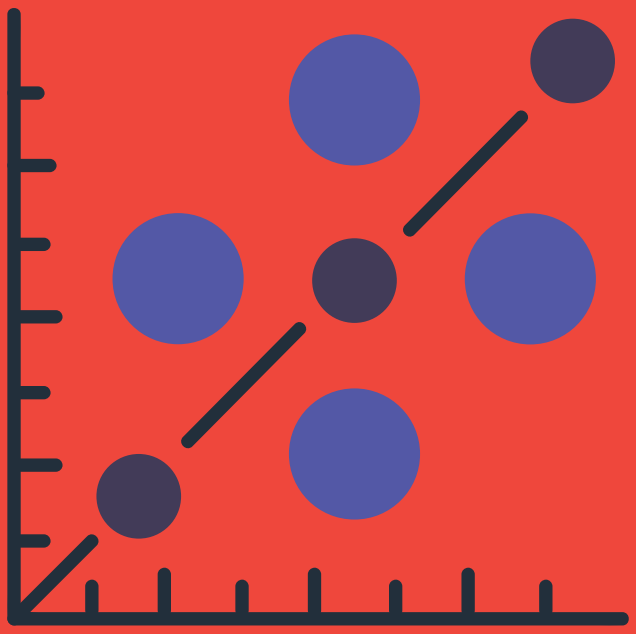
TALCA
UNIVERSIDAD
CHILE



Relación entre Variables

MATEMÁTICAS

VICERRECTORÍA DE PREGRADO
CIVE DIRECCIÓN DE CICLO DE VIDA ESTUDIANTIL
CIMA UNIDAD DE ACOMPAÑAMIENTO ESTUDIANTIL
PACE PROGRAMA DE ACCESO A LA EDUCACIÓN SUPERIOR
Ministerio de Educación

RUTA DE APRENDIZAJE

- El aprendizaje esperado en este documento es conocer medidas numéricas y gráficas para expresar la relación entre dos variables, así analizar el efecto de una sobre la otra.

ÍNDICE

INTRODUCCIÓN

En diversas áreas frecuentemente es necesario conocer la relación existente entre variables, para analizar el efecto que pudiese tener una sobre la otra. Por ejemplo, en salud, se podría tener interés en estudiar si la edad afecta la presión sanguínea, si existe relación en la ingesta de algún nutriente con el aumento de peso, si la concentración de algún medicamento altera la frecuencia cardiaca, etc. En economía las relaciones entre variables entrega información para aumentar las ganancias, reducir costos, estimar la demanda, entre otros.

Para llevar a cabo este objetivo, se utilizan algunas herramientas estadísticas apropiadas, por ejemplo, si el interés es solo describir la relación entre variables cuantitativas, se utiliza la **covarianza y/o gráfico de dispersión**. Si además se desea conocer el grado de asociación lineal entre las variables, se debe utilizar el **coeficiente de correlación de Pearson**.

Covarianza muestral

La covarianza se define como el valor esperado de las variaciones de dos variables con respecto a sus valores esperados. En palabras sencillas, **la covarianza es un indicador que muestra la existencia de asociación entre dos variables** (Kim, 2018).

Se calcula de la siguiente manera:

$$COV(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Otra manera de escribir esta fórmula es:

$$COV(X,Y) = \frac{(\sum_{i=1}^{n} x_i y_i) - n\overline{xy}}{n-1}$$

Interpretación de la covarianza

El valor de covarianza pertenece a los números reales y como indicador de asociación se puede interpretar de la siguiente manera:

- $COV(X,Y) > 0$, indica asociación **positiva** de las variables.
- $COV(X,Y) = 0$, indica asociación **nula o no asociación** de las variables.
- $COV(X,Y) < 0$, indica asociación **negativa** de las variables.

Recordando

La covarianza indica la **existencia de asociación**, si es que existe, pero **NO** qué tan asociadas están las variables en estudio.

Gráfico de dispersión

La manera más sencilla de analizar la relación entre dos variables es el gráfico de correlación o de dispersión. La forma de construir este gráfico es la siguiente:

- En cada uno de los ejes perpendiculares se coloca una de las variables estudiadas.
- La variable en el eje horizontal es la independiente "x" y en el eje vertical es la dependiente "y".
- La escala por eje oscila desde el valor mínimo y el máximo de la variable, sin necesidad de iniciar de cero, las escalas se proporcionan de tal manera que ambas tengan igual longitud.
- Se escribe cada unidad observada, representándola por un punto en la intersección de perpendiculares imaginarias, levantadas en los valores que le corresponden al individuo para cada variable.

Se logra así un gráfico de puntos cuya distribución nos informa sobre la existencia de correlación (Taucher, 2014).

"El gráfico de puntos nos revela la correlación, cuando los puntos se disponen en una nube elíptica y oblicua con respecto a los ejes. La correlación puede ser positiva o negativa. **Es positiva** cuando a valores bajos de x le corresponden valores bajos de y, y a valores altos de x le corresponden valores altos de y. **Es negativa** si al aumentar los valores de x los valores de y disminuyen.

La ausencia de correlación se manifiesta en el gráfico por una disposición circular, horizontal o vertical de los puntos" (Taucher, 2014).

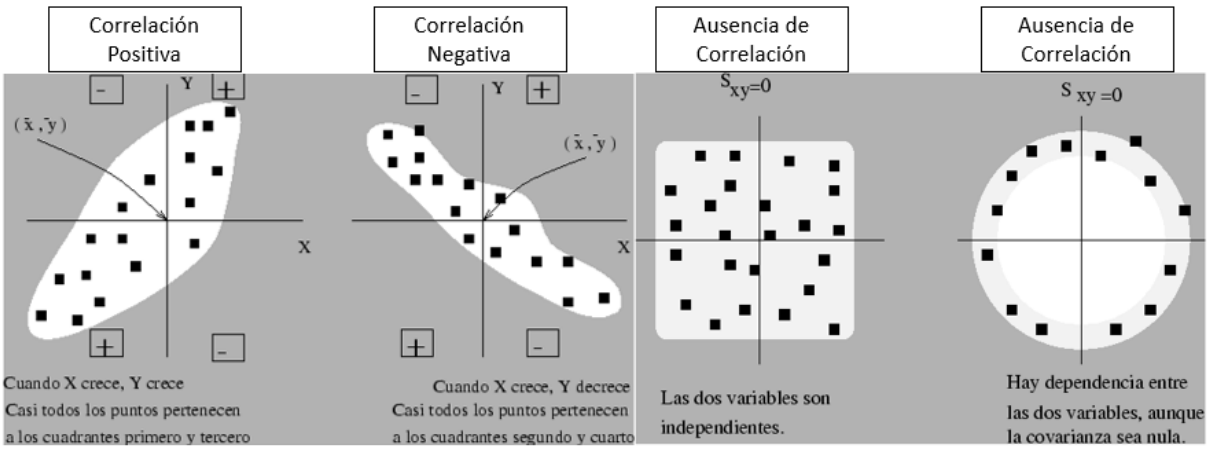


Figura 1

Interpretación del diagrama de dispersión

Si es complicado observar la existencia de correlación, se puede trazar líneas perpendiculares a los ejes, en los valores correspondientes a las medianas de las variables, esto adjudica dos mitades hacia la izquierda de la vertical y dos mitades hacia la derecha. Luego, se cuentan los puntos en cada uno de los cuadrantes obtenidos, si en dos cuadrantes diagonalmente opuestos la cantidad de puntos es superior a la que se encuentra en el otro sentido, decimos que hay correlación. Por ejemplo, en la imagen 2, los cuadrantes diagonales tienen 8 y 1 dato respectivamente, como 8 es superior a 1 se asume que existe correlación.

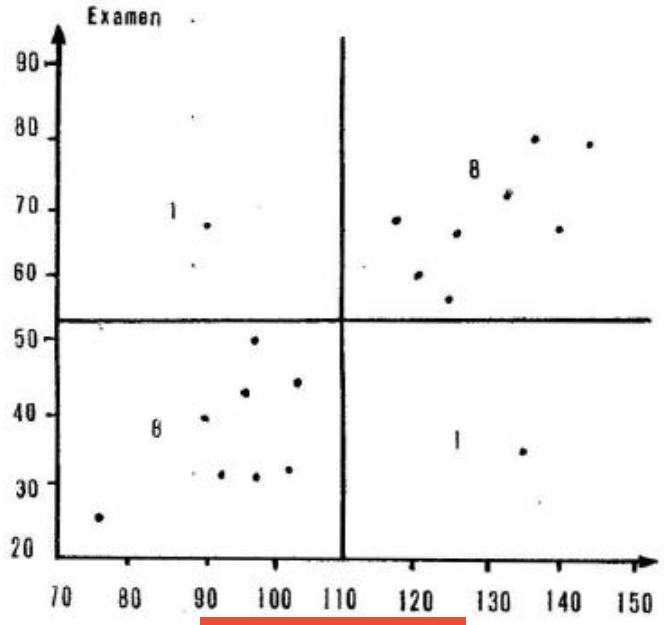


Figura 2

Diagrama de dispersión con ejes perpendiculares para analizar la correlación (Taucher, 2014).

Coeficiente de correlación "r"

Los gráficos son un método aproximado para medir la correlación, la medida más adecuada para apreciar el nivel de relación es el **coeficiente de correlación r de Pearson**.

Los requisitos para que el coeficiente de correlación sea una buena medida son:

- La correlación teórica sea una línea recta.
- Que las variables tengan una distribución normal bivariada.

Normalmente, se asume que estos requisitos se cumplen, pero si evidentemente no se llegasen a cumplir, hay otros métodos llamados **"no paramétricos"** que se pueden usar para medir correlación (Taucher, 2014).

La fórmula para el cálculo de r es:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{1=1}^{n} x_i \sum_{1=1}^{n} y_i}{\sqrt{(n\Sigma x_i^2 - (\Sigma x_i)^2)(n\Sigma y_i^2 - (\Sigma y_i)^2)}}$$

Otra manera de escribir esta fórmula es:

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{(\Sigma(x - \bar{x})^2)(\Sigma(y - \bar{y})^2)}}$$

Interpretación de la correlación

La medida de correlación puede variar entre -1 y 1, la correlación es más estrecha si se encuentra más cercana a esos valores. De esta manera se tiene:

- $\boldsymbol{r \to -1}$, existe relación **negativa** o **inversa** entre las variables de estudio.
- $\boldsymbol{r \to 0}$, **NO** existe relación entre las variables de estudio.
- $\boldsymbol{r \to 1}$, existe relación **positiva** o **directa** entre las variables de estudio.

Y se interpreta como el porcentaje de asociación entre las variables de estudio (Taucher, 2014).

Recordando

La covarianza indica la **existencia de asociación** y la correlación indica el **grado de asociación lineal.**

EJERCICIOS RESUELTOS

A continuación, se presentan ejercicios resueltos con sus procedimientos, en estos se sugiere hacer lo siguiente:

- Lee comprensivamente.
- Revisa el paso a paso.
- Destaca lo que te resulte importante.
- Destaca lo que te genere dudas y luego consulta al tutor.

1. **Se investiga si la capacidad vital depende de la edad en niños, para eso se toma una muestra de 8 niños de diversas edades con los siguientes resultados:**

Tabla 1: Problema página 104 de Bioestadística para carreras de la salud (Taucher, 2014)

Niño	Edad (años)	Capacidad Vital
1	4	0,79
2	5	0,93
3	6	1,15
4	7	1,29
5	8	1,47
6	9	1,71
7	10	1,87
8	11	1,99

a) **Realiza un gráfico de dispersión de los datos obtenidos.**

b) **Calcula la covarianza e interpretar los resultados.**

c) **Calcula la correlación de Pearson e interpretar los resultados.**

Desarrollo

a) Gráfico de dispersión

Paso 1: definir la variable dependiente e independiente

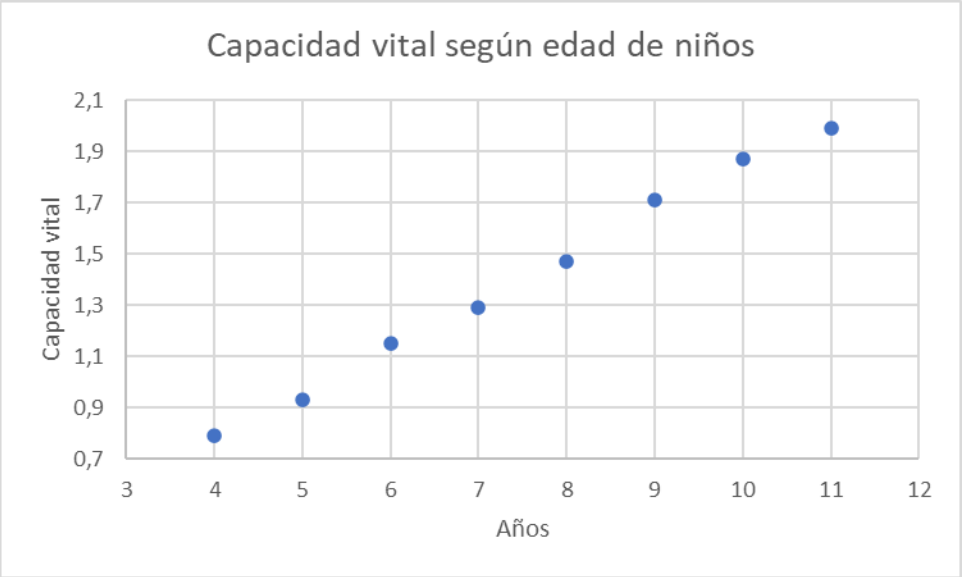
Como se estudia si la capacidad vital depende de la edad, la variable dependiente "y" es la capacidad vital y la variable independiente "x" es la edad.

Paso 2: establecer mínimos y máximos a utilizar

Como en el eje horizontal "x" el mínimo es 4, se inicia el gráfico en el 3. En el eje vertical "y" el menor valor es $0{,}79$, por lo que se opta iniciar en el $0{,}7$.

Paso 3: realizar el gráfico

Colocar en el plano los puntos según el par ordenado de los datos.



Paso 4: analizar el gráfico

Los puntos del gráfico parecen formar una recta creciente, por lo que se podría pensar que existe una relación directa o positiva entre la capacidad vital y la edad en niños.

b) Covarianza

Para esto se utiliza la fórmula:

$$COV(X,Y) = \frac{(\sum_{i=1}^{n} x_i y_i) - n\overline{xy}}{n-1}$$

Paso 1: calcular $\sum_{i=1}^{n} x_i y_i$

Niño	**Edad x (años)**	**Capacidad Vital y**	$x \cdot y$
1	4	0,79	$4 \cdot 0{,}79 = 3{,}16$
2	5	0,93	$5 \cdot 0{,}93 = 4{,}65$
3	6	1,15	$6 \cdot 1{,}15 = 6{,}9$
4	7	1,29	$7 \cdot 1{,}29 = 9{,}03$
5	8	1,47	$8 \cdot 1{,}47 = 11{,}76$
6	9	1,71	$9 \cdot 1{,}71 = 15{,}39$
7	10	1,87	$10 \cdot 1{,}87 = 18{,}7$
8	11	1,99	$11 \cdot 1{,}99 = 21{,}89$
TOTAL	60	11,2	91,48

Luego, $\Sigma x_i \cdot yi = 91{,}48$

Paso 2: calcular el promedio de cada variable

Recordar que la fórmula para el promedio es:

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

Luego se tiene:

Niño	Edad x (años)	Capacidad Vital y
1	4	0,79
2	5	0,93
3	6	1,15
4	7	1,29
5	8	1,47
6	9	1,71
7	10	1,87
8	11	1,99
TOTAL	60	11,2

Teniendo claro que $n = 8$, ya que es el tamaño de la muestra tenemos que:

$\bar{x} = \frac{60}{8} = 7{,}5$ y $\bar{y} = \frac{11{,}2}{8} = 1{,}4$

Paso 3: calcular $n\overline{xy}$

Reemplazando los valores tenemos:

$$n\overline{xy} = 8 \cdot 7{,}5 \cdot 1{,}4$$

$$n\overline{xy} = 84$$

Paso 4: reemplazar los valores obtenidos en los pasos anteriores en la fórmula de covarianza

Recordar que la fórmula de covarianza es

$$COV(X,Y) = \frac{(\sum_{i=1}^{n} x_i y_i) - n\overline{xy}}{n-1}$$

Reemplazando se tiene:

$$COV(X,Y) = \frac{91{,}48 - 84}{8 - 1}$$

Recordando

≈: aproximadamente.

$$COV(X,Y) = \frac{7{,}48}{7}$$

$$COV(X,Y) \approx 1{,}069$$

Paso 5: interpretar el resultado

Como el valor de la covarianza es positiva, entonces se podría suponer que la relación entre la capacidad vital y la edad de un niño es directa.

c) Correlación

Para calcular correlación se utiliza la siguiente fórmula:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{1=1}^{n} x_i \sum_{1=1}^{n} y_i}{\sqrt{(n\Sigma x_i^2 - (\Sigma x_i)^2)(n\Sigma y_i^2 - (\Sigma y_i)^2)}}$$

Paso 1: calcular $n\sum_{i=1}^{n} x_i y_i$

Recordemos por el ítem anterior que $\Sigma x_i \cdot yi = 91{,}48$

Luego,

$$n\Sigma x_i \cdot yi = 8 \cdot 91{,}48$$
$$n\Sigma x_i \cdot yi = 731{,}84$$

Paso 2: calcular $\Sigma x_i \Sigma y_i$

Es decir, multiplicar el total de la suma de los x e y

$$\Sigma x_i \Sigma y_i = 60 \cdot 11{,}2$$

$$\Sigma x_i \Sigma y_i = 672$$

Paso 3: determinar_Σx_i^2 y Σy_i^2

Niño	Edad x (años)	Capacidad Vital y	$x \cdot y$	x^2	y^2
1	4	0,79	3,16	$4^2 = 16$	$0{,}79^2 = 0{,}6241$
2	5	0,93	4,65	$5^2 = 25$	$0{,}93^2 = 0{,}8649$
3	6	1,15	6,9	$6^2 = 36$	$1{,}15^2 = 1{,}3225$
4	7	1,29	9,03	$7^2 = 49$	$1{,}29^2 = 1{,}6641$
5	8	1,47	11,76	$8^2 = 64$	$1{,}47^2 = 2{,}1609$
6	9	1,71	15,39	$9^2 = 81$	$1{,}71^2 = 2{,}9241$
7	10	1,87	18,7	$10^2 = 100$	$1{,}87^2 = 3{,}4969$
8	11	1,99	21,89	$11^2 = 121$	$1{,}99^2 = 3{,}9601$
TOTAL	60	11,2	91,48	492	17,0176

Luego, $\Sigma x_i^2 = 492$ y $\Sigma y_i^2 = 17{,}0176$

Paso 4: calcular $(\Sigma x_i)^2$ y $(\Sigma y_i)^2$

$$(\Sigma x_i)^2 = 60^2$$
$$(\Sigma x_i)^2 = 3600$$

$$(\Sigma y_i)^2 = 11{,}2^2$$
$$(\Sigma y_i)^2 = 125{,}44$$

Paso 5: reemplazar los resultados obtenidos en la fórmula

Recordemos que la fórmula es:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{(n\Sigma x_i^2 - (\Sigma x_i)^2)(n\Sigma y_i^2 - (\Sigma y_i)^2)}}$$

Reemplazando:

$$r = \frac{731{,}84 - 672}{\sqrt{(8 \cdot 492 - 3600)(8 \cdot 17{,}0176 - 125{,}44)}}$$

$$r = \frac{59{,}84}{\sqrt{(3936 - 3600)(136{,}1408 - 125{,}44)}}$$

$$r = \frac{59{,}84}{\sqrt{(336)(10{,}7008)}}$$

$$r = \frac{59{,}84}{\sqrt{3595{,}4688}}$$

$$r = \frac{59{,}84}{59{,}96}$$

$$r \approx 0{,}998$$

Paso 6: interpretar el resultado

Como $r \approx 0{,}998$ es cercano a 1, se infiere que la relación lineal entre la capacidad vital y la edad de los niños es alta y directa.

PRUEBA TUS CONOCIMIENTOS

A continuación, se presentan ejercicios propuestos para que puedas resolver y practicar, recuerda hacer lo siguiente:

- Resuélvelos siguiendo los pasos utilizados en los problemas resueltos.
- Si es necesario apóyate con los apuntes.
- Si surgen dudas, registrarlas para luego consultar con el tutor.
- ¡Buen trabajo!

1. **Un grupo de profesionales especialistas en salud mental de un hospital psiquiátrico, donde los pacientes permanecen mucho tiempo, deseaba estimar el nivel de respuesta de pacientes retraídos en un programa de terapia de remotivación. Para ello, se contaba con una prueba estandarizada, pero era incosteable y tardaba para administrarla. Para superar este obstáculo, el grupo desarrolló una prueba que era mucho más fácil de aplicar. Para probar la utilidad del nuevo instrumento para medir el nivel de respuesta del paciente, el grupo decidió estudiar la relación entre las calificaciones obtenidas con la nueva prueba y las calificaciones obtenidas con la prueba estandarizada. Para esto se seleccionaron 11 pacientes que habían rendido la nueva prueba para que hicieran la prueba estandarizada, los resultados obtenidos se presentan en la siguiente tabla** (Wayne 1991)**.**

Tabla 2 Calificaciones obtenidas por los pacientes en las pruebas nueva y estandarizada.

Número de paciente	Calificación obtenida en la nueva prueba (X)	Calificación obtenida en la prueba estandarizada (Y)
1	50	61
2	55	61
3	60	59
4	65	71
5	70	80
6	75	76
7	80	90
8	85	106
9	90	98
10	95	100
11	100	114

a) Realice un gráfico de dispersión de los datos obtenidos.
b) Calcula la covarianza e interpretar los resultados.
c) Calcula la correlación de Pearson e interpretar los resultados.

2. **Se llevó a cabo un experimento para estudiar el efecto de cierto medicamento para disminuir la frecuencia cardiaca en adultos. La variable independiente es la dosis en miligramos del medicamento, y la variable dependiente es la diferencia entre la frecuencia cardiaca más baja después de la administración del medicamento y un control antes de administrarlo. Se reunieron los siguientes datos.**

   a) Realice un gráfico de dispersión de los datos obtenidos.
   b) Calcula la covarianza e interpretar los resultados.
   c) Calcula la correlación de Pearson e interpretar los resultados

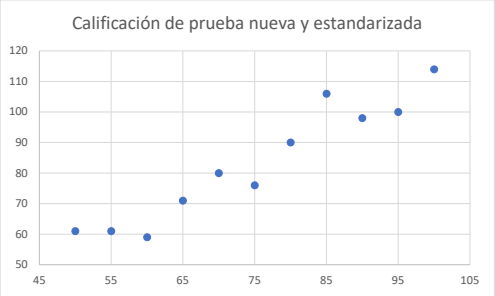
Dosis (mg) X	**Disminución de la frecuencia cardiaca (latidos/min)** Y
0,5	10
0,75	8
1,00	12
1,25	12
1,5	14
1,75	12
2,00	16
2,25	18
2,5	17
2,75	20
3,00	18
3,25	20
3,5	21

(Wayne 1991)

Respuestas

1.

a)

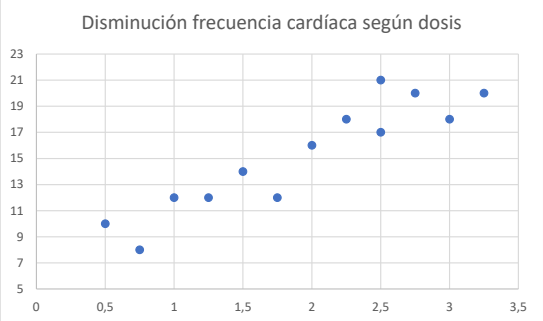


b) 309

c) 0,956

2.

a)



b) 3,394

c) 0,921

SÍNTESIS

Para analizar la relación entre variables existe **métodos gráficos** como el gráfico de dispersión y **métodos cuantitativos** como los siguientes:

Covarianza

- $COV(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$

Correlación de Pearson

- $r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{(n\Sigma x_i^2-(\Sigma x_i)^2)(n\Sigma y_i^2-(\Sigma y_i)^2)}}$

BIBLIOGRAFÍA

Kim, H.-Y. (2018). Statistical notes for clinical researchers: covariance and correlation. *Restorative Dentistry & Endodontics, 43*(1), 1-7. Obtenido de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5816993/pdf/rde-43-e4.pdf

Lind, D., Wathen, S., & Marchal , W. (2012). *Estadística aplicada a los negocios y la economía .* México: The McGraw-Hill Companies, Inc.

Taucher, E. (2014). *Bioestadística .* Ocho Libros Editores Ltda.

Wayne, D. (1991). *Bioestadística base para el análisis de las ciencias y la salud.* México : Limusa S.A. .

¿Quieres recibir orientación para optimizar tu estudio en la universidad?

CONTAMOS CON PROFESIONALES EXPERTOS EN
EL APRENDIZAJE QUE TE PUEDEN ORIENTAR

SOLICITA NUESTRO APOYO

Sitio Web de CIMA

Ver más fichas

Solicita más información

TALCA UNIVERSIDAD CHILE

CIMA UNIDAD DE ACOMPAÑAMIENTO ESTUDIANTIL