

FORMACIÓN PROFESIONAL · INVESTIGACIÓN COMERCIAL · UNIDAD 9

Depuración, codificación y descripción

Lunes 3 de mayo de 2027. Adrián ha grabado las ocho primeras fichas de la encuesta de seguimiento y los kits del último pedido ya tienen una media. Pero un hogar pidió 9 kits y Unai quiere borrar la fila. ¿Qué se depura, con qué reglas y cómo se describe lo que queda?

PARA EMPEZAR

En esta unidad aprenderás a…

- Elaborar un diccionario de datos y distinguir «no contesta», «no procede» y cero.
- Depurar un registro con reglas escritas: duplicados, filtros, rangos y extremos comprobados.
- Construir tablas de frecuencias con su base y elegir el gráfico adecuado.
- Calcular media, mediana, moda, rango y desviación típica muestral con fórmulas visibles.
- Describir la forma de una distribución con un gráfico y un coeficiente con su convención.

SITUACIÓN DE PARTIDA

«Es un caso raro: bórralo»

Lunes 3 de mayo de 2027. La encuesta telefónica de seguimiento a los hogares que probaron los kits en abril ha empezado bien: Adrián ha pasado a la hoja de cálculo las **ocho primeras fichas** en papel y la variable que más interesa a Itxaso, los **kits del último pedido**, ya tiene una media. Pero Adrián frunce el ceño: hay un hogar que pidió **9 kits**.

Unai propone borrar esa fila: «Es un caso raro; si lo dejamos, el informe sale inflado». Itxaso quiere saber cuántos kits lleva un pedido normal para preparar las cajas isotérmicas de los repartos. Nora necesita la descripción el **viernes 14 de mayo** y no quiere decidir con un número que nadie sepa explicar. Te encarga que revises el registro, escribas las reglas de depuración y prepares una descripción comprobable.

❓ ¿Qué debe decidir o comprobar Nora antes de usar la media de kits por pedido para preparar las cajas?

❓ ¿Qué sabemos del pedido de 9 kits y qué estamos suponiendo cuando decimos que «es un caso raro»?

❓ ¿Qué prueba permitiría revisar la primera respuesta, la de borrar la fila?

❓ Y si el hogar no hubiera contestado, ¿sería lo mismo que cero kits?

MAPA DE LA UNIDAD

Lo que vamos a ver

9.1 Diccionario y validación

9.2 Ausencias, duplicados y filtros

9.3 Frecuencias y gráficos

9.4 Tendencia central y dispersión

9.5 Forma de la distribución

Este nueve cambia la media



Este nueve cambia bastante la media.

¿Es un error? 🤔



Primero lo comprobaré en el registro.

CLAVE

Un dato no se elimina porque cambie el resultado esperado, sino porque **una regla escrita y una comprobación** demuestran que es un error. El tique T-2027/04415 confirma un pedido real de 9 kits: se conserva.

Saber qué significa cada celda

Diccionario de datos

Describe cada variable con su nombre, tipo, unidad, códigos o rango válido y la regla con la que se comprueba. Se versiona: si cambia un código, cambia la versión y se anota el motivo.

Qué es una fila

La **unidad de análisis**: en la encuesta de seguimiento, una entrevista a un hogar, identificado con un código y nunca con el nombre. De ella dependen el recuento y los denominadores.

Tipos de variable

Cualitativas **nominales** (canal: se cuentan, no se promedian) u **ordinales** (valoración de 1 a 5: admiten mediana). Cuantitativas **discretas** (kits) o **continuas** (euros, minutos).

Respuestas abiertas

Categorías exhaustivas y excluyentes, con código y definición. Si «otros» pasa de una de cada diez, se revisa la lista; dos personas codifican una parte y comparan.

CLAVE

Quien graba, quien depura y quien analiza **leen lo mismo** en cada celda: el tipo decide la tabla, el gráfico y las medidas.

9.1 · DOCUMENTO 1

Diccionario de datos EK-2027/DD-v1.0

Extracto del diccionario de la encuesta de seguimiento (ficticio).

VARIABLE	TIPO Y UNIDAD	CÓDIGOS O RANGO	REGLA DE VALIDACIÓN
hogar	texto (clave)	H-001 a H-100	existe en el marco; único
kits_ult	discreta, kits	1 a 20; 98; 99	entero; 98 si P7 = No
canal_ult	nominal	1 tienda; 2 web; 98; 99	98 si P7 = No
pedidos_abr	discreta, pedidos	0 a 30; 99	0 es un valor válido

Validación: lo más barato es que el error no entre. La columna kits_ult admite solo enteros de 1 a 20, 98 y 99, y en el formulario la pregunta solo aparece si P7 = Sí. Después de grabar se comprueban tipo, rango, coherencia y unicidad.

CLAVE

Los datos en bruto se guardan aparte y sin tocar (**2027-05-03_seguimiento_kits_brutos_v1**); la depuración trabaja sobre una copia versionada.

Tres cosas distintas

🔇

No contesta (99)

La pregunta se aplicaba y no hay respuesta: es una **ausencia**. No se sustituye por cero ni por la media; se cuenta y se informa.

🚫

No procede (98)

El filtro salta la pregunta: quien nunca ha pedido no tiene «último pedido». Ese hogar queda **fuera de la base** de la variable.

0️⃣

Cero

El hogar contestó y el valor es 0. En «pedidos en abril» es válido y cuenta en la media; en «kits del último pedido» es imposible para quien ha pedido y se revisa.

CLAVE

98 y 99 son **etiquetas, no cantidades**. Promediadas con las ocho filas: 226 ÷ 8 = **28,25 kits** por pedido, un número que no significa nada. Con R04 ya fuera (v2), una condición: =PROMEDIO.SI(C2:C8;"<98") → **5**; en el bruto (C2:C9), 4,83.

9.1 · LA NORMA DICE

Minimización y exactitud

⚖️

Los datos personales serán «adecuados, pertinentes y limitados a lo necesario en relación con los fines para los que son tratados» (c, minimización) y «exactos y, si fuera necesario, actualizados» (d, exactitud).

Reglamento (UE) 2016/679 (RGPD), de 27 de abril de 2016, art. 5.1 · Consulta la versión vigente

SUJETO OBLIGADO

Quien trata los datos de la encuesta: la base usa el código H-042 y no el nombre del hogar.

NO OBLIGA A…

…a borrar un dato porque estorbe: un dato inexacto se rectifica con su fuente, como el tique o la ficha en papel.

CLAVE

Ojo con la IA: no pegues en un asistente datos de hogares identificables.

Las ocho primeras fichas, en bruto

Tal como Adrián las grabó el lunes 3 de mayo; la observación copia lo que anotó el entrevistador (ficticio; 98 = no procede; 99 = no contesta).

REGISTRO	HOGAR	KITS_ULT (KITS)	OBSERVACIÓN DE LA FICHA
R01	H-003	2	
R02	H-011	4	
R03	H-017	4	03/05, 10:12
R04	H-017	4	03/05, 10:12
R05	H-024	6	
R06	H-030	99	colgó antes de la pregunta
R07	H-042	9	
R08	H-051	98	P7 = No (nunca ha pedido)

CLAVE

Depurar no es dejar los datos como nos gustan: es dejarlos **como son**, y explicar cada cambio.

Escritas antes de mirar los resultados

👯

Duplicados

Mismo hogar y misma hora = la misma ficha grabada dos veces: se conserva la primera. =CONTAR.SI.CONJUNTO($B$2:B2;B2;$D$2:D2;D2)>1 marca la segunda si D guarda fecha y hora (aquí, la observación). Otra fecha: se consulta.

🚫

No procede y no contesta

Las filas con 98 salen de la base y se cuentan aparte. Las filas con 99 no se rellenan: se informa de la **tasa de respuesta de la variable**.

📏

Coherencia y rango

Lo que no cumple el diccionario se contrasta con la ficha en papel.

🔍

Extremos

Un valor dentro del rango pero alejado del resto se comprueba en el original: error de grabación, se corrige con la fuente; real, se conserva; sin comprobar, se analiza con él y sin él.

CLAVE

Las reglas se aplican **en orden sobre la copia**; el bruto sigue intacto.

Registro de depuración EK-2027, v1 → v2

N.º	FILA	REGLA APLICADA	ACCIÓN Y EVIDENCIA
1	R04	Duplicado de R03 (hogar y hora)	Fuera de la v2; sigue en el bruto
2	R08	No procede (P7 = No)	Fuera de la base de kits_ult
3	R06	No contesta	Ausencia; se informa
4	R07	Extremo comprobado	Se conserva: tique T-2027/04415

Quedan 2, 4, 4, 6, 9 (n = 5). Denominadores: 8 registros, 7 hogares distintos, 6 hogares a los que se aplicaba la pregunta y 5 respuestas válidas.

CLAVE

Tasa de respuesta de la variable: **5 ÷ 6 = 83,3 %**. Rellenar la ausencia con la media (5) inventa un dato: s bajaría de 2,646 a √(28 ÷ 5) ≈ 2,366 kits sin tener más información.

La regla de los cuartiles

Q1 − 1,5·RIC · Q3 + 1,5·RIC

Primer cuartil	=CUARTIL.INC(datos;1)	**Q1 = 4**
Tercer cuartil	=CUARTIL.INC(datos;3)	**Q3 = 6**
Rango intercuartílico	Q3 − Q1	**RIC = 2**
Límites de revisión	4 − 1,5 · 2 · 6 + 1,5 · 2	**1 y 9**
El 9 de H-042	justo en el límite, no fuera	**la regla solo avisa**

Adrián lo revisa porque cambia mucho la media, y eso basta. Itxaso localiza el tique T-2027/04415: pedido real de 9 kits para una comida familiar.

CLAVE

Un extremo comprobado y real **se conserva**. Lo que decide es la evidencia, no lo raro que parezca.

🙋 PREGUNTA PARA LA CLASE

¿Sobre cuántos hogares se calcula la **tasa de respuesta** de kits_ult?

A Sobre los 8 registros: 5 ÷ 8 = 62,5 %

B Sobre los 6 hogares a los que se aplicaba la pregunta: 5 ÷ 6 = 83,3 %

C Sobre los 7 hogares distintos: 5 ÷ 7 = 71,4 %

✔ Los 8 registros incluyen un duplicado (R04) y los 7 hogares, uno al que no se aplicaba la pregunta (R08, no procede). Queda la ausencia de R06 frente a 5 respuestas válidas (actividad 5).

9.3 · TABLA 1

Kits del último pedido (base: 5 hogares con dato)

KITS	HOGARES (N_i)	% (F_i)	% ACUMULADO
2	1	20,0	20,0
4	2	40,0	60,0
6	1	20,0	80,0
9	1	20,0	100,0
Total	**5**	**100,0**	

Frecuencia absoluta n_i; relativa $f_i = n_i \div n$; la acumulada solo tiene sentido si los valores tienen orden. Canal del último pedido, sobre los mismos 5: web 3 (60,0 %) y tienda 2 (40,0 %). Calidad del campo, otra tabla con otra base: de 7 hogares, 5 con dato (71,4 %), 1 no contesta (14,3 %) y 1 no procede (14,3 %).

CLAVE

«Un porcentaje sin su base es una opinión con decimales». «Web: 37,5 %» divide 3 entre los 8 registros: **no es el porcentaje de nada**.

9.3 · TABLA 2

Qué gráfico para qué variable

VARIABLE	PREGUNTA	GRÁFICO ADECUADO	EVITA
Nominal u ordinal	¿qué categoría pesa más?	barras (orden de la escala si es ordinal)	sectores con muchas porciones
Discreta, n pequeño	¿dónde se concentran?	diagrama de puntos	histograma de 5 casos
Continua	¿qué forma tiene?	histograma con intervalos iguales	barras sueltas por valor
Dos cuantitativas	¿van juntas?	dispersión	dos gráficos separados

Título informativo, unidad en el eje, base (n) y fuente con fecha; en barras, eje desde cero. Con dos categorías basta una frase: «3 de 5 pidieron por la web».

CLAVE

En tu sector. Marketing y Publicidad: el recuerdo de un anuncio, en barras con «base: personas que lo vieron». Gestión de Ventas y Espacios Comerciales: artículos por tique, en barras por valor con la base de tiques del periodo, para dimensionar cajas.

Desviación típica muestral de los cinco pedidos

Media: $\bar{x}$ = 25 ÷ 5 = 5 kits. Mediana: posición (5 + 1) ÷ 2 = 3.ª, 4 kits. Moda: 4 kits. Rango: 9 − 2 = 7 kits.

HOGAR	X (KITS)	X − $\bar{X}$	(X − $\bar{X}$)²
H-003	2	−3	9
H-011	4	−1	1
H-017	4	−1	1
H-024	6	1	1
H-042	9	4	16
Suma	**25**	**0**	**28**

Las desviaciones suman 0. s^2 = 28 ÷ 4 = 7; s = √7 ≈ 2,646 kits. CV = 2,646 ÷ 5 = 52,9 %. En la hoja: =PROMEDIO → 5; =MEDIANA → 4; =MODA.UNO → 4; =MAX−MIN → 7; =DESVEST.M → 2,646.

CLAVE

Si el tique hubiera dicho 3 kits (2, 3, 4, 4, 6): media 3,8; mediana 4; rango 4; s = √2,2 ≈ 1,483. **Si algún resultado no cambia, hay un número tecleado** donde debería haber una fórmula.

¿Muestra o población?



s · muestral (n − 1)

Los datos son una muestra y describes la variabilidad de la población de la que salen: dividir entre n la subestimaría. Los cinco hogares son una parte de los que probaron los kits: s = √(28 ÷ 4) ≈ **2,646** (=DESVEST.M).

VS



σ · poblacional (N)

Los datos son toda la población de interés (por ejemplo, los cinco pedidos de una semana y nada más). Aquí σ = √(28 ÷ 5) = √5,6 ≈ **2,366** (=DESVEST.P), que infravalora la dispersión.

CLAVE

Escríbelo siempre: «s (muestral, n − 1)» o «σ (poblacional, N)». Con los mismos datos, las dos fórmulas dan resultados distintos.

Coeficiente de asimetría con su convención

$$G1 = g1 \cdot \sqrt{[n(n-1)]} \div (n-2)$$

$m_2 = \Sigma(x - \bar{x})^2 \div n$	28 ÷ 5	**5,6**
$m_3 = \Sigma(x - \bar{x})^3 \div n$	(−27 − 1 − 1 + 1 + 64) ÷ 5 = 36 ÷ 5	**7,2**
g1 (de momentos)	7,2 ÷ 5,6^(3/2) = 7,2 ÷ 13,252	**≈ 0,543**
G1 (hoja de cálculo)	=COEFICIENTE.ASIMETRIA · 0,543 · √20 ÷ 3	**≈ 0,810**
Curtosis (G2, exceso)	=CURTOSIS · con cinco datos no se interpreta	**≈ 0,673**

Tres señales que se miran juntas: el gráfico, la media frente a la mediana (5 y 4) y un coeficiente con su convención. Ninguna basta sola.

CLAVE

Sin el 9 (2, 4, 4, 6): media = mediana = 4 y G1 = 0. **Toda la asimetría la produce un solo pedido**, comprobado y real: «G1 (hoja de cálculo) ≈ 0,810, n = 5».

OJO PROFESIONAL

La nota de Unai, antes de depurar

Martes 4 de mayo de 2027. Contiene cinco errores; dos líneas son correctas y otra es consecuencia de las anteriores.

N.º	APARTADO	CONTENIDO	REVISIÓN
2	Códigos	No contesta (R06) y no procede (R08) grabados como 0 kits	❌ 99 y 98 son etiquetas, no ceros
3	Duplicados	R04 se conserva: «es otra fila de la hoja»	❌ mismo hogar y hora que R03
4	Extremos	R07 (9 kits) se elimina por atípico; no se consulta el tique	❌ el tique confirma 9 kits
5	Medidas	2, 4, 4, 4, 6, 0, 0: media 2,86; mediana 4; σ 2,10 (=DESVEST.P)	❌ datos mal depurados y σ para una muestra
6	Canal	Web: 3 de 8 registros (37,5 %)	❌ base: 5 hogares con dato, 60,0 %
8	Conclusión	«El hogar típico pide menos de 3 kits»	↳ consecuencia de la línea 5

Las líneas 1 (variable y 8 registros grabados el 03/05/2027) y 7 (datos en bruto guardados sin tocar) son correctas.

CLAVE

Nota reescrita: 2, 4, 4, 6, 9 (n = 5 de 6 aplicables); media 5, mediana 4, s (n − 1) ≈ 2,646; web 3 de 5 (60,0 %); **pedido típico de 4 kits** y pedidos grandes reales.

Las cinco compras de la prueba



El equipo de análisis, para Nora

Viernes 14 de mayo de 2027: descripción de los kits del último pedido, con datos depurados y registro de cambios. Caso ficticio.

1. **Decisión, datos y alcance.** Qué caja isotérmica estándar prepara Itxaso y cuándo hace falta una segunda. 8 registros → 5 hogares con dato válido. Alcance: **estos cinco hogares**, una muestra de los que probaron los kits; se describen sin estimar con precisión la media de todos.

2. **Procedimiento.** 2, 4, 4, 6, 9 kits: media 25 ÷ 5 = 5; mediana (3.ª posición) 4; moda 4; rango 7; $s = \sqrt{(28 \div 4)} \approx 2{,}646$; $G1 \approx 0{,}810$. Sin 98 ni 99, sin el duplicado R04 y con R07 comprobado.

3. **Comprobación e interpretación.** Diagrama de puntos: tres pedidos en 4 o menos y uno aislado en 9. Media mayor que mediana y G1 positivo: cola derecha; sin el 9, media 4, mediana 4 y G1 = 0.

4. **Qué permite y qué no.** El pedido típico es de 4 kits y existen pedidos grandes reales. No permite afirmar que «los hogares piden 5 kits de media» ni estimar un porcentaje de pedidos grandes con cinco casos.

✔ Caja estándar para **4 kits** y una segunda caja para pedidos mayores, revisable con los datos de la encuesta completa. La media (5) sirve para totales: cinco pedidos son 25 kits que preparar.

🙋 PREGUNTA PARA LA CLASE

Miércoles 12 de mayo. En el importe del último pedido aparece **380 €** junto a 21, 32, 38 y 44. ¿Qué debe hacer Adrián?

A Dejarlo: la última vez el dato grande era verdad

B Borrarlo: sube la media

C Comprobarlo en el albarán antes de decidir

✔ El albarán dice 38,00 €: error de grabación. Se corrige con la fuente y se anota en el registro. La media baja de 103,00 € a 34,60 € y la mediana sigue en 38 €. «Ni borrar ni conservar por costumbre: comprobar».

Datos depurados y descripción

Con los datos ficticios o propios autorizados de tu investigación:

1

Diccionario y copias

Unidad de análisis, tipo, unidad, códigos y reglas de validación; datos en bruto intactos y copia depurada con versión.

→ 2

Depura y tabula

Registro de depuración (fila, regla, acción y evidencia) y tabla de frecuencias con base y un gráfico adecuado.

→ 3

Describe y decide

Media, mediana, moda, rango y s con fórmulas visibles y celda de control; forma con gráfico y coeficiente con su convención; decisión, responsable y fecha de revisión.

CLAVE

Comprueba tu paso: **ningún código convertido en cero**, denominadores visibles y la hoja se recalcula al cambiar un dato. Modelo: 2, 4, 4, 6, 9; s = √7.

APRENDE CON LA IA

La respuesta que recibió Unai

Unai preguntó si vale la pena calcular la mediana además de la media. Pulsa cada tarjeta.

«La mediana siempre coincide con la media»

BULO

Con 2, 4, 4, 6, 9 la media es 5 y la mediana, 4. Solo coinciden sin el 9 (2, 4, 4, 6): las dos valen 4.

«La mediana no cambia con ningún dato»

BULO

Resiste mejor los extremos, pero no es inmune: sin el 2 (4, 4, 6, 9) pasa de 4 a (4 + 6) ÷ 2 = 5.

«Calcula solo la media, que es la medida estándar»

BULO

Con cola derecha, la mediana describe el pedido típico (4) y la media sirve para totales (25 kits). Itxaso necesita las dos y el máximo comprobado.

«El extremo mueve la media un kit entero y no mueve la mediana»

REALIDAD

Sin el 9, la media baja de 5 a 4 y la mediana sigue en 4. Las respuestas rotundas («siempre», «nunca») se comprueban con un contraejemplo.

REPASO

Ideas clave

- Diccionario de datos: unidad de análisis, tipo, unidad, códigos, rango y regla de cada variable.
- Bruto intacto; la depuración, sobre una copia versionada y con reglas escritas antes de mirar los resultados.
- Denominadores siempre visibles: 8 registros, 7 hogares, 6 aplicables, 5 válidas; tasa de respuesta 5 ÷ 6 = 83,3 %.
- 2, 4, 4, 6, 9: media 5, mediana 4, moda 4, rango 7 y $s = \sqrt{7} \approx 2{,}646$ kits ($n - 1$).
- Con cola derecha, la mediana da el pedido típico (4 kits) y la media sirve para totales.
- No contesta (99), no procede (98) y cero son distintos: los códigos son etiquetas; promediarlos da 28,25 kits, que no significa nada.
- Duplicado = mismo hogar y misma hora; un extremo se comprueba en el original y, si es real, se conserva (tique T-2027/04415).
- Cada porcentaje, sobre la base de su pregunta: web 3 de 5 (60,0 %), no 3 de 8 (37,5 %).
- Forma: gráfico, media frente a mediana y coeficiente con su convención (G1 de la hoja ≈ 0,810, n = 5); sin el 9, G1 = 0.



¡Ahora te toca a ti!

Repasa, practica y comprueba lo que has aprendido.

📝 **Resumen**

✅ **Test**

👆 **Interactivos**

Libro, presentaciones, resúmenes y tests gratis en **josesande.com**