Xavier Bardina
Mercè Farré

ESTADÍSTICA DESCRIPTIVA

Universitat Autònoma de Barcelona
Servei de Publicacions
Bellaterra, 2009

DADES CATALOGRÀFIQUES RECOMANADES PEL SERVEI DE BIBLIOTEQUES DE LA UNIVERSITAT AUTÒNOMA DE BARCELONA

Bardina, Xavier

Estadística descriptiva / Xavier Bardina i Mercè Farré. — Bellaterra : Universitat Autònoma de Barcelona. Servei de Publicacions, 2009. — (Manuals de la Universitat Autònoma de Barcelona ; 54)

ISBN 978-84-2590-7

I. Col·lecció
1. Estadística — Mètodes gràfics

519.2

Aquest llibre s'ha publicat amb la col·laboració de la Generalitat de Catalunya

Primera edició: 2009

Edició i impressió:
Universitat Autònoma de Barcelona
Servei de Publicacions
Edifici A
08193 Bellaterra (Cerdanyola del Vallès). Spain
Tel.: 93 581 10 22. Fax: 93 581 32 39
sp@uab.cat
http://publicacions.uab.cat/

ISBN 978-84-490-2590-7
Dipòsit legal: B.14589-2009

Imprès a Espanya. Printed in Spain

Índex

Presentació

Aquest manual és un text introductori a l'estadística descriptiva. Per llegir-lo no cal tenir cap coneixement previ ni sobre probabilitat ni sobre estadística, i totes les demostracions es poden seguir amb el nivell matemàtic del batxillerat científic.

Tots els conceptes i resultats que s'estudien van acompanyats de força exemples i exercicis resolts. S'ha procurat en l'exposició que el text reculli tots els comentaris i aclariments que es donarien si s'expliqués aquesta matèria a la pissarra en una aula amb estudiants. El text està pensat per ser explicat durant un semestre acadèmic.

Al final de cada capítol hi ha una llista d'exercicis, alguns resolts i d'altres proposats. Aquests exercicis són un test perquè el lector pugui veure si ha entès els conceptes que s'han introduït.

El text consta de 7 capítols o temes. Un primer capítol de preliminars en què s'explica la diferència entre l'*estadística descriptiva*, que és l'objecte d'estudi d'aquests apunts, i l'*estadística inferencial*. En el segon capítol es veu detalladament l'estadística descriptiva d'una variable. És a dir, s'estudien tot un conjunt de taules, gràfics i mesures que ens permeten resumir la informació que contenen les dades, però estudiant cada variable separadament. En els tres capítols següents s'estudia l'estadística descriptiva bivariant, segons els tipus de variables que volem estudiar conjuntament. Així, en el tercer capítol s'estudia com comparar una variable numèrica en dos o més grups, en el capítol 4 s'analitza la relació entre dues variables categòriques, mitjançant les taules de contingència, i en el capítol 5 es veuen les eines necessàries per estudiar la relació entre dues variables numèriques. Un cas particular que mereix ser estudiat a part, és quan una de les dues variables és el temps. En aquest cas parlem de *sèries temporals*, i en el sisè

capítol d'aquests apunts farem una introducció a aquest tema. Finalment, en el darrer capítol introduïm els índexs simples i els sintètics.

El paquet estadístic que hem utilitzat com a eina de càlcul ha estat l'SPSS. En alguns llocs del manual indiquem com fer els càlculs amb aquest paquet estadístic, però el lector podrà fàcilment adaptar-ho a altres paquets.

El lector pot trobar semblances entre algunes parts d'aquest text i alguns capítols del text *Estadística: un curs introductori per a estudiants de ciències socials i humanes*, que trobareu referenciat a la bibliografia. No obstant això, aquest manual aprofundeix més en els conceptes de l'estadística descriptiva i conté demostracions que aquell text no conté perquè s'orienta a un curs introductori d'estadística aplicada a les ciències socials.

Capítol 1

Preliminars

En aquest primer capítol s'explica la diferència entre *estadística descriptiva* i *estadística inferencial* i es treballen dos aspectes crucials: la classificació de les variables, i les notacions científiques i els arrodoniments correctes.

Els punts d'aquest apartat, si bé preliminars, són fonamentals per a la correcta aplicació de tots els mètodes presentats en capítols posteriors.

El fet és que, per exemple, dades que no són numèriques, com les que classifiquen els individus en funció de la seva situació laboral, sovint es codifiquen numèricament quan s'introdueixen en una base de dades; aleshores els programes estadístics les tracten com a tals i donen resultats ben absurds. És l'usuari, i no pas el programa, qui decideix els procediments aplicables en funció de la tipologia de les dades.

1.1 L'estadística: introducció

L'estadística tracta de metodologies i eines per tal de recollir dades, resumir-les, explorar-les, analitzar-les i, si és possible, extreure'n conclusions a un nivell més ampli que el del context experimental.

Les dades de les quals s'ocupa l'estadística són les que presenten variabilitat o *incertesa*. Els estudiosos, tècnics i científics o investigadors de les diferents disciplines (humanitats, ciències socials i experimentals), quan recullen dades del seu interès, es troben sovint amb resultats que varien molt d'un individu o objecte a un altre, aparentment sense una funcionalitat que ho expliqui: estem parlant d'incertesa.

En el llenguatge estadístic planer, una **població** és un conjunt d'individus o objectes, i una **mostra** és un subconjunt d'aquests. De manera més rigorosa, convé identificar població amb tot el conjunt d'observacions possibles de la variable que hom estigui estudiant.

L'estadística té dues vessants: l'**estadística descriptiva** i l'**estadística inferencial**.

La descriptiva s'ocupa de resumir i explorar les dades, mitjançant tabulació, gràfics i càlcul de característiques o indicadors descriptius.

L'estadística inferencial pretén extrapolar els resultats de les dades obtingudes en una mostra aleatòria (escollida a l'atzar), a un conjunt més gran o població.

Amb un parell d'exemples veurem ben clara la diferència entre els dos tipus d'estadística. Quan fem un **cens**, volem obtenir informació de tota la població (o del màxim d'individus possible). El nostre objectiu serà resumir la informació que hem obtingut, i per tant estarem fent **estadística descriptiva**. En canvi, quan es fa una **enquesta** primer cal resumir la informació de les dades, i per tant estem fent estadística descriptiva, però el nostre objectiu no és aquest, sinó que volem poder dir coses de tota la població, no només de la mostra, i per tant necessitem l'**estadística inferencial**. Precisament, *inferir* vol dir *passar d'una part a un tot.*

En aquest curs, veurem només l'estadística descriptiva, la qual té interès per si mateixa i és un pas previ per a l'estadística inferencial.

1.2 Tipus de variables. Escales de mesura

Una **variable** és tota mesura que és susceptible de prendre més d'un valor: no és una constant. D'una variable se'n diu estadística o també aleatòria quan d'antuvi (abans d'obtenir les observacions) hi ha un cert grau d'incertesa en els valors que s'observaran.

Hi ha variables **numèriques o quantitatives**, com per exemple l'edat, i **no numèriques o qualitatives**, com per exemple el sexe. Les variables qualitatives també s'anomenen *atributs.* Els valors d'una variable qualitativa sovint s'anomenen també *modalitats.*

Una classificació de les variables numèriques distingeix entre variables **discretes** i **contínues**. Una variable és contínua si teòricament pot ser mesurada en una escala tan fina com vulguem; a la pràctica, que la mesura sigui més o menys acurada depèn dels instruments utilitzats, de com es codifiquen les dades i de la precisió que vol l'investigador. Exemples de

variables contínues són el temps, el pes dels individus, etc. El temps, per exemple, pot ser mesurat en mil·lennis, segles, anys, semestres, setmanes, dies, hores, segons, dècimes de segon, centèsimes de segon, etc.

Les variables contínues poden prendre una quantitat no numerable (no comptable) de valors. Les variables que no són contínues s'anomenen discretes, i poden prendre només una quantitat numerable de valors. Les variables que compten el nombre d'individus que tenen certa característica són un exemple clar de variables discretes. Per exemple, nombre de consumidors habituals d'alcohol, nombre de votants de cert partit, etc.

Quan es vol avaluar el resultat d'un experiment o d'una observació, l'**escala de mesura** depèn del tipus de fenomen, dels aparells utilitzats per fer els mesuraments, de la capacitat de recol·lecció d'informació, de les necessitats de l'investigador, etc. Distingirem entre:

- **L'escala nominal**. Quan estem observant un fenomen, l'escala nominal de mesura és la que defineix com a valors de la variable observada un conjunt de modalitats. Les modalitats han de ser mútuament excloents o disjuntes i han de cobrir tots els possibles valors (ser exhaustives). Exemples: sexe, estat civil, etc. són nominals de manera natural. D'altres fenòmens, que podrien ser mesurats en escales més fines, es poden reduir a l'escala nominal agrupant els valors en classes o categories mútuament excloents i exhaustives, si a l'investigador li interessa recollir únicament aquesta informació. La divisió en categories pot no ser única. Per exemple, l'afinitat política a Catalunya es pot classificar com: nacionalista d'esquerres, nacionalista de dretes, no nacionalista d'esquerres, no nacionalista de dretes i altres sentiments; però hi hauria altres classificacions possibles. *L'escala nominal només permet la classificació dels individus.* És la més simple.

- **L'escala ordinal**. És un pas més de l'escala nominal que pot utilitzarse per a observacions que prenen valors que admeten un ordre natural. Aquest no és el cas del sexe o de l'estat civil. L'escala ordinal estableix un ordre en les categories. Per exemple, l'estatus social admet una mesura ordinal: classe alta, classe mitjana i classe baixa. *L'escala ordinal permet classificar i establir ordres.*

- **L'escala numèrica**. Quan un fenomen pot ser mesurat de manera numèrica: temps, pes, nombre de casos, índex d'atur, taxa de delinqüència, preu de l'aigua, salari, nombre de fills, etc., se sol adoptar aquesta escala perquè és més fina que les anteriors. *L'escala numèrica,*

a més de classificar i establir ordres, permet operar amb els valors de la variable: sumar-los, calcular-ne valors mitjans i, en general, operar aritmèticament amb aquests.

1.3 Arrodoniments i notació científica

Quan es fan operacions numèriques, cal ser acurat amb els arrodoniments, per tal d'evitar l'acumulació d'errors. La regla és arrodonir a la xifra més propera. Per exemple, 346.8* arrodonit a enters és 347, l'enter més proper. Anàlogament, 123.3467 té set xifres significatives, quatre de les quals són decimals; si el volem expressar en cinc xifres significatives (només dos decimals), el nombre arrodonit correctament és 123.35. Anàlogament, si volem arrodonir 165 400 a milers, posarem 165 milers.

La regla consisteix, doncs, a observar el primer dígit a la dreta de la posició a la qual es vol arrodonir.

Si el dígit situat a la dreta és 0, 1, 2, 3 o 4, es manté el dígit anterior, és a dir, s'arrodoneix "per sota"; així:

312.3 arrodoniment enter: 312
0.453 arrodoniment amb dos decimals: 0.45

Si el dígit situat a la dreta és 6, 7, 8, o 9, s'afegeix un 1 al dígit anterior, és a dir, s'arrodoneix "per sobre"; així:

312.6 arrodoniment enter: 313
0.458 arrodoniment amb dos decimals: 0.46

Finalment, si el dígit següent és només un 5, aleshores hi ha diferents criteris, un dels quals és, per exemple, arrodonir una vegada per sota i la següent per sobre, en cas que es facin diversos arrodoniments.

Aquestes directrius s'han de tenir en compte quan es presenten els resultats finals i quan, en un procés, es fan nombrosos càlculs consecutius amb la consegüent acumulació d'errors.

La notació científica consisteix a fixar un nombre de xifres significatives i expressar la resta en potències de 10. Per exemple, $864\,000\,000 = 864 \times 10^6 = 8.64 \times 10^8 = ...$ i $0.0003456 = 3456 \times 10^{-7} = 3.456 \times 10^{-4} = ...$

La notació científica la utilitzen les calculadores, els fulls de càlcul (Excel, per exemple) i els paquets estadístics (SPSS i d'altres). L'expressió que solen presentar per pantalla és, per exemple:

$7.1712\,\mathrm{E}-0.6$ vol dir $7.1712 \times 10^{-6} = 0.0000071712\,,$
$7.1712\,\mathrm{E}+0.6$ vol dir $7.1712 \times 10^{6} = 7\,171\,200\,.$

* S'utilitza la notació internacional: el punt separa les xifres decimals.

Capítol 2

Estadística descriptiva d'una variable

L'objectiu de l'estadística descriptiva és resumir, de la manera més concisa, entenedora i visual possible, tota o bona part de la informació rellevant que contenen les dades.

Si bé alguns conceptes s'entenen millor quan comparem dues variables o una mateixa variable en dos grups diferents, en aquest capítol intentarem limitar-nos al tractament aïllat de les dades d'una única variable, tot presentant els procediments i les eines més bàsics.

Per establir les bases d'una producció i interpretació correctes, en gairebé tots els apartats es comenta el fet diferencial que suposa l'opció de *ponderar o no ponderar les dades*, i es posa de manifest que s'obtenen resultats diferents en funció de si les *dades estan agrupades o no* en classes o intervals.

És molt important presentar correctament la informació resumida, sigui gràficament, mitjançant taules, o sobre la base d'un conjunt d'indicadors. No ens referim a una bona presentació des del punt de vista estètic, que també és important, sinó a la correcció ètica i professional: fer referència a les fonts d'informació, no ometre cap informació rellevant, facilitar la comprensió del lector i no manipular l'observador amb escales magnificades ni amb altres proeses gràfiques.

Les seccions d'aquest capítol comencen per la tabulació de les dades, segueixen amb les representacions gràfiques i continuen amb la presentació, el càlcul i la interpretació dels indicadors descriptius més importants, els idonis per a cada tipus de variable. La **secció** 2.4, dedicada a l'anàlisi exploratòria, i les últimes seccions tracten d'alguns complements, entre d'altres