ESTADISTICA.AR

Relación entre dos variables

Estadística · Semana 4 · Unidad 2: Estadística descriptiva

1 / 29

NOTAS · 1

Pensada para 90 a 120 minutos. Hasta ahora se describió una variable por vez; desde hoy, dos a la vez. Cambia la pregunta: ya no es cómo es una variable, sino si se mueve con otra.

PREGUNTA PARA LA CLASE

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

2 / 29

NOTAS · 2

Que arriesguen y, sobre todo, que digan cómo lo averiguarían: qué contarían, qué compararían. La respuesta está en la tabla de contingencia de la parte 1.



NOTAS · 3

Qué herramienta se usa depende, como siempre, del tipo de las dos variables.



SEMANA 4 · DOS VARIABLES CUALITATIVAS

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.

5 / 29

NOTAS · 5

Cada persona cae en una sola celda: la tabla cruza las categorías de las dos variables. Los totales de los márgenes son las frecuencias marginales: la distribución de cada variable por separado.

Se toma de 25 a 64 años para que no pesen los que todavía estudian ni los jubilados.

SEMANA 4 · DOS VARIABLES CUALITATIVAS

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.

6 / 29

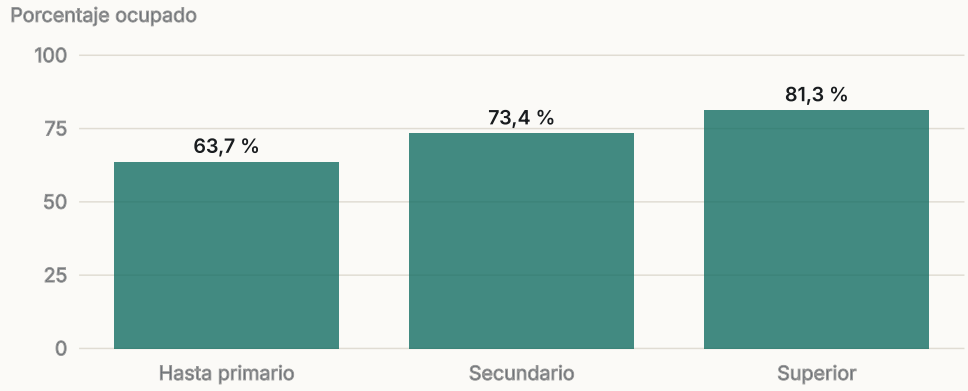
NOTAS · 6

Los tres usan la misma celda, pero dividen por cosas distintas. Antes de leer un porcentaje de una tabla de doble entrada, hay que preguntar: ¿de qué total?

SEMANA 4 · DOS VARIABLES CUALITATIVAS

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.

7 / 29

NOTAS · 7

La respuesta a la pregunta del principio: sí. Del 63,7 % con hasta primario al 81,3 % con estudios superiores. Si las dos variables no tuvieran relación, las tres barras serían iguales al porcentaje marginal, 75,4 %.

Que haya relación no dice por qué: puede ser que estudiar abra puertas, o que quienes tienen más recursos estudien más y también consigan trabajo. Es el tema de la parte 5.

PREGUNTA PARA LA CLASE

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

8 / 29

NOTAS · 8

El primero es por fila: 244 de 300, 81,3 %. El segundo, por columna: 244 de 561, 43,5 %.

Suenan parecidas y responden preguntas distintas. El error de confundirlas es frecuente, también en los medios.

SEMANA 4 · DOS VARIABLES CUALITATIVAS

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.

9 / 29

NOTAS · 9

Una regla práctica: la variable explicativa va en las filas y los porcentajes suman 100 en cada fila. Si va en las columnas, por columna.

2

La paradoja de Simpson

Cuando el total dice lo contrario que las partes

10 / 29

SEMANA 4 · LA PARADOJA DE SIMPSON

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.

11 / 29

NOTAS · 11

Pedir que miren solo la última fila: ¿cuál tratamiento es mejor? La percutánea. Ahora las dos primeras: la cirugía abierta gana con los cálculos chicos y también con los grandes.

No es un error de cuentas: los números son de un estudio real, y el ejemplo es clásico.

SEMANA 4 · LA PARADOJA DE SIMPSON

¿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.

12 / 29

NOTAS · 12

Con cálculos grandes el éxito es menor con cualquier tratamiento. Como la cirugía abierta tuvo muchos más, su total baja, aunque dentro de cada tamaño le vaya mejor.

El tamaño del cálculo es una tercera variable que se relaciona con las otras dos: decide qué tratamiento se elige y también el resultado. Se la llama variable de confusión.

SEMANA 4 · LA PARADOJA DE SIMPSON

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.

13 / 29

NOTAS · 13

El práctico tiene el mismo fenómeno con dos hospitales y pacientes leves y graves.

3

Dos variables cuantitativas

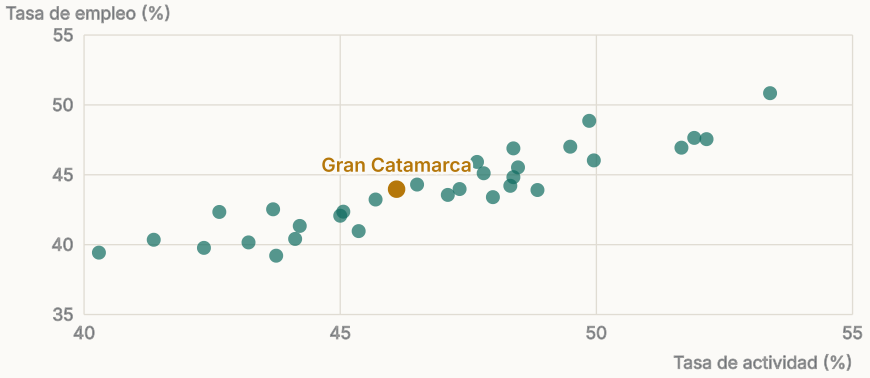
Dispersión, covarianza y correlación

14 / 29

SEMANA 4 · DOS VARIABLES CUANTITATIVAS

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.

15 / 29

NOTAS · 15

La tasa de actividad es el porcentaje de la población que trabaja o busca trabajo; la de empleo, el que trabaja. Los puntos suben juntos: donde más gente participa del mercado de trabajo, más gente trabaja.

Para describirlo se mira la dirección (sube o baja), la forma (recta o curva) y la fuerza (qué tan juntos están los puntos alrededor de una línea).

SEMANA 4 · DOS VARIABLES CUANTITATIVAS

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.

16 / 29

NOTAS · 16

Jujuy está arriba en las dos y suma mucho; Santiago del Estero, abajo en las dos, y también suma. Si la mayoría de los productos son positivos, la relación es directa.

SEMANA 4 · DOS VARIABLES CUANTITATIVAS

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.

17 / 29

NOTAS · 17

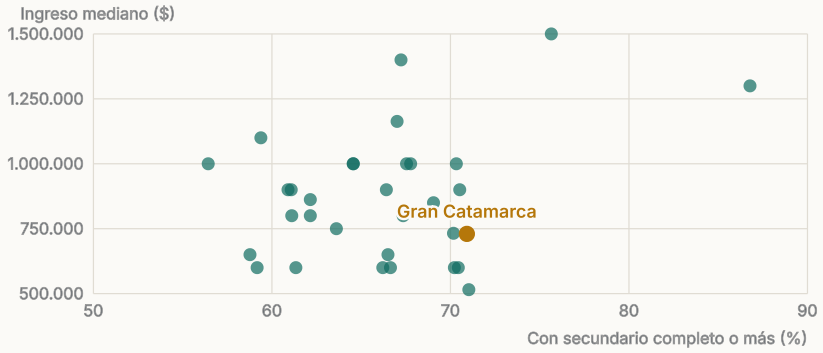
La covarianza en «puntos porcentuales al cuadrado» no se interpreta. Dividir por los dos desvíos la convierte en un número sin unidad: es la correlación.

Con seis puntos la correlación cambia mucho si se agrega o se saca uno: con los 32 es más estable.

SEMANA 4 · DOS VARIABLES CUANTITATIVAS

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.

18 / 29

NOTAS · 18

Hay algo de tendencia hacia arriba, pero los puntos están muy dispersos: conocer la educación de un aglomerado ayuda poco a adivinar su ingreso mediano.

Ojo: es una correlación entre aglomerados, no entre personas. Que los aglomerados más educados no ganen mucho más no dice nada sobre si a una persona le conviene estudiar. Es la falacia ecológica.

SEMANA 4 · DOS VARIABLES CUANTITATIVAS

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.

19 / 29

NOTAS · 19

El simulador de correlación de estadistica.ar permite ver nubes con distintos r y probar el efecto de un atípico: sirve para practicar a adivinar r a ojo, que es el primer ejercicio del práctico.

4

La recta de mínimos cuadrados

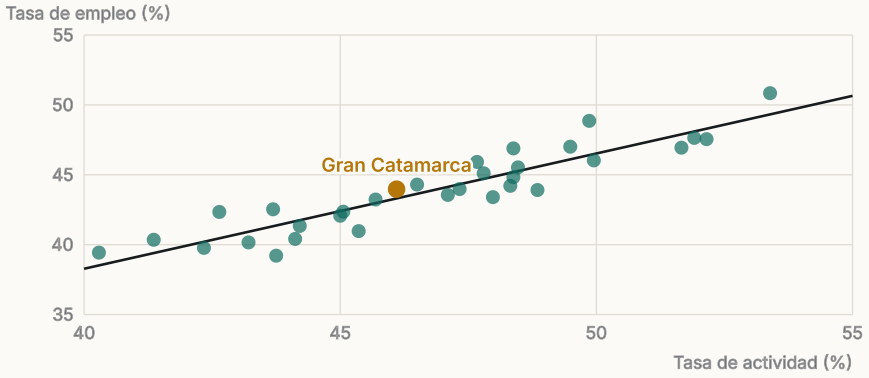
Predecir una variable con la otra

20 / 29

SEMANA 4 · LA RECTA DE MÍNIMOS CUADRADOS

La recta de mínimos cuadrados

$\hat{y}$ = 5,28 + 0,82 · actividad



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

21 / 29

NOTAS · 21

De todas las rectas posibles, es la que hace más chica la suma de los residuos al cuadrado: las distancias verticales de cada punto a la recta. Por eso se llama de mínimos cuadrados.

Las fórmulas: $b = s_{xy} / s_x^2$ y $a = \bar{y} - b \cdot \bar{x}$. La recta siempre pasa por el punto de las medias.

SEMANA 4 · LA RECTA DE MÍNIMOS CUADRADOS

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.

22 / 29

NOTAS · 22

Catamarca está por encima de la recta: tiene más empleo del que corresponde a su nivel de actividad, porque su desocupación es más baja que el promedio.

SEMANA 4 · LA RECTA DE MÍNIMOS CUADRADOS

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.

23 / 29

NOTAS · 23

Con la correlación débil del ingreso y la educación, $R^2 = 0{,}09$: la recta explicaría solo el 9 %.

«Explica» es un término estadístico: dice cuánto se reduce el error al predecir con la recta. No quiere decir que una variable cause la otra.

SEMANA 4 · LA RECTA DE MÍNIMOS CUADRADOS

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.

24 / 29

NOTAS · 24

La regresión se retoma en la unidad 5, con inferencia: si la pendiente es distinta de cero en la población, y con qué margen de error.

5

Correlación y causalidad

Lo que una asociación no dice

25 / 29

SEMANA 4 · CORRELACIÓN Y CAUSALIDAD

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.

26 / 29

NOTAS · 26

El práctico tiene un ejercicio para discutir con casos así. La paradoja de Simpson es un caso de tercera variable.

SEMANA 4 · CORRELACIÓN Y CAUSALIDAD

¿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.

27 / 29

NOTAS · 27

El estudio de los cálculos renales fue observacional: los médicos eligieron el tratamiento según el caso, y eso produjo la paradoja. En un ensayo con sorteo no habría pasado.

La EPH también es observacional: la relación entre educación y empleo no prueba que estudiar cause empleo, aunque haya buenas razones para pensarlo.

SEMANA 4 · CORRELACIÓN Y CAUSALIDAD

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.

28 / 29

NOTAS · 28

La tercera fila es lo que se hizo la semana pasada con el ingreso por categoría ocupacional: también es una relación entre dos variables.

SEMANA 4 · RELACIÓN ENTRE DOS VARIABLES

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

29 / 29

NOTAS · 29

Para el práctico: calculadora, y la base de los 32 aglomerados para el ejercicio con software.