Trabajo práctico 8: Muestreo y distribuciones muestrales

AL TERMINAR ESTA GUÍA

- Seleccionar una muestra aleatoria simple con una tabla de números aleatorios y una muestra sistemática.
- Repartir una muestra entre estratos con afijación igual, proporcional y de Neyman, y estimar con el estimador estratificado.
- Construir la distribución muestral de la media y comprobar que es insesgada y que su varianza es σ^2/n.
- Distinguir el error estándar del desvío estándar y usar el teorema central del límite para calcular probabilidades sobre la media y la proporción muestrales.

PARA EL DOCENTE

Los ejercicios 1 a 4 son para el encuentro práctico; el 3 es el que mejor muestra qué es una distribución muestral y conviene hacerlo completo, en grupos, repartiendo las 16 muestras. El 4 necesita la tabla de la normal. El 5 necesita computadora: sirve para un encuentro virtual o como tarea, y se complementa con el simulador del teorema central del límite.

El 6 es para discutir en grupos. Para el 1 se puede usar cualquier fila de la tabla de números aleatorios del sitio: cada grupo saca una muestra distinta y se comparan las medias en el pizarrón.

EJERCICIO 1 Cuarenta escuelas rurales

A mano

Un departamento tiene 40 escuelas rurales, numeradas del 01 al 40 en el listado del ministerio. Se quiere visitar 8 para relevar el estado de los edificios. La matrícula de cada escuela, en el orden del listado, es:

Escuelas 01 a 10	15	31	119	31	9	35	9	7	55	56
Escuelas 11 a 20	58	25	30	18	63	23	23	119	26	6
Escuelas 21 a 30	53	6	7	49	6	18	24	41	48	58
Escuelas 31 a 40	142	6	62	133	16	131	72	54	33	23

Para sortear se usa esta fila de una tabla de números aleatorios, que sigue de la primera línea a la segunda:

Dígitos 1 a 25	22653	28684	91341	10784	27314

Dígitos 26 a 50	70794	77994	65628	02441	17797

a) ¿Cuál es la población, la unidad de muestreo y el marco muestral? ¿Por qué no se visitan las 40?

> RESOLUCIÓN
>
> La población son las 40 escuelas rurales del departamento; cada escuela es la unidad de muestreo (y también la de análisis), y el marco es el listado del ministerio. Se visita una muestra porque relevar un edificio cuesta tiempo y viajes: con 8 escuelas bien elegidas se puede estimar lo que pasa en todas. Si el listado estuviera desactualizado —una escuela cerrada, otra nueva que no figura—, el problema sería del marco y ningún sorteo lo arreglaría.

b) Seleccioná una muestra aleatoria simple de 8 escuelas: leé los dígitos de a pares, de izquierda a derecha, quedate con los números del 01 al 40 y salteá los repetidos. ¿Qué escuelas salen?

> RESOLUCIÓN
>
> Salen las escuelas 22, 32, 34, 11, 07, 27, 31, 28. Hicieron falta 20 pares: se descartaron 11 números fuera de rango (el primero, 65) y el 07, que ya había salido.
>
> Se saltea el repetido porque el muestreo es sin reposición: una escuela se visita una sola vez. Con la tabla, cada escuela tiene la misma probabilidad de salir, y también cada conjunto de ocho: eso es lo que define al muestreo aleatorio simple.

c) Seleccioná ahora una muestra sistemática de 8 escuelas. ¿Cada cuántas escuelas se elige una? Si el arranque aleatorio, entre 1 y 5, salió 3, ¿qué escuelas salen?

> RESOLUCIÓN
>
> $$k = \frac{N}{n} = \frac{40}{8} = 5$$
>
> Se elige una de cada 5, empezando por la 03: 03, 08, 13, 18, 23, 28, 33, 38. Es más fácil de aplicar que la aleatoria simple —un solo número al azar—, y funciona bien si el orden del listado no tiene un patrón que se repita cada 5 escuelas. Si el listado estuviera ordenado por tamaño, la muestra quedaría repartida entre chicas y grandes, que es una ventaja.

d) Calculá la matrícula media de cada muestra y la de las 40 escuelas. ¿Por qué ninguna muestra da exactamente la media de la población?

> RESOLUCIÓN
>
> $$\bar{x}_{\text{MAS}} = 52{,}38 \qquad \bar{x}_{\text{sist}} = 54{,}88 \qquad \mu = \frac{1740}{40} = 43{,}50$$
>
> La media de la población es un parámetro: hay uno solo y, en la práctica, no se conoce. Cada muestra da un estadístico distinto, porque incluye escuelas distintas. Esa diferencia es el error de muestreo; no es una equivocación, sino la consecuencia de mirar solo 8 de 40. Lo que sí se puede hacer es medir cuánto varía de muestra a muestra, y eso es el error estándar.

EJERCICIO 2 El gasto en transporte por zonas

A mano

Una encuesta quiere estimar el gasto mensual en transporte de los 18.000 hogares de una ciudad, con una muestra de 200 hogares. Los hogares se dividen en tres zonas, y por un estudio anterior se conoce aproximadamente cuánto varía el gasto dentro de cada una (el desvío estándar, en miles de pesos):

Zona	Hogares	Desvío del gasto
Urbana	9.900	18
Periurbana	5.400	29
Rural	2.700	42

a) ¿Por qué conviene estratificar por zona en lugar de tomar una muestra aleatoria simple de toda la ciudad?

> RESOLUCIÓN
>
> Porque el gasto en transporte cambia mucho de una zona a otra —en la rural los viajes son más largos— y dentro de cada zona los hogares se parecen más entre sí. Estratificar garantiza que las tres zonas estén representadas y saca del error de muestreo la variación entre zonas: con el mismo tamaño de muestra, la estimación es más precisa. También permite dar resultados por zona.

b) Repartí los 200 hogares entre las zonas con afijación igual, proporcional y óptima de Neyman.

> RESOLUCIÓN
>
> Proporcional: $n_h = n\,\dfrac{N_h}{N}$ Neyman: $n_h = n\,\dfrac{N_h S_h}{\sum N_h S_h}$
>
Zona	N_h · S_h	Igual	Proporcional	Neyman
Urbana	178.200	67	110	79
Periurbana	156.600	67	60	70
Rural	113.400	66	30	51
Total	**448.200**	**200**	**200**	**200**
>
> La igual reparte lo mismo a cada zona, sin mirar su tamaño: sobrerrepresenta a la rural, que tiene el 15 % de los hogares. La proporcional respeta el peso de cada zona. La de Neyman, además, le da más muestra a la zona donde el gasto varía más: la rural pasa de 30 a 51 hogares, porque ahí cada hogar aporta más información. En Neyman, los decimales se redondearon de modo que la suma siga dando 200.

c) Se hizo la encuesta con afijación proporcional. Las medias de gasto en cada zona dieron, en miles de pesos: urbana 46,5 (desvío 19); periurbana 65,4 (desvío 26); rural 85,4 (desvío 45). Estimá el gasto medio de la ciudad y su error estándar.

> RESOLUCIÓN
>
> $$\bar{x}_{\text{st}} = \sum W_h \bar{x}_h = 0{,}55 \cdot 46{,}5 + 0{,}30 \cdot 65{,}4 + 0{,}15 \cdot 85{,}4 = 58{,}01$$

$$EE(\bar{x}_{\text{st}}) = \sqrt{\sum W_h^2 \frac{s_h^2}{n_h}} = \sqrt{0{,}55^2 \cdot \frac{19^2}{110} + 0{,}30^2 \cdot \frac{26^2}{60} + 0{,}15^2 \cdot \frac{45^2}{30}} = 1{,}878$$

El gasto medio estimado es de 58,0 mil pesos por hogar, con un error estándar de 1,88 mil. Cada zona pesa según su proporción de hogares (W_h = N_h / N). Con afijación proporcional, esta media ponderada coincide con el promedio simple de los 200 hogares (58,01): la muestra ya tiene las proporciones de la población. Con la igual o la de Neyman, no; ahí promediar sin ponderar daría una estimación sesgada.

EJERCICIO 3 Todas las muestras posibles

A mano

En una cuadra viven cuatro familias. La cantidad de hijos de cada una es:

Familia	Hijos
A	6
B	5
C	3
D	1

Se toma una muestra de dos familias y se calcula el promedio de hijos.

a) Calculá la media μ y la varianza σ² de la población.

RESOLUCIÓN

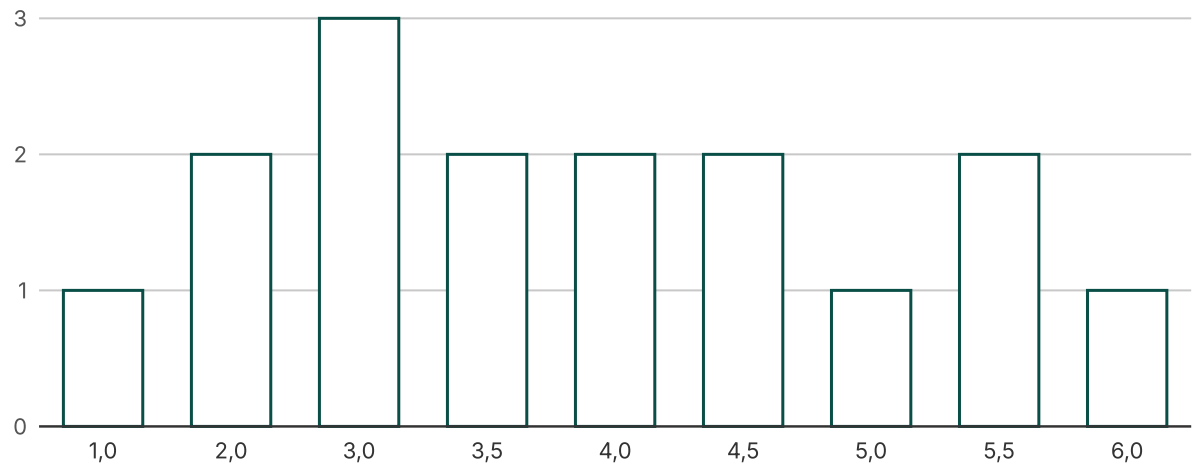
$$\mu = \frac{6+5+3+1}{4} = 3{,}75 \qquad \sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = 3{,}6875$$

Son parámetros: se pueden calcular porque esta población es tan chica que se conoce entera. La varianza se divide por N, no por N − 1, porque no es una estimación.

b) Con reposición, hay 16 muestras ordenadas posibles de dos familias (AA, AB, …, DD). Listalas, calculá la media de cada una y armá la distribución de frecuencias de la media muestral.

RESOLUCIÓN

Media muestral	Muestras	Probabilidad
1,0	1	1/16
2,0	2	2/16
3,0	3	3/16
3,5	2	2/16
4,0	2	2/16
4,5	2	2/16
5,0	1	1/16
5,5	2	2/16
6,0	1	1/16
Total	**16**	**1**



Eje vertical: cantidad de muestras.

Esta es la distribución muestral de la media: todos los valores que puede dar x̄ y con qué probabilidad, si cada muestra es igualmente probable. Se concentra en el centro y es más simétrica que la población: las muestras con dos valores extremos del mismo lado son pocas.

c) Calculá la media y la varianza de esa distribución. Comparalas con μ y con σ²/n.

RESOLUCIÓN

$$E(\bar{x}) = 3{,}75 = \mu \qquad V(\bar{x}) = 1{,}8438 = \frac{\sigma^2}{n} = \frac{3{,}6875}{2}$$

El promedio de todas las medias muestrales es la media de la población: la media muestral es un estimador insesgado. Y su varianza es la de la población dividida por n: promediar achica la variabilidad. El desvío de esa distribución, σ/√n, es el error estándar de la media.

d) Repetí sin reposición: hay 6 muestras posibles (AB, AC, AD, BC, BD, CD). ¿Sigue valiendo E(x̄) = μ? ¿Y la varianza?

RESOLUCIÓN

Muestra	Media
AB	5,5
AC	4,5
AD	3,5
BC	4,0
BD	3,0
CD	2,0

$E(\bar{x}) = 3{,}75 \qquad V(\bar{x}) = 1{,}2292 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} = \frac{3{,}6875}{2} \cdot \frac{2}{3}$

La media sigue siendo insesgada, pero la varianza es menor: al no repetir familias, la muestra cubre más de la población. El factor (N − n)/(N − 1) es la corrección por población finita. Con una muestra chica de una población grande vale casi 1, y por eso se suele omitir; acá, con la mitad de la población en la muestra, pesa mucho.

EJERCICIO 4 La espera en la guardia

A mano

En la guardia de un hospital, el tiempo de espera hasta la primera atención tiene media 45 minutos y desvío estándar 24 minutos. La distribución es muy asimétrica a la derecha: la mayoría espera poco y algunos esperan horas.

Cada mes, la dirección del hospital revisa una muestra aleatoria de 36 pacientes y calcula el tiempo de espera promedio.

a) ¿Cuál es el error estándar de la media? ¿Qué diferencia hay entre ese número y el desvío estándar de 24 minutos?

RESOLUCIÓN

$$EE(\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{24}{\sqrt{36}} = 4 \text{ minutos}$$

El desvío estándar (24 min) mide cuánto varía la espera de un paciente a otro. El error estándar (4 min) mide cuánto varía el promedio de 36 pacientes de una muestra a otra. Son cosas distintas: el desvío es una propiedad de la población y no cambia con n; el error estándar se achica cuando la muestra crece.

b) ¿Se puede usar la normal para la media muestral, si los tiempos de espera son tan asimétricos?

RESOLUCIÓN

Sí, por el teorema central del límite: la media de una muestra grande tiene distribución aproximadamente normal aunque la población no la tenga. Con n = 36 suele alcanzar, salvo poblaciones extremadamente asimétricas. Lo que no se puede es usar la normal para un paciente individual: la espera de uno solo sigue siendo asimétrica.

$$\bar{x} \dot{\sim} N(\mu = 45;\ EE = 4)$$

c) ¿Cuál es la probabilidad de que el promedio de un mes supere los 55 minutos, si el funcionamiento de la guardia no cambió? ¿Y de que quede por debajo de 40?

RESOLUCIÓN

$$P(\bar{x} > 55) = P\left(Z > \frac{55 - 45}{4}\right) = P(Z > 2{,}50) = 0{,}0062$$

$$P(\bar{x} < 40) = P\left(Z < \frac{40 - 45}{4}\right) = P(Z < -1{,}25) = 0{,}1056$$

Se tipifica con el error estándar, no con el desvío: la variable es el promedio. Un mes con promedio mayor a 55 minutos sería muy raro (0,6 %) si nada cambió, y es una señal para revisar qué pasó. Este razonamiento es la base de las pruebas de hipótesis.

d) ¿De cuántos pacientes tendría que ser la muestra para que el error estándar sea de 3 minutos como mucho?

RESOLUCIÓN

$$\frac{24}{\sqrt{n}} \leq 3 \ \Rightarrow\ n \geq \left(\frac{24}{3}\right)^2 = 64$$

Para achicar el error estándar a la mitad hay que cuadruplicar la muestra, porque n está bajo una raíz. Pasar de 36 a 64 pacientes cuesta mucho más de lo que mejora la precisión.

e) En la misma guardia, el 30 % de los pacientes se va sin ser atendido. Si se revisa una muestra de 200 pacientes, ¿cuál es el error estándar de la proporción muestral? ¿Cuál es la probabilidad de que en la muestra se vaya más del 35 %?

RESOLUCIÓN

$$EE(\hat{p}) = \sqrt{\frac{p\,(1-p)}{n}} = \sqrt{\frac{0{,}30 \cdot 0{,}70}{200}} = 0{,}0324$$

$$P(\hat{p} > 0{,}35) = P\left(Z > \frac{0{,}05}{0{,}0324}\right) = P(Z > 1{,}54) = 0{,}0618$$

La proporción muestral también es una media —de unos y ceros— y el teorema central del límite vale igual. Se puede usar la normal porque n·p = 60 y n·(1 − p) = 140 son mayores que 5.

EJERCICIO 5 Miles de muestras de la EPH

Con software

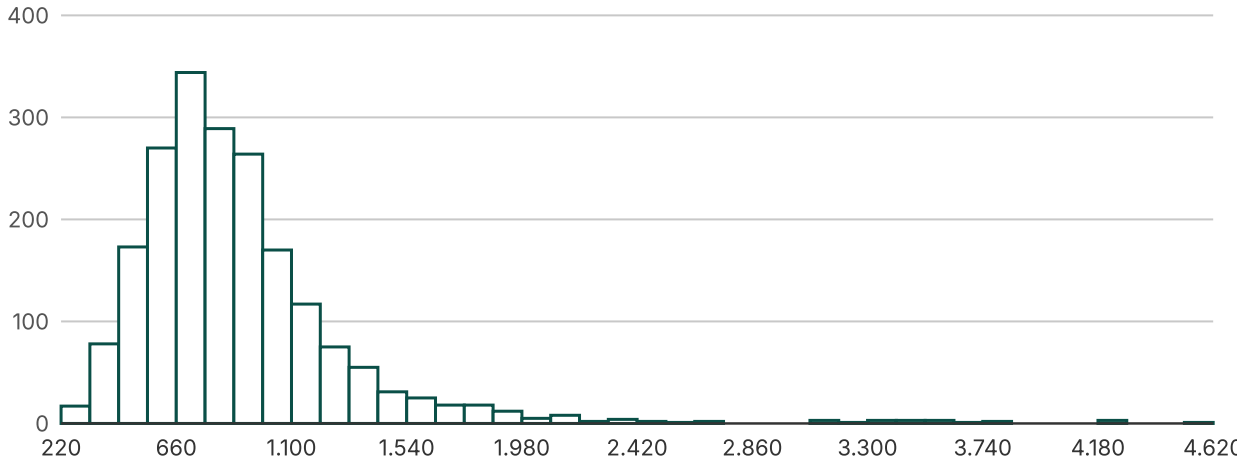
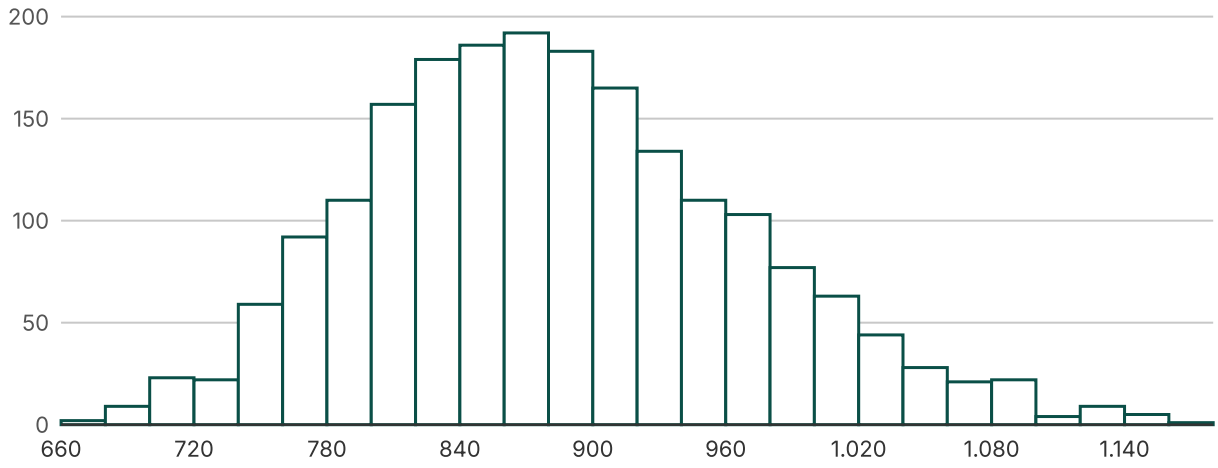
Se usa la base de la semana 1: «Ocupación e ingresos en Gran Catamarca», extracto de la EPH del primer trimestre de 2026 (estadistica.ar/datos/eph-catamarca-2026t1). Quedate con las personas ocupadas que declararon ingreso (ESTADO = 1 y P21 > 0).

Para este ejercicio, esas personas son la población: se conocen todas, y se puede comparar lo que dicen las muestras con el valor verdadero. En R, sample(x, n) toma una muestra sin reposición y replicate(2000, mean(sample(x, n))) repite el experimento.

a) ¿Cuántas personas tiene la población? Calculá la media, la mediana y el desvío estándar del ingreso (P21). ¿Qué forma tiene la distribución?

> RESOLUCIÓN
>
> Son 594 personas. La media es de $881.992, la mediana de $700.000 y el desvío estándar (dividiendo por N, porque es la población) de $934.992. La media queda por encima de la mediana: la distribución es asimétrica a la derecha, con una cola de ingresos altos (es el histograma de la semana 2).

b) Tomá 2.000 muestras de 5 personas y otras 2.000 de 100 personas. Para cada tamaño, calculá la media de cada muestra y hacé un histograma de las 2.000 medias. ¿Qué cambia entre uno y otro?

> RESOLUCIÓN
>
> En R: x ← eph$P21[eph$ESTADO == 1 & eph$P21 > 0]; m5 ← replicate(2000, mean(sample(x, 5))); m100 ← replicate(2000, mean(sample(x, 100))); hist(m5); hist(m100). Una corrida de 2.000 muestras de 5 personas (ingreso medio, en miles de pesos):
>
> 
>
>
> Eje vertical: cantidad de muestras.
>
> Y de 100 personas:
>
> 
>
>
> Eje vertical: cantidad de muestras.

Los dos histogramas se centran en la media de la población, $881.992: la media muestral es insesgada. Con n = 5 las medias todavía son asimétricas —una muestra que incluye un ingreso muy alto tira el promedio para arriba— y muy dispersas. Con n = 100 el histograma es una campana mucho más angosta. Es el teorema central del límite: la media se vuelve normal aunque los ingresos no lo sean.

c) Calculá el desvío estándar de las 2.000 medias para cada tamaño y comparalo con σ/√n. ¿Hace falta la corrección por población finita?

RESOLUCIÓN

$$\frac{\sigma}{\sqrt{5}} = 418.141 \qquad \frac{\sigma}{\sqrt{100}} = 93.499 \qquad \frac{\sigma}{\sqrt{100}}\sqrt{\frac{N-100}{N-1}} = 85.338$$

En la corrida de arriba, el desvío de las medias dio $436.896 con n = 5 y $85.011 con n = 100. Con 100 personas sobre 594, la muestra es el 17 % de la población y la corrección por población finita se nota: el error estándar corregido, $85.338, es el que coincide con la simulación. Con n = 5 la corrección casi no cambia nada (vale 0,997). Cada corrida da números algo distintos, pero siempre cerca de estos.

d) ¿Qué proporción de la población gana un millón de pesos o más? Si se toma una muestra de 100 personas, ¿cuál es el error estándar de la proporción muestral? Comprobalo con una simulación.

RESOLUCIÓN

$$p = 0{,}2980 \qquad EE(\hat{p}) = \sqrt{\frac{p\,(1-p)}{100}}\sqrt{\frac{N-100}{N-1}} = 0{,}0417$$

En R: sd(replicate(2000, mean(sample(x, 100) >= 1e6))). El desvío de las proporciones simuladas tiene que andar cerca de 0,042. La proporción es la media de una variable que vale 1 si la persona gana un millón o más y 0 si no: todo lo que vale para la media vale también para ella.

EJERCICIO 6 ¿Qué diseño conviene?

Para discutir

Para cada situación, decidí qué diseño de muestreo usarías (aleatorio simple, sistemático, estratificado o por conglomerados) o qué problema tiene el que se propone. Pensá en el marco muestral disponible y en el costo.

a) Se quiere estimar la proporción de estudiantes secundarios de la provincia que fuma. Hay un listado de escuelas, pero no de estudiantes.

RESOLUCIÓN

Por conglomerados: se sortean escuelas (o cursos dentro de las escuelas) y se encuesta a todos los estudiantes de las elegidas. No hace falta el listado de estudiantes, y los viajes se concentran en pocas escuelas. El costo es precisión: los estudiantes de una misma escuela se parecen entre sí, y con la misma cantidad de encuestas el error es mayor que con una muestra aleatoria simple. Conviene estratificar las escuelas por departamento o por gestión antes de sortear.

b) Un control de calidad revisa una de cada 50 historias clínicas archivadas, en el orden del archivo.

> RESOLUCIÓN
>
> Sistemático: es simple y deja la muestra repartida a lo largo de todo el archivo. Funciona si el orden no tiene un ciclo que coincida con el salto de 50. Si las historias estuvieran archivadas por fecha, la muestra cubriría todos los períodos, lo que es una ventaja. El arranque tiene que ser aleatorio, entre 1 y 50.

c) Una radio hace una encuesta en su página web: «¿Está de acuerdo con el aumento del boleto?». Respondieron 3.000 personas y el 85 % dijo que no.

> RESOLUCIÓN
>
> No es una muestra probabilística: responde quien quiere (autoselección) y quien escucha esa radio y entra a la página. Los que están enojados con el aumento tienen más ganas de responder. Las 3.000 respuestas no arreglan el sesgo: una muestra grande y sesgada da una estimación precisa de algo que no es lo que se quería medir. No se puede calcular un error estándar que tenga sentido.

d) Se quiere estimar el ingreso medio de los hogares de la provincia y también dar resultados para cada departamento, incluidos los más chicos.

> RESOLUCIÓN
>
> Estratificado por departamento. Con afijación proporcional, los departamentos chicos tendrían tan pocos casos que no se podría decir nada de ellos: hay que darles más muestra que la que les corresponde por su tamaño y después ponderar para el total provincial. Es lo que hacen las encuestas oficiales: estratificar y combinar con conglomerados (manzanas, viviendas) para abaratar el trabajo de campo.

e) Para conocer la opinión sobre el transporte público, un encuestador entrevista a las personas que esperan en la parada de la plaza principal un martes a la mañana.

> RESOLUCIÓN
>
> Muestreo por conveniencia, no probabilístico. Quedan afuera los que no usan el colectivo (justamente, tal vez, porque el servicio es malo), los que viajan desde otras paradas y los que viajan a otra hora. El marco implícito —quienes pasan por esa parada a esa hora— no es la población de interés. Puede servir para una prueba piloto del cuestionario, pero no para estimar.

---

estadistica.ar/docentes/trabajos-practicos/semana-8 · Las lecturas de cada tema y la hoja de fórmulas están en el sitio.