BIOESTADÍSTICA BÁSICA

Pruebas Paramétricas:

DISTRIBUCIÓN T/ F/ CORRELACIÓN PEARSON

CLASE 4

LIBRO DE EPIDEMIOLOGÍA Y ESTADÍSTICA (ANTONIO VILLA ET AL)

CAPITULO 16 - 17 -18

PRUEBAS PARAMÉTRICAS

- Las pruebas paramétricas son una herramienta estadística que se utiliza para el análisis de los factores de la población.
- Su cálculo implica una estimación de los parámetros de la población con base en muestras estadísticas.
- Determinar las hipótesis planteadas.

COMPARACIÓN DE MEDIAS

USE LA DISTRIBUCIÓN Z

Si la desviación estándar es conocida o la muestra es >30

$$\overline{X} \pm z \frac{\sigma}{\sqrt{n}}$$

USE LA DISTRIBUCIÓN t

Si la desviación estándar no es conocida o la muestra es <30

$$\overline{X} \pm t \frac{s}{\sqrt{n}}$$

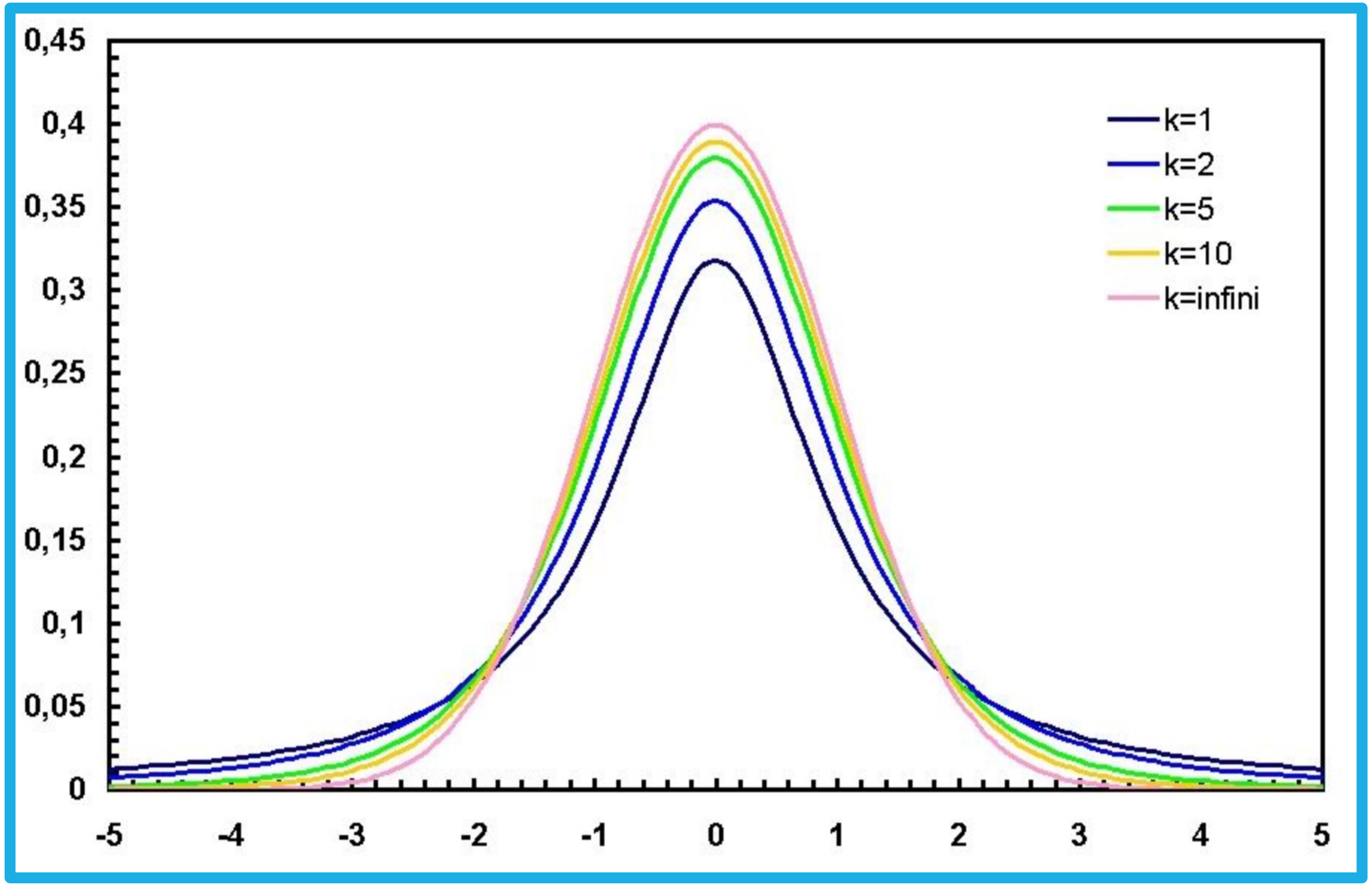
PERO ANTES, SE DEBE CONOCER A LA T DE STUDENT

DISTRIBUCIÓN T

Cuando se requieren comparaciones de dos promedios se utliza la distribución t.

A diferencia de la normal (Z), existen múltiples distribuciones t para cada tamaño.

Aquí aparece el concepto GRADOS DE LIBERTAD: El número de formas independientes por las que un sistema dinámico puede moverse.



La distribución T difiere de la de Z en que la varianza de t depende del tamaño de la muestra y siempre es mayor a uno.

Grados de libertad: n-1

CONSTRASTES DE HIPÓTESIS

UNA MUESTRA

DOS MUESTRAS

UNA MUESTRA: Compara la media con un valor establecido, no con otra muestra. A

One Sample Mean

Test Statistic when σ known

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

$\bar{x}$: sample mean
μ : population mean
σ : population standard deviation
n : sample size

Test Statistic when σ unknown

$$t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)} \quad , \quad df = n - 1$$

$\bar{x}$: sample mean
μ : population mean
s : sample standard deviation
n : sample size

HIPÓTESIS

$H_o: \mu = \mu_o$

$H_1: \mu < \mu_o$ Unilateral izquiera (inferior)

$H_1: \mu > \mu_o$ Unilateral derecha (superior)

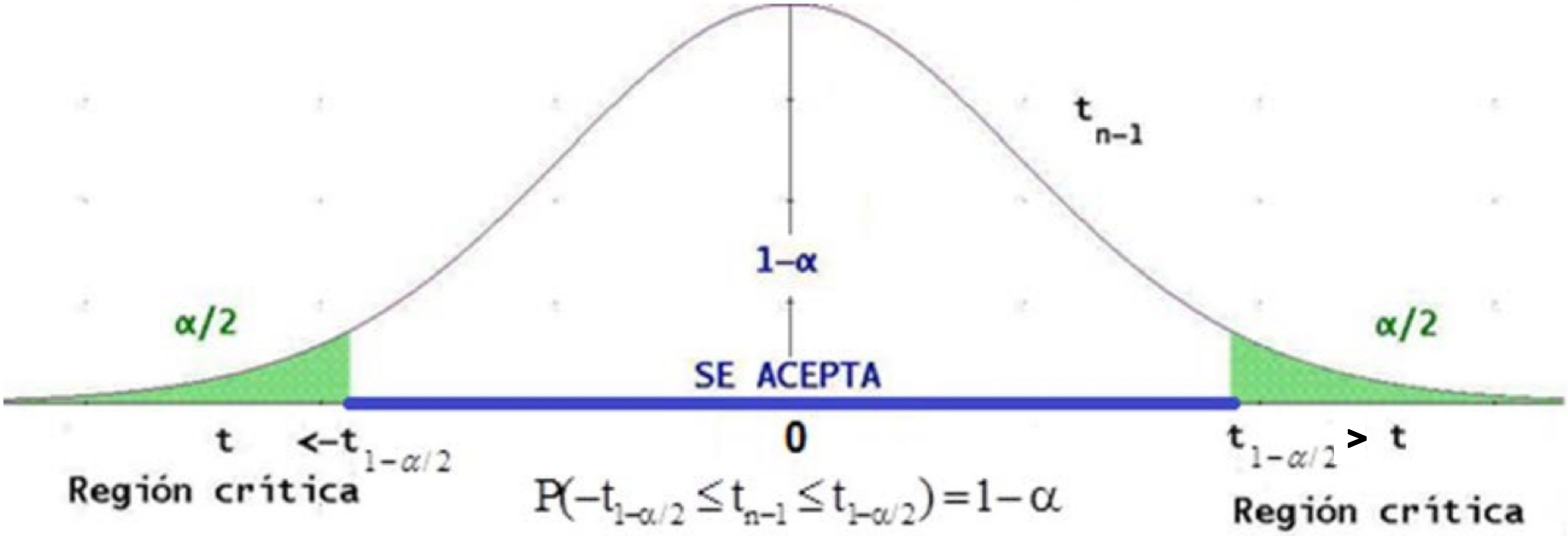
$H_1: \mu \neq \mu_o$ Bilateral

REGLAS DE DECISIÓN

Bilateral	Unilateral Derecha	Unilateral Izquierda
$To \geq Tc$ $-To \leq -Tc$	$To > Tc$	$-To < -Tc$

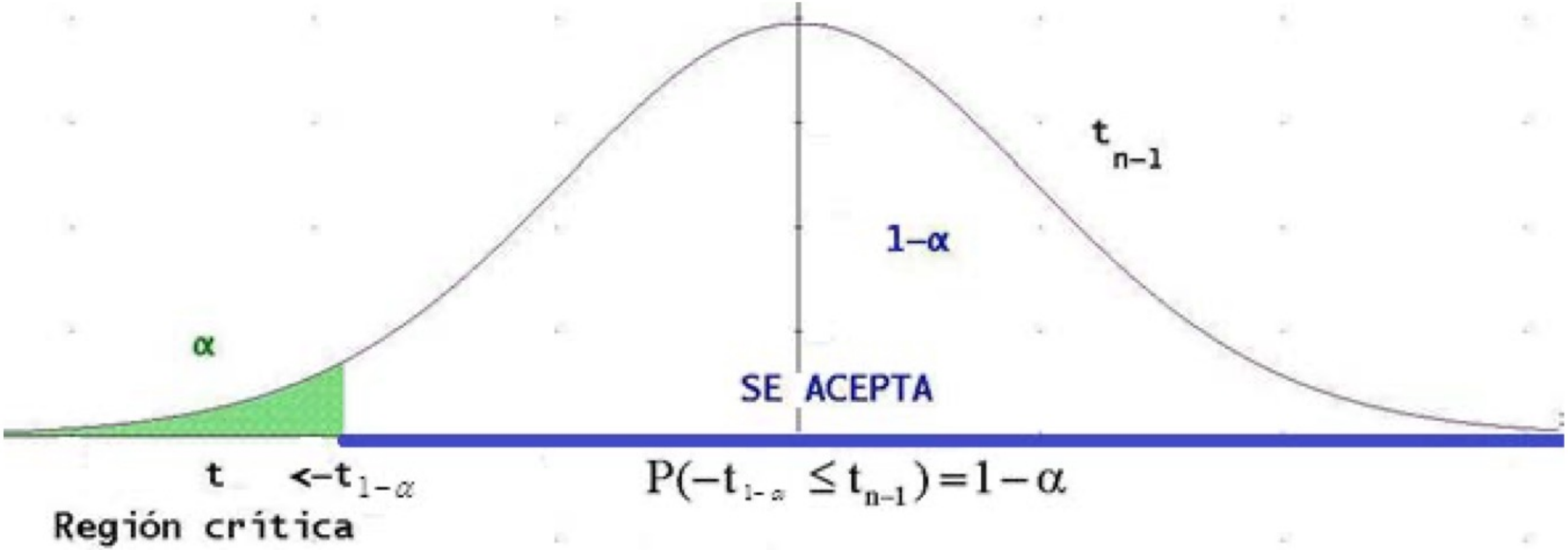
Contraste bilateral (2 colas)

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$



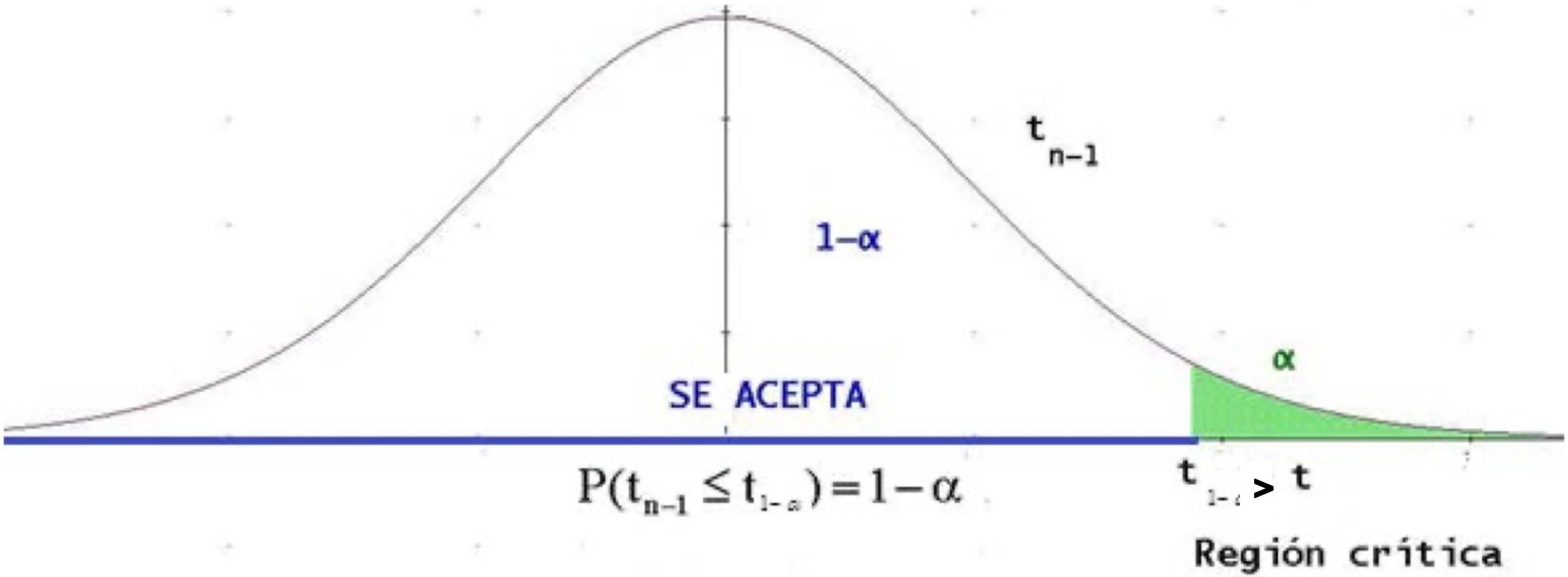
Contraste unilateral (cola de la izquierda)

$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0$$



Contraste unilateral (cola de la derecha)

$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$



La dirección médica de una clínica privada en Guayaquil, toma una muestra aleatoria de 16 mediciones sobre el tiempo de hospitalización, la media es de 5.4 días y una sd de 3.1 días. La dirección médica supone que el promedio de tiempo de hospitalización es mayor de 5 días. ¿Esta afirmación es correcta? Utiliza un nivel de significancia del 5%.

$$H_0: \mu \leq 5$$

$$H1: \mu > 5$$

Un farmaceútico en Cuenca, toma una muestra aleatoria de 12 sobres de un medicamento. El peso promedio de cada sobre es de 15.97grs, y la sd es de 0,15. Una farmaceútica reconocida afirma que el peso promedio mínimo del medicamento es de 16grs por sobre.
¿Esta afirmación es correcta? Utiliza un nivel de confianza del 90%.

$$H_0: \mu \geq 16$$

$$H1: \mu < 16$$

PRUEBA DE HIPOTESIS SOBRE LA MEDIA DE UNA DISTRIBUCIÓN NORMAL, VARIANZA DESCONOCIDA

LA EEQ publica cifras del número anual de Kilowatt-hora que gastan varios aparatos eléctrodomésticos. Se afirma que una aspiradora gasta un promedio de 46 kilowatt-hora al año. Si una **muestra aleatoria** de 12 hogares que se incluye en un estudio planeado indica que las aspiradoras gastan un promedio de 42 kilowatt-hora al año con una desviación estándar de 11.9 kilowatt-hora, ¿esto **sugiere** con un nivel de significancia de 0.05 que las aspiradoras gastan, en promedio, **menos de** 46 kilowatt-hora anualmente? Suponga que la población de kilowatt-hora es normal.

HIPÓTESIS

H0: $\mu = 46kw/h$

H1: $\mu < 46kw/h$

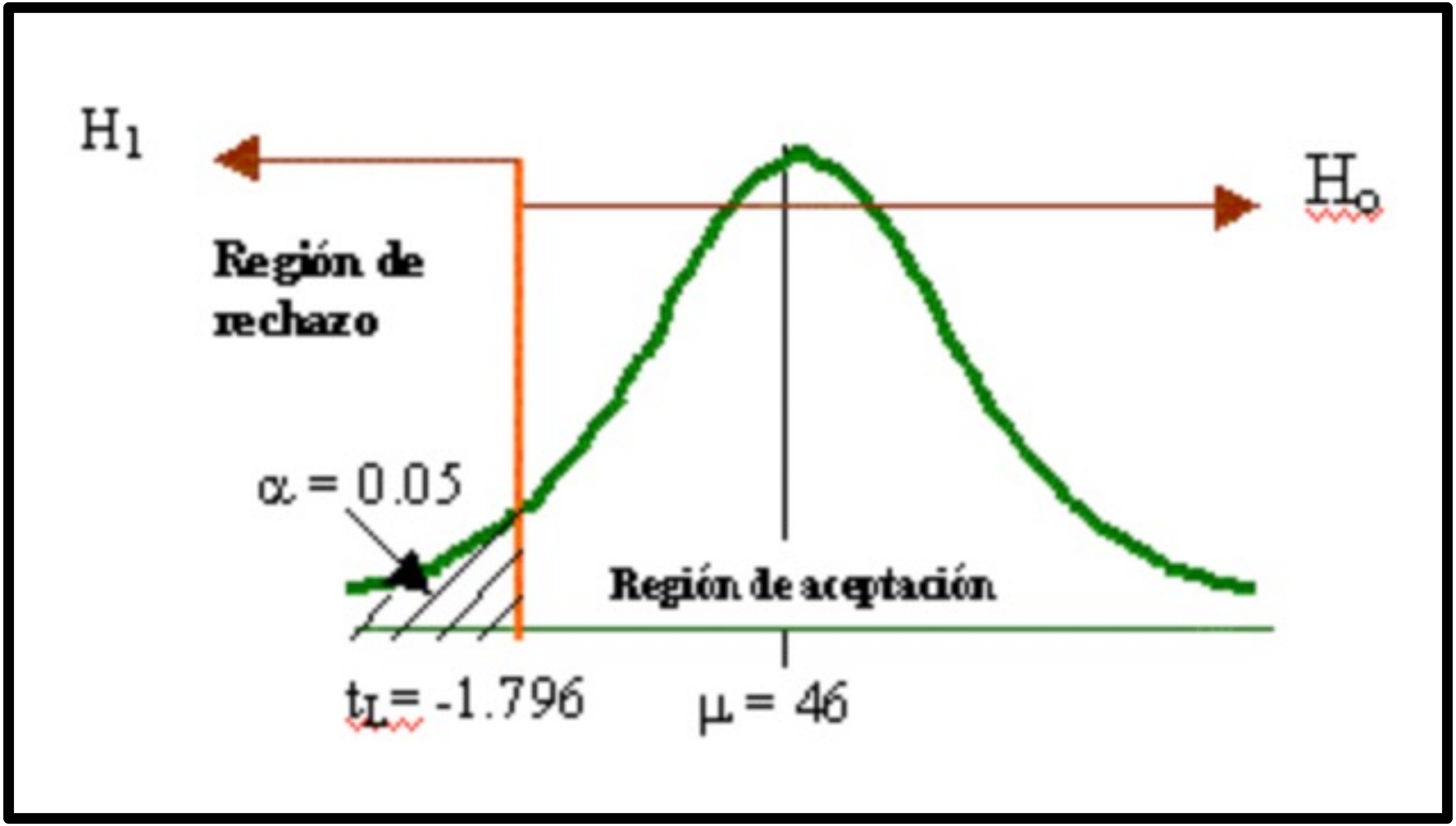
REGLAS DE DECISIÓN

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$t = \frac{42-46}{11.9 \div \sqrt{12}} = -1,16$$

T crítico

t test M5.xlsx


H_1
H_o
Región de rechazo
$\alpha = 0.05$
Región de aceptación
$t_L = -1.796$
$\mu = 46$

Como –1.16 > -1.796, por lo tanto se acepta la H_0.

Se concluye que con un nivel de significancia del 0.05, el número promedio de kilowwatt-hora que gastan al año las aspiradoras, sí es = 46.

DOS MUESTRAS: Compara la media entre dos muestras (A y B).

PAREADAS E INDEPENDIENTES

Grupos pareados/ correlacionados: dispone de dos conjuntos de datos cuantitativos continuos que pertencen a un mismo grupo de individuos.

Grupos independientes: dispone de datos cuantitativos continuos de dos diferentes grupos de individuos, no están correlacionados.

Two Sample Means

Test Statistic when σ known

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$F = \frac{s_1^2}{s_2^2} = \frac{\textit{larger variance}}{\textit{smaller variance}}$$

If $F \le t^*_{(\alpha, df)}$, then pool → **HOMOCEDASTICIDAD**

If $F > t^*_{(\alpha, df)}$, then don't pool → **HETEROCEDASTICIDAD**

Test Statistic when σ unknown With pooled variances

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\left(\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}\right)\left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)}$$

$$df = n_1 + n_2 - 2$$

Libro Epidemiología pág 257

Test Statistic when σ unknown with un-pooled variances

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{s_2^2}{n_2}\right)^2}$$

Libro Epidemiología pág 266

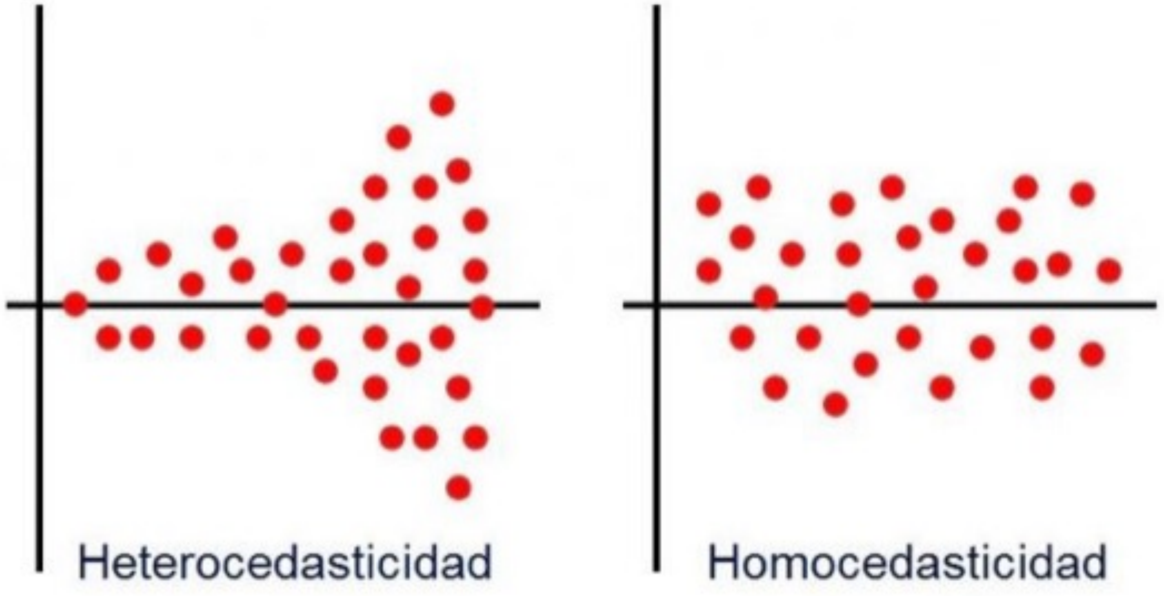
$\bar{x}$ 1 y 2 = Media muestral del grupo 1 y grupo 2

n 1 y 2 = número de observaciones del grupo 1 y grupo 2

s^2 1 y 2 = varianza del grupo 1 y grupo 2

CONSIDERAR:

HETEROCEDASTICIDAD: → Se conoce con el nombre de Prueba t heterocedasticidad (test de Welch).



ENTONCES, SI EL EJERCICIO TIENE: 2 MUESTRAS + 2 MEDIAS MUESTRALES + 2 VARIANZAS

1. DETERMINAR NORMALIDAD DEL DATO

2. OBTENER EL F TEST PARA DETERMINAR HOMOCEDASTICIDAD O HETEROCEDASTICIDAD

3. EN FUNCIÓN DEL PASO 2, ELEGIR LA FÓRMULA CORRECTA DE T.

4. EN FUNCIÓN DEL PASO 2 Y 3, ELEGIR LA FÓRMULA DE LOS GRADOS DE LIBERTAD

5. COMPARAR EL T OBS CON EL T CRITICO

EXCEL =INV.T(0,05, GL)

PASO 2

F-TEST (RAZÓN DE VARIANZAS)
F de Snedecor

El F-test, también conocido como contraste de la razón de varianzas, contrasta la hipótesis nula de que dos poblaciones normales tienen la misma varianza.

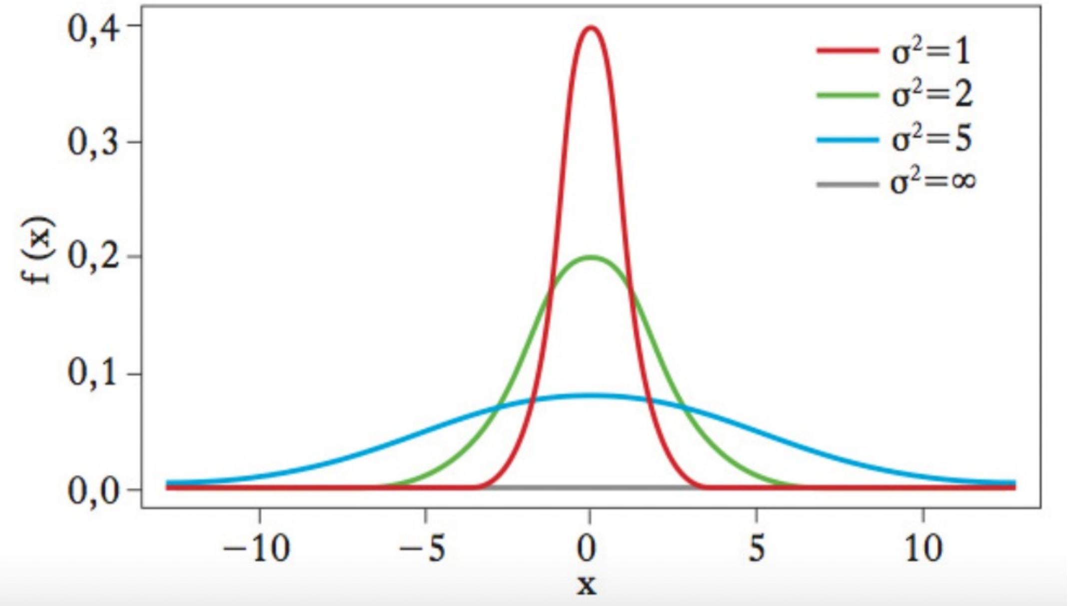
H0: Homocedasticidad

H1: Heterocedasticidad

El F-test estudia el cociente de varianzas

LA PRUEBA F QUE DIVIDE LA VARIANZA MÁS GRANDE ENTRE LA VARIANZA MAS PEQUEÑA

$$F_{obs} = Varianza\ más\ grande \div Varianza\ más\ pequeña$$

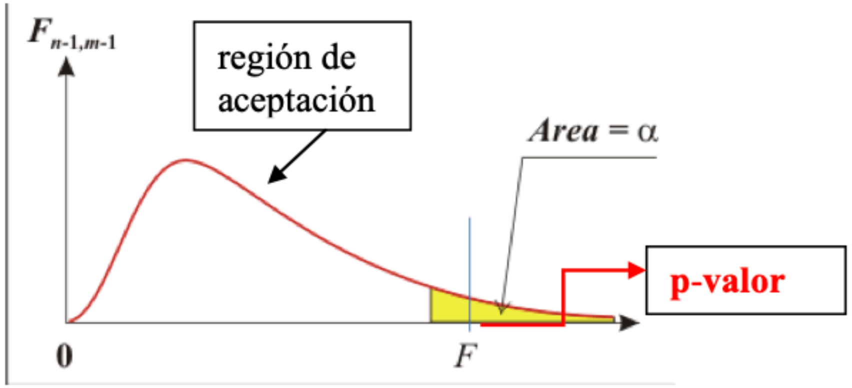


Para decidir si existe homocedasticidad, el valor de F_{obs} debe compararse con el valor $F_{crítico}$

Si $F_{obs} > F_{crítico}$

Rechazo la H0 y concluyo que hay heterocedasticidad.

Gl de F= n-1



Se desea evaluar si existe homocedasticidad entre las varianzas de dos grupos a los cuales, después, se les evaluará si existe una posible diferencia estadística entre sus promedios.

1. Primer grupo, esta formado por 16 personas, con una varianza de 7,75
2. Segundo grupo, esta formado por 11 personas, con una varianza de 2,18

$$F_{obs} = 7{,}75/2{,}18 = 3{,}56$$

Al grupo con mayor varianza n=16 le corresponden: gl = n-1   16-1 = 15   n1=15

Al grupo con menor varianza n=11 le corresponden gl = n-1   11-1 = 10   n2=10

t test.xlsx

Fobs > Fcrítico

Heterocedasticidad = Test de Welch

PASO 3

ELEGIR LA FÓRMULA CORRECTA PARA EL T OBSERVADO

$$t = \frac{\left(\bar{x}_1 - \bar{x}_2\right)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

PASO 4

ELEGIR LA FÓRMULA CORRECTA PARA EL T CRITICO

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{s_2^2}{n_2}\right)^2}$$

CALCULAR EN EXCEL =INV.T(0,05, GL)

La fórmula para calcular los grados de libertad para la prueba t de Welch se tiene en cuenta la diferencia entre las dos desviaciones estándar.

Sin embargo, si las dos muestras tienen las mismas desviaciones estándar, entonces los grados de libertad para la prueba t de Welch serán exactamente los mismos que los grados de libertad para la prueba t de Student

PASO 5

CONTRASTAR LA HIPÓTESIS

tob > t cri rechazo la H0
- tob < -t cri rechazo la H0

$$H_0 = \overline{XA} = \overline{XB}$$

Igualdad de medias

$$H_1 = \overline{XA} \neq \overline{XB}$$

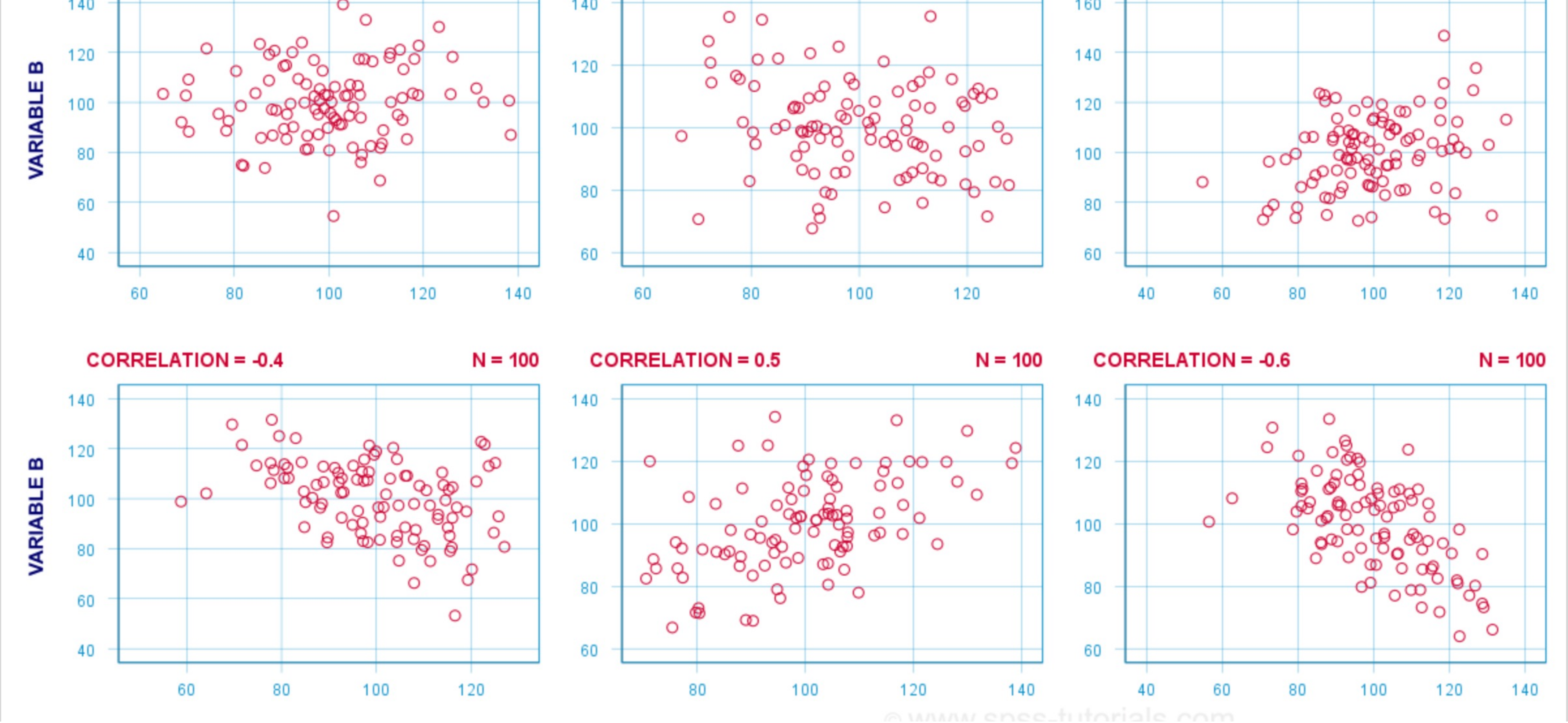
Diferencia de medias

APLICA PARA LOS DOS (HOMOCEDASTICIDAD Y HETEROCEDASTICIDAD)

ENLACE DE T TEST

https://www.cienciadedatos.net/documentos/12_t-test

EJERCICIO EN EXCEL



COEFICIENTES DE CORRELACIÓN

PEARSON

- **Relación o Asociación lineal entre dos variables cuantitativas u ordinales**.

- Los coeficientes de correlación de valores absolutos que oscilan entre -1 y 1. Hay que tomar en cuenta que tienen signos.

- Donde -1 y 1 indican correlaciones perfectas, negativas y positivas respectivamente
- 0 indica correlación nula.

OJO: CORRELACION NO ES CAUSALIDAD

TOMAR EN CUENTA

La magnitud o fuerza de la relación está especificada por el valor numérico del coeficiente, y el signo refleja la dirección de tal valor.

Dependiendo el signo (- o +), se especificará una relación perfecta negativa o positiva.

$\rightarrow$ r es un coeficiente (no tiene unidades)

$\rightarrow$ $-1 \leq r \leq 1$

Si $r \approx 1$ existe una correlación directa fuerte

Si $r \approx -1$ existe una correlación inversa fuerte

Si $r = 1$ o $r = -1$ hay una correlación funcional

Si $r \approx 0$ no existe una correlación lineal

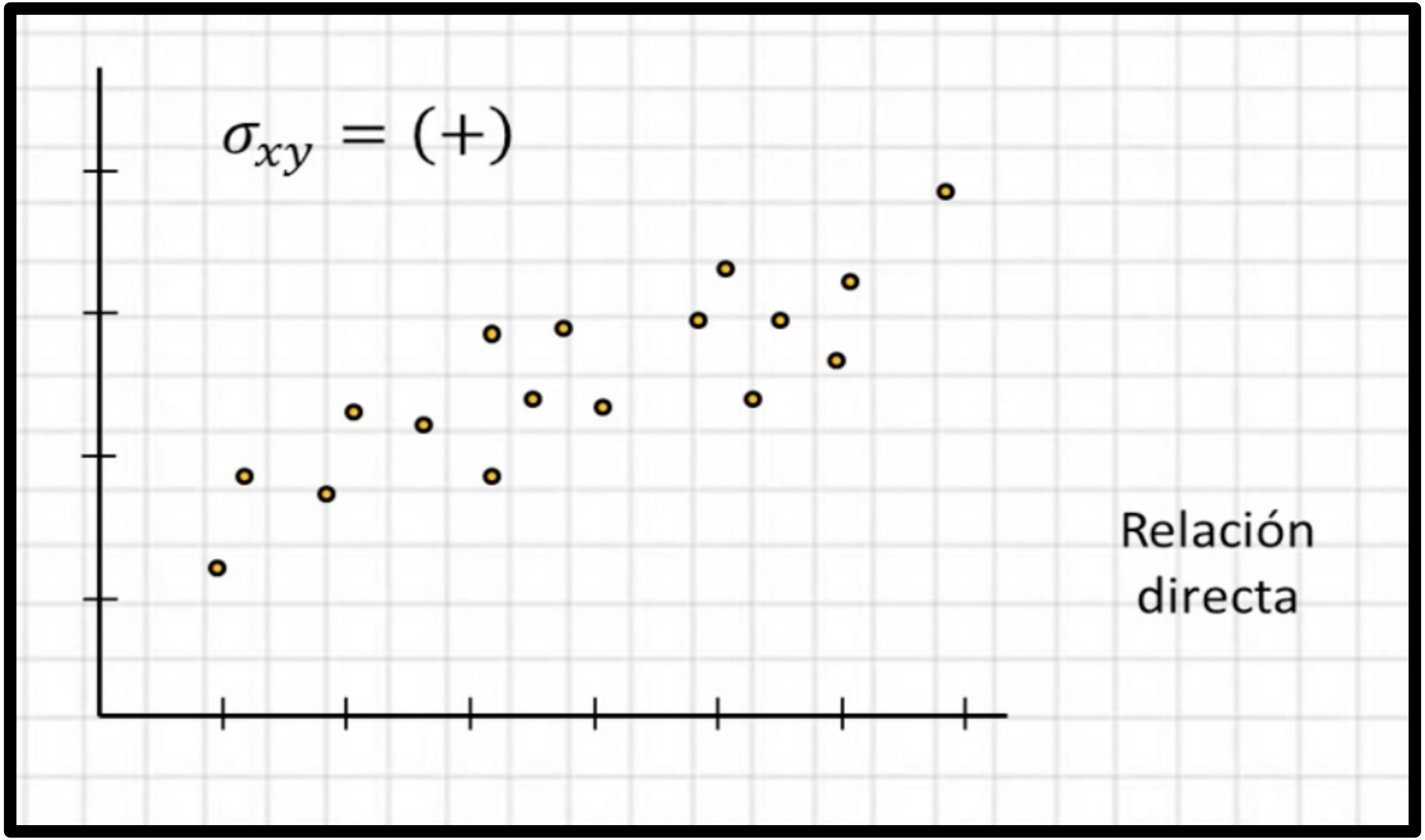
COVARIANZA

Es el valor que refleja en qué cuantía dos variables aleatorias varían de forma conjunta respecto a sus medias.

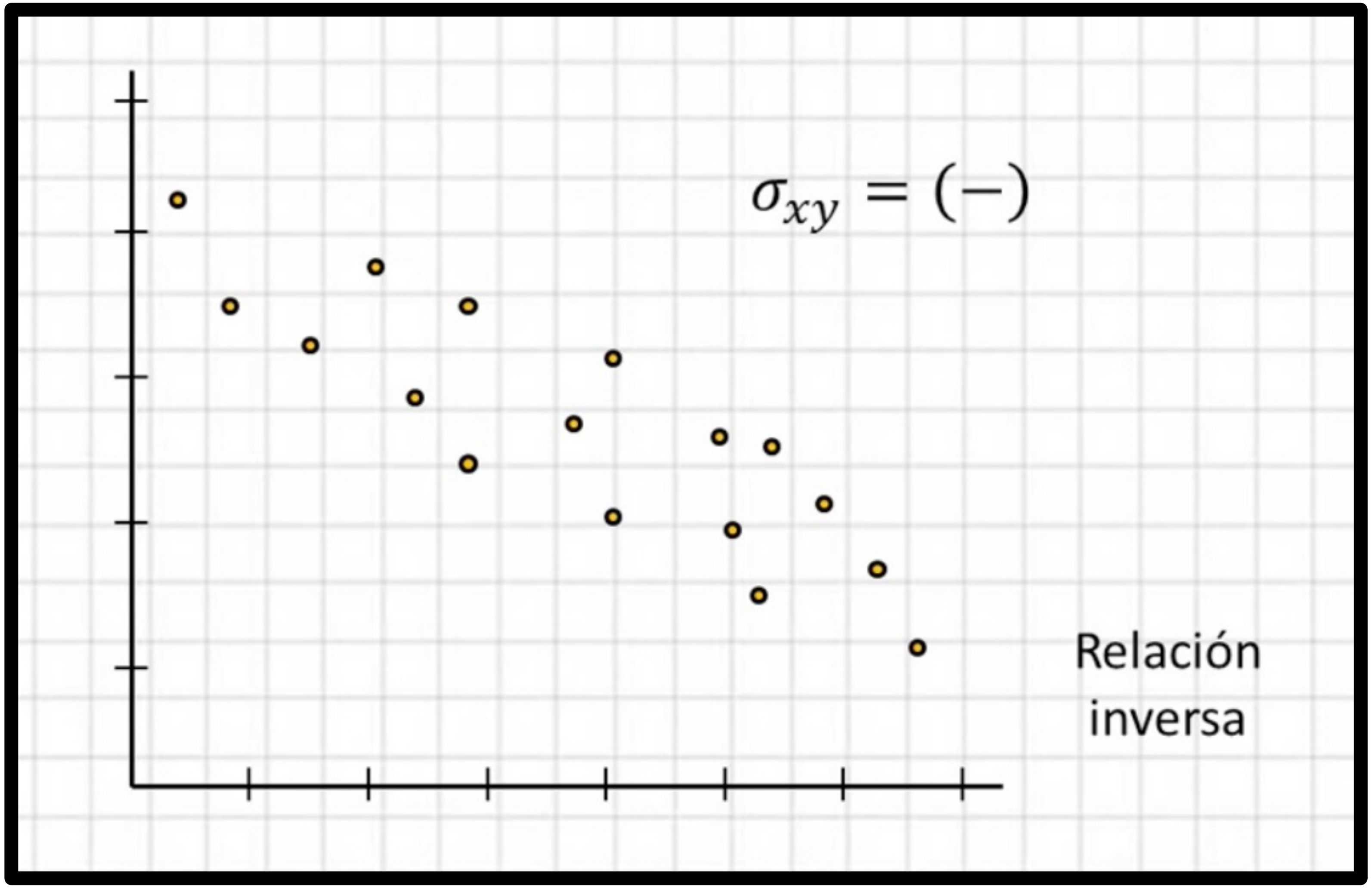
Nos permite saber cómo se comporta una variable en función de lo que hace otra variable.

Por ejemplo: cuando X sube ¿Cómo se comporta Y?

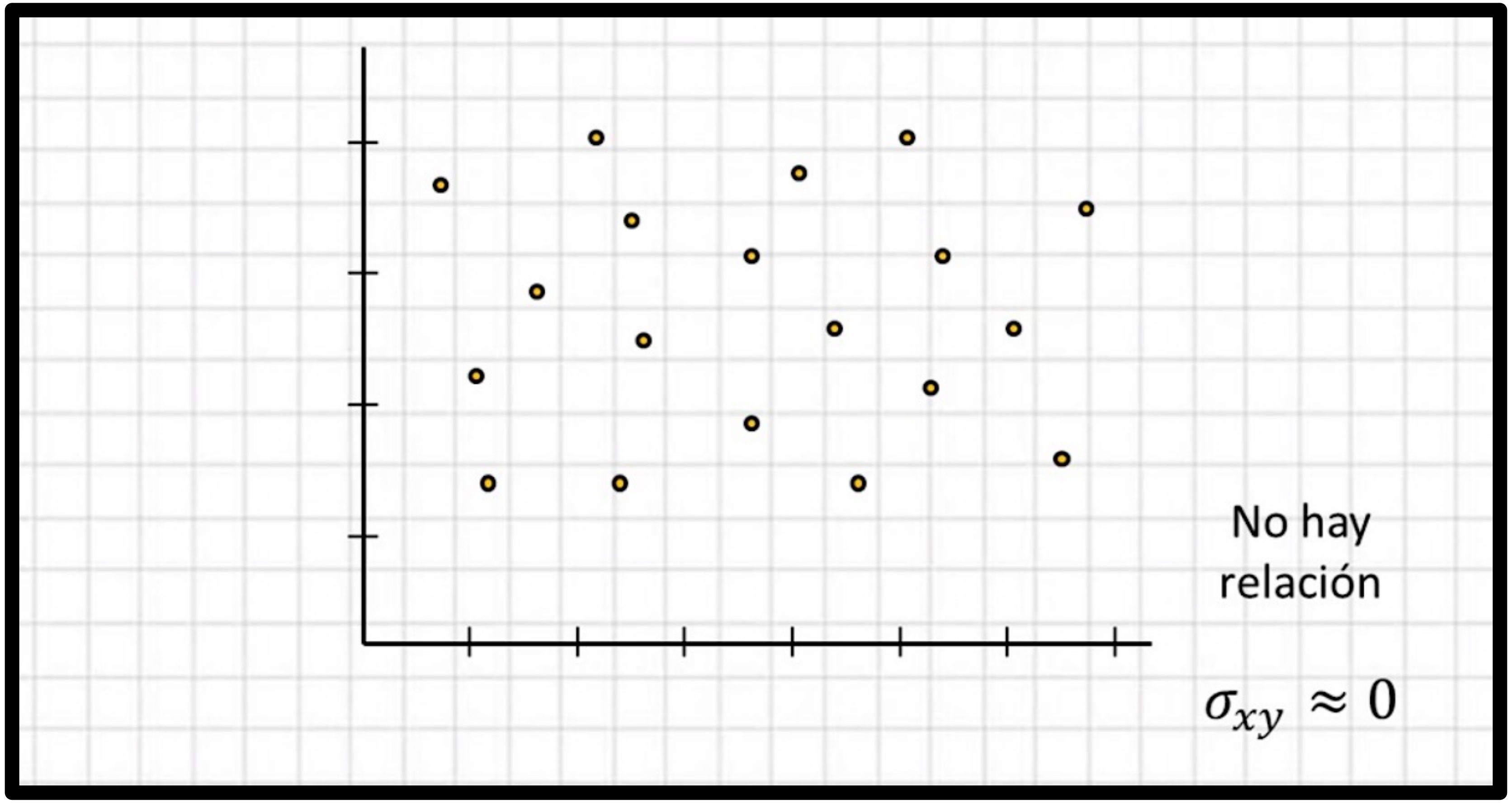
Cuando entre las dos variables hay una **relación directa**, la **covarianza** da un valor **positivo**



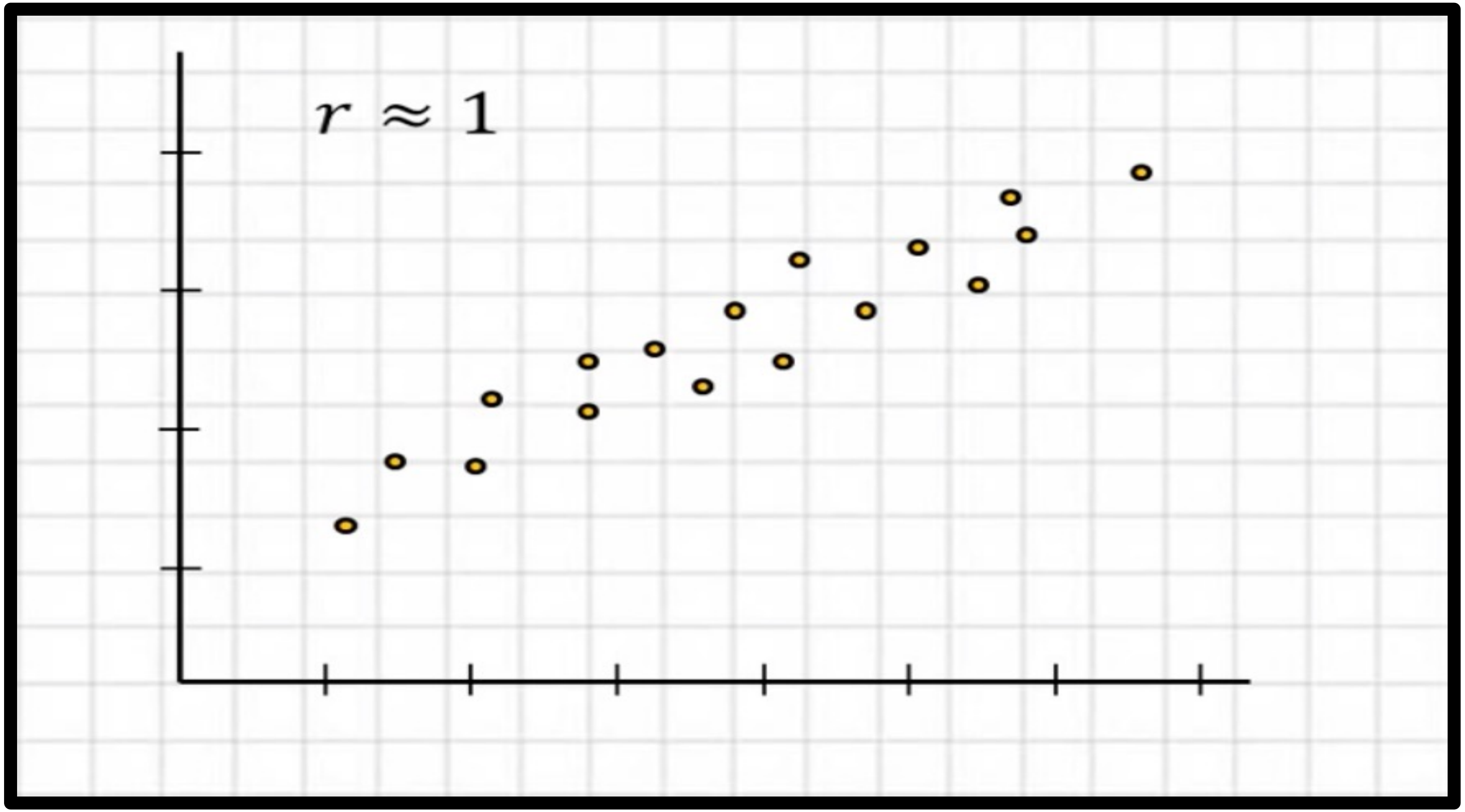
Cuando entre las dos variables hay una **relación inversa**, la **covarianza** da un valor **negativo**



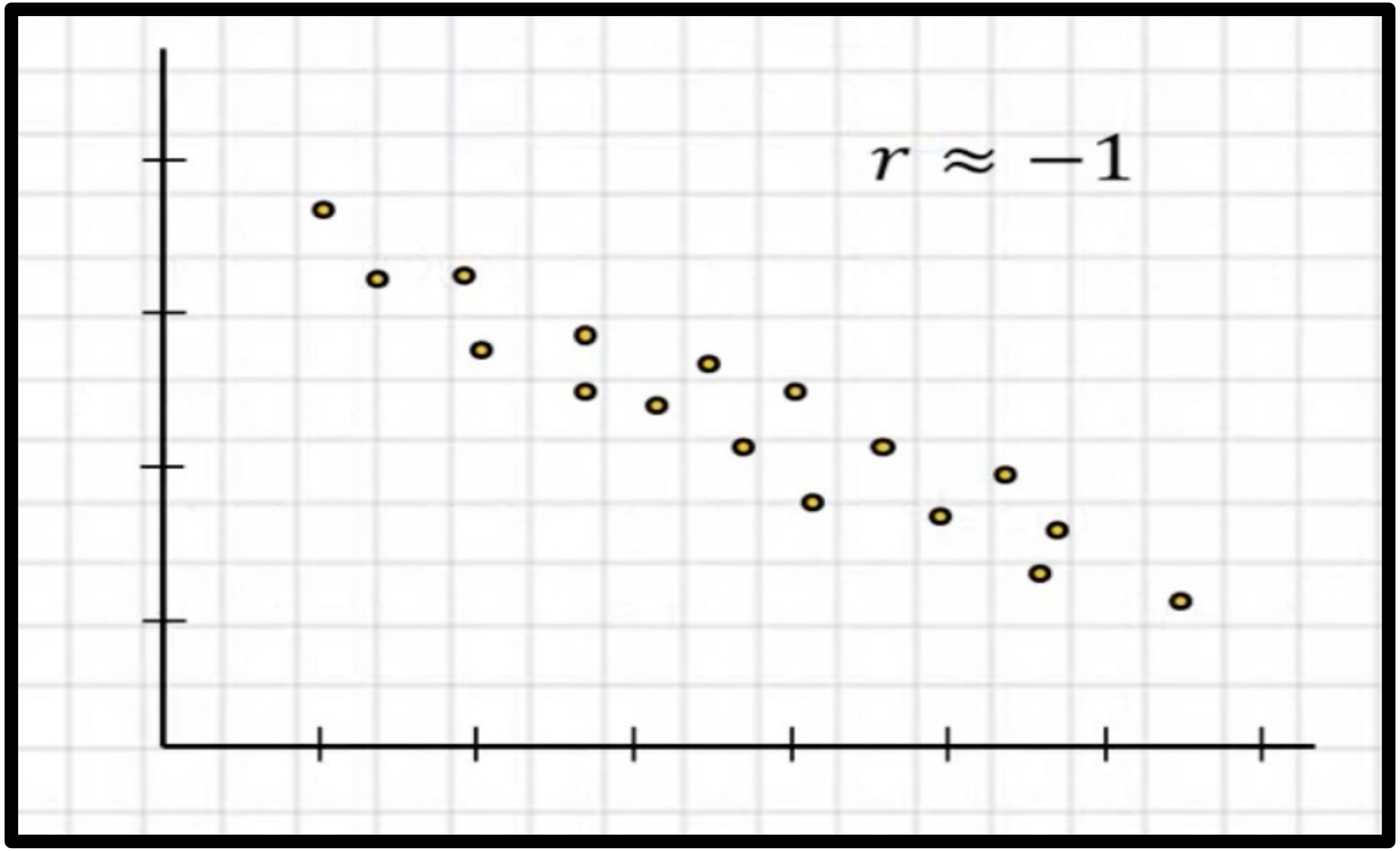
Cuando entre las dos variables **no hay una relación**, la **covarianza** da un valor **en torno a cero**



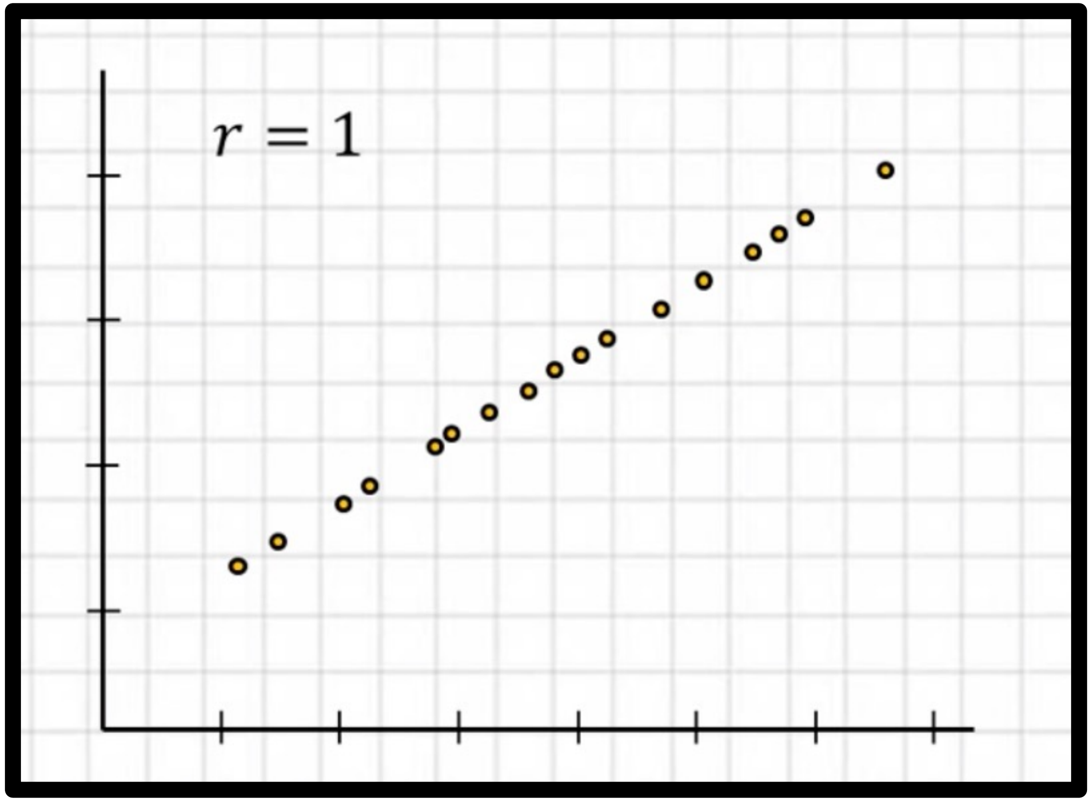
Si $r \approx 1$ existe una correlación directa fuerte

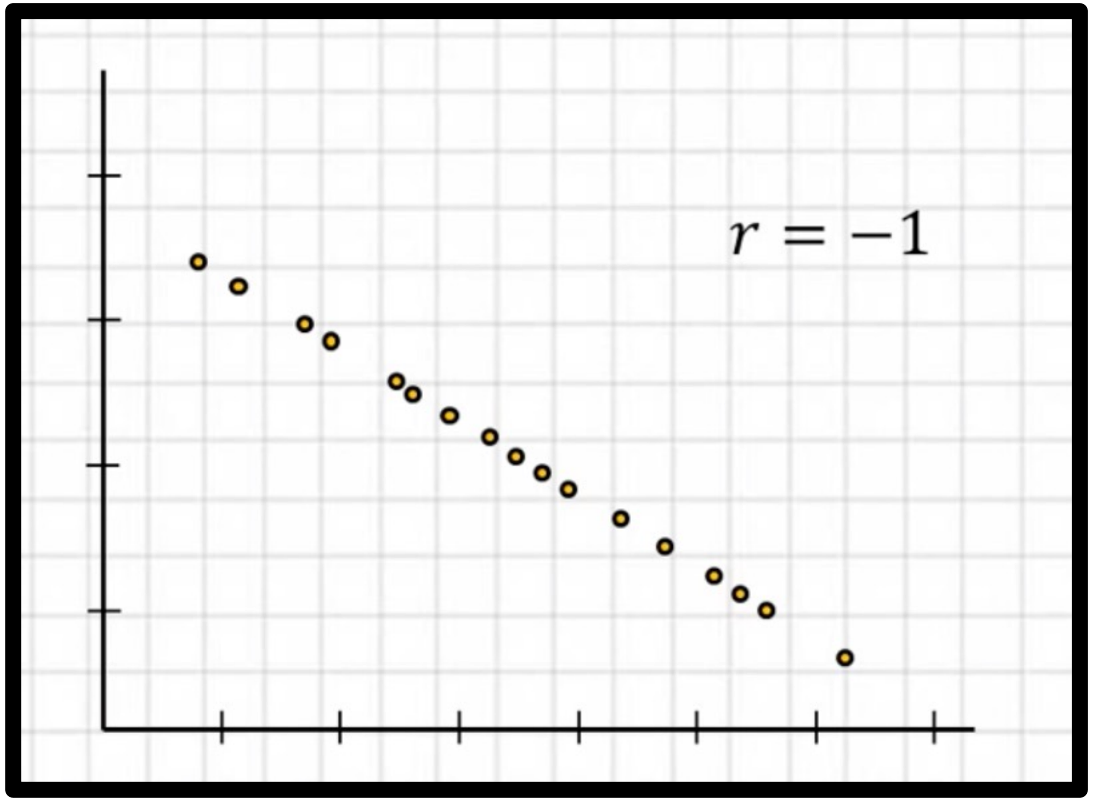


Si $r \approx -1$ existe una correlación inversa fuerte

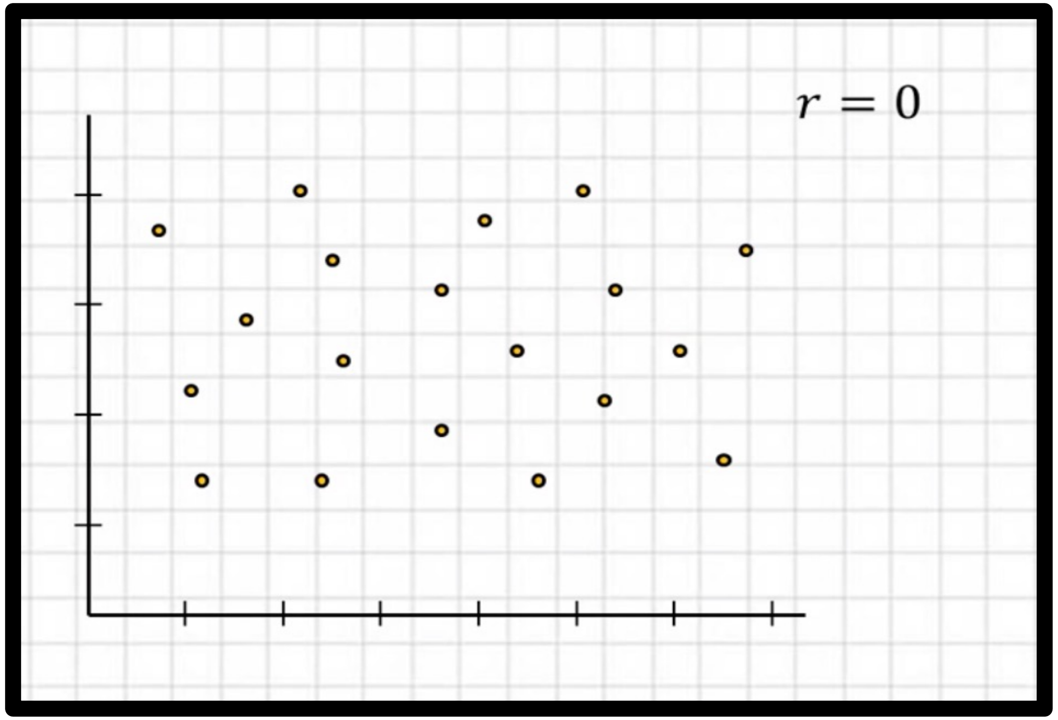


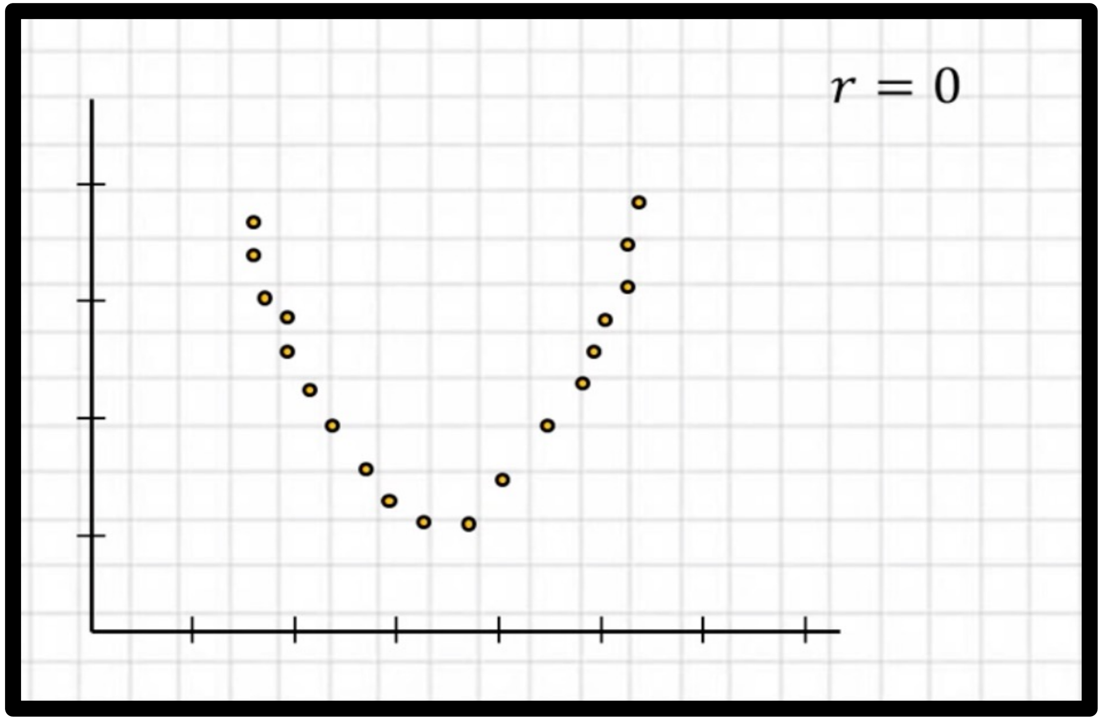
Si $r = 1$ o $r = -1$ hay una correlación funcional





Si $r \approx 0$ no existe una correlación lineal





COEFICIENTE PARAMÉTRICAS

Se usa cuando las variables a comparar son CUANTITATIVAS.

- ✓ PEARSON

COEFICIENTE NO PARAMÉTRICAS

Se usa cuando una o ambas variables son de escala ordinal, o cuando las variables cuantitativas no siguen una distribución normal.

- ✓ SPEARMAN

CONSIDERAR LO SIGUIENTE:

FORMULACIÓN DE HIPÓTESIS

Ho: No hay correlación

H1: Si hay correlación

NIVEL DE SIGNIFICANCIA

5%, p=0.05

INDICE Y EL NIVEL DE CORRELACIÓN

0.00-0.099 = nula

0.10-0.40 = escasa

0.40-0.60= moderada

0.60-0.80= buena

0.80-1.0= fuerte

PEARSON

$$r_{xy} = \frac{\dfrac{\sum XY}{N} - \overline{X}\,\overline{Y}}{S_x S_y}$$

$$S_x = \sqrt{\frac{\sum X^2}{N} - \overline{X}^2}$$

$$S_x = \sqrt{\frac{\sum Y^2}{N} - \overline{Y}^2}$$

X	Y	X^2	Y^2	XY
105	4			
116	8			
103	2			
124	7			
137	9			
126	9			
112	3			
129	10			
118	7			
105	6			
1175	**65**			

Correlaciones.xlsx

X	Y	X^2	Y^2	XY
105	4	11025	16	420
116	8	13456	64	928
103	2	10609	4	206
124	7	15376	49	868
137	9	18769	81	1233
126	9	15876	81	1134
112	3	12544	9	336
129	10	16641	100	1290
118	7	13924	49	826
105	6	11025	36	630
1175	65	139245	489	7871

$$\bar{X} = \frac{\sum X}{N} =$$

$$\bar{Y} = \frac{\sum Y}{N} =$$

$$\bar{X} = \frac{\sum X}{N} = \frac{1175}{10} = 117.5$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{65}{10} = 6.5$$

X	Y	X^2	Y^2	XY
105	4	11025	16	420
116	8	13456	64	928
103	2	10609	4	206
124	7	15376	49	868
137	9	18769	81	1233
126	9	15876	81	1134
112	3	12544	9	336
129	10	16641	100	1290
118	7	13924	49	826
105	6	11025	36	630
1175	65	139245	489	7871

$$S_x = \sqrt{\frac{\sum X^2}{N} - \bar{X}^2} =$$

$$S_x = \sqrt{\frac{\sum Y^2}{N} - \bar{Y}^2} =$$

$$S_x = \sqrt{\frac{\sum X^2}{N} - \bar{X}^2} = \sqrt{\frac{139245}{10} - 117.5^2} = 10.874$$

$$S_x = \sqrt{\frac{\sum Y^2}{N} - \bar{Y}^2} = \sqrt{\frac{489}{10} - 6.5^2} = 2.579$$

X	Y	X^2	Y^2	XY
105	4	11025	16	420
116	8	13456	64	928
103	2	10609	4	206
124	7	15376	49	868
137	9	18769	81	1233
126	9	15876	81	1134
112	3	12544	9	336
129	10	16641	100	1290
118	7	13924	49	826
105	6	11025	36	630
1175	65	139245	489	7871

$$r_{xy} = \frac{\frac{\sum XY}{N} - \overline{X}\overline{Y}}{S_x S_y} =$$

$$r_{xy} = \frac{\frac{\sum XY}{N} - \overline{X}\overline{Y}}{S_x S_y} = \frac{\frac{7871}{10} - 117.5 * 6.5}{10.874 * 2.579} = 0.8327$$

EJERCICIO EN EXCEL

MUJERES INDÍGENAS RURALES