Argentina programa 4.0

Universidad Nacional de San Martín

Módulo 2 - Ciencia de Datos

Semana 1. Elementos de matemática y probabilidad

Clase anterior:

- Funciones (por lo general, contínuas)
- Pendiente, máximos y mínimos
- Vectores, Matrices
- ¿Qué usos tienen en ciencia de datos e inteligencia artificial?
- ¿Cómo esas cosas se conectan?

Elementos de matemática y probabilidad

Un preludio a la ciencia de datos y al aprendizaje automático

- ¿Qué es un modelo?
- ¿Cómo encontrar un modelo que representa un conjunto de datos?
- Probabilidad (y modelos)
- Distribuciones de probabilidad, percentiles, etc.
- Probabilidad condicional

Datos y modelos

Motivación:

- **La IA se ocupa esencialmente de los datos: aprender de los datos, predecir a partir de los datos, generar nuevos datos, etc.**
- **Modelo: descripción matemática de los datos (fundamentales o no)**
- **Conceptos/operaciones matemáticas fundamentales:**
    - Funciones continuas (cálculo)
    - Álgebra lineal (matrices, etc.)
    - Probabilidad
- **Estas nociones son importantes para comprender la ciencia de datos en general y el aprendizaje automático**

Argentina programa 4.0

¿Qué es un modelo?

¿Cómo encuentro un modelo que representa datos?

Universidad Nacional de San Martín

Escuela de Ciencia y Tecnología ECyT_UNSAM

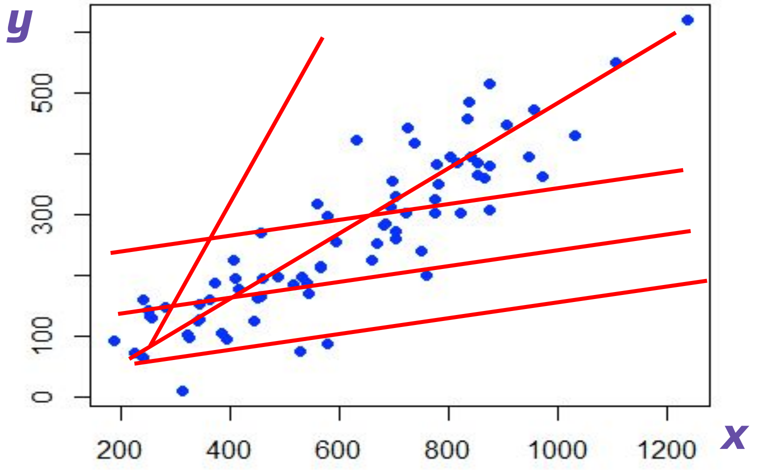
Secretaría de Economía del Conocimiento

Datos y modelos

- **Normalmente los datos son discretos:**
  - Ejemplo: pares de números
    pueden visualizarse como un gráfico de dispersión
- **Frecuentemente los modelos vienen dados por funciones contínuas**
  - Ej.: función continua de una variable ***f(x)***
  - Puede representarse en un gráfico ***y = f(x)***

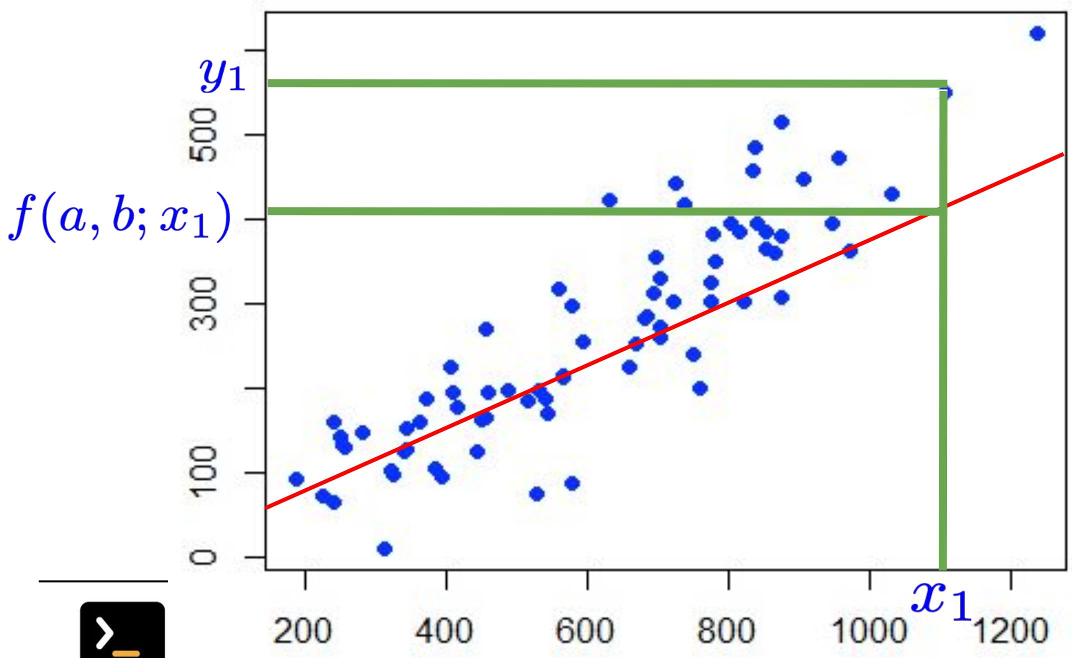
    variable
  - Ejemplo de modelo: **y = a x + b**

    parámetros
  - Parámetros determinados a partir de los datos



Ajustando una función a los datos

- Encontrar un modelo (función) que represente (aproximadamente) los datos
- Minimizar la distancia de un modelo a los datos



En este caso

$$f(a,b;x) = ax + b$$

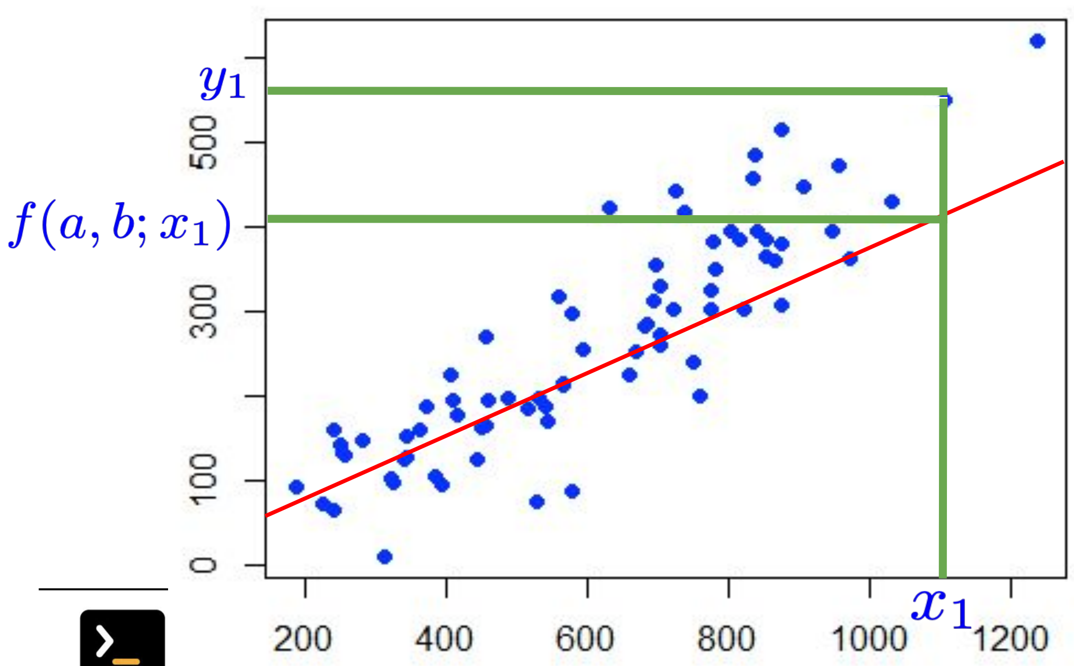
Distancia entre la **previsión del modelo** y el dato:

$$f(a,b;x_1) - y_1$$

$$= ax_1 + b - y_1$$

Ajustando un modelo a los datos

- Encontrar un modelo (función) que represente (aproximadamente) los datos
- Minimizar la distancia de un modelo a los datos



Distancia **cuadrática** entre todos los puntos y la **predicción:**

$$
\begin{aligned}
&(f(a,b;x_1) - y_1)^2 + \\
&(f(a,b;x_2) - y_2)^2 + \\
&(f(a,b;x_3) - y_3)^2 + \\
&\dots \\
&= D(a,b)
\end{aligned}
$$

Ajustando un modelo a los datos

- Encontrar un modelo (función) que represente (aproximadamente) los datos
- Minimizar la distancia de un modelo a los datos

El **modelo** que mejor representa los datos (dentro de esa categoría de modelos) es el que minimiza la función $D(a,b)$

Una vez que obtenemos *a* y *b*, podemos hacer **predicciones** para nuevos valores de *x*

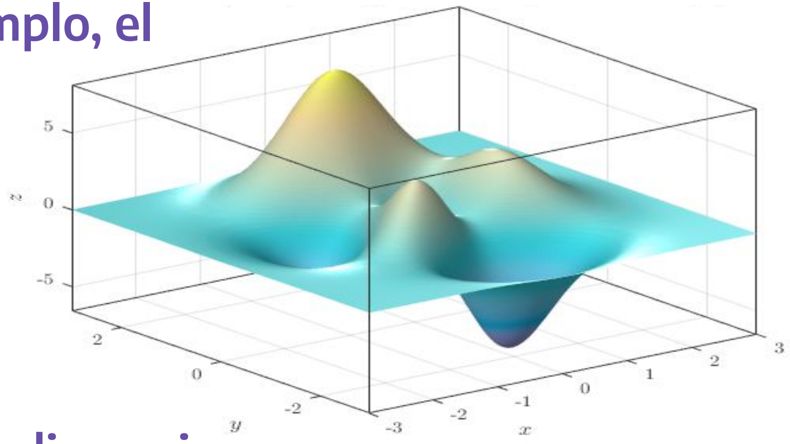
En el caso simple de la función ***y = a x + b*** es muy fácil encontrar *a* y *b* a partir del conjunto de todos los x_i e y_i (derivadas, etc.)
ver expresión en el notebook

Distancia **cuadrática** entre todos los puntos y la **predicción:**

$$(f(a,b;x_1)-y_1)^2+$$
$$(f(a,b;x_2)-y_2)^2+$$
$$(f(a,b;x_3)-y_3)^2+$$
$$\cdots$$
$$=D(a,b)$$

Recordando: Maximización y minimización

Minimizar (o maximizar) una función (por ejemplo, el beneficio o la distancia cuadrática):



- Puntos extremos
- Pendientes nulas en todas las direcciones (gradiente)
- Muy utilizado en aprendizaje automático
- Encontrar máximos y mínimos en muchas dimensiones puede ser muy complicado
- En este caso de ajuste de función, las dimensiones son los parámetros del modelo

Ajustando un modelo a los datos

- Encontrar un modelo (función) que represente (aproximadamente) los datos
- Minimizar la distancia de un modelo a los datos

El **modelo** que mejor representa los datos (dentro de esa categoría de modelos) es el que minimiza

$$D(a, b)$$

Una vez que obtenemos *a* y *b*, podemos hacer **predicciones** para nuevos valores de *x*

En el caso simple de la función ***y = a x + b***
es muy fácil encontrar *a* y *b* a partir del conjunto de todos los x_i e y_i (derivadas, etc.)
ver expresión en el notebook

En este caso en particular hay una solución exacta y su expresión es:

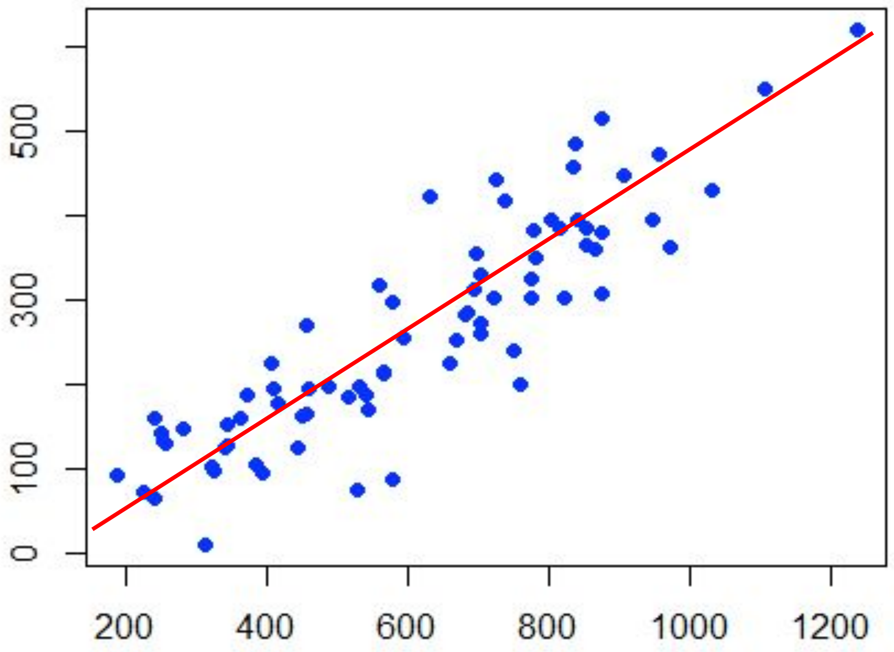
$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$

Ver ejemplo en el notebook

¡Aprendiendo de los datos!

- A partir de los datos, obtuvimos un **modelo** que los representa: el que tiene los parámetros que minimizan la distancia cuadrática promedio a los datos (por eso se llama de "método de mínimos cuadrados")
- El modelo **generaliza** una relación y permite hacer **predicciones**



- Se puede pensar que el modelo se **entrenó** con los dados (**fiteando**) y ahora puede **predecir** un valor de *y* para cualquier *x*
- ¡Un algoritmo y un ejemplo simple de aprendizaje automático a partir de datos!

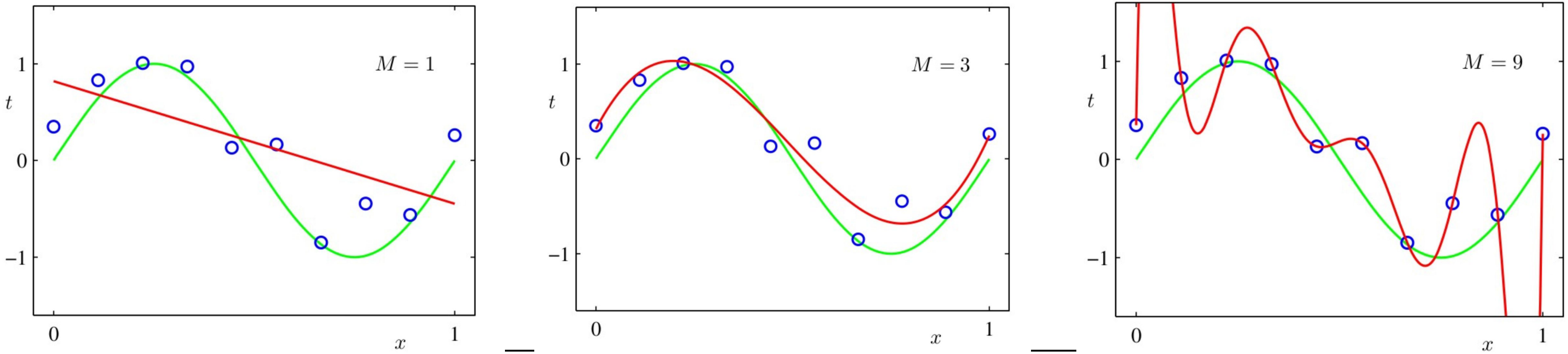
¿Cómo esto se generaliza?

- Aquí el modelo era lineal, $y = ax + b$ (por eso se le llama **regresión linear**)
- Podemos tener funciones más complejas, como cuadrática, polinómica, etc.

Vector de coeficientes

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M$$

- En general tienen más parámetros (= más libertad para ajustar los datos)
- ¡Hay que minimizar en muchas dimensiones!



¿Cómo esto se generaliza?

- **Aquí el modelo dependía de una única variable, *x*, y queremos predecir un único valor, *y*.**
- **En general la dependencia puede ser en muchas variables**
    - Ejemplo: valor del inmueble puede depender del área, número de ambientes, región, edad, etc.
    - x1, x2, x3, etc. (u otras notaciones)
- **Podemos querer modelizar/predecir várias cantidades**
    - Ejemplo: altura y peso en función de la edad (y otras variables)
- **En vez de valores precisos, muchas veces queremos predecir/entender distribuciones de probabilidad**

Aplicaciones más avanzadas

- Ejemplo con una función lineal, baja dimensionalidad de los datos (2) y pocos parámetros (2)
- Algunos sistemas son altamente no lineales, tienen una alta dimensionalidad y un número extremadamente grande de parámetros.
- Por ejemplo, las redes neuronales convolucionales estándar pueden tener 100.000.000 de parámetros libres.
- Un poco más complicado....
- Ya lo veremos, especialmente en el módulo 3...

Argentina programa 4.0

Probabilidades

Probabilidad

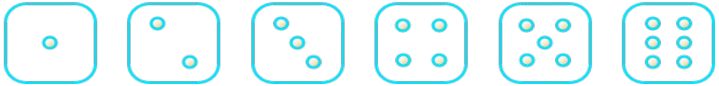
- Concepto fundamental en nuestro cotidiano, sin que tomemos mucha conciencia
- No suele ser tratado adecuadamente en la educación general
- También fundamental para entender la ciencia de datos y la inteligencia artificial
- Como en todo, se usa un lenguaje y herramientas matemáticas, pero hay un marco conceptual que intentaremos hacer intuitivo
- La "matemática aburrida y difícil" nos llevará al maravilloso mundo de la ciencia de datos y la inteligencia artificial...

Probabilidad

- **Nomenclatura: aleatorio, probabilístico, estocástico**
- **Fuentes de incertidumbre**
  - Observabilidad incompleta
  - Modelización incompleta
  - Imprevisibilidad (por ejemplo, el estado de ánimo)
  - Estocasticidad inherente (por ejemplo, QM)
- **Probabilidades simples**
  - Probabilidad de que ocurra algo (binaria)
  - Probabilidad entera (discreta) [más estudiada, intuitiva]
  - Probabilidad continua [más útil, realista, frecuente]

Variables aleatorias discretas

- Realización: resultado de un proceso aleatorio (por ejemplo, tirar un dado)
- Muestreo: número de realizaciones
- Ejemplo: 10.000 lanzamientos
- Resultado:

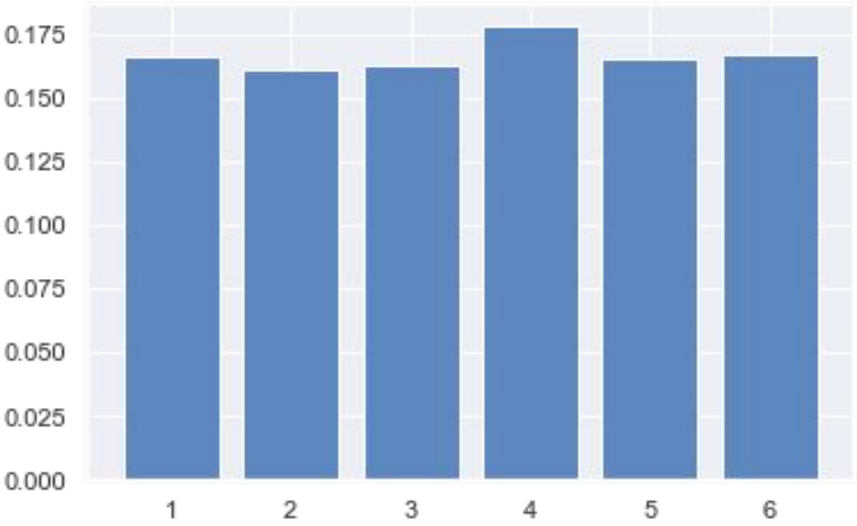


Posibles resultados

Frecuencias relativas:

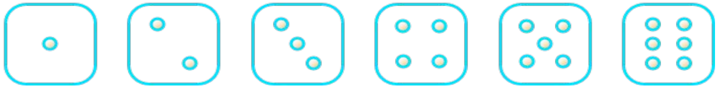
veces que apareció el valor

número de lanzamientos



Variables aleatorias discretas

- Realización: resultado de un proceso aleatorio (por ejemplo, tirar un dado)
- Muestreo: número de realizaciones
- Ejemplo: 10.000 lanzamientos
- Distribución probabilística subyacente



Posibles resultados

Función distribución de probabilidad

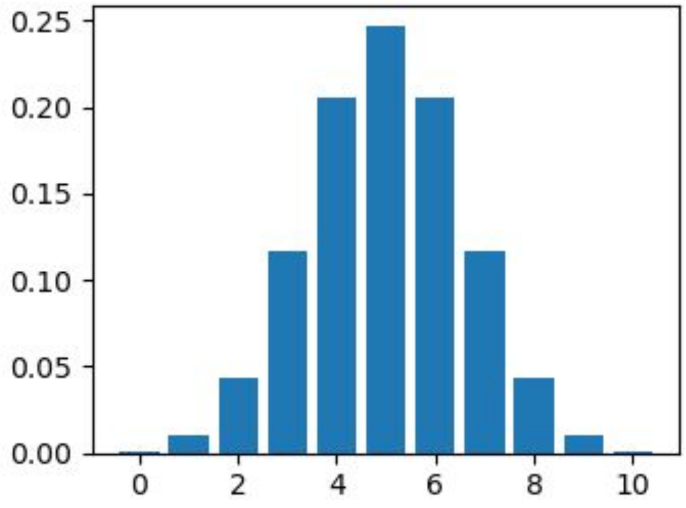
Probabilidades suman 1:

$$\sum_{x \in \mathrm{x}} P(x) = 1$$

Función distribución de probabilidad

- Normalmente cuando pensamos em procesos aleatórios, pensamos en equiprobabilidad (ejemplo: dados)
- Las probabilidades puede ser distintas según el número que se obtiene
- Ejemplo: lanzamos diez veces una moneda, y contamos la cantidad *n* de veces que obtenemos cara. La variable *n* va a tomar los valores {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10} con distinta probabilidad.

A cualquier variable aleatoria discreta le podemos asociar una **función distribución de probabilidad** (PDF, de *Probability Distribution Function*)
Para cada valor, le asigna la probabilidad de que la variable tome dicho valor.
Esta probabilidad la podemos entender como la frecuencia relativa de ese valor.



Ejemplo: distribución binomial

Y si en vez de probabilidades iguales (cara o seca), tenemos probabilidades distintas de que algo ocurra o no?

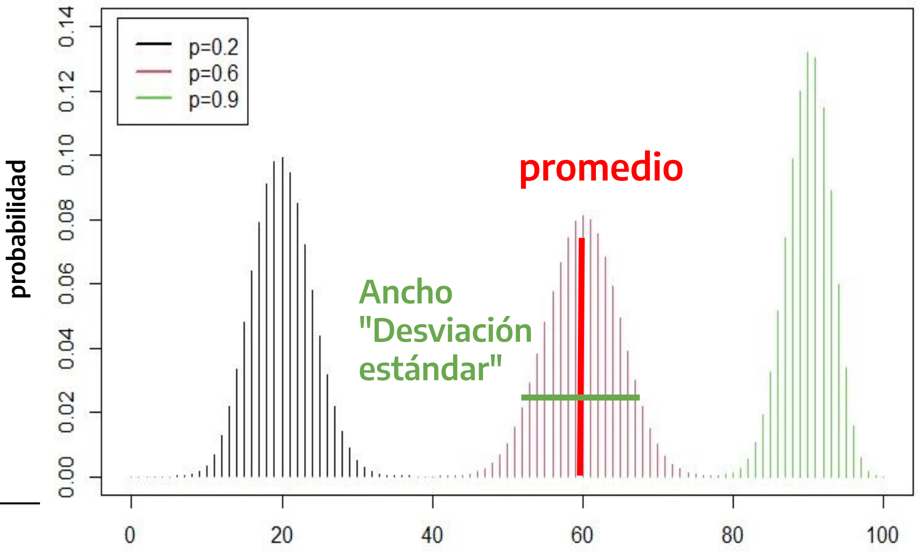
Ejemplo: probabilidad de éxito y fracaso (respuesta a una pregunta "sí" o "no")

$P(E) = p$ a la probabilidad de éxito, $P(F) = q$ a la probabilidad de fracaso. Como la suma tiene que ser uno, $q = 1 - p$

Si hacemos el mismo proceso n veces (y las probabilidades son independientes), la probabilidad de obtener x éxitos es:

$$f(x) = \frac{n!}{x!(n-x)!}\, p^x q^{n-x}$$

Observación: aquí la variable x va de 0 a n

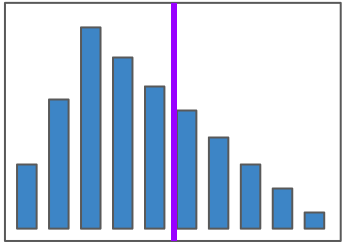


Cuantificando distribuciones (y datos)

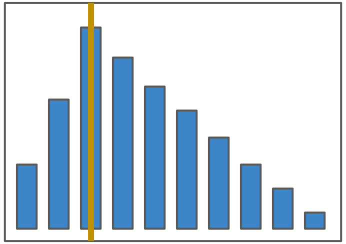
Medidas que se pueden hacer en distribuciones ¡Y en datos!

- Promedio: suma de todos los valores/N
- Mediana: valor que separa los datos en 2 mitades
- Moda: valor más probable

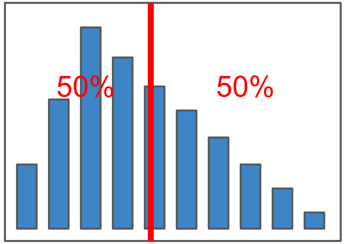
$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



Media



Moda



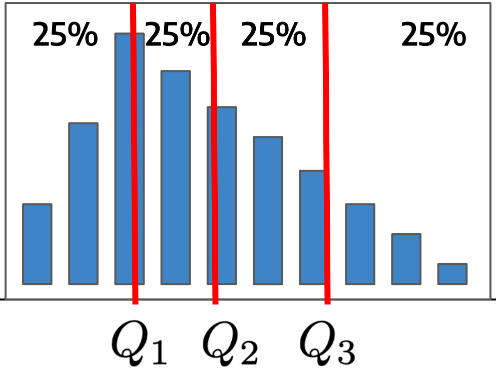
Mediana

Cuantificando distribuciones (y datos)

Medidas que se pueden hacer en distribuciones ¡Y en datos!

Medidas del "ancho":

- **Varianza:** $\mathrm{Var} = \dfrac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ldots}{n}$
- **Desviación estándar** $\sigma = \sqrt{\mathrm{Var}}$
- **Cuantiles: contienen fracciones de (la probabilidad de) los datos**
  - Ejemplo: cuartiles, tienen ¼ de los datos
    Entre el primero y el tercer cuartil está la región con el 50% de probabilidad

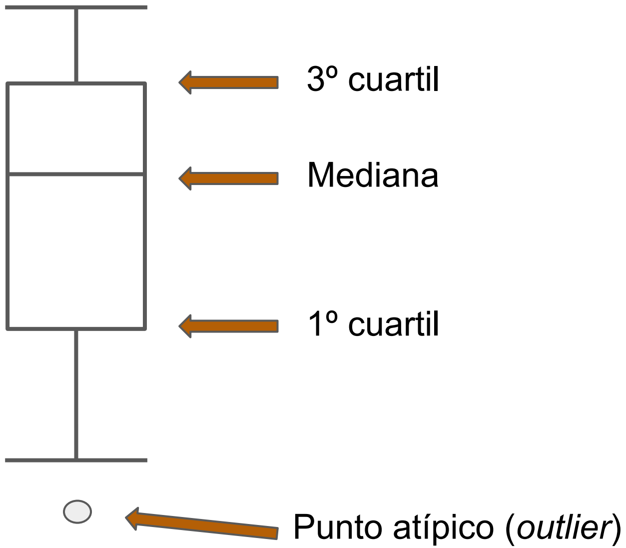


Resumen de distribuciones

PDF: información sobre como se distribuyen los datos

Se pueden condensar en algunas características, como el promedio o mediana y el ancho, por la desviación estándar o los cuartiles.

Una manera cómoda y rápida de visualizar varios de los indicadores que estuvimos viendo es usando los diagrama de caja (que llamamos *box plot*, como en inglés). El mismo resume las propiedades de una serie de datos, mostrando la mediana, máximo, mínimos y cuartiles:

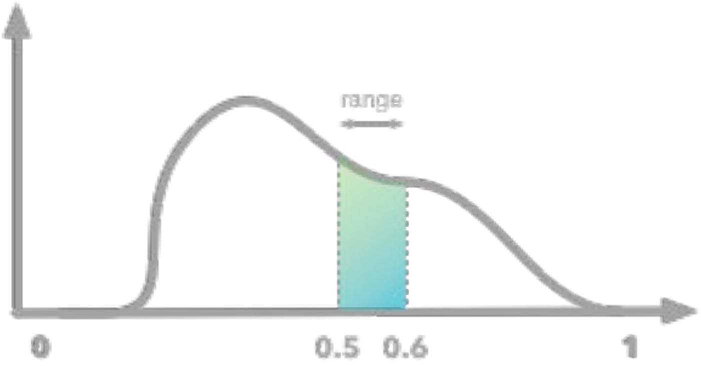


También se pueden medir otras cantidades como la asimetría, etc.

Densidad de probabilidad - función de distribución

- Una variable continua tendrá una probabilidad continua
- p(x) es la densidad de probabilidad

$p(x)$

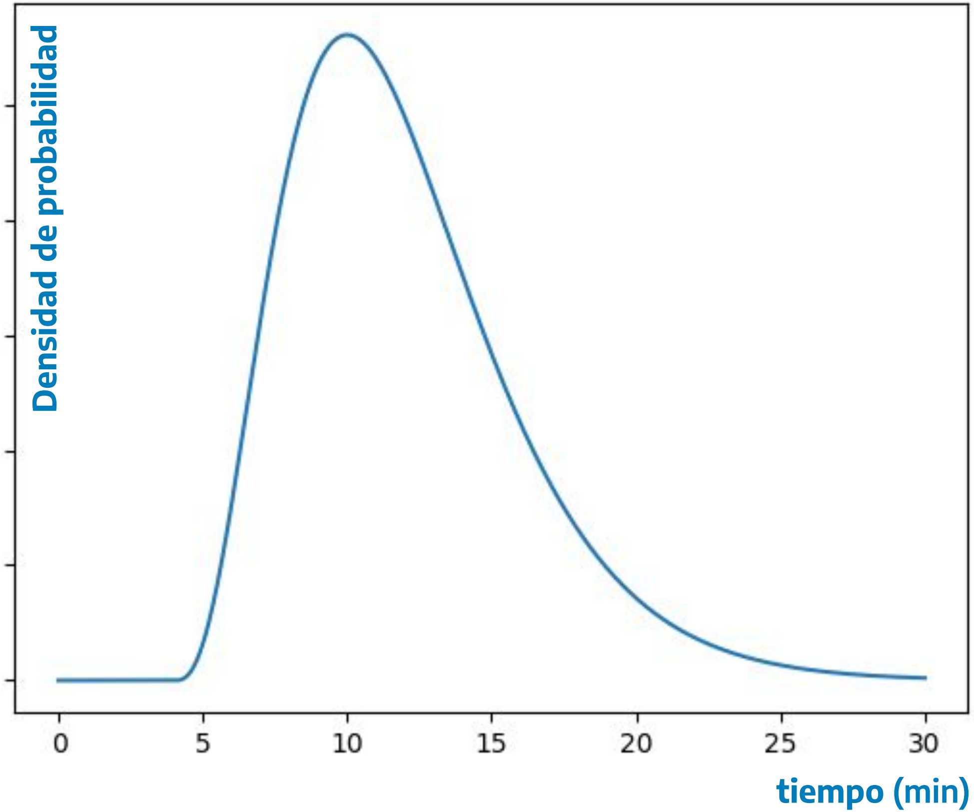


x

- Se puede calcular una probabilidad para un **intervalo de valores**: área bajo la curva (integral)

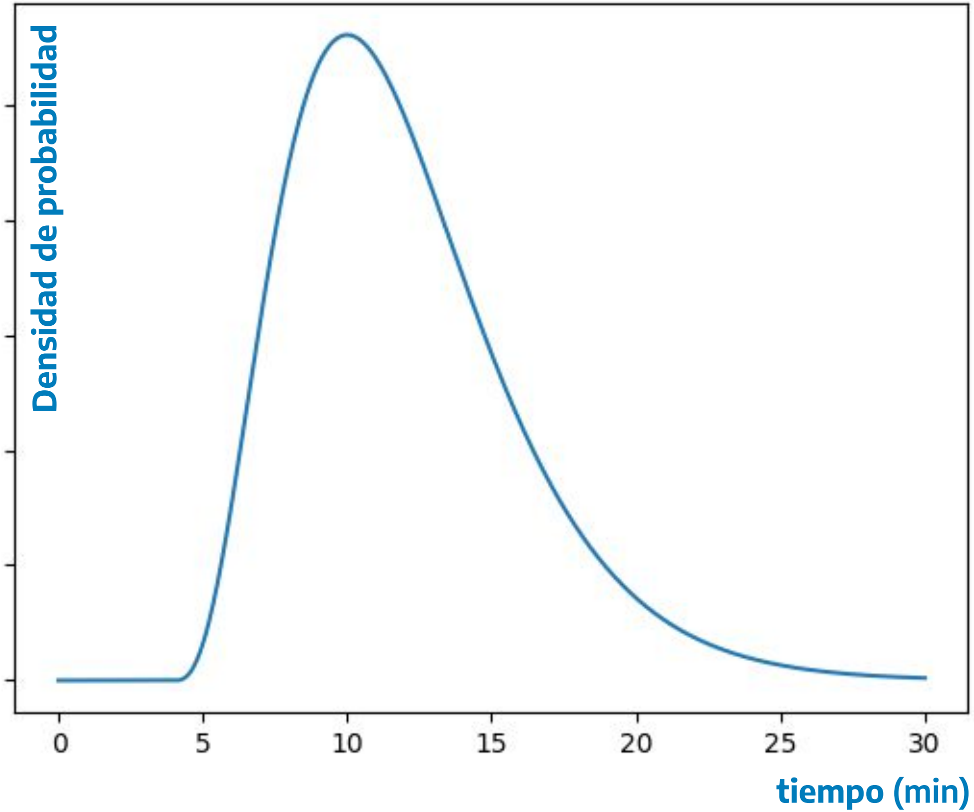
Un ejemplo ficticio: tiempo de llegada

- Función densidad de probabilidad del tiempo que se tarda para llegar a algún lado
- Observación: este es un modelo. Se podría probar si representa bien los datos y obtener los parámetros en alguna situación real



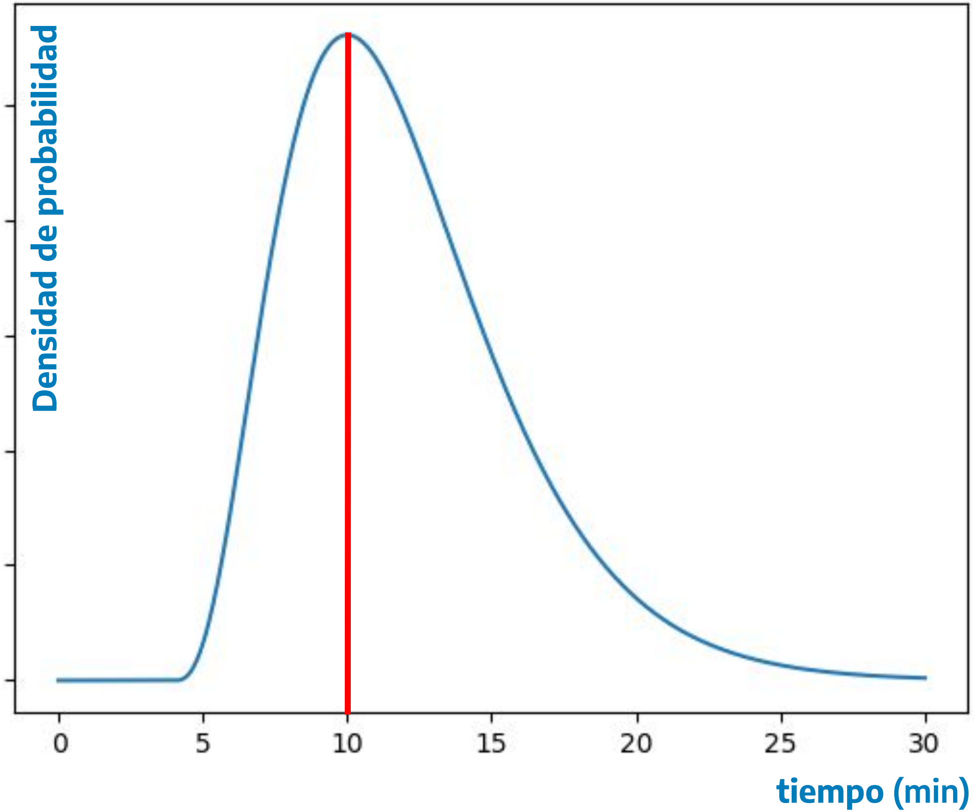
Un ejemplo ficticio: tiempo de llegada

- Función densidad de probabilidad del tiempo que se tarda para llegar a algún lado
- Observación: distribución sesgada



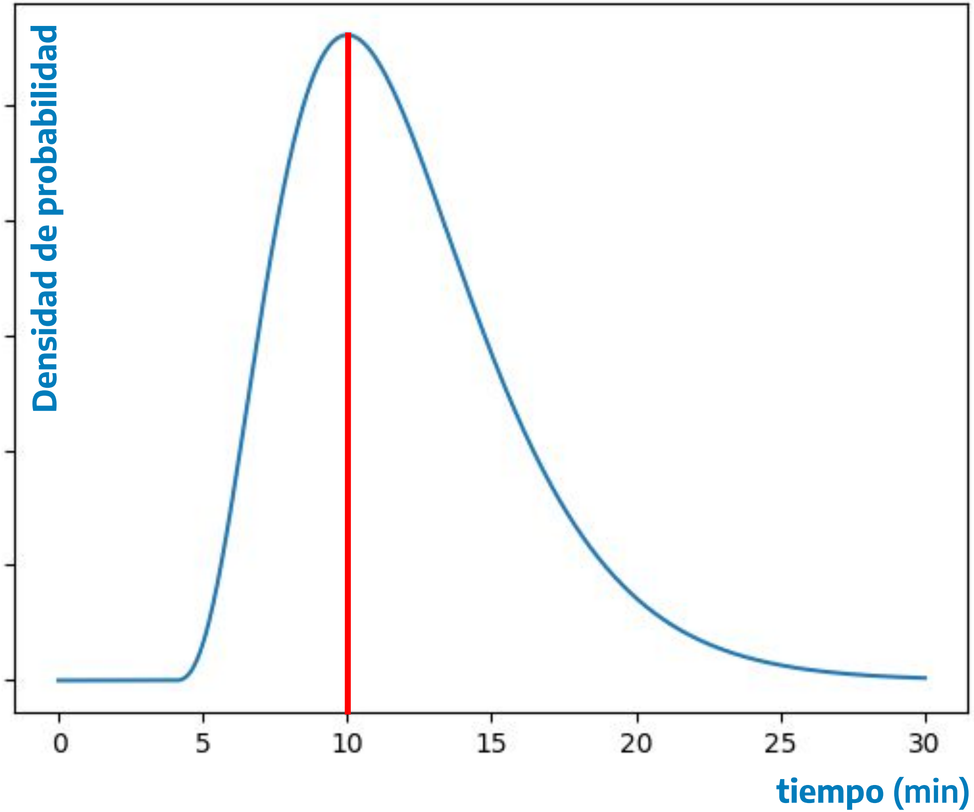
Ejemplo de PDF: tiempo de llegada

- Función densidad de probabilidad del tiempo que se tarda para llegar a algún lado
- ¿Cuál es la duración más frecuente (probable) del viaje?



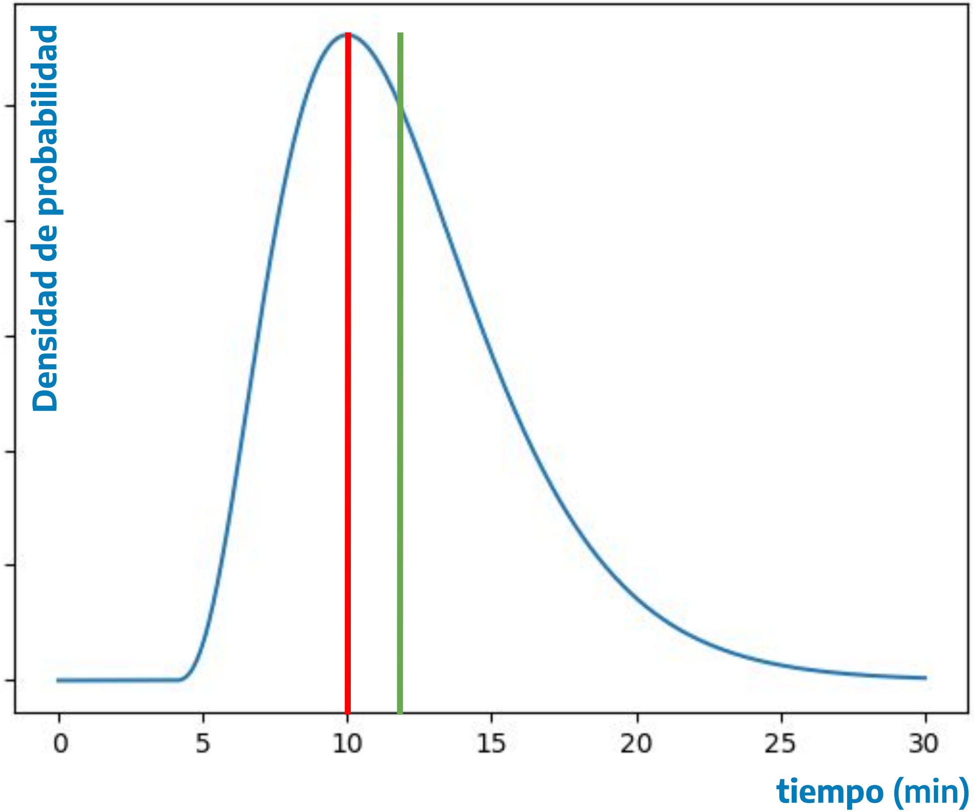
Un ejemplo ficticio: tiempo de llegada

- Función densidad de probabilidad del tiempo que se tarda para llegar a algún lado
- ¿Cuál es la duración más probable del viaje?
- Moda o pico de la distribución
- En este caso: 10 min



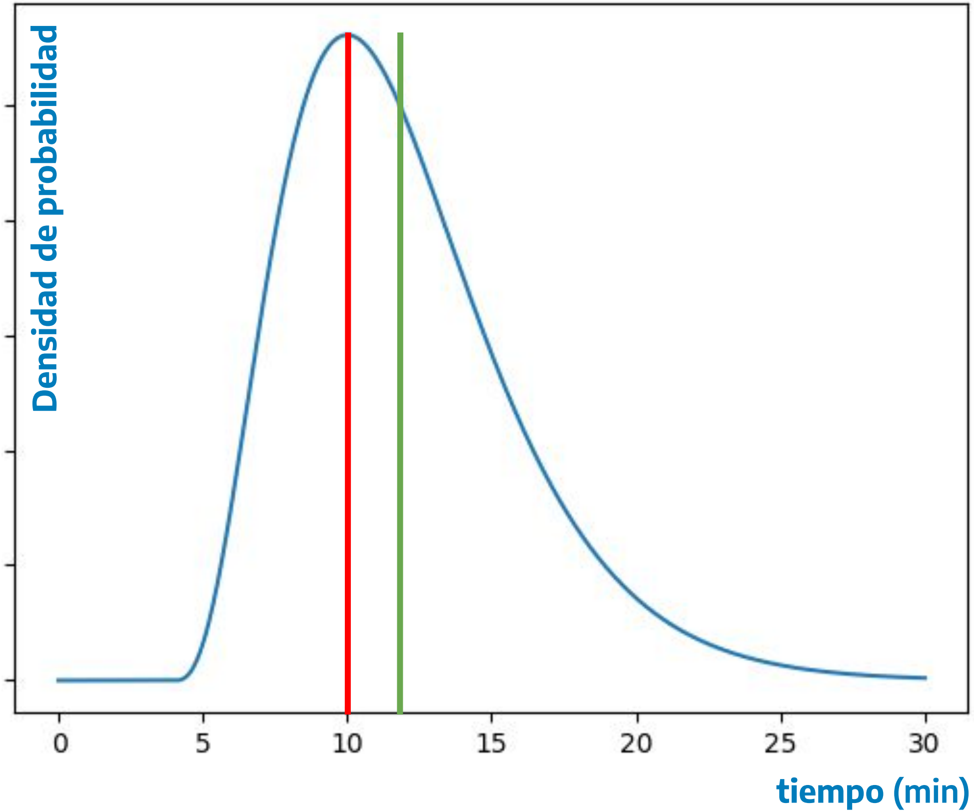
Un ejemplo ficticio: tiempo de llegada

- Función densidad de probabilidad del tiempo que se tarda para llegar a algún lado
- ¿Cuál es la duración más probable del viaje?
- ¿Cuánto tiempo en promedio voy a tardar?



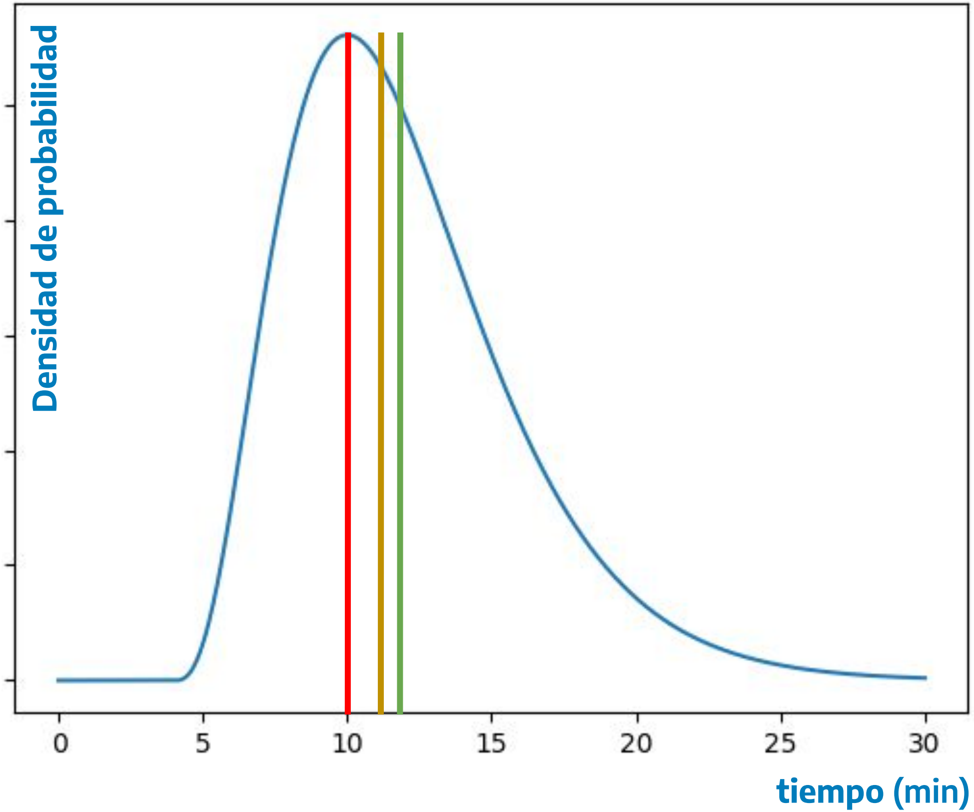
Un ejemplo ficticio: tiempo de llegada

- Función densidad de probabilidad del tiempo que se tarda para llegar a algún lado
- ¿Cuál es la duración más probable del viaje?
- ¿Cuánto tiempo en promedio voy a tardar?
- Valor promedio
- En este caso: 12 min



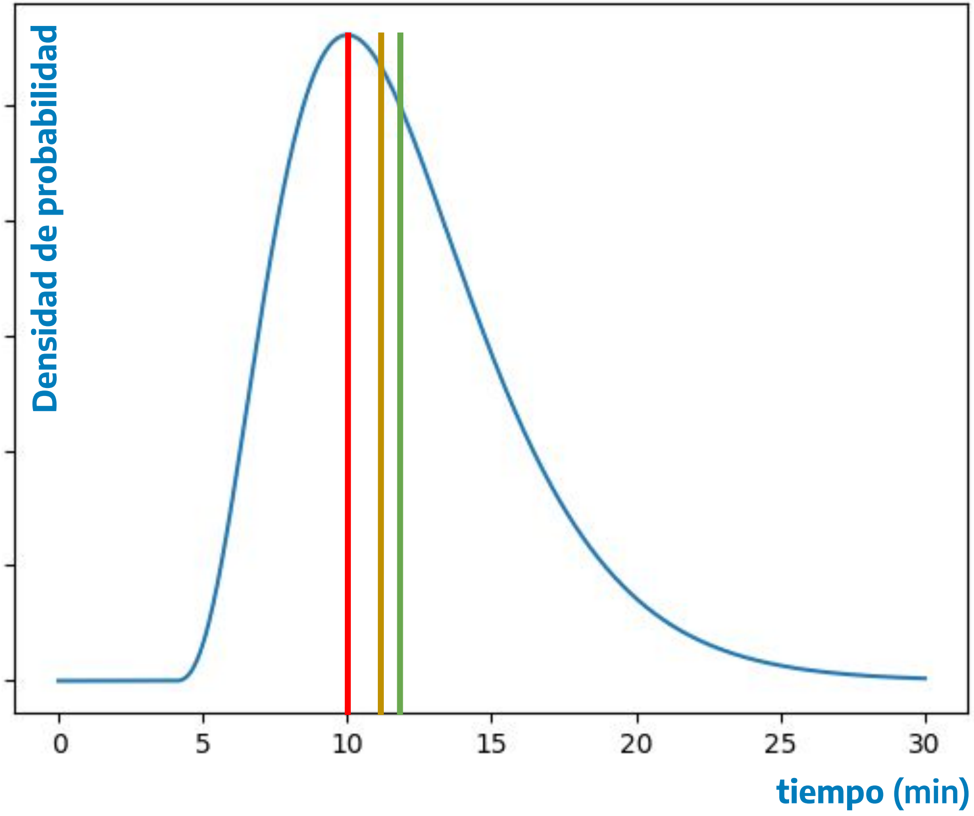
Un ejemplo ficticio: tiempo de llegada

- Función densidad de probabilidad del tiempo que se tarda para llegar a algún lado
- ¿Cuál es la duración más probable del viaje?
- ¿Cuánto tiempo en promedio voy a tardar?
- ¿Cuál es el tiempo a partir del cual voy a llegar después el 50% de las veces?



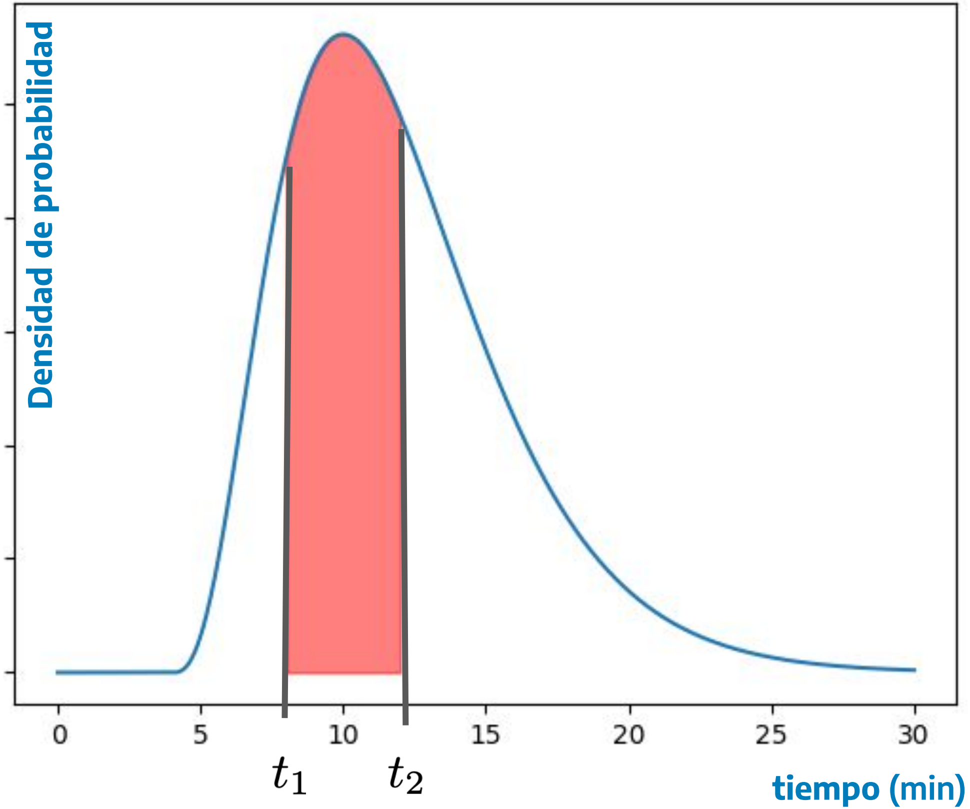
Un ejemplo ficticio: tiempo de llegada

- Función densidad de probabilidad del tiempo que se tarda para llegar a algún lado
- ¿Cuál es la duración más probable del viaje?
- ¿Cuánto tiempo en promedio voy a tardar?
- ¿Cuál es el tiempo a partir del cual voy a llegar después el 50% de las veces?
- Mediana (11.3 min)



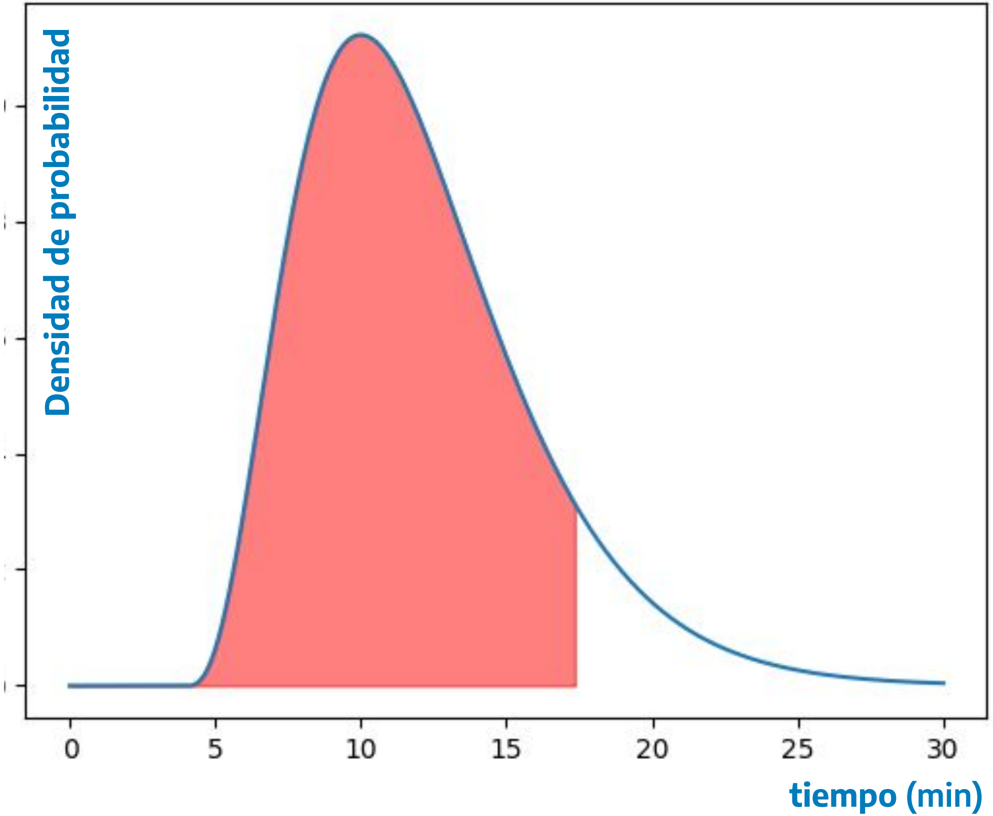
Un ejemplo ficticio: tiempo de llegada

- Función densidad de probabilidad (PDF) del tiempo que se tarda para llegar
- ¿Por qué densidad de probabilidad?
- ¿Puedo predecir la probabilidad de una hora exacta?
  ¿O de un rango de tiempo?
- La probabilidad de llegar entre los tiempos t1 y t2 el el área bajo la curva de la PDF



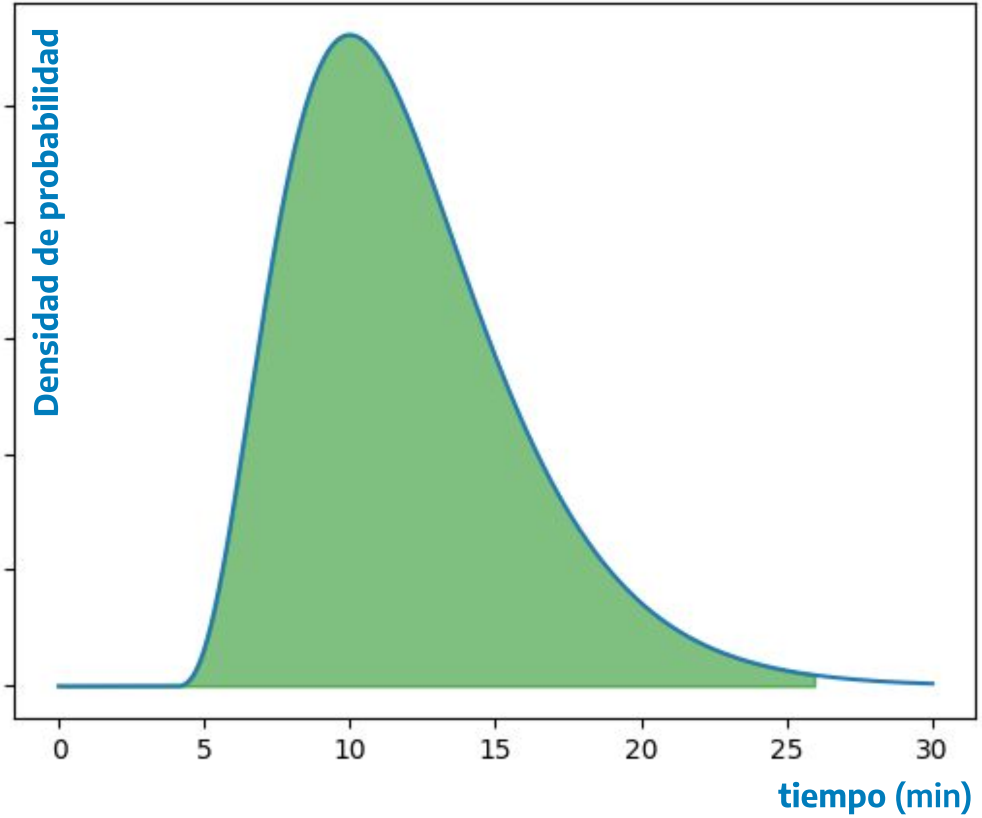
Un ejemplo ficticio: tiempo de llegada

- Función densidad de probabilidad del tiempo que se tarda para llegar a algún lado
- ¿Qué es llegar puntual?
- Digamos: llegar antes de un determinado horário
- Para llegar puntual, tengo que salir con un **margen**
- Si quiero ser puntual 90% de las veces, tengo que buscar la duración que contiene 90% del área bajo la curva
- El margen es de 7,36 min (a más que el tiempo promedio)



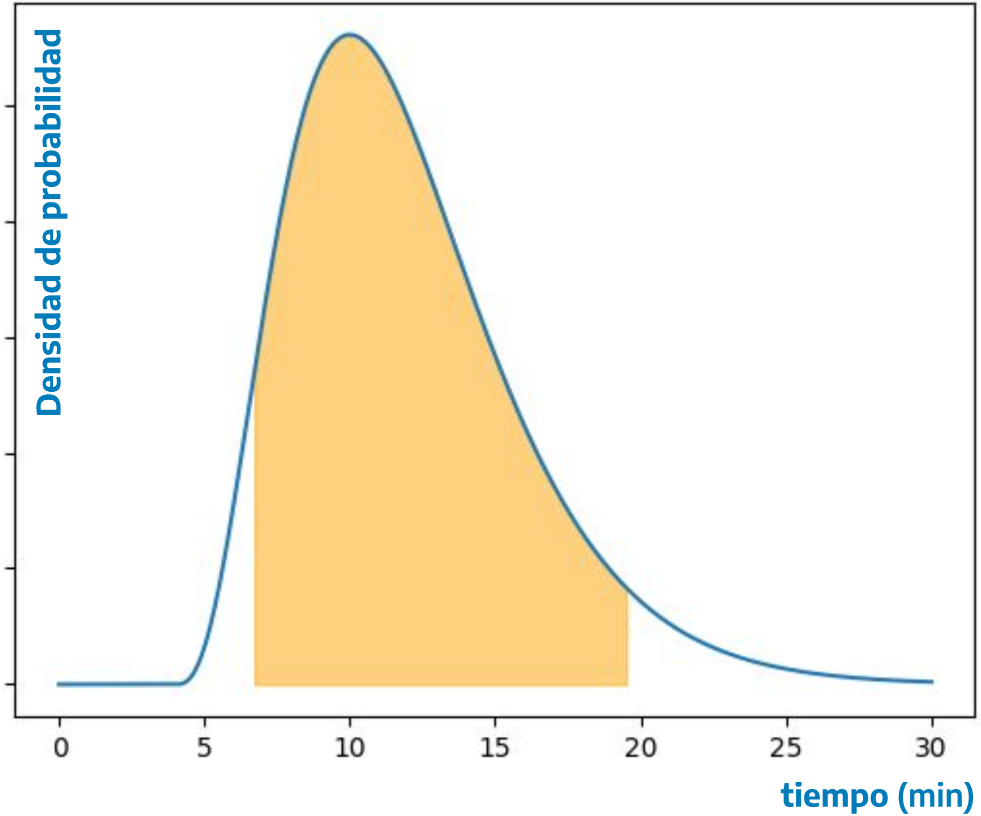
Un ejemplo ficticio: tiempo de llegada

- Función densidad de probabilidad del tiempo que se tarda para llegar a algún lado
- ¿Qué es llegar puntual?
- Digamos: llegar antes de un determinado horário
- Si quiero ser puntual 99.5% de las veces, tengo que buscar la duración que contiene 99.5% del área bajo la curva
- El margen es de 16 min



Un ejemplo ficticio: tiempo de llegada

- Función densidad de probabilidad del tiempo que se tarda para llegar a algún lado
- ¿Qué es llegar puntual?
- Digamos: **alrededor** de un cierto horario
- Entonces quiero duraciones que abarquen, digamos 90% del área de la curva, alrededor del pico



Distribución Normal o Gaussiana

La distribución normal o Gaussiana es la más común entre todas las distribuciones de densidad de probabilidad utilizadas en Estadística. Tiene importantes aplicaciones en la modelización de variables estadísticas asociadas a los elementos de una población.

Ejemplos:

- Medidas físicas del cuerpo humano en una población (altura, peso, etc..,)
- Medidas de calidad en muchos procesos industriales
- Errores en las observaciones astronómicas

<u>Teorema del límite central:</u> Cuando los resultados de un **conjunto de datos** se deben a una **combinación muy grande de factores independientes**, que actúan sumando sus efectos, siendo cada efecto individual de poca importancia respecto al conjunto, es esperable que los **resultados** de ese conjunto sigan una **distribución normal.**

Distribución Normal o Gaussiana

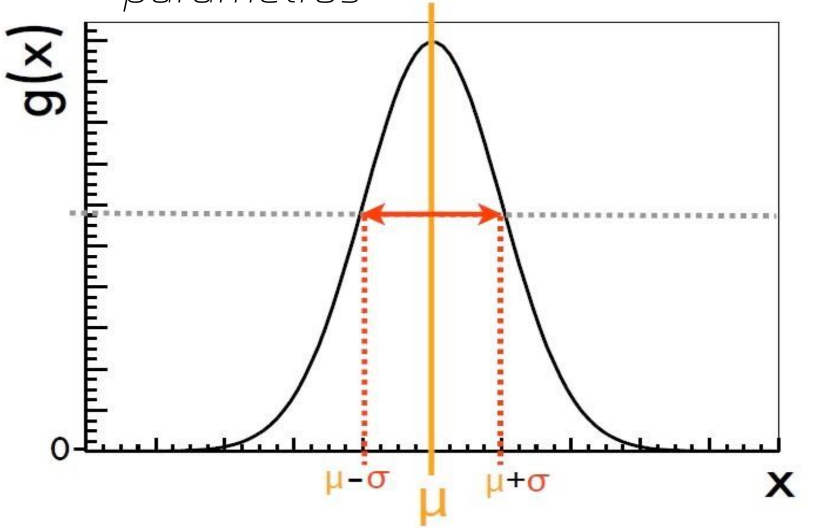
PDF gaussinana:

$$g(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

argumento

parametros

- Modo (pico), μ: valor más probable
- Región de confianza
    - Rango que abarca una probabilidad determinada (percentil)
    - Desviación típica σ: contiene el 68% de probabilidad



- Valor medio, μ (la mediana es más representativa si hay valores atípicos o colas)

Toda la forma de la PDF contiene información útil (por ejemplo, múltiples picos)

La pdf contiene esencialmente toda la información que podemos obtener de un proceso estocástico

Resumen

- **Datos:**
    - Juntar información sobre algo (observar la realidad)
    - En general discretos (conjuntos de puntos/vectores)
- **Estadísticas de los datos:**
    - Promedios, medianas, desviación estándar, cuartiles, etc.
    - Caracterizar los datos
    - Sacar conclusiones
- **Modelos:**
    - Generalizar/aprender
- **Muchos modelos son probabilísticos (suelen ser los mejores/más completos)**
- **¡¿Listos para arrancar con la ciencia de datos?!**