Curso de impulso a la investigación
Hospital Vega Baja

Principios básicos de una base de datos

Ana Escrig y Carlos Vergara

16/11/2023

Servicio de Estudios Estadísticos

estadística@fisabio.es

@estadistica-fisabio

We ❤ R

- **Ana Escrig Pérez**
  - ana.escrig@fisabio.es
  - @BarruAna

- **Carlos Vergara Hernández**
  - carlos.vergara@fisabio.es
  - @carlosvergara

Introducción

Las **hojas de cálculo** son frecuentemente utilizadas para la entrada y el almacenamiento de datos.

Es importante **organizar adecuadamente** los datos para evitar errores y facilitar su posterior análisis.

A continuación, mostramos algunos **aspectos y recomendaciones a tener en cuenta** para crear una buena base de datos en hojas de cálculo o similares.

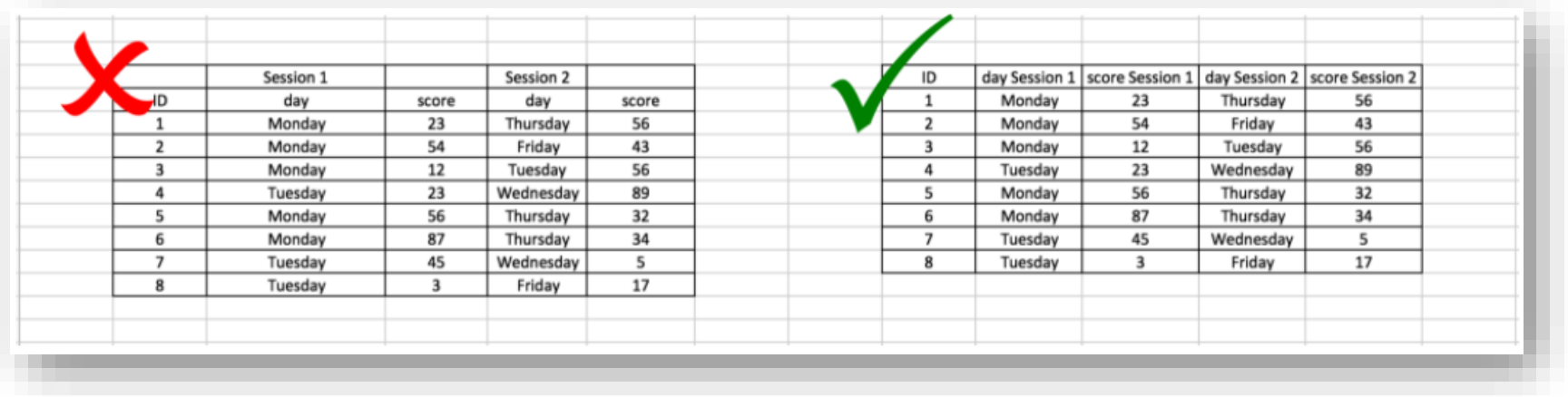
Recomendaciones para crear una base de datos

1. Introducir todos los datos en la misma tabla o marco de datos

- Lo ideal es organizar todos los datos en una sóla tabla, con filas correspondientes a individuos y columnas correspondientes a variables.
- La tabla debe completarse celda por celda, comenzando desde la esquina superior izquierda y sin omitir líneas.
- La primera fila debe contener los nombres de las variables. No utilizar más de una fila para los nombres de las variables.

Recomendaciones para crear una base de datos

1. Todos los datos en la misma tabla o marco de datos



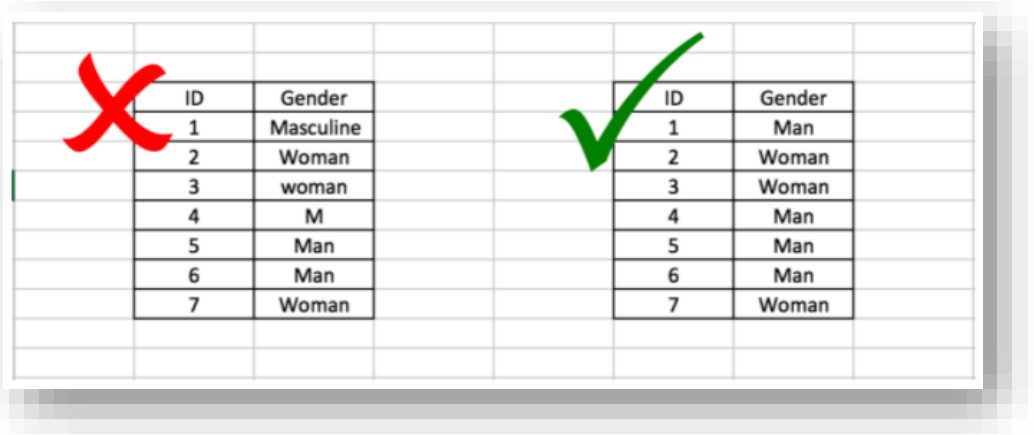
	Session 1		Session 2	
ID	day	score	day	score
1	Monday	23	Thursday	56
2	Monday	54	Friday	43
3	Monday	12	Tuesday	56
4	Tuesday	23	Wednesday	89
5	Monday	56	Thursday	32
6	Monday	87	Thursday	34
7	Tuesday	45	Wednesday	5
8	Tuesday	3	Friday	17

ID	day Session 1	score Session 1	day Session 2	score Session 2
1	Monday	23	Thursday	56
2	Monday	54	Friday	43
3	Monday	12	Tuesday	56
4	Tuesday	23	Wednesday	89
5	Monday	56	Thursday	32
6	Monday	87	Thursday	34
7	Tuesday	45	Wednesday	5
8	Tuesday	3	Friday	17

2. Consistencia (I)

- **Codificar las variables categóricas de forma consistente.**

Ejemplo: Para una variable categórica como el sexo de un individuo, utilizar un único valor común para los hombres (“Hombre”) y un único valor común para las mujeres (“Mujer”).



ID	Gender
1	Masculine
2	Woman
3	woman
4	M
5	Man
6	Man
7	Woman

ID	Gender
1	Man
2	Woman
3	Woman
4	Man
5	Man
6	Man
7	Woman

No escribir a veces “H”, otras “hombre” y otras “Hombre”

2. Consistencia (II)

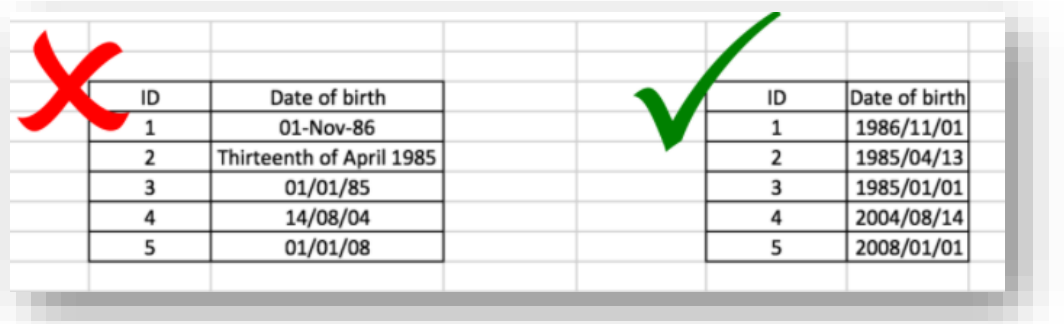
- **Indicar los valores faltantes de forma consistente.** Utilizar siempre el mismo código o dejar siempre las celdas en blanco para indicar valores faltantes.

- Si por algún motivo los datos se van a estructurar **en múltiples hojas o tablas de datos, ser consistente en el diseño,** todas las hojas deben tener la misma estructura de variables.

Recomendaciones para crear una base de datos

2. Consistencia (III)

- Utilizar un **formato consistente para todas las fechas**, preferiblemente el formato estándar AAAA-MM-DD (*por ejemplo, 2015-08-01*).



ID	Date of birth
1	01-Nov-86
2	Thirteenth of April 1985
3	01/01/85
4	14/08/04
5	01/01/08

ID	Date of birth
1	1986/11/01
2	1985/04/13
3	1985/01/01
4	2004/08/14
5	2008/01/01

No escribir a veces “8/1/2015” y otras “8-1-15”

- Tener **cuidado con los espacios adicionales dentro de las celdas**. Una celda en blanco o vacía es diferente de una celda que contiene un sólo espacio.

“hombre” es diferente a “hombre ”

3. Nombre adecuado para las variables (I)

- **No duplicar nombres de variables**
- **No utilizar espacios en nombres de variables ni archivos.** En su lugar, utilizar guiones bajos (_), punto (.) o guiones (-) (elegir uno y ser consistente).
- **Evitar acentos y otros caracteres especiales** ($, @, %, #, &, *, /, . . .).
- Establecer **nombres breves, pero con significado**, no demasiado cortos.

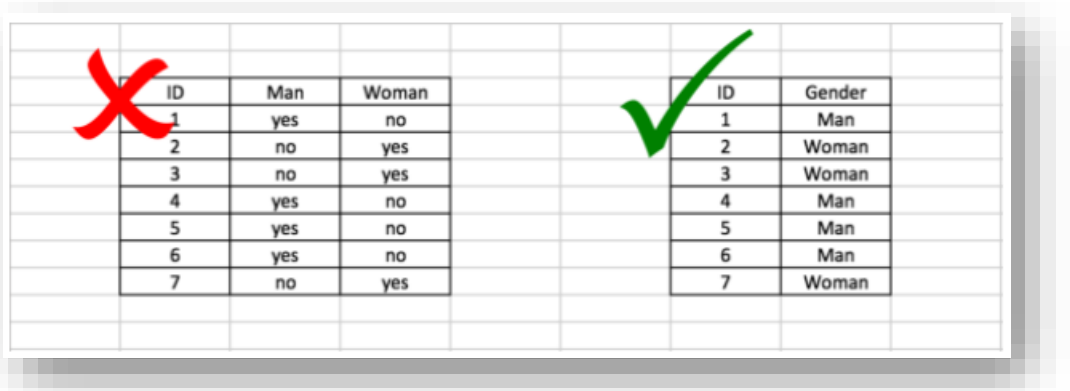
3. Nombre adecuado para las variables (II)

Adecuado	**Adecuado alternativo**	**Evitar**
Max_temp_C	MaxTemp	Maximum Temp (ºC)
Precipatation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell type
Observation_01	first_observation	1st Obs.

4. Una única variable en cada columna (I)

Una variable corresponde a una característica o métrica de un individuo, que puede tener diferentes niveles o valores (*por ejemplo, el sexo o la edad de un paciente*).

- Para estas **variables, en las que a los individuos les corresponde un sólo valor, se debe crear una sola columna por variable.**



ID	Man	Woman
1	yes	no
2	no	yes
3	no	yes
4	yes	no
5	yes	no
6	yes	no
7	no	yes

ID	Gender
1	Man
2	Woman
3	Woman
4	Man
5	Man
6	Man
7	Woman

Una variable = una columna

4. Una única variable en cada columna (II)

Cuando un mismo individuo puede tener varios niveles al mismo tiempo, se debe crear una columna para cada nivel.

Por ejemplo, si un paciente puede tener varios síntomas al mismo tiempo (dolor de cabeza, de piernas, . . .), crear una columna por cada síntoma y completarlas con sí o no.

ID	Source of pain
1	foot and head
2	head and elbow
3	elbow
4	
5	foot and head
6	foot
7	elbow
8	no pain

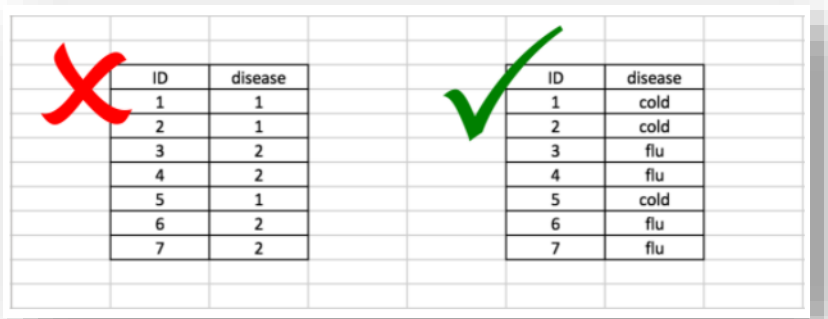
ID	pain	source_foot	source_head	source_elbow
1	yes	yes	yes	no
2	yes	no	yes	yes
3	yes	no	no	yes
4	yes			
5	yes	yes	yes	no
6	yes	yes	yes	no
7	yes	no	no	yes
8	no			

Evitar múltiples

5. No codificar las variables cualitativas

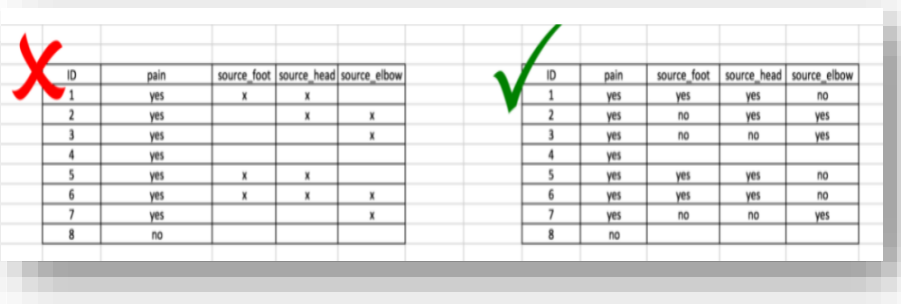
Evitar la codificación de variables cualitativas.

Una codificación 1/2 para el género es inútil, es preferible el uso de “mujer” y “hombre”. Si realmente no se puede evitar utilizar códigos, usar únicamente 1/0 para indicar presencia/ausencia o sí/no.



ID	disease
1	1
2	1
3	2
4	2
5	1
6	2
7	2

ID	disease
1	cold
2	cold
3	flu
4	flu
5	cold
6	flu
7	flu



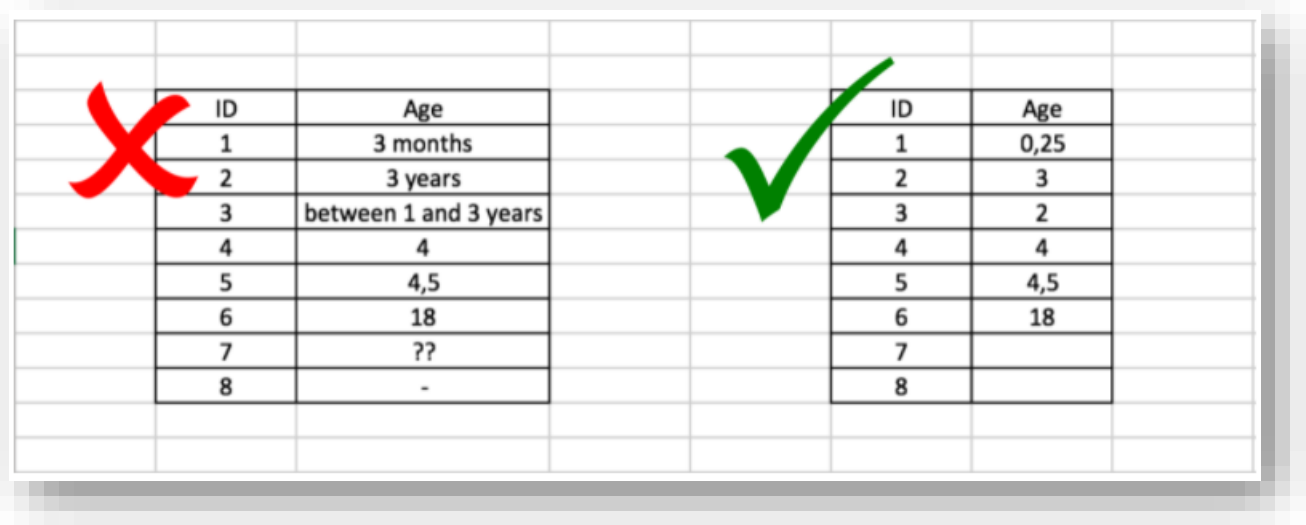
ID	pain	source_foot	source_head	source_elbow
1	yes	x	x	
2	yes		x	x
3	yes			x
4	yes			
5	yes	x	x	
6	yes	x	x	x
7	yes			x
8	no			

ID	pain	source_foot	source_head	source_elbow
1	yes	yes	yes	no
2	yes	no	yes	yes
3	yes	no	no	yes
4	yes			
5	yes	yes	yes	no
6	yes	yes	yes	no
7	yes	no	no	yes
8	no			

Recomendaciones para crear una base de datos

6. Mantener las variables numéricas como numéricas y mantener siempre las mismas unidades

Por ejemplo, la edad de los pacientes deber ser un número, no “25 años”, sólo “25”.



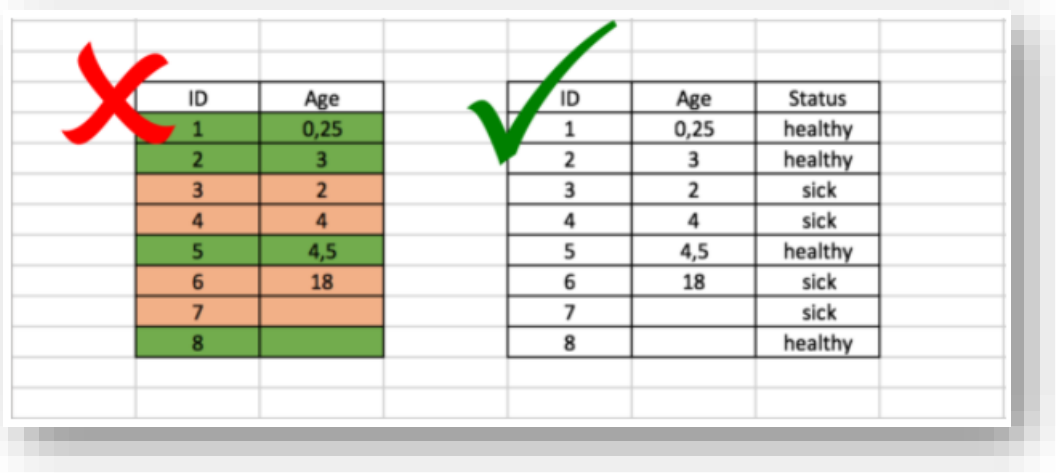
ID	Age
1	3 months
2	3 years
3	between 1 and 3 years
4	4
5	4,5
6	18
7	??
8	-

ID	Age
1	0,25
2	3
3	2
4	4
5	4,5
6	18
7	
8	

Variables numéricas únicamente números

7. No utilizar colores ni resaltado para codificar datos

No deben utilizarse colores ni resaltado de celdas para indicar alguna característica. En su lugar, debe crearse una nueva variable en una columna adicional que represente la característica en cuestión.

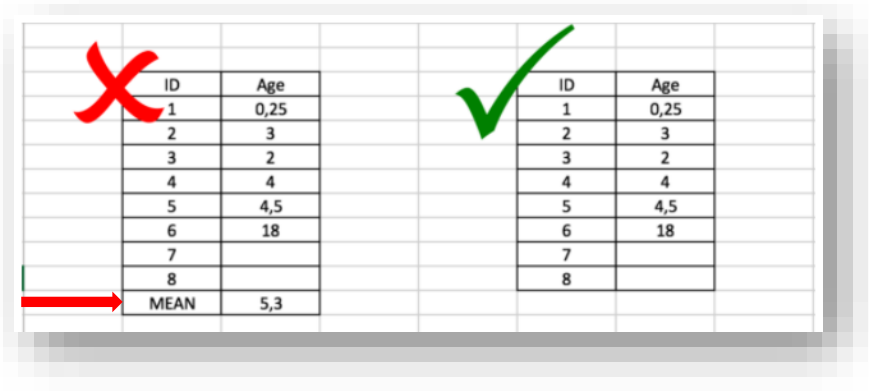


ID	Age
1	0,25
2	3
3	2
4	4
5	4,5
6	18
7	
8	

ID	Age	Status
1	0,25	healthy
2	3	healthy
3	2	sick
4	4	sick
5	4,5	healthy
6	18	sick
7		sick
8		healthy

8. No realizar cálculos ni comentarios en los archivos de datos brutos (I)

El archivo de datos debe contener **sólo los datos y nada más**: sin cálculos, sin gráficos, ni explicaciones o anotaciones...



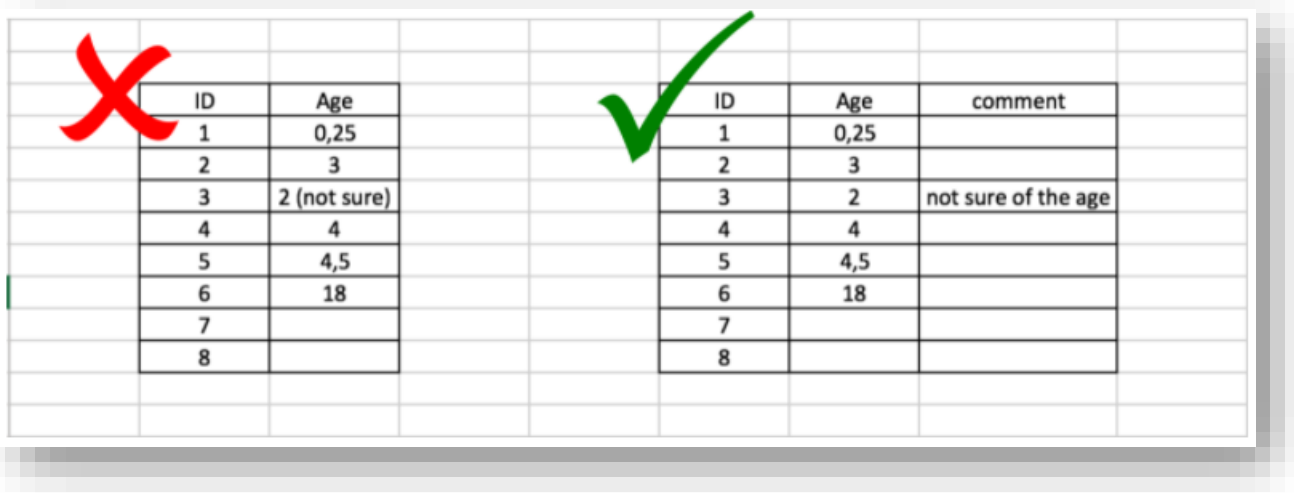
ID	Age
1	0,25
2	3
3	2
4	4
5	4,5
6	18
7	
8	
MEAN	5,3

ID	Age
1	0,25
2	3
3	2
4	4
5	4,5
6	18
7	
8	

*Si los datos se encuentran en Excel y se quieren realizar algunos análisis, debe hacerse **una copia** del archivo y realizar los cálculos y gráficos en la copia*

8. No realizar cálculos ni comentarios en los archivos de datos brutos (II)

Si se necesitan añadir comentarios o anotaciones, puede agregarse una columna especial para ellos.

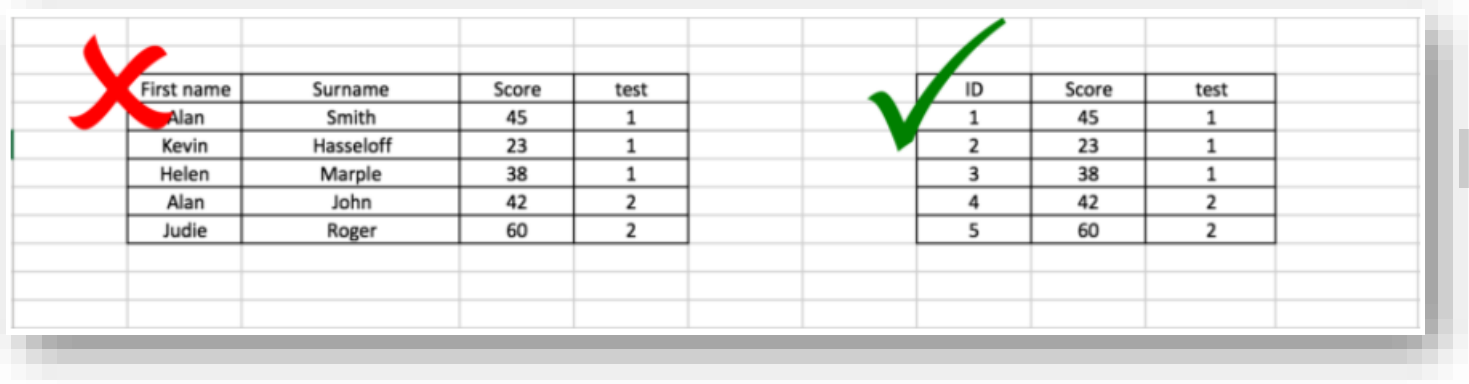


ID	Age
1	0,25
2	3
3	2 (not sure)
4	4
5	4,5
6	18
7	
8	

ID	Age	comment
1	0,25	
2	3	
3	2	not sure of the age
4	4	
5	4,5	
6	18	
7		
8		

9. Anonimizar la base de datos

No debe aparecer la identidad de las personas en la base de datos. En su lugar, añadir un identificador y guardar la correspondencia en otra base de datos, con los nombres, expedientes médicos. . . y cualquier otra información que no sea relevante para el análisis estadístico.



First name	Surname	Score	test
Alan	Smith	45	1
Kevin	Hasseloff	23	1
Helen	Marple	38	1
Alan	John	42	2
Judie	Roger	60	2

ID	Score	test
1	45	1
2	23	1
3	38	1
4	42	2
5	60	2

ANONIMIZAR

10. Crear un diccionario de datos (I)

Es útil tener un **archivo** separado **que explique el significado de las variables y facilite su comprensión** a las personas que analicen los datos.

Puede presentarse en otra tabla o hoja de datos para que el analista pueda utilizarlo en los análisis.

	A	B	C	D
1	name	plot_name	group	description
2	mouse	Mouse	demographic	Animal identifier
3	sex	Sex	demographic	Male (M) or Female (F)
4	sac_date	Date of sac	demographic	Date mouse was sacrificed
5	partial_inflation	Partial inflation	clinical	Indicates if mouse showed partial pancreatic inflation
6	coat_color	Coat color	demographic	Coat color, by visual inspection
7	crumblers	Crumblers	clinical	Indicates if mouse stored food in their bedding

10. Crear un diccionario de datos (II)

Contenido de un diccionario

- El **nombre** exacto de la **variable** en el archivo de datos.
- Una explicación del **significado** de cada variable.
- Las **unidades** de medida.
- Una versión del **nombre** de las variables que podría utilizarse **para la visualización de los resultados** en tablas o gráficos (enunciado…).
- **Rango de valores** esperados (mínimos y máximos).
- Cualquier otra información que deba tenerse en cuenta en el análisis.

11. Hacer copias de seguridad

Realizar copias de seguridad periódicas de los datos (en diferentes ubicaciones).

Se evitan sustos:

Si accidentalmente se sobrescribe algún dato, se podrá recuperar.

12. Utilizar validación de datos para evitar errores (I)

Es importante asegurarse de que los datos no contienen errores.

En el caso de **Excel**, la opción de **validación de datos*** permite controlar el tipo de datos o los valores que los usuarios pueden introducir en cada celda.

Para **especificar los posibles valores** de una variable:

- Seleccionar la columna correspondiente a la variable.
- En la barra del menú, elegir “Datos” -> “Validación de datos”.
- Elegir un criterio de validación adecuado.

Un número entero o decimal en algún rango, una lista de posibles valores o texto, con una longitud limitada.

*http://bit.ly/excel_dataval

12. Utilizar validación de datos para evitar errores (II)

Al mismo tiempo, en el caso de **Excel**, puede establecerse un tipo de datos particular para cada variable, por ejemplo, texto.

Para **especificar el tipo de datos** de una variable:

- Seleccionar la columna de la variable.
- Pulsar con el botón derecho del ratón sobre la columna y seleccionar “Formato de celdas”.
- Elegir el formato.

Referencias

Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: 10.1080/00031305.2017.1375989
https://rtask.thinkr.fr/the-ten-commandments-for-a-well-formatteddatabase/