Banjercito

Diplomado en Ciencia de Datos

Notas del Curso

Módulo III — Semana 2

Instructor: Alan Badillo Salas

Agosto 2024

1 Introducción

En la semana 1 hemos visto los tipos de análisis que se pueden realizar, tanto cuantitativos como cualitativos y las medidas de tendencia central como la media, la mediana, la moda, los cuartiles y los percentiles.

Los análisis cuantitativos se orientan a analizar datos numéricos, mientras los cualitativos a establecer relaciones profundas entre los datos. Las medidas de tendencia central permiten describir a los datos numéricos para poder entender su comportamiento. Por ejemplo, la media nos da un valor ideal de lo que debería valer un grupo de datos, la mediana nos indica que si es menor al promedio hay más valores pequeños que grandes, representando un sesgo a la izquierda y sino que hay más valores grandes que pequeños y un sesgo a la derecha. Mientras que, los cuartiles y percentiles nos permiten entender cómo va evolucionando la población a un porcentaje dado, para poder estimar intervalos de confianza sobre dónde se encuentran los diferentes promedios de una población.

Esta semana estudiaremos la varianza, desviación estándar, la covarianza y la correlación entre los datos. Esto nos permitirá entender la naturaleza de cómo los datos se apegan o alejan del valor central de la media.

1.1 Contenido de la Semana 2

Esta semana revisaremos los siguientes temas:

1. Ejes de análisis
   (a) Variables independientes
   (b) Media muestral y poblacional
2. Correlación de datos
   (a) Varianza
   (b) Desviación Estándar
   (c) Covarianza
   (d) Correlación

2 Ejes de análisis

Cuando analizamos datos generalmente lo hacemos por muestras, cada muestra es como una fila de diferentes datos combinados, por ejemplo, la edad, género, peso, estatura y salario de una persona. Esto lo podemos ver como una tupla o grupo de datos:

$$persona = (edad, genero, peso, estatura, salario) \tag{1}$$

Cuando tenemos muchas muestras podemos numerar cada muestra obteniendo un conjunto de muestras, por ejemplo, para un estudio sobre $1,000$ personas tendríamos:

$$\begin{aligned}
persona_1 &= (edad_1, genero_1, peso_1, estatura_1, salario_1) \\
persona_2 &= (edad_2, genero_2, peso_2, estatura_2, salario_2) \\
persona_3 &= (edad_3, genero_3, peso_3, estatura_3, salario_3) \\
&... \\
persona_{500} &= (edad_{500}, genero_{500}, peso_{500}, estatura_{500}, salario_{500}) \\
persona_{501} &= (edad_{501}, genero_{501}, peso_{501}, estatura_{501}, salario_{501}) \\
&... \\
persona_{998} &= (edad_{998}, genero_{998}, peso_{998}, estatura_{998}, salario_{998}) \\
persona_{999} &= (edad_{999}, genero_{999}, peso_{999}, estatura_{999}, salario_{999}) \\
persona_{1000} &= (edad_{1000}, genero_{1000}, peso_{1000}, estatura_{1000}, salario_{1000})
\end{aligned} \tag{2}$$

Así cada uno de los $1,000$ registros tendrá asociado una edad, un género, un peso, una estatura y un salario. Por ejemplo, para la $persona_{500}$ y $persona_{700}$ podríamos tener:

$$\begin{aligned}
persona_{500} &= (23, hombre, 72.5, 1.85, 23400) \\
&... \\
persona_{700} &= (26, mujer, 53.4, 1.55, 28200)
\end{aligned} \tag{3}$$

Entonces, cada valor dispuesto en la tupla o grupo de valores de la muestra pertenecerá a un eje de análisis, el cuál determinará una cualidad o característica de la muestra, teniendo así para este ejemplo cinco diferentes ejes de análisis: *Edad*, *Género*, *Peso*, *Estatura* y *Salario*. Estos ejes determinarán cómo están constituidas las personas, es decir, cómo se comportan las muestras respecto a cada valor que determina al conjunto total de muestras.

2.1 Variables independientes

Cada eje de análisis lo podemos pensar como una variable independiente que toma valores sobre todas las muestras. Esto significa, que podemos ver una muestra en términos de variables, por ejemplo:

$$\begin{aligned} p_1 &= (e_1, g_1, s_1, t_1, r_1) \\ p_2 &= (e_2, g_2, s_2, t_2, r_2) \\ p_3 &= (e_3, g_3, s_3, t_3, r_3) \\ &... \\ p_{500} &= (e_{500}, g_{500}, s_{500}, t_{500}, r_{500}) \\ p_{501} &= (e_{501}, g_{501}, s_{501}, t_{501}, r_{501}) \\ &... \\ p_{998} &= (e_{998}, g_{998}, s_{998}, t_{998}, r_{998}) \\ p_{999} &= (e_{999}, g_{999}, s_{999}, t_{999}, r_{999}) \\ p_{1000} &= (e_{1000}, g_{1000}, s_{1000}, t_{1000}, r_{1000}) \end{aligned} \tag{4}$$

Donde cada muestra p_i está representada por las variables $(e_i, g_i, s_i, t_i, r_i)$, observando que el peso fue representada por la variable s, aunque pudimos elegir cualquier otra o seguir usando $peso_i$. Lo mismo para estatura que se presentó por t y el salario por r.

Para hablar de la *i-ésima* muestra, es decir, la muestra de índice i, usamos el subíndice i, por ejemplo, la *i-ésima* persona es p_i y la edad de la *i-ésima* persona es e_i. Así, $e_1, e_2, e_3, ..., e_{500}$ representan las edades de las primeras 500 personas, mientras que $s_1, s_2, s_3, ..., s_{100}$ serán los pesos de las primeras 100 personas.

Cada eje de análisis se puede entender en términos de las variables independientes como la tupla o grupo de todas las variables asociadas a cada muestra, es decir:

- **Edad** — $\vec{E} = (e_1, e_2, e_3, ..., e_{500}, e_{501}, ..., e_{998}, e_{999}, e_{1000})^T$
- **Género** — $\vec{G} = (g_1, g_2, g_3, ..., g_{500}, g_{501}, ..., g_{998}, g_{999}, g_{1000})^T$
- **Peso** — $\vec{S} = (s_1, s_2, s_3, ..., s_{500}, s_{501}, ..., s_{998}, s_{999}, s_{1000})^T$

- **Estatura** — $\vec{T} = (t_1, t_2, t_3, ..., t_{500}, t_{501}, ..., t_{998}, t_{999}, t_{1000})^T$
- **Salario** — $\vec{R} = (r_1, r_2, r_3, ..., r_{500}, r_{501}, ..., r_{998}, r_{999}, r_{1000})^T$

Ahora cada eje de análisis se puede entender como una variable aleatoria que toma los valores aleatorios para cada muestra. Se dice que toma los valores aleatorios porque en realidad no sabemos qué valores tomará de una muestra a otra, por ejemplo, para las edades podríamos tener:

$$\vec{E} = (18, 34, 55, ..., 23, 17, ..., 64, 19, 21)^T \tag{5}$$

Representando así una variable aleatoria con edades que no muestran un comportamiento claro entre la primera muestra, la segunda, tercera, etc. Incluso si hubieramos numerado en otro orden las muestrás los valores no mostrarían algún indicio de tomar algún valor. A veces nos referiremos al eje X o al vector de valores del eje como $\vec{X}$ y las componentes o valores como x_i entendiendo que es el *i-ésimo* valor de la muestra i para el eje X. Por ejemplo, e_{200} representará el valor para el eje E de la edad para la muestra 200.

2.2 Media muestral y poblacional

La media o valor promedio para un eje de análisis se puede calcular en términos muestrales cuando el número de muestras es pequeño (menor a 100 muestras) o en términos poblacionales, cuando el número de muestras es suficientemente grande (más de 100 muestras).

La media muestral se calcula tomando $N-1$ muestras, donde N es el número de muestras y la media poblacional se calcula tomando N muestras. Así tenemos:

$$\overline{x_{muestral}} = \frac{1}{N-1} \sum_{i=1}^{N} x_i \tag{6}$$

Que será un valor más aproximando a la media poblacional cuándo hay pocas muestras. Y también tenemos la media poblacional que es más usual:

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{7}$$

Entonces, si hay menos de 100 muestras en nuestro análisis usaremos la media muestral en lugar de la poblacional, en general asumiremos en la ecuaciones $N-1$ en lugar de N.

Por ejemplo, para las edades tendríamos que la media es:

$$\begin{aligned} \overline{e} &= \frac{1}{1000} \sum_{i=1}^{1000} e_i \\ &= \frac{e_1 + e_2 + e_3 + ... + e_{500} + e_{501} + ... + e_{998} + e_{999} + e_{1000}}{1000} \\ &= \frac{18 + 34 + 55 + ... + 23 + 17 + ... + 64 + 19 + 21}{1000} \end{aligned} \tag{8}$$

Así $\overline{e}$ representa el valor medio o valor promedio de las edades. Es decir, el valor que se encuentra en medio de todos los valores y por lo tanto caracteriza a todas las edades como un único valor. Por ejemplo, si $\overline{e} = 24.5$ representaría que la población en promedio tiene 24.5 años, es decir, que la población en general tiene entre 24 y 25 años, aunque esto podría no ser cierto. Por ejemplo, si una persona tiene 1 año y otra 99 años, la edad promedio será de $\frac{1+99}{2} = 50$, y nadie tiene 50 años, pero en promedio sí. Por lo que debemos determinar que tan dispersos están los datos respecto a este valor medio para que tenga más sentido.

3 Correlación de datos

Hemos visto que un eje de datos X se puede entender como una variable aleatoria $\vec{X}$ y que de dicha variable aleatoria o eje se puede calcular $\overline{x}$ que es el valor medio o promedio.

También observamos que el valor medio $\overline{x}$ podría tener más sentido si supieramos que tanto se alejan los datos hacia este valor. Por ejemplo, si todos los datos fueran iguales o cercanos al valor promedio, podríamos afirmar que todas las personas tienen la misma edad o el mismo peso, según el eje que se esté analizando. Pero si los datos fueran lejanos al promedio como la edad de 1 año y la edad de 99 años que se alejan bastante del promedio de 50 años, entonces no podríamos hacer conclusiones tan fuertes sobre la población en general.

Veamos un ejemplo más realista, supongamos que en la calle entrevistamos a $1,000$ personas y les preguntamos su edad, algunos contestarán que es de 10, 20, 50 o incluso 90 años, teniendo edades de todos los rangos, pero si entrevistamos a $1,000$ personas en una universidad es probable que muchos tengan edades más cercanas a 18, 19, 20, 21, 22 y 23 años, ya que la mayoría de la población serán alumnos cursando sus estudios de licenciatura, aunque alguno que otro contestará que tiene 50 u 80 años, por ejemplo, algún profesor.

En el ejemplo anterior podemos observar que si los datos se acercan al promedio, por ejemplo, a los 20 años para el caso de los alumnos, será porque hay una verdad oculta o un comportamiento oculto entre las muestras de datos.

Otro ejemplo es el de la producción de tornillos que deberían medir $1cm$. Si tomaramos una muestra de $10,000$ tornillos fabricados, tendríamos que muchos miden cerca de $1cm$, aunque alguno que otro podría medir $1.1cm$ o $0.9cm$ y en algún caso muy extremo incluso $1.5cm$. Quizás diez de diez mil tornillos estén mal fabricados, representando el 0.1% de la población. ¿Qué tan significativo es esto?

3.1 Varianza

La varianza es una medida de proximidad entre los datos al valor medio o promedio. Esta se cacula sumando el área del dato hacia la media. Es decir, que si calculamos el área de un dato hacia el valor promedio tendríamos:

$$A_i = (x_i - \overline{x})^2 \tag{9}$$

Esto significa que para la muestra i, la *i-ésima* área entre el valor x_i y el valor promedio $\overline{x}$ es la distancia del dato al valor medio elevado al cuadrado, generando así el valor A_i que representa el área de alejamiento entre el dato y el promedio.

Algunos valores estarán más alejados del promedio que otros, los que estén suficientemente cerca tendrá un área cercana a cero y los datos alejados tendrán un área muy grande.

Si calculamos el valor promedio de todas estas áreas, obtendrémos un valor conocido como **la varianza** de los datos, y representará el área promedio de que tanto se alejan los datos a su valor central o valor medio.

Observemos que si la varianza es cercana a cero, todos los valores estarán cercanos al promedio y si es muy grande entonces los datos estarán muy alejados del valor promedio.

La forma de calcular la varianza es $VAR(X) = \frac{1}{N}\sum_{i=1}^{N} A_i$, pero en términos directos y sabiendo que el área de la *i-ésima* muestra es $A_i = (x_i - \overline{x})^2$, tenemos:

$$VAR(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{10}$$

Y este valor será muy útil para entender cómo se desvian los datos respecto al valor medio. La varianza es difícil de interpretar ya que solo sabemos que si es cercana a cero es porque los datos están cercanos al promedio y que si es un valor muy grande es porque los datos se alejan mucho del promedio. Pero, ¿Qué tan cercano o lejano está un dato en realidad?

3.2 Desviación Estándar

Para saber qué tanto se acerca o aleja un dato al valor medio $\overline{x}$, podemos calcular la raíz cuadrada de la varianza. Este valor representará no un área, sino un segmento o lado del cuadra que representará un valor lineal de qué tan alejado está un dato.

Es decir, si calculamos la raíz cuadrada de la varianza, obtendremos una valor llamado la desviación estándar y con esto sabremos qué tan alejado está un dato linealmente hacia su valor medio o promedio. Con esto podremos saber si un

dato se aleja tantos años de la edad promedio o tantos centímetros de la altura promedio de un tornillo, según el eje de análisis.

La forma de calcular la desviación estándar es:

$$DESV.EST(X) = \sqrt{VAR(X)} \tag{11}$$

Representando la desviación de los datos hacia el promedio. Generalmente usamos el símbolo σ para repsentar la desviación estándar y μ para representar la media o valor promedio.

Entonces, la pareja (μ, σ) describe el promedio y la varianza para un eje de análisis. Por ejemplo, si analizamos las edades podríamos tener $(\mu = 24.5, \sigma = 1.5)$ que significará que la edad promedio es de 24.5 años y la desviación estándar es de 1.5 años. Esto significa que la población promedio está cerca de los 24.5 años y podríamos encontrar al 33% de la población 1.5 años arriba del promedio (26 años) y otro 33% de la población abajo del promedio (23 años).

¿Cómo sabemos que el 33% de la población arriba del promedio se encuentra a una desviación estándar del promedio y que otro 33% está abajo del promedio a una desviación estándar el promedio?

A esto se le conoce como intervalo de confianza, y si la población se comporta de forma normal (bajo una distribución normal que se verá más adelante), entonces, podemos estimar el porcentaje de la población que estará desviado de forma estándar respecto al promedio, como se muestra en la siguiente tabla:

Población	**Intervalo**
66%	$(\mu - \sigma, \mu + \sigma)$
95%	$(\mu - 2 \cdot \sigma, \mu + 2 \cdot \sigma)$
99.5%	$(\mu - 3 \cdot \sigma, \mu + 3 \cdot \sigma)$

Table 1: Pocentajes de la población según su desviación estándar (intervalo)

Inferior	**Superior**
$\mu - \sigma \rightarrow 33\%$	$\mu + \sigma \rightarrow 33\%$
$\mu - 2 \cdot \sigma \rightarrow 47.5\%$	$\mu + 2 \cdot \sigma \rightarrow 47.5\%$
$\mu - 3 \cdot \sigma \rightarrow 49.75\%$	$\mu + 3 \cdot \sigma \rightarrow 49.75\%$

Table 2: Pocentajes de la población según su desviación estándar (inferior y superior)

Con esto podemos determinar el intervalo donde esperamos que la población se encuentre. ¿Tendrá alguna relación con los percentiles?

3.3 Covarianza

Cuando analizamos un eje, podemos pensarlo en términos de su variable independiente, de su media y su varianza, es decir, en la terna (X, μ_x, σ_x).

Si analizamos otro eje por ejemplo Y podemos pensar en su terna (Y, μ_y, σ_y).

Una forma de analizar dos ejes es intentar entender una relación entre los datos de ambos ejes, respecto a su media. Si por ejemplo, para una muestra i, el dato de un eje X está arriba del promedio, entonces $x_i - \mu_x$ será positivo, si por el contrario se encuentra debajo del promedio entonces $x_i - \mu_x$ será negativo.

De forma similar podemos pensar en el dato del un eje Y para la misma muestra i, que podría encontrase arriba de su promedio teniendo $y_i - \mu_y$ positiva o debajo de su promedio $y_i - \mu_y$ negativa.

Entonces, si multiplicamos ambos valores $(x_i - \mu_x) \cdot (y_i - \mu_y)$, tendremos que si ambos están al mismo tiempo arriba de su promedio respectivo del eje, entonces el producto será positivo $\{(+) \cdot (+) \to (+)\}$, o si ambos están debajo de su promedio respectivo del eje, entonces el producto también será positivo $\{(-) \cdot (-) \to (+)\}$, esto significa que ambos datos de la muestra sobre cada uno de sus ejes está arriba o abajo del promedio al mismo tiempo. Si sumaramos todas los productos, sobre todas las muestras, entonces obtendríamos un valor positivo que indicaría que los datos de ambos ejes están al mismo tiempo debajo o arriba de su promedio.

Veamos un caso práctico, supongamos que tenemos dos ejes de análisis S el eje del peso y T el eje de la estatura para una persona, y tenemos 10 muestras:

	Peso (S)	**Estatura** (T)
1	56.7	1.58
2	63.2	1.65
3	58.1	1.61
4	76.4	1.78
5	72.3	1.72
6	85.9	1.74
7	56.9	1.60
8	48.8	1.54
9	54.6	1.59
10	71.3	1.67

Table 3: 10 muestras en los ejes S (Peso) y T (Estatura)

Ahora calculemos el valor promedio para el eje S, es decir, la media μ_s:

$$\begin{aligned}\mu_s &= \frac{1}{N}\sum_{i=1}^{N} s_i \\ &= \frac{56.7 + 63.2 + 58.1 + ... + 54.6 + 71.3}{10} \\ &= 64.42\end{aligned} \tag{12}$$

De forma similar podemos calcular el valor promedio para el eje T, es decir, la media μ_t:

$$\begin{aligned}\mu_s &= \frac{1}{N}\sum_{i=1}^{N} s_i \\ &= \frac{1.58 + 1.65 + 1.61 + ... + 1.59 + 1.67}{10} \\ &= 1.648\end{aligned} \tag{13}$$

Con estas medias podemos construir las diferencias entre cada dato menos su promedio respectivo, es decir, $(s_i - \mu_s)$ y $(t_i - \mu_t)$:

	$(s_i - \mu_s)$	$(t_i - \mu_t)$
1	-7.72	-0.07
2	-1.22	0.00
3	-6.32	-0.04
4	11.98	0.13
5	7.87	0.07
6	21.48	0.09
7	-7.52	-0.05
8	-15.62	-0.11
9	-9.82	-0.06
10	6.88	0.02

Table 4: Diferencia entre cada dato y su media para los ejes S (Peso) y T (Estatura)

Si observamos los signos, podemos ver que la mayoría de los datos coincide en signos:

	$(s_i - \mu_s)$	$(t_i - \mu_t)$
1	(-)	(-)
2	(-)	$(\cdot)$
3	(-)	(-)
4	(+)	(+)
5	(+)	(+)
6	(+)	(+)
7	(-)	(-)
8	(-)	(-)
9	(-)	(-)
10	(+)	(+)

Table 5: Signos de la diferencia entre cada dato y su media para los ejes S (Peso) y T (Estatura)

Esto significa que si sumamos cada multiplicación, esperaremos que este valor sea positivo.

Pero, ¿Qué pasará cuando los signos sean contrarios, es decir, que $(x_i - \mu_y) \cdot (y_i - \mu_y) < 0$?

En el caso que el producto de $(x_i - \mu_y) \cdot (y_i - \mu_y)$ sea negativo, será porque el dato para la muestra i sobre el eje X está arriba del promedio, mientras que el dato para eje Y sobre la misma muestra está debajo de su promedio respectivo o al revés, el dato para el eje X está abajo del promedio, mientras que el dato para eje Y sobre la misma muestra está arriba de su promedio respectivo.

Esto significa que uno será positivo y otro negativo $(+)\cdot(-) \to (-)$ o $(-)\cdot(+) \to (-)$. Cuando la covarianza es negativa podemos ver que las muestras indican que en un eje los valores están debajo de su promedio, mientras que los otros están por arriba o al revés. Esto implica una relación inversa.

La ecuación general para calcular la covarianza entre dos ejes es obtener el promedio de las multiplicaciones de las diferencias entre cada dato menos su promedio respectivo al eje:

$$COV(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x) \cdot (y_i - \mu_y) \tag{14}$$

Y podemos interpretar la varianza como sigue:

- **Valor muy positivo** — significa que los datos de un eje están sobre su promedio, mientras que los datos del otro eje también están sobre

su promedio y también significa que los datos de un eje están bajo su promedio mientras que los datos del otro eje también están debajo de su promedio. Esto forma una **relación directa** y significa que cuando un dato está a la alza sobre un eje, en el otro también lo estará y cuando un dato está a la baja en un eje, en el otro también estará a la baja. Por ejemplo, si el peso de una persona es mayor al promedio, la estatura de la persona también será mayor al promedio, y si el peso de una persona está debajo de su promedio, la estatura de la persona también estará bajo su promedio.

- **Valor muy negativo** — significa que los datos de un eje están sobre su promedio, mientras que los datos del otro eje no están están sobre su promedio, por el contrario, están debajo de su promedio. Esto forma una **relación inversa** y significa que cuando un dato está a la alza sobre un eje, en el otro está a la baja y cuando un dato está a la baja en el eje, en el otro se encuentra a la alza. Por ejemplo, si el ancho de banda aumenta, el tiempo de descarga disminuye y si el ancho de banda disminuye debajo de su promedio, el tiempo de descarga aumentará arriba de su promedio.

- **Valor cercano a cero** — significa que ya sea positivo o negativo, hay casi las mismas multiplicaciones positivas que negativas. Esto significa que cuando un eje se encuentra arriba de su promedio, el otro también algunas veces, pero otras veces abajo de su promedio. Esto genera que se cancelen sumas positivas y negativas indicando que no hay una covarianza fuerte entre ambos ejes. Generalmente esto implica que no hay correlación entre los ejes, es decir, que un eje no explica al otro eje de manera directa o indirecta, y a esto se le llama **relación nula** o correlación nula.

Pero, ¿Qué tan positivo o negativo debe ser el valor para decir que hay una correlación fuerte entre ambos ejes? Y ¿Qué significa que dos ejes estén correlacionados?

3.4 Correlación

La correlación entre dos ejes es el grado de similitud entre los datos de cada una de las muestras respecto a la media de su eje. Es decir, si pensamos una hipótesis en términos de estar sobre el promedio o debajo del promedio de un eje, entonces responderíamos preguntas como:

- ¿La persona es más grande o más chica que la edad promedio?
- ¿La persona pesa más o menos que el peso promedio?
- ¿La persona mide más o menos que la estatura promedio?
- ¿La persona gana más o menos que el salario promedio?

En términos de estar arriba o abajo del promedio, cada eje no nos dice nada muy significativo. Pero haciendo el análisis bivariado (sobre dos ejes), encontramos que un eje podría explicar al otro eje de forma directa o indirecta, permitiéndonos hacer hipótesis más complejos como:

- Si la persona es más grande que la edad promedio, ¿Su peso será mayor al peso promedio?
- Si la persona es más chica que la edad promedio, ¿Su estatura será menor a la estatura promedio?
- Si una persona gana más que el promedio, ¿Su edad será menor a la edad promedio?

Estas preguntas forman una correlación entre los datos, que puede ser directa, si la covarianza entre los ejes es muy positiva o puede ser indirecta, si la covarianza entre los ejes es muy negativa.

La mejor forma de medir la correlación entre dos ejes es normalizando los datos, es decir, dividiendo la covarianza entre las desviaciones estándares de cada ejes. De esta forma y asombrosamente, el intervalo quedará entre $[-1, +1]$, donde un valor de -1 será una correlación inversa perfecta y $+1$ será una correlación directa perfecta. Un valor cercano a cero seguirá significando que los dos ejes no están correlacionados o que tienen una correlación nula.

Es se puede ver como:

$$CORR(X, Y) = \frac{COV(X, Y)}{\sigma_x \cdot \sigma_y} \tag{15}$$

Entonces podemos resumir los términos de una correlación como:

$CORR(X, Y)$	**Interpretación**
-1	Correlación inversa perfecta
-0.8	Correlación inversa fuerte
-0.5	Correlación inversa débil
-0.3	Correlación inversa insuficiente
-0.1	Correlación inversa nula
0	Correlación nula
+0.1	Correlación directa nula
+0.3	Correlación directa insuficiente
+0.5	Correlación directa débil
+0.8	Correlación directa fuerte
+1	Correlación directa perfecta

Table 6: Tabla de significado de la correlación entre dos ejes X y Y

Esta correlación también se conoce como *Correlación de Pearson* y es un estadístico muy utilizado para hacer pruebas de independencia entre los ejes. Si

se encuentra una correlación fuerte o perfecta entre dos ejes ya sea directa o inderecta, significará que un eje puede ser explicado a través del otro y viceversa.

4 Caso de Estudio

```
La edad, el peso y la estatura si influyen
```

Vamos a estudiar el comportamiento del salario en las personas usando el siguiente modelo:

$$Salario = \frac{C_1 \cdot Estatura}{C_2 \cdot Edad + C_3 \cdot Peso} \tag{16}$$

Este modelo intenta explicar que el *Salario* de una persona será directamente proporcional a la estatura de una persona, e inversamente proporcional a la edad y peso de la persona. Las constantes C_1, C_2, C_3 necesitan ser ajustadas para un experimento.

Este modelo representa un ejemplo de un indicador que necesitaremos construir en la vida real para lograr predecir un comportamiento, por ejemplo, cuando suponemos que las ventas de un producto son directamente proporcionales a las reseñas de los clientes e inversamente proporcional al precio y el tiempo de entrega. Es decir, suponemos que a mejor reseñas de un producto, mayores ventas (relación directa), pero a mayor precio y tiempo de entrega menores ventas (relación inversa).

Así podemos encontrar muchos casos de estudio en cada sector que nos enfoquemos, desde manejo de dinero y clientes, hasta experimentos con bacterías y reactores.

Este caso de estudio lo podemos ver como:

$$r_i = \frac{C_1 \cdot t_i}{C_2 \cdot e_i + C_3 \cdot s_i} \tag{17}$$

Donde r_i representa el salario de la *i-ésima* muestra, t_i la estatura de esa misma muesta y e_i y s_i la edad y peso respectivamente.

Entonces podemos simular el caso de estudio mediante los siguientes pasos:

1. Generar muestras que contengan datos para la edad, peso, estatura y salario de cada persona. Siguiendo la hipótesis de que el salario será alto para personas altas y bajo para personas grandes y de mayor peso. En otro experimento construiremos los datos sin seguir la hipótesis para comparar.

2. Construir el indicador del salario siguiendo el modelo, usando diferentes C_1, C_2, C_3.

3. Graficar en pares los ejes contra el salario, es decir, edad y salario, peso y salario, estatura y salario y finalmente el indicador y el salario.

4. Medir la media, varianza y desviación estándar de cada eje.

5. Medir la covarianza de los ejes contra el salario.

6. Medir la correlación de los ejes contra el salario.

Finalmente analizaremos los resultados del caso de estudio e interpretaremos que pasa cuando cambiamos las constantes C_1, C_2, C_3 y cuando no seguimos la hipótesis.