



13 de febrero de 2020

# Estadística para Data Science

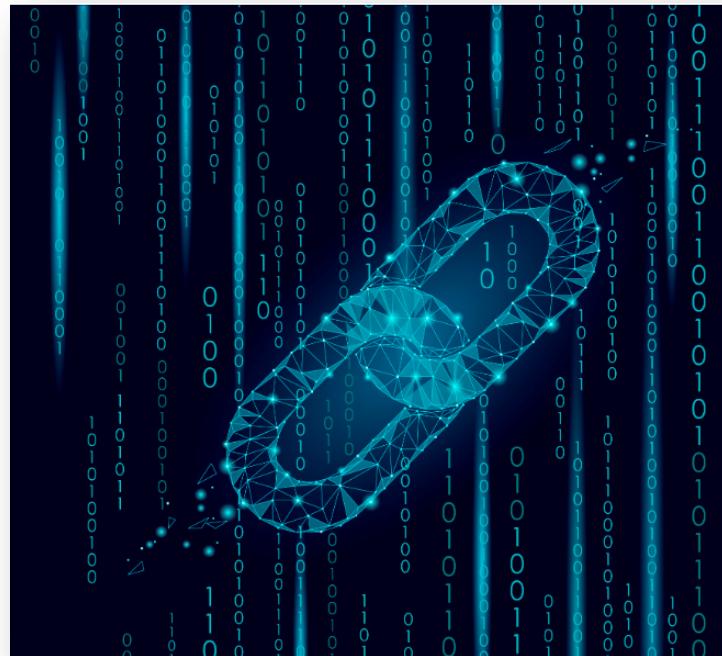
## Sesión 1: Introducción a la estadística

Jesús Hernando Corrochano

# ¿Quienes SOMOS?



- 1. Programa / Planificación**
- 2. ¿Qué es la estadística?**
- 3. Método Estadístico**
- 4. Establecer....**
- 5. Variables**
- 6. Conceptos**
- 7. Frecuencias**
- 8. Trabajo en Equipo**
- 9. Gráficos**
- 10.Trabajo en Equipo**





	Sesión 1 13/2	Sesión 2 20/2	Sesión 3 27/2	Sesión 4 5/3	Sesión 5 12/3	Sesión 6 19/3	Sesión 7 26/3	Sesión 8 2/3
Introducción a la estadística								
Introducción a la combinatoria y la probabilidad								
Estadística descriptiva								
Regresión y correlación								
Estadística inferencial								
Probabilidad Total. Teorema de Bayes. Test A/B								



# Estadística para Data Science

# El Dato como VALOR empresarial



# LOS 4 VECTORES DEL MUNDO DIGITAL



## INNOVACIÓN

Modulación de los cambios  
Aportación novedad continua

## EXPERIENCIA

El móvil como elemento central  
Contenido cambiante  
De productos a servicios

## OPERACIÓN

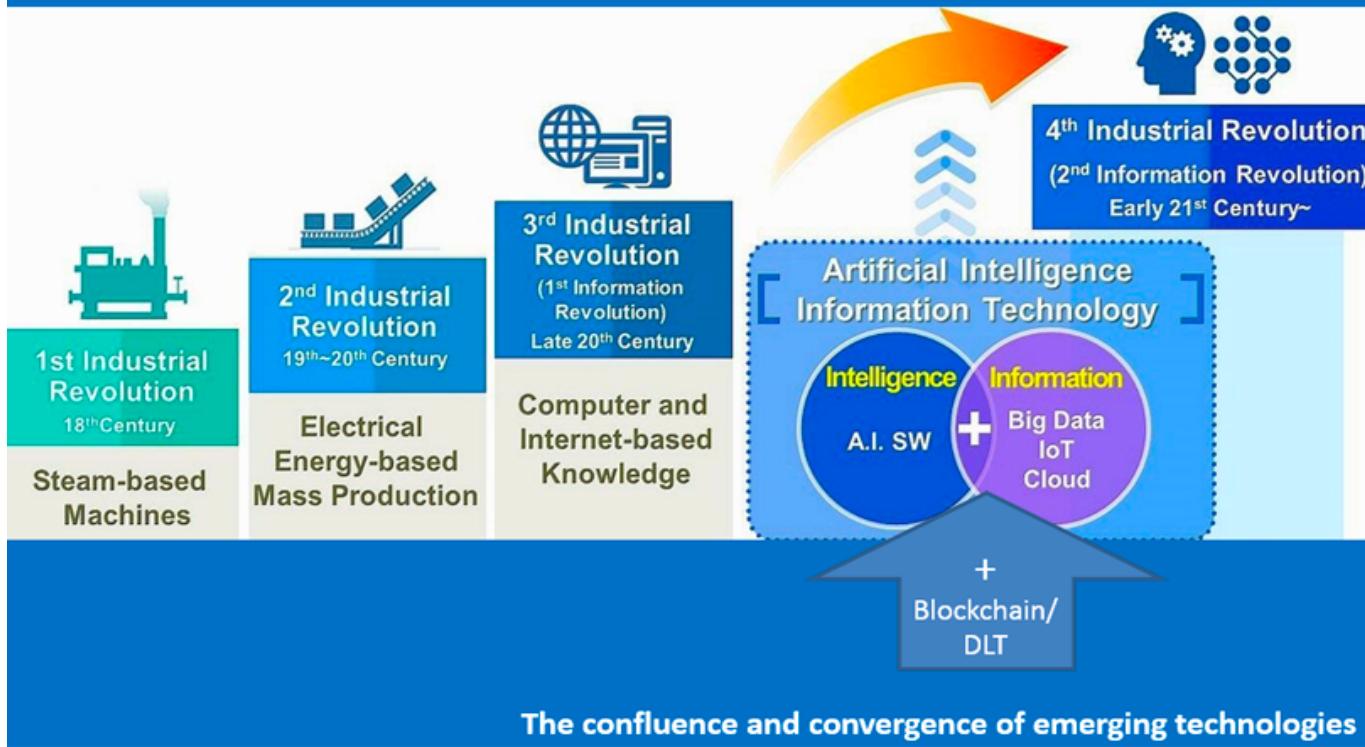
Optimización  
Simplicidad  
Agilidad  
Conocimiento

## MODELOS DE NEGOCIO

Estrategia Digital  
Reinvención Negocio  
Repensar modelos entrega  
Nuevos negocios digitales  
Reinvención de industrias

## ● The knowledge revolution?

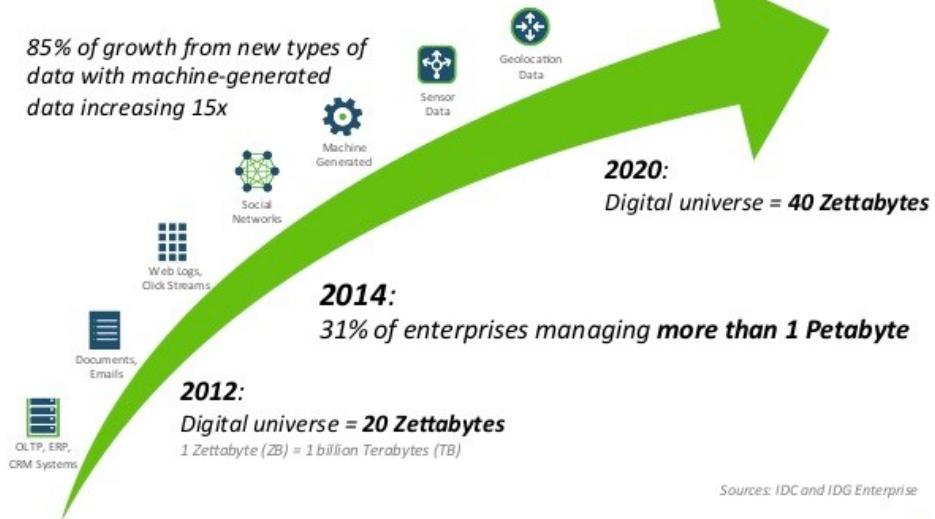
### The Fourth Industrial Revolution



## The knowledge revolution?

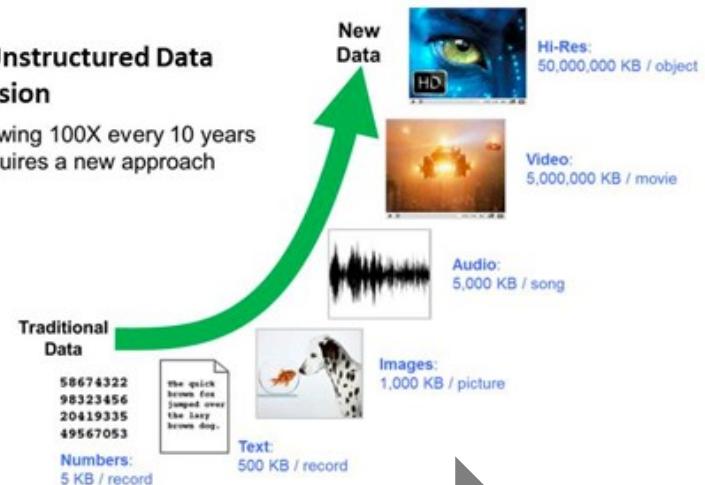
### Data Continues to Grow Sharply

85% of growth from new types of data with machine-generated data increasing 15x



### The Unstructured Data Explosion

- Growing 100X every 10 years
- Requires a new approach



Datos

Información

Conocimiento

# The knowledge revolution?

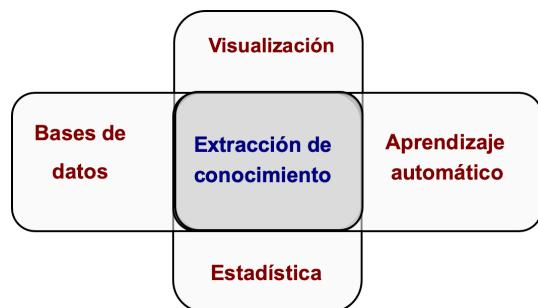
La sociedad de la información e Internet han generado una explosión de datos

- No hay suficientes personas que pueda analizar tal cantidad de datos
- Potencia de computación disponible
- El desarrollo de software es un cuello de botella

Extraer conocimiento a través de ejemplos es atractivo

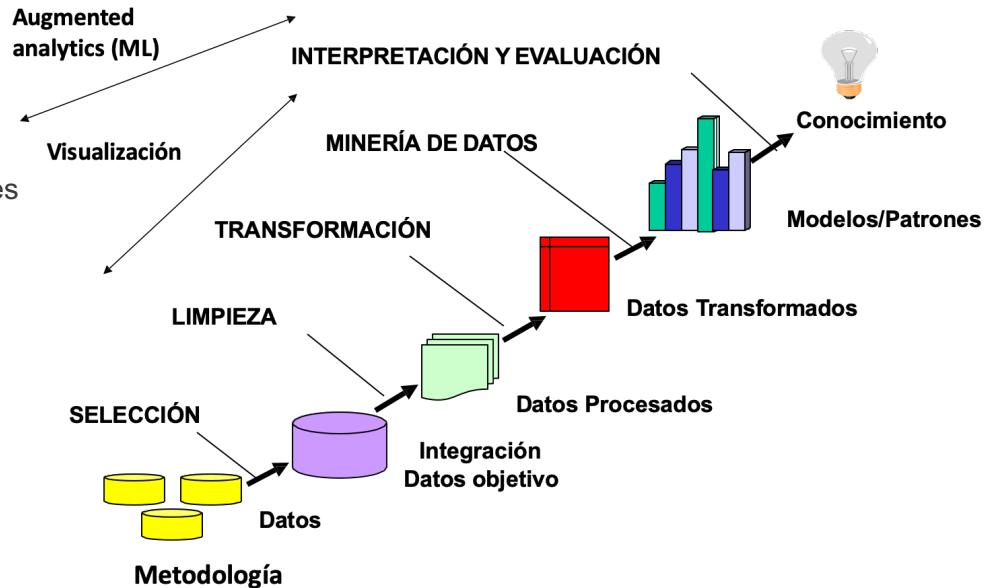
- Aprendizaje de la experiencia para tomar decisiones
- Servicios comerciales, financieros, etc. tienden hacia la personalización: adaptación al individuo

Es muy importante la compresión/interpretación, y la claridad de la salida



## El Proceso de KDD

KDD = Knowledge Discovery in Databases





- Predicción de carga
  - ¿Cuánta energía se va a consumir en los próximos días?
- Agencia tributaria
  - ¿Cuál es el perfil de los "defraudadores"?
  - ¿Se pueden subdividir en grupos homogéneos y caracterizar los diferentes tipos de contribuyentes?
  - ¿Cuáles están más alejados de cada grupo?
- Herramienta de investigación. Ej.: imágenes:
  - Dada una imagen tomada por un telescopio, ¿soy capaz de detectar y clasificar objetos interesantes?
  - Alerta de fuegos, fugas de combustible, militares, etc.
- Mejora de procesos industriales...
  
- Web Mining: análisis de páginas para extraer automáticamente información
- e-Mining: análisis de las interacciones de los clientes con mis páginas
  
- Tipo de información que busco:
  - Qué tipo de clientes tengo
  - Cómo interacciona cada tipo de cliente con las páginas Web
  - Qué banners son los que siguen mis clientes (publicidad)
  - Descubrimiento de patrones de compra/navegación
- Herramientas de gestión automática del correo



- Herramienta de investigación científica o análisis. Ej.: imágenes (TB/hora):
  - Dada una imagen tomada por un telescopio, soy capaz de detectar y clasificar objetos interesantes?
  - Sensores remotos, telescopios
  - Genética, biología
  - Meteorología, modelos climáticos
- Toma de decisiones
  - ¿Cuándo concedo un crédito hipotecario? ¿por cuánto? ¿qué tipo de solicitante no devolverá el crédito?
  - Un cliente de tarjeta de crédito está realizando una compra, ¿pagará? ¿se la han robado?
- Prevención
  - Diagnóstico precoz de enfermedades
  - Fallos en procesos industriales
- Marketing y ventas
  - Hábitos y fidelidad de clientes. ¿Cuál es el perfil de los clientes que se gastan al mes más de 1.000 €?
  - Análisis de compras. ¿Qué productos de nuestra empresa es el que compran los clientes junto al detergente?
  - Análisis de perfil más adecuado para publicidad directa.

## ● ¿Qué es la estadística?

Según Murray Spiegel:

*“La Estadística está ligada con los métodos científicos en la toma, organización, recopilación, presentación y análisis de datos, tanto para la deducción de conclusiones como para tomar decisiones razonables de acuerdo a tales análisis”*

### **Estadística Descriptiva:**

Es la que trata solamente de describir, diciendo cómo es una población, qué características tiene o qué se encontró al estudiarla, sin tratar de sacar conclusiones o inferencias.

### **Estadística Inductiva o Inferencial:**

Trata de las condiciones bajo las cuales las conclusiones obtenidas al estudiar una población pueden ser válidas para toda la población o universo.

# Método Estadístico

## Fases

se lleva a cabo siguiendo las etapas habituales en el llamado **método científico** cuyas etapas son:

**Planteamiento del problema:** consiste en definir el objetivo de la investigación y precisar el universo o población.

**Recogida de la información:** consiste en recolectar los datos necesarios relacionados al problema de investigación.

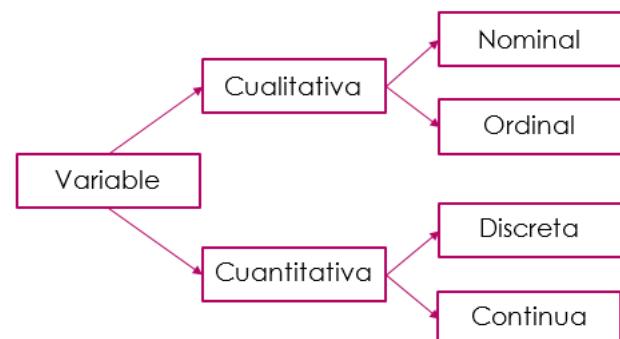
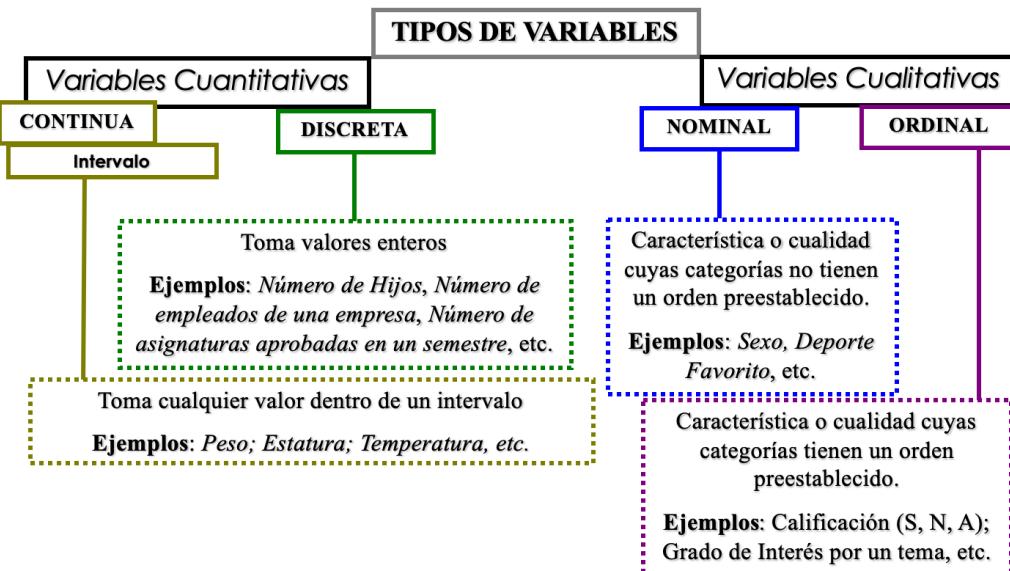
**Análisis descriptivo:** consiste en resumir los datos disponibles para extraer la información relevante en el estudio.

**Inferencia estadística:** consiste en suponer un modelo para toda la población partiendo de los datos analizados para obtener conclusiones generales.

**Diagnóstico:** consiste en verificar la validez de los supuestos del modelo que nos han permitido interpretar los datos y llegar a conclusiones sobre la población



# Variables



# Conceptos

**Universo:** Totalidad de individuos o elementos en los cuales puede presentarse determinada característica susceptible a ser estudiada.

- No siempre es posible estudiarlo en su totalidad.
- Puede ser finito o infinito, y en el caso de ser finito, puede ser muy grande y no poderse estudiar en su totalidad. Por eso es necesario escoger una parte de ese universo, para llevar a cabo el estudio.

## Población:

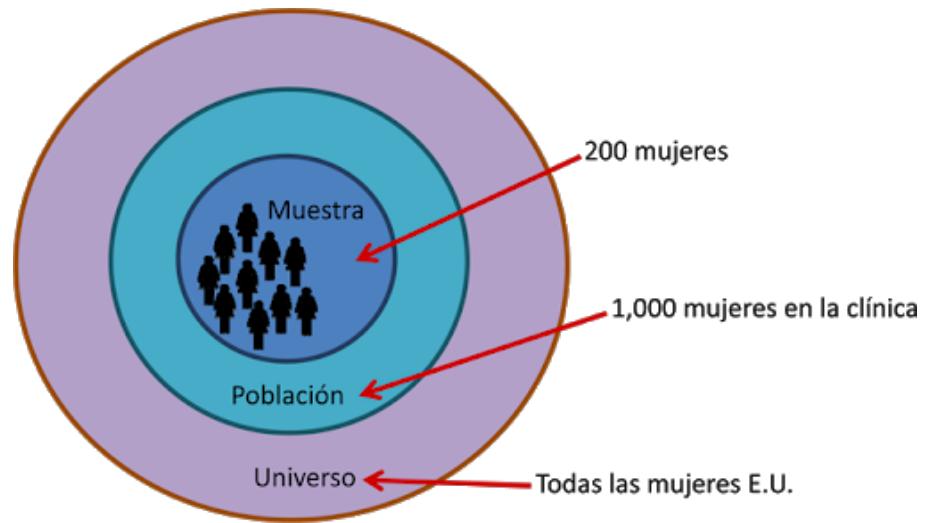
Grupo: del cual se desea algo (obtener información).

- Parte del universo en la cual vamos a basar nuestro estudio, según las características de nuestra investigación.
- Conjunto de todos los casos que concuerdan con una serie de especificaciones.
- Se debe definir la unidad de análisis, “¿Quiénes van a ser medidos?”. Para esto se debe precisar el problema a investigar y los objetivos de la investigación

**Muestra:** En lugar de examinar el grupo entero llamado población o universo, se examina una parte de éste a la que se llama muestra.

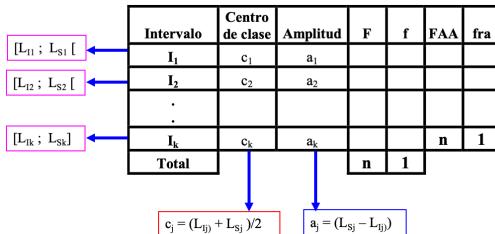
- Muestra probabilísticas: Todos los elementos de la población tienen la misma posibilidad de ser escogidos. Esto se logra a través de una selección aleatoria y/o mecánica de las unidades de análisis.
- Muestras no probabilísticas: Los elementos se seleccionan según los criterios de la persona encargada de hacer la muestra.

## Relación entre población y muestra



# Frecuencias

Intervalo (i)	Edades	Frecuencia Absoluta (fi)	Frecuencia Acumulada (Fi)	Frecuencia Relativa (hi)	Frecuencia Relativa Acumulada (Hi)
1	1 - 10	7	7	7 / 42 = 0,17	7 / 42 = 0,17
2	11 - 20	6	7+6= 13	6 / 42 = 0,14	13 / 42 = 0,31
3	21 - 30	8	13+8= 21	8 / 42 = 0,19	21 / 42 = 0,5
4	31 - 40	6	21+6= 27	6 / 42 = 0,14	27 / 42 = 0,64
5	41 - 50	5	27+5= 32	5 / 42 = 0,12	32 / 42 = 0,76
6	51 - 60	4	32+4= 36	4 / 42 = 0,1	36 / 42 = 0,86
7	61 - 70	4	36+4= 40	4 / 42 = 0,1	40 / 42 = 0,95
8	71 - 80	2	40+2= 42	2 / 42 = 0,05	42 / 42 = 1
Amplitud 9		N: 42		hi = fi / N	Hi = Fi / N



Peso	Marca	Recuento	f relativa (%)	F acumulada	F. a. relativa (%)
[36,40)	38	2	2,22	2	2,22
[40,44)	42	3	3,33	5	5,56
[44,48)	46	8	8,89	13	14,44
[48,52)	50	10	11,11	23	25,56
[52,56)	54	13	14,44	36	40,00
[56,60)	58	20	22,22	56	62,22
[60,64)	62	12	13,33	68	75,56
[64,68)	66	9	10,00	77	85,56
[68,72)	70	7	7,78	84	93,33
[72,76)	74	4	4,44	88	97,78
[76,80)	78	2	2,22	90	100,00
		90	100,00		

Llamaremos:

- ✓ A las fronteras del intervalo, *límites inferior y superior* de clase y los denotaremos por  $l_i$ ,  $L_i$  respectivamente.
- ✓ **Marca de clase ( $c_i$ )** al punto medio del intervalo, es decir, al promedio aritmético entre el límite inferior y el superior:  $c_i = \frac{L_i + l_i}{2}$ . Es el valor que tomaremos como representativo del intervalo o clase.
- ✓ **Amplitud ( $a_i$ )** es la diferencia entre el extremo superior e inferior:  $a_i = L_i - l_i$ .
- ✓ Al número de observaciones de una clase se le llama *frecuencia de clase ( $n_i$ )*. Si dividimos esta frecuencia por el número total de observaciones, se obtiene la *frecuencia relativa de clase (fi)*, y del mismo modo que lo hacímos para datos sin agrupar definimos ( $N_i$ ) y ( $F_i$ ).

## Cómo construir una distribución de frecuencias agrupada en intervalos

1. Empezamos determinando el recorrido de la variable ( $Re$ ) o *rango* de valores que tenemos en la muestra. Se define como la diferencia entre el mayor y el menor valor de la variable.
2. *Número de clases*. Depende del tamaño de la muestra. Para muestras de tamaño moderado  $n$  menor que 50, se suele elegir un número de clases o intervalos igual a  $\sqrt{n}$ . Para muestras mayores se utiliza la **fórmula de Sturges**  $\frac{\log(n)}{\log(2)} + 1$ , en general el número de intervalos no debe sobrepasar de 15 o 20, en casos de muestras muy grandes.
3. Determinamos la *amplitud de los intervalos*. Es más cómodo que la amplitud de todas las clases sea la misma (siempre que sea posible y excepto el primero y el último), si es así  $a_i = a = Re/n^q$  intervalos.
4. Tomaremos como regla general, a no ser que se indique lo contrario, hacer que el intervalo esté cerrado por la izquierda y abierto por la derecha (excepto el último intervalo).

## Ejercicio Grupal

Completa los datos que faltan en la tabla.

$x_i$	$n_i$	$f_i$	$N_i$	$F_i$
10	2	0.05	2	0.05
13	4	0.1	6	0.15
16			16	0.4
19	15			
22	6	0.15	37	0.925
25				

Completa los datos que faltan en la tabla.

$[l_i, L_i[$	$n_i$	$f_i$	$N_i$
$[0, 10[$	60		60
$[10, 20[$		0.4	
$[20, 30[$	30		170
$[30, 40[$		0.1	
$[40, 50]$			200

Trabajo en Equipo



## ● Gráficos

A un nivel estadístico y matemático, denominados **gráfica** a aquella representación visual a partir de la cual pueden representarse e interpretarse valores generalmente numéricos.

De entre las múltiples informaciones extraíbles de la observación de la gráfica podemos encontrar la existencia de relación entre variables y el grado en que se da, las frecuencias o la proporción de aparición de determinadas valores.

Esta representación visual sirve de apoyo a la hora de mostrar y comprender de manera sintetizada los datos recabados durante la investigación, de manera que puede tanto los investigadores que llevan a cabo el análisis como otros puedan comprender los resultados y resulte sencillo utilizarlo como referencia, como información a tener en cuenta o como punto de contraste ante la realización de nuevas investigaciones y meta-análisis.

<https://matematicasmodernas.com/tipos-de-graficas-estadisticas/>

<http://decodigo.com/2019/03/librerias-mas-usadas-python.html>

# Gráficos

## Trabajo en Equipo



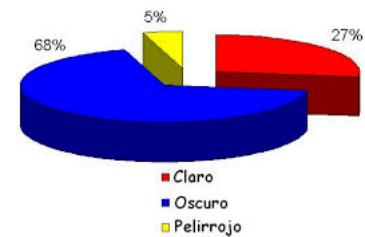
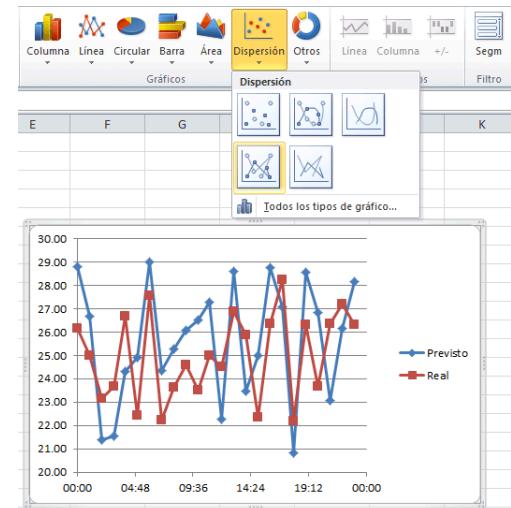
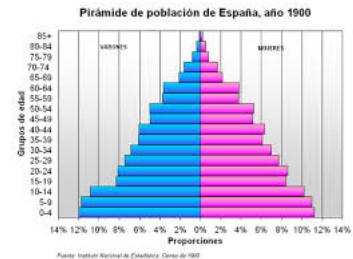
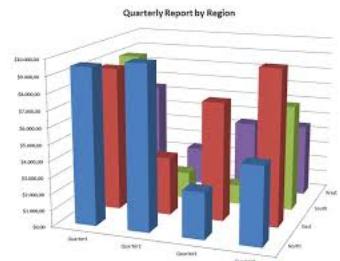
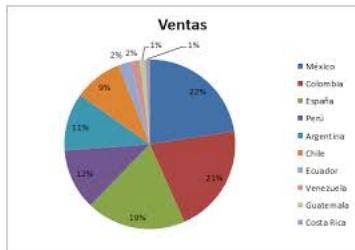
## EJERCICIOS:



Ejercicios en Clase

### EJERCICIO 2:

¿Qué es el chartismo?



## ● Para pensar....



El gobierno desea saber si el número medio de hijos por familia ha descendido respecto a la década anterior. Para ello se ha encuestado a 50 familias respecto al número de hijos y se ha obtenido los datos siguientes.

2 4 2 3 1 2 4 2 3 0 2 2 2 3 2 6 2 3 2 2 3 2 3 3 3 4 3 3 4 5 2 0 3 2 1 2 3 2 2 3 1  
4 2 3 2 4 3 3 2 2 1.

- a) Construye la tabla de frecuencias con estos datos.
- b) ¿Cuántas familias tienen exactamente 3 hijos?
- c) ¿Qué porcentaje de familias tienen exactamente 3 hijos?
- d) ¿Qué porcentaje de familias de la muestra tiene más de dos hijos? ¿Y menos de tres?
- e) Construye el gráfico que consideres más adecuado con las frecuencias no acumuladas.
- f) Construye el gráfico que consideres más adecuado con las frecuencias acumuladas.

La distribución de los salarios en la industria turística española es la que figura en la tabla. Calcula:

- a) El salario medio por trabajador (marca de clase del último intervalo, 20000).
- b) El salario más frecuente.
- c) El salario tal que la mitad de los restantes sea inferior a él.

$[l_b, L_i[$	$n_i$
[0,1500[	2145
[1500, 2000[	1520
[2000, 2500[	840
[2500, 3000[	955
[3000, 3500[	1110
[3500, 4000[	2342
[4000, 5000[	610
[5000, 10000[	328
$\geq 10000$	150

## ● Para pensar....



El gobierno desea saber si el número medio de hijos por familia ha descendido respecto a la década anterior. Para ello se ha encuestado a 50 familias respecto al número de hijos y se ha obtenido los datos siguientes.

2 4 2 3 1 2 4 2 3 0 2 2 2 3 2 6 2 3 2 2 3 2 3 3 3 4 3 3 4 5 2 0 3 2 1 2 3 2 2 3 1  
4 2 3 2 4 3 3 2 2 1.

- Construye la tabla de frecuencias con estos datos.
- ¿Cuántas familias tienen exactamente 3 hijos?
- ¿Qué porcentaje de familias tienen exactamente 3 hijos?
- ¿Qué porcentaje de familias de la muestra tiene más de dos hijos? ¿Y menos de tres?
- Construye el gráfico que consideres más adecuado con las frecuencias no acumuladas.
- Construye el gráfico que consideres más adecuado con las frecuencias acumuladas.

La distribución de los salarios en la industria turística española es la que figura en la tabla. Calcula:

- El salario medio por trabajador (marca de clase del último intervalo, 20000).
- El salario más frecuente.
- El salario tal que la mitad de los restantes sea inferior a él.

$[l_b, L_i[$	$n_i$
[0,1500[	2145
[1500, 2000[	1520
[2000, 2500[	840
[2500, 3000[	955
[3000, 3500[	1110
[3500, 4000[	2342
[4000, 5000[	610
[5000, 10000[	328
$\geq 10000$	150



Calle Rufino González 25  
28037 Madrid  
+34810527241  
[www.mioti.es](http://www.mioti.es)

