

ESTADÍSTICA BÀSICA

Curs 2000-2001

**Francesc Carmona
Alexandre Sánchez**

Departament d'Estadística



UNIVERSITAT DE BARCELONA

Sèrie de Quaderns Docents del Departament d'Estadística, Núm. 11

Barcelona, 31 d'octubre de 2000

ÍNDEX

1	INTRODUCCIÓ.....	5
1.1	EL CONCEPTE D'ESTADÍSTICA.....	7
1.2	ALGUNES DEFINICIONS BÀSIQUES.....	8
1.3	VARIABLES ESTADÍSTIQUES.....	8
1.4	ESCALES DE MESURA.....	9
1.5	EL MOSTRATGE	9
1.5.1	Conceptes fonamentals en el mostratge.....	9
1.5.2	Procés de mostratge.....	10
1.5.3	Procediments del mostratge.....	10
1.6	LES DADES: SÈRIES ESTADÍSTIQUES.....	15
1.7	L'ESTADÍSTICA EN LA INVESTIGACIÓ DE MERCATS.....	15
1.7.1	Tipus de disseny pels estudis de mercats.....	16
1.7.2	Les fonts d'informació.....	16
1.7.3	Tipus d'enquestes.....	17
1.7.4	Relació entre la naturalesa de la investigació i els tipus d'informació.....	18
1.8	EXERCICIS	19
2	ORGANITZACIÓ DE DADES UNIVARIANTS.....	21
2.1	INTRODUCCIÓ	23
2.2	DADES DISCRETES.....	23
2.3	ATRIBUTS.....	27
2.4	DADES CONTÍNUES	29
2.5	DADES AGRUPADES.....	30
2.6	SÈRIES CRONOLÒGIQUES	34
2.7	EXERCICIS	36
3	ESTADÍSTICS	39
3.1	MESURES DE POSICIÓ, DE DISPERSIÓ I DE FORMA.....	41
3.1.1	Mesures de posició.....	41
3.1.2	Mesures de dispersió.....	42
3.1.3	Mesures de forma	42
3.2	MESURES DE TENDÈNCIA CENTRAL	43
3.3	MESURES DE DISPERSIÓ.....	45
3.4	MESURES DE FORMA	46
3.4.1	Mesures d'asimetria.....	46
3.4.2	Mesures de kurtosi.....	46
3.5	DIAGRAMA DE TIJA I FULLES.....	48
3.6	PERCENTILS.....	49
3.7	ESTADÍSTICS D'ORDRE.....	50
3.7.1	Estadístics de centralitat.....	51
3.7.2	Estadístic de dispersió (rang interquartilic)	51
3.7.3	Estadístic d'asimetria	51
3.7.4	Valors atípics (outliers)	51
3.7.5	Distribució normal.....	51
3.8	GRÀFICS BASATS EN ESTADÍSTICS D'ORDRE.....	52
3.8.1	Sumari dels 5 números.....	52
3.8.2	Diagrama de caixa (Boxplot).....	53
3.8.3	Diagrama de tija i fulles.....	54
3.9	EXERCICIS	55
4	DESCRIPCIÓ DE DADES BIVARIANTS.....	59
4.1	DADES BIVARIANTS.....	61
4.2	TAULES CREUADES.....	62
4.2.1	Construcció d'una taula a partir de les dades originals.....	63
4.2.2	Recodificació de variables.....	64
4.3	REPRESENTACIONS GRÀFIQUES PER A DADES BIVARIANTS.....	66

4.4	CARACTERÍSTIQUES NUMÈRIQUES PER A DADES BIVARIANTS.....	68
4.4.1	Mesures per a dades quantitatives no agrupades.....	68
4.4.2	Mesures d'associació per a taules creuades.....	69
4.4.3	Mesures d'associació basades en l'estadístic khi-quadrat	70
4.4.4	Mesures d'associació basades en la reducció proporcional de l'error.	70
4.4.5	Mesures d'associació per a variables mesurades en escala ordinal.....	71
4.5	INTERPRETACIÓ DE TAULES DE TABULACIÓ CREUADA	71
4.5.1	Ús de percentatges.....	71
4.5.2	Elaboració de les relacions.....	72
5	INTRODUCCIÓ A LA REGRESSIÓ	75
5.1	INTRODUCCIÓ	77
5.2	LA REGRESSIÓ SIMPLE	78
5.3	EL MÈTODE DELS MÍNIMS QUADRATS.....	78
5.4	PREDICCIONS.....	81
5.5	MESURES DE L'AJUST	81
5.5.1	Descomposició de la variabilitat.....	82
5.5.2	Coeficient de determinació	82
5.6	CÀLCULS DE REGRESSIÓ AMB CALCULADORA	83
5.7	EXERCICIS	86
	BIBLIOGRAFIA.....	89

1 Introducció

1.1 El concepte d'Estadística

Quan ens enfrontem per primer cop a l'estudi de l'Estadística ho fem amb una idea prèvia, i sovint equivocada, del que anem a tractar. Suposem que li preguntem a una persona que passa pel carrer:

“- Sap de que tracta l'Estadística?”

No costa res imaginar l'entrevistat que contesta:

“- És clar, l'estadística tracta de...”

No està gens clar però, què és el que respondrà a continuació.

Per uns l'Estadística fa referència a les dades amb què es treballa, les taules i gràfics amb què ens bombardegen tots els mitjans de comunicació.

Altres veuran en l'Estadística aquesta eina, mig màgica mig maligna, de que es valen els polítics per provar coses com que “tots hem guanyat”, menys l'Estadística, ja que les enquestes han fallat totes.

Si amb paciència mirem d'agrupar les respostes, deixant de banda les de caràcter anecdòtic, podrem formar a grans trets dos grans grups de respostes :

Les que consideren l'estadística com una col·lecció de fets i dades resumides i classificades d'acord a un determinat criteri. Per referir-nos a aquesta accepció es reserva el terme *d'estadística*. Una *estadística* és una col·lecció de dades classificada segons un criteri determinat.

Les que veuen en l'Estadística un mètode per tractar i analitzar la informació continguda en les dades.

Serà en aquesta darrera accepció en que ens basarem en aquest curs, i que considera l'Estadística com *un instrument dedicat a la recol·lecció, organització i anàlisi d'observacions o dades numèriques*. Tampoc cal oblidar que l'Estadística, a més d'un mètode, és una branca de la ciència íntimament lligada amb les matemàtiques, que estudia lleis i fenòmens que depenen de l'atzar.

Centrant-nos en el concepte d'Estadística com un mètode podem distingir dues orientacions en la seva pràctica que ens permeten distingir entre *Estadística Descriptiva* i *Estadística Inferencial*

L'Estadística descriptiva proporciona

mètodes per ordenar, agrupar, classificar i representar les dades.

característiques numèriques que permeten resumir les dades amb la intenció de condensar la informació amb la màxima claredat i la mínima pèrdua.

La Inferència Estadística és un conjunt de tècniques que permeten induir objectivament conclusions sobre tot un col·lectiu (població) a partir d'un coneixement parcial d'aquest (mostra). Pretén establir un model general a partir d'un conjunt petit de dades.

1.2 Algunes definicions bàsiques

Al descriure la funció de l'Estadística han aparegut alguns termes que, si bé poden resultar familiars, tenen un significat molt precís en aquest context. Per aquest motiu, començarem per definir amb precisió alguns dels termes que emprarem sovint d'ara endavant.

Individu : és la unitat més simple subjecte a observació i registre de dades. Pot correspondre o no a un individu físic com una persona o una peça. Per exemple, en un estudi per determinar si les empreses tenen un servei propi de missatgers l'individu seria l'empresa.

Població : és el conjunt de referència (real o hipotètic) de tots els individus sobre els que volem estudiar un cert fenomen i amb alguna característica comú observable. Es delimita en funció de propietats dels individus. Un cop més el sentit estadístic no sempre coincideix amb el sentit usual de la paraula. La mida de la població és el número d'individus que la formen. Distingim entre poblacions finites i infinites.

Mostra : és un subconjunt d'individus observats que vol ser representatiu de la població a estudi. **Valors mostrals** són els valors numèrics observats per als diferents individus de la mostra. **Mida mostral** és el número d'individus seleccionats per a la mostra.

Parlarem d'un **cens** quan la mida de la mostra coincideix pràcticament amb la de la població.

Variable : és una característica numèrica o atribut observada sobre els individus.

1.3 Variables estadístiques

Ja s'ha definit més amunt el significat de variable. En la pràctica de l'Estadística resulta de gran importància el tipus de variable amb que es treballa, així com l'escala en que es mesura aquesta. Això condiciona el tipus d'informació que aporta i els tipus d'anàlisi que podem fer sobre ella..

El *suport* d'una variable és el conjunt de valors que pot prendre *a priori*, és a dir abans que l'haguem mesurat. Per exemple, si comptem el nombre de vehicles de transport d'una companyia el seu suport seran els enters positius i el zero, però no pot prendre valors negatius.

D'acord amb el tipus de suport, les variables es classifiquen en:

Quantitatives : són aquelles que prenen valors numèrics. Poden ser

Discretes : prenen valors numèrics en un conjunt finit o numerable.

Exemples: núm. de germans, núm. de cares fins que surti creu, unitats venudes, ...

Contínues : prenen valors numèrics en un conjunt infinit no numerable.

Exemples: temps que dura un procés, pes en kg d'una persona, longitud en mm d'una fulla, ...

Qualitatius: s'anomenen així perquè es refereixen a *qualitats* dels objectes estudiats (estat civil, sexe, color, marca...) i prenen valors que no són numèrics. També s'anomenen *atributs*.

1.4 Escales de mesura

Una idea molt lligada al tipus de variable, i també de gran importància a l'hora de decidir com s'analitzen les dades és la de *l'escala*, és a dir, la forma en que es mesura o quantifica la informació.

Les variables qualitatives es poden mesurar en escales nominals o ordinals. Les quantitatives en escales d'interval o de quocient. La taula següent resumeix els principals tipus d'escales i les seves característiques.

VARIABLES	ESCALA	CARACTERÍSTIQUES	EXEMPLES
Qualitatius	Nominal	Categories no numèriques sense cap relació d'ordre.	Estat civil, Sexe, Grup sanguini, Marca, Color
Qualitatius	Ordinal	Els valors són categories no numèriques amb una relació d'ordre.	Nivell d'Estudis Classe social Categoria (Hotel)
Quantitatius	Interval	Existeix una unitat numèrica constant. Només podem comparar valors per la diferència entre ells	Talles de roba, Talla de sabates, Temperatura en °C Comparació : “Tres talles més”
Quantitatius	Quocient	Com l'anterior, però amb un zero absolut. Podem comparar valors per quocient (doble, meitat)	Nº de fills, sou, edat. Comparació : “A guanya el doble que B”

1.5 El mostratge

Excepte en la investigació de laboratori o en el cas que una població sigui prou petita, ens veiem obligats a treballar amb mostres extretes d'una població.

La característica més important d'aquestes mostres ha de ser, juntament amb la seva representativitat, que garanteixi que els resultats obtinguts sobre ella siguin generalitzables, inferibles a tota la població. Es dedueix doncs, fàcilment, la importància d'una correcta extracció de la mostra, que d'ara endavant anomenarem mostratge.

1.5.1 Conceptes fonamentals en el mostratge

Els conceptes fonamentals en el mostratge són:

Element mostrat: Un element és cada una de les unitats bàsiques sobre les que es vol obtenir informació.

Població: Una població o univers, com també s'anomena, és el conjunt de tots els elements definits abans de la selecció de la mostra. La població s'ha de definir en termes de (a) elements, (b) unitats de mostratge, (c) abast i (d) temps.

Unitat mostral: Una unitat mostral és la unitat bàsica que està disponible per ser seleccionada en alguna fase del mostratge. Conté els elements de la població que poden formar part de la mostra.

Exemple:

En un estudi sobre la pràctica d'esports a l'aire lliure:

- L'element és cada persona entre 18 i 45 anys que practica esport a l'aire lliure.
- Les unitats mostrals poden ser estadis, camps d'esport, circuits de *jogging*, etc.

Abast: L'abast és la regió geogràfica objecte d'estudi.

Temps: El temps és el període temporal de l'estudi i, habitualment, coincideix amb el de recollida de la mostra.

Marc mostral: El marc és la llista que conté tots els elements de la població, sense repeticions i sense elements aliens a aquesta.

Exemple:

Si volem fer un estudi entre la gent en edat de votar, el cens i el marc no coincideixen. Caldria eliminar del cens tots els individus amb menys de 18 anys.

Població en estudi: Una població en estudi és el conjunt d'elements del qual s'agafa la mostra. Sovint, les dificultats d'ordre pràctic fan que la mostra real s'agafi d'una població una mica diferent de la que es defineix *a priori*. En el marc mostral falten elements de la població.

Exemple:

Fem una enquesta telefònica i a la llista de socis d'un club falten els telèfons d'alguns.

1.5.2 Procés de mostratge

Amb els conceptes del mostratge que acabem de definir, podem descriure de forma general els passos per seleccionar una mostra.

- Pas 1.** Definició de la població. Això inclou els elements, unitats de mostratge, abast i temps.
- Pas 2.** Identificació del marc mostral del qual se seleccionarà la mostra.
- Pas 3.** Decisió sobre la mida de la mostra. En aquest pas es determina el número d'elements que s'inclouran a la mostra. És un problema difícil de decidir quan una mostra és massa gran o massa petita.
- Pas 4.** Selecció d'un procediment específic mitjançant el qual es determinarà la mostra. Exactament, es tracta de prendre la decisió respecte a quins elements de la població s'inclouran a la mostra.
- Pas 5.** Selecció física de la mostra basant-se en el procediment escollit en el pas 4.

1.5.3 Procediments del mostratge

Existeixen molts i diferents procediments pels quals els investigadors poden seleccionar les seves mostres, però inicialment podem diferenciar dues menes de mostratge:

Probabilístic. On tots els elements de la població tenen una probabilitat coneguda de ser seleccionats per a la mostra. La selecció es fa amb regles de decisió matemàtiques que no permeten cap mena de discreció a l'investigador o entrevistador de camp. Es pot dir a terme si es disposa d'un marc mostral que permeti *sortejar* l'aparició dels elements a la mostra.

Observem que diem “una probabilitat coneguda” i no “una probabilitat igual”. El mostratge probabilístic amb igualtat d'oportunitats per a tots els elements és només un cas particular que rep el nom de *mostratge aleatori simple*.

L'avantatge dels mètodes de mostratge probabilístic és que tenen una base teòrica en la teoria de mostres i l'estadística matemàtica, a partir de les quals podem calcular el grau fins el que el valor de la mostra pot diferir del valor de la població d'interès. Aquesta diferència s'anomena *error mostral*. Així podem acotar l'error mostral o avaluar la precisió de les estimacions.

L'inconvenient principal d'aquests mètodes és el seu cost. També cal destacar la dificultat de disposar d'un marc mostral acurat.

No probabilístic. On la selecció d'un element de la població per a que formi part de la mostra es basa, en part, en el criteri de l'investigador o de l'entrevistador de camp. No existeix probabilitat coneguda sobre la oportunitat de selecció de cada element particular de la població. En conseqüència, no podem calcular l'error mostral i no coneixem si els valors estimats calculats amb la mostra són exactes o no.

El cost i les dificultats de disseny són més reduïts que amb els mètodes probabilístics. Poden donar bons resultats, però també poden produir errors que, amés, són difícils d'acotar i que no sempre s'eliminen augmentant la mida de la mostra.

MÈTODES NO PROBABILÍSTICS

- **Mostres per conveniència**

Com el seu nom indica, s'agafaran unitats mostrals que ens facilitin la seva mesura, que siguin accessibles. Això redueix els costos del mostratge, però la mostra escollida no sol ser massa representativa i per tant les estimacions dels paràmetres poden ser ben diferents dels seus valors poblacionals reals. A més a més no és possible mesurar la precisió de les estimacions o acotar l'error comès.

Permet aconseguir informació *ràpida i bruta*, adient per a estudis exploratoris.

Exemples:

- (1) Sol·licitar la opinió de persones que transiten per un centre comercial.
- (2) Utilitzar un grup proper a l'investigador per a fer una enquesta o consulta.
- (3) Entrevistes a gent del carrer per a un canal de televisió.

- **Mostratge per criteri**

L'investigador determina els elements que s'inclouran en l'estudi en base al seu criteri, és a dir, escull els individus que creu que proporcionen les unitats més representatives. Si el criteri és bo, les mostres seran més representatives que en el mostratge per conveniència, però tampoc podem acotar l'error mostral o mesurar la precisió.

Exemple:

Si s'han d'escollir ciutats per fer proves d'un nou producte, una elecció a l'atzar podria donar ciutats massa petites o poc representatives. L'investigador pot seleccionar les que creu més adients.

- **Mostres compostes d'unitats-tipus**

La població estudiada es descomposa en grups homogenis i s'agafa una única unitat-tipus dins de cada grup.

Exemple:

Per estudiar unes determinades característiques dels automòbils, dividim la població de turismes en grups o categories com (a) utilitaris, (b) familiars, (c) gran-turisme, (d) monovolums, etc. i escollim un vehicle de cada grup que analitzarem en profunditat.

- **Mètode de les quotes**

L'objectiu és reproduir, a escala, la població en la mostra. Es fixen unes determinades proporcions o quotes que del total de la mostra corresponen a unitats d'unues determinades característiques de control. Les unitats s'escullen lliurement però de forma que hi apareguin els percentatges fixats de cada característica o grup específic de població.

Aquestes característiques poden ser el sexe, classe d'edat, nivell d'estudis, etc.

Aquest mètode és similar al mostratge aleatori estratificat, on també es considera la població dividida en estrats i on s'agafa un nombre d'individus d'acord amb el pes de cada estrat en la població. Ara bé, mentre que en el mostratge aleatori estratificat s'agafen els individus a l'atzar dins de cada estrat, aquí es poden escollir per conveniència, la qual cosa, si bé ho fa més ràpid i barat, també pot induir errors, biaixos i minves de representativitat. En tot cas, és el procediment més utilitzat en investigació comercial.

Si les quotes fan referència a més d'una característica, la selecció dels individus pot no ser tan senzilla com sembla a primera vista, ja que cal verificar que compleixin simultàniament varies condicions. En aquest cas parlarem de *quotes creuades*. Si per simplicitat s'agafa cada quota per separat parlarem de *quotes marginals*.

Exemple:

Suposem que considerem les següents característiques:

- 1. Sexe → 2 categories Home/Dona
- 2. Nivell d'estudis → 3 categories Primaris/Medis/Superiors
- 3. Grups d'edat → 5 categories [0,18), [18-30), [30-45), [45-65), [65,∞)

Si volem entrevistar 25 persones per dur a terme un estudi d'intenció de compra, entre majors de 18 anys, podem fer-ho, per quotes marginals, fent servir el full següent:

Característica	nº d'entrevistes totals: 25	Entrevistes realitzades
Sexe		
Home	13	1234567....13
Dona	12	1234567...12
Edat		
18-30	7	1234567
30-45	8	12345678
45-65	4	1234
>65	6	123456
Nivell d'Estudis		
Primaris	10	12345678910
Mitjans	10	12345678910
Superiors	5	12345

Les quotes creuades requereixen que l'enquestador combini simultàniament els tres criteris. Per exemple, en una mostra de 190 persones on s'assignessin les mateixes quotes que en l'estudi d'abans, però simultàniament, tindríem:

Edat	Home			Dona		
	P	M	S	P	M	S
18-30	15	10	5	15	10	5
30-45	15	10	5	15	10	5
45-65	9	6	3	9	6	3
>65	8	4	2	8	4	8

- **Mostratge per rutes aleatòries**

L'entrevistador ha de seguir un camí determinat a l'atzar en la selecció de la mostra. Aquest procediment, si està ben dissenyat, és pràcticament aleatori. Normalment s'utilitza en combinació amb el mostratge per quotes.

Exemple:

En una enquesta electoral l'entrevistador segueix un carrer i escull a l'atzar un número, un pis i una porta per seleccionar l'individu que contestarà l'enquesta.

- **Bola de neu**

Aquest mètode consisteix en escollir una petita sotsmostra, per algun dels mètodes anteriors, per a continuació demanar a cada component que suggereixen altres components, que haurien de ser inclosos en la mostra.

El mètode té l'avantatge de l'especificitat, motiu pel qual, si la població no és nombrosa o està molt especialitzada, es facilita molt la selecció dels individus en la mostra.

És un bon mètode per estudis exploratoris i quan s'entrevisten experts o poblacions marginals.

Exemple:

En un estudi sobre drogues, entrevistem a un grup reduït de drogoaddictes i aquests ens remeten als seus coneguts.

MÈTODES PROBABILÍSTICS O ALEATORIS

- **Mostratge aleatori simple**

La característica d'aquest mostratge és que totes les unitats tenen en un moment donat la mateixa probabilitat de passar a formar part de la mostra. Pot fer-se amb reemplaçament o sense.

En el mostratge aleatori simple *amb reemplaçament* totes les unitats de la població tenen la mateixa probabilitat de ser seleccionades per formar part de la mostra. Aquesta probabilitat no varia a mida que s'efectua el mostratge, ja que un individu pot, teòricament, ser seleccionat més d'una vegada.

En el mostratge aleatori simple *sense reemplaçament*, si la població és finita, la probabilitat que s'esculli una unitat és la mateixa per a totes les unitats que resten en cada extracció i depèn de les que s'han escollit anteriorment per formar part de la mostra.

Per seleccionar una mostra per aquest mètode d'una població finita s'acostumen a utilitzar taules de números aleatoris.

A la pràctica, si la mida de la població és gran amb relació a la mida de la mostra, és indiferent realitzar el mostratge amb o sense reemplaçament, però l'anàlisi resulta més simple quan suposem reemplaçament.

Exemple:

Una enquesta electoral basada en el cens d'una ciutat gran.

Si la població és petita considerarem el mostratge sense reemplaçament si no volem que un individu pugui aparèixer més d'un cop en la mostra.

Exemple:

Una enquesta sobre el socis d'un club amb menys de 1000 carnets.

- **Mostratge aleatori estratificat**

El mostratge aleatori simple s'ha d'utilitzar quan els elements de la població són homogenis respecte a la característica a estudiar, és a dir, quan no coneixem *a priori* quins elements de la població tindran valors alts o petits. Quan disposem d'informació sobre la població que influeix en la característica que estudiem, convé considerar-la per seleccionar la mostra.

Exemple:

En les enquestes d'opinió sabem que les respostes són diferents en funció del sexe, edat, professió, etc.

Ens interessa que la mostra tingui una composició anàloga a la població. Així, la població mare es divideix en classes o estrats *homogenis* i dins de cadascú d'ells es fa un mostratge aleatori simple. La mostra s'agafa assignant un número o quota de membres a cada estrat, de manera que la mostra garanteixi una presència adequada de cada estrat. Això s'aconsegueix dividint la mida total de la mostra (a) proporcionalment a la mida relativa de cada estrat en la població (*afixació proporcional*) o (b) proporcionalment a la variabilitat de l'estrat (*afixació òptima*).

- **Mostratge per conglomerats**

Quan la selecció individual dels elements que componen la mostra és molt costosa, el problema es facilita escollint aleatòriament uns quants grups, és a dir, la unitat de mostratge és un grup amb molts elements que s'anomenen *conglomerats*.

Per poder aplicar aquest sistema, cal que la població estigui dividida en grups sense intersecció per poder escollir una mostra d'aquests i estudiar-ne tots els components o una mostra aleatòria simple.

Es divideix la població en conglomerats i seleccionem com unitats mostrals totes les unitats d'una mostra de conglomerats. Cal que cada conglomerat estigui constituït per unitats elementals heterogènies respecte a la variable que ens interessa. Si és així, el mostratge resulta tan eficient com el mostratge aleatori simple, en cas contrari l'error pot augmentar.

Sovint els conglomerats corresponen a àrees o zones geogràfiques ben definides (províncies, pobles, illes de cases,...). En aquest cas el mostratge també rep el nom de **mostratge per àrees**.

- **Mostratge per etapes o polietàpic**

El sorteig es fa en dues o més etapes. Primer s'extreu un cert nombre de conglomerats, es fa una nova selecció de conglomerats dins de cadascú d'ells i així successivament.

- **Mostratge aleatori sistemàtic**

Si la població presenta un ordre total, no relacionat amb les variables o característiques estudiades, podem escollir una unitat a l'atzar entre les k primeres i a partir d'aquesta de k en k .

1.6 Les dades: Sèries estadístiques

Les dades són el resultat de mesurar les variables estadístiques sobre els individus de la mostra. En conseqüència, una classificació de les dades reflectirà clarament la de les variables que mesurem. Hi ha però, un punt de vista que no s'ha tingut en compte al classificar les variables, i és el de l'instant o interval de temps en que es recullen.

En funció de la importància del temps en el nostre estudi parlarem de:

Dades instantànies. Són dades, agrupades o no, recollides en un instant o interval de temps fixat. També podem considerar aquí les dades a on el temps no té massa importància. A partir de les dades instantànies es solen elaborar les **sèries de freqüències** que reben aquest nom perquè hom les fa comptant les vegades que apareix cada valor, o que cada valor cau dins d'un interval, però sense fixar-se en quin moment apareix cada valor. Per exemple. tirem un dau 25 cops i obtenim els resultats següents:

1 2 5 4 3 1 3 6 4 4 3 1 4 2 3 6 4 6 3 5 6 1 2 3 6

Si els ordenem i comptem quants cops ha aparegut cada puntuació obtenim la sèrie següent:

Puntuació	1	2	3	4	5	6
Freqüència	4	3	6	5	2	5

Dades cronològiques. A cada valor de la sèrie li correspon l'instant del temps en que s'ha mesurat o, simplement, un interval de temps.

Exemple:

any	1995	1996	1997	1998
vendes	24,5	25,8	32,6	36,9

El tractament de sèries cronològiques pressuposa generalment uns intervals de temps fixes.

1.7 L'Estadística en la INVESTIGACIÓ DE MERCATS

Un dels camps principals d'aplicació de l'Estadística és en la investigació comercial. En aquest context apareixen alguns termes i conceptes, que des de la generalitat de l'Estadística, adquireixen un significat particular. Alguns d'ells es presenten a continuació.

1.7.1 Tipus de disseny pels estudis de mercats

Estudis exploratoris

Serveixen per plantejar hipòtesis per a la investigació, descobrir idees o fets rellevants i apuntar solucions o explicacions. Són una fase preliminar o preparatòria dels altres tipus d'estudis ja que permeten la identificació dels problemes i la seva formulació més precisa.

Estudis descriptius

El seu principal objectiu és descriure les característiques d'una determinada situació (comercial, del mercat o de l'entorn) amb l'anàlisi de diferents variables i obtenir apreciacions sobre el seu comportament, en especial el grau d'associació entre aquestes variables.

Estudis explicatius o causals

Orientats a investigar les relacions causa-efecte d'un model no determinista, permeten precisar i controlar els efectes a partir d'una causa coneguda.

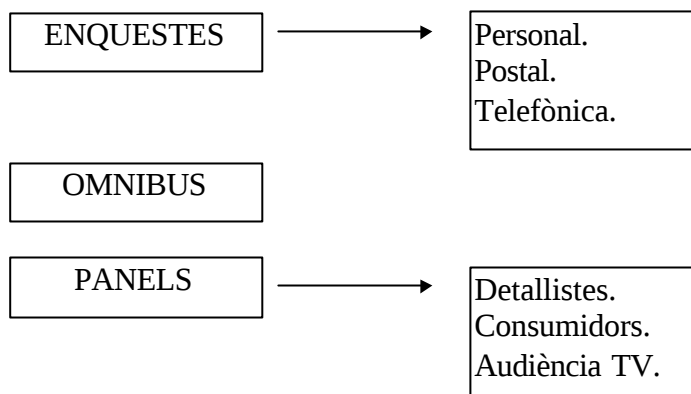
1.7.2 Les fonts d'informació

Les **dades primàries** són les que s'obtenen de forma específica per a la investigació que es vol dur a terme. Solen ser adients però costoses d'obtenir. Es poden obtenir de dues formes bàsiques : per **observació** o per **comunicació**.

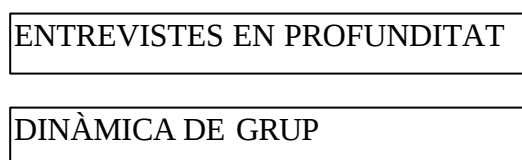
Les **dades secundàries**, són aquelles que ja estan disponibles. Aquestes darreres poden ser internes o externes.

DADES EXTERNES	DADES INTERNES
Socio-econòmiques i demogràfiques. Actituds i opinions. Coneixement o informació. Intencions. Motivacions: Necessitats funcionals. Necessitats simbòliques. Necessitats vivencials. Conductes.	Vendes. Consumidors. Productes. Preus. Comunicació. Equip comercial.

Estudis quantitius



Estudis qualitius



1.7.3 Tipus d'enquestes

	Postal	Telefònica	Entrevista personal
Complexitat de la informació	Mitjana	Escassa	Elevada
Disponibilitat de la informació	no disponible immediatament	disponible immediatament	disponible immediatament
Temps disponible	Molt	Poc	Mitjà
Recursos que disposem (financers/físics)	Escassos	Mitjans	Molts
Taxa de respostes	Escasses	Moderades	Altes

1.7.4 Relació entre la naturalesa de la investigació i els tipus d'informació

	Recursos disponibles escassos	Recursos disponibles importants
Exploratòria	Informacions secundàries. Panel d'experts.	Informacions secundàries. Panel d'experts ampliat. Entrevista de grup. Observació.
Descriptiva	Informacions secundàries. Enquesta.	Informacions secundàries. Enquesta ampliada. Observació.
Explicativa	Informacions secundàries. Panel d'experts. Enquesta.	Informacions secundàries. Experimentació. Observació. Panel d'experts. Entrevista de grup. Enquesta ampliada.

1.8 EXERCICIS

En els següents exercicis, tracteu d'identificar per cada cas:

la *població* de referència
els *individus* observats
les *variables* mesurades, tot especificant
tipus de variable (contínua, discreta, atributs)
recorregut de la variable
mostra, *mida mostral* i *valors mostrals*, si és possible precisar-los.

1. Volem conèixer el nombre de germans que tenen els alumnes d'una determinada escola. Només fem la consulta a la nostra classe, dins l'horari de la nostra assignatura, el dia d'avui i tot preguntant als assistents quants germans tenen.
2. Volem conèixer la facturació bruta i el nombre d'empleats de l'empresa IBEM Espanya, empresa d'informàtica, en els darrers 10 anys. L'empresa té una antiguitat de 99 anys.
3. L'indicador de la borsa russa la darrera setmana ha estat:

Divendres	Dilluns	Dimarts	Dimecres	Dijous	Divendres
244,26	241,38	239,10	241,96	244,26	245,55

4. La companyia ALEPH vol saber com es concentren les vendes entre els seus clients. Per fer-ho, els classifica segons llurs compres estiguin entre 0 i 5 milions, entre 5 i 10, entre 10 i 20 o més milions. Obtenen el següent resultat:

0-5	5-10	10-15	15-20	>20
8	12	50	32	4

5. a) La OMS (Organització Mundial de la Salut) realitzà una estadística sobre 30.000 nens d'arreu el món on s'estudià el pes i la talla corporal a l'any d'edat.
b) En un segon estudi, la OMS recopilà el percentatge de mortalitat infantil de 70 països i la seva renda per càpita.
6. Un alumne afeccionat al futbol comptabilitza els gols que va marcant el seu jugador preferit a cada partit de la lliga i al llarg de tota la temporada.
7. Per encàrrec de l'empresa STAT S.A. fem una selecció de personal altament qualificat en mètodes quantitatius. Després de publicar un anunci als diaris més importants es reben els *curricula*. El cap de personal vol conèixer quants són homes i quants dones, si estan casats i les aspiracions econòmiques que tenen

2 Organització de dades univariants

2.1 Introducció

Una **sèrie estadística** és un conjunt d'observacions o mesures efectuades sobre una mostra d'una població, d'acord a una o més variables.

Col·loquialment, el terme sèrie estadística es coneix per “una estadística”.

El resultat d'una sèrie estadística es presenta sota dos formats:

en forma de taula

en forma de gràfic

La seva presentació depèn del tipus de variable a consideració.

Dins de les sèries estadístiques, en un primer nivell, podem distingir entre:

Sèries simples i sèries agrupades: les mesures s'han pres en un sol instant de temps.

Són sèries instantànies.

Exemple: enquesta sobre el número de germans.

Sèries cronològiques: on les mesures s'han pres en diferents instants de temps.

Exemple: facturació de l'empresa STAT S.A.

2.2 Dades discretes

A un grup de 20 estudiants d'una classe d'una determinada escola se'ls ha preguntat quants germans i germanes tenen, amb el següent resultat:

valors	freq.
0	4
1	5
2	6
3	3
4	1
5	1
Total =	20

A la primera columna estan els diferents valors observats, ordenats del més petit al més gran. A la segona, quantes vegades s'ha observat cada un d'aquests valors. Aquest segon nombre és la *frequència absoluta* d'aparicions.

Si una sèrie estadística correspon a una variable discreta amb recorregut limitat es diu que és de *dades de tipus II*.

Noteu que no es realitza cap agrupació, són les observacions les que es presenten repetides.

x_i	n_i	N_i	f_i	$f_i\%$	F_i	$F_i\%$
0	4	4	0,20	20%	0,20	20%
1	5	9	0,25	25%	0,45	45%
2	6	15	0,30	30%	0,75	75%
3	3	18	0,15	15%	0,90	90%
4	1	19	0,05	5%	0,95	95%
5	1	20	0,05	5%	1,00	100%
Totals	20		1,00	100%		

on x_i són els valors observats
 n_i les freqüències absolutes
 N_i les freqüències absolutes acumulades
 f_i les freqüències relatives
 $f_i\%$ les freqüències relatives percentuals
 F_i les freqüències relatives acumulades i
 $F_i\%$ les freqüències relatives acumulades percentuals

Donat un valor a , les freqüències acumulades ajuden a quantificar els individus que verifiquen la condició

X més petit o igual que a

Representació gràfica

1. Diagrama de varetes

Una forma molt usual de representar dades discretes és dibuixar els punts en un gràfic cartesià, amb l'alçada de la freqüència absoluta i unint-los amb el eix d'abscisses per una línia.

2. Diagrama de barres

La base de les barres és constant, d'amplada arbitrària, però fixa per a tots els valors. L'alçada correspon a la freqüència absoluta dels valors observats.

3. Pictogrames

Es repeteix una determinada figura o símbol de forma proporcional a la freqüència de cada valor observat. Són utilitzats sovint en mitjans de comunicació, propaganda, etc.

4. Diagrama de barres compostes

En les seves diferents opcions (apilades, aparellades, múltiples) són adients per comparar gràficament dues o més sèries.

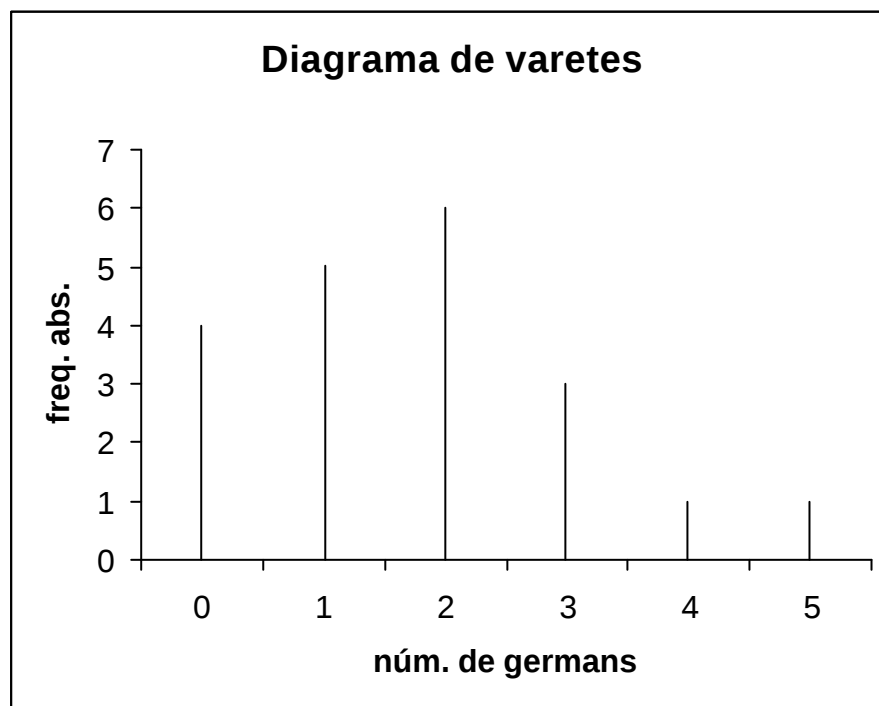
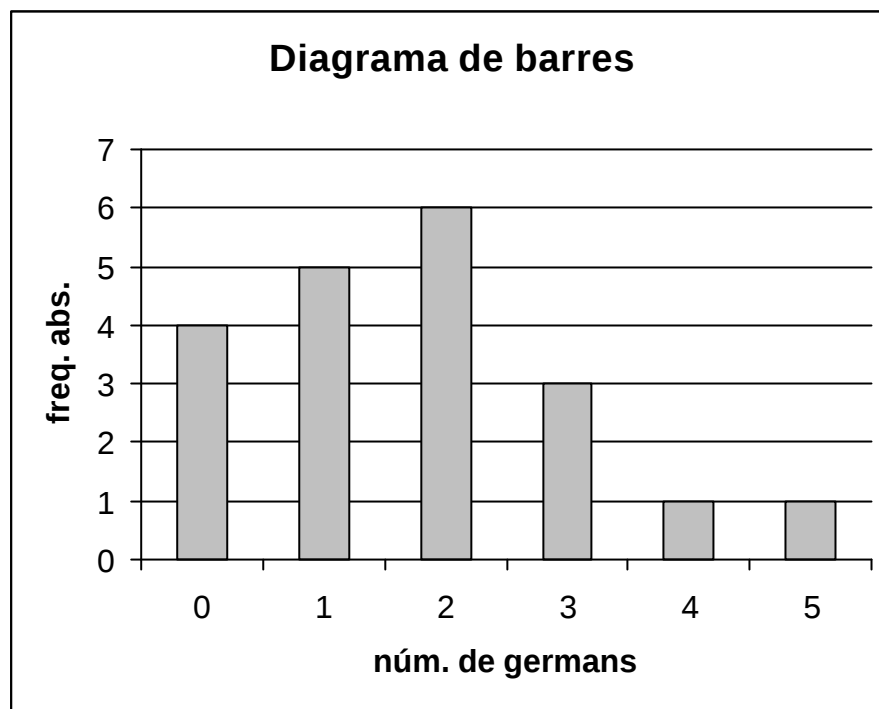
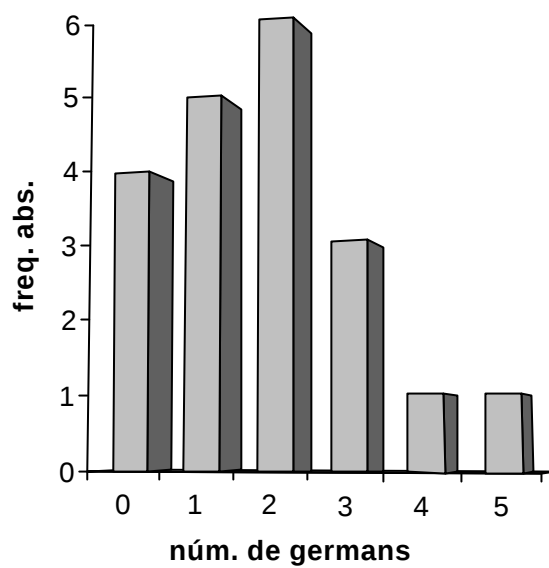
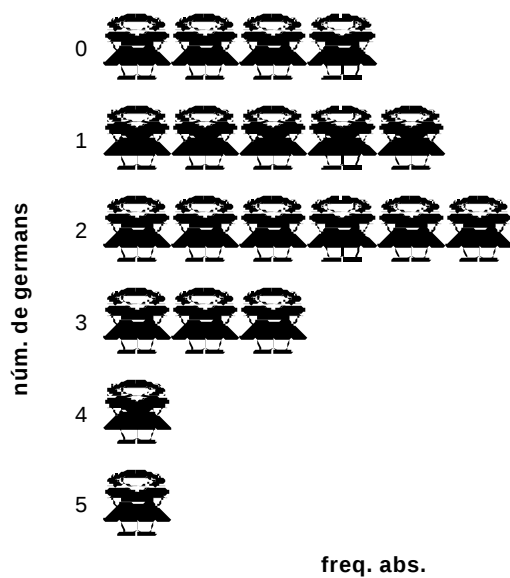


Diagrama de barres 3D



Pictograma



2.3 Atributs

Les *dades de tipus IV* corresponen a estudis estadístics on la variable observada qualifica *atributs*. La forma natural de descriure aquestes dades és classificar-les en *categories* o classes definides d'acord a aquests atributs, comptabilitzant per cada una de les diferents categories, quantes observacions posseeixen l'atribut que les defineix.

Veiem un exemple:

Una marca d'importació de cotxes treu al mercat un nou model. D'aquest model, se'n fabriquen quatre colors: Gris marengo, Vermell, Blau metal·litzat i Blanc. Després de mig any de vendes, i de cara a la segona tanda de fabricació, demana les dades de venda als seus distribuïdors amb aquest resultat:

color	Marengo	Blau	Vermell	Blanc
unitats	230	170	80	40

La variable mesurada és *qualitativa*, s'observen atributs, concretament els colors dels individus (els cotxes venuts). Hem dividit la població (tots els cotxes venuts) en quatre *categories excloents*, definides pel color.

Les quatre quantitats (230, 170, 80 i 40) s'anomenen *freqüència absoluta* de la categoria, de forma totalment equivalent a la freqüència d'un valor discret en el cas de les dades de tipus II o la d'una classe en les de tipus III.

És important recordar que donada una variable qualitativa qualsevol, les diferents categories han d'ésser *excloents* per parlar de freqüència absoluta. En cas contrari podríem assignar una mateixa observació a dues categories.

Definida la freqüència absoluta d'una classe, podem estendre les definicions que fèiem abans de freqüències acumulades, relatives i percentuals a les classes.

Amb l'exemple del color dels cotxes:

color	n_i	N_i	f_i	$f_i\%$	F_i	$F_i\%$
Marengo	230	230	0,44	44%	0,44	44%
Blau	170	400	0,33	33%	0,77	77%
Vermell	80	480	0,15	15%	0,91	91%
Blanc	40	520	0,07	7%	1,00	100%
Totals	520		1,00	100%		

Encara que, en aquest cas, les freqüències acumulades no tenen gaire sentit, ja que la acumulació es fa amb un ordre arbitrari.

Representació gràfica

De forma semblant a les dades de tipus II, podem representar les dades de tipus IV amb els següents diagrames:

Diagrama de barres

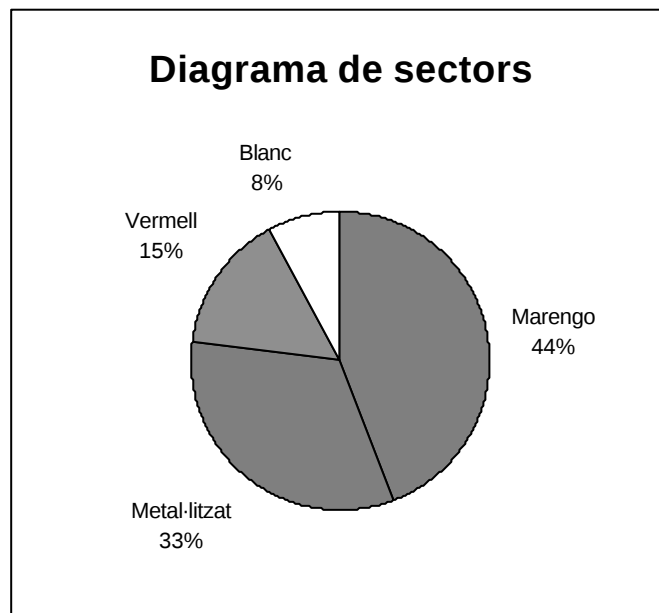
Malgrat que per dades d'escala nominal és estrictament correcte, no és gaire aconsellable utilitzar els diagrames de barres per aquest tipus de dades. La raó és que inconscientment s'associa la presència de l'eix horitzontal a algun tipus d'ordre entre els atributs. En tot cas, es recomana ordenar els atributs en funció de les freqüències absolutes, de major a menor. En aquest cas el diagrama de barres s'anomena *diagrama de Pareto*.

Si la escala de la variable observada és ordinal, no hi ha cap inconvenient.

Per representar dades nominals, el més adient és representar-les amb un diagrama de sectors.

Diagrama de sectors

També conegut com *diagrama de pastís* per traducció literal de l'anglès 'piechart'. Donat un cercle de radi arbitrari, a cada valor es fa correspondre una porció d'àrea marcada per un determinat angle, que és proporcional a la seva freqüència relativa. L'àrea total són 360°.



2.4 Dades contínues

Si una sèrie estadística és la d'una variable *contínua* i la mostra és de *mida moderada*, direm que es tracta de *dades de tipus I*.

Són sèries on, per el petit nombre de dades de la mostra, *no cal agrupar les dades*.

A efectes gràfics les possibilitats clàssiques són limitades i aquest tipus de dades no es poden tabular.

Exemple:

En un control de qualitat d'una refineria de sucre, s'ha mesurat el pes de 15 caixes de paquets de sucre. Cada caixa conté 25 paquets que, en principi, han de ser d'un quilo. Les observacions són:

caixa	pes	caixa	pes	caixa	pes
1	25,2	6	24,9	11	24,9
2	24,9	7	24,9	12	24,7
3	25,1	8	24,7	13	25,2
4	25,0	9	25,0	14	25,0
5	24,9	10	25,0	15	25,0

La població estudiada és el conjunt de caixes del magatzem.

La mostra són les 15 caixes observades i els valors mostrals el pes d'aquestes.

La variable és contínua.

Important! Cada observació té anotada de forma individual el seu corresponent valor. Preneu nota que *en cap cas hem agrupat* les observacions.

Un diagrama força adequat, que estudiarem més endavant, és el anomenat de *tija i fulles* (stem-leaf display). Per a dades més nombroses també pot ser útil el gràfic anomenat de caixa (box-plot).

2.5 Dades agrupades

Per a les sèries estadístiques de variables quantitatives contínues amb un recorregut força ampli, l'observació dels valors mostrals individuals quasi bé no proporciona cap informació. Cal agrupar aquestes dades en intervals, també anomenats *classes*.

De les dades així agrupades em direm *dades de tipus III*.

Aquesta agrupació és també vàlida per a variables discretes amb, com a mínim, una dotzena de valors observats diferents.

Forma d'agrupar les dades:

Una classe agrupa sota una mateixa etiqueta un o més valors possibles de la variable.

Llavors, es comptabilitzen quantes observacions estan incloses dins d'aquesta classe, és a dir, s'anota la *freqüència absoluta* de la classe.

De la mateixa manera que en les dades qualitatives, les categories construïdes mitjançant les classes *han d'ésser excloents*.

Veiem el següent exemple:

Per encàrrec del gerent de l'empresa STAT S.A. es vol estudiar el perfil dels sous nets mensuals dels treballadors de l'empresa, inclosos els directius. En total hi ha 61 persones. Algunes de les dades, en pessetes, són les següents:

67.723, 78.750, 123.997, 254.010, 78.750, 72.900,
....., 220.900, 427.901, 89.900

Una forma usual de procedir és classificar els sous en categories formant, per exemple, les 8 categories següents:

Interval	Freqüència
50-100	17
100-150	19
150-200	13
200-250	7
250-300	3
300-350	1
350-400	0
400-450	1
Total:	61

Si parlem en general, es consideren, com en les dades de tipus II, els conceptes de *freqüència absoluta*, *relativa* i *percentual*, que ara aniran referides a categories construïdes mitjançant *classes*.

Les classes poden incloure més d'un valor possible. Frequentment s'empra una *marca de classe* per designar a la classe sencera, que acostuma a coincidir amb el valor central de l'interval.

A l'hora de construir la taula existeix un problema: Quantes classes formem i de quina manera establim els intervals?

Procediment sistemàtic de construcció

1. Determineu el mínim i el màxim dels valors observats per tal de calcular el recorregut o *rang de variació* de les dades, és a dir:

$$\text{rang de variació} = \text{valor més gran} - \text{valor més petit}$$

2. Decidiu el nombre d'intervals i la seva amplada.

Per això, una regla empírica pot ser fer al voltant de 15 intervals, mai més de 30 ni menys de 6, normalment s'utilitzen entre 7 i 20.

Encara que no és obligatori, és aconsellable que l'amplada sigui per tots igual. En aquest cas, l'amplada es calcula simplement com

$$\text{amplada} = \text{rang de variació} / \text{nombre d'intervals}$$

3. Calculeu els extrems dels intervals, també anomenats fronteres de classe, tot recordant que els intervals sempre han de ser excloents. En aquest sentit, hi ha una norma generalment acceptada de forma que els intervals adjacents són semioberts, és a dir, l'extrem inferior no pertany a l'interval.

És aconsellable que les marques no coincideixin amb cap dels valors observats.

4. Calculeu les marques de classe de cada interval:

$$\text{marca} = (\text{valor més gran de l'interval} + \text{valor més petit}) / 2$$

5. Construïu la taula de freqüències.

El punt 2 és molt problemàtic. Un criteri addicional que pot ajudar a l'hora de construir els intervals és escollir-ne un nombre tal que per un 90% de les classes hi hagin com a mínim 5 representants.

D'altra banda, diversos autors han proposat fórmules per calcular el número d'intervals en funció de la mida de la mostra:

$$\text{número de classes} = 2,5\sqrt[4]{n}$$

$$10 \times \log_{10} n \quad \text{Dixon and Kronmal (1965)}$$

$$2\sqrt{n} \quad \text{si } n < 50 \quad \text{Velleman (1976)}$$

$$1 + \log_2 n \quad \text{si } n = 2^s \quad \text{Sturges (1926)}$$

Histogrammes

Un histograma consisteix en un conjunt de rectangles juxtaposats amb les següents característiques:

Les bases dels rectangles es recolzen sobre l'eix horitzontal. Aquest eix té una escala adequada i uniforme.

El punt central de cada base coincideix amb la marca de classe.

L'amplada de cada rectangle és igual a l'amplada de l'interval corresponent.

L'àrea és proporcional a la freqüència absoluta de la classe.

Quan l'amplada és igual per a totes les classes, el que facilita molt la comprensió de la representació, el costum és agafar l'alçada del rectangle exactament igual a la freqüència absoluta de la classe. En cas contrari, tindrem que ajustar les alçades d'acord al punt 4. Per exemple, podem dibuixar l'alçada que resulta al dividir la freqüència absoluta per la longitud de l'interval.

Polígons de freqüències

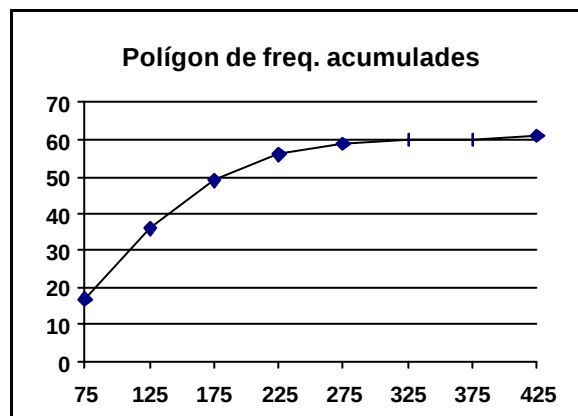
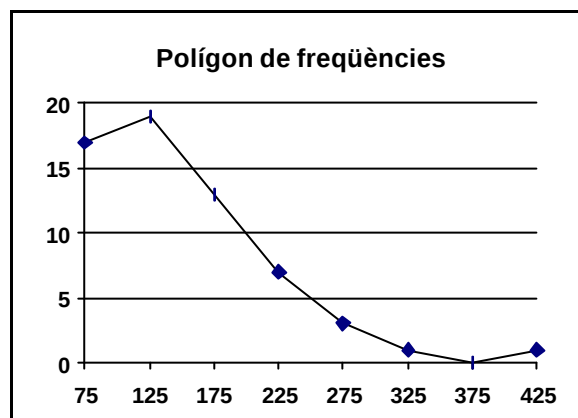
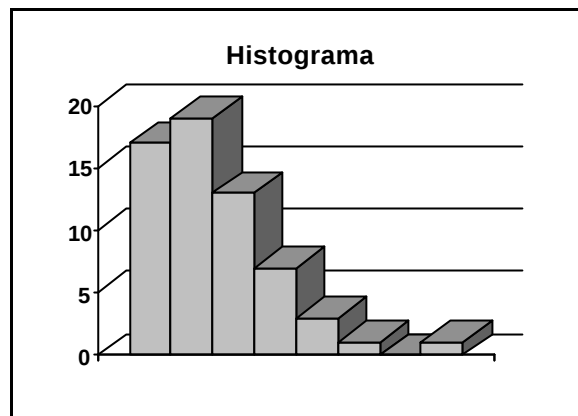
Es basa en unir els punts definits per la marca de classe de cada interval com primera coordenada i com segona la seva corresponent freqüència de classe.

També pot obtenir-se unint els punts mitjos situats a la part superior de les barres de l'histograma.

Es poden representar també polígons de freqüències acumulades, en lloc de les absolutes.

Ja s'ha comentat abans la utilitat de les freqüències acumulades: permeten respondre fàcilment preguntes de la forma “quantes observacions verifiquen ésser més petites que un cert valor?”

En el exemple dels sous mensuals nets de l'empresa STAT S.A. podem fer aquests gràfics:



2.6 Sèries cronològiques

Una sèrie cronològica és aquella on les dades s'han mesurat en diferents unitats de temps.

La variable observada en la sèrie tant pot ser quantitativa (discreta o contínua) com qualitativa.

Des del punt de vista matemàtic, s'estudia la variable X com una funció del temps:

$$X = F(t)$$

Considerem el següent exemple:

En l'empresa STAT S.A. s'han facturat en els sis darrers anys les següents quantitats (en milions de PTA):

Any	1993	1994	1995	1996	1997	1998
Fact.	278,3	267,4	287,2	292,3	299,2	303,1

Aquí la unitat de temps és l'any, però en general no té per què ser així. Altres unitats vàlides poden ser: biennis, triennis, quinquennis, lustres o segles, o bé trimestres, mesos, dies, hores o segons.

La forma més usual de representació gràfica és força simple:

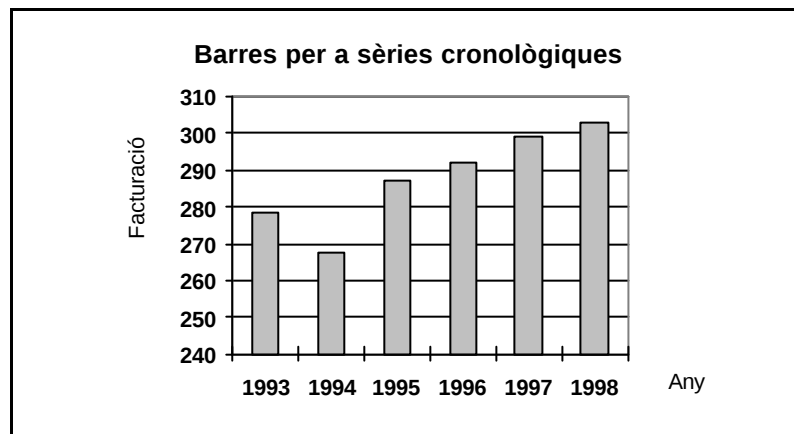
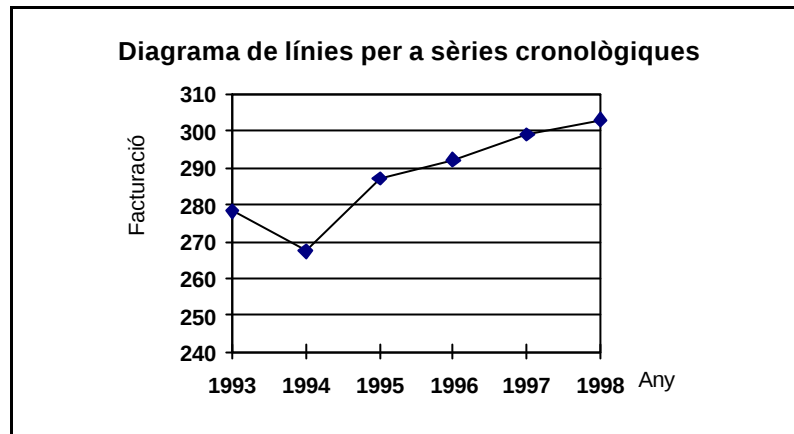
En el pla, l'eix horitzontal indicarà el temps (és molt important explicitar la unitat de temps)

L'eix vertical indica el valor de la variable observada.

Amb aquesta representació podem optar per:

Diagrama de línies (millor)
Diagrama de barres

Amb l'exemple anterior obtenim els gràfics:



El diagrama de línies és un gràfic molt similar a un polígon de freqüències, però no s'han de confondre, els tipus de dades que representen l'un i l'altre són totalment diferents.

L'estudi de les sèries cronològiques es mereix una atenció especial. És molt important l'anàlisi del comportament al llarg del temps, que pot tenir dos components:

Tendències a llarg termini: que poden ser creixents o decreixents. Per exemple, en la sèrie de l'empresa STAT S.A. la tendència és creixent.

Moviments cíclics, també anomenats oscil·lacions. Poden ser periòdics (cada mes, cada any, etc.)

2.7 EXERCICIS

1. Es pregunta a una mostra de persones quantes vegades ha vist les darreres sis setmanes un determinat programa de televisió que s'emet setmanalment. Les dades són les següents:

núm. de vegades	0	1	2	3	4	5	6
núm. d'individus	3	5	7	9	6	7	4

Construeix una taula estadística on apareguin les freqüències absolutes, les freq. relatives i les freq. relatives acumulades de la variable “nombre de vegades que ha vist el programa”, així com el diagrama de barres corresponent a les freq. absolutes.

2. Enquestades 50 persones pel seu temps d'espera (en minuts) en una parada d'autobús, s'obtingueren els següents resultats:

2	4	8	1	0	10	9	9	7	2	5	6
4	2	7	1	6	3	8	10	3	0	6	10
1	5	2	5	1	3	1	3	6	8	5	4
4	6	3	6	0	7	9	3	2	7	8	0
7	7										

Construeix una taula estadística on apareguin les freq. absolutes, les relatives i les acumulades tant relatives com absolutes. Construeix els diagrames de barres i de varetes per a les freq. absolutes.

3. Una empresa que desitja treure al mercat una modalitat nova de joguina didàctica dirigida a nens en edat pre-escolar, efectua un sondeig sobre una mostra de 30 nens/es d'edat pre-escolar tot preguntant als pares quantes joguines didàctiques té el seu fill. Els resultats són:

6	3	6	1	5	2	5	2	1	1	6	3
4	2	4	5	3	2	4	3	6	6	2	4
2	5	5	4	2	1						

Tabuleu i representeu de la forma més adient una gràfica que resumeixi les dades. Quin és el percentatge de nens amb com a mínim 3 joguines?

4. El nombre de dies de treball perduts per malaltia durant un determinat trimestre pels 35 treballadors/es d'una cadena de muntatge són els següents:

2	1	0	1	1	3	0	0	2	7	5	0
1	3	0	0	4	1	2	4	0	5	3	0
6	0	4	0	2	6	2	3	0	1	1	

Tabuleu i representeu de la forma més adient una gràfica que resumeixi les dades. Quin és el percentatge de treballadors amb 4 o més absències?

5. Amb les dades proporcionades amb la taula següent, completa la sèrie de freqüències absolutes n_i , absolutes acumulades N_i , relatives f_i i relatives acumulades F_i :

x_i	n_i	N_i	f_i	F_i
10	2	2	0,05	0,05
13	4	6	0,10	0,15
16		16		0,40
19	15			
22	6	37	0,15	0,925
25				1

6. Investigats els preus per habitació de 50 hotels d'una ciutat, s'han obtingut els següents resultats (en milers de PTA):

7	3	5	4	5	7	4	7,5	8	5	5	7,5
3	7	10	15	5	7,5	12	8	4	5	3	5
10	3	4	5	7	5	3	4	7	4	7	5
4	7	10	7,5	7	8	7,5	7	7,5	8	7	12
8	14										

Representeu la distribució dels preus:

- Agrupats en una sèrie de freqüències
- Agrupats en sis intervals d'igual amplada

7. Les dades que es detallen tot seguit corresponen als litres de combustible gastats pels vehicles de transport d'una determinada empresa al llarg d'una setmana:

35	30	12	46	49	54	56	59	60	70	67	75
93	18	32	44	46	50	55	58	21	60	72	67
77	36	45	47	73	69	24	63	74	65	77	56
41	80	52	65	103	45	48	56	58	74	80	71
24	54	48	39	44	88	54	70	74	71	28	45
68											

Tabuleu i representeu la distribució de freqüències de la forma més adient.

8. Els desplaçaments dels operaris/es d'una factoria per tal d'arribar al seu lloc de treball ha estat:

Km	més d'1	més de 15	més de 30	més de 45	més de 60	més de 75
núm. oper.	91	30	13	6	5	4

Tabuleu i representeu la distribució de freqüències de la forma més adient.
Quants operaris/es realitzen un desplaçament comprés entre 30 i 60 km?

9. Classificades les províncies de tot l'estat pel nombre d'habitants (dades de 1986), s'obtingueren els següents resultats:

núm. d'habitants	núm. de províncies
1-100.000	3
100.000-250.000	8
250.000-500.000	16
500.000-750.000	10
750.000-1.000.000	4
1.000.000-2.000.000	8
2.000.000-3.000.000	1
3.000.000-4.000.000	0
4.000.000-5.000.000	2

Dibuixeu l'histograma corresponent i amb les marques de classe adequades.

3 Estadístics

3.1 Mesures de posició, de dispersió i de forma

Un cop hem recollit i ordenat les dades, el pas següent consisteix en caracteritzar-les d'alguna manera, és a dir, buscar mesures que ens permetin, a través seu:

- Descriure de forma condensada les dades recollides.
- Diferenciar la població a la que pertany la mostra d'altres poblacions similars.

Per fer-ho ens basem principalment en tres tipus de mesures:

- Mesures de posició
- Mesures de dispersió
- Mesures de forma

Aquestes mesures seran funció dels valors mostrals i s'anomenen estadístics.

3.1.1 Mesures de posició

En trobem de dues menes :

Mesures de tendència central

Es basen en l'obtenció d'un valor que es troba, en algun sentit, en una posició centrada respecte les dades.

Les mesures de tendència central que tenen major interès són:

Mitjana aritmètica.

Valor mig de les dades.

Mediana.

És el valor central de les dades, és a dir, un número tal que la meitat de les dades són inferiors i l'altra meitat són superiors.

Moda.

És el valor (o valors) més freqüent.

També es poden calcular altres mitjanes com la mitjana ponderada o la geomètrica.

Altres mesures de posició

El valor que busquem es troba en una posició qualsevol d'interès per a nosaltres. Per exemple, el valor sota del qual es troba el 95% de les dades.

Les mesures de posició que estudiarem són:

- Els percentils.
- Altres quantils (decils, quartils,...).

3.1.2 Mesures de dispersió

Mesuren la dispersió de les dades respecte d'algun punt, normalment d'una mesura de tendència central com la mitjana aritmètica.

Les més importants a retenir són:

La variància.

Valor mitjà de les desviacions (al quadrat) de les dades respecte a la seva mitjana.

La desviació típica.

És l'arrel quadrada de l'anterior.

Coeficient de variació.

És el quocient entre la desviació típica i la mitjana.

Altres.

El *Rang* és la diferència entre el mínim i el màxim.

El *Rang interquartílic* és la diferència entre els quartils superior i inferior.

3.1.3 Mesures de forma

Ens donen alguna orientació sobre la forma de la representació gràfica.

Mesures d'asimetria

Indiquen en quin grau la corba (el polígon de freqüències o l'histograma) es simètrica o no.

Coeficient d'asimetria. Quan val 0 les dades són simètriques, quan és positiu la cua superior té més àrea que la inferior i el contrari en cas de que sigui negatiu.

Mesures d'apuntament

Indiquen què tan apuntada ("punxeguda") és una corba (tot comparant-la, habitualment, amb la distribució de densitat Normal).

La més usual és el *Coeficient de kurtosi o apuntament*. Indica què tan apuntada és la distribució de les nostres dades en relació a la llei Normal. Per a la normal val 0. Quan el coeficient és inferior a 0 la corba és plana i quan és superior la corba és molt apuntada.

3.2 Mesures de tendència central

Mitjana aritmètica

La mesura més comú del centre d'unes dades és el valor mig dels seus valors o mitjana aritmètica.

Si $x_1, x_2, \dots, x_n$ són les n observacions, la seva mitjana és

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

o de forma més compacte

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

La taula resum dels estadístics mostra com es calcula la mitjana en els casos de dades agrupades.

Mediana

Una altra forma de descriure el centre d'unes dades és utilitzar el seu valor central. La mediana és un número tal que la meitat de les dades són inferiors i l'altra meitat són superiors.

Per tal de calcular-la s'ordenen totes les dades de forma creixent.

Si el número d'observacions és imparell, la mediana és l'observació central de la llista ordenada. La posició central és la número $(n+1)/2$ des del començament de la llista i l'observació que es trobi en aquesta posició és la mediana.

Si el número d'observacions n és parell, aleshores la mediana és la mitjana de les dues observacions centrals de la llista ordenada. Serà doncs, la mitjana de les observacions que es trobin en les posicions $n/2$ i $(n/2)+1$ des del començament de la llista ordenada.

La taula resum dels estadístics mostra el càlcul especialment delicat de la mediana en el cas de dades tipus III.

Comparació entre la mitjana i la mediana

És possible que hi hagi una diferència molt notable entre la mitjana i la mediana d'unes dades. De vegades, uns pocs valors extrems a un costat de les dades poden arrossegar la mitjana de forma prou important, mentre que la mediana resta quasi invariable davant la presència o absència d'aquests valors atípics. Aquesta propietat de ser robustes és la principal característica de tots els estadístics d'ordre, com la mediana, que estudiarem més endavant. Això no vol dir que la mediana sigui sempre millor que la mitjana, el que convé és oferir les dues mesures, conjuntament amb un gràfic.

La mitjana i la mediana de un conjunt de dades amb distribució simètrica estan molt a prop. Si la distribució és *exactament* simètrica, són iguals. En una distribució asimètrica, la mitjana queda desplaçada cap a la cua més llarga. Moltes de les variables econòmiques o de població són asimètriques a la dreta, hi ha molts valors petits i uns pocs de grans. Els pocs valors grans tiren de la mitjana, de forma que està per sobre de la mediana. En aquests casos cal donar la mediana com mesura central.

Mesures de tendència central

Estadístic	Sèries tipus I	Sèries tipus II	Sèries tipus III
Mitjana aritmètica	$\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$	$\bar{x}_n = \frac{1}{n} \sum_{j=1}^k x_j n_j$	$\bar{x}_n = \frac{1}{n} \sum_{j=1}^k \tilde{x}_j n_j$
Mediana	Si n és imparell, $Me = x_{(m)}$ on $m = \frac{n+1}{2}$ Si n és parell, $Me = \frac{x_{(m)} + x_{(m+1)}}{2}$ on $m = \frac{n}{2}$		Si j és l'interval medià, $L_{j-1} + \frac{a_j}{n_j} \left(\frac{n}{2} - N_{j-1} \right)$
Moda		El valor x_j amb major freqüència absoluta	Si j és l'interval modal, $L_{j-1} + a_j \frac{h_{j+1}}{h_{j-1} + h_{j+1}}$
Mitjana ponderada $\bar{x}_p$	$\frac{\sum_{i=1}^n x_i p_i}{\sum_{i=1}^n p_i}$	$\frac{\sum_{j=1}^k x_j p_j n_j}{\sum_{j=1}^k p_j}$	$\frac{\sum_{j=1}^k \tilde{x}_j p_j n_j}{\sum_{j=1}^k p_j}$
Mitjana geomètrica $\bar{x}_g$	$\sqrt[n]{x_1 x_2 \cdots x_N}$	$\sqrt[n]{x_1^{n_1} x_2^{n_2} \cdots x_k^{n_k}}$	$\sqrt[n]{\tilde{x}_1^{n_1} \tilde{x}_2^{n_2} \cdots \tilde{x}_k^{n_k}}$
Mitjana harmònica $\bar{x}_h$	$\frac{1}{\bar{x}_h} = \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}$	$\frac{1}{\bar{x}_h} = \frac{1}{n} \sum_{j=1}^k \frac{1}{x_j} n_j$	$\frac{1}{\bar{x}_h} = \frac{1}{n} \sum_{j=1}^k \frac{1}{\tilde{x}_j} n_j$
Mitjana quadràtica $\bar{x}_q$	$\frac{1}{n} \sum_{i=1}^n x_i^2$	$\frac{1}{n} \sum_{j=1}^k x_j^2 n_j$	$\frac{1}{n} \sum_{j=1}^k \tilde{x}_j^2 n_j$

Sèries tipus I: Els valors observats són $x_1, x_2, \dots, x_n$. La mida de la mostra n .

Sèries tipus II: Els valors observats són $x_1, x_2, \dots, x_k$, amb les freqüències absolutes $n_1, n_2, \dots, n_k$. La mida de la mostra és $n = n_1 + n_2 + \dots + n_k$. Les freqüències acumulades són $N_1, N_2, \dots, N_k$.

Sèries tipus III: Els intervals són $(L_0, L_1], (L_1, L_2], \dots, (L_{k-1}, L_k]$. Les marques de classe s'indiquen amb $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_k$, i les freqüències amb la mateixa notació que per les sèries tipus II. L'amplada de la classe és $a_j = L_j - L_{j-1}$ i l'alçada $h_j = n_j / a_j$.

En tots els casos, $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ és la mostra ordenada de forma creixent. Per a les sèries tipus II, amb els valors repetits tantes vegades com indiqui la freqüència absoluta de cadascú.

3.3 Mesures de dispersió

Rang o recorregut

La forma més simple d'indicar la dispersió d'unes dades és calcular el seu rang o recorregut, és a dir, la diferència entre les observacions màxima i mínima.

Variància

La variància s^2 d'un conjunt d'observacions és la mitjana dels quadrats de les desviacions de les observacions respecte a la seva mitjana. La fórmula és la següent:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}$$

o, d'una forma més simple,

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

La *desviació típica* s , també anomenada *desviació estàndard*, és l'arrel quadrada positiva de la variància s^2 .

La desviació típica té les mateixes unitats que les observacions, mentre que la variància està mesurada en unitats al quadrat.

El *coeficient de variació* és el quocient de la desviació típica i la mitjana, expressat en tan per cent. És una mesura adimensional i, per tant, permet la comparació de les dispersions de conjunts de valors amb diferents unitats.

Variància corregida

És molt fàcil comprovar que la suma de les desviacions és sempre zero:

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

de forma que la última desviació es pot calcular quan es coneixen les altres $n-1$. Només $n-1$ de les desviacions poden variar lliurement i, per això el número $n-1$ s'anomena *graus de llibertat* de la variància o de la desviació típica. Per aquest motiu, donada una mostra d'observacions, és millor calcular la variància corregida:

$$\hat{s}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

o la desviació típica corregida $\hat{s}$.

Moltes calculadores ofereixen el càlcul de les variàncies i desviacions típiques corregides o no. Per a una mostra anotarem els valors corregits.

En tot cas, tan s com $\hat{s}$, mesuren la dispersió respecte a la mitjana. S'utilitzaran només quan escollim la mitjana com mesura de centre.

Si no hi ha dispersió, $\hat{s} = 0$. Això passa únicament quan totes les observacions tenen el mateix valor. Al contrari, $\hat{s} > 0$ i quan més disperses estiguin les observacions respecte a la seva mitjana, més gran es fa.

Com passava amb la mitjana, també $\hat{s}$ està fortament influenciada per les observacions extremes. Unes poques observacions atípiques poden fer que $\hat{s}$ sigui molt gran. Més endavant, en aquest mateix capítol, proposarem mesures de dispersió més robustes, com el rang interquartílic.

En general, utilitzarem $\bar{x}$ i $\hat{s}$ només per distribucions simètriques i, particularment, quan les dades provenen d'una distribució normal. En tot cas, és obvi, que de vegades hi ha males interpretacions perquè dos números no poden resumir bé un conjunt de dades. Recordeu d'acompanyar els estadístics amb un o més gràfics.

3.4 Mesures de forma

3.4.1 Mesures d'asimetria

En un conjunt simètric de dades de tipus III, el estadístic $\sum_{i=1}^n (x_i - \bar{x}_n)^3$ és zero. Amb dades asimètriques aquesta suma creixerà. Una mesura adimensional d'asimetria és:

Estadístic	Sèries tipus I	Sèries tipus II	Sèries tipus III
Coefficient d'asimetria	$\frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^3}{s^3}$	$\frac{\frac{1}{n} \sum_{j=1}^k (x_j - \bar{x}_n)^3 n_j}{s^3}$	$\frac{\frac{1}{n} \sum_{j=1}^k (\tilde{x}_j - \bar{x}_n)^3 n_j}{s^3}$

Una altra mesura d'asimetria menys utilitzada és el coeficient d'asimetria de Pearson:

$$\frac{3(\bar{x}_n - Me)}{s}$$

Aquest coeficient resulta de la relació empírica $3(\bar{x}_n - Me) = \bar{x}_n - Mo$ que es verifica per a dades unimodals i en forma de campana.

3.4.2 Mesures de kurtosi

Estadístic	Sèries tipus I	Sèries tipus II	Sèries tipus III
Coefficient de kurtosi	$\frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^4}{s^4}$	$\frac{\frac{1}{n} \sum_{j=1}^k (x_j - \bar{x}_n)^4 n_j}{s^4}$	$\frac{\frac{1}{n} \sum_{j=1}^k (\tilde{x}_j - \bar{x}_n)^4 n_j}{s^4}$

Donat que aquest coeficient val 3 per a la distribució Normal, alguns autors i programes d'ordinador defineixen el coeficient d'apuntament com

$$\frac{m_4}{s^4} - 3$$

on m_4 és el moment centrat de quart ordre, és a dir, el numerador dels estadístics definits a la taula anterior.

Mesures de dispersió

Estadístic	Sèries tipus I	Sèries tipus II	Sèries tipus III
Rang	$R = x_{(n)} - x_{(1)}$		$R = L_k - L_0$
Variància	$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$ $= \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}_n^2$	$s^2 = \frac{1}{n} \sum_{j=1}^k (x_j - \bar{x}_n)^2 n_j$ $= \frac{1}{n} \sum_{j=1}^k x_j^2 n_j - \bar{x}_n^2$	$s^2 = \frac{1}{n} \sum_{j=1}^k (\tilde{x}_j - \bar{x}_n)^2 n_j$ $= \frac{1}{n} \sum_{j=1}^k \tilde{x}_j^2 n_j - \bar{x}_n^2$
Desviació típica	$s = \sqrt{s^2}$		
Desviació típica corregida	$\hat{s} = \sqrt{\frac{n}{n-1}} s^2$		
Coefficient de variació	$CV = \frac{s}{\bar{x}_n} 100$		
Mitjana de les desviacions absolutes	$\frac{1}{n} \sum_{i=1}^n x_i - \bar{x}_n $	$\frac{1}{n} \sum_{j=1}^k x_j - \bar{x}_n \cdot n_j$	$\frac{1}{n} \sum_{j=1}^k \tilde{x}_j - \bar{x}_n \cdot n_j$
Mediana de les desviacions absolutes	$Me\{ x_i - Me \}$	$Me\{ x_{(i)} - Me \}$	

Sèries tipus I: Els valors observats són $x_1, x_2, \dots, x_n$. La mida de la mostra n .

Sèries tipus II: Els valors observats són $x_1, x_2, \dots, x_k$, amb les freqüències absolutes $n_1, n_2, \dots, n_k$. La mida de la mostra és $n = n_1 + n_2 + \dots + n_k$. Les freqüències acumulades són $N_1, N_2, \dots, N_k$.

Sèries tipus III: Els intervals són $(L_0, L_1], (L_1, L_2], \dots, (L_{k-1}, L_k]$. Les marques de classe s'indiquen amb $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_k$, i les freqüències amb la mateixa notació que per les sèries tipus II. L'amplada de la classe és $a_j = L_j - L_{j-1}$ i l'alçada $h_j = n_j / a_j$.

En tots els casos, $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ és la mostra ordenada de forma creixent. Per a les sèries tipus II, amb els valors repetits tantes vegades com indiqui la freqüència absoluta de cadascú.

3.5 Diagrama de tija i fulles

Es tracta d'un procediment semi-gràfic per tal de presentar la informació per a sèries de dades quantitatives, que és especialment útil quan la mida de la mostra no és excessivament gran, és el diagrama de tija i fulles de Tukey. Té l'avantatge de ser molt menys arbitrari que un histograma i molt més senzill en la seva construcció. El procediment és el següent:

- a) Arrodonim les dades observades a dues o tres xifres significatives.
 - b) Disposem aquestes xifres en una taula amb dues columnes separades per una línia vertical:
 - b1) A la primera columna posarem la primera o dues primeres xifres que formen la tija.
 - b2) A la segona columna escriurem les fulles o últimes xifres, repetides tantes vegades com apareguin a les dades.
 - b3) Cada tija defineix una classe i s'escriu una sola vegada. El número de fulles és la freqüència absoluta de la classe. Malgrat això i quan el número de dades és prou gran, podem definir varies classes per una mateixa tija, simplement agrupant les fulles d'alguna forma, per exemple així:
 $\{0,1\}, \{2,3\}, \{4,5\}, \{6,7\}, \{8,9\}$.
 - c) El gràfic es completa afegint les freqüències acumulades des dels extrems i cap a la mediana, que s'assenyala amb un parèntesi en la freqüència absoluta de la classe que la conté.
- Encara que el concepte de valor atípic el definirem més endavant, volem assenyalar que els valors atípics no s'inclouen en el gràfic.

Exemple:

Considerem el següent conjunt de dades tipus III,

53,0	70,2	84,3	55,3	78,5	63,5	71,4	53,4	82,5	67,3	69,5	73,0
55,7	85,8	95,4	51,1	74,4	54,1	77,8	52,4	69,1	53,5	64,3	82,7
55,7	70,5	87,5	50,7	72,3	59,5						

El gràfic de tija i fulles corresponent és aquest

n=30		
depths		(unit=1)
11	5	01233345559
(5)	6	34799
14	7	00123478
6	8	22457
1	9	5

Els principals avantatges d'aquest tipus de gràfic són:

- La informació original pràcticament es conserva i es pot reproduir.
- Les dades queden ordenades.
- Podem observar simetries, amplades, concentracions i errors de les dades.
- És un mètode robust.

3.6 Percentils

El concepte de percentil pot definir-se en dos sentits diferents.

Primer podem considerar que es vol calcular el percentil d'un valor x , és a dir, el percentatge d'observacions que tenen un valor inferior a x .

D'altra banda, podem buscar el percentil 30, per exemple, si es vol trobar el valor x (observat o no) tal que el 30% de les observacions són inferiors a ell.

En el primer sentit, el percentil d'una observació x es calcula així:

$$\frac{(\text{número d'observacions inferiors o iguals a } x) \cdot 100}{n}$$

En el segon, el percentil p es calcula així:

a) Primer calcularem $p \frac{n}{100}$

b) El resultat anterior ens indica en quina observació o entre quines observacions (farem la mitjana entre les dues) es troba el percentil p , si considerem les observacions ordenades de menor a major.

La idea de la mediana com valor que parteix les dades per la meitat fa que coincideixi amb el percentil 50. De forma semblant, si definim els quartils com els valors que parteixen la mostra en quatre parts, obtenim els percentils 25, 50 i 75, i el segon quartil és la mediana. Així podem també definir els decils $D_1, D_2, \dots, D_9$ com els percentils 10, 20, ..., 90.

Per tot això, i si volem que les definicions de mediana i quartils que hem fet en apartats anteriors coincideixin amb els percentils corresponents, la regla per calcular percentils que hem enunciat no és prou acurada. Hem de calcular el percentil p a partir de la mostra ordenada. La posició ideal del percentil p es calcula més acuradament amb les següents formes alternatives:

i) $n \frac{p}{100} + \frac{1}{2}$

ii) $(n+1) \frac{p}{100}$

iii) $\left(n + \frac{1}{3}\right) \frac{p}{100} + \frac{1}{3}$

Així, aquests valors ens indiquen la posició en la mostra ordenada o entre quines posicions (farem la mitjana dels dos) tenim el percentil p .

3.7 Estadístics d'ordre

La mediana forma part d'un conjunt d'estadístics, genèricament s'anomenen *Estadístics d'ordre*, que extrauen la informació de la posició de la dada en la mostra ordenada i no del valor concret de la dada. D'aquesta forma s'aconsegueixen estadístics més robustos, és a dir, menys sensibles a valors mostrals extrems. Anem a explicar la teoria i primeres aplicacions d'aquests estadístics d'ordre.

Donada una sèrie de dades, resultat d'un mostratge aleatori simple,

$$x_1 \quad x_2 \quad \dots \quad x_n$$

es defineix la mostra ordenada com:

$$x_{(1)} \quad x_{(2)} \quad \dots \quad x_{(n)}$$

on cada un dels valors rep el nom d'estadístic d'ordre.

La llista de valors:

$$1 \quad 2 \quad \dots \quad n$$

rep el nom de *rang creixent*, i la llista en sentit oposat:

$$n \quad n-1 \quad \dots \quad 1$$

de *rang decreixent*.

Es verifica que:

$$\text{rang creixent} + \text{rang decreixent} = n + 1$$

També es defineix la *profunditat* d'un estadístic d'ordre com:

$$\begin{aligned} \text{profunditat} (x_{(i)}) &= \min\{ \text{rang creixent} (x_{(i)}), \text{rang decreixent} (x_{(i)}) \} \\ &= \min\{ i, n-i+1 \} \end{aligned}$$

Per exemple:

$$\begin{aligned} \text{prof. de la mediana} &= \frac{n+1}{2} \quad \left\{ \begin{array}{l} \text{si } n = 2k-1, \text{ mediana} = x_{(k)} \\ \text{si } n = 2k, \quad \text{mediana} = \frac{x_{(k)} + x_{(k+1)}}{2} \end{array} \right. \\ \text{prof. dels extrems} &= 1 \quad \text{Mín.} = x_{(1)} \quad \text{Màx.} = x_{(n)} \end{aligned}$$

$$\text{prof. dels quartils} = \frac{\lceil \text{prof. de la mediana} \rceil + 1}{2}$$

on $[x]$ és el enter més gran que no excedeix x i els quartils són el quartil inferior Q_1 i el quartil superior Q_3 . La mediana pot ser Me o Q_2 .

També podem definir d'altres estadístics d'ordre com els *octaus* en la forma

$$\text{prof. dels octaus} = \frac{[\text{prof. dels quartils}] + 1}{2}$$

Sobre els estadístics d'ordre també podem definir i calcular mesures de tendència central, de dispersió i de forma.

3.7.1 Estadístics de centralitat

$$\begin{aligned} \text{mediana} &= Me \\ \text{trimediana} &= \frac{1}{4} Q_1 + \frac{1}{2} Me + \frac{1}{4} Q_3 \end{aligned}$$

3.7.2 Estadístic de dispersió (rang interquartilic)

$$\begin{aligned} \text{dispersió} &= (\text{quartil superior}) - (\text{quartil inferior}) \\ d_Q &= Q_3 - Q_1 \end{aligned}$$

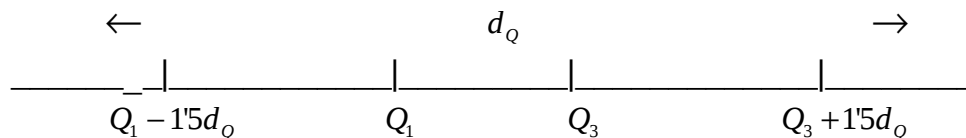
3.7.3 Estadístic d'asimetria

Coefficient de biaix interquartilic

$$BQ = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} \quad \text{on } Q_2 = Me$$

3.7.4 Valors atípics (outliers)

Els estadístics d'ordre abans definits permeten fixar les barreres fora de les que considerarem els valors com valors atípics (en anglès *outliers*) o “fora del normal”:



Tot valor per sota de $Q_1 - 1.5d_Q$ o per sobre de $Q_3 + 1.5d_Q$ és un valor atípic.

3.7.5 Distribució normal

En el cas d'una variable amb distribució de probabilitat normal, de mitjana m i desviació estàndard s , els valors dels principals estadístics d'ordre són:

$$Q_1 - 1'5d_Q = m - 2,698s$$

$$Q_1 = m - 0,6745s$$

$$\text{Mediana} = Me = Q_2 = m$$

$$Q_3 = m + 0,6745s$$

$$Q_3 + 1'5d_Q = m + 2,698s$$

Per entendre millor el que són els valors atípics, podem calcular la probabilitat de ser valor atípic en el cas d'una variable aleatòria amb distribució normal:

$$P[\text{valor atípic}] = 0,00698$$

És a dir, en una població normal 7 de cada 1000 individus són atípics, segons la definició que hem donat.

D'altra banda, la comparació del rang interquartílic amb el valor exacte de la llei normal ens proporciona una altra mesura de dispersió anomenada *pseudovariància*:

$$\text{pseudosigma} = \frac{\text{rang interquartílic}}{1,349}$$

$$\text{pseudovariància} = (\text{pseudosigma})^2$$

3.8 Gràfics basats en estadístics d'ordre

3.8.1 Sumari dels 5 números

# n		
Prof. de la mediana	mediana	
Prof. dels quartils	quartil inf.	quartil sup.
1	extrem inf.	extrem sup.

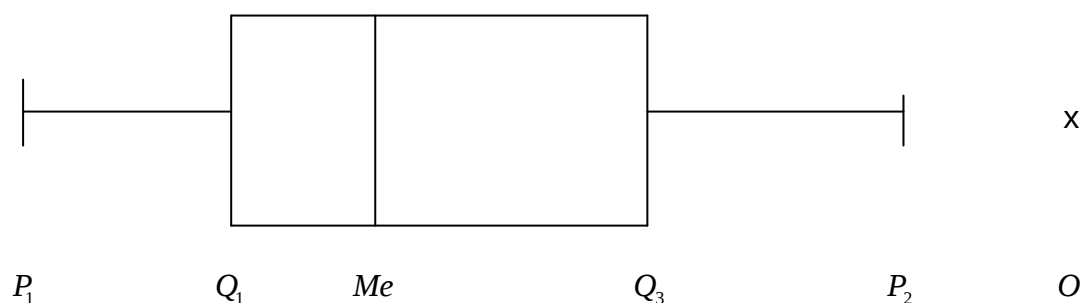
No es tracta exactament d'un gràfic, però la disposició d'aquests cinc estadístics ens permet una primera aproximació a la centralitat, dispersió i rang dels valors de la mostra.

Exemple:

Amb les dades de la pàgina 48, el sumari dels cinc números és el següent

#	30		
15,5		69,3	
8	55,3		77,8
1	50,7		95,4

3.8.2 Diagrama de caixa (Boxplot)



En aquest diagrama representem una caixa d'alçada arbitrària i de llargada fixada per els quartils.

Encara que $Q_1 - 1,5d_Q$ i $Q_3 + 1,5d_Q$ són els punts de tall per considerar un valor com atípic (outlier), en el diagrama de caixa els punts P dibuixats són els valors observats més llunyans que no són atípics.

Els valors atípics es marquen amb un senyal, en aquest cas una creu.

El diagrama de caixa mostra les següents característiques de les dades:

- localització → mediana
- dispersió → rang interquartilic
- asimetria → posicions relatives de la mediana i els quartils.
- cues → línies fins a P_1 i P_2
- valors atípics (outliers)

A més, aquests gràfics són robustos, propietat que hereten dels estadístics d'ordre. Això significa que són relativament resistents en front de petites modificacions de les dades mostrals.

Exemple:

Per a les dades de la pàgina 48, els valors dels estadístics d'ordre que intervenen en el diagrama de caixa són

$$Me = 69,3 \quad Q_1 = 55,3 \quad Q_2 = 77,8$$

$$P_1 = 50,7 \text{ ja que } Q_1 - 1,5d_Q = 55,3 - 1,5 \cdot 22,5 = 21,55$$

$$P_2 = 95,4 \text{ ja que } Q_2 + 1,5d_Q = 77,8 + 1,5 \cdot 22,5 = 111,55$$

En aquest cas no hi ha valors atípics.

3.8.3 Diagrama de tija i fulles

El diagrama de tija i fulles que hem explicat en un apartat anterior, conté alguns elements dels estadístics d'ordre. En concret, la profunditat del major valor de cada classe, excepte la classe que conté la mediana que s'assenyala amb parèntesis sobre la freqüència d'aquesta classe.

Els valors atípics s'exclouen d'aquesta representació.

Exemple:

n=30		
depths		(unit=1)
11	5	01233345559
(5)	6	34799
14	7	00123478
6	8	22457
1	9	5

3.9 EXERCICIS

- En set moments diferents del dia s'observa el nombre de clients que havia en un supermercat. Les dades són: 2, 5, 7, 3, 4, 9. Calculeu:
 - Les mitjanes aritmètica, geomètrica, quadràtica i harmònica.
 - La mediana i la moda.
- Calculeu el recorregut, la mitjana aritmètica, la moda i la mediana de les dades dels següents exercicis:

a) Exercici 1 del tema 2	b) Exercici 2 del tema 2
c) Exercici 3 del tema 2	d) Exercici 5 del tema 2
e) Exercici 6 del tema 2	f) Exercici 7 del tema 2
- Calculeu el recorregut, la mitjana aritmètica, la moda i la mediana de les dades del exercici 9 del tema 2 i amb la taula de 9 intervals presentada. Agrupeu posteriorment les dades en només 5 intervals: $1 \cdot 10^6$, $10^6 - 2 \cdot 10^6$, $2 \cdot 10^6 - 3 \cdot 10^6$, $3 \cdot 10^6 - 4 \cdot 10^6$, $4 \cdot 10^6 - 5 \cdot 10^6$, i repetiu el càlcul de la mitjana. Quina de les dues mitjanes és més acurada?
- Per tal de conèixer la potencialitat de compra d'una determinada àrea metropolitana, s'ha mostregjat la renda per càpita i s'ha obtingut una taula que resumeix les dades. Malauradament, un error informàtic ha destruït les dades originals i de la taula només es conserva una fotocòpia en molt mal estat. És possible reconstruir l'estudi amb ella?

La taula és la següent:

$(L_{i-1}, L_i]$	n_i	N_i	f_i	F_i	$\tilde{x}_i$	a_i	h_i
20 - 50	2	2			35.000		
? - 60		8		0,04			
60 - ?				0,12		10.000	
			0,13		75.000		
	44						0,0022
100 - ?		119					0,001
? - 150			0,11				
150 - ?		171					
? - 250		188				50.000	
? - 850	12	200			550.000		

on $(L_{i-1}, L_i]$ són els extrems inferior i superior de l'interval de la classe i (en milers de pessetes)

n_i les freqüències absolutes
 N_i les freqüències absolutes acumulades
 f_i les freqüències relatives
 F_i les freqüències relatives acumulades
 $\tilde{x}_i$ les marques de classe
 a_i l'amplada de l'interval
 h_i l'alçada de la classe i

Completa la taula, calculeu els valors de la mitjana, la moda, la mediana, el primer quartil i els percentils 40 i 99.

5. En 1798, el científic anglès Henry Cavendish determinà la densitat de la Terra amb molta precisió. Quan es fan mesures complicades, es aconsellable repetir l'operació varies vegades i treballar amb la mitjana de totes elles. Cavendish repetí la seva mesura 29 vegades. Aquí són els resultats que va obtenir (en aquestes dades la densitat de la Terra s'expressa com un múltiple de la densitat de l'aigua):

5,50 5,61 4,88 5,07 5,26 5,55 5,36 5,29 5,58 5,65
 5,57 5,53 5,62 5,29 5,44 5,34 5,79 5,10 5,27 5,39
 5,42 5,47 5,63 5,34 5,46 5,30 5,75 5,68 5,85

Representeu gràficament les dades de la manera que considereu més convenient. La forma de la distribució, permet utilitzar $\bar{x}$ i $\hat{s}$ per descriure-la? Trobeu $\bar{x}$ i $\hat{s}$. Amb tot el que acabeu de fer, quina és la vostra estimació de la densitat de la Terra a partir d'aquestes mesures?

6. Les tres factories d'una indústria han produït el darrer any el següent nombre de motocicletes, detallat per trimestres:

	Factoria 1	Factoria 2	Factoria 3
Trimestre 1	600	650	550
Trimestre 2	750	1200	900
Trimestre 3	850	1250	1050
Trimestre 4	400	800	650

- a) Representeu gràficament la taula anterior de la forma més adient.
 b) Obteniu la producció mitjana trimestral de cada factoria i de tota la indústria.
 c) Calculeu la producció mitjana diària de cada factoria i de tota la indústria, si tenim en compte que va haver respectivament 68, 78, 54 i 74 dies laborables per trimestre.
7. Les següents taules recullen el nombre de peces defectuoses trobades per dos equips en els controls de qualitat efectuats sobre mostres de mida estàndard, durant una setmana:

	Peces defectuoses	0	1	2	3	4	5
Controls efectuats	Equip A	2	7	15	4	2	1
	Equip B	2	5	8	8	5	2

Calculeu la moda i la mediana de les sèries per separat i conjuntament. Representeu les dades de la forma més adient.

8. La taula següent detalla la distribució corresponent al capital invertit per 420 empreses de la construcció:

Capital (milions de PTA)	nombre d'empreses
menys de 5	12
de 5 a 13	66
de 13 a 20	212
de 20 a 30	84
de 30 a 50	30
de 50 a 100	14
més de 100	2

Utilitzeu com a marca de classe del primer i darrer interval 4 i 165 respectivament i calculeu la mitjana aritmètica, la moda i la mediana.
Representeu també l'histograma de freqüències.

4 Descripció de dades bivariants

4.1 Dades bivariants

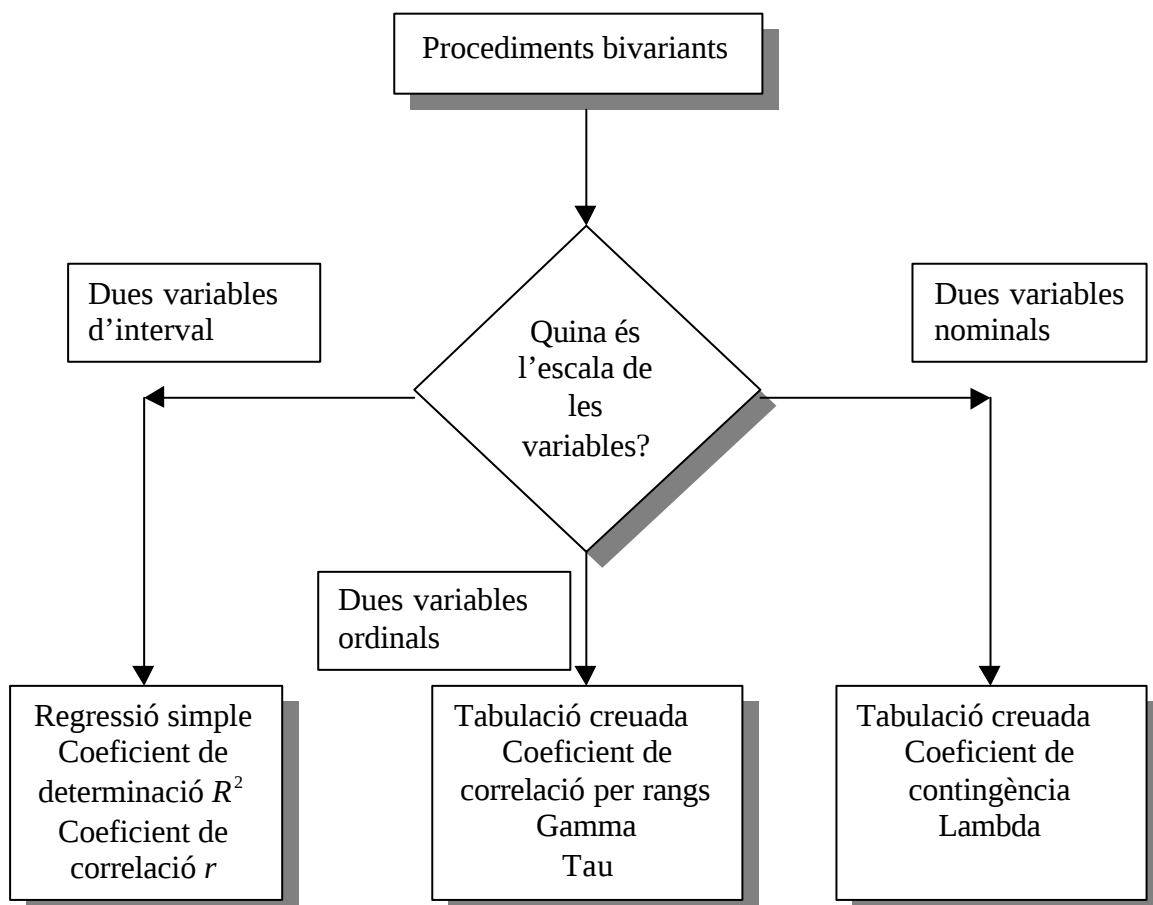
Tot el que s'ha vist fins ara fa referència a l'anàlisi estadística d'una sola variable. Sovint, però, se n'observa més d'una alhora (edat, sexe, marques, hàbits, actituds,...) i interessa, no només descriure-les per separat, sinó poder respondre qüestions del tipus:

- Existeix alguna relació entre les variables o les característiques?
- En cas d'existir, què tan forta és?
- Es pot predir una variable a partir de l'altra?

L'anàlisi de dades bidimensionals o bivariants es duu a terme mitjançant un esquema semblant al de les dades univariants, és a dir, es consideren la *tabulació*, la *representació gràfica* i les *característiques numèriques*.

En aquest capítol presentarem els procediments d'anàlisi de dades que són apropiats per a les relacions bivariants. La tècnica apropiada depèn del nivell de l'escala de les variables involucrades i del desig de l'investigador d'obtenir un estadístic descriptiu. També tractarem alguns temes relacionats amb la interpretació de les taules de tabulació creuada.

Com hem dit, els procediments d'anàlisi de dades bivariants depenen essencialment del tipus d'escala de les variables. La següent figura resumeix la descriptiva de l'anàlisi bivariant:



Quan es consideren dues variables d'escala per interval, encara que la presentació de les dades sigui amb una taula, no és una taula de freqüències pròpiament dites, sinó que consisteix simplement en l'enumeració de les parelles de punts; cada fila representa un individu i cada columna una variable mesurada sobre ell. Per exemple, si s'estudia l'alçada i el pes dels nadons en un hospital, les dades s'organitzen en una taula d'aquesta mena:

Nen	Alçada (cm)	Pes (g)
Joan Pérez	47	3100
Anna Rius	52	3800
Pere Ruiz	50	3300
...	...	...

Aquest tipus de dades normalment no s'agrupen en classes i directament es poden representar en forma de diagrames de dispersió o *scatterplots*. Quan les dades són plenament aleatòries, fruit d'una mostra aleatòria simple, podem estudiar el grau de dependència o independència lineal amb el *coeficient de correlació*. Quan existeix una relació causal i el principal objectiu és la predicció de valors d'una variable en funció de l'altra, una tècnica per estudiar la relació entre les variables és la regressió, en particular la regressió simple, que mereix tot el capítol següent.

Així doncs, com això ho estudiarem més endavant, ara tractarem la tabulació de dades.

4.2 Taules creuades

Quan es considera la variació conjunta de dues variables amb escala nominal u ordinal, se sol agrupar les observacions en una *taula de freqüències bivariants* on es presenta la distribució de freqüències corresponent a les parelles de valors o categories de les dues variables. Els valors (per variables discretes), les categories (en el cas de variables qualitatives) o els intervals de variació (en variables quantitatives on es fa agrupació) d'una de les variables es disposen en la primera fila de la taula i els de l'altra en la primera columna. Això proporciona les anomenades *taules creuades* com la que presentem a continuació:

	Classe social				
Utilitza A	Alta	Mitjana-Alta	Mitja.-Baixa	Baixa	Total
Sí	164	203	114	6	487
No	163	406	752	192	151
Total	327	609	866	198	2000

En aquesta taula es relacionen dues variables qualitatives, classe social, i possessió d'un determinat bé com un automòbil o una tarja de crèdit.

Destaquem que una variable quantitativa amb escala per interval, si agrupem els valors en classes o intervals, té el mateix tractament que una variable amb escala ordinal.

4.2.1 Construcció d'una taula a partir de les dades originals

Un aspecte que no té res a veure amb l'anàlisi, però que és de gran importància pràctica, és la construcció de la taula de doble entrada a partir de la base de dades original, ja que si s'ha de fer "a ma" pot ser una feina molt feixuga.

La majoria dels programes estadístics disposen d'una opció anomenada

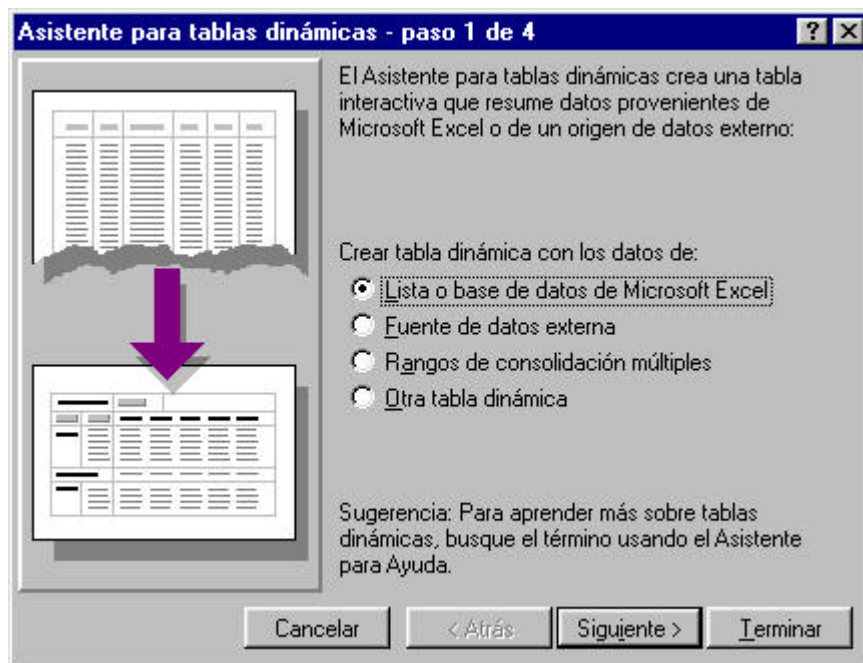
CROSSTABULATION

o TABULACIÓ CREUADA que permet construir una taula de freqüències, com les que ja hem vist, a partir d'unes dades originals encolumnades amb l'aspecte següent :

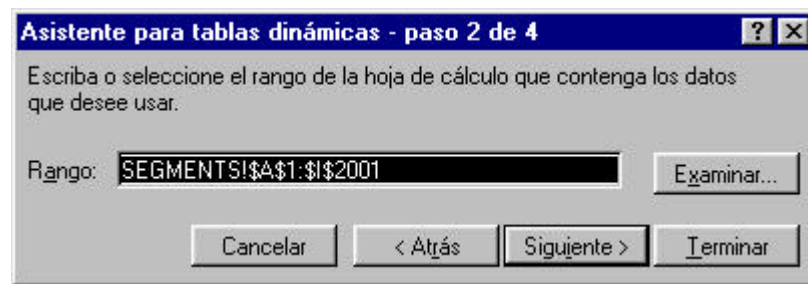
REGIO	EDAT	SEXE	CLASSOC	ESTUDIS	A	B	C	COMPT.
1	4	1	3	4	2	1	1	1
2	1	1	3	4	2	2	2	1
1	1	1	2	3	2	2	2	1

Si es vol fer amb EXCEL, es pot fer de manera molt senzilla amb l'opció "Tablas dinámicas", que permet creuar dues o més variables i presentar una taula resum. Ara bé, si es vol fer amb EXCEL convé afegir una nova columna "fictícia" que contingui un 1 a cada cel·la, ja que quan es fa la tabulació dinàmica cal utilitzar una variable per comptar el número de valors de cada combinació que apareix. En concret, per construir la taula de l'exemple de la relació entre "Classe social" i "Ús del servei A" crearem una variable fictícia plena de 1 que anomenarem "Comptador" i definirem la taula dinàmica de la forma següent:

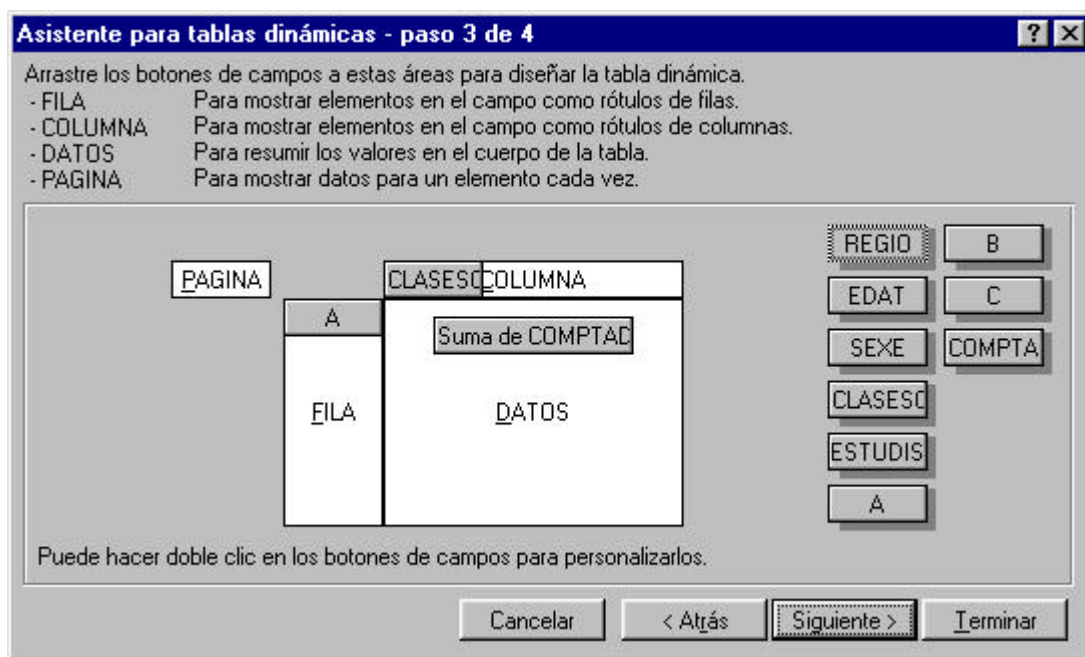
1. Indicarem que volem fer la taula a partir de les dades del full



2. Per defecte el programa selecciona tot el full i agafa com nom de les variables les de la primera fila



3. Definirem quines variables van a les files o columnes de la taula i posarem la variable “Comptador” en la part central (“Datos”).



4. Al prémer “Terminar” el programa generará la taula que busquem :

Suma de COMPTADOR	CLASE SOCIAL				
A	1	2	3	4	TOTAL
1	164	203	114	6	487
2	163	406	752	192	1513
TOTAL	327	609	866	198	2000

4.2.2 Recodificació de variables

Si una o ambdues variables no estan codificades, cal fer-ho abans de fer la tabulació creuada.

Amb el paquet estadístic STATGRAPHICS es fa servir l'opció 6. *Recode Data* del menú *P. Categorical data analysis*. Al entrar en aquesta opció el programa demana la variable que es vol recodificar, el criteri per als extrems dels intervals i opcionalment el nom d'un vector que contingui els límits dels intervals en els quals es vol basar la recodificació. Es pot deixar aquest darrer camp en blanc i fer aquesta especificació interactivament.

Recode Data
=====
Data variable: Salari
Lower limit condition: Greater
Upper limit condition: Equal or less
Unspecified values: Same
Limit specification variable:

1. Especificacions per recodificar les variables.

Si es vol especificar interactivament els intervals, només cal deixar el darrer camp en blanc i en prémer *F6* es passa a una altra pantalla on s'indicarà els intervals i els nous valors que se'ls vol assignar. L'assignació es pot fer de forma arbitrària o indicant només els dos primers intervals i escollint, amb *F5*, l'opció *Repeat Pattern*. En aquest darrer cas es farà la codificació de tots els valors seguint l'esquema definit per als dos primers intervals. Un cop definits els intervals cal prémer *F6* i, si no hi ha problemes, es transformaran els valors en els indicats per a cada interval i es demanarà el nom de la variable on es volen gravar. Resulta evident que les variables que es poden recodificar per aquest programa són les quantitatives o qualitatives amb codificació numèrica.

El paquet de programes SPSS amb la opció *Recodificar→En distintas variables* del menú *Transformar* reassigna els valors de variables existents (qualitatives o quantitatives) o agrupa rangs de valors existents en nous valors per a una nova variable. Per exemple, podria agrupar els salaris en una nova variable que contingui categories de rangs de salaris, o també, pot transformar una variable de cinc categories a una altra de només tres.

Així, pot recodificar les variables numèriques i “de cadena”. Pot recodificar variables numèriques en variables “de cadena”, i viceversa. Si seleccionem *múltiples variables*, totes han de ser del mateix tipus. No es poden recodificar juntes les variables numèriques i “de cadena”.

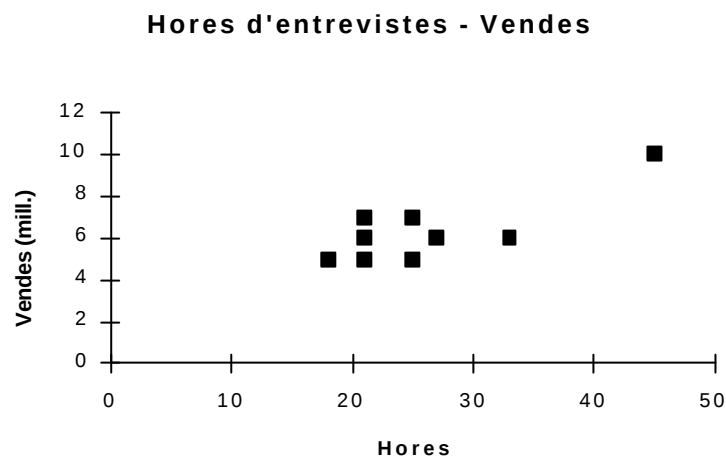
Per especificar com han de recodificar-se els valors, polsarem en *Valores antiguos y nuevos*. Si ho desitgem, podem definir un subconjunt de casos per recodificar-los.

4.3 Representacions gràfiques per a dades bivariants

Com en el cas de les dades univariants la representació gràfica està lligada a la forma en la que estiguin tabulades les dades.

Diagrama de dispersió (Scatterplot)

Consisteix en un gràfic amb dos eixos on cada individu es representa amb un punt que té per coordenades els valors de les dues variables mesurades en els eixos. Aquest gràfic és adient per a les taules amb valors quantitius del tipus on no es disposa de freqüències sinó dels valors individuals.



Histograma tridimensional.

Aquesta tècnica és adient per tractar dues variables quantitatives agrupant-les en intervals. També la podem fer servir amb una variable quantitativa agrupada i les classes definides per una variable qualitativa, nominal u ordinal. Els valors de cada variable es disposen en un dels eixos i s'aixeca una barra d'alçada proporcional a la freqüència sobre el rectangle que determinen.

Relació entre classe social i us de tarja de crèdit

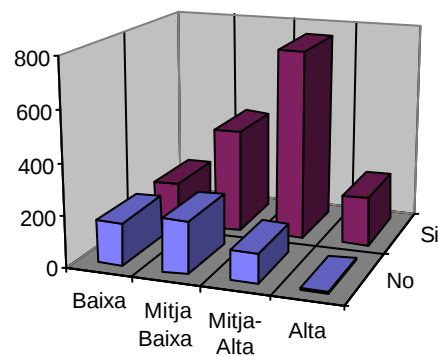


Diagrama de barres múltiple

Si en la llegenda del diagrama de barres s'indiquen els valors d'una variable qualitativa, aquesta tècnica s'utilitza per representar conjuntament els diagrames de barres, un per a cada categoria de la variable considerada.

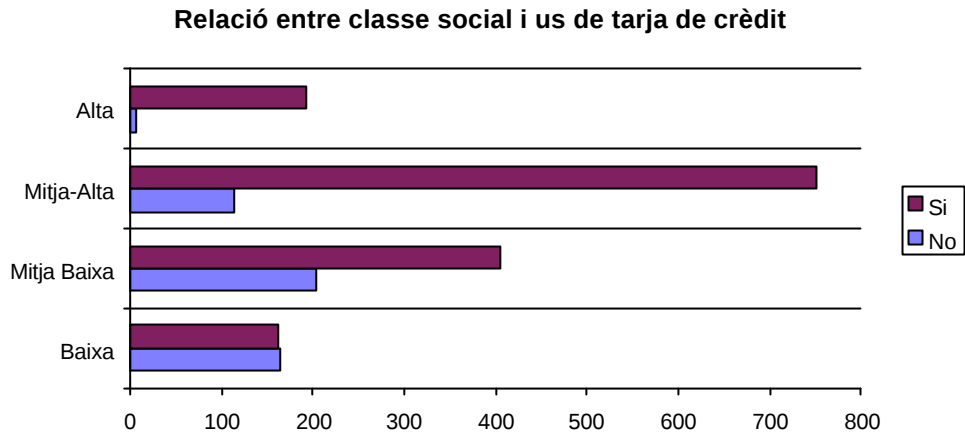
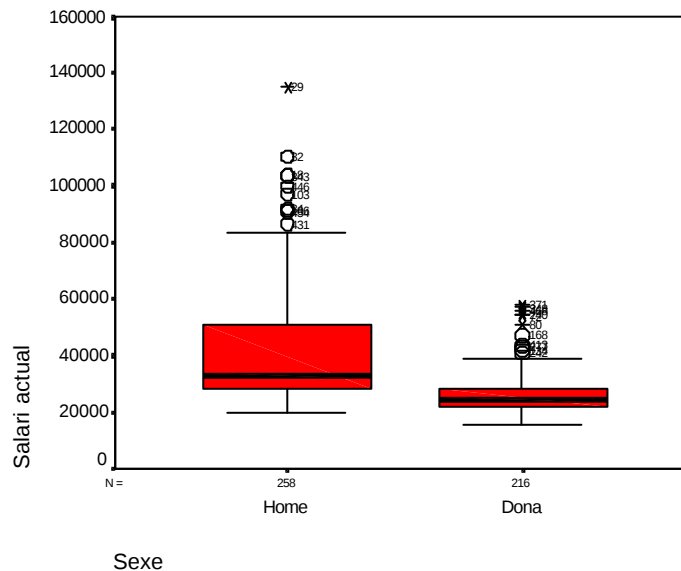


Diagrama de caixes múltiples

Aquesta tècnica és adient per a una variable quantitativa (sense agrupació), per exemple les unitats (paquets) de pa de motlle venuts, i una qualitativa, per exemple quatre expositors diferents. Permet la comparació de la variable quantitativa en les diferents categories definides per la variable qualitativa. És una representació gràfica molt relacionada amb l'Anàlisi de la Variància d'un factor (veure el capítol corresponent).



D'altres gràfics

Hi ha d'altres tipus de gràfics, com els de barres apilades, que són adients en certes situacions i amb cert tipus de dades: com les freqüències relatives. L'oferta de gràfics en un full de càlcul com EXCEL és molt variada. En tot cas, hem de procurar fer servir el gràfic més senzill i que expliqui millor les dades.

4.4 Característiques numèriques per a dades bivariants

Les diverses mesures de síntesi que podem calcular per a dades bivariants també depenen del tipus de variable amb que es tracti i del tipus de taula. Diferenciarem dos grans grups: Les que fan referència a dades bivariants quantitatives no agrupades i les que s'obtenen a partir de taules creuades.

4.4.1 Mesures per a dades quantitatives no agrupades

Quan es representen les dades bivariants en un diagrama de dispersió o *Scatterplot*, es pot, o no, observar en els punts una certa tendència a agrupar-se segons una forma determinada, sovint una el·lipse més o menys aixafada. Dues de les mesures més populars per a dades bivariants serveixen per quantificar el grau de *dependència lineal* entre les dues variables: la covariància mostral i el coeficient de correlació lineal.

Si disposem de n observacions $(x_i, y_i) i = 1, \dots, n$ de dues variables X, Y , es defineix la **covariància mostral** entre les dues variables com la mitjana dels productes de les desviacions de les dades respecte de la seva mitjana:

$$s_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Si els punts són repetits, farem servir les freqüències a la fórmula, a l'estil de les dades de tipus II. D'altra banda, també podem corregir la covariància mostral amb els graus de llibertat de la mostra i dividir per $n - 1$ en l'anomenada covariància mostral corregida.

La covariància és una mesura de variació conjunta de les dues variables. Si és positiva s'interpreta com que a més grans els valors de la primera variable, més grans els de la segona. Si és negativa s'interpreta a l'inrevés, és a dir, per valors més grans de la primera esperem obtenir valors més grans de la segona. Un inconvenient d'aquesta mesura és que, al expressar-se com producte de les unitats de mesura de les variables de partida es fa difícil dir, donat un valor, què tan gran és la variació conjunta que representa.

Una mesura adimensional de variació conjunta la dona el **coeficient de correlació lineal** definit com:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Aquest coeficient varia entre -1 i +1. Valors pròxims a +1 indiquen una correlació lineal positiva forta i valors pròxims a -1 una correlació lineal negativa forta. Valors propers a 0 s'entenen com que no hi ha correlació lineal.

En general, si r és major que 0.8 (sense considerar el signe), la relació entre les variables és forta, si r està en 0.8 i 0.4, la relació és entre moderada i feble; i si r és menor que 0.4, la relació és molt feble o inexistent.

Per a les covariancies i les correlacions sol utilitzar-se una notació matricial que simplifica la manipulació, sobretot quan es considera un conjunt de parelles de variables. Es defineix la matriu simètrica de variancies i covariancies mostrals com:

$$\mathbf{S} = \begin{pmatrix} s_X^2 & s_{XY} \\ s_{XY} & s_Y^2 \end{pmatrix}$$

i la matriu de correlacions mostrals com:

$$\mathbf{R} = \begin{pmatrix} 1 & r_{XY} \\ r_{XY} & 1 \end{pmatrix}$$

En la pràctica, els paquets estadístics acostumen a proporcionar el triangle superior o inferior d'aquestes matrius.

Tot i que el STATGRAPHICS no té un apartat explícit per calcular mesures de variació conjunta per a dades bivariants, es poden calcular aquestes dues mesures a través de les opcions del menú *Q. Multivariate methods*. L'opció 1. *Correlation Analysis* subministra la matriu de correlacions de dues o més variables. L'opció 2. *Covariance Analysis* dona la matriu de covariancies. En tots dos casos el format de sortida és similar, obtenint-se en cada casella tres valors (dos en el cas de la matriu de covariancies) que són: el coeficient calculat, la mida de la mostra i un valor per determinar si aquest coeficient és o no significativament diferent de zero (prova d'hipòtesi).

En el SPSS, l'opció *Correlaciones*→*Bivariadas...* del menú *Analizar* proporciona els coeficients de correlació lineal o de Pearson de dues o més variables.

4.4.2 Mesures d'associació per a taules creuades

Les taules creuades serveixen per estudiar la relació entre dues variables qualitatives o variables que han estat prèviament categoritzades. Per "estudiar la relació" s'acostuma a entendre estudiar la **independència** i l'**associació** entre les variables. Per a l'estudi de la independència calen les anomenades *proves khi-quadrat*, amb aquests tests es mostra com acceptar o rebutjar que dues variables són independents. Ara bé, quan no ho són, què es pot dir del tipus de relació que hi ha entre elles o de la intensitat d'aquesta relació? Les **mesures d'associació** són estadístics que pretenen aportar informació sobre aquests dos aspectes. Aquestes mesures són de difícil interpretació, ja que no totes volen dir el mateix en tots els casos. A continuació es comenta breument cada grup:

Mesures d'associació per a variables mesurades en escala **nominal**.

Mesures basades en l'estadístic khi-quadrat

Mesures basades en la reducció proporcional de l'error

Mesures d'associació per variables mesurades en escala **ordinal**.

Mesures basades en parelles concordants i discordants

En el següent quadre podem veure les mesures d'associació que genera la opció *Display Statistics* de *Crosstabulation* del STATGRAPHICS.

Summary Statistics for Contingency Tables (Page 1)			
Chi-square	D.F.	Significance	
30.7988	6	2.76930E-5	
Statistic	Symmetric	With rows dependent	With columns dependent
Lambda	0.12772	0.15429	0.10363
Uncertainty Coeff.	0.04552	0.05151	0.04078
Somer's D	0.14891	0.14050	0.15839
Eta		0.19642	0.19806
Summary Statistics for Contingency Tables (Page 2)			
Statistic	Value	Significance	
Contingency Coeff.	0.31279		
Cramer's V	0.23286		
Conditional Gamma	0.20880		
Pearson R	0.15805		
Kendall's Tau B	0.14917	0.00368	
Kendall's Tau C	0.15640		

2. Mesures d'associació obtingudes amb la opció *Display Statistics* del STATGRAPHICS

4.4.3 Mesures d'associació basades en l'estadístic khi-quadrat

L'estadístic khi-quadrat utilitzat per contrastar la independència entre les variables, no és una bona mesura d'associació, ja que el seu valor depèn de la mida mostral i del grau d'independència entre les variables. A més, atès que només depèn de les freqüències observades i esperades, taules ben diferents poden tenir els mateixos valors khi-quadrat. Els estadístics basats en khi-quadrat miren d'eliminar la influència de la mida mostral efectuant diverses correccions que pretenen normalitzar-lo de forma que caigui entre 0 i 1. Dos dels més importants que proporciona el STATGRAPHICS són:

El coeficient de contingència
La V de Cramer

Tots dos prenen valors entre 0 i 1, però el primer no assoleix mai el valor màxim de un i, a més a més, el valor màxim depèn del nombre de files i de columnes de la taula.

4.4.4 Mesures d'associació basades en la reducció proporcional de l'error.

Aquests coeficients, a diferència dels anteriors, tenen una interpretació clara. Es basen en què tan bo es pot predir el valor d'una variable dependent, conegut el de la variable independent. Les mesures comparen els errors en dues situacions, una on es fa servir la variable independent per a la predicció de la dependent i l'altre on no s'utilitza. El STATGRAPHICS proporciona com mesura d'aquesta mena el:

Coefficient Lambda

4.4.5 Mesures d'associació per a variables mesurades en escala ordinal

Moltes mesures ordinals d'associació es basen en comparar parelles de casos. Considerem, per exemple, la taula següent on es presenten tres individus sobre una mostra de 100 persones amb les variables de nivell d'estudis i nivell d'ingressos mesurades en forma ordinal.

Individu	Nivell d'ingressos	Nivell d'estudis
1	Baix	Mig
2	Mig	Alt
3	Alt	Mig
...	...	...

Els individus 1 i 2 formen una parella **concordant**, en el sentit que el valor de cada variable és més gran (o més petit) en un cas i en l'altre. Els individus 2 i 3 formen una parella **discordant** en el sentit que quan una de les variables és més gran (o més petita) en un cas que en l'altre variable és més petita (o més gran). Si la majoria de les parelles d'individus són concordants es diu que l'associació és positiva, mentre que si són discordants es parla d'associació negativa. Si ambdós tipus de parelles són igualment probables s'entén que no hi ha associació.

Moltes mesures d'associació es basen en la concordança o discordança de les parelles de valors. Entre les més importants podem assenyalar:

La D de Sommer

La Tau B de Kendall

La Tau C de Kendall.

4.5 Interpretació de taules de tabulació creuada

La utilització de taules de tabulació creuada en la pràctica de moltes disciplines aplicades, com la Investigació de Mercats, és molt important i requereix una profundització en la interpretació dels resultats i les relacions. Alguns aspectes fonamentals relacionats amb la interpretació d'aquestes taules són l'ús de percentatges i l'elaboració de les relacions descobertes amb la introducció de variables addicionals.

4.5.1 Ús de percentatges

En els histogrames i diagrames de barres, també en el càlcul de l'estadístic khi-quadrat, s'acostumen a utilitzar les freqüències absolutes de cada cel·la. Però les freqüències brutes o absolutes són difícils d'interpretar. Generalment, amb el propòsit de simplificar la interpretació de les relacions significatives, és molt útil modificar les dades de la taula i presentar-les en forma de percentatge. El problema es presenta perquè l'investigador té tres formes alternatives per calcular els percentatges: (1) percentatge de la fila, de manera que els percentatges en cada fila sumin 100%; (2) percentatge de la columna, de

manera que els percentatges en cada columna sumin 100%; ó (3) percentatge de la cel·la, de manera que els percentatges de totes les cel·les sumin 100%.

Exemple:

A. Freqüències absolutes			
Ingrés	Tipus de vi consumit		Total
	Barat	Millor	
Menys de 13500 e	75	10	85
13500 euros i més	40	80	120
Total	115	90	205

B. Percentatges de les cel·les			
Ingrés	Tipus de vi consumit		Total
	Barat	Millor	
Menys de 13500 e	36.6%	4.9%	41.5%
13500 euros i més	19.5%	39.0%	58.5%
Total	56.1%	43.9%	100%

C. Percentatges de les files			
Ingrés	Tipus de vi consumit		Total
	Barat	Millor	
Menys de 13500 e	88.2%	11.8%	100%
13500 euros i més	33.3%	66.7%	100%

D. Percentatges de les columnes		
Ingrés	Tipus de vi consumit	
	Barat	Millor
Menys de 13500 e	65.2%	11.1%
13500 euros i més	34.8%	88.9%
Total	100%	100%

Quin d'aquests percentatges és de major utilitat per al investigador? Podem tractar de determinar una regla general examinant la taula anterior. És clar que els percentatges de la cel·la ens diuen molt poc sobre la relació entre l'ingrés i el tipus de vi consumit. Generalment, són útils només en la identificació de la mida dels segments particulars, i no per ajudar-nos a comprendre les relacions. Si mirem els percentatges de la columna, podem treure les següents conclusions: el 65.2% d'aquells que consumeixen vi barat guanyen menys de 13500 euros, mentre que només el 34.8% dels que consumeixen vi barat guanyen 13500 euros o més. Aquest tipus de conclusió sembla indicar que el tipus de vi consumit té un efecte sobre l'ingrés obtingut. Òbviament, això no és el que volem dir en aquest cas. És lògic esperar que el nivell d'ingrés tingui un cert efecte en el tipus de vi consumit i no l'oposat. Al examinar els percentatges de la fila, podem adonar-nos del tipus de relació que cercàvem. En aquest cas podem obtenir les següents conclusions: el 88.2% dels que guanyen menys de 13500 euros consumeixen vi barat, mentre que l'11.8% d'aquestes persones consumeixen un vi millor. Aquesta conclusió implica que l'ingrés afecta el tipus de vi consumit.

La regla és *calcular els percentatges en la direcció del factor causal*. Per tant, si el factor causal és el factor fila, calcularem els percentatges de la fila i viceversa.

4.5.2 Elaboració de les relacions

Ha de tenir-se molta cura quan fem servir les taules de tabulació creuada. És possible que la veritable relació sigui més complexa que la que pot identificar-se adequadament amb la tabulació creuada de dues variables. Aquest cas es presenta especialment si l'usuari considera que existeix una relació causal entre les variables. Es hem de preguntar: es manté aquesta relació quan es consideren d'altres variables? Per exemple, com es relacionen el tipus de vi consumit i l'ingrés amb els diferents grups d'edat o diferents regions del país, o diferents sexes, etc. El que necessitem és profunditzar en la

relació de les taules de dues vies. Això s'aconsegueix amb la tabulació creuada de les mateixes variables, com abans, excepte que ara s'ha d'obtenir una tabulació creuada per a cada una de les categories d'una altra variable o variables. D'aquesta manera podem veure si existeix una relació diferent per als homes i per a les dones, etc.

L'associació inicial entre les dues variables s'anomena *associació total* o *d'ordre zero*. Les taules que mostren l'associació entre les dues variables dins de les categories d'altres variables reben el nom de *taules condicionals* i revelen *associacions condicionals*. Les variables sobre les que la tabulació creuada és condicional reben el nom de *variables de control*. Així, podem calcular la tabulació creuada de l'ingrés i el tipus de vi consumit, tot mantenint el control sobre l'edat. Quan s'elaboren taules condicionals amb base una variable de control, s'anomenen *taules condicionals de primer ordre*. També és possible controlar més d'una variable. En aquest cas, s'elaboren taules condicionals per als elements dins de les diferents combinacions de categories de control. Quan considerem dues variable de control, les taules creuades de les variables a estudi sota control reben el nom de *taules condicionals de segon ordre*, i així successivament. Quantes més variables de control, major serà el número de taules obtingudes i més difícil la interpretació dels resultats. Per exemple, si tinguéssim tres variables de control amb 2,3 i 4 categories, respectivament, necessitem $2 \times 3 \times 4 = 24$ taules condicionals (i una mida mostral adequada). Per tant, cal agregar les variables de control amb prudència. Nosaltres analitzarem només el procés d'elaboració de taules condicionals de primer ordre.

Els possibles resultats de l'elaboració depenen de la conclusió inicial. Suposem que la taula inicial d'ordre zero ens porta a la conclusió d'una relació entre les variables. Llavors, de les taules condicionals es podria: (1) sostenir la conclusió original de que existeix una relació, (2) especificar una relació diferent per categories de control o (3) identificar la relació inicial com espúria. D'altra banda, la taula original ens ha pogut portar a la conclusió de la no existència d'una relació. Llavors, de les taules condicionals es podria (1) sostenir la conclusió inicial de no existència d'una relació o (2) identificar una relació. El següent esquema és un resum de les possibilitats:

Quan es troba una relació d'ordre zero

Sostenir la conclusió inicial d'una relació existent.

Especificació de la relació inicial.

Una bona pregunta que ha de formular-se sobre una relació de tabulació creuada és la següent: sota quines condicions es sosté aquesta relació?

Identificació d'una relació espúria.

Quan no es troba una relació d'ordre zero

Sostenir la conclusió inicial de no relació.

Identificar una relació.

Finalment, cal comentar que un petit número de variables pot proporcionar un gran número de taules. Els investigadors no poden creuar tot amb tot i buscar relacions en els llistats d'ordinador. Cal tenir un model del problema i algunes idees sobre les interrelacions de les variables pertinents. Una bona definició del problema i una especificació de les necessitats d'informació pot ser de gran utilitat en el procés d'anàlisi de dades.

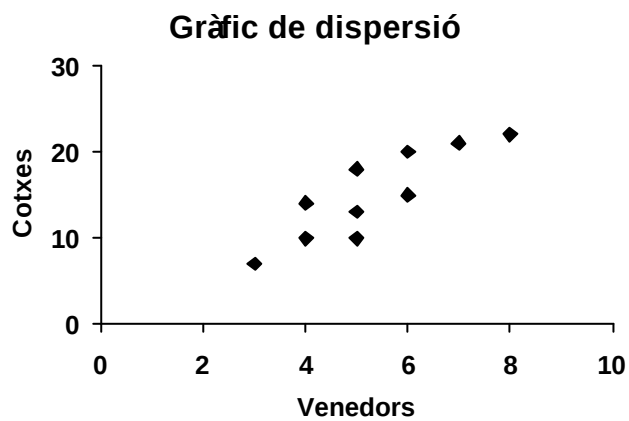
5 Introducció a la regressió

5.1 Introducció

Una empresa de compra-venda d'automòbils organitza exposicions els caps de setmana i contracta un número variable de venedors que oscil·la entre 4 i 8. El gerent d'aquesta empresa vol estudiar la relació entre el número de venedors i el número de cotxes venuts ja que, si és possible, podria preveure les vendes a partir del número de venedors que contracta. Per aclarir-ho el gerent va examinar el registre de vendes dels darrers quatre mesos i va localitzar un període de 10 setmanes durant les quals no hi va haver cap incentiu especial ni a la venda ni a la compra. El nombre de cotxes venuts durant aquest període i el nombre de venedors empleats en cada cas va ser el següent:

Setmana	Venedors	Cotxes
1	5	10
2	6	20
3	5	18
4	4	10
5	3	7
6	4	14
7	7	21
8	6	15
9	5	13
10	8	22

Per examinar aquesta relació és de molta utilitat començar per dibuixar un diagrama de dispersió:



Aquest gràfic fa palesa una relació força evident entre el número de venedors i les vendes, relació que hom podia esperar. Anem a quantificar-la.

5.2 La regressió simple

La regressió és un mètode d'anàlisi que s'aplica a l'estudi de la relació entre una variable quantitativa que anomenarem dependent o resposta i un conjunt de variables anomenades independents o regressores. Els objectius de la regressió són dos:

Mostrar la forma com les variables regressores expliquen la variable resposta.

Fer pronòstics dels valors de la variable resposta, quan coneixem els valors de les variables regressores.

El cas més senzill, per començar a treballar en aquest mètode, és quan tenim una variable resposta i una única variable regressora. Això és la regressió simple.

L'exemple amb el que hem començat aquest capítol correspon a una regressió simple, on el número de venedors són les dades de la variable regressora o independent i el número de cotxes venuts són les dades observades de la variable resposta o dependent.

Dins la regressió simple, podem considerar la forma més senzilla de relació, és a dir, pensem en una relació lineal i ajustem una recta a les dades.

La equació d'una recta és

$$y = b_0 + b_1x$$

on b_0 és el punt d'intercepció o ordenada en l'origen, és a dir, el valor de y quan $x = 0$, i b_1 és el pendent de la recta, el canvi en y per unitat de canvi en x .

Aquesta recta la podem dibuixar 'a ull' en la gràfica de dispersió o la podem trobar a partir de dos punts (dos punts determinen una única recta), però volem ser més precisos i buscarem un procediment matemàtic per definir la millor recta possible.

5.3 El mètode dels mínims quadrats

El procediment que s'utilitza per trobar la recta que millor s'ajusta al núvol de punts és essencialment una formalització de l'ajust visual. Quan ajustem visualment una recta a un conjunt de dades, desplaçem la regla fins que considerem que hem minimitzat les desviacions dels punts respecte de la línia.

Podem definir la desviació en el i -èsim punt com el segment vertical entre el punt i la recta, és a dir

$$y_i - (b_0 + b_1x_i)$$

Només ens queda decidir el criteri de minimització d'aquestes desviacions.

Utilitzarem el que es coneix com *principi de mínims quadrats*, que consisteix en minimitzar la suma dels quadrats de les desviacions. Així escollirem els valors de b_0 i b_1 que minimitzin

$$\sum_{i=1}^n [y_i - (b_0 + b_1x_i)]^2$$

La solució, que s'obté amb els mètodes ordinaris del càlcul diferencial, és la següent:

	$\hat{b}_1 = \frac{S_{xy}}{S_x} = \frac{s_{xy}}{s_x^2} \quad \text{i} \quad \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$	
on		
	$S_x = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2$	$s_x^2 = \frac{1}{n} S_x$
i		
	$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)$	$s_{xy} = \frac{1}{n} S_{xy}$

Amb aquestes estimacions dels paràmetres calcularem les prediccions que proporciona la recta de millor ajust per a cada punt

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$$

i amb elles obtenim l'anomenada *suma de quadrats dels errors*, que és el mínim aconseguït per aquest mètode

$$SQE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Exemple:

A la següent pàgina tenim els càlculs de la recta de regressió, les prediccions, les desviacions i la suma de quadrats dels errors per a les dades de les 10 setmanes.

L'equació de la recta de regressió és

$$y = -0,821 + 2,985x$$

o també

$$y - 15 = 2,985(x - 5,3)$$

El mateix resultat s'obté si fem la regressió amb les dades de la variable x centrades, és a dir, amb les desviacions $x_i - \bar{x}$.

Càlculs de regressió simple

x_i	y_i	x_i^2	$x_i y_i$	$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
5	10	25	50	14,10	-4,10	16,85
6	20	36	120	17,09	2,91	8,47
5	18	25	90	14,10	3,90	15,18
4	10	16	40	11,12	-1,12	1,25
3	7	9	21	8,13	-1,13	1,29
4	14	16	56	11,12	2,88	8,30
7	21	49	147	20,07	0,93	0,86
6	15	36	90	17,09	-2,09	4,37
5	13	25	65	14,10	-1,10	1,22
8	22	64	176	23,06	-1,06	1,12
$\sum x_i = 53$	$\sum y_i = 150$	$\sum x_i^2 = 301$	$\sum x_i y_i = 855$		$\sum (y_i - \hat{y}_i) = 0$	$\sum (y_i - \hat{y}_i)^2 =$
$\bar{x} = 5,3$	$\bar{y} = 15$					58,896

$$\hat{b}_1 = \frac{855 - \frac{1}{10} 53 \cdot 150}{301 - \frac{1}{10} (53)^2} = 2,9850746$$

$$\hat{b}_0 = 15 - \hat{b}_1 \cdot 5,3 = -0,820896$$

5.4 Prediccions

Així doncs, amb les estimacions $\hat{b}_0$ i $\hat{b}_1$ es calculen les *prediccions* $\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$ i amb aquestes, els *residus*

$$e_i = y_i - \hat{y}_i$$

L'anàlisi dels residus permet estudiar la validesa del model.

Suposem que el model de la regressió simple és correcte per a un determinat problema, llavors és possible que ens interressi la estimació del valor mitjà de la variable resposta Y per a un valor donat de $x = x_0$.

La millor predicció d'aquest valor és

$$\hat{y}_0 = \hat{b}_0 + \hat{b}_1 x_0$$

Observació: Cal que el valor x_0 estigui dintre del rang de les dades x per tal de garantir al màxim la validesa del model. A més, estadísticament es pot justificar el fet que les millors prediccions es poden fer a prop de $\bar{x}$.

Exemple:

La predicció per a la resposta mitjana, és a dir, la previsió de la mitjana de cotxes venuts per a un valor de la variable regressora de $x_0 = 5$ venedors és

$$\hat{y}_0 = \hat{b}_0 + \hat{b}_1 x_0 = -0,820896 + 2,9859746 \cdot 5 = 14,10$$

5.5 Mesures de l'ajust

L'avaluació global de l'ajust de la regressió es pot fer amb la *SQE* o, millor, amb la variància mostral dels residus $\frac{1}{n} \sum_{i=1}^n e_i^2$. Però els residus no són tots independents, sinó que estan lligats per dos equacions de forma que utilitzarem l'anomenada *variància residual*:

$$\hat{s}^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2 = SQE / (n-2)$$

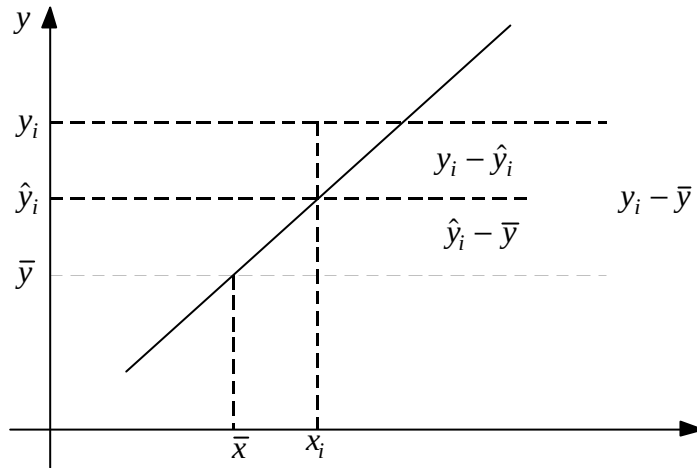
La variància residual és un índex de la precisió del model, però aquesta mesura depèn de les unitats de la variable resposta i no és útil per comparar rectes de regressió de variables diferents.

Una altra mesura de l'ajust requereix una adequada descomposició de la variància de la variable resposta.

5.5.1 Descomposició de la variabilitat

La idea principal consisteix en descomposar les desviacions de les respostes respecte la seva mitjana així:

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$



La suma de quadrats verifica la següent igualtat:

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

ja que $\sum e_i \hat{y}_i - \bar{y} \sum e_i = 0$.

La descomposició de la suma de quadrats de les observacions y_i en dos termes independents s'interpreta així: la variabilitat de la variable Y es descomposa en un primer terme que reflexa la variabilitat no explicada per la regressió, que és deguda a l'atzar, i el segon terme que conté la variabilitat explicada o eliminada per la regressió i pot interpretar-se com la part determinista de la variabilitat de la resposta.

Podem veure que:

$$\text{Variació total} = VT = \sum (y_i - \bar{y})^2 = S_y$$

$$\text{Variació no explicada} = VNE = \sum (y_i - \hat{y}_i)^2 = \sum e_i^2 = SQE$$

$$\text{Variació explicada} = VE = \sum (\hat{y}_i - \bar{y})^2 = \hat{b}_1^2 S_x$$

$$VT = VNE + VE$$

5.5.2 Coeficient de determinació

Una mesura més adequada de l'ajust és la proporció de variabilitat explicada que definim com *coeficient de determinació* així:

$$R^2 = \frac{VE}{VT} = 1 - \frac{SQE}{S_y}$$

Aquesta mesura es pot utilitzar en qualsevol tipus de regressió. Però en el cas particular de la regressió lineal amb una recta tenim

$$R^2 = \frac{\hat{b}_1^2 S_x}{S_y} = \frac{S_{xy}^2}{S_x S_y} = \frac{s_{xy}^2}{s_x^2 s_y^2} = r^2$$

que és el quadrat del *coeficient de correlació lineal* entre dues variables.

El coeficient de determinació R^2 és una mesura de la bondat de l'ajust, $0 \leq R^2 \leq 1$, mentre que el coeficient de correlació és una mesura de la dependència lineal entre les dues variables, quan són aleatòries i només hi ha una variable regressora. Observem que en aquest cas

$$SQE = S_y(1 - r^2)$$

Exemple :

Per a les dades que hem estat treballant, tenim que

$$SQE = 58,896$$

$$VT = S_y = 238$$

$$R^2 = 1 - \frac{58,896}{238} = 0,7525$$

5.6 Càlculs de regressió amb calculadora

Moltes calculadores disposen de la possibilitat de fer càlculs de regressió. Amb una calculadora científica “normaleta” es poden fer la majoria dels càlculs descrits en aquest capítol amb gran senzillesa. Les explicacions que donarem a continuació es basen en una *Casio fx-570s* però són aplicables, amb petites variacions a la majoria de calculadores.

1. Activació del modus LR (linear regression)

La primera cosa que cal fer és posar la calculadora en la modalitat de treballar amb dues variables i preparar els càlculs de regressió.

SHIFT MODE 3

A continuació és convenient esborrar totes les dades estadístiques anteriors que hi pugui haver

SHIFT AC

2. Entrada de les dades

L'entrada de dades s'acostuma a fer d'alguna forma especial, ja que s'han d'introduir els valors per parelles. Suposem que volem entrar les dades de l'exemple d'aquest capítol:

Les dades són (5, 10),(6, 20),(5, 18),(4, 10),(3, 7),(4,14),(7,21),(6,15),(5,13),(8,22).

Entrarem:

5 x_D, y_D 10 DATA 6 x_D, y_D 20 DATA 5 x_D, y_D 18 DATA 4 x_D, y_D 10 DATA 3 x_D, y_D 7
DATA
4 x_D, y_D 14 DATA 7 x_D, y_D 21 DATA 6 x_D, y_D 15 DATA 5 x_D, y_D 13 DATA 8 x_D, y_D 22
DATA

A partir d'aquesta entrada podem recuperar els coeficients de la recta de regressió, les característiques de les variables x i y i el coeficient de correlació r . També podem utilitzar aquests resultats per fer previsions i calcular els residus, si ho desitgem.

3. Resultats

Recta de regressió

$\overline{\text{SHIFT}}$ A	→	Terme independent:	-0,82089552
$\overline{\text{SHIFT}}$ B	→	Pendent:	2,985074627

Ajust lineal

$\overline{\text{SHIFT}}$ r	→	Coefficient de correlació lineal:	0,867490531
-----------------------------	---	-----------------------------------	-------------

Característiques de les variables

$\overline{\text{SHIFT}}$ $\bar{x}$	→	Mitjana mostral de x :	5,3
$\overline{\text{SHIFT}}$ $x s_n$	→	Desv. estàndard mostral de x :	s_x 1,417744688
$\overline{\text{SHIFT}}$ $\bar{y}$	→	Mitjana mostral de y :	15
$\overline{\text{SHIFT}}$ $y s_n$	→	Desv. estàndard mostral de y :	s_y 4,878524367

Amb aquests resultats podem obtenir fàcilment moltes de les característiques que hem estudiat de la regressió:

El coeficient de determinació R^2 és pot calcular com el quadrat del coeficient de correlació:

$$R^2 = (r)^2 = (0,867490531)^2 = 0,752539821$$

La variància residual és

$$\hat{s}^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2 = \text{SQE} / (n-2)$$

i la suma de quadrats dels residus $\text{SQE} = S_y \cdot (1 - r^2)$.

De manera que podem calcular

$$\sum (y_i - \bar{y})^2 = S_y = n \cdot s_y^2 = n \cdot \{y s_n\}^2 = 10 \times 4,878524^2 = 238$$

$$\text{SQE} = n \cdot \{y s_n\}^2 \cdot (1 - r^2) = 10 \times 4,878524^2 \times (1 - 0,752539821) = 58,895522$$

i per tant, l'anomenat error estàndard de la regressió és:

$$\sqrt{\hat{s}^2} = \sqrt{55,8955 / (10 - 2)} = \sqrt{7,36194} = 2,7132896$$

4. Predicció

Si l'ajust és raonable i el model sembla prou vàlid, podem fer-lo servir per fer prediccions.

Moltes calculadores tenen la tecla $\boxed{\hat{y}}$ amb la que podem fer les prediccions per a qualsevol valor x_0 . Així si volem predir el número de cotxes venuts una setmana que es contractin 5 venedors farem:

$$\hat{y}(5) = -0.820896 + 2.98507 \times 10 = 14,10448$$

Fent-ho amb la calculadora només cal fer 5 $\boxed{\hat{y}}$ i obtenim 14,10447761.

5.7 EXERCICIS

1. L'empresa VentCorr, dedicada a la venda per correspondència, vol esbrinar si hi ha relació entre el pes del correu que reben a diari i el nombre de comandes efectuades. Per a això, durant 10 dies mesura el pes del correu i les comandes, obtenint els resultats de la taula següent:

Dia	Pes correu	Comandes
1	70,6	254
2	48,4	199
3	19,8	64
4	39,8	162
5	50,1	195
6	68,5	238
7	55,8	219
8	73,3	256
9	96,3	348
10	93,5	377

- a) Representeu les dades en un diagrama de dispersió per tal de determinar si hi ha o no relació entre les dues variables.
- b) Ajusteu un model de regressió lineal simple. Discutiu l'adequació del model a les dades.
- c) Quines seran les comandes previstes per a un dia en que el pes del correu rebut sigui de 60?
2. Una companyia vol establir un índex ben correlacionat amb les vendes per fer-lo servir en les seves previsions. Per a això disposa de les dades següents, corresponents al període 1987-1998,

Any	Vendes Unitàries	Despeses en aixecament d'habitatges (milions d'unitats)	Despeses en adquisició de bens imperibles (\$*10 ⁹)
1987	249980	1354.3	53.4
1988	304558	1298.2	59.2
1989	291727	1472.9	66.0
1990	308512	1165.0	70.5
1991	356190	1291.6	72.6
1992	391844	1507.7	82.6
1993	475486	1466.8	90.0
1994	422125	1432.1	90.5
1995	425690	2052.2	103.5
1996	514623	2356.6	130.3
1997	567639	2045.5	130.3
1998	525443	1337.7	127.5

- Quin dels dos indicadors presenta major correlació mostral amb les vendes de la companyia?
- Un cop determinat l'índex més correlacionat amb les vendes, fer-lo servir per ajustar una recta de regressió entre aquest i les vendes i efectueu la previsió per als anys 1999 i 2000.
- Discutiu el possible interès d'utilitzar ambdós coeficients per preveure les vendes a través de la regressió múltiple.

3. L'empresa BONSNEGOCIS comercialitza una línia de productes adreçada a famílies de renda mitjana. Per tal d'avaluar el seu mercat potencial ha dut a terme un estudi on, entre d'altres variables, s'han pres les dades de consum per família i any, així com els de renda familiar disponible. Diguem: X = "Renda familiar anual en milions de ptes.", Y = "Consum per família i any. ptes."
- Per a 22 famílies enquestades s'obtingué la taula següent.

Y	X	Y	X	Y	X	Y	X
4.9	4.6	7.0	9.0	5.7	10.0	6.5	13.1
6.2	11.9	5.3	5.7	5.1	6.5	5.2	6.9
5.8	9.0	5.2	5.3	5.3	8.9	5.2	6.6
5.0	4.4	6.1	11.3	5.1	7.0	5.6	9.6
6.3	13.1	5.4	6.9	5.6	8.5	5.8	11.1
5.9	10.4	6.0	9.6				

- Representeu les dades en un diagrama de dispersió i estudeu la correlació entre les dues variables. És el coeficient de correlació prou gran?
 - Construiu un model de regressió lineal que expressi l'associació entre el consum i renda.
 - Estimeu els coeficients del model anterior i obteniu una estimació de la variància dels errors.
 - Feu una previsió del consum familiar per una família amb una renda de $X = 15$.
 - Calculeu els residus. Hi ha valors atípics?
4. Es vol estudiar la influència del sexe i els anys d'experiència per mesurar els nivells de salaris assolits pels executius de vendes d'una gran companyia multinacional. Per a això s'ha obtingut la taula següent:

Anys d'experiència	1	2	3	4	5	6
Sou (homes)	20710	23160	23210	24140	25590	25760
Sou (dones)	19510	20440	21340	21760	22750	23200

- Elaboreu un model de regressió per explicar el sou a partir dels anys d'experiència.
- Dibuixeu les rectes de regressió corresponents als homes i a les dones. Hi ha diferència en els pendents?

6. Una empresa farmacèutica vol relacionar les hores que inverteixen els venedors en visitar clients i les vendes efectives d'aquests venedors. Després de fer el seguiment durant un mes de les vendes, obté la taula 1.

Taula 1: Relació d'hores de visita i vendes obtingudes							
Hores	Vendes	Hores	Vendes	Hores	Vendes	Hores	Vendes
90	503	77	495	126	700	101	548
125	499	130	823	142	866	87	528
124	695	95	487	146	589	115	566
85	534	77	496	114	642	93	466
133	881	139	939	118	1861	107	708
128	850	77	527				

- a) Representeu les dades en un diagrama de dispersió i estudeu la correlació entre les hores i les vendes efectuades. És el coeficient de correlació prou gran?
- b) Construïu un model de regressió lineal que permeti preveure els valors de les vendes a partir dels de les hores.
- c) Quin valor de vendes s'espera obtenir si un venedor ha estat visitant 100 hores?
7. Es va fer un experiment en un supermercat per observar la relació entre la quantitat d'espai disponible per a exhibició assignat a una marca de cafè i les vendes setmanals. Es va distribuir una quantitat d'espai de 3, 6 o 9 metres quadrats de forma aleatòria durant 12 setmanes i es va deixar constant, en 3 metres quadrats la quantitat assignada a les altres marques. Les dades obtingudes al cap de les 12 setmanes varen ser:

Espai (metres q.)	Vendes (unitats)	Espai (metres q.)	Vendes (unitats)
6	526	6	434
3	421	3	443
6	581	9	590
9	630	6	570
3	412	3	346
9	560	9	672

- a) Ajusteu un model de regressió lineal per explicar les vendes a partir de l'espai destinat a l'exposició. Si el coeficient de determinació es baix proveu amb un model quadràtic obtingut mitjançant regressió múltiple.
- b) Trobeu la predicció suposat que s'assignen 6 metres quadrats.
- c) Espereu que la relació continuï essent lineal per un ampli recorregut de valors de X , per exemple de 1 a 30?

BIBLIOGRAFIA

- Batista, J.M. i Valls, M. (1985) "Técnicas Gráficas en Análisis Exploratorio de Datos". *Qüestió* V.9 nº 3, Barcelona.
- Batista, J.M. i Valls, M. (1985) "Nuevas Técnicas de Análisis Estadístico de Datos: Tabulación y Síntesis Numérica". *Qüestió* V.9 nº 2, Barcelona.
- Guillén, M. (1992) *Anàlisi i descripció de dades*. Departament d'Econometria, Estadística i Economia Espanyola. Universitat de Barcelona.
- Hoaglin, D.C., Mosteller, F. i Tukey, J.W. (1983) *Understanding Robust and Exploratory Data Analysis*. J. Wiley, New York.
- Institut d'Estadística de Catalunya (1990) *Anuari Estadístic de Catalunya*.
- Mendenhall, W. (1990) *Estadística para administradores*. Grupo Editorial Iberoamericana, México.
- Montgomery, D.C. and Peck, E.A. (1992) *Introduction to Linear Regression Analysis*. J. Wiley, New York.
- Moore, D.S. (1998) *Estadística aplicada básica*. Antoni Bosch, Barcelona.
- Peña, D. (1987) *Estadística: Modelos y Métodos. 2. Modelos lineales y series temporales*. Alianza Universidad Textos, Madrid.
- Sánchez, A., Calvo, M. y Ruiz de Villa, C. (1994) *Pràctiques d'Estadística amb STATGRAPHICS*. Publicacions del Departament d'Estadística nº 6. Universitat de Barcelona.
- Tomeo, V. y Uña, I. (1989) *Diez lecciones de Estadística descriptiva*. Editorial AC, Madrid.
- Tukey, J.W. (1977) *Exploratory Data Analysis*. Addison-Wesley, Londres.
- Wonnacott & Wonnacott. *Estadística Básica Práctica*. Ed. Limusa, México.

