L'anàlisi exploratòria de dades

PID_00270394

Albert Padró-Solanet i Grau

Temps mínim de dedicació recomanat: 4 hores

Universitat Oberta de Catalunya

Albert Padró-Solanet i Grau

Professor dels Estudis de Dret i Ciència Política de la UOC. Màster en Gestió Pública UAB. Màster en Ciència Política UAB. Llicenciat en Filosofia UAB.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Albert Padró-Solanet (2020)

Primera edició: febrer 2020

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

Índex

Introducció

Aquest mòdul complementa el que es dedica a la metodologia quantitativa de l'assignatura *Metodologia de les Ciències Socials* del grau de Criminologia. Per una banda, vol ser un aprofundiment i una reformulació de conceptes fonamentals del mòdul com la conceptualització, l'operacionalització i el mesurament per ajudar que siguin utilitzats conscientment en la pràctica de la recerca empírica. Per altra banda, és un repàs dels conceptes bàsics de l'anàlisi exploratòria de dades que ajudi l'estudiant a refrescar i a descobrir els instruments que l'estadística ha dissenyat per ajudar a entendre i donar sentit a les dades recollides en la realitat empírica.

Objectius

Amb la lectura d'aquest mòdul l'estudiant assolirà els objectius següents:

1. Entendre la relació de l'operacionalització i mesurament de les variables amb la conceptualització teòrica.

2. Distingir el nivell de mesurament de les variables i entendre els seus condicionants.

3. Entendre els conceptes de fiabilitat i validesa aplicats a les mesures dels conceptes criminològics.

4. Reconèixer que els instruments estadístics descriptius estan associats al nivell de mesurament de les variables.

5. Conèixer els principals instruments de l'estadística descriptiva univariant segons el nivell de mesurament de les variables: els gràfics univariants, les mesures de centralitat i mesures de dispersió de les variables.

6. Conèixer els instruments que mostren l'associació entre diferents variables en els diferents nivells de mesurament.

1. La conceptualització i l'operacionalització de les variables

1.1. La conceptualització

La primera tasca de tota recerca científica consisteix en aclarir les nostres idees de com funcionen els fenòmens que volem estudiar. És normal que els **conceptes** utilitzats en la parla ordinària (les paraules o els símbols que utilitzem per referir-nos a idees o representacions mentals) tinguin una multiplicitat de significats diferents que, de vegades, fins i tot poden ser contradictoris. El filòsof idealista Hegel veia en aquesta característica del llenguatge (sobretot, l'alemany) una virtut, ja que recollia la dinàmica dialèctica de la realitat. Però, la ciència empírica requereix que els conceptes que utilitza siguin unívocs (tinguin un únic significat) i siguin precisos per poder construir explicacions de la realitat clares i que, posteriorment, permetin construir hipòtesis que es puguin testar empíricament. Si no sabem amb claredat què és el que esperem, com podrem saber si el que pensàvem de la realitat és correcte o no? S'ha de saber amb claredat a què es refereixen els conceptes amb els que intentem explicar la realitat que interessa a la criminologia: Què és un delicte? Què és reincidència? Què és la desorganització social? Què és sentiment d'inseguretat? Etcètera.

Sobre el concepte de delicte

En el llenguatge ordinari, el concepte de **delicte** es pot fer servir d'una forma vaga per referir-se a una acció que va contra la convenció; en un lloc un determinat comportament és delictiu i en un altre, no.

Per exemple, per a molts sicilians, les infraccions de trànsit realment no són delictes ja que, per a ells, no són normes sinó recomanacions.

Però aquesta ambigüitat pot crear problemes per entendre amb precisió quins són els arguments que es fan servir per explicar les raons dels delictes i els pronòstics. Cada recerca ha de fer aquest esforç de conceptualització dels termes que utilitza. Sobretot si es tracta d'una recerca que innova en relació amb les conceptualitzacions que han fet les recerques prèvies. Naturalment, si la recerca no té com a punt central la innovació conceptual tendirà a utilitzar els conceptes desenvolupats per les recerques prèvies, de manera que sigui més eficient i comparable amb els treballs anteriors.

En una recerca sobre els factors que afecten les sentències per **delictes sexuals a menors** dictades per les audiències provincials espanyoles entre 2014 i 2016, la conceptualització de delicte no és problemàtica perquè es refereix directament als articles del Codi Penal espanyol que els tracten. De fet, si adoptem el punt de vista de la recerca, el Codi Penal pot ser vist com un enorme esforç per conceptualitzar els comportaments que atempten contra el bé públic en una determinada comunitat política. Però en altres recerques que vulguin tractar de, per exemple, **delictes violents** hem de fer un exercici per aclarir de quin tipus de violència es tracta, ja que com deixen clars els manuals de retòrica, la violència, com qualsevol altre concepte en mans de la retòrica, pot voler dir qualsevol cosa. En el límit, no saludar la veïna podria ser considerada una conducta violenta.

Sobre el concepte de reincidència

El concepte de **reincidència** es refereix a la repetició d'un tipus de conducta delictiva. La taxa de reincidència és la proporció d'un determinat tipus de delinqüent (per exemple, els delinqüents que es troben en llibertat provisional) que torna a cometre delictes en un temps determinat. Per calcular aquesta mesura s'utilitzen els registres de detencions o reingressos a la presó, però aquesta forma de mesurar el fenomen és problemàtica perquè hi ha una part del comportament delictiu que no queda registrat per part de la policia. Per tant, s'han de pensar formes específiques de ferho d'acord amb els objectius del tipus d'estudi (Capdevila Capdevila & Ferrer Puig, 2009; Capdevila Capdevila, Ferrer Puig & Luque Reina, 2005). Alguns estudis utilitzen la reincidència autoinformada: la reincidència que informen els mateixos infractors per pal·liar el problema de l'existència de registres de bona part dels delictes (naturalment, refiar-se de les confessions dels delinqüents, encara que sigui en entrevistes anònimes, també és problemàtic). Existeixen diferents conceptes de reincidència depenent del punt del sistema de justícia penal que vulgui ser mesurat i avaluat. La reincidència policial es refereix a una nova detenció; la reincidència judicial es refereix a un nou processament; la reincidència penal a una nova pena o mesura cautelar; la reincidència jurídica es refereix a un nou fet delictiu del mateix títol del Codi Penal. L'equip d'Eulàlia Luque va utilitzar el concepte de reincidència penitenciària:

«el reingrés en un centre penitenciari de persones que prèviament han estat sotmeses (almenys una vegada) a una pena de presó.»

La ciència és una tasca col·lectiva. La comunitat científica s'especialitza en camps específics i comprova si les proves fetes per altres científics són reproduïbles i si són o no correctes. Intenta proposar explicacions i proves millors i que resolguin problemes que pensin que no s'han resolt satisfactòriament prèviament. Aquest treball col·laboratiu també necessita claredat, univocitat i precisió en els conceptes.

1.2. L'operacionalització: dels conceptes a les variables

Aquesta tasca d'aclariment dels conceptes s'anomena **conceptualització** i és prèvia a la tasca de definir com aquests conceptes poden ser **mesurats empíricament**. La tasca de definir la forma com s'han de mesurar els conceptes s'anomena **operacionalització**.

L'operacionalització recull les instruccions que indiquen com s'ha d'etiquetar, mesurar o identificar un concepte.

L'operacionalització converteix els conceptes teòrics en variables empíriques de les quals podem tenir diferents indicadors.

Taula 1. Conceptes, variables i indicadors

Concepte	Variable	Indicador
Estabilitat residencial	Quantitat de canvi a la població d'un veïnat	Taxa de canvi de població en un any = (nous residents + residents migrats)/total població veïnat
Heterogeneïtat ètnica	Diversitat de la composició ètnica d'un veïnat	% de població no ciutadana espanyola en un veïnat

Concepte	Variable	Indicador
Control local	Capacitat d'un veïnat de de supervisar els propis membres o forasters	Taxa de pertinença dels veïns en associacions formals i informals (a partir de la pregunta d'enquesta: A quines associacions del veïnat pertany?)

Sobre el concepte de desorganització social

Kornhaurser (1978: 63) defineix el concepte de **desorganització social** com la situació que «existeix en primera instància quan l'estructura i la cultura d'una comunitat és incapaç d'implementar i d'expressar els valors dels seus propis residents.» El concepte de desorganització social vol capturar la idea que existeixen comunitats humanes, veïnats, que estructuralment no són capaços de combatre la delinqüència i assolir l'aspiració a un millor entorn. A una comunitat desorganitzada li manca tot el que caracteritza una comunitat organitzada:

1) solidaritat o un consens sobre normes i valors essencials (els residents valoren les mateixes coses, com l'absència de delinqüència);

2) cohesió, o un lligam fort entre els veïns (els veïns es coneixen i es valoren entre ells);

3) integració, la interacció social regular.

La intuïció darrere d'aquest concepte pot ser clara, però l'operacionalització per comprovar la teoria és complexa. Per exemple, Sampson i Groves (1989) van crear índexs als barris que tinguessin en compte l'estatus socioeconòmic, l'heterogeneïtat ètnica, la mobilitat residencial, la disrupció familiar i la urbanització, de la mateixa forma que mesures de desorganització social, etc., per tal de comprovar la relació amb les taxes de criminalitat. En aquests casos tan complexos, la definició de com s'operacionalitza el concepte de desorganització social, és una forma de definir-lo (Kubrin & Wo, 2015).

El cas de la **justícia penal** és especial, perquè l'operacionalització del delicte és el que fa el Codi Penal. Els jutges i magistrats són l'equip entrenat per atribuir a cada denúncia uns valors específics del codi penal a cada sentència que pronuncien. Després de fer el judici i escoltar les parts, fiscals i advocats de la defensa, es diu si s'ha comès un delicte i, si és així, quin tipus de delicte s'ha comès i es valoren les circumstàncies en les que s'ha comès, d'aquí se'n deriven les penes que s'imposen als acusats i les compensacions a les víctimes. L'existència de tot aquest enorme aparell d'operacionalització i de mesurament fa que sembli, des del punt de vista de la recerca en criminologia i en la justícia penal, que l'operació de mesurament en aquest cas és trivial, però aquesta trivialitat és aparent, perquè els costos de mesurar correctament han estat assumits prèviament per un conjunt d'organitzacions entrenades i especialistes en mesurar aquest delicte de forma vàlida i fiable: cossos de seguretat, jutjats i personal legal. Precisament la pregunta que es fa respecte dels funcionament del sistema de justícia penal és sobre la validesa i fiabilitat de les mesures que resulten.

Per exemple, un dels primers estudis sobre les sentències va demanar a diferents jutges que resolguessin un mateix cas i van observar les àmplies diferències que existien en les sentències que van pronunciar.

De vegades, la millor estratègia per resoldre els problemes de conceptualització, d'aclarir-los, fer-los unívocs i menys abstractes consisteix precisament en definir el concepte a través de la seva operacionalització; per això es parla de la **definició operativa dels conceptes**. L'operacionalització és com una **recepta de cuina** que pot seguir qualsevol altre investigador per comprovar que es poden obtenir les mateixes troballes d'una recerca. L'operacionalització està connectada amb l'enfocament teòric, és una part essencial de la metodologia i permet contrastar si es millora el coneixement dels fenòmens que volem estudiar; per tant, forma part d'una estratègia per permetre l'acumulació del coneixement científic en una àrea.

Quan l'investigador **operacionalitza** un concepte en **variables** ens diu de quina forma s'ha de mesurar el concepte teòric en les **unitats d'anàlisi, observacions, individus, elements o casos**. De vegades aquestes unitats d'anàlisi són persones que són entrevistades en una enquesta. L'operacionalització de la mala conducta en els presos consisteix en la formulació de la pregunta que es realitzarà per mesurar aquesta mala conducta.

Per exemple: 'alguna vegada ha atacat algun altre pres que no l'havia agredit abans?'.

Altres vegades són agregats de persones i la informació que obtenim es refereix a aquests agregats.

Un exemple d'agregat pot ser una presó. El concepte de conflictivitat en una presó pot ser operacionalitzat a través del percentatge mensual d'interns que són tractats a la infermeria per lesions traumàtiques en relació al total d'interns.

Inevitablement, l'operacionalització forma part d'aquesta decisió de disseny d'investigació sobre quines són les unitats d'anàlisi. Els investigadors s'hi juguen bona part de la rellevància de la seva recerca quan trien unes unitats o altres d'anàlisi per testar les seves teories o explicacions de la realitat.

Problema del mostreig

El problema de decidir quins són els casos adequats per respondre les preguntes que formula una recerca es coneix com el *problema del mostreig*.

2. El nivell de mesurament: quina precisió tenim en la mesura de les variables?

Quan mesurem els conceptes teòrics a les unitats d'anàlisi (o casos, individus...) ho podem fer amb diferents graus de precisió. Aquest grau de precisió s'anomena **nivell de mesurament**.

El nivell de mesurament pot ser **qualitatiu** o **quantitatiu**. En el mesurament qualitatiu, els mesuraments que atribuïm a les unitats d'anàlisi (els valors de les variables en cada unitat d'anàlisi) no poden ser interpretats com una expressió quantitativa de la presència de l'atribut en la unitat d'anàlisi.

Per exemple, quan una persona es posa en una balança i veiem que pesa 60 kg, aquesta característica és numèrica. Sabem que aquesta persona pesa la meitat d'una altra persona que pesa 120 kg i el doble d'una altra que pesa 30 kg.

Aquest és un exemple de **mesurament quantitatiu**. Però no sempre tenim aquest mateix nivell de precisió en els mesuraments.

Per exemple, podem classificar els mateixos individus que abans hem pesat en homes i dones. Aquesta variable és força diferent de l'anterior. Està clar que el número amb el que identifiquem cada cas en la nostra matriu de dades (1, dona; 2, home) no té cap significat autènticament numèric: un home no és el mateix que dues dones (per molt masclista que sigui!).

Quan els identifiquem amb aquests valors només estem dient que cada individu **pertany a una classe**.

Els mesuraments depenen de les operacionalitzacions dels conceptes i aquestes del paper dels conceptes en les explicacions que es volen testar. Fins fa uns anys, la classificació diàdica (en dos valors: home i dona) semblava «natural» i «inevitable». Els caràcters sexuals primaris servien per classificar la població en dues categories d'acord amb la seva capacitat de tenir fills i alimentar-los. Avui dia, en molts entorns culturals i polítics es veu com una reducció artificial de la realitat del gènere i la classificació és com a mínim triàdica, que permet que hi hagi part de la població que no es vulgui classificar en cap de les categories anteriors. En una recerca que utilitzés el concepte psicosociològic de «femineïtat», no es conformaria en una classificació simple, ni tan sols triàdica, i podria utilitzar una bateria de preguntes per tal d'establir el grau o la intensitat de femineïtat dels individus de la mostra, de manera que la mesura es convertís en quantitativa.

En general, el nivell de mesurament no és una qüestió «natural» i «inevitable» del concepte mesurat. Depèn de quin és el grau de precisió requerida per respondre les preguntes de la nostra recerca, de la nostra capacitat per capturar la informació i dels recursos disponibles.

Els nivells de mesurament formen una **escala**. Cada nivell superior ens ofereix un grau d'informació més acurada respecte les variables que volem mesurar en els nostres casos.

2.1. Nivell de mesurament qualitatiu nominal

El **nivell de mesurament nominal** o **categòric** només ens permet classificar els casos en diferents grups. Aquest nivell de mesurament és el del gènere en la discussió anterior: dones i homes.

Per obtenir una classificació es requereixen regles sistemàtiques per distingir entre un fenomen d'un altre i establir criteris per determinar les fronteres entre les diferents classes o els diferents grups.

Per exemple, com hem anat comentant, el Codi Penal és una eina de classificació dels diferents tipus de delictes que ofereix aquestes regles sistemàtiques per distingir-los i marcar fronteres entre ells. Un altre exemple de variable nominal és la població de residència dels individus d'una mostra. En una recerca on aquesta pregunta fos important (perquè, per exemple, es volen conèixer els nivells de seguretat percebuda dels residents) és possible que s'hagin d'establir regles clares per resoldre possibles casos ambigus. Com s'ha de classificar algú que resideix entre setmana en una població i només el cap de setmana torna a la residència on està legalment registrat? Encara un altre exemple, un cas que ha entrat en el sistema judicial pot estar: arxivat, sobresegut, en tràmit, absolt, o condemnat.

Des del punt de vista logicomatemàtic, la classificació utilitza el **principi d'identitat**: els casos de la mateixa categoria són **idèntics** en relació amb la propietat que es mesura i són **diferents** dels casos que estan a les altres categories.

Cada membre de la categoria «dones» és idèntic en la seva «femineïtat» que tots els altres membres de la categoria o classe. De la mateixa manera, un resident del barri de Sant Roc, és idèntic en la seva residència a tots els altres residents del barri de Sant Roc i és diferent dels residents en els altres barris de la mostra.

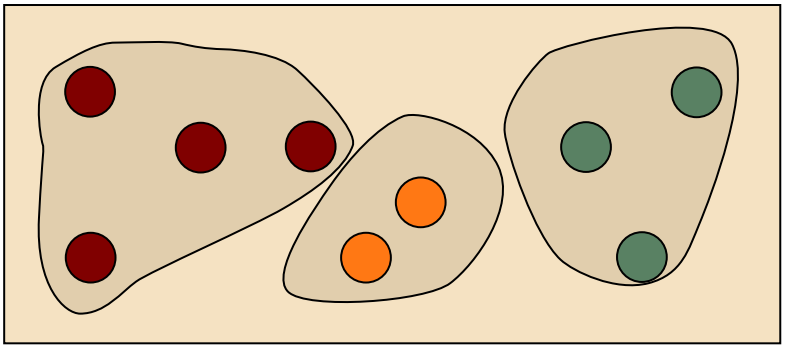
Formalment, si hi ha dues categories A i B: $A \neq B$

El conjunt de les categories en el nivell de mesura nominal han de complir dues propietats lògiques:

1) Les categories **han de ser mútuament excloents**. Això vol dir que no poden existir membres que pertanyin simultàniament a dues categories. Si ets home, no pots ser dona. Si ets resident de Sant Roc, no ets resident de Sant Joan. No hi pot haver encavallament entre les categories.

2) Les categories **han de ser col·lectivament exhaustives**. És a dir, tots els membres de la mostra han d'estar a alguna categoria, no n'hi pot haver cap que no estigui classificat en una categoria. Entre totes les categories s'engloben tots els casos de la mostra.

Figura 1



La **dicotomia**, que ens diu si un individu forma part d'una categoria o no en forma part, és el tipus originari de les classificacions més complexes. La variable binària de sexe (home o dona) la podem reduir a dues dicotomies (homes; no homes) o (dones; no dones). Lògicament, qualsevol classificació pot ser reduïda a un conjunt de dicotomies.

El nivell de mesurament nominal ens ofereix un coneixement limitat de molts fenòmens que volem estudiar. Això també implica una limitació en el tipus d'anàlisis estadístiques possibles amb aquesta informació. El nivell nominal es troba en el nivell inferior de l'escala de mesurament. Per sobre, però encara en el nivell qualitatiu, es troba el nivell ordinal.

2.2. Nivell de mesurament qualitatiu ordinal

En el nivell de mesurament ordinal s'afegeix un element al nivell nominal: existeix un ordre clar entre les categories. Cada element està només en una categoria i entre totes les categories engloben tots els elements (les categories són mútuament excloents i col·lectivament exhaustives). Però ara podem dir que hi ha una categoria que és més que totes les altres en la propietat que es mesura; després, n'hi ha una altra que és menys que la categoria anterior en aquesta propietat o variable, però amb més propietat que les altres... i així anar fent fins a completar totes les categories. Hi ha un ordre.

Si tenim quatre categories, podem ordenar-les d'acord amb un criteri de més/menys.

- Per exemple, els delictes del Codi Penal poden ser classificats per ordre de seriositat. En les recerques criminològiques existeixen moltes mesures que es recullen a les variables del nivell ordinal.
- Per exemple, el nivell d'educació es pot recollir com a: 1. «Cap educació formal», 2. «Educació primària», 3. «Educació secundària», 4. «Educació superior». A cada categoria hi ha un contacte més gran amb el sistema educatiu, però ho mesurem amb unes categories àmplies, on els valors numèrics atribuïts a cada categoria no indiquen una quantitat, sinó que simplement expressen un ordre.

- Típicament, moltes de les respostes a les preguntes de les enquestes d'opinió estan mesurades en el nivell ordinal (les anomenades escales de Likert). Per exemple, les opcions de resposta a la pregunta «En quina mesura està d'acord amb l'afirmació "el meu barri és segur?"»: 1. Molt en desacord; 2. En desacord; 3. No ho sé; 4. D'acord; 5. Molt en desacord. Pels valors de les categories només sabem que els que responen 1 estan menys d'acord amb l'afirmació que els de la categoria 2, però, a part d'això, no podem ser més precisos. No podem dir-ho amb una quantitat numèrica.

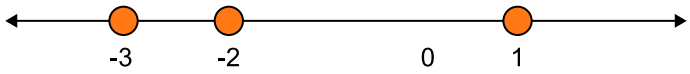
Si disposem de més informació sobre la presència d'una propietat en les unitats d'anàlisi podem pujar als nivells de mesurament quantitatiu de l'escala.

2.3. Nivell de mesurament quantitatiu d'interval

El **nivell de mesurament quantitatiu d'interval** s'utilitza quan no només som capaços de classificar els nostres casos (gent o esdeveniments), sinó que podem establir diferències de grau entre aquests casos respecte de la propietat que estem mesurant.

Una escala d'interval requereix que els intervals mesurats tinguin una mateixa unitat de mesura i que, per tant, la diferència en l'escala entre 2 i 1 sigui la mateixa diferència que entre 2 i 3; i la diferència entre 1 i 3 sigui dues vegades la diferència entre 1 i 2 o entre 2 i 3. No només diem que existeix un ordre. La distància lògica entre atributs o propietats pot ser expressada de forma significativa per intervals estàndards.

Figura 2



Les mesures d'interval que normalment s'utilitzen a les ciències socials consisteixen en mesures que provenen d'índexs compostos i estandarditzats, com el popular coeficient d'intel·ligència.

En criminologia, podríem obtenir aquest tipus d'índexs o escales –mesurades quantitativament a nivell d'interval– en mesures de perillositat o de risc de reincidència en joves. Es podrien basar en l'agregació de una certa quantitat d'ítems: respostes a qüestionaris que mesuren la impulsivitat juntament amb avaluacions dels factors de risc o vulnerabilitat social. El resultat final seria una mesura en la que sabríem amb precisió que, per exemple, els valors més elevats ens indiquen persones amb més risc de no assolir una rehabilitació i, per tant, es podran establir els protocols adients per tractar-los.

La diferència entre els mesuraments de nivell quantitatiu d'interval respecte del nivell de mesurament següent, el nivell quantitatiu de raó, és que, en el nivell d'interval no es coneix el nivell zero de presència de la propietat o atribut.

En l'exemple anterior de l'índex de factors de risc, sabem que cada cop que ens desplacem en l'índex en una unitat, estem avançant cap a un risc més gran o retrocedint cap a un risc més petit (normalment, en aquests índexs el valor zero és el valor més freqüent en la població). El que no sabem és quin valor té l'índex pel valor 0 de risc real.

Aquesta qüestió, que és important des del punt de vista conceptual, també té conseqüències en el tipus d'operacions matemàtiques amb sentit que es poden realitzar amb aquestes mesures. En aquests tipus de mesures només es poden fer sumes i restes, no multiplicacions ni divisions.

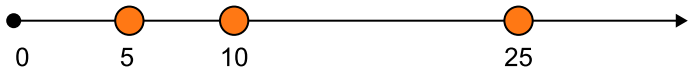
L'exemple més típic de les escales d'interval són les escales de temperatura en graus Fahrenheit o en graus Celsius. Totes dues escales són d'interval: no tenen el 0 en el mateix lloc, ni les distàncies entre graus són les mateixes. Però es pot passar de l'una a l'altra amb una fórmula matemàtica, una transformació lineal, clara: ºF = ºC 1,8 + 32. En les dues escales, sabem que si puja la temperatura 10º (entre 0º i 10º, posem) l'increment de temperatura haurà estat dels mateixos graus que si puja entre 10º i 20º. La suma i la resta en aquestes escales no és un problema. El problema apareix quan es vol multiplicar o dividir: 20 ºC de temperatura és el doble de temperatura que 10 ºC? No, perquè 0 ºC no és l'absència total de calor. En l'escala absoluta de temperatura kelvin, on el 0 sí que és l'absència de calor, els 0 ºC són 273 K i 10 ºC són 283 K; per tant, el doble de temperatura que 10 ºC serien 566 K, i no 20 ºC!

2.4. El nivell de mesurament quantitatiu de raó

El nivell de mesurament escala de raó es troba en el graó més alt de l'escala de mesurament.

El nom d'escala de raó significa que les distàncies entre les categories, no només tenen sentit en els intervals, sinó també en les **proporcions**: el valor de 10 és el doble de 5 i aquest és 5 vegades més petit que 25.

Figura 3



Moltes de les mesures de la criminologia estan mesurades amb el nivell quantitatiu de raó.

Per exemple, el nombre de vegades que un delinqüent ha estat arrestat prèviament, es mesura en l'escala quantitativa de raó. Si algú ha estat arrestat 3 vegades, ha estat 3 vegades menys arrestat que un altre que ha estat arrestat 9 vegades.

Aquesta comparació és possible perquè en aquesta variable el valor zero no és arbitrari. El nombre d'actes delictius en les enquestes d'autoinforme delictiu o el nombre de victimitzacions en les enquestes de victimització també són mesurades en aquest nivell. Altres mesures com l'edat mesurada en anys o la renda personal, també són variables quantitatives de raó.

De la mateixa manera, els valors de les variables que són l'agregat d'esdeveniments o de característiques dels individus que els componen en diferents unitats geogràfiques (districtes, municipis, comarques, províncies, regions, estats...) també es mesuren amb l'escala quantitativa de raó.

Per exemple, la taxa de criminalitat o la taxa de reincidència tenen aquest nivell de mesura.

Observació

Noteu que variables que en l'àmbit individual normalment són mesurades amb el nivell qualitatiu nominal (per exemple, l'ètnia o el gènere), en els agregats esdevenen variables d'escala de raó: per exemple, la taxa de masculinitat o el percentatge de població gitana. Això confirma, des d'un altre punt de vista, que el nivell de mesurament depèn de l'estratègia de recerca que triem; que el nivell de mesurament no és una cosa intrínseca de la natura del concepte que volem mesurar.

2.5. Implicacions del nivell de mesurament en el tipus de tractament

A l'hora de fer les anàlisis estadístiques, conèixer el nivell de mesurament de les variables és crucial per saber quins tipus d'anàlisis tenen sentit. Les tècniques específiques per descriure i per relacionar les variables depenen del seu nivell de mesurament.

Clarament, amb les variables de tipus nominal, calcular mesures de resum com la mitjana és totalment absurd. Però de vegades ens trobem en situacions on variables que realment estan mesurades amb el nivell ordinal són tractades com si fossin variables de tipus quantitatiu d'interval, especialment si hi ha molts nivells de l'escala ordinal. Aquest tipus de pràctica és més freqüent en les recerques més antigues, fetes quan encara no s'havien desenvolupat eines específiques per tractar variables de tipus qualitatiu, ja que aquestes van anar endarrerides respecte de les tècniques orientades a variables quantitatives o numèriques típiques de ciències socials primerenques i molt influents, com l'economia. Si es té en compte el que s'està fent i s'avisa el lector de l'anàlisi, una pràctica d'aquest estil pot ser acceptable. Naturalment, sempre és millor utilitzar les tècniques que són més adients per a cada nivell de mesurament.

D'una banda, en les variables de tipus numèric o quantitatiu es pot distingir entre variables que només poden tenir valors discrets (típicament en les variables de recompte) i altres variables que poden tenir valors continus. Aquesta qüestió no és molt rellevant, però pot provocar expressions curioses del tipus 'els fets delictius mitjans en una població són 3,7 per 1.000'. D'altra banda, s'ha de tenir molt clar que les variables numèriques amb valors discrets no han de ser confoses amb les mesurades amb el nivell ordinal.

En general, sempre és convenient en la recerca mesurar els conceptes en el nivell de mesurament més elevat. Si a l'hora de presentar o analitzar els resultats es veu que hem recollit massa informació, sempre serem a temps de disminuir el nivell de mesurament, agrupant els valors i convertint les variables numèriques en variables ordinals.

De vegades, la conveniència de disminuir el nivell de mesurament no és degut a la necessitat de simplificar sinó de la mateixa natura del concepte que volem mesurar.

Per exemple, la maduresa d'una persona podem mesurar-la amb els anys. Però si estem interessats en l'experiència segurament l'edat no serà una bona mesura (una mesura vàlida) o una mesura suficient. Potser és millor compondre l'edat amb l'experiència vital per classificar millor els individus. Per exemple, categoritzar com a **joves** aquells individus que no han superat els 30 anys i que encara no han treballat de forma remunerada; mentre que es classifiquen com a **adults** els individus que superin els 25 anys i que ja hagin treballat de forma remunerada. Aquí estaríem disminuint el nivell de mesurament per tenir més precisió a l'hora de capturar el concepte de maduresa.

3. La qualitat de les mesures: la fiabilitat i la validesa

3.1. La fiabilitat

Una mesura és **fiable** si produeix el mateix resultat quan el procés de mesurament es repeteix.

Exemple

Per exemple, si ens pesem cada matí abans de dutxar-nos i cada vegada tenim un pes força diferent –perquè resulta que tenim una bàscula no gaire bona i el pes varia segons l'equilibri dels dos peus al plat– no tindrem una mesura fiable del nostre pes (també podem dir que el nostre instrument de mesura, la bàscula, no és fiable).

Un altre exemple quotidià

Per què es recomana mesurar la febre dels nens petits amb termòmetre a l'anus més que no pas amb el termòmetre a la boca? Doncs, precisament perquè una mesura és més fiable que l'altra. No és segur que el nen mantingui quiet el termòmetre sota la llengua durant tot el temps de prendre la mesura.

A les ciències socials hi ha moltes fonts que entorpeixen la fiabilitat en el mesurament de les dades empíriques. Vegem-ne dos exemples amb dades de registre i amb dades d'enquestes.

En el camp dels registres oficials, en un registre policial dels delictes comesos hi pot haver diferents raons que porten a incorreccions. Hi pot haver errors en la codificació de les dades. Però també hi pot haver problemes de saturació de la feina en alguns moments, de forma que de vegades no quedin correctament registrats tots els delictes. És possible que un mateix delicte pugui ser classificat de forma diferent segons els criteris que utilitzi l'encarregat d'entrar les dades. El resultat és que la mesura dels delictes pot no ser fiable.

Per comprovar la fiabilitat de les mesures es fan servir diferents procediments.

- En el **procediment de test-retest**, repeteix el mesurament una segona vegada per veure fins a quin punt les mesures preses en un primer moment canvien quan es tornen a mesurar els mateixos individus.
- En la **prova de fiabilitat de divisió per la meitat** (Split-Half Check) es divideix la mostra i a cada part s'hi aplica un mesurament diferent d'un mateix concepte, les diferències entre els resultats obtinguts serviran de mesura de la fiabilitat d'aquestes mesures. Aquest mètode es va desenvolupar per mesurar la fiabilitat dels tests psicològics. Els tests es divideixen en dues parts, cada part es passa a una meitat de la mostra i es veu si els resultats dels tests tenen el mateix grau de correlació, indicant que estan mesurant una mateixa característica.

3.2. La validesa

Mentre que la fiabilitat fa referència a la precisió amb la que una mesura empírica tendeix a recollir un determinat concepte, la **validesa** es refereix a si la mesura realment mesura el concepte que es vol mesurar.

Hipotèticament, el problema d'una manca de fiabilitat el podríem resoldre realitzant molts mesuraments i fent-ne la mitjana. Naturalment, una mesura que no és fiable no és vàlida, perquè ens pot donar uns valors que es troben allunyats dels autèntics valors de la variable. Però al mateix temps, hi pot haver mesures que siguin fiables i no siguin vàlides. És a dir, hi pot haver mesures que tendeixen a donar-nos sempre els mateixos resultats (són fiables), però no són vàlides perquè no varien d'acord amb el concepte que es creu que mesuren. Per explicar aquesta relació entre validesa i fiabilitat es fa servir la imatge de la capacitat d'encertar una diana d'un tirador (la imatge ja va ser utilitzada en un comentari de Galtung sobre l'error aleatori de mesurament). Un tirador que no té pols és el mateix que una mesura que no és fiable: els seus trets estan dispersos perquè una tremolor o una tensió no controlada en el moment de disparar fan que el tret es desviï de la fita.

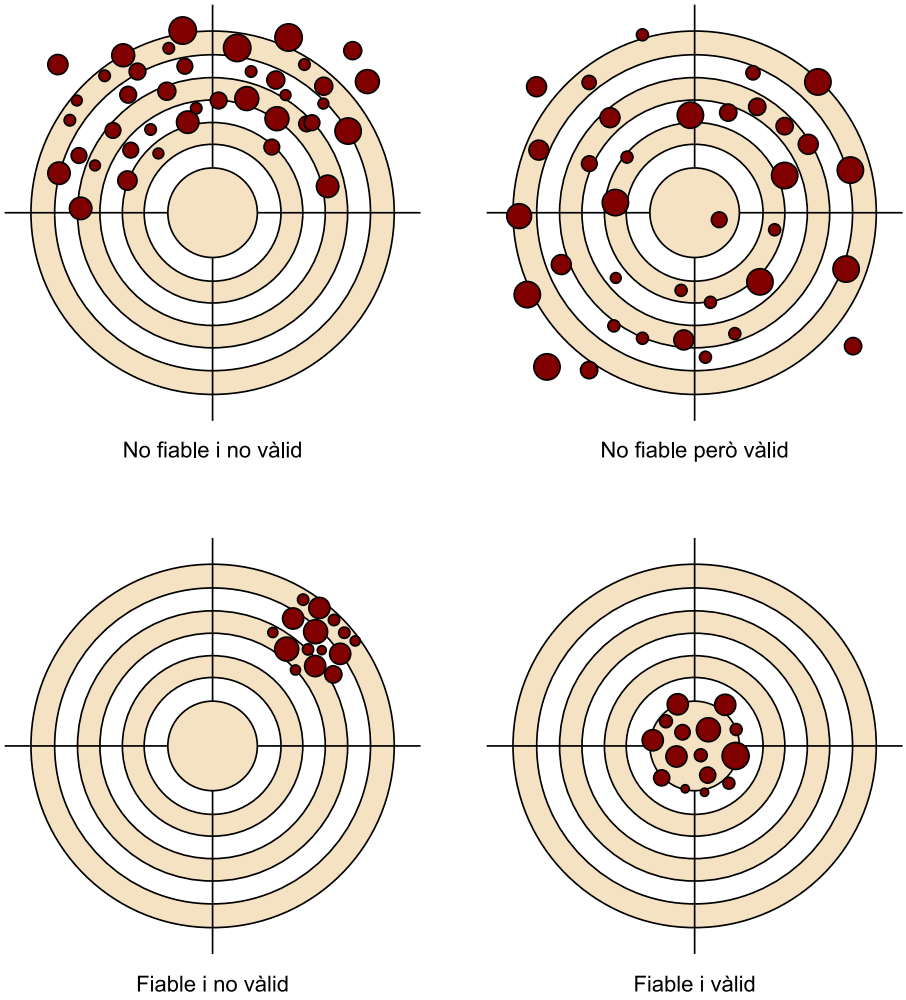
A la **figura 4** veiem a la primera filera els trets d'un tirador que no és fiable: en tots dos blancs els trets estan dispersos. En el blanc de la dreta, els trets es distribueixen al voltant de la diana. Els trets apuntaven l'objectiu –eren «vàlids», en aquest sentit–, però no eren fiables. Al blanc de l'esquerra el tirador tampoc no és fiable, però, el que és pitjor, no apunta correctament la diana (potser la mira de l'escopeta està desviada). A les imatges de sota tenim els resultats d'un bon tirador: és precís en totes dues dianes; però apunta correctament a la dreta i de forma desviada a l'esquerra.

Una **mesura vàlida** ha de tenir totes dues propietats: ser precisa (**fiable**) i mesurar el que se suposa que ha de mesurar (**vàlida**).

El tirador de la dreta de la primera filera no és un bon tirador, encara que apunti correctament la diana: una mesura que és vàlida però no és fiable, no és una bona mesura.

La distinció entre **error de mesura sistemàtic** i **error de mesura aleatori** serveix per explicar aquesta relació. Als gràfics de la primera columna tenim errors sistemàtics a l'hora de mesurar el concepte que ens interessa: en tots dos casos no estem apuntant al concepte que volem mesurar. El segon gràfic de la primera filera mostra només un error aleatori. Si podem mesurar un nombre suficientment gran de vegades, podrem descomptar aquest error.

Figura 4. Relació entre fiabilitat i validesa



L'existència d'**error aleatori** és l'origen de l'enfocament estadístic. A les ciències socials, l'error aleatori no només es relaciona amb l'error de mesura producte de l'instrument d'anàlisi (com passa a les ciències físiques), sinó amb la complexitat inherent dels fenòmens que estudien: hi ha molts factors que simultàniament afecten els valors de les diferents variables. L'**error sistemàtic** existeix quan alguns d'aquests factors tenen un impacte en la mesura que no és neutre. Naturalment, les qüestions dels errors són relatives als conceptes que volen ser mesurats (la diana del blanc).

Seguint l'exemple anterior, quan demanem que la gent s'ubiqui en una escala ideològica, és possible que no estiguem mesurant correctament les preferències ideològiques, és a dir, tenim un error sistemàtic, producte de la crispació política. Però, precisament per aquesta raó, aquesta mesura podria convertir-se en una bona mesura de la crispació política!

La relació entre el concepte teòric i la mesura se suposa teòricament. La mesura, per ella mateixa, no és vàlida o invàlida. Ho és per la interpretació que en fa l'investigador.

Un exemple de problemes de validesa d'una mesura és el d'usar el nombre de denúncies de maltractament de gènere per mesurar el nombre de maltractaments. A partir del moment en què el tema del maltractament va ser reconegut públicament com una xacra, es va observar que el nombre de denúncies va créixer, en part per les facilitats per fer aquesta denuncia (això és el que cercaven aquest tipus de mesures que volien ajudar a desvetllar la violència que passava desapercebuda i servir per poder posar mesures preventives). Des d'aquest punt de vista, ara es disposava d'una millor mesura de la violència

de gènere. El problema va ser que, tal com va declarar la fiscal en cap de Catalunya, un cop reconegut el delicte de maltractament, aquest es va passar a argumentar amb molta més freqüència en els casos de divorci perquè d'aquesta manera l'advocat que l'esgrimia tenia un avantatge davant del cònjuge. El resultat és que, per aquesta interferència, ja no estàvem segurs de tenir una millor mesura de la violència de gènere.

Per evitar incórrer en problemes de validesa sempre s'ha d'estar alerta de la possibilitat que aquests s'estiguin produint, tant en la construcció i utilització dels instruments de mesura de les variables com a l'hora d'interpretar les mesures. S'ha de pensar contínuament de quina manera alguns factors poden interferir en la mesura del concepte teòric: com podria ser invalidada una mesura?

Però, a més d'aquest consell a la precaució, com es pot comprovar la validesa de les mesures? En alguns camps es fan estudis de validesa de forma sistemàtica.

- Per exemple, en el camp de la salut es fan estudis de validació dels diagnòstics realitzats en el sistema de salut. Es realitzen autòpsies a una mostra de pacients d'hospitals per determinar les causes reals de les morts i les comparen amb com els havien diagnosticat. Així és possible comprovar si funcionen correctament les noves eines de diagnosi o, també, és possible mesurar quin és el percentatge de casos que van ser tractats de forma incorrecta.
- En els estudis postelectorals fets als Estats Units i al Regne Unit s'han fet comprovacions de validesa de la variable de participació. Com que els registres de participació són públics, es comprova si la gent que ha estat enquestada ha fet el ha dit que va fer: votar o no votar. D'aquesta manera es pot tenir un perfil dels grups de ciutadans que més menteixen a l'hora de dir si han votat o no (entre un 23 i un 28% dels que diuen que han votat a les enquestes, en realitat, no ho han fet!).
- De forma semblant, en les enquestes preelectorals les respostes són reanalitzades un cop que s'han fet les eleccions per veure quins grups de votants tendeixen a amagar la seva intenció de vot (el que s'anomena popularment el vot ocult). Amb aquesta informació les empreses d'opinió poden millorar les seves prediccions en les posteriors eleccions. Naturalment, aquests mètodes d'ajustament són vàlids mentre l'entorn polític sigui estable, en el moment en el que canvia, totes aquestes referències prèvies i correccions de les mesures es poden tornar en contra.

Aquests són exemples de **validació externa** de les mesures de les variables. En la validació externa s'intenta que alguna prova del món real garanteixi que s'està mesurant correctament el concepte que volem mesurar. Si comprovem el grau en el que la nostra variable ens permet obtenir mesures del món real, podrem anomenar aquesta prova de validesa com a validesa pragmàtica o predictiva. No sempre es troben mesures externes de control de validesa. Una solució és cercar altres mesures que se suposa que estan relacionades amb el mateix concepte que volem mesurar i així es veu si les dues mesures estan correlacionades. El fet que totes dues estiguin correlacionades, és una pista per suposar que la mesura que hem fet és una mesura correcta del concepte que intentem mesurar. Aquest tipus de prova de validesa pot ser anomenada validesa interna o construïda i serveix de base per a la construcció de mesures de conceptes amb indicadors múltiples.

Per exemple, podem tenir tres mesures diferents de l'existència de corrupció política en un país: la primera, la percepció de corrupció entre els treballadors públics segons els homes de negocis (aquesta és la base de l'índex de Transparency International); la segona, pot ser el nombre de casos de corrupció destapats pels mitjans de comunicació; i, finalment, la tercera mesura pot ser la discrepància entre els costps pressupostats i els costos efectivament realitzats en les obres públiques. Totes tres mesures se suposa que estan relacionades amb el nivell de corrupció política en un país, però totes tres mesuren el concepte des de punts de vista i amb mètodes molt diferents. Per tant, no esperem que tinguin exactament els mateixos valors, tot i que sí que esperem que, en tant que mesu-

ren el mateix concepte de corrupció, estiguin mínimament correlacionades. Si les tres mesures anessin en conjunt en un mateix sentit (hi hagués prou correlació entre elles), tindríem més confiança en què estem mesurant un concepte difícil de forma correcta. Fins i tot, podríem plantejar-nos construir un índex que agregués i resumís les diferents mesures en una d'única. Si, en canvi, alguna d'aquestes mesures fos molt discrepant, ens veuríem obligats a revisar-la i fins i tot descartar-la, perquè **no és** una **mesura vàlida** del concepte.

Finalment, la mesura més general de validesa de les variables és la **validesa aparent** (*face validity*), que és un terme per dir que una mesura és correcta perquè ens ho sembla com a investigadors especialitzats en un fenomen.

Veiem els valors de la mesura i els casos als quals corresponen i ens sembla que són correctes. Naturalment, no es tracta d'una prova molt forta, ja que només val en tant que els lectors hi estiguin d'acord.

Per exemple, quant a les puntuacions dels índexs de Transparency International veiem que Itàlia o Espanya queden per sota dels països escandinaus en percepció de corrupció, però per sobre de Nigèria o Mèxic; com que ens sembla que aquesta ordenació és versemblant, podem dir que l'índex és **aparentment vàlid**.

Els problemes potencials de validesa i de fiabilitat de les nostres mesures han d'aparèixer en qualsevol recerca.

4. L'anàlisi exploratòria de les dades

4.1. La descripció univariant de les dades

Abans de mesurar les relacions entre les variables, és convenient fer una descripció univariant de les variables que s'utilitzen a la recerca. Les descripcions ens permeten entendre quina és l'estructura de les variables i ens serviran per entendre que es produeixin algunes relacions entre diferents variables. Per altra banda, les descripcions univariants són molt importants com a control de la informació de la nostra base de dades, ja que ens permeten detectar possibles errors en les codificacions de les variables.

Per a l'investigador sempre és crucial visualitzar les distribucions univariants de les seves variables.

L'estadística ens proporciona eines per **resumir** la informació de les distribucions. Serveix per veure els boscos (característiques generals del conjunt de dades) que hi ha darrera dels arbres (els casos individuals). Serveix per a evitar que els casos que cridin més l'atenció –per la raó que sigui– influeixin desproporcionadament en la valoració del grup. És un control per a evitar equivocar-nos i arribar a conclusions errònies en relació a una informació numèrica nombrosa.

L'enfocament exploratori de dades posa l'èmfasi en conèixer el màxim possible de coses a través de les dades. Com més se sàpiga de les dades, més útils seran per desenvolupar la teoria. Dos principis han de regir l'exploració de les dades:

1) Primer, s'ha de ser escèptic cap a les mesures de resum. Cap d'elles pot sintetitzar completament la informació continguda a les variables.

2) En segon lloc, s'ha de ser obert de mires respecte del que ens poden ensenyar les dades. No hem de donar per suposat que coneixem com són les dades. Moltes vegades poden descobrir coses inesperades, noves, respecte de les característiques o explicacions dels fenòmens. Els anglosaxons d'aquesta capacitat de fer descobriments per atzar, inesperadament, en diuen *serendipity* i és una qualitat important en la recerca científica tenir aquesta disposició a estar obert davant de les noves idees o els factors que poden ajudar a explicar els fenòmens d'interès.

Quan es descriuen estadísticament les variables, hi ha tres qüestions que ens interessen per sintetitzar les distribucions:

- Quins són els valors més característics on es troben majoritàriament els casos?
- Com són de característics o representatius aquests valors?
- Quina és la forma de la distribució de les dades?

1) En termes estadístics, el **resum** d'una variable és la mesura de **centralitat** o de localització al voltant de la que es troben els casos. És el valor més representatiu de la distribució de dades d'una variable.

2) La **dispersió** serveix per respondre a la pregunta sobre fins a quin punt els valors centrals són representatius. Si la majoria dels casos no són com els centrals, si no es troben agrupats i més aviat estan repartit entre tots els valors que pot adquirir la variable, els estadístics de centralitat ens diran poca cosa de la distribució que volem conèixer.

3) Finalment, tenim la **forma** de les distribucions. Com es distribueixen els valors al voltant dels valors centrals? Es tracta d'una distribució simètrica o asimètrica? Es tracta d'una distribució més aviat uniforme i rectangular o hi ha una moda? Hi ha més d'una moda?

Els gràfics

Els **gràfics** ens ofereixen una **descripció més completa de les distribucions** que no pas els estadístics de centralitat i dispersió. Per extreure el màxim d'informació de les variables, les anàlisis exploratòries posen èmfasi en les representacions gràfiques. Ens donen simultàniament una aproximació als tres aspectes importants d'una distribució: forma, dispersió i localització. Sembla un tòpic el proverbi xinès que diu que una imatge val més que deu mil paraules, però en l'anàlisi exploratòria de dades representar gràficament les dades és prescriptiu.

Els estadístics per resumir les variables estan associats al nivell de mesurament d'aquestes variables. A la **taula 2** es llisten els estadístics de centralitat i de dispersió i els gràfics adients per als diferents nivells de mesurament.

Taula 2. Nivell de mesura de les variables i descriptius de les distribucions

Nivell mesura variable	Descripció	Centralitat	dispersió	Gràfics
• Qualitativa • Escala nominal	• Valors no numèrics • Absència d'ordre	• Moda (Mo)	• Raó de variació • Núm. de categories • Índex Hirsch-Herfindahl	• Diagrama de barres • Gràfic de sectors
• Qualitativa • Escala ordinal	• Valors no numèrics • Presència d'ordre	• Mediana (Med) • Moda (Mo)	• Rang (mínim-max) • Quartils • Percentils • Rang interquartílic	• Diagrama de barres • Box-plot (caixa)

Nivell mesura variable	Descripció	Centralitat	dispersió	Gràfics
• Quantitativa • Escala interval i raó	• Quantitativa discreta (quantitat finita o numerable de valors numèrics) • Quantitativa contínua (qualsevol valor numèric en l'interval) • Escala d'interval (només té sentit diferència entre valors) • Escala raó (a més de diferència, raó entre valors)	• Mitjana (μ) • Mediana (Med) • Moda (Mo)	• Rang • Quartils • Percentils • Rang interquartílic • Desviació típica • Variança • Coeficient de variació	• Histograma • Box-plot (caixa) • Gràfic de tronc-i-fulles • (Stem-and-leaf)

4.1.1. La matriu de dades

El punt de partida en tota anàlisi estadística és la **matriu de dades** en la qual tenim en les fileres la llista de les nostres **unitats d'anàlisi** (els casos, les observacions, els elements o els individus) sobre els quals hem mesurat les nostres variables.

Taula 3. Matriu de dades

Observacions / Variables	Variable 1	Variable 2	...	Variable *M*
Obs. 1				
Obs. 2				
...				
Obs. *N*				

Per exemple, a la **taula 4** es representa una matriu de dades en la qual les unitats d'anàlisi són els districtes de Barcelona. Aquesta és una taula molt petita, però ens serveix d'il·lustració a l'hora de fer l'anàlisi descriptiva. En les fileres tenim els districtes. En les columnes tenim llistades algunes variables relatives, per una banda, a l'enquesta de victimització de la ciutat de l'any 2017 (l'índex de victimització, l'índex de denúncia, la percepció de la seguretat a la ciutat i la percepció de la seguretat al barri); per altra banda, tenim dades del sistema de seguretat pública de Barcelona amb les denúncies per infracció de l'ordenança de convivència ciutadana, els incidents per degradació de l'espai públic, els incidents en la convivència veïnal, els incidents per activitats molestes a l'espai públic i, finalment, els incidents per activitats indegudes a l'espai públic. Les mesures del sistema de seguretat són de l'any 2018 i totes estan estandarditzades en tant per mil habitants en el districte (font: Ajuntament de Barcelona. Gerència de Seguretat i Prevenció).

Taula 4. Matriu de dades de seguretat de l'Ajuntament de Barcelona

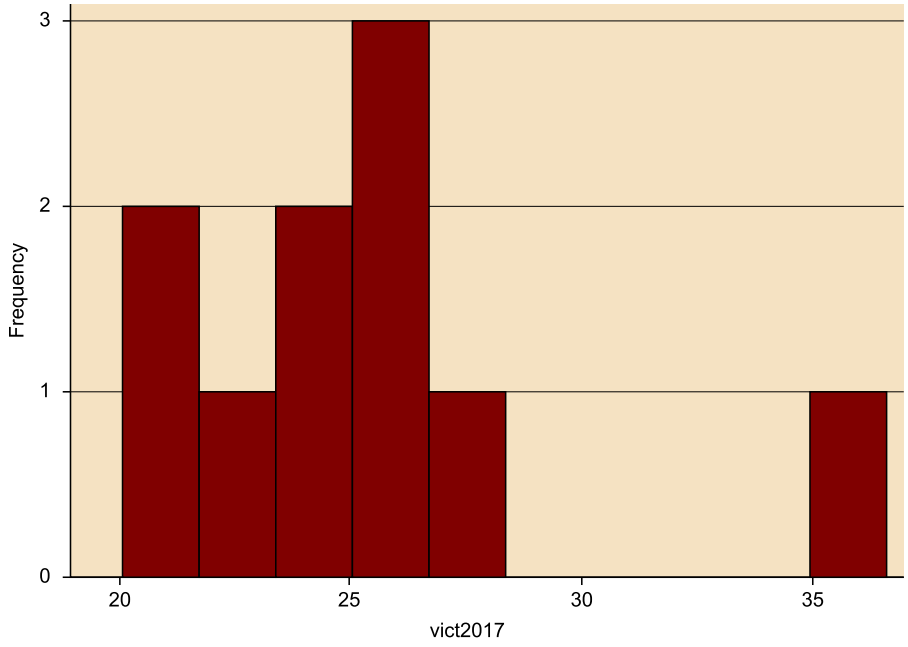
	victim	denun	segciut	segbarri	denunciesOC	incdegrEsp	IncidConv	Molest	ActIndeg
1. Ciutat Vella	36,6	15,1	6,2	5,2	46,15	4,23	56,64	110,53	18,18

	victim	denun	segciut	segbarri	denunciesOC	incdegrEsp	IncidConv	Molest	ActIndeg
2. Eixample	28,2	17,4	6,3	6,9	2,08	1,90	26,39	24,45	9,64
3. Sants-Montjuïc	24,7	22,6	6,1	6,2	2,31	1,95	22,80	27,05	3,90
4. Les Corts	20,1	23,4	6,4	7,3	0,59	1,41	13,58	16,85	2,36
5. Sarrià-Sant Gervasi	25,7	34,6	6,0	6,8	1,47	1,52	18,75	16,84	1,70
6. Gràcia	22,7	22,4	6,3	7,0	4,29	1,97	25,09	25,96	2,24
7. Horta-Guinardó	21,2	24,1	6,2	6,3	0,27	1,23	17,74	13,00	1,51
8. Nou Barris	24,1	21,9	6,2	5,9	0,29	1,68	19,70	20,77	2,22
9. Sant Andreu	26,1	29,4	6,2	6,3	0,83	1,91	15,61	17,39	1,68
10. Sant Martí	25,8	21,8	6,2	6,1	5,92	1,90	18,05	25,23	5,13

4.1.2. Les mesures de centralitat

Les mesures de centralitat més conegudes són la moda (Mo) que és el valor més freqüent d'una distribució. En el cas dels districtes de Barcelona respecte de l'índex de victimització de l'enquesta de victimització del 2017 (victim), no hi ha un valor que es repeteixi més d'una vegada.

Figura 5. Histograma índex victimització BCN 2017



Però amb l'ajuda de l'histograma de la **figura 5** es pot veure com hi ha un pic amb 3 districtes amb un índex de victimització en el rang 25-26.

La **mitjana** aritmètica (μ) és la suma de tots els valors de la distribució dividida pel nombre de casos o unitats de mesura.

$$\mu_x = \frac{(x_1 + x_2 + \ldots + x_n)}{n} = \frac{\sum x_i}{n}$$

En l'índex de victimització de l'any 2017 (victim), la mitjana és de 25,52.

La **mediana** és el valor que, si ordenem tots els casos de més petit a més gran, té el 50% dels casos amb valors superiors i el 50% dels casos amb valors inferiors. Com que tenim 10 districtes, el valor mitjà de la distribució no és un dels valors d'un dels districtes, sinó el punt mig del rang entre dos districtes Sarrià-Sant Gervasi i Sants-Montjuïc: 25.2.

La mediana es troba a la vora de la mitjana, perquè es tracta d'una distribució relativament simètrica. La mitjana tendeix a moure's cap a la dreta, per l'efecte del districte de Ciutat Vella, que té un índex de victimització extraordinàriament elevat. En general, la mediana és una mesura de centralitat més «**robusta**» que no pas la mitjana, que es veu molt afectada per l'existència de casos amb valors molt diferents de la resta de la distribució. En aquest cas, Ciutat Vella.

4.1.3. Les mesures de dispersió

Existeixen molts estadístics de dispersió de les variables mesurades numèricament.

El **rang** mesura la distància entre el valor màxim i el mínim.

L'índex de victimització de l'any 2017 (victim):

Rang = Max – Min = 36.6 – 20.1 = 16.5

El **rang interquartílic** representa la distància entre el primer quartil (el valor que té el 25% dels casos amb valors inferiors i el 75% dels casos amb valors superiors) i el tercer quartil (el valor que té el 75% dels casos amb valors inferiors i el 25% dels casos amb valors superiors):

Rang interquartílic:

valor $Q_3 – Q_1 = 26.1 – 26.1 = 3.4$

La **variança** és la suma ponderada pel total de casos de les diferències al quadrat entre els valors que adquireix la variable i la seva mitjana. La base és la diferència entre la mitjana i cada valor, per evitar que, quan sumem totes aquestes diferències que algunes vegades són positives i altres negatives (la mitjana

és un valor que es troba al mig de la distribució), s'elevi al quadrat la diferència. Això fa que tots els valors ara siguin positius i es puguin sumar per tenir una estimació de la dispersió dels casos al voltant de la mitjana:

Variança:

$$S^2 = \frac{\sum(\mu_x - x_i)^2}{n} = 21.01$$

La **desviació típica** és l'estadístic de dispersió més popular. És l'arrel quadrada de la variança. La variança ens proporciona una bona mesura de la dispersió, però té el problema que la seva unitat de mesura és el quadrat del de la variable. Això fa que no pugui ser interpretat en l'escala de la distribució dels casos. La desviació típica resol aquest problema amb l'arrel quadrada.

Desviació típica:

$$S = \left(\frac{\sum(\mu_x - x_i)^2}{n}\right)^{\frac{1}{2}} = 4.58$$

La desviació típica de l'índex de victimització en els districtes de Barcelona és 4.58, que significa una bona proporció dels casos.

La desviació típica és molt convenient quan volem comparar distribucions d'una mateixa variable entre grups o poblacions diferents. Podem veure si, per exemple, la distribució de les qualificacions de dues classes són molt diferents, tot i que a totes dues classes la mitjana sigui la mateixa. Però si volem comparar les distribucions de variables diferents la desviació típica no és adequada. El **coeficient de variació** resol el problema dividint la desviació típica per la mitjana. És una mesura de dispersió adimensional: sense unitats de mesura, de manera que permet comparar la dispersió de qualsevol tipus de distribució.

Coeficient de variació:

$$CV = \frac{s}{\mu_x} = 0.18$$

4.1.4. Mesures de forma

Existeixen el coeficient d'asimetria i el coeficient d'apuntament que comparen les distribucions amb la distribució normal.

La mesura més simple i útil de l'asimetria de les variables és la diferència entre la mitjana i la mediana dividida per la desviació típica.

Asimetria:

$$A = \frac{\mu_x - Med}{s}$$

4.1.5. Les taules de freqüència i les representacions gràfiques

Una forma de representar el conjunt de les dades d'una variable és a través de la **taula de freqüències**.

La **taula de freqüències** llista les freqüències de cada valor de la variable.

Aquesta solució és convenient per a les variables mesurades qualitativament, que acostumen a tenir un nombre limitat de valors o categories; però per a les variables de tipus quantitatiu (especialment si la variable és contínua), cal fer una agrupació de les categories en rangs.

La **taula 5** mostra la taula de freqüència de la variable comunitat autònoma de l'estudi de les sentències per delictes sexuals a menors en les audiències provincials espanyoles entre els anys 2014 i 2016.

Taula 5. Taula de freqüència comunitat autònoma. Sentències per delictes sexuals a menors 2014-2016

Comunitat autònoma	Freq.	Percent	Cum.
Andalusia	378	16.12	16.12
Aragó	63	2.69	18.81
Astúries	41	1.75	20.55
Balears	117	4.99	25.54
Canàries	172	7.33	32.88
Cantàbria	37	1.58	34.46
Castella-Lleó	78	3.33	37.78
Castella-Manxa	106	4.52	42.3
Catalunya	330	14.07	56.38
València	260	11.09	67.46
Extremadura	93	3.97	71.43
Galícia	102	4.35	75.78
Madrid	287	12.24	88.02
Múrcia	116	4.95	92.96
Navarra	35	1.49	94.46

Comunitat autònoma	Freq.	Percent	Cum.
País Basc	108	4.61	99.06
La Rioja	11	0.47	99.53
Ceuta	2	0.09	99.62
Melilla	9	0.38	100

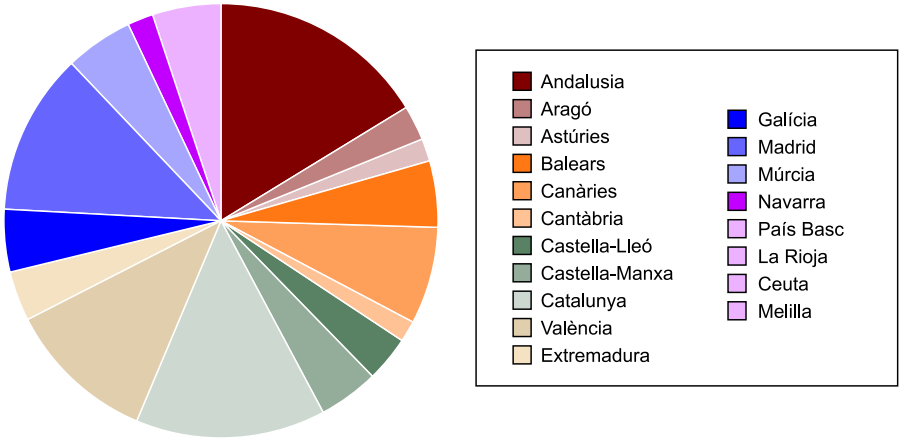
Els **diagrames de barres** són convenients per les variables qualitatives ordinals o categòriques. La **taula 6** és una taula de freqüència que reprodueix la informació de la variable comunitat autònoma (variable mesurada a **nivell nominal**) de la recerca sobre la sentències per abusos sexuals que apareix a la **taula 5**. En aquesta taula s'hi ha afegit un diagrama de barres i les categories s'han ordenat en ordre descendent de grandària. D'aquesta forma, la informació continguda es pot veure fàcilment d'un cop d'ull.

Taula 6. Taula de freqüències amb diagrama de barres. Sentències per delictes sexuals a menors 2014-2016

Comunitat aut	Freq.	
Andalusia	378	**
Catalunya	330	**
Madrid	287	**************************************
València	260	**********************************
Canàries	172	**********************
Balears	117	***************
Múrcia	116	***************
País Basc	108	**************
Castella-Manxa	106	**************
Galícia	102	*************
Extremadura	93	************
Castella-Lleó	78	**********
Aragó	63	********
Astúries	41	*****
Cantàbria	37	*****
Navarra	35	*****
La Rioja	11	*
Melilla	9	*
Ceuta	2	

La **figura 6** és un **diagrama de sectors** que presenta alternativament la informació de les taules 4 i 5. Aquest tipus de gràfic visualment és atractiu i, per aquesta raó, és utilitzat freqüentment en els reportatges periodístics. Però com es pot apreciar a la figura 6, moltes vegades **no és la millor forma de presentar acuradament la informació**. La figura 6 ens proporciona una visió de la fragmentació de les dades en les 17 comunitats, però **és confusa** i no s'aproxima a la qualitat de la representació de la taula 6 que combina taula de freqüència i diagrama de barres ordenades per freqüència de les categories.

Figura 6. Diagrama de sectors. Sentències sexuals a menors



El bon gràfic

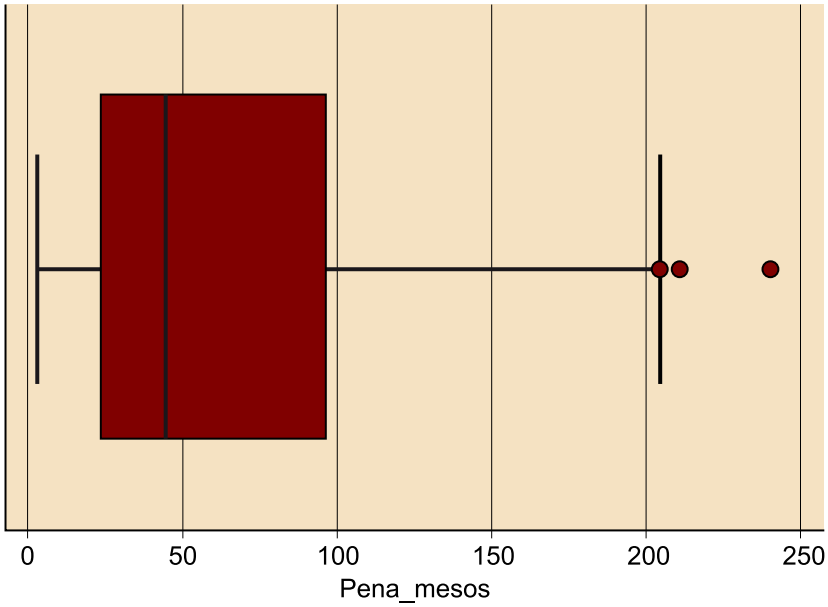
En general, en les **representacions gràfiques** intentarem de ser el més clars i precisos possible. D'aquesta manera, millorarem en l'eficàcia de la comunicació de la informació. El millor gràfic és el que és capaç de transmetre la informació rellevant en el mínim temps, espai i tinta. Els gràfics eficients són una mostra de respecte per als lectors de les recerques. El bon gràfic ha de mostrar les dades i ha d'induir a pensar en el que és substancial. S'ha d'evitar fer representacions que indueixin a una percepció distorsionada de les dades. El bon gràfic representa molta informació en poc espai i proporciona coherència a un gran nombre de dades. També pot presentar les dades a diferents nivells de detall.

Els diagrames de sectors estan **completament contraindicats** per representar les **variables de tipus ordinal** perquè fan perdre la noció d'ordre en les categories. En les cròniques periodístiques de les enquestes d'opinió, on una bona part de les preguntes són escales d'actituds ordinals, per millorar en la «varietat» en la presentació i així evitar «l'avorriment» del lector, es tendeixen a utilitzar els diagrames sectors, que amaguen alguna informació més rellevant de les variables: la distribució de l'opinió entre els que afavoreixen o no una alternativa.

Per a les variables ordinals, la representació gràfica preferida hauria de ser el **diagrama de barres**, tot i que, tal com recull la taula 2, també es poden utilitzar els diagrames de caixa o box-plot. El **box-plot** és una representació gràfica de la localització del primer, segon i tercer quartils d'una variable i del seu màxim i mínim. Aquest tipus de representació de les variables és molt **compacte** i és molt útil quan volem **comparar distribucions** de diferents variables. El box-plot no només s'utilitza amb les variables mesurades a nivell ordinal, sinó que també és molt recomanable per representar les variables numèriques.

A la base de dades de les sentències per delictes sexuals a menors d'edat, tenim la **variable numèrica** de les **penes de presó** mesurades en mesos, a les que han estat condemnats els ofensors que han estat trobat culpables.

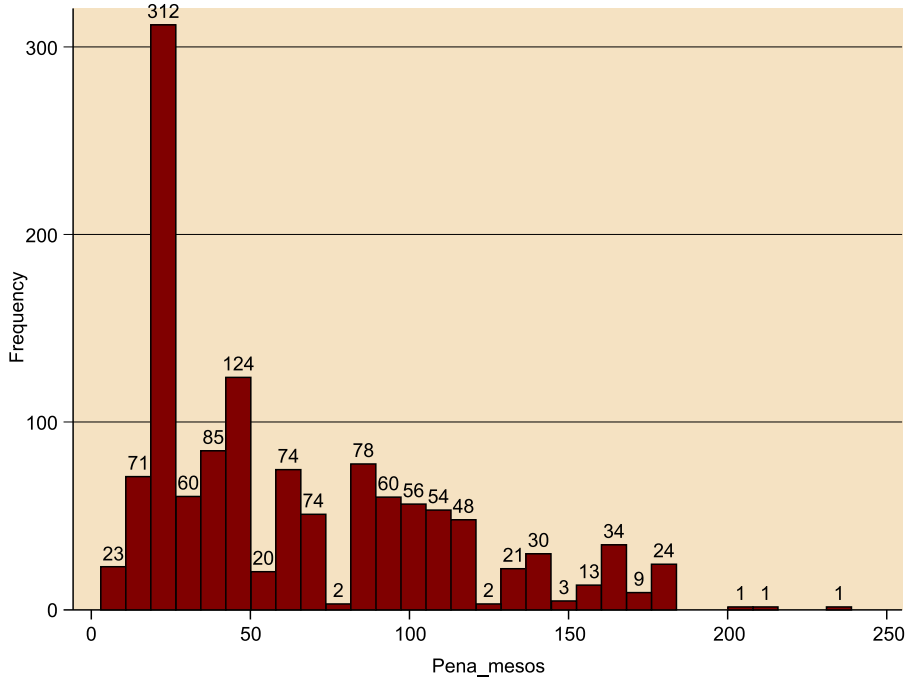
Figura 7. Box-plot pena de presó (mesos)



A la figura 7 es representa el box-plot de la variable pena de presó. La caixa central del gràfic delimitada pel primer quartil (24 mesos) i el tercer quartil (96 mesos), conté el 50% dels casos. La línia vertical que divideix la caixa central correspon a la situació de la mediana (48 mesos, que és el segon quartil). Per sobre de la mediana, tenim el 50% dels casos i, per sota seu, tenim l'altre 50% dels casos. Les línies primes verticals que estan connectades amb la caixa per una línia horitzontal als dos costats de la caixa central (de vegades anomenades bigotis) marquen el límit dels casos que tenen per sota (3 mesos) i per sobre (210 mesos), el 5% dels casos. Els punts que apareixen a la dreta són els casos extrems, els *outliers*.

Si sabem llegir el box-plot és molt fàcil adonar-se que es tracta d'una distribució molt asimètrica, ja que la majoria dels valors es troben agrupats en valors baixos de la distribució. Naturalment, la representació del box-plot és esquemàtica (i aquest és l'avantatge quan volem comparar les distribucions de moltes variables), però no ens permet tenir el nivell de detall en la representació d'una variable de tipus quantitatiu que ens permet l'histograma (ja n'hem vist un a la figura 5).

Figura 8. Histograma pena de presó (mesos)



La figura 8 és l'histograma de la mateixa variable (pena de presó en mesos) que el box-plot de la figura 7. Les barres de l'histograma recullen agrupaments de valors de la variable i han estat etiquetades pel nombre de casos que contenen. Des d'aquest punt de vista, la informació és comparable a una taula de freqüència. Quan comparem la representació de l'histograma amb el box-plot podem comprovar fins a quin punt ens ofereix més detall i complexitat.

4.2. L'anàlisi estadística de dues variables

Per comprovar les relacions entre les variables hem de tenir en compte el seu nivell de mesura, de la mateixa manera que a les mesures descriptives univariants.

4.2.1. Diferència entre mitjanes i ANOVA

La relació entre dues variables més fàcil d'imaginar és la que s'estableix entre una **variable independent categòrica** i una **variable dependent numèrica**. Aquesta situació pot correspondre a la situació d'un experiment en el qual la variable independent (categòrica dicotòmica –amb dos valors possibles–) indica que s'ha rebut el tractament o no.

Per exemple, la variable dependent numèrica és el creixement de les plantes d'una parcel·la i la variable independent (o variable de tractament) és haver rebut una dosi de fertilitzant o no. Si el fertilitzant té efecte (és a dir, si la variable independent té relació amb la variable dependent) esperem que les plantes que han rebut el tractament tinguin un creixement mitjà superior. Per comprovar que el tractament té efecte, haurem de mesurar el creixement mitjà a les parcel·les sense tractament i el creixement mitjà a les parcel·les amb tractament. La **diferència de les mitjanes** dels dos grups de parcel·les serà l'efecte del tractament. No totes les parcel·les són iguals, n'hi ha de més assolellades que altres, n'hi ha amb terra més ric que altres però, si en fem moltes i les distribuïm ale-

atòriament entre el grup de tractament i de control, podrem establir sense dubtes l'efecte mitjà del tractament sobre el creixement de les plantes.

Si estem treballant amb una mostra aleatòria i volem inferir si la diferència entre les mitjanes que hem observat existeixen en la població, es pot fer un test de significativitat estadística. L'estadístic de la diferència entre les mitjanes segueix una distribució t-Student que és fàcil de calcular i que està incorporada en els paquets estadístics més bàsics.

Quan tenim la variable independent categòrica que té valors múltiples (és multinomial; per exemple, la confessió religiosa), es poden analitzar les diferències entre les diferents categories i la variable dependent numèrica. Si estem treballant amb una mostra aleatòria i volem fer el test de significativitat, haurem de fer una anàlisi de la variança (**ANOVA**, en les sigles en anglès). L'estadístic F, que relaciona la variació dels valors dels individus dins de les categories amb la variació total dels individus, permet establir els valors crítics a partir dels quals es pot dir amb un cert grau de certesa (nivell de confiança) que les dues variables estan relacionades.

4.2.2. Relacions entre variables numèriques: el gràfic de dispersió

Quan la variable independent i la variable dependent son numèriques, la millor manera d'establir si hi ha una relació és representar-les en un **gràfic de dispersió**. Cada observació del nostre estudi s'identifica per una coordenada cartesiana i pot ser representada en el pla. L'existència de relacions s'observa perquè hi ha patrons definits en la situació dels punts.

Figura 9. Absència de relació i relació lineal positiva amb poca dispersió

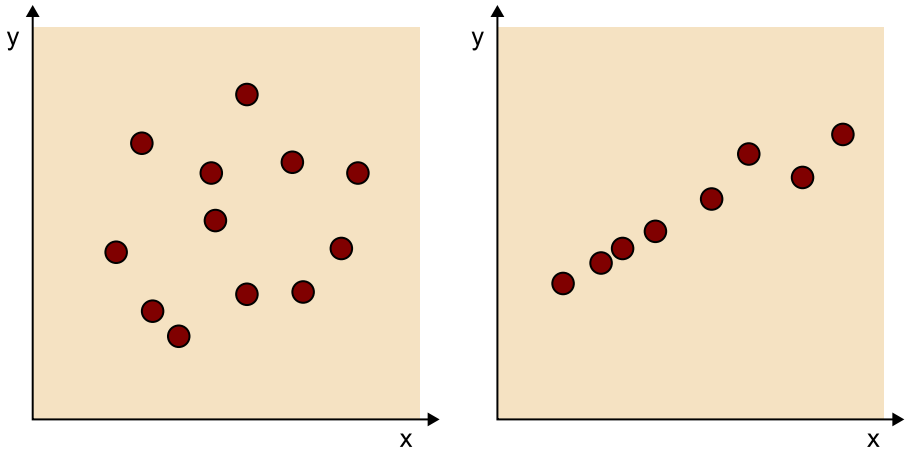
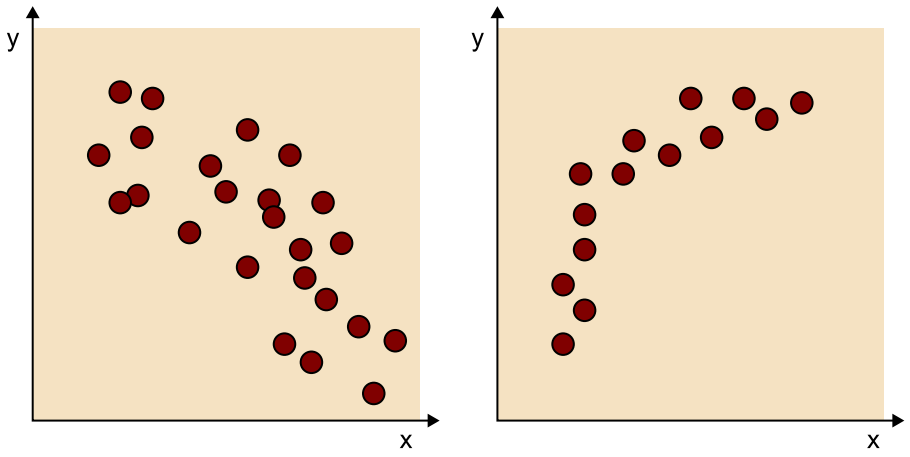


Figura 10. Relació lineal negativa amb dispersió i relació curvilínia



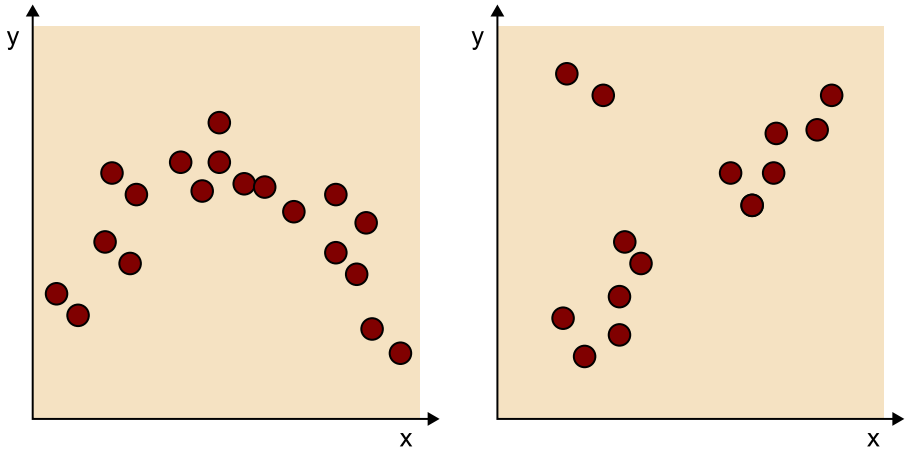
Als sis panells de les **figures 9**, **10** i **11**, podem observar diferents tipus de relacions entre variables numèriques. En el **primer panell** (**figura 9** esquerra) es veu que els casos es reparteixen sense ordre en el pla. A valors grans a l'eix de la x podem trobar valors alts o baixos de l'eix de la y i, a l'inrevés, a valors baixos de la x podem trobar qualsevol tipus de valor a l'eix de la y. No coneixem millor el valor de la y d'un cas quan en coneixem el seu valor a x. En canvi, en el **segon panell** (**figura 9** dreta), ens trobem amb una posició completament diferent: si coneixem els valors de la variable independent x, podem conèixer amb molta precisió quin serà el seu valor a la y.

Per exemple, imaginem que la variable x són les hores d'estudi a matemàtiques i la variable dependent y són les qualificacions obtingudes a les proves finals. Si la relació entre les dues variables és la descrita en el **panell 1**, quan sabem la quantitat d'hores estudiant no podem conèixer quina és la nota final que s'obtindrà. La situació és completament diferent si la relació és la que descriu el **panell 2**: si conec quant temps un estudiant ha dedicat a les matemàtiques, conec gairebé perfectament la nota que obtindrà.

En el **panell 3** (**figura 10** esquerra) la relació és negativa. Quan creix la variable independent, decreixen els valors de la variable dependent. Més hores d'estudi no serveixen per tenir més nota (potser aquí tenim una reversió de la relació de causalitat: els estudiants que estan més bloquejats amb les matemàtiques hi dediquen més temps, però realment són els que tenen pitjors resultats!). A més, la relació entre les variables és molt menys precisa. El núvol de punts és més dispers, no adopta la forma clara de línia que mostrava el **panell 2**.

Finalment, el **panell 4** (**figura 10** dreta) mostra que els gràfics de dispersió poden mostrar patrons de relacions entre les variables més complexos i matisats que les relacions lineals. En el gràfic es veu que la relació és positiva, però l'efecte del canvi a la variable independent sobre la variable dependent és molt fort en els valors més baixos, mentre que per als valors alts de x, la variable independent gairebé no té efecte. En l'exemple, això vol dir que les primeres hores dedicades a estudiar matemàtiques tenen un efecte gran sobre la qualificació obtinguda, però que, a partir d'un cert moment, estudiar més hores hi contribueix molt menys.

Figura 11. Relació no monòtona i constatació de buits i de casos extrems (*outliers*)



El **panell 5** (**figura 11** esquerra) mostra una relació encara més interessant entre les dues variables. La relació entre les variables és positiva en els valors baixos de la *x* i és negativa a partir d'un cert moment. Això vol dir que la relació no és monòtona, no sempre va en una direcció. En el nostre exemple, hi ha un moment en el que més hores d'estudi són contraproduents, no és que l'efecte d'una hora d'estudi més tingui un efecte menor (efecte marginal decreixent, en llenguatge dels economistes), sinó que, directament, una hora més d'estudi fa treure pitjors resultats a les avaluacions.

Finalment, el **panell 6** (**figura 11** dreta) mostra com els diagrames de punts són molt útils per indicar anomalies a les relacions entre les variables. Poden ser identificats buits en les relacions: hi ha alguns valors d'algunes de les variables que no estan presents. També són bons per visualitzar l'existència de casos extrems (*outliers*).

En fer la descripció d'un gràfic ens podem fixar en tres aspectes:

- Quina és la tendència que es pot detectar? Positiva, negativa, curvilínia.
- Quina és la fortalesa d'aquesta relació?
- Quin és el grau de dispersió dels casos? Hi ha casos extraordinaris (*outliers*)? Quins són? Quines raons semblen explicar-los? (molt sovint aquests casos són errors en la introducció de les dades).

En conjunt, els gràfics de punts són un instrument imprescindible per comprovar les relacions que existeixen entre les variables numèriques. El nivell de mesura numèric de les variables ens dona molta informació dels casos individuals i això permet que les relacions entre les variables puguin ser molt afinades. La visualització de les relacions ens permet identificar anomalies (els casos extrems, els casos que s'aparten de la relació normal) que ens poden portar a pensar i desenvolupar o refinar les explicacions dels fenòmens que estem estudiant. En general, l'èmfasi en la visualització de les dades està relacionada amb la *serendipity*, un neologisme anglès que planteja la capacitat de fer descobertes inesperades al llarg de la nostra recerca.

4.2.3. Coeficient de correlació lineal: *r* de Pearson

Per mesurar el grau d'associació lineal entre dues mesures numèriques es fa servir el coeficient de correlació lineal de Pearson, *r*. És una mesura direccional que es troba acotada entre –1 i 1. El signe ens diu si la relació és negativa o positiva. Com més gran és el valor absolut de *r*, s'acosta més a 1, la relació lineal entre les dues variables és més forta. S'ha de ser conscient que relacions no lineals (com la del panell 5) tot i que siguin clares i fortes, no apareixeran o se subestimaran en els valors de *r*. Els paquets estadístics ens ofereixen mesures de significativitat estadística.

4.2.4. Anàlisi de regressió simple (o bivariant)

Quan, més enllà de la fortalesa de la relació entre dues variables i la direcció d'aquesta relació, volem examinar la relació entre dues variables numèriques entre les que suposem que existeix una relació de causalitat (o una relació estructural), fem servir l'**anàlisi de regressió**.

A l'anàlisi de regressió s'entén que la variable dependent *y*, i la variable independent *x* poden ser connectades matemàticament amb diferents formes funcionals.

La més simple és la forma de la línia recta:

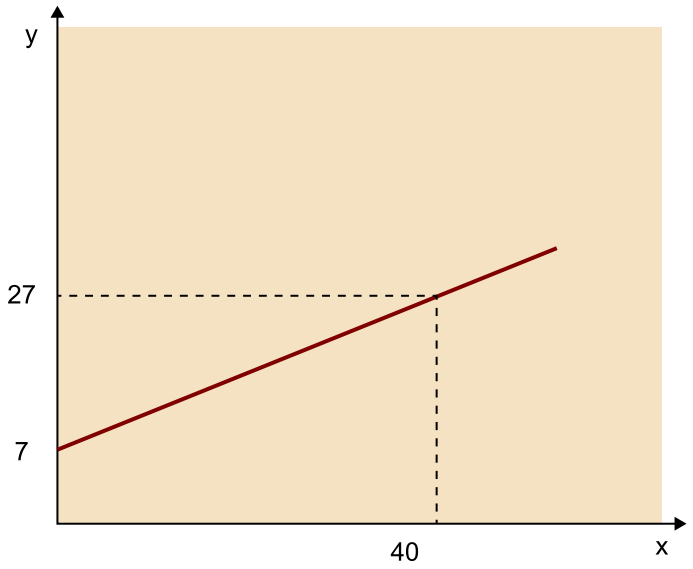
$$y = a + b\,x$$

Amb aquesta funció es poden plantejar relacions relativament complexes entre les variables. Per il·lustrar què hi ha en aquesta fórmula podem utilitzar un exemple políticament incorrecte.

Un estirabot del reconegut misogin Josep Pla diu que «l'edat ideal de la parella d'un home té la meitat d'anys de l'home més set anys». Així, un home de 40 anys hauria de fer parella amb una dona de 27. Aquesta relació es pot formular com una línia recta que estableix els anys de la dona (*y*) en funció dels anys de l'home (*x*):

Anys dona = 7 + ½ anys home

Figura 12. Relació anys dones respecte anys homes



En l'anàlisi de regressió, se suposa que la relació entre la variable independent (l'explicativa) i la variable dependent (a explicar) no és completament determinista: hi ha molts altres factors que expliquen els valors que assoleix la variable dependent (a més de la variable que estem tenint en compte). Llavors, esperem que existeixi una variació probabilista per cada valor del factor explicatiu. El factor d'error (e) captura l'efecte de tots aquests factors i ens diu que els valors de y que observem són superiors o inferiors als valors que prediu la relació i estan representats per la recta de regressió:

$$y = a + b\,x + e$$

En l'anàlisi de regressió bivariant podem esbrinar la relació entre les dues variables i els diagnòstics habituals que ens ofereixen els paquets estadístics ens permeten veure com és de forta i significativa aquesta relació.

4.3. Relacions entre variables qualitatives

4.3.1. La taula de contingència

Quan tenim variables qualitatives o categòriques (nominals o ordinals), la **taula de contingència** és l'instrument bàsic utilitzat per representar la relació. Aquesta eina és comparable al diagrama de dispersió utilitzat per les variables mesurades numèricament. Ens proporciona un grau semblant de riquesa i de matís en la percepció de la relació. En les anàlisis de les enquestes és típic que les dades es recullin amb aquest nivell de mesura.

Una taula de contingència és una **taula de doble entrada** que facilita l'examen de les relacions entre les variables.

Normalment, les categories de la variable independent (*y*) es posen a les columnes i les categories de la variable dependent (*x*), a les fileres de la taula.

Taula 7. Taula de contingència acusat víctima i fase del judici al qual arriba el cas

	Relació acusat víctima				
Fase del judici	**Casos perduts**	**Conegut/mestre**	**Desconegut**	**Pares/familiars**	**Total**
Sobreseïment	2	0	11	2	15
Arxivament	6	14	10	17	47
Judici faltes	2	0	3	0	5
Judici oral	0	3	13	14	30
Total	10	17	37	33	97

La **taula 7** mostra la relació entre la variable tipus de relació entre la víctima i l'acusat en els delictes sexuals a menors en els casos jutjats a l'Audiència Provincial de Lleida l'any 2011 i la fase del judici al que varen arribar. La primera casella de l'esquerra (primera columna, primera filera) ens diu que hi va haver 2 casos en els quals no es va poder saber la relació entre víctima i acusat, en els quals hi va haver sobreseïment.

La darrera columna i la darrera filera són especials perquè tenen la suma de cada filera i de cada columna, respectivament. S'anomenen els marginals. En realitat, la darrera columna és la distribució de les categories de la variable dependent (el nivell de judici al qual s'arriba), i la darrera filera és la distribució de les categories de la variable independent (la relació entre víctima i ofensor).

Per veure l'efecte de la variable independent sobre la dependent s'ha d'eliminar l'efecte de les diferents grandàries de les categories (hi ha molts més pares o familiars i desconeguts que no pas coneguts o mestres i casos perduts). Per aquesta raó, s'ha de calcular el percentatge de cada casella a la seva columna. És a dir, el valor de la casella pel marginal de la fila. Això ho veiem a la **taula 8**.

Taula 8. Relació acusat víctima i fase del judici al qual arriba el cas

	Relació acusat víctima				
Fase del judici	**Casos perduts**	**Conegut/mestre**	**Desconegut**	**Pares/familiars**	**Total**
Sobreseïment	20%	0%	30%	6%	15.5%
Arxivament	60%	82%	27%	51%	48.4%
Judici faltes	20%	0%	8%	0%	5.1%
Judici oral	0%	17%	35%	42%	30.9%

	Relació acusat víctima				
N	10	17	37	33	97

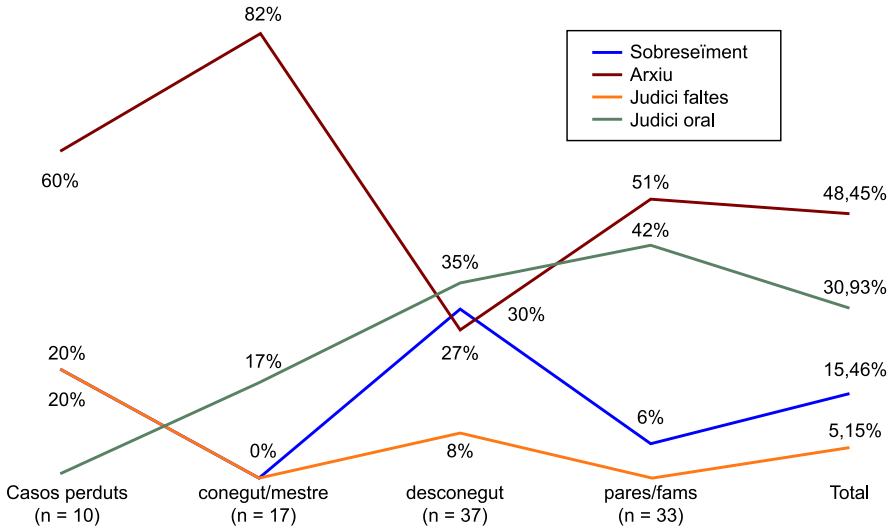
Com es llegeix la taula de contingència? L'**efecte** de la variable independent, la relació acusat víctima, sobre la variable dependent, la fase del judici a la que s'arriba, es veu **comparant els percentatges per columna de cada filera**. És a dir, triant una filera (per exemple, Arxivament) i veient els percentatges de «desconeguts» que hi ha per cada fase del judici. Així, els casos arxivats on l'ofensor és desconegut de la víctima són un 27%, molt per sota dels pares o familiars que són el 51%.

Si comparem els percentatges de cada categoria de la variable dependent (files) amb el seu marginal, veurem si cada casella està per sobre o per sota de la proporció en el conjunt de casos.

Quan llegim els percentatges per columna, el que estem mirant és la composició de la categoria de la variable independent sobre la dependent: com es distribueixen els casos en els que els ofensors són pares o familiars de la víctima respecte a la fase a la que arriben en el judici.

La comparació de les diferències entre els percentatges de columna permet calcular la diferència entre els percentatges per a cada categoria. Aquesta diferència és l'efecte de la variable independent amb la dependent. Per representar aquesta diferència entre les diferents caselles de la taula es pot dibuixar un **diagrama de pendents**, que es representa en un gràfic de línies. La **figura 13** presenta la informació de la **taula 8** en la forma d'un diagrama de pendents. D'un cop d'ull, es pot veure de quina forma la relació entre víctima i ofensor afecta la probabilitat d'accedir a una fase superior en el judici. Per exemple, destaca el fet que la major part dels casos on l'ofensor és conegut o és el mestre de la víctima, el cas se sobreseu.

Figura 13. Gràfic de pendents: Relació víctima-ofensor i fase del judici



Per tal de veure si són estadísticament significatives les relacions, es pot fer un test de significativitat de les diferències entre les proporcions. Però, en una taula normal hi ha moltes caselles i això vol dir que es poden fer moltes proves diferents. Si volem un test de la relació entre dues variables categòriques, aniria molt bé tenir un test global.

4.3.2. El test d'independència khi quadrat (χ^2)

El test d'independència **khi-quadrat** ens permet fer aquest test global de la relació entre dues variables categòriques. El test es basa en la discrepància entre els casos que s'observen a les diferents caselles de la taula en relació amb els casos que s'esperaria observar si les dues variables fossin independents. Si no hi hagués relació entre les variables, hi hauria poca discrepància. La distribució de valors de l'estadístic de **khi quadrat** ens diu quina és la probabilitat d'aquestes discrepàncies si les dades provinguessin d'una població en la qual les dues variables no estiguessin relacionades. Si el valor de khi quadrat supera un valor crític, es pot afirmar, en un determinat nivell de confiança, que les dues variables no són independents.

4.3.3. El coeficient de contingència i la V de Cramer

El test de khi quadrat només ens diu si es pot rebutjar la independència entre les variables, però no ens diu la fortalesa de la relació quan la comparem amb la relació que existeix amb altres variables. El **coeficient de contingència** (CC) o la **V de Cramer** són estadístics basats en la variable khi quadrat que estan acotats entre 0 i 1. El problema del CC és que no arriba generalment al màxim d'1, mentre que la V de Cramer, sí.

4.3.4. La lambda (λ)

Alternativament, la **lambda** (λ) és una mesura d'associació basada en la lògica de la reducció proporcional de l'error quan alguna de les variables té un **nivell de mesura nominal**. La reducció proporcional de l'error ens diu en quina mesura deixem d'equivocar-nos en atribuir un valor en una variable a un individu quan coneixem prèviament el seu valor en una altra variable. Es tracta, per aquesta raó, d'una mesura direccional: assumeix que hi ha una variable independent i una altra de dependent. L'ANOVA o la regressió lineal funcionen amb aquesta lògica.

4.3.5. La gamma (γ) i les tau (τ) de Kendall

La **gamma** (γ) i les diferents **tau** (τ) **de Kendall** són mesures d'associació de **variables ordinals** basades en la reducció proporcional de l'error. El càlcul es basa en la proporció de parelles concordants (que indiquen una relació positiva entre les variables) i parelles discordants (que n'indiquen una de negativa) a una taula de contingència. Les mesures són direccionals i estan acotades entre

-1 i 1. Les varietats de les tau de Kendall depenen del tipus de taula en les que es fa el càlcul, quadrades o rectangulars (és a dir, on el nombre de categories de les dues variables són iguals o diferents).

Glossari

coeficient de variació m $CV = \frac{s}{\mu_x}$

desviació típica f $s = \left(\frac{\sum (\mu_x - x_i)^2}{n} \right)^{1/2}$

índex HH m L'índex Herfindahl-Hirschman mesura la concentració de les variables categòriques (es va desenvolupar per mesurar la concentració industrial en diferents sectors) $HH = \sum p_i 2$, on pi és la proporció de cada categoria en la variable.

mitjana f $\mu_x = \frac{(x_1 + x_2 + \ldots + x_n)}{n} = \frac{\sum (x_i)}{n}$. Mesura de centralitat del nivell de mesurament quantitatiu. Hi poden haver versions «robustes» de la mesura que eviten els efectes dels valors extrems. Per exemple, la mitjana truncada (que elimina el 5 % dels valors superiors i el 5 % dels valors inferiors). És un sistema utilitzat en les competicions esportives (p. ex., gimnàstiques o de salts de trampolí) per evitar l'impacte de vots de jutges que vulguin afectar el resultat final amb una valoració desproporcionada.

moda f Valor més freqüent. És l'única mesura de centralitat de les variables categòriques nominals, però pot ser utilitzada en altres escales de mesurament. En les distribucions de probabilitat es poden distingir diferents modes: les generals i les locals.

quartil o decil m Valor que té una proporció d'individus de la distribució per sota o per sobre. Per exemple, el primer quartil Q1 és el valor que té el 25 % dels casos de la distribució per sota i el 75 % dels casos per sobre. El sisè decil D6 és el valor que té el 60 % dels casos per sota i el 40 % dels casos per sobre.

rang m Mesura de la distància entre el valor màxim i el mínim.

rang interquartílic m Diferència entre els valors del tercer i primer quartil d'una distribució (rang interquartílic = Q_3-Q_1).

raó de variació f $RV = \frac{1 - n_{Mo}}{n}$. Una mesura de dispersió de les variables categòriques, on n_{Mo} és la freqüència de la categoria modal.

variància f $s^2 = \frac{\sum (\mu_x - x_i)^2}{n}$

Bibliografia

Referències i lectures recomanades

Bayens, G. J.; Roberson, C. (2011). *Criminal Justice Research Methods: Theory and Practice* (2a ed.). CRC Press.

Capdevila Capdevila, M.; Ferrer Puig, M. (2009). *Tasa de reincidencia penitenciaria 2008* (p. 237) [Ámbit Social i Criminològic]. Barcelona: Centre d'Estudis Jurídics i Formació Especialitzada.

Capdevila Capdevila, M.; Ferrer Puig, M.; Luque Reina, E. (2005). *La reincidencia en el delito en la justicia de menores*(p. 276). Barcelona: Centre d'Estudis Jurídics i Formació Especialitzada.

Kubrin, C. E.; Wo, J. C. (2015). «Social Disorganization Theory's Greatest Challenge: Linking Structural Characteristics to Crime in Socially Disorganized Communities». A: A. R. Piquero (ed.). *The Handbook of Criminological Theory*. Chichester: West Sussex; Malden, MA: John Wiley & Sons.

Piquero, A. R.; Weisburd, D. (2009). *Handbook of Quantitative Criminology*. Nova York / Dordrecht / Heidelberg / Londres: Springer.

Weisburd, D.; Britt, C. (2007). *Statistics in Criminal Justice*. Springer.