Trabajo práctico 1: Los datos y sus variables

AL TERMINAR ESTA GUÍA

- Reconocer en un relevamiento la población, la muestra, la unidad de análisis, las variables, el parámetro y el estadístico.
- Clasificar variables por tipo y escala de medición, y saber qué operaciones tienen sentido con cada una.
- Pasar información suelta a una matriz de datos y documentarla en un diccionario de variables, con los datos faltantes marcados.
- Abrir una base real, leer sus códigos antes de calcular y distinguir el error aleatorio del sesgo.

PARA EL DOCENTE

Los ejercicios 1 a 3 son para el encuentro práctico y casi no llevan cuentas; el 3 funciona bien en parejas. El 4 necesita computadora: sirve para un encuentro virtual o como tarea, y prepara la base que se usa en las semanas siguientes.

El 5 y el 6 son para discutir en grupos. El 6 vuelve en la unidad de muestreo: conviene retomarlo ahí.

EJERCICIO 1 Cortes de agua en un municipio

A mano

Un municipio del interior tiene 14.000 hogares con conexión a la red de agua. Para saber cuántos sufrieron cortes el último mes, la oficina de estadística sorteó 500 hogares del padrón de la cooperativa de agua y fue a encuestarlos. En cada hogar respondió una persona mayor de edad, que contestó si hubo cortes, cuántos y cuántas horas duró el más largo.

De los 500 hogares, 259 dijeron haber tenido al menos un corte.

a) ¿Cuál es la población? ¿Cuál es la muestra? ¿Cuál es la unidad de análisis, y quién es el que responde?

> RESOLUCIÓN
>
> La población son los 14.000 hogares conectados a la red; la muestra, los 500 que se sortearon y se encuestaron. La unidad de análisis es el hogar: de él se dice si tuvo cortes.
>
> El que responde es una persona del hogar, que habla por todos. Conviene no confundir las dos cosas: la persona es el informante, pero la fila de la base es el hogar.

b) Nombrá las variables que se relevaron y decí de qué tipo es cada una.

RESOLUCIÓN

Si hubo cortes o no: cualitativa, con dos categorías (dicotómica). La cantidad de cortes: cuantitativa discreta, porque se cuenta. Las horas del corte más largo: cuantitativa continua, porque se mide y podría tomar cualquier valor en un intervalo.

c) ¿Qué proporción de los hogares encuestados tuvo cortes? Expresala también en porcentaje.

RESOLUCIÓN

$$\hat{p} = \frac{259}{500} = 0{,}518 = 51{,}8\ \%$$

d) El intendente declara: «El 51,8 % de los hogares del municipio tuvo cortes». ¿Ese número es un parámetro o un estadístico? ¿Cuántos hogares del municipio estimás que tuvieron cortes?

RESOLUCIÓN

Es un estadístico: se calculó con los 500 hogares de la muestra. El parámetro es la proporción de hogares con cortes entre los 14.000, y no se conoce: para saberlo habría que preguntarles a todos. La frase del intendente presenta el estadístico como si fuera el parámetro; lo correcto es decir «se estima que».

$$0{,}518 \cdot 14.000 \approx 7.252 \text{ hogares}$$

Cada hogar encuestado «representa» a N/n hogares de la población. Es la misma idea que el ponderador de la EPH del ejercicio 4.

e) ¿Qué fracción de los hogares se encuestó? Si se hubieran sorteado otros 500 hogares, ¿el porcentaje daría lo mismo? ¿Qué cambia de una muestra a otra y qué no?

RESOLUCIÓN

$$\frac{n}{N} = \frac{500}{14.000} = 0{,}036 \approx 3{,}6\ \%$$

Con otra muestra el porcentaje cambiaría un poco: el estadístico varía de muestra en muestra. El parámetro no cambia, porque es una propiedad de la población. Cuánto puede variar el estadístico, y cómo acotarlo, es el tema de la segunda mitad de la materia.

EJERCICIO 2 La historia clínica

A mano

La ficha de ingreso de un centro de salud registra estos datos de cada paciente:

Variable	Cómo se registra
Edad	Años cumplidos
Sexo	Varón o mujer
Grupo sanguíneo	0, A, B o AB
Temperatura	En °C, con un decimal
Peso	En kg, con un decimal
Consultas en el último año	Cantidad
Nivel educativo	Primario, secundario o superior; completo o no
Dolor	De 0 (nada) a 10 (el peor imaginable)
Obra social	Sí o no
DNI	Número de documento
Satisfacción con la atención	Muy mala, mala, regular, buena o muy buena

a) Clasificá cada variable: cualitativa o cuantitativa; si es cualitativa, nominal u ordinal; si es cuantitativa, discreta o continua. Indicá también su escala de medición: nominal, ordinal, de intervalo o de razón.

RESOLUCIÓN

Variable	Tipo	Escala
Edad	Cuantitativa continua (se registra discreta)	Razón
Sexo	Cualitativa nominal	Nominal
Grupo sanguíneo	Cualitativa nominal	Nominal
Temperatura	Cuantitativa continua	Intervalo
Peso	Cuantitativa continua	Razón
Consultas en el último año	Cuantitativa discreta	Razón
Nivel educativo	Cualitativa ordinal	Ordinal
Dolor	Cualitativa ordinal (a menudo se trata como cuantitativa)	Ordinal
Obra social	Cualitativa nominal (dicotómica)	Nominal
DNI	Identificador: no es una variable para analizar	Nominal
Satisfacción con la atención	Cualitativa ordinal	Ordinal

Hay casos discutibles, y está bien que se discutan. La edad es continua —el tiempo pasa sin saltos—, pero al registrarla en años cumplidos queda discreta. El dolor de 0 a 10 es ordinal: no está claro que pasar de 2 a 4 sea lo mismo que pasar de 7 a 9, aunque en la práctica se lo promedia.

b) ¿Por qué la temperatura en grados Celsius no es de razón, y el peso sí? ¿Tiene sentido decir que un paciente con 40 °C tiene «el doble de temperatura» que algo a 20 °C?

RESOLUCIÓN

En una escala de razón el cero significa ausencia de lo que se mide: 0 kg es no pesar nada, y por eso 80 kg es el doble que 40 kg. El cero de la escala Celsius es arbitrario —el punto de congelamiento del agua—, así que los cocientes no tienen sentido: 40 °C no es el doble de calor que 20 °C. Sí tienen sentido las diferencias: de 36 a 38 hay lo mismo que de 38 a 40. Eso es una escala de intervalo.

c) El DNI se escribe con números y el grupo sanguíneo puede ser «0». ¿Los hace eso cuantitativos? Para cada variable de la tabla, ¿qué operación tiene sentido: contar, ordenar, restar o dividir?

RESOLUCIÓN

No: que algo se escriba con números no lo hace cuantitativo. Un número de DNI no mide nada, y sacar el DNI promedio de los pacientes no tiene sentido. El 0 del grupo sanguíneo es una etiqueta: dice que la sangre no tiene los antígenos A ni B, y no se suma ni se ordena con nada.

Las operaciones se acumulan con la escala. En las nominales (sexo, grupo sanguíneo, obra social) solo se puede contar cuántos hay en cada categoría. En las ordinales (nivel educativo, dolor, satisfacción) además se puede ordenar. En la temperatura además se puede restar. En las de razón (edad, peso, consultas) además se puede dividir.

d) La edad también se puede preguntar en tramos: menos de 20, de 20 a 39, de 40 a 59, 60 y más. ¿Qué tipo de variable queda? ¿Qué se pierde? ¿Se puede volver atrás?

RESOLUCIÓN

Queda una cualitativa ordinal: los tramos se pueden ordenar, pero ya no se puede calcular la edad media ni saber si alguien tiene 21 o 39. Se pierde información y no se puede recuperar. Al revés sí se puede: con la edad en años se arman los tramos que se quieran después. Por eso conviene registrar la variable con el mayor detalle posible y agrupar al analizar.

EJERCICIO 3 De las fichas a la matriz de datos

A mano

Un equipo de enfermería anotó en papel los datos de seis pacientes, con sus palabras y sin un formato fijo:

Ficha 1. Varón de 41 años. Completó la primaria y no siguió. No tiene hijos. Pesa 69,8 kg. No tiene obra social.

Ficha 2. Mujer de 70 años. Empezó el secundario y no lo terminó. Tiene cuatro hijos. Pesa 61,7 kg. Tiene obra social.

Ficha 3. Mujer de 32 años. Terminó el secundario. Tiene un hijo. Pesa 59,7 kg. No tiene obra social.

Ficha 4. Mujer de 40 años. Completó la primaria y no siguió. No tiene hijos. Pesa 67,9 kg. Tiene obra social.

Ficha 5. Varón de 54 años. Está cursando una carrera en la universidad. Tiene un hijo. Pesa 87,8 kg. Tiene obra social.

Ficha 6. Mujer de 76 años. Completó la primaria y no siguió. Tiene dos hijos. No se registró el peso. Se atiende en el hospital público, sin obra social.

a) Armá la matriz de datos: una fila por paciente y una columna por variable. Decidí cómo cargar cada respuesta, con códigos numéricos donde haga falta.

RESOLUCIÓN

ID	SEXO	EDAD	NIVEL	HIJOS	PESO	OBRA
1	1	41	1	0	69,8	0
2	2	70	2	4	61,7	1
3	2	32	3	1	59,7	0
4	2	40	1	0	67,9	1
5	1	54	4	1	87,8	1
6	2	76	1	2	−9	0

Una columna ID permite volver a la ficha original si algo no cierra. Los nombres cortos, en mayúsculas y sin espacios, son los que aceptan sin protestar R, SPSS y Excel. Las frases de cada ficha se tradujeron a categorías: «está cursando en la universidad» es superior incompleto; «se atiende en el hospital público» es no tener obra social.

b) Escribí el diccionario de variables: para cada columna, su nombre, qué mide, el tipo de variable y los valores posibles o la unidad.

RESOLUCIÓN

Columna	Qué mide	Tipo	Valores o unidad
ID	Número de ficha	Identificador	1 a 6
SEXO	Sexo	Cualitativa nominal	1 = varón · 2 = mujer
EDAD	Edad	Cuantitativa	Años cumplidos
NIVEL	Máximo nivel educativo	Cualitativa ordinal	1 = primario completo · 2 = secundario incompleto · 3 = secundario completo · 4 = superior incompleto · 5 = superior completo
HIJOS	Cantidad de hijos	Cuantitativa discreta	0 o más
PESO	Peso	Cuantitativa continua	Kilos, un decimal · −9 = sin dato
OBRA	Tiene obra social	Cualitativa nominal	0 = no · 1 = sí

Los códigos del nivel educativo siguen el orden de los niveles. No es obligatorio, pero ayuda: con este diccionario, «secundario completo o más» es NIVEL ≥ 3. El ejercicio 4 muestra lo que pasa cuando no se respeta el orden.

c) La ficha 6 no trae el peso. ¿Qué cargás en esa celda? ¿Por qué no un 0?

> RESOLUCIÓN
>
> Un código de dato faltante, que figure en el diccionario: la celda vacía, NA en R o un valor imposible como −9, declarado como faltante en SPSS. Un 0 sería un peso, y uno imposible; al promediar, arrastraría la media hacia abajo sin que nadie lo note. Lo que no se sabe tiene que quedar marcado como que no se sabe.

d) Con la matriz armada, contestá: ¿cuántos pacientes tienen obra social? ¿Cuántos terminaron al menos el secundario? ¿Cuántas de las mujeres tienen hijos?

> RESOLUCIÓN
>
> Tienen obra social 3 de los 6 (OBRA = 1). Terminaron al menos el secundario 2 (NIVEL ≥ 3). De las 4 mujeres, 3 tienen hijos (SEXO = 2 y HIJOS > 0).
>
> Con las fichas en papel, cada pregunta obliga a releer las seis. Con la matriz, cada una es un filtro sobre una o dos columnas: para eso se arma.

EJERCICIO 4 Una base real: la EPH de Gran Catamarca

Con software

La base «Ocupación e ingresos en Gran Catamarca» es un extracto real de la Encuesta Permanente de Hogares del INDEC, primer trimestre de 2026. Se baja de estadistica.ar/datos/eph-catamarca-2026t1, donde está la descripción de cada variable y de sus códigos. El separador es punto y coma.

Se puede usar R, Python, Excel, InfoStat o SPSS. Antes de calcular nada, leé la descripción de la base: la mitad del trabajo con datos reales es entender qué significa cada código.

a) ¿Cuántas filas y cuántas columnas tiene la base? ¿Cuál es la unidad de análisis? ¿Cuántas viviendas y cuántos hogares distintos hay?

> RESOLUCIÓN
>
> Tiene 1.339 filas y 12 columnas. La unidad de análisis es la persona: cada fila es una persona.
>
> Las personas están agrupadas: CODUSU identifica la vivienda (hay 426 distintas) y CODUSU junto con NRO_HOGAR, el hogar (hay 429). Que haya más hogares que viviendas quiere decir que en algunas viviendas vive más de un hogar. En R: nrow(unique(d[c("CODUSU", "NRO_HOGAR")])).

b) Clasificá las doce variables. Separá los identificadores y los ponderadores, que no son variables para describir.

RESOLUCIÓN

Variable	Qué es	Tipo
CODUSU	Identificador de la vivienda	Identificador
NRO_HOGAR	Número de hogar dentro de la vivienda	Identificador
COMPONENTE	Número de persona dentro del hogar	Identificador
AGLOMERADO	Aglomerado urbano (22 = Gran Catamarca)	Cualitativa nominal (constante en este extracto)
CH04	Sexo	Cualitativa nominal
CH06	Edad en años cumplidos	Cuantitativa (razón)
NIVEL_ED	Máximo nivel educativo	Cualitativa ordinal
ESTADO	Condición de actividad	Cualitativa nominal
CAT_OCUP	Categoría ocupacional	Cualitativa nominal
P21	Ingreso de la ocupación principal	Cuantitativa continua (razón)
PONDERA	Ponderador de población	Ponderador
PONDIIO	Ponderador del ingreso	Ponderador

Casi todas las cualitativas están cargadas con números: el tipo de una variable no se lee en la columna sino en el diccionario.

c) Armá la tabla de frecuencias de ESTADO. ¿Qué significan los códigos 0 y 4? Calculá el promedio de la columna ESTADO: ¿qué quiere decir ese número?

RESOLUCIÓN

ESTADO	Personas	%
0 · entrevista individual no realizada	4	0,3
1 · ocupado	631	47,1
2 · desocupado	26	1,9
3 · inactivo	545	40,7
4 · menor de 10 años	133	9,9
Total	**1.339**	**100,0**

El 0 son 4 personas de las que no se obtuvo la entrevista individual: es un dato faltante. El 4 son los menores de 10 años, a los que la EPH no les pregunta por el trabajo. El promedio de ESTADO da 2,13, y no quiere decir nada: es una variable nominal y sus códigos son etiquetas, no cantidades.

d) Algunos códigos no son valores. ¿Cuántas personas tienen CH06 = −1? ¿Cuántas tienen P21 = 0 y cuántas P21 = −9? ¿Qué significa cada uno?

RESOLUCIÓN

CH06 = −1 lo tienen 7 personas: quiere decir «menos de un año», no una edad negativa. En P21 hay 715 ceros, que significan «no corresponde» (la persona no está ocupada, no es que gane cero), y 30 códigos −9 de no respuesta.

Antes de cualquier cuenta hay que decidir qué hacer con ellos: el −1 de la edad se recodifica a 0 años cumplidos, y los 0 y −9 de P21 quedan afuera al describir el ingreso. Promediar la columna tal como viene mezcla montos con códigos.

e) ¿Cuántas personas terminaron al menos el secundario? Un compañero filtró NIVEL_ED ≥ 4. Mirá en la descripción qué significa el código 7 y corregí la cuenta.

RESOLUCIÓN

Con NIVEL_ED ≥ 4 salen 819 personas, pero el 7 es «sin instrucción»: el nivel más bajo tiene el código más alto. Las que terminaron al menos el secundario son las de los códigos 4, 5 y 6: 765. El filtro del compañero sumó 54 personas sin instrucción.

Es el error más difícil de ver, porque la cuenta no da nada raro. En una variable ordinal, los códigos no siempre siguen el orden de las categorías: hay que leer el diccionario antes de filtrar u ordenar.

f) Entre las personas de 10 años o más con entrevista (ESTADO 1, 2 o 3), ¿qué porcentaje está ocupado? Calculalo contando filas y después ponderando por PONDERA. ¿Cuál de los dos números estima la tasa de Gran Catamarca? ¿Cuántas personas representa la base?

RESOLUCIÓN

Contando filas, 52,5 % de las 1.202 personas están ocupadas; ponderando, 50,9 %.

Cada persona encuestada representa a PONDERA personas de la población, y PONDERA suma 236.275: es lo que la encuesta estima que vive en Gran Catamarca. Los dos porcentajes son estadísticos, pero el que estima el parámetro —el porcentaje de ocupados en toda la población— es el ponderado, porque la muestra no eligió a todas las personas con la misma probabilidad. El que cuenta filas describe a los encuestados.

EJERCICIO 5 ¿De dónde sale el dato?

Para discutir

Para cada pregunta, decidí de qué tipo de fuente saldría la respuesta: un censo, una encuesta por muestreo o un registro administrativo. Justificá, y decí qué limitación tiene esa fuente.

a) ¿Cuántos nacimientos hubo en la provincia el año pasado, por departamento de residencia de la madre?

RESOLUCIÓN

Un registro administrativo: las estadísticas vitales, que salen de las actas del Registro Civil. Cubre todos los nacimientos inscriptos, sin muestra. Sus limitaciones son las de la carga: nacimientos que se inscriben tarde, o un domicilio mal registrado, que manda el nacimiento a otro departamento.

b) ¿Cuál fue la tasa de desocupación de Gran Catamarca en el último trimestre?

RESOLUCIÓN

Una encuesta: la EPH. No hay un registro de personas desocupadas —el que busca trabajo no se anota en ningún lado— y un censo cada trimestre es impagable. La tasa es una estimación, con un error muestral que hay que informar.

c) ¿Cuántas personas viven en cada localidad de la provincia, por edad y sexo?

RESOLUCIÓN

Un censo: es la única fuente que llega a todas las localidades, incluso a las más chicas, donde una muestra tendría muy pocos casos. Su limitación es la frecuencia: se hace cada diez años, y entre censos hay que usar proyecciones.

d) ¿Cuántos estudiantes de esta facultad trabajan más de 20 horas por semana?

RESOLUCIÓN

Una encuesta a los estudiantes, o una pregunta agregada al formulario de inscripción. Los registros de alumnos de la facultad existen, pero no tienen esa variable: un registro solo trae lo que el trámite necesita saber. Si se hace una encuesta, hay que cuidar que respondan también los que tienen menos tiempo, que son justamente los que trabajan.

e) Con los tres tipos de fuente a la vista: ¿cuál tiene error muestral? ¿Puede tener errores una fuente que cubre a todos?

RESOLUCIÓN

Solo la encuesta tiene error muestral, porque solo ella mira una parte. Pero el censo y los registros también tienen errores: personas que no se censaron, preguntas mal respondidas, datos mal cargados. Cubrir a todos elimina el error de muestreo, no los demás.

EJERCICIO 6 ¿Qué opina la ciudad?

Para discutir

Se quiere saber qué proporción de los vecinos de una ciudad está a favor de cerrar el centro al tránsito de autos. En cada caso, decidí si el resultado sirve para estimar esa proporción y, si no sirve, qué falla.

a) Una radio pide a los oyentes que voten por WhatsApp. Llegan 2.400 mensajes y el 81 % está a favor. ¿Ese 81 % estima la opinión de la ciudad? ¿Se arreglaría con 10.000 mensajes?

RESOLUCIÓN

No. Votan los que escuchan esa radio y, entre ellos, los que tienen ganas de opinar: suelen ser los que tienen una postura fuerte. Es un sesgo de selección, por autoselección. Más mensajes no lo arreglan: 10.000 personas elegidas del mismo modo dan una estimación igual de desviada, solo que más precisa alrededor del valor equivocado.

b) Una consultora sortea 400 números de la guía de teléfonos fijos y llama.

RESOLUCIÓN

El sorteo está bien, pero la lista de la que se sortea no cubre a la población: hoy muchos hogares no tienen teléfono fijo, sobre todo los más jóvenes. Es un sesgo de cobertura: el marco muestral deja afuera a una parte de la población, que puede opinar distinto.

c) El cuestionario pregunta: «¿No le parece que el centro sería más seguro y más lindo sin autos?».

RESOLUCIÓN

La pregunta empuja hacia el sí: es un sesgo de medición, por una pregunta inductiva. Aunque la muestra fuera perfecta, el instrumento mide otra cosa. Una versión neutral sería: «¿Está a favor o en contra de cerrar el centro al tránsito de autos?».

d) Dos encuestas bien hechas, cada una con una muestra aleatoria de 400 vecinos, dan 54 % y 49 % a favor. ¿Alguna de las dos está sesgada?

RESOLUCIÓN

No necesariamente. Dos muestras aleatorias distintas dan resultados distintos: es el error aleatorio, o muestral. A diferencia del sesgo, no empuja siempre para el mismo lado y se achica al aumentar el tamaño de la muestra. El sesgo es un error sistemático: no se va con más casos, solo con un mejor diseño.

---

estadistica.ar/docentes/trabajos-practicos/semana-1 · Las lecturas de cada tema y la hoja de fórmulas están en el sitio.