ESTADISTICA.AR

Relación entre dos variables

Estadística · Semana 4 · Unidad 2: Estadística descriptiva

PREGUNTA PARA LA CLASE

En Gran Catamarca, ¿las personas con más estudios trabajan más?

Hoy

1

Dos variables cualitativas

La tabla de contingencia

2

La paradoja de Simpson

Cuando el total dice lo contrario que las partes

3

Dos variables cuantitativas

Dispersión, covarianza y correlación

4

La recta de mínimos cuadrados

Predecir una variable con la otra

5

Correlación y causalidad

Lo que una asociación no dice

1

Dos variables cualitativas

La tabla de contingencia

La tabla de contingencia

Personas de 25 a 64 años encuestadas en Gran Catamarca, por nivel educativo y condición de ocupación.

Nivel educativo	Ocupada	No ocupada	Total
Hasta primario completo	58	33	91
Secundario, incompleto o completo	259	94	353
Superior, incompleto o completo	244	56	300
Total	561	183	744

«No ocupada» junta a las desocupadas y a las inactivas. Fuente: INDEC, Encuesta Permanente de Hogares, primer trimestre de 2026. Gran Catamarca.

Tres porcentajes distintos

Marginal

Del total, está ocupado el **75,4 %**.

561 de 744: una variable sola.

Condicional por fila

De quienes tienen estudios superiores, está ocupado el **81,3 %**.

244 de 300: dentro de un nivel.

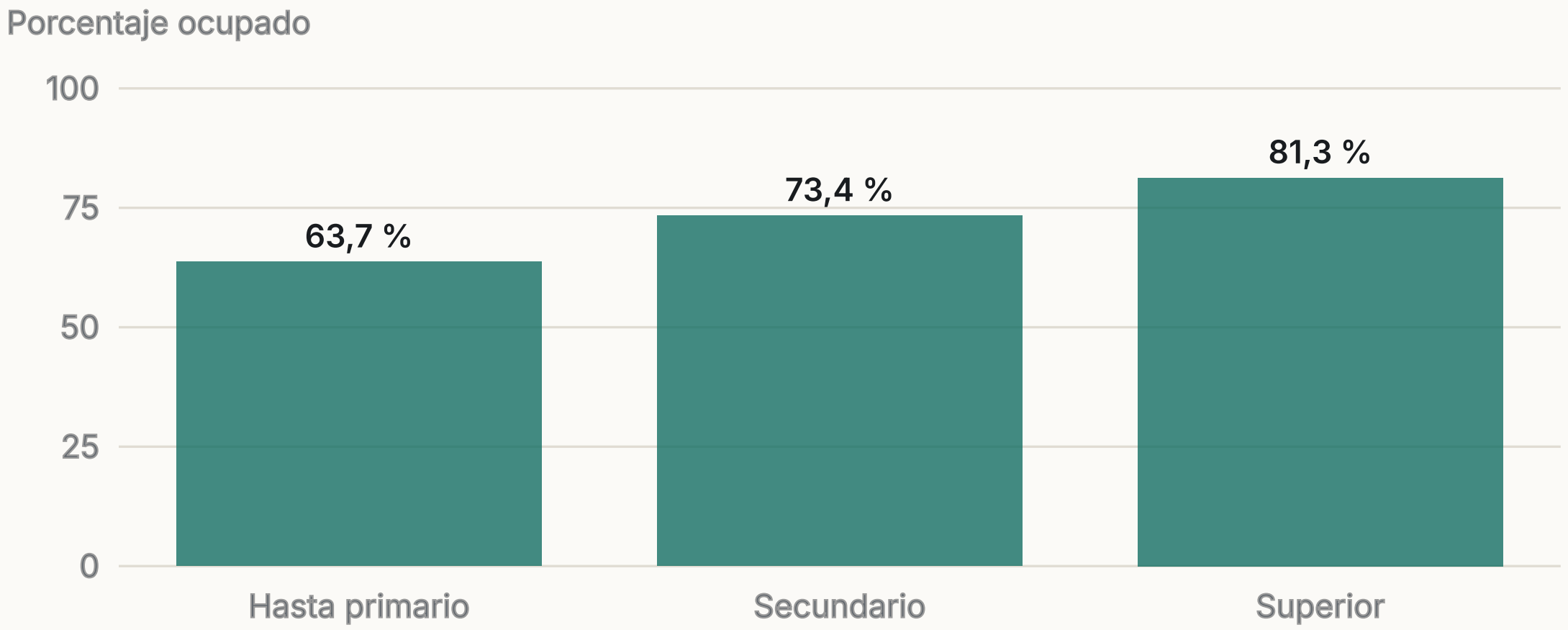
Conjunta

Tiene estudios superiores **y** está ocupado el **32,8 %** del total.

244 de 744: una celda sobre el total.

La ocupación según el nivel educativo

Porcentaje de ocupadas en cada nivel: porcentajes por fila de la tabla.



Fuente: INDEC, Encuesta Permanente de Hogares, primer trimestre de 2026. Gran Catamarca.

PREGUNTA PARA LA CLASE

¿Qué porcentaje de las personas con estudios superiores está ocupado? ¿Y qué porcentaje de las ocupadas tiene estudios superiores?

En qué sentido calcular los porcentajes

Los porcentajes se calculan **dentro de cada categoría de la variable que explica**, y se comparan entre esas categorías.

Acá

El nivel educativo explica la ocupación: porcentajes por fila, uno por nivel.

Si se lo da vuelta

Por columna se describe a los ocupados: qué nivel tienen. Es otra pregunta.

2

La paradoja de Simpson

Cuando el total dice lo contrario que las partes

Dos tratamientos para los cálculos renales

Pacientes con éxito del tratamiento, según el tamaño del cálculo.

	Cirugía abierta	Nefrolitotomía percutánea
Cálculos chicos	81 de 87 (93 %)	234 de 270 (87 %)
Cálculos grandes	192 de 263 (73 %)	55 de 80 (69 %)
Todos	273 de 350 (78 %)	289 de 350 (83 %)

Fuente: Charig y otros (1986), British Medical Journal 292: 879-882.

¿Cómo puede ser?

75 %

de los pacientes de cirugía abierta tenían cálculos grandes, los difíciles

23 %

de los de la percutánea

El total mezcla grupos que no son comparables: la cirugía abierta recibió los casos más difíciles.

Lo que enseña la paradoja

Una tercera variable

Puede cambiar, y hasta dar vuelta, la relación entre otras dos.

Comparar lo comparable

Dentro de cada grupo: cálculos chicos con chicos, grandes con grandes.

Pensar antes de sumar

Qué variables podrían estar detrás, antes de juntar las tablas.

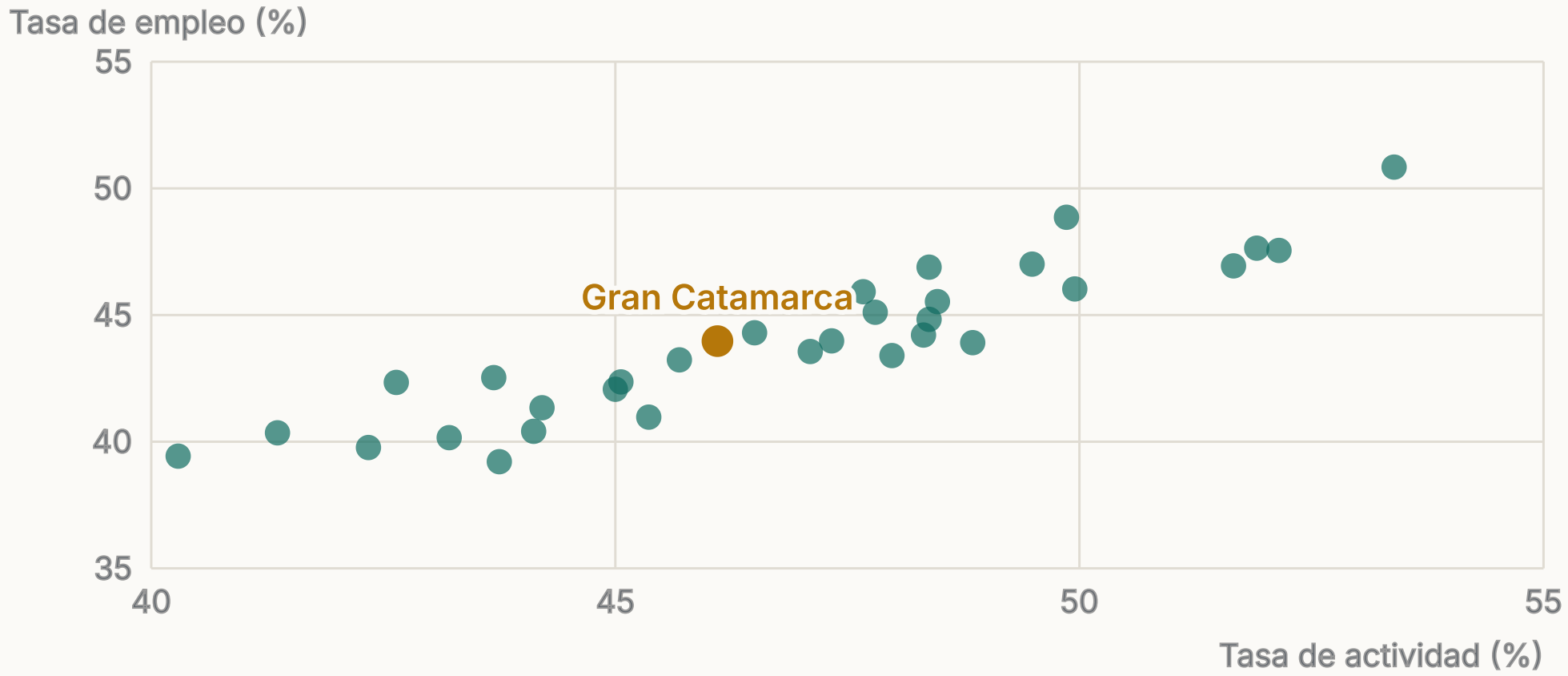
3

Dos variables cuantitativas

Dispersión, covarianza y correlación

El diagrama de dispersión

Cada punto es un aglomerado urbano: su tasa de actividad y su tasa de empleo.



Fuente: INDEC, Encuesta Permanente de Hogares, primer trimestre de 2026. Los 32 aglomerados; base en estadistica.ar/datos/indicadores-aglomerados.

La covarianza a mano: los seis del NOA

Aglomerado	Actividad, x	Empleo, y	x − x̄	y − ȳ	Producto
Gran Catamarca	46,10	43,97	−0,58	−0,68	0,40
Gran Tucumán	47,33	43,98	0,65	−0,67	−0,44
Salta	45,69	43,23	−0,99	−1,42	1,41
Jujuy	49,86	48,86	3,18	4,21	13,38
Santiago del Estero	42,64	42,34	−4,04	−2,31	9,34
La Rioja	48,47	45,53	1,79	0,88	1,57
Media y suma	46,68	44,65			25,66

El producto es positivo cuando los dos desvíos tienen el mismo signo: los dos arriba de su media, o los dos abajo.

Covarianza y correlación

Covarianza

$s_{xy} = \Sigma (x_i - \bar{x})(y_i - \bar{y}) / (n - 1) = 25{,}66 / 5 = 5{,}13$

Su signo da la dirección; su tamaño depende de las unidades.

Correlación de Pearson

$r = s_{xy} / (s_x \cdot s_y) = 5{,}13 / (2{,}51 \cdot 2{,}31) = 0{,}89$

Sin unidad, siempre entre −1 y 1.

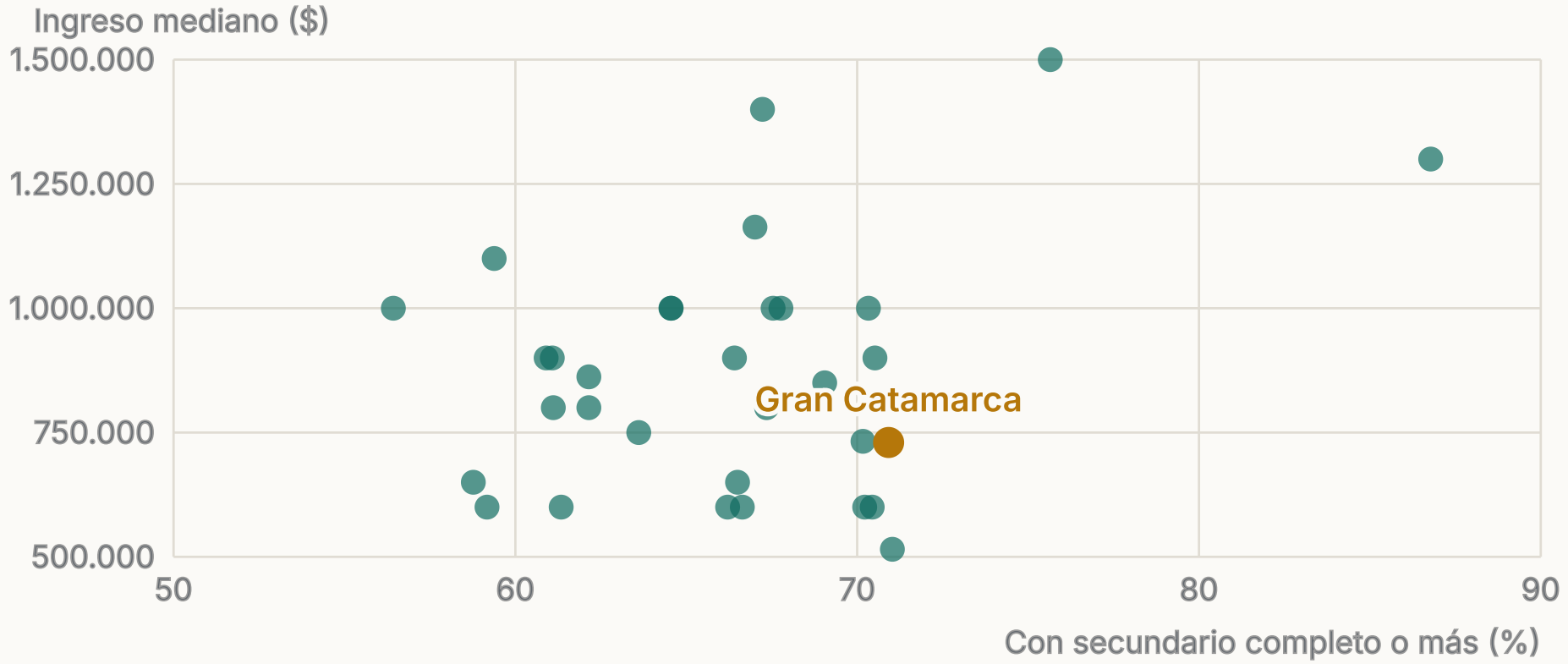
Con los 32

$r = 0{,}92$

Una relación lineal directa y fuerte.

Una correlación débil: r = 0,30

Cada punto es un aglomerado: el porcentaje de personas con al menos secundario completo y el ingreso mediano.



Fuente: INDEC, Encuesta Permanente de Hogares, primer trimestre de 2026. Los 32 aglomerados; base en estadistica.ar/datos/indicadores-aglomerados.

Cómo leer r

El signo

Positivo: suben juntas. Negativo: cuando una sube, la otra baja.

El tamaño

Cerca de ±1, los puntos casi sobre una recta; cerca de 0, sin relación lineal.

Como guía: débil hasta 0,3; fuerte desde 0,7.

Solo lo lineal

Una relación curva fuerte puede dar r cerca de 0.

Siempre mirar el gráfico antes que el número.

Los atípicos

Un solo punto lejano puede subir o bajar mucho a r.

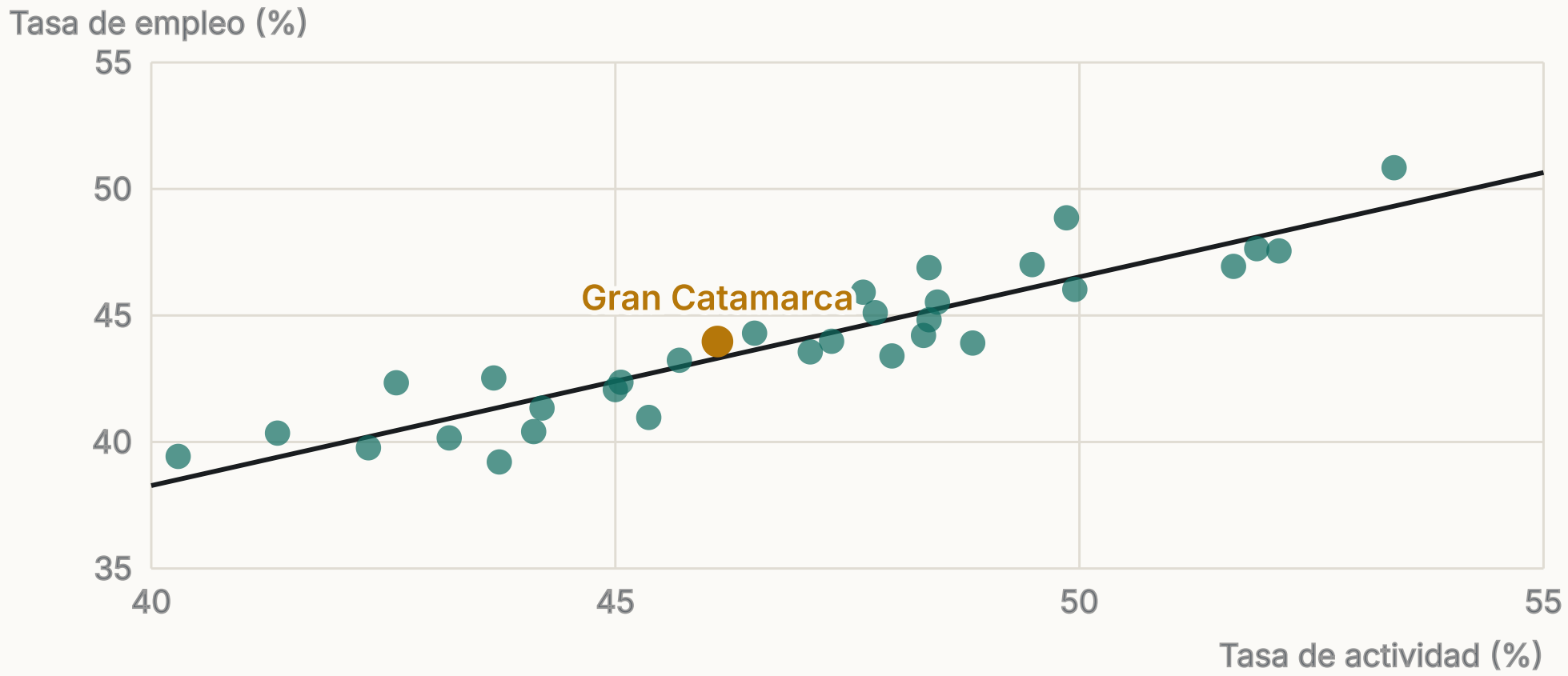
4

La recta de mínimos cuadrados

Predecir una variable con la otra

La recta de mínimos cuadrados

$\hat{y} = 5{,}28 + 0{,}82 \cdot \text{actividad}$



Fuente: INDEC, Encuesta Permanente de Hogares, primer trimestre de 2026. Los 32 aglomerados; base en estadistica.ar/datos/indicadores-aglomerados.

Qué dicen sus números

La pendiente, b

Por cada punto más de actividad, la recta predice **0,82 puntos** más de empleo.

La ordenada, a

El empleo predicho con actividad cero: 5,28.

No tiene sentido: ningún aglomerado está cerca de 0.

Gran Catamarca

Predicho: 43,31 %.
Observado: 43,97 %.

El residuo, 0,66 puntos, es lo que la recta no explica.

El coeficiente de determinación

$R^2 = 0{,}85$

es r al cuadrado: $0{,}92^2$

85 %

de la variabilidad del empleo entre aglomerados se explica por su relación lineal con la actividad

El resto, lo que la recta no explica, está en los residuos.

Cuidados al predecir

No extrapolar

La recta vale en el rango de los datos: de 40 a 54 % de actividad.

No es reversible

La recta de y en x no sirve para predecir x a partir de y.

Mirar los residuos

Si forman una curva o un embudo, la recta no es el modelo adecuado.

5

Correlación y causalidad

Lo que una asociación no dice

Correlación no implica causalidad

Una tercera variable

Los helados y los ahogados suben juntos: los dos dependen del calor.

La causa al revés

Más bomberos, más daños: los incendios grandes atraen más bomberos.

La selección

Quienes eligen un tratamiento pueden ser distintos de quienes no.

El azar

Con muchas variables, alguna correlación aparece por casualidad.

¿Cuándo se puede hablar de causa?

Estudio observacional

- Se mira lo que pasa, sin intervenir.
- Los grupos pueden diferir en muchas cosas.
- Muestra **asociación**; la causa, solo con mucho cuidado.

Experimento aleatorizado

- Se **sortea** quién recibe cada tratamiento.
- El sorteo reparte parejas las otras variables, conocidas o no.
- Una diferencia en el resultado se puede atribuir al tratamiento.

Qué usar según las dos variables

Variables	Tabla o gráfico	Medida
Dos cualitativas	Tabla de contingencia; barras por grupo	Porcentajes condicionales
Dos cuantitativas	Diagrama de dispersión	Correlación de Pearson; recta y R^2
Una cualitativa y una cuantitativa	Cajas lado a lado	Medias o medianas por grupo

Para saber si una relación vista en la muestra existe en la población hace falta inferencia: la prueba chi cuadrado y la de la correlación vienen en la unidad 5.

Para llevarse

- En una tabla de contingencia, los porcentajes se calculan dentro de cada categoría de la variable que explica.
- Una tercera variable puede dar vuelta una relación: la paradoja de Simpson.
- La correlación mide la fuerza de una relación lineal, de −1 a 1. Antes del número, el gráfico.
- La recta de mínimos cuadrados predice y a partir de x; R^2 dice qué parte de la variabilidad explica.
- Correlación no es causalidad: para eso hace falta un experimento, o mucho cuidado.

Guía de trabajo práctico 4
estadistica.ar/docentes/trabajos-practicos/semana-4

Calculadora de correlación y regresión
estadistica.ar/herramientas/calc-correlacion-regresion

Simulador de correlación
estadistica.ar/herramientas/sim-correlacion

Paradoja de Simpson
estadistica.ar/conceptos/paradoja-de-simpson