Argentina programa 4.0

Universidad Nacional de San Martín

Módulo 2 - Ciencia de Datos

Semana 1. Elementos de matemática y probabilidad

Argentina programa 4.0

Universidad Nacional de San Martín

Módulo 2 - Ciencia de Datos

Semana 1. Elementos de matemática y probabilidad

Argentina programa 4.0

Elementos de matemática y probabilidad

- Elementos de Cálculo y Álgebra. Funciones. Vectores y Matrices. Nociones de derivadas e integrales
- Definición de probabilidad. Probabilidad conjunta, marginal y condicional Leyes de la probabilidad
- La interpretación frecuentista y bayesiana de la probabilidad
- Distribuciones especiales: Binomial, Poisson, Gaussiana
- Estimadores, estimación de máxima verosimilitud

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**

¿Qué es un modelo?

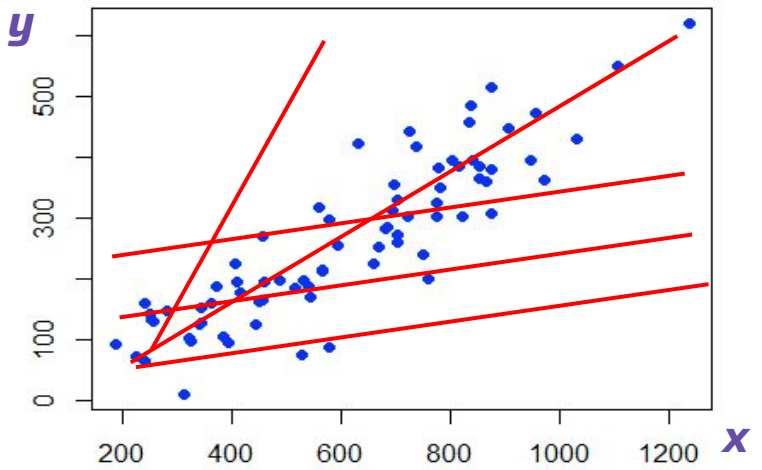
¿Cómo encuentro un modelo que representa datos?

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

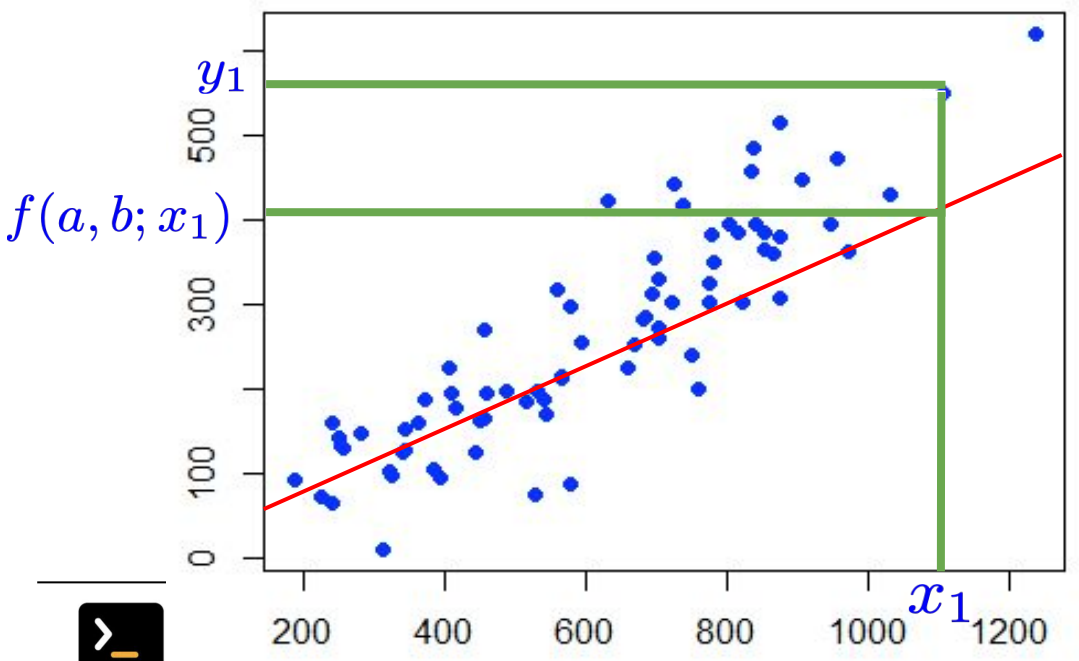
    variable

    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

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

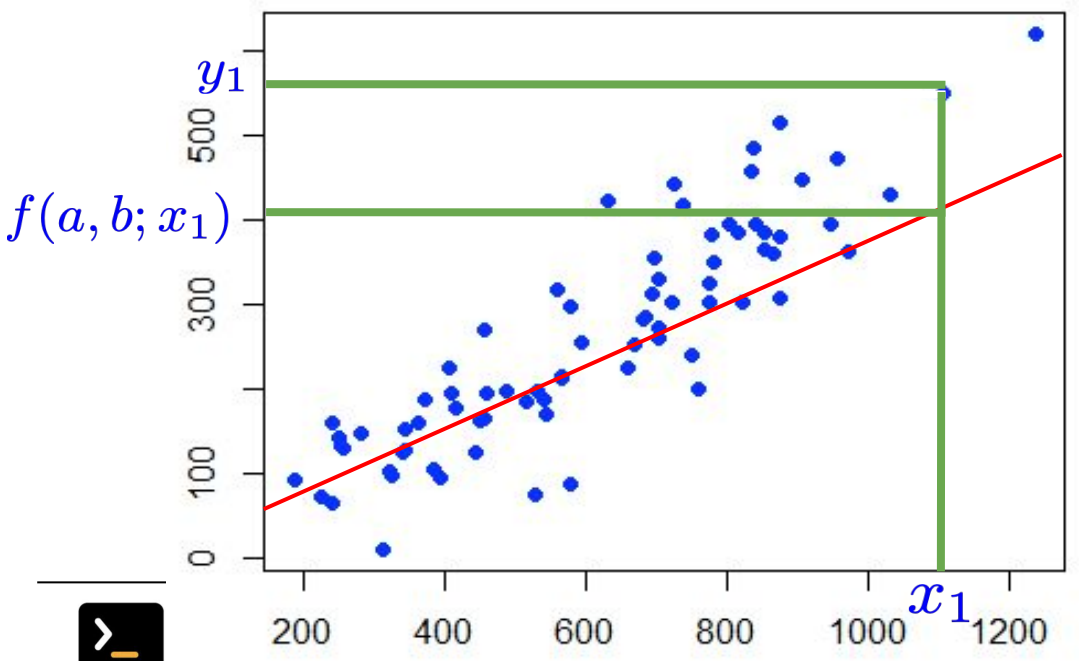
Distancia entre la previsión del modelo y el dato:

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

$$= ax_1 + b - y_1$$

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:

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

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*

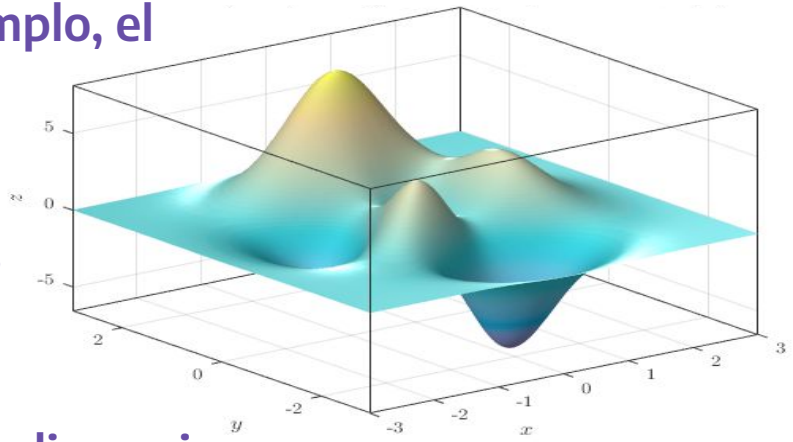
En el caso simple de la función ***y = a x + b***
es muy fácil encontrar *a* y *b* a partir del conjunto de todos los x_i e y_i (derivadas, etc.)
ver expresión en el notebook

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

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

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:

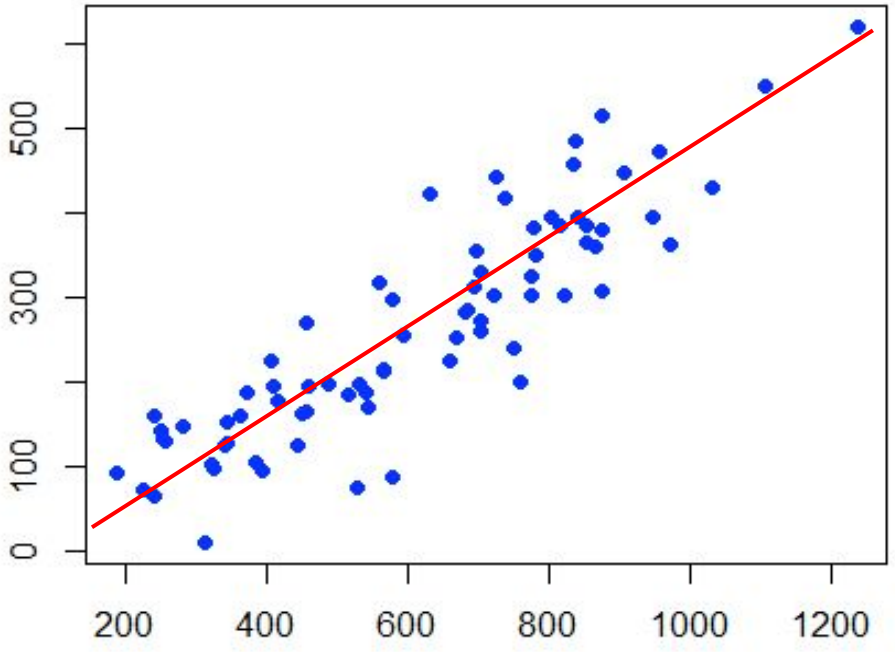
$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

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!

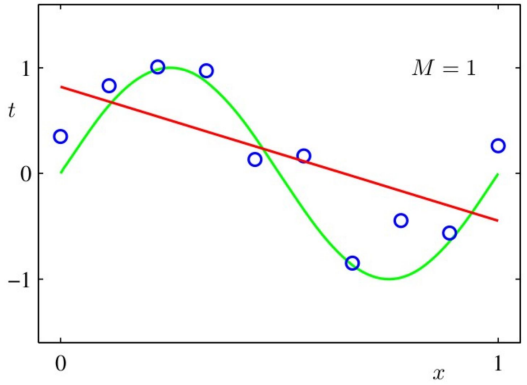
¿Cómo esto se generaliza?

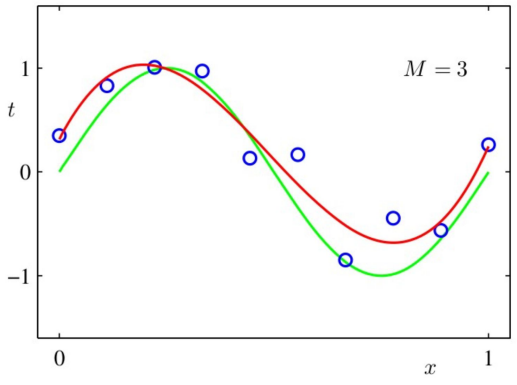
- Aquí el modelo era lineal, $y = ax + b$ (por eso se le llama **regresión linear**)
- Podemos tener funciones más complejas, como cuadrática, polinómica, etc.

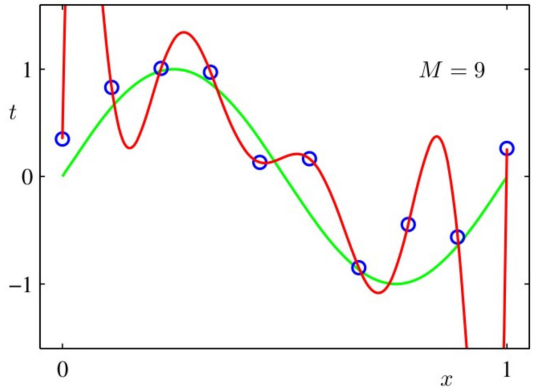
Vector de coeficientes

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

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







¿Cómo esto se generaliza?

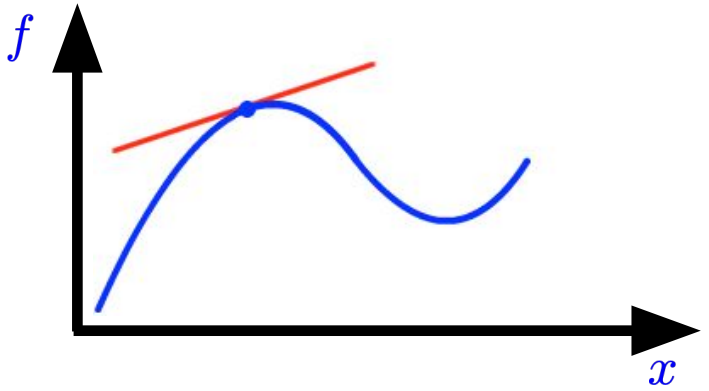
- **Aquí el modelo dependía de una única variable, *x*, y queremos predecir un único valor, *y*.**
- **En general la dependencia puede ser en muchas variables**
  - Ejemplo: valor del inmueble puede depender del área, número de ambientes, región, edad, etc.
  - x1, x2, x3, etc. (u otras notaciones)
- **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

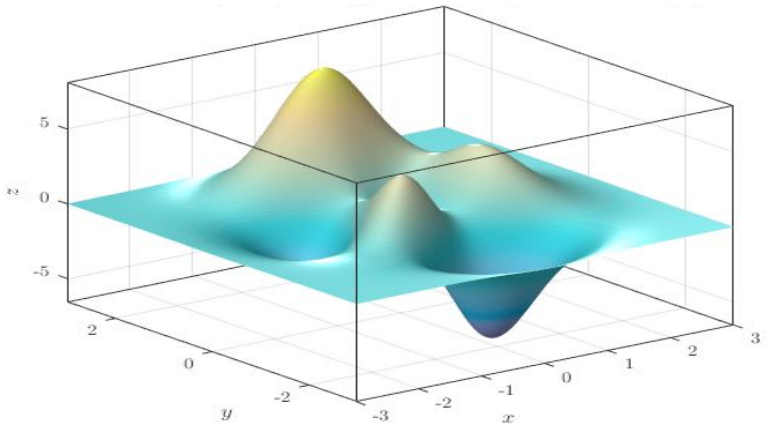
- Ejemplo con una función lineal, baja dimensionalidad de los datos (2) y pocos parámetros (2)
- Algunos sistemas son altamente no lineales, tienen una alta dimensionalidad y un número extremadamente grande de parámetros.
- Por ejemplo, las redes neuronales convolucionales estándar pueden tener 100.000.000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

Pendiente de una curva

- **La derivada es la pendiente de la recta tangente a la curva**
  - es una nueva función derivada
  - notación habitual $\frac{df}{dx}$ o $f'(x)$
  - recta tangente en el punto p: $y = a\,x + b$, donde $a = f'(p)$



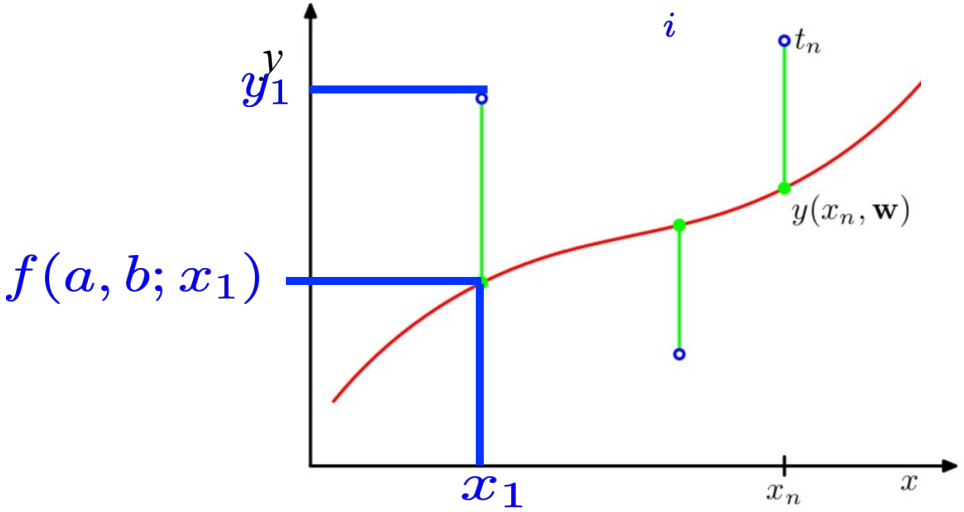
- Si la función tiene más variables
- pendiente de la recta tangente en cada plano: (f,x), (f,y)
- derivada a lo largo de cada variable



Ejemplo: regresión por mínimos cuadrados

- **"Algoritmo más sencillo de aprendizaje automático"**
  - Encontrar un modelo (función) que represente (aproximadamente) los datos
  - Minimizar la distancia de un modelo a los datos

$$D^2(a,b;\{y\},\{x\}) = \sum_i \left(f(a,b;x_i) - y_i\right)^2$$



Ejemplo: regresión por mínimos cuadrados

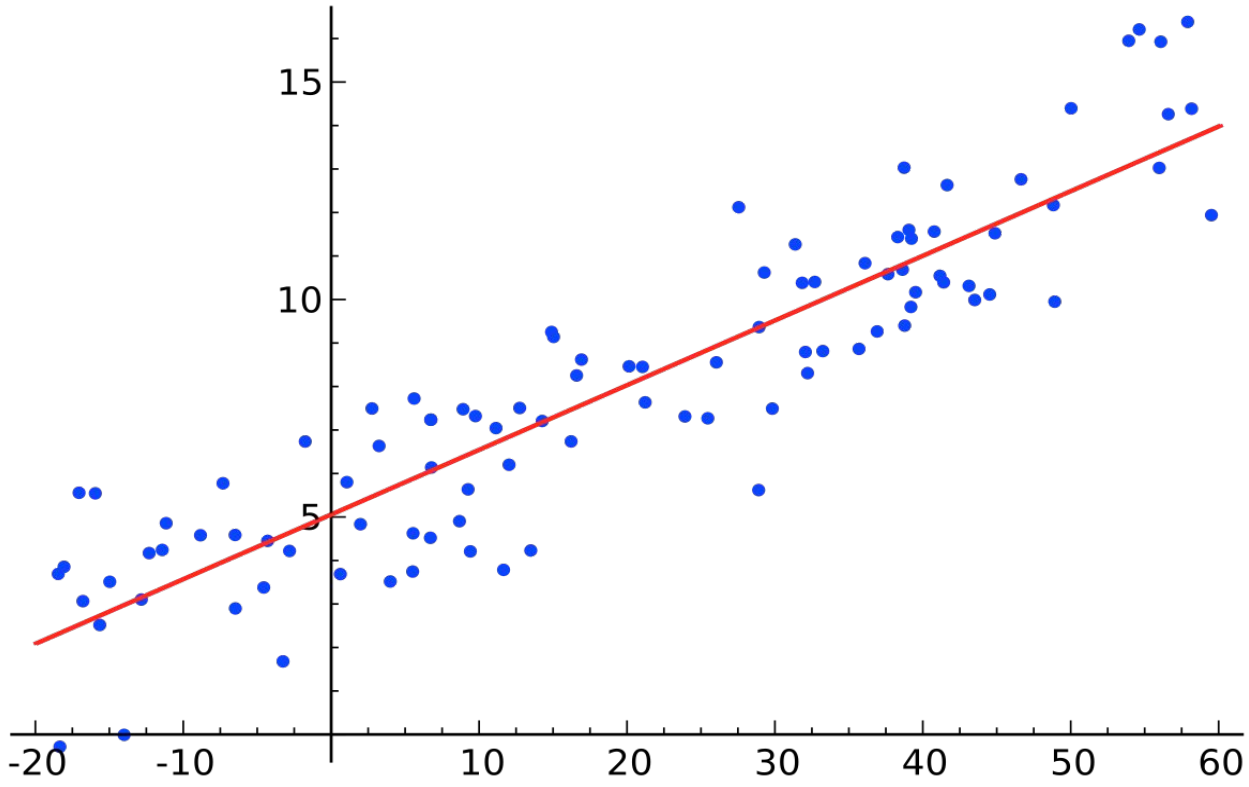
- **"Algoritmos más sencillos de aprendizaje automático"**
  - Minimizar la distancia de un modelo a los datos

$$D^2(a,b;\{y\},\{x\}) = \sum_i \left(f(a,b;x_i) - y_i\right)^2$$

  - Ejemplo: relación lineal y = a x + b
  - Cálculo (derivadas) + matrices (inversa, transpuesta): < 10 líneas de código, encuentra a y b

- **¡Esto es un aprendizaje automático a partir de datos!**

Ejemplo: regresión por mínimos cuadrados



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

- Programación: determinista y segura
- La mayoría de las decisiones se basan en una evaluación probabilística (consciente o no)
- Fundamental para nuestra comprensión del mundo
- Concepto subyacente en la IA
- Poco intuitivo
- Abordado superficialmente y mal tratado en la mayoría de los cursos

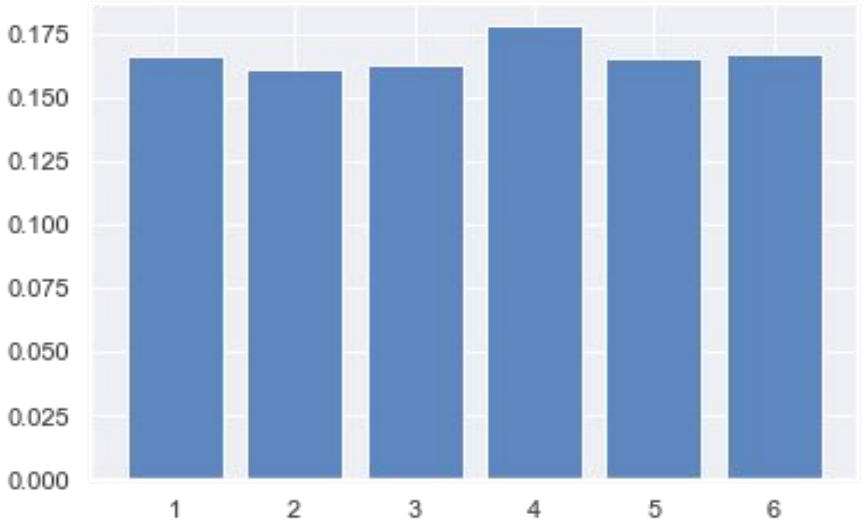
Probabilidad

- **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:

$$\frac{\#\text{ veces que apareció el valor}}{\#\text{ número de lanzamientos}}$$

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

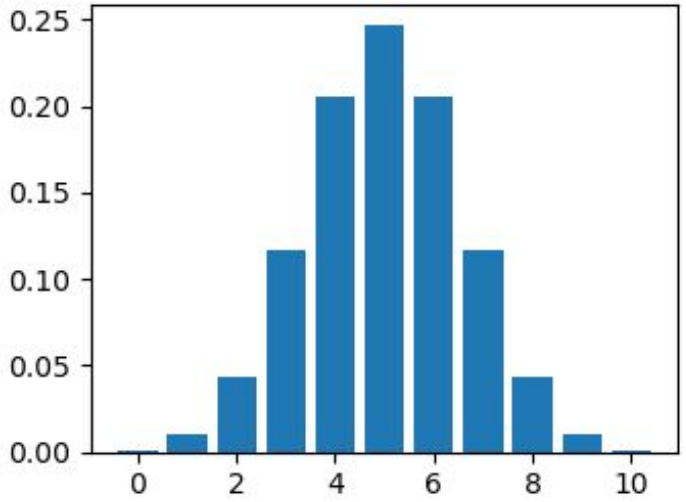
Probabilidades suman 1:

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

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?

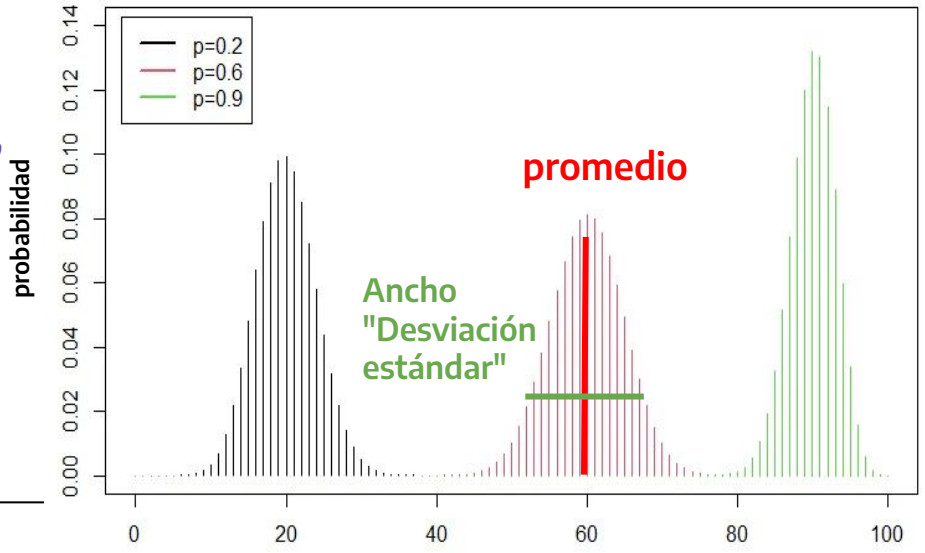
Ejemplo: probabilidad de éxito y fracaso (respuesta a una pregunta "sí" o "no")

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:

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

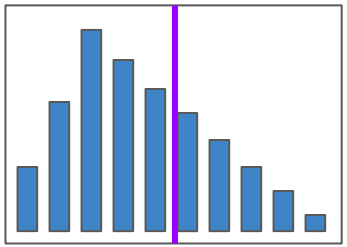
Observación: aquí la variable *x* va de 0 a *n*



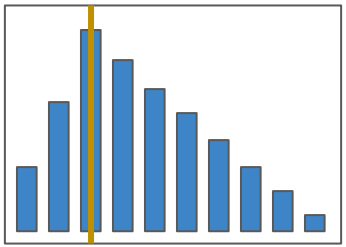
Cuantificando distribuciones (y datos)

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

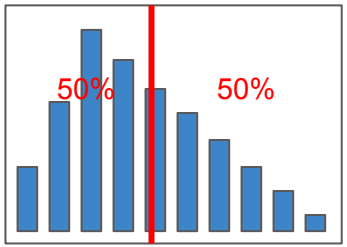
- Promedio: suma de todos los valores/N $\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Mediana: valor que separa los datos en 2 mitades
- Moda: valor más probable



Media



Moda



Mediana

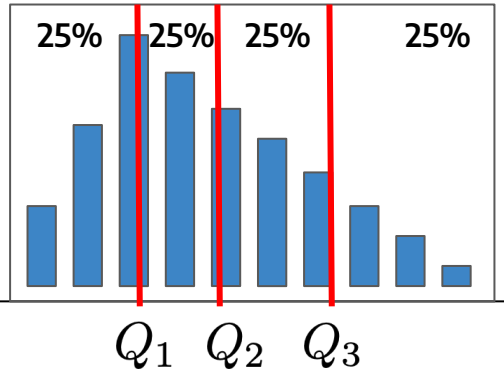
Cuantificando distribuciones (y datos)

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

Medidas del "ancho":

- **Varianza:** $\mathrm{Var} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ldots}{n}$

- **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

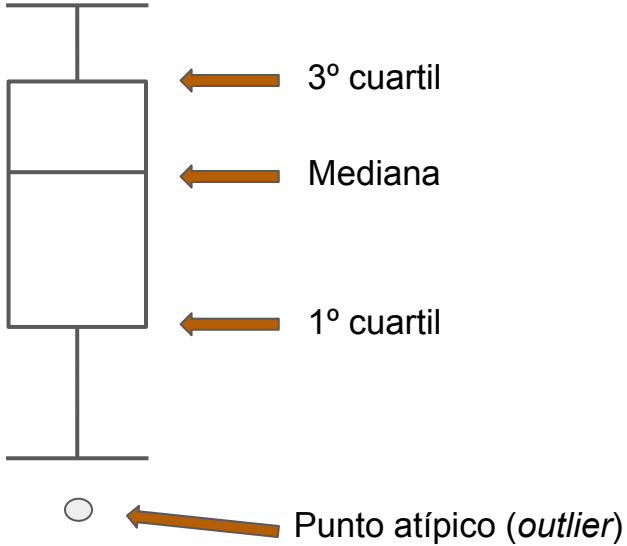


Resumen de distribuciones

PDF: información sobre como se distribuyen los datos

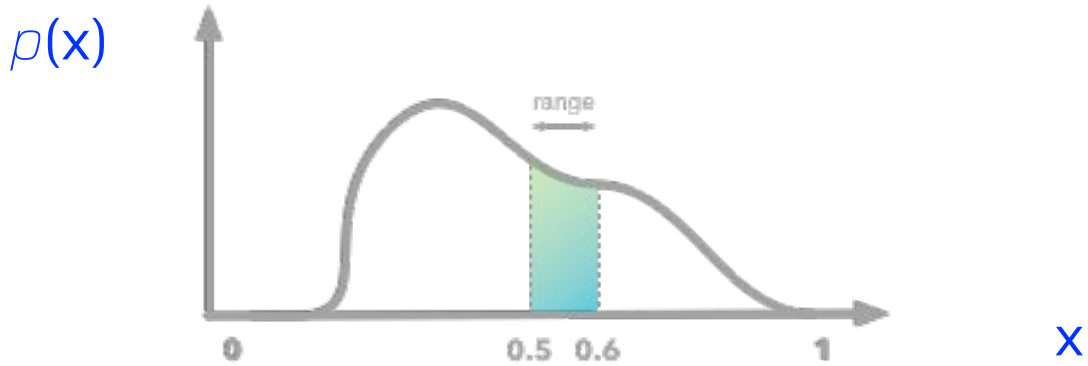
Se pueden condensar en algunas características, como el promedio o mediana y el ancho, por la desviación estándar o los cuartiles.

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:



Densidad de probabilidad - función de distribución

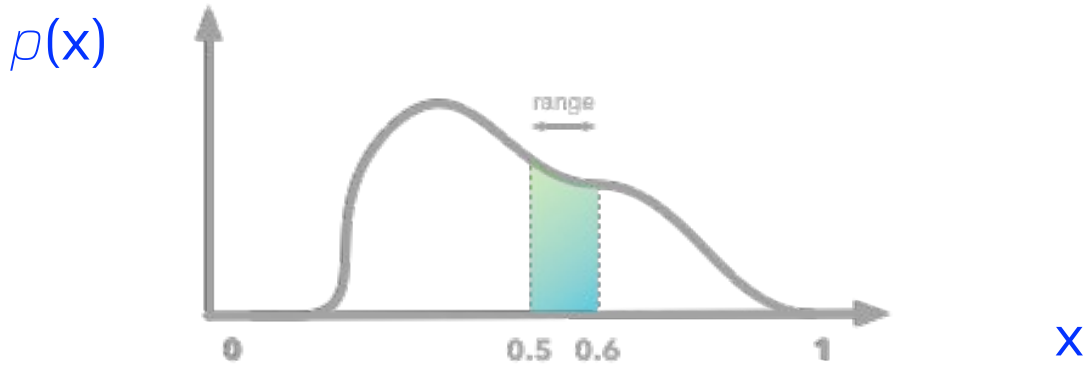
- Una variable continua tendrá una probabilidad continua
- p(x) es la densidad de probabilidad



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

Densidad de probabilidad - función de distribución

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



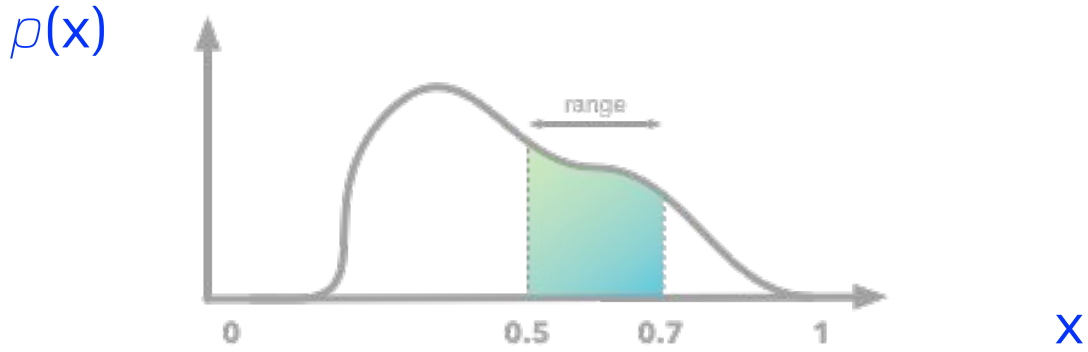
- Se puede calcular una probabilidad para un intervalo de valores

$$P = \int_A^B p(x)dx$$

integral (área bajo la curva)

Densidad de probabilidad - función de distribución

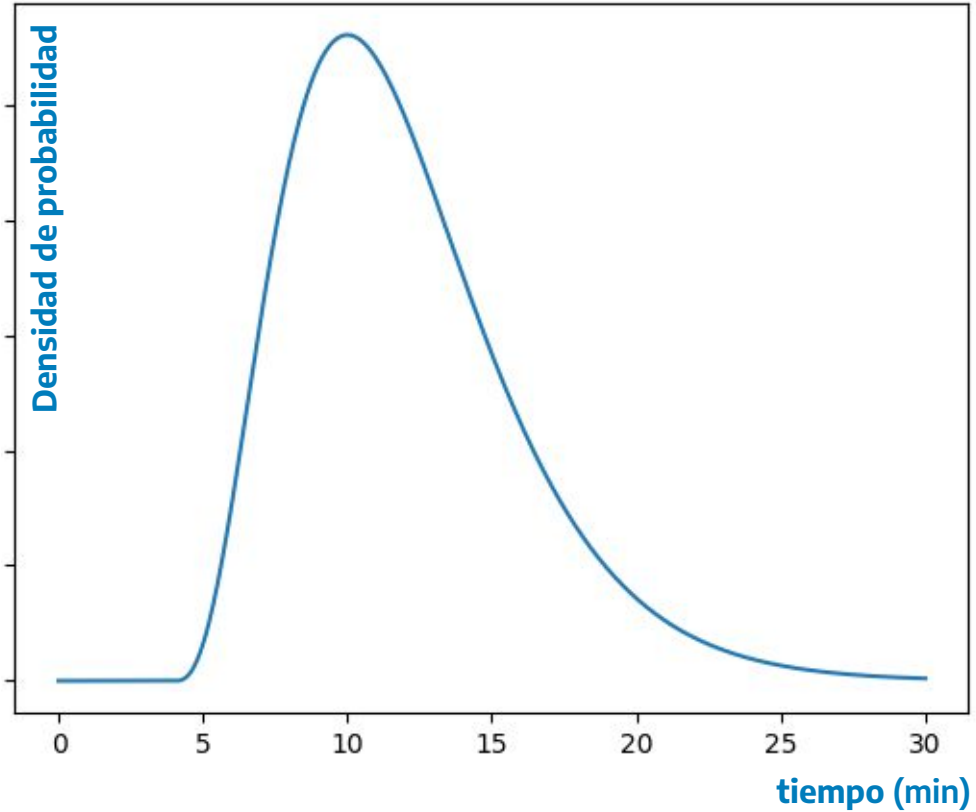
- Una variable continua tendrá una probabilidad continua
- p(x) es la densidad de probabilidad



- Se puede calcular una probabilidad para un intervalo de valores
- Las probabilidades deben sumar 1: la integral sobre todos los valores posibles es la unidad (normalización) $\int_{-\infty}^{+\infty} p(x)dx = 1$

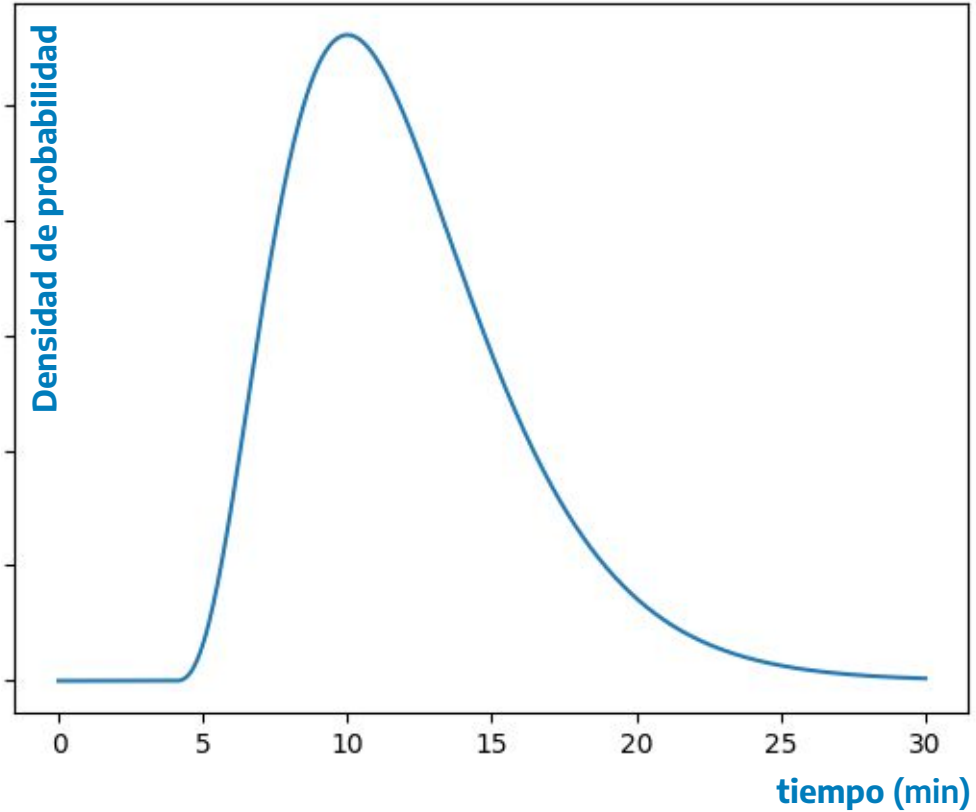
Un ejemplo ficticio: tiempo de llegada

- 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



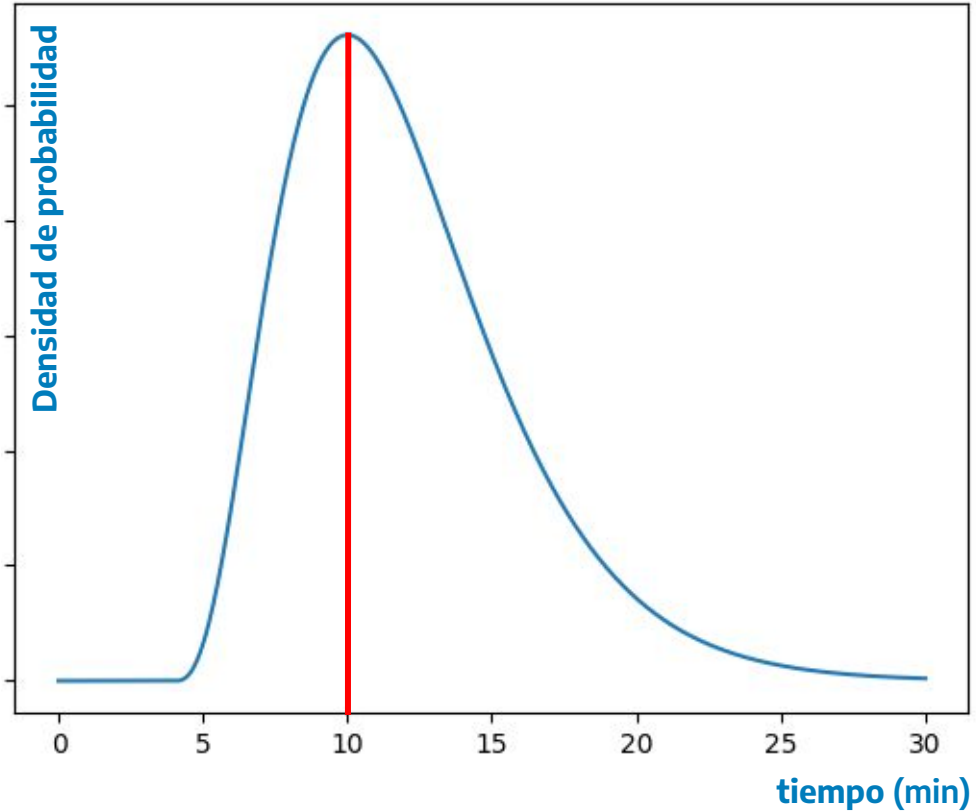
Un ejemplo ficticio: tiempo de llegada

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



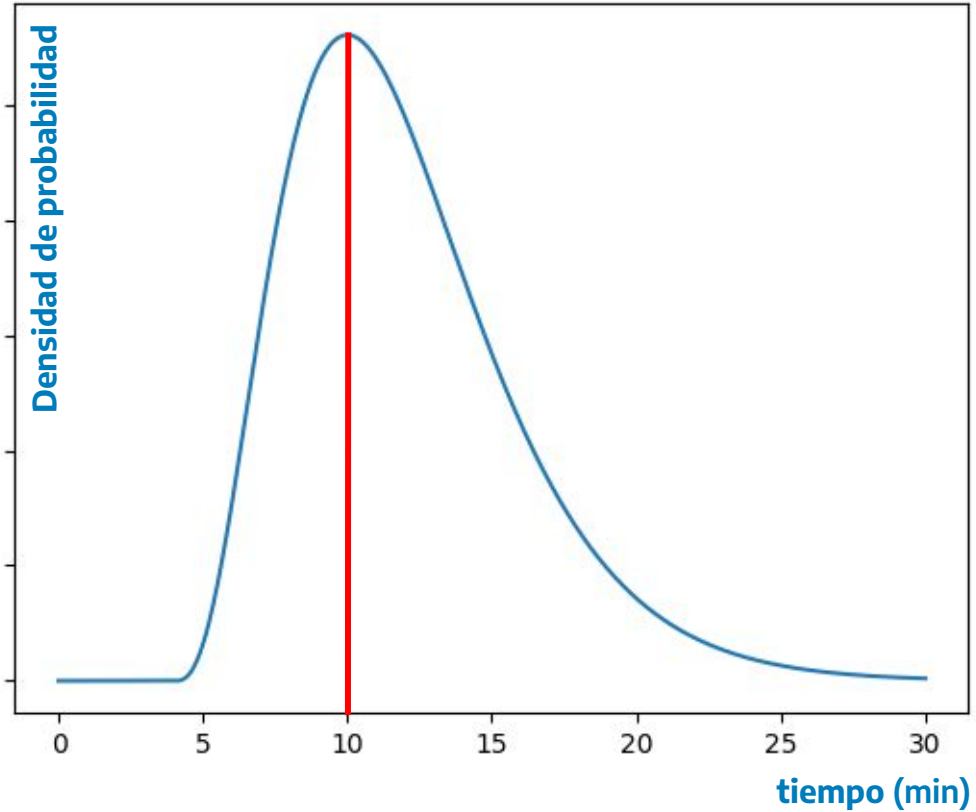
Ejemplo de PDF: tiempo de llegada

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



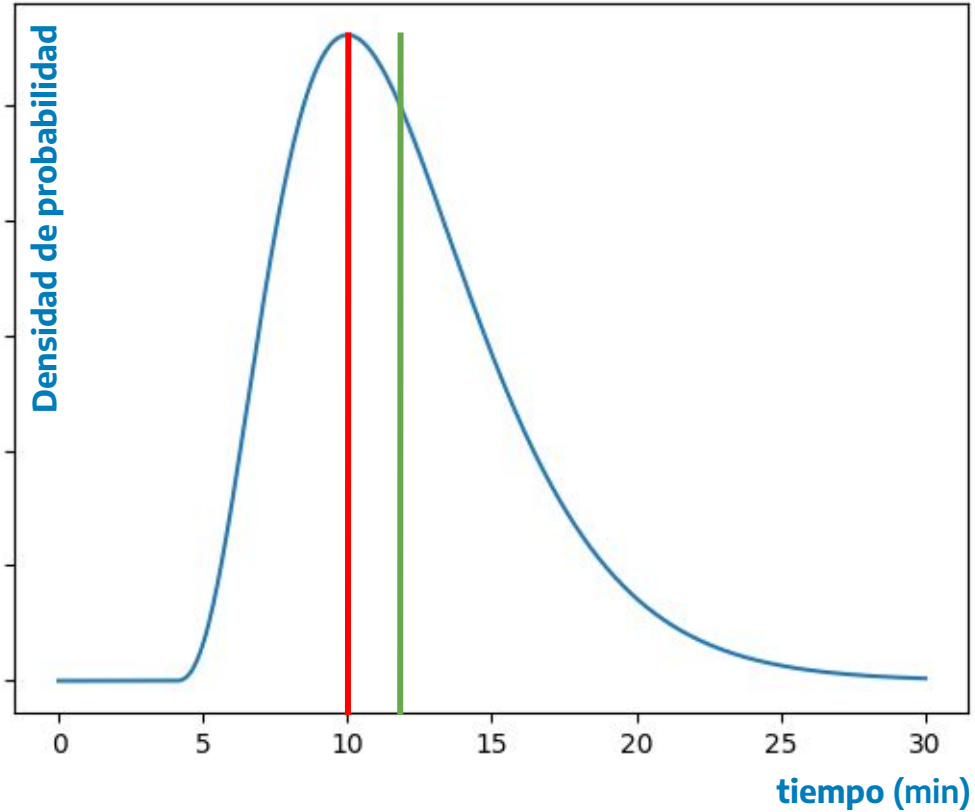
Un ejemplo ficticio: tiempo de llegada

- 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



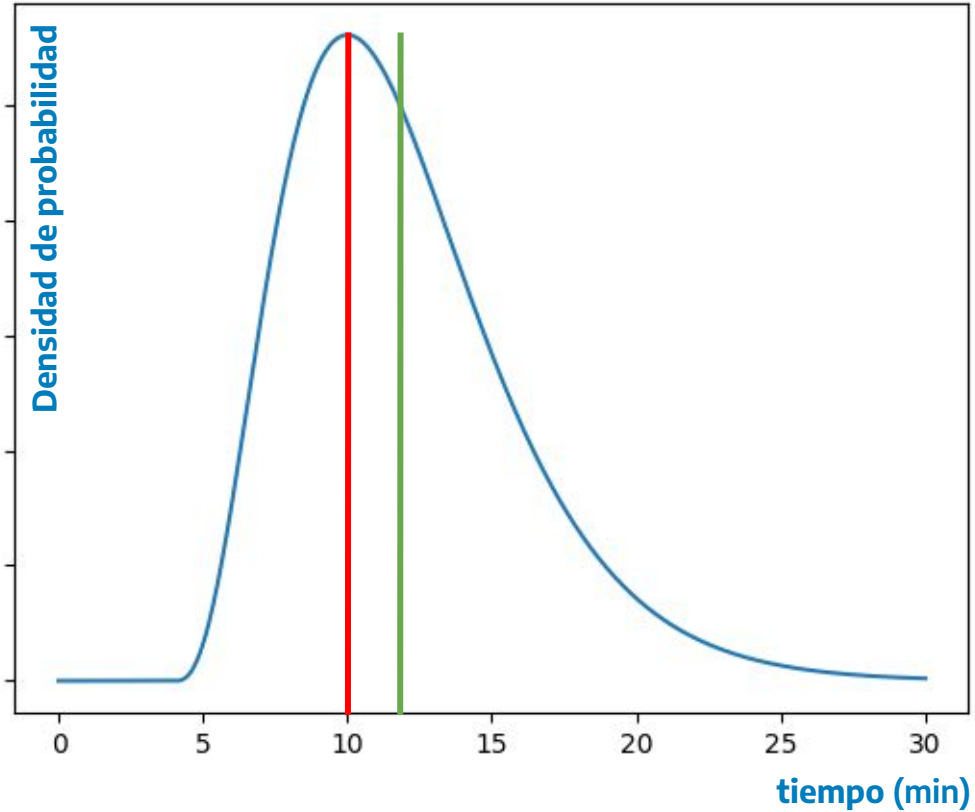
Un ejemplo ficticio: tiempo de llegada

- 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?



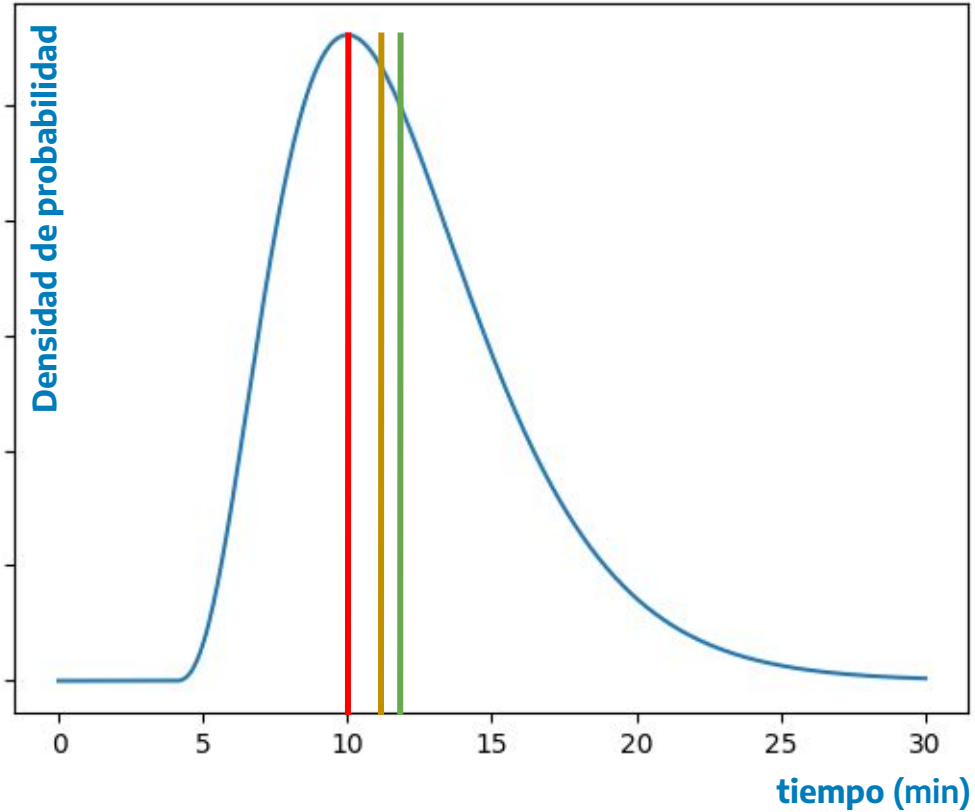
Un ejemplo ficticio: tiempo de llegada

- 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



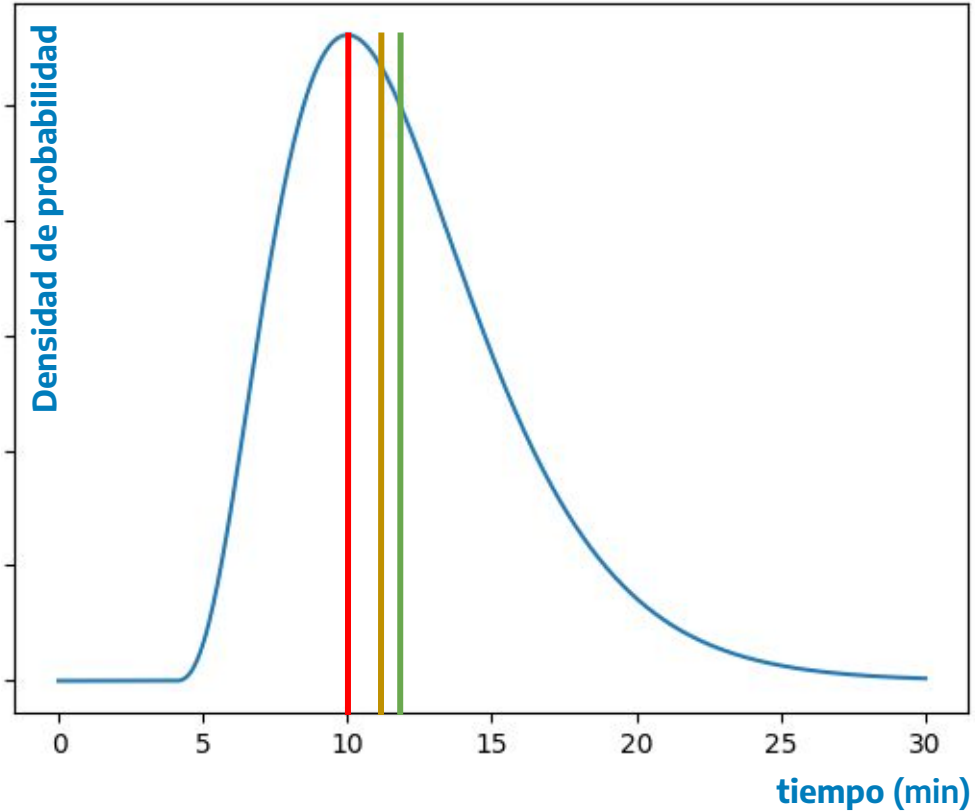
Un ejemplo ficticio: tiempo de llegada

- 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?



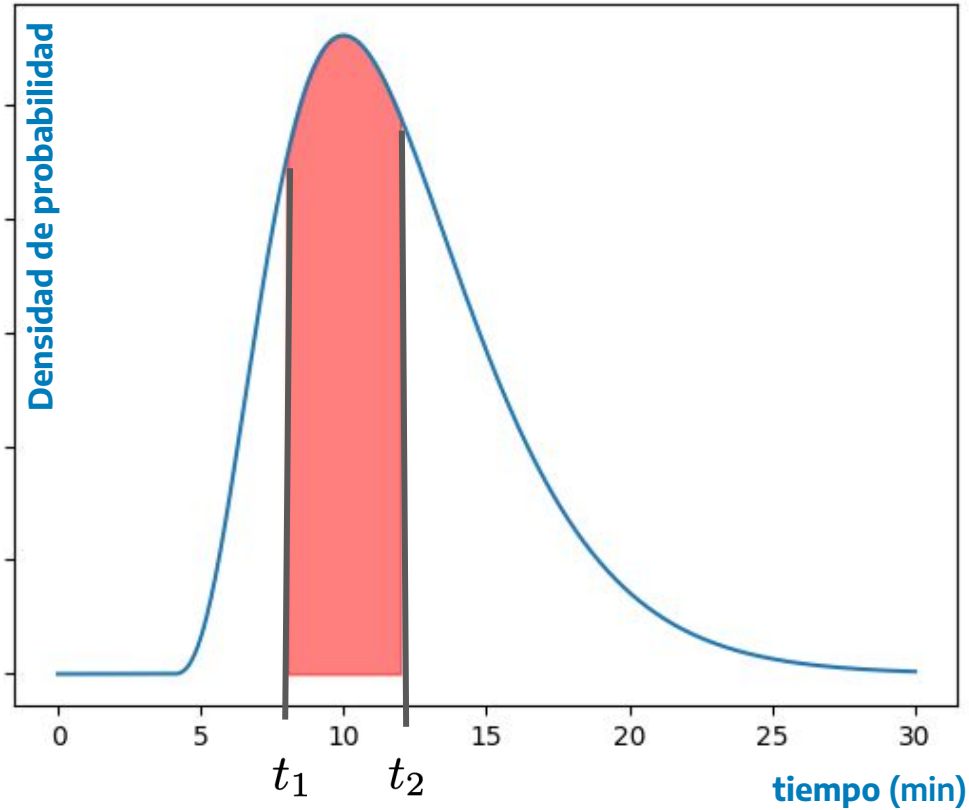
Un ejemplo ficticio: tiempo de llegada

- 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)



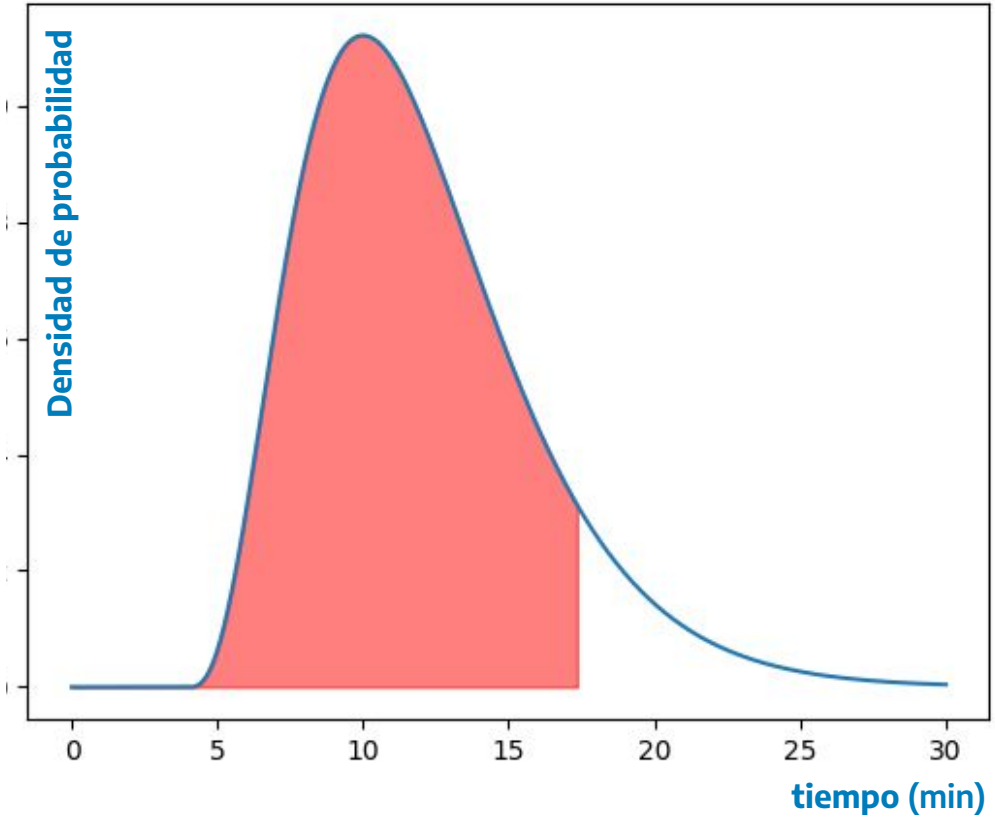
Un ejemplo ficticio: tiempo de llegada

- 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**



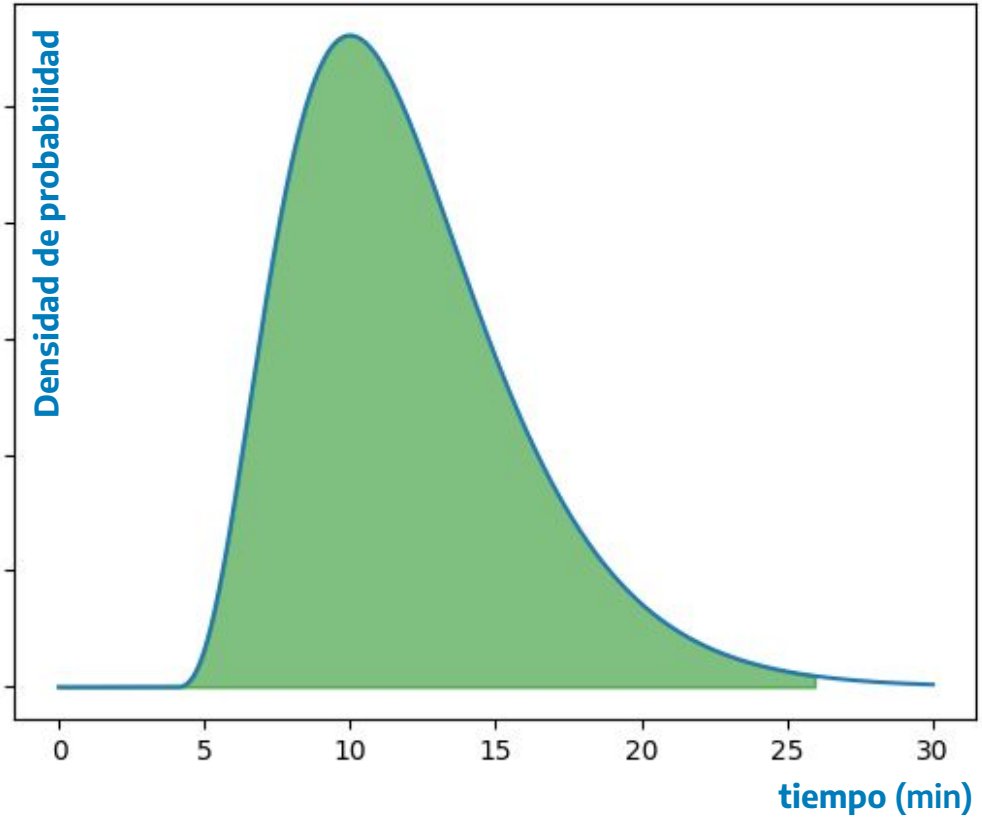
Un ejemplo ficticio: tiempo de llegada

- 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)



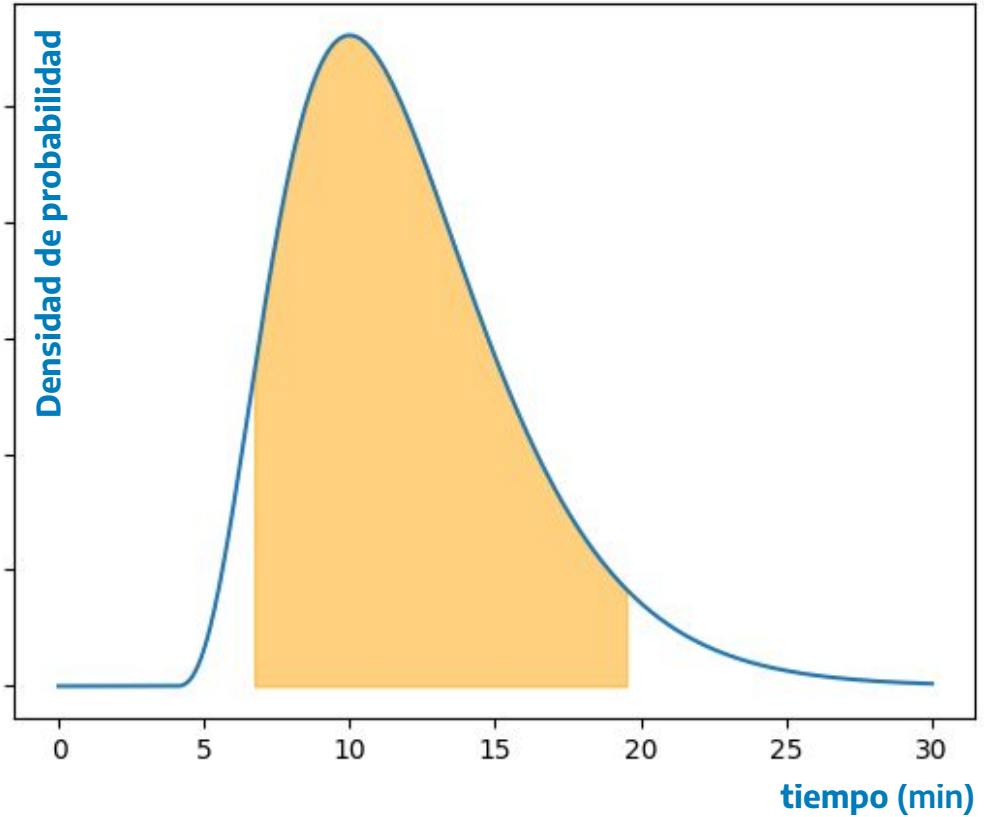
Un ejemplo ficticio: tiempo de llegada

- 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% de las veces, tengo que buscar la duración que contiene 99% del área bajo la curva
- El margen es de 14 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

Teorema del límite central: 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.**

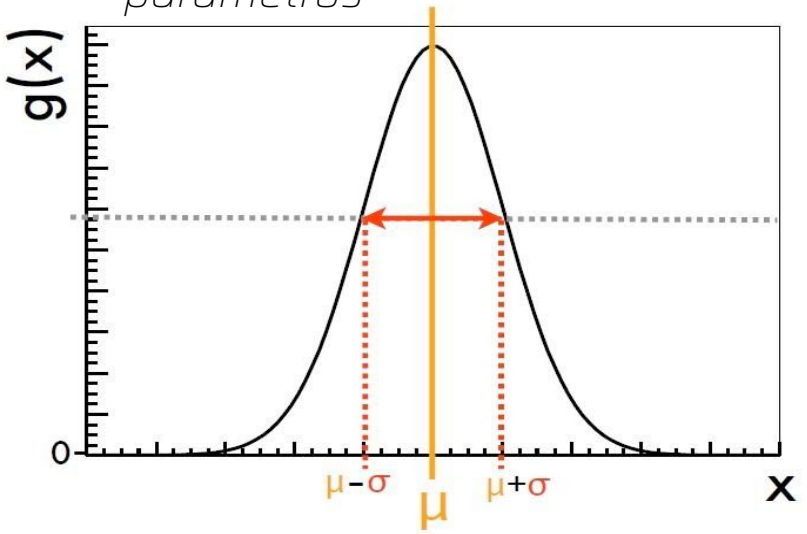
Distribución Normal o Gaussiana

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)

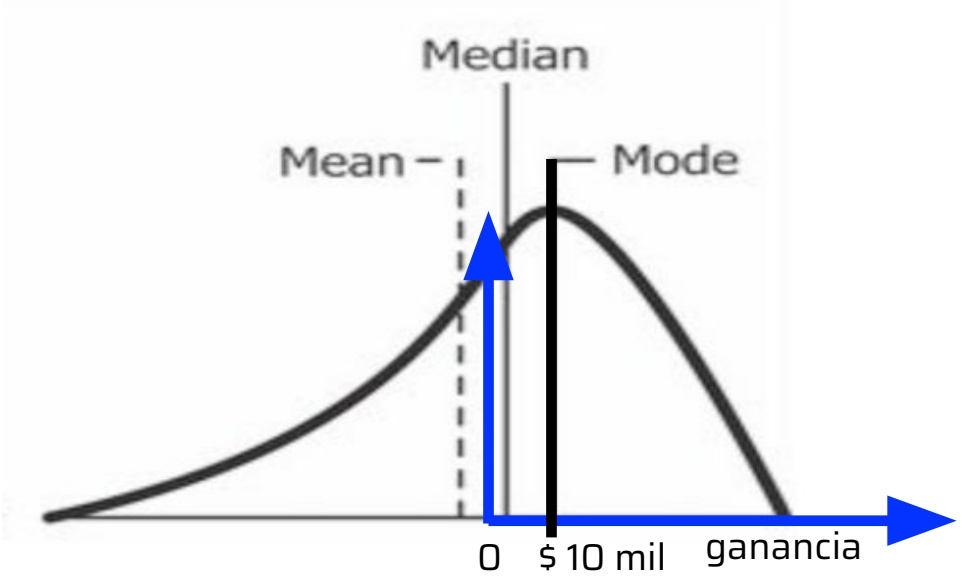
Toda la forma de la PDF contiene información útil (por ejemplo, múltiples picos)

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

Información codificada en una pdf

Ejemplo: pdf sesgado

- 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 probabilístico/estocástico

Probabilidad con más variables

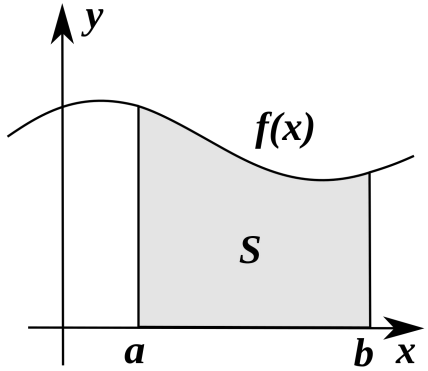
- **Probabilidad conjunta P(x,y)**
- **(probabilidad de que ocurran dos cosas)**
  - Si los sucesos son independientes P(x,y) = P(x) x P(y)
- **Probabilidad marginal P(x)**
  - Probabilidad de que ocurra x, independientemente de y
  - Si es discreta: $$P(x) = \sum_y P(x, y)$$
  - Caso continuo: $$p(x) = \int p(x, y) dy$$

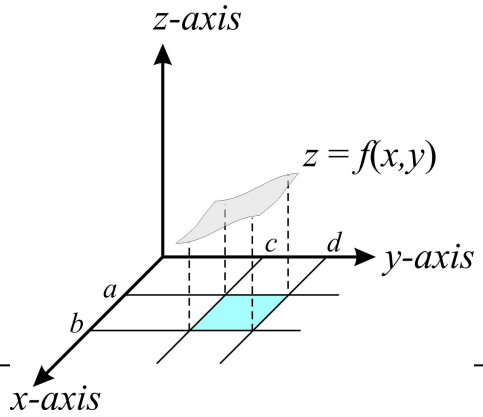
Área bajo la curva

- **El área bajo la curva se denomina integral**
  - Notación habitual $S = \int_a^b f(x)dx$
  - Si el límite de integración es una variable, la integral es también una nueva función

$$S(z) = \int_0^z f(x)dx$$

  - Si la función es sobre 2 dimensiones, el área se convierte en un volumen, y así sucesivamente
  - se utiliza prácticamente en todas partes en ML





Probabilidad con más variables

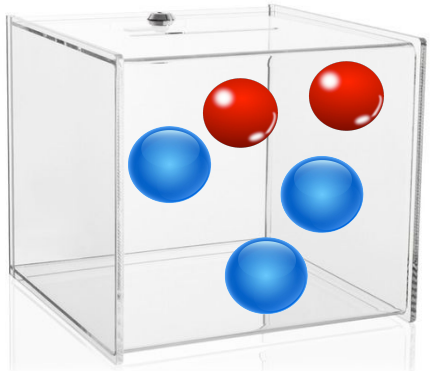
- **Probabilidad conjunta P(x,y)**
  (probabilidad de que ocurran dos cosas)
  - Si los sucesos son independientes P(x,y) = P(x) x P(y)
- **Probabilidad marginal P(x)**
  - Probabilidad de que ocurra x, independientemente de y
- **Probabilidad condicional P(x|y)**
  - Probabilidad de que ocurra x, dado que ocurre y (es decir, para un valor dado de y)
  - Si los sucesos son independientes P(x|y) = P(x)

Probabilidad Condicional

- Probabilidad de que ocurra x, dado que ocurre y (es decir, para un valor dado de y): P(x|y)
- Si los sucesos son independientes P(x|y) = P(x)

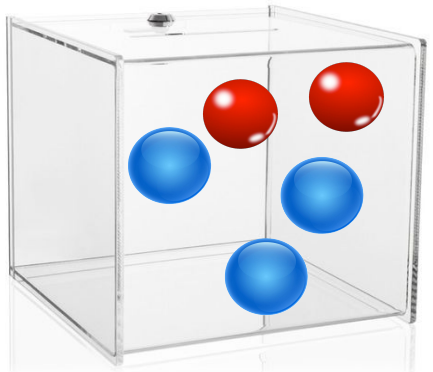
Probabilidad Condicional

- Probabilidad de que ocurra x, dado que ocurre y (es decir, para un valor dado de y): P(x|y)
- Si los sucesos son independientes P(x|y) = P(x)
- En caso contrario, ¡no!



Probabilidad Condicional

- Probabilidad de que ocurra x, dado que ocurre y (es decir, para un valor dado de y): P(x|y)
- Si los sucesos son independientes P(x|y) = P(x)
- En caso contrario, ¡no!
  - P(rojo)=0,4
  - P(rojo|rojo)=0,25
  - P(rojo|azul) = 0,5
- En general P(x,y)=P(x|y)P(y)



Un ejemplo detallado: enfermedad en mujeres y hombres

	Hombre	Mujer
Enfermos	200	100
Sanos	300	400

$P(X,Y)$, X = [H,M], Y = [E, S]

La frecuencia nos da (una estimación de) la probabilidad

$P(H,E) = 200/1000 = 0.2$ (probabilidad conjunta)

Un ejemplo detallado: enfermedad en mujeres y hombres

	Hombre	Mujer
Enfermos	200	100
Sanos	300	400

$P(X,Y)$, X = [H,M], Y = [E, S]

La frecuencia nos da la probabilidad

$P(H,E)$ =200/1000 = 0.2

Probabilidades marginales

- $P(E)$ = 300/1000 = 0.3
- $P(H)$ = 500/1000 = 0.5

Probabilidad condicional

$P(E|H)$ = 200/500 = 0.4

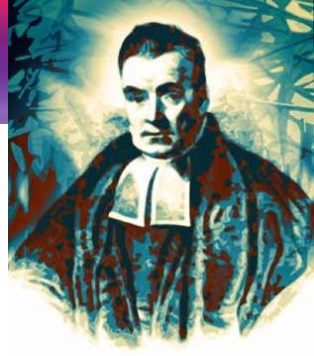
Aplicando la regla

$$P(\mathrm{E}|\mathrm{H}) = \frac{P(\mathrm{E},\mathrm{H})}{P(\mathrm{H})}$$

= 0.2/0.5 = 0.4

(Se extiende fácilmente a distribuciones de probabilidad!)

Estadística bayesiana



- Probabilidad frecuentista: frecuencia con la que ocurren los acontecimientos (¡necesita muchas repeticiones!)
- Probabilidad bayesiana: grado de creencia
- Enfoque potente: inferencia, confianza, uso de información previa, actualizaciones
- Muy utilizado en física y, más recientemente, en IA
- Teorema de Bayes:
  - A partir de las probabilidades condicional y marginal

$$P(\mathrm{X},\mathrm{Y}) = P(\mathrm{X}|\mathrm{Y}) \times P(\mathrm{Y})$$

$$P(\mathrm{X},\mathrm{Y}) = P(\mathrm{Y}|\mathrm{X}) \times P(\mathrm{X})$$

$$P(\mathrm{X}|\mathrm{Y}) = \frac{P(\mathrm{Y}|\mathrm{X})P(\mathrm{X})}{P(\mathrm{Y})}$$

Ejemplo: ¿Estoy infectado?

- La prueba para detectar una enfermedad es sensible en un 90%: tasa de verdaderos positivos (TPR) = 0,9
- Es específica en un 80%: Tasa de verdaderos negativos (TNR) = 0,8 (FPR = 0,2)
- Prevalencia (fracción de infectados): 5%
- Entonces, ¿si doy positivo, significa que hay un 90% de probabilidades de que esté infectado?

$$P(\text{Infectado}|\text{Positivo}) = \frac{P(\text{Positivo}|\text{Infectado})P(\text{Infectado})}{P(\text{Positivo})}$$

$$= \frac{P(\text{Positivo}|\text{Infectado})P(\text{Infectado})}{P(\text{Positivo}|\text{Infectado})P(\text{Infectado}) + P(\text{Positivo}|\text{NoInfectado})P(\text{NoInfectado})}$$

$$= \frac{0.9 \times 0.05}{0.9 \times 0.05 + 0.2 \times 0.95} \simeq 19\%$$

El caso Sally Clark

- 1er hijo fallecido a las 8 semanas (síndrome de muerte súbita del lactante - SMSL)
- 2º hijo fallecido a las 11 semanas
- Probabilidad de SMSL: 1/8543
- Probabilidad de 2 fallecimientos ~ 1/73 millones
- ¡Sally fue condenada y encarcelada sólo basándose en estadísticas!

El caso Sally Clark

- **Error estadístico 1: ¡probabilidades no independientes!**

$$P(2\ SMSL) = P(1\ SMSL|1\ SMSL) \neq P(1\ SMSL) \times P(1\ SMSL)$$

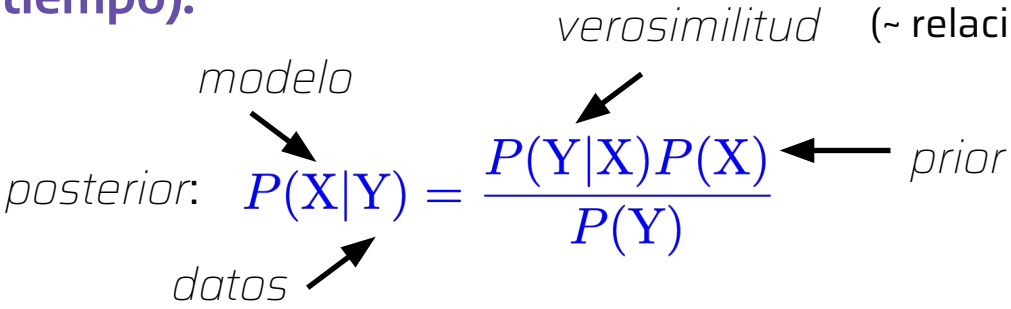
- **Error estadístico 2: lo que importa es P(Culpable|2 Muertes)**
  - Hay que tener en cuenta P(2 Asesinatos)
  - Probabilidades estimadas de que sea culpable:

$$P(2\ SMSL)/P(2\ Asesinatos) \sim 4.5{:}1 - 9{:}1$$

- **¡Sally fue liberada!**

Modelización de datos

- **La estadística bayesiana es extremadamente poderosa para modelizar datos (por ejemplo, el modelo de los precios de las acciones en función del tiempo).**



En la práctica $P(\mathrm{X}|\mathrm{Y}) \propto P(\mathrm{Y}|\mathrm{X})P(\mathrm{X})$

Proporcionalidad constante derivada de la normalización (las probabilidades suman 1)

- Maximizando la posterior con respecto a *X* se obtiene el mejor modelo (dentro de una familia) que describe los datos
- Se obtiene el pdf completo de *X*: forma, percentiles / barras de error....

Para llevarse a casa

- **Métodos bayesianos son muy poderosos**
  - inferencia
  - usar información a priori
  - confianza
  - muy usados en aprendizaje automático: regresión, barras de error, etc.
- **Aplicaciones del mundo real utilizan todos los conceptos de estas clases**
- **Ahora, cuando oigas hablar de vectores, matrices, minimización, integrales, probabilidades, pdf, percentil, TPR, FPR, TNR, probabilidad condicional, inferencia bayesiana, etc., recuerda esta clase**

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?!**

Preguntas son bien venidas!

“Todas las preguntas son un llamado a conocer el mundo. No existen las preguntas tontas”

Carl Sagan

(El mundo y sus demonios. La ciencia como una luz en la oscuridad)

Agregar?

Generación de números aleatórios

Barras de error: determinación, significado, representación gráfica

¿Algo de álgebra lineal?

Prolegomena/Foreword

- Welcome (few mins) [happy to take questions in Portuguese or Spanish]
- [boring mathematics, will take is to the delighting AI world][not codes nor calculations, but concepts]
- [we know you are high level executives][many connected to computer science]
- Describe the course, dynamics, main objectives, etc.: next class
- Different dynamics, 1h, no break, questions (even more informal)
- Here, prep session, refresh, concepts and jargon (Feynmann)
- Algorithms/codes/mathematics/numerics
- Build mathematical models (this is what physicists are good at!) [mathematicians x engineers]
- Understand the very basics mathematical concepts (not to be afraid) [so that we don't stop each time, but will be recap/repeated]
- Build up intuition / intuitive meaning (non rigorous)
- [we do all this math unconsciously all the time][AI/ML also helps us understand how we think!]
- Mostly vocabulary, “machine gun of concepts and nomenclature, so they will appear familiar when you hear them again in the AI context.

Adaptado de

C. Bonifazi, I. Pedrón, F. Aguero

Distribuciones con variables aleatorias discretas

Variable aleatoria discreta: Una variable X es discreta si los valores que toma son numerables (finitos o infinitos). A esos valores los podemos representar con:

$$\{x_1, x_2, x_3, \ldots, x_n\}$$

En este caso tenemos n valores que se comportan de alguna forma siguiendo alguna distribución.

Ejemplos:

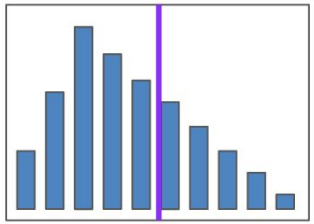
- Número de acciones vendidas de una empresa
- Número de errores de transmisión en un proceso
- Número de veces en que sale cara en una secuencia de tiradas

Parámetros Estadísticos

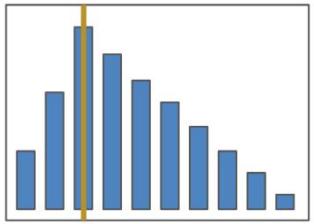
Media: $$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Mediana: $$\mathrm{M}_e(x) = \begin{cases} x_{(n+1)/2} & \text{si } n \text{ es impar} \\ \dfrac{x_{n/2} + x_{n/2+1}}{2} & \text{si } n \text{ es par} \end{cases}$$

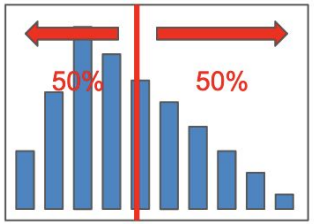
Moda: $\mathrm{M}_o(x) =$ valor que mas se repite de un conjunto de datos



Media



Moda



Mediana

Distribuciones con variables aleatorias discretas

Problema 1: Lanzamos un dado perfecto 240 veces, anotamos el resultado obtenido en la cara superior obteniendo los siguientes resultados:

Cara superior	1	2	3	4	5	6
Núnero de veces	40	39	42	38	42	39

Como sabemos que fueron 240 veces, podemos escribir una tabla de frecuencias relativas de la siguiente forma:

Cara superior	1	2	3	4	5	6
Frecuencia relativa	0,1667	0,1625	0,1750	0,1583	0,1750	0,1625

Esta **frecuencia relativa** la podemos interpretar como la **función distribución de probabilidad,** o sea es la probabilidad de que la variable tome dicho valor

Distribuciones con variables aleatorias discretas

La **función de distribución de probabilidad** pierde la información del número de medidas/valores que fueron utilizados en el estudio realizado. Cuanto más datos sean utilizados más precisa será la determinación de las probabilidades.

En el caso del problema, podemos calcular los resultados esperados:

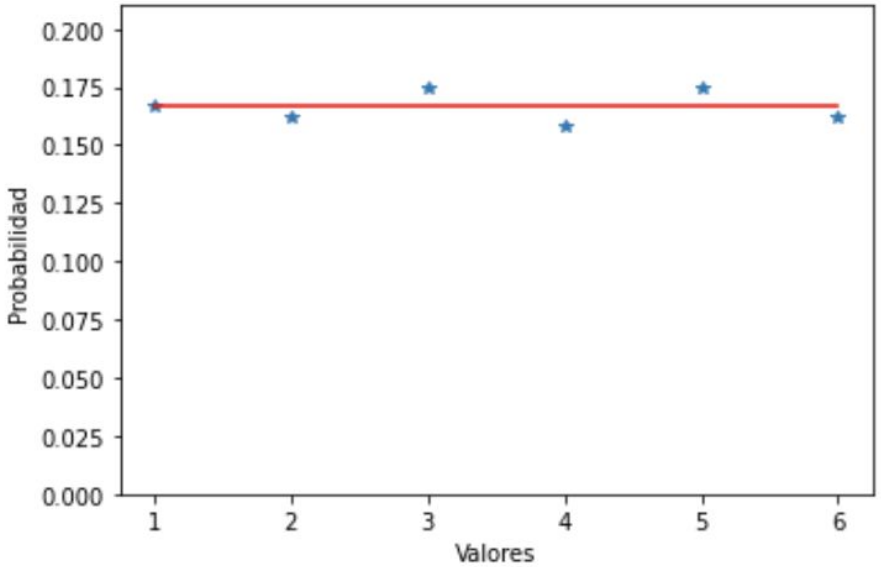
Cara superior	1	2	3	4	5	6
Frecuencia relativa	0,1667	0,1625	0,1750	0,1583	0,1750	0,1625
Probabilidad	1/6	1/6	1/6	1/6	1/6	1/6

1/6 = 0,16666 ....

Distribuciones con variables aleatorias discretas

La **función de distribución de probabilidad** pierde la información del número de medidas/valores que fueron utilizados en el estudio realizado. Cuanto más datos sean utilizados más precisa será la determinación de las probabilidades.

En el caso del problema, podemos calcular los resultados esperados:

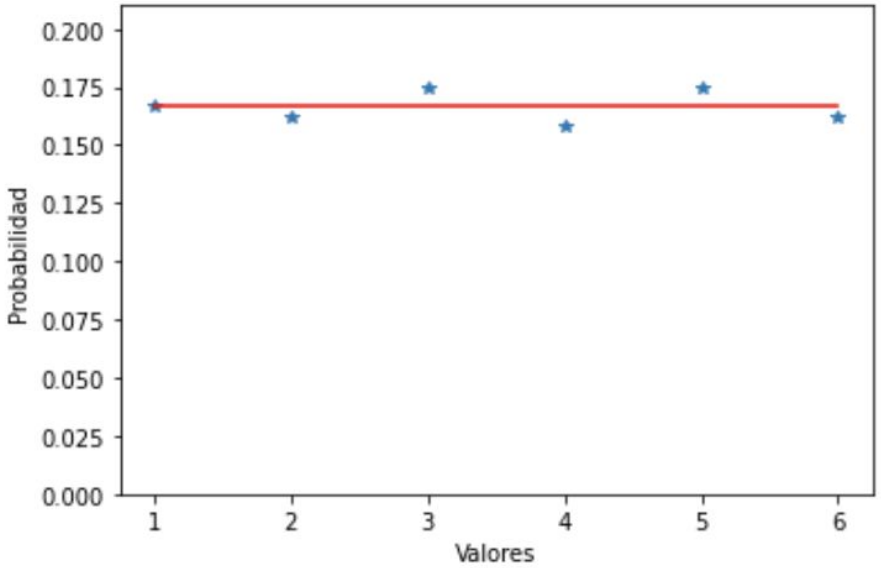


= 1/6

Distribuciones con variables aleatorias discretas

La función de distribución de probabilidad pierde la información del número de medidas/valores que fueron utilizados en el estudio realizado. Cuanto más datos sean utilizados más precisa será la determinación de las probabilidades.

En el caso del problema, podemos calcular los resultados esperados:



Distribución Uniforme

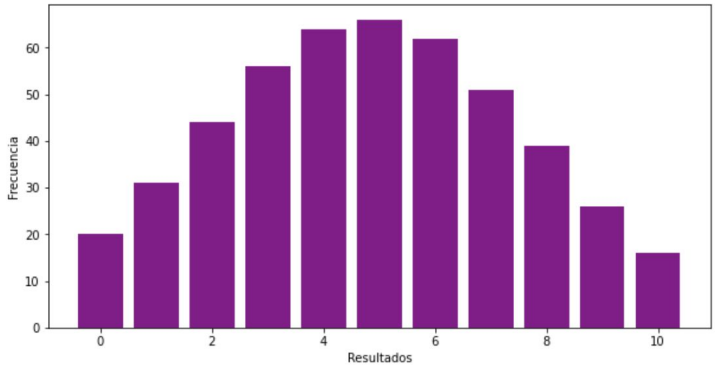
= 1/6

Si realizamos 1000 lanzamientos de un dado perfecto, ¿cuántas veces esperamos sacar el número 5?

¿ y el número 3?

Distribuciones con variables aleatorias discretas

Problema 2: Los resultados obtenidos por los 475 alumnos de primer año se obtuvieron las siguientes notas con sus frecuencias.



Resultado	F. relativa
0	0,042
1	0,065
2	0,093
3	0,118
4	0,135
5	0,139
6	0,130
7	0,107
8	0,082
9	0,055
10	0,034

Caracteristicas:

- Es simétrica
- Media = Mediana = Moda = 5

Distribución Normal

Las observaciones con más frecuencia o probabilidad están alrededor del valor central. O sea, las observaciones con menos frecuencia o probabilidad se encuentran lejos del valor central.

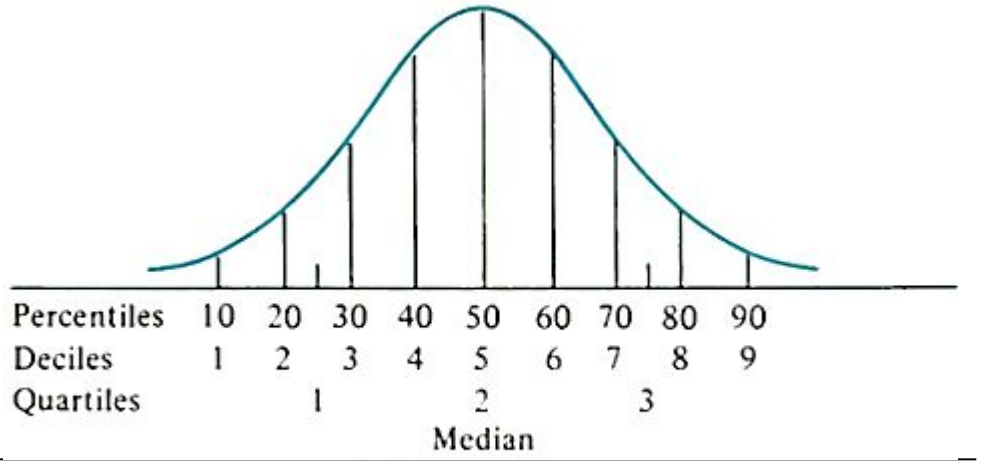
Variables aleatorias discretas

Otra herramienta útil para cuantificar la dispersión de los datos son los cuantiles.

Cuantil: el cuantil de orden p de una distribución (con 0 < p < 1) es el valor de la variable x□ que marca un corte de modo que una proporción p de los datos es menor o igual que x . Por ejemplo, el cuantil de orden 0,93 dejaría un 93% de valores por debajo, y el cuantil de orden 0,50 se corresponde con la mediana de los datos.

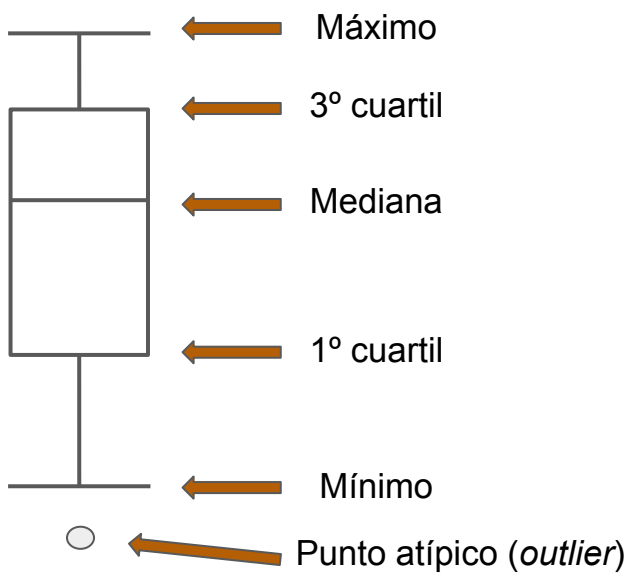
Los cuantiles más usuales son:

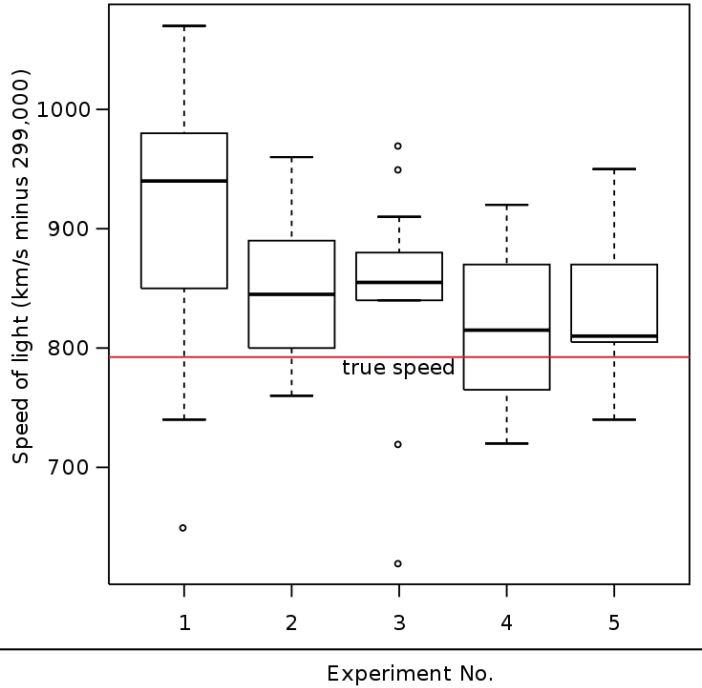
- Cuartiles: dividen a los datos en cuatro partes (son los cuantiles 0,25; 0,50 y 0,75).
- Percentiles: dividen a los datos en 100 partes.



Box plot (diagrama de caja)

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:

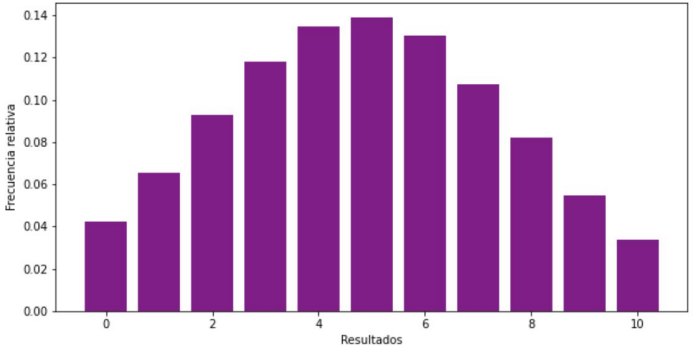


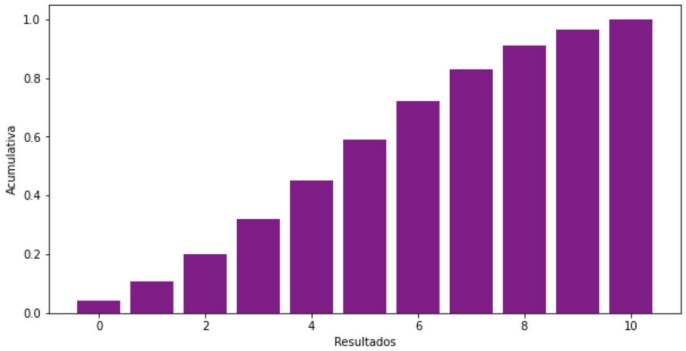


Distribuciones con variables aleatorias discretas

Función acumulativa:

$$F(x) = P(X \leq x_i) = \sum_{x_i} p_i$$





Resultado	F. relativa	Acumul.
0	0,042	0,042
1	0,065	0,107
2	0,093	0,200
3	0,118	0,318
4	0,135	0,453
5	0,139	0.592
6	0,130	0,722
7	0,107	0,829
8	0,082	0,911
9	0,055	0,966
10	0,034	1

Distribución de Probabilidad discreta

Definición: Si X es una variable aleatoria discreta con valores:

$$\{x_1, x_2, x_3, \ldots, x_n\}$$

la función $f(x)$ para cada x_i se denomina función de distribución de probabilidad de x, si y sólo si sus valores $f(x)$ satisfacen las siguientes condiciones:

$$f(x) \geq 0$$

a) $$0 \leq f(x_i) \leq 1$$

b) $$\sum_i f(x_i) = 1$$

c)

La distribución acumulada $F(x)$ de una variable aleatoria discreta X, cuya distribución de probabilidad es $f(x)$ está dada por:

$$F(x) = P(X \leq x) = \sum_{y \leq x} f(y)$$

Parámetros Estadísticos

Esperanza matemática: Sea X una variable aleatoria con distribución de probabilidad $f(x)$. La media o valor esperado de X es:

$$\mu = E(X) = \sum_x xf(x)$$

Varianza: Medida del cuadrado de la distancia promedio entre la media y cada elemento de la población. Si X es una variable aleatoria con una distribución de probabilidad, $f(x)$, y media μ. La varianza de X es calculada por medio de:

$$\sigma^2 = E\left[(X-\mu)^2\right] = \sum_x (x-\mu)^2 f(x)$$

Desviación Estándar:

- Es una medida de dispersión de la misma dimensión de la variable y que representa por medio de la letra σ.
- Raíz cuadrada positiva de la varianza; una medida de la dispersión, expresada en las mismas unidades que los datos originales y no en las unidades cuadradas de la varianza.

Distribución Uniforme

Definición: Si la variable aleatoria X asume los valores x_1, x_2,.....x_k, con iguales probabilidades, entonces la distribución discreta uniforme es:

$$f(x;k) = \frac{1}{k} \qquad \mu = \frac{\sum_{i=1} x_i}{k} \qquad \sigma^2 = \frac{\sum_{i=1}(x_i - \mu)^2}{k}$$

$$x = x_1, x_2, \ldots, x_k$$

De esta forma, para el caso genérico donde x_1 = a y x_k = b, con iguales probabilidades para todos los valores tenemos: k = b - a +1

$$\mu = \frac{k-1}{2} + a \qquad \sigma^2 = \frac{k^2 - 1}{12}$$

- Ejemplo de los dados perfectos (Problema 1):

a = 1, b = 6 ⇒ k = b - 1 + 1 ⇒ k = 6; μ = (6 -1)/2 + 1 = 5/2 + 1 = 3,5

σ = (6^2 - 1)/12 = 35/12 = 2,9

Proceso de Bernouilli

Definición: Tenemos dos procesos excluyentes, que llamaremos de éxito (E) y fracaso (F), y que son independientes entre sí. Las probabilidades de éxito y fracaso son constantes.

P(E) = p a la probabilidad de éxito
P(F) = q a la probabilidad de fracaso

Definimos una variable aleatoria tal que

$$\begin{cases} x_i = 1 & \text{si el resultado es de éxito} \\ x_i = 0 & \text{si el resultado es de fracaso} \end{cases}$$

entonces: P(E) = P (X=1) = p

P(F) = P (X=0) = q

$\Rightarrow q = 1 - p$

Valor medio: $\mu = p$

Varianza: $\sigma^2 = p \cdot q$

Distribución Binomial

Definición: Consideremos realizar n veces un proceso de Bernoulli de forma independiente y suponiendo que la probabilidad de éxito p permanece constante en cada uno de esos procesos. Definimos la variable aleatoria X como el número de éxitos resultantes en las n veces que realizamos el proceso. Entonces, X tendrá una distribución Binomial, dada por:

$$f(x) = \binom{n}{x} p^x q^{n-x} \qquad \text{si } x = x_i$$

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

Valor medio: $\mu = n \cdot p$ Varianza: $\sigma^2 = n \cdot p \cdot q$

Ejemplos

Problema 3: Teniendo una pregunta objetiva de 5 opciones, ¿cuál sería la probabilidad de que acierte y la probabilidad de que me equivoque al adivinar?

Problema 4: A partir de un estudio realizado por una asociación de conductores de autopista mostró que el 60% de los mismos utilizan el cinturón de seguridad correctamente. Si se selecciona una muestra de 10 conductores en una autopista. ¿Cuál es la probabilidad de que exactamente 7 de ellos lleven el cinturón de seguridad?

Ejemplos

Problema 4-bis: A partir de un estudio realizado por una asociación de conductores de autopista mostró que el 60% de los mismos utilizan el cinturón de seguridad correctamente. Si se selecciona una muestra de 10 conductores en una autopista. ¿Cuál es la probabilidad de que al menos 7 de ellos lleven el cinturón de seguridad?

Distribución de Poisson

Definición: Podemos entender la distribución de Poisson como un caso particular de la distribución Binomial en el que la media μ = $n.p$ es muy pequeña con respecto al número de pruebas. Algunos ejemplos:

- **el número de accidentes de tráfico en una ciudad por semana**
- **el número de llamadas que llegan a un centro de atención al cliente**
- **el número de emergencias que llegan al sector de urgencias de un hospital**

Como podemos ver, la probabilidad de que alguna de estas cosas suceda es muy pequeña y por otro lado, tenemos un valor muy grande de n, como por ejemplo el número de personas que tenemos en una ciudad. Distribución de los "sucesos raros".

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Valor medio: $\mu = \lambda$

Varianza: $\sigma^2 = \lambda$

Distribución de Poisson

Problema 5: Un analista de empresas ha pronosticado que el 3.5% de las pequeñas empresas van a quebrar en 2023. Para una muestra de 100 pequeñas empresas, estime la probabilidad de que al menos 3 de ellas entren en quiebra sabiendo que la predicción del experto es correcta.

Distribución de Probabilidad continua

Variable aleatoria continua: Se dice que una variable aleatoria X es continua si su conjunto de posibles valores es todo un intervalo (finito o infinito) de números reales.

Ejemplos:

- Tiempo de retraso con el que un alumno puede llegar al aula de clases
- El peso o altura de los pacientes en un centro hospitalario
- Cantidad de agua que hay en una botella

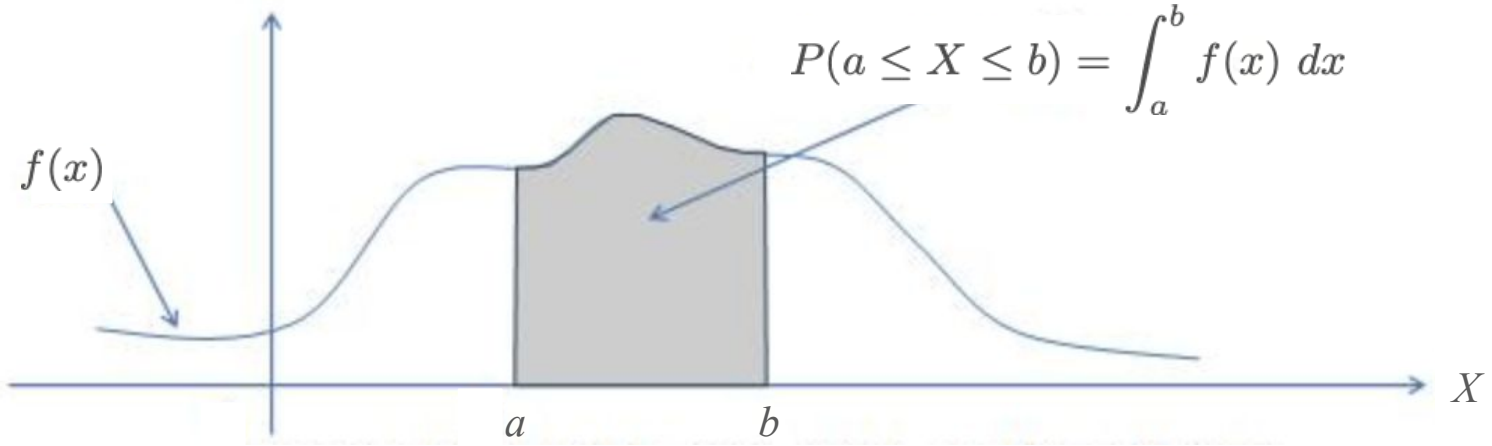
$$f(x) \geq 0 \quad \forall\, x \in \mathbb{R}$$

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

Función de densidad de una variable aleatoria continua: La función f(x) es una función de densidad de probabilidad de una variable aleatoria continua X, definida sobre el conjunto de los números reales, si:

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) = \int_a^b f(x)\, dx$$

Función de densidad de probabilidad continua

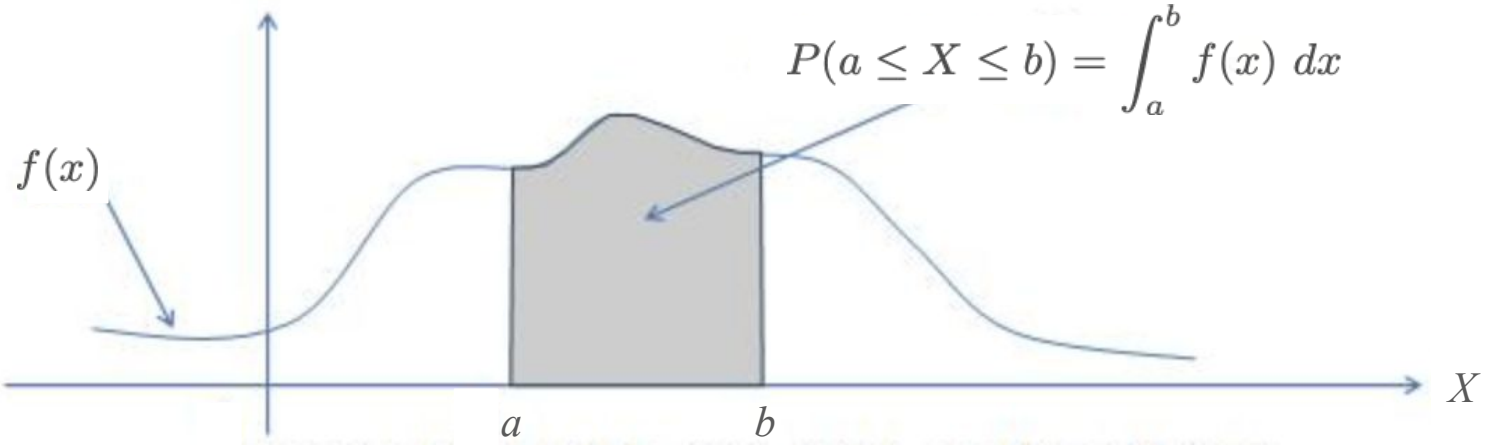


Área bajo la curva de f(x) en el intervalo [a,b]

Función de distribución acumulada: La distribución acumulada $F(x)$ de una variable aleatoria continua X, con una función de densidad $f(x)$ es:con una función de densidad $f(x)$ es:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

Función de densidad de probabilidad continua



Área bajo la curva de f(x) en el intervalo [a,b]

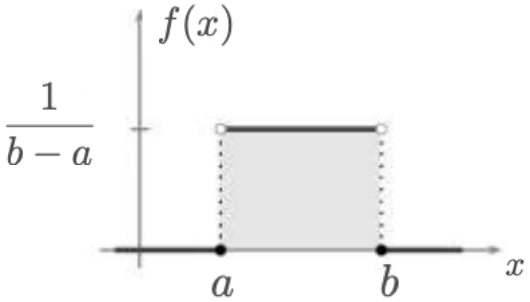
Valor esperado: $E(X) = \int_{-\infty}^{\infty} x f(x)\ dx$

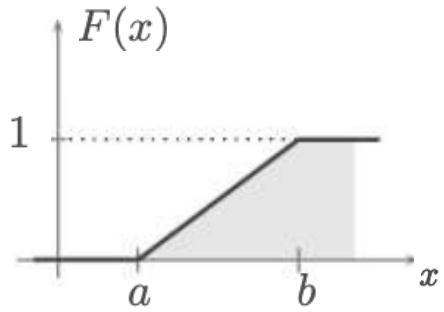
Varianza: $V(X) = \int_{-\infty}^{\infty} (x - E(x))^2 f(x)\ dx$

Distribución Uniforme

Definición: Diremos que una variable aleatoria continua X se distribuye uniformemente en un intervalo [a,b] cuando:

$$f(x) = \frac{1}{b-a} \quad \text{si } a \leq x \leq b$$





Valor medio: $\mu = \frac{a+b}{2}$

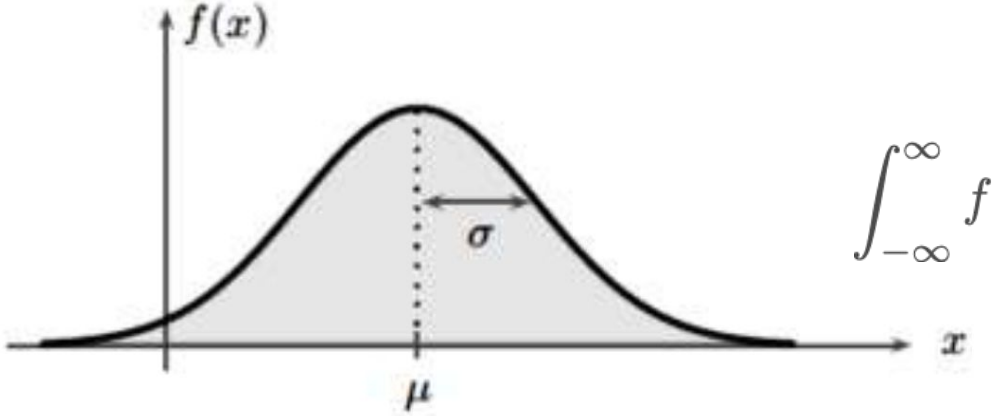
Varianza: $\sigma^2 = \frac{(b-a)^2}{12}$

Distribución Normal o Gaussiana

Definición: Decimos que una variable aleatoria continua X sigue una distribución normal con parámetros μ y σ^2 si:

$$X \to \mathrm{N}(\mu, \sigma^2) \qquad f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Como podemos v tros μ y σ^2, que coinciden co



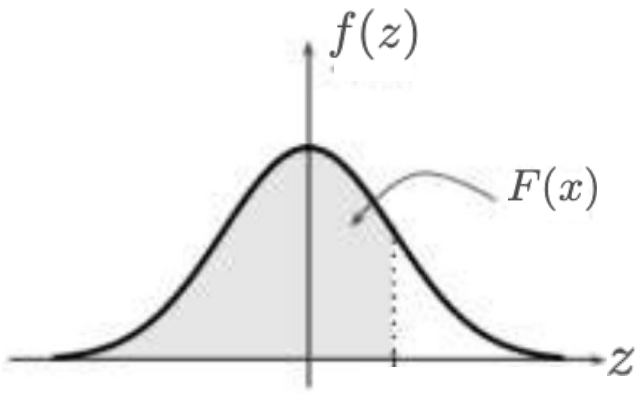
$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

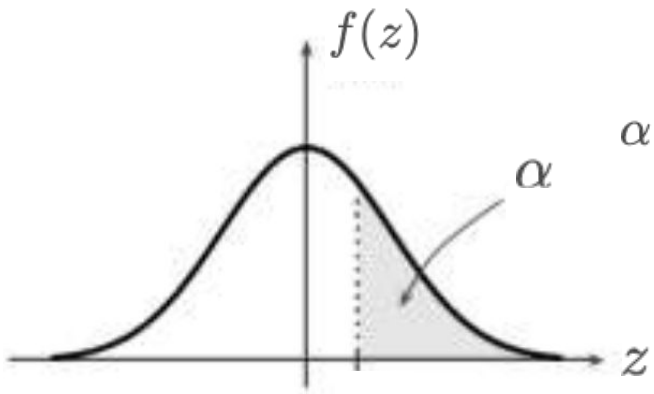
Distribución Normal o Gaussiana

Probabilidad acumulada de la Normal

$X \to \mathrm{N}(\mu, \sigma^2)$ $\quad Z = \frac{X-\mu}{\sigma}$ $\quad Z \to \mathrm{N}(0,1)$ Función estandarizada

$$F(x) = P(Z \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$





$\alpha = 1 - F(x)$

Distribución Normal o Gaussiana

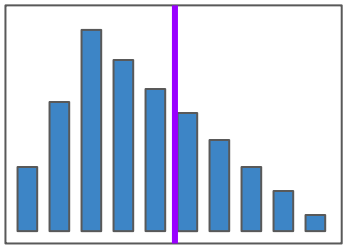
Problema 6: En una ciudad se estima que la temperatura máxima en el mes de junio sigue una distribución normal, con media 23ºC y desviación típica 5ºC. Calcular el número de días del mes en los que se espera alcanzar máximas entre 21ºC y 27º.

Variables aleatorias discretas

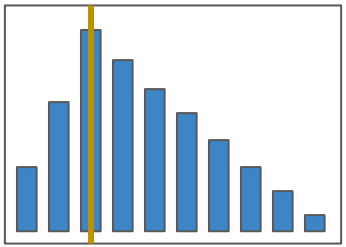
Moda: es el valor que más aparece en un set de datos, o el valor con mayor probabilidad en una distribución de probabilidad.

Ej: x = (1, 1, 3, 4, 5, 5, 6, 8, 8, 8, 11) **8**

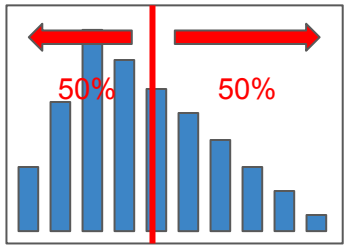
Los tres parámetros que vimos (media, mediana y moda) dan una idea del valor "central" de un conjunto de datos o distribución:



Media



Moda



Mediana

Variables aleatorias discretas

También hay parámetros estadísticos que nos permiten cuantificar qué tanto difieren los datos de los valores centrales (cuánta *dispersión* hay). La más usada es la **varianza**.

<u>Varianza:</u> se define como el valor medio del cuadrado de la desviación de los datos respecto de su promedio. Es decir,

$$\mathrm{Var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

También se suele usar la desviación estándar (o desvío estándar), que es notada con σ, que es simplemente la raíz cuadrada (positiva) de la varianza. Tiene la ventaja de estar en las mismas unidades que la variable (si x tiene unidades de metros, σ también).