# Análisis de regresión

17 de agosto
(semana 2)

# Plan de trabajo

1. Estudio de pares de variables cualitativa-cuantitativa
2. Estudio de pares de variables cuantitativas
3. Uso de R para el modelo ANOVA a una vía

**Nota:**

- Ya está disponible el taller 1.
- Mismo monitor: Jesús David Castro, jecastroa@unal.edu.co. En correos poner asunto: "DUDA REGRESION" al inicio del asunto.
- **No olviden mi horario de atención**: miérc./viernes de 11 a 12:30pm en mi oficina (325-404). Martes 10-12 virtual (con cita previa).
- La primera entrega del trabajo final será para el domingo 4 de septiembre (ya está en Drive).
- En la semana del 29 de agosto al 2 de septiembre no habrá clase. Pendientes de una actividad a desarrollar.

2 var. cualitativas:

Exploratoria: Tablas de contingencia

Inferencial: Prueba Chi-cuadrado

$$\begin{cases} H_0: \text{las variables son independ.} \\ H_1: \text{" " no son independ} \end{cases}$$

$$J_n = \frac{\sum\sum (O_{ij} - E_{ij})^2}{E_{ij}}$$

De ahora en adelante, no vamos a encontrar un UMP

$\tau$: "Rechazar Ho si $j_n > \chi^2_{\alpha}((r-1)(c-1))$

Ingredientes para una prueba de hipótesis

1. Estadística cuya distribución bajo Ho es conocida

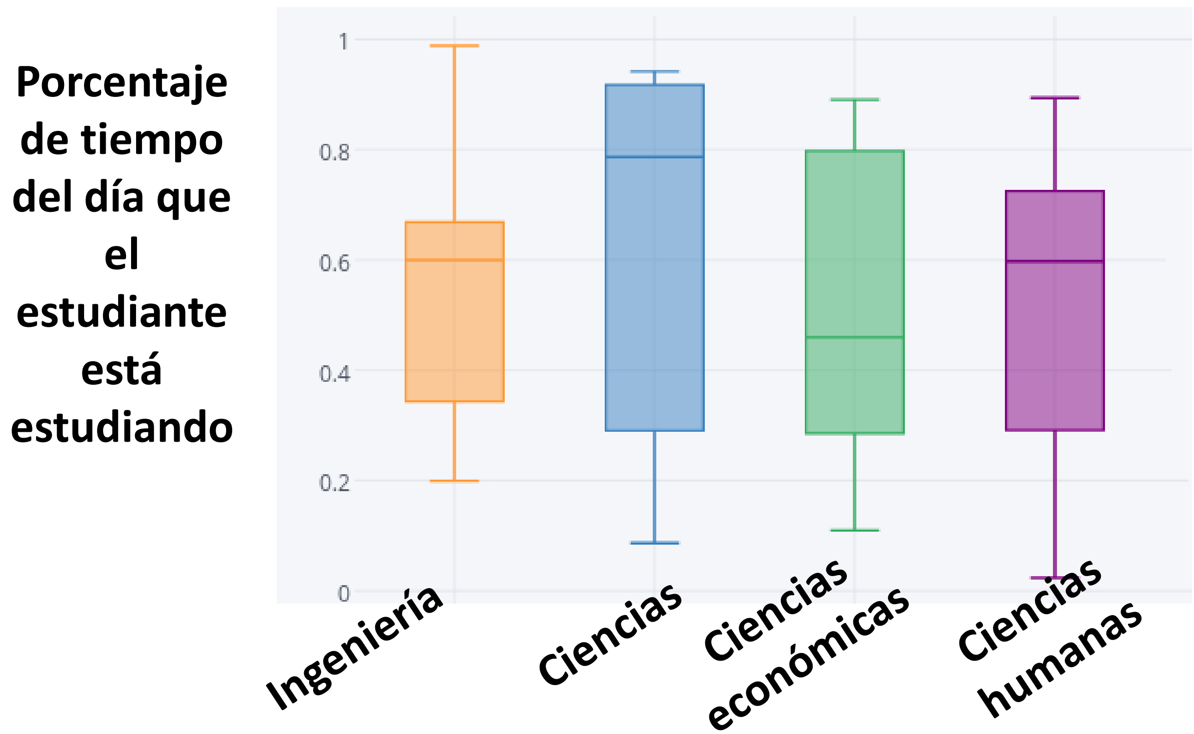
$$J_n \xrightarrow[n\to\infty]{H_0} \chi^2((r-1)(c-1))$$

2. Qué evidencia de esa estadística me permite rechazar Ho

# Relación entre una variable cualitativa y una cuantitativa

# Exploración (1 v. cuantitativa y 1 v. cualitativa)

- Las categorías de la variable cualitativa dividen al conjunto de análisis en **clases o subgrupos**.
- Se hace un diagrama (histograma o boxplot) para los valores de la variable cuantitativa para cada clase o subgrupo y se comparan entre sí (medidas de tendencia central, dispersión, simetría, curtosis).




Fuente: https://stackoverflow.com/questions/53767621/boxplot-with-pandas-in-python

ANOVA: Analysis Of Variance

ANAVA: Análisis de varianza

# ANOVA one way o de un factor

- **Factor:** Variable cualitativa que segmenta a la población de estudio.
- **Nivel:** Cada valor diferente que se considera del factor.
- **Tratamiento:** combinación de niveles de diferentes factores a los que se somete un grupo de unidades.
- En el ejemplo anterior, el factor es la facultad y los niveles correspondientes son las facultades de ingeniería, de ciencias, de ciencias económicas, y de ciencias humanas.

- Sea $Y_{ij}$ : "Variable aleatoria correspondiente a la respuesta del $j$-ésimo individuo dentro del $i$-ésimo nivel".

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \ \forall j = 1, 2, \ldots, n_i; \ i = 1, 2, \ldots, k$$

**Gran media:** valor medio que se tendría si no se sometiese a ningún tratamiento

**Efecto debido al i-ésimo nivel**

**Término de error de dicha observación**

$Y_{ij}$: Tiempo de estudio del j-ésimo individuo de la i-ésima facultad

$$Y_{ij} \sim N(\mu+\tau_i, \sigma^2)$$

$$Y_{ij} = \underbrace{\mu}_{\text{Media o esperanza general}} + \underbrace{\tau_i}_{\text{Cambio media por ser de la facultad } i} + \underbrace{\varepsilon_{ij}}_{\text{error asociado al individuo } j \text{ del grupo } i}, \qquad \varepsilon_{ij} \overset{iid}{\sim} N(0,\sigma^2)$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n_i} Y_{ij}}{N} \qquad \hat{\tau}_i = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} - \hat{\mu}$$

# Supuestos del modelo ANOVA

## *Supuestos del modelo ANOVA*

- $N := \sum_{i=1}^{k} n_i$ (Total muestra)
- $\sum_{i=1}^{k} \tau_i = 0$ (Modelo sobredeterminado)
- $\{\varepsilon_{ij}\}$ son una m.a. $N\left(0, \sigma^2\right)$
  - × Las $\{Y_{ij}\}$ no son una m.a.
  - × $Y_{ij} \sim N\left(\mu + \tau_i, \sigma^2\right), \forall i, j.$
  - × Hay igual varianza entre los grupos (homoscedasticidad)

| Niv.1 | Niv.2 | | Niv.$k$ |
|---|---|---|---|
| $Y_{11}$ | $Y_{21}$ | | $Y_{k1}$ |
| $Y_{12}$ | $Y_{22}$ | | $\vdots$ |
| $\vdots$ | $Y_{23}$ | $\cdots$ | $Y_{kn_k}$ |
| $Y_{1n_1}$ | $\vdots$ | | |
| | $Y_{2n_2}$ | | |

# Teorema de descomposición de varianza

Si se definen:

$$Y_{..} := \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij},\ \bar{Y}_{..} = \frac{Y_{..}}{N};\ Y_{i.} := \sum_{j=1}^{n_i} Y_{ij} \text{ y } \bar{Y}_{i.} = \frac{Y_{i.}}{n_i}, \forall i.$$

Entonces:

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{..}\right)^2}_{SC_{total}} = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i.}\right)^2}_{SC_{error} \text{ o } SC_{dentro}} + \underbrace{\sum_{i=1}^{k} n_i\left(\bar{Y}_{i.}-\bar{Y}_{..}\right)^2}_{SC_{trat} \text{ o } SC_{entre}}.$$

Si $Y_{ij}$ es una m.a

$SC_{TOTAL} \approx SC_{DENTRO}$

Además,

$$SC_{total} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^{\,2} - \frac{Y_{..}^{\,2}}{N} \text{ y}$$

$$SC_{trat} = \sum_{i=1}^{k} n_i\left(\bar{Y}_{i.}-\bar{Y}_{..}\right)^2 = \sum_{i=1}^{k}\frac{Y_{i.}^{\,2}}{n_i} - \frac{Y_{..}^{\,2}}{N}$$

Si hay efecto de los tratam

$SC_{DENTRO}$ (pequeño)

$SC_{entre}$ (grande)

# Idea de la prueba

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{..}\right)^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i.}-\left[\bar{Y}_{..}-\bar{Y}_{i.}\right]\right)^2$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[\left(Y_{ij}-\bar{Y}_{i.}\right)^2-2\left(Y_{ij}-\bar{Y}_{i.}\right)\left(\bar{Y}_{..}-\bar{Y}_{i.}\right)+\left(\bar{Y}_{..}-\bar{Y}_{i.}\right)^2\right]$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i.}\right)^2-2\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i.}\right)\left(\bar{Y}_{..}-\bar{Y}_{i.}\right)+\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\bar{Y}_{..}-\bar{Y}_{i.}\right)^2$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i.}\right)^2-2\sum_{i=1}^{k}\left(\bar{Y}_{..}-\bar{Y}_{i.}\right)\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i.}\right)+\sum_{i=1}^{k}n_i\left(\bar{Y}_{..}-\bar{Y}_{i.}\right)^2$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i.}\right)^2-2\sum_{i=1}^{k}\left(\bar{Y}_{..}-\bar{Y}_{i.}\right)\underbrace{\left\{n_i\bar{Y}_{i.}-n_i\bar{Y}_{i.}\right\}}_{0}+\sum_{i=1}^{k}n_i\left(\bar{Y}_{..}-\bar{Y}_{i.}\right)^2$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i.}\right)^2+\sum_{i=1}^{k}n_i\left(\bar{Y}_{..}-\bar{Y}_{i.}\right)^2$$

# Hipótesis de ANOVA

Considere el sistema:

$$\begin{cases} H_0 : \tau_1 = \tau_2 = \ldots = \tau_k = 0 \\ \textit{versus} \\ H_1 : \exists \tau_r \neq 0 \end{cases}$$

→ $\{Y_{ij}\}$ son m.a.s

→ Var. cualit. $\underline{\underline{\text{no}}}$ afecta a var. cuantit

## *Teorema:*

Bajo $H_0$, se tiene que:

$$\frac{SC_{error}}{\sigma^2} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2}{\sigma^2} \sim \chi^2\left(N-k\right), \text{ y}$$

$$\frac{SC_{trat}}{\sigma^2} = \frac{\sum_{i=1}^{k} n_i\left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2}{\sigma^2} \sim \chi^2\left(k-1\right); \text{ siendo v. a. independientes.}$$

# Idea de la prueba (I)

$S_i^2 = \frac{\sum_{j=1}^{n_i} \left( Y_{ij} - \overline{Y}_{i\cdot} \right)^2}{n_i - 1}$ es un estimador de la varianza del nivel i $\left( \sigma^2 \right)$,

luego, $\frac{\left( n_i - 1 \right) S_i^2}{\sigma^2} \sim \chi^2 \left( n_i - 1 \right), \ \forall i.$

Como las muestras en cada nivel son independientes de los demás niveles,

$$\sum_{i=1}^{k} \frac{\left( n_i - 1 \right) S_i^2}{\sigma^2} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left( Y_{ij} - \overline{Y}_{i\cdot} \right)^2}{\sigma^2} = \frac{SC_{error}}{\sigma^2} \sim \chi^2 \left( \sum_{i=1}^{k} \{ n_i - 1 \} \right) \overset{d}{=} \chi^2 \left( N - k \right)$$

# Idea de la prueba (II)

Bajo $H_0$, **no** hay diferencia en el valor medio de los niveles, así que $\bar{Y}_{..}$ es el estimador de la media común de todos, luego

$S^2 = \frac{SC_{total}}{N-1} = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{..}\right)^2$ es el estimador insesgado de $\sigma^2$ usando toda la muestra, luego $\frac{(N-1)}{\sigma^2}S^2 = \frac{SC_{total}}{\sigma^2} \sim \chi^2(N-1)$.

Finalmente, $\frac{SC_{trat}}{\sigma^2} = \frac{SC_{total}}{\sigma^2} - \frac{SC_{error}}{\sigma^2}$; y se puede probar que $SC_{trat} \perp SC_{error}$ (lo haremos más adelante), por ende, $\frac{SC_{trat}}{\sigma^2} \sim \chi^2(k-1)$.

# Prueba F de una ANOVA

## *Prueba F de una ANOVA*

Considere el sistema: $\begin{cases} H_0 : \tau_1 = \tau_2 = \ldots = \tau_k = 0 \\ versus \\ H_1 : \exists \tau_r \neq 0 \end{cases}$.

Si se define $CM_{trat} = SC_{trat} \big/ k-1$ y $CM_{error} = SC_{error} \big/ \cancel{k-1}$ N-k

$$F_C := \frac{CM_{trat}}{CM_{error}} \overset{H_0 \text{ cierta}}{\sim} F\left(k-1, N-k\right).$$

El test $\tau$:"Rechazar $H_0$ si $f_C > F_{1-\alpha}\left(k-1, N-k\right)$" es un test del $100 \cdot \alpha\%$ de significancia.

El p-valor de este test es $p-value = p\left[F_C > f_C \middle| H_0\right]$.

# Resumen de ANOVA

| Fuente de variabilidad | Suma de cuadrados | Grados de libertad | Cuadrado medio | $F_0$ | Valor-$p$ |
|---|---|---|---|---|---|
| Tratamientos | $SC_{TRAT} = \sum_{i=1}^{k} \frac{Y_{i\bullet}^2}{n_i} - \frac{Y_{\bullet\bullet}^2}{N}$ | $k-1$ | $CM_{TRAT} = \frac{SC_{TRAT}}{k-1}$ | $\frac{CM_{TRAT}}{CM_E}$ | $P(F > F_0)$ |
| Error | $SC_E = SC_T - SC_{TRAT}$ | $N-k$ | $CM_E = \frac{SC_E}{N-k}$ | | |
| Total | $SC_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}^2 - \frac{Y_{\bullet\bullet}^2}{N}$ | $N-1$ | | | |

Fuente: Análisis y diseño de experimentos (Gutierrez et al.)

## *Porcentaje de varianza explicado*

$R^2 := \frac{SC_{trat}}{SC_{total}}$ representa el porcentaje de varianza explicado por el modelo y es una manera de ver qué tan fuerte es la relación.

# ¿Qué pasa si se rechaza la hipótesis nula?

- Quiere decir que el efecto de algún nivel del factor es significativo e influye sobre la media de las variables, ¿pero cuál?

- En otras palabras, se desea evaluar el sistema:

$$\begin{cases} H_0 : \mu_j - \mu_l = 0 \ \left(\tau_j - \tau_l = 0\right) \\ versus \\ H_1 : \ \mu_j - \mu_l \neq 0 \ \left(\tau_j - \tau_l \neq 0\right) \end{cases}, \ \forall j,l \text{ con } j \neq l$$

$$\text{Test: "Rechazar } H_0 \text{ si } \left|\bar{y}_{j\cdot} - \bar{y}_{l\cdot}\right| > t_{1-\frac{\alpha^*}{2}}\left(N-k\right) \cdot \sqrt{cm_{error}\left(\frac{1}{n_j} + \frac{1}{n_l}\right)}\text{"}$$

Cada test es un test de diferencia de medias en muestras independientes.

Sin embargo, hay $\dfrac{k\left(k-1\right)}{2}$ diferencias a considerar, así que es necesario ajustar la significancia global.

# ¿Qué pasa si se rechaza la hipótesis nula?(II)

- Hay varios métodos para hacer esa cantidad de pruebas de hipótesis controlando la Family-wise error rate (FWER) o la significancia global.

  - Método de Bonferroni (Lo vimos en clase. Es muy conservador).
  - Método de Tukey
  - Método de Scheffé
  - Método de Benjamini-Hochberg

# Validación del modelo

***Residuales***

Los residuales del modelo, $\{r_{ij}\}$ se definen como los componentes no explicados por el modelo. Es decir:

$$r_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i\cdot}$$

- Aunque no es posible considerar que los residuales tienen exactamente el mismo comportamiento probabilístico que los errores del modelo, se espera que ellos den indicios sobre las propiedades deseables de los errores.

# Validación del modelo (II)

- Validación de la homoscedasticidad (homogeneidad de varianzas)
  - Método gráfico (graficar los residuales versus las predicciones).
  - Prueba de hipótesis de Bartlett o de Levene.
  - Si la hipótesis de homoscedasticidad no se verifica, es necesario usar el modelo de Welch.

- Validación de la independencia (no correlación serial)
  - Método gráfico (graficar los residuales versus el orden temporal, si lo hay).

- Validación de la distribución normal con media cero
  - QQ plots
  - Prueba de normalidad

- TODOS los supuestos del modelo se deben verificar, si no, es necesario reajustar el modelo o usar otro.

# Algunas observaciones finales

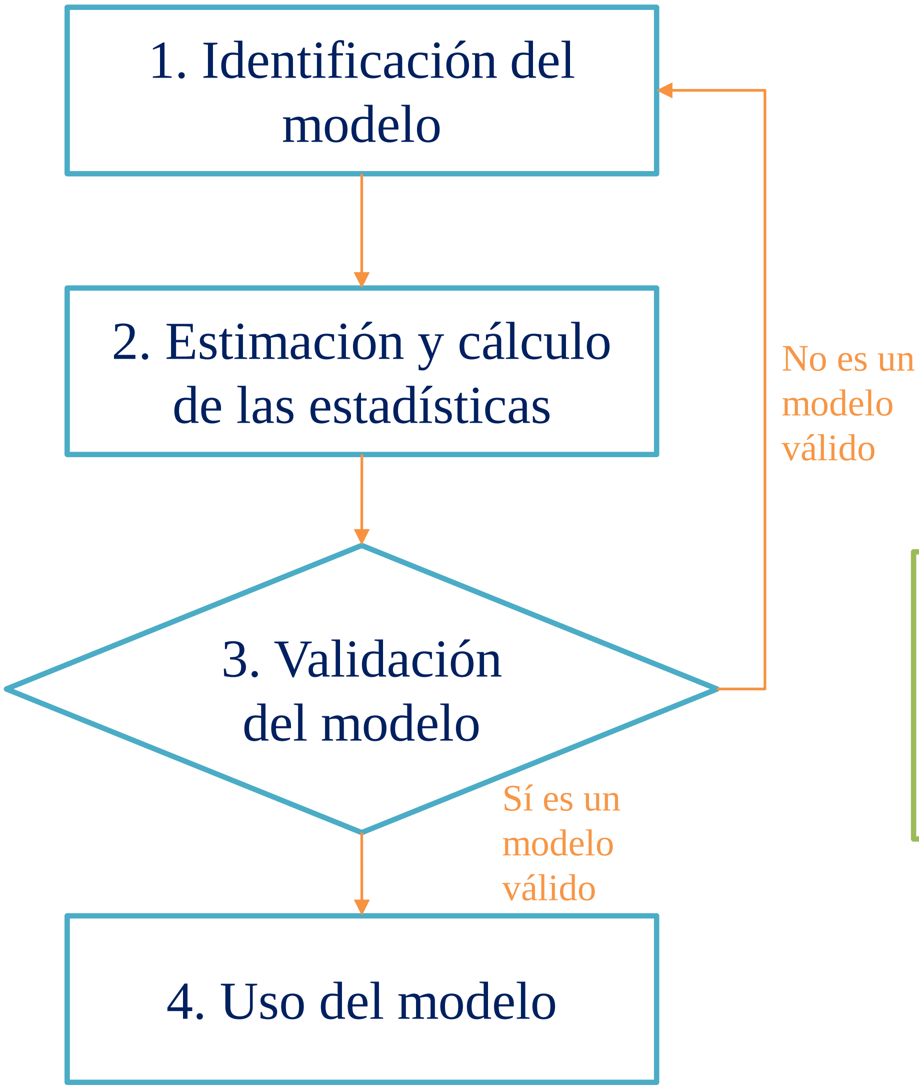
- La prueba ANOVA se basa en la idea de un diseño experimental completamente al azar. Es robusta a la ausencia de normalidad, pero no tanto a la heteroscedasticidad.

- Si los supuestos no se tienen, es necesario replantear el modelo:
  - Si no hay homoscedasticidad, usar la prueba de Welch.
  - Si no hay normalidad ni homoscedasticidad, se puede transformar los datos (transformaciones estabilizadoras de varianza) o
  - Usar la prueba de Kruskal-Wallis, o
  - Usar remuestreo.

- El modelo ANOVA se puede extender a 2 (o más) factores.

$$Y_{ijl} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijl}, \ \forall l = 1, \ldots, n_{ij}; \ i = 1, \ldots, k_1; j = 1, \ldots, k_2$$

# Algunas observaciones finales (II)




**Validación interna:**
Cumplimiento supuestos
**Validación externa:**
Representación adecuada del fenómeno

Tarea: Ajustar una distrib. a los datos

1. Identificación — Histograma o un diag. de caja

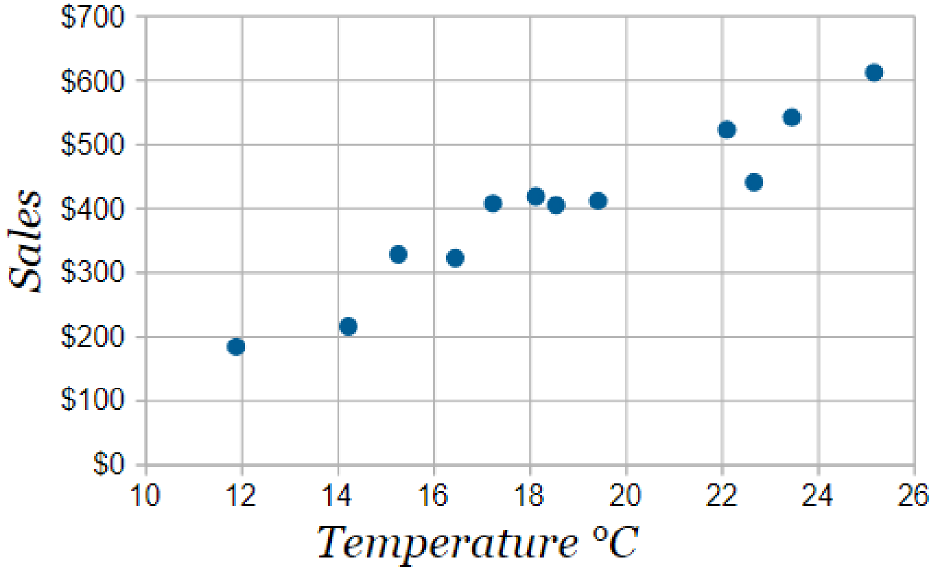
2. Estimación — Encontrar las estimaciones para los parámetros
↓
3. Validación* — Prueba de bondad de ajuste
↓ sí
4. Uso

No

# Relación entre dos variables cuantitativas

# Diagrama de dispersión y coeficiente de correlación

- Vimos en estadística descriptiva que el diagrama de dispersión nos permitía visualizar si dos variables cuantitativas tenían o no una relación entre ellas.



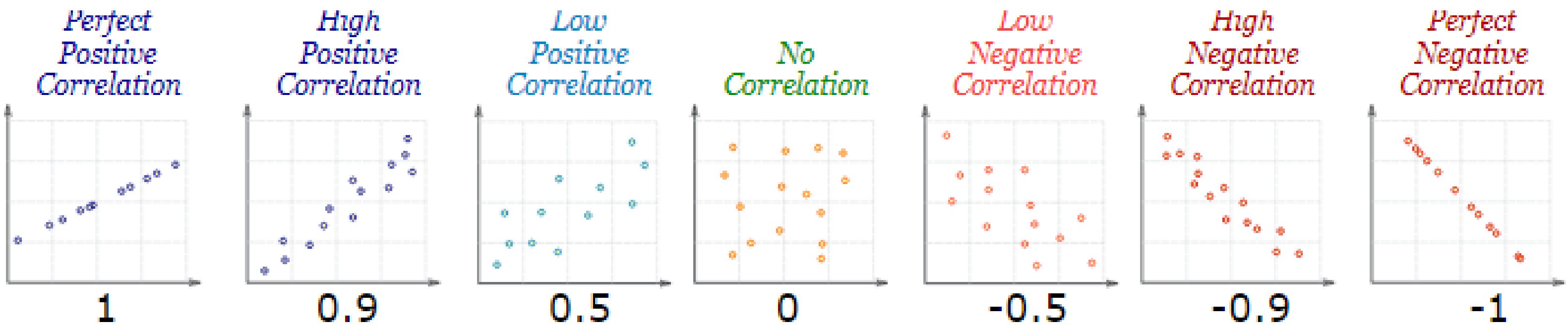

Fuente: https://www.mathsisfun.com/data/scatter-xy-plots.html

- Pero, ¿cómo podemos cuantificar si una relación entre dos variables es fuerte o débil? Para ello, utilizamos el **coeficiente de correlación lineal** de Pearson.
- Recuerden que relación **NO** necesariamente implica causalidad entre las dos variables.

# Diagrama de dispersión y coeficiente de correlación

## *Coeficiente de correlación lineal de Pearson*

Es un parámetro que mide qué tan fuerte es la relación lineal entre dos variables. Toma valores entre -1 y 1, donde -1 indica una relación lineal inversa y 1 indica una perfecta relación lineal directa.




Fuente: https://www.mathsisfun.com/data/correlation.html

- ¡Ojo! El coeficiente de correlación solo mide relaciones lineales. Si las dos variables están relacionadas de manera no lineal, es posible que este coeficiente no detecte esa relación.
- [http://guessthecorrelation.com/](http://guessthecorrelation.com/)