

TEXTOS DOCENTS

356

ANÀLISI DE DADES I INTRODUCCIÓ A LA PROBABILITAT

David Márquez
Carles Rovira

Departament de Probabilitat, Lògica i Estadística



UNIVERSITAT DE BARCELONA



TEXTOS DOCENTS

356

ANÀLISI DE DADES I INTRODUCCIÓ A LA PROBABILITAT

David Márquez
Carles Rovira

Departament de Probabilitat, Lògica i Estadística

Publicacions i Edicions



UNIVERSITAT DE BARCELONA



INTRODUCCIÓ

Aquest manual recull uns apunts preparats per fer l'assignatura *Anàlisi de dades i introducció a la probabilitat* del Grau de Matemàtiques de la Universitat de Barcelona.

Es tracta d'una assignatura de primer curs, on l'estudiant està en el pas de l'institut a la universitat, dels estudis de batxillerat als universitaris. Aquest pas sempre és un pas difícil i complex, i més encara quan l'ensenyament triat és un grau en matemàtiques.

Aquest manual, com l'assignatura, té tres objectius bàsics: els dos primers són generals de les matemàtiques i l'últim és particular de la probabilitat i l'estadística.

1. Mostrar a l'alumnat com s'han de treballar les matemàtiques.
2. Acostumar l'alumnat al llenguatge i al rigor matemàtic.
3. Donar uns coneixements de probabilitat i estadística que sovint no han estat gaire treballats durant l'etapa del batxillerat.

ÍNDEX

1	Mostreig	1
1.1	Població i mostra	1
1.2	Mostreig aleatori simple	2
1.3	Mostreig sistemàtic	3
1.4	Mostreig estratificat	3
1.5	Mostreig per conglomerats	4
1.6	Mostreig polietàpic	5
1.7	Altres tipus de mostreig	5
1.7.1	Mostreig per quotes	6
1.7.2	Mostreig de la bola de neu	6
2	Estadística descriptiva	7
2.1	Tipus de variables	7
2.2	Taules de freqüències	9
2.3	Representació gràfica de dades	12
2.3.1	Variables qualitatives	12
2.3.2	Variables quantitatives	13
2.4	Representació numèrica: estadístics	15
2.4.1	Mesures de tendència central	16
2.4.2	Mesures de dispersió	18
2.4.3	La dispersió relativa	21
2.4.4	Desigualtat de Txebitxev	22
2.4.5	Estandarditzar	23
2.4.6	Mesures de posició	24
2.4.7	Mesures de forma	26
2.5	Altres mètodes de representació gràfica	27
2.5.1	El diagrama de caixa	27
2.5.2	Diagrama de tija i fulles	28
2.6	Complements	29
2.6.1	Transformacions d'una variable	29
2.6.2	Mitjana harmònica per a les relacions inverses	30
2.7	Exercicis	32

3	Estadística descriptiva bivariant	35
3.1	Dades no aparellades	35
3.2	Taules de contingència	37
3.3	Representacions gràfiques	39
3.4	Mesures de dependència	40
3.5	Models de regressió	44
3.5.1	Càlcul de la recta de regressió	44
3.5.2	Qualitat de l'ajustament	48
3.5.3	Anàlisi dels residus	49
3.6	Exercicis	52
4	Combinatòria	55
4.1	La regla del producte	55
4.2	Tècniques per comptar	56
4.2.1	Variacions sense repetició. Permutacions	57
4.2.2	Variacions amb repetició	58
4.2.3	Combinacions	59
4.2.4	Combinacions amb repetició	60
4.2.5	Altres models	61
4.3	Nombres combinatoris	62
4.4	Exercicis	64
5	Càlcul de Probabilitats	67
5.1	Espai mostral i esdeveniments	67
5.1.1	Experiment aleatori	67
5.1.2	Espai mostral	68
5.1.3	Esdeveniments i operacions	68
5.1.4	Particions	70
5.2	Probabilitat	70
5.2.1	Definició clàssica de probabilitat	71
5.2.2	Definició axiomàtica de probabilitat en espais finits o numerables	75
5.2.3	Propietats	77
5.2.4	Definició axiomàtica general de probabilitat	79
5.3	Probabilitat condicionada	80
5.3.1	Definició de probabilitat condicionada	80
5.3.2	Esdeveniments independents	83
5.3.3	Teorema de probabilitats compostes	86
5.3.4	Probabilitats totals	87
5.3.5	Fórmula de Bayes	89
5.4	Exercicis	91
6	Les distribucions binomial i normal	95
6.1	La distribució binomial	95
6.1.1	Representació de la distribució binomial	97
6.2	La distribució normal	100

6.2.1	Distribució normal estàndard	102
6.3	Exercicis	103
	Bibliografia	105

CAPÍTOL I

1. MOSTREIG

En aquest capítol tractarem un dels aspectes més importants de l'estadística: com es tria una mostra d'una població. La tria d'aquesta mostra és una tema molt complex i important perquè els resultats de l'estadística es basen en el fet d'haver triat una bona mostra que representi correctament el conjunt de la població. Els apartats que conté aquest capítol són els següents:

1.1 Població i mostra

1.2 Mostreig aleatori simple

1.3 Mostreig sistemàtic

1.4 Mostreig estratificat

1.5 Mostreig per conglomerats

1.6 Mostreig polietàpic

1.7 Altres tipus de mostreig

1.1 Població i mostra

Anomenarem **població** un conjunt d'elements amb certa homogeneïtat dels quals es vol estudiar alguna característica. Una població està formada per **elements** o **individus**.

Quan podem disposar d'un llistat de tots els elements o individus de la població direm que tenim un **cens** de la població.

Habitualment no és possible estudiar tots els elements d'una població, encara que disposem d'un cens, i haurem de seleccionar d'alguna manera un conjunt d'elements o individus

representatius de la població que estem estudiant. Aquest subconjunt representatiu és el que coneixem per **mostra** d'una població.

Els motius pels quals no podem estudiar tots els elements de la població són molt variats i diversos. Tot seguit n'expliquem alguns a mode d'exemple:

- Pot ser inviable econòmicament estudiar tota la població.
- L'estudi pot implicar la destrucció dels elements estudiats, com per exemple la durada d'una bombeta o una televisió.
- Podem tenir una població que no podem analitzar completament, com en el cas de voler estudiar les peces defectuoses que produeix un cert procés.
- El temps que requereix l'estudi de la població completa és impossible.

Abans de començar a descriure els diferents tipus de mostreig i per recalcar la importància d'escollir una mostra prou representativa de la població, explicarem un exemple real i típic d'errada a l'hora de triar la mostra.

L'any 1936 un dels estudis realitzats sobre qui seria el president dels EUA va pronosticar que Roosevelt seria clarament derrotat. Aquest estudi, a més, s'havia obtingut mitjançant una mostra molt gran. En canvi, Roosevelt va guanyar de manera molt aclaparadora. L'error va ser que l'enquesta sobre qui seria el president dels EUA es va fer per telèfon, en una època en què el fet de tenir telèfon significava pertànyer a un estatus social que no representava de cap manera tot el conjunt de la població dels EUA.

1.2 Mostreig aleatori simple

És un tipus de mostreig molt important i molt utilitzat en la inferència estadística. Direm que tenim una **mostra aleatòria simple** quan l'hàgim obtinguda emprant les dues propietats següents:

- Cada element de la població té la mateixa probabilitat de ser triat per formar part de la mostra.
- Els elements es trien un a un i amb reemplaçament, de manera que la població és idèntica a cada extracció.

La primera condició ens garanteix la representativitat de la mostra, que no hi ha individus privilegiats. Quant a la segona ens assegura que les eleccions són independents entre si. D'altra banda, si la grandària de la població (N) és molt gran respecte a la de la mostra (n) és pràcticament irrellevant fer el mostreig amb reposició o sense.

Per seleccionar una mostra d'una població finita per aquest mètode es fa, per exemple, de la manera següent: s'assigna un nombre a cada individu de la població i s'escriuen N paperetes amb els nombres de l'1 a N , es barregen i traiem les n paperetes que han de formar la mostra, una a una i tornant a reintegrar la papereta després de cada extracció.

També es pot fer posant en una urna N boles numerades de l'1 a N i fent n extraccions amb reemplaçament. Una altra manera d'elegir una mostra d'aquest tipus és utilitzant els anomenats nombres aleatoris i que es poden triar mitjançant l'ordinador.

Aquest mètode garanteix una aleatorietat màxima i té molt bones propietats matemàtiques. Un problema que comporta és que la posada a la pràctica una vegada triada la mostra pot requerir molt temps i, sobretot, una despesa econòmica molt gran —els elements seleccionats poden estar molts dispersos.

1.3 Mostreig sistemàtic

Aquest mostreig és un mètode molt simple per obtenir una mostra.

Suposem que volem obtenir una mostra de grandària n . En primer lloc, numerem els individus de la població de l'1 al N , on N és la grandària de la població. Fem el quocient N/n i agafem la part entera del nombre resultant, anomenem-lo $k = [N/n]$. La mostra sistemàtica es fa elegint a l'atzar un nombre entre 1 i k . Aquest serà el primer element de la mostra. La resta d'individus de la mostra es trien sumant k a aquest primer element. És a dir, si el primer element és n_1 , els altres seran $n_1 + k$, $n_1 + 2k$, i així successivament fins a $n_1 + (n - 1)k$. Obtindrem el nombre necessari d'individus per formar la mostra i tots els elements de la llista hauran tingut possibilitats de ser triats.

Si l'ordre de la llista és aleatori, aquest procediment és més o menys equivalent al mostreig aleatori simple, encara que resulta més fàcil portar-lo a terme.

Exemple: Volem seleccionar 8 estudiants d'entre els 900 que formen una facultat. Numerem els 900 estudiants per ordre de llista pels seus cognoms. Agafem un nombre a l'atzar entre 1 i $[900/8] = 112$, suposem que ha sortit el 27. Aleshores els estudiants elegits són els que ocupen els llocs 27, $27 + 112 = 139$, $139 + 112 = 251$, 363, 475, 585, 699 i 811.

1.4 Mostreig estratificat

En els dos tipus de mostreig que hem exposat fins ara no s'ha pres en consideració la possibilitat que tinguem cert coneixement previ sobre certes característiques de la població. És a dir, els dos procediments anteriors prioritzen l'aleatorietat sobre qualsevol aspecte i pressuposen que la població és homogènia respecte a la característica que estem estudiant. Quan disposem d'una informació sobre la població convé tenir-la en compte en seleccionar la mostra. Un exemple típic són les enquestes d'opinió, en què els enquestats són heterogenis pel que fa al gènere, l'edat, la professió, la classe social, etc. i això ho volem tenir present a l'hora de triar la mostra.

El que pretenem fent un mostreig estratificat és que la mostra tingui una composició anàloga a la població. Si tenim una població en què els seus elements es poden dividir en **classes** o **estrats** —grups disjunts—, direm que fem un mostreig estratificat si la mostra es pren assignant un nombre o una quota de membres a cada estrat per mantenir dins la mostra

la mateixa proporcionalitat que dins la població. Aquesta estratificació està motivada perquè la resposta a la variable que estem estudiant pot ser diferent segons l'estrat i volem tenir-la representada proporcionalment. Finalment, dins de cada estrat, els elements s'escolliran mitjançant mostreig aleatori simple o mostreig sistemàtic.

Exemple: Es vol fer un estudi de l'ús de l'ordinador per part dels estudiants per a tasques relacionades amb els estudis en un grau de Matemàtiques determinat amb 900 estudiants. Sabem que en el primer curs no hi ha cap assignatura d'informàtica però a segon n'hi ha una; a tercer, dues, i a quart, depèn molt de les optatives que l'estudiant esculli. Resulta obvi que la resposta dependrà del curs que estigui fent l'estudiant i per tenir en compte aquesta informació farem un mostreig estratificat tenint present els estudiants que té cada curs. Suposem que a primer hi ha 300 estudiants; a segon, 250; a tercer, 200, i a quart, 150. Si la mostra volem que tingui una grandària de 100 individus i que mantingui aquestes quotes de representativitat, escollirem 33 estudiants de primer, 28 de segon, 22 de tercer i 17 de quart. Dins de cada curs els estudiants s'escolliran a l'atzar.

1.5 Mostreig per conglomerats

Hi haurà situacions en què serà difícil aplicar un mostreig aleatori simple –o sistemàtic– o un mostreig estratificat. No disposarem d'una llista dels elements, ni d'estrats, o tot i disposar d'una llista no serà viable econòmicament portar a terme un mostreig aleatori simple o sistemàtic. En aquests casos pot passar que els elements de la població estiguin agrupats per conglomerats, on entendrem per **conglomerats** agrupacions naturals d'individus tant físiques com geogràfiques com d'altres tipus en què es distribueixen els elements de la població que volem estudiar. Si cada un d'aquests conglomerats representa correctament la població, seleccionarem primer alguns d'aquests conglomerats a l'atzar mitjançant un mostreig aleatori simple sobre els conglomerats, i després o bé analitzarem tots els elements dels conglomerats triats, o bé realitzarem un mostreig aleatori simple dels seus elements.

Aquest tipus de mostreig s'anomena mostreig per conglomerats i té el gran avantatge que facilita molt la recollida de dades o informació. El problema apareix si els conglomerats no són homogenis entre si, ja que aleshores la mostra no serà representativa. Els conglomerats poden no ser homogenis perquè són diferents entre ells o bé perquè no contenen tota la diversitat de la població.

Exemple: Imaginem que volem fer un estudi sobre els estudiants de grau de Matemàtiques de les facultats de l'Estat espanyol. En el cas que poguéssim fer una llista de tots els estudiants, si fem un mostreig aleatori simple o sistemàtic d'aquesta llista, el treball tindria un cost econòmic i de temps enorme perquè podria ser que ens haguéssim de desplaçar a una facultat per un sol estudiant. Si podem acceptar que cada grau és un conglomerat que representa correctament tota la població d'estudiants de grau de Matemàtiques, podem seleccionar a l'atzar –per mostreig aleatori simple– uns quants graus de tot l'Estat i, dins d'aquests, fer un mostreig aleatori simple o sistemàtic.

Finalment, destacaria les diferències entre un mostreig estratificat i per conglomerats. L'estratificat funciona millor com més grans són les diferències entre els estrats i més homogenis siguin aquests estrats internament. En canvi, els conglomerats van molt millor si hi ha poques diferències entre ells i són molt heterogenis internament. Tot i les diferències existents entre els dos models, de vegades pot resultar molt interessant fer un mostreig utilitzant els dos procediments barrejats.

1.6 Mostreig polietàpic

Aquest tipus de mostreig recull el procediment que s'acaba de comentar en l'apartat anterior, una combinació dels mètodes del mostreig estratificat i del mostreig per conglomerats. Barrejant aquests dos mètodes aconseguim una millor representativitat de la mostra i una recollida d'informació molt més senzilla.

En segons quins casos, com en l'exemple següent, podem utilitzar primer el mostreig per conglomerats i després el mostreig estratificat.

Exemple: En el mateix exemple que acabem d'exposar en la secció anterior sobre el mostreig per conglomerats pot ocórrer que l'estudi que volem portar a terme tingui una resposta diferent si l'estudiant està fent el primer curs del grau que si està fent el segon, el tercer o el quart. Aleshores, després d'haver triat unes facultats determinades mitjançant el mostreig per conglomerats, ens pot interessar escollir els estudiants de la mostra mantenint la proporció per cursos que té la població. En aquest cas, dins de cada conglomerat farem una estratificació per cursos.

En d'altres casos primer farem un mostreig estratificat i després per conglomerats com en el següent exemple.

Exemple: Volem estudiar certes característiques dels alumnes de les escoles de Barcelona. Per limitar els costos decidim analitzar només algunes escoles. Ara bé, no podem utilitzar el mostreig per conglomerats simplement seleccionant algunes escoles a l'atzar perquè hi ha moltes escoles que no són representatives del total d'alumnes de les escoles de Barcelona. En primer lloc, haurem de fer una divisió per estrats tenint en compte el nivell econòmic dels barris de Barcelona i el tipus d'escola –pública, concertada o privada. Quan tinguem els estrats i la proporció que hem de reproduir per mantenir la divisió dins la població passarem a elegir pel mostreig aleatori simple alguns dels conglomerats, les escoles, de dins els estrats.

1.7 Altres tipus de mostreig

Fins ara tots els tipus de mostreig que hem explicat presenten fortament algun tipus d'aleatorietat. En canvi, existeixen altres formes de mostreig que no tenen aquest component d'atzar o hi és poc present.

1.7.1 Mostreig per quotes

Aquest mostreig és molt similar al mostreig estratificat però, a diferència d'aquest, l'elecció dels elements dins de cada estrat no és aleatòria.

Suposem que la població que volem estudiar pot dividir-se en diferents estrats. Aleshores la nostra mostra haurà de mantenir la representativitat d'aquests estrats. Fins aquí coincideix amb el mostreig estratificat. La diferència radica en el fet que l'elecció dels elements a partir d'ara no té per què ser aleatòria, una vegada determinada la quota d'individus dins de cada estrat, l'investigador que fa el treball pot triar els representants d'aquests estrats com vulgui i desitgi.

1.7.2 Mostreig de la bola de neu

Molt indicat per a treballs sobre poblacions estranyes, minoritàries, disperses o clandestines. L'investigador que fa el treball i que ha d'elegir la mostra decideix incloure certs individus dins la mostra per les seves característiques o perquè compleixen certs requeriments. Aquest grup inicial d'individus ens servirà per anar localitzant més individus que compleixen aquests requisits i que també s'inclouran a la mostra i així successivament.

CAPÍTOL II

2. ESTADÍSTICA DESCRIPTIVA

Un estudi estadístic té bàsicament tres etapes: recollir les dades, descriure-les i analitzar-les. En aquest tema explicarem com s'han de descriure.

Ho farem a partir de les observacions d'una variable. Tant pot ser que tinguem només les observacions d'una mostra de la població com que tinguem dades sobre tota la població.

Aquest tema es compon de les seccions següents:

2.1 Tipus de variables

2.2 Taules de freqüències

2.3 Representació gràfica de dades. Variables qualitatives. Variables quantitatives.

2.4 Representació numèrica: estadístics. Mesures de tendència central. Mesures de dispersió. La dispersió relativa. Desigualtat d Txebitxev. Estandarditzar. Mesures de posició. Mesures de forma.

2.5 Altres mètodes de representació gràfica. Diagrama de caixa. Diagrama de tija i fulles.

2.6 Complements. Transformació d'una variable. Mitjana harmònica per les relacions inverses.

2.1 Tipus de variables

Les variables poden ser de diversos tipus. Les podem dividir en dues grans classes:

1. **Variables qualitatives.** Es refereixen a una característica no numèrica de l'individu. En principi no s'expressen numèricament, però a la pràctica moltes vegades es codifiquen numèricament per facilitar-ne el tractament. Per exemple, el gènere d'un individu és una variable qualitativa que pot prendre dos valors: M (masculí) i F (femení), però moltes vegades es codifica amb un 1 (M) i un 0 (F).

2. **Variables quantitatives.** Són les que es refereixen a característiques dels individus que s'expressen numèricament. Dintre d'aquestes en podem trobar de **discretes** –quan només poden prendre un nombre discret de valors– o de **contínues** –quan poden prendre qualsevol valor dins d'un interval.

Exemple: Si estudiem la població formada pels alumnes d'una universitat. Algunes variables que podem estudiar són: gènere (qualitativa), nombre de fills (quantitativa discreta) o alçada (quantitativa contínua).

Exemple: Considerem com a població els avions de passatgers del món. Algunes variables: antiguitat (quantitativa discreta), model (qualitativa) o quilòmetres recorreguts el darrer any (quantitativa contínua).

Exemple: Considerem com a població els estudiants d'uns estudis virtuals d'informàtica. Tenim els resultats d'una mostra agafada a l'atzar on se'ls demana la informació següent:

1. SEX: Sexe (1-Home, 2-Dona)
2. EDAT: Edat
3. INT: Temps de connexió setmanal a Internet en minuts
4. TV: Temps de televisió setmanal en minuts
5. OCI: Despesa mensual en oci en euros
6. ING: Ingressos mensuals en euros
7. SOC: Nivell sociològic de $0, \dots, 10$ (0-Molt conservador, 10-Molt progressista)
8. LLEN: Llengua utilitzada habitualment (1-català, 2-castellà, 3-bilingüe català/castellà, 4-altres)

Hi ha variables qualitatives (SEX, LLEN), variables quantitatives discretes (EDAT, SOC) i variables quantitatives contínues (INT, TV, OCI, ING).

A partir d'un fitxer real amb aquestes dades de 325 individus donarem exemples dels diversos mètodes que estudiarem en els propers dos capítols. Els resultats seran els obtinguts utilitzant algun tipus de programari estadístic.

En aquest capítol suposem que tenim ja unes dades, d'una única variable i que el que volem fer és descriure-les. Ho farem emprant dos tipus de tècniques: representant-les gràficament i buscant valors numèrics que les resumeixin.

Suposem, per tant, que tenim n observacions: $x_1, x_2, \dots, x_n$, d'una certa variable x . Aquestes observacions poden ser tota la població o una mostra representativa d'aquesta població.