USAP
UNIVERSIDAD DE SAN PEDRO SULA

3. Correlación

Introducción

En los negocios, no todo es el producto, pueden existir factores relacionados o externos que modifiquen cómo se distribuye un producto.

De igual manera, la estadística no todo está supeditado al uso de una variable, también existen técnicas para analizar más de una variable de forma simultánea e interrelacionada.

Existen varias técnicas para hacer análisis de la relación entre dos variables, algunas de ellas pueden ser tan simples como el diagrama de dispersión; universalmente aceptada en todo tipo de empresas, hasta la regresión lineal simple o múltiple.

Diagrama de dispersión

Es la presentación gráfica que muestra la relación de dos variables. Al estar involucradas dos variables, una de ellas se considera la independiente y la otra la dependiente. Al igual que en Matemáticas, la independiente corresponde a la variable X y la dependiente corresponde a la variable Y.

Ver la tendencia que muestra el diagrama puede dar una idea al usuario de cuál es la correlación que se puede esperar en la muestra y dependiendo del caso proyectarlo hacia la población.

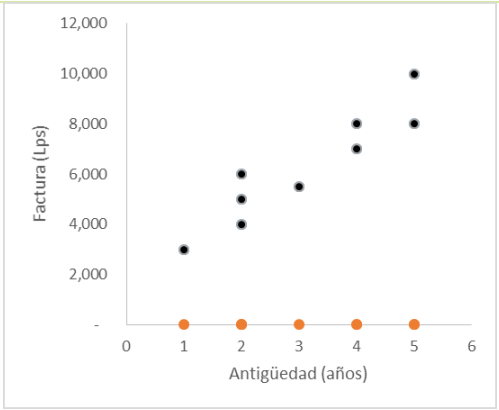
Ejemplo 3.1

1. La Empresa MOTORSI se da mantenimiento preventivo a vehículos turismo. Se tomó u na muestra para evaluar si el valor del pago tiene alguna relación con la antigüedad de los clientes. Se tomó una muestra de 9 clientes que visitaron MOTORSI la semana pasada y a través de un diagrama de dispersión evaluar su comportamiento. El resultado de la muestra es el siguiente:

AÑOS	FACTURACIÓN
1	3,000
2	4,000
5	8,000
4	8,000
2	5,000
3	5,500
4	7,000
2	6,000
5	10,000

Desarrollo

En un plano cartesiano se grafica en el eje X la antigüedad del cliente y en el eje Y el valor facturado en la última visita. Utilizando Excel, se muestra la siguiente gráfica.



Se puede observar que mientras más antigüedad tienen los clientes, es mayor el valor de la facturación.

Lo más usado para el trazo de las gráficas son los paquetes estadístico; el más común es Microsoft Excel.

COMANDO EN EXCEL

Para para elaborar un Diagrama de dispersión en Excel:

1) Insertar →  Gráfico de dispersión (x,y)

Análisis de correlación

El análisis de correlación es el estudio de la relación entre variables numéricas. Es lo mismo que se observó en el diagrama de dispersión con base numérica.

> "ANÁLISIS DE CORRELACIÓN: Grupo de técnicas para medir la asociación entre dos variables." (Lind |Marchal |Wathen, 2008, p.459).

El primer paso para hacer el análisis de correlación es el cálculo del coeficiente de correlación, técnica descubierta por Carl Pearson, que estandariza la medida de las variables hasta crear un intervalo que oscila entre -1 y 1.

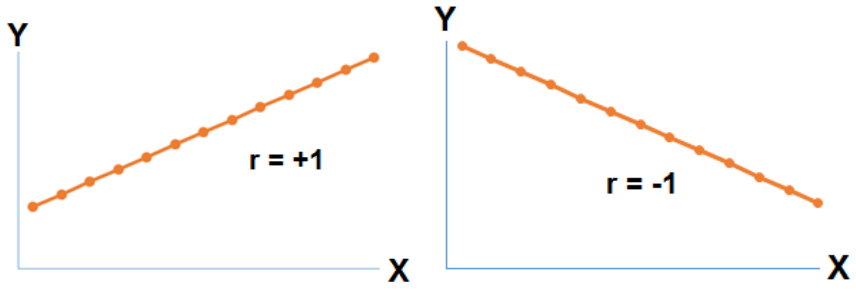
Coeficiente de correlación

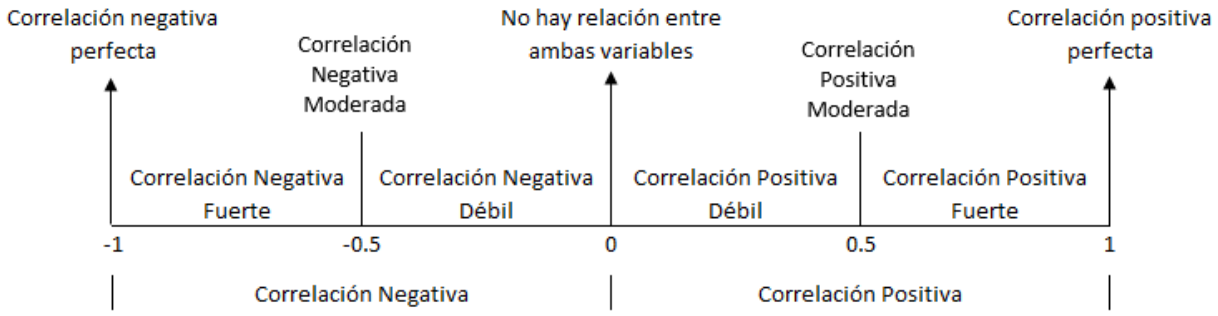
Es la técnica para probar la fuerza de la relación entre dos variables continuas, en la cual una es independiente y la otra es dependiente.

El coeficiente de correlación es conocido como ***r-Pearson*** o simplemente ***r***, que puede adoptar un valor entre -1 y 1. Las características de un coeficiente de correlación son:

a. $r = -1$ ó $r = 1$ Correlación perfecta.
b. $r = 0$ No hay relación lineal
c. $r > 0$ Correlación positiva
d. $r < 0$ Correlación negativa

Al generar un diagrama de dispersión, con variables continuas, se puede visualizar la tendencia que tendría una recta que haya sido creada a través del análisis de todos los puntos del diagrama. Si la correlación es positiva, la tendencia orienta el diagrama con una recta positiva y creciente y en la relación negativa, la tendencia la orienta hacia una recta negativa o decreciente.





"COEFICIENTE DE CORRELACIÓN: Medida de la fuerza de la relación lineal entre dos variables." (Lind |Marchal |Wathen, 2008, p.462).

La fórmula del coeficiente de correlación es una combinación de Media aritmética, desviación estándar y el tamaño de la muestra.

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{(n-1)s_X s_Y}$$

Donde:

- X : Cada una de las observaciones de la variable independiente.
- Y : Cada una de las observaciones de la variable dependiente
- $\bar{X}$: Media aritmética muestral de la variable independiente
- $\bar{Y}$: Media aritmética muestral de la variable dependiente
- s_X : Desviación estándar de la variable independiente
- s_Y : Desviación estándar de la variable dependiente
- n : Tamaño de la muestra

A manera de repaso, las fórmulas de la media aritmética y la desviación estándar son:

$$\bar{X} = \frac{\sum X_i}{n}$$

$$s_X = \sqrt{\frac{\sum(X_i - \bar{X})^2}{(n-1)}}$$

Ejemplo 3.2

1. En la empresa Sara se venden unidades de aire acondicionado; se ha observado que a mayor cantidad de llamadas de los vendedores durante el mes, mayor cantidad de compra de unidades de aire acondicionado.
   Se tomó una muestra de las ventas realizadas por 6 de los vendedores de planta y se quiere comparar la cantidad de llamadas realizadas durante el mes y las ventas facturadas.

   Los resultados de la muestra fueron los siguientes:

AGENTE	LLAMADAS	UNIDADES VENDIDAS
Tomás García	20	30
José Girón	40	60
Gregorio Figueroa	30	60
Carlos Ramírez	10	40
Miguel Godoy	20	50
Marcos Reyes	20	30

   Calcular el coeficiente de correlación

   a. Diagrama de dispersión:
      El diagrama de dispersión en Excel se visualiza de la siguiente manera:

   b. Media aritmética de ambas muestras

$$\bar{X} = \frac{20 + 40 + 30 + 10 + 20 + 20}{6} = \frac{140}{6} = 23.0$$

$$\bar{Y} = \frac{30 + 60 + 60 + 40 + 50 + 30}{6} = \frac{270}{6} = 45.0$$

   c. Variación y Desviación estándar de ambas muestras

AGENTE	LLAMADAS	UNIDADES VENDIDAS	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X}) * (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
Tomás García	20	30	$(20 - 23)^2 =$ -3	$(20 - 45)^2 =$ -15	45	9	225
José Girón	40	60	$(40 - 23)^2 =$ 17	$(60 - 45)^2 =$ 15	255	289	225
Gregorio Figueroa	30	60	$(30 - 23)^2 =$ 7	$(60 - 45)^2 =$ 15	105	49	225
Carlos Ramírez	10	40	$(10 - 23)^2 =$ -13	$(40 - 45)^2 =$ -5	65	169	25
Miguel Godoy	20	50	$(20 - 23)^2 =$ -3	$(50 - 45)^2 =$ 5	-15	9	25
Marcos Reyes	20	30	$(20 - 23)^2 =$ -3	$(30 - 45)^2 =$ -15	45	9	225
Σ	140	270			500	534	950

$$s_X = \sqrt{\frac{534}{6-1}} = \sqrt{106.7} = 10.3$$

$$s_Y = \sqrt{\frac{950}{6-1}} = \sqrt{190.0} = 13.8$$

d. Coeficiente de correlación

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{(n-1)s_X s_Y}$$

$$r = \frac{500}{(6-1)(10.3)(13.8)}$$

$$r = \frac{500}{712.25} = 0.702$$

La correlación entre ambas variables es positiva y fuerte.
El hacer llamadas telefónicas a los posibles clientes nos llevó a un incremento en las ventas.

Coeficiente de determinación

El obtener una respuesta como moderada, fuerte o perfecta no dice mucho en términos numéricos, ya que se puede interpretar con una respuesta ambigua. Para apoyar la respuesta, se utiliza el Coeficiente de determinación que proporciona un resultado en porcentaje, el cual es más fácil de interpretar. Se calcula elevando al cuadrado el coeficiente de correlación.

USAP
UNIVERSIDAD DE SAN PEDRO SULA

"COEFICIENTE DE DETERMINACIÓN: Proporción de la variación total en la variable dependiente Y que se explica, o contabiliza, por la variación en la variable independiente X." (Lind |Marchal |Wathen, 2008, p.465).

Ejemplo 3.3

1. Calcular el coeficiente de determinación de una muestra de dos variables, cuyos coeficiente de correlación es 0.702

Desarrollo

$$r^2 = (0.702)^2$$
$$r^2 = 0.4928$$

Existe una correlación de 49% entre ambas variables.

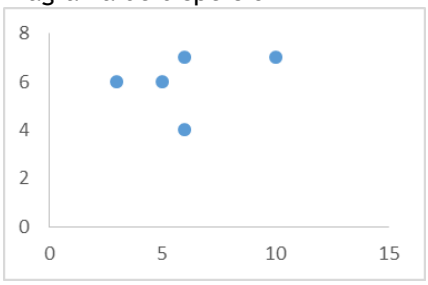
2. En una muestra de 5 elementos, los datos de la variable independiente y la dependiente se obtuvieron de la siguiente forma:

X	6	5	3	6	10
Y	4	6	6	7	7

Calcular el coeficiente de determinación.

Desarrollo

a) Diagrama de dispersión



b) Media aritmética de X y de Y

$$\bar{X} = \frac{6+5+3+6+10}{5} = \frac{30}{5} = 6$$

$$\bar{Y} = \frac{4+6+6+7+7}{5} = \frac{30}{5} = 6$$

c) Variación y varianza de X y Y

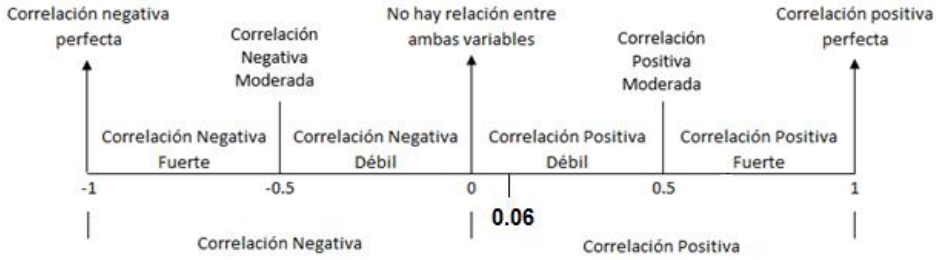
X	Y	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X}) * (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
6	4	$(6-7)^2 =$ -1	$(4-6)^2 =$ -2	2	1	4
8	5	$(-7)^2 =$ 1	$(5-6)^2 =$ -1	-1	1	1
3	7	$(3-7)^2 =$ -4	$(7-6)^2 =$ 1	-4	16	1
7	7	$(7-7)^2 =$ -	$(7-6)^2 =$ 1	-	-	1
11	7	$(11-7)^2 =$ 4	$(7-6)^2 =$ 1	4	16	1
35	30			1	34	8

$$s_X = \sqrt{\frac{34}{5-1}} = \sqrt{8.5} = 2.92$$

$$s_Y = \sqrt{\frac{8}{5-1}} = \sqrt{2} = 1.4$$

d) Coeficiente de correlación

$$r = \frac{1}{(5-1)(2.92)(1.4)} = \frac{1}{16} = 0.06$$



Existe relación positiva débil entre ambas variables.

e) Coeficiente de determinación

$$r^2 = (0.06)^2 = 0.004$$

Solo se puede suponer un 0.4% de correlación, lo que indica que la variable X no influye en el resultado de la variable Y.

Prueba de la importancia del coeficiente de correlación

Aunque un coeficiente de determinación sea alto, el resultado hace referencia a una muestra; para inferir sobre los resultados de la población, se recurre a la prueba de hipótesis; es decir, se somete el coeficiente de correlación a una prueba con el estadístico t.

El fin de la prueba es llegar a concluir que el coeficiente de correlación de la población es 0; es decir, que las variables no se relacionan. La hipótesis adecuada para este tipo de es:

$$H_0: \rho = 0$$
$$H_a: \rho \neq 0$$

Es de hacer notar que la hipótesis nula es una igualdad, por lo tanto, la prueba se debe realizar para 2 colas. La fórmula del estadístico de prueba es:

$$t = \frac{r(n-2)}{\sqrt{1-r^2}} \qquad con\ n-2\ grados\ de\ libertad$$

Si bien es cierto, la prueba se realiza con una sola muestra; pero, el análisis es con 2 variables, así que se hace para n-2 grados de libertad (1 grado de libertad por cada variable).

Para probar la importancia del coeficiente de correlación en la población, se prueba la hipótesis utilizando el método de los 5 pasos.

Ejemplo 3.4

1. En la empresa Sara se venden unidades de aire acondicionado; se ha observado que a mayor cantidad de llamadas de los vendedores durante el mes, mayor cantidad de compra de unidades de aire acondicionado.

   Se tomó una muestra de las ventas realizadas por 6 de los vendedores de planta y se quiere comparar la cantidad de llamadas realizadas durante el mes y las ventas facturadas; el coeficiente de correlación obtenido fue de 0.702. Se va a probar si existe relación entre las variables con un nivel de confianza del 95%.

AGENTE	LLAMADAS	UNIDADES VENDIDAS
Tomás García	20	30
José Girón	40	60
Gregorio Figueroa	30	60
Carlos Ramírez	10	40
Miguel Godoy	20	50
Marcos Reyes	20	30

Desarrollo

PASO 1: Hipótesis nula y alternativa

$$H_0: \rho = 0$$
$$H_a: \rho \neq 0$$

PASO 2: Nivel de significancia

$$\alpha = 0.05$$

PASO 3: Estadístico de prueba

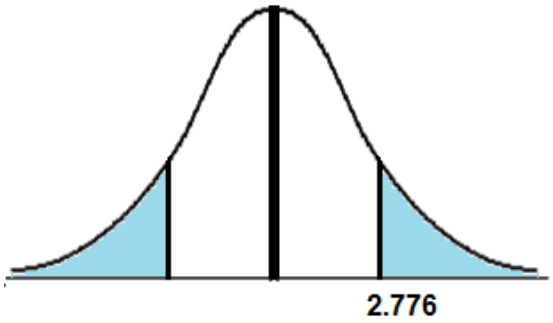
$$t = \frac{r(n-2)}{\sqrt{1-r^2}}$$

PASO 4: Regla de decisión

$H_0: \rho = 0$
$2\ colas\ (distribución\ t)$
$\alpha = 0.05$
$n = 6$
$gl = 6 - 2 = 4$

	Intervalo de confianza, *c*					
	80%	90%	95%	98%	99%	99.9%
	Nivel de significancia para una prueba de una cola, α					
gl	0.100	0.050	0.025	0.010	0.005	0.0005
	Nivel de significancia para una prueba de dos colas, α					
	0.200	0.10	0.05	0.02	0.01	0.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869

$$t = 2.776$$



PASO 5: Toma de decisión

$\boldsymbol{r = 0.702}$

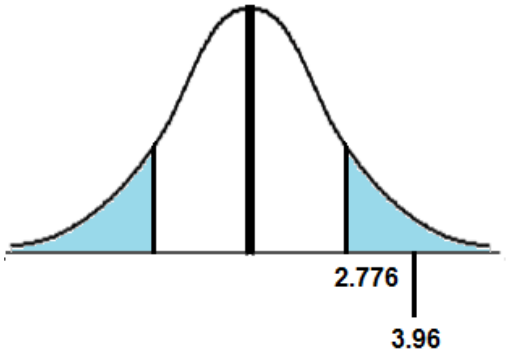
$\boldsymbol{n = 6}$

$$\boldsymbol{t = \frac{r(n-2)}{\sqrt{1-r^2}}}$$

$$\boldsymbol{t = \frac{0.702(6-2)}{\sqrt{1-(0.702)^2}}}$$

$$\boldsymbol{t = \frac{2.81}{0.71}}$$

$$\boldsymbol{t = 3.96}$$



La hipótesis nula se rechaza
La correlación de la población no es 0
Existe relación entre las variables

2. El departamento de producción de Celltronics International desea explorar la relación entre el número de empleados que trabajan en una línea de ensamble parcial y el número de unidades producido. Como experimento, se asignó a dos empleados al ensamble parcial. Su desempeño fue de 15 productos durante un periodo de una hora. Después, cuatro empleados hicieron los ensambles y su número fue de 25 durante un periodo de una hora. El conjunto completo de observaciones pareadas se muestra a continuación.

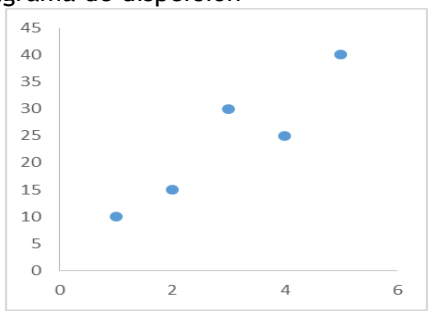
a) Trace un diagrama de dispersión.
b) Con base en el diagrama de dispersión, ¿parece haber alguna relación entre el número de ensambladores y la producción?
c) Calcular el coeficiente de correlación
d) Calcular el coeficiente determinación
e) Probar la importancia del coeficiente de correlación con un nivel de confianza del 95%.

Número de ensambladores	Producción en una hora (unidades)
2	15
4	25
1	10
5	40
3	30

USAP
UNIVERSIDAD DE SAN PEDRO SULA

Desarrollo

a) Diagrama de dispersión



b) Con base en el diagrama de dispersión, parece que a mayor cantidad de ensambladores, mayor producción.

c) Coeficiente de correlación

- Media aritmética de cada muestra

$$\bar{X} = \frac{2+4+1+5+3}{5} = 3$$

$$\bar{Y} = \frac{15+25+10+40+30}{5} = 24$$

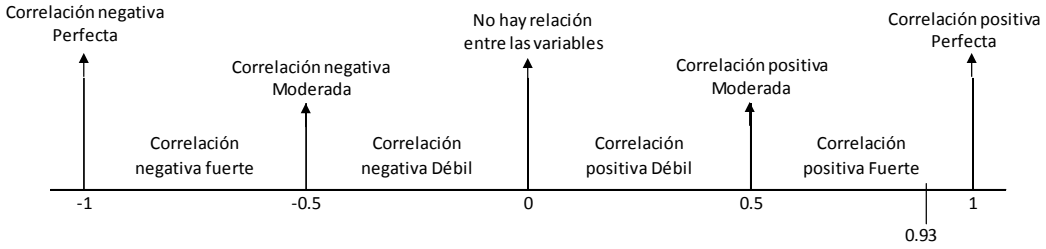
- Variación y desviación estándar de ambas muestras

Número de ensambladores (X)	Producción en una hora (Y)	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X}) * (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
2	15	-1	-9	9	1	81
4	25	1	1	1	1	1
1	10	-2	-14	28	4	196
5	40	2	16	32	4	256
3	30	0	6	0	0	36
3	**24**			**70**	**10**	**570**

$$s_X = \sqrt{\frac{10}{5-1}} = \sqrt{2.5} = 1.58$$

$$s_Y = \sqrt{\frac{570}{5-1}} = \sqrt{142.5} = 11.94$$

$$r = \frac{70}{(5-1)(1.58)(11.94)} = \frac{70}{75.46} = 0.93$$



Existe una correlación positiva fuerte.

d) Coeficiente de determinación

$$r^2 = (0.93)^2 = 0.86$$

Se puede suponer que hay una correlación del 86% entre ambas variables; lo que indica que la variable X influye en la variable Y.

e) Prueba de la importancia del coeficiente de correlación

PASO 1: Hipótesis nula y alternativa

$$H_0: \rho = 0$$
$$H_a: \rho \neq 0$$

PASO 2: Nivel de significancia

$$\alpha = 0.05$$

PASO 3: Estadístico de prueba

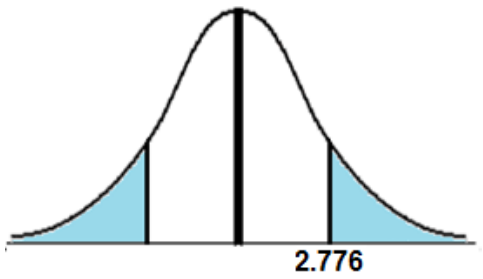
$$t = \frac{r(n-2)}{\sqrt{1-r^2}}$$

PASO 4: Regla de decisión

$H_0: \rho = 0$
$2\ colas\ (distribución\ t)$
$\alpha = 0.05$
$n = 5$
$gl = 5 - 1 = 4$

	Intervalo de confianza, *c*					
	80%	90%	95%	98%	99%	99.9%
	Nivel de significancia para una prueba de una cola, α					
gl	0.100	0.050	0.025	0.010	0.005	0.0005
	Nivel de significancia para una prueba de dos colas, α					
	0.200	0.10	0.05	0.02	0.01	0.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869

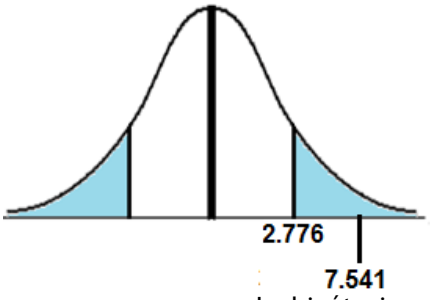
$t = 2.776$



PASO 5: Toma de decisión

$r = 0.93$
$n = 5$

$$t = \frac{r(n-2)}{\sqrt{1-r^2}}$$
$$t = \frac{0.93(5-2)}{\sqrt{1-(0.93)^2}}$$
$$t = \frac{2.79}{0.37}$$
$$t = 7.541$$



La hipótesis nula se rechaza

Existe relación entre el número de ensambladores y la producción por hora

Ejercicio

1. Se dan las siguientes hipótesis
   $H_0: \rho \leq 0$
   $H_a: \rho > 0$
   Una muestra aleatoria de 12 observaciones pareadas indicó una correlación de 0.32. ¿Se puede concluir que la correlación en la población es mayor que cero? Utilice 0.05 como nivel de significancia.
2. Se dan las siguientes hipótesis
   $H_0: \rho \leq 0$
   $H_a: \rho > 0$
   Una muestra aleatoria de 15 observaciones pareadas tiene una correlación de -0.46. ¿Se puede concluir que la correlación en la población es menor que cero? Con un nivel de significancia de 0.05.
3. La Refinería de Puesto Cortés estudia la relación entre el precio de la gasolina y el número de galones vendidos. Para una muestra de 20 gasolineras el martes pasado, la correlación fue de 0.78. con un nivel de significancia de 0.01, ¿Será mayor que cero la correlación en la población?
4. Un estudio de 20 instituciones financieras en todo el mundo revelo que la correlación entre sus activos y las utilidades antes del pago de impuestos es 0.86. Con un nivel de significancia de 0.05, ¿se puede concluir que hay una correlación positiva en la población?
5. El departamento de Servicios Estudiantiles de una Universidad local desea demostrar la relación entre al número de cervezas que consume un estudiante y su contenido de alcohol en la sangre. Una muestra de 9 estudiantes participó en un estudio en el cual a cada uno se le asignó, al azar, un número de latas de cerveza de 12 onzas que debía beber. 30 minutos después de consumir su número asignado de cervezas, un miembro del equipo evaluador midió su contenido de alcohol en la sangre. La información muestral fue la siguiente:

Estudiante	Cervezas	Contenido de alcohol en la sangre
1	6	0.10
2	7	0.09
3	7	0.09
4	4	0.10
5	5	0.10
6	3	0.07
7	3	0.10
8	6	0.12
9	6	0.09

a. Elaborar el diagrama de dispersión
b. Determinar el coeficiente de correlación
c. Establecer el coeficiente de determinación
d. Con un nivel de significancia de 0.01, ¿Es razonable concluir que hay una relación positiva en la población entre el número de cervezas consumidas y el contenido de alcohol en la sangre?

BIBLIOGRAFÍA

- Lind, D.A., Marchal, W.G., Wathen, S.A. (15). (2012). *Estadística Aplicada a los Negocios y la Economía*. México: McGraw-Hill

- David M. Levine, Timothy C. Krehbiel, Mark L. Berenson. 2006. *Estadística para Administración*. (4° edición). Naucalpan de Juárez, México.: Pearson Prentice Hall