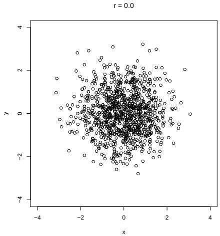
# MAT-041: Correlación lineal

**Felipe Osorio**

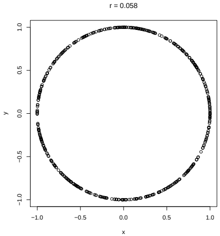
fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

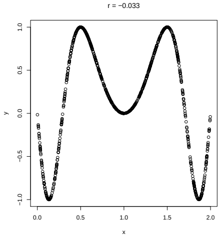
# Correlación: Midiendo asociación lineal



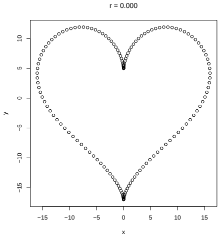

(a) $r = 0.000$




(b) $r = 0.058$

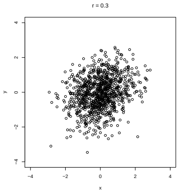



(c) $r = -0.033$

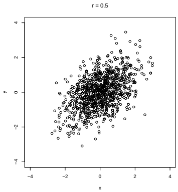



(d) $r = 0.000$

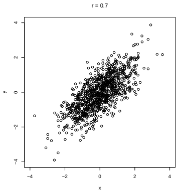
# Correlación: Midiendo asociación lineal



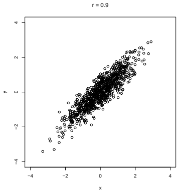

(a) $r = 0.30$



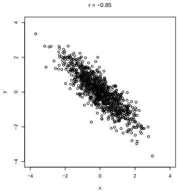

(b) $r = 0.50$



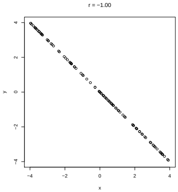

(c) $r = 0.70$




(d) $r = 0.90$




(e) $r = -0.85$




(f) $r = -1.00$

### Definición 1 (Covarianza):

Para el conjunto $(x_1, y_1), \ldots, (x_n, y_n)$, se define la covarianza como una medida de variabilidad conjunta de dos variables cuantitativas, como:

$$\mathsf{cov}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

*Observación:*

Evidentemente, $\mathsf{cov}(\boldsymbol{x}, \boldsymbol{x}) = \mathsf{var}(\boldsymbol{x}) = s_x^2$.

# Propiedades de la covarianza

**Propiedades:**

(a)

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}.$$

(b)

$$\mathsf{cov}(a\boldsymbol{x} + b, c\boldsymbol{y} + d) = ac\,\mathsf{cov}(\boldsymbol{x}, \boldsymbol{y}).$$

EX UMBRA IN SOLEM

**Definición 2 (Correlación):**

La correlación entre $\boldsymbol{x} = (x_1, \dots, x_n)^\top$ e $\boldsymbol{y} = (y_1, \dots, y_n)^\top$ es la covarianza de sus versiones estandarizadas. Es decir,

$$\begin{aligned}
\operatorname{cor}(\boldsymbol{x}, \boldsymbol{y}) &= \frac{1}{n-1} \sum_{i=1}^{n} \Big(\frac{x_i - \overline{x}}{s_x}\Big)\Big(\frac{y_i - \overline{y}}{s_y}\Big) \\
&= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \\
&= \frac{\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{\operatorname{var}(\boldsymbol{x}) \operatorname{var}(\boldsymbol{y})}}.
\end{aligned}$$

*Observación:*

$\operatorname{cor}(\boldsymbol{x}, \boldsymbol{y})$ es una medida adimensional.

EX UMBRA IN SOLEM

**Propiedades:**

(a)
$$\mathsf{cor}(a\boldsymbol{x} + b, c\boldsymbol{y} + d) = \pm\,\mathsf{cor}(\boldsymbol{x}, \boldsymbol{y}).$$

(b)
$$\mathsf{var}(\boldsymbol{x} + \boldsymbol{y}) = \mathsf{var}(\boldsymbol{x}) + \mathsf{var}(\boldsymbol{y}) + 2\,\mathsf{cov}(\boldsymbol{x}, \boldsymbol{y}).$$

(c)
$$\{\mathsf{cor}(\boldsymbol{x}, \boldsymbol{y})\}^2 \leq 1.$$

*Observación:*

- Evidentemente, $-1 \leq \mathsf{cor}(\boldsymbol{x}, \boldsymbol{y}) \leq 1$.
- Cuando $\mathsf{cor}(\boldsymbol{x}, \boldsymbol{y}) = 0$, diremos que $\boldsymbol{x}$ e $\boldsymbol{y}$ son no correlacionados.

**Definición 3 (Coeficiente de correlación de Spearman):**

Suponga los datos pareados $(x_1, y_1), \ldots, (x_n, y_n)$. Sea $R_i$, $S_i$ los rangos de $x_i$ e $y_i$, respectivamente $(i = 1, \ldots, n)$. Entonces el coeficiente de correlación de Spearman es dado por

$$r_{\mathsf{S}} = \frac{\sum_{i=1}^n (R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^n (R_i - \overline{R})^2 \sum_{i=1}^n (S_i - \overline{S})^2}}.$$

*Observación:*

Sea

$$D = \sum_{i=1}^n (R_i - S_i)^2,$$

y suponga que no existen empates entre los $x$'s e $y$'s, entonces podemos escribir

$$r_{\mathsf{S}} = 1 - \frac{6D}{n^2 - 1}.$$

# Ejemplo: Distorsiones de Lenna[1]

(a) $r = 1.000$

(b) $r = 0.903$

(c) $r = 0.991$

(d) $r = 0.915$

(e) $r = 0.983$

(f) $r = 0.799$

[1] (a) Original, (b) sal y pimienta, (c) filtro mediana, (d) ruido speckle, (e) filtro Lee, (f) imagen saturada.

# Ejemplo: Distorsiones de Lenna

```
> library(SpatialPack)
# https://github.com/faosorios/SSIM/blob/master/data/lena.rda
> load("lena.rda") # carga datos de Lena

# aplica distorsiones y filtros
> lena.05 <- clipping(lena, low = 0.5) # saturación
> lena.sp <- imnoise(lena, type = "saltnpepper")
> lena.speckle <- imnoise(lena, type = "speckle")
> lena.med <- denoise(lena.sp, type = "median") # filtro mediana
> lena.lee <- denoise(lena.speckle, type = "Lee") # filtro de Lee

# calculando correlaciones
> x <- as.vector(lena) # 262144 observaciones
> cor(x, x)
[1] 1
> cor(x, as.vector(lena.05))
[1] 0.7997093
> cor(x, as.vector(lena.sp))
[1] 0.9028631
> cor(x, as.vector(lena.med))
[1] 0.9907281
> cor(x, as.vector(lena.speckle))
[1] 0.9154696
> cor(x, as.vector(lena.lee))
[1] 0.9829129
```

# Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

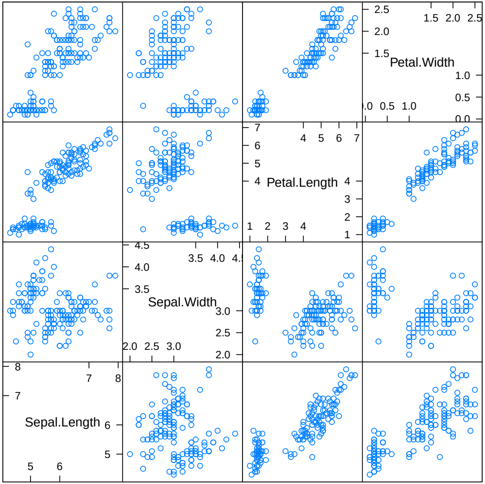
(a) setosa

(b) versicolor

(c) virginica

# Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Edgar Anderson's Iris Data



Deseamos estudiar $p$ variables (características) de interés asociadas a una muestra aleatoria $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ donde cada $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ es un vector $p$-dimensional.

Podemos disponer la información en una matriz

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix}.$$

Análogamente a la media y varianza muestrales $\overline{x}$ y $s^2$, podemos definir sus contrapartes multivariadas como:

$$\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i,$$

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top.$$

que representan el vector de medias y la matriz de covarianza, respectivamente.

*Observación:*

En este caso, tenemos $\boldsymbol{S} = (s_{ij})$, donde

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j),$$

con $\overline{x}_i = \big(\sum_{k=1}^{n} x_{ki}\big)/n$.

Los elementos anteriores permiten definir la matriz de correlación entre las $p$ variables, como:

$$\boldsymbol{R} = (r_{ij})$$

donde

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2 \sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}.$$

*Observación:*

Defina $\boldsymbol{D} = \operatorname{diag}(s_{11}, s_{22}, \dots, s_{pp})$, de este modo, podemos definir

$$\boldsymbol{R} = \boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}.$$

# Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Datos observados:

Mediciones (cm) del largo y ancho de los sépalos y el largo y ancho de pétalos para 50 flores desde 3 especies de Iris (setosa, virginica y versicolor).

Base de datos:

```
# Datos de flores Iris
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
1            5.1         3.5          1.4         0.2    setosa
2            4.9         3.0          1.4         0.2    setosa
3            4.7         3.2          1.3         0.2    setosa
4            4.6         3.1          1.5         0.2    setosa
5            5.0         3.6          1.4         0.2    setosa
...

149          6.2         3.4          5.4         2.3 virginica
150          5.9         3.0          5.1         1.8 virginica
```

# Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Datos observados:

Mediciones (cm) del largo y ancho de los sépalos y el largo y ancho de pétalos para 50 flores desde 3 especies de Iris (setosa, virginica y versicolor).

Base de datos:

```
# Datos de flores Iris
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
1            5.1         3.5          1.4         0.2    setosa
2            4.9         3.0          1.4         0.2    setosa
3            4.7         3.2          1.3         0.2    setosa
4            4.6         3.1          1.5         0.2    setosa
5            5.0         3.6          1.4         0.2    setosa
...

149          6.2         3.4          5.4         2.3 virginica
150          5.9         3.0          5.1         1.8 virginica
```

# Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Matriz de Correlación ($\boldsymbol{R}$):

| | Largo Sépalo | Ancho Sépalo | Largo Pétalo | Ancho Pétalo |
|---|---|---|---|---|
| Largo Sépalo | 1.000 | -0.118 | 0.872 | 0.818 |
| Ancho Sépalo | -0.118 | 1.000 | -0.428 | -0.366 |
| Largo Pétalo | 0.872 | -0.428 | 1.000 | 0.963 |
| Ancho Pétalo | 0.818 | -0.366 | 0.963 | 1.000 |

Cálculo en **R**:

```
> cor(iris[,1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length       1.0000     -0.1176       0.8718      0.8179
Sepal.Width       -0.1176      1.0000      -0.4284     -0.3661
Petal.Length       0.8718     -0.4284       1.0000      0.9629
Petal.Width        0.8179     -0.3661       0.9629      1.0000
```

# Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Matriz de Correlación ($\boldsymbol{R}$):

| | Largo Sépalo | Ancho Sépalo | Largo Pétalo | Ancho Pétalo |
|---|---|---|---|---|
| Largo Sépalo | 1.000 | -0.118 | 0.872 | 0.818 |
| Ancho Sépalo | -0.118 | 1.000 | -0.428 | -0.366 |
| Largo Pétalo | 0.872 | -0.428 | 1.000 | 0.963 |
| Ancho Pétalo | 0.818 | -0.366 | 0.963 | 1.000 |

Cálculo en **R**:

```
> cor(iris[,1:4])
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length       1.0000     -0.1176       0.8718      0.8179
Sepal.Width       -0.1176      1.0000      -0.4284     -0.3661
Petal.Length       0.8718     -0.4284       1.0000      0.9629
Petal.Width        0.8179     -0.3661       0.9629      1.0000
```

# Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Se obtuvo además el vector de medias ($\overline{\boldsymbol{x}}$) y la matriz de Covarianza ($\boldsymbol{S}$):[2]

```
> z <- cov.wt(iris[,1:4])
> z
$cov
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    0.6856935  -0.0424340    1.2743154   0.5162707
Sepal.Width    -0.0424340   0.1899794   -0.3296564  -0.1216394
Petal.Length    1.2743154  -0.3296564    3.1162779   1.2956094
Petal.Width     0.5162707  -0.1216394    1.2956094   0.5810063

$center
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
    5.843333     3.057333     3.758000     1.199333

$n.obs
[1] 150
```

[2] Análogamente podemos usar `cov(iris[,1:4])`.