Jordi Esquirol Caussa, Elisabeth Herrero Vila, Josep Sánchez Aldeguer

Metodologia i cstadística per a professionals de la salut

III. L'ANÀLISI ESTADÍSTICA

Escola Universitària d'Infermeria i de Fisioteràpia «Gimbernat»

Escola Universitària de Fisioteràpia

Universitat Autònoma de Barcelona
Servei de Publicacions
Bellaterra, 2012

Primera edició: juny de 2012

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/

Imprès a Espanya. Printed in Spain

Dipòsit legal: B-10975-2012
ISBN 978-84-490-2865-6

ÍNDEX

AUTORS

Jordi Esquirol Caussa

Doctor per la Universitat Autònoma de Barcelona, al programa de Medicina Interna. Màster en Bioètica i Dret: Problemes de Salut i Biotecnologia (Universitat de Barcelona). Màster en Medicina Preventiva i Promoció de la Salut (Universitat de Barcelona). Màster en Gerontologia Clínica (Universitat Autònoma de Barcelona). Responsable del Servei Universitari de Recerca en Fisioteràpia i professor de Salut Pública, Metodologia, Administracions sanitàries i Ètica (Escola Universitària Gimbernat, adscrita a la Universitat Autònoma de Barcelona).

Elisabeth Herrero Vila

Metge especialista en Medicina Familiar i Comunitària. Màster en Medicina Preventiva i Promoció de la Salut (Universitat de Barcelona). Màster en Gestión Clínica y Asistencial de Atención Primaria (Universitat de Barcelona). Màster en Gerontologia Clínica (Universitat Autònoma de Barcelona). Diploma de Postgrau en Psicopatologia Clínica (Universitat de Barcelona).

Josep Sánchez Aldeguer

Doctor en Medicina per la Universitat Autònoma de Barcelona. Professor de la Facultat de Medicina de la Universitat Autònoma de Barcelona. Professor de Metodologia, Administracions sanitàries, Farmacologia i Geriatria a l'Escola Universitària Gimbernat (adscrita a la Universitat Autònoma de Barcelona). Responsable de la Coordinadora de Metges de Residències Geriàtriques del Vallès.

Metodologia i Estadística per a professionals de la salut

Bàsica
Pràctica

III. L'ANÀLISI ESTADÍSTICA

Jordi Esquirol Caussa
Elisabeth Herrero Vila
Josep Sánchez Aldeguer

INTRODUCCIÓ A LA SECCIÓ III

En aquesta tercera secció s'exposen els mètodes d'anàlisi estadístic més bàsics i més utilitzats en els estudis científics en ciències de la salut.

No es detallen *totes* les tècniques d'anàlisi estadístic, doncs per això hi ha tractats i professionals dedicats exclusivament a això, però sí els mètodes més bàsics de l'anàlisi estadística descriptiva i de l'estadística inferencial univariant, bivariant i multivariant.

Les exposicions matemàtiques es donen "partint de zero", i no es suposa cap coneixement estadístic previ. Tampoc és necessari conèixer la utilització de cap programa informàtic específic, i tots els exemples han estat realitzats amb un full de càlcul (Microsoft Excel 2007®), a l'abast de la majoria.

Els exemples que il·lustren totes les explicacions són extrets de la pràctica clínica i són de tema sanitari, per fer que el lector arribi a la seva comprensió d'una manera més fàcil (això no implica que els resultats de tots els exemples donats puguin ser extrapolats a la població general, doncs en la majoria la mostra no era representativa de la població).

Esperem que les exposicions siguin útils per al dia a dia professional o per als estudis, i que la seva lectura pugui fer-se de manera còmoda i amena.

Abans de començar l'anàlisi

Per poder obtenir uns resultats fiables d'un estudi científic cal seguir un seguit d'etapes abans de començar l'anàlisi de les dades pròpiament dita.

Els objectius de l'estudi han de ser clars i estar concretament formulats al protocol o memòria de l'estudi.

També s'han d'haver definit de manera clara i concreta totes les variables que intervindran en l'anàlisi de les dades: el tipus i el mètode de mesura de les variables ha de ser especificat al protocol: l'anàlisi estadístic forma part de totes les etapes de l'estudi científic, fins i tot les més precoces.

Un dels punts més importants a tenir en compte per analitzar les dades extretes d'una mostra és l'habilitat de triar quin és l'anàlisi estadística que cal emprar per a cada una de les preguntes de la investigació. És fonamental tenir una eina específica per a cada tasca concreta, conèixer la seva utilitat, saber quan utilitzar-la i com fer-ho destrament.

Però, abans de decidir realitzar l'anàlisi cal tenir les dades suficientment aptes i preparades per a ser tractades. Un cop recollides les dades de la mostra, cal filtrar-les, organitzar-les, resumir-les i només llavors podrem analitzar-les.

Preparació de les dades

Quan estem al davant de les dades recollides per un estudi científic, abans de realitzar directament l'anàlisi estadística és necessari haver comprovat i preparat les dades per al seu maneig.

Probablement aquestes dades ens arribin totes incloses en un full de càlcul (o les haguem d'introduir nosaltres mateixos), on les files representen els individus analitzats i cada una de les columnes representen una variable; la intersecció entre una fila i una columna defineix una cel·la que conté el valor d'aquella variable per aquell individu.

Però amb això no n'hi ha prou per començar l'anàlisi; abans de continuar, haurem d'analitzar el format de cada una de les variables i la seva naturalesa, definir de manera correcta i concreta cada una de les variables i de les seves possibles categories (si s'escau), i cercar la presència de valors impossibles i de valors perduts (veure "Definició de les variables: concepte i tipus" al llibre corresponent a la Secció II. Bases de l'Estadística).

Possiblement, per facilitar l'anàlisi haguem de transformar o crear algunes variables noves a partir de les ja existents; podrem transformar variables existents o crear noves variables a partir del càlcul numèric d'una o diverses variables quantitatives i crear variables qualitatives (categòriques) a partir de variables tant quantitatives com qualitatives.

D'aquests procediments se n'anomena *transformar* o *recodificar* les variables en les pròpies variables (en elles mateixes, modificant-ne el contingut) o en variables de nova creació (més recomanable). Els processos més freqüentment emprats en aquest sentit són:

- Calcular noves variables quantitatives a partir d'altres variables quantitatives definides anteriorment.

  Per exemple, a partir de les variables *Pes* i *Alçada*, ambdues numèriques contínues (en Kg. i en cm., respectivament), crear una nova variable quantitativa contínua que anomenarem *IMC* (amb una Definició de la variable: *Índex de Massa Corporal*), que prendrà els valors amb unitats Kg/m^2 a partir de la fórmula:

  $$IMC = {Pes}/{Alçada^2}$$

- Categoritzar una variable quantitativa contínua en una nova variable qualitativa (veure "Categorització de les variables quantitatives al llibre corresponent a la Secció II. Bases de l'Estadística).

  A partir de la variable creada *IMC*, podrem categoritzar-la en una nova variable categòrica anomenada *IMC5gr* en, per exemple, cinc categories diferents, segons els intervals de valors:

  - *IMC* igual o menor a 18,00: Categoria 1 (Definició de categoria: *Baix pes*)
  - *IMC* entre 18,01 i 24,00: Categoria 2 (Definició de categoria: *Normopes*)
  - *IMC* entre 24,01 i 30,00: Categoria 3 (Definició de categoria: *Sobrepès I*)
  - *IMC* entre 30,01 i 40,00: Categoria 4 (Definició de categoria: *Sobrepès II*)
  - *IMC* igual o superior a 40,01: Categoria 5 (Definició de categoria: *Obesitat Mòrbida*)

- Crear noves categories d'agrupació d'una variable categòrica amb altres categories.

> A partir de la nova variable *IMC5gr*, podrem també categoritzar-la en una nova variable categòrica anomenada *IMC3gr* en, per exemple, tres categories diferents, segons els intervals de valors:
>
> - *IMC* igual o menor a 18,00: Categoria 1 (Definició de categoria: *Baix pes*)
> - *IMC* entre 18,01 i 24,00: Categoria 2 (Definició de categoria: *Normopes*)
> - *IMC* igual o superior a 24,01: Categoria 3 (Definició de categoria: *Obesitat*)

Després de completar el procés de definició, creació i modificació de les variables, cal seguir un procediment de *control de qualitat* de les dades, per assegurar que les dades que analitzarem són de qualitat suficient per oferir resultats fiables.

Control de qualitat de les dades

Aquesta fase és indispensable per poder realitzar una anàlisi correcta i amb resultats fiables de l'estudi científic; cal seguir una sèrie de passos:

1. Definició de *valors perduts* (*missings*): observar les cel·les en què no hi ha cap valor determinat, generalment cal deixar aquestes cel·les buides (poden també aparèixer amb un signe "-"), i generalment no s'han de codificar amb un "0" (no és el mateix no tenir cap informació per aquella variable en aquell individu, que tenir un valor numèric "0" o tenir un "NS/NC" o "no sap, no contesta"). El valor es pot haver perdut durant la transcripció de les dades o pot no haver existit mai. Sovint es considera acceptable una quantitat inferior al 5% per valors perduts en una variable.
2. Detecció de *valors erronis*: els valors erronis poden ser causats bàsicament per:
    a. *Valors impossibles*: valors que no tenen sentit, com edats de persones superiors a les biològicament possibles, pesos negatius, talles impossibles, temperatures o nombres de fills aberrants , etc.
    b. *Valors fora de rang*: són valors que, tot i ser possibles, són fora dels límits definits en la investigació (criteris d'inclusió o d'exclusió); per exemple, si fem un estudi sobre pacients

pediàtrics que hi hagi alguna edat superior a 18 anys, o que en un estudi en embarassades hi hagi codificat un home.

c. *Incompliment de zeros estructurals*: hi ha casos impossibles o contradictoris *per se*, com per exemple que hi hagi un valor positiu en la variable "diagnòstic de càncer prostàtic" en una persona catalogada de "femení"; no hi pot haver cap valor d'aquest tipus en les nostres dades.

d. *Variables alfanumèriques*: molt sovint aquest tipus de variables són una font d'errors i de mala qualitat de les dades: a més de ser molt sensibles als errors tipogràfics, els programes informàtics interpreten com a diferents les distintes pulsacions de teclat (per exemple, poden interpretar com a diferents categories les pulsacions "femeni", "femení", "Femeni", "Femení" i qualsevol altra variació o error de pulsació; per això es té tendència a atorgar nombres a les categories i definir la categoria a part, veure "Definició de les variables: concepte i tipus" al llibre corresponent a la Secció II. Bases de l'Estadística).

3. Procurar *recuperar* valors erronis, si és possible: hauríem d'identificar l'error i posteriorment intentar-lo corregir per no haver-lo de codificar com a *valor perdut*; per intentar recuperar les dades errònies haurem de contactar amb la persona que ha recollit i registrat les dades i/o amb qui les hagi introduït, consultar els originals de la recollida de les dades per provar de detectar i corregir els errors de digitació, intentar registrar de nou la dada errònia si és possible contactar amb el participant a l'estudi i és possible repetir la mesura. Si no és possible cap d'aquestes opcions, caldrà considerar perduda (*missing*) aquella dada.

Per començar l'anàlisi de les dades recollides i ja preparades, hem de recordar també uns altres punts bàsics:

- Estar-ne segur que la pregunta d'investigació, les hipòtesis i/o els objectius de la recerca estan definits de manera clara i objectiva (veure "El marc pràctic" al llibre corresponent a la Secció I. Conceptes bàsics de Metodologia científica).
- Comprovar que comprenem perfectament els tipus de dades i de variables que estem gestionant (veure "Definició de les variables: concepte i tipus" al llibre corresponent a la Secció II. Bases de l'Estadística).
- Assegurar-se que la tècnica estadística que ens disposem a emprar és l'adequada per contestar la pregunta d'investigació.
- Comprovar les limitacions de l'anàlisi: si, donada la mostra que hem recollit, els resultats poden ser generalitzats a tota la població.

Anàlisi univariant

És clar que amb la recollida de la informació i la comprovació i preparació de les dades no acaba el procés de l'estudi, sinó que cal "llegir" les dades de manera que ens puguin donar una informació més comprensible i útil per a extreure'n conclusions i per a prendre decisions clíniques amb una base sòlida.

D'això se'n diu fer-ne l'anàlisi estadístic *apropiat*; hi ha molts tipus d'anàlisi diferents, i triar l'adequat és el que ens permetrà interpretar els resultats de la manera més adequada possible.

Després de tot el procediment de preparació de les dades caldrà identificar el tipus de variables que hem d'analitzar; cada tipus de variable presenta unes possibilitats diferents de gràfics i de mètodes d'anàlisi (es poden consultar els esquemes resum de les principals parts i dels principals mètodes de l'anàlisi estadística a partir de l'annex de la pàgina 207).

El primer pas de l'anàlisi de les variables de l'estudi és l'anomenat **anàlisi descriptiu univariant** (es poden consultar els esquemes bàsics de l'anàlisi univariant a l'annex de la pàgina 208 i de l'anàlisi inferencial univariant a la pàgina 209), això és, l'anàlisi de cada una de les variables de manera individual i analitzar la variabilitat per a cada variable de manera aïllada (veure "La variabilitat" al llibre corresponent a la Secció II. Bases de l'Estadística).

Posteriorment, es podrà realitzar l'anàlisi bivariant i/o multivariant, per analitzar si hi ha relacions entre les variables. Cada tipus diferent de variable té un procediment diferent d'anàlisi, atenent a les seves característiques i a la informació que se'n pot extreure.

A efectes de simplificar i per facilitar la comprensió de l'anàlisi univariant, hem estructurat aquest procés dividint-lo en els diferents tipus de variables i, dins de cada tipus, s'especifica el procés de tabulació, anàlisi bàsic i representació gràfica.

Es pot consultar un resum / esquema de l'anàlisi descriptiu de les variables a l'Annex de la pàgina 208.

Variables Qualitatives (Categòriques)

Les variables qualitatives (categòriques), descrites a "Variables Qualitatives (Categòriques)" del llibre corresponent a la Secció II. Bases de l'Estadística, mesuren característiques en els individus analitzats.

Les dades cauen en grups o categories, que poden ser *ordinals* (si les diferents categories tenen un ordre jeràrquic) o *nominals* (si no tenen un ordre o relació lògica de jerarquització entre les categories); si només tenen dues categories es denominen variables *dicotòmiques*, o *politòmiques* si presenten més de dues categories diferents.

Tabulació

Com s'introdueix a "Classificació i tabulació de la informació" al llibre corresponent a la Secció II. Bases de l'Estadística, l'anàlisi de les dades incloses en una variable qualitativa categòrica generalment s'inicia amb una ordenació de la informació per categories. El primer pas a realitzar és:

- Variable categòrica nominal: elaborar una taula que contingui totes les dades ordenades per categories; l'ordre en què es col·loquin les categories de la variable no és important, en aquest cas.
- Variable categòrica ordinal (o quantitativa discreta categoritzada, pàgina 25): tenir en compte quan elaborem la taula amb els elements ordenats per categories, que l'ordre entre les categories en sí mateix també és important, a diferència de les variables nominals.

Ràpidament es pot veure de manera aproximada quines són les categories que amb més freqüència es donen, i quines són les menys freqüents, especialment quan es fa un recompte del nombre d'elements que han caigut a cada categoria (Taula 1).

Posteriorment es realitza un recompte de les unitats presents en cada una de les categories i es comença l'elaboració del què anomenem *taula de freqüències* o *distribució de freqüències*, on cada categoria o classe està determinada per la *freqüència absoluta*, la *freqüència relativa* i les *freqüències acumulades* (en variables ordinals) (Taula 2).

D	D	B	C	B	B
D	D	D	C	D	U
D	C	D	U	D	C
U	B	B	U	B	D
D	B	B	D	U	B
C	B	C	D	D	C
D	C	D	D	C	B
D	D	D	B	U	C
B	C	B	U	D	D

B	B	B	B	B	B
B	B	B	B	B	B
B	B	D	D	D	D
D	D	D	D	D	D
D	D	D	D	D	D
D	D	D	D	D	D
U	U	U	U	U	U
U	C	C	C	C	C
C	C	C	C	C	C

Localització de la lesió	Nombre de lesions
Braç (B)	14
Dits (D)	22
Ulls (U)	7
Cama (C)	11
Total	54

Taula 1: Ordenació i recompte de les unitats en cada categoria per iniciar l'elaboració de la taula de freqüències de l'exemple de "Classificació i tabulació de la informació" del llibre corresponent a la Secció II. Bases de l'Estadística.

Anàlisi bàsica

L'estadística emprada per les variables categòriques analitza bàsicament la freqüència amb què es dóna cada una de les categories en la mostra.

A partir de les freqüències relatives de cada categoria es pot observar la proporció de cada una d'elles; a partir d'aquestes proporcions podem fer estimacions, comparacions i cercar relacions entre els grups.

La taula o *distribució de freqüències* és una manera d'organitzar les dades per tal d'expressar la freqüència en què es donen observacions en cada una de les classes, mostrant el patró de la distribució d'una manera més fàcilment interpretable; la taula de freqüències és una ordenació de les dades en forma de taula, assignant les freqüències corresponents a cada dada o categoria. Es recomana el seu ús quan hi ha gran quantitat de dades en categories diferents.

Al determinar quants elements pertanyen a cada classe o categoria, podem amb facilitat establir la freqüència en què cada classe determinada es presenta.

La taula o distribució de freqüències resultant ens dóna:

- La *freqüència absoluta*
- La *freqüència relativa*
- La *freqüència acumulada*