Tècniques quantitatives per a l'Administració pública

PID_00267224

Camilo Cristancho Mantilla

Temps mínim de dedicació recomanat: 5 hores

Universitat Oberta de Catalunya

Camilo Cristancho Mantilla

Investigador postdoctoral en el Departament de Ciències Polítiques de la Universitat de Barcelona. És membre del grup de recerca Qualitat de la Democràcia i la seva recerca més recent se centra en desigualtats i processos d'influència política. El seu treball es basa en mètodes de ciència social computacional, experiments i anàlisi estadística.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Daniel Rajmil (2019)

Primera edició: setembre 2019

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

Índex

Introducció

L'Administració pública s'enfronta a problemes i desafiaments d'enormes proporcions i complexitat, com ara els dèficits d'habitatge, la pobresa, la inseguretat o l'atur, entre molt altres. Això implica la necessitat de comprendre molt bé els problemes per a poder aprofitar de la millor manera els fons públics i buscar les millors solucions. Generalment, l'observació d'aquestes situacions implica comprendre amb dades agregades el comportament o les circumstàncies dels individus, les famílies o els grups d'interès, per exemple les escoles o els barris. També comprèn la necessitat d'avaluar com canvien aquests comportaments i circumstàncies davant de determinades intervencions socials, com ara els programes de govern o davant dels canvis institucionals propiciats per les lleis i les regulacions.

El gestor públic necessita tenir les eines adequades per a afrontar aquests reptes en la mesura en què és responsable de donar les millors solucions i de garantir a més l'eficiència de polítiques finançades amb impostos. Afortunadament, l'estadística proveeix les eines per a interpretar amb rigor aquestes situacions complexes. D'una banda, l'ús d'estadístiques permet fer l'observació i el seguiment d'un gran nombre de característiques o atributs dels subjectes d'estudi (individus, famílies, col·lectius, etc.). Per a això, es creen variables a partir dels conceptes d'interès que mesuren rigorosament els atributs. D'altra banda, les estadístiques permeten interpretar les variables i avaluar-les agregadament per a comprendre fenòmens col·lectius i patrons temporals. D'aquesta manera, és possible determinar les necessitats de la despesa social, crear mesuraments per a definir criteris d'assignació de recursos i calcular l'acompliment de les inversions socials. Aquest tipus de tasques implica tenir la capacitat d'examinar una gran quantitat de variables simultàniament per a ordenar i donar sentit a situacions socials complexes.

A més a més, el gestor públic té la responsabilitat de seguir i avaluar l'evidència disponible sobre altres intervencions públiques. Això implica comprendre l'evidència o les afirmacions en la recerca acadèmica o aplicada que es publica en informes oficials, avaluacions, comunicats de premsa i articles acadèmics. Per això és necessari desenvolupar habilitats per a avaluar les conclusions empíriques i la validesa dels mètodes de recerca. Això és especialment rellevant en les avaluacions d'impacte que proporcionen informació sobre els efectes que un programa pot tenir sobre la població beneficiària. Els mètodes per a conèixer si aquests efectes són atribuïbles al programa solen ser força complexos, però aquesta guia proporciona les bases per a aproximar-se a aquest tipus de treballs estadístics. A més, té com a objectiu desenvolupar les habilitats bàsiques per a fer anàlisis de gran utilitat, que són la base d'estudis més sofisticats.

Aquesta guia pretén introduir tècniques d'anàlisi estadística d'una manera aplicada als problemes de la gestió pública i de la manera més pràctica possible. Això significa que s'ha tractat de simplificar al màxim les nocions matemàtiques i teòriques amb la finalitat de presentar els mètodes aplicats utilitzant dades reals o fictícies per a possibilitar càlculs simples.

La guia es divideix en cinc seccions. En la primera, s'introdueixen conceptes de mesurament, com els tipus de variables i la gestió de les dades. En la segona secció, es presenten conceptes i tècniques d'anàlisi descriptiva per a una sola variable. En la tercera secció, es tracta el tema de les relacions entre variables categòriques; en aquest punt, s'introdueix el tema del disseny de recerca i la prova d'hipòtesi. En la quarta secció, s'introdueixen les tècniques d'anàlisi de relacions entre variables contínues, i en la cinquena i última secció, s'exposa el model de regressió. Diverses tècniques d'anàlisi per a casos particulars s'han omès per qüestions de simplicitat o espai, però se suggereixen algunes fonts complementàries per a ampliar els temes.

1. Mesurament

Les activitats de planejament i control pròpies de l'Administració pública requereixen observar i monitorar els recursos i els resultats de la gestió. Això implica comprendre els fenòmens d'interès mitjançant representacions numèriques que permetin mesurar-los sistemàticament. Una **mesura** és una assignació de nombres (o operacionalització) a un fenomen que ens interessa analitzar.

El **mesurament** és la base de l'anàlisi estadística quan permet transformar conceptes com ara la salut pública, l'assoliment educatiu o l'eficiència recaptatòria en quantitats específiques que es poden avaluar mitjançant la comparació.

1.1. Tipus de variables

Una **variable** és una característica o condició que pot canviar o assumir diferents valors. Un **valor** és una quantitat específica que és possible per a una variable. Quan observem un fenomen d'interès ens centrem en una característica que pot variar de valor entre els subjectes o en el temps. Començarem definint tres tipus de variables: quantitatives, categòriques i ordinals.

Les **variables quantitatives** prenen valors numèrics. Hi ha aspectes que es poden mesurar a partir de quantitats observables, com ara els pressupostos o les despeses, els quals es representen en quantitats de diners. Les variables quantitatives poden tenir **valors discrets** o **continus**. En el cas dels discrets, només poden prendre valors numèrics específics (per exemple, edat, diners, nombre de persones en un programa, etc.). En el cas dels valors continus, la variable pot prendre qualsevol valor numèric (per exemple, l'altura, el pes, el percentatge d'execució d'un projecte, etc.).

Les **variables categòriques** s'utilitzen per a descriure fenòmens d'interès que no es poden mesurar en quantitats numèriques o en el cas que ens interessi només comprendre'ls en termes de categories. Per exemple, per a conèixer les característiques de la població sol ser interessant mesurar la composició de grups d'acord amb atributs particulars, com ara el gènere, la raça o la religió. Malgrat que aquestes variables no tenen escales de mesurament quantitatives, és possible mesurar-les en termes de categories; per exemple: femení *versus* masculí; blanc *versus* no blanc; catòlics, protestants, jueus, musulmans i altres. El gestor públic necessita saber com mesurar, descriure i analitzar aquests grups estadísticament quan determinen quantitats d'interès, per exemple l'efecte di-

ferencial de les polítiques sobre grups poblacionals. Aquestes variables també es coneixen com a **variables qualitatives**. Per a definir les categories s'utilitzen diferents criteris:

- Criteri únic, com a usuari *versus* no usuari.
- Més d'un criteri. Taxonomies o tipologies, és a dir, conjunts de categories que poden ser mútuament excloents (per exemple, sota la línia de pobresa, en risc de pobresa, fora de risc) o col·lectivament exhaustives (per exemple, estudia, estudia i treballa, treballa, jubilada).

Les **variables ordinals** representen categories que estan ordenades. En serien un exemple els graus d'educació (EGB, ESO, educació secundària no obligatòria, formació professional o cicle formatiu, diplomatura universitària o primer cicle, llicenciatura o grau universitari, màster, postgrau o doctorat). El mesurament ordinal permet classificar objectes o esdeveniments per categories i ordenar-los per graus. Es pot associar un nombre a cada cas, i aquest nombre no solament indica la categoria a la qual pertany, sinó com es relaciona amb les altres categories.

1.2. Dades

Un element central de l'anàlisi estadística és la **gestió de la informació**. Com hem d'emmagatzemar les nostres observacions per a poder analitzar la informació? Quina és la millor manera de registrar les variables que hem mesurat?

> El més important és tenir clar què mesurem i d'on hem pres aquesta dada. Per a això, hem de tenir molt clar quina és la nostra **unitat d'anàlisi**.

Podem fer observacions d'individus, parelles, famílies, grups, organitzacions, etc., i de cadascuna d'aquestes unitats d'anàlisi en podem mesurar diferents atributs (variables). La manera més senzilla d'organitzar les nostres observacions és mitjançant la construcció d'una **taula** o **matriu de dades**, que és un conjunt organitzat de dades en una estructura tabular (similar a un full de càlcul) i constitueix l'element fonamental d'una base de dades. Consisteix a organitzar la informació en una quadrícula en la qual les observacions o registres són cadascuna de les files en les quals es divideix la taula, i les variables o camps són cadascuna de les seves columnes. Amb aquesta organització, és possible registrar cada valor d'una variable per a una observació en cada cel·la o intersecció entre fila i columna.

És important tenir en compte que cada observació sempre ha de tenir un **identificador únic** (un nombre o codi que representi cada individu, parella, família, organització o unitat d'anàlisi utilitzada). Això és indispensable per a

identificar-lo correctament, atès que pot estar registrat de múltiples maneres (per exemple, amb els noms escrits d'una manera diferent –majúscules, segon nom, abreujaments, etc.– en la mateixa o en diferents taules.

A la taula 1 es mostra un fragment de l'estudi del CIS «3242. Macrobarómetro de març de 2019». (http://cis.es/cis/opencm/es/1_encuestas/estudios/ver.jsp?estudio=1444). Es tracta d'una enquesta feta a 16.194 persones que representen la població amb dret a vot a les eleccions generals i que resideix a Espanya. Podem observar l'estructura de matriu de dades en la qual cada fila correspon a un individu entrevistat i cada columna a una variable. La primera columna és una variable que identifica cada individu amb un número únic de qüestionari. Les següents columnes contenen informació d'interès sobre els atributs de l'individu, la seva percepció sobre el principal problema a Espanya i sobre la seva intenció de vot a les eleccions generals del 2019. No obstant això, podem veure que aquesta informació necessita una interpretació per a conèixer els noms de les variables i les etiquetes de cada categoria. Per a això, és necessari disposar d'informació addicional que ens permeti interpretar les dades i transformar-les per a la seva anàlisi. Més específicament, necessitem conèixer els noms, els tipus i les definicions de les variables, les unitats amb les quals es mesuren o les escales en les quals es classifiquen i els noms o les etiquetes per a cada categoria. Aquesta informació es guarda en un document addicional de **metadades** (dades sobre les dades) que conté informació sobre les variables i les preguntes o fonts utilitzades. Utilizant aquest document (http://cis.es/cis/export/sites/default/-archivos/marginales/3240_3259/3242/cues3242.pdf podem processar la informació de la taula 1 i obtenir les dades tal com es mostren en la taula 2.

Taula 1. Taula de dades sense processar.

CUES	P23	P22	P25	Estudis	P601	P10
1	81	1	2	6	13	94
2	74	2	3	2	1	1
3	43	2	1	6	1	97
4	58	2	1	5	19	2
5	65	2	2	5	1	94
6	57	1	1	6	1	2
7	47	1	1	6	8	12
8	64	2	2	6	8	1
9	68	1	2	2	1	11
10	53	2	4	4	1	11

Fragment de l'estudi 3242 del CIS *Macrobarómetro*.

A la taula 2 podem veure clarament quines variables tenim a la nostra disposició i els valors que pren cada variable per a cadascun dels deu individus del fragment de l'estudi. Hem d'observar que el fet que les categories es representin mitjançant valors numèrics o mitjançant les seves etiquetes o noms no canvia el tipus de variable. L'ús de les etiquetes facilita l'anàlisi, mentre que l'ús de codis numèrics s'utilitza per a optimitzar la mida dels arxius de dades. Els programes estadístics tracten de diferents maneres les dades emmagatzemades, siguin codis o etiquetes o els dos tipus d'informació.

Taula 2. Taula de dades processades.

Id	Edat	Sexe	Situació_laboral	Estudis	Principal_problema	Intenció_vot
1	81	Home	Jubilat/da o pensionista (anteriorment ha treballat)	Superiors	Els/les polítics/ques en general, els partits i la política	Encara no ho té decidit
2	74	Dona	Pensionista (anteriorment no ha treballat, mestressa de casa, etc.	Primària	L'atur	PP
3	43	Dona	Treballa	Superiors	L'atur	No votarà
4	58	Dona	Treballa	F.P.	La violència contra la dona	PSOE
5	65	Dona	Jubilat/da o pensionista (anteriorment ha treballat)	F.P.	L'atur	Encara no ho té decidit
6	57	Home	Treballa	Superiors	L'atur	PSOE
7	47	Home	Treballa	Superiors	Els problemes d'índole econòmica	EH Bildu
8	64	Dona	Jubilat/da o pensionista (anteriorment ha treballat)	Superiors	Els problemes d'índole econòmica	PP
9	68	Home	Jubilat/da o pensionista (anteriorment ha treballat)	Primària	L'atur	EAJ-PNB
10	53	Dona	Aturat/da i ha treballat abans	Secundària 2.a etapa	L'atur	EAJ-PNB

Fragment de l'estudi 3242 del CIS *Macrobarómetro*.

L'Enquesta contínua de llars de l'Institut Nacional d'Estadística (INE) proporciona un exemple de document de metodologia amb tota la informació necessària per a interpretar les dades de les enquestes. Utilitzarem aquesta enquesta en el següent apartat, per la qual cosa és recomanable descarregar les dades i els documents de l'estudi del 2018 a la pàgina de l'INE (http://www.ine.es/dyngs/inebase/es/operacion.htm?c=estadistica_c&cid=1254736176952&menu=resultats&secc=1254736195203&idp=125473557298). El fitxer de dades en format de valors separats per comes (.csv) es pot importar des de fulls de càlcul de Microsoft Excel, LibreOffice Calc o Google.

2. Anàlisi descriptiva univariant

2.1. La descripció estadística i la inferència

L'estadística **descriptiva** és un conjunt de mètodes d'organització i resum de les dades. La seva principal funció és descriure les propietats agregades d'un conjunt de dades.

D'aquesta manera, és possible comprendre un fenomen a partir de l'agregació d'observacions individuals i resumir adequadament les diverses característiques d'aquest conjunt. En aquest punt, és crucial notar la diferència entre la població que es vol estudiar i el conjunt d'observacions que és possible fer sobre una part d'aquesta població. Un valor descriptiu per a una població es denomina **paràmetre**, i un valor descriptiu per a una mostra és una **estadística**.

Les **estadístiques inferencials** són mètodes per a utilitzar la mostra amb la finalitat de treure conclusions generals (inferències) sobre la població. L'objectiu és utilitzar estadístiques de la mostra per a fer inferències sobre paràmetres poblacionals.

Atès que una mostra és típicament una part de la població, les dades de la mostra proporcionen informació limitada sobre aquesta població. Per aquesta raó, les estadístiques de la mostra són generalment representacions imperfectes dels paràmetres poblacionals corresponents. Per exemple, un paràmetre és el percentatge d'adults que viuen en parella sense fills a Espanya, i l'estadística és el percentatge de 100.000 adults que viuen en parella sense fills de l'Enquesta contínua de llars de l'INE.

Les diferències que existeixen, per casualitat, entre la mostra estadística i el paràmetre de població es coneixen com a **error de mostratge**. L'error de mostratge d'una estadística és igual a la diferència que existeix quan usem una estadística de mostra per a predir el valor d'un paràmetre de població. Definir i mesurar l'error de mostratge és una part important de l'estadística inferencial.

La diferència entre l'estadística inferencial i l'estadística descriptiva és l'ús d'un **model de probabilitat**. Un model és una descripció (matemàtica) de les connexions entre les variables d'interès. Es tracta d'una simplificació de la realitat,

i per tant mai és «correcte» o «fals», però pot ser més o menys útil. La inferència estadística i algunes nocions bàsiques sobre la teoria de la probabilitat es presenten en els següents apartats.

L'estadística descriptiva ens permet resumir les dades amb taules i gràfics (tant per a variables quantitatives com categòriques). Les **descripcions numèriques** informen sobre les tendències, la variabilitat i la posició (per a variables quantitatives). Les **descripcions bivariants** ens informen sobre com es relacionen dues variables (per a variables quantitatives o categòriques).

En les següents seccions veurem com reconèixer diferents tipus de dades, produir una varietat de mesures estadístiques, interpretar aquestes mesures estadístiques bàsiques i saber quines mesures són adequades per a quins tipus de dades.

2.2. Distribucions de freqüència

Després de recopilar dades, la primera tasca per a un investigador és organitzar i simplificar aquestes dades perquè sigui possible obtenir una visió general dels resultats. Aquest és l'objectiu de les tècniques estadístiques descriptives. El primer mètode per a simplificar i organitzar les dades és construir una distribució de freqüència.

Una **distribució de freqüència** és una tabulació organitzada que mostra exactament quants individus estan situats en cada categoria en l'escala de mesurament. Presenta una imatge organitzada de tot el conjunt de valors i mostra on es troba cada individu en relació amb uns altres en la distribució.

Una taula de distribució de freqüència consta d'almenys dues columnes: una que enumera les categories en l'escala de mesurament (X) i una altra per a la freqüència (f). A la columna X, els valors s'enumeren de major a menor, sense ometre'n cap. Per a la columna de freqüència, es determinen els comptes per a cada valor (amb quina freqüència ocorre cada valor de X en el conjunt de dades). Aquests comptes són les freqüències per a cada valor de X. La suma de les freqüències ha de ser igual al nombre d'observacions (N).

La taula 3 mostra vint observacions de l'«Encuesta continua de hogares» preses a l'atzar. Si ens interessa investigar com afecta el nombre d'habitacions de l'habitatge algun resultat d'interès de política pública (per exemple, el rendiment educatiu), hem de començar descrivint quantes habitacions tenen els habitatges a la nostra població d'interès. En concret, com varia el nombre d'habitacions de l'habitatge.

Taula 3. Fragment de l'Enquesta contínua de llars 2018 (INE).

id_hab	ca	idq_pv	tipushab	metroshab	reghab	habhab	densitathab
073443	05	35	1	0180	1	5	60
059321	13	28	5	0138	2	5	46
024757	10	46	5	0090	2	6	22,5
065139	19	52	5	0092	2	5	92
068892	01	18	2	0100	1	6	100
090857	11	10	5	0118	1	6	59
032499	13	28	5	0080	2	3	40
025713	02	50	5	0065	1	4	65
052257	02	50	5	0074	1	4	24,7
030525	01	18	5	0086	2	5	43
024614	10	46	4	0099	3	5	16,5
055340	11	10	5	0100	4	5	33,3
021115	14	30	2	0070	1	5	35
079361	01	11	5	0058	2	5	29
051262	10	46	5	0068	1	5	13,6
072956	03	33	1	0090	1	4	90
000874	07	05	5	0100	1	6	100
064281	16	48	1	0182	1	9	60,7
082938	13	28	5	0138	1	6	69
051712	16	48	5	0069	2	4	34,5

Font: http://www.ine.es

A més de les dues columnes per a les categories a l'escala de mesurament i per a la freqüència (el nombre d'observacions en cada categoria) que acabem d'explicar, una taula de distribució de freqüència també pot incloure una tercera columna que indiqui el percentatge acumulat. A la columna de les categories, els valors poden aparèixer ordenats (en el cas de variables ordinals) o seguir un ordre alfabètic o d'interès substantiu, sense ometre'n cap. Per a la columna de freqüència, es determinen els comptes per a cada valor (amb quina freqüència ocorre cada valor de la variable d'interès en el conjunt de dades); aquests comptes són les freqüències per a cada valor de la variable d'interès. La suma de les freqüències ha de ser igual al nombre d'observacions.

La taula 4 mostra la distribució de freqüències de la variable «nombre d'habitacions» (Habvi).

Taula 4. Distribució de freqüències del nombre d'habitacions.

habvi	Freqüència	Percentatge	Percentatge acumulat
1	98	0,1	0,1
2	1.009	1	1,1
3	3.336	3,32	4,42
4	14.296	14,22	18,64
5	41.455	41,23	59,87
6	23.733	23,61	83,47
7	9.566	9,51	92,99
8	4.039	4,02	97,01
9	1.857	1,85	98,85
10	782	0,78	99,63
11	202	0,2	99,83
12	80	0,08	99,91
13	50	0,05	99,96
14	17	0,02	99,98
15	10	0,01	99,99
16	6	0,01	99,99
17	1	0	100
18	2	0	100
19	1	0	100
20	1	0	100
22	1	0	100
Total	100.542	100	

Enquesta contínua de llars 2018 (INE).

Quan una taula de distribució de freqüència enumera totes les categories individuals es diu una **distribució de freqüència regular**. No obstant això, en certes ocasions, un conjunt de categories cobreix una àmplia gamma de valors. En aquestes situacions, una llista de tots els valors de la variable d'interès seria força llarga, massa llarga per a ser una presentació «simple» de les dades. Per a corregir aquesta situació, s'utilitza una taula de **distribució de freqüència agrupada**. Atès que hi ha poques llars amb habitatges amb més de deu habitacions (menys del 2 %, tal com es pot observar a la columna de percentatge acumulat), aquestes es van agrupar en una categoria de 10 o més habitacions (taula 5).

Taula 5. Distribució de freqüències agrupada del nombre d'habitacions.

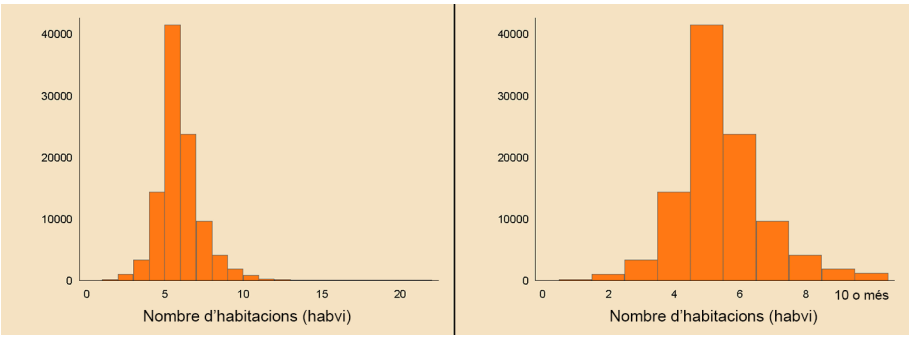
habvi	Freqüència	Percentatge	Percentatge acumulat
1	98	0,1	0,1
2	1,009	1	1,1
3	3.336	3,32	4,42
4	14.296	14,22	18,64
5	41.455	41,23	59,87
6	23.733	23,61	83,47
7	9.566	9,51	92,99
8	4.039	4,02	97,01
9	1.857	1,85	98,85
10 o més	1.153	1,15	100
Total	100.542	100	

Enquesta contínua de llars 2018 (INE).

Una manera alternativa de representar la distribució de freqüències és mitjançant un gràfic. La convenció és representar els valors de la variable d'interès en l'eix *X* i les freqüències en l'eix *Y*. En cas que estiguem interessats a representar una variable categòrica, és possible utilitzar un diagrama de barres.

En un **histograma**, una barra està centrada sobre cada categoria (o interval de classe), de manera que l'altura de la barra correspon a la freqüència i l'ample s'estén fins als límits reals, de manera que les barres adjacents es toquen. La figura 1 mostra els histogrames per a les distribucions de freqüències i freqüències agrupades.

Gràfic 1. Histogrames del nombre d'habitacions.



Quan les categories es mesuren en una escala nominal o ordinal, la representació gràfica de la distribució de freqüències ha de ser un gràfic de barres. Un **gràfic de barres** és com un histograma, amb la diferència que es deixen espais entre les barres adjacents. La principal diferència entre un histograma i un diagrama de barres és que un histograma només es fa servir per a traçar la freqüència d'ocurrències d'un valor en un conjunt de dades contínues que s'ha

dividit en classes, anomenades *intervals*. Els gràfics de barres, d'altra banda, es poden emprar per a molts altres tipus de variables, inclosos els conjunts de dades ordinals i nominals.

Els gràfics de distribució de freqüència són útils perquè mostren el conjunt complet de valors. D'un cop d'ull, és possible determinar els valors més alts, els valors més baixos i on se centren els valors. El gràfic també mostra si els valors estan agrupats o dispersos en un rang ampli. Per aquesta raó, ens interessa la forma de la distribució. La forma de la distribució ens permet veure, per exemple, en quina mesura hi ha simetria pel que fa a un eix vertical. Una distribució és simètrica si el costat esquerre del gràfic és (aproximadament) una imatge mirall del costat dret. Un exemple d'una distribució simètrica és la distribució normal, en forma de campana. En canvi, les distribucions estan esbiaixades quan les puntuacions s'acumulen a un costat de la distribució i deixen una «cua» d'uns pocs valors extrems a l'altre costat. Analitzar la forma de la distribució ens indica en quina mesura la nostra població està esbiaixada cap als extrems.

En el nostre cas, podem veure que si no agrupem les dades (quadre esquerre del gràfic 1) tenim una cua cap a la dreta amb molt pocs habitatges que tenen entre 10 i 22 habitacions. D'això se'n diu una distribució esbiaixada positivament, atès que els valors tendeixen a acumular-se en el costat esquerre de la distribució amb la cua disminuint cap a la dreta. En cas que ens interessi agrupar les llars amb més de 10 habitacions, podem veure que la distribució és simètrica (quadre de la dreta del gràfic 1).

2.3. Mesures de centralitat

Les estadístiques descriptives més comunes són les mesures de tendència central.

> Tal com el seu nom indica, les **mesures de tendència central** intenten situar el punt central o mitjà en un grup de dades. En termes generals, la tendència central és una mesura estadística que determina un valor únic que descriu amb precisió el centre de la distribució i representa la distribució completa dels valors. L'objectiu de la tendència central és identificar el valor únic que sigui el millor representant per a tot el conjunt de dades.

En identificar el «**valor mitjà**», la tendència central permet als investigadors resumir o condensar un gran conjunt de dades en un sol valor. Per tant, la tendència central serveix com una estadística descriptiva perquè permet als investigadors descriure o presentar un conjunt de dades de manera concisa i

molt simplificada. A més, és possible comparar dos (o més) conjunts de dades simplement comparant el valor mitjà (tendència central) per a un conjunt enfront del valor mitjà per a un altre conjunt.

És essencial que la tendència central estigui determinada per un procediment objectiu i ben definit per a poder expressar exactament com s'ha obtingut el valor mitjà i poder duplicar el procés. Cap procediment únic produeix sempre un valor representatiu. Hi ha tres tècniques comunament utilitzades per a mesurar la tendència central:

- la **mitjana**, que és la mitjana aritmètica de les observacions;
- la **mediana**, que és l'observació que cau exactament en el centre del grup, i
- la **moda**, que és el valor que es produeix amb més freqüència.

Aquestes mesures de resum es denominen *estadístiques*, entenent per *estadística* qualsevol quantitat numèrica el valor de la qual està determinat per les dades.

La **mitjana** és la mesura de tendència central més utilitzada. El càlcul de la mitjana es fa sobre variables quantitatives (amb valors numèrics o d'intervals). La mitjana s'obté calculant la suma, o el total, per al conjunt complet de dades i després dividint aquesta suma pel nombre d'observacions. Suposem que tenim una mostra de *n* observacions els valors de les quals designem *x1*, *x2*, *...xn*. El valor mitjà de la mostra és:

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

Conceptualment, la mitjana també es pot definir com la quantitat que rep cada individu quan el total es divideix igualment entre tots; o bé com el punt d'equilibri de la distribució, perquè la suma de les distàncies per sota de la mitjana és exactament igual que la suma de les distàncies per sobre de la mitjana.

Per a il·lustrar el càlcul de la mitjana, la taula 6 mostra els Pressupostos Generals de l'Estat per a 2018. Per a conèixer el valor mitjà del pressupost per a cadascuna de les polítiques, hem de sumar el pressupost total i dividir-lo pel nombre de polítiques. És a dir, 368,323 milers de milions d'euros dividits en 27 polítiques, que donen com resultat una mitjana de 13,642.59 euros per partida.

Taula 6. Pressupostos Generals de l'Estat per a 2018.

Polítiques	Milions d'euros
1. Accés a l'habitatge i foment de l'edificació.	481
2. Altres actuacions de caràcter econòmic	639
3. Recerca militar	679

Font: Estadístiques dels Pressupostos Generals de l'Estat (www.sepg.pap.hacienda.gob.es).

Polítiques	Milions d'euros
4. Òrgans constitucionals, Govern i altres	681
5. Cultura	887
6. Comerç, turisme i pimes.	896
7. Administració financera i tributària	1.390
8. Política exterior	1.581
9. Justícia	1.780
10. Subvencions al transport	2.139
11. Educació	2.541
12. Serveis socials i promoció social	2.631
13. Sanitat	4.251
14. Infraestructures	5.437
15. Foment de l'ocupació	5.716
16. Indústria i energia	5.771
17. Recerca civil	6.379
18. Agricultura, Pesca i Alimentació	7.707
19. Defensa	8.401
20. Seguretat ciutadana i institucions penitenciàries	8.418
21. Altres prestacions econòmiques	14.385
22. Gestió i administració de la seguretat social	17.297
23. Desocupació	17.702
24. Serveis de caràcter general	24.643
25. Deute públic	31.547
26. Transferències a altres administracions públiques	49.510
27. Pensions	144.834

Font: Estadístiques dels Pressupostos Generals de l'Estat (www.sepg.pap.hacienda.gob.es).

No obstant això, veiem que la política de «Pensions» és molt superior a la resta i que les d'«Accés a l'habitatge», «Recerca militar», «Altres actuacions de caràcter econòmic» i «Òrgans constitucionals, govern i altres» són molt inferiors. Aquests valors extrems, de vegades denominats *valors atípics*, tenen una influència desproporcionada sobre la mitjana i, per tant, poden afectar la forma com la mitjana representa les dades. En aquests casos, la mitjana es desplaçarà cap als extrems (es desplaçarà cap a la cua) i no proporcionarà un valor «central». Això implica que la mitjana no sempre funciona com una mesura de tendència central i és necessari disposar de procediments alternatius disponibles.

La segona mesura de tendència central és la mediana. La **mediana** es defineix com el punt del mig de la llista quan els valors en una distribució s'enumeren en ordre de menor a major. Aquesta divideix les puntuacions, de manera que el 50% dels valors en la distribució tinguin valors iguals o menors a la mediana. El càlcul de la mediana requereix valors que es puguin ordenar (de menor a major) i que es mesurin en una escala ordinal, d'interval o de relació. La mediana del pressupost de 2018 seria, doncs, de 5.437 euros. El valor de la política que es troba en el centre de la distribució, la número 14, és «Infraestructures» —que es pot calcular com a (27+1)/2). En aquest cas es té un nombre imparell de polítiques i el càlcul és senzill. Amb un nombre parell de valors, han d'ordenar-se els valors i la mediana és a mig camí entre les dues puntuacions mitjanes (és a dir, la mitjana).

Un avantatge de la mediana és que no es veu afectada per valors extrems. Per tant, la mediana tendeix a romandre en el «centre» de la distribució, fins i tot quan hi ha alguns valors extrems o quan la distribució és molt esbiaixada. En aquestes situacions, la mediana serveix com una bona alternativa a la mitjana.

La tercera mesura de tendència central és la moda. La **moda** es defineix com la categoria o valor més freqüent en la distribució. En un gràfic de distribució de freqüències, la moda és la categoria o puntuació corresponent al punt més alt o màxim de la distribució.

Si volguéssim descriure la distribució del nombre d'habitacions per llar que hem vist a l'apartat anterior, seria possible calcular la moda per a les dades de la taula 4. Podem veure fàcilment que el nombre d'habitacions amb més freqüència és cinc. En aquest cas, també podríem calcular la mitjana. Trobem que el nombre mitjà d'habitacions per llar a la nostra mostra és de 5,4. La moda s'usa habitualment com una mesura complementària de tendència central que acompanya la mitjana o la mediana.

D'altra banda, si volem descriure les dades agregades en categories (taula 5), veiem que ja ens és possible calcular la mitjana perquè no tenim valors numèrics, sinó una categoria de «10 o més». No obstant això, veiem que en aquestes dades la categoria que es repeteix amb més freqüència són les llars de cinc habitacions. El principal avantatge de la moda és que és l'única mesura de tendència central que es pot utilitzar per a dades mesurades en una escala categòrica.

També, però, hem de considerar que és possible que una distribució tingui més d'una moda. En el cas que hi hagués moltes llars petites (de tres habitacions), però també moltes llars de sis, tindríem una distribució amb dos valors que es repeteixen amb la major freqüència. Aquesta distribució es diu *bimodal*. És important tenir en compte que una distribució pot tenir una sola mitjana i una sola mediana, però dues o més modes. El terme *moda* s'usa sovint per a

descriure un pic en una distribució que no és realment el punt més alt. Per tant, una distribució pot tenir una moda major en el pic més alt i una moda menor en un pic secundari en una ubicació diferent.

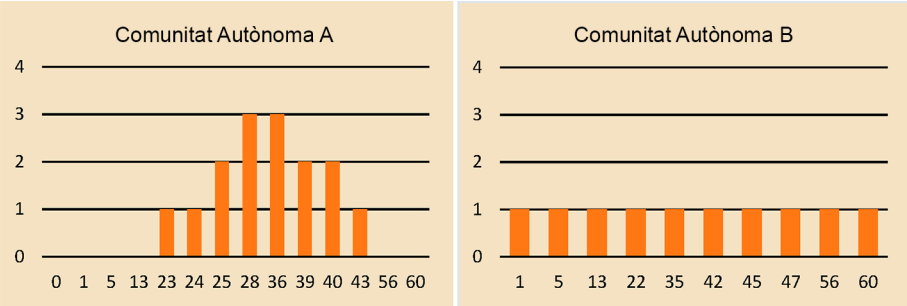
2.4. Mesures de dispersió

Les **mesures de dispersió** (també conegudes com a **variació**) ens informen sobre el grau en el qual les dades s'agrupen sobre la mitjana. També es pot entendre com un indicador de com es distribueixen els valors en una distribució.

La dispersió serveix com una mesura descriptiva i com un component important de la majoria de les estadístiques inferencials. Com a estadística descriptiva, mesura el grau en el qual els valors es distribueixen o agrupen en una distribució. Una mesura de dispersió generalment acompanya una mesura de tendència central com a estadística descriptiva bàsica per a un conjunt de valors. La tendència central descriu el punt central de la distribució i la dispersió descriu com els valors es distribueixen al voltant d'aquest punt central.

En el context de les estadístiques inferencials, la dispersió proporciona una mesura de la precisió amb què un valor o observació individual representa tota la població. Les dues mostres del gràfic 2 tenen una mitjana de 32.6. Quan la dispersió de la població és petita, totes les puntuacions s'agrupen juntes, i una mesura de tendència central proporcionarà una bona representació de tot el conjunt (gràfic 2A). Per contra, quan la dispersió és gran i els valors estan distribuïts àmpliament, és fàcil que un o dos valors extrems donin una imatge distorsionada de la població general (gràfic 2B).

Gràfic 2. Exemples de diferents nivells de dispersió.



La dispersió es pot mesurar de diferents maneres: el rang, la desviació estàndard (o variància), el coeficient de variació, el rang interquartílic o els valors Z. En tots aquests casos, la dispersió es determina mesurant la distància. A continuació, es presenten les tres primeres maneres de mesurar.

El **rang** és la distància total coberta per la distribució, des del valor més alt fins al valor més baix (utilitzant els límits reals superior i inferior del rang). El rang és altament sensible als valors atípics, però no es veu afectat per la forma de la

distribució. En l'exemple del gràfic 2, el rang per a la comunitat autònoma A és de 20 (valor màxim-valor mínim), mentre que en la comunitat autònoma B té un valor de 59.

La **desviació estàndard** mesura la distància estàndard entre un valor i la mitjana. El càlcul de la desviació estàndard es pot resumir com un procés de sis passos:

- Calcular la mitjana.
- Calcular la desviació (distància des de la mitjana) per a cada valor.
- Elevar al quadrat cada desviació.
- Sumar les diferències al quadrat.
- Dividir la suma pel nombre d'observacions.
- Calcular l'arrel quadrada de la variància.

Aquests passos es poden resumir en la fórmula següent:

$$Si = \sqrt{\frac{\sum_{i=1}^{N} (Xi - \mu)^2}{N}}$$

Per a l'exemple del gràfic 2, la desviació estàndard és de 6.83 per a la comunitat autònoma A i de 16.35 per a la comunitat autònoma B. Els passos de càlcul es presenten en la taula 7.

Taula 7. Càlcul de la desviació estàndard per a dues mostres.

Comunitat Autònoma A					**Comunitat Autònoma B**			
Xi	μ	Xi - μ	$(Xi - \mu)^2$		Xi	μ	Xi - μ	$(Xi - \mu)^2$
22	32.6	-10.6	112.4		1	32.6	-31.6	998.6
24	32.6	-8.6	74.0		5	32.6	-27.6	761.8
25	32.6	-7.6	57.8		13	32.6	-19.6	384.2
25	32.6	-7.6	57.8		22	32.6	-10.6	112.4
28	32.6	-4.6	21.2		35	32.6	2.4	5.8
28	32.6	-4.6	21.2		42	32.6	9.4	88.4
28	32.6	-4.6	21.2		45	32.6	12.4	153.8
36	32.6	3.4	11.6		47	32.6	14.4	207.4
36	32.6	3.4	11.6		56	32.6	23.4	547.6
36	32.6	3.4	11.6		60	32.6	27.4	750.8
39	32.6	6.4	41.0					
39	32.6	6.4	41.0					
40	32.6	7.4	54.8					

Comunitat Autònoma A					Comunitat Autònoma B			
40	32.6	7.4	54.8					
43	32.6	10.4	108.2					

	CA A	CA B
Suma valor quadrat de les diferències	699.6	4010.4
Suma quadrat de les diferencies / N (variància)	46.6	267.4
Desviació estàndard (arrel quadrada de la variància)	6.83	16.35

3. Relacions entre variables categòriques

3.1. Relacions entre variables i verificació d'hipòtesi

Fins ara s'han presentat mètodes d'anàlisi estadística que s'apliquen a una sola variable. No obstant això, per a la majoria de les anàlisis és d'interès no solament descriure les variables per separat, sinó poder respondre a preguntes sobre les relacions entre les variables. Per exemple, hi ha alguna relació entre les variables o característiques? En cas que hi hagi una relació, fins a quin punt és forta? Es pot utilitzar una variable per a predir-ne una altra?

L'anàlisi de dades en dues dimensions o bivariants es duu a terme mitjançant enfocaments similars als de les dades univariants. Es fan tabulacions, representacions gràfiques i s'analitzen les característiques numèriques. El propòsit d'aquest tipus d'anàlisi és descriure fenòmens d'interès (què està passant) per a poder observar una associació i establir regularitats o patrons, o explicar la relació entre dues variables (per què passa).

En general, entre les variables que s'analitzen se sol diferenciar entre:

- **Variable dependent**: és la variable que es vol predir o explicar, en general representada per la lletra *Y*. També pot ser descrita com el resultat d'un valor conegut de la variable independent.

- **Variable independent**: la variable (o conjunt de variables) que prediu o explica la variable *Y*. En general es representa amb la lletra *X*.

La variable dependent és aleatòria, és a dir, per cada valor donat a la variable independent, hi ha molts possibles resultats per a la variable dependent.

En la mesura en què s'aconsegueix entendre què causa variabilitat en un fenomen d'interès és possible proposar o avaluar polítiques públiques amb més certesa sobre els resultats possibles. Per a això, és necessari tenir molt clar un model d'anàlisi, és a dir, establir quina és la variable d'interès que es vol explicar o predir i quina o quines són les variables que expliquen aquests canvis. Aquesta variable que es vol explicar o predir és la variable dependent. En la gestió pública, la variable dependent poden ser els resultats d'una política, els béns socials o el grau d'èxit d'un programa (per exemple, assoliment educatiu, nivell d'atur, esperança de vida, etc.). La variable que l'explica és la variable independent. També es denomina *variable explicativa*, *causal* o *exògena*. Aquesta

pot ser un atribut propi de l'objecte d'estudi, com la composició de les llars o la densitat de població, o una variable que es pot canviar o controlar, com ara el pressupost o el nivell de competència d'un actor responsable de la política.

Una vegada establertes les variables d'interès (per exemple, l'esperança de vida i l'impost sobre les begudes ensucrades) és necessari definir clarament com creiem (o com esperem) que estan relacionades. Per a això, definim una hipòtesi. La **hipòtesi** és un enunciat declaratiu que indica explícitament la relació que s'espera trobar entre les variables (per exemple, esperem que un augment en l'impost sobre les begudes ensucrades faci créixer l'esperança de vida). A partir d'una hipòtesi que pugui ser contrastada és possible concloure sobre la relació que esperem entre les variables. És a dir, podem avaluar en quin mesura l'evidència empírica disponible (les dades que hem observat) permet rebutjar o no la hipòtesi .

Lectura recomanada

Una explicació més detallada sobre el disseny de recerca i les teories i hipòtesis en les ciències socials, la podeu trobar a:

I. Crespo; E. Anduiza; M. Méndez (2009). *Metodología de la ciencia política*. Madrid: Centro de Investigaciones Sociológicas («Cuadernos Metodológicos», 28).

Aquest tipus d'anàlisi entre dues variables s'anomena **anàlisi bivariable** (o bivariant). La taula 8 resumeix les tècniques d'anàlisis apropiades per a cada tipus de variables.

Taula 8. Tècniques d'anàlisi per tipus de variable.

		Variable independent (VI)	
		Categòrica	Contínua
Variable dependent (VD)	Categòrica	Taules de contingència	(Regressió logística)
	Contínua	Diferència de mitjanes	Correlació Regressió

A continuació, es descriuen els procediments d'anàlisi de dades que són apropiats per a les relacions que involucren variables categòriques, és a dir, les taules de contingència.

3.2. Taules de contingència: cel·les, columnes, files i marginals

Una **taula de contingència** és una tabulació creuada que resumeix simultàniament dues variables d'interès. S'utilitza quan alguna o les dues variables tenen escales categòriques o ordinals. El propòsit és analitzar la variació conjunta de les dues variables. Se solen agrupar les observacions en una taula de freqüències bivariants on es presenta la distribució de freqüències corresponent a les parelles de valors o categories de les dues variables.

Per a construir la taula de contingència és necessari establir quines són les variables dependent i independent. En aquest cas volem explicar el tipus de prestació (VD) i com varia per gènere (VI). Per a això és necessari organitzar els percentatges per a les categories de la VI. D'aquesta manera és possible compa-

rar els tipus de prestació entre les categories home i dona. Convencionalment, la VI va a les columnes i la VD va a les files (percentatges per columnes), i s'indica el nombre d'observacions per a cadascuna de les categories de la VI.

La taula 9 presenta el percentatge de beneficiaris de prestacions per desocupació i tipus de prestació per gènere (SEPE, gener del 2019, total nacional).

Taula 9. Beneficiaris de prestacions per desocupació i tipus de prestació per gènere.

	Homes	**Dones**	**Marginal**
Prestació contributiva	46%	41%	43%
Subsidi per desocupació	39%	40%	39%
Renda agrària	4%	4%	4%
Subsidi agrari	4%	7%	5%
Renda activa d'inserció	7%	9%	8%
Programa d'activació per a l'ocupació	0%	0%	0%
Total	**872.946**	**1.025.423**	**1.898.369**

Font: http://www.sepe.es

En termes generals, es pot dir que hi ha una relació si els valors de la VD són diferents segons els valors de la VI, però hi ha tres interpretacions possibles de la taula. La més simple és per categories de la VI. Podem veure, per exemple, que un 46% dels homes reben una prestació contributiva. La segona interpretació es pot fer per comparació entre les categories de la VD. Es pot veure que hi ha més homes que reben la prestació contributiva que dones o que les dones reben la renda activa d'inserció en una proporció superior a la dels homes. La tercera comparació és amb el valor global de la mostra (marginal). Es pot observar, per exemple, que entre els homes hi ha un percentatge més alt que rep una prestació contributiva que en el global de la mostra (46% > 43%).

3.3. Mesures del grau d'associació entre variables

A partir de les anàlisis, és possible arribar a conclusions sobre la direcció i la força de la relació entre les variables. Per a les variables nominals n'hi ha prou amb descriure quins valors de la VI estan associats amb quins valors de la VD. En el cas de tenir variables ordinals, és possible parlar de la direcció, és a dir, de relacions positives o negatives.

D'altra banda, la força de la relació indica en quina mesura difereixen els valors de la VD en les categories de la VI. La relació és perfecta quan tots els valors d'una categoria de la VI van associats a una categoria diferent de la VD. La força de la relació no està determinada pel nivell de significació estadística.

Podem calibrar la força de la relació a partir de la diferència entre els percentatges corresponents a les diferents categories de la variable independent. Per exemple, la diferència per gènere en la prestació contributiva és de cinc punts (46-41, relació forta), mentre que no hi ha diferències entre homes i dones per a la Renda agrària ni per al Programa d'activació per a l'ocupació. Les diferències per gènere per al Subsidi per desocupació, el Subsidi agrari o la Renda activa d'inserció són més febles perquè només assoleixen els tres punts percentuals.

Addicionalment, hem de saber fins a quin punt la diferència entre les freqüències observades i les esperades és prou gran per a dir que existeix una relació amb la població.

3.4. La prova de significació khi quadrat

La hipòtesi que hem plantejat és que hi ha diferències entre gèneres en les prestacions per desocupació. El següent pas és avaluar l'evidència empírica per a avaluar si les diferències observables són prou grans per a dir que existeix una relació amb la població. Si es compleix aquesta condició, és possible afirmar que el resultat és estadísticament significatiu. Això vol dir que no és probable que s'hagi produït a l'atzar.

Per a això s'utilitzarà la prova de significació khi quadrat:

$$\chi^2 = \sum \frac{(Fo - Fi)^2}{Fi}$$

on *Fo* es refereix a la freqüència observada i *Fe* a la freqüència esperada. La freqüència esperada es calcula així:

$$Fe=[(\text{marginal fila/total})*(\text{marginal columna/total})]*\text{total} = (\text{mar ginal fila}) * (\text{mar ginal columna}) / \text{total}$$

La taula 10 és una taula de contingència amb dues categories de la VD i dues de la VI, i els passos per a calcular els valors de khi quadrat.

Taula 10. Càlcul de khi quadrat.

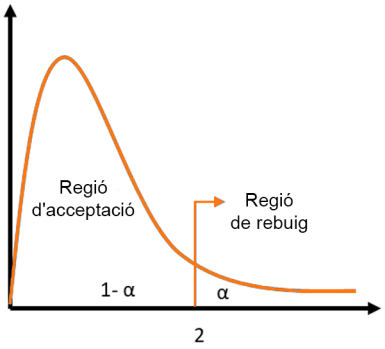
	X1	X2	Marginal columna			Fo	Fe	Fo - Fe	$(Fo - Fe)^2$	$(Fo - Fe)^2/Fe$
Y1	5	15	20		Y1 X1	15	10	5	25	2,5
Y2	15	5	20		Y1 X2	5	10	-5	25	2,5
Marginal fila	20	20	40		Y2 X1	15	10	5	25	2,5
					Y2 X2	5	10	-5	25	2,5
									Suma	**10**

Una vegada que sabem que l'estadístic khi quadrat (χ^2) té un valor de 10, és necessari calcular els graus de llibertat a partir dels atributs de les variables (nombre de files i nombre de columnes en la taula creuada).

Graus llibertad = (nombre de files - 1)*(nombre de columnes - 1)

En l'exemple de la taula 10, tenim dues files i dues columnes, de manera que els graus de llibertat serien d'1. Amb aquesta informació, és possible calcular el nivell de significació (α) a partir de la distribució de khi quadrat (o distribució de Pearson). Aquesta és una distribució de probabilitat contínua amb un paràmetre que representa els graus de llibertat de la variable aleatòria. Amb la informació disponible a la taula de distribució de probabilitat khi quadrat, és possible concloure que la probabilitat d'obtenir un valor 10 o superior amb 1 grau de llibertat és menor que p = 0.005. Per tant, el valor és estadísticament significatiu, perquè és menor que 0.01.

Gràfic 3. Distribució de probabilitat khi quadrat i taula inversa.



Graus llibertat	0.1	0.05	0.025	0.01	0.005
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75

El nivell de significació indica el nivell de confiança que volem que tinguin els càlculs de la prova; és a dir, si volem tenir un nivell de confiança del 95%, el valor de significació (α) ha de ser del 0,05. Són comuns els nivells de significació del 0,05, 0,01 i 0,1. Habitualment, no és suficient dir que una relació és significativa, sinó que s'especifica en quin nivell ho és. A les taules de resultats, se sol indicar mitjançant asteriscos: * p <0,05, ** p <0,01 i *** p <0,001. ***

És necessari tenir en compte els següents punts per a interpretar correctament el nivell de significació:

- No és reversible: p = 0,01 no vol dir que hi ha un 99% de probabilitat que existeixi una relació.
- Està basat en el cas que el mostratge és aleatori.
- No diu res sobre altres fonts d'error possibles (per exemple, biaixos en la mostra o errors de mesura).
- No implica que el resultat sigui important en sentit pràctic (significació substantiva).
- No indica la força de la relació entre les variables.
- Com més gran és la mostra, més fàcil és trobar relacions estadísticament significatives.

- No indica que la relació sigui causal.

4. Relacions entre variables contínues

4.1. Diferències de mitjanes i prova *T*

Un repte comú a l'Administració pública és poder arribar a conclusions sobre les diferències entre dues mostres per a analitzar els efectes de les mesures o intervencions polítiques. La comparació implica concloure si els valors mesurats per a una mostra són diferents dels de l'altra mostra en mitjana. Per exemple, un tècnic ambiental necessitarà saber si els nivells de contaminació de l'aire d'un municipi van millorar després d'haver regulat la circulació de vehicles antics. Per a això, serà necessari fer una comparació entre abans i després de la regulació, o entre dos municipis semblants que només difereixen en l'existència d'aquesta regulació.

En situacions com aquestes, en les quals es necessita saber si dues mostres són diferents en les seves mitjanes o proporcions de mostra, la tècnica apropiada és una **prova de diferència de mitjanes**, que permet fer anàlisis per a variables d'interès contínues entre grups (variable categòrica).

Quan es busca comprovar si existeixen diferències entre dues mitjanes mostrals, l'objectiu és determinar si ambdues mitjanes es podrien haver extret de la mateixa població, o si les dues són tan diferents que no es podrien haver extret de la mateixa població. Per a molts problemes de gestió o avaluació de polítiques, s'espera que els valors per a una mostra siguin diferents dels de l'altra.

Igual que amb un altre tipus d'anàlisi, la pregunta de recerca ha de ser molt clara. És necessari definir la raó per la qual s'espera una diferència entre els dos grups (què fa que un grup sigui diferent d'un altre?) i la direcció esperada de la diferència.

Per exemple, si comparem l'acompliment de dos emprenedors que han rebut fons públics per al desenvolupament empresarial, la hipòtesi ha de reflectir l'expectativa que l'acompliment dels emprenedors que han rebut les ajudes serà més gran que el d'aquells que no les han rebudes. La lògica general és poder confirmar que els grups difereixen i en quina mesura la inversió en programes o canvis institucionals està relacionada amb una diferència en els resultats. Se segueix la lògica d'un experiment en el qual el grup que rep el finançament s'assembla a un grup que rep un tractament i el que no el rep s'usa com un grup de control que serveix per a conèixer l'efecte del tractament. Aquesta lògica implica que els grups són idèntics i que la selecció d'estar en els grups de tractament o control és aleatòria. El procediment d'anàlisi consisteix llavors a definir la variable d'interès per als dos grups.

En aquest cas prendrem la utilitat neta dels emprenedors en els cinc primers anys de funcionament. La hipòtesi que volem rebutjar és que les mitjanes d'utilitat neta són semblants per als emprenedors que van rebre el finançament públic i per als que no el van rebre. Per a això seguirem cinc passos:

1) Calcular les mitjanes i desviacions estàndard de la VD per als dos grups.

Grup	Mitjana	Desviació estàndard	Nombre d'emprenedors
Emprenedors amb finançament públic (tractament)	52.1	45.1	22
Emprenedors sense finançament públic (control)	27.1	26.4	22

2) Calcular el paràmetre estimat (estadístic), és a dir, la diferència entre tots dos grups.

Estimació mitjana = mitjana control - mitjana tractament = 27.1 – 52.1 = -25

Es pot observar que els emprenedors que han rebut finançament públic tenen una utilitat neta en cinc anys superior als qui no l'han rebut. No obstant això, no es pot concloure que aquesta diferència sigui estadísticament significativa, atès que cada grup té la seva pròpia variabilitat.

3) Calcular un error estàndard general o «agrupat» per a tots dos grups.

Error estàndard (mitjana control - mitjana tractament) =

$$\sqrt{\frac{S_{iControl}}{N_{Control}} + \frac{S_{iTrat\ am\ iinto}}{N_{Trat\ am\ iinto}}} = \sqrt{\frac{26.4^2}{22} + \frac{45.1^2}{22}} = 11.14$$

A partir d'aquí, volem conèixer la probabilitat que els grups hagin estat extrets de la mateixa població. Per a això és necessari recórrer a mètodes paramètrics com el basat en la distribució T-Student. La distribució T-Student és similar a la distribució normal. Té com a paràmetres la mitjana i la variància, i depèn de la mida de la mostra a través dels graus de llibertat.

4) Calcular l'estadístic *t* usant l'estimació mitjana i l'error estàndard:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{SE} = \frac{0 - 25}{11.14} = 2.24$$

5) Buscar la puntuació *t* de -2.24 a la taula *T* per a dues cues, per a 21 graus de llibertat (22 observacions menys un).

Web recomanada

Hi ha eines en línia que consulten els valors de les taules de diferents estadístics. Un exemple el teniu a https://www.uv.es/ceaces/scrips/tablas/tastud.htm

La probabilitat que la diferència entre els dos grups sigui igual a 0 és de 0.036. Això implica que l'efecte del finançament públic dels emprenedors té un efecte estadísticament significatiu sobre la seva utilitat neta en els cinc primers anys.

4.2. Tipus de diferències de mitjanes

El mètode presentat a l'apartat anterior per a una prova de diferència de mitjanes és només una de les tres proves d'aquest tipus. Es pot aplicar quan les dues mostres són independents i s'assumeix que les dues variàncies de població són diferents. No obstant això, aquesta condició no es dona sempre i per a això tenim dues altres proves: una per a mostres independents amb variàncies iguals i una altra per a mostres dependents. Per a decidir quina prova cal aplicar a una pregunta de recerca en particular, és necessari comprendre la **diferència entre mostres independents i dependents**.

Les **mostres independents** són aquelles en les quals les observacions de les dues mostres no estan «aparellades» de cap manera. El millor procediment per a obtenir mostres independents és mitjançant tècniques de mostratge aleatori. Per exemple, si un analista encarregat dels programes d'ajuda a la dependència selecciona aleatòriament dues mostres d'una base de dades de la seva comunitat autònoma, cadascuna de les quals consta de 500 beneficiaris, la probabilitat que hi hagi un aparellament un a un entre els casos de la mostra és mínima. Així, el primer beneficiari de la mostra A pot ser una dona de 70 anys que viu en el municipi X i cobra l'ajuda per incapacitat, mentre que la primera observació en la mostra B pot ser un home de 87 anys, del municipi Z, que cobra l'ajuda per raons d'edat. Les altres observacions de cada mostra haurien de ser igualment diverses en termes dels atributs dels individus.

Per contra, en **les mostres dependents**, cada observació d'una mostra té un parell similar a l'altra. Una prova per a conèixer l'abans i el després d'un tractament (per exemple, un canvi legislatiu o una intervenció social) generaria mostres dependents si els mateixos individus s'observessin abans i després.

Les tècniques per a l'anàlisi de diferències de mitjanes amb mostres independents amb variàncies iguals i per a mostres dependents segueixen lògiques similars.

Lectura recomanada

Les tècniques per a l'anàlisi de diferències de mitjanes amb mostres independents amb variàncies iguals i per a mostres dependents no s'expliquen en aquesta guia, però les podeu veure detalladament a:

D. A. Lind; W. G. Marchal; S. A. Wathen (2017). *Pruebas de hipótesis de dos muestras*. A: D. A. Lind; W. G. Marchal; S. A. Wathen. *Estadística aplicada a los negocios y a la economía* . Mèxic: McGraw-Hill/Interamericana Editors, SA de CV.

4.3. L'anàlisi de correlació

El propòsit de l'anàlisi de correlació és representar l'associació entre dues variables quantitatives mitjançant un conjunt de tècniques numèriques que inclou l'anàlisi gràfica i l'ús d'indicadors. L'**anàlisi de correlació** serveix per a indagar si els canvis en els valors d'una variable estan associats amb els valors d'una altra variable. Si això ocorre, és possible concloure que ambdues variables estan correlacionades o bé que hi ha correlació entre aquestes.

Més concretament, el principal objectiu de l'anàlisi de correlació consisteix a determinar els següents elements en la relació entre les variables analitzades:

- Si existeix o no una associació entre les variables, és a dir, si la variació en els valors d'una d'aquestes està relacionada amb la variació en els valors de l'altra.

- Si la relació que existeix entre les variables és directa o inversa (o, dit d'una altra manera, positiva o negativa), és a dir, si en augmentar el valor de la variable independent el valor de la variable dependent també augmenta o, per contra, disminueix.

- Determinar la intensitat de la relació entre ambdues variables, és a dir, la magnitud de la variació que experimenta la variable dependent quan la variable independent canvia.

Per a determinar els elements que caracteritzen l'associació que existeix entre dues variables, en general, el primer pas és traçar les dades en un **diagrama de dispersió**. Aquest procediment proporciona una representació visual de la relació entre les variables. El següent pas sol ser calcular el **coeficient de correlació**, que brinda una mesura quantitativa de la força de la relació entre dues variables.

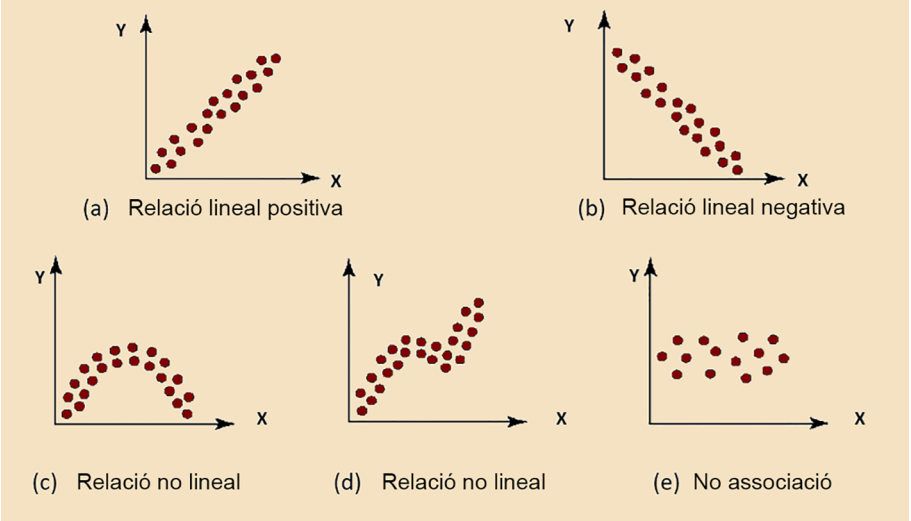
La visualització de la relació: el gràfic de dispersió

Quan es pren una mostra de dues variables per a cada observació de la població o de la mostra, s'obté una sèrie de parells de dades. Aquestes parelles tenen la forma (X,Y) i es poden representar com a punts en un plànol bidimensional o plànol cartesià; la representació gràfica de les parelles es coneix com a *diagrama de dispersió* i és una eina molt utilitzada per a conèixer la tendència de les dades abans d'aprofundir en l'estudi de la correlació i fins i tot de l'anàlisi de regressió, que s'explicarà en la següent unitat.

Cada individu en el conjunt de dades està representat per un punt en el diagrama de dispersió. És pràctica comuna situar la variable dependent en l'eix vertical o *Y*, i la variable independent en l'eix horitzontal o *X*. Per a elaborar un diagrama de dispersió simplement n'hi ha prou amb localitzar en l'eix corresponent els valors de les variables *X* i *Y* que s'observen simultàniament per a cada individu de la població o mostra.

El diagrama de dispersió resultant serveix per a avaluar visualment si les variables poden estar relacionades i quin tipus de relació hi ha entre aquestes. En el gràfic 4 es poden veure cinc exemples de relacions.

Gràfic 4. Exemples de relacions bivariants per a variables contínues.



Fixeu-vos que, segons quina sigui la dispersió de les dades (núvol de punts) en el plànol cartesià, es pot donar el cas que la relació sigui positiva (o directa) o negativa (o inversa). La relació serà positiva si quan *X* augmenta, el valor de *Y* també augmenta, i serà negativa si quan *X* augmenta, el valor de *Y* disminueix. En els casos (a) i (b) la relació que es presenta és clarament positiva i negativa, respectivament. Addicionalment, en tots dos casos la relació és lineal, és a dir, la relació de canvi entre ambdues variables és similar a una línia recta el pendent o inclinació de la qual dependrà del signe de la relació. No obstant això, en altres casos no és tan evident si la relació és positiva o negativa, com passa en les relacions no lineals. En l'exemple (c), la relació és primer positiva i després de cert valor de *X* es converteix en negativa. En el cas (d), la relació, que primer és positiva, es converteix en negativa i finalment passa a ser positiva novament segons que augmenten els valors de *X*. En tots dos casos, la relació no es podria representar amb una línia recta, per la qual cosa són no lineals —el cas (c) representaria una relació quadràtica i el (d) una relació cúbica. També es pot donar el cas que no es pugui identificar clarament un patró en els núvols de punts, la qual cosa indicaria que no existeix en realitat una associació entre les variables, com s'exemplifica en el cas (i).

Tanmateix, el diagrama solament ens pot donar una «sensació» visual sobre l'associació entre dues variables, però en realitat no mesura la força de tal associació. Per a mesurar la força de la relació entre *X* i *Y* podem calcular el coeficient de correlació.

Per a il·lustrar aquest tipus de diagrames, analitzarem, amb dades de l'Ajuntament de Barcelona, l'associació que existeix entre la població immigrant d'un barri i el nombre d'aturats en aquest mateix barri.

La immigració

La immigració és un assumpte públic important, especialment en moments en els quals els canvis demogràfics poden representar tant una amenaça com una oportunitat per al mercat laboral dels països i la sostenibilitat dels seus sistemes de pensions, i també en contextos en els quals l'arribada massiva d'immigrants pot significar una important càrrega per a les societats d'acollida. Tot i que la percepció de la població sobre la relació entre immigració i desocupació sol ser negativa, múltiples estudis demostren que els efectes negatius de la immigració sobre el mercat laboral són nuls o, en el pitjor dels casos, molt petits, i que tendeixen a desaparèixer amb el pas del temps. D'altra banda, el que s'ha demostrat, no obstant això, és que els immigrants són part de la població més afectada per les crisis econòmiques com la que va patir Espanya a partir del 2008, per la qual cosa no és una sorpresa trobar una correlació positiva entre el nombre d'immigrants i el nombre d'aturats en una població específica.

En relació amb l'escassa influència d'aquest fenomen en el mercat de treball, recomanem llegir l'entrada «De los efectos de la inmigración sobre el mercado de trabajo», publicada per Lidia Farré en el blog *Nada es gratis* (http://nadaesgratis.es/lidia-farre/de-los-efectos-de-la-inmigracion-sobre-el-mercado-de-trabajo). Pel que fa a l'afectació de l'atur en la població immigrant vegeu, per exemple, «Los inmigrantes son los más afectados por la crisis y el paro, según Ranstad», publicat a *El Economista.es* en plena crisi el 2009 (https://www.eleconomista.es/economia/noticias/1590521/10/09/los-extranjeros-son-los-mas-afectados-por-la-crisis-y-el-paro.html), o l'estudi de l'OCDE que ressenyava més recentment *El Mundo* a l'article «La crisis se ensaña más con los inmigrantes que con los españoles» (https://www.elmundo.es/espana/2014/12/01/547b907ce2704e77408b4593.html).

Per a analitzar aquesta relació a Barcelona, s'examinen les dades del nombre d'immigrants i el nombre d'aturats per als 73 barris de la ciutat. La taula 11 conté una mostra d'aquests barris, només els cinc barris més grans i els cinc més petits segons la mida de la població.

Taula 11. Immigració i atur per a una mostra de deu barris de Barcelona.

		Població immigrant		**Població en atur**	
Barri	**Població**	Nombre d'immigrants	Nombre d'immigrants per cada 1.000 habitants	Nombre d'aturats	Nombre d'aturats per cada 1.000 habitants
La Clota	590	23	39,0	18	30,5
La Marina del Prat Vermell - ZonaFranca	1.146	35	30,5	125	109,1
Vallbona	1.354	65	48,0	116	85,7
Can Peguera	2.216	68	30,7	121	54,6
Baró de Viver	2.511	84	33,5	128	51,0
El Raval	47.274	4.976	105,3	3.565	75,4
La Vila de Gràcia	50.670	3.055	60,3	2.196	43,3

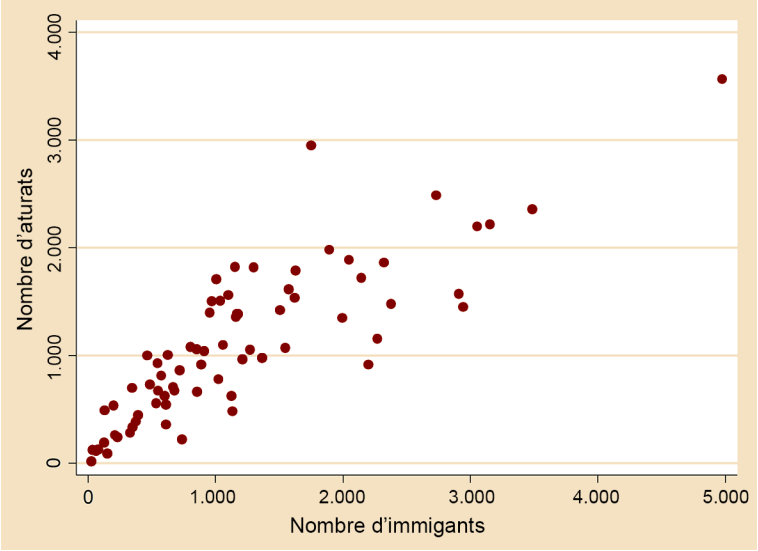
Font: Ajuntament de Barcelona (2019), Portal d'estadístiques, xifres per barris. http://www.bcn.cat/estadistica/castella/dades/barris/index.htm

		Població immigrant		Població en atur	
La Sagrada Família	51.349	3.155	61,4	2.216	43,2
Sant Andreu	56.695	1.750	30,9	2.947	52,0
La Nova Esquerra del'Eixample	57.676	3.486	60,4	2.357	40,9

Font: Ajuntament de Barcelona (2019), Portal d'estadístiques, xifres per barris. http://www.bcn.cat/estadistica/castella/dades/barris/index.htm

En el gràfic 5 es pot veure el diagrama de dispersió que mostra la relació entre ambdues variables per als 73 barris de la ciutat de Barcelona.

Gràfic 5. Diagrama de dispersió d'immigració i atur per barris de Barcelona.



El patró que segueixen les dades sembla suggerir una relació positiva entre ambdues variables, encara que no tots els punts es troben sobre una línia recta. Per aquesta raó, és necessari mesurar la força i la direcció d'aquesta relació entre dues variables mitjançant el coeficient de correlació.

El càlcul del coeficient de correlació *r* de Pearson

La mesura estadística més utilitzada per a representar la relació lineal entre dues variables és el coeficient de correlació de Pearson, que descriu la força de l'associació entre dues variables i es designa amb la lletra *r*.

Per calcular el coeficient de correlació de Pearson entre *X* i *Y* s'utilitza la relació entre la covariància entre ambdues variables i el producte de les desviacions estàndards d'aquestes. La covariància és un valor que indica el grau de variació conjunta de dues variables aleatòries respecte a les seves mitjanes i es calcula com la mitjana del producte de les desviacions de cada variable pel que fa a la seva mitjana:

Nota

Igual que en el cas de la desviació estàndard, es divideix entre *n-1* en el cas de tenir dades mostrals o entre *n* en el cas de dades poblacionals.

$$cov(x, i) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(i_i - \bar{i})$$

Si bé el resultat no limitat en cap rang específic i el nombre en si mateix no té cap interpretació, quan la covariància produeix un valor positiu, les variables tendeixen a canviar en la mateixa direcció, mentre que quan produeix un negatiu tendeixen a canviar en la direcció oposada.

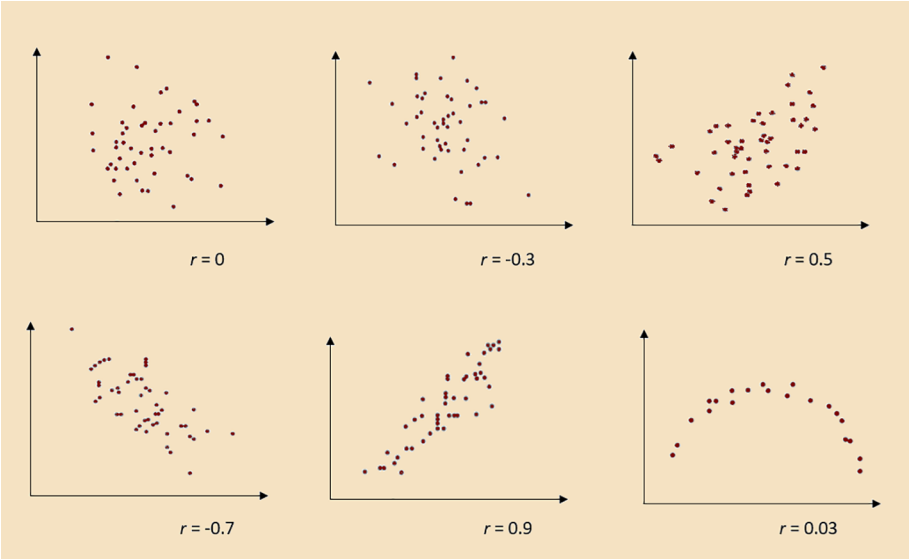
A diferència de la covariància, la *r* de Pearson dona un valor limitat a un rang i permet una interpretació fàcil. Aquest indicador compara la quantitat de variabilitat conjunta entre *X* i *Y* mesurada per la covariància amb la quantitat en què *X* i *Y* varien per separat. De manera que la fórmula per a calcular el coeficient de correlació és:

$$r = \frac{cov(X, Y)}{S_{YS\,X}}$$

El valor del coeficient de correlació de Pearson *r* pot prendre valors des de menys un fins a un, és a dir, $-1 \leq r \leq 1$. Com més proper a un sigui el valor del coeficient en qualsevol direcció, més forta serà l'associació lineal entre les dues variables. Per contra, com més proper a 0 sigui el coeficient de correlació, més feble és l'associació entre ambdues variables. Un coeficient de correlació d'1 o bé de -1 indica una correlació perfecta, és a dir, el diagrama de dispersió representaria un núvol de punts en una línia recta perfecta amb pendent positiu o negatiu, respectivament. Si és igual a 0 es conclourà que no existeix cap relació lineal entre ambdues variables.

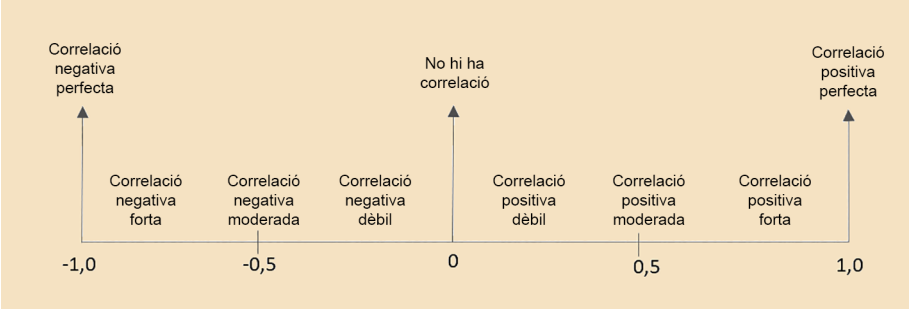
Noteu que el coeficient de correlació de Pearson es refereix únicament a l'associació lineal entre dues variables. Addicionalment, la força de la correlació que indica el coeficient no depèn de la direcció. Per exemple, un coeficient de correlació proper a 0 (0.07, per exemple) indica que la relació lineal és molt feble, i s'arriba a la mateixa conclusió si *r* fos -0.07. Els coeficients -0.94 i 0.94 indiquen una correlació molt forta entre les dues variables. El gràfic 6 mostra exemples de coeficients de Pearson obtinguts per a diferents diagrames de dispersió. És interessant veure que en darrer cas el valor de *r* és gairebé 0. Això passa perquè malgrat que hi ha una associació entre ambdues variables, l'associació no és lineal i el coeficient de correlació de Pearson només identifica correlacions lineals.

Gràfic 6. Exemples de coeficients de Pearson per a diferents patrons de dispersió.



No existeix una regla precisa per a afirmar si la correlació entre les variables es pot considerar forta o feble, ja que la qualificació depèn del rigor de l'estudi i l'experiència de l'investigador per a jutjar els resultats d'acord amb les expectatives plantejades. No obstant això, es poden seguir certes convencions per a resumir la força i la direcció del coeficient de correlació que poden ajudar a la interpretació d'un coeficient de correlació determinat, com es veu en el gràfic 7.

Gràfic 7. Interpretació d'un coeficient de correlació de Pearson.



En resum, les característiques del coeficient de correlació són les següents:

- Mostra la direcció i la força de la relació lineal (recta) entre dues variables quantitatives.
- Varia de -1 a 1, tots dos inclusivament.
- Un valor proper a 0 indica que hi ha poca associació entre les variables.
- Un valor proper a 1 indica una associació directa o positiva entre les variables.
- Un valor proper a -1 indica una associació inversa o negativa entre les variables.
- La unitat de mesura de X i Y no exerceix cap paper en la interpretació de r.

Com a conclusió, hem de recordar que una associació, per forta que sigui, no implica necessàriament causalitat. Per exemple, encara que es pugui demostrar que els ingressos de professors i el nombre de pacients en institucions psiquiàtriques han augmentat proporcionalment i que el coeficient de correlació entre ambdues variables sigui positiu, no es pot concloure que una variable causi l'altra. Així mateix, s'ha demostrat que, encara que hi ha una correlació positiva entre el cost de l'acomiadament i el nivell d'ocupació, tal associació no implica que la reducció del cost de l'acomiadament generi nous llocs de treball. Les relacions d'aquest tipus en les quals sembla que hi ha una relació causal que en realitat no existeix es denominen **correlacions espúries**. El que es pot concloure quan es tenen dues variables amb forta correlació és que hi ha una relació o associació entre ambdues, no que un canvi en una ocasiona un canvi en l'altra.

Abaratiment de l'acomiadament

Un estudi del Centre de Recerca en Economia Internacional (CREI) de la Universitat Pompeu Fabra adverteix que abaratir l'acomiadament no afavoreix per si sol l'ocupació, sinó que «depèn crucialment» del grau d'incertesa de l'empresari davant de la resolució d'un possible conflicte judicial. Maïa Güell, autora de l'estudi, ha argumentat que aquests costos poden tenir «efectes ambigus, neutres o fins i tot positius» sobre la taxa d'ocupació. Vegeu el següent article a *La Vanguardia* (31/08/2010): https://www.lavanguardia.com/economia/fiscalidad-empresa/20100831/53992656625/un-estudio-cuestiona-que-abaratar-el-despido-genere-automaticamente-empleo.html

Per a continuar amb l'exemple de l'apartat anterior, calculem el coeficient de correlació de Pearson entre el nombre d'immigrants i el nombre d'aturats per als 73 barris de la ciutat de Barcelona:

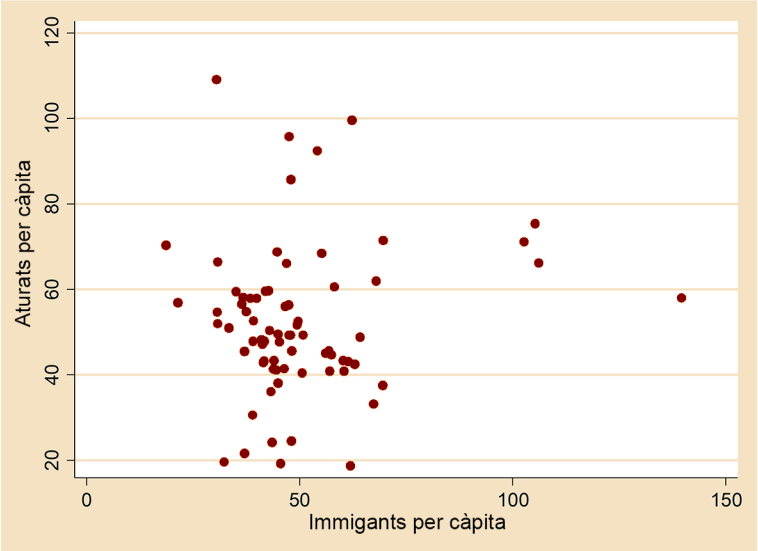
$$r = \frac{cov(X, Y)}{S_{YS\,X}} = \frac{570.688}{925,2 \times 711,4} = 0,84$$

Amb aquest valor, es pot interpretar que existeix una forta correlació positiva entre el nombre d'immigrants i el nombre d'aturats a Barcelona.

Com s'ha dit abans, aquesta interpretació no implica que hi hagi una relació causal entre ambdues variables. A més, és necessari analitzar si la correlació calculada pot ser una correlació espúria. En aquest cas, si la relació entre el nombre d'immigrants i d'aturats està influenciada per altres factors que encara no s'han analitzat, pot ser que la correlació trobada no representi realment la intensitat de la relació entre ambdues variables. Un factor que, per exemple, pot afectar aquesta relació és la mida dels barris. Si als barris amb més població és més probable que hi hagi més immigrants i al mateix temps més aturats, la mida dels barris pot estar intervenint en la relació i distorsionant la veritable intensitat d'aquesta.

Per a controlar per la mida dels barris, és possible fer el mateix tipus de diagrames i el càlcul de la *r* de Pearson utilitzant el nombre d'aturats i el nombre d'immigrants per càpita, és a dir, per cada 1.000 habitants a cada barri. El gràfic 8 mostra el diagrama de dispersió.

Gràfic 8. Diagrama de dispersió d'immigració i atur per cada 1.000 habitants per barris de Barcelona.



Comparat amb el gràfic anterior, en el qual es mostrava una clara relació positiva entre el nombre d'aturats i el nombre d'immigrants, amb les dades per càpita és més difícil identificar visualment la direcció de la relació entre ambdues variables, si és que se'n pot identificar cap.

$$r = \frac{cov(X, Y)}{S_{YS\,X}} = \frac{38{,}56}{18{,}94 \times 17{,}47} = 0{,}116$$

De fet, si es calcula el coeficient de correlació amb les dades per cada 1.000 habitants, s'observa que la correlació és positiva, però molt feble.

Aquesta observació reflecteix la importància d'avançar en l'anàlisi i no treure conclusions simplement basant-se en la inspecció visual d'un diagrama de dispersió o d'un coeficient de correlació. És fonamental analitzar dos tipus de problemes.

- En primer lloc, l'existència de particularitats de les dades que poden afectar els resultats de l'anàlisi, com ara alguns problemes de mesurament i l'existència de dades atípiques o *outliers*. En el nostre exemple, d'una banda, no tota la immigració està registrada formalment, i pot ser que els registres estadístics no siguin tan precisos com haurien de ser, i, d'altra banda, l'existència de dades atípiques en què la immigració i l'atur són particularment alts, com és el cas del barri del Raval (és el punt més allunyat de l'origen en el primer diagrama de dispersió), pot estar generant un coe-

ficient de correlació excepcionalment alt quan no es controla per la mida de la població.

- En segon lloc, és fonamental analitzar la naturalesa de la relació i els altres factors que la poden afectar. A més de la mida de la població, altres variables poden afectar la intensitat de l'associació entre el nombre d'immigrants i el nombre d'aturats als barris de Barcelona, com ara les característiques econòmiques d'aquests barris i l'oferta d'ocupació, entre molts altres.

5. El model de regressió lineal

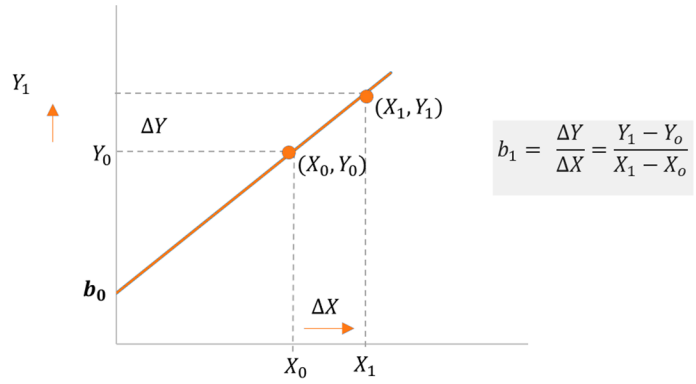
A l'apartat anterior, s'ha mostrat que el diagrama de dispersió i l'anàlisi de correlació lineal ofereixen una idea bastant aproximada sobre el tipus d'associació que existeix entre dues variables i la intensitat d'aquesta associació. No obstant això, totes dues anàlisis són limitades quan es requereix una descripció més precisa d'aquesta relació i no tenen capacitat predictiva. Per això, una vegada s'ha fet una anàlisi de correlació lineal, el següent pas és desenvolupar una equació matemàtica que permeti estimar el valor d'una variable dependent sobre la base del valor d'una altra o altres variables independents. Aquest procediment es coneix com a **anàlisi de regressió** i, quan només s'analitza la relació entre dues variables, es denomina **regressió lineal simple**.

La regressió lineal simple pretén trobar la recta que millor representi tots els punts representats en un diagrama de dispersió. L'equació d'una línia recta es representa així:

$$Y = b_0 + b_1 X$$

on el coeficient b_0 representa la intersecció de la línia recta amb l'eix vertical, és a dir, el valor de la variable Y quan la variable X pren un valor igual a 0, mentre que b_1 representa el pendent o grau d'inclinació de la línia recta, és a dir, el canvi mitjà que es produeix en Y quan la variable X varia en una unitat.

Gràfic 9. Recta de regressió lineal.



Coneixent el valor d'aquests dos coeficients, és possible descriure la relació entre X i Y, i també estimar amb alguna precisió el valor que prendria Y davant de diferents valors de X.

En la situació poc probable en què tots els punts del núvol de dispersió se situessin a sobre d'una mateixa línia recta, trobar l'equació que millor representa la relació entre X i Y, és a dir, trobar els paràmetres b_0 i b_1, no seria un proble-

ma, ja que n'hi hauria prou d'unir els punts per a obtenir la recta amb més bon ajust a la línia de punts. Però en un núvol de punts més realista amb més dispersió, és possible traçar una infinitat de línies rectes diferents. L'anàlisi de regressió lineal simple tracta de trobar la recta que representi millor el conjunt de dades observades, és a dir, la que s'ajusti a una descripció més precisa del núvol de punts representats en el diagrama de dispersió.

Una vegada determinada aquesta equació, aquesta es pot usar per a predir els valors de *Y* que haurien d'ocórrer amb valors donats de *X*. És a dir, l'equació de regressió $Y = b_0 + b_1 + X$ es pot usar per a predir un valor *y* individual que s'espera que ocorri amb un valor *x* observat:

$$\hat{y}_i = b_0 + b_1 x$$

Aquests valors es coneixen com a *valors predits* o *esperats* de *Y* perquè són el que la línia ens porta a esperar que estiguin associats amb els valors de *X*. El símbol $\hat{y}_i$, s'usa per a representar un valor de *Y* que es prediu usant l'equació de regressió, perquè puguem distingir-lo d'un valor de *Y* real.

5.1. El càlcul dels coeficients de regressió

Com s'ha explicat a l'apartat anterior, la regressió lineal simple busca trobar la recta que representi millor tots els punts representats en un diagrama de dispersió. No obstant això, múltiples línies que aproximarien raonablement les dades observades es podrien traçar en aquest diagrama. Com es determina quina és la recta que ajusta millor les dades?

Per a determinar l'equació de la línia recta que s'ajusta millor a les observacions, l'anàlisi de regressió simple utilitza el mètode de **mínims quadrats ordinaris**. En aquest mètode, s'empren les dades de la mostra per a estimar els paràmetres b_0 i b_1 que minimitzen la suma dels quadrats de les desviacions entre els valors observats de la variable dependent, representats per $_{\mathrm{y\,i}}$, i els valors estimats de la variable dependent a partir dels valors estimats de b_0 i b_1, representats per $^{\mathrm{y}\,\hat{}}{}_i$.

Definim e_{i} com l'error associat amb l'observació *i*. Els errors o residus es defineixen així:

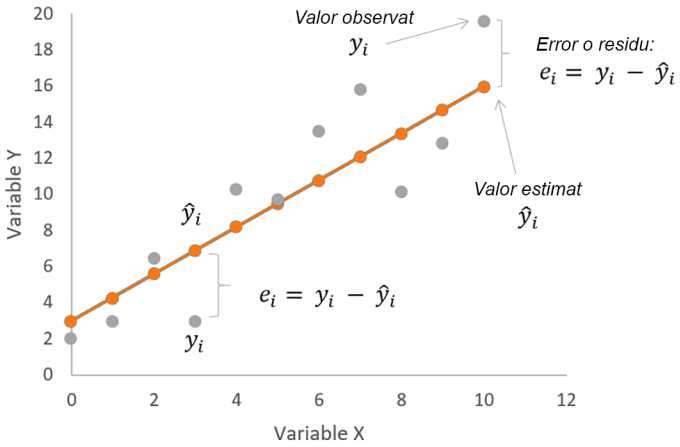
$$e_i = y_i - \hat{y}_i$$

El que es busca és que la desviació que s'obté entre la diferència vertical dels valors «reals» y_i i els valors «estimats» $\hat{y}_i$ sigui la menor possible. Atès que les diferències entre tots dos valors poden ser positives o negatives, no és correcte

simplement comparar la sumatòria de les desviacions de diferents rectes estimades. Per a garantir la minimització dels errors, el que es busca és que la suma dels quadrats de les desviacions, és a dir, $\sum i_i^2 = \left(y_i - \hat{y}_i\right)^2$, sigui mínima.

Una propietat dels errors és que el seu valor esperat és igual a 0 $\mathrm{E}(e_i) = 0$, la qual cosa significa que la mitjana dels errors és igual a 0. Això es pot entendre més fàcilment veient la representació dels errors en el gràfic 10.

Gràfic 10. Valors observats i estimats en la regressió lineal.



Denotarem la mitjana de X com a $\bar{x}$, la mitjana de Y com a $\bar{Y}$, les seves respectives desviacions estàndard com a S_y i S_x, i el coeficient de correlació de Pearson com a r. Es pot demostrar que els paràmetres b_0 i b_1 que minimitzen el quadrat de les desviacions es poden calcular com a:

$$\boldsymbol{b_1 = r\ \frac{S_i}{S_X}} \qquad \boldsymbol{b_0 = \bar{Y} - b_1 \bar{X}}$$

Un parell d'observacions. La línia de regressió de mínims quadrats sempre passa pel punt ($\bar{X}$, $\bar{Y}$), és a dir, per la mitjana d'ambdues variables. A més, noteu que el signe del coeficient de correlació coincideix amb el signe del paràmetre b_1, el pendent de la recta. En conseqüència, si el coeficient de correlació indica que la relació entre X i Y és positiva, el pendent de la recta serà positiva i el signe de b_1 serà positiu. El mateix en cas que la r de Pearson sigui negativa.

És important subratllar que els resultats del model de regressió lineal amb el mètode de mínims quadrats ordinaris s'obtenen sota una sèrie de supòsits, necessaris per a poder fer proves d'hipòtesis de la població a partir de mostres:

- La relació ens les variables X i Yés lineal en els paràmetres. Aquest és un supòsit que no resulta molt restrictiu, ja que permet incorporar relacions

no lineals entre les variables en introduir funcions (o transformacions) no lineals de *X* i/o *Y*.

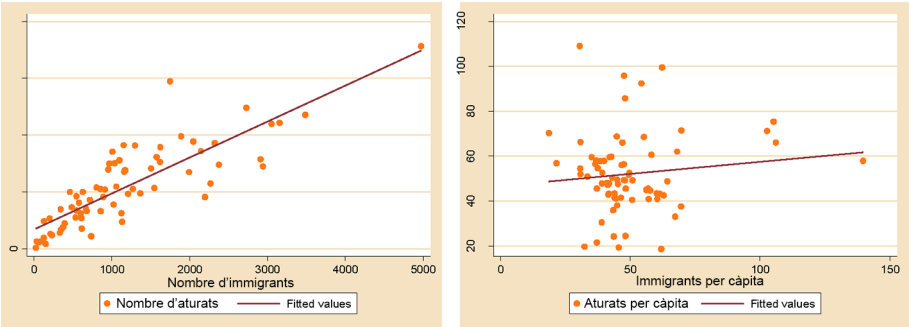
- Els errors o residus es distribueixen normalment al voltant de la recta de regressió poblacional.

- Les variàncies dels errors són les mateixes en tots els valors de *X* (propietat que es coneix com a *homoscedasticitat*).

- Els errors o residus són independents entre si: no es mostra cap patró definit (propietat que es coneix com a *no autocorrelació*).

Per a il·lustrar el càlcul dels paràmetres de la línia de regressió, és possible seguir l'exemple de la relació entre el nombre d'immigrants i el nombre d'aturats als barris de Barcelona. El gràfic 11 mostra els dos diagrames de dispersió analitzats anteriorment, inclosa ara la línia de regressió en tots dos casos.

Gràfic 11. Línia de regressió per a la relació entre immigració i atur a Barcelona.



Com es podia anticipar a partir dels coeficients de correlació calculats, en tots dos casos s'observa una relació positiva, però amb un pendent molt més inclinat en el primer cas, en el qual no es controla per l'efecte de la mida de la població a cada barri. Per a veure la magnitud de les diferències en cada cas, estimem l'equació de la regressió lineal en tots dos casos assumint que la variable *Y* és el nombre d'aturats mentre que *X* és el nombre d'immigrants.

Per a les dades del nombre d'aturats i del nombre d'immigrants per barri:

$$\boldsymbol{b}_1 = \boldsymbol{r}\ \frac{\boldsymbol{S}_y}{S_x} = 0{,}84 \times \frac{950{,}22}{711{,}41} = 0.632$$

$$\boldsymbol{b}_0 = \bar{\boldsymbol{Y}} - \boldsymbol{b}_1\bar{\boldsymbol{X}} = 1.080{,}15 - 0.632 \times 1.166{,}62 = 342{,}78$$

$$\boldsymbol{Y} = 342{,}78 + 0.632\boldsymbol{X}$$

Per a les dades del nombre d'aturats per càpita i nombre d'immigrants per càpita per barri, en tots dos casos per cada 1.000 habitants:

$$\boldsymbol{b}_1 = \boldsymbol{r}\ \frac{\boldsymbol{S}_y}{S_x} = 0,116 \times \frac{17,47}{18,94} = 0,107$$

$$\boldsymbol{b}_0 = \bar{\boldsymbol{Y}} - \boldsymbol{b}_1 \bar{\boldsymbol{X}} = 52.1 - 0.107 \times 50,1 = 46,71$$

$$\boldsymbol{Y} = 46,71 + 0,107\boldsymbol{X}$$

Les dues equacions estimades són molt diferents perquè les dades utilitzades també ho són. Com s'interpreten aquests resultats? La primera equació indica que, d'acord amb la magnitud de l'intercepto, la mitjana del nombre d'aturats als barris de Barcelona, quan el nombre d'immigrants és igual a 0, és de 342.78. En aquest cas, la interpretació de la intersecció és simple, perquè és fàcil imaginar-se una situació poc probable però factible en la qual algun barri no té cap immigrant. Amb un altre tipus de dades la interpretació de la intersecció és menys evident i no se sol comentar gaire. Es pot donar el cas d'anàlisis que produeixin una intersecció negativa amb dades que no solen ser negatives, o que produeixin valors que és poc probable que ocorrin (com ara preus zero quan s'estima la demanda de qualsevol producte segons el seu preu o el de la competència).

El paràmetre més important és el pendent de la recta. En el primer cas, l'equació implica que un increment d'1 immigrant a la ciutat està associat a un increment de 0.63 aturats. En el segon cas, el paràmetre estimat indica que cada increment d'1 immigrant per cada 1.000 habitants està associat amb un increment de 0.107 aturats per cada 1.000 habitants. El pendent en aquest cas és molt menor, però no es poden comparar directament ambdues estimacions, ja que les dades utilitzades en cada cas són diferents i, per tant, la interpretació també ho és. El que sí que es pot comparar és el grau de significació estadística dels pendents en els dos models i la bondat d'ajust de tots dos.

5.2. El nivell de significació dels coeficients de regressió

Una vegada estimats els paràmetres, s'ha de contrastar si aquests paràmetres són significativament diferents de 0, és a dir, si la relació estimada és estadísticament significativa o no. Tal com s'ha fet a les anàlisis anteriors, es tracta d'un tipus de prova d'hipòtesi que es coneix com a **contrast de significació**.

Segons s'aplica en l'anàlisi de regressió, la hipòtesi que es contrasta és que el paràmetre estimat és igual a 0 (Ho = b_1 és igual a 0). En el cas del pendent, per exemple, que el paràmetre b_1 és igual a 0, és a dir, que no existeix associació entre la variable *X* i la variable *Y*. El paràmetre calculat és el valor de b_1 calculat per la regressió amb les dades de la mostra, mentre que el valor hipotètic és 0.

Com s'ha explicat a l'apartat «Diferencies de mitjanes i prova *T*», l'estadístic *t* és un exemple d'un estadístic de prova la distribució del qual sota la hipòtesi nul·la és coneguda. La fórmula general per a una prova *t* és:

$$t = \frac{\text{par àmetre cal culat - val or hipotètic}}{\text{error estàndar del estimad or}}$$

Els graus de llibertat per a aquesta prova són el nombre d'observacions menys dues (n-2).

Per a entendre com es fa aquest contrast, es presenta el resultat de l'estadístic t, i també la taula ANOVA (anàlisi de variància). Aquests s'obtenen amb programes informàtics per a l'anàlisi estadística, ja que calcular-los manualment és complicat.

S'ha utilitzat el programa Stata, però qualsevol programa estadístic amb models de regressió lineal (R, SPSS, Excel) pot produir resultats similars.

La següent imatge mostra el resultat de la regressió per a les dades del primer model, que relaciona el nombre d'aturats i el nombre d'immigrants per barri.

Resultats de la regressió. Dades brutes.

Source	SS	df	MS	Number of obs	=	73
				F(1, 71)	=	176.14
Model	26331446.6	1	26331446.6	Prob > F	=	0.0000
Residual	10613696.8	71	149488.687	R-squared	=	0.7127
				Adj R-squared	=	0.7087
Total	36945143.3	72	513126.991	Root MSE	=	386.64

paro	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
inmig	.6320525	.0476234	13.27	0.000	.5370942	.7270108
_cons	342.7879	71.65544	4.78	0.000	199.911	485.6647

Després de comprovar que els valors estimats dels paràmetres són els que s'han calculat abans (Y= 342.78+0.632X), és possible veure el valor de l'estadístic t i el seu corresponent p-valor. Com és usual en aquest tipus de proves, a un nivell de significació del 5%, es pot rebutjar la hipòtesi nul·la que el pendent d'aquesta equació és igual a 0, ja que el p-valor és menor que el nivell de significació triat.

Per a les dades del nombre d'aturats per càpita i del nombre d'immigrants per càpita per barri, en tots dos casos per cada 1.000 habitants, també es comprova que els paràmetres calculats són els mateixos que produeix el programa (Y= 46.71+0.107X) però, a diferència de l'anterior, ara no podem rebutjar la hipòtesi nul·la que el pendent d'aquesta equació és igual a 0, ja que el p-valor és superior al nivell de significació triat.

Resultats de la regressió. Dades ponderades per cada 10.000 habitants.

Source	SS	df	MS	Number of obs	=	73
				F(1, 71)	=	0.98
Model	302.485644	1	302.485644	Prob > F	=	0.3263
Residual	21981.9546	71	309.604994	R-squared	=	0.0136
				Adj R-squared	=	-0.0003
Total	22284.4402	72	309.506114	Root MSE	=	17.596

paropc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
inmigpc	.107462	.1087193	0.99	0.326	-.1093181	.3242421
_cons	46.71821	5.823272	8.02	0.000	35.10693	58.32948

Això significa que, malgrat que la recta que s'observa en el diagrama de dispersió sembla representar una relació positiva entre ambdues variables, aquesta relació no és estadísticament significativa.

5.3. La bondat d'ajust

A més d'estimar el valor dels paràmetres de la recta de regressió i el seu nivell de significació, també s'ha d'avaluar el grau en el qual la recta s'ajusta al núvol de punts i, per tant, la capacitat de l'equació per a fer estimacions correctes, la qual cosa es coneix com la **bondat d'ajust de la regressió**, que es mesura principalment amb el coeficient de determinació.

Un aspecte útil de la regressió és que pot dividir la variació observada a Y (o variància total) en dues parts: la variació en els valors predits $\hat{Y}$ i la variació en els errors de predicció e_i.

La variació de Y es denomina *suma de quadrats de la regressió* o *suma dels quadrats totals*, que denotarem com a SSY, i es defineix com la suma de les desviacions al quadrat de la variable dependent Y respecte a la mitjana de la mateixa variable, $\bar{Y}$. Quan es calcula en una mostra:

$$SSY = \sum\left(y_i - \bar{Y}\right)^2$$

La suma de quadrats totals SSY es pot dividir en dues parts: la suma dels quadrats explicada pel model (SSY') i la suma de quadrats dels errors (SSE):

- La suma de quadrats explicats pel model és la suma de les desviacions al quadrat dels valors predits $\hat{y}_i$ i la mitjana d'aquests valors:

$$SSY' = \sum\left(\hat{y}_i - \bar{\hat{y}}\right)^2$$

- La suma del quadrat dels errors, com hem definit abans, és la suma dels errors de predicció al quadrat:

$$\text{SSE} = \sum e_i^2 = \sum \left(y_i - \hat{y}\right)^2$$

Així, la suma de quadrats de la regressió SSY es pot escriure com a:

$$\text{SSY} = \text{SSY}' + \text{SSE}$$

És a dir, la capacitat de l'equació de regressió per a predir amb precisió els valors de *Y* es mesura calculant primer la proporció de la variabilitat dels valors de *Y* estimats amb l'equació de regressió i la proporció que no es pot predir o explicar amb el model. Mentre que SSY és la variació total de *Y*, SSY' és part d'aquesta variació explicada pel model i SSE és la part de la variació no explicada.

Per tant, la proporció de variació explicada es pot calcular com a:

$$Proporción explicada = \frac{SSY'}{SSY}$$

De la mateixa manera, la proporció no explicada és:

$$Proporción no explicada = \frac{SSE}{SSY}$$

La proporció explicada pel model és el que es coneix com el **coeficient de determinació** de la regressió, el qual expressa el percentatge de variació de la variable dependent causat o atribuït a la variació de la variable independent.

Existeix una relació important entre la proporció de variació explicada i la correlació de Pearson: el coeficient de determinació es pot calcular com el quadrat del coeficient de correlació (r^2). Per això, la notació del coeficient de determinació generalment és R^2. Atès que aquest coeficient es calcula com el quadrat del coeficient de correlació de Pearson, el seu valor està limitat en l'interval entre 0 i 1. Com la *r* de Pearson, en la mesura en què els punts s'apropin a la recta, el coeficient de determinació serà més proper a 1, i si els punts s'allunyen de la recta, el coeficient de correlació serà més proper a 0.

Per a veure com funciona la bondat d'ajust en l'exemple que hem utilitzat, usarem de nou el resultat que produeix Stata, ja que normalment les proporcions explicades i no explicades no es calculen manualment, sinó que s'obtenen de la taula ANOVA que produeix la instrucció de regressió.

ANOVA. Dades brutes.

Source	SS	df	MS		
Model	26331446.6	1	26331446.6	Number of obs	= 73
Residual	10613696.8	71	149488.687	F(1, 71)	= 176.14
				Prob > F	= 0.0000
Total	36945143.3	72	513126.991	R-squared	= 0.7127
				Adj R-squared	= 0.7087
				Root MSE	= 386.64

paro	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
inmig	.6320525	.0476234	13.27	0.000	.5370942	.7270108
_cons	342.7879	71.65544	4.78	0.000	199.911	485.6647

Per a les dades del primer model que relaciona el nombre d'aturats i el nombre d'immigrants per barri:

$$\boldsymbol{Proporcióexplicada} = \frac{\boldsymbol{SSY'}}{\boldsymbol{SSY}} = \frac{26331446{,}6}{36945143{,}3} = 0{,}7127$$

$$\boldsymbol{R}^2 = \boldsymbol{r}^2 = 0{,}84^2 = 0{,}7127$$

D'acord amb el coeficient de determinació en aquest model, la variació en el nombre d'immigrants a la ciutat explica el 71.3 % de la variació en el nombre d'aturats. Com veiem a la següent taula, el coeficient de determinació en el segon model només és de 0.0136. Això implica que la variació en el nombre d'immigrants per cada 1.000 habitants a la ciutat només explica l'1.36% de la variació en el nombre d'aturats per càpita.

ANOVA. Dades ponderades per cada 10.000 habitants.

Source	SS	df	MS		
Model	302.485644	1	302.485644	Number of obs	= 73
Residual	21981.9546	71	309.604994	F(1, 71)	= 0.98
				Prob > F	= 0.3263
Total	22284.4402	72	309.506114	R-squared	= 0.0136
				Adj R-squared	= -0.0003
				Root MSE	= 17.596

paropc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
inmigpc	.107462	.1087193	0.99	0.326	-.1093181	.3242421
_cons	46.71821	5.823272	8.02	0.000	35.10693	58.32948

En conseqüència, és possible concloure que la bondat d'ajust del primer model és millor que la bondat d'ajust del segon. De nou, no es pot concloure que el primer model és «millor», perquè l'objectiu de tots dos models és explicar una variable dependent diferent. Però així com el primer té més poder predictiu, en el segon cas direm que el model estimat no és útil per a predir la variació observada en el nombre d'aturats per cada 1.000 habitants als barris de Barcelona.

5.4. Introducció a l'anàlisi multivariant

És clar que la majoria de les relacions d'interès involucren més de dues variables. En la pràctica, incorporarem tants controls com creiem necessari. Un model de regressió lineal múltiple és més adequat per a anàlisi *ceteris paribus*, ja que permet controlar explícitament molts dels factors que afecten simultàniament la variable dependent. A més, ens dona més flexibilitat per a modelar la relació entre Y i les X. Quan hi ha més d'una variable independent x, la més utilitzada és la regressió múltiple. La lògica és similar a la de la regressió univariant.

Lectura recomanada

Per a saber més sobre la regressió múltiple podeu consultar:

D. A. Lind; W. G. Marchal; S. A. Wathen (2017). *Análisis de regresión múltiple*. A: D. A. Lind; W. G. Marchal; S. A. Wathen. *Estadística aplicada a los negocios y la economía* . Mèxic: McGraw-Hill/Interamericana Editors, SA de CV.

Exercicis d'autoavaluació

1. Defineix les variables per als conceptes següents:

- Ingressos
- Pressupost
- Població objectiu d'una política educativa

Quin tipus de variable és cadascuna?

2. Per a les dades de la taula 2.Taula de dades processades (Fragment de l'estudi «3242 del Macrobaròmetre de març del 2019» del CIS:

- Quin tipus de variable és la variable «Edat»? De quina altra manera es podria representar aquesta variable?
- Quin tipus de variable és la variable «Principal_problema»?
- Quin és el valor de la variable «Estudis» per a l'observació 8?

3. Tens una llista d'observacions del nombre de veïns que han anat a l'Ajuntament per dia a demanar solució als seus problemes en els últims dos mesos.

10	11	9	9	5	9	10	11	11	6
5	9	10	10	11	7	6	3	11	4
10	9	6	7	10	8	7	6	10	6
9	3	8	8	10	11	10	7	5	6
5	6	10	6	8	6	10	11	9	9
11	6	10	8	11	9	6	7	11	10
6	11	10	6	10	9	8	10	9	7
9	9	7	11	7	10	9	8	8	10
3	2	9	11						

Construeix una taula de distribució de freqüències i una representació gràfica de les dades. Amb aquesta descripció, escriu una interpretació per a incloure aquesta estadística en l'informe trimestral de l'Ajuntament.

4. Les següents dades representen el nombre de setmanes que van necessitar set empreses per a obtenir les seves llicències d'activitat.

2, 11, 5, 7, 6, 7, 4

- Calcula el temps mitjà per a obtenir una llicència d'activitat.
- Calcula la mitjana de la mostra.
- Calcula la moda de la mostra.

5. Amb la finalitat de fer recomanacions sobre un projecte legislatiu, es vol estudiar fins a quin punt existeix relació entre el copagament i la valoració dels serveis sanitaris. S'ha fet una enquesta a una mostra aleatòria de 100 individus, tal com es mostra en la següent taula de freqüències observades. Confirma aquesta evidència la hipòtesi plantejada amb un nivell de confiança del 95%?

Valoració del servei	**Copagament sanitari**		**Total**
	Sí	No	
Dolent	20	25	45

Valoració del servei	Copagament sanitari		Total
Bo	10	45	55
Total	30	70	100

Solucionari

1.

Ingressos:

- Nivell d'ingrés subjectiu (alt, mitjà, baix): variable ordinal.
- Nivell d'ingrés objectiu (0-800 €; 801-1.000 €; 1,001-1,200 €; 1,201-1,800 €; 1,801-3,000 €; 3,001-6,000.€; més de 6,000 €): variable ordinal.
- Ingressos mensuals (euros per mes): variable quantitativa contínua.

Pressupost:

- Grau d'execució pressupostària com a percentatge del pressupost total (valor entre 0 i 100%): variable quantitativa contínua.

Població objectiu d'una política educativa:

- Elegibilitat_ Elegible /No elegible: variable categòrica.
- Algun criteri d'elegibilitat (per exemple: Nota mitjana – valor entre 0 i 10): variable contínua.

2. La variable «Edat» és de tipus quantitatiu. També es podria representar com una variable categòrica en la qual s'agrupin valors per rangs d'edat (menys de 18 anys; entre 18 i 25 anys; entre 26 i 40 anys; entre 40 i 65 anys; més de 65 anys).

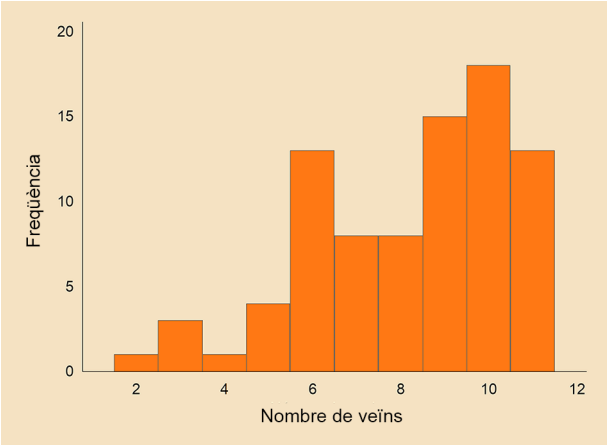
La variable «Principal_problema» és una variable categòrica.

El valor de la variable «Estudis» per a l'observació 8 és estudis superiors, tal com es pot veure en la intersecció de la fila 8 i la columna «Estudis» de la taula 2.

3. La taula de distribució de freqüències per al nombre de veïns que han anat a l'Ajuntament per dia per a demanar solució als seus problemes en els últims dos mesos:

Nombre veïns	Freqüència	Percentatge	Percentatge acumulat
2	1	1.19	1.19
3	3	3.57	4.76
4	1	1.19	5.95
5	4	4.76	10.71
6	13	15.48	26.19
7	8	9.52	35.71
8	8	9.52	45.24
9	15	17.86	63.1
10	18	21.43	84.52
11	13	15.48	100
Total	**84**	**100**	

La representació gràfica d'aquestes dades és un histograma:



Podem observar que el nombre de veïns que van visitar l'Ajuntament en els últims tres mesos varia entre un mínim de dos i un màxim d'onze veïns per dia. També podem concloure que la majoria dels dies el nombre de veïns que ha visitat l'Ajuntament en els últims tres mesos és superior a sis.

4. Per a calcular la mitjana hem de sumar els valors i dividir-los pel nombre d'observacions. Tenim: (2 + 11 + 5 + 7 + 6 + 7 + 4)/7 = 42/7 = 6. Així, el temps mitjà que va caldre per a obtenir una llicència d'activitat és de 6 setmanes.

Per a calcular la mediana, primer hem d'ordenar les dades en ordre creixent:

2, 4, 5, 6, 7, 7, 11

Com que la mida de la mostra és 7, es dedueix que la mediana de la mostra és el quart valor més petit. El nombre mitjà de la mostra de setmanes que va caldre per a obtenir una llicència d'activitat és m = 6 setmanes.

Per a calcular la moda fem una taula de freqüència i veiem que el nombre de setmanes que es repeteix més és 7. La moda per a obtenir una llicència d'activitat és de 7 setmanes.

Setmanes	Freqüència
2	1
4	1
5	1
6	1
7	2
11	1
Total	**7**

5. El primer pas per a provar la hipòtesi que el copagament sanitari està relacionat amb la valoració dels serveis és calcular les freqüències esperades:

Valoració del servei	Copagament sanitari		Total		Valoració del servei	Copagament sanitari	
	Sí	No				Sí	No
Dolent	20	25	45		Dolent	13.5	31.5
Bo	10	45	55		Bo	16.5	38.5

Valoració del servei	Copaga-ment sanitari		Total		Valoració del servei	Copaga-ment sanitari	
Total	30	70	100				

Posteriorment es calcula l'estadístic khi quadrat:

$$\chi^2 = \sum \frac{(Fo - Fe)^2}{Fe} = 3.13 + 2.56 + 1.34 + 1.10 = 8.13$$

Els graus de llibertat són: (n-1) x (m-1) = 1 x 1 = 1

Mirant a la taula khi quadrat obtenim que la probabilitat d'obtenir un valor 8,13 o més gran amb 1 grau de llibertat és menor que p = 0,005. Podem arribar a la conclusió que el valor és estadísticament significatiu, ja que és menor que 0,01. Amb aquesta informació podem concloure que la valoració del servei varia negativament amb el copagament sanitari, és a dir, els qui es veuen afectats pel copagament donen menys valor al servei que els qui no s'han vist afectats per aquest.

Graus llibertat	0.1	0.05	0.025	0.01	0.005
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75

Bibliografia

Crespo, I.; Anduiza, E.; Méndez, M. (2009). *Metodología de la ciencia política*. Madrid: Centro de Investigaciones sociológicas («Cuadernos Metodológicos», 28).

Manual que explica amb molt detall i claredat els conceptes de la metodologia de recerca en les ciències socials. Els exemples que ofereix sobre temes de ciències polítiques expliquen clarament les definicions i les aplicacions pràctiques de les teories, els conceptes, les estratègies de recerca, les dades i la contrastació d'hipòtesis.

Glass, G. V.; Stanley, J. C.; Gómez, E. G.; Guzmán, E. (1986). *Métodos estadísticos aplicados a las ciencias sociales*. Mèxic: Prentice-Hall Hispanoamericana.

Manual clàssic d'estadística amb explicacions molt completes i detallades sobre estadística descriptiva i inferencial. Amb les seves cinc-centes pàgines mostra els detalls de les tècniques i proporciona exercicis per a cada tema.

Hernández, J. J. C. (2007). *Conceptos básicos de estadística para ciencias sociales*. Madrid: Delta Publicaciones.

Llibre introductori que serveix com a guia per a triar els mètodes estadístics apropiats per a cada tipus de problema en les ciències socials. Interessant pel l'èmfasi a aprendre per a què serveix i com és possible aplicar l'estadística.

Lind, D. A.; Marchal, W. G.; Wathen, S. A. (2017). *Estadística aplicada a los negocios y la economía*. Mèxic: McGraw-Hill/Interamericana Editors, SA de CV.

Manual d'estadística molt complet i amb un tractament rigorós de la teoria. És molt detallat en les explicacions i conté exemples i exercicis per a desenvolupar en Excel. Els objectius d'aprenentatge són útils a l'hora de programar el procés d'aprenentatge.

Meier, K.; Brudney, J.; Bohte, J. (2012). *Applied statistics for public and nonprofit administration*. Boston: Wadsworth, Cengage Learning.

Un llibre indispensable per a l'estudi de la gestió pública pel seu enfocament intuïtiu i els seus exemples aplicats. Desenvolupa els conceptes estadístics sense entrar en els detalls matemàtics complexos.

Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences*. Nova York: Routledge.

Un manual avançat sobre estadística aplicada. És un complement per als qui vulguin continuar aprenent noves tècniques que no tracta aquest material o vulguin aprofundir en els conceptes i mètodes. Els seus exemples i exercicis s'expliquen per al programa estadístic SPSS.