ESTADÍSTICA SANITÀRIA

Mètodes. Mitjanes i variacions. Tendències i correlacions. Mostres estadístiques. Gràfiques

Óscar Zurriaga

Departament de Medicina Preventiva i Salut Pública, Ciències de l'Alimentació, Toxicologia i Medicina Legal. Universitat de València

Estadística

És la ciència de la

Descripció

- **sistematització, recollida, ordenació i presentació** de les **dades** referents a un fenomen que presenta variabilitat o incertesa per a estudiar-les metòdicament a fi de

Probabilitat

- **deduir les lleis** que regeixen aquests fenòmens

Inferència

- i poder així fer previsions sobre aquests fenòmens, prendre **decisions** o arribar a **conclusions**.

Passos en un estudi estadístic

- **Plantejar hipòtesis sobre una *població***
  - **Els fumadors tenen *més baixes* laborals que els no fumadors.**
  - En quin sentit? Més nombre? Temps mitjà?

- **Decidir quines dades cal recollir (disseny d'experiments)**
  - Quins individus formaran part de l'estudi (*mostres*)
    - Fumadors i no fumadors en edat laboral.
    - Criteris d'exclusió. Com es trien? Es descarten els qui pateixen malalties cròniques?
  - Quines dades cal recollir dels individus (*variables*)
    - Nombre de baixes.
    - Temps de duració de cada baixa.
    - Sexe? Sector laboral? Altres factors?

- **Recollir les dades (*mostreig* o *mostratge*)**
  - Estratificat? Sistemàticament?

- **Descriure (resumir) les dades obtingudes**
  - Temps mitjà de baixa en fumadors i no fumadors (*estadístics*).
  - % de baixes per fumadors i sexe (*freqüències*), gràfics...

- **Efectuar una inferència sobre la població**
  - Els fumadors estan de baixa almenys 10 dies/any més *de mitjana* que els no fumadors.

- **Quantificar la confiança en la inferència**
  - Nivell de confiança del 95%
  - Significació del contrast: p=2%

Variables

- Una **variable** és una característica observable que varia entre els diversos individus d'una població. La informació que tenim de cada individu és resumida en **variables**.

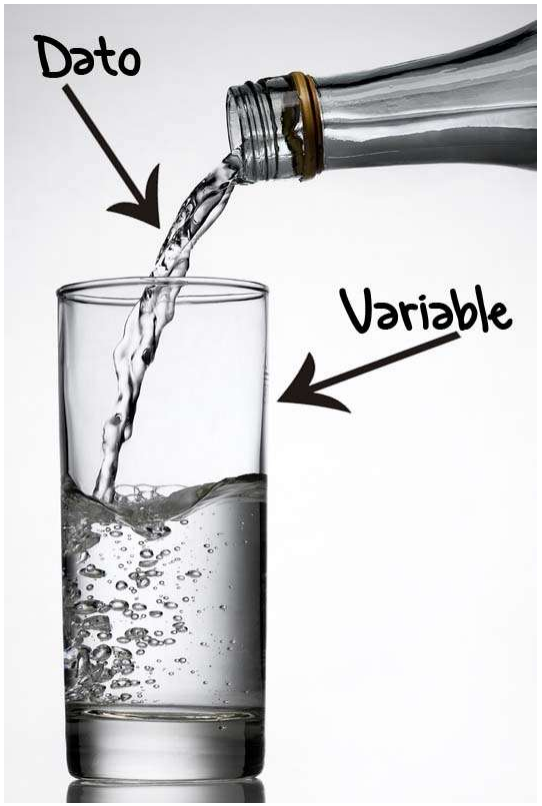
- En els individus de la *població*, de l'un a l'altre ***és variable***:

{A, B, AB, O} ← var. qualitativa (grup sanguini)

{Deprimit, indiferent, molt feliç} ← var. ordinal

{0, 1, 2, 3...} ← var. numèrica discreta

{1,62; 1,74...} ← Var. numèrica contínua

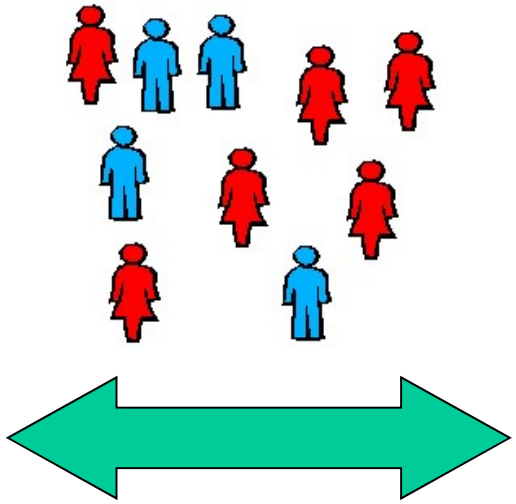


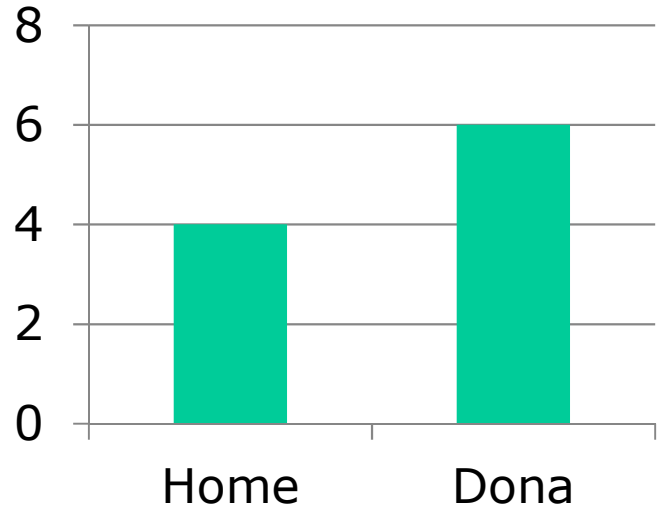
Tipus de variables

- Qualitatives
  Si els valors (*modalitats*) no es poden associar naturalment a un nombre (no es poden fer operacions algebraiques amb aquests valors).
  - Nominals: si els valors no es poden ordenar.
    - Sexe, grup sanguini, religió, nacionalitat, fumar (sí/no).
  - Ordinals: si els valors es poden ordenar.
    - Milloria a un tractament, grau de satisfacció, intensitat del dolor.
- Quantitatives o numèriques
  Si els valors són numèrics (té sentit fer operacions algebraiques amb aquests valors).
  - Discretes: si pren valors enters.
    - Nombre de fills, nombre de cigarrets, nombre d'aniversaris.
  - Contínues: si entre dos valors són possibles infinits valors intermedis.
    - Alçada, pressió intraocular, dosis de medicament administrat, edat.

Presentació ordenada de dades

Sexe	Freq.
Home	4
Dona	6





- Les taules de freqüència i les representacions gràfiques són dues maneres ***equivalents*** de presentar la informació. Totes dues exposen ordenadament la informació recollida en una mostra.

Taules de freqüència

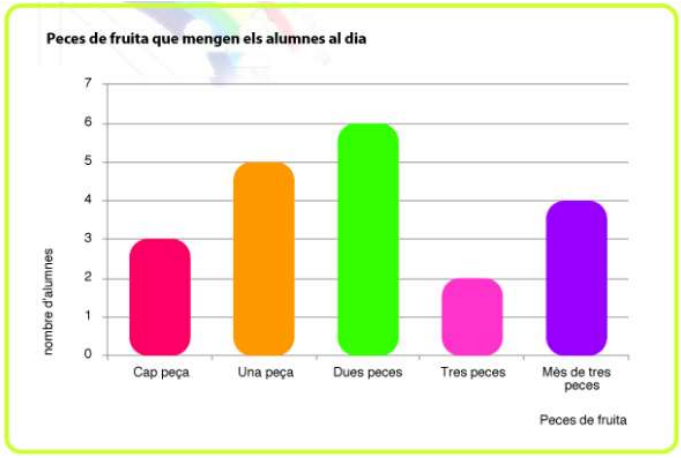
- Exposen la informació recollida en la mostra de manera que no es perda informació.
  - Freqüències absolutes: comptabilitzen el nombre d'individus de cada modalitat.
  - Freqüències relatives (percentatges): ídem, però dividit pel total.
  - Freqüències acumulades: només tenen sentit per a variables ordinals i numèriques
    - Molt útils per a calcular quantils
      - Quin percentatge d'individus té menys de 3 fills? Solució: 83,8%
      - Entre 4 i 6 fills? Solució 1a: 8,4%+3,6%+1,6% = **13,6%** Solució 2a: 97,3% - 83,8% = **13,5%**

Número de hijos

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	419	27,6	27,8	27,8
	1	255	16,8	16,9	44,7
	2	375	24,7	24,9	69,5
	3	215	14,2	14,2	83,8
	4	127	8,4	8,4	92,2
	5	54	3,6	3,6	95,8
	6	24	1,6	1,6	97,3
	7	23	1,5	1,5	98,9
	Ocho o más	17	1,1	1,1	100,0
	Total	1509	99,5	100,0	
Perdidos	No contesta	8	,5		
Total		1517	100,0		

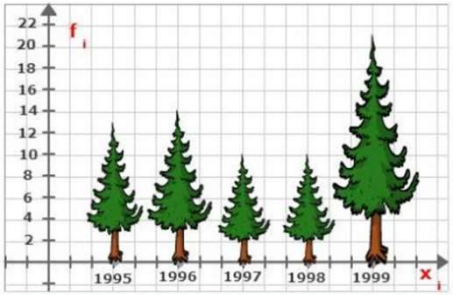
Gràfiques per a variables qualitatives

- Diagrames de barres
  - Altures proporcionals a les freqüències (absolutes o relatives).
  - Es poden aplicar també a variables discretes.

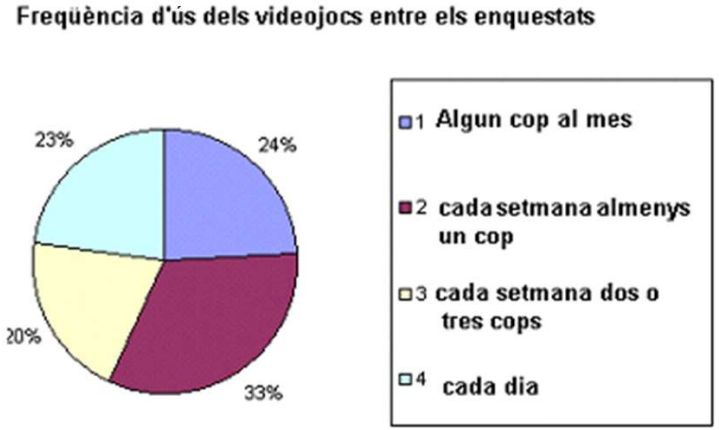


- Diagrames de sectors (pastissos)
  - No s'han d'usar amb variables ordinals.
  - L'àrea de cada sector és proporcional a la freqüència (absolutes o relatives).

- Pictogrames
  - Fàcils d'entendre.
  - L'àrea de cada modalitat ha de ser proporcional a la freqüència.

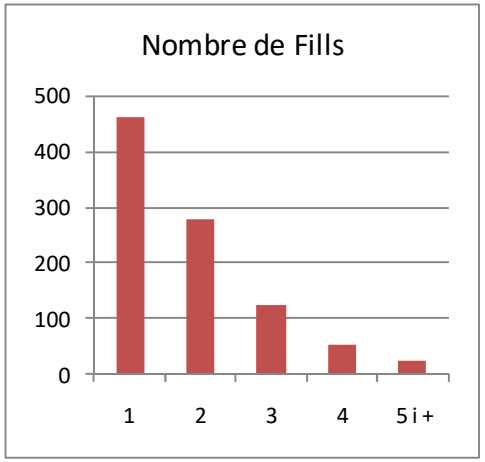


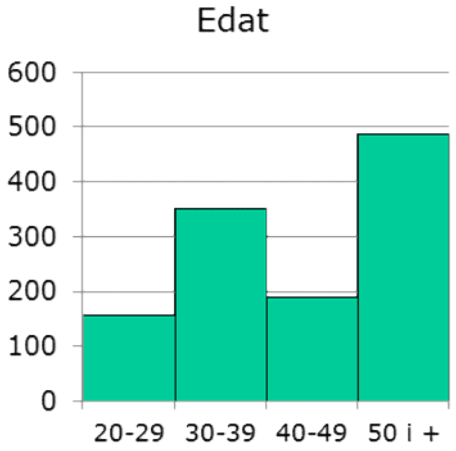
Superfície repoblada d'una comarca (en milers d'hectàrees durant el període 1995-1999.



Gràfiques diferencials per a variables numèriques

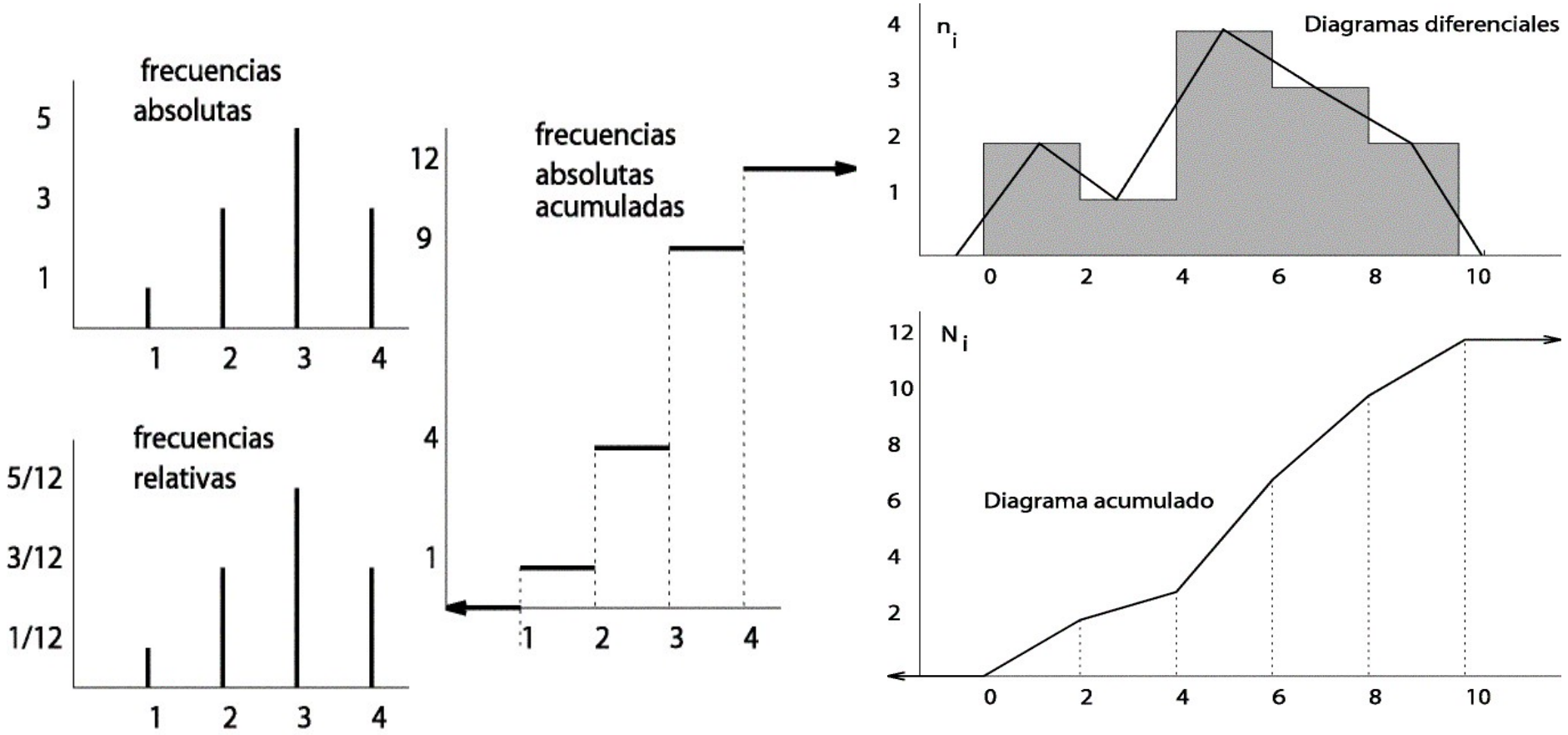
- Són diferents segons que les variables siguen discretes o contínues. Serveixen amb freqüències absolutes o relatives.
  - **Diagrames de barres per a variables discretes**
    - Amb un espai entre barres (indiquen valors que no són possibles).
  - **Histogrames per a variables contínues**
    - L'àrea que hi ha sota l'histograma entre dos punts indica la quantitat (percentatge o freqüència) d'individus en l'interval.





Diagrames integrals

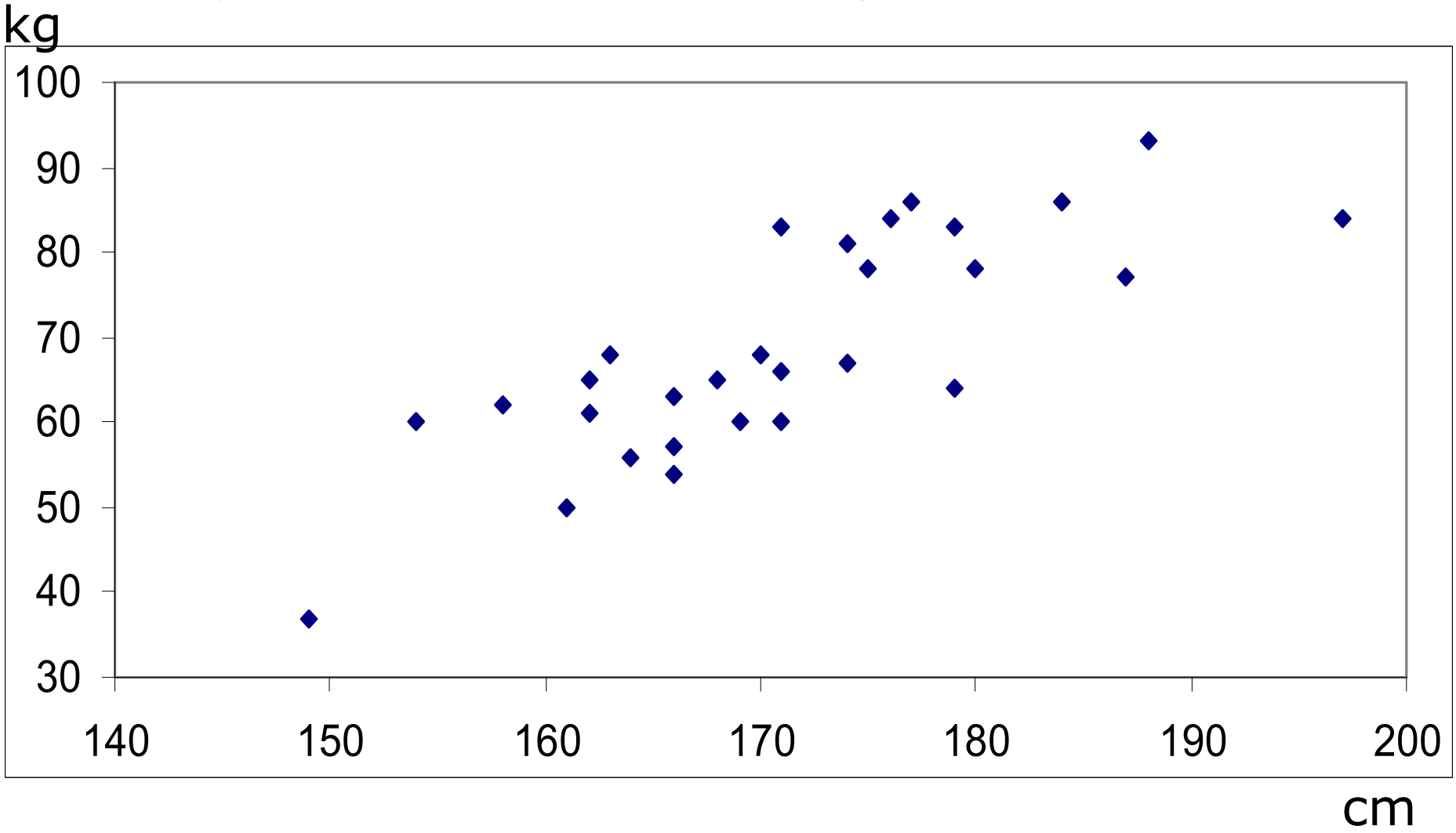
- Cadascun dels diagrames anteriors té el diagrama integral corresponent. Es tracen a partir de les freqüències acumulades. Indiquen, per a cada valor de la variable, la quantitat (freqüència) d'individus que tenen un valor inferior o igual.



Tendències i correlacions

Diagrames de dispersió o núvol de punts

Alçada i pes de 30 individus: diagrama de dispersió



Coeficient de correlació lineal de Pearson

- El coeficient de correlació lineal de Pearson de dues variables, ***r***, indica si els punts tenen tendència a disposar-se alineadament (excepte rectes horitzontals i verticals).

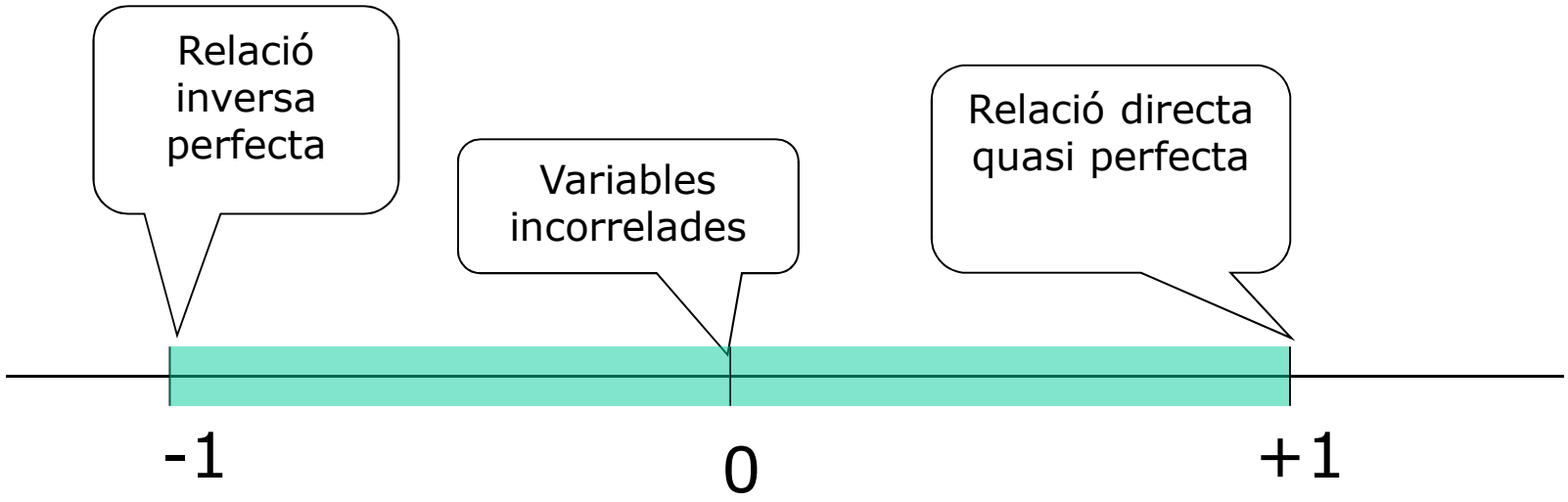
- El signe indica si la possible relació és directa o inversa.

- *r* és útil per a determinar si hi ha relació lineal entre dues variables, però no és útil per a altres tipus de relacions (quadràtica, logarítmica...).

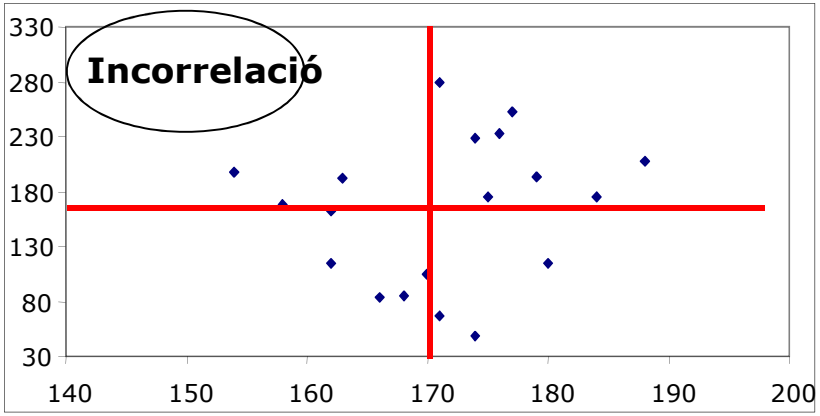
Propietats de *r*

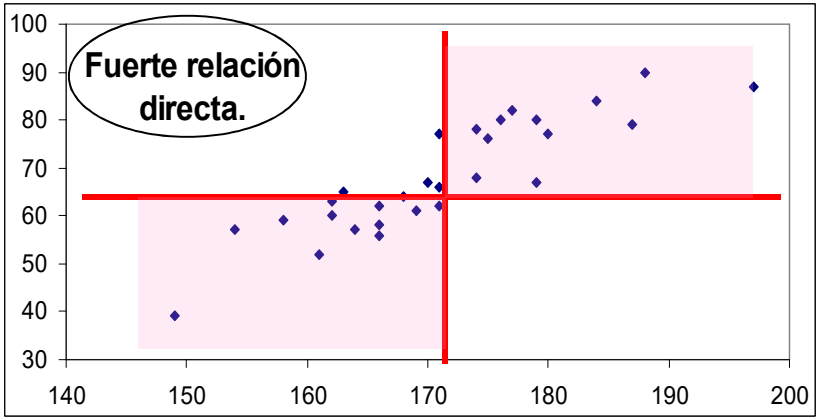
- És adimensional.
- Només pren valors en [-1,1].
- Les variables són incorrelades ⇔ $r=0$
- Relació lineal perfecta entre dues variables ⇔ $r=+1$ o $r=-1$
  - N'excloem els casos de punts alineats horitzontalment o verticalment.
- Com més a prop estiga *r* de +1 o -1, millor serà el grau de relació lineal.

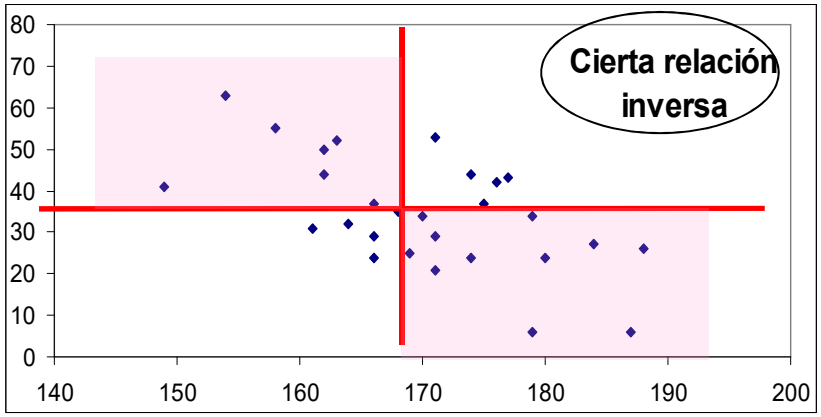
  (sempre que no hi haja observacions anòmales)



Relació directa i inversa








Correlació
no implica causalitat

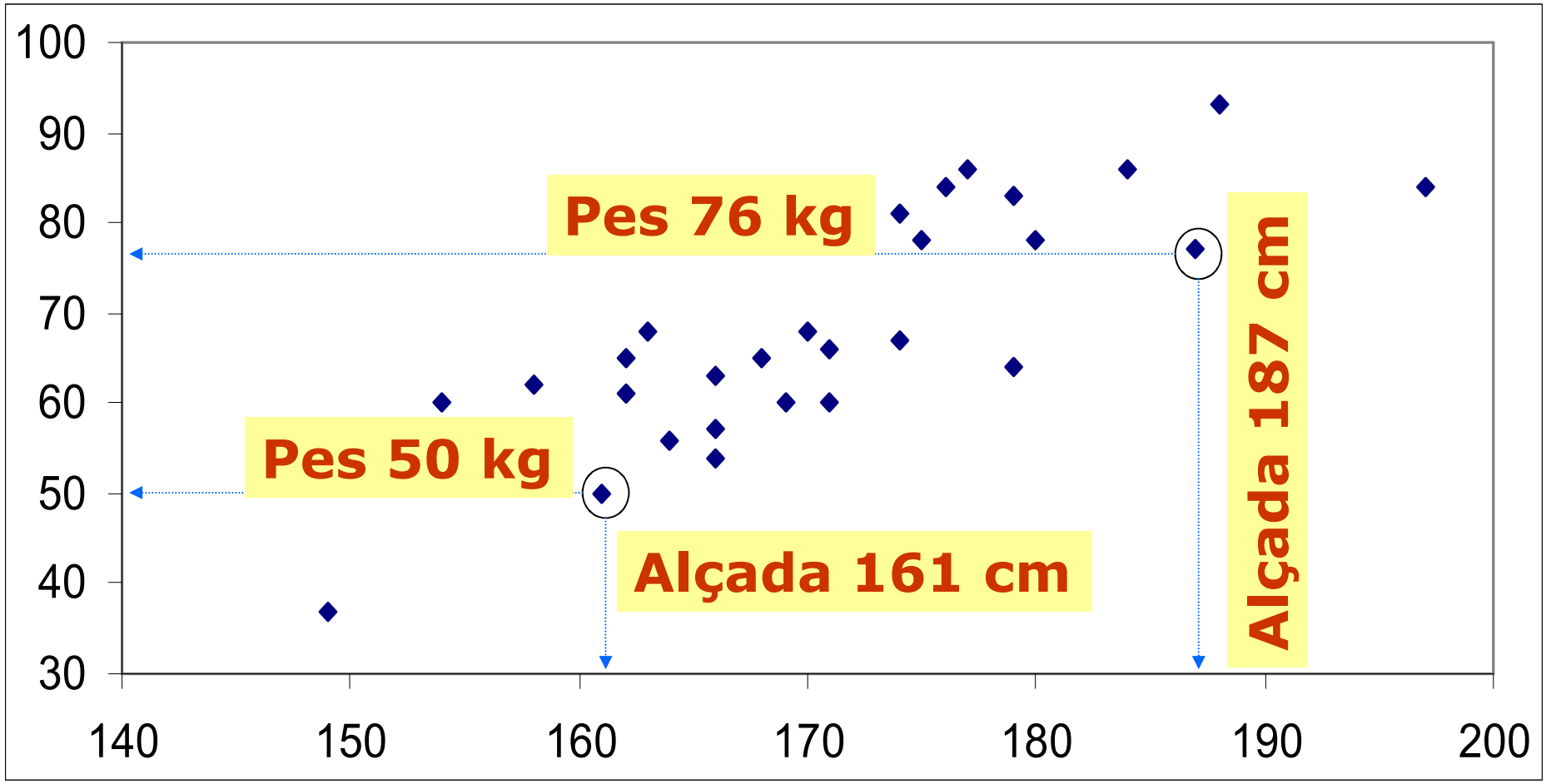
Regressió

- L'anàlisi de regressió serveix per a predir una mesura en funció d'una altra mesura (o d'altres mesures).
  - Y = variable dependent
    - predita
    - explicada
  - X = variable independent
    - predictora
    - explicativa
  - És possible descobrir una relació?
    - $Y = f(X)$ + error
      - f és una funció d'un tipus determinat
      - L'error és aleatori, menut, i no depèn de X

Tendències i correlacions

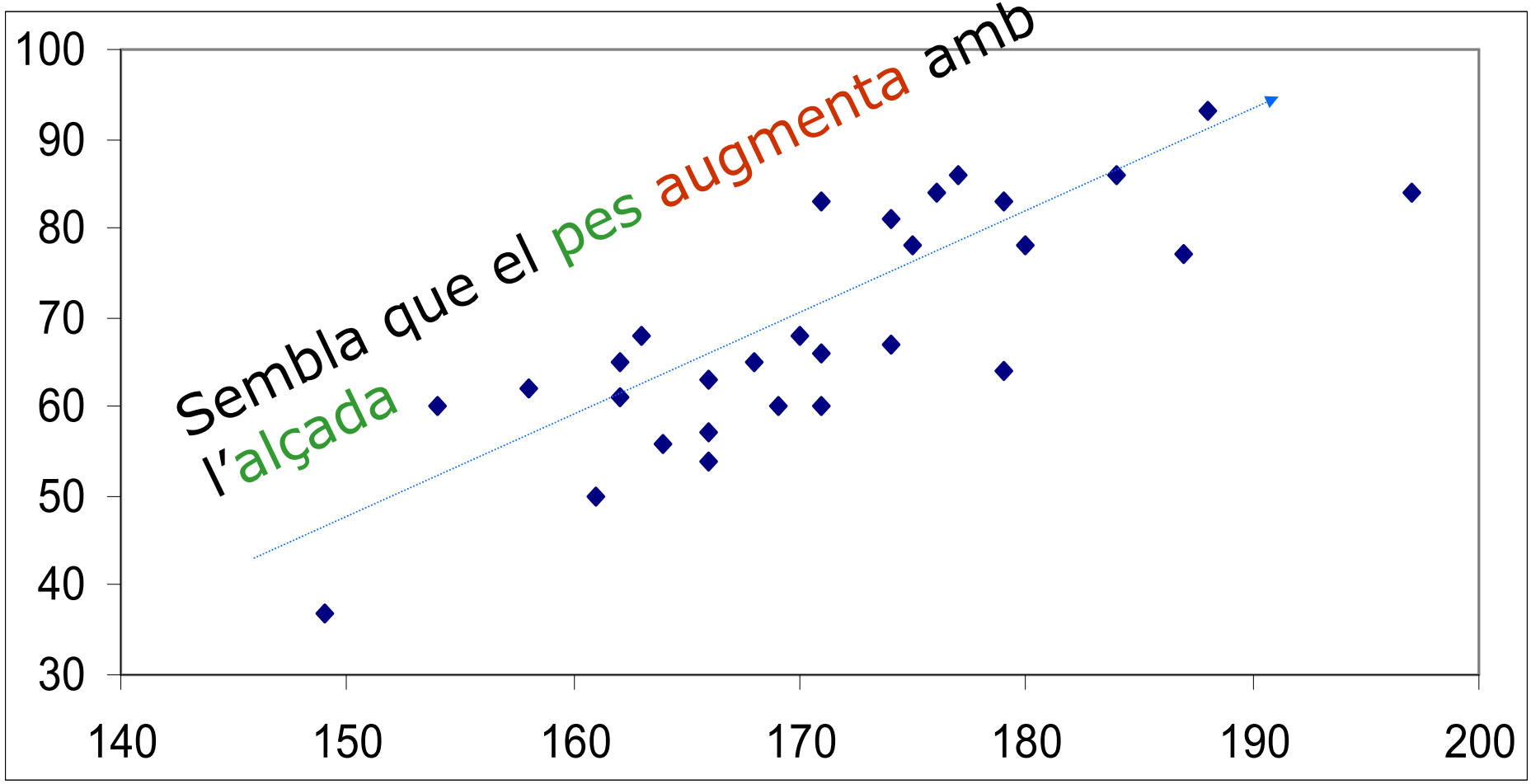
Diagrames de dispersió o núvols de punts

Alçada i pes de 30 individus: diagrama de dispersió



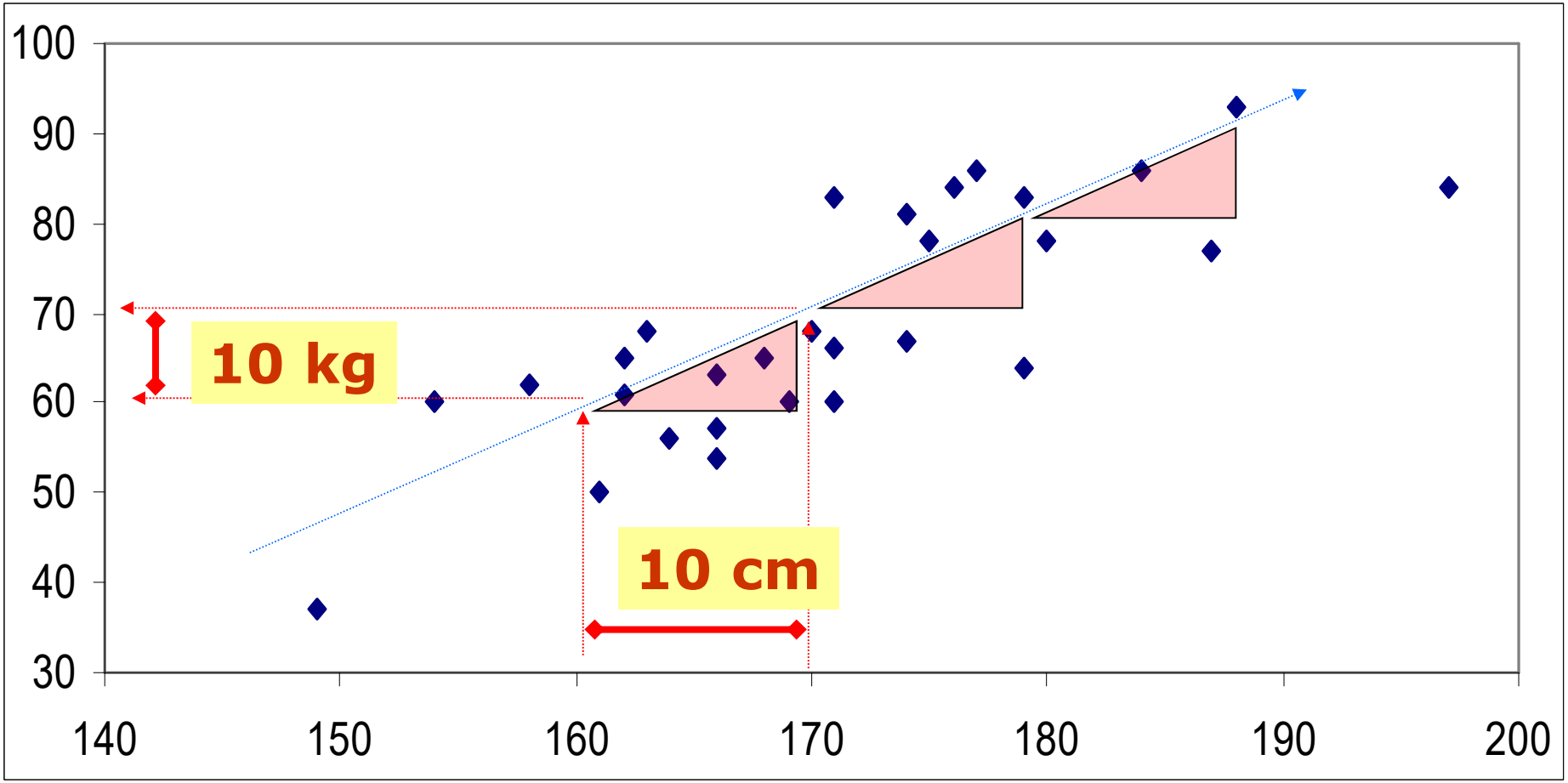
Relació entre variables

Alçada i pes de 30 individus: diagrama de dispersió

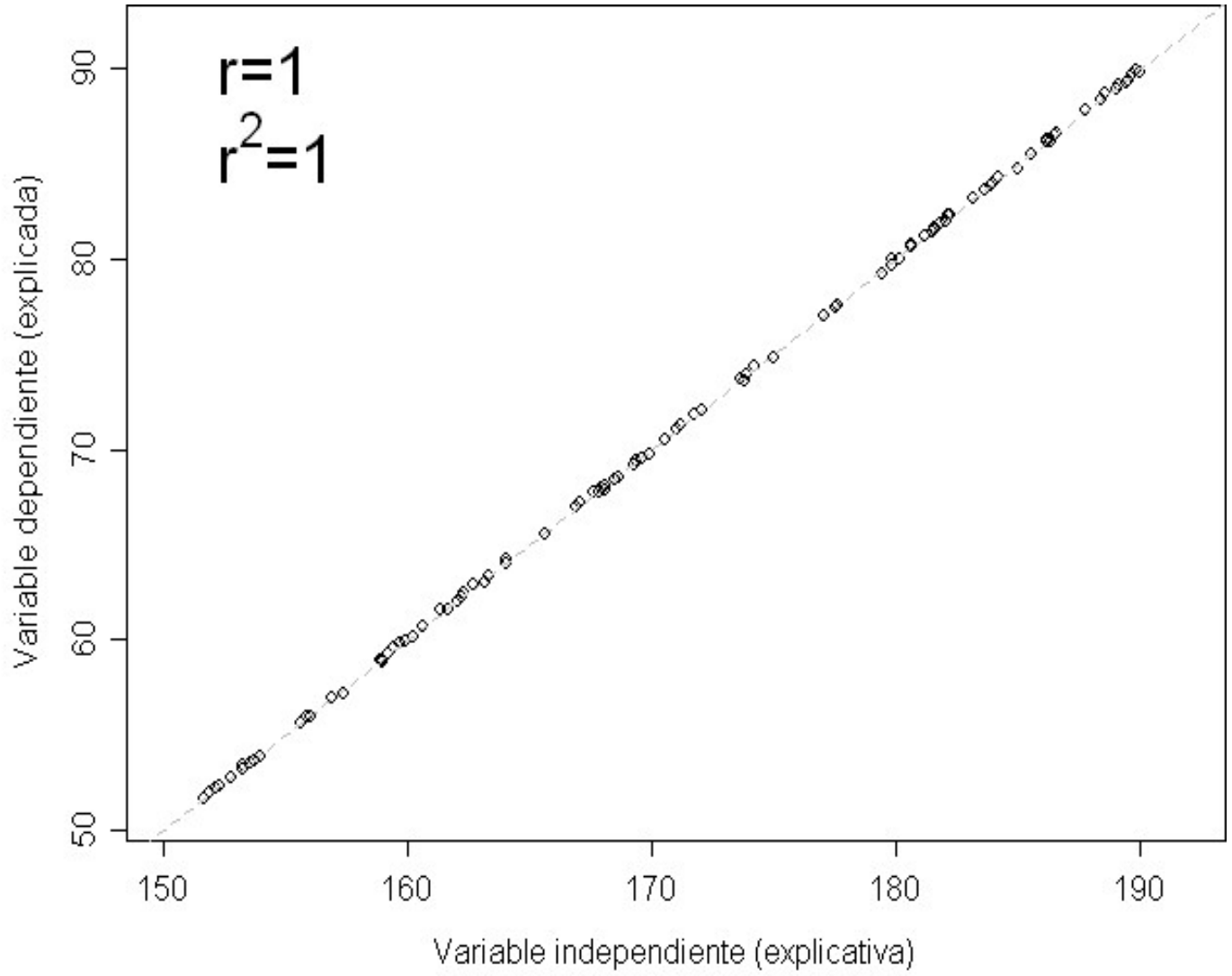


Predicció d'una variable en funció d'una altra

Aparentment el pes augmenta 10 kg per cada 10 cm d'alçada...
És a dir,
el pes augmenta una unitat per cada unitat d'alçada.



Animació: diagrama de dispersió i evolució de *r*



Població i mostra

- **Població**

És el conjunt sobre el qual estem interessats a obtenir conclusions (inferir).

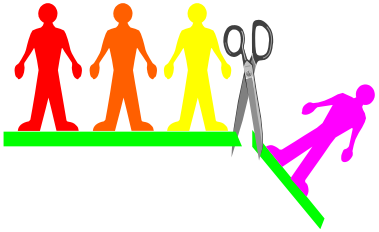
  - Normalment és massa gran per a poder abastar-lo.

- **Mostra** (*sample*)

És un subconjunt de la població al qual tenim accés i sobre el qual realment fem les observacions (mesuraments).

Hauria de ser *representatiu*

  - Està format per membres *seleccionats* de la població (individus, unitats experimentals).



Mostres i mostreig

- Les poblacions estan formades per individus, però seria millor denominar-les unitats de mostreig o unitats d'estudi:
  - Persones, famílies, hospitals, països…

- La població ideal que es vol estudiar es denomina població objectiu.
  - No és fàcil estudiar-la per complet. Ens hi aproximem mitjançant mostres que donen idealment la mateixa probabilitat a cada individu de ser elegit.
  - Tampoc no és fàcil triar mostres de la població objectiu:
    - Si usem el telèfon, excloem de la mostra els qui no en tenen.
    - Si triem indiv. pel carrer, oblidem els qui estan treballant...

- El grup que en realitat podem estudiar (*v. g.*, els qui tenen telèfon) es denomina població d'estudi.

Fonts de biaix

- Les poblacions objectiu i d'estudi poden diferir quant a les variables que estudiem.
  - El nivell econòmic en la població d'estudi és més alt que en la població objectiu...
  - Els individus que es trien al carrer poden ser de més edat (més freqüència de jubilats, per ex.)…
  - En aquest cas, es diu que les mostres que es trien estaran esbiaixades (biaix de selecció).
- Hi ha altres fonts d'error/biaix
  - No respondre a enquestes compromeses
    - Consum de drogues, violència de gènere, pràctiques poc ètiques…
  - Mentir en les preguntes *delicades*.

Tècniques de mostreig

- Mostrejos probabilístics
  - Es coneix la probabilitat que un individu siga triat per a la mostra.
  - Menys probabilitat de biaix.
  - S'hi pot usar estadística matemàtica.

- Mostrejos no probabilístics
  - La probabilitat no es coneix.
  - Són mostrejos que segurament amaguen biaixos.
  - En principi, els resultats no es poden extrapolar a la població.
    - Tot i així, una bona part dels estudis que es publiquen fan servir aquesta tècnica.

- Mostrejos probabilístics
  - Aleatori simple
  - Sistemàtic
  - Estratificat
  - Per grups

Mostreig aleatori simple (m. a. s.)

- Es trien individus de la població d'estudi de manera que tots tenen la mateixa probabilitat de ser escollits, fins a assolir la grandària mostral desitjada.

- Es pot fer partint de llistes d'individus de la població i escollint individus aleatòriament (nombres aleatoris).

- L'aplicació d'aquest mostreig normalment té un cost bastant alt.

- En general, les tècniques d'inferència estadística suposen que la mostra s'ha creat usant m. a. s.

Mostreig sistemàtic

- Tenim una llista dels individus de la població d'estudi. Si volem una mostra d'aquesta població d'una grandària determinada, es trien individus igualment espaiats de la llista, en què el primer és escollit a l'atzar.
- ATENCIÓ: si a la llista hi periodicitats, la mostra estarà esbiaixada.
  - Un cas real: es va triar una de cada cinc cases per a un estudi de salut pública en una ciutat on les cases es distribueixen en illes de cinc cases. Van sortir amb molta freqüència les de les cantonades, que reben més sol, estan més ben ventilades...

Mostreig estratificat

- S'aplica quan sabem que hi ha certs factors (variables, subpoblacions o estrats) que poden influir en l'estudi i volem assegurar-nos que hi ha una quantitat mínima d'individus de cada tipus:
  - Homes i dones,
  - Joves, adults i ancians…

- Es fa un m. a. s. dels individus de cada estrat.

- Quan els resultats s'extrapolen a la població, cal tenir en compte la grandària relativa de l'estrat respecte al total de la població.

Mostreig per grups o conglomerats

- Quan és difícil tenir una llista de tots els individus que formen part de la població d'estudi, però sí que sabem que es troben agrupats naturalment en grups.

- Es trien diversos d'aquests grups a l'atzar i, una vegada triats alguns, podem estudiar tots els individus dels grups escollits o bé continuar aplicant dins seu més mostrejos per grups, per estrats, aleatoris simples...
  - Per a saber l'opinió dels metges del SNS, podem triar diverses regions d'Espanya, dins d'aquestes, diverses comarques, i dins de les comarques, diversos centres de salut, i...

- Quan els resultats s'extrapolen a la població, cal tenir en compte la grandària relativa d'uns grups respecte dels altres.
  - Regions amb diferent població poden tenir probabilitats diferents de ser escollides, comarques, hospitals grans davant de més menuts...

Estimació

- Estimador

  És una quantitat numèrica calculada sobre una mostra i que esperem que siga una bona aproximació d'una certa quantitat amb el mateix significat en la població (paràmetre).

- Poden ser
  - Puntuals
  - Per interval de confiança

Estimació puntual i per intervals

- Es denomina estimació puntual d'un paràmetre l'oferit per l'estimador sobre una mostra.

- Es denomina estimació confidencial o interval de confiança per a un nivell de confiança 1-α donat, un interval que ha sigut construït de tal manera que, amb freqüència 1-α, realment conté el paràmetre.

  - La probabilitat d'error (no contenir el paràmetre) és α.
    - Probabilitat d'error de tipus I o nivell de significació.
    - Valors típics: α=0,10; **0,05**; 0,01

  - En general, la grandària de l'interval disminueix amb la grandària mostral i augmenta amb 1-α.

  - En tot interval de confiança hi ha una notícia bona i una de roïna:
    - La bona: hem fet servir una tècnica que en un alt percentatge de casos sí que l'encerta.
    - La roïna: no sabem si l'ha encertada en el nostre cas.

- Exemple: una mostra de $n = 100$ individus d'una població té una mitjana de pes de 60 kg i una desviació de 5 kg.
  - Aquestes quantitats poden considerar-se aproximacions (estimacions puntuals)
    - 60 kg estima μ
    - 5 kg estima σ
    - 5/arrel(n)= 0,5 estima l'error estàndard (típic) EE
      - Aquestes són les anomenades estimacions puntuals: un nombre concret calculat sobre una mostra és aproximació d'un paràmetre.
  - Una estimació per interval de confiança ofereix un interval com a resposta. A més, podem assignar-li una probabilitat aproximada que mesure la nostra confiança en la resposta:
    - Hi ha una confiança del 68% que μ estiga en $60 \pm 0{,}5$
    - Hi ha una confiança del 95% que μ estiga en 60 ± 1


HE HECHO UN ESTUDIO ESTADÍSTICO QUE MUESTRA LA BAJADA EN EL PIB REAL DEL PAÍS DURANTE EL MUNDIAL DE FÚTBOL
¡LA CORRELACIÓN ES ASOMBROSA! CREO QUE LO PRESENTARÉ AL FMI
¿POR QUÉ HAY DÍAS EN QUE NO HAY DATOS?
ESTABA VIENDO EL PARTIDO