



Análisis del conjunto de datos.

Alberto Morillo Alujas.
Tests and Trials.

Tests and Trials, S.L.
Parida Caparells 97, Edifici CEEI 2, Mòdul 8, 25 192 Lleida

1



Análisis del conjunto de datos.

1. La presentación de un problema.

- Generación de hipótesis de trabajo.

Tests and Trials, S.L.
Parida Caparells 97, Edifici CEEI 2, Mòdul 8, 25 192 Lleida

2



Análisis del conjunto de datos.



Parida Caparells 97, Edifici CEEI 2, Mòdul 8, 25 192 Lleida

3



Análisis del conjunto de datos.

1. La presentación de un problema.

- Generación de hipótesis de trabajo.
- Introducción de datos.

Tests and Trials, S.L.
Parida Caparells 97, Edifici CEEI 2, Mòdul 8, 25 192 Lleida

4

Análisis del conjunto de datos.



2. ¿Donde almacenar los datos de nuestra investigación?.

- ¿Hojas de cálculo?.
- ¿Bases de datos (SGBD)?.

Análisis del conjunto de datos.



2.2. Sobre las bases de datos:

- Principio 1
- Principio 2
- Principio 3

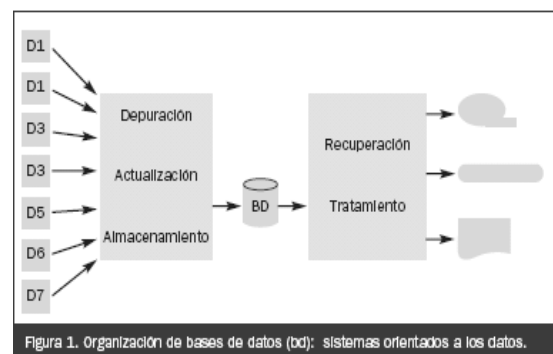
Análisis del conjunto de datos.



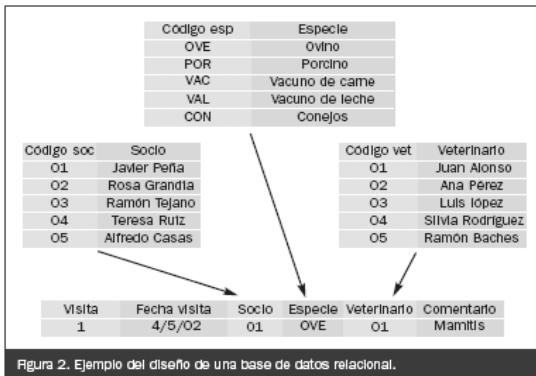
3. Características de las bases de datos relacionales:

- Definición.
- Organización.

Análisis del conjunto de datos.



Análisis del conjunto de datos.



Tests and Trials, S.L.
Parida Caparells 97, Edifici CEET 2, Mòdul 8, 25 192 Lleida

9

Análisis del conjunto de datos.



3. Características de las bases de datos relacionales:

- Definición.
- Organización.
- Tablas diccionario.
- Reglas de Codd. (ver http://es.wikipedia.org/wiki/Edgar_F._Codd)
(http://es.wikipedia.org/wiki/Normalizaci%C3%B3n_de_una_base_de_datos)

Tests and Trials, S.L.
Parida Caparells 97, Edifici CEET 2, Mòdul 8, 25 192 Lleida

10

Análisis del conjunto de datos.



4. Funciones de las bases de datos relacionales:

- Adición modificación o anulación de registros.
- Localización de datos.
- Ordenación de datos.
- Unión de datos de varias tablas.
- Exportación e importación de datos.

Tests and Trials, S.L.
Parida Caparells 97, Edifici CEET 2, Mòdul 8, 25 192 Lleida

11

Análisis del conjunto de datos.



5. Depuración de datos.

- Libres de errores.
- Captura desde una base de datos.
- Procesos lógicos de la depuración:
 - Especial para cada variable.
 - Intervalos de valores.
 - Variables relacionadas.
 - Formatos de variables.

Tests and Trials, S.L.
Parida Caparells 97, Edifici CEET 2, Mòdul 8, 25 192 Lleida

12



5. Depuración de datos.

- Procesos lógicos de la depuración:
 - Valores en blanco y perdidos (blank and missing values).
- Estrategias de depuración:
 - Marcar con signo negativo en las variables numéricas y el guión en las alfanuméricas.
 - Notaciones de número de incidencias.
 - Listar los registros con incidencias.



5. Depuración de datos.

- Estrategias de depuración:
 - Localizar los valores originales.
 - Corregirlos.
 - Volver a corregir.
 - ¿Qué hacer con los valores erróneos tras la depuración?



6. Codificación de datos.

- “Transformar mediante las reglas de un código la formulación de un mensaje”.



Tabla 1. Codificación de la variable “RAZA”.

RAZA	CODIGO
Large White	01
Landrace	02
Pietrain	03
Duroc	04

Análisis del conjunto de datos.



Tabla 2. Valores de la variable “RAZAVER”
(Raza de los verracos).

RAZA
Large White
Landrace
Pietrain
Duroc
Land Race
large white
Pietren

Tests and Trials, S.L.
Parida Caparells 97, Edifici CEEI 2, Mòdul 8, 25 192 Lleida

17

Análisis del conjunto de datos.



6. Codificación de datos.

- Variables binarias.
- Variables ficticias o dummy variables.

Tests and Trials, S.L.
Parida Caparells 97, Edifici CEEI 2, Mòdul 8, 25 192 Lleida

18

Análisis del conjunto de datos.



6. Codificación de datos.

- Codificación de variables de respuestas cerradas
- Organización de valores.
 - Todos los datos deben ser numéricos.
 - Situación de las variables.
 - Los códigos de las variables deben ser mutuamente excluyentes.

Tests and Trials, S.L.
Parida Caparells 97, Edifici CEEI 2, Mòdul 8, 25 192 Lleida

19

Análisis del conjunto de datos.



6. Codificación de datos.

- Los códigos de las variables deben ser mutuamente excluyentes.

Pregunta 6. ¿está usted actualmente casado, no casado o tiene hijos?. Marque una respuesta.

Actualmente casado	1
No casado	2
Con hijos	3

Tests and Trials, S.L.
Parida Caparells 97, Edifici CEEI 2, Mòdul 8, 25 192 Lleida

20



6. Codificación de datos.

- Los códigos de las variables deben ser mutuamente excluyentes.

Pregunta 6. ¿está usted actualmente casado, viudo, divorciado, separado o nunca se ha casado? Marque una respuesta.

Actualmente casado	1
Viudo	2
Divorciado	3
Separado	4
Nunca casado	5



6. Codificación de datos.

- Cada variable debe ser codificada de tal forma que ofrezca la máxima información posible.



7. *Missing values, zero values y blank values.*

- Blank values (8).
- Zero values (0).
- Missing values (9).



8. Variabilidad de las mediciones.

- Biología.
- Observador.
- Instrumento.

Análisis del conjunto de datos.



8. Variabilidad de las mediciones.

- Biología = mínimas variables subjetivas.
- Observador = siempre el mismo a la misma hora.
- Instrumento = mismo aparato de medida.
- Exactitud del aparato de medida acorde.
- Variables subjetivas con patrones.
- Formación de los investigadores en el uso de aparatos.

Tests and Trials, S.L.
Parida Caparells 97, Edifici CEEI 2, Mòdul 8, 25 192 Lleida

25

Análisis del conjunto de datos.



8. Variabilidad de las mediciones.

- Contratación de los medidores automáticos.
- Varias medidas del mismo individuo.
- Calibración de los aparatos.
- Técnicas de enmascaramiento.

Tests and Trials, S.L.
Parida Caparells 97, Edifici CEEI 2, Mòdul 8, 25 192 Lleida

26

Análisis del conjunto de datos.



9. Selección de las variables.

- Variables que permiten la aplicación del protocolo.
- Variables que permiten evaluar las oscilaciones de la respuesta.
- Variables modificadoras del efecto.
- Variables intermedias del efecto.
- Variables descriptoras de los individuos.
- Otras variables (secundarias).

Tests and Trials, S.L.
Parida Caparells 97, Edifici CEEI 2, Mòdul 8, 25 192 Lleida

27

Análisis del conjunto de datos.



10. Nominación de las variables.

- Se puede utilizar cualquier letra del alfabeto tanto en mayúsculas como en minúsculas.
- Se admite el uso del guión bajo (_).
- Igualmente se pueden usar las cifras 0, 1, 2, 3, 4, 5, 6, 7, 8 y 9.
- Un nombre de variable no puede contener puntos ni espacios en blanco.

Tests and Trials, S.L.
Parida Caparells 97, Edifici CEEI 2, Mòdul 8, 25 192 Lleida

28



10. Nominación de las variables.

- Un nombre de variable no puede contener puntos ni espacios en blanco.
- El primer carácter de una variable debe ser una letra o un guión bajo.
- Se distinguen las mayúsculas de las minúsculas para identificar dos variables distintas.



10. Nominación de las variables.

- No escribir nombres de variables de más de 255 caracteres (y os recomendamos que no sea mayor de 8 caracteres ya que muchos programas no lo admiten, vease SPSS).
- El nombre de una variable debe identificar lo más claramente posible su contenido.
- No hacer uso de la posibilidad de distinguir dos variables usando mayúsculas o minúsculas en alguno de los caracteres.



11. Visualización de datos: distribución de frecuencias y medidas descriptivas.

- Visualización de datos variable a variable.
- Aproximación de los datos a la Ley Normal.



11. Distribución de frecuencias:

- Ordenación de los valores de una variable siguiendo unas categorías (escala nominal u ordinal), números o intervalos de clase (escala métrica).

Análisis del conjunto de datos.



11.Distribución de frecuencias con variable categórica:

Tabla 1. Nivel de estudios de los granjeros de una empresa de producción pollos.

Nivel de estudios	Casos	ph	%
Universitarios	3	0,04	4,05 %
Bachiller	24	0,32	32,43 %
ESO o primarios	45	0,61	60,81 %
Sin estudios	2	0,03	2,70 %
Totales	74	1	100%

Tests and Trials, S.L.
Partida Caparells 97, Edifici CEEI 2, Mòdul 8, 25 192 Lleida

33

Análisis del conjunto de datos.



11.Distribución de frecuencias de variables cuantitativas: stem & leaf.

Tabla 2. Stem & leaf de la muestra de pesos al destete.

Pesos al destete de una muestra del fichero, primeros 14 datos.	
Stem	Leaf
4	6
5	4, 5, 5, 8, 8, 8, 9
6	0, 0, 0, 0, 0, 0

Tests and Trials, S.L.
Partida Caparells 97, Edifici CEEI 2, Mòdul 8, 25 192 Lleida

34

Análisis del conjunto de datos.



11.Distribución de frecuencias de variables cuantitativas: stem & leaf.

- Caso especial de dos variables en un stem & leaf (<http://regentsprep.org/Regents/math/data/stemleaf.htm>)
- Ventajas de stem & leaf:
 - Nos muestra todos los datos de la distribución.
 - Podemos ver donde se concentran los datos o donde no hay valores.
 - Podemos ver la amplitud de la distribución.
 - Vemos los valores alejados.
 - Vemos la forma de la distribución.

Tests and Trials, S.L.
Partida Caparells 97, Edifici CEEI 2, Mòdul 8, 25 192 Lleida

35

Análisis del conjunto de datos.



11.Representación gráfica de los casos de las variables:

- Seguir las normas de crear los gráficos.

Tests and Trials, S.L.
Partida Caparells 97, Edifici CEEI 2, Mòdul 8, 25 192 Lleida

36

Análisis del conjunto de datos.



Tabla 3. Tipos de variables y sus representaciones gráficas.

Tipo de variable	Representaciones gráficas
Categórica ó nominal	- diagrama de sectores - diagrama de barras - diagrama de Pareto
Categorías ordenadas y cuantitativas discretas	- diagrama de barras - diagrama de líneas
Cuantitativas continuas	- stem and leaf - histogramas - polígonos de frecuencias

Tests and Trials, S.L.
Parida Caparrells 97, Edifici CEET 2, Mòdul 8, 25 192 Lleida

37

Análisis del conjunto de datos.



11. Ordenación de los datos frente a momentos

- Percentiles
- Cuartiles.
- Deciles.
- Amplitud intercuartílica.
- Five numbers, five-number (Tukey): mínimo, máximo, cuartiles 25, 50 y 75.

Tests and Trials, S.L.
Parida Caparrells 97, Edifici CEET 2, Mòdul 8, 25 192 Lleida

38

11. Ordenación de los datos frente a momentos

- Boxplot

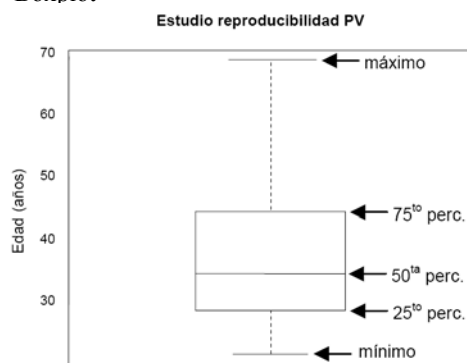


Figura 1: gráfico de caja.

- ¿Qué técnica estadística usar con los diferentes tipos de variables?



Variable "X" Independiente →				
Variable "Y" dependiente ↓		Categorica C=2	Categorica C>=2	Cuantitativa
	Categorica C=2	Pruebas χ^2 Comparacion de dos proporciones	Pruebas χ^2	Regresion logistica
	Categorica C>=2	Pruebas χ^2	Pruebas χ^2	Regresion logistica
	Cuantitativa	Comparacion de dos medias (t de Student)	Analisis de la variancia	Regresion simple

Tests and Trials, S.L.
Parida Caparrells 97, Edifici CEET 2, Mòdul 8, 25 192 Lleida

40



- a.m.alujas@testsandtrials.com
- Asunto: Taller de estadística.