Plantejament d'un estudi quantitatiu

Manuel Terrádez Gurrea

PID_00235996

Temps mínim previst de lectura i comprensió: **4 hores**

UOC
Universitat Oberta de Catalunya

Índex

Introducció..... 5

Objectius..... 6

1. Plantejament inicial..... 7

2. Objectiu i hipòtesi d'investigació..... 11

3. Justificació i viabilitat..... 13

4. Planificació del projecte..... 14

5. Mostreig..... 17

6. Anàlisi de la qualitat de la informació..... 21

7. Preparació de les dades..... 26

8. Estructura d'un informe..... 28

9. Exemples i casos pràctics..... 30
9.1. Control estadístic de qualitat 30
9.2. Anàlisi de riscos en banca 32
9.3. Enquestes de satisfacció 35
9.4. Indicadors sintètics 37
9.5. Anàlisi de supervivència en medicina 38

Resum..... 42

Bibliografia..... 43

Introducció

Vivim una època en què el coneixement basat en les dades permet aplicar anàlisis quantitatives en múltiples àmbits en els quals es genera informació que pot ajudar a resoldre problemes o necessitats; per això, per a investigadors de molt diverses disciplines és important conèixer les principals tècniques quantitatives.

La correcta planificació d'un projecte quantitatiu no és quelcom trivial, per la qual cosa és convenient conèixer els conceptes associats a aquest tipus d'estudis, així com també les principals fases que cal seguir per a dur a bon port el projecte.

És per això que a continuació mostrarem quines són les principals necessitats que poden motivar la realització d'un estudi quantitatiu. Així mateix, explicarem com traduir les necessitats en objectius i hipòtesis de treball i presentarem uns breus apunts sobre la justificació d'un projecte quantitatiu. En el quart apartat d'aquest mòdul, un dels més extensos, ens centrarem en la planificació del projecte, on s'expliquen les diverses fases que s'han d'emprendre des de les propostes de diversos autors. Després, introduirem els conceptes bàsics de la teoria del mostreig i els principals tipus de mostreigs. Posteriorment, explicarem algunes de les claus de l'anàlisi de la qualitat de la informació. També tractarem aspectes tan diversos com la preparació, selecció i transformació de les dades per al seu ús posterior. A continuació, descriurem quina estructura ha de tenir un informe per a plasmar i difondre els resultats del projecte. I, per acabar, presentarem una sèrie d'exemples i casos pràctics d'aplicació en diversos camps dels conceptes que s'han introduït al llarg del mòdul.

Objectius

Després de la lectura d'aquest mòdul l'estudiant serà capaç de:

1. Entendre per què és important realitzar estudis quantitatius.

2. Distingir quan un estudi quantitatiu pot resoldre una necessitat sorgida en un altre àmbit.

3. Planificar adequadament un projecte quantitatiu.

4. Comprendre les fases més importants d'un estudi quantitatiu: la selecció i extracció de la mostra, l'anàlisi de la qualitat de la informació i la preparació de les dades.

5. Aprendre a plasmar els resultats d'un estudi quantitatiu en un informe.

1. Plantejament inicial

Un **estudi quantitatiu** és aquell que usa l'emmagatzematge de dades, el mesurament numèric i l'anàlisi estadística per a provar hipòtesis o establir patrons de comportament.

Lectura recomanada

Es pot trobar una definició més completa d'un estudi quantitatiu a:

Hernández, R.; Fernández, C.; Baptista, P. (2010). *Metodología de la investigación*. México: McGraw-Hill.

Un estudi quantitatiu té sentit si es planteja per a resoldre una necessitat o per a abordar un problema que hagi sorgit en qualsevol àmbit.

Avui dia, amb la proliferació de dades de diversos tipus –moltes d'elles disponibles gratuïtament en les anomenades *bases de dades obertes* (*open data*)–, és inevitable caure en la temptació de descarregar-ne moltes i plantejar-se fer un estudi a partir d'aquestes dades, i no d'una pregunta o necessitat concreta, una situació que queda palesa en articles com «Are we mining data instead of answering questions?».

Una altra situació que es dóna amb certa freqüència, i a la qual segur que s'ha enfrontat alguna vegada gairebé qualsevol professional de l'estadística, és que un investigador es presenti amb una sèrie de dades que ha obtingut en el seu treball habitual i plantegi una pregunta del tipus «digues-me quines conclusions puc obtenir amb aquestes dades».

Aquests dos comportaments que acabem d'exemplificar són molt lícits i de cap manera menyspreables, ja que en algunes ocasions proporcionen anàlisis interessants perquè el treball de coneixement de les dades i les seves relacions pot arribar a ser molt productiu. No obstant això, la veritable essència d'un estudi quantitatiu és, com hem esmentat més amunt, mirar de donar solució (o, almenys, aportar una mica de llum perquè s'hi pugui arribar) a problemes que poden sorgir en qualsevol àmbit de la societat, la indústria, altres ciències, etc.

Per tant, la forma lògica i idònia de procedir en les primeres fases de l'estudi és la següent:

- Un investigador (i no l'estadístic, tot i que òbviament un estadístic és també un investigador, i com a tal pot tenir necessitats pròpies del seu àmbit a les quals pot donar resposta mitjançant estudis quantitatius, però en aquest context considerem l'estadístic com a professional aliè al camp de recerca en el qual sorgeix la necessitat que es vol resoldre amb l'estudi) es planteja un problema o una necessitat que pensa que es pot resoldre mitjançant l'obtenció de dades i la seva posterior anàlisi.

- L'investigador detecta quines són les dades que necessitaria per a donar resposta al seu problema, davant la qual cosa hi ha dues possibilitats:

  - Les dades ja existeixen i estan disponibles, bé sigui gratuïtament en alguna base de dades oberta, bé sigui previ pagament o sol·licitud, ja que pertanyen a una empresa privada o a un organisme públic.

  - Les dades no existeixen, i per tant cal dissenyar un experiment o una enquesta per a obtenir-les. En aquest cas, és molt important que el disseny es realitzi conjuntament entre l'investigador i l'estadístic, perquè un error molt habitual és que no es tingui en compte el professional quantitatiu en aquesta fase i es cometin errors de disseny que posteriorment condicionen l'anàlisi de la informació.

- Es recopilen les dades.

També és important destacar que, a l'hora de fixar l'objectiu o finalitat de l'estudi, és fonamental ser precisos i raonables, perquè de vegades succeeix que els objectius es plantegen de forma vaga i inconcreta (la qual cosa dificulta el disseny de l'experiment), o que persegueixen donar resposta a necessitats d'un abast tal, que resulten molt complicades d'abordar amb temps i recursos limitats.

En moltes ocasions, l'èxit d'un estudi quantitatiu depèn en major mesura de formular l'objectiu de manera adequada (és a dir, plantejar la pregunta correcta) que de refinar els resultats de l'algorisme estadístic corresponent, i és per això que es fa necessari dedicar el temps i l'esforç que calguin per a fer un plantejament apropiat.

Presentem a continuació una sèrie d'exemples d'aplicacions de l'estadística (algunes de clàssiques, unes altres de molt recents) en diversos camps, que sorgeixen de necessitats o problemes plantejats:

- Control estadístic de qualitat. Els processos productius acostumen a tenir una gran complexitat i una molt alta dependència de la qualitat obtinguda, per la qual cosa és necessari sotmetre'ls a un monitoratge continu que permeti reaccionar amb rapidesa davant possibles desviacions pel que fa a les característiques de qualitat requerides.

- Prediccions meteorològiques. La meteorologia té una gran influència en múltiples indústries (agricultura, turisme, etc.), i és per això que des de fa molt de temps s'ha mirat de predir el clima, però està subjecte a tal quantitat de factors que resulta molt complex fer-ho fins i tot a molt curt termini.

- Previsió d'existències o demanda. Un problema important a què s'enfronten la majoria d'empreses de fabricació de productes és la determinació el més ajustada possible de la seva demanda, perquè d'això dependran tant els pressupostos (i, a partir d'aquests, les possibilitats d'inversió)

com la gestió de les existències o l'emmagatzematge dels sobrants, que generalment té un cost elevat.

- Predicció de valors borsaris. Els inversors de borsa estan interessats a predir el comportament futur de les accions per a mirar d'invertir en aquelles que s'espera que pugin de valor, i per a retirar les inversions en aquelles que s'espera que baixin.

- Anàlisi de riscos en banca. Un problema cabdal de les entitats financeres és minimitzar la morositat generada per l'impagament dels préstecs i crèdits concedits. Per a aconseguir-ho, es mira de determinar quins clients tenen un millor perfil basant-se en les seves característiques socioeconòmiques i en el seu comportament previ (si es tracta de clients coneguts), per a concedir crèdit als que tenen un perfil més adequat.

- Enquestes de satisfacció. Qualsevol empresa o organisme que proporciona un producte o servei es planteja (o hauria de fer-ho) quin és el nivell de satisfacció dels seus usuaris amb l'objectiu de fidelitzar els clients actuals, obtenir-ne de nous, ampliar la gamma de productes, modificar les característiques dels ja existents, etc.

- Determinació de l'autor d'un text. De vegades els lingüistes es pregunten a qui es pot atribuir un text antic del qual es desconeix l'autoria, cosa que es pot arribar a determinar basant-se en les similituds que té amb altres textos coneguts de diversos autors.

- *Moneyball*: estadística i esport. Curiosament, l'esport és un dels àmbits en què, malgrat disposar d'una quantitat ingent de dades, l'explotació estadística és més recent, ja que tradicionalment ha estat sotmès a consideracions més subjectives i expertes, de l'estil del talent dels jugadors, l'experiència dels entrenadors o la intuïció dels observadors. No obstant això, després de l'enorme repercussió que va tenir el cas d'un equip de beisbol dels EUA (Oakland's Athletics), en el qual l'ús d'anàlisis estadístiques per a la selecció de la plantilla va reportar un èxit sense precedents, s'està generalitzant la seva aplicació a gairebé tots els esports.

- Anàlisi de supervivència en medicina. Davant el diagnòstic de malalties greus, és habitual que els metges provin diferents tractaments en els pacients amb l'objectiu de determinar quins els produeixen una millor resposta, tant des del punt de vista de major supervivència com també de mantenir una millor qualitat de vida. En l'anàlisi de supervivència, la variable d'interès és «temps transcorregut fins a l'aparició d'un esdeveniment» (*time-to-event*), i encara que és habitual que l'esdeveniment estudiat sigui la defunció d'un individu (d'aquí el terme *supervivència*), podria ser un altre tipus de fet com, per exemple, la fallada d'una màquina (anàlisi de fiabili-

tat), la venda d'un immoble, l'entrada d'una trucada a un centre d'atenció telefònica, etc.

- Creació d'indicadors sintètics. Són moltes les disciplines en les quals tenen molta rellevància les comparatives mitjançant l'elaboració de rànquings (productivitat de països, qualitat d'universitats, rendibilitat d'empreses, etc.), però davant la multiplicitat d'indicadors per a efectuar els mesuraments de cada àmbit d'interès, sorgeix la necessitat de resumir-los en uns quants valors que condensin la major part d'informació, els anomenats *indicadors sintètics*.

2. Objectiu i hipòtesi d'investigació

Com ja s'ha comentat, és molt important definir amb claredat els objectius de l'estudi que es vol abordar. És habitual que una investigació que impliqui certa complexitat abordi més d'un objectiu alhora, si bé acostuma a haver-hi un únic objectiu principal i, probablement, alguns objectius secundaris addicionals. A més, la forma de concretar els objectius acostuma a comportar l'elaboració d'unes hipòtesis de treball (que representen una declaració formal d'uns resultats) que en finalitzar l'estudi caldrà concloure si s'han acceptat o s'han rebutjat.

Un **objectiu** és un determinat propòsit o resultat que pot aconseguir-se i definir-se de manera que serveixi de base a un pla d'acció.

Els objectius han de ser tangibles i mesurables, per a poder avaluar el grau de compliment, i generalment (encara que no sempre) es formulen associats a un termini.

Hi ha diverses formes de classificar els objectius:

- **Segons la seva naturalesa**: objectius d'innovació (busquen que es produeixi un canvi positiu), de control (busquen impedir un canvi negatiu) o d'aprofundiment en un tema acadèmic (busquen fer evolucionar teories, pràctiques, millorar l'eficiència de processos, entre altres).

Exemples

En enquestes de satisfacció es pot buscar un augment de la satisfacció general o una reducció de les reclamacions. En estudis mèdics es pot buscar una millora en els beneficis d'un fàrmac o una disminució dels seus efectes secundaris.

- **Segons el seu àmbit d'actuació**: tàctics (busquen canvis concrets i a curt termini) o estratègics (busquen canvis estructurals i a llarg termini).

Exemples

En estudis meteorològics es busca tant la predicció de la temperatura que farà l'endemà com l'avaluació dels efectes a llarg termini del canvi climàtic. En estudis de control de qualitat es vol valorar si un lot és vàlid o no, però també establir els límits de tolerància per a avaluar futurs lots.

Ara bé, la traducció d'un problema o d'una situació que es vol resoldre a un objectiu d'investigació i a unes hipòtesis de treball no sempre és trivial. Comença amb la formulació de la pregunta clau que l'investigador té interès per

a resoldre, és a dir, el motiu principal de la investigació. Posteriorment, s'ha d'expressar en termes d'investigació tornant a formular la pregunta en un o més objectius i, amb posterioritat, exposar aquests objectius en una terminologia precisa i científica, i formular-ne hipòtesis de treball. Aquest procés s'anomena *definició del problema*.

El procés de **definició del problema** inclou diverses fases relacionades entre les quals hi ha les següents:

1) Comprendre la situació de partida: identificar indicis clau.

En aquesta fase sempre és important, igual que en estudis no quantitatius, documentar-se sobre altres treballs previs que hagin abordat problemes similars i que hagin estat difosos en publicacions acadèmiques o un altre tipus de mitjans.

2) Identificar problemes (actuals o potencials) a partir dels indicis detectats.

3) Redactar la pregunta clau i associar-hi els objectius d'investigació corresponents.

4) Determinar les unitats d'anàlisi.

La unitat d'anàlisi per a un estudi indica què o qui ha de proporcionar la informació i en quin nivell d'agregació. Els investigadors especifiquen si es recaptaran dades sobre individus (clients, empleats i propietaris), llars (famílies, famílies extenses, etc.), organitzacions (empreses i unitats de negoci), departaments (vendes, finances, etc.), àrees geogràfiques o objectes (productes, anuncis, etc.), etc.

5) Determinar les variables pertinents.

Una variable és qualsevol característica mesurable que està subjecta a variació; mostra diferències en el valor, que poden ser de força (magnitud) o direcció (signe). En una investigació quan una variable s'observa o es manipula, es denomina variable experimental.

6) Formular les preguntes o les hipòtesis d'investigació.

Les preguntes d'investigació expressen els objectius de l'estudi en termes d'inquietuds que poden ser analitzades en la investigació. Les hipòtesis, per la seva banda, són més específiques i rigoroses que les preguntes d'investigació.

3. Justificació i viabilitat

És important remarcar que a l'hora de plantejar un projecte d'investigació, i sobretot si involucra l'obtenció de recursos (ja siguin públics o privats), cal justificar-lo convenientment, destacant-ne la rellevància i els beneficis que se n'obtindran.

La justificació pot venir per aspectes teòrics (aportacions metodològiques, per exemple), pràctics (noves aplicacions per a resoldre problemes reals) o mixtos (aplicació a un camp nou de metodologies ja conegudes en altres àmbits), per la seva rellevància social (beneficia un col·lectiu concret), per a obrir noves línies d'investigació, etc.

També és habitual dur a terme una anàlisi en profunditat del marc teòric (publicacions, informes, etc.) ja existent sobre el tema a tractar, com a base per a justificar la seva pertinència o el motiu pel qual s'han de destinar esforços i recursos a l'estudi que es vol fer. Una anàlisi exhaustiva de l'estat de la qüestió sempre ajuda a emmarcar i contextualitzar el projecte.

Un altre aspecte important és la viabilitat de la investigació. Ja hem comentat que els objectius plantejats han de ser raonables, dit això en sentit ampli, és a dir, cal ser conscient del temps i dels recursos (humans, tècnics i econòmics) disponibles a l'hora d'afrontar un projecte.

4. Planificació del projecte

És important tenir clar que cada estudi o projecte d'investigació mira de resoldre un problema concret i, per tant, té les seves pròpies característiques diferenciadores, amb la qual cosa pot requerir un disseny específic no directament intercanviable amb altres projectes. Per aquest motiu és rellevant recollir la màxima informació prèvia sobre el problema objecte de la investigació, i que no sigui possible una estandardització total de les fases que s'han de dur a terme.

No obstant això, existeixen diverses metodologies que miren d'estructurar i sistematitzar les etapes més freqüents que s'han d'abordar a l'hora de dur a terme un estudi quantitatiu. Una de les més reconegudes és CRISP-DM, que es pot resumir de la manera següent:

CRISP-DM

Per a més informació sobre CRISP-DM 1.0, podeu consultar la guia *Step-by-step data mining guide*.

1) Comprensió del negoci o problema.

a. Especificació dels objectius qualitatius.
b. Avaluació de la situació actual.
c. Especificació dels objectius quantitatius.
d. Desenvolupament del pla de projecte.

2) Comprensió de les dades.

a. Recollida de les dades inicials.
b. Descripció de les dades.
c. Exploració de dades.
d. Especificació de la qualitat de la informació.

3) Preparació de les dades.

a) Selecció de les dades que s'utilitzaran.
b) Neteja de les dades.
c) Construcció de noves variables.
d) Integració de dades de diverses fonts.
e) Ús d'un format correcte de les dades.

4) Modelització.

a. Selecció d'una tècnica adequada.
b. Generació del disseny del test.
c. Desenvolupament del model.
d. Avaluació del model.

5) Validació del model.

a. Avaluació dels resultats.
b. Revisió del procés.
c. Determinació dels propers passos.

6) Validació del model.

a. Creació d'un pla d'explotació.
b. Creació d'un pla de revisió i manteniment.
c. Obtenció d'informes finals de resultats.
d. Revisió del projecte.

De forma similar es pronuncien altres autors, com Fayyad, a l'hora de ressaltar les principals fases d'aquest tipus d'estudis, si bé destaquen que no es tracta d'un procés lineal, sinó iteratiu i subjecte a múltiples revisions de cadascun dels passos anteriors:

Lectura recomanada

Una explicació més exhaustiva del procés i les seves fases es pot trobar a:

Fayyad, U; Piatetsky-Shapiro, G; Smyth, P. (1996). «From Data Mining to Knowledge Discovery in Databases». *AI Magazine* (vol. 17, núm 3, pàg. 37-54).

1) Comprensió del camp d'aplicació.

a. Incorporació de coneixement previ.
b. Identificació de l'objectiu del projecte des del punt de vista de l'investigador.

2) Creació d'una base de dades.

a. Selecció de les dades d'interès.

3) Preparació de les dades.

a. Neteja de les dades.
b. Preprocessament de les dades.
c. Eliminació del soroll.
d. Tractament de les dades perdudes.
e. Reducció de la dimensió.
f. Transformació de variables.

4) Modelització.

a. Selecció del model adequat per a l'objectiu.
b. Anàlisi exploratòria de les dades.
c. Selecció d'hipòtesi.

5) Cerca de patrons.

a. Interpretació dels patrons.
b. Visualització.

6) Explotació dels resultats.

a. Ús directe.
b. Incorporació a altres sistemes ja existents.
c. Documentació i publicació.

Pel que fa als treballs relacionats amb enquestes (estudis de mercat, sondejos d'opinió, etc.), que són els més freqüents en l'àmbit que ens ocupa, si bé es poden emmarcar dins dels estudis quantitatius generals, tenen algunes característiques pròpies, per la qual cosa la seva planificació inclou certes particularitats que miren de quedar reflectides en la relació que indiquem a continuació.

1) Etapa prèvia o d'estudi.

a. Definició d'objectius.
b. Estudi de les fonts d'informació disponibles.
c. Anàlisi i diagnòstic de la informació recollida.
d. Plantejament de l'operació i disseny mostral.

2) Etapa preparatòria dels treballs de camp.

a. Preparació i edició de materials.
b. Selecció i formació del personal.
c. Organització i assignació de funcions.
d. Previsió i solució d'incidències.

3) Treball de camp.

4) Tractament de la informació.

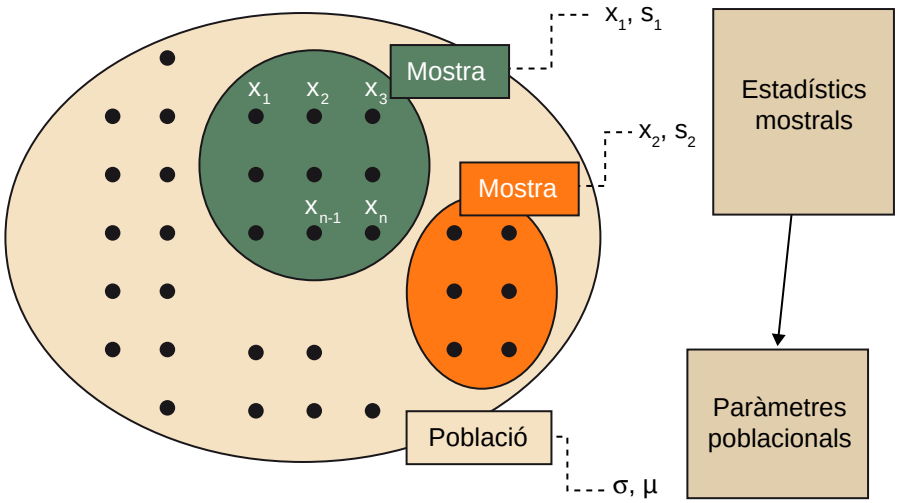
a. Introducció de dades.
b. Depuració.
c. Codificació.
d. Enregistrament.
e. Validació.
f. Avaluació.
g. Tabulació de resultats.
h. Anàlisi estadística univariable.
i. Anàlisi estadística multivariable.
j. Interpretació de resultats.
k. Presentació i publicació de resultats.
l. Limitacions del treball i pròxims passos.

5. Mostreig

Una de les branques més clàssiques i importants de l'estadística és la coneguda com a *inferència*. En aquesta branca, partint d'una població, s'obté una mostra (és a dir, un subconjunt de la població), i dels estadístics que es calculen mitjançant les dades obtingudes s'infereixen (és a dir, es dedueixen) valors dels paràmetres poblacionals.

Com que estudiarem mètodes les dades dels quals són obtingudes mitjançant l'observació, acostuma a ser necessari prendre una mostra de la població, perquè aquesta rares vegades és observable, ja sigui per la seva grandària o per l'esforç (en cost o temps) que suposaria fer-ho.

La realitat és que quan prenem una mostra podem conèixer-ne els valors representatius (estadístics) però mai no coneixerem realment com és la nostra població; tanmateix, sí que podem fer una estimació de com és. Gràficament, la forma de procedir seria la següent:



A continuació donarem una definició més formal dels elements que formen part d'aquest procés:

- **Població**: és el conjunt d'individus que constitueixen l'univers objecte d'un determinat estudi, i sobre els quals es desitja obtenir determinades conclusions. Anomenarem *N* la grandària de la població (que pot ser desconeguda).

- **Mostra**: és un subconjunt representatiu de la població. Anomenarem *n* la grandària de la mostra.

- **Paràmetre poblacional**: és una característica de la població, desconeguda, que es pretén estimar.

- **Estadístic**: és un valor (o combinació de valors) observable de la mostra.

- **Estimador**: és un estadístic que s'utilitza per a estimar el valor d'un paràmetre poblacional.

En aquest apartat s'explicaran diferents mètodes de mostreig que poden utilitzar-se per a definir el **pla de mostreig**. És molt important realitzar correctament el pla de mostreig (que inclou decidir no només el mètode de mostreig, sinó també la grandària de mostra, el mètode de recollida o observació, etc.), ja que si es fa un mal mostreig pot ser que s'arribi a conclusions que no tinguin validesa.

Exemple

Un exemple de mostreig fet incorrectament queda recollit en l'article "¿Por qué es más fácil obtener plaza en un sorteo si te apellidas Martínez?".

És molt important també que la mostra sigui representativa (o, per a ser més precisos, estadísticament representativa) de la població, és a dir, que no tingui biaixos. Perquè una mostra es consideri representativa hauria de contenir les característiques rellevants de la població en les mateixes proporcions que hi estan incloses.

Els **biaixos** són els errors que es cometen a l'hora d'obtenir una mostra: entre ells, els més importants són el biaix de selecció, quan els elements de la mostra no s'han seleccionat correctament, i el biaix de no resposta, quan no es pot obtenir la informació d'algun element mostral.

Els mètodes de mostreig es classifiquen generalment en dos grans grups: probabilístics i no probabilístics.

En els **mostreigs no probabilístics**, els elements de la població es trien seguint una lògica «dirigida», no es trien a l'atzar; per tant, no cal conèixer quina és la probabilitat de pertinença a la mostra de cadascun dels individus de la població. Això ens porta a no poder mesurar l'error que es comet en fer les prediccions. Dins d'aquest tipus de mostreig podem incloure els mostreigs per conveniència (per exemple, l'entrevistador se situa en un punt concret i entrevista la gent que passa pel seu costat), per judici (s'escull un comitè d'experts en un tema concret), accidentals (es provoca una situació), i per bola de neu (es trien els primers enquestats i després aquests faciliten els següents elements de la mostra entre els seus coneguts o afins).

De totes maneres, ens centrarem en els **mostreigs probabilístics**, per ser els més habituals i també els més rigorosos. En ells, la selecció dels elements es fa a l'atzar i es coneix prèviament quina és la probabilitat de pertànyer a la mostra.

Dins d'aquest tipus, destaquem els següents mètodes de mostreig:

- **Mostreig aleatori simple**. És el més habitual i senzill, i a més és molt eficaç. Pot ser amb o sense reemplaçament (segons si un element ja triat es reintegra en la població –i, per tant, pot tornar a ser seleccionat– o no). Els elements són equiprobables i s'obtenen per pur atzar (amb mètodes com, per exemple, el clàssic bombo o la generació de números aleatoris).

- **Mostreig estratificat**. S'utilitza quan hi ha grups (estrats) en la població la representativitat de la qual es vol respectar en la mostra, per la qual cosa se selecciona una mostra aleatòria dins de cada estrat (que pot respectar o no la proporció d'elements d'aquest estrat en la població). El mètode de selecció en cada estrat pot ser diferent, però aquesta selecció ha de ser independent entre estrats.

- **Mostreig per conglomerats**. També s'utilitza quan hi ha grups en la població, però en aquest cas el mostreig es duu a terme en dues etapes: primer se selecciona a l'atzar una mostra de conglomerats, i després se selecciona una mostra aleatòria d'elements de cada conglomerat.

- **Mostreig per quotes**. En aquest cas es distribueixen els individus de la població en una sèrie de categories i se n'assigna un nombre a cada categoria, de manera que la proporció d'individus de cada categoria en la mostra sigui similar a la proporció dins de la població.

- **Mostreig sistemàtic**. En aquest mètode se selecciona a l'atzar només la primera unitat, mentre que les restants se seleccionen a intervals fixos partint de la primera.

Els mètodes de mostreig no són incompatibles entre ells, sinó que es poden combinar, originant mostreigs en etapes múltiples.

Un altre aspecte clau a l'hora de realitzar el mostreig és determinar la **grandària mostral** més adequada. Òbviament, com major sigui la grandària de mostra més precisió es podrà aconseguir en les estimacions, però cal buscar un equilibri entre aquest objectiu i el cost que suposa augmentar la grandària.

La grandària mostral òptima està íntimament lligada a uns altres dos conceptes: la **precisió** (mesurada mitjançant el marge de l'error –també anomenat *error d'estimació* o *error mostral*– o l'interval en el qual esperem que es trobi el nostre paràmetre poblacional) i el **nivell de confiança** (certesa que realment el paràmetre que busquem estigui dins de l'interval definit pel marge d'error).

Precisió, confiança i grandària de la mostra sempre van associades i, per tant, modificar qualsevol dels tres paràmetres altera els restants; per exemple, augmentar la confiança o la precisió sense modificar l'altre paràmetre obliga a augmentar la grandària de la mostra.

Sobre aquest aspecte hi ha una bibliografia molt àmplia que estableix les fórmules per a obtenir la grandària de mostra òptima segons el tipus de mostreig i l'estimador utilitzat, i fins i tot en els últims temps han proliferat múltiples calculadores en línia amb aquesta finalitat com Calculadora para obtener el tamaño de una muestra o Calculators.

De totes maneres, no hem d'oblidar que si bé és molt important seleccionar la mostra de forma adequada, no ho és menys definir correctament la població, perquè d'altra manera no serà possible assegurar que la mostra sigui representativa de la població.

Altres conceptes associats a aquest tipus d'experiments són els de fiabilitat i validesa.

La **fiabilitat** fa referència al nivell d'aplicabilitat i consistència d'un estudi i, en concret, dels seus mètodes, condicions i resultats. Una forma de realitzar estudis aplicables és seguir una planificació sistemàtica amb un mètode rigorós (des del punt de vista científic).

Un estudi determinat té **validesa interna** quan les relacions causa-efecte que es produeixen són les planificades per l'investigador, és a dir, no existeixen variables estranyes que afectin l'experiment (o, almenys, el seu impacte està controlat). D'altra banda, la **validesa externa** és la capacitat d'un estudi per a generalitzar o extrapolar les seves conclusions a altres contextos diferents a aquell sota el qual s'ha realitzat.

6. Anàlisi de la qualitat de la informació

La producció d'informes estadístics de qualitat depèn en gran mesura de la qualitat de les dades de partida i, per tant, és molt important avaluar-la de forma sistemàtica. Així mateix, és important avisar els possibles usuaris o revisors dels nostres informes estadístics de les dades utilitzades i del seu nivell de qualitat.

Amb molta freqüència i per diversos motius (procedència de diferents fonts o períodes temporals, etc.), les dades que volem utilitzar en el nostre estudi no estan prou preparades per a començar a analitzar-les, per la qual cosa cal fer un procés d'anàlisi de la seva qualitat i de neteja o transformació per a poder treure'n la major i millor quantitat d'informació possible.

Eurostat, l'agència europea d'estadística, va proposar en 2005 una relació d'indicadors per a avaluar la qualitat de la informació.

Lectura recomanada

Per a més detalls es pot consultar:

Eurostat (2007). *Handbook on data quality assessment methods and tools*.

Aspecte	Codi	Indicador	Importància
Rellevància	R1	Índex de satisfacció de l'usuari	3
	R2	Taxa d'estadístiques disponibles	1
Precisió	P1	Coeficient de variació	1
	P2	Taxa de resposta per unitat	2
	P3	Taxa de resposta per ítem	2
	P4	Taxa i ràtio d'imputació	2
	P5	Taxa de classificació errònia	2
	P6	Taxa de cobertura geogràfica	1
	P7	Grandària mitjana de revisions	1
Puntualitat	T1	Puntualitat de publicació	1
	T2	Retard fins a disponibilitat de resultats preliminars	1
	T3	Retard fins a disponibilitat de resultats definitius	1
Accessibilitat	A1	Nombre de publicacions	1
	A2	Nombre d'accessos a bases de dades	1
	A3	Taxa de completesa	3
Integritat	I1	Grandària de sèries temporals comparables	1
	I2	Nombre de sèries temporals comparables	1
	I3	Taxa de diferències respecte estàndards	3

Aspecte	Codi	Indicador	Importància
	I4	Asimetries	1
Coherència	C1	Taxa d'estadístiques que satisfan els requeriments	3

Cal destacar que és una llista genèrica, però que es pot adaptar a cada cas concret, és a dir, pot ser que no tots aquests indicadors siguin aplicables al nostre estudi, i fins i tot que n'hi hagi uns altres que siguin rellevants i no estiguin a la llista. De fet, cada base de dades necessita el seu propi banc de proves i, per això, és una tasca difícil de sistematitzar, ja que l'avaluació de la qualitat està íntimament lligada al context en el qual s'utilitzaran les dades i al propòsit de la investigació.

Cal no perdre de vista tampoc que aquest tipus d'avaluació de la qualitat ha de realitzar-se de forma eficient, perquè no ens faci invertir més temps i esforç del necessari.

No obstant això, l'esmentat organisme europeu va establir el procés d'avaluació de la qualitat de dades d'una forma sistemàtica, que es pot resumir de la següent manera:

1) Revisar objectius de qualitat de dades i disseny mostral.

2) Realitzar revisions preliminars de les dades.

3) Seleccionar el model estadístic adequat.

4) Verificar les hipòtesis.

5) Obtenir conclusions.

Ens centrarem en les dues primeres tasques, que són les prèvies a l'anàlisi estadística, perquè les restants seran tractades en altres materials de l'assignatura.

1) Revisar els objectius de qualitat de dades (OQD) i el disseny mostral.

Aquesta fase té com a principal objectiu revisar la consistència dels OQD, el disseny mostral i qualsevol documentació sobre la recopilació de dades. Si els OQD no han estat desenvolupats, cal definir el mètode estadístic i especificar els límits tolerables en els errors de decisió.

Com a activitats concretes en aquesta fase podem destacar les següents:

- Revisar els objectius de l'estudi. Els objectius de l'estudi, dels quals ja hem parlat en seccions anteriors, són ara revisats perquè ens permetin dissenyar

un context en el qual recollir i analitzar les dades, i identificar clarament la població d'estudi i les possibles subpoblacions d'interès.

- Traslladar objectius a hipòtesis estadístiques. És habitual definir el que es denomina una *hipòtesi nul·la* (o *condició base*), que es dóna per positiva en absència d'evidència en contra, i una *hipòtesi alternativa*. També és habitual definir un paràmetre poblacional d'interès, un valor numèric amb el qual aquest paràmetre vol ser comparat, i una relació entre tots dos del tipus «igual», «major que» o «menor que».

- Establir límits d'errors de decisió. Les mesures d'estadística inferencial acostumen a anar acompanyades d'una especificació de l'error, bé sigui directa, o bé basada en un nivell de confiança o tolerància de l'estimació. En aquest context sorgeixen els anomenats *error de tipus I* o *fals positiu* (error comès en rebutjar una hipòtesi nul·la certa) i *error de tipus II* o *fals negatiu* (error comès en no rebutjar una hipòtesi nul·la falsa). Altres conceptes associats són el *nivell de significació*, que és la probabilitat de cometre un error de tipus I, o la *potència* d'un test, que és la inversa de la probabilitat de cometre un error de tipus II.

- Revisar el disseny mostral. En aquest punt és important revisar si el disseny mostral triat és adequat per a complir els objectius de l'estudi.

2) Fer revisions preliminars de les dades.

Aquesta fase té com a principal objectiu generar sortides numèriques i gràfiques per a descriure les dades i utilitzar aquesta informació a fi d'obtenir coneixement sobre l'estructura de les dades i identificar possibles patrons i relacions.

En aquesta fase podem destacar les activitats concretes següents:

- Revisar informes de qualitat.
- Calcular estadístics descriptius.
- Utilitzar gràfics i altres tipus d'eines de visualització de dades.

Considerem important aprofundir més en aquesta segona fase de preparació de les dades, ja que és força més rellevant del que moltes vegades es creu. Tant en projectes d'investigació com en projectes empresarials, acostuma a representar prop del 80% del temps del projecte.

Podem destacar les etapes següents:

1) Disseny de l'enquesta/experiment.

Aquest aspecte és molt important, perquè la manera de recollir les dades pot condicionar la seva futura qualitat i aplicabilitat. Es recomana demanar assessorament estadístic a l'hora de realitzar el disseny.

2) Anàlisi univariable.

L'anàlisi univariable de les variables ens permet no només conèixer les característiques més importants de la distribució, sinó també detectar possibles errors.

3) Tractament de les dades no disponibles.

Si hi ha dades no disponibles (*missing values*), cal assegurar-se que no existeixi un biaix important i després decidir què fer-ne:

- Eliminar la variable completa si el percentatge de dades que falten és molt alt.
- Eliminar els registres que tinguin alguna dada no informada.
- Preferiblement, imputar els valors que faltin per a algun valor representatiu (mitjana, moda) o utilitzar tècniques més sofisticades d'imputació.

4) Tractament de dades atípiques.

És molt important detectar les dades atípiques (*outliers*), ja que poden tenir una gran influència en el resultat final dels models o estudis. És recomanable no tenir en compte aquestes dades a l'hora de construir el model, encara que això també depèn d'altres factors.

5) Codificació de la informació.

Abans d'utilitzar tècniques multivariables és important fer un correcte tractament de la informació:

- Codificar correctament les variables categòriques.
- Transformar les variables contínues de manera que les seves escales siguin semblants.
  - Normalitzar (canvi d'escala).
  - Estandarditzar (centrat i escalat).
  - Altres transformacions (logaritme, arrel quadrada, etc.).
- Generar noves variables.
  - Crear variables *dummy* (0/1) si es considera necessari.

– Realitzar operacions amb variables (sumes, diferències, ràtios, mitjanes, etc.).

6) Documentació.

És molt important documentar de forma adequada tots els passos realitzats en la fase de preparació de la informació per a poder realitzar una futura consulta, tant per a nosaltres mateixos com per a qualsevol altra persona que l'hagi de revisar. De fet, és recomanable documentar els passos durant el procés i no deixar aquesta tasca per al final del projecte, ja que això pot representar de vegades una pèrdua d'informació important.

7. Preparació de les dades

La preparació de les dades, a més de no ser gens fàcil d'estandarditzar, és una tasca que no es fa una sola vegada al principi d'un projecte, ja que en molts estudis s'ha de repetir unes quantes vegades durant l'anàlisi, a mesura que apareixen nous problemes o quan ens adonem d'alguns problemes que ja existien però que havien passat desapercebuts en un primer diagnòstic.

Wickham defineix una base de dades com «un conjunt de valors, ja siguin nombres o text, organitzats de tal manera que cada valor pertany a una variable i a una observació». I denomina *dades ordenades* (en anglès, *tidy data*) a una «forma estàndard d'organitzar els valors de les dades dins d'una base de dades», de manera que cada variable és una columna, cada observació és una fila i cada taula està formada només per un tipus d'unitats d'observació, és a dir, que en una mateixa taula no ha d'haver-hi dades de diferents tipus.

Lectura recomanada

Una explicació més exhaustiva es pot trobar a:

Wickham, H (2014). «Tidy Data». *Journal of Statistical Software* (núm. 59, pàg. 1-23).

Aquesta és una convenció típica entre els estadístics i els analistes d'informació que facilita molt l'anàlisi quantitativa de les dades. Tanmateix, encara que això no és tan important, l'ordre de les variables segueix normalment una sèrie de regles: les anomenades dimensions (variables fixes, normalment de tipus text, conegudes prèviament) van al principi, mentre que les variables de mesura vénen després, normalment agrupades segons les relacions que hi hagi entre elles.

En qualsevol cas, una eina clau quan parlem de dades és el **diccionari de dades**, un document que conté totes les variables del nostre estudi: nom, tipus, format, escala, unitats, domini (valors possibles) i qualsevol altra informació rellevant.

Com a resum dels aspectes concrets que considerem que cal revisar com a mínim per a garantir una correcta qualitat de la informació, en destacaríem els següents:

- Revisió de possibles pèrdues d'informació (o, per contra, creació de multiplicitats) si s'ha hagut de fer un encreuament entre diferents arxius.

- Camps clau: són els camps que ens permeten identificar de forma unívoca cada registre.

- *Missing values*: es tracta de valors no disponibles, és a dir, dades que no s'han pogut obtenir, bé perquè no existeixen o bé perquè ha estat impossible recuperar-les.

- Possibles biaixos: cal evitar-los.
- Duplicats: hi ha camps que no els admeten (per exemple, els camps clau).
- Formats: és important que en una mateixa variable no hi hagi dades en formats diferents.
- *Outliers*: n'hem parlat més amunt.
- Valors fora de rang: valors majors o menors que els límits que admet la variable, i que normalment indiquen errors.
- Integritat entre les dades internes.
- Coherència amb possibles fonts externes de dades.

L'anàlisi univariable/bivariable és sempre important com a primer pas, però moltes vegades cal anar més enllà: la majoria de processos i comportaments en la vida real són «multivariables» (complexos, diversos, multidimensionals), i així s'han d'analitzar.

En aquest punt també considerem important advertir del risc del programari senzill i accessible. Amb els avenços tècnics i informàtics, han aparegut molts programes i paquets estadístics que permeten, sense gaire formació prèvia, analitzar dades i obtenir resultats, la qual cosa representa un avantatge considerable. No obstant això, s'ha de tenir en compte que, sense un correcte coneixement (o assessorament professional) de les tècniques estadístiques, existeix un risc important de fer una anàlisi incorrecta o de realitzar una interpretació inadequada dels resultats obtinguts.

8. Estructura d'un informe

La forma de presentar els resultats i les conclusions dels estudis estadístics és molt semblant a la d'altres camps científics. En general, hem de mirar de descriure els resultats i les conclusions de les nostres anàlisis de manera que ho pugui entendre una persona sense coneixements estadístics; si no som capaços d'explicar-ho de manera senzilla, potser no és prou valuós.

L'estructura d'un informe podria ser la següent:

1) Resum

Fer una descripció clara de la qüestió d'interès, les dades utilitzades per a mirar de respondre-la i les conclusions de l'anàlisi. Afegir els estadístics (estimadors, intervals de confiança, valors de probabilitat) més importants; no cal donar gaire detall, però tampoc podem obviar cap problema significatiu que hàgim trobat. L'objectiu principal és disposar de tota la informació clau en el resum, mentre que la resta de detalls de suport apareixeran en altres apartats de l'informe.

2) Fonaments

Fer una descripció de la motivació científica de l'estudi. No cal donar gaire detall, però sí tots els fets que han format part del procés de decisió.

3) Fonts de dades

Descriure les fonts de dades i els mètodes de mostreig, si són coneguts, així com també les variables disponibles i el seu significat per a l'anàlisi. Ressaltar els patrons de dades que falten, així com també els possibles factors de confusió.

4) Mètodes estadístics

Descriure els mètodes utilitzats per a l'anàlisi a dos nivells:

- Fer una descripció tècnica, incloent-hi referències bibliogràfiques. Descriure el programari utilitzat i els mètodes per a garantir la correcció dels models desenvolupats. Explicar com s'han gestionat els principals problemes, com ara el tractament de les dades no disponibles o les dades atípiques, entre d'altres.

- Explicar de manera senzilla la lògica que hi ha darrere les tècniques utilitzades. Fer interpretacions dels paràmetres estimats. Si cal, explicar per què no s'han utilitzat altres tipus de tècniques.

5) Resultats

Oferir de forma gradual els resultats obtinguts, començant amb les estadístiques descriptives i continuant amb els models.

6) Discussió

Presentar les conclusions que es poden derivar de les anàlisis. Suggerir futurs estudis. Ressaltar les limitacions de les dades i de l'anàlisi.

7) Annex

Taules i altres resultats que no s'hagin aportat en l'apartat de resultats.

És important adreçar el lector cap als resultats més rellevants. Després de dedicar molt de temps a analitzar les dades, hem d'oferir una breu revisió dels aspectes de major rellevància. Els diagnòstics estadístics, als quals probablement hem trigat molt a arribar, moltes vegades es poden resumir en una simple frase. S'han de reflectir els principals resultats i impressions sobre les dades, no tots els detalls de l'anàlisi.

9. Exemples i casos pràctics

En aquest apartat presentarem una sèrie d'exemples i casos pràctics d'aplicació dels conceptes introduïts en el mòdul en diversos camps, els quals ja han estat breument esbossats en la «Introducció».

9.1. Control estadístic de qualitat

El control de qualitat de processos productius és una àrea en la qual tradicionalment s'han utilitzat tècniques estadístiques per a diverses comeses, i s'ha generat fins i tot una disciplina anomenada *control estadístic de qualitat* o *control estadístic de processos*.

Les aplicacions són múltiples, des dels estudis de repetibilitat i reproductibilitat (en els quals es busca calibrar un instrument de mesura) fins als de capacitat de processos (que avaluen fins a quin punt un determinat procés és capaç de satisfer uns requisits de qualitat prèviament establerts), passant pels gràfics de control més clàssics (l'objectiu dels quals és dissenyar un sistema d'informació contínua que permeti detectar amb antelació l'aparició de causes especials de variabilitat, identificar-ne l'origen i incorporar mesures correctores).

Necessitat

El director de producció d'una fàbrica de vidres per a automòbils està amoïnat perquè ha augmentat el percentatge de peces produïdes que són rebutjades pels concessionaris de venda de cotxes, ja que no compleixen les mesures previstes.

Per a això, encarrega al director de qualitat un estudi que permeti establir si les característiques de qualitat han variat respecte a les especificacions inicials i, en cas afirmatiu, conèixer el motiu més probable d'aquesta variació.

Objectiu i hipòtesi

En aquest cas, ens centrarem en l'objectiu de dissenyar i implantar un gràfic de control perquè suposa un projecte ampli, que moltes vegades comporta objectius secundaris com, per exemple, establir la capacitat del procés.

El disseny d'un gràfic de control no persegueix altra cosa que representar gràficament l'evolució en el temps d'un estadístic obtingut a partir de mostres del procés preses periòdicament. Un element clau del mateix és establir els anomenats *límits de control*, que es representen amb línies, una per damunt i una altra per sota de la línia central (habitualment la mitjana); es produeix un

senyal de falta de control (és a dir, un avís que el procés podria no estar funcionant com cal, ja que el comportament és molt poc probable en condicions normals) quan un punt representat se situa fora dels límits de control.

Planificació

Si ens basem en la proposta de Fayyad, citada anteriorment, en aquest cas les fases serien les següents:

1) Comprensió del camp d'aplicació.

Com gairebé sempre, és fonamental que hi hagi un treball conjunt entre els experts en l'àmbit d'estudi i els analistes quantitatius, perquè uns puguin comprendre correctament les necessitats dels altres i plasmar-les convenientment.

Els fonaments teòrics del control estadístic de qualitat no són excessivament complexos, ja que es basen en la clàssica inferència estadística, però sempre existeixen aspectes específics i particulars que poden condicionar l'anàlisi, per la qual cosa és important tenir-los en compte.

2) Creació d'una base de dades.

La majoria de processos industrials de fabricació són complexos i, per tant, difícils de mesurar. En el cas que ens ocupa, un vidre d'automòbil ha de complir una sèrie de requisits de grandària, gruix, duresa, flexibilitat, transparència, etc., la qual cosa provoca que una correcta selecció de les dades d'interès sigui prioritària, perquè cadascuna de les característiques es mesura en una fase del procés i per una maquinària diferent en la majoria dels casos.

3) Preparació de les dades.

En els inicis del control estadístic de processos, els mesuraments es realitzaven a peu de fàbrica i per persones per a garantir la immediatesa, i es traslladaven a un full de paper. Això provocava una alta taxa d'errors en la recollida de dades i, en conseqüència, que la fase de creació de la base de dades i neteja de la mateixa fos tediosa.

Actualment, la majoria de processos estan controlats per sofisticats aparells electrònics que realitzen la recollida de dades, amb la qual cosa les possibilitats d'error s'han reduït considerablement.

4) Técnica de modelització.

Com ja s'ha comentat, en aquest cas l'elecció de la tècnica de modelització acostuma a ser un gràfic de control, si bé n'hi ha de múltiples tipus (univariables o multivariables, per variables o per atributs, amb grandària de mostra constant o variable, etc.) i l'elecció del més adequat serà com sempre un punt important del projecte.

5) Cerca de patrons.

La cerca de patrons és clau en aquest tipus d'estudis, perquè es tracta de localitzar comportaments anòmals per a identificar les seves causes i corregir-les. Per a això, primer caldrà establir quin és el patró normal i també què s'entén per comportament anòmal, el qual es definirà habitualment en termes de la probabilitat (molt baixa) que passi en condicions normals.

La importància de la visualització dels resultats aquí és evident perquè, com ja hem comentat, el suport principal dels mateixos és un gràfic, encara que aquest pugui estar recolzat per dades numèriques.

6) Explotació dels resultats.

Una vegada obtingut un primer gràfic de control, aquest es reutilitza per al seguiment continu del procés, si bé periòdicament el seu disseny haurà de ser revisat per a tenir en compte variacions que puguin haver-se produït en el procés després de l'estudi inicial.

Mostreig

En aquest tipus de problemes, el mostreig més habitual és el sistemàtic; és a dir, s'escull una peça d'inici i a partir d'ella s'escullen les següents a intervals fixos, ja siguin temporals, o de nombre de peces.

9.2. Anàlisi de riscos en banca

Necessitat

Com ja hem comentat, en aquest cas la necessitat que genera l'anàlisi quantitativa és minimitzar la morositat provocada per l'impagament dels préstecs i crèdits concedits, ja que això té un impacte important en el compte de resultats dels bancs, no només el directe motivat per la minva d'ingressos, sinó també l'indirecte, derivat del fet que els possibles impagaments s'han d'aprovisionar; és a dir, s'ha de reservar capital per a afrontar les pèrdues futures que generaran els possibles impagaments, i així evitar la fallida del banc.

Objectiu i hipòtesi

La necessitat plantejada en el punt anterior es tradueix en una pregunta que de forma molt bàsica podria ser la següent: a qui he de prestar diners? I de forma una mica més elaborada podria ser: qui és més probable que em retorni els diners d'acord als terminis i interessos estipulats?

Per això, l'objectiu és mirar de determinar quins dels clients susceptibles de rebre un préstec (o un altre producte d'actiu, com un crèdit, una targeta, etc.) tenen un millor perfil de pagament, basant-se en les seves característiques socioeconòmiques i en el seu comportament previ (si es tracta de clients coneguts).

Un banc disposa d'informació dels préstecs concedits i de certes característiques dels seus clients, tant d'aquelles que s'han preguntat en el moment de la concessió (generalment, informació socioeconòmica –edat, estat civil, professió, antiguitat en l'ocupació, etc.– o d'ingressos –nòmina– i despeses –lloguer, altres càrregues o quotes prèvies–), com de comportaments previs de pagament si ja eren clients (rebuts domiciliats, compliment dels pagaments de quotes d'altres productes, ingressos recurrents, dipòsits o altres productes d'estalvi, etc.).

Tot això per al cas de persones físiques, perquè si es tracta d'empreses la informació és diferent (balanç o compte de resultats –facturació, beneficis, liquiditat, etc.–, qüestionaris qualitatius –sector en el qual exerceix la seva activitat, nombre d'empleats, tipus de productes o serveis que ofereix, etc.–, i també, per descomptat, comportaments previs amb l'entitat), però assimilable.

La hipòtesi subjacent aquí és la següent: si establim un perfil de client amb bona qualificació basat en informació passada, podrem veure fins a quin punt s'assemblen els nous clients a aquest perfil, i d'aquesta forma discriminar quins tindran més probabilitat de ser bons pagadors.

Planificació

En aquesta ocasió seguirem les fases proposades pel CRISP-DM:

1) Comprendre el negoci o problema.

En aquest cas la comprensió del problema és capital, perquè l'activitat bancària és un negoci molt variat i particular: cada producte té unes característiques diferents (uns són renovables amb certa periodicitat i uns altres no, les formes de pagament són diverses), com també en té cada tipus de client (persones físiques enfront de persones jurídiques, particulars enfront d'autònoms, microempreses enfront de pimes enfront de grans corporacions). És molt important que en aquest punt els analistes quantitatius recopilin la major informació possible i de fonts molt diverses com, per exemple, els analistes d'operacions de les oficines bancàries.

2) Comprendre les dades.

La informació de què disposa un banc sobre els seus clients és molt rica i diversa, però també està subjecta a múltiples problemes relatius a la qualitat de les dades, motivats per fonts diverses, períodes temporals diferents, formats heterogenis, introducció de dades manuals, etc. És, doncs, molt important que s'estableixin controls estrictes i rigorosos sobre la informació disponible en les bases de dades internes.

Tanmateix, la comprensió de la tipologia de dades per part de l'analista quantitatiu és també un aspecte rellevant.

3) Preparar les dades.

Molt lligat amb l'etapa anterior, en aquesta s'ha d'aprofitar la informació recaptada i, a partir d'una anàlisi de les característiques de cada variable disponible (anàlisi univariable i/o bivariable), seleccionar aquelles que tenen una qualitat acceptable i descartar les que n'estan mancades (variables amb molts valors perduts, amb comportaments anòmals, etc.), generar noves variables (ràtios, ponderacions, combinacions, etc.).

4) Modelitzar.

En aquest tipus de models, un aspecte molt important és definir els criteris per a construir la variable depenent. És obvi que si el que es pretén és predir la possible morositat dels clients, la variable depenent ha d'estar relacionada amb la mora (i, per tant, serà binària, ja que només pot tenir dos possibles valors: SÍ o NO). No obstant això, existeixen diferents formes de definir la morositat, per la qual cosa una correcta elecció pot facilitar-ne la modelització.

Prèviament al desenvolupament del model, també és important una etapa de selecció de variables i reducció de la dimensió, ja que normalment la quantitat de factors de què es disposa és molt elevada. Aquí s'acostuma a fer ús d'una anàlisi bivariable (per a determinar quines de les variables tenen una major associació amb la mora de forma individual) i d'una anàlisi de correlacions (per a evitar introduir en el model factors amb alta correlació que puguin originar un problema de col·linealitat).

D'altra banda, quant a la tècnica estadística adequada, en són vàlides moltes (arbres de classificació, anàlisi discriminant, xarxes neuronals, etc.), però la més estesa en el sector financer és la regressió logística, per la seva senzillesa i perquè s'adapta bé al problema definit: variable depenent binària i resposta en termes de probabilitat.

5) Validar el model.

a) Avaluar els resultats.

Aquesta fase, que sempre és molt important, consisteix a utilitzar tècniques de validació per a comprovar que el model funcionarà bé per a l'ús previst, és a dir, la predicció. Per això, cal descartar possibles problemes de sobreajustament, i per a fer-ho es poden emprar les típiques tècniques de validació com ara *split* de mostres, validació encreuada, *bootstrapping*, validació *out-of-time* o *out-of-sample*, etc.

6) Explotar el model.

Una vegada desenvolupat i validat un model, s'ha d'incorporar als sistemes operacionals del banc per a poder ser utilitzat en futures concessions d'operacions d'actiu. Per això, en el moment de la seva implantació cal realitzar proves d'usuari que permetin comprovar que el model desenvolupat s'ha implementat en els sistemes amb correcció.

Mostreig

En aquest tipus de models hi ha diverses possibilitats: no és infreqüent que un banc utilitzi totes les seves dades disponibles per a construir el model, i en aquest cas no existeix mostreig sinó que s'utilitza tota la població.

En cas d'utilitzar una mostra, el mostreig pot ser aleatori simple sense reemplaçament, o bé pot ser estratificat, si es desitja respectar les proporcions d'algun factor de la població en la mostra. Per exemple, el segment de client o el tipus de producte.

9.3. Enquestes de satisfacció

Necessitat

El gerent d'una empresa de telefonia està amoïnat perquè ha observat, en les estadístiques internes que li facilita el departament comercial, que el balanç de clients (és a dir, la diferència entre el nombre de nous clients i el nombre de clients antics que han deixat de ser-ho) és negatiu. Per a això, sol·licita al director de màrqueting que elabori una estratègia per a canviar aquesta situació.

Objectiu i hipòtesi

El director de màrqueting es planteja, doncs, dos objectius: d'una banda, mirar de fidelitzar els clients actuals, i, per l'altra, intentar ajudar el departament comercial a captar nous clients i ajudar el departament de publicitat a enfocar les noves campanyes. Per a això dissenya dues enquestes: una l'objectiu de la qual és conèixer quin és el nivell de satisfacció dels clients actuals, per a

evitar-ne la fugida; i una altra l'objectiu de la qual és conèixer quina és la percepció que es té de les tarifes i del servei per part dels no clients, per a intentar detectar quins podrien estar interessats a contractar-lo.

Aquí és important destacar que, atès que es tracta de dos objectius que, si bé poden resultar complementaris, són diferents, no és factible (almenys de forma senzilla) dissenyar un estudi que permeti cobrir-los tots dos alhora.

Planificació

El disseny de l'enquesta és molt diferent per als casos que ens ocupen:

En el primer cas es tractaria de preguntes amb les típiques escales de satisfacció segons nivells d'acord amb els diversos productes i serveis contractats, l'atenció del servei postvenda, les tarifes aplicades, etc. Així mateix, la companyia disposa d'informació sobre els clients mitjançant la qual poder segmentar-los (edat, antiguitat com a client, zona de residència, tipus de contracte, franges horàries de major ús, etc.).

En el segon cas es tractaria de preguntes més obertes, per a mirar de detectar opinions sobre la companyia (coneixement espontani de la mateixa, referències de coneguts que són clients, percepció de l'adequació de les seves tarifes, percepció de la imatge de marca, etc.) però sense induir les respostes.

El tractament de la informació recaptada també serà diferent en alguns aspectes:

En el primer cas caldrà encreuar la informació oferta pels clients amb les dades internes, per a detectar possibles inconsistències. Així mateix, és freqüent descartar certs casos en els quals les respostes no tenen amb prou feines variabilitat: sempre es proporciona la millor valoració, o la pitjor, o la mitjana.

En el segon cas és molt probable que existeixin diversos comportaments típics en enquestes massives que provoquen que calgui descartar certes respostes: falta d'informació, incoherència entre les respostes, opinions difícils de codificar, etc.

Mostreig

Per a la primera enquesta, l'objectiu de la qual és conèixer la satisfacció dels clients actuals, el director de màrqueting decideix aplicar un mostreig aleatori simple. A més, per a fomentar la resposta, l'empresa sorteja un telèfon intel·ligent entre els clients que responguin l'enquesta.

Per a la segona enquesta, el disseny del mostreig és molt més obert i complex, ja que la població no és coneguda i, per tant, no és fàcil aplicar un mostreig probabilístic. Després de valorar altres alternatives (enviament massiu, mos-

treig per conveniència), el director de màrqueting opta per un mostreig per bola de neu, ja que es pretén fer ús de les xarxes socials per a buscar públic jove i se sap que aquest tipus de tècnica funciona raonablement bé en aquests entorns. A més, una enquesta en línia acostuma a ser més barata que una per correu o telefònica, i en aquests casos, com que la taxa de resposta és baixa, acostuma a ser habitual una grandària mostral molt alta.

9.4. Indicadors sintètics

Necessitat

Un alcalde està interessat a obtenir una valoració de l'activitat cultural de la seva ciutat i dels barris que la componen, per a comparar-la amb la d'altres ciutats del voltant, reforçar els serveis culturals en aquells barris on es detectin mancances i controlar l'evolució d'aquesta valoració al llarg del temps.

Objectiu i hipòtesi

L'objectiu en aquest cas consisteix a obtenir un indicador únic que resumeixi la informació de les activitats culturals de què disposa l'ajuntament per a poder establir les comparatives desitjades mitjançant un rànquing únic.

Planificació

Seguint novament les fases proposades per Fayyad:

1) Comprensió del camp d'aplicació.

Com ja s'ha comentat en més d'una ocasió, aquesta fase és molt important. A més, tractant-se d'un concepte tan ampli com la cultura, és convenient fixar prèviament el concepte i l'abast de l'estudi. Per això, el més habitual és comptar amb l'opinió i l'assessorament dels experts en la matèria i, per descomptat, també recollir informació d'estudis similars que s'hagin realitzat amb anterioritat.

2) Creació d'una base de dades.

Existeixen multitud de dades sobre les activitats culturals que es duen a terme en una ciutat, des del nombre de biblioteques, sales de teatre, cinemes, institucions culturals i l'ús de les mateixes (llibres prestats, assistència a espectacles, etc.), fins a la despesa pública destinada a cultura o el volum de producció de llibres o mitjans de comunicació.

3) Preparació de les dades.

Aquesta fase sempre consumeix molt de temps, però del seu nivell de qualitat depèn en gran part l'èxit final del projecte.

En aquest cas resultarà important que la informació recollida sigui homogènia a tots els barris i en els períodes temporals, que no hi hagi gaires dades no disponibles, que les unitats en les quals es recullen les dades siguin, si no iguals, sí assimilables, etc.

4) Modelització.

Atès que l'objectiu principal de l'estudi, des d'un punt de vista estadístic, és la reducció de la dimensió, la tècnica adequada serà molt probablement l'anàlisi de components principals, si bé es pot recolzar en un altre tipus de tècniques com, per exemple, l'anàlisi clúster.

5) Cerca de patrons.

En aquest tipus de projectes és molt rellevant aconseguir que les dimensions obtingudes en la fase anterior siguin interpretables des d'un punt de vista qualitatiu. Per això, és recomanable un treball conjunt entre l'expert en l'àmbit d'estudi i l'analista quantitatiu.

També és molt important visualitzar els resultats mitjançant gràfics tradicionals o noves tècniques de visualització que permetin el seu ràpid i fàcil enteniment per part d'un públic no expert en tècniques quantitatives.

6) Explotació dels resultats.

Una vegada obtinguts els resultats quantitatius, es tractaria d'incorporar-los a les estadístiques periòdiques generades pel servei estadístic de l'ajuntament, de manera que siguin accessibles, bé per a públic restringit o bé per a tots els ciutadans.

Mostreig

En aquest tipus de casos no acostuma a ser habitual fer mostreig, ja que l'ajuntament disposa d'un servei estadístic que recull tota la informació necessària per a poder realitzar l'estudi i, per tant, la població a analitzar són tots els barris o districtes que formen la ciutat.

9.5. Anàlisi de supervivència en medicina

Necessitat

Un laboratori farmacèutic té interès a analitzar l'efecte d'un dels seus fàrmacs a l'hora d'augmentar el temps de vida i millorar-ne la qualitat en pacients diagnosticats amb un tipus de càncer.

Objectiu i hipòtesi

L'objectiu principal és comparar el temps de vida (supervivència) de diferents grups als quals es prescriuen tractaments diferents. El cas més habitual i senzill és disposar de dos grups: un als membres del qual s'administra el fàrmac i un altre als qui no, anomenat *grup de control*.

Altres objectius secundaris són: analitzar possibles efectes secundaris i establir factors de risc (diferents del fàrmac utilitzat) que millorin o empitjorin el pronòstic de supervivència.

Planificació

Seguint de nou les fases del CRISP-DM, que, com ja s'ha comentat, és una metodologia de planificació de projectes quantitatius, tindríem:

1) Comprendre el negoci o problema.

Una vegada més, la comprensió del problema és capital, encara més que en els altres exemples pel fet d'estar tractant amb vides humanes.

Un handicap important en aquest tipus d'estudis és també la utilització de l'argot típic d'un col·lectiu, en aquest cas el mèdic, que no acostuma a ser de domini comú, almenys per part d'analistes quantitatius. És per tant clau que hi hagi una primera fase d'enteniment entre els objectius marcats per l'investigador principal i els plantejats per l'expert estadístic.

Un altre inconvenient acostuma a ser la dificultat d'obtenció de dades. Cal tenir en compte que no es tracta d'experiments senzills i controlables, atès que la realitat de cada pacient és diferent, la qual cosa complica l'homogeneïtat de la informació.

2) Comprendre les dades.

Molt lligat amb el punt anterior, el correcte enteniment de les dades obtingudes, els seus biaixos i limitacions, són un aspecte a tenir en compte.

Per exemple, en aquest tipus d'estudis és habitual que els temps d'inici de l'estudi (del tractament, si és el cas) siguin diferents per a cada individu, que hi hagi subjectes que siguin exclosos per circumstàncies alienes a l'estudi (per exercir el dret a la protecció de dades, per deixar el tractament voluntàriament,

per una defunció per causes diferents de les de la malaltia objecte d'estudi, etc.). Totes aquestes circumstàncies influeixen de manera rellevant en les decisions que s'han de prendre des d'un punt de vista quantitatiu.

3) Preparar les dades.

En aquesta fase pren especial rellevància la integració de dades de diverses fonts, ja que és probable que el diagnòstic i el seguiment de la malaltia no es faci al mateix centre mèdic, que es recullin també dades de qüestionaris qualitatius juntament amb mesures clíniques, que hi hagi judicis subjectius per part tant del metge com del pacient, que l'obtenció de la informació sigui progressiva (amb la qual cosa a l'inici del projecte no es disposa de totes les dades), que alguna variable clau no es pugui aconseguir (i per tant s'utilitzi alguna variable o indicador que exerceix com a aproximació d'una altra variable d'interès que no s'hagi pogut aconseguir en el seu lloc), entre d'altres.

4) Modelitzar.

A diferència de la majoria d'exemples restants, en aquest tipus de casos la selecció de la tècnica adequada no és un problema perquè l'anomenada *anàlisi de supervivència* compleix amb els requisits per a ser utilitzada i és l'estàndard habitual, gairebé únic; i va sorgir per a donar resposta a aquestes necessitats.

No obstant això, sí que és molt important el disseny de l'experiment. Com ja hem comentat en punts anteriors, aquest tipus d'estudi pateix diverses limitacions i hi ha el risc que a causa d'això les conclusions obtingudes no siguin fiables. La correcta elecció del grup de control, per exemple, serà clau per a l'èxit de l'experiment.

És freqüent que també hi hagi un biaix de selecció de la població, ja que determinats laboratoris treballen habitualment amb uns metges o hospitals concrets, i, per tant, no és senzill obtenir dades prou àmplies i variades.

5) Validar el model.

Aquesta fase no difereix significativament pel que fa a altres exemples comentats.

6) Explotar el model.

La correcta difusió i explotació d'aquest tipus de models és molt important per a la comunitat investigadora i la societat en general.

Mostreig

Com ja s'ha comentat, existeixen certes dificultats que impedeixen realitzar un mostreig probabilístic en aquest tipus d'estudis: escassa grandària de la població, restringit a uns quants (o pot ser que un únic) centres hospitalaris, etc.

És freqüent, per tant, que es treballi amb tota la població disponible i no amb una mostra.

Resum

En aquest mòdul hem revisat tots els aspectes relacionats amb el plantejament d'un estudi quantitatiu.

Els primers apartats introdueixen el lector en les necessitats que poden motivar un estudi quantitatiu i la seva traducció en objectius i hipòtesis de treball.

Un dels capítols centrals, tant en extensió com en importància, és el dedicat a la planificació del projecte, on s'expliquen les diverses fases que s'han d'emprendre des de les propostes de diversos autors que, com es pot comprovar, són força similars entre elles. Convé no oblidar que aquestes fases no cal que es desenvolupin de forma lineal, sinó que el més habitual és fer-ho de forma cíclica.

Posteriorment s'introdueixen els conceptes bàsics de la teoria del mostreig i els principals tipus de mostreig, ja que cal recordar que en moltes ocasions un estudi quantitatiu ha de recórrer a la selecció d'una mostra per a analitzar la informació.

Els dos apartats següents se centren en dues fases molt importants a l'inici dels projectes quantitatius i íntimament lligades entre elles: l'anàlisi de la qualitat de la informació i la preparació de les dades. De la correcta i completa realització d'aquestes fases depèn en gran part l'èxit final de l'estudi.

Amb posterioritat es dedica un breu capítol a un possible model d'estructura d'informe que permeti plasmar i difondre els resultats del projecte.

Finalment es desenvolupen una sèrie d'exemples i casos pràctics d'aplicació dels conceptes introduïts en el mòdul en diversos camps.

Bibliografia

Diversos Autors (2000). *CRISP-DM 1.0. Step-by-step data mining guide*. IBM Corporation.

Emerson, S. S. (2013). *Organizing your approach to a data analysis*.

Eurostat (2007). *Handbook on data quality assessment methods and tools*.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. (1996). «From Data Mining to Knowledge Discovery in Databases». *AI Magazine* (vol. 17, núm. 3, pàg. 37-54).

Hernández, R.; Fernández, C.; Baptista, P. (2010). *Metodología de la investigación*. Mèxic: McGraw-Hill.

Pérez López, C. (2010). *Técnicas de muestreo estadístico*. Madrid: Garceta.

Wickham, H. (2014). «Tidy Data». *Journal of Statistical Software* (núm. 59, pàg. 1-23).