**Correlación lineal con Python**

1. **Introducción**

Método estadístico que permite **cuantificar la relación lineal existente entre dos variables**.

Existen varios **estadísticos**, llamados **coeficientes de correlación lineal**, desarrollados con el objetivo de **medir este tipo de asociación**, algunos de los más empleados son

* Pearson
* Spearman
* Kendall

Con frecuencia, **los estudios de correlación lineal preceden a análisis más complejos**, como la creación de [modelos de regresión](https://www.cienciadedatos.net/documentos/py10-regresion-lineal-python.html). Primero, se analiza si las variables están correlacionadas y, en caso de estarlo, se procede a generar modelos.

Es importante destacar que, **la existencia de correlación entre dos variables, no implica necesariamente causalidad**. La asociación observada puede deberse a un tercer factor ([confounder](https://en.wikipedia.org/wiki/Confounding)).

**Correlación lineal con Python**

Existen múltiples implementaciones que permiten calcular correlaciones lineales en Python, tres de las más utilizadas están disponibles en las librerías:

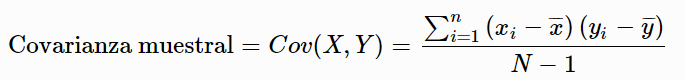
* **SciPy**,
* **Pandas** y
* **Pingouin**.

**2. Coeficientes de correlación lineal**

**2.2 Covarianza**

Para **estudiar la relación lineal** existente **entre dos variables continuas** es necesario disponer de parámetros que permitan cuantificar dicha relación.

Uno de estos parámetros es la **covarianza**, que **mide el grado de variación conjunta de dos variables aleatorias**.



donde x¯¯ e y¯¯ son la media de cada variable, y xi e yi son el valor de las variables para la observación i.

* Valores positivos 🡪 dos variables cambian en la misma dirección y,
* valores negativos 🡪 lo hacen en direcciones opuestas.

**La principal limitación:** de la covarianza es que, su magnitud, depende de las escalas en que se miden las variables estudiadas.

Esto implica que no puede utilizarse para comparar el grado de asociación entre pares de variables medidas en distintas escalas.

Una forma de **evitar esta limitación y poder hacer comparaciones** consiste en **estandarizar la covarianza**, generando lo que se conoce como **coeficientes de correlación**.

**2.3 Coeficientes de correlación lineal**

Son estadísticos que cuantifican la asociación lineal entre dos variables numéricas. Existen diferentes tipos, de entre los que destacan el Pearson, Rho de Spearman y Tau de Kendall. Todos ellos comparten que:

* valor comprendido en [+1 , -1].
  + +1 🡪 correlación positiva perfecta
  + -1 🡪 correlación negativa perfecta.
* Se emplean como **medida de la fuerza de asociación entre dos variables (tamaño del efecto**):
  + 0: asociación nula.
  + 0.1: asociación pequeña.
  + 0.3: asociación mediana.
  + 0.5: asociación moderada.
  + 0.7: asociación alta.
  + 0.9: asociación muy alta.

Las **principales diferencias entre estos tres coeficientes** son:

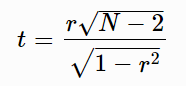
* Pearson:
  + variables cuantitativas que tienen una distribución normal o próxima a la normal.
  + Es más sensible a los valores extremos que las otras dos alternativas.
* Spearman:
  + variables cuantitativas (continuas o discretas).
  + En lugar de utilizar directamente el valor de cada variable, los datos son ordenados y reemplazados por su respectivo orden [ranking](https://es.wikipedia.org/wiki/Ranking).
  + método no paramétrico muy utilizado cuando no se satisface la condición de normalidad necesaria para aplicar la correlación de Pearson.
* Kendall:
  + alternativa no paramétrica que, al igual que la correlación de Spearman, utiliza la ordenación de las observaciones [ranking](https://es.wikipedia.org/wiki/Ranking).
  + recomendable cuando se dispone de pocos datos y muchos de ellos ocupan la misma posición en el rango, es decir, cuando hay muchas ligaduras.

**2.4 Significancia estadística**

Además del valor obtenido para el coeficiente de correlación, es necesario calcular su significancia estadística.

Por muy cercano que sea el valor del coeficiente de correlación a +1 o −1, si no **es significativo**, no se dispone de evidencias suficiente para afirmar que existe una correlación real, ya que el valor observado podría deberse a simple aleatoriedad.

El **test paramétrico de significancia estadística** empleado **para el coeficiente de correlación** es el **t-test**. Donde el estadístico t se obtiene acorde a la ecuación:



* r es el calor del coeficiente de correlación y
* N es el número de observaciones disponibles.

Los grados de libertad se calculan como df = N−2.

# ¿Qué es una prueba t? ¿Y por qué es como decirle a un niño que limpie ese desorden en la cocina?

# <https://blog.minitab.com/en/statistics-and-quality-data-analysis/what-is-a-t-test-and-why-is-it-like-telling-a-kid-to-clean-up-that-mess-in-the-kitchen>

**¿Qué son los Valores T y los Valores P en Estadística?** <https://blog.minitab.com/es/que-son-los-valores-t-y-los-valores-p-en-estadistica>

*(concepto de* ***grado de libertad****:* [*https://blog.minitab.com/es/que-son-los-grados-de-libertad-en-estadistica*](https://blog.minitab.com/es/que-son-los-grados-de-libertad-en-estadistica)*)*

*RCORDAR EJEMPLO: 7 días, 7 sombreros, 7-1=6 días de “libertad de elegir sombrero”*

*Los grados de libertad se definen frecuentemente como el número de observaciones (piezas de información) en los datos que pueden variar libremente al estimar parámetros estadísticos.*

*el número de grados de libertad es igual al número de “observaciones” menos el número de relaciones requeridas entre las observaciones*

En este test, se considera como:

* hipótesis nula (H0) que las variables son independientes (coeficiente de correlación poblacional = 0)
* hipótesis alternativa (Ha), que sí existe relación (coeficiente de correlación poblacional ≠ 0)

También se puede calcular la significancia de un coeficiente de correlación recurriendo a métodos no paramétricos como el [**bootstrapping**](https://www.cienciadedatos.net/documentos/pystats04-bootstrapping-python.html)**.**

**2.5 Tamaño de efecto**

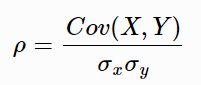
La correlación lineal, además del valor del coeficiente de correlación y de sus significancia, también tiene un tamaño de efecto asociado conocido como **coeficiente de determinación R2**:

se interpreta como la cantidad de varianza de Y explicada por X.

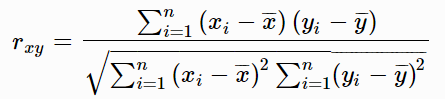
* En el caso del coeficiente de Pearson y el de Spearman, R2 se obtiene elevando al cuadrado el coeficiente de correlación.
* En el caso de Kendall no se puede calcular de este modo.

**Coef de Pearson**

es la covarianza estandarizada



La anterior ecuación se corresponde con el coeficiente de Pearson poblacional (ρ). En la práctica, raramente se tiene acceso a toda la población, por lo que su valor se estima a partir de una muestra mediante el **coeficiente de Pearson muestral (r)**



Las condiciones que se deben de cumplir para que el coeficiente de correlación de Pearson sea válido son:

* La relación que se quiere estudiar es de tipo lineal (de lo contrario, el coeficiente de Pearson no la puede detectar).
* Las dos variables deben de ser numéricas.
* Normalidad: ambas variables se tienen que distribuir de forma normal. En la práctica, se suele considerar válido aun cuando se alejan moderadamente de la normalidad.
* Homocedasticidad: la varianza de Y debe ser constante a lo largo de la variable X.

Esto se puede contrastar si en un scatterplot los valores de Y mantienen la misma dispersión en las distintas zonas de la variable X.

**Características**

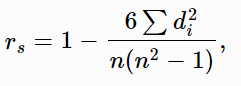
* [-1, +1], siendo +1 correlación lineal positiva perfecta y -1 correlación lineal negativa perfecta.
* Es independiente de las escalas en las que se midan las variables.
* No varía si se aplican transformaciones a las variables.
* No tiene en consideración que las variables sean dependientes o independientes.
* El coeficiente de correlación de Pearson no equivale a la pendiente de la recta de regresión.
* **Es sensible a outliers**, por lo que se recomienda, en caso de poder justificarlos, excluirlos antes de realizar el cálculo.

**Coef de Spearman**

Es el equivalente al coeficiente de Pearson pero con una **previa transformación de los datos a**[**ranking**](https://es.wikipedia.org/wiki/Ranking)🡪 en lugar de utilizar directamente el valor de cada variable, los datos son ordenados y reemplazados por su respectivo orden (primero, segundo, tercero...).

**Alternativa no paramétrica al coeficiente de Pearson** **cuando**:

* los valores son ordinales, o bien,
* cuando los valores son continuos pero no satisfacen la condición de normalidad



Siendo di la distancia entre los rangos de cada observación (xi−yi)

*n* el número de observaciones.

Condiciones: que la relación entre las variables sea monótona, es decir, que cuando una variable crece la otra también lo hace o cuando una crece la otra decrece (que la tendencia sea constante).

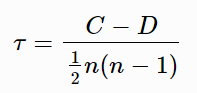
Al trabajar con el orden de las observaciones en lugar de su valor real, tiene la característica de ser menos sensible que *Pearson* a valores extremos.

**Coef de Tau de Kendall**

La correlación de *Kendall* es un método no paramétrico que, al igual que la correlación de Spearman, **utiliza la ordenación de las observaciones**[**ranking**](https://es.wikipedia.org/wiki/Ranking).

Es **otra alternativa al coeficiente de correlación de Pearson cuando no se cumple la condición de normalidad**.

**Suele utilizarse en lugar del coeficiente de Spearman cuando el número de observaciones es pequeño o los valores se acumulan en una región** por lo que el número de ligaduras al generar el ranking es alto.



C el número de pares concordantes, aquellos en los que el rango de la segunda variable es mayor que el rango de la primera variable.

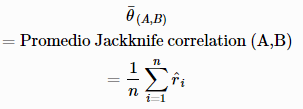
D el número de pares discordantes, cuando el rango de la segunda es igual o menor que el rango de la primera variable.

**Jackknife correlation**

El **coeficiente de correlación de Pearson** resulta efectivo en ámbitos muy diversos. Sin embargo, tiene la desventaja de no ser robusto frente a outliers (a pesar de que se cumpla la condición de normalidad).

* Si dos variables tienen un pico o un valle común en una única observación, por ejemplo por un error de lectura, la correlación va a estar dominada por este registro a pesar de que entre las dos variables no haya correlación real alguna.
* Lo mismo puede ocurrir en la dirección opuesta. Si dos variables están altamente correlacionadas excepto para una observación en la que los valores son muy dispares, entonces la correlación existente quedará enmascarada.

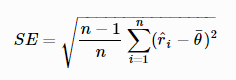
**Una forma de evitarlo es recurrir a la Jackknife correlation**, que consiste en calcular todos los posibles coeficientes de correlación entre dos variables si se excluye cada vez una de las observaciones. El promedio de todas las Jackknife correlations calculadas atenúa en cierta medida el efecto del outlier.



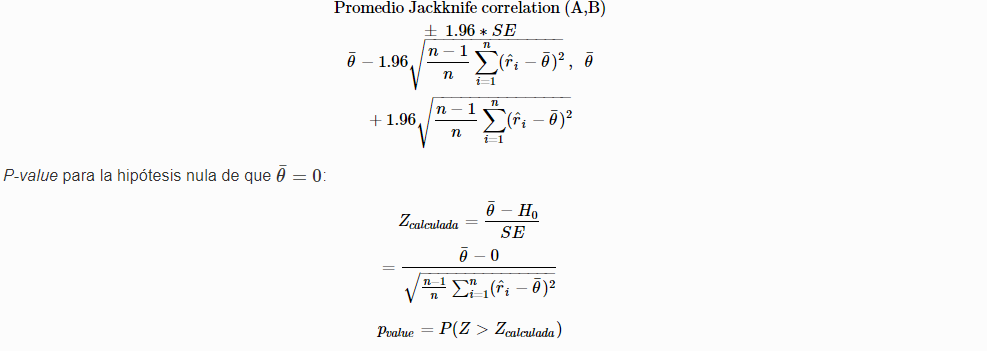
n es el número de observaciones

r^i es el coeficiente de correlación de *Pearson* estimado entre las variables A y B, habiendo excluido la observación i.

Además del promedio, se puede estimar su error estándar (SE) y así obtener intervalos de confianza para la *Jackknife correlation* y su correspondiente *p-value*.



Intervalo de confianza del 95% (Z=1.96)



Cuando se emplea este método es conveniente calcular la diferencia entre el valor de correlación obtenido por Jackknife correlation (θ¯) y el que se obtiene si se emplean todas las observaciones (r¯).

A esta diferencia se le conoce como **Bias**. Su magnitud es un indicativo de cuanto está influenciada la estimación de la correlación entre dos variables debido a un valor atípico u outlier.



Si se calcula la diferencia entre cada correlación (r^i) estimada en el proceso de *Jackknife* y el valor de correlación (r^) obtenido si se emplean todas las observaciones, se puede identificar que observaciones son más influyentes.

Cuando el estudio requiere minimizar al máximo la presencia de falsos positivos, a pesar de que se incremente la de falsos negativos, se puede seleccionar como valor de correlación el menor de entre todos los calculados en el proceso de *Jackknife*.



RESUMEN:

A pesar de que el método de Jackknife permite aumentar la robustez de la correlación de Pearson, si los outliers son muy extremos su influencia seguirá siendo notable. Siempre es conveniente una representación gráfica de los datos para poder identificar si hay valores atípicos y eliminarlos. Otras alternativas robustas son la correlación de Spearman o el método de Bootstrapping.

**Correlación parcial**

Como se ha explicado, la correlación estudia la relación (lineal o monotónica) existente entre dos variables.

**Puede ocurrir que, la relación que muestran dos variables, se deba a una tercera variable que influye sobre las otras dos**. A este fenómeno se le conoce como **confounding.**

* Por ejemplo: si se correlaciona el tamaño del pie de una persona con su inteligencia, se encuentra una correlación positiva alta. Sin embargo, dicha relación se debe a una tercera variable que está relacionada con las otras dos, la edad.

**La correlación parcial permite estudiar la relación lineal entre dos variables bloqueando el efecto de una tercera (o más) variables**. Si el valor de correlación de dos variables es distinto al valor de correlación parcial de esas mismas dos variables, cuando se controla una tercera, significa que la tercera variable influye en las otras dos. La función en partial\_corr() del paquete pingouin permite estudiar correlaciones parciales.