

Tecnológico Nacional De México Instituto Tecnológico De Tijuana

Subdirección Académica

Departamento De Sistemas Y Computación

**SEMESTRE**:

Enero – junio 2020

**CARRERA**:

Ing. Tecnologías De Información Y Comunicación

**NOMBRE DEL TRABAJO:**

Correlación

**UNIDAD A EVALUAR:**

Unidad 1

**NOMBRE DEL ALUMNO:**

Garcia Bautista Ana Laura # 15210793

**MAESTRO (A):**

Romero Hernández José Christian

¿Qué es?

Es una técnica de análisis de información con base estadística y, por ende, matemática. Consiste en analizar la relación entre, al menos, dos variables - p.e. dos campos de una base de datos o de un log o raw data-. El resultado debe mostrar la fuerza y el sentido de la relación.

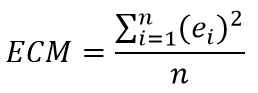
Para analizar la relación entre variables se utilizan los llamados «coeficientes de correlación». Se realizan sobre sobre variables cuantitativas o cualitativas. Ello determinará si se calcula o bien el coeficiente de correlación de [Pearson](https://es.wikipedia.org/wiki/Coeficiente_de_correlaci%C3%B3n_de_Pearson), el de [Spearman](https://es.wikipedia.org/wiki/Coeficiente_de_correlaci%C3%B3n_de_Spearman" \t "_blank), o el de [Kendall](https://en.wikipedia.org/wiki/Kendall_rank_correlation_coefficient). Esto si estamos hablando de correlaciones bivariadas. Existen otras como pueden ser las correlaciones o las medidas de distancia o disimilaridad de intervalos, recuentos o binarias (p.e. [distancia euclídea](https://es.wikipedia.org/wiki/Distancia_euclidiana), euclídea al cuadrado, [Chebyshev](https://en.wikipedia.org/wiki/Chebyshev_distance" \t "_blank), Bloque, [Minkovsky](https://en.wikipedia.org/wiki/Minkowski_distance" \t "_blank), etc.)

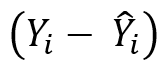
Es una técnica ampliamente documentada, con múltiples fuentes de información abiertas para que cualquiera pueda acceder a sus principios y realizar sus propios análisis.

Hasta el momento la aplicación de la correlación ha sido amplia y diversa en diferentes campos como ciencias naturales, economía, psicología, etc. y por supuesto, en investigaciones de todo tipo. En lo que se refiere al campo de la **seguridad de la información** los fundamentos son los mismos, aunque por el momento aún se está desarrollando.

Fundamentos matemáticos de la correlación:

Analizaremos el caso más típico de dos variables cuantitativas. La relación puede ser de tipo lineal, polinómica, logarítmica, etc. Para ello se utiliza el Método de Mínimos Cuadrados. Consiste en minimizar el error de la relación calculada frente al hecho real, es decir, minimizar la desviación típica de los residuos de la regresión (Error Cuadrático Medio).



Siendo: ei la diferencia para cada elemento entre la observación real y el dato estimado , y n el número de elementos observados.

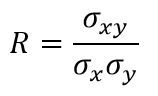
De ahí, aplicando un ajuste lineal,

Ecuación Lineal Multidimensional

En el caso más simple: con una variable dependiente y otra independiente tendríamos

Ecuación Lineal Bidimensional

De esta surge el Coeficiente de Correlación de Pearson, el más conocido de los coeficientes de correlación:



Siendo: σxy la covarianza de las variables X e Y,σx la desviación típica de la variable X,

y σy la desviación típica de la variable Y.

Los valores que puede tomar R son: -1≤R≤1

* R=1: existe una relación positiva perfecta.

También se pueden valer de los siguientes rangos:

* R=-1: existe una relación negativa perfecta.
* -1<R<-0,5: existe relación negativa fuerte.
* R=-0,5: existe una relación negativa moderada.
* -0,5<R<0: existe relación negativa débil.
* R=0: no existe relación, no existe relación lineal, Y no depende linealmente de X.
* 0<R<0,5: existe relación positiva débil.
* R=0,5: existe una relación positiva moderada.
* 0,5<R<1: existe relación positiva fuerte.
* R=1: existe una relación positiva perfecta.

Para analizar la relación entre variables existen múltiples métodos, ejemplos de ellos son:

* [Regresión Lineal](https://en.wikipedia.org/wiki/Linear_regression).
* [Regresión No Lineal](https://en.wikipedia.org/wiki/Nonlinear_regression).
* [Regresión Logística](https://en.wikipedia.org/wiki/Logistic_regression).
* [Modelos Probit](https://en.wikipedia.org/wiki/Probit_model).
* [Tablas de Contingencia](https://es.wikipedia.org/wiki/Tabla_de_contingencia).
* Modelos Logarítmicos.

Ejemplos de fuentes de Información con explicaciones sobre la correlación:

* [MIT]: [Correlation and Regression](http://web.mit.edu/10.001/Web/Course_Notes/Statistics_Notes/Correlation/lec23.html" \t "_blank)
* [MIT]: [Correlation and Regression (2)](http://www.ocw.nur.ac.rw/NR/rdonlyres/Brain-and-Cognitive-Sciences/9-07Spring-2004/FCA56A18-7E3A-44E5-BF5B-8154F37E1D7E/0/16_corr_regres3.pdf" \t "_blank)
* [MIT]: [Regression Analysis: Method of Least Squares](http://web.mit.edu/10.001/Web/Course_Notes/Statistics_Notes/Correlation/node3.html)
* [MIT]: [Correlation Analysis](http://web.mit.edu/10.001/Web/Course_Notes/Statistics_Notes/Correlation/node2.html" \t "_blank)
* [MIT]: [Tests for the Regression Equation](http://web.mit.edu/10.001/Web/Course_Notes/Statistics_Notes/Correlation/node4.html)
* [MIT]: [Probabilistic Systems Analysis and Applied Probability, Derived distributions, convolution; covariance and correlation](http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-041sc-probabilistic-systems-analysis-and-applied-probability-fall-2013/unit-ii/lecture-11/MIT6_041SCF13_L11.pdf)
* [RAE]: [Correlación](http://lema.rae.es/drae/?val=correlaci%C3%B3n)
* [UOC]: [Regresión Lineal](http://www.academia.edu/13222607/CORRELACI%C3%93N_LINEAL_Y_AN%C3%81LISIS_DE_REGRESI%C3%93N_MAPA_CONCEPTUAL)
* [WIKIPEDIA]: [Correlation and dependence](https://en.wikipedia.org/wiki/Correlation_and_dependence" \t "_blank)

¿Qué se necesita para realizar una correlación?, ¿cuáles son los requisitos mínimos?:

Conocer los datos y la tipología de las variables: bien de tipo cuantitativo o cualitativo. Dado que, en función del tipo de las variables del conjunto de datos, se aplicarán una serie de coeficientes u otros.

Conocer qué método es el más adecuado para su aplicación: de nada vale conocer el script que ejecuta una correlación, si no se puede basar matemáticamente cómo desarrollar los cálculos.

Conocer, comprender e interpretar los resultados que los indicadores ofrecen: parte fundamental del análisis, basada en la diferencia existente entre los conceptos de causalidad y correlación que subyace en el análisis, y que suelen confundirse. El mero cálculo de un coeficiente, no determinado por una causalidad, no tiene porqué ser válido. Es decir «la existencia de correlación, no implica causalidad», como se refleja en [esta viñeta](https://xkcd.com/552/):

1. Lo primero, un **«conjunto de datos»**.
2. En dicho conjunto el requisito es tener mínimo dos «variables» (éste es el nombre técnico), según unas técnicas u otras se han denominado «dominios», «campos» de una base de datos y similares.
3. Conocer las bases matemáticas y estadísticas del análisis de datos, en concreto:

