

PENSAMENT ESTADÍSTIC

Anàlisi de dades via SPSS i introducció a alternatives de software lliure com Jamovi i JASP

Martí Casals. 2022

Introduïm l'Estadística: Preparats per recordar alguns conceptes?

Presentació estudiants

- Què i on has après l'Estadística fins ara?
- Per què t'has apuntat al curs?
- Utilitzes l'estadística normalment? I a quin camp?
Quines necessitats tens?
- Quin *software* utilitzes normalment?
- Què esperes aprendre?

Programa del curs

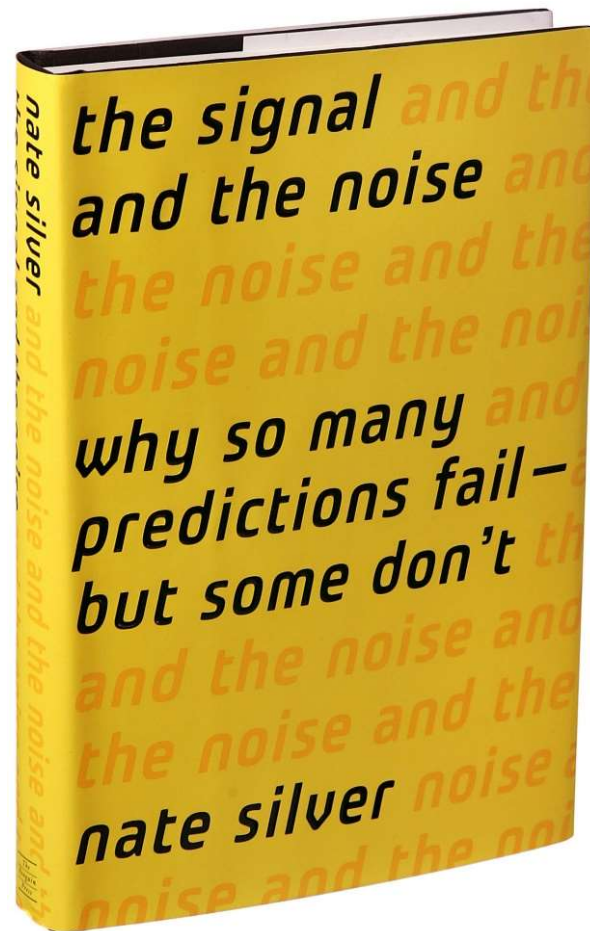
- Dia 1: Introduir-se en el món de l'Estadística i funcions del *Wrangling Data via SPSS*
- Dia 2: L'estadística descriptiva i les seves possibilitats.
- Dia 3: Alguns conceptes d'Inferència Estadística
- Dia 4: Altres programes estadístics: JASP i Jamovi.

Objectius del curs

- Repassar i fer una pinzellada de l'estadística que heu après fins ara
- Recordar conceptes clau d'estadística, abusos, errors estadístics i les maneres de resoldre'ls a l'hora de publicar articles científics
- Conèixer els diferents programaris estadístics (SPSS, JASP, Jamovi) per poder analitzar dades reals.
- Introduir el pensament estadístic

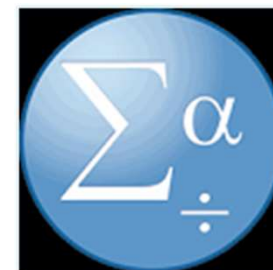
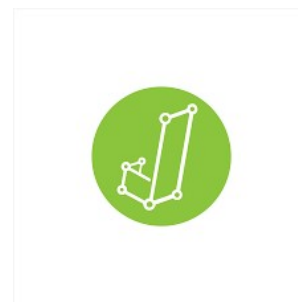
Estadística només números?

- The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning <Nate Silver; The signal and the noise>



Motivació del curs

STAT'S DON'T LIE



L' Estadística

Origen: recopilació i registre de dades i fets d'interès, sobretot per part dels estats (dades tipus demogràfiques, socials, econòmiques, vitals, etc).

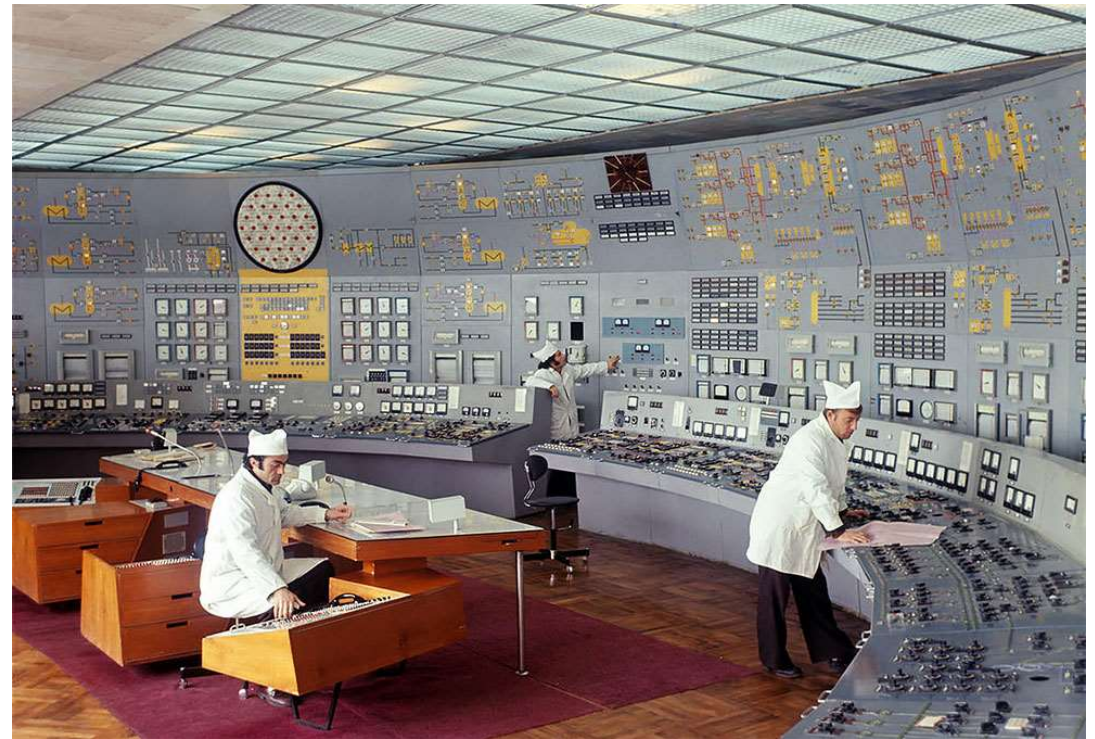


2 conceptes clau: les **DADES** (des del seu disseny per a l'extracció fins les conclusions a partir d'elles), però també la **INCERTESA**!



Tenim suficient amb les dades?

DATA (ALONE) = NOISE
DATA + CONTEXT = INFORMATION
EXPERIMENTATION + ERROR = EXPERIENCE
INFORMATION + EXPERIENCE = KNOWLEDGE
KNOWLEDGE + HUMILITY = WISDOM



“Without data, you're just another person with an opinion.” <W. Edwards Deming>

Diferents definicions: Estadística

- Ciència que utilitza les dades per entendre la incertesa sobre un procés i prendre decisions informades fent ús de la teoria de la probabilitat
<Anabel Forte 2019>
- Ciència on es recull, organitza, resumeix, presenta i analitza les dades (no només utilitzar-los) i situacions d'incertesa, així com extreure conclusions vàlides i prendre decisions raonables en base a aquests anàlisis" -*M. R. Spiegel, Theory and Problems of Statistics*

Especialitzacions: Bioestadística



Statistics applied to biomedical problems



Decision making in the face of uncertainty or variability



Design and analysis of experiments, detective work in observational studies (in epidemiology, outcomes research, etc.)



Attempt to remove bias or find alternative explanations to those posited by researchers with vested interests

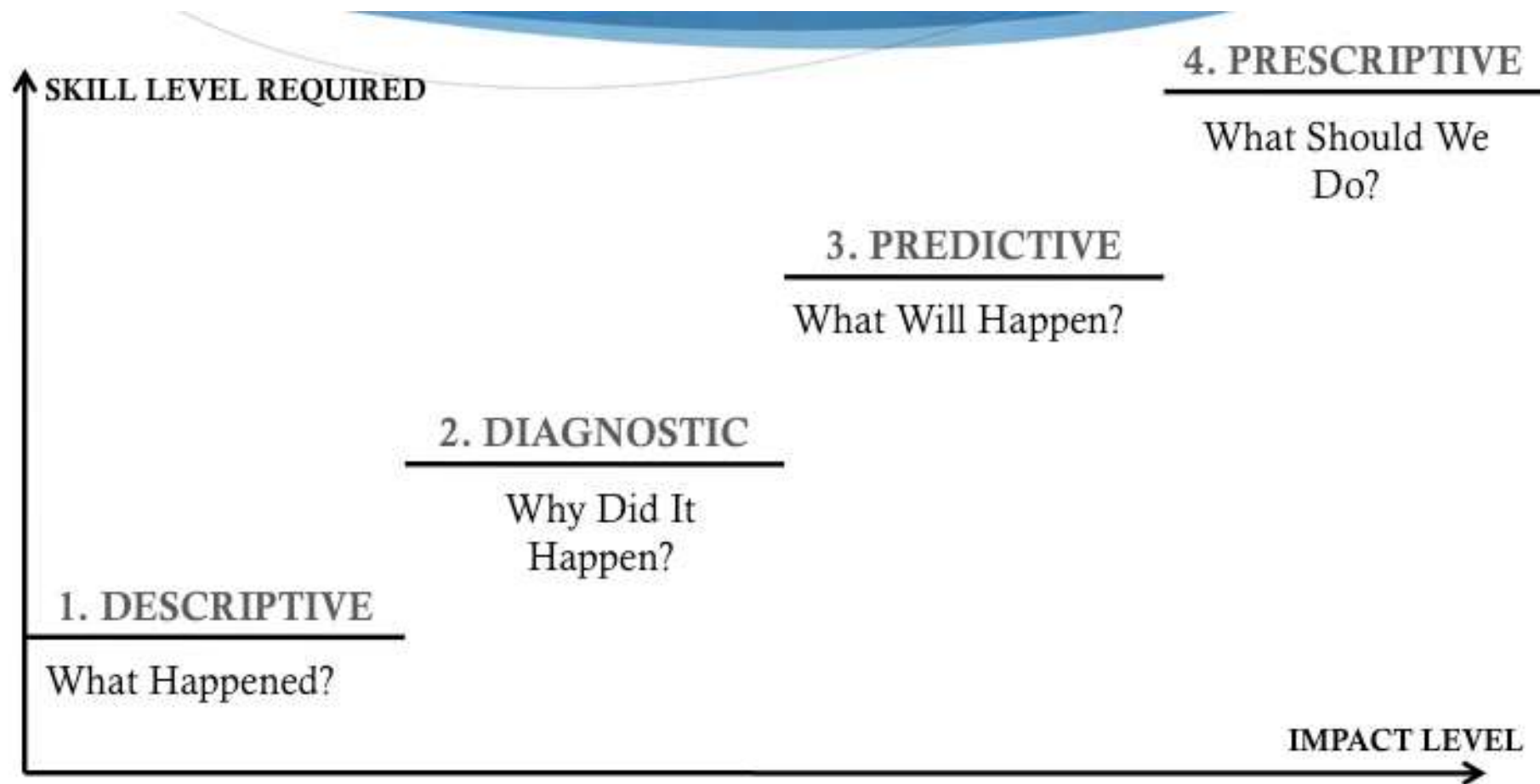


Experimental design, measurement, description, statistical graphics, data analysis, inference

To optimize its value, biostatistics needs to be fully integrated into biomedical research and we must recognize that experimental design and execution are all important. <[Frank E Harrell 2019](#)>



Objectius i procediments de l'estadística



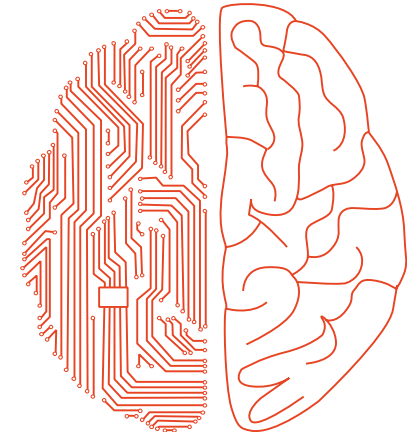
Objectius i procediments de l'estadística

<i>Goal</i>		<i>Statistics</i>
Descript		Estimation <i>Summary measures, Graphics, Confidence Intervals</i>
Compare		Hypothesis testing
Relate or predict		Modelling or Regression

Schooling, C. Mary, and Heidi E. Jones. "Clarifying questions about "risk factors": predictors versus explanation." *Emerging themes in epidemiology* 15.1 (2018): 1-6.

Statistical thinking and statistical science

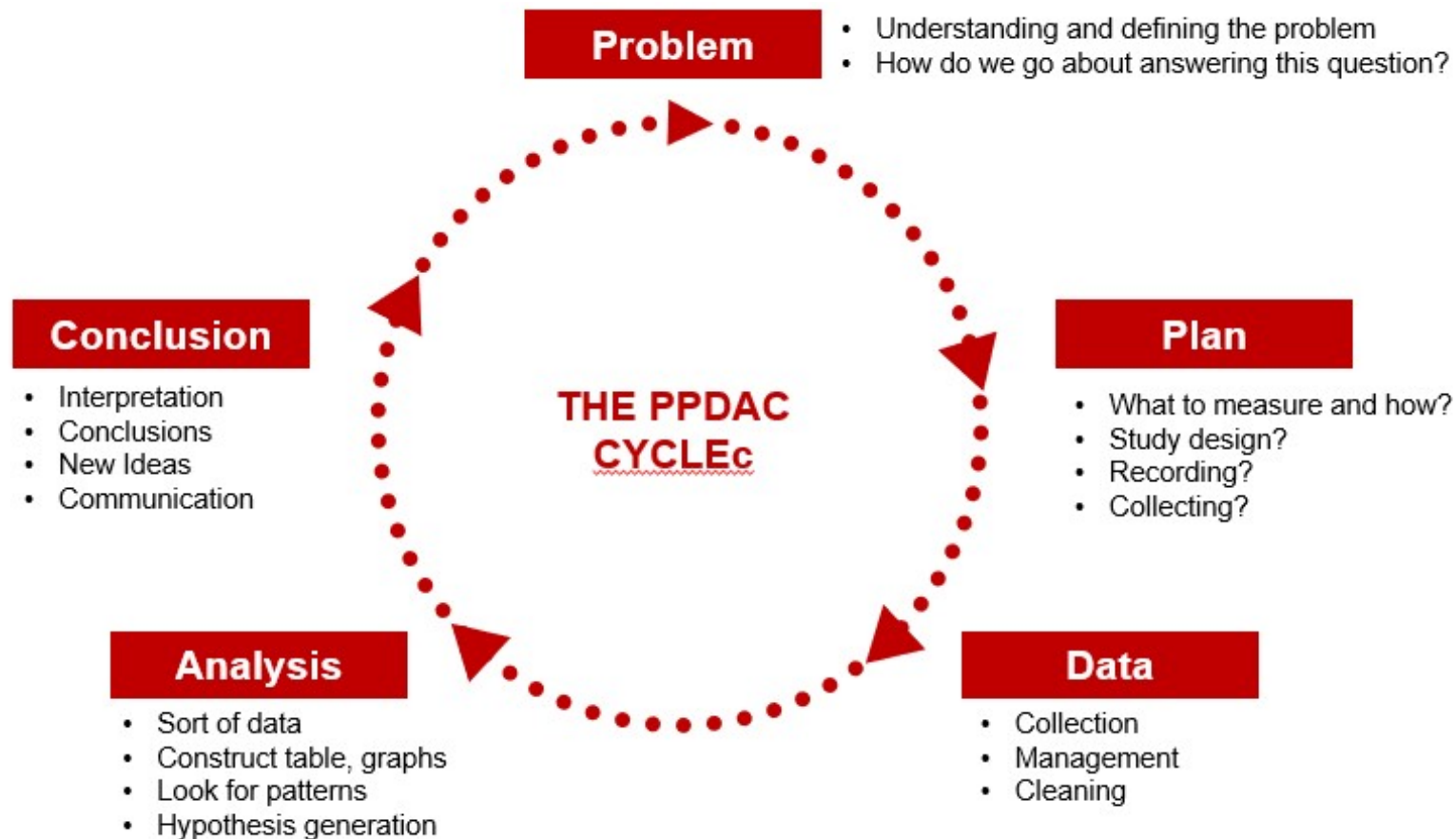
Statistical thinking is a way of understanding a complex world by describing it in relatively simple terms that nonetheless capture essential aspects of its structure, and that also provide us with some idea of **how uncertain we are about our knowledge**.



Statistical Science is the discipline of learning about the world from data, typically involving a **problem-solving cycle** such as **PPDAC**. <David Spiegelhalter>



Application of PPDAC cycle: A new scientific method applied also in the industry



Source: adapted from David Spiegelhalter

Conceptes Generals d'Estadística

Population: It represents the largest group of individuals who want to study and generally usually inaccessible.

Sample: Subset of the population in which measurements are done. This sample should be representative of the original population (any individual has equal opportunity to be selected).

Variable: Feature measurable and observable that represents a concept of study

Measure: Procedure for assign quantitative or qualitative values to the characteristics of objects, people or events. If these procedures are not well measured the validity of the results is not guaranteed.

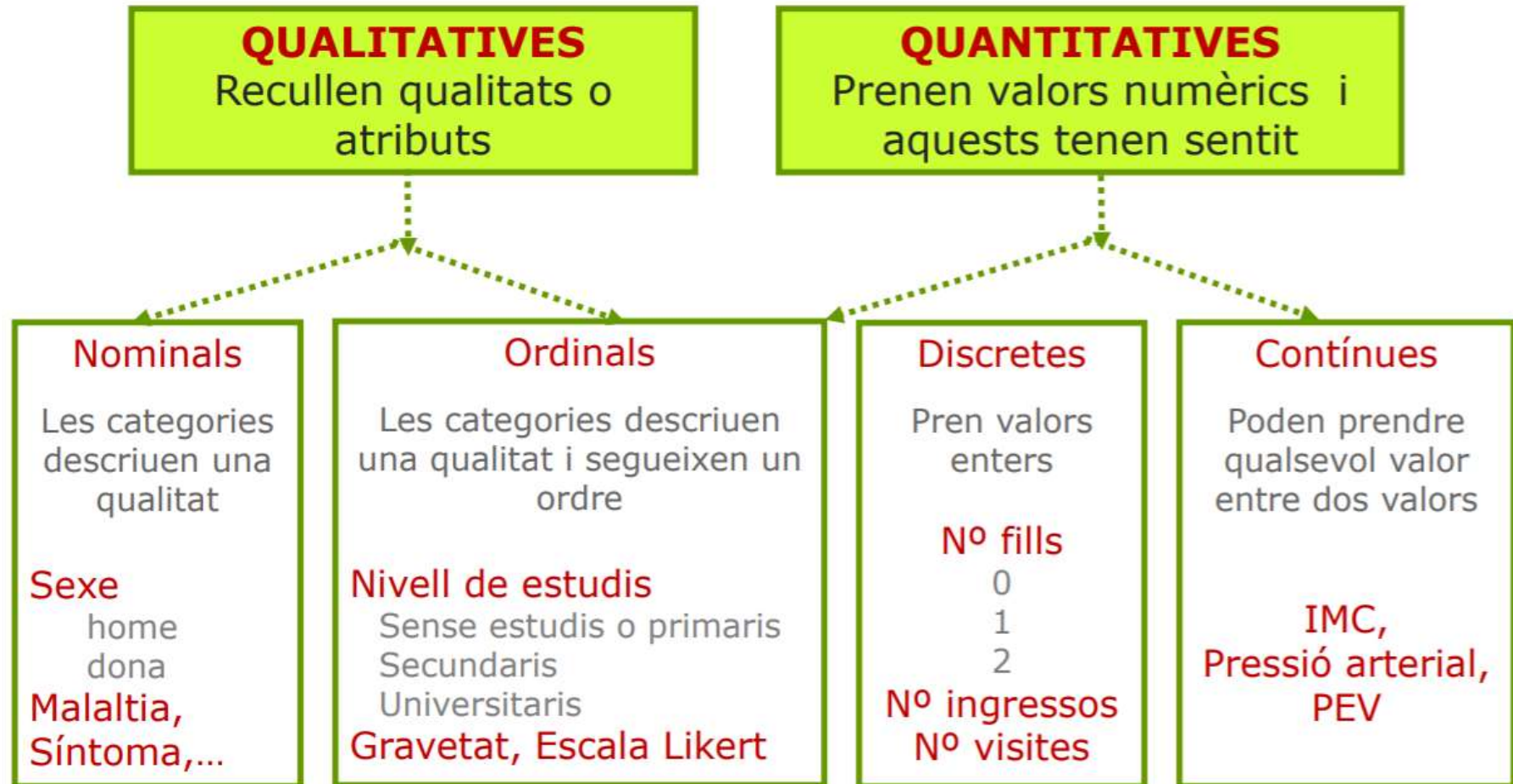
Mostreig: Mostra representativa

Sampling: a sample should have the same characteristics as the population it is **representing**.

Methods:

- **simple random sample** (each sample of the same size has an equal chance of being selected)
- **stratified sample** (divide the population into groups called strata and then take a sample from each stratum)
- **cluster sample** (divide the population into strata and then randomly select some of the strata. All the members from these strata are in the cluster sample.)
- **systematic sample** (randomly select a starting point and take every n-th piece of data from a listing of the population)

Tipus de variables



Tipus variables segons nivell de medició

VARIABLES	ESCALA	CARACTERÍSTIQUES	EXEMPLES
Qualitatives	Nominal	Categories no numèriques sense cap relació d'ordre.	Estat civil, Sexe, Grup sanguini, Marca, Color
Qualitatives	Ordinal	Els valors són categories no numèriques amb una relació d'ordre.	Nivell d'Estudis Classe social Categoria (Hotel)
Quantitatives	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”
Quantitatives	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”

Tipus variables segons la classificació en un estudi

Response, dependent or outcome variable:

- One that answer the research question

Explain, independent or exposure variables:

- They are those that are related to the causes of the events we want to study

Confounding or effect modifier variables:

- Are those that can affect the relation between exposure and outcome variables

Universal variables:

- Are those that can be exposures or confounders that always have to be considered. For example: sex, age, residence location, ethnic, etc.

PENSAMENT ESTADÍSTIC

Anàlisi de dades via SPSS i introducció a alternatives de software lliure com Jamovi i JASP

Introduïm l'SPSS: Preparats per coneixel més a fons?

Tenim ben organitzades les Dades?

THE AMERICAN STATISTICIAN
2018, VOL. 72, NO. 1, 2–10
<https://doi.org/10.1080/00031305.2017.1375989>



 OPEN ACCESS  Check for updates

Data Organization in Spreadsheets

Karl W. Broman^a and Kara H. Woo^b

^aDepartment of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI; ^bInformation School, University of Washington, Seattle, WA

ABSTRACT

Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this article offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, do not leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row), create a data dictionary, do not include calculations in the raw data files, do not use font color or highlighting as data, choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text files.

ARTICLE HISTORY

Received June 2017
Revised August 2017

KEYWORDS

Data management; Data organization; Microsoft Excel; Spreadsheets

Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. *The American Statistician*, 72(1), 2-10.

Archivo Editar Ver Datos Transformar Analizar Marketing directo Gráficos Utilidades Ventana Ayuda													
1: CODIGOJUGADOR 52													Visible: 14 de
	Temporada	CODIGOJUGADOR	Pos	Altura	RAZA	edad	Injury	num_lesions	Anatomical_site	Patologia	Severity	RTP_lesio	TL
1	1993-1994	52	Pivot	206	B	18	Sí	2	Neck	Lesiones ligamentosa...	minimal (1-3 days)	2	Si
2	1993-1994	24	Alero	205	B	31	Sí	2	Ankle	Muscular_Tendinoses	mild (4-7 days)	5	Si
3	1993-1994	24	Alero	205	B	31	Sí	2	Thoracic spine	Muscular_Tendinoses	minimal (1-3 days)	3	Si
4	1993-1994	58	Alero	208	B	32	Sí	2	Hip and Groin	Muscular_Tendinoses	mild (4-7 days)	4	Si
5	1993-1994	43	Base	187	B	21	Sí	2	Trunk and Abdomen	Muscular_Tendinoses	minimal (1-3 days)	2	Si
6	1993-1994	6	Alero	199	B	34	Sí	2	Neck	Muscular_Tendinoses	minimal (1-3 days)	3	Si
7	1993-1994	52	Pivot	206	B	18	Sí	2	Lumbar spine	Muscular_Tendinoses	minimal (1-3 days)	2	Si
8	1993-1994	24	Alero	205	B	31	Sí	2	Wrist and hand	Otras	minimal (1-3 days)	3	Si
9	1993-1994	24	Alero	205	B	31	Sí	2	Foot	Otras	minimal (1-3 days)	2	Si
10	1993-1994	16	Alero	201	B	27	Sí	2	Lower leg	Otras	Severe (>28 days)	33	Si
11	1994-1995	64	Pivot	208	M	25	Sí	2	Foot	Lesiones óseas	minimal (1-3 days)	2	Si
12	1994-1995	46	Base	193	B	29	Sí	2	Wrist and hand	Lesiones óseas	minimal (1-3 days)	3	Si
13	1994-1995	64	Pivot	208	M	25	Sí	2	Knee	Lesiones óseas	moderate (8-28)	22	Si
14	1994-1995	60	Pivot	207	B	26	Sí	2	Ankle	Lesiones ligamentosa...	moderate (8-28)	9	Si
15	1994-1995	46	Base	193	B	29	Sí	2	Knee	Muscular_Tendinoses	moderate (8-28)	9	Si
16	1994-1995	46	Base	193	B	29	Sí	2	Ankle	Muscular_Tendinoses	minimal (1-3 days)	1	Si
17	1994-1995	6	Alero	199	B	35	Sí	2	Lower leg	Muscular_Tendinoses	moderate (8-28)	26	Si
18	1994-1995	46	Base	193	B	29	Sí	2	Thigh	Muscular_Tendinoses	moderate (8-28)	13	Si
19	1994-1995	46	Base	193	B	29	Sí	2	Thigh	Muscular_Tendinoses	minimal (1-3 days)	1	Si
20	1994-1995	6	Alero	199	B	35	Sí	2	Wrist and hand	Muscular_Tendinoses	mild (4-7 days)	6	Si

Font: T.Bovè et al (*Apunts Sports Medicine*; 2019)

“Show me your equations and models and I'll ask to see your data. Show me your data and won't need to see your equations and models. --(if Fred Brooks were a data scientist)” <**Roger D.Peng**>

Encuesta general USA 1991.sav - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

0:

	sexo	raza	región	feliz	vida	herma	hijos	educ	edad	ed
1	Mujer	Blanca	Nor-E	Muy feliz	Excitante	1	2	12	61	No p
2	Mujer	Blanca	Nor-E	Bastante	Excitante	2	1	20	32	
3	Hombre	Blanca	Nor-E	Muy feliz	No proced	2	1	20	35	
4	Mujer	Blanca	Nor-E	No conte	Rutinaria	2	0	20	26	
5	Mujer	Negra	Nor-E	Bastante	Excitante	4	0	12	25	No
6	Hombre	Negra	Nor-E	Bastante	No proced	7	5	10	59	
7	Hombre	Negra	Nor-E	Muy feliz	Excitante	7	3	10	46	
8	Mujer	Negra	Nor-E	Bastante	No proced	7	4	16	Nn	

Vista de datos / Vista de variables

SPSS El procesador está preparado

Encuesta general USA 1991.sav - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

0:

	sexo	raza	región	feliz	vida	herma	hijos	educ	edad	ed
1	2	1	1	1	1	1	2	12	61	
2	2	1	1	2	1	2	1	20	32	
3	1	1	1	1	0	2	1	20	35	
4	2	1	1	9	2	2	0	20	26	
5	2	2	1	2	1	4	0	12	25	
6	1	2	1	2	0	7	5	10	59	
7	1	2	1	1	1	7	3	10	46	
8	2	2	1	2	0	7	4	16	99	

Vista de datos / Vista de variables

SPSS El procesador está preparado

Encuesta general USA 1991.sav - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?



	Nombre	Tipo	Anch	Deci	Etiqueta	Valo
1	sexo	Numérico	1	0	Sexo del encuestado	{1, Hombre}..
2	raza	Numérico	1	0	Raza del encuestado	{1, Blanca}...
3	región	Numérico	8	0	Región de los Estados Unidos	{1, Nor-Este}.
4	feliz	Numérico	1	0	Nivel de felicidad	{0, No procec
5	vida	Numérico	1	0	¿Su vida es excitante o aburrida?	{0, No procec
6	hermanos	Numérico	2	0	Número de hermanos y hermanas	{98, No sabe
7	hijos	Numérico	1	0	Número de hijos	{8, Ocho o m
8	educ	Numérico	2	0	Número de años de escolarización	{97, No proce
9	edad	Numérico	2	0	Edad del encuestado	{98, No sabe



Vista de datos

Vista de variables



SPSS El procesador está preparado

Agrupació de classe

- Poden agrupar-se els valors de les variables en classes (intervals).

Podem transformar una variable quantitativa a una variable qualitativa!!! No a l'inrevès!

Exemples:

Edats:

- Menys de 20 anys, de 20 a 50 anys, més de 50 anys

Lesions:

- Menys de 3 lesions, De 3 a 5, 6 o más lesions

IMC:

- Normal ($< 25 \text{ kg/m}^2$), Sobrepès (25- 29,99), Obesitat (30 –34,99) y Obesitat tipus II (35- 39,99).

Codificació de la variable

- Es buena idea codificar las variables como números para poder procesarlas con facilidad en un ordenador.
- Es conveniente asignar “etiquetas” a los valores de las variables para recordar qué significan los códigos numéricos.
 - Sexo (Cualit: Códigos arbitrarios)
 - 1 = Hombre
 - 2 = Mujer
 - Felicidad: Respetar un orden al codificar.
 - 1 = Muy feliz
 - 2 = Bastante feliz
 - 3 = No demasiado feliz
- Se pueden asignar códigos a respuestas especiales como
 - 0 = No sabe
 - 99 = No contesta...
- Estas situaciones deberán ser tenidas en cuentas en el análisis. Datos perdidos ('missing data')



How IBM SPSS Statistics can help academia make better decisions with data

Videos:

<https://study.sagepub.com/aldrich3e/student-resources/step-by-step-spss%C2%AE-tutorial-videos>

<https://www.thestatisticalmind.com/spss-video-tutorials/>

Tutorials:

<https://www.spss-tutorials.com/basics/>

Data Science comprised of:

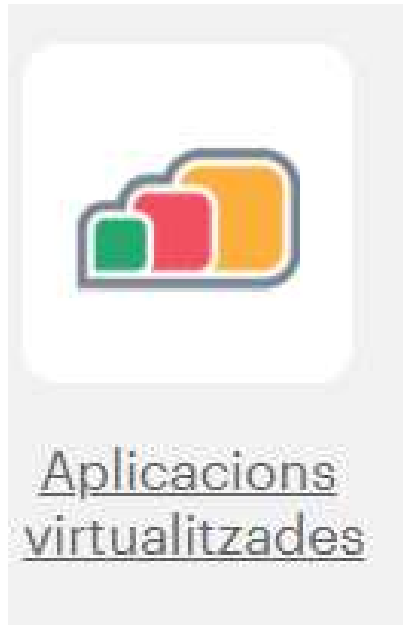
1. Data Gathering, Preparation, and Exploration

2. Data Representation and Transformation

3. Computing with Data

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4), 745-766.

Entrar a l'SPSS. Revisar funcions *wrangling data*



- Data/Select Cases
- Data/Split File
- Transform/Compute Variable
- Transform/Recode Variables
- Transform/Date and Time Wizzard

Data wrangling or data cleaning is important to prepare students for dealing with the whole data analysis cycle. <M. Aerts, 2021>

Practica 2: Summary Data - Jugadors NBA

Appl. Sci. 2019, 9(3), 500; <https://doi.org/10.3390/app9030500>

Open Access Article

Mortality of NBA Players: Risk Factors and Comparison with the General US Population

Jose A. Martínez ¹ ✉, Klaus Langohr ² ✉, Julián Felipo ³ ✉ and Martí Casals ^{4,5,*} ✉

¹ Department of Business Economics, Universidad Politécnica de Cartagena, 30201 Cartagena, Spain

² Department of Statistics and Operations Research, Universitat Politècnica de Catalunya/Barcelonatech, 08034 Barcelona, Spain

³ Newsroom, Basketball Departament, Mundo Deportivo, 08036 Barcelona, Spain

⁴ Sport and Physical Activity Studies Centre (CEEAF), University of Vic—Central University of Catalonia (UVic-UCC), 08500 Catalonia, Spain

⁵ Medical Department, Futbol Club Barcelona, Barça Innovation Hub, 08028 Barcelona, Spain

* Author to whom correspondence should be addressed.



Un estudio asocia la mortalidad de ex jugadores NBA a estatura y etnia

- Ha sido realizado por investigadores españoles y permite realizar estimaciones sobre la esperanza de vida de los baloncestistas

