

---

---

# *PRÁCTICA SESIÓN 5: CALIDAD DE DATOS*

PEDRO M. VALERO MORA

## ESQUEMA

1.	Técnicas básicas para control de datos .....	2
2.	Técnicas básicas para 1 variable categórica.....	3
3.	Técnicas básicas para 2 variables categóricas .....	4
4.	Técnicas básicas para una variable numérica.....	7
5.	Técnicas para dos variables numéricas .....	8
6.	Valores perdidos.....	9

---

---

# 1. Técnicas básicas para control de datos

*usando el SPSS para controlar la calidad*

- Utilizaremos técnicas que en principio son para cálculo estadístico pero que nos dan pistas de los posibles problemas
- A menudo los análisis estadísticos llevan a problemas de calidad en lugar de a resultados
- El número de pruebas hipotético sería muy grande por eso normalmente nos ceñimos a un número limitado

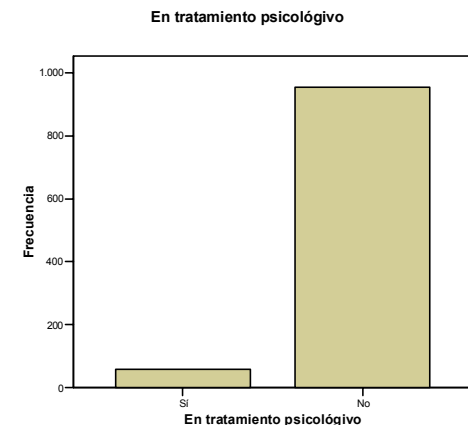
---

---

## 2. Técnicas básicas para 1 variable categórica

*Lo más sencillo*

- Lo mejor es utilizar diagramas de barras si hay pocos valores diferentes y tablas de frecuencias si hay muchos. Se utiliza el comando Frecuencias en el menú Analizar, submenú estadísticos descriptivos
- Podemos ver si hay valores sin etiqueta, si hay más valores que los permitidos, si alguna categoría tiene muchos o pocos valores aparentemente, etc.



---

---

### 3. Técnicas básicas para 2 variables categóricas

*utilizando tablas*

- Existen gráficos para esta situación pero no son muy útiles porque los valores interesantes suelen tener frecuencias muy bajas y puede ser difícil detectarlos. Lo mejor es utilizar tablas. Se utiliza el comando tablas de contingencia en el menu Analizar, submenú estadísticos descriptivos
- Ejemplo Edad por número de hijos. Se pueden ver combinaciones de categorías con muy pocos valores

y que resulten cerca de lo imposible o implausible,

Tabla de contingencia Edad del encuestado \* Número de hijos

Recuento		Número de hijos								Total	
		0	1	2	3	4	5	6	7		
Edad del encuestado	18	2	1	0	0	0	0	0	0	0	3
	19	6	1	2	0	0	0	0	0	0	9
	20	11	7	0	0	0	0	0	0	0	18
	21	33	5	0	0	0	0	0	0	0	38
	22	26	3	4	0	1	1	0	0	0	35
	23	18	7	2	0	0	0	0	0	0	27
	24	12	6	4	2	0	0	0	0	0	24
	25	20	4	1	3	0	0	0	0	0	28
	26	19	5	6	2	1	0	0	0	0	33
	27	13	9	5	0	1	0	1	0	0	29
	28	15	4	9	1	4	0	0	0	0	33
	29	13	4	6	1	3	0	0	0	0	27
	30	10	10	6	6	3	0	0	0	0	35
	31	11	7	9	6	2	0	0	0	0	35
	32	13	7	18	1	1	1	1	1	0	43
	33	16	8	8	4	1	1	0	0	0	38
	34	7	7	14	4	1	1	0	0	0	34
	35	14	10	14	11	4	1	0	0	1	55
	36	8	7	13	5	1	0	0	1	0	35
	37	9	6	13	3	1	0	0	0	1	33
	38	7	7	11	10	2	0	0	0	0	37
	39	5	7	16	5	2	0	0	0	0	35
	40	7	6	15	7	0	0	1	0	0	36
	41	8	9	12	5	2	1	0	1	0	38
	42	8	7	8	2	1	2	1	1	0	30
	43	5	6	10	6	5	1	0	0	0	33
	44	3	3	14	5	3	1	1	0	0	30
	45	5	3	8	4	0	0	1	0	0	21
	46	1	1	6	4	0	0	0	0	0	12
	47	4	6	8	8	1	0	0	0	1	28

Extraño?

- 
- 
- El problema de las tablas grandes es organizarlas para que se vean bien en la pantalla. El truco es poner la variable con más valores en las filas y la que menos en las columnas. Eso se adapta mejor al ordenador

---

---

## 4. Técnicas básicas para una variable numérica

*No son habituales en encuestas pero a veces hay*

- Miramos máximos y mínimos y a veces histogramas. Con el histograma se pueden ver los máximos y mínimos (aunque a veces con dificultad) y algunas otras cosas. Los mínimos y los máximos están en el menu Analizar, Submenú Estadísticos Descriptivos, comando Descriptivos. Los histogramas están en el menú Gráficos, comando Histograma
- Los histogramas son muy parecidos a los diagramas de barras cuando hay muchos casos diferentes en la variable categóricas (no obstante, no hay que confundirlos entre sí)

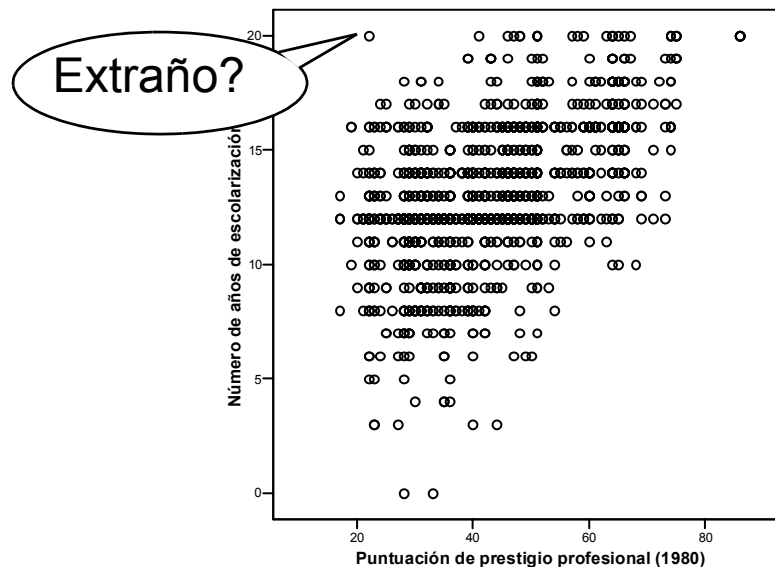
---

---

## 5. Técnicas para dos variables numéricas

*Sólo veremos hasta dos variables pero podrían ser más*

- Lo más apropiado son los diagramas de dispersión.
- Los valores destacados indican posibles errores en los datos



---

---

## 6. Valores perdidos

- La mayoría de los procedimientos en SPSS ofrecen información sobre los valores perdidos
- Un procedimiento que no da información son las tablas de cruces de frecuencias. Para conseguir que aparezcan se puede utilizar la sintaxis del SPSS:
  - Pegar el comando en lugar de apretar el botón de Aceptar
  - Luego añadir en el comando */missing include*
  - Después ejecutarlo para ver el resultado

### ACTIVIDADES

---

**EJERCICIO 6.1** En el archivo de datos de ENCUESTA GENERAL USA 1991 examinar la explicación de las variables

**EJERCICIO 6.2** En el archivo anterior comprobar si el número de hermanos para los sujetos parece correcto.

**EJERCICIO 6.3** Comprobar el número de hijos a ver si parece correcto.

- 
- 
- EJERCICIO 6.4** Comprobar los años de escolarización de los individuos, los padres y el cónyuge para ver si parecen correctos.
- EJERCICIO 6.5** Comprobar otras variables a ver si los resultados parecen correctos en general.
- EJERCICIO 6.6** Comprobar si el número de hijos por la edad parece correcto
- EJERCICIO 6.7** Comprobar si hay variables que tengan muchos valores perdidos
- EJERCICIO 6.8** Una pregunta conceptual, ¿de todas esas variables cuál producirá valores perdidos con mas probabilidad?
- EJERCICIO 6.9** ¿Hay valores faltantes en la pregunta sobre problemas con drogas ilegales? ¿Con qué variable podríamos hacer la comparación?
- EJERCICIO 6.10** En el archivo de datos de GSS93 para datos perdidos.sav están los datos acerca de una encuesta.Revisa los nombres de las variables para entender el estudio.

- 
- 
- EJERCICIO 6.11** ¿Puedes detectar alguna variable en el archivo anterior cuyos máximos o mínimos sean difíciles de creer?
- EJERCICIO 6.12** El cruce de la variable edad y la variable hijos produce algunos resultados curiosos. ¿Podrías indicar cuáles son?
- EJERCICIO 6.13** ¿Crees que las frecuencias para los signos del zodiaco aparentan ser correctas?
- EJERCICIO 6.14** La pregunta el Hombre evolucionó de los animales, ¿produce una distribución de frecuencias que parece correcta?
- EJERCICIO 6.15** La pregunta el efecto invernadero está provocado por el agujero de ozono ¿produce una distribución de frecuencias que parece correcta?
- EJERCICIO 6.16** De las variables en el fichero, examina dos que creas que producirían más valores perdidos que las otras y comparalas con dos que no producirían muchos valores perdidos. ¿Qué variables son? ¿Has acertado y producen muchos valores perdidos?
- EJERCICIO 6.17** Hay dos preguntas relacionadas con aspectos de la vida sexual (amantes y frecsex) de los sujetos. ¿Podrías examinar si los que no contestan a una de esas dos preguntas tampoco contesta a la otra?