

Clase 2: Repaso de estadística bivariada

Análisis Avanzado de Datos

Gabriel Sotomayor



Recordatorio de la clase anterior

¿Por qué usamos modelos estadísticos en Ciencias Sociales?

- **Capturar y reducir la complejidad:** Los modelos permiten vincular datos con teorías, ayudando a interpretar la realidad social.
- **Formalizar y probar teorías:** Dando precisión y permitiendo identificar relaciones causales y predecir fenómenos.
- **Énfasis en la explicación sociológica:** Las técnicas son una herramienta para la investigación social. No basta con explicar la varianza de una variable dependiente, sino la capacidad e explicar las relaciones teóricamente.



Evaluaciones

Tarea 1: 2 de septiembre

- Gestión de datos
- Estadística bivariada

Prueba 1: 9 de Septiembre

- Uso de modelos en ciencias sociales
- Estadística bivariada
- Regresión lineal simple



Objetivo de la sesión

Revisar el estudio de relaciones entre variables con estadística bivariada.



Relaciones entre variables

En sociología frecuentemente queremos contestar preguntas acerca de la relación entre variables, tales como la relación entre la escolaridad de los padres y la de los hijos, el ingreso y la probabilidad de participar en una protesta o entre el sexo y las horas dedicadas al trabajo doméstico.

Para esto necesitamos tener mediciones de ambas variables en la misma unidad (personas, comunas, hogares, etc) para poder observar su variación conjunta.



Ejemplo: Brecha Salarial de género

A lo largo de la clase trabajaremos con el ejemplo de la brecha salarial de género de cada comuna, la cual se calcula con la siguiente fórmula:

$$\frac{\text{Salario Promedio Hombres} - \text{Salario Promedio Mujeres}}{\text{Salario Promedio Hombres}}$$

Por ejemplo si en una comuna el salario promedio de los hombres es de \$400.000 y el de las mujeres es 300.000

$$\text{Brecha Salarial de Género} = \frac{100.000}{400.000} \times 100 = 0.25 \times 100$$



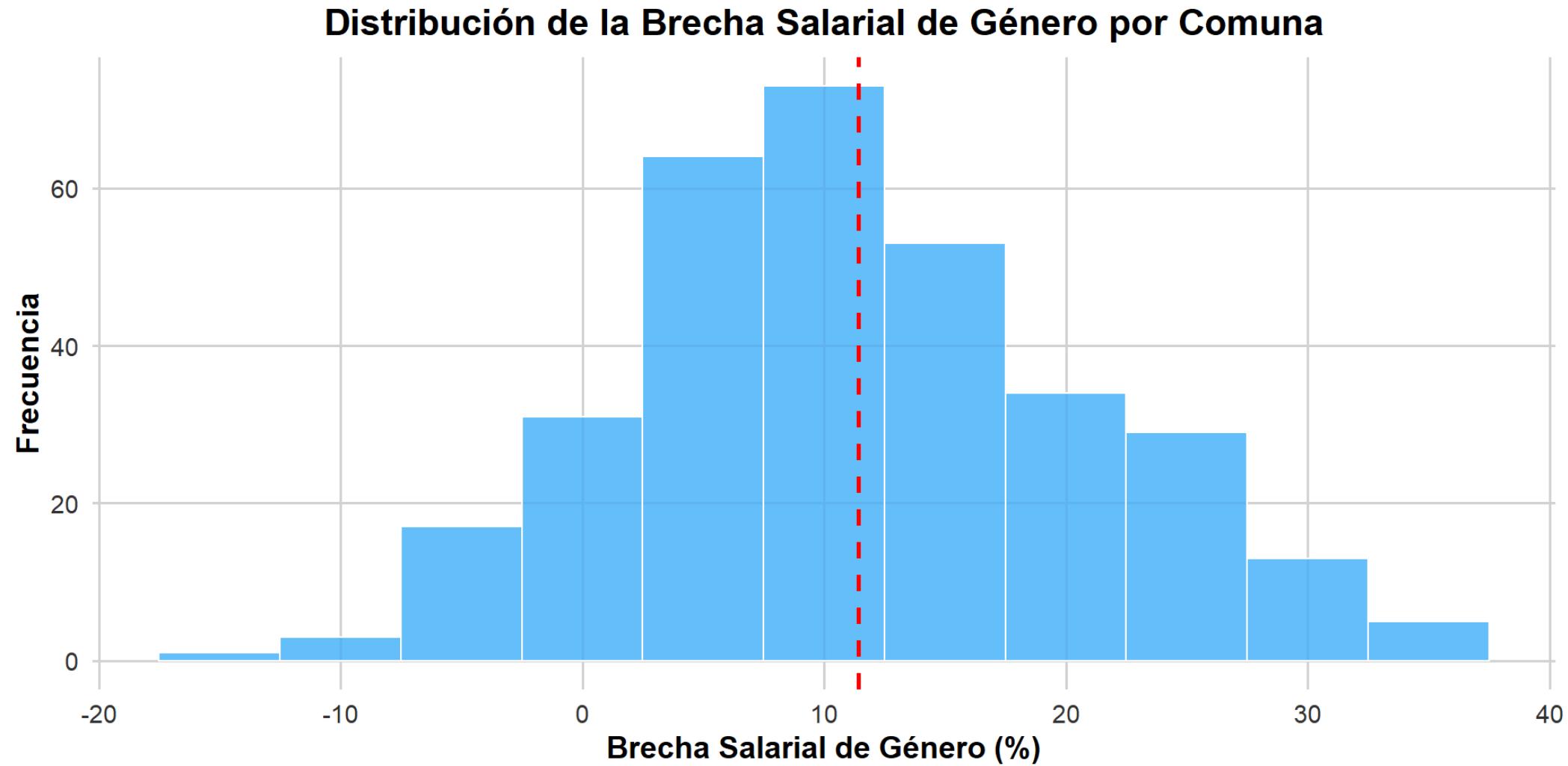
Estadísticos descriptivos

Medidas de tendencia central: Valores situados al centro de las distribuciones que representan espacios donde los datos tienden a agruparse (Media, Mediana, Moda).

Medidas de Dispersión: Describen la variabilidad de los datos de una distribución (Rango, Varianza, Desviación estándar).



Brecha salarial de género comunal



Media y Varianza

Media: Suma de las puntuaciones dividida por el número de observaciones. Sensible a los casos extremos.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Varianza: Es una medida de dispersión que representa el promedio de las distancias al cuadrado de cada dato respecto al promedio. Al elevar al cuadrado las distancias, se evita que los signos negativos y positivos se cancelen entre sí.

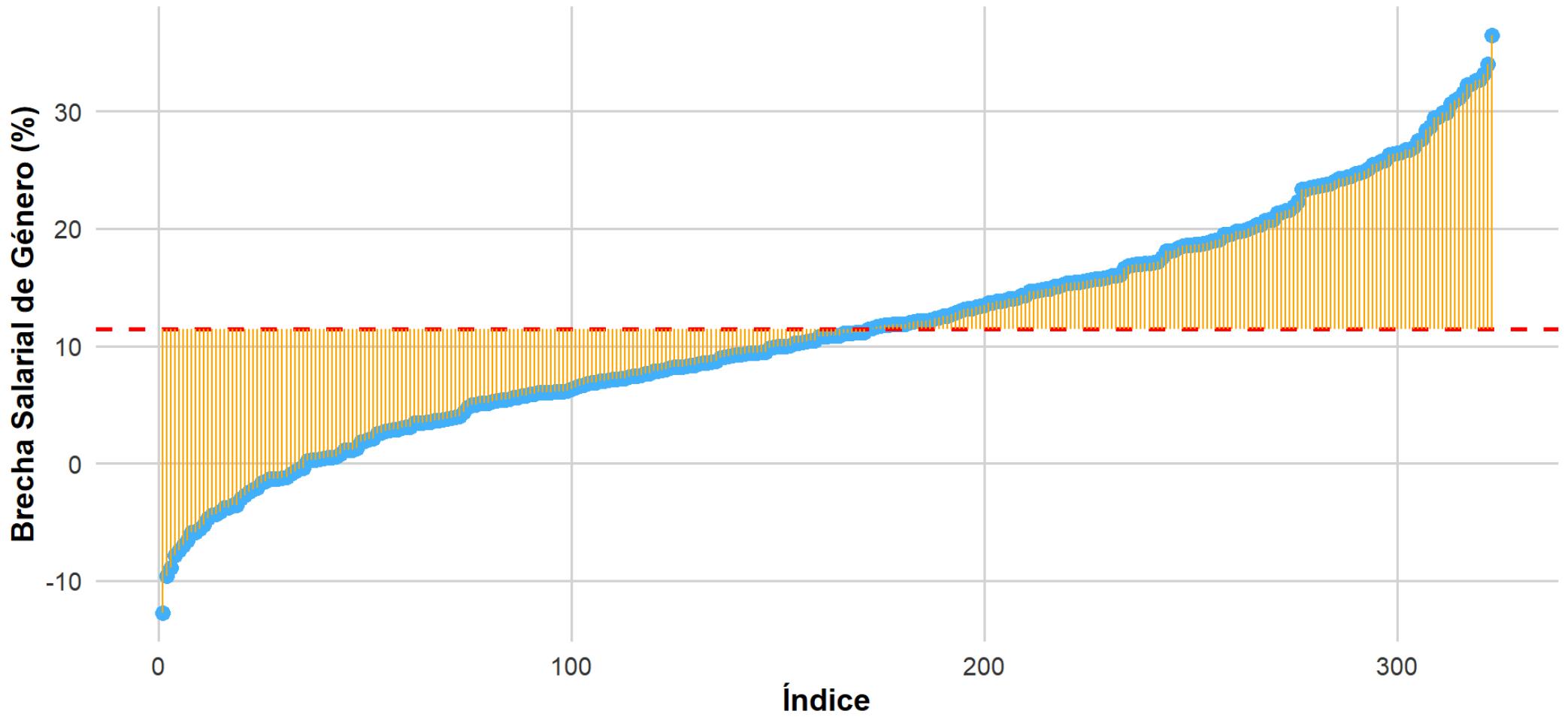


$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$



Represeñación de la varianza

Distancia de cada punto a la Media de la Brecha Salarial



Cálculo de la varianza

| | comuna | Observaciones | Desviaciones | Desviaciones.al.cuadrado |
|----|---------------|---------------|-----------------------|--------------------------|
| 1 | iquique | 20.36 | 20.36 - 11.42 = 8.94 | (8.94)^2 = 79.88 |
| 2 | alto hospicio | 26.27 | 26.27 - 11.42 = 14.84 | (14.84)^2 = 220.27 |
| 3 | pozo almonte | 30.94 | 30.94 - 11.42 = 19.51 | (19.51)^2 = 380.66 |
| 4 | huara | 21.9 | 21.9 - 11.42 = 10.48 | (10.48)^2 = 109.79 |
| 5 | pica | 20.72 | 20.72 - 11.42 = 9.29 | (9.29)^2 = 86.38 |
| 6 | antofagasta | 26.37 | 26.37 - 11.42 = 14.95 | (14.95)^2 = 223.39 |
| 7 | mejillones | 32.57 | 32.57 - 11.42 = 21.15 | (21.15)^2 = 447.16 |
| 8 | sierra gorda | 31.13 | 31.13 - 11.42 = 19.71 | (19.71)^2 = 388.32 |
| 9 | taltal | 24.29 | 24.29 - 11.42 = 12.86 | (12.86)^2 = 165.5 |
| 10 | calama | 32.29 | 32.29 - 11.42 = 20.86 | (20.86)^2 = 435.33 |

$$\text{Varianza} = \frac{28961.14}{(323 - 1)} = 89.94$$



Desviación estándar

Desviación estándar: Es la raíz cuadrada de la varianza. Es comúnmente utilizada en otros cálculos y es más fácil de interpretar ya que esta aproximadamente en la unidad de medida original. Es la que mejor da cuenta de la dispersión (es decir de las distancias de los casos al promedio).

$$\text{Desviación estándar} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$



¿Que variables influyen en la amplitud de la brecha salarial de género comunal?

Pudimos observar que la brecha salarial de género tiene bastante variación entre distintas comunas del país.

Ahora vamos a revisar la relación entre la magnitud de la brecha y el nivel de ruralidad y de educación.

¿Cómo creen que se relacionan estas variables con la magnitud de la brecha salarial de género? ¿Porque?



Rol de las variables

Variable dependiente o respuesta: Es la variable de interés en el estudio, aquella cuya variación se desea comprender. Comúnmente, se representa en el eje de las ordenadas (eje Y).

Variable independiente o explicativa: Es la variable que influye o explica los cambios en la variable respuesta. Habitualmente, se representa en el eje de las abscisas (eje X).

Ojo: si bien lo términos pueden sugerir causalidad esto no necesariamente es así (en la mayoría de los casos no lo es).



Gráfico de dispersión

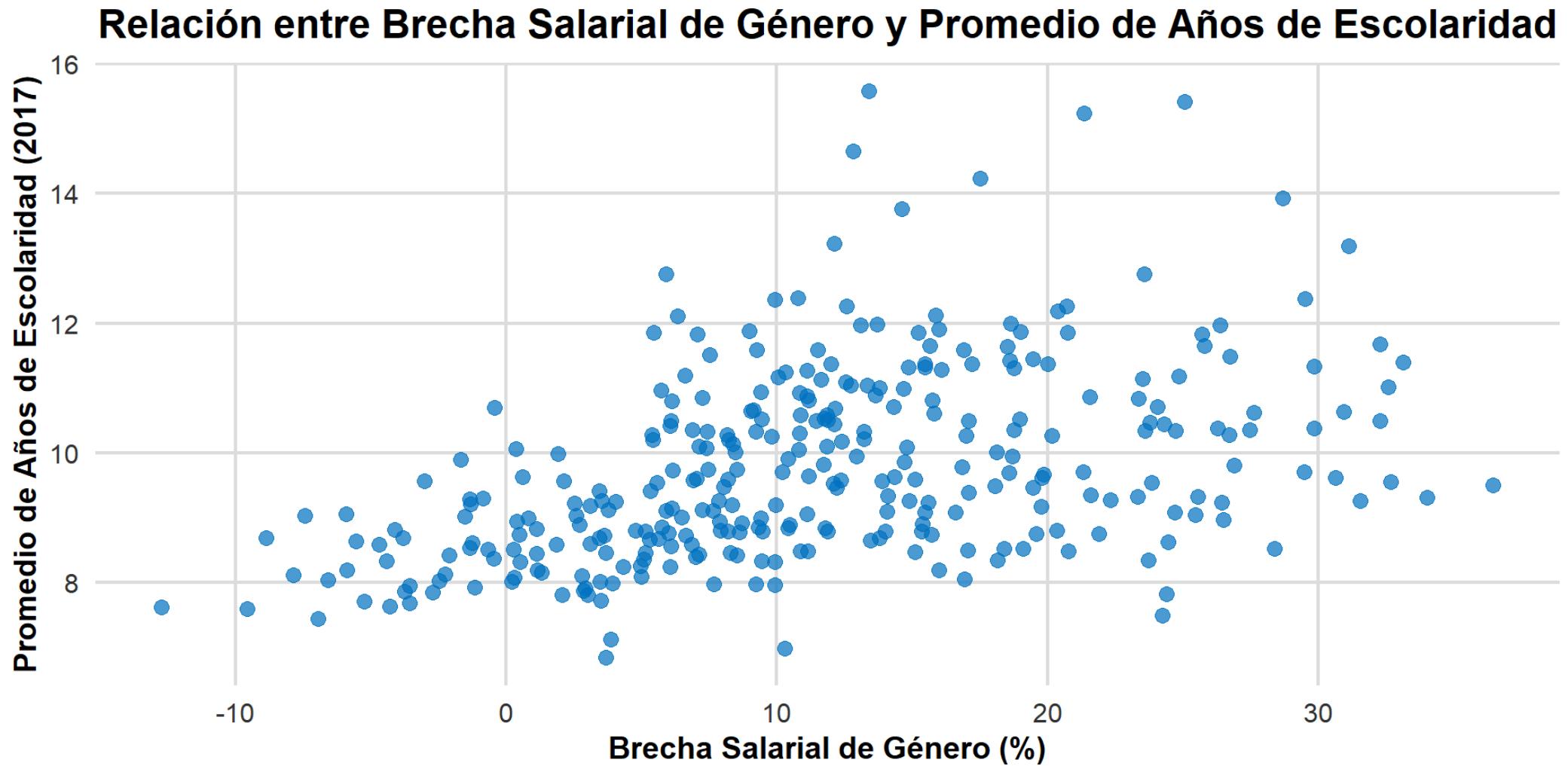


Gráfico de dispersión

Un **gráfico de dispersión** muestra la relación entre dos variables cuantitativas medidas en los mismos individuos. Los valores de una variable aparecen en el eje de las abscisas y los de la otra en el eje de las ordenadas. Cada individuo aparece como un punto del diagrama. Su posición depende de los valores que toman las dos variables en cada individuo.

En cualquier gráfico de datos, identifica el aspecto general y las desviaciones del mismo. Puedes describir el aspecto general de un diagrama de dispersión mediante la **forma**, la **dirección** y la **fuerza** de la relación.

Un tipo importante de desviación son las observaciones atípicas, valores individuales que quedan fuera del aspecto general de la relación.



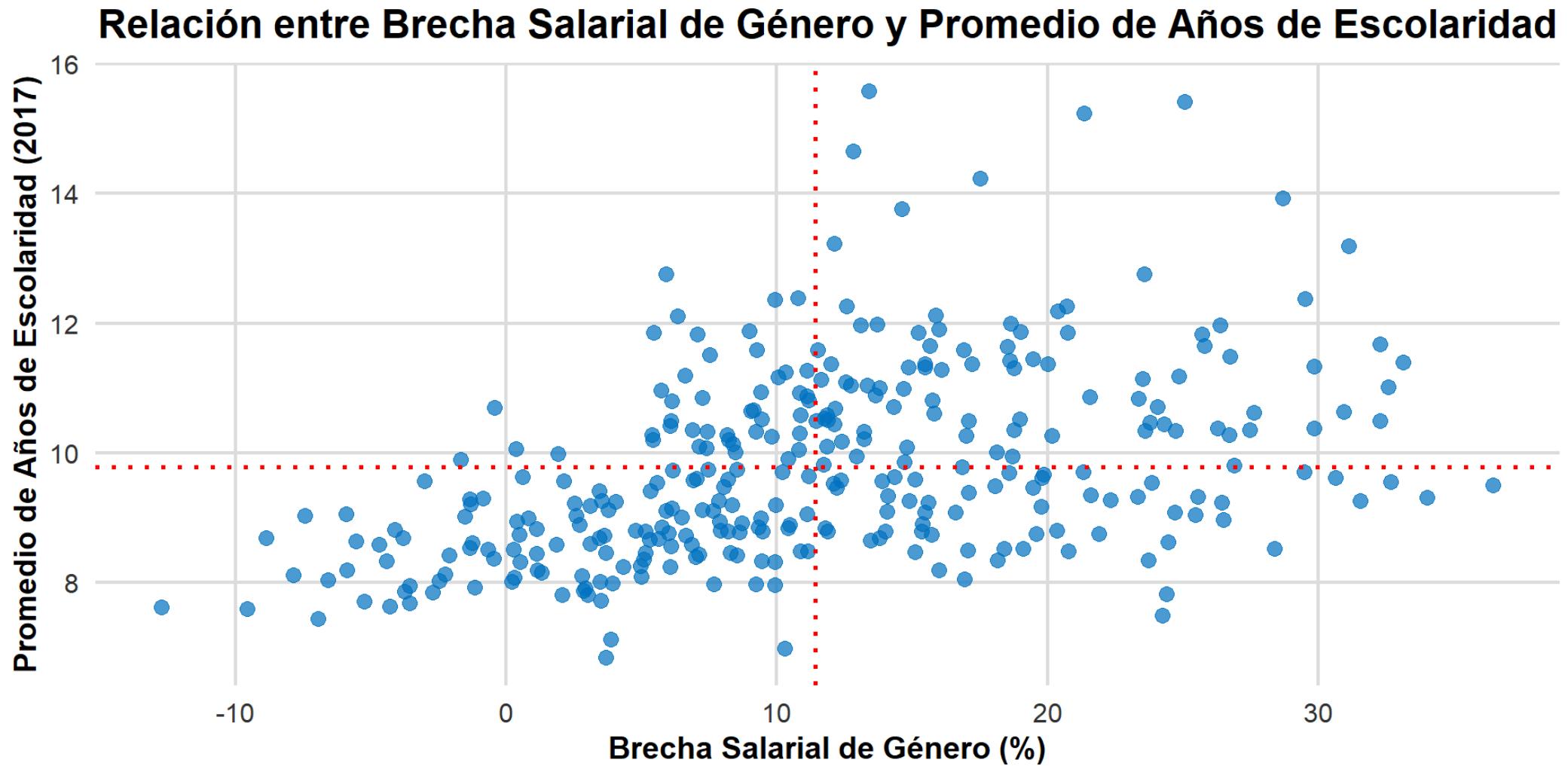
Asociación positiva y negativa

Dos variables están asociadas **positivamente** cuando valores superiores a la media de una de ellas tienden a ir acompañados de valores también situados por encima de la media de la otra variable, y cuando valores inferiores a la media también tienden a ocurrir conjuntamente.

Dos variables están asociadas **negativamente** cuando valores superiores a la media de una de ellas tienden a ir acompañados de valores inferiores a la media de la otra variable, y viceversa.



Gráfico de dispersión



Correlación

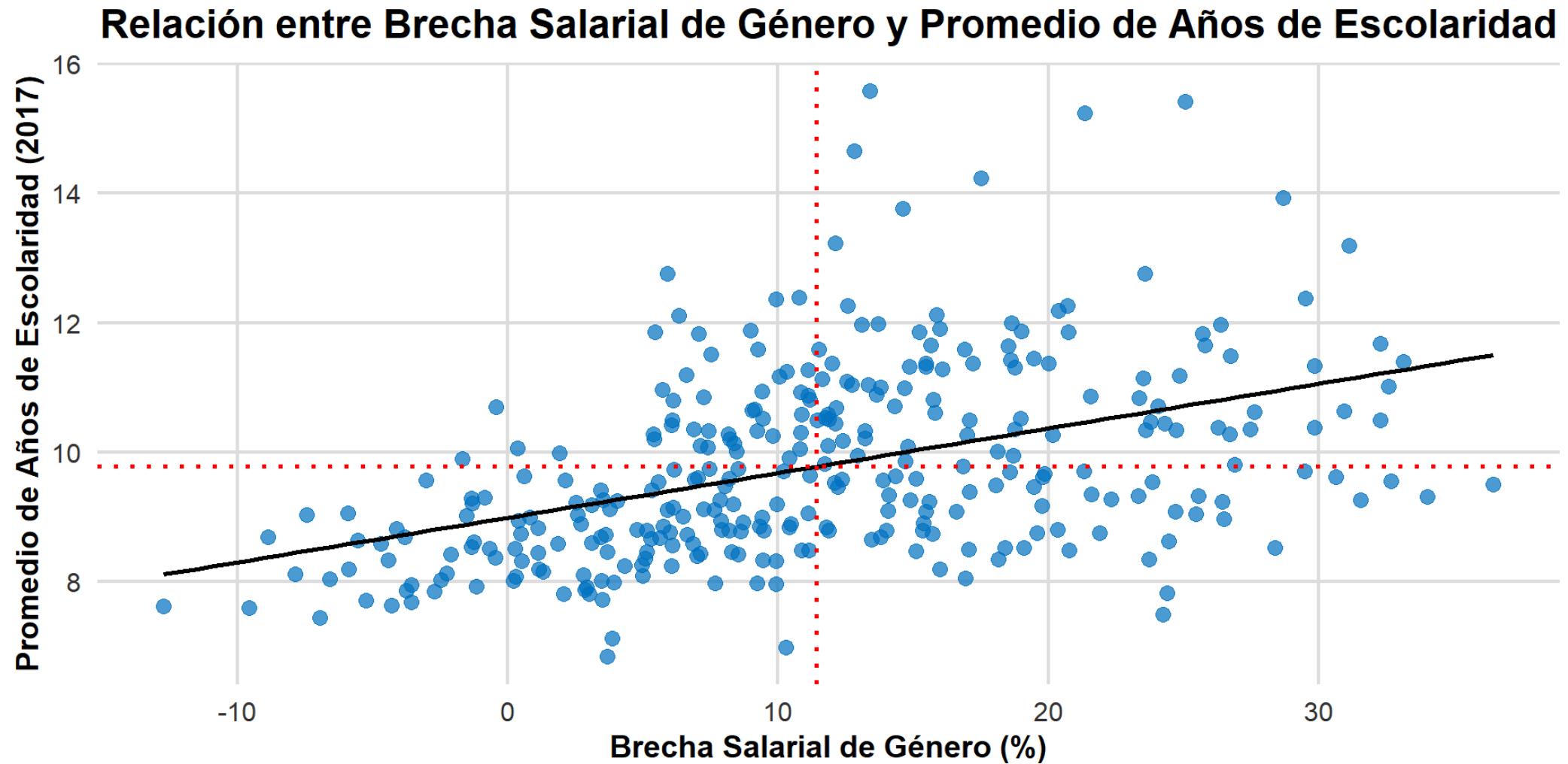
La correlación mide la fuerza y la dirección de la relación lineal entre dos variables cuantitativas. La correlación se simboliza con la letra r.

Si tenemos datos de dos variables x e y para n individuos. Los valores para el primer individuo son x_1 e y_1 , para el segundo son x_2 e y_2 , etc. Las medias y las desviaciones típicas de las dos variables son \bar{x} y s_x para los valores de x, e \bar{y} y s_y para los valores de y. La correlación r entre x e y es:

$$r = \frac{1}{n - 1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$



Gráfico de correlación



Ejemplo de cálculo

| | brecha | esc | | brecha_est | | esc_esta |
|---|--------------------------|-------|---------------------------------|--------------------------------|--|----------|
| 1 | 20.36 | 12.19 | $(20.36 - 11.42) / 9.48 = 0.94$ | $(12.19 - 9.77) / 1.48 = 1.64$ | | |
| 2 | 26.27 | 10.38 | $(26.27 - 11.42) / 9.48 = 1.56$ | $(10.38 - 9.77) / 1.48 = 0.41$ | | |
| 3 | 30.94 | 10.63 | $(30.94 - 11.42) / 9.48 = 2.06$ | $(10.63 - 9.77) / 1.48 = 0.58$ | | |
| 4 | 21.90 | 8.75 | $(21.9 - 11.42) / 9.48 = 1.1$ | $(8.75 - 9.77) / 1.48 = -0.7$ | | |
| 5 | 20.72 | 12.26 | $(20.72 - 11.42) / 9.48 = 0.98$ | $(12.26 - 9.77) / 1.48 = 1.68$ | | |
| 6 | 26.37 | 11.97 | $(26.37 - 11.42) / 9.48 = 1.58$ | $(11.97 - 9.77) / 1.48 = 1.49$ | | |
| | | | Prod_est | | | |
| 1 | $(0.94) * (1.64) = 1.54$ | | | | | |
| 2 | $(1.56) * (0.41) = 0.64$ | | | | | |
| 3 | $(2.06) * (0.58) = 1.2$ | | | | | |
| 4 | $(1.1) * (-0.7) = -0.77$ | | | | | |
| 5 | $(0.98) * (1.68) = 1.65$ | | | | | |
| 6 | $(1.58) * (1.49) = 2.35$ | | | | | |

$$\text{Correlación} = \frac{142.84}{(323 - 1)} = 0.44$$



Características de la correlación (I)

Simetría en las Variables: La correlación no distingue entre variables explicativas y respuesta; es indiferente cuál se llame x o y.

Requisito Cuantitativo: Las dos variables deben ser cuantitativas para que los cálculos de la correlación tengan sentido. No se puede calcular la correlación entre una variable cuantitativa y una categórica.

Independencia de Unidades: Como la correlación utiliza valores estandarizados, no cambia si se modifican las unidades de medida de las variables. La correlación es un valor sin unidades.

Significado del Signo:

- Correlación positiva: Indica una asociación positiva entre las variables.
- Correlación negativa: Indica una asociación negativa.



Características de la correlación (II)

Rango de la Correlación: La correlación siempre toma valores entre -1 y 1 .

- Cercanía a 0 : Indica una relación lineal débil.
- Cercanía a ± 1 : Indica una relación lineal fuerte. Un valor de ± 1 indica una relación lineal perfecta.

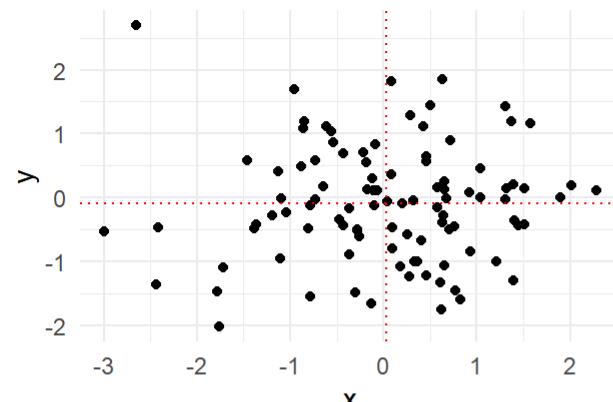
Limitación a Relaciones Lineales: La correlación sólo mide la fuerza de relaciones lineales, no describe adecuadamente las relaciones curvilíneas, aunque estas sean fuertes.

Sensibilidad a Observaciones Atípicas: La correlación puede verse fuertemente afectada por valores atípicos, lo que puede distorsionar la percepción de la relación entre las variables. Es importante utilizar la correlación con precaución cuando se detectan atípicos.

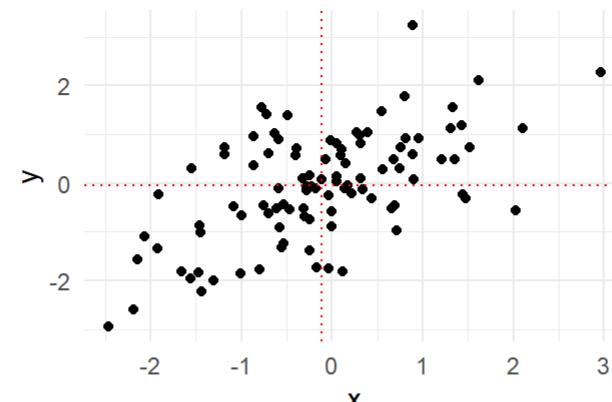


Graficos de dispersión y correlación

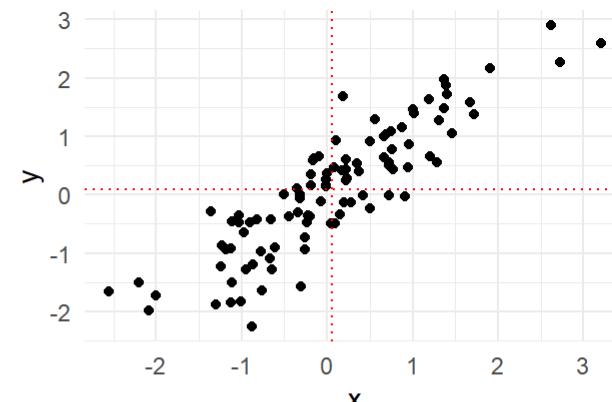
Correlación $r = 0$



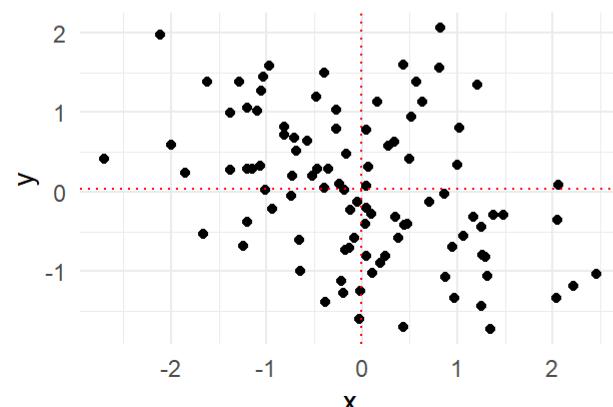
Correlación $r = 0.5$



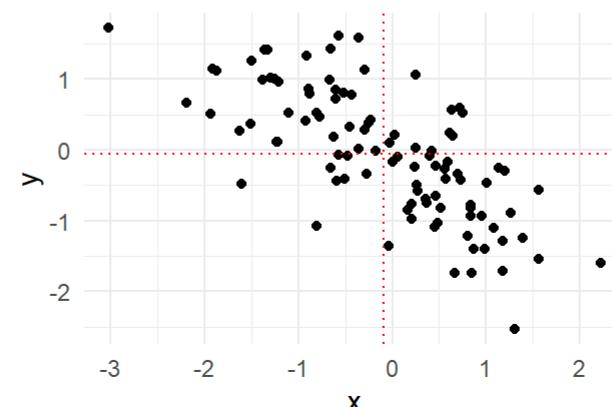
Correlación $r = 0.9$



Correlación $r = -0.3$



Correlación $r = -0.7$



Correlación $r = -0.99$

