

Tècniques d'anàlisi de dades per a l'empresa

Ivan Soler-Ramos

Joan Torrent-Sellens

25 hores

Ivan Soler-Ramos

Llicenciat en Sociologia (UB). Post-grau en Minería de dades (UOC). Investigador del grup de recerca interdisciplinària sobre les TIC, i2TIC, i expert en anàlisi de dades, estadística i econometria. 10 anys d'experiència en recerca de mercats en el sector privat.

Joan Torrent-Sellens

Doctor en Societat de la Informació i el Coneixement (IN3-UOC). Màster en Anàlisi d'economia aplicada (UAB). Llicenciat en Ciències Econòmiques i Empresarials (UAB). Professor dels Estudis d'Economia i Empresa de la UOC, i director de la UOC Business School. Director del grup de recerca interdisciplinària sobre les TIC, i2TIC, i investigador sobre l'economia internacional, espanyola i catalana.

Índex

Introducció	5
Objectius	7
1. Plantejament d'una recerca	8
2. Les dades a les metodologies quantitatives	9
2.1. L'origen de les dades.....	9
2.2. Dades primàries i dades secundàries.....	10
2.3. Tècniques de recollida d'informació.....	11
3. Les variables	12
3.1. Concepte de <i>variable</i>	12
3.2. Variables quantitatives i variables qualitatives.....	12
3.3. Variables independents i variables dependents.....	13
4. Població i mostra	14
5. Anàlisi exploratòria de dades	15
5.1. Anàlisi exploratòria univariant.....	15
5.1.1. Anàlisi exploratòria univariant amb variables quantitatives.....	15
5.1.2. Anàlisi exploratòria univariant amb variables qualitatives.....	24
5.2. Anàlisi exploratòria bivariant.....	30
5.2.1. Transformació de variables quantitatives en variables qualitatives.....	31
5.2.2. Anàlisi exploratòria bivariant amb variables qualitatives.....	32
Resum	38
Glossari	39
Bibliografia	41

Introducció: la necessitat de les dades

Des de l'empresa més petita fins a la més gran, totes necessiten dades i totes generen dades. Generar dades i interpretar-les i analitzar-les correctament és essencial per a la generació de valor de l'empresa. Conèixer l'entorn i els mercats en què s'ubica l'activitat empresarial, interpretar-ne els canvis i preveure'n el futur, i analitzar el procés intern de generació de valor de l'empresa, són alguns dels exemples en què l'anàlisi de dades pot ser molt útil.

De fet, perquè una empresa funcioni correctament en el mercat és essencial que prengui bones decisions. Això depèn de múltiples qüestions, però en aquest curs ens centrarem en una, una de realment important: el paper de les dades en la presa de decisions.

Les decisions, perquè siguin bones, s'han de guiar per un coneixement tan profund com sigui possible sobre allò que s'ha de decidir. I la informació, la bona informació, és essencial per a assolir el coneixement que ens condueix cap a una presa de decisions correcta. Alhora, una de les millors maneres d'obtenir una bona informació és fer una anàlisi de dades acurada.

Fins ara hem utilitzat tres conceptes –*dades*, *informació* i *coneixement*– que massa habitualment s'usen de manera indistinta i d'una manera apropiada.

Les dades són les unitats mínimes i, per si mateixes, són irrelevantes per a la presa de decisions. És mitjançant el processament d'aquestes dades quan hi donem significat, i aquest significat l'anomenem *informació*. És a dir, la informació són les dades que han adoptat un significat.

Aquest significat l'assolim processant les dades, però no solament mitjançant càlculs matemàtics o estadístics. Les dades poden acabar convertint-se en informació si les resumim, les contextualitzem o bé les categoritzem.

Finalment, aquestes dades amb significat es converteixen en coneixement quan es combinen amb les altres informacions de què disposem, amb la nostra experiència o amb el nostre raonament. El procés de conversió dades/informació/coneixement és imprescindible per a prendre decisions correctament a l'empresa.

Molt sovint la informació necessària per a prendre decisions no està estructurada, ni enregistrada, ni disponible. Molts petits empresaris coneixen de

memòria els seus clients, saben què volen, com ho volen, quan ho volen, i també coneixen els seus competidors. Saben com funciona el seu sector i quines són les característiques dels seus mercats. Aquesta informació valuosa, però, no està disponible per a altres persones i processos, cosa que pot ser perjudicial per a l'empresa. Una estructuració d'aquesta informació mitjançant dades sempre serà útil per a optimitzar els processos de decisió i facilitarà el creixement de l'empresa. I, encara més, una empresa que es dediqui a recollir les dades i no en faci cap tipus d'anàlisi malbarata un actiu essencial perquè funcioni correctament en el mercat. Així, per exemple, pot perdre l'oportunitat d'identificar noves vetes de mercat (*market niche*), noves maneres de diversificar-ne els productes, noves vies d'innovació o d'internacionalització, i així millorar-ne la posició relativa en el mercat.

Acabem de constatar que l'anàlisi de dades és primordial en el procés de conversió de dades en informació i d'informació en coneixement. Aquest procés és imprescindible per a prendre decisions correctament, que al final, redunda en un millor funcionament de l'empresa. Precisament, el curs que ara comencem pretén fer una primera aproximació a l'anàlisi de dades a l'empresa i donar a l'estudiant uns primers instruments per a fer aquesta anàlisi. No es tracta d'un curs d'estadística, ni d'un curs de fulls de càlcul. Pretenem fer una introducció rigorosa, però pràctica, a l'anàlisi de dades per a l'empresa, utilitzant exemples sobre la informació de mercat.

Objectius

Els objectius d'aquest curs són els següents:

1. Distingir entre *dades*, *informació* i *coneixement*, i copsar com l'anàlisi de dades és imprescindible per a prendre decisions correctament a l'empresa.
2. Entendre què és una recerca i com es pot plantejar.
3. Interpretar i distingir els diferents tipus de variables, i aprendre a dissenyar un procés de recerca.
4. Comprendre els fonaments pràctics de l'anàlisi de dades exploratori univariante i bivariante.
5. Fer una primera diagnosi sobre les implicacions de l'anàlisi de dades en la presa de decisions de l'empresa.

1. Plantejament d'una recerca

El primer pas per a obtenir la informació necessària i fer anàlisi de dades és **plantejar una recerca**. Des de la més senzilla fins a la més complexa, tota recerca ha de respondre les qüestions següents:

Hem de tenir clar què volem investigar? És a dir, quin tipus de fenomen volem investigar i per què el volem investigar? Què farem amb la informació que n'obtenim?

També cal dissenyar com aconseguir aquesta informació? Per això, hem de saber quines dades seran necessàries per a poder recollir-les i transforma-les en informació i, finalment, en coneixement.

Per tant, hem d'establir el mètode, la metodologia i, finalment, les eines, per a poder recollir les dades que ens interessin.

En resum, les principals preguntes que ens haurem de fer a l'hora de plantejar una recerca són:

- Quin és l'objecte de la recerca?
- Quin n'és l'objectiu?
- Quines són les dades que necessitarem per a donar-hi resposta?
- On són aquestes dades?
- Han estat recollides amb anterioritat?
- Estan disponibles?

La decisió sobre la metodologia de recerca...

... està condicionada per les possibilitats que ens ofereixen les dades i pels recursos de què disposem per a analitzar-les.

2. Les dades a les metodologies quantitatives

2.1. Dades primàries i dades secundàries

Bàsicament, les dades poden ser de dos tipus: **dades primàries i dades secundàries**. Les dades primàries són aquelles que hem recollit en una primera instància, és a dir, amb els instruments que nosaltres mateixos hem definit. Les dades secundàries són dades recollides (i processades) prèviament per altres persones o institucions.

La diferència essencial entre les dades primàries i les secundàries no respon al tipus d'informació que se'n pot extreure. Si ens hi fixem, tota dada secundària va ser en el moment de la recollida una dada primària i tota dada primària serà, un cop recollida i processada, una dada secundària per als altres. La diferència entre les dades primàries i les dades secundàries s'esdevé a través de l'encaix entre la informació que necessitem i la que tenim disponible. Quan volem una informació que no tenim disponible, llavors cal fer una recerca amb dades primàries.

Per tant, a l'hora de plantejar una recerca hem de respondre la pregunta següent:

- La informació que necessitem, la tenim disponible en alguna font fiable?

Si la resposta és afirmativa, el sentit comú ens portarà a utilitzar la dada ja processada, la dada secundària, ja que usualment la dada primària es troba en una fase inicial i necessita més esforç per a recollir-la. Tot i amb això, sovint succeeix que la informació que necessitem no està disponible, no hi tenim accés per la raó que sigui, o perquè ningú l'ha recollit abans.

Alguns breus apunts sobre les dades secundàries

L'anàlisi de dades secundàries és molt important en els passos inicials d'una recerca com a manera de contextualitzar i augmentar el nostre coneixement previ a la recollida primària d'informació. L'anàlisi de dades secundàries ens ajuda a identificar variables clau, seleccionar mostres, formular dissenys de recerca o suggerir nous enfocaments.

El principal avantatge de les dades secundàries és l'accessibilitat i la immediatesa, atès que ja han estat processades i recollides.

També tenen alguns inconvenients: poden no encaixar bé amb el problema que hem d'estudiar, no tenir la precisió desitjada o bé estar referenciades en unitats inadequades o en períodes diferents. Per tant, és essencial verificar-les.

Per exemple...

Les dades de l'IPC són secundàries perquè les elabora l'Institut Nacional d'Estadística. En canvi, si fèssim una recerca sobre la dinàmica dels preus a la nostra empresa, aquestes dades serien primàries.

Un altre exemple...

No té gaire sentit que algú reculli la informació referent a les opinions i a les preferències dels nostres clients i si ho fa, segurament, no ens la lliurarà així com així.

2.2. L'origen de les dades

Segons l'origen de les dades, es poden classificar en els tipus següents:

1. **Dades internes:** quan la font de dades és a dins de la nostra organització. Aquests tipus de dades poden ser molt interessants per a fer una anàlisi que optimitzi el funcionament de la nostra empresa. Alguns dels avantatges que presenten aquest tipus de dades són l'accessibilitat fàcil i el baix cost que representen.
2. **Dades externes:** són dades que podem obtenir d'una altra font. Aquí s'inclouen totes les dades primàries que recollim per resoldre les nostres necessitats d'informació. Hi ha diversos tipus de fonts de dades externes. Si les dividim en funció de l'accessibilitat, obtenim:

2.1. **Fonts obertes:** són dades externes que estan disponibles de manera oberta, en biblioteques o bé a Internet. Aquestes dades responen a publicacions generades per les administracions públiques, en llibres o bé les que podem obtenir en qualsevol institut d'estadística.

2.2. **Fonts semiobertes:** són dades que generen associacions empresarials, gremis o federacions d'empreses i que, generalment, s'ofereixen als seus integrants o associats. Segons si som integrants o no de l'organització, les dades seran una font oberta (per a nosaltres) o tancada (per als altres).

2.3. **Fonts tancades:** habitualment són dades estandarditzades que una empresa ha anat recollint i que processa i, finalment, ven a qui vulgui disposar d'aquesta informació. Habitualment, l'adquisició d'aquestes dades és força costosa. Algunes de les dades de font tancada més habituals fan referència a informació de mercat relativa a dades de consumidors, a l'estat de diferents sectors d'activitat, a dades de comunicació i audiència, i a altres elements de valor de l'empresa.

2.4. **Fonts sindicades:** són dades de pagament, o bé de pagament per cada ús, o bé de pagament per subscripció. Són recollides i processades per empreses que per iniciativa pròpia fan estudis de mercat i publiquen els informes corresponents. Aquesta informació és de pagament i qualsevol persona hi pot accedir. Per exemple, els estudis que du a terme AC Nielsen són alguns del més coneguts.

Algunes possibles fonts de dades internes...

... són aquelles que es generen amb l'activitat comercial (distribució de vendes per productes, zones o clients), amb l'activitat productiva (costos, existències) o bé aquelles resultants de la informació comptable i financera, per esmentar-ne algunes de les més importants.

Visiteu els webs següents

www.idescat.cat

www.ine.es

<http://epp.eurostat.ec.europa.eu/>

Visiteu el web

<http://es.nielsen.com>

2.3. Tècniques de recollida d'informació

Les tècniques de recollida d'informació depenen: 1) dels objectius que volem assolir amb la recerca; 2) de la realitat que volem estudiar; 3) del tipus de recerca que en conseqüència volem iniciar; i 4) del tipus d'anàlisi de dades que farem un cop obtinguda la informació.

Segons si són dades primàries o secundàries, algunes de les tècniques de recollida d'informació més freqüents són:

1. Dades primàries:

1.1. La **recollida d'informació per observació** és aquell procediment en què s'obté la informació per observació directa dels fenòmens d'interès: l'investigador enregistra la informació en les anotacions. La recollida de dades per observació és una tècnica de recollida d'informació pròpiament si s'orienta a un objectiu de recerca, està planificada i cerca obtenir respostes a proposicions generals.

1.2. L'**enquesta** és la tècnica de recollida d'informació més habitual en la metodologia quantitativa. L'enquesta utilitza procediments estandaritzats d'interrogació amb la finalitat d'aconseguir mesures quantitatives sobre les característiques subjectives i/o objectives d'una població.

2. Dades secundàries:

2.1. Les **estadístiques** corresponen a la informació proporcionada pels instituts d'estadística oficial.

3. Les variables

3.1. Concepte de variable

Arribats a aquest punt, a continuació farem un breu repàs dels conceptes més bàsics relacionats amb l'**estadística descriptiva**.

L'estadística descriptiva o anàlisi exploratòria de dades és la forma més bàsica de l'estadística i té un notable contingut instrumental.

No és objecte d'aquest curs esdevenir un manual d'estadística, però sí que com a eina que necessitarem per a analitzar dades primàries, és interessant que fem un breu repàs dels principals conceptes que necessitarem treballar molt orientat a la pràctica.

El primer que hem de tenir en compte és que per a obtenir la informació que volem hem de fer operatius els conceptes. És a dir, hem de ser capaços de transformar els conceptes, les preguntes essencials que ens fem, en forma de variables. És a dir, hem de fer operatives les dimensions de la realitat perquè es pugin treballar de manera empírica.

Així doncs, cal definir el concepte de **variable**.

Una **variable** no és res més que una de les diferents dimensions de la realitat que volem conèixer. Parlarem de **valor de la variable** quan ens referim a la mesura que la variable presenta. Parlarem de **variació de la variable** quan ens referim a la diferència que hi ha entre els valors d'una variable.

Tractarem l'estadística de manera molt intuïtiva...

... sense profunditzar en les matemàtiques que hi ha al darrere, tot aplicant d'una manera pràctica conceptes bàsics d'anàlisi.

3.2. Variables quantitatives i variables qualitatives

Des del contingut d'una variable, és a dir des del valor que representa, i d'una manera simple, és possible agrupar dos grans tipus de variables:

1. **Variable quantitativa**: variable que representa una xifra.
2. **Variable qualitativa**: variable que representa una categoria.

Al seu torn, les variables quantitatives es poden dividir en els tipus següents:

1. **Variable quantitativa contínua:** quan el valor no representa cap interrupció més enllà de les que l'instrument de mesurament ens planteja.
2. **Variable quantitativa discreta:** aquelles que presenten una interrupció entre els seus valors perquè els intermedis no existeixen.

Les variables qualitatives es poden dividir en els tipus següents

1. **Variable qualitativa ordinal:** són aquelles que presenten un ordre establert dins d'una escala. Per exemple, la típica valoració d'opinió d'1 a 10. A la pràctica, i pel tema que ens ocupa, no hi ha diferència entre les variables qualitatives ordinals i les variables quantitatives discretes.
2. **Variable qualitativa nominal:** en aquest cas les variables representen categories que no tenen relació de mesura entre elles.

Un cas específic i molt habitual de variable qualitativa nominal és l'anomenada **variable dicotòmica**. Aquesta variable solament pren dos valors: habitualment 1 o 2, o bé 0 o 1. S'utilitza per a assenyalar la presència o l'absència d'una característica, o l'afirmació o la negació a una pregunta d'opinió d'un qüestionari.

3.3. Variables dependents i variables independents

Finalment, hi ha una última diferència que cal abordar quan tractem variables. Es tracta de la funció que poden representar en la nostra anàlisi. Segons aquesta aproximació, les variables poden ser dependents i independents:

1. Les **variables dependents** són variables que suposem que varien en funció d'una altra variable.
2. Les **variables independents** són variables que suposem que no evolucionen per cap altra variable, tot i que la seva variació sí que incideix en l'evolució d'altres variables dependents.

Per exemple, imaginem que mesurem l'alçada dels nens d'un col·legi i volem esbrinar com varien. Per a fer-ho, disposem de 2 variables: l'edat del nen i la seva alçada. La variable independent seria l'edat i la variable dependent seria l'alçada. És a dir, la seqüència lògica del pensament ens porta a interpretar directament que l'alçada del nen varia amb la seva edat, no pas al contrari.

Cal afegir, però, que hi ha molts casos en què aquesta relació de dependència no és gens clara i cal fer anàlisis més sofisticades per tal de treure'n l'entrellat. Aquests tipus d'anàlisi cauen fora de l'abast d'aquest curs.

Per exemple, l'alçada

Si mesurem en centímetres l'alçada d'un individu, evidentment aquesta presenta interrupcions perquè no és infinita.

Com a exemple

El nombre de fills d'una persona.

Per exemple, els colors d'un cotxe

Si assignem 1 al color verd i 2 al color vermell, evidentment no podem interpretar que el color vermell és el doble del color verd.

4. Població i mostra

Un cop tenim decidides les dades que necessitarem i la seva tipologia, a continuació hem de veure on les hem de trobar. Hi ha diversos tipus de llocs on trobar-les, però en aquest curs ens centrarem en les dades obtingudes segons la declaració d'un entrevistat. Per això utilitzarem **enquestes a mostres representatives i fiables d'una població**.

S'entén per **població** el conjunt d'individus que disposen de les característiques que necessitem observar.

S'entén per **mostra** un subconjunt d'aquesta població que representa a escala les característiques principals de la població.

Quan la grandària de la població és gran i no podem arribar a tothom, aleshores hem de definir una mostra. Per a fer-ho hi ha diferents tècniques, però nosaltres ens centrarem en dues condicions: la **representativitat** i la **fiabilitat** de la mostra.

Una **mostra és representativa** quan representa les característiques de la població objecte d'estudi. És a dir, quan obtenim un conjunt d'individus inferior a la població i que la representen. A més, amb el disseny de la mostra hem de suposar que cometem un error. Aquest error l'anomenem **error mostral**.

Una **mostra és fiable** quan l'error mostral que comporta treballar amb una mostra és assumible. El fet que una mostra sigui o no sigui fiable depèn essencialment del criteri de l'investigador i del tipus de recerca.

Metodològicament, per a poder treballar amb una mostra hi ha diverses tècniques. A continuació n'esmentem dues:

- La mostra aleatòria o probabilística simple (MAS): en aquesta tècnica tots els individus que formen la població tenen les mateixes probabilitats de ser escollides com a unitats mostrals.
- La **mostra estratificada**: en aquesta tècnica la població se segmenta en diferents estrats o classes i en cadascun s'escull una mostra aleatòria. Els elements presents en cada estrat han de ser més similars entre si que respecte a la població. Per tant, es tracta de buscar la màxima homogeneïtat en cada estrat i la màxima heterogeneïtat entre els diferents estrats. A cada estrat li correspon una **submostra**, i el conjunt de les submostres constitueix el total de la mostra poblacional.

Per exemple

Podem obtenir dades mesurant alguna de les característiques d'un producte o servei que produïm. Però, per conèixer l'opinió que tenen els nostres clients del nostre producte o servei, els hem de preguntar. I això s'ha de fer de manera estructurada.

5. Anàlisi exploratòria de dades

Una vegada les dades que necessitem per a la nostra recerca estan identificades i recollides, a continuació ja estarem en disposició d'iniciar l'anàlisi de dades. Per a l'anàlisi de dades, emprarem l'estadística descriptiva, també anomenada *anàlisi exploratòria de dades*.

L'**anàlisi exploratòria de dades** és un conjunt de tècniques que tenen com a finalitat descriure, presentar i, finalment, analitzar les dades que volem observar en una població.

Per a assolir els objectius...

... de l'anàlisi exploratòria de dades utilitzarem les taules, els gràfics i els resums d'estadístics.

L'anàlisi exploratòria de dades s'anomena **univariant** quan mesura una de les característiques dels individus de la població. I s'anomena **bivariant** quan mesura una de les característiques de la població en funció d'una altra característica.

Per a treballar amb les dades, aquestes s'organitzen seguint una estructura particular. Habitualment, les emprades pels fulls de càlcul i pels programes d'anàlisi estadística. En general, cada unitat mostral (registre/individu) ocupa una fila diferent i les diferents variables (és a dir, les dimensions de la realitat que volem observar) es presenten en columnes diferents.

A més, en l'anàlisi de dades mitjançant una enquesta hem de tenir en compte el següent:

- La quantitat i la qualitat de les respostes que volem obtenir no està necessàriament relacionada amb la dimensió de la mostra.
- Les dades seran més útils a mesura que les puguem comparar amb altres de similars o amb altres poblacions de referència.

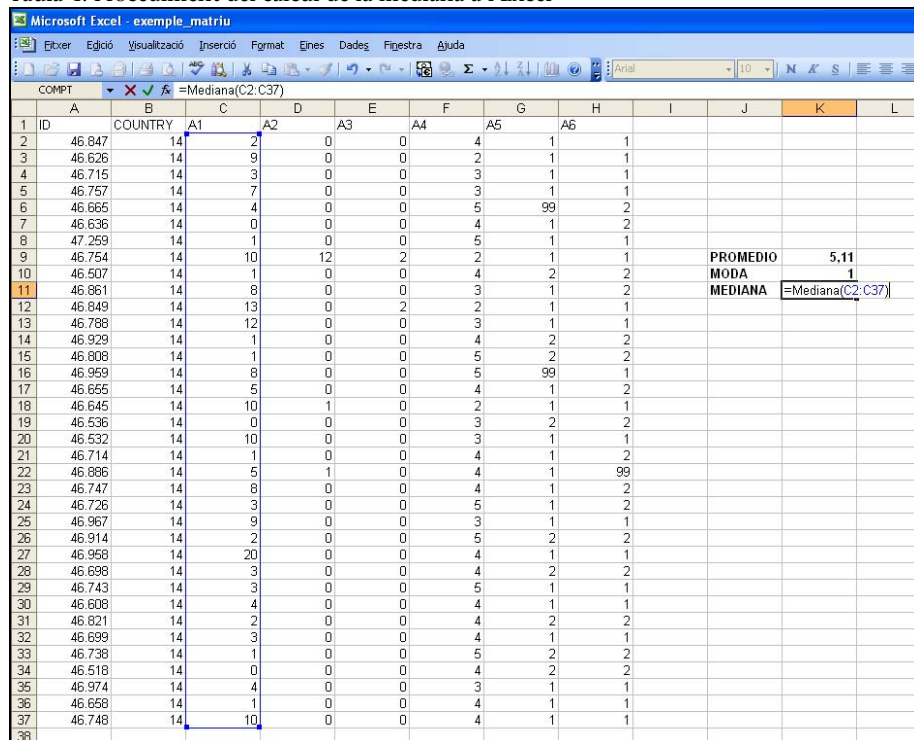
5.1. Anàlisi exploratòria univariant

L'estadística exploratòria univariant mesura una característica de l'individu que pot presentar diverses modalitats. Segons la informació que contingui el tipus de variable que volem analitzar, l'anàlisi descriptiva que durem a terme serà d'un tipus o d'un altre. Al llarg d'aquest curs tractarem les anàlisis més habituals en estadística exploratòria.

5.1.1. Anàlisi exploratòria univariant amb variables quantitatives

La **mediana** és el valor que divideix tota una sèrie de dades en dues parts exactament iguals. És a dir, on la quantitat de dades inferiors a la mediana és igual a la quantitat de dades superiors a la mediana. A l'Excel, la funció *MEDIANA* [MEDIANA (número 1, número 2...)] permet calcular la mediana d'un conjunt de dades (taula 4).

Taula 4. Procediment del càlcul de la mediana a l'Excel



	A	B	C	D	E	F	G	H	I	J	K	L
1	ID	COUNTRY	A1	A2	A3	A4	A5	A6				
2	46.847	14	2	0	0	4	1	1				
3	46.626	14	9	0	0	2	1	1				
4	46.715	14	3	0	0	3	1	1				
5	46.757	14	7	0	0	3	1	1				
6	46.665	14	4	0	0	5	99	2				
7	46.636	14	0	0	0	4	1	2				
8	47.259	14	1	0	0	5	1	1				
9	46.754	14	10	12	2	2	1	1			PROMEDIO	5,11
10	46.507	14	1	0	0	4	2	2			MODA	1
11	46.861	14	8	0	0	3	1	2			MEDIANA	=Mediana(C2:C37)
12	46.849	14	13	0	2	2	1	1				
13	46.788	14	12	0	0	3	1	1				
14	46.929	14	1	0	0	4	2	2				
15	46.808	14	1	0	0	5	2	2				
16	46.959	14	8	0	0	5	99	1				
17	46.655	14	5	0	0	4	1	2				
18	46.645	14	10	1	0	2	1	1				
19	46.536	14	0	0	0	3	2	2				
20	46.532	14	10	0	0	3	1	1				
21	46.714	14	1	0	0	4	1	2				
22	46.886	14	5	1	0	4	1	99				
23	46.747	14	8	0	0	4	1	2				
24	46.726	14	3	0	0	5	1	2				
25	46.967	14	9	0	0	3	1	1				
26	46.914	14	2	0	0	5	2	2				
27	46.958	14	20	0	0	4	1	1				
28	46.698	14	3	0	0	4	2	2				
29	46.743	14	3	0	0	5	1	1				
30	46.608	14	4	0	0	4	1	1				
31	46.821	14	2	0	0	4	2	2				
32	46.699	14	3	0	0	4	1	1				
33	46.736	14	1	0	0	5	2	2				
34	46.518	14	0	0	0	4	2	2				
35	46.974	14	4	0	0	3	1	1				
36	46.658	14	1	0	0	4	1	1				
37	46.748	14	10	0	0	4	1	1				
38												

Les mesures de dispersió s'utilitzen profusament en l'anàlisi de dades i, sovint, serveixen per a completar la informació que ens aporten les mesures de tendència central.

Les **mesures de dispersió** ens indiquen la distància mitjana de les dades respecte a la mitjana aritmètica. En aquest curs tractarem el **rang**, la **desviació mitjana**, la **variància** i la **desviació estàndard o típica**.

El **rang** és la mesura de desviació que té un càlcul més senzill. Simplement, es tracta de la diferència entre el valor de la distribució més gran i el més petit. Se sol anomenar amb la lletra *R*.

La **desviació mitjana** és la mitjana de les diferències entre els valors de la variable i la mitjana aritmètica en valor absolut. A l'Excel, la funció *DESVPROM* [DESVPROM (número 1, número 2...)] permet calcular la desviació mitjana d'un conjunt de dades (taula 5).

Taula 5. Procediment del càlcul de la desviació mitjana a l'Excel

C	D	E	F	G	H	I	J	K	L
2	0	0	4	1	1				
9	0	0	2	1	1				
3	0	0	3	1	1				
7	0	0	3	1	1				
4	0	0	5	99	2				
0	0	0	4	1	2				
1	0	0	5	1	1				
10	12	2	2	1	1		PROMEDIO	5,11	
1	0	0	4	2	2		MODA	1	
8	0	0	3	1	2		MEDIANA	4	
13	0	2	2	1	1		DESVIACIÓN MEDIA	=DESV.PROM(C2:C37)	
12	0	0	3	1	1				
1	0	0	4	2	2				
1	0	0	5	2	2				
8	0	0	5	99	1				
5	0	0	4	1	2				
10	1	0	2	1	1				
0	0	0	3	2	2				
10	0	0	3	1	1				
1	0	0	4	1	2				
5	1	0	4	1	99				
8	0	0	4	1	2				
3	0	0	5	1	2				
9	0	0	3	1	1				
2	0	0	5	2	2				
20	0	0	4	1	1				
3	0	0	4	2	2				
3	0	0	5	1	1				
4	0	0	4	1	1				
2	0	0	4	2	2				
3	0	0	4	1	1				
1	0	0	5	2	2				
0	0	0	4	2	2				
4	0	0	3	1	1				
1	0	0	4	1	1				
10	0	0	4	1	1				

La **variància** (S^2). Per assegurar que les diferències entre la mitjana i els punts d'un valor siguin positius és possible elevar-les al quadrat. La variància és el resultat de la divisió de la sumatòria de les distància entre cada dada i la mitjana aritmètica del conjunt de dades elevades al quadrat. A l'Excel, la funció *VAR* [*VAR* (número 1, número 2...)] permet calcular la variància d'un conjunt de dades (taula 6).

Taula 6. Procediment del càlcul de la variància a l'Excel

C	D	E	F	G	H	I	J	K	L
2	0	0	4	1	1				
9	0	0	2	1	1				
3	0	0	3	1	1				
7	0	0	3	1	1				
4	0	0	5	99	2				
0	0	0	4	1	2				
1	0	0	5	1	1				
10	12	2	2	1	1		PROMEDIO	5,11	
1	0	0	4	2	2		MODA	1	
8	0	0	3	1	2		MEDIANA	4	
13	0	2	2	1	1		DESVIACIÓN MEDIA	3,75	
12	0	0	3	1	1		VARIANZA	=VAR(C2:C37)	
1	0	0	4	2	2				
1	0	0	5	2	2				
8	0	0	5	99	1				
5	0	0	4	1	2				
10	1	0	2	1	1				
0	0	0	3	2	2				
10	0	0	3	1	1				
1	0	0	4	1	2				
5	1	0	4	1	99				
8	0	0	4	1	2				
3	0	0	5	1	2				
9	0	0	3	1	1				
2	0	0	5	2	2				
20	0	0	4	1	1				
3	0	0	4	2	2				
3	0	0	5	1	1				
4	0	0	4	1	1				
2	0	0	4	2	2				
3	0	0	4	1	1				
1	0	0	5	2	2				
0	0	0	4	2	2				
4	0	0	3	1	1				
1	0	0	4	1	1				
10	0	0	4	1	1				

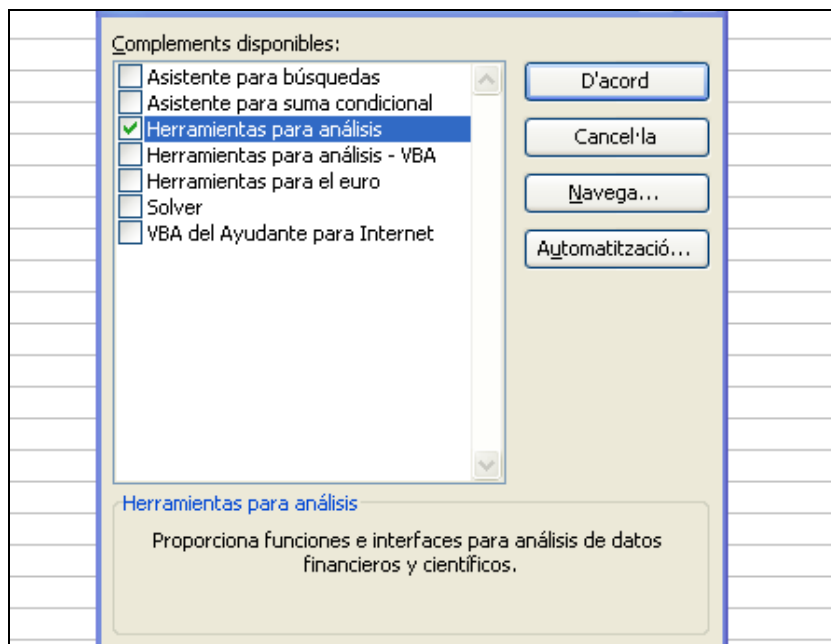
La variància té el problema que eleva també al quadrat les unitats de les dades originals. És a dir, si estem mesurant, per exemple, segons o minuts obtenim la variància en segons o minuts. Per tal d'evitar-ho, sovint, s'utilitza la **desviació estàndard o típica**. La desviació estàndard és l'arrel quadrada de la variància. A l'Excel, la funció *DEVEST* [DEVEST (número 1, número 2...)] permet calcular la desviació estàndard d'un conjunt de dades (taula 7).

Taula 7. Procediment del càlcul de la desviació estàndard a l'Excel

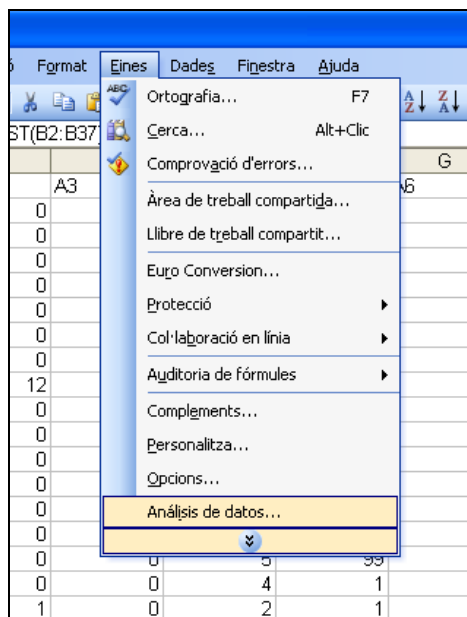
C	D	E	F	G	H	I	J	K
0	0	4	1	1				
0	0	2	1	1				
0	0	3	1	1				
0	0	3	1	1				
0	0	5	99	2				
0	0	4	1	2				
0	0	5	1	1				
12	2	2	1	1		PROMEDIO	5,11	
0	0	4	2	2		MODA	1	
0	0	3	1	2		MEDIANA	4	
0	2	2	1	1		DESVIACIÓN MEDIA	3,75	
0	0	3	1	1		VARIANZA	21	
0	0	4	2	2		DESVIACIÓN ESTANDAR	=DESVEST(B2:B37)	
0	0	5	2	2				
0	0	5	99	1				
0	0	4	1	2				
1	0	2	1	1				
0	0	3	2	2				
0	0	3	1	1				
0	0	4	1	2				
1	0	4	1	99				
0	0	4	1	2				
0	0	5	1	2				
0	0	3	1	1				
0	0	5	2	2				
0	0	4	1	1				
0	0	4	2	2				
0	0	5	1	1				
0	0	4	1	1				
0	0	4	2	2				
0	0	4	1	1				
0	0	5	2	2				
0	0	4	2	2				
0	0	3	1	1				
0	0	4	1	1				
0	0	4	1	1				

Aquesta informació, i alguna més, la podem obtenir d'una forma més fàcil i directa si utilitzem l'eina [ANÀLISI DE DADES] que ens ofereix el Microsoft Excel. Amb aquesta eina podem obtenir, simultàniament, totes aquestes estadístiques descriptives que hem vist fins ara.

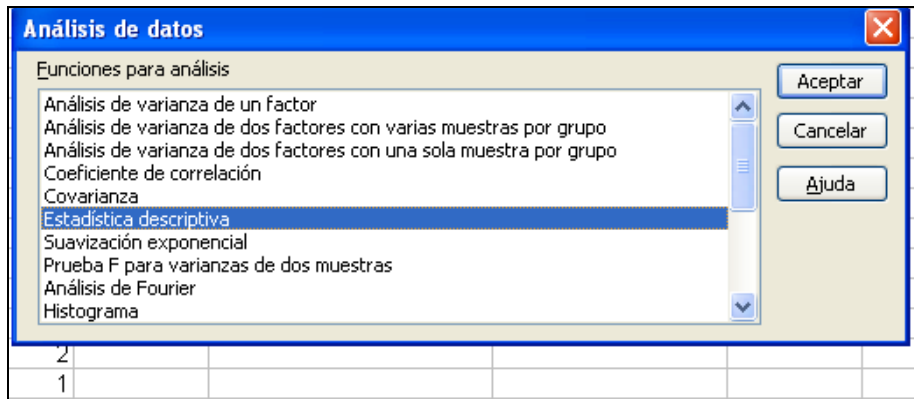
Per poder utilitzar aquesta eina hem de verificar si ha estat habilitada a l'ordinador. En primer lloc, hem d'entrar al menú [EINES] i observar si ens apareix l'opció [ANÀLISI DE DADES]. Si aquesta opció no està present, l'hem d'activar. Per a fer-ho cal seleccionar l'opció [COMPLEMENTOS] dins del menú [EINES] i seleccionar [HERRAMIENTAS PARA ANÁLISIS].



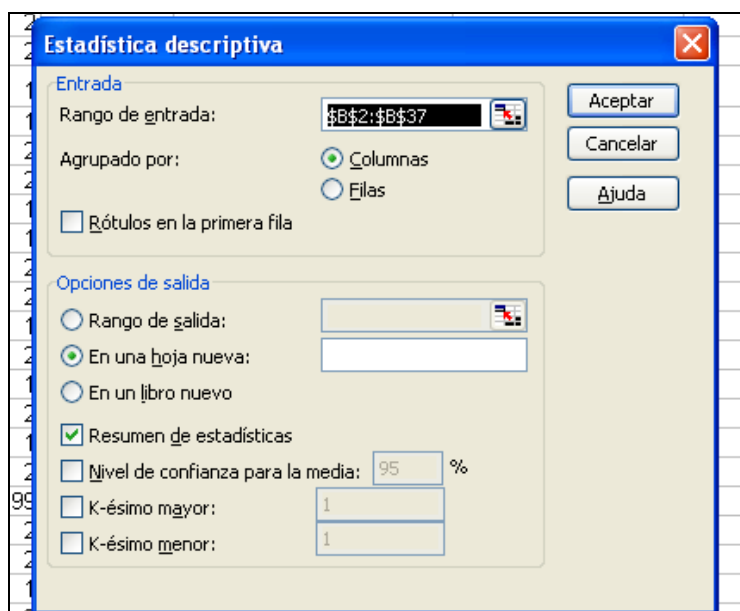
Posteriorment, hem de fer clic a [D'ACORD], i verificar que, al menú [EINES], aquesta opció hi surt.



Una vegada aquesta eina està habilitada, hi farem un clic a sobre i es desplegarà un menú on seleccionarem l'opció [ESTADÍSTICA DESCRIPTIVA].



Després ens apareixerà la pantalla següent:



En aquesta pantalla ens apareixen tot un seguit d'opcions que hem de tenir en compte. Hi ha opcions d'entrada, és a dir, les opcions que es refereixen a la manera com les dades entren a l'ordinador i opcions de sortida, és a dir, de quina manera ens tornarà l'ordinador les dades ja processades.

Pel que fa a les opcions d'entrada, tenim:

- [RANGO DE ENTRADA]. Hem d'indicar quines dades voldrem analitzar. Les podem seleccionar amb el ratolí.
- [AGRUPADO POR]. Aquí haurem de dir si les dades estan col·locades en columnes o en files. En aquest cas la variable A1 està situada en una columna.

Pel que fa a les opcions de sortida, tenim:

- [RANGO DE SALIDA], [EN UNA HOJA NUEVA], [EN UN LIBRO NUEVO]: són opcions que es refereixen al lloc on volem que l'Excel situï

les dades un cop estan processades. En aquest cas li demanem que les obri en un nou full.

- [RESUMEN DE ESTADÍSTICOS]: haurem de marcar aquesta opció perquè ens doni un resum dels estadístics que hem vist fins ara (i alguns altres).

Un vegada seleccionades les opcions, farem clic a [D'ACORD], i ens apareixerà en un nou full el resum següent:

Taula 7. Resum d'estadístics descriptius a l'Excel

	A	B	C
1	Columna1		
2			
3	Media	5,111111111	
4	Error típico	0,765089244	
5	Mediana	3,5	
6	Moda	1	
7	Desviación estándar	4,590535467	
8	Varianza de la muestra	21,07301587	
9	Curtosis	1,51090243	
10	Coefficiente de asimetría	1,174008367	
11	Rango	20	
12	Mínimo	0	
13	Máximo	20	
14	Suma	184	
15	Cuenta	36	
16			
17			
18			
19			

Així doncs, hem vist dues maneres d'aproximar-nos a l'anàlisi quantitativa exploratòria univariant utilitzant l'Excel. Usar l'eina [ANALISIS DE DATOS] és més complet, ràpid i fàcil que no pas utilitzar cada una de les funcions una a una.

Tot i amb això, a vegades serà preferible utilitzar una única funció. Hem de tenir en compte que si fem alguna modificació en les dades d'entrada o en la sortida processada, si utilitzem l'eina [ANALISIS DE DATOS] aquests canvis no es veuran reflectits i, per tant, haurem de tornar a repetir l'anàlisi. Això no succeeix si utilitzem la funció individual en una cel·la del full de dades.

5.1.2. Anàlisi exploratòria univariant amb variables qualitatives

L'anàlisi exploratòria més habitual amb variables qualitatives és la taula de freqüències.

Aquesta taula ens pot presentar els estadístics següents:

Freqüència absoluta: és el nombre d'individus que comparteixen la mateixa modalitat de la variable.

Freqüència relativa: és la proporció d'individus que comparteix la mateixa modalitat de la variable.

Freqüència absoluta acumulada: és el nombre d'individus que comparteixen una modalitat o alguna de menor.

Freqüència relativa acumulada: és la proporció d'individus que comparteixen una modalitat o alguna de menor.

Les freqüències acumulades tenen sentit en cas que les modalitats de les variables siguin ordenables. Per exemple, si prenem una variable quantitativa discreta com és el nombre de fills, pot tenir sentit observar-ne les freqüències acumulades.

És a dir si:

- El 20% d'una població no té cap fill.
- El 35% té un fill.
- El 20% té dos fills.

Podria ser una informació rellevant conèixer que $\frac{3}{4}$ parts ($20\% + 35\% + 20\% = 75\%$) de la població estudiada té 2 fills o menys.

En canvi, si la característica que hem d'analitzar és el nom del fill més gran, cap de les freqüències acumulades no tindrien sentit, ja que no és possible determinar quin nom és més gran o més petit que un altre.

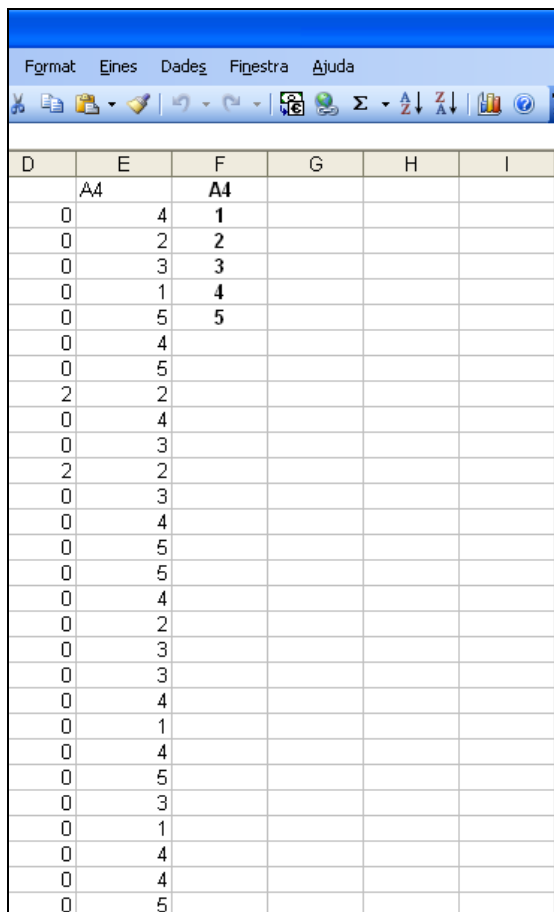
És a dir si:

- El 20% dels nens és diuen Marta.
- El 35% es diu Jordi.
- El 20% es diu Joan.

Podríem dir que el 75% del nens de la població objecte d'estudi es diuen Marta o Jordi o Joan, però no hi hauria *a priori* cap relació entre les modalitats de les variables.

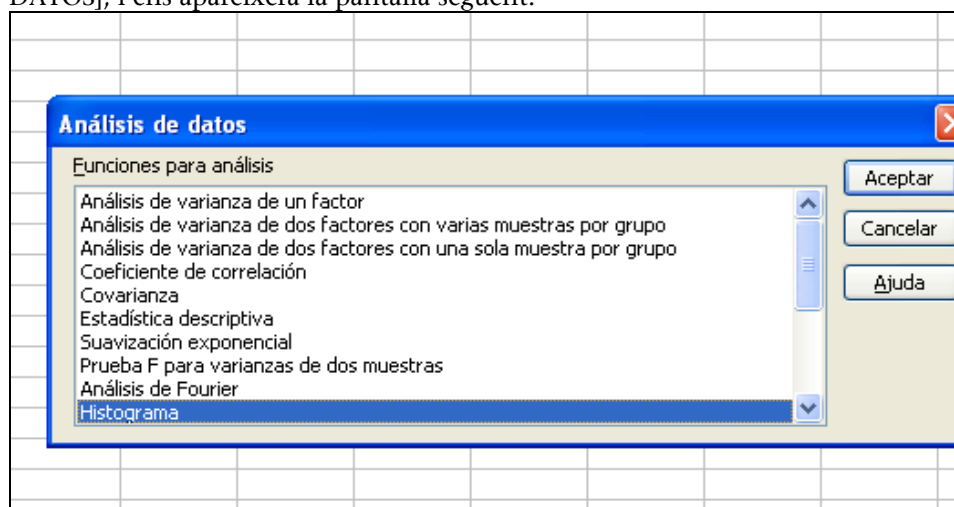
Veiem, a continuació, com es construeix una taula de freqüències al Microsoft Excel.

El primer que farem és escriure en una columna el nom de la variable i, successivament, les diferents modalitats que apareixen en la variable. Com es veu a la imatge d'exemple, la variable A4 en la columna E presenta 5 modalitats diferents (1, 2, 3, 4 i 5).



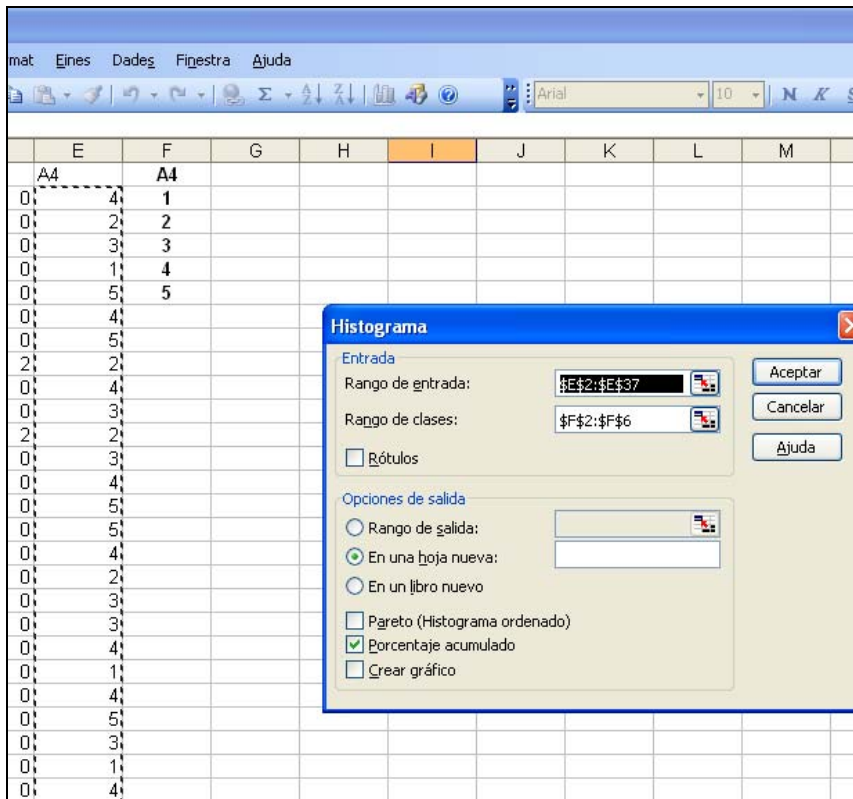
D	E	F	G	H	I
	A4	A4			
0	4	1			
0	2	2			
0	3	3			
0	1	4			
0	5	5			
0	4				
0	5				
2	2				
0	4				
0	3				
2	2				
0	3				
0	4				
0	5				
0	5				
0	4				
0	2				
0	3				
0	3				
0	4				
0	1				
0	4				
0	5				
0	3				
0	1				
0	4				
0	4				
0	5				

Ara tornarem a seleccionar dins del menú [EINES], l'opció [ANALISIS DE DATOS], i ens apareixerà la pantalla següent:



En aquesta pantalla haurem de seleccionar l'opció [HISTOGRAMA] i [ACEPTAR].

Un cop a dintre de l'opció [HISTOGRAMA] ens apareixerà la pantalla següent:



En aquesta pantalla trobem unes opcions molt semblants a les que hem vist amb anterioritat. Així, tornen a aparèixer opcions d'entrada, és a dir, les opcions que es refereixen a la manera com les dades entren a l'ordinador i opcions de sortida, és a dir, de quina manera ens tornarà l'ordinador les dades un cop processades.

Opcions d'entrada:

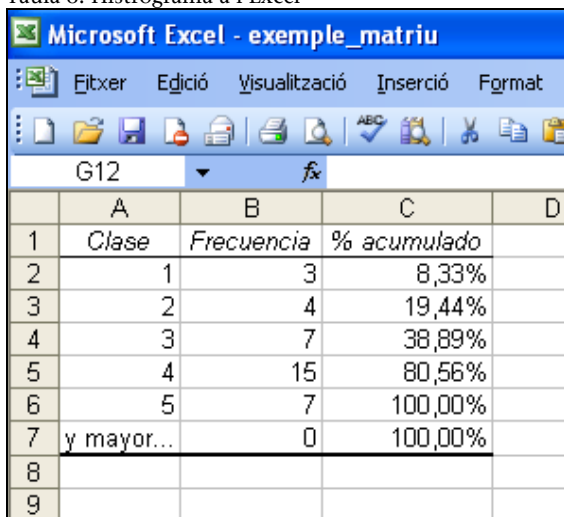
- [RANGO DE ENTRADA]: hem d'indicar quines dades voldrem analitzar. Les podrem seleccionar amb el ratolí.
- [RANGO DE CLASES]: aquí hem d'indicar les diferents modalitats que pren la variable A4. En aquest cas són aquelles situats en la columna F (\$F\$2:\$F\$6).

Opcions de sortida:

- [RANGO DE SALIDA], [EN UNA HOJA NUEVA], [EN UN LIBRO NUEVO]: són opcions que es refereixen a on volem que l'Excel situï les dades un cop estan processades. Novament, li demanarem que les obri en un nou full. En aquest cas marcarem l'opció [PORCENTAJE ACUMULADO].

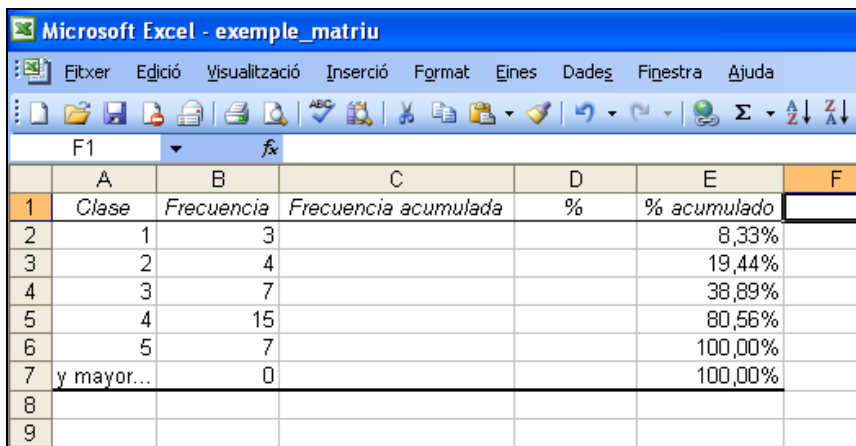
Un cop acceptem ens apareixerà en un nou full la taula següent (taula 8):

Taula 8. Histograma a l'Excel



	A	B	C	D
1	Clase	Frecuencia	% acumulado	
2	1	3	8,33%	
3	2	4	19,44%	
4	3	7	38,89%	
5	4	15	80,56%	
6	5	7	100,00%	
7	y mayor...	0	100,00%	
8				
9				

Com l'Excel no ens ofereix la freqüència absoluta acumulada ni la freqüència relativa, les haurem de construir manualment per tal de completar la taula de freqüències. Així, el primer que cal fer és obrir dues noves columnes mitjançant l'opció [COLUMNES] dins del menú [INSERCIÓ], assegurant que s'insereixin entre la [FRECUENCIA] i el [% ACUMULADO]. A la primera fila de la columna C escriurem [FRECUENCIA ACUMULADA] i a la primera fila de la columna D [%].



	A	B	C	D	E	F
1	Clase	Frecuencia	Frecuencia acumulada	%	% acumulado	
2	1	3			8,33%	
3	2	4			19,44%	
4	3	7			38,89%	
5	4	15			80,56%	
6	5	7			100,00%	
7	y mayor...	0			100,00%	
8						
9						

A la cel·la C2 copiarem la freqüència absoluta (anotarem [=B2]) i a la cel·la C3 sumarem la freqüència absoluta de la segona modalitat (anotarem [=C2+B3]).

	A	B	C	D	E	F
1	Clase	Frecuencia	Frecuencia acumulada	%	% acumulado	
2	1	3	=B2		8,33%	
3	2	4			19,44%	
4	3	7			38,89%	
5	4	15			80,56%	
6	5	7			100,00%	
7	y mayor...	0			100,00%	
8						
9						

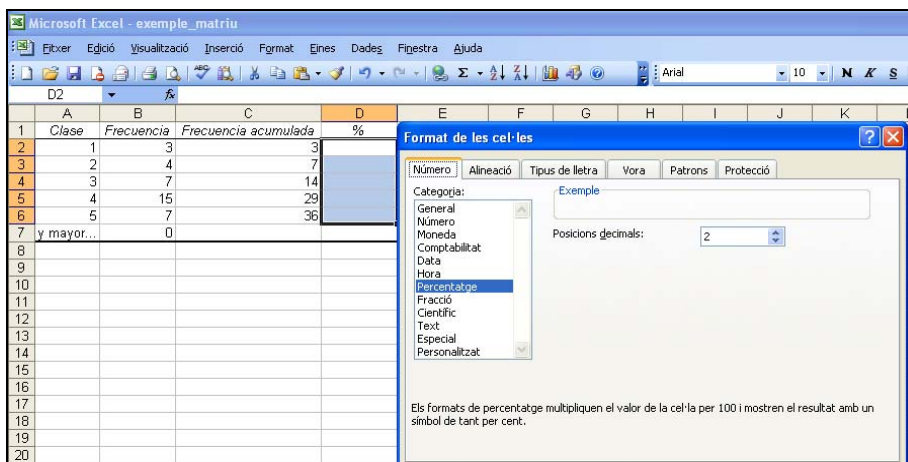
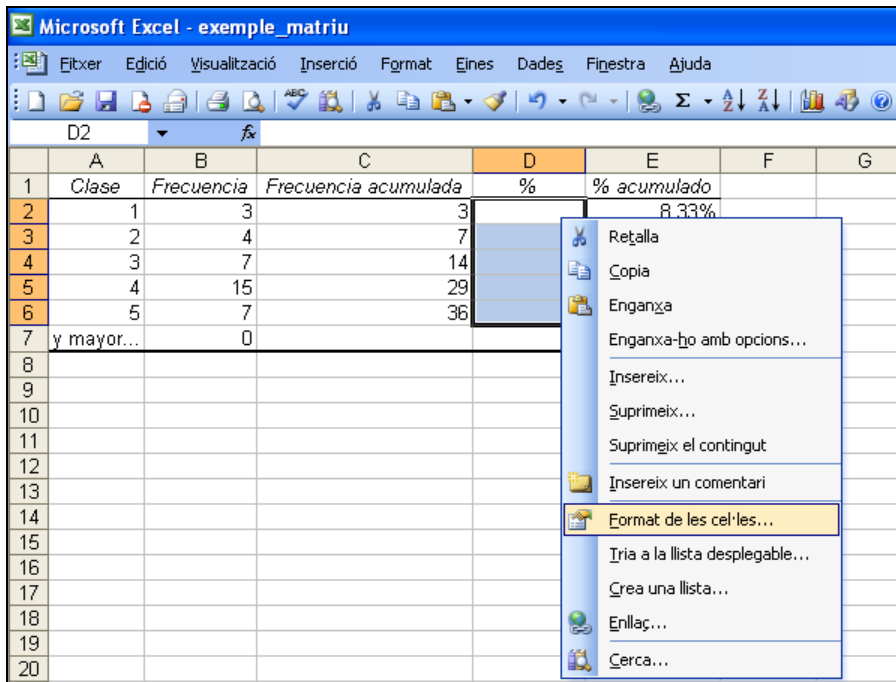
	A	B	C	D	E	F
1	Clase	Frecuencia	Frecuencia acumulada	%	% acumulado	
2	1	3	3		8,33%	
3	2	4	=C2+B3		19,44%	
4	3	7			38,89%	
5	4	15			80,56%	
6	5	7			100,00%	
7	y mayor...	0			100,00%	
8						
9						

Un cop efectuat aquest pas, tan sols caldrà arrossegar amb el ratolí la marca situada a la part inferior dreta de la cel·la i ja tindrem les freqüències acumulades.

	A	B	C	D	E	F
1	Clase	Frecuencia	Frecuencia acumulada	%	% acumulado	
2	1	3	3		8,33%	
3	2	4	7		19,44%	
4	3	7	14		38,89%	
5	4	15	29		80,56%	
6	5	7	36		100,00%	
7	y mayor...	0			100,00%	
8						
9						

Arribats a aquest punt, construirem la freqüència relativa, que s'expressa en percentatge. Per tant, el primer que farem serà definir el format de les cel·les com a percentatge. Això ho podem fer seleccionant les cel·les que ocuparan aquestes freqüències i obrint el menú contextual fent clic amb el botó dret del ratolí. Aquí seleccionarem l'opció [FORMAT DE LES CEL·LES] i, un cop a

la pantalla següent dins la pestanya [NÚMERO], l'opció [PERCENTATGE]. Deixarem les dues posicions decimals que per defecte ofereix l'Excel.



Finalment, completarem la columna (D) de la freqüència relativa, calculant la proporció de B2 respecte al total de casos. Això ho farem mitjançant el càlcul $[=B2/SUMA(b\$2:b\$6)]$. A B2 trobem la freqüència absoluta de la classe (o modalitat) 1 de la variable A4, que dividim per la suma (utilitzant la funció [SUMA] de l'Excel) de les freqüències de totes les modalitats de la variable. És a dir, el total de casos. Cal destacar que afegirem el signe [\$] després de les indicacions de la columna B per tal de fixar l'operació [SUMA].

	A	B	C	D	E	F
1	Clase	Frecuencia	Frecuencia acumulada	%	% acumulado	
2	1	3	3	=B2/ suma(b\$2:b\$6)		
3	2	4	7		19,44%	
4	3	7	14		38,89%	
5	4	15	29		80,56%	
6	5	7	36		100,00%	
7	y mayor...	0			100,00%	
8						
9						

El darrer pas, i de la mateixa manera que ho hem fet amb anterioritat, arrossegarem la marca de la part inferior de la dreta de la cel·la per tal de completar la columna. I ja tindrem la taula de freqüències completa (taula 9):

Taula 9. Taula de freqüències a l'Excel

	A	B	C	D	E	F
1	Clase	Frecuencia	Frecuencia acumulada	%	% acumulado	
2	1	3	3	8,33%	8,33%	
3	2	4	7	11,11%	19,44%	
4	3	7	14	19,44%	38,89%	
5	4	15	29	41,67%	80,56%	
6	5	7	36	19,44%	100,00%	
7	y mayor...	0			100,00%	
8						
9						

5.2. Anàlisi exploratòria bivariant

L'anàlisi exploratòria bivariant, igual que l'univariant, es fa tant amb variables quantitatives com amb variables qualitatives. Ens centrarem en l'anàlisi bivariant amb variables qualitatives i, específicament, aprendrem a fer taules de contingència. Les raons són principalment dues: 1) les taules de contingència s'utilitzen profusament en l'anàlisi de dades; i 2) hem de saber que qualsevol variable quantitativa pot transformar-se en una variable qualitativa.

És important recordar que...

... aquest curs és introductor i, per tant, solament pretenem assolir uns coneixements bàsics d'anàlisi de dades. En aquest sentit, l'anàlisi exploratòria bivariant amb variables quantitatives restarà fora dels nostres objectius.

5.2.1. Transformació de variables quantitatives en variables qualitatives

La transformació de variables qualitatives en variables quantitatives té com a raó principal simplificar la interpretació de les variables, de tal manera que la classificació en categories faciliti la presa de decisions.

El procés de conversió d'una variable quantitativa en una variable qualitativa s'anomena *categorització*.

Aquesta categorització es pot dur a terme seguint criteris que ja s'han establert prèviament i que faciliten una lectura comparativa de les dades. La conversió també es pot fer per raons teòriques de qualsevol tipus, o bé per a facilitar a la persona que fa l'anàlisi de dades la presa de decisions.

Hi ha diverses tècniques de categorització de dades quantitatives, però en aquest curs només en veurem un exemple molt senzill. Per fer-ho, reprendrem l'exemple que hem utilitzat a l'anàlisi exploratòria univariant, i continuarem utilitzant com a instrument el full de càlcul Excel.

El nostre punt de partida l'estableix la distribució de la variable A1, de la qual hem calculat els estadístics descriptius en els apartats anteriors. Un d'ells ha estat la mitjana. Així, la mitjana aritmètica de la variable A1 és ($\bar{x}=5,11$).

Podríem, doncs, categoritzar la variable A1 en una nova variable, que anomenarem *A1_COD* segons si els valors de A1 són més petits o més grans que la mitjana aritmètica. En la taula 10 es pot observar el procediment i el resultat en forma de nova variable qualitativa d'aquesta categorització.

Taula 10. Categorització d'una variable a l'Excel

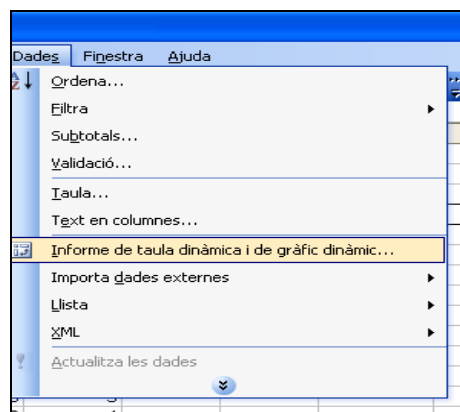
	A	B	C	D	E	F	G	H	I	J
1	ID	A1	A1_cod	A2	A3	A4	A4			
2	46.847	2	Inferior	0	0	4	1			
3	46.626	9	Superior	0	0	2	2			
4	46.715	3	Inferior	0	0	3	3	PROMEDIO		5,11
5	46.757	7	Superior	0	0	1	4			
6	46.665	4	Inferior	0	0	5	5			
7	46.636	0	Inferior	0	0	4				
8	47.259	1	Inferior	0	0	5				
9	46.754	10	Superior	12	2	2				
10	46.507	1	Inferior	0	0	4				
11	46.861	8	Superior	0	0	3				
12	46.849	13	Superior	0	2	2				
13	46.788	12	Superior	0	0	3				
14	46.929	1	Inferior	0	0	4				
15	46.808	1	Inferior	0	0	5				
16	46.959	8	Superior	0	0	5				
17	46.655	5	Inferior	0	0	4				
18	46.645	10	Superior	1	0	2				
19	46.536	0	Inferior	0	0	3				
20	46.532	10	Superior	0	0	3				
21	46.714	1	Inferior	0	0	4				
22	46.886	5	Inferior	1	0	1				
23	46.747	8	Superior	0	0	4				
24	46.726	3	Inferior	0	0	5				
25	46.967	9	Superior	0	0	3				
26	46.914	2	Inferior	0	0	1				
27	46.958	20	Superior	0	0	4				
28	46.698	3	Inferior	0	0	4				
29	46.743	3	Inferior	0	0	5				
30	46.608	4	Inferior	0	0	4				
31	46.821	2	Inferior	0	0	4				
32	46.699	3	Inferior	0	0	4				
33	46.738	1	Inferior	0	0	5				
34	46.518	0	Inferior	0	0	4				
35	46.974	4	Inferior	0	0	3				
36	46.658	1	Inferior	0	0	4				
37	46.748	10	Superior	0	0	4				
38										

5.2.2. Anàlisi exploratòria bivariant amb variables qualitatives

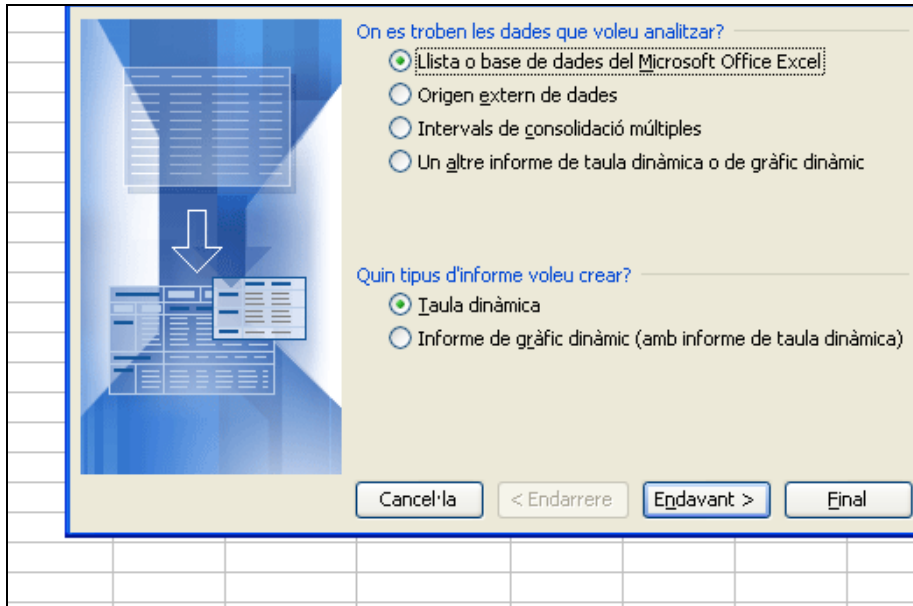
L'anàlisi exploratòria bivariant amb variables qualitatives més habitual es fa mitjançant la construcció d'una taula de contingència. Per a poder treballar amb variables qualitatives a l'Excel necessitem utilitzar les taules dinàmiques.

Per tal de construir una taula dinàmica seguirem les passes següents:

1. Al menú [DADES] seleccionarem l'opció [INFORME DE TAULA DINÀMICA I DE GRÀFIC DINÀMIC]. Ens apareixerà un quadre de diàleg on indicarem la localització de les dades que volem analitzar i el tipus d'informació que volem crear.



2. En el nostre cas seleccionarem les opcions per defecte. És a dir, l'origen de les dades serà una [LLISTA O BASE DE DADES DEL MICROSOFT OFFICE EXCEL] i la informació que volem crear serà una [TAULA DINÀMICA].



3. Seguidament, veurem l'auxiliar de taules i gràfics dinàmics. Amb el ratolí seleccionarem les dades que volem incloure a la taula dinàmica. El tercer pas ens demanarà on volem situar la taula creada. Si l'obviem marcant [FINAL], ens situarà la taula dinàmica en un nou full. Aquesta és l'opció que hem seleccionat.

Microsoft Excel - exemple_matriu

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	ID	A1	A1_cod	A2	A3	A4	A4						
2	46.847	2	Inferior	0	0	4	1						
3	46.626	9	Superior	0	0	2	2						
4	46.715	3	Inferior	0	0	3	3		PROMEDIO	5,11			
5	46.757	7	Superior	0	0	1	4						
6	46.665	4	Inferior	0	0	5	5						
7	46.636	0	Inferior	0	0	4							
8	47.259	1	Inferior	0	0	5							
9	46.754	10	Superior	12	2	2							
10	46.507	1	Inferior	0	0	4							
11	46.861	8	Superior	0	0	3							
12	46.849	13	Superior	0	2	2							
13	46.788	12	Superior	0	0	3							
14	46.929	1	Inferior	0	0	4							
15	46.808	1	Inferior	0	0	5							
16	46.959	8	Superior	0	0	5							
17	46.655	5	Inferior	0	0	4							
18	46.645	10	Superior	1	0	2							
19	46.536	0	Inferior	0	0	3							
20	46.532	10	Superior	0	0	3							
21	46.714	1	Inferior	0	0	4							
22	46.886	5	Inferior	1	0	1							
23	46.747	8	Superior	0	0	4							
24	46.726	3	Inferior	0	0	5							
25	46.967	9	Superior	0	0	3							
26	46.914	2	Inferior	0	0	1							
27	46.958	20	Superior	0	0	4							
28	46.698	3	Inferior	0	0	4							
29	46.743	3	Inferior	0	0	5							
30	46.608	4	Inferior	0	0	4							
31	46.821	2	Inferior	0	0	4							
32	46.699	3	Inferior	0	0	4							
33	46.738	1	Inferior	0	0	5							
34	46.518	0	Inferior	0	0	4							
35	46.974	4	Inferior	0	0	3							
36	46.658	1	Inferior	0	0	4							
37	46.748	10	Superior	0	0	4							
38													

Auxiliar de taules i gràfics dinàmics - pas 2 de 3

On són les dades que voleu fer servir?

Interval: 'Exemple UQL!\$A\$1:\$F\$37'

Buttons: Cancel·la, < Endarrere, Endavant >, Final

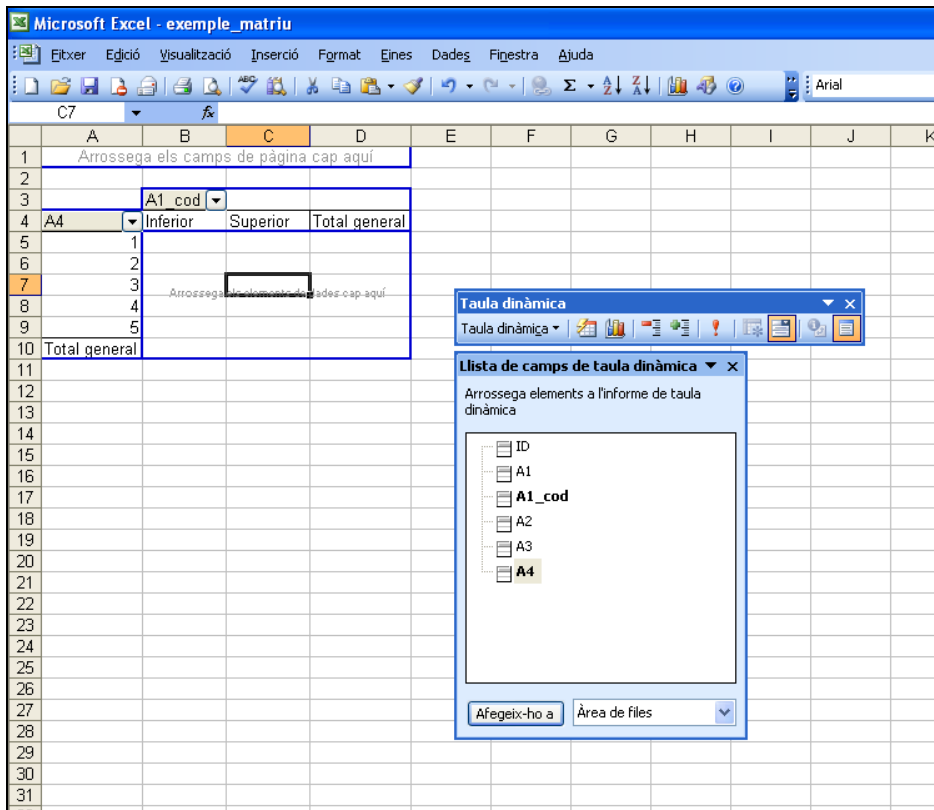
4. Arribats a aquest punt, ens apareixeran en un nou full els quadres de diàleg que ens guiaran per poder construir la taula dinàmica.
5. En primer lloc, hem de seleccionar les variables que la compondran i situar-les a la taula. Arrossegarem una de les variables i la situarem en columna, i després arrossegarem l'altra cap a les files.

En el nostre exemple situem la variable A1_COD, la variable qualitativa que hem categoritzat prèviament, en columna i la variable A4 en fila. D'aquesta manera ens serà fàcil observar com varia A4 en funció de les categories establertes (inferior i superior a la mitjana) d'A1_COD.

Per a facilitar l'anàlisi i la presa de decisions és habitual...

... situar en columnes la teòrica variable independent o bé la variable on volem observar la variació d'una altra segons les categories de la primera.

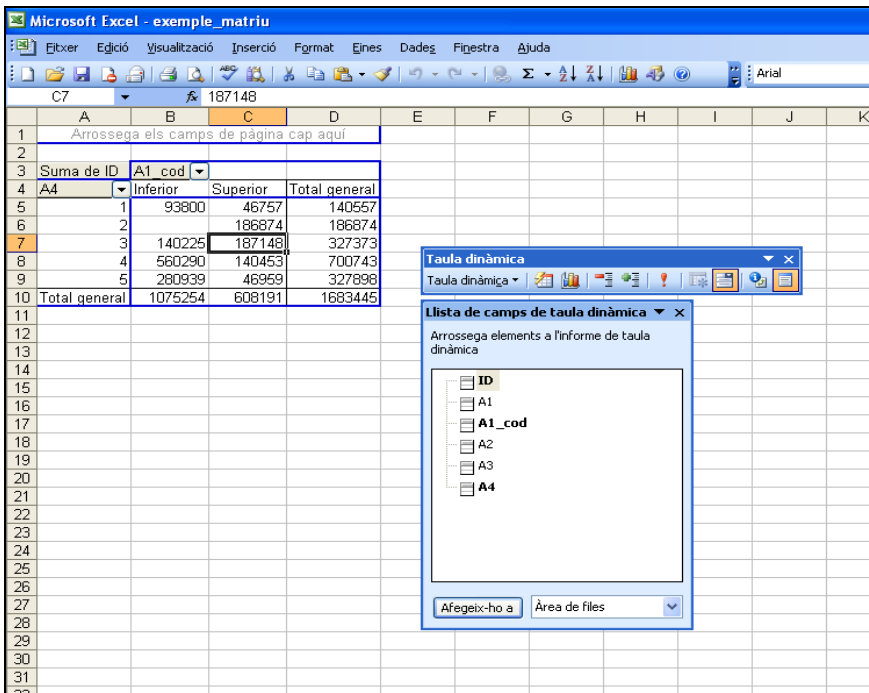
The screenshot shows the Microsoft Excel interface with a dynamic table being created. The table range is A3:A16. The wizard 'Llista de camps de taula dinàmica' is open, showing a list of fields: ID, A1, A1_cod, A2, A3, and A4. The 'A1_cod' field is selected for the column and 'A4' for the row. The 'Taula dinàmica' task pane is also visible.



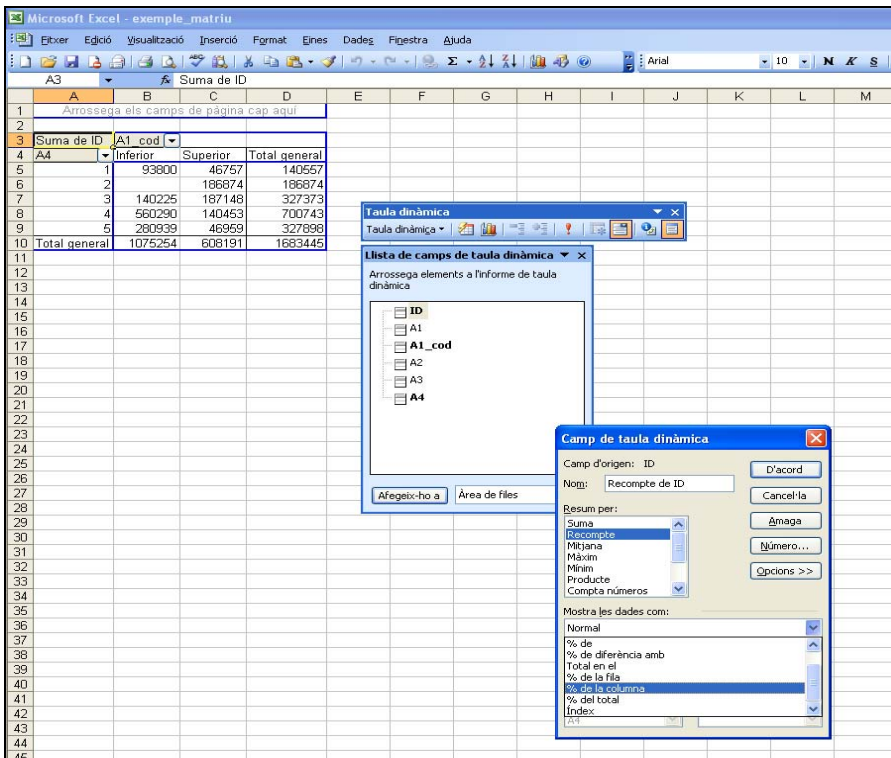
6. Per tal de completar la taula de contingència, hem de dir a l'ordinador què és el que volem fer amb aquestes variables. És a dir, quina informació en volem treure. La lectura de dades que volem obtenir l'hem d'indicar a la taula dinàmica.

En el nostre exemple, i com es pot observar a la figura següent, veiem com hem arrossegat la variable ID a la cel·la A3. Aquesta variable s'anomena *identificador* i és una variable que identifica tots i cadascun dels diferents registres (files). Per tal de poder identificar el registre, cada valor (modalitat) de la variable ha de ser diferent. Operar amb una variable identificador també és molt útil per a fer diferents operacions. A la pràctica comptar identificadors és el mateix que comptar individus (registres).

Per defecte, l'Excel dona [SUMA de ID] com a resum de la variable. Però en aquest cas, el que volem és veure la distribució dels individus en una variable en funció de les categories d'una altra. Així, la suma dels valors dels identificadors dels individus no és una ajuda, al contrari. No els hem de sumar, els hem de comptar.



7. Amb l'objectiu de comptar els individus farem doble clic a la cel·la A3 i s'obrirà un quadre de diàleg anomenat [CAMP DE TAULA DINÀMICA]. En aquest quadre seleccionarem un resum per [RECOMPTE]. Seguidament, farem clic a [OPCION], que és on podrem seleccionar la manera com volem que ens siguin mostrades les dades. Com que el que volem és veure la distribució d'A4 segons A1_COD, el que farem és escollir el [% DE LA COLUMNA].



8. Finalment, obtenim la nostra taula de contingència (taula 11). La lectura que hem de fer és vertical, ja que el que apareix són els percentatges de la columna.

Si observem detingudament les dades, podem observar que la distribució de la variable A4 és força diferent per als individus amb valors d'A1 inferior a la mitjana que per als individus que tenen valors d'A1 superior a la mitjana. Els individus amb valors petits a A1 tendeixen més a presentar valors més grans a A4.

La informació que hem obtingut mitjançant el processament de dades, i que està present a la taula de contingència, ens pot portar a pensar que hi ha una relació entre les respostes de les variables tractades.

Raonar aquesta informació, tot completant-la amb anàlisis estadístiques de més profunditat, molt possiblement optimitzarà qualsevol presa de decisions, més enllà del coneixement que hàgim generat en l'anàlisi de la taula de contingències.

A títol pràctic

El treball amb taules (i gràfics) dinàmiques amb l'Excel és molt senzill i intuïtiu.

Les proves per assaig i error directament relacionades amb les taules, i l'auxiliar de taules dinàmiques, són una excel·lent manera de comprendre ràpidament les possibilitats d'aquesta eina.

Taula 11. Taula de contingències a l'Excel

	A	B	C	D	E	F
1						
2						
3	Recompte de ID	A1_cod				
4	A4	Inferior	Superior	Total general		
5	1	8,70%	7,69%	8,33%		
6	2	0,00%	30,77%	11,11%		
7	3	13,04%	30,77%	19,44%		
8	4	52,17%	23,08%	41,67%		
9	5	26,09%	7,69%	19,44%		
10	Total general	100,00%	100,00%	100,00%		
11						
12						
13						
14						

Resum

En aquest curs introductori sobre l'anàlisi de dades hem treballat conceptes estadístics bàsics que ens permeten tenir una primera visió de com les dades poden donar suport a la presa de decisions a l'empresa. L'anàlisi de dades és un gran instrument per a la presa de decisions, i aquesta és imprescindible per al bon funcionament de l'empresa en el mercat.

En primer lloc, hem repassat alguns aspectes estadístics clau. Hem diferenciat entre *dades*, *informació* i *coneixement*, conceptes que es refereixen a realitats molt diferents i que, massa habitualment, s'utilitzen de manera indistinta. Hem conegut les preguntes bàsiques que ens hem de plantejar a l'hora d'iniciar una recerca suportada en dades. I, finalment, també hem treballat què són les dades i com es poden diferenciar.

A partir d'aquí, i com que aquest curs es basa en la metodologia quantitativa d'anàlisi de dades, hem conegut què és una variable i quins són els diferents tipus. Hem treballat competències sobre com diferenciar entre població i mostra, i a fer les anàlisis univariants i bivariants senzilles més habituals, tant pel que fa a les variables quantitatives com a les variables qualitatives.

En conjunt, al llarg d'aquest curs hem fet una revisió introductòria a l'estadística descriptiva, que ha de permetre, una vegada conclosa la comprensió dels materials didàctics i fetes les activitats, tenir una primera visió de com orientar les respostes a les necessitats d'informació de l'empresa, i conèixer uns primers instruments per a fer anàlisi estadística bàsica aplicada a la presa de decisions.

Glossari

anàlisi univariant *f* Anàlisi que mesura una de les característiques dels individus d'una població.

anàlisi bivariant *f* Anàlisi que mesura una de les característiques d'un conjunt d'individus en funció d'una altra característica.

desviació estàndard *f* Arrel quadrada de la variància.

desviació mitjana *f* Mitjana de les diferències entre els valors de la variable i la mitjana aritmètica en valor absolut.

freqüència absoluta *f* Nombre d'individus que comparteixen la mateixa modalitat de la variable.

freqüència absoluta acumulada *f* Nombre d'individus que comparteixen una modalitat o alguna de menor.

freqüència relativa *f* Proporció d'individus que comparteix la mateixa modalitat de la variable.

freqüència relativa acumulada *f* Proporció d'individus que comparteixen una modalitat o alguna de menor.

mediana (Me) *f* Valor que divideix tota una sèrie de dades en dues parts exactament iguals.

mitjana aritmètica *f* Valor resultant de dividir el sumatori d'un conjunt de dades pel nombre total de dades. Quan la mitjana aritmètica es refereix a una població s'utilitza el símbol (μ) i quan es refereix a una mostra s'utilitza \bar{x} .

moda (Mo) *f* Valor que més vegades es repeteix, és a dir, aquell que té la freqüència més elevada.

mostra *f* Subconjunt d'una població que en representa les característiques principals a escala.

població *f* Conjunt d'individus que disposen de les característiques que necessitem observar.

rang (R) *m* Diferència entre el valor més gran i el valor més petit d'una distribució.

taula de contingència *f* Taula per a enregistrar i analitzar la relació entre dues o més variables, habitualment de naturalesa qualitativa nominal o ordinal.

valor d'una variable *m* Mesura que una variable presenta.

variable *f* Reducció d'una dimensió d'una realitat.

variable dependent *f* Variable que suposem que varia en funció d'una altra variable.

variable dicotòmica *f* Cas específic (i molt habitual) de variable qualitativa nominal. Aquesta variable solament pren dos valors i s'utilitza habitualment per a mostrar la presència o l'absència d'una característica.

variable independent *f* Variable que funciona com a *input* del sistema la variació de la qual incideix en la variació d'altres variables dependents.

variable qualitativa *f* Variable que representa una categoria.

variable qualitativa ordinal *f* Variable qualitativa que presenta un ordre establert dins una escala.

variable qualitativa nominal *f* Variable qualitativa que representa categories sense relació de mesura entre elles.

variable quantitativa *f* Variable que representa una xifra.

variable quantitativa contínua *f* Variable que no representa cap interrupció més enllà de les que imposa l'instrument de mesura.

variable quantitativa discreta *f* Variable que presenta una interrupció entre els valors, perquè els valors intermedis no existeixen.

variació d'una variable *f* Diferència entre els valors d'una variable.

variància (S^2) *f* Resultat de la divisió entre el sumatori de les distàncies entre cada dada i la mitjana aritmètica de les dades, elevades al quadrat.

Bibliografia

Levin, J. (1997). *Fundamentos de estadística en la investigación social*. Mèxic, DF: Oxford University Press.

Montero Lorenzo, J. M. (2009). *Problemas resueltos de estadística descriptiva para ciencias sociales*. Madrid: Paraninfo.

Otamendi, F. J.; Diaz Chao, A. (2011). *Estadística para emprendedores: Lecciones prácticas y casos con Minitab*. Madrid: Addlink Media.