Índice general

Lección 4. Proyección ortogonal, media y desviación típica

Marcos Bujosa

19 de septiembre de 2026

Resumen

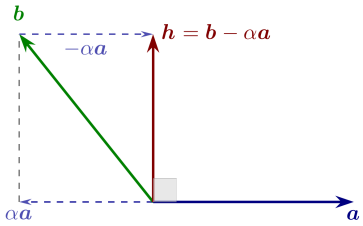
ἀγεωμέτρητος μηδεὶς εἰσίτω ("Nadie no-geómetra entre" *en el frontispicio de la academia de Platón*)

La media aritmética y la desviación típica como objetos geométricos en $\mathbb{R}^n$. La ortogonalidad como principio de cálculo.

- (slides) — (html) — (pdf)

1 De dónde venimos: la proyección ortogonal

En la lección 3 construimos un vector $\boldsymbol{h} = \boldsymbol{b} - \alpha\boldsymbol{a}$ ortogonal a $\boldsymbol{a}$:



$$\langle \boldsymbol{a}, \boldsymbol{h}\rangle = 0 \implies \alpha = \frac{\langle \boldsymbol{a}, \boldsymbol{b}\rangle}{\langle \boldsymbol{a}, \boldsymbol{a}\rangle}; \qquad \boldsymbol{b} = \underbrace{\alpha\boldsymbol{a}}_{\text{"sombra"}} + \underbrace{\boldsymbol{h}}_{\perp \boldsymbol{a}}.$$

Dijimos que cuando $\boldsymbol{a} = \mathbf{1}$ *"ese α resultará ser la media"*. El vector $\alpha\boldsymbol{a}$, que llamamos "sombra" de $\boldsymbol{b}$ sobre $\boldsymbol{a}$, es lo que ahora llamaremos *proyección ortogonal*. Veremos que:

- La *media aritmética* es fruto de una proyección ortogonal (sobre $\mathcal{L}(\mathbf{1})$).
- La *desviación típica* es una norma (distancia).

El principio que une ambos conceptos es la *ortogonalidad*.

De dónde venimos: la proyección ortogonal

En la lección 3, al demostrar la desigualdad de Cauchy–Schwarz, construimos el vector $\boldsymbol{h} = \boldsymbol{b} - \alpha\boldsymbol{a}$ eligiendo α para forzar la ortogonalidad $\boldsymbol{h} \perp \boldsymbol{a}$. La condición $\langle \boldsymbol{a}, \boldsymbol{h}\rangle = 0$ determinaba

unívocamente

$$\alpha = \frac{\langle \boldsymbol{a},\, \boldsymbol{b}\rangle}{\langle \boldsymbol{a},\, \boldsymbol{a}\rangle},$$

y el vector $\alpha\boldsymbol{a}$ era la "sombra" de $\boldsymbol{b}$ sobre la dirección de $\boldsymbol{a}$. Anunciamos entonces que cuando $\boldsymbol{a} = \mathbf{1}$ (el vector cuyas componentes son todas 1), ese escalar α resultaría ser la media aritmética de las componentes de $\boldsymbol{b}$.

El vector $\alpha\boldsymbol{a}$, con $\alpha = \frac{\langle \boldsymbol{a},\boldsymbol{b}\rangle}{\langle \boldsymbol{a},\boldsymbol{a}\rangle}$, que en la lección 3 construimos como "sombra" de $\boldsymbol{b}$ sobre la dirección de $\boldsymbol{a}$, es lo que a partir de ahora llamaremos *proyección ortogonal* de $\boldsymbol{b}$ sobre $\mathcal{L}(\boldsymbol{a})$ (el conjunto de todos los múltiplos de $\boldsymbol{a}$; geométricamente, una recta en $\mathbb{R}^n$ que pasa por el origen). El nombre refleja que el vector diferencia $\boldsymbol{h} = \boldsymbol{b} - \alpha\boldsymbol{a}$ es ortogonal a $\boldsymbol{a}$, que es precisamente la condición que usamos para determinar α.

En la lección 3 anunciamos también que "ajustar por mínimos cuadrados es minimizar la distancia $\|\boldsymbol{y} - \widehat{\boldsymbol{y}}\|$". La proyección ortogonal es, por construcción, el múltiplo de $\boldsymbol{a}$ más cercano a $\boldsymbol{b}$, es decir, $\alpha\boldsymbol{a}$ es el punto de $\mathcal{L}(\boldsymbol{a})$ el que minimiza $\|\boldsymbol{b} - \alpha\boldsymbol{a}\|$. No hace falta ningún cálculo adicional: la ortogonalidad *es* la condición de mínimo.

Hoy cumplimos esa promesa. Pero antes de hacerlo, necesitamos ajustar el producto escalar que usamos: el producto escalar estándar de $\mathbb{R}^n$ no es el adecuado para trabajar en estadística. Ya anticipamos al final de la lección 2 que existía una "norma estadística" con un factor $\frac{1}{n}$, y que serviría para que el vector $\mathbf{1}$ tuviera norma 1. Ahora precisamos esa idea y la elevamos a producto escalar. En cuanto lo ajustemos, todo encajará.

2 El producto escalar en la estadística

Con el producto escalar estándar, $\|\mathbf{1}\|_e^2 = \langle \mathbf{1}, \mathbf{1}\rangle_e = \sum_{i=1}^n 1 = n \neq 1$.

Exigencia estadística: la media de un vector constante c debe ser c. Equivalentemente: $\|\mathbf{1}\| = 1$.

La solución: un nuevo producto escalar con factor de corrección $\frac{1}{n}$:

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle_s = \frac{1}{n}\,\langle \boldsymbol{x}, \boldsymbol{y}\rangle_e = \frac{1}{n}\sum_{i=1}^{n} x_i y_i.$$

Ahora $\|\mathbf{1}\|_s^2 = \langle \mathbf{1}, \mathbf{1}\rangle_s = \frac{1}{n}n = 1$, como queríamos.

- Hereda las propiedades de producto escalar: simetría, linealidad, definido positivo.
- La norma asociada: $\|\boldsymbol{x}\|_s = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x}\rangle_s} = \sqrt{\frac{1}{n}\sum x_i^2}$.
- La ortogonalidad: $\boldsymbol{x} \perp \boldsymbol{y} \Leftrightarrow \langle \boldsymbol{x}, \boldsymbol{y}\rangle_e = 0 \Leftrightarrow \langle \boldsymbol{x}, \boldsymbol{y}\rangle_s = 0$.
- Por tanto, el Teorema de Pitágoras, Cauchy–Schwarz y la desigualdad triangular de la lección 3 siguen siendo válidos con $\langle \cdot, \cdot\rangle_s$ y $\|\cdot\|_s$.

El producto escalar de la estadística

¿Por qué necesitamos un producto escalar distinto? La respuesta viene de una exigencia estadística muy natural, que ya anticipamos al final de la lección 2.

La media aritmética de un vector constante igual a c debe ser c (el valor que mejor representa el centro de la distribución de los datos); en particular, la media del vector $\mathbf{1} = (1, 1, \ldots, 1)$ debe ser 1.

Esto se logra empleando un nuevo producto escalar, que aplica un factor de corrección $1/n$. Con dicho producto escalar, $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_s = \frac{1}{n} \sum x_i y_i$, la norma de $\mathbf{1}$ es efectivamente 1:

$$\|\mathbf{1}\|_s^2 = \langle \mathbf{1}, \mathbf{1} \rangle_s = \frac{1}{n} \sum_{i=1}^{n} 1 = 1 \qquad \Longleftrightarrow \qquad \|\mathbf{1}\|_s = 1$$

Esta norma no solo garantiza que $\|\mathbf{1}\|_s = 1$, además hace que la media aritmética de $\boldsymbol{y}$ sea el producto escalar (de la estadística) entre $\boldsymbol{y}$ y $\mathbf{1}$:

$$\langle \boldsymbol{y}, \mathbf{1} \rangle_s = \frac{1}{n} \sum y_i = \mu_{\boldsymbol{y}}.$$

Una observación importante: la condición de ortogonalidad no cambia. $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_s = 0$ si y solo si $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_e = 0$, porque el factor $\frac{1}{n} > 0$ nunca se anula. Así que todas las nociones de ortogonalidad que hemos desarrollado en las lecciones 2 y 3 siguen siendo válidas, sin ninguna modificación.

Más en general: las demostraciones del teorema de Pitágoras, la desigualdad de Cauchy–Schwarz y la desigualdad triangular que realizamos en la lección 3 no utilizan ninguna propiedad específica del producto escalar euclídeo $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_e$. Solo utilizan tres propiedades abstractas: simetría ($\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$), linealidad en cada argumento y positividad ($\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$, con igualdad solo si $\boldsymbol{x} = \mathbf{0}$). El producto escalar estadístico $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_s = \frac{1}{n} \langle \boldsymbol{x}, \boldsymbol{y} \rangle_e$ hereda las tres propiedades del euclídeo; por tanto, *todos los resultados de la lección 3 siguen siendo válidos* con este nuevo producto escalar y su norma asociada $\| \cdot \|_s$, sin necesidad de modificar ninguna demostración.

Nota. En las lecciones 2 y 3 sembramos esta distinción: hablamos de la "norma euclídea" $\sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_e}$ y de la "norma estadística" $\sqrt{\frac{1}{n} \langle \boldsymbol{x}, \boldsymbol{x} \rangle_e}$. A partir de ahora, cuando trabajemos con datos estadísticos, usaremos siempre el producto escalar con factor $\frac{1}{n}$. El factor $\frac{1}{n}$ que aparece en la varianza, en la covarianza y en la fórmula de la correlación no es una convención arbitraria: es la consecuencia de exigir que $\|\mathbf{1}\|_s = 1$.

3 La media: dos caras de la misma moneda

Primera cara: la media como producto escalar con $\mathbf{1}$.

$$\mu_{\boldsymbol{y}} \;=\; \langle \boldsymbol{y}, \mathbf{1} \rangle_s \;=\; \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Segunda cara: la media como proyección ortogonal sobre $\mathcal{L}(\mathbf{1})$ (los múltiplos de $\mathbf{1}$).

Buscamos α tal que $(\boldsymbol{y} - \alpha\mathbf{1}) \perp \mathbf{1}$:

$$\langle \boldsymbol{y} - \alpha\mathbf{1},\, \mathbf{1}\rangle_s = 0 \implies \langle \boldsymbol{y}, \mathbf{1}\rangle_s - \alpha\,\langle \mathbf{1}, \mathbf{1}\rangle_s = 0 \implies \alpha = \frac{\langle \boldsymbol{y}, \mathbf{1}\rangle_s}{\langle \mathbf{1}, \mathbf{1}\rangle_s} = \frac{\mu_{\boldsymbol{y}}}{1} = \mu_{\boldsymbol{y}}.$$

(*Ambas caras colapsan*: $\alpha = \mu_{\boldsymbol{y}}$.)

Llamaremos *vector de medias* $\overline{\boldsymbol{y}} = \mu_{\boldsymbol{y}}\mathbf{1}$ al vector constante cuyas componentes son todas iguales a $\mu_{\boldsymbol{y}}$. El *vector de medias* $\overline{\boldsymbol{y}}$ es la proyección ortogonal de $\boldsymbol{y}$ sobre $\mathcal{L}(\mathbf{1})$.

La media: dos caras de la misma moneda

La media aritmética admite dos lecturas geométricas, y conviene tenerlas claras porque reaparecerán en cada paso del curso.

Primera cara: la media como producto escalar.

Con el producto escalar estadístico, la media de $\boldsymbol{y}$ es directamente su producto escalar con $\mathbf{1}$:

$$\mu_{\boldsymbol{y}} = \langle \boldsymbol{y}, \mathbf{1}\rangle_s = \frac{1}{n}\sum_{i=1}^{n} y_i \cdot 1 = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Esta lectura es algebraica: la media es una operación lineal sobre $\boldsymbol{y}$.

Segunda cara: la media como proyección ortogonal.

Llamamos $\mathcal{L}(\mathbf{1})$ al conjunto de todos los múltiplos de $\mathbf{1}$; geométricamente, es una recta en $\mathbb{R}^n$ que pasa por el origen. Queremos encontrar el punto de $\mathcal{L}(\mathbf{1})$ más cercano a $\boldsymbol{y}$, es decir, el escalar α tal que $\alpha\mathbf{1}$ minimiza la distancia $\|\boldsymbol{y} - \alpha\mathbf{1}\|_s$. Como aprendimos en la lección 3, el punto más cercano de una recta a un vector es la proyección ortogonal: la diferencia $\boldsymbol{y} - \alpha\mathbf{1}$ (que en unas transparencias llamaremos *vector en desviaciones*) debe ser ortogonal a $\mathbf{1}$.

Imponemos la condición de ortogonalidad:

$$\langle \boldsymbol{y} - \alpha\mathbf{1},\, \mathbf{1}\rangle_s = 0.$$

Usando la linealidad del producto escalar:

$$\langle \boldsymbol{y}, \mathbf{1}\rangle_s - \alpha\,\langle \mathbf{1}, \mathbf{1}\rangle_s = 0 \implies \alpha = \frac{\langle \boldsymbol{y}, \mathbf{1}\rangle_s}{\langle \mathbf{1}, \mathbf{1}\rangle_s} = \frac{\mu_{\boldsymbol{y}}}{1} = \mu_{\boldsymbol{y}}.$$

El resultado es el mismo: $\alpha = \mu_{\boldsymbol{y}}$. Pero la vía geométrica revela *por qué* la media tiene esa forma: es la única manera de que $\boldsymbol{y} - \alpha\mathbf{1}$ sea ortogonal a $\mathbf{1}$, y esa ortogonalidad es lo que caracteriza al punto más cercano.

Llamamos *vector de medias* al vector $\overline{\boldsymbol{y}} = \mu_{\boldsymbol{y}}\mathbf{1}$: es la proyección de $\boldsymbol{y}$ sobre $\mathcal{L}(\mathbf{1})$ y, por tanto, el vector constante más próximo a $\boldsymbol{y}$; sus componentes son todas iguales a $\mu_{\boldsymbol{y}}$.

Conviene dar nombre también a lo que queda fuera de la recta. El *complemento ortogonal* de $\mathcal{L}(\mathbf{1})$, que denotamos $\mathcal{L}(\mathbf{1})^{\perp}$, es el conjunto de todos los vectores ortogonales a $\mathbf{1}$; es decir —recordando que $\mu_{\boldsymbol{v}} = \langle \boldsymbol{v}, \mathbf{1}\rangle_s$—, el conjunto de vectores de media nula: $\boldsymbol{v} \in \mathcal{L}(\mathbf{1})^{\perp} \Leftrightarrow \mu_{\boldsymbol{v}} = 0$.

La vía analítica, presentada aquí como alternativa.

El mismo resultado se obtiene minimizando directamente $f(\alpha) = \|\boldsymbol{y} - \alpha\mathbf{1}\|_s^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \alpha)^2$. Derivando respecto a α e igualando a cero:

$$\frac{df}{d\alpha} = \frac{1}{n}\sum_{i=1}^{n} 2(y_i - \alpha)(-1) = 0 \implies \sum_{i=1}^{n}(y_i - \alpha) = 0 \implies \alpha = \frac{1}{n}\sum_{i=1}^{n} y_i = \mu_{\boldsymbol{y}}. \tag{1}$$

La condición de primer orden nos da un punto crítico, pero no nos dice si se trata de un mínimo o de un máximo. La vía geométrica no tenía esta ambigüedad: la proyección ortogonal es, por construcción, el punto de $\mathcal{L}(\mathbf{1})$ (es decir, el vector constante) más *cercano* a $\boldsymbol{y}$, de modo que el carácter mínimo de la solución quedaba garantizado desde el principio. En la vía analítica, en cambio, debemos comprobarlo explícitamente mediante la condición de segundo orden. Calculando la segunda derivada (que vio en cálculo):

$$\frac{d^2 f}{d\alpha^2} = \frac{1}{n}\sum_{i=1}^{n} 2 = 2 > 0.$$

La segunda derivada es estrictamente positiva (e independiente de α), lo que confirma que la función f es convexa y que el punto crítico encontrado es efectivamente un mínimo global.

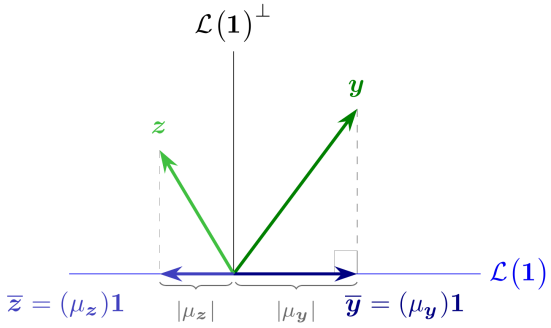
El resultado (1) coincide con el obtenido por la vía geométrica. Pero nótese que la condición $\sum_{i=1}^n (y_i - \alpha) = 0$ no es sino $\langle \boldsymbol{y} - \alpha\mathbf{1}, \mathbf{1}\rangle_s = 0$ escrita en términos de sumatorios: la condición de primer orden del problema de minimización *es* la condición de ortogonalidad. Ambas vías son la misma cosa vista desde ángulos distintos. La vía geométrica es más directa, y su lógica se generalizará sin esfuerzo a la regresión múltiple, donde la vía analítica se vuelve considerablemente más engorrosa.

> *Mensaje central*: calcular la media aritmética es, sin saberlo, resolver un problema de proyección ortogonal. Más adelante veremos que la regresión lineal simple es el mismo procedimiento, pero proyectando sobre un plano en lugar de sobre una recta. Y la regresión múltiple, sobre un subespacio de dimensión mayor. La lógica es siempre la misma: imponer ortogonalidad.

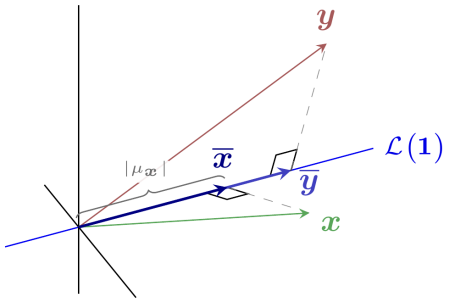
Para profundizar:

- Strang, G. *Introduction to Linear Algebra* (2016), §4.2. El tratamiento de la proyección sobre una recta como caso particular de mínimos cuadrados.
- Bujosa, M. *Curso de Álgebra Lineal*, capítulo 11: libro online.

4 Ilustración geométrica de la proyección



El vector $\boldsymbol{y}$ (verde oscuro) se proyecta sobre $\mathcal{L}(\mathbf{1})$ (la recta de los vectores constantes, en azul). La "sombra" $\overline{\boldsymbol{y}} = \mu_{\boldsymbol{y}}\mathbf{1}$ es su correspondiente vector de medias, cuya longitud es $|\mu_{\boldsymbol{y}}|$. El vector $\boldsymbol{z}$ (verde claro) es otro ejemplo con media negativa.



Representación de dos vectores, $\boldsymbol{y}$ (rojo) y $\boldsymbol{x}$ (verde claro), y sus proyecciones ortogonales sobre $\mathcal{L}(\mathbf{1})$; es decir, sus correspondientes vectores de medias $\overline{\boldsymbol{y}}$ y $\overline{\boldsymbol{x}}$. Los ejes de color negro son perpendiculares a $\mathbf{1}$, es decir, pertenecen al *complemento ortogonal* $\mathcal{L}(\mathbf{1})^\perp$: el conjunto de todos los vectores perpendiculares a $\mathbf{1}$.

Ilustración geométrica de la proyección

Las figuras anteriores ilustran la geometría de la proyección ortogonal sobre $\mathcal{L}(\mathbf{1})$ en espacios de dimensiones 2 y 3 (donde podemos dibujar).

Primera figura. La recta azul es $\mathcal{L}(\mathbf{1}) = \{(a, a) \mid a \in \mathbb{R}\}$. La recta negra vertical es su complemento ortogonal $\mathcal{L}(\mathbf{1})^\perp$ (el conjunto de vectores perpendiculares a $\mathbf{1}$). El vector $\overline{\boldsymbol{y}} = \mu_{\boldsymbol{y}}\mathbf{1}$ (azul) es la proyección ortogonal (la "sombra") del verde oscuro $\boldsymbol{y}$. La figura muestra también un vector $\boldsymbol{z}$ con media negativa: por ello su proyección cae en el lado opuesto del origen.

Segunda figura (representación en un espacio de dimensión mayor). La recta azul es el conjunto de vectores constantes $\mathcal{L}(\mathbf{1})$. El complemento ortogonal $\mathcal{L}(\mathbf{1})^\perp$ es ahora un *plano* que pasa por el origen y es perpendicular a la recta azul. Los vectores $\boldsymbol{y}$ (rojo) y $\boldsymbol{x}$ (verde) tienen proyecciones distintas sobre la recta azul, correspondientes a sus respectivas medias.

Una observación importante sobre los límites de la visualización: en la práctica trabajamos con n observaciones, así que los vectores de datos viven en $\mathbb{R}^n$ con n posiblemente muy grande. No podemos dibujar $\mathbb{R}^{100}$, pero la geometría es la misma: $\mathcal{L}(\mathbf{1})$ sigue siendo una recta, su complemento ortogonal sigue siendo un subespacio de dimensión $n-1$[1], y la proyección sigue siendo el vector de medias. Aunque las figuras anteriores son representaciones literales en $\mathbb{R}^2$ y $\mathbb{R}^3$, podemos interpretarlas como representaciones esquemáticas de la geometría general.

4.1 Ilustración con la visión estadística

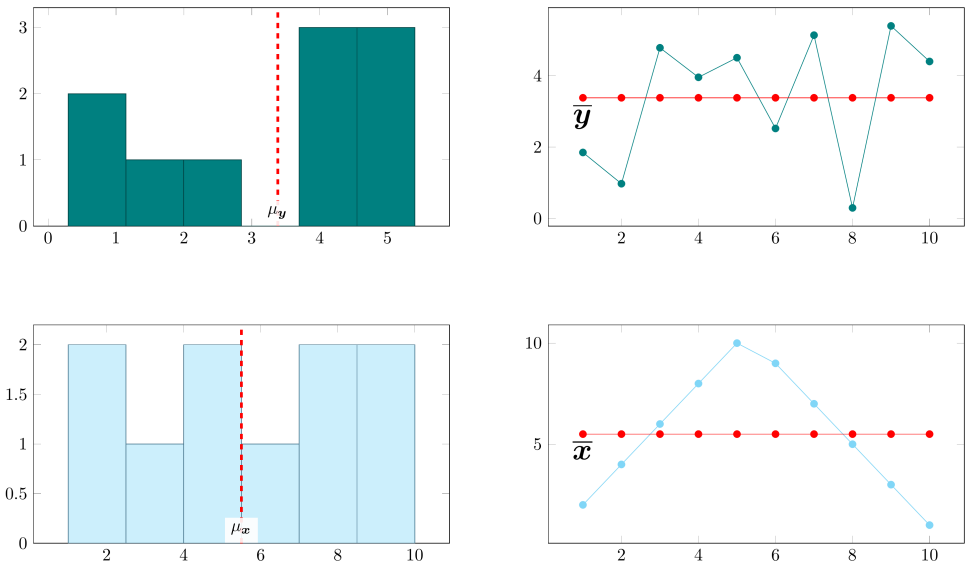


Figura 1: Dos lecturas de la media, lado a lado, con $n = 10$ datos (dos muestras distintas, arriba y abajo). Izquierda: histograma con la media marcada (lectura estadística: el centro de la distribución). Derecha: el mismo vector de datos, componente a componente, junto al vector de medias (constante, en rojo) — la proyección sobre $\mathcal{L}(\mathbf{1})$ (lectura geométrica).

$\mu_{\boldsymbol{y}}$: *un punto* en el histograma; $\quad \overline{\boldsymbol{y}} = \mu_{\boldsymbol{y}}\mathbf{1}$ *un vector constante* en $\mathbb{R}^n$.

Ilustración con la visión estadística

Una tercera figura, con $n = 10$ datos en cada ejemplo, conecta explícitamente esta lectura geométrica con la lectura de la estadística descriptiva ya familiar para el lector. A la izquierda de cada fila aparece el histograma del vector de datos correspondiente, con la media marcada mediante una línea vertical: la lectura habitual de la media como "centro" de la distribución. A la derecha de cada fila aparece el mismo vector de datos, ahora representado componente a componente (el valor de cada y_i frente a su índice i), junto al vector de medias $\overline{\boldsymbol{y}} = \mu_{\boldsymbol{y}}\mathbf{1}$, dibujado como una línea horizontal constante a la altura $\mu_{\boldsymbol{y}}$: la lectura geométrica desarrollada en esta lección, con el vector constante más próximo al vector de datos.

[1] Usamos aquí la palabra "dimensión" de forma completamente informal, apoyándonos en la intuición visual de recta (una dirección) y plano (dos direcciones). En la lección 5 precisaremos esta idea con más calma, al explicar por qué en $\mathbb{R}^2$ la correlación entre dos vectores es siempre ± 1.

Esta figura hace explícito el mensaje central de esta transparencia: la *media* $\mu_{\boldsymbol{y}}$ es un número (que indica el centro de la distribución del histograma); el *vector de medias* $\overline{\boldsymbol{y}}$ es un vector, con tantas componentes como datos (la línea horizontal roja de la derecha, con sus $n = 10$ puntos). Son objetos relacionados pero de naturaleza distinta y conviene no confundirlos: la media es un escalar; el vector de medias es el vector resultante de la proyección ortogonal sobre los múltiplos de $\mathbf{1}$.

5 El vector en desviaciones es ortogonal a 1

El vector en desviaciones (*respecto a la media*) es

$$\boldsymbol{y} - \overline{\boldsymbol{y}} \;=\; \boldsymbol{y} - \mu_{\boldsymbol{y}}\mathbf{1}.$$

Por construcción, $(\boldsymbol{y} - \overline{\boldsymbol{y}}) \perp \mathbf{1}$, es decir:

$$\langle \boldsymbol{y} - \overline{\boldsymbol{y}},\, \mathbf{1}\rangle_s = 0 \iff \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_{\boldsymbol{y}}) = 0.$$

Por tanto, la media del vector en desviaciones, $\mu_{(\boldsymbol{y}-\overline{\boldsymbol{y}})}$*, es cero.*

Esta no es una propiedad que surja a posteriori; esta es *la condición de ortogonalidad* que caracteriza a la media.

El vector en desviaciones respecto a la media es ortogonal a 1

La condición de ortogonalidad $(\boldsymbol{y} - \overline{\boldsymbol{y}}) \perp \mathbf{1}$ tiene una traducción estadística inmediata sobre la media del vector en desviaciones:

$$\langle \boldsymbol{y} - \overline{\boldsymbol{y}},\, \mathbf{1}\rangle_s = 0 \iff \mu_{(\boldsymbol{y}-\overline{\boldsymbol{y}})} = 0,$$

donde hemos usado que si $\mu_{\boldsymbol{v}}$ es la notación para la media de un vector $\boldsymbol{v}$, entonces la media del vector en desviaciones es $\mu_{(\boldsymbol{y}-\overline{\boldsymbol{y}})} = \langle \boldsymbol{y} - \overline{\boldsymbol{y}}, \mathbf{1}\rangle_s$.

Este resultado no es más que un caso particular de la caracterización de $\mathcal{L}(\mathbf{1})^{\perp}$ como *el conjunto de vectores de media cero* que vimos más arriba: el vector $\boldsymbol{y} - \overline{\boldsymbol{y}}$ pertenece a $\mathcal{L}(\mathbf{1})^{\perp}$ porque, por construcción, es ortogonal a $\mathbf{1}$ y, por tanto, tiene media cero.

La media del vector en desviaciones es cero. En los cursos de estadística descriptiva esto se presenta como una "propiedad de la media"; aquí vemos que es mucho más que eso: es la condición geométrica que la *caracteriza*.

Resumiendo:

La estadística descompone el vector de datos $\boldsymbol{y}$ en dos componentes ortogonales:

$$\boldsymbol{y} = \underbrace{\overline{\boldsymbol{y}}}_{\text{Vector de medias}} + \underbrace{(\boldsymbol{y} - \overline{\boldsymbol{y}})}_{\text{Vector en desviaciones}},$$

donde el *vector de medias* es un vector constante: $\overline{\boldsymbol{y}} = \mu_{\boldsymbol{y}}\mathbf{1}$; es decir, $\overline{\boldsymbol{y}} \in \mathcal{L}(\mathbf{1})$; y donde la constante $\mu_{\boldsymbol{y}}$ se denomina *media aritmética*.

Avance: la regresión lineal simple descompondrá $\boldsymbol{y}$ en dos componentes ortogonales:

$$\boldsymbol{y} = \underbrace{\widehat{\boldsymbol{y}}}_{\text{Ajuste}} + \underbrace{\widehat{\boldsymbol{e}}}_{\text{Vector de residuos}},$$

donde $\widehat{\boldsymbol{y}} \in \mathcal{L}(\mathbf{1}, \boldsymbol{x})$; es decir, $\widehat{\boldsymbol{y}}$ es una combinación lineal de $\mathbf{1}$ y $\boldsymbol{x}$. Consecuentemente habrá *dos* condiciones de ortogonalidad (una por cada regresor).

La regresión lineal múltiple es similar, con tantas condiciones de ortogonalidad como regresores. *La lógica será siempre la misma.*

Este patrón se repetirá en toda la regresión. Cuando ajustemos una recta de regresión, los residuos serán ortogonales a *cada uno* de los regresores. Con un regresor (la constante $\mathbf{1}$), hay una única condición: $\sum e_i = 0$. Con dos regresores ($\mathbf{1}$ y $\boldsymbol{x}$), habrá dos condiciones: $\sum e_i = 0$ y $\sum x_i e_i = 0$. Con k regresores habrá k condiciones. La lógica es siempre la misma: los errores son ortogonales a cada uno de los regresores.

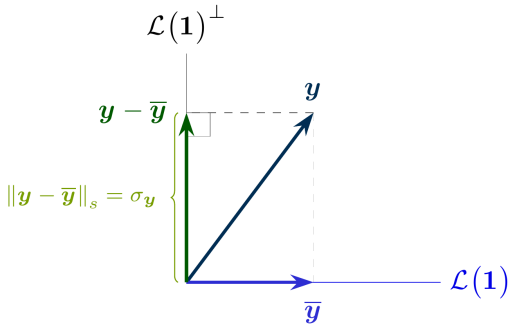
Esta multiplicidad de condiciones de ortogonalidad es lo que hace que la regresión sea una *proyección*: el vector de residuos es ortogonal al subespacio generado por los regresores. El caso de la media es el caso más simple: un único regresor ($\mathbf{1}$), proyección sobre una recta, una sola condición de ortogonalidad.

En el "avance" mostrado en la transparencia se emplea, a modo de anticipo, la notación $\widehat{\boldsymbol{y}}$ y $\widehat{\boldsymbol{e}}$ con circunflejo (o "gorro") que formalizaremos en la lección 6 (regresión lineal simple).

6 Desviación típica: como norma del vector en desviaciones

Definición. La *desviación típica* de $\boldsymbol{y}$ es la norma del componente $\boldsymbol{y} - \overline{\boldsymbol{y}}$ ortogonal a $\mathbf{1}$:

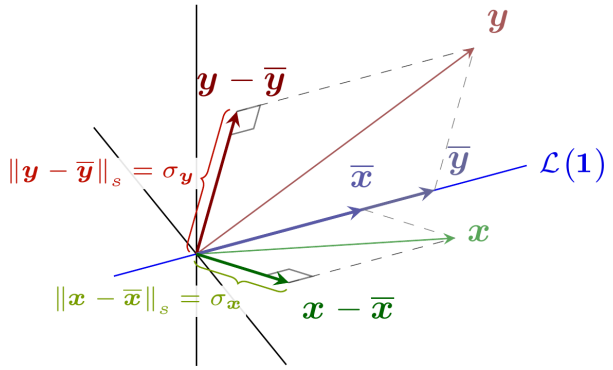
$$\sigma_{\boldsymbol{y}} = \|\boldsymbol{y} - \overline{\boldsymbol{y}}\|_s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_{\boldsymbol{y}})^2}.$$



Como $\overline{\boldsymbol{y}} \perp (\boldsymbol{y} - \overline{\boldsymbol{y}})$, el *Teorema de Pitágoras* de la lección 3 aplicado a $\boldsymbol{y} = \overline{\boldsymbol{y}} + (\boldsymbol{y} - \overline{\boldsymbol{y}})$ da:

$$\|\boldsymbol{y}\|_s^2 = \|\overline{\boldsymbol{y}}\|_s^2 + \|\boldsymbol{y} - \overline{\boldsymbol{y}}\|_s^2 = \|\overline{\boldsymbol{y}}\|_s^2 + \sigma_{\boldsymbol{y}}^2.$$

La lección 5 desarrollará esta identidad en profundidad.



En $\mathbb{R}^3$: proyecciones de $\boldsymbol{y}$ y $\boldsymbol{x}$ sobre $\mathcal{L}(\mathbf{1})$ y sobre $\mathcal{L}(\mathbf{1})^\perp$. Las normas de las proyecciones sobre $\mathcal{L}(\mathbf{1})^\perp$ son las desviaciones típicas respectivas.

La desviación típica como norma del vector en desviaciones respecto a la media

Una vez que tenemos la proyección $\overline{\boldsymbol{y}} = \mu_{\boldsymbol{y}}\mathbf{1}$ y el vector en desviaciones $\boldsymbol{y}-\overline{\boldsymbol{y}}$, la norma de dicho vector es un objeto natural. Mide cuánto se aleja $\boldsymbol{y}$ de su proyección, es decir, cómo de lejos está $\boldsymbol{y}$ del vector constante más próximo $\overline{\boldsymbol{y}}$. Dicho de otra forma: cuánto se dispersan las componentes de $\boldsymbol{y}$ respecto a su media $\mu_{\boldsymbol{y}}$.

Desarrollando la norma:

$$\sigma_{\boldsymbol{y}} = \|\boldsymbol{y}-\overline{\boldsymbol{y}}\|_s = \sqrt{\langle \boldsymbol{y}-\overline{\boldsymbol{y}}, \boldsymbol{y}-\overline{\boldsymbol{y}}\rangle_s} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu_{\boldsymbol{y}})^2}.$$

Esta es exactamente la fórmula de la desviación típica que se define en los cursos de estadística descriptiva. Aquí vemos que no es una fórmula ad hoc: es la norma de un vector, el vector en desviaciones respecto a la media.

La primera figura de esta sección muestra la geometría en $\mathbb{R}^2$: el vector $\boldsymbol{y}$ (verde) se descompone en su proyección $\overline{\boldsymbol{y}}$ (sobre la recta azul $\mathcal{L}(\mathbf{1})$) y el vector en desviaciones $\boldsymbol{y}-\overline{\boldsymbol{y}}$ (sobre la recta negra $\mathcal{L}(\mathbf{1})^\perp$). La desviación típica $\sigma_{\boldsymbol{y}}$ es la longitud de la componente vertical, es decir, la norma del *vector en desviaciones* respecto a la media.

Como $\overline{\boldsymbol{y}} \perp (\boldsymbol{y}-\overline{\boldsymbol{y}})$ (por construcción), el Teorema de Pitágoras en $\mathbb{R}^n$ (demostrado en la lección 3, con enunciado genérico $\boldsymbol{x}\perp\boldsymbol{z} \Rightarrow \|\boldsymbol{x}+\boldsymbol{z}\|^2 = \|\boldsymbol{x}\|^2+\|\boldsymbol{z}\|^2$) aplicado a la descomposición $\boldsymbol{y} = \overline{\boldsymbol{y}} + (\boldsymbol{y}-\overline{\boldsymbol{y}})$ nos da:

$$\|\boldsymbol{y}\|_s^2 = \|\overline{\boldsymbol{y}}\|_s^2 + \|\boldsymbol{y}-\overline{\boldsymbol{y}}\|_s^2 = \mu_{\boldsymbol{y}}^2 + \sigma_{\boldsymbol{y}}^2,$$

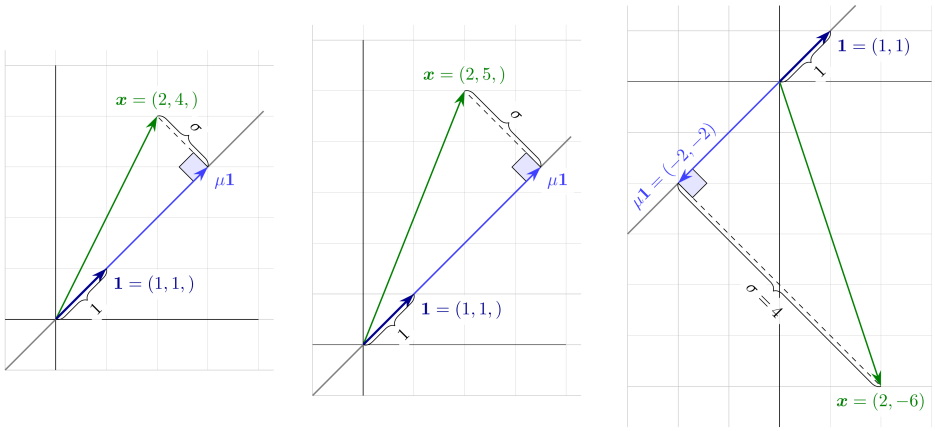
donde el paso $\|\overline{\boldsymbol{y}}\|_s^2 = \mu_{\boldsymbol{y}}^2$ usa la homogeneidad de la norma (vista en la lección 2): $\|\mu_{\boldsymbol{y}}\mathbf{1}\|_s = |\mu_{\boldsymbol{y}}|\cdot\|\mathbf{1}\|_s = |\mu_{\boldsymbol{y}}|\cdot 1 = |\mu_{\boldsymbol{y}}|$, y al elevar al cuadrado $|\mu_{\boldsymbol{y}}|^2 = \mu_{\boldsymbol{y}}^2$.

En la lección 5 desarrollaremos esta identidad con todo detalle: de ella saldrán la varianza, la covarianza y la correlación.

La segunda figura muestra la misma geometría para dos vectores distintos en $\mathbb{R}^3$. Cada uno se descompone en su componente en $\mathcal{L}(\mathbf{1})$ (su media, en azul) y su componente en $\mathcal{L}(\mathbf{1})^\perp$ (su vector centrado, en verde/rojo). La norma de esta segunda componente es la desviación típica respectiva.

Para recordar. La desviación típica es una norma. El vector de medias es una proyección y la media aritmética es el valor de las componentes del vector de medias. Todos son objetos geométricos elementales que ya conocíamos; la novedad es reconocerlos en las fórmulas estadísticas habituales.

7 Ejemplos numéricos en $\mathbb{R}^2$



Ejemplos numéricos en $\mathbb{R}^2$

Verificamos los tres ejemplos de las figuras.

Ejemplo 1. $\boldsymbol{x} = (2, 4)$.

Media: $\mu_{\boldsymbol{x}} = \frac{1}{2}(2 + 4) = 3$.

Vector de medias: $\overline{\boldsymbol{x}} = 3\mathbf{1} = (3, 3)$.

Vector en desviaciones: $\boldsymbol{x} - \overline{\boldsymbol{x}} = (2 - 3, 4 - 3) = (-1, 1)$.

Comprobación de ortogonalidad: $\langle(-1, 1), (1, 1)\rangle_s = \frac{1}{2}(-1 + 1) = 0$.

Desviación típica: $\sigma_{\boldsymbol{x}} = \|(-1, 1)\|_s = \sqrt{\frac{1}{2}(1 + 1)} = \sqrt{1} = 1$.

Pitágoras: $\|\boldsymbol{x}\|_s^2 = \frac{1}{2}(4 + 16) = 10$; $\mu_{\boldsymbol{x}}^2 + \sigma_{\boldsymbol{x}}^2 = 9 + 1 = 10$.

Ejemplo 2. $\boldsymbol{x} = (2, 5)$.

Media: $\mu_{\boldsymbol{x}} = \frac{7}{2} = 3{,}5$.

Vector de medias: $\overline{\boldsymbol{x}} = (3{,}5,\ 3{,}5)$.

Vector en desviaciones: $\boldsymbol{x} - \overline{\boldsymbol{x}} = (-1{,}5,\ 1{,}5)$.

Comprobación de ortogonalidad: $\langle(-1{,}5,\ 1{,}5), (1, 1)\rangle_s = \frac{1}{2}(-1{,}5 + 1{,}5) = 0$.

Desviación típica: $\sigma_{\boldsymbol{x}} = \sqrt{\frac{1}{2}(2{,}25 + 2{,}25)} = \sqrt{2{,}25} = 1{,}5$.

Ejemplo 3. $\boldsymbol{x} = (2, -6)$.

Media: $\mu_{\boldsymbol{x}} = \frac{1}{2}(2 - 6) = -2$.

Vector de medias: $\overline{\boldsymbol{x}} = (-2, -2)$.

Vector en desviaciones: $\boldsymbol{x} - \overline{\boldsymbol{x}} = (4, -4)$.

Comprobación de ortogonalidad: $\langle (4, -4), (1, 1) \rangle_s = \frac{1}{2}(4 - 4) = 0$.

Desviación típica: $\sigma_{\boldsymbol{x}} = \sqrt{\frac{1}{2}(16 + 16)} = \sqrt{16} = 4$.

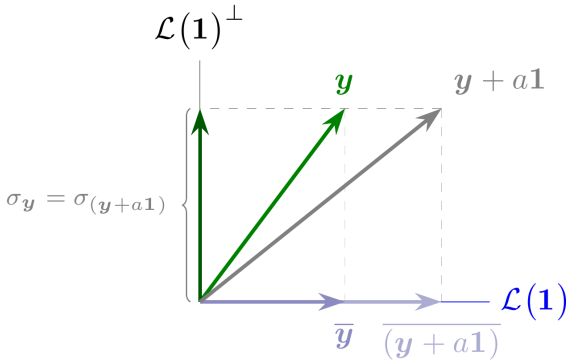
Pitágoras: $\|\boldsymbol{x}\|_s^2 = \frac{1}{2}(4 + 36) = 20$; $\mu_{\boldsymbol{x}}^2 + \sigma_{\boldsymbol{x}}^2 = 4 + 16 = 20$.

Observación. En los tres casos la norma del vector $\mathbf{1}$ es $\|(1, 1)\|_s = \sqrt{\frac{1}{2}(1 + 1)} = 1$, como exigimos. Y el ángulo que forma $\mathbf{1}$ con el eje horizontal es de 45°, de modo que la proyección sobre $\mathcal{L}(\mathbf{1})$ es efectivamente la diagonal del plano. El tercer ejemplo es especialmente ilustrativo: la media es negativa ($\mu = -2$), así que el vector de medias apunta en el sentido negativo de la diagonal.

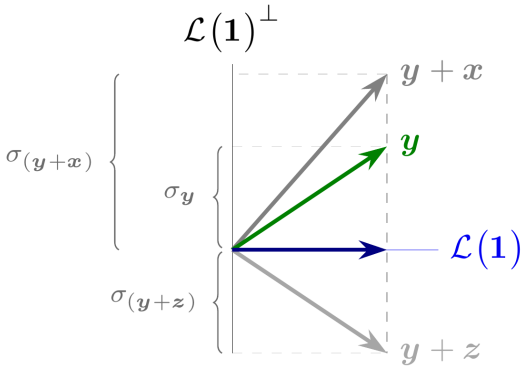
Es buen momento para mirar el notebook Python.

Las figuras de hoy tienen su versión interactiva en $\mathbb{R}^3$: rote los vectores y compruebe que la desviación típica es la norma del vector en desviaciones.

8 Propiedades geométricas



Sumar una constante no nula ($\boldsymbol{y} + a\mathbf{1}$): cambia la media μ, pero *nunca* la desviación típica σ.



Sumar un vector de media cero (i.e., perpendicular a $\mathbf{1}$): no cambia la media μ, pero *generalmente cambia* la desviación típica σ.

Propiedades geométricas

Las figuras anteriores ilustran dos propiedades geométricas de la media y la desviación típica que son inmediatas desde la perspectiva de la proyección, pero que en los cursos de estadística descriptiva se demuestran algebraicamente.

Propiedad 1: sumar una constante no cambia la desviación típica.

Si $\boldsymbol{w} = \boldsymbol{y} + a\mathbf{1}$ para algún escalar a, entonces:

- La proyección de $\boldsymbol{w}$ sobre $\mathcal{L}(\mathbf{1})$ es $\overline{\boldsymbol{w}} = (\mu_{\boldsymbol{y}} + a)\mathbf{1}$: la media aumenta en a.
- Los vectores en desviaciones respecto a sus respectivas medias son iguales: $\boldsymbol{w} - \overline{\boldsymbol{w}} = (\boldsymbol{y} + a\mathbf{1}) - (\mu_{\boldsymbol{y}} + a)\mathbf{1} = \boldsymbol{y} - \mu_{\boldsymbol{y}}\mathbf{1} = \boldsymbol{y} - \overline{\boldsymbol{y}}$.

Geométricamente, $a\mathbf{1}$ está en $\mathcal{L}(\mathbf{1})$, así que sumar $a\mathbf{1}$ desplaza el vector *a lo largo de la recta* $\mathcal{L}(\mathbf{1})$, sin modificar la componente perpendicular. La desviación típica, que es la norma de esa componente perpendicular, no varía.

Propiedad 2: sumar un vector de media cero puede cambiar la desviación típica.

Sea $\boldsymbol{x}$ un vector de media cero, $\mu_{\boldsymbol{x}} = 0$; entonces $\boldsymbol{x} \perp \mathbf{1}$, es decir, $\boldsymbol{x} \in \mathcal{L}(\mathbf{1})^{\perp}$. Al sumárselo a $\boldsymbol{y}$:

- La proyección de $\boldsymbol{y} + \boldsymbol{x}$ sobre $\mathcal{L}(\mathbf{1})$ sigue siendo $\overline{\boldsymbol{y}}$: la media no cambia, porque $\boldsymbol{x}$ no tiene componente en $\mathcal{L}(\mathbf{1})$.
- Su vector en desviaciones es $(\boldsymbol{y} - \overline{\boldsymbol{y}}) + \boldsymbol{x}$: la componente perpendicular generalmente sí cambia.

La desviación típica de $\boldsymbol{y} + \boldsymbol{x}$ depende de cómo se combine $\boldsymbol{y} - \overline{\boldsymbol{y}}$ con $\boldsymbol{x}$ dentro del subespacio $\mathcal{L}(\mathbf{1})^{\perp}$. Puede ser mayor, menor o igual que $\sigma_{\boldsymbol{y}}$, según la dirección de $\boldsymbol{x}$.

La segunda figura de esta sección ilustra las dos posibilidades con *dos* vectores de media cero distintos, $\boldsymbol{x}$ y $\boldsymbol{z}$: el vector $\boldsymbol{y} + \boldsymbol{x}$ tiene mayor desviación típica que $\boldsymbol{y}$, mientras que $\boldsymbol{y} + \boldsymbol{z}$ tiene la misma. Los tres vectores, en cambio, tienen la misma proyección sobre $\mathcal{L}(\mathbf{1})$ (la misma media).

Para profundizar. Estas dos propiedades son la versión geométrica de lo que en estadística se llama "invarianza de la varianza frente a traslaciones" y "la varianza mide dispersión respecto a la media". (La varianza no es más que $\sigma_{\boldsymbol{x}}^2$, el cuadrado de la desviación típica que acabamos de definir; la lección 5 la presentará con ese nombre y la estudiará con detalle.) Desde la geometría, son casi obvias: la varianza es el cuadrado de la norma de la componente perpendicular a $\mathcal{L}(\mathbf{1})$; sumar algo paralelo a $\mathbf{1}$ no cambia esa componente; sumar algo perpendicular a $\mathbf{1}$ generalmente la cambia.

- Wooldridge, J. *Introductory Econometrics* (2020), Apéndice C: invarianza de la media y la desviación típica ante transformaciones lineales, desde la estadística clásica.

9 Recapitulación

Objeto/Concepto	Definición geométrica/algebraica	Fórmula en estadística descriptiva
Media $\mu_{\boldsymbol{y}}$	$\langle \boldsymbol{y}, \mathbf{1} \rangle_s$	$\frac{1}{n}\sum y_i$
Vector de medias $\overline{\boldsymbol{y}}$	Proyección de $\boldsymbol{y}$ sobre $\mathcal{L}(\mathbf{1})$	—
Vector en desviaciones	$\boldsymbol{y} - \overline{\boldsymbol{y}}$, con $(\boldsymbol{y} - \overline{\boldsymbol{y}}) \perp \mathbf{1}$	—
Desviación típica $\sigma_{\boldsymbol{y}}$	$\|\boldsymbol{y} - \overline{\boldsymbol{y}}\|_s$	$\sqrt{\frac{1}{n}\sum (y_i - \mu_{\boldsymbol{y}})^2}$
Teorema de Pitágoras (*)	$\|\boldsymbol{y}\|_s^2 = \|\overline{\boldsymbol{y}}\|_s^2 + \sigma_{\boldsymbol{y}}^2$	$\frac{1}{n}\sum y_i^2 = \mu_{\boldsymbol{y}}^2 + \sigma_{\boldsymbol{y}}^2$

(*) Lo veremos con más detalle en la siguiente lección.

El principio unificador: la media $\mu_{\boldsymbol{y}}$ es, a la vez, el producto escalar $\langle \boldsymbol{y}, \mathbf{1} \rangle_s$ y el coeficiente de la proyección ortogonal de $\boldsymbol{y}$ sobre $\mathcal{L}(\mathbf{1})$. La ortogonalidad lo define todo.

Recapitulación

El mensaje central de esta sesión es que dos de los estadísticos más elementales —la media μ y la desviación típica σ— son objetos geométricos: el coeficiente de una proyección ortogonal (por cuanto queda multiplicado el vector $\mathbf{1}$) y una norma, respectivamente.

Esto no es una mera reformulación. Tiene consecuencias prácticas:

1. La condición $(\boldsymbol{y} - \overline{\boldsymbol{y}}) \perp \mathbf{1}$ —es decir, $\sum(y_i - \mu_{\boldsymbol{y}}) = 0$— no es una propiedad de la media: es la *definición* de la media como proyección ortogonal sobre los vectores constantes $\mathcal{L}(\mathbf{1})$.
2. El Teorema de Pitágoras $\|\boldsymbol{y}\|_s^2 = \|\overline{\boldsymbol{y}}\|_s^2 + \sigma_{\boldsymbol{y}}^2$ no es una fórmula a memorizar: es la consecuencia inevitable de la ortogonalidad entre $\overline{\boldsymbol{y}}$ y $\boldsymbol{y} - \overline{\boldsymbol{y}}$.
3. El factor $\frac{1}{n}$ en el producto escalar estadístico no es arbitrario: es la consecuencia de exigir $\|\mathbf{1}\|_s = 1$, que a su vez es la condición para que la media de $\mathbf{1}$ sea 1.

La lección 5 llevará este marco geométrico más lejos: la varianza (con más profundidad que la mención de hoy), la covarianza como producto escalar de vectores centrados, y la correlación como el coseno del ángulo entre esos vectores centrados. El ángulo, que introdujimos en la lección 3, vuelve a ser protagonista.

10 Guiño a la sesión siguiente

Lección 5: varianza, covarianza y correlación.

- La *varianza* $\sigma_{\boldsymbol{y}}^2 = \|\boldsymbol{y} - \overline{\boldsymbol{y}}\|_s^2$: Tma. de Pitágoras con más profundidad.
- La *covarianza* entre $\boldsymbol{x}$ e $\boldsymbol{y}$: el producto escalar de sus vectores en desviaciones $(\boldsymbol{x} - \overline{\boldsymbol{x}})$ e $(\boldsymbol{y} - \overline{\boldsymbol{y}})$.
- La *correlación* $\rho_{\boldsymbol{xy}}$: el coseno del ángulo entre esos vectores en desviaciones.

El ángulo que introdujimos en la lección 3 vuelve al centro del escenario.

Y vista la correlación tendremos todas las piezas para abordar la regresión lineal simple.

Guiño a la sesión siguiente

Hoy hemos visto que la media y la desviación típica son una proyección y una norma. Eso ya es mucho. Pero hay más.

En la lección 5 llevaremos la misma lógica más lejos. Centrar los vectores significa trabajar con $\boldsymbol{x} - \overline{\boldsymbol{x}}$ e $\boldsymbol{y} - \overline{\boldsymbol{y}}$: las componentes perpendiculares a las proyecciones sobre $\mathcal{L}(\mathbf{1})$, una por cada variable. El producto escalar de esos dos *vectores en desviaciones* es la *covarianza*:[2]

$$\sigma_{\boldsymbol{xy}} = \langle \boldsymbol{x} - \overline{\boldsymbol{x}},\, \boldsymbol{y} - \overline{\boldsymbol{y}} \rangle_s = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{\boldsymbol{x}})(y_i - \mu_{\boldsymbol{y}}).$$

Y el coseno del ángulo entre esos vectores en desviaciones es el coeficiente de *correlación de Pearson*:

$$\rho_{\boldsymbol{xy}} = \cos\theta = \frac{\langle \boldsymbol{x} - \overline{\boldsymbol{x}}, \boldsymbol{y} - \overline{\boldsymbol{y}} \rangle_s}{\|\boldsymbol{x} - \overline{\boldsymbol{x}}\|_s \, \|\boldsymbol{y} - \overline{\boldsymbol{y}}\|_s} = \frac{\sigma_{\boldsymbol{xy}}}{\sigma_{\boldsymbol{x}} \sigma_{\boldsymbol{y}}}.$$

El ángulo, que en la lección 3 introdujimos como herramienta geométrica, resulta ser el coeficiente de correlación. La desigualdad de Cauchy–Schwarz, que en la lección 3 garantizaba que el coseno está en $[-1, 1]$, garantiza aquí que $\rho \in [-1, 1]$. Todo es coherente.

Con la correlación en la mano, estaremos listos para abordar la regresión lineal simple: veremos que el coeficiente de regresión es, esencialmente, la correlación reescalada por las desviaciones típicas. Y el R^2 de la regresión resultará ser $\rho^2 = \cos^2\theta$: el cuadrado del coseno del ángulo entre los vectores en desviaciones.

En los libros de estadística, $\rho_{\boldsymbol{xy}}$ aparece a veces como $r_{\boldsymbol{xy}}$ cuando se quiere subrayar su carácter de estimador muestral; en este curso usaremos siempre ρ.

[2] Adelantamos ya las fórmulas exactas, a modo de anticipo; la lección 5 las presentará como definiciones formales y se detendrá con calma en justificar por qué tienen precisamente esa forma.