el cas contrari), com veurem tot seguit. Pel que fa a les fórmules del càlcul mostral, cal diferenciar entre les que utilitzen l'aproximació a partir de dades contínues, i per tant d'on poden extreure's indicadors de centralitat i dispersió (en el cas que ens interessa, la desviació típica). En el cas d'utilitzar la desviació típica d'un univers per aproximar-nos a la seua estimació mostral, caldrà disposar abans de la seua mesura, la qual cosa implica extraure una mesura respecte de la variable central a partir d'un estudi pilot, o bé de dades secundàries. No obstant, aquesta opció sol ser secundària donat que la major part d'enquestadores aposten pel model de les proporcions, amb el qual no cal conèixer el valor de la desviació típica de la variable central -que d'altra banda no és l'única de l'estudi- sinó més bé posen damunt la taula la possibilitat de major incertesa davant una situació probabilística dicotòmica, això és, un 0,5 sobre 1 de que ocòrrega un esdeveniment ($\hat{p}$) i un 0,5 de que ocòrrega el contrari ($\hat{q}$). En l'opció proporcional també cap la possibilitat que els valors de $\hat{p}$ i $\hat{q}$ siguen diferents de 0,5, i en estos casos, donada la menor heterogeneïtat de la mostra, la mida final també serà menor. La tercera dada que cal conèixer és la del nivell de

[55] El límit de 100.000 unitats com a factor de tria entre universos infinits i finits no era tal en Cochran (1953). Més bé, considerava que una població era infinita si el número de persones en la mostra era menys del 5% que el total de persones en l'univers. Per tal de simplificar, com que el 5% de 100.000 persones ja sumen 5.000 persones, cosa que sol ser difícil d'assumir en una investigació empírica, s'estima el canvi de fórmula a població infinita.

[56] En els casos de mostreig estratificat, sistemàtic i per conglomerats, les fórmules són més complexes. Val a dir que en els casos pràctics (per exemple, els baròmetres del Centro de Investigciones Sociológicas) el càlcul de la mida de la mostra es fa a partir d'un mostreig aleatori simple, tot i que es tracta d'un mostreig polietàpic que utilitza procediments estratificats, per conglomerats i d'aleatorització.

confiança. Com s'ha apuntat anteriorment, el nivell de confiança és una mesura estandarditzada sobre la qual hi ha bastant acord a assenyalar 2σ com a mesura estàndard, la qual cosa es tradueix en un nivell de confiança (Z) del 95,5%, o el que és el mateix, que de cada 100 enquestes sobre la mateixa mostra, només un 4,5% donarien resultats fora de l'interval de confiança resultant d'una mesura. Nivells més baixos de confiança donarien mostres més xicotetes, mentre que nivells més alts augmentarien la mida de la mostra fins a extrems no assumibles amb els recursos propis d'una empresa o institució acadèmica Per últim, quedaria per conèixer l'error estimat. La mesura de l'error ofereix el marge d'error màxim que estem disposats a assumir. Com s'explicava anteriorment, marges d'error inferiors a 2% disparen la mida de la mostra, per la qual cosa se sol situar l'estàndard entre el 2% i el 3%, tot i que no és estrany trobar-se enquestes on l'error s'aproxima al 5%.

L'aproximació que fa Cochran, que podem veure desenvolupada en Scheaffer *et al.* (2012: p. 75 i ss.) es basa en l'estimació de la variància d'$\bar{y}$ (D) a partir de dues desviacions típiques de la seua estimació, de manera que:

$$D = \frac{B^2}{4}$$

On B és el resultat de l'error d'estimació, que en la majoria dels problemes se'ns dóna com a número sencer.

A partir del valor D es pot calcular la mida de la mostra necessària per tal de dur a terme un càlcul inferencial basat en la mitjana a partir de la següent fórmula:

$$n = \frac{Ns^2}{(N-1)D + s^2}$$

De la mateixa manera, el càlcul per a un càlcul inferencial basat en les propocions es basaria en la següent fórmula:

$$n = \frac{N\hat{p}\hat{q}}{(N-1)D + \hat{p}\hat{q}}$$

D'on, en ambdues fórmules, es pot substituir en el denominador $(N-1)$ per N sempre que la població siga suficientment gran, és a dir, sempre que no calga fer la correcció per a la població finita observada anteriorment.

Una simplificació d'estes fórmules es pot trobar en bibliografia en castellà i en anglès (Sierra Bravo, 2001: p. 227 i ss.; Cea, 2012: p. 170; Spiegel, 2018: p. 205 i ss.) i, amb altra notació matemàtica, en Moore (2007: p. 354; 2007: p. 502). Així, es pot diferenciar entre dues fórmules, com avançàvem anteriorment: d'una banda, l'estimació de la mida mostra a partir de dades contínues (és a dir, a partir de la seua desviació típica), que es calcularia a partir de la fórmula:

$$n = \frac{Z^2 \hat{s}^2}{e^2}$$

On n és la mida de la mostra; Z el nivell de confiança tipificat; s és la desviació típica; i e és l'error).

La segona via és l'estimació de la mida de la mostra a partir del càlcul de les proporcions, en aquest cas a partir de la fórmula:

$$n = \frac{Z^2 \hat{p}\hat{q}}{e^2}$$

On $\hat{p}$ representa la probabilitat que ocòrrega un determinat esdeveniment i $\hat{q}$ la probabilitat inversa).

Per a les poblacions finites, Cochran aplica el factor de correcció de mostres finites, la qual cosa implica posar en contacte la fórmula anterior amb el factor $\frac{1}{N-1}$ (on N representa la mida exacta o més ben aproximada de l'univers), de la qual cosa es desprèn per al càlcul amb dades contínues:

$$n = \frac{Z^2 \hat{s}^2 N}{e^2(N-1) + Z^2 \hat{s}^2}$$

I per a les proporcions:

$$n = \frac{Z^2 \hat{P} \hat{Q} N}{e^2 (N-1) + Z^2 \hat{P} \hat{Q}}$$

Per suposat, aïllant l'error podem aproximar-lo a partir del nivell de confiança i la mida estimada de la mostra, de la mateixa manera que podem obrar amb la resta de factors de l'equació de Cochran, de manera tal que per al càlcul per la via de la mitjana es podria calcular el nivell de confiança a partir de la següent fórmula:

$$Z = \frac{e\sqrt{n}}{s}$$

Que en el cas de mostres finites quedaria:

$$Z = \frac{e\sqrt{n(N-1)}}{s\sqrt{N-n}}$$

De la mateixa manera, el càlcul per a l'aproximació per la via de les proporcions seria:

$$Z = \frac{e\sqrt{n}}{\sqrt{\hat{p}\hat{q}}}$$

Que per a població finita resultaria en:

$$Z = \frac{e\sqrt{n(N-1)}}{\sqrt{\hat{p}\hat{q}(N-n)}}$$

Per contra, l'error mostral es podria calcular, en el cas de les poblacions infinites i aproximació per la mitjana en:

$$e = \frac{Zs}{\sqrt{n}}$$

Que en poblacions finites seria:

$$e = Zs\sqrt{\frac{N-n}{n(N-1)}}$$

I per a l'aproximació per la via de les proporcions i poblacions infinites:

$$e = \frac{Z\sqrt{\hat{p}\hat{q}}}{\sqrt{n}}$$

I en la població finita:

$$e = Z\sqrt{\frac{\hat{p}\hat{q}(N-n)}{n(N-1)}}$$

En la pràctica, i això és una cosa que caldrà treballar a classe, la persona encarregada del càlcul de la mida de la mostra juga amb les xifres per a quadrar un error no massa elevat amb una mida suficientment gran com per a que s'ajuste al pressupost de què es disposa per a fer-la realitat. La pràctica investigadora ofereix ferramentes per tal de situar, en funció de quin siga el límit de l'univers, una quantitat més o menys estable de mostra a partir de la qual l'error oscil·la en uns termes raonables. Amb estes indicacions, aleshores, es pot dir que franquegem la barrera d'allò demostrable empíricament i entrem en un terreny axiològic en què s'ha de sospesar de quina manera ens aproximem a les millors dades disponibles amb un pressupost donat.

3. Bibliografia

Agresti, Alan (2018). *Statistical Methods for the Social Sciences*. Harlow: Pearson.

Alasuutari, Pertti, Bickman, Leonard i Brannen, Julia (Eds.) (2008). *The SAGE handbook of social research methods*. Lonndon: Sage.

Almazán, Alejandro, Arribas, José, Camarero, Luis, Mañas, Beatriz i Vallejos, Antonio (2015). *Análisis estadístico para la investigación social*. Madrid: Ibergarceta Publicaciones.

Aronson, Jeffrey (2001). Francis Galton and the invention of terms for quantiles. *Journal of Clinical Epidemiology*, 54(12), 1191-1194.

Asensi-Artiga, Vivina i Parra-Pujante, Antonio (2002). El método científico y la nueva filosofía de la ciencia. *Anales de Documentación*, 5, 9-19.

Babbie, Earl (2013). *The practice of social research*. Belmont: Wadsworth.

Bachelard, Gaston (2002[1938]). *The formation of the Scientific Mind. A Contribution to a Psychoanalysis of Objective Knowledge*. Manchester: Clinamen Press.
nn
Benjamin, Daniel *et al.* (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6-10.

Blalock Jr., Hubert (1979). *Social Statistics*. New York: Mc Graw Hill.

Bourdieu, Pierre, Chamboredon, Jean Claude i Passeron, Jean Claude (2002). *El oficio de sociólogo. Presupuestos epistemológicos.* Buenos Aires: Siglo XXI.

Bourdieu, Pierre (2001). *Poder, derecho y clases sociales*. Bilbao: Desclée de Brouwer.

Bowley, Arthur i Burnett-Hurst, Alexander (1915). *Livelihood and poverty: a study in the economic conditions of working-class households in Northampton, Warrington, Stanley and Reading*. London: G. Bell and Sons.

Brain, Robert (2001). The ontology of the questionnaire: Max Weber on measurement and mass investigation. *Studies in History and Philosophy of Science*, 32(4), 647-684.

Bunge, Mario (2004). *La investigación científica. Su estrategia y su filosofía.* México DF: Siglo XXI.

Byrman, Alan (2004). *Social Research Methods*. Oxford: Oxford University Press.

Camarero, Luis, Almazan, Alejandro, Arribas, José et al. (2013). *Estadística para la investigación social*. Madrid: Ibergarceta Publicaciones.

Campbell, Donald i Fiske, Donald (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81-105.

Campos, Raymundo i Monroy-Gómez-Franco, Luis (2016). La relación entre crecimiento económico y pobreza en México. *Investigación económica*, 75(298), 77-113.

Cea, María Ángeles (2012). *Fundamentos y aplicaciones en metodología cuantitativa*. Madrid: Síntesis.

Cochran, William (1953). *Sampling Techniques*. New York: John Wiley and Sons.

Cohen, Jacob (1990). Things I have learned (so far). *American Psychologist*, 45(12), 1304-1312.

Cohen, Jacob (1994). The earth is round (p<.05). *American Psychologist*, 49(12), 997-1003.

Comissió de Normalització Lingüística de la Facultat de Ciències Econòmiques i Empresarials (1996). *Vocabulari d'estadística*. Barcelona: Publicacions de la Universitat de Barcelona.

Comte, Auguste (1848). *Discours sur l'ensemble du positivism*. Paris: Société Positiviste Internationale.

Comte, Auguste (1851-1954). *Système de politique positiviste, ou traité de sociologie, institutant la religion de l'humanité*. Paris: E. Thunot et cie.

Corbetta, Piergiorgio (2007). *Metodología y técnicas de investigación social.* Madrid: Mc Graw Hill

Creswell, John (2014). *Research Design. Qualitative, Quantitative and Mixed Methods Approaches*. London: Sage.

Denzin, Norman (1970). *The Research Act: A Theoretical Introduction to Sociological Methods*. London: Butterworth.

Desrosières, Alain (1998). *The Politics of Large Numbers. A History of Statistical Reasoning.* Cambridge: Harvard University Press.

Díaz de Rada, Vidal (2015). *Manual de trabajo de campo en la encuesta: (presencial y telefónica)*. Madrid: Centro de Investigaciones Sociológicas.

Dilthey, Wilhelm (1949 [1883]). *Introducción a las ciencias del espíritu*. México DF: México.

Duque, Ignacio (2003). El momento fundacional de las ciencias sociales españolas contemporáneas y el Ateneo como crisol, escena y pósito: paradojas y perspectivas de su horizonte teórico, investigaciones concretas e intervención social reformista. En AADD, *Centenario de la" información de 1901" del Ateneo de Madrid sobre "Oligarquía y Caciquismo*" (pp. 245-330), Madrid: Fundamentos.

Durkheim, Émile (1895). *Les règles de la méthode sociologique*. Paris: Félix Alcan.

Echegaray, Lázaro (2018). *Historia de la investigación social*. Madrid: Esic Editorial.

Eisenhart, Churchill (1974). *The development of the concept of the best mean of a set of measurements from antiquity to the present day*. Manuscrit inèdit. Recuperat de http://galton.uchicago.edu/~stigler/eisenhart.pdf.

Fisher, Ronald (1922). On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85 (1), 87-94.

Fisher, Ronald (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher, Ronald i Yates, Frank (1974). *Statistical tables: For biological, agricultural and medical research*. Edinburgh: Oliver and Boyd.

Francés, Francisco, Alaminos, Antonio, Penalva, Clemente i Santacreu, Óscar (2014). *El proceso de medición de la realidad social: La investigación a través de encuestas*. Cuenca: Pydlos ediciones.

Frankfort-Nachmias, Chava i Leon-Guerrero, Anna (2018). *Social statistics for a diverse society*. Thousand Oaks: Pine Forge Press.

Galton, Francis (1907a). Vox Populi. *Nature*, 75 (1952), 450–451.

Galton, Francis (1907b). The Ballot Box. *Nature*, 75 (1949), 509-51

García, Manuel (1974). *Sobre el método. Problemas de la investigación empírica en sociología*. Madrid: Centro de Investigaciones Sociológicas.

Ghasemi, Asghar i Zahediasl, Saleh (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International journal of endocrinology and metabolism*, 10(2), 486-489.

Gigerenzer, Gerd i Murray, David (1987). *Cognition as intuitive statistics*. Hillsdale: Lawrence Erlbaum Associates.

Gigerenzer, Gerd i Marewski, Julian (2015). Surrogate Science: The Idol of a Universal Method for Scientific Inference. *Journal of Management*, 41(2), 421-440.

Giner, Jordi (2015). *Retorn de persones retirades d'origen britànic residents a la Marina Alta*. Tesi Doctoral. València: Universitat de València.

Glass, Gene i Stanley, Julian (1974). *Métodos estadísticos aplicados a las ciencias sociales*. Madrid: Prentice Hall.

González, Pedro (2000). Medir en las ciencias sociales. En Manuel García Ferrando, Jesús Ibáñez i Francisco Alvira, *El Análisis de la realidad social. Métodos y técnicas de investigación social* (pp. 344-407). Madrid: Alianza.

Haberman, Shelby (1973). The Analysis of Residuals in Cross-Classified Tables. *Biometrika*, 29(1), 205-220.

Healey, Joseph (2016). *The essentials of statistics: A tool for social research*. Boston: Cengage Learning.

Hernández, Roberto, Fernández, Carlos i Baptista, Pilar (2014). Metodología de la investigación. México DF: McGraw-Hill.

Higgins, Peter (2008). *Number Story. From Counting to Cryptography*. Londres: Springer Verlag.

Hotchkiss, Michael (2005). *Princeton sociologist Walter Wallace dies at age 88*. Recuperada de https://www.princeton.edu/news/2015/12/08/princeton-sociologist-walter-wallace-dies-age-88.

Ibáñez, Jesús (2002). Perspectivas de la investigación social: el diseño en las tres perspectivas. En Manuel García, Jesús Ibáñez i Francisco Alvira (eds.): *El análisis de la realidad social. Métodos y técnicas de investigació* (pp. 57-98). Madrid: Alianza

Ioannidis, John P. (2005). Why most published research findings are false. *Plos Medicine*, 2(8), e124.

Kalton, Graham (1983). *Introduction to survey sampling*. London: Sage.

Kaufman, Michael (2003). Robert K. Merton, Versatile Sociologist and Father of the Focus Group, Dies at 92. *The New York Times*, 24 febrer, secció B, 7.

Kendall, Maurice i Babington, Bernard (1938). *Tables of Random Sampling Numbers*, Cambridge: Cambridge University Press.

Kendall, Maurice (1960). Studies in the History of Probability and Statistics. Where Shall the History of Statistics Begin? *Biometrika*, 47 (3/4):447-449.

Korstanje, Maximiliano (2009). Magia y estadística. Rituales sociales contra la incertidumbre. *Aposta. Revista de Ciencias Sociales*, 41, 1-26.

Kuhn, Thomas (1962). *The Structure of Scientific Revolutions*. Chicago: The University of Chicago Press.

Landro, Alberto i González, Mirta (2012). Bernoulli, De Moivre, Bayes, Price y los fundamentos de la inferencia inductiva. *Cuadernos del CIMBAGE*, 15, 33-56.

La-Roca, Francesc (2006). *Estadística aplicada a les ciències socials*. València: Publicacions de la Universitat de València.

Lazarsfeld, Paul (1958). Evidence and inference in social research. *Daedalus*, 87(4), 99-130.

Leti, Giuseppe (2000). The birth of statistics and the origins of the new natural sciences. *Metron,* 58, 185–211.

Lie, Einar (2002). The Rise and Fall of Sampling Surveys in Norway, 1875–1906. *Science in Context*, 15(3), 385-409.

Lisón, Carmelo (1968). Una gran encuesta de 1901-1902 (Notas para la Historia de la Antropología Social en España). *Revista Española de la Opinión Pública*, 12, 83-151.

López-Roldán, Pedro i Fachelli, Sandra (2015). *Metodología de la Investigación Social Cuantitativa*. Bellaterra (Cerdanyola del Vallès): Dipòsit Digital de Documents, Universitat Autònoma de Barcelona.

Martínez, Juan Ignacio (2018). El diseño de gráficos y tablas. En Félix Requena i Luis Ayuso (coords.) *Estrategias de investigación en las ciencias sociales* (pp. 259-280). València: Tirant lo Blanch.

Mateos-Aparicio, Gregoria (2002). Historia de la probabilidad (desde sus orígenes a Laplace) y su relación con la historia de la teoría de la decisión. En Asociación de Historia de la Estadística y de la Probabilidad de España: *Historia de la Probabilidad y de la Estadística* (pp. 1-18). Madrid: Alfa Centauro.

McDonald Lynn (1994). *The women founders of the social sciences*. Ottawa: Carleton University Press.

McMullen, Launce (1939). “Student” as a man. *Biometrika*, 30(3-4), 205-210.

Merton, Robert (1949). *Social Theory and Social Structure.* New York: The Free Press.

Merton, Robert (1965). *On the Shoulders of Giants. A Shandean Postcript.* New York: The Free Press.

Moore, David (2007). *The basic practice of statistics*. New York: W.H. Freeman and Company.

Moore, David, Notz, William i Fligner, Michael (2018). *The basic practice of statistics*. New York: W.H. Freeman and Company.

National Bureau of Standards (1952). *Tables of the binomial probability distribution*. Washington: US Government Printing Office.

Neyman, Jerzy i Pearson, Egon (1933). *The testing of statistical hypotheses in relation to probabilities a priori*. *Mathematical Proceedings of the Cambridge Philosophical Society*, 29(4), 492-510.

Neyman, Jerzy (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97(4), 558-625.

Neyman, Jerzy (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London*, 236(767), 333-380.

Ornstein, Michael (2013). *A companion to survey research*. London: Sage Public.

Ortí, Alfonso (1995). La confrontación de modelos y niveles epistemológicos en la génesis e historia de la investigación social. En Juan Manuel Delgado y Juan Gutiérrez *Métodos y técnicas cualitativas de investigación en ciencias sociales*, (pp. 87-99). Madrid: Síntesis.

Pearson, Karl (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*. A, 185, 71-110.

Pearson, Karl (1896). Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. *Philosophical Transactions of the Royal Society of London*, 187, 253-318.

Pearson, Karl (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 5 (302), 157-175.

Pearson, Karl (1904). *On the theory of contingency and its relation to association and normal correlation*. London: Dulau and Co.

Pearson, Karl (1920). Notes on the history of correlation. *Biometrika*, 13(1), 25-45.

Pearson, Egon (1938). "Student" as statistician. *Biometrika*, 30(3-4), 210-250.

Rodríguez, Jacinto, Ferreras, María Luisa i Núñez, Adoración (1991). Inferencia estadística, niveles de precisión y diseño muestral. *Revista Española de Investigaciones Sociológicas*, 54, 139-162.

Rosenberg, Morris (1968). *The Logic of Survey Analysis*. New York, Basic Books Inc.

Saint-Simon, Claude-Henri (1803). *Lettres d'un habitant de Genève à ses contemporains.* Autoeditat.

Sánchez, Juan José (2001). Estadística, orden natural y orden social. *Papers: revista de sociologia*, 63, 33-46.

Scheaffer, Richard, Mendenhall, William, Ott, Lyman i Gerow, Kenneth (2012). *Elementary survey sampling*. Boston: Cengage Learning.

Sierra Bravo, Restituto (2001). *Técnicas de investigación social. Teoría y ejercicios.* Madrid, Paraninfo.

Spiegel, Murray (1990). *Estadística*. Madrid: McGraw-Hill.

Spiegel, Murray i Stephens, Larry (2018). *Statistics*. New York: McGraw-Hill.

Squire, Peverill (1988). Why the 1936 Literary Digest Poll Failed. *The Public Opinion Quarterly*, 52(1), 125-133.

Stevens, Stanley (1951). Mathematics, Measurement and Psychophysics. En Stanley Stevens (ed.) *Handbook of Experimental Psychology* (pp. 1-30.), New York: Wiley.

Stigler, Stephen M. (1989). Francis Galton's Account of the Invention of Correlation. *Statistical Science*, 4(2), 73-79.

Surowiecki, James (2005). *The wisdom of crowds*. New York: Anchor Books.

Wallace, Walter (1971). *The Logic of Science in Sociology*. Chicago: Aldine Atherton.

Weiss, Hilde (1979). Karl Marx's "Enquête ouvrière". En Tom Bottomore (dir.) *Karl Marx* (pp. 172-184), Oxford: Blackwell.

Wilson, Edwin B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209-212.

4. Annexos

4.1. Taula de probabilitats de la distribució binomial

$$f(x) \sim B(n,p)$$

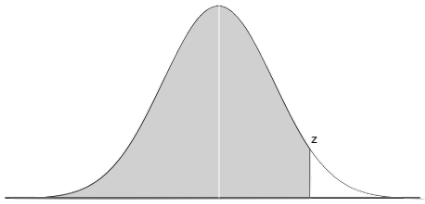
$$p(x=k) = \binom{n}{k} p^k q^{n-k}$$

n	p / k	0,01	0,05	0,10	0,15	0,20	0,25	0,30	0,333	0,35	0,40	0,45	0,49	,50
2	0	0,9801	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4444	0,4225	0,3600	0,3025	0,2601	0,2500
	1	0,0198	0,0950	0,1800	0,2550	0,3200	0,3750	0,4200	0,4444	0,4550	0,4800	0,4950	0,4998	0,5000
	2	0,0001	0,0025	0,0100	0,0225	0,0400	0,0625	0,0900	0,1111	0,1225	0,1600	0,2025	0,2401	0,2500
3	0	0,9703	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2963	0,2746	0,2160	0,1664	0,1327	0,1250
	1	0,0294	0,1354	0,2430	0,3251	0,3840	0,4219	0,4410	0,4444	0,4436	0,4320	0,4084	0,3823	0,3750
	2	0,0003	0,0071	0,0270	0,0574	0,0960	0,1406	0,1890	0,2222	0,2389	0,2880	0,3341	3674	0,3750
	3	0,0000	0,0001	0,0010	0,0034	0,0080	0,0156	0,0270	0,0370	0,0429	0,0640	0,0911	0,1176	0,1250
4	0	0,9606	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1975	0,1785	0,1296	0,0915	0,0677	0,0625
	1	0,0388	0,1715	0,2916	0,3685	0,4096	0,4219	0,4116	0,3951	0,3845	0,3456	0,2995	0,2600	0,2500
	2	0,0006	0,0135	0,0486	0,0975	0,1636	0,2109	0,2646	0,2963	0,3105	0,3456	0,3675	0,3747	0,3750
	3	0,0000	0,0005	0,0036	0,0115	0,0256	0,4609	0,0756	0,0988	0,1115	0,1536	0,2005	0,2400	0,2500
	4	0,0000	0,0000	0,0001	0,0005	0,0016	0,0039	0,0081	0,0123	0,0150	0,0256	0,0410	0,0576	0,0625
5	0	0,9510	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1317	0,1160	0,0778	0,0503	0,0345	0,0312
	1	0,0480	0,2036	0,3280	0,3915	0,4096	0,3855	0,3602	0,3292	0,3124	0,2592	0,2059	0,1657	0,1562
	2	0,0010	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3292	0,3364	0,3456	0,3369	0,3185	0,3125
	3	0,0000	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1646	0,1811	0,2304	0,2757	0,3060	0,3125
	4	0,0000	0,0000	0,0004	0,0022	0,0064	0,0146	0,0284	0,0412	0,0488	0,0768	0,1128	0,1470	0,1562
	5	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0041	0,0053	0,0102	0,0185	0,0283	0,0312
6	0	0,9415	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0878	0,0754	0,0467	0,0277	0,0176	0,0156
	1	0,0571	0,2321	0,3543	0,3993	0,3932	0,3560	0,3025	0,2634	0,2437	0,1866	0,1359	0,1014	0,0938
	2	0,0014	0,0305	0,0984	0,1762	0,2458	0,2966	0,3241	0,3292	0,3280	0,3110	0,2780	0,2437	0,2344
	3	0,0000	0,0021	0,0146	0,0415	0,0819	0,1318	0,1852	0,2195	0,2355	0,2765	0,3032	0,3121	0,3125
	4	0,0000	0,0001	0,0012	0,0055	0,0154	0,0330	0,0595	0,0823	0,0951	0,1382	0,1861	0,2249	0,2344
	5	0,0000	0,0000	0,0001	0,0004	0,0015	0,0044	0,0102	0,0165	0,0205	0,0369	0,0609	0,0864	0,0938
	6	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0014	0,0018	0,0041	0,0083	0,0139	0,0156
7	0	0,9321	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0585	0,0490	0,0280	0,0152	0,0090	0,0078
	1	0,0659	0,2573	0,3720	0,3960	0,3670	0,3115	0,2471	0,2048	0,1848	0,1306	0,0872	0,0603	0,0574
	2	0,0020	0,0406	0,1240	0,2097	0,2753	0,3115	0,3177	0,3073	0,2985	0,2613	0,2140	0,1740	0,1641
	3	0,0000	0,0036	0,0230	0,0617	0,1147	0,1730	0,2269	0,2561	0,2679	0,2903	0,2918	0,2786	0,2734
	4	0,0000	0,0002	0,0026	0,0109	0,0287	0,0577	0,0972	0,1280	0,1442	0,1935	0,2388	0,2676	0,2734
	5	0,0000	0,0000	0,0002	0,0012	0,0043	0,0115	0,0250	0,0384	0,0466	0,0774	0,1172	0,1543	0,1641
	6	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0036	0,0064	0,0084	0,0172	0,0320	0,0494	0,0547
	7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0005	0,0006	0,0016	0,0037	0,0068	0,0078
8	0	0,9227	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0390	0,0319	0,0168	0,0084	0,0046	0,0039
	1	0,0746	0,2793	0,3826	0,3847	0,3355	0,2670	0,1977	0,1561	0,1373	0,0896	0,0548	0,0352	0,0312
	2	0,0026	0,0515	0,1488	0,2376	0,2936	0,3115	0,2965	0,2731	0,2587	0,2090	0,1569	0,1183	0,1094
	3	0,0001	0,0054	0,0331	0,0839	0,1468	0,2076	0,2541	0,2731	0,2786	0,2787	0,2568	0,2273	0,2188
	4	0,0000	0,0004	0,0046	0,0185	0,0459	0,0865	0,1361	0,1707	0,1875	0,2322	0,2627	0,2730	0,2734
	5	0,0000	0,0000	0,0004	0,0026	0,0092	0,0231	0,0467	0,0683	0,0808	0,1239	0,1719	0,2098	0,2188
	6	0,0000	0,0000	0,0000	0,0002	0,0011	0,0038	0,0100	0,0171	0,0217	0,0413	0,0703	0,1008	0,1094
	7	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0012	0,0024	0,0033	0,0079	0,0164	0,0277	0,0312
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0002	0,0007	0,0017	0,0033	0,0039
9	0	0,9135	0,6302	0,3874	0,2316	0,1342	0,0751	0,0404	0,0260	0,0207	0,0101	0,0046	0,0023	0,0020
	1	0,0830	0,2985	0,3874	0,3679	0,3020	0,2253	0,1556	0,1171	0,1004	0,0605	0,0339	0,0202	0,0176
	2	0,0034	0,0629	0,1722	0,2597	0,3020	0,3003	0,2688	0,2341	0,2162	0,1612	0,1110	0,0776	0,0703
	3	0,0001	0,0077	0,0446	0,1069	0,1762	0,2336	0,2668	0,2731	0,2716	0,2508	0,2119	0,1739	0,1641
	4	0,0000	0,0006	0,0074	0,0283	0,0661	0,1168	0,1715	0,2048	0,2194	0,2508	0,2600	0,2506	0,2461
	5	0,0000	0,0000	0,0008	0,0050	0,0165	0,0389	0,0735	0,1024	0,1181	0,1672	0,2128	0,2408	0,2461
	6	0,0000	0,0000	0,0001	0,0006	0,0028	0,0087	0,0210	0,0341	0,0424	0,0743	0,1160	0,1542	0,1641
	7	0,0000	0,0000	0,0000	0,0000	0,0003	0,0012	0,0039	0,0073	0,0098	0,0212	0,0407	0,0635	0,0703
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0009	0,0013	0,0035	0,0083	0,0153	0,0176
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0001	0,0003	0,0008	0,0016	0,0020
10	0	0,9044	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0173	0,0135	0,0060	0,0025	0,0012	0,0010
	1	0,0914	0,3151	0,3874	0,3474	0,2684	0,1877	0,1211	0,0867	0,0725	0,0403	0,0207	0,0114	0,0098
	2	0,0042	0,0746	0,1937	0,2759	0,3020	0,2816	0,2335	0,1951	0,1757	0,1209	0,0763	0,0495	0,0439
	3	0,0001	0,0105	0,0574	0,1298	0,2013	0,2503	0,2668	0,2601	0,2522	0,2150	0,1665	0,1267	0,1172
	4	0,0000	0,0010	0,0112	0,0401	0,0881	0,1460	0,2001	0,2276	0,2377	0,2508	0,2384	0,2130	0,2051
	5	0,0000	0,0001	0,0015	0,0085	0,0264	0,0584	0,1029	0,1366	0,1536	0,2007	0,2340	0,2456	0,2461
	6	0,0000	0,0000	0,0001	0,0012	0,0055	0,0162	0,0368	0,0596	0,0689	0,1115	0,1596	0,1966	0,2051
	7	0,0000	0,0000	0,0000	0,0001	0,0008	0,0031	0,0090	0,0163	0,0212	0,0425	0,0746	0,1080	0,1172
	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0030	0,0043	0,0106	0,0229	0,0389	0,0439
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0005	0,0016	0,0042	0,0083	0,0098
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0010

Font: National Bureau of Standards (1952: p. 1 i ss.)

4.2. Taula de probabilitats de la distribució normal tipificada

$$\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx$$



$$p(Z \leq z)$$

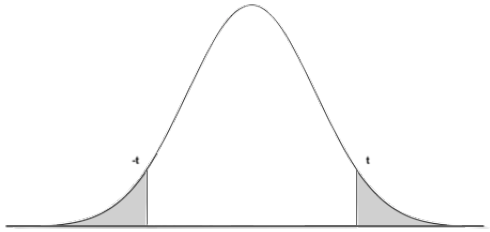
z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5754
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7258	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7518	0,7549
0,7	0,7580	0,7612	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7996	0,8023	0,8051	0,8079	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9430	0,9441
1,6	0,9452	0,9463	0,9474	0,9485	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9700	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9762	0,9767
2,0	0,9773	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9865	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9980	0,9980	0,9981
2,9	0,9981	0,9982	0,9983	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	1,0000	1,0000	1,0000
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

∝	0,25	0,20	0,15	0,10	0,05	0,025	0,01	0,005	0,001	0,0005
Z	0,674	0,841	1,036	1,282	1,645	1,960	2,326	2,576	3,091	3,291

Font: Fisher i Yates (1974: p. 45) i Moore (2007: p. 687)

4.3. Taula de probabilitats de la distribució T de Student

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma\left(\frac{v}{2}\right)} 1 + \left(\frac{t^2}{v}\right)^{-\frac{v+1}{2}}$$

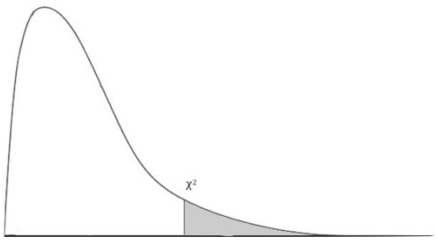


Nivell de significació per a proves d'una cua						
	0,1	0,05	0,025	0,01	0,005	0,0005
Nivell de significació per a proves de dues cues						
gl	0,2	0,1	0,05	0,02	0,01	0,001
1	3,078	6,314	12,706	31,821	63,657	636,619
2	1,886	2,920	4,303	6,965	9,925	31,598
3	1,638	2,353	3,182	4,541	5,841	12,941
4	1,533	2,132	2,776	3,747	4,604	8,610
5	1,476	2,015	2,571	3,365	4,032	6,859
6	1,440	1,943	2,447	3,143	3,707	5,959
7	1,415	1,895	2,365	2,998	3,499	5,405
8	1,397	1,860	2,306	2,896	3,355	5,041
9	1,383	1,833	2,262	2,821	3,250	4,781
10	1,372	1,812	2,228	2,764	3,169	4,587
11	1,363	1,796	2,201	2,718	3,106	4,437
12	1,356	1,782	2,179	2,681	3,055	4,318
13	1,350	1,771	2,160	2,650	3,012	4,221
14	1,345	1,761	2,145	2,624	2,977	4,140
15	1,341	1,753	2,131	2,602	2,947	4,073
16	1,337	1,746	2,120	2,583	2,921	4,015
17	1,333	1,740	2,110	2,567	2,898	3,965
18	1,330	1,734	2,101	2,552	2,878	3,922
19	1,328	1,729	2,093	2,539	2,861	3,883
20	1,325	1,725	2,086	2,528	2,845	3,850
21	1,323	1,721	2,08	2,518	2,831	3,819
22	1,321	1,717	2,074	2,508	2,819	3,792
23	1,319	1,714	2,069	2,500	2,807	3,767
24	1,318	1,711	2,064	2,492	2,797	3,745
25	1,316	1,708	2,060	2,485	2,787	3,725
26	1,315	1,706	2,056	2,479	2,779	3,707
27	1,314	1,703	2,052	2,473	2,771	3,690
28	1,313	1,701	2,048	2,467	2,763	3,674
29	1,311	1,699	2,045	2,462	2,766	3,659
30	1,310	1,697	2,042	2,457	2,760	3,646
40	1,303	1,684	2,021	2,423	2,704	3,551
60	1,296	1,671	2,000	2,390	2,660	3,460
120	1,289	1,658	1,980	2,358	2,617	3,373
∞	1,282	1,645	1,960	2,326	2,576	3,291

Font: Fisher i Yates (1974: p. 46)

4.4. Taula de probabilitats de la distribució khi quadrat

$$f(x|v) = \frac{x^{\frac{v-2}{2}} e^{\frac{-x}{2}}}{2^{\frac{v}{2}} \Gamma\left(\frac{v}{2}\right)}$$

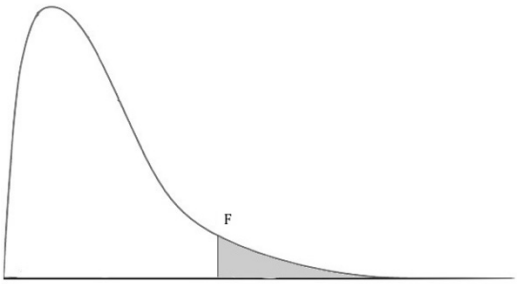


df	0,99	0,98	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,000157	0,000628	0,00393	0,0158	0,0642	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,0201	0,0404	0,103	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815
3	0,115	0,185	0,352	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,341	16,268
4	0,297	0,429	0,711	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,465
5	0,554	0,752	1,145	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,086	20,517
6	0,872	1,134	1,635	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	1,239	1,564	2,167	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	1,646	2,032	2,733	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125
9	2,088	2,532	3,325	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	2,558	3,059	3,940	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	3,053	3,609	4,575	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	3,571	4,178	5,226	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	4,107	4,765	5,892	7,042	8,634	9,926	12,340	15,119	16,985	19,812	22,362	25,472	27,688	34,528
14	4,660	5,368	6,571	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,123
15	5,229	5,985	7,261	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,697
16	5,812	6,614	7,962	9,312	11,152	12,624	15,338	18,418	20,465	23,542	26,296	29,633	32,000	39,252
17	6,408	7,255	8,672	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,790
18	7,015	7,906	9,390	10,865	12,857	14,440	17,338	20,601	22,760	25,989	28,869	32,346	34,805	42,312
19	7,633	8,567	10,117	11,651	13,716	15,352	18,338	21,689	23,900	27,204	30,144	33,687	36,191	43,820
20	8,260	9,237	10,851	12,443	14,578	16,266	19,337	22,775	25,038	28,412	31,410	35,020	37,566	45,315
21	8,897	9,915	11,591	13,240	15,445	17,182	20,337	23,858	26,171	29,615	32,671	36,343	38,932	46,797
22	9,542	10,600	12,338	14,041	16,314	18,101	21,337	24,939	27,301	30,813	33,924	37,659	40,289	48,268
23	10,196	11,293	13,091	14,848	17,187	19,021	22,337	26,018	28,429	32,007	35,172	38,968	41,638	49,728
24	10,856	11,992	13,848	15,659	18,062	19,943	23,337	27,096	29,553	33,196	36,415	40,270	42,980	51,179
25	11,524	12,697	14,611	16,473	18,940	20,867	24,337	28,172	30,675	34,382	37,652	41,566	44,314	52,620
26	12,198	13,409	15,379	17,292	19,820	21,792	25,336	29,246	31,795	35,563	38,885	42,856	45,642	54,052
27	12,879	14,125	16,151	18,114	20,703	22,719	26,336	30,319	32,912	36,741	40,113	44,140	46,963	55,476
28	13,565	14,847	16,928	18,939	21,588	23,647	27,336	31,391	34,027	37,916	41,337	45,419	48,278	56,893
29	14,256	15,574	17,708	19,768	22,475	24,577	28,336	32,461	35,139	39,087	42,557	46,693	49,588	58,302
30	14,953	16,306	18,493	20,599	23,364	25,508	29,336	33,530	36,250	40,256	43,773	47,962	50,892	59,703

Font: Fisher i Yates (1974: p. 47)

4.5. Taula de probabilitats de la distribució F de Fisher-Snedecor

$$f(x; gl_1, gl_2) = \frac{\sqrt{\frac{(gl_1 x)^{gl_1} gl_2{}^{gl_2}}{(gl_1 x + gl_2)^{(gl_1+gl_2)}}}}{x\mathrm{B}\left(\frac{gl_1}{2}, \frac{gl_2}{2}\right)}$$

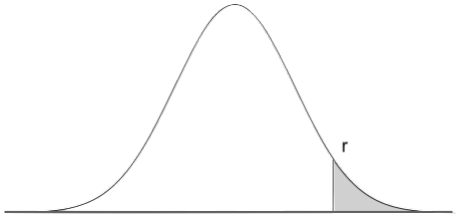


Per a $\propto = 0{,}05$

		Graus de llibertat del numerador (gl_2)									
		1	2	3	4	5	6	8	16	24	∞
Graus de llibertat del denominador (gl_1)	1	161,4	199,5	215,7	224,6	230,2	234,0	238,9	243,9	249,0	254,3
	2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
	3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
	4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
	5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
	6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
	7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
	8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
	9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
	10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
	11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
	12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
	13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
	14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
	15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
	16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
	17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
	18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
	19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
	20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
	21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
	22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
	23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
	24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
	25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
	26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
	27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
	28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
	29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
	30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
	40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
	60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
	120	3,92	3,07	2,68	2,45	2,29	2,17	2,02	1,83	1,61	1,25
	∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

Font: Fisher i Yates (1974: p. 53)

4.6. Taula de probabilitats de la correlació r



n \ p	Probabilitat per dalt de r									
	0,20	0,10	0,05	0,025	0,02	0,01	0,005	0,0025	0,001	0,0005
3	0,8090	0,9511	0,9877	0,9969	0,9980	0,9995	0,9999	1,0000	1,0000	1,0000
4	0,6000	0,8000	0,9000	0,9500	0,9600	0,9800	0,9900	0,9950	0,9980	0,9990
5	0,4919	0,6870	0,8054	0,8783	0,8953	0,9343	0,9587	0,9740	0,9859	0,9911
6	0,4257	0,6084	0,7293	0,8114	0,8319	0,8822	0,9172	0,9417	0,9633	0,9741
7	0,3803	0,5509	0,6694	0,7545	0,7766	0,8329	0,8745	0,9056	0,9350	0,9509
8	0,3468	0,5067	0,6215	0,7067	0,7295	0,7887	0,8343	0,8697	0,9049	0,9249
9	0,3208	0,4716	0,5822	0,6664	0,6892	0,7498	0,7977	0,8359	0,8751	0,8983
10	0,2998	0,4428	0,5494	0,6319	0,6546	0,7155	0,7646	0,8046	0,8467	0,8721
11	0,2825	0,4187	0,5214	0,6021	0,6244	0,6851	0,7348	0,7759	0,8199	0,8470
12	0,2678	0,3981	0,4973	0,5760	0,5980	0,6581	0,7079	0,7496	0,7950	0,8233
13	0,2552	0,3802	0,4762	0,5529	0,5745	0,6339	0,6835	0,7255	0,7717	0,8010
14	0,2443	0,3646	0,4575	0,5324	0,5536	0,6120	0,6614	0,7034	0,7501	0,7800
15	0,2346	0,3507	0,4409	0,5140	0,5347	0,5923	0,6411	0,6831	0,7301	0,7604
16	0,2260	0,3383	0,4259	0,4973	0,5177	0,5742	0,6226	0,6643	0,7114	0,7419
17	0,2183	0,3271	0,4124	0,4821	0,5021	0,5577	0,6055	0,6470	0,6940	0,7247
18	0,2113	0,3170	0,4000	0,4683	0,4878	0,5425	0,5897	0,6308	0,6777	0,7084
19	0,2049	0,3077	0,3887	0,4555	0,4747	0,5285	0,5751	0,6158	0,6624	0,6932
20	0,1991	0,2992	0,3783	0,4438	0,4626	0,5155	0,5614	0,6018	0,6481	0,6788
21	0,1938	0,2914	0,3687	0,4329	0,4513	0,5034	0,5487	0,5886	0,6346	0,6652
22	0,1888	0,2841	0,3598	0,4227	0,4409	0,4921	0,5368	0,5763	0,6219	0,6524
23	0,1843	0,2774	0,3515	0,4132	0,4311	0,4815	0,5256	0,5647	0,6099	0,6402
24	0,1800	0,2711	0,3438	0,4044	0,4219	0,4716	0,5151	0,5537	0,5986	0,6287
25	0,1760	0,2653	0,3365	0,3961	0,4133	0,4622	0,5052	0,5434	0,5879	0,6178
26	0,1723	0,2598	0,3297	0,3882	0,4052	0,4534	0,4958	0,5336	0,5776	0,6074
27	0,1688	0,2546	0,3233	0,3809	0,3976	0,4451	0,4869	0,5243	0,5679	0,5974
28	0,1655	0,2497	0,3172	0,3739	0,3904	0,4372	0,4785	0,5154	0,5587	0,5880
29	0,1624	0,2451	0,3115	0,3673	0,3835	0,4297	0,4705	0,5070	0,5499	0,5790
30	0,1594	0,2407	0,3061	0,3610	0,3770	0,4226	0,4629	0,4990	0,5415	0,5703
40	0,1368	0,2070	0,2638	0,3120	0,3261	0,3665	0,4026	0,4353	0,4741	0,5007
50	0,1217	0,1843	0,2353	0,2787	0,2915	0,3281	0,3610	0,3909	0,4267	0,4514
60	0,1106	0,1678	0,2144	0,2542	0,2659	0,2997	0,3301	0,3578	0,3912	0,4143
80	0,0954	0,1448	0,1852	0,2199	0,2301	0,2597	0,2864	0,3109	0,3405	0,3611
100	0,0851	0,1292	0,1654	0,1966	0,2058	0,2324	0,2565	0,2786	0,3054	0,3242
1000	0,0266	0,0406	0,0520	0,0620	0,0650	0,0736	0,0814	0,0887	0,0976	0,1039

Font: Moore (2007: p. 693).