

Unidad 1

# Introducción

- La humanidad siempre ha intentado explicar el mundo en que vivimos.
- Los primeros intentos atribuían los fenómenos físicos a causas sobrenaturales (magia y religión).
- Con el desarrollo de la ciencia, se lograron explicaciones objetivas y más precisas.
- Explicar la realidad en que vivimos es un proceso que aún continúa...

En general, se puede representar a los fenómenos que ocurren en la naturaleza como sigue:



La representación anterior es conocida como *modelo*.

*Un modelo es una abstracción que simplifica la realidad y permite estudiarla con mayor detalle.*

# Tipos de fenómenos

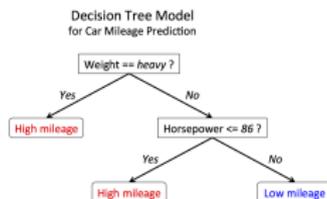
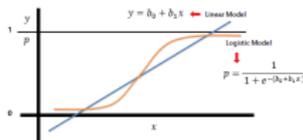
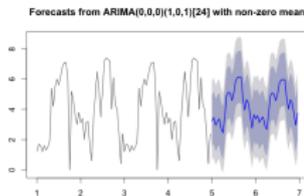
Existen dos corrientes científicas para describir cómo ocurren los fenómenos:

- **Determinismo:** las condiciones en las que se da el fenómeno determinan completamente su resultado.
- **Indeterminismo:** se admite la existencia de fenómenos en los que no es posible saber de antemano cuál será su resultado, aún controlando las condiciones en las que se desarrolla.

## P. S. Laplace

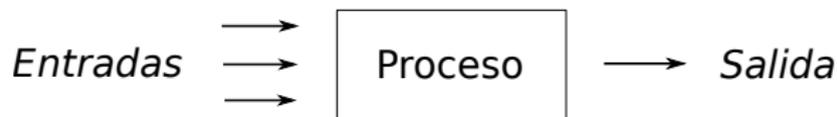
*Si fuera posible conocer todas las leyes que rigen todos los procesos que ocurren en el universo y el estado actual de todos los objetos que lo componen además, si fuera posible analizar tales datos, se podría abarcar en una sola fórmula los movimientos de todos los objetos en el universo; nada resultaría incierto y tanto el futuro como el pasado estarían presentes ante nuestros ojos.*

*Desarrollo de expresiones matemáticas para describir, en algún sentido, la relación que existe entre un conjunto de variables que describen las condiciones en que se desarrolla un fenómeno y una variable asociada su resultado.*



# Modelos deterministas

En el caso determinista, las entradas determinan completamente la salida:



Si  $x_1, \dots, x_p$  son las entradas y  $y$  la salida o respuesta, podemos modelar matemáticamente el proceso como:

$$y = f(x_1, \dots, x_p)$$

El objetivo es determinar una buena elección para  $f$ , se pueden utilizar distintos criterios para elegir a un candidato.

## Ejemplo. Caída libre

Cuenta la historia que Galileo Galilei descubrió experimentalmente que la distancia que recorre un cuerpo en caída libre, despreciando la resistencia del aire, está dada por:

$$d(t) = \frac{1}{2}gt^2$$

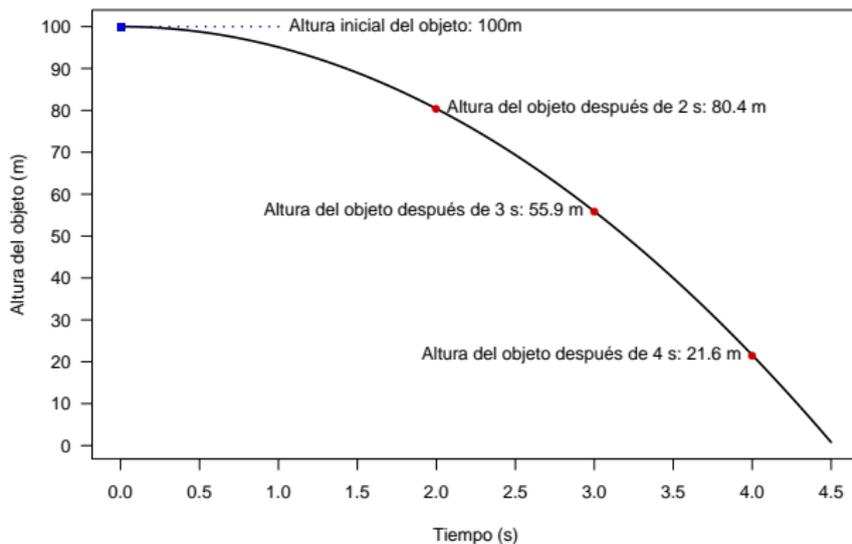
donde:

- $t$  tiempo transcurrido,
- $g = 9.8ms^{-2}$  (aceleración de la gravedad),
- $d(t)$  distancia recorrida al tiempo  $t$ .

# Caída libre (cont.)

## Ejemplo

Se deja caer un objeto desde 100 m altura. ¿Cuál será la altura del objeto después de  $t = 2, 3, 4$  s?



# Crecimiento poblacional

- El crecimiento exponencial se puede utilizar para modelar el crecimiento o decrecimiento de una población.
- Si el crecimiento/decrecimiento se da a una tasa proporcional al tamaño de la población, entonces se puede modelar el tamaño de la población al tiempo  $t$  como

$$N(t) = N_0 \exp^{-\lambda t}$$

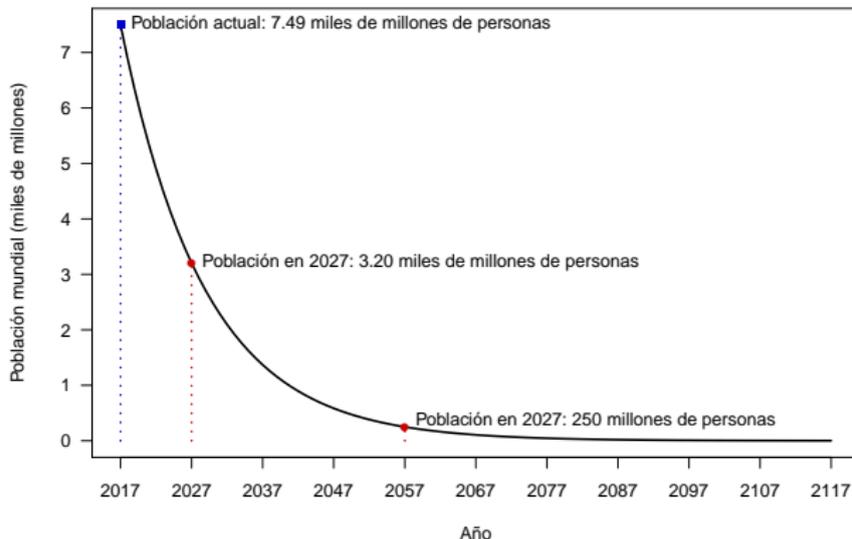
donde:

- $N_0$  es la población inicial (tiempo  $t = 0$ ),
- $t$  es el tiempo transcurrido,
- $\lambda$  es la tasa de crecimiento/decrecimiento,
- $N(t)$  es la población al tiempo  $t$ .

# Crecimiento poblacional (cont.)

## Ejemplo

La población de mundial el 02/02/2017 es aproximadamente 7,494 millones de personas. Si en la población se detuvieran los nacimientos y la tasa de mortalidad se mantuviera constante en 8.5% anual:



# Modelos Estocásticos

En el caso estocástico, el objetivo es modelar matemáticamente la relación entre las condiciones en que se desarrolla un fenómeno y la incertidumbre de sus posibles respuestas.



Si  $x_1, \dots, x_p$  son las entradas y  $y$  la salida o respuesta, podemos modelar matemáticamente el proceso como:

$$y \sim F(x_1, \dots, x_p)$$

El objetivo es encontrar un buen modelo probabilista  $F$ , estamos ante un problema típico de inferencia estadística.

## Ejemplo. El estado del tiempo

Es posible modelar el estado del tiempo como una *cadena de Markov*. Consideremos tres estados posibles  $S = \{1, 2, 3\}$  donde: 1 es soleado, 2 es nublado, 3 es lluvioso. Suponemos que la matriz de transición  $P$  está dada por:

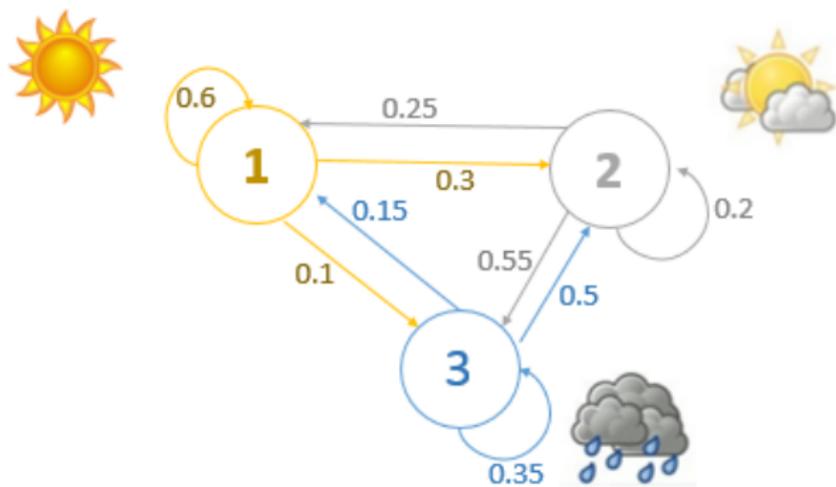
$$P = \begin{bmatrix} 0.60 & 0.30 & 0.10 \\ 0.25 & 0.20 & 0.55 \\ 0.15 & 0.50 & 0.35 \end{bmatrix}$$

La  $i, j$ -ésima entrada de  $P$  representa la probabilidad de que el tiempo de mañana sea  $j$  dado que hoy es  $i$ ,  $i, j = 1, 2, 3$ . Por ejemplo:

$$P(\text{soleado}|\text{soleado}) = 0.60$$

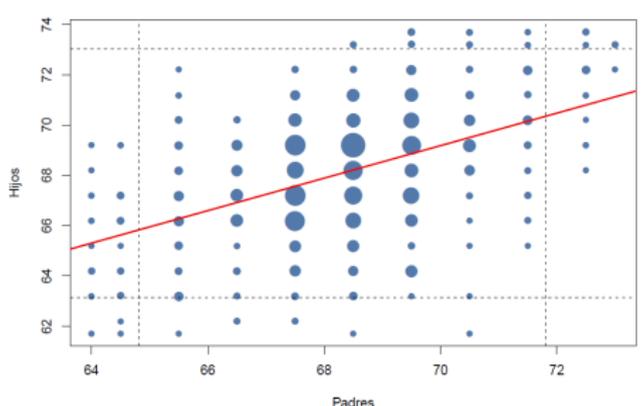
$$P(\text{nublado}|\text{lluvioso}) = 0.50$$

$$P(\text{lluvioso}|\text{soleado}) = 0.10$$



# Ejemplo. Regresión de Galton

*Francis Galton en el siglo XIX observó la relación que había entre la altura de padres e hijos, notó que aunque existe una tendencia de que un padre alto tenga un hijo alto y padre bajo tenga un hijo bajo, la distribución de las estaturas, no cambia drásticamente de una generación a otra (a este fenómeno se le conoce como regresión a la media).*



Los modelos matemáticos se usan para describir la relación funcional entre las variables de entrada  $(x_1, x_2, \dots, x_p)$  y la variable respuesta  $y$ .

$$Y = f(x_1, x_2, \dots, x_p) \quad (\text{Determinista})$$

$$Y \sim F(x_1, x_2, \dots, x_p) \quad (\text{Probabilístico})$$

- ¿Qué función  $f$  (o  $F$ ) tomar?
- ¿El modelo es adecuado? ¿Existe el mejor modelo?
- ¿Qué supuestos debemos tomar en cuenta?
- ¿Las variables  $x_i$  explican a  $y$ ?

# Los modelos de regresión

Nuestro interés se centrará en los modelos probabilistas y en específico en el modelo de regresión lineal.

$$E(Y|\mathbf{X}) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

Donde  $\mathbf{X} = (X_1, \dots, X_p)$ .

- El término *regresión* viene la regresión a la media del ejemplo de Galton. El término *lineal* se refiere a que la forma funcional del modelo es lineal en los parámetros  $\beta$ .
- Cuando sólo se tiene una variable de entrada,  $p = 1$ , el modelo es de regresión lineal simple, cuando hay más de una variable de entrada el modelo es de regresión lineal múltiple.

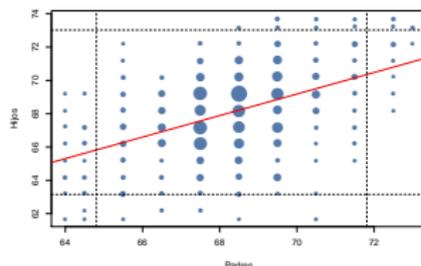
# El modelo de regresión lineal simple

En el ejemplo de Galton es posible utilizar un modelo de regresión lineal simple para explicar las alturas de los hijos a partir de las alturas de los padres.

$$E(Y|X) = \beta_0 + \beta_1 X$$

donde:

- $Y$  es la altura del hijo,
- $X$  es la altura del padre,
- $\beta_0$  y  $\beta_1$  son los parámetros del modelo.



## El modelo de regresión lineal simple (cont.)

- Podemos notar como para un valor dado de  $X$ , no hay un valor único de  $Y$ , sino que  $Y$  varía alrededor de su valor promedio y lo hace aleatoriamente.
- Denotamos por  $\epsilon$  al error estadístico que representa la desviación de  $Y$  de su media, para un valor de  $X$  fijo.

Entonces podemos representar el modelo de regresión lineal simple como sigue

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Lo usual es asumir  $\epsilon \sim (0, \sigma^2)$ , de manera que

$$Y|X \sim (\beta_0 + \beta_1 X, \sigma^2).$$

- El supuesto anterior es muy fuerte dado que asume una varianza constante para cualquier valor de  $X$ .

# Interpretación del modelo

- El objetivo es hacer inferencias sobre los parámetros desconocidos  $\beta_0$ ,  $\beta_1$  y  $\sigma^2$ .
- Para un valor de  $X$  fijo, el valor esperado de  $Y$  es  $\beta_0 + \beta_1 X$ .
- El parámetro  $\beta_1$  representa el incremento en el valor esperado de  $Y$  por un incremento unitario en  $X$ .
- Para un valor de  $X$  fijo,  $\sigma^2$  es la variabilidad de  $Y$  alrededor de su valor esperado, es decir,  $\sigma^2$  es la varianza de  $Y$  dado  $X$ .
- ¿Qué hay de la interpretación de  $\beta_0$ ? Las ecuaciones de regresión solamente son válidas dentro del rango de los valores observados en las entradas.

# Aplicaciones de los modelos de regresión

- 1 Las ventas de un producto pueden ser predichas a partir del gasto en publicidad.
- 2 El desempeño de un trabajador en un empleo puede ser predicho a partir de las respuestas de una prueba de aptitudes.
- 3 El tamaño del vocabulario de un niño puede ser predicho a partir de la edad del niño y el grado de escolaridad de sus padres.
- 4 El salario de un trabajador puede ser predicho a partir de su edad, escolaridad y sector de ocupación.
- 5 El volumen de madera de un árbol puede ser determinado a partir de variables como el diámetro del tronco y la altura.

## George Box

*Essentially, all models are wrong, but some are useful.*

