TEMA 2

FUNDAMENTOS DE ESTADÍSTICA

ESTADÍSTICA DESCRIPTIVA UNIDIMENSIONAL I

Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

ÍNDICE

OBJETIVOS ESPECÍFICOS

- Profundizar en estadística descriptiva y controlar la información, realizando análisis y extrayendo las primeras conclusiones de los datos.
- Elegir, correctamente, el gráfico adecuado para su correcta visualización e interpretación.
- Aplicar la estadística descriptiva a situaciones de la vida real y profesional.

2. ESTADÍSTICA DESCRIPTIVA UNIDIMENSIONAL

La estadística descriptiva se ocupa de tomar los datos de un conjunto dado, **organizarlos en tablas o representaciones gráficas** y del cálculo de unos números que **nos informen, de manera global, del conjunto estudiado.**

La estadística descriptiva unidimensional o univariante **se centra en el análisis de una única característica** o cualidad del individuo. Las características a analizar presentan *k* variables (modalidades o características), que son exhaustivas y mutuamente excluyentes.

2.1. DISTRIBUCIÓN DE FRECUENCIAS, TABLAS ESTADÍSTICAS Y MÉTODOS GRÁFICOS

La distribución de frecuencias o tabla de frecuencias **es una ordenación** en forma de tabla de los datos estadísticos, asignando a cada dato su frecuencia correspondiente.

2.1.1. FRECUENCIA ABSOLUTA

La **frecuencia absoluta** de un valor o modalidad de una variable estadística es el **número de veces** que el valor o modalidad aparece en la población objeto de estudio. Se representa por n_i.

La **suma** de las frecuencias absolutas es igual al número total de datos, que se representa por ***N*** (también se dice que ***N*** es el número de observaciones o frecuencia total).

Su fórmula es:

$$\sum_{i=1}^{k} n_i = N$$

La propiedad de la frecuencia absoluta la verificaremos a través del siguiente sumatorio:

$$\sum_{i=1}^{k} n_i = n_1 + n_2 + n_3 + \ldots + n_k = N$$

donde: $0 \leq n_i \leq N$ y $i = 1, 2, 3, \ldots \in \mathbf{N}$

EJEMPLO

Las edades de los participantes de un concurso de ajedrez son: 18, 13, 12, 14, 11, 12, 15, 20, 18, 14, 15, 11, 10, 10, 11, 13, 15, 16, 12, 11.

La frecuencia absoluta de 10 es 2, porque la edad de 10 años se repite 2 veces. La frecuencia absoluta de 15 es 3, porque la edad de 15 años se repite 3 veces.

2.1.2. FRECUENCIA RELATIVA

La **frecuencia relativa** es el cociente entre la frecuencia absoluta de un determinado valor y el número total de datos, o dicho de otro modo, **es el resultado de dividir la frecuencia absoluta entre el tamaño de la población.** Se puede expresar en tantos por ciento y se representa por f_i.

La suma de las frecuencias relativas es igual a 1.

Su fórmula es:

$$\sum_{i=1}^{k} f_i = \frac{n_i}{N}$$

La propiedad de la frecuencia relativa la verificaremos a través del siguiente sumatorio:

$$\sum_{i=1}^{k} f_i = f_1 + f_2 + f_3 + \ldots + f_k = 1$$

donde: $\mathbf{0 \leq f_i \leq 1}$

EJEMPLO

Las edades de los 20 participantes de un concurso de ajedrez son: 18, 13, 12, 14, 11, 12, 15, 20, 18, 14, 15, 11, 10, 10, 11, 13, 15, 16, 12, 11.

La frecuencia relativa de 10 es 2/20, que es la frecuencia absoluta/total de la cantidad de los datos. La frecuencia relativa de 15 es 3/20, que es la frecuencia absoluta/ total de la cantidad de los datos.

2.1.3. FRECUENCIA ACUMULADA

La **frecuencia acumulada** es la suma de las frecuencias absolutas o relativas de todos los valores inferiores o iguales al valor considerado. Se representan con las letras mayúsculas N_i y F_i, respectivamente.

IMPORTANTE

La frecuencia relativa (o la frecuencia relativa acumulada) se suele expresar o representar en **tanto por ciento.**

EJEMPLO

Las edades de los 20 participantes de un concurso de ajedrez son: 18, 13, 12, 14, 11, 12, 15, 20, 18, 14, 15, 11, 10, 10, 11, 13, 15, 16, 12, 11.

- Frecuencia Absoluta Acumulada
  - La frecuencia absoluta acumulada de 10 es 2, ya que el primer elemento no varía de la frecuencia absoluta.
  - La frecuencia absoluta acumulada de 15 es 16, que es el resultado de sumar todas las frecuencias hasta el elemento 15.
- Frecuencia Relativa Acumulada
  - La frecuencia relativa acumulada de 10 es 2/20, que es igual que la frecuencia absoluta dividido entre el total de la cantidad de los datos. Al igual que antes, el primer elemento no varía.
  - La frecuencia relativa acumulada de 15 es 16/20=0.8, que es la frecuencia relativa del elemento 15 dividido entre el total de la cantidad de los datos.

2.1.4. TABLAS ESTADÍSTICAS

Como puedes observar, el cálculo de las frecuencias acumuladas, parece complejo de calcular a primera vista, pero en realidad es sencillo.

Para simplificar y dar transparencia al cálculo tanto de los componentes básicos y principales de estadística así como a cálculos más complejos que se verán más adelante, **se usan las tablas de frecuencias.**

2.1.4.1. TABLAS ESTADÍSTICAS PARA VARIABLES DISCRETAS (DATOS NO AGRUPADOS)

Veamos el ejemplo en el que se acaban de calcular todas las frecuencias, pero esta vez se agruparán y resumirán en una tabla de frecuencias.

Teníamos la siguiente entrada de información: 18, 13, 12, 14, 11, 12, 15, 20, 18, 14, 15, 11, 10, 10, 11, 13, 15, 16, 12, 11.

Luego, la tabla para datos no agrupados o variables discretas es:

x_i	n_i	N_i	f_i	F_i
10	2	2	2/20 = 0,1	2/20 = 0,1
11	4	6	4/20 = 0,2	6/20 = 0,3
12	3	9	3/20 = 0,15	9/20 = 0,45
13	2	11	2/20 = 0,1	11/20 = 0,55
14	2	13	2/20 = 0,1	13/20 = 0,65
15	3	16	3/20 = 0,15	16/20 = 0,8
16	1	17	1/20 = 0,05	17/20 = 0,85
18	2	19	2/20 = 0,1	19/20 = 0,95
20	1	20	1/20 = 0,05	20/20 = 1

Así, se obtiene de manera sencilla y rápida el cálculo de las 4 tipologías de cálculo de las frecuencias (absoluta y relativa) tanto acumulativa como no.

Como "buena costumbre" realizaremos un sumatorio al final del cálculo, que servirá de medida de comprobación de que los cálculos se han realizado correctamente. En él se busca el cumplimiento de las propiedades de las frecuencias.

En el siguiente ejemplo se añadirá esta línea adicional.

EJEMPLO

En una empresa quieren ver la edad media de sus trabajadores/as. Para ello, se han registrado los siguientes datos de edades: 32, 31, 28, 29, 33, 32, 31, 30, 31, 31, 27, 28, 29, 30, 32, 31, 31, 30, 30, 29, 29, 30, 30, 31, 30, 31, 34, 33, 33, 29, 29.

Tabla para datos no agrupados

x_i	Recuento	n_i	N_i	f_i	F_i
27	I	1	1	1/31 = 0.032	1/31 = 0.032
28	II	2	3	2/31 = 0.065	3/31 = 0.097
29	VI	6	9	6/31 = 0.194	9/31 = 0.290
30	VII	7	16	7/31 = 0.226	16/31 = 0.516
31	VIII	8	24	8/31 = 0.258	24/31 = 0.774
32	III	3	27	3/31 = 0.097	27/31 = 0.871
33	III	3	30	3/31 = 0.097	30/31 = 0.968
34	I	1	**31**	1/31 = 0.032	31/31 = **1**
$\sum$		**31**		**1**	

2.1.4.2. TABLAS ESTADÍSTICAS PARA VARIABLES CONTINUAS (DATOS AGRUPADOS)

Cuando la variable es continua o cuando la variable discreta toma un gran número de valores, se utiliza la **distribución de frecuencias agrupadas** o tabla con datos agrupados.

Se pueden agrupar los valores en intervalos de **idéntica o distinta amplitud** denominados clases y, a cada una de ellas, se le asigna su frecuencia correspondiente.

En este tipo de tabla, hay que tener en cuenta los siguientes conceptos:

- **Límites de la clase.** El límite se representa por L_i. Cada clase está delimitada por el límite inferior de la clase y el límite superior de dicha clase. Se representa respectivamente por L_{i+1} para el límite superior y L_{i-1} para el límite inferior.
- **Amplitud de la clase.** La amplitud de la clase se representa por ai , que es la diferencia entre el límite superior e inferior de la clase. Su fórmula es:

$$a_i = L_{i+1} - L_{i-1}$$

- **Marca de la clase.** La marca de clase C_i es el punto medio de cada intervalo y es el valor que representa a todo intervalo para el cálculo de algunos parámetros. Su fórmula es:

$$C_i = (L_{i-1} - L_i) / 2$$

EJEMPLO

En una cosecha, la recolecta de trigo obtenida en Kg. ha sido: 3, 15, 24, 28, 33, 35, 38, 42, 43, 38, 36, 34, 29, 25, 17, 7, 34, 36, 39, 44, 31, 26, 20, 11, 13, 22, 27, 47, 39, 37, 34, 32, 35, 28, 38, 41, 48, 15, 32, 13.

Calcula las frecuencias y determina cuál es la amplitud para la marca de clase C_2.

En este caso, se tienen muchos datos con muy poca frecuencia (24 números se repiten sólo una vez), por lo que se elige agrupar la información en intervalos, para que los cálculos sean más sencillos.

Tabla para datos agrupados

Int	C_i	n_i	N_i	f_i	F_i
[0,5)	2.5	1	1	1/40 = 0.025	1/40 = 0.025
[5,10)	7.5	1	2	1/40 = 0.025	2/40 = 0.050
[10,15)	12.5	3	5	3/40 = 0.075	5/40 = 0.125
[15,20)	17.5	3	8	3/40 = 0.075	8/40 = 0.200
[20,25)	22.5	3	11	3/40 = 0.075	11/40 = 0.2775
[25,30)	27.5	6	17	6/40 = 0.150	17/40 = 0.425
[30,35)	32.5	7	24	7/40 = 0.175	24/40 = 0.600
[35,40)	37.5	10	34	10/40 = 0.250	34/40 = 0.850
[40,45)	42.5	4	38	4/40 = 0.100	38/40 = 0.950
[45,50)	47.5	2	40	2/40 = 0.050	40/40 = 1
Σ		40		1	

Para calcular la amplitud de la marca de clase C_2, hay que situarse en la segunda fila donde comprobamos que, para la marca de clase, el valor es de 7.5. Este valor está situado dentro del intervalo [5, 10). Por tanto, se puede calcular la amplitud de la siguiente forma:

$$a_i = L_{i+1} - L_{i-1} \quad \rightarrow \quad a_2 = L_{2+1} - L_{2-1} = L_3 - L_1 = 10 - 5 = 5$$

2.1.4.3. ELECCIÓN EN EL NÚMERO DE CLASES

El **número de clases o intervalos se puede obtener a través de dos fórmulas.** La segunda fórmula se la conoce también como la ***Regla de Sturges.***

- Si *N* no es muy grande: $\sqrt{N}$
- Si *N* es grande: $1 + 3.22 \cdot \log_{10} N$

EJEMPLO

Se tiene una localidad con 100 personas y otra con 1.000.000 habitantes. Se necesita calcular el número de intervalos que hay que utilizar para cada población.

- Si $N = 100$ entonces se puede utilizar la primera fórmula:

$$\sqrt{N} = \sqrt{100} = 10$$

- Si $N = 1000000$ entonces se debería aplicar la **Regla de Sturges,** esto es:

$$1 + 3.22 \cdot \log_{10} N = 1 + 3.22 \cdot \log_{10} 1000000 = 20.32 \approx 20$$

Los intervalos han de ser razonables, ni muy grandes ni muy pequeños para que no haya pérdida de información. Siempre hay que observar los datos de la frecuencia acumulada, cómo están distribuidos.

Si los datos son homogéneos, modificaremos la amplitud para que los datos estén lo más dispersos y conseguir que la heterogeneidad sea la mayor posible.

RECUERDA

Hay que tener en cuenta que, **cuando se trabaja con intervalos y tienen distinta amplitud, las fórmulas y consideraciones se vuelven más complejas.** Aunque si la prioridad es obtener una mayor precisión en el análisis, entonces en esos casos, compensa agrupar la información o los datos con distinta amplitud de intervalos.

En ***Matplotlib*** se llama ***bins*** a la elección del **número de intervalos** para hacer las gráficas. Para ampliar información, te recomendamos la lectura de los siguientes enlaces:

www.datatofish.com
www.statisticshowto.com

2.1.5. MÉTODOS GRÁFICOS

Las distintas representaciones gráficas están previstas para usarse con un tipo de datos, esto es, no todos los gráficos sirven para representar cualquier dato.

Los gráficos tienen como objetivo explicar un dato "a golpe de vista", por tanto en la búsqueda se debe usar el gráfico apropiado, con el mejor diseño y color para que ayude a entender y aclarar conceptos y mejorar la toma de decisiones.

SABÍAS QUE...

El paquete ***Matplotlib*** es una biblioteca bastante completa para crear visualizaciones estáticas, animadas e interactivas en Python.

El enlace a la documentación y a la galería con ejemplos es el siguiente:

www.matplotlib.org

Con *matplotlib* también se pueden hacer **visualizaciones con mapas** (con Basemap y mplot3D para tres dimensiones), donde existe la posibilidad de rotar la figura e incluso hacer zoom en la propia visualización.

Por otro lado, ***Seaborn*** es una librería para Python que permite generar gráficos elegantes de forma sencilla. *Seaborn* está basado en *matplotlib* y proporciona una interfaz de alto nivel que es realmente sencilla de aprender. Es una interfaz para crear gráficos estadísticos explicativos y atractivos: el objetivo es visualizar datos complejos de forma sencilla y extraer conclusiones.

El enlace a la documentación y a la galería con ejemplos es el siguiente:

www.seaborn.pydata.org

In [ ]:

```
## Importación paquete Matplotlib
import matplotlib.pyplot as plt

## Importación Seaborn
import seaborn as sns
```

2.1.5.1. DIAGRAMA DE BARRAS

Un diagrama de barras se utiliza para representar datos cualitativos o cuantitativos de tipo discreto.

Se representa sobre unos ejes de coordenadas. En el eje de abscisas, se colocan los valores de la variable, y sobre el eje de ordenadas, las frecuencias absolutas o relativas o acumuladas. Los datos se representan mediante barras de una altura proporcional a la frecuencia.

Sus características son:

- Barras con anchuras iguales.
- Espacio entre barras iguales.
- Altura proporcional a ni o fi.
- El eje OY es ilimitado (a veces se representa con flecha).
- El eje OX es limitado.

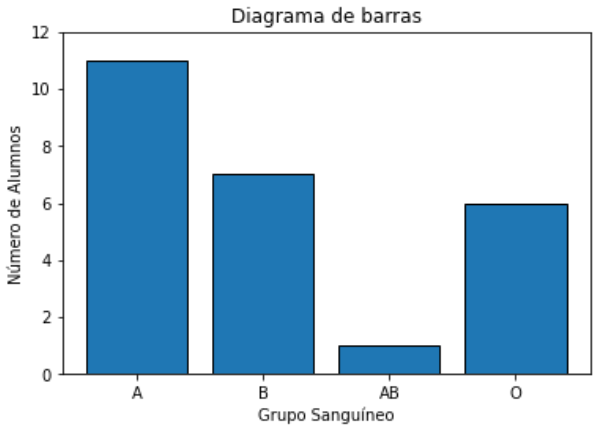
EJEMPLO

Un estudio realizado a 25 alumnos/as de una clase de informática para determinar su grupo sanguíneo ha dado el siguiente resultado: 11 del grupo A, 7 del grupo B, 1 del grupo AB y 6 del grupo O.

In [ ]:

```
# Gráfico de barras
plt.title("Diagrama de barras")
plt.xlabel("Grupo Sanguíneo")
plt.ylabel("Número de Alumnos")

grupo = ['A', 'B', 'AB', 'O']
alumnos = [11, 7, 1, 6]

plt.bar(range(4), alumnos, edgecolor='black')
plt.xticks(range(4), grupo, rotation=360)

#limit
plt.ylim(min(alumnos)-1, max(alumnos)+1)

plt.show()
```

Out [ ]:



2.1.5.2. DIAGRAMA DE FRECUENCIAS ACUMULADAS

El diagrama de frecuencias acumuladas se utiliza para **representar datos de tipo discreto.**

Este gráfico se corresponde con la función constante entre cada dos valores de la variable a representar e igual en cada tramo a la frecuencia relativa (o absoluta) acumulada hasta el menor de los dos valores de la variable que construyen el tramo en el que es constante.

Si las frecuencias representadas son las absolutas, entonces el mayor valor que se toma, el eje OY es el tamaño muestral *N*.

Sus características son:

- Cerrado por la izquierda (sin punto) y abierto por la derecha (con punto).
- Espacio entre líneas.
- Altura proporcional a N_i o F_i.
- El eje de la OX es ilimitado.

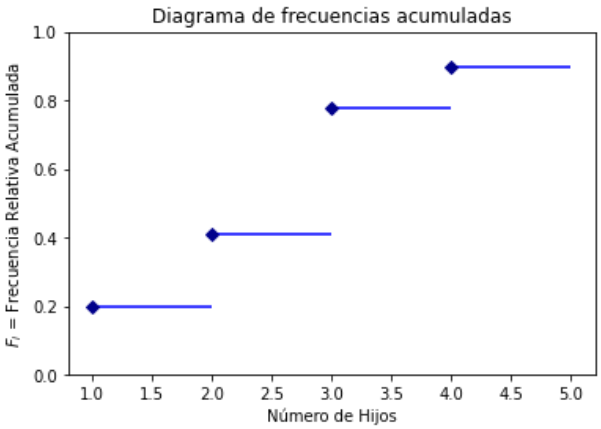
EJEMPLO

En un estudio sobre el nº de hijos en una población, donde las madres no han tenido más de cuatro hijos (0, 1, 2, 3 ó 4), se ha obtenido la siguiente frecuencia relativa acumulada: 0.2, 0.41, 0.78 y 0.9.

In [ ]:

```
# Lineal Graphic
plt.title("Diagrama de frecuencias acumuladas")
plt.ylabel("$F_{i}$ = Frecuencia Relativa Acumulada")
plt.xlabel("Número de Hijos")

# Limits
plt.ylim(0, 1)

# Data
x= [1,   2,    3,    4   ]
y= [0.2, 0.41, 0.78, 0.9 ]

plt.hlines(y=0.2, xmin=1, xmax=2, color='blue')
plt.hlines(y=0.41, xmin=2, xmax=3, color='blue')
plt.hlines(y=0.78, xmin=3, xmax=4, color='blue')
plt.hlines(y=0.9, xmin=4, xmax=5, color='blue')

plt.plot(x, y, 'D', color='darkblue')

plt.show()
```

Out [ ]:



OBSERVACIÓN

Para indicar que el punto está incluido se utilizan los puntos en intervalos. La línea indica que está abierto y, por tanto, no se incluye el punto.

2.1.5.3. HISTOGRAMA

Un histograma es una representación gráfica de una variable en forma de barras. En el eje de abscisas, se construyen unos rectángulos que tienen por base la amplitud del intervalo, y por altura, la frecuencia absoluta de cada intervalo.

La superficie de cada barra es proporcional a la frecuencia de los valores representados.

Los histogramas **se utilizan para variables continuas o para variables discretas,** con un gran número de datos, y que se han agrupado en clases. A continuación, vamos a mostrar dos ejemplos, el primero para variables discretas y el segundo para variables continuas.

EJEMPLO

Un estudio sobre la edad de 130 personas de todas las edades (a partir del año de edad) ha recogido la siguiente información:

1,1,2,3,3,5,7,8,9,10,10,11,11,13,13,15,16,17,18,18,18,19,20,21,21,23,24,24,25,25,25,25,26, 26,26,27,27,27,27,27,29,30,...

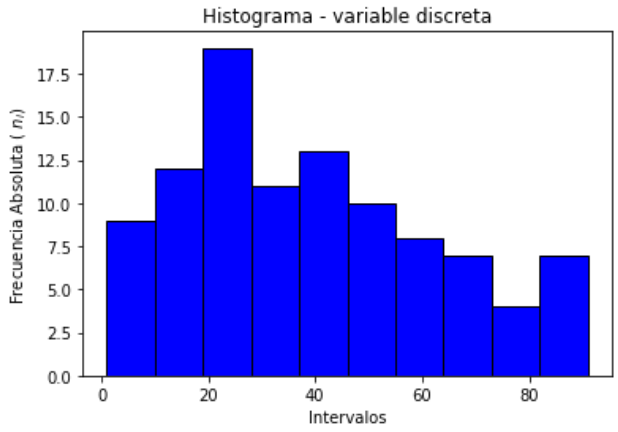
Realiza el gráfico del histograma.

In [ ]:

```
## Histograma - variable discreta
plt.title("Histograma - variable discreta")
plt.xlabel("Intervalos")
plt.ylabel("Frecuencia Absoluta ( $n_{i}$) ")
```

```
# Raw data
x  =  [1,1,2,3,3,5,7,8,9,10,10,11,11,13,13,15,16,17,18,18,18,19,20,
       21,21,23,24,24,25,25,25,25,26,26,26,27,27,27,27,27,29,30,30,
       31,33,34,34,34,35,36,36,37,37,38,38,39,40,41,41,42,43,44,45,
       45,46,47,48,48,49,50,51,52,53,54,55,55,56,57,58,60,61,63,64,
      65,66,68,70,71,72,74,75,77,81,83,84,87,89,90,90,91
    ]

# Plot the distribution of data
plt.hist(x, bins=10, color='b', edgecolor='black')

plt.show()
```

Out [ ]:



OBSERVACIÓN

Para saber el número de bins (o intervalos) que le corresponde, revisa el apartado ***Elección en el número de clases.***

EJEMPLO

Se ha recogido el salario bruto anual de 46 trabajadores/as de una empresa para analizar la distribución. El salario se ha expresado en miles, es decir, la expresión 12.5 equivale a 12.500 euros brutos anuales.

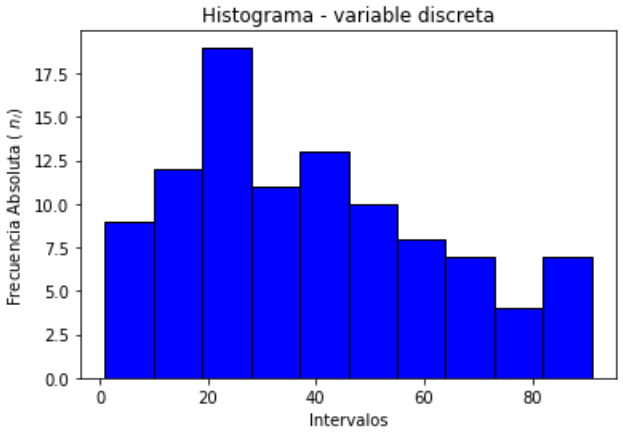
Los datos son los siguientes:

12.5, 15.3, 13.7, 20.5, 19.1, 20.2, 11.3, 19.4, 11.8, 12.9, 19.2, 13, 12.5, 10.2, 16.5, 19.3, 19, 18.5, 19.3, 20.7, 19.9, 36, 35.4, 19.5, 19.7, 17.3, 16.8, 11.9, 33.5, 28.5, 25.3, 35, 37.3, 22.5, 24.7, 23.7, 21.5, 22.4, 25.3, 24.7, 26, 28, 26.3, 31.2, 31, 34.5.

In [ ]:

```
## Histograma - variable continua
plt.title("Histograma - variable continua")
plt.ylabel("Frecuencia Absoluta ( $n_{i}$) ")
plt.xlabel("Salario")

# Data in numpy array
x = np.array([12.5, 15.3, 13.7, 20.5, 19.1, 20.2, 11.3, 19.4, 11.8, 12.9, 19.2,
              13, 12.5, 10.2, 16.5, 19.3, 19, 18.5, 19.3, 20.7, 19.9, 36, 35.4,
              19.5, 19.7, 17.3, 16.8, 11.9, 33.5, 28.5, 25.3, 35, 37.3, 22.5,
              24.7, 23.7, 21.5, 22.4, 25.3, 24.7, 26, 28, 26.3, 31.2, 31, 34.5])

# Plot the distribution of numpy data
plt.hist(x, bins=15, align='left', color='b', edgecolor='black')

plt.show()
```

Out [ ]:



Observa que los intervalos de amplitud por defecto se hacen iguales. Hay situaciones en las que no se recomienda realizar el mismo tamaño de amplitud y conviene realizar distintos tamaños de amplitud.

Para construir un **histograma con intervalo de amplitud diferente,** hay que calcular las alturas de los rectángulos del histograma.

Se representa la altura del intervalo como h_i. Siendo la frecuencia del intervalo ni y la amplitud del intervalo a_i. Entonces la fórmula es:

$$h_i = \frac{n_i}{a_i}$$

Donde la fórmula de la amplitud es:

$$a_i = L_{i+1} - L_{i-1}$$

Este cálculo será igual tanto para frecuencias absolutas como para las frecuencias relativas, acumuladas o no.

EJEMPLO

En la siguiente tabla se muestran las calificaciones (suspenso, aprobado, notable y sobresaliente) obtenidas por un grupo de 50 alumnos/as. Las notas obtenidas son:

0,1.1,2.2,3.8,4,0,4,1,4,8,4,9,5.2,5.3,5.6,5.5,5,5,5,5,6,6,6,6.5,6,6,6,6.3,6.4,6.5,6.5,6.6,7,7,7.6, 7.3,7.5,7.7,8.1,8.2,8,8.5,8.5,8.6,8.7,8.89,9.2,9.4,9.5,10

Realizaremos la tabla de frecuencias, calculando la altura y, después, construiremos el histograma asociado.

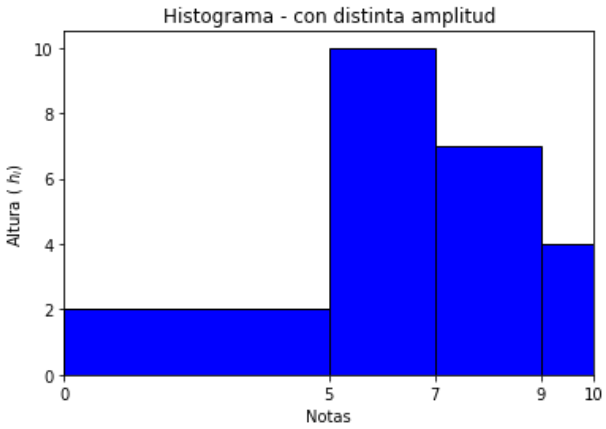
c_i	n_i	a_i	h_i
[0, 5)	10	5 - 0 = 5	10 / 5 = 2
[5, 7)	20	7 - 5 = 2	20 / 2 = 10
[7, 9)	15	9 - 7 = 2	15 / 2 = 7.5
[9, 10)	5	10 - 9 = 1	5 / 1 = 5

In [ ]:

```
## Histograma - con distinta amplitud
plt.title("Histograma - con distinta amplitud")
plt.ylabel("Altura ( $h_{i}$) ")
plt.xlabel("Notas")

# Axis OX
notas = [0, 5, 7, 9, 10]
plt.xticks(notas)
plt.xlim(0, 10)

# Raw data
# no se necesita todo el dataset sólo la parte de la frecuencia representativa
hi = np.array([1, 2.2,
               5.2, 5.3, 5.6, 5.5, 5, 5, 5, 5, 6,  6,
               7, 7, 7, 7.7, 8.1, 8.2, 8,
               9.2, 9.4, 9.5, 10])
```

```
bins list = [0, 5, 7, 9, 10]
# El primer bin [0, 5) incluye el 0 y excluye el 5. Todos son iguales excepto el
# último bin, que incluye ambos intervalos (el 9 y el 10)

plt.hist(hi, bins = bins_list, color='b', edgecolor='black')

plt.show()
```

Out [ ]:

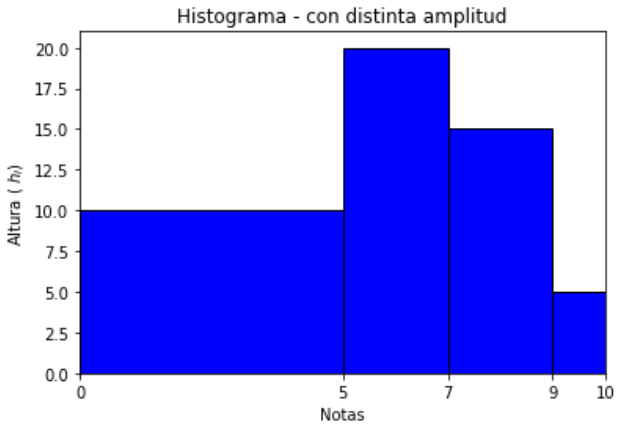


Realizaremos el **mismo gráfico sin tener en cuenta que la amplitud es distinta,** por lo que lo haremos sin el cálculo de la altura.

In [ ]:

```
## Histograma - con distinta amplitud
plt.title("Histograma - con distinta amplitud")
plt.ylabel("Altura ( $h_{i}$) ")
plt.xlabel("Notas")

# Axis OX
notas = [0, 5, 7, 9, 10]
plt.xticks(notas)
plt.xlim(0, 10)

# Raw data
x_notas =  np.array([0,1.1,2.2,3.8,4,0,4,1,4,8,4,9,5.2,5.3,5.6,5.5,5,5,5,5,6,6,6,6.5,
           6,6,6,6.3,6.4,6.5,6.5,6.6,7,7,7.6,7.3,7.5,7.7,8.1,8.2,8,8.5,8.5,
           8.6,8.7,8.89,9.2,9.4,9.5, 10])

bins_list = [0, 5, 7, 9, 10]
# El primer bin [0, 5) incluye el 0 y excluye el 5. Todos son iguales excepto el
# último bin, que incluye ambos intervalos (el 9 y el 10)


plt.hist(x_notas, bins = bins_list, color='b', edgecolor='black')

plt.show()
```

Out [ ]:



Si observas los dos gráficos juntos, verás mejor que el segundo puede llevar a error la apreciación de aprobados y notables.

Observa también cómo la interpretación del gráfico al no estar bien construido genera más dificultad.

2.1.5.4. POLÍGONO DE FRECUENCIAS

Un polígono de frecuencias se forma uniendo los extremos de las barras mediante segmentos. También se puede realizar trazando los puntos que representan las frecuencias y uniéndolos mediante segmentos.

Se aplica tanto en variables discretas como continuas. Teniendo en cuenta que para construir el polígono de frecuencias con datos agrupados, se toma la marca de clase que coincide con el punto medio de cada rectángulo.

Suele representarse junto con el diagrama de barras o el histograma, para ver mejor los puntos de inflexión del gráfico.

Sus características son:

- Barras con anchuras iguales.
- Sin espacio entre barras.
- Altura proporcional a n_i o f_i.
- Sin límites en el eje.
- Con límites en eje *OX*.

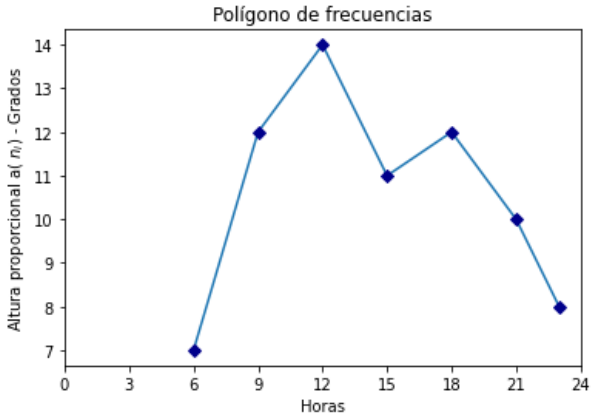
EJEMPLO

Las temperaturas en un día de otoño de la ciudad de Burgos han sufrido las siguientes variaciones: 7º, 12º, 14º, 11º. 12º, 10 º y 8º (en centígrados) correspondientes a las siguientes horas: 6, 9, 12, 15, 18, 21 y 24.

In [ ]:

```
# Polígono de frecuencias
plt.title("Polígono de frecuencias")
plt.ylabel("Altura proporcional a( $n_{i}$) - Grados ")
plt.xlabel("Horas")

# Axis - Tratamiento eje OX
plt.xlim(0, 24)
horas = [0, 3, 6, 9, 12, 15, 18, 21,  24]
plt.xticks(horas)

# Data
x = [6, 9, 12, 15, 18, 21,  23]
y = [7, 12, 14, 11, 12, 10,  8]

# Graphs
plt.plot(x, y)
plt.plot(x, y, 'D', color='darkblue')

plt.show()
```

Out [ ]:



Polígono de frecuencias e histograma de frecuencias

En muchas ocasiones, el polígono no se representa sólo, se utiliza de forma conjunta con las frecuencias.

Las características de este tipo de gráfico son:

- Barras con anchuras iguales.
- Sin espacio entre barras.
- Altura proporcional a n_i o f_i.

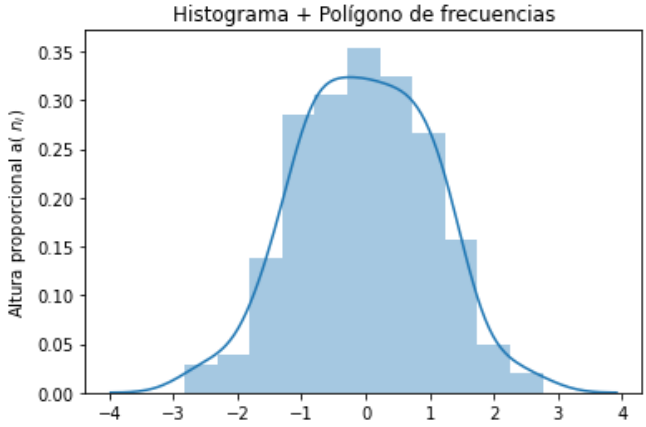
EJEMPLO

A partir de una muestra de 200 elementos, extraídos de forma aleatoria, se construye el histograma de frecuencias y su respectivo polígono.

In [ ]:

```
# Poligono de frecuencias
plt.title("Histograma + Polígono de frecuencias")
plt.ylabel("Altura proporcional a( $n_{i}$) ")

# Data
x = np.random.randn(200)

# Plot
kwargs = {'cumulative': False}
sns.distplot(x, hist_kws=kwargs, kde_kws=kwargs)

plt.show()
```

Out [ ]:



Polígono de frecuencias e histograma de frecuencias acumulado

Al igual que en el apartado anterior, el polígono suele dibujarse unido a la gráfica de frecuencias acumuladas.

Las características de este tipo de gráficos con:

- Barras con anchuras iguales.
- Sin espacio entre barras.
- Altura proporcional a *Ni* o *Fi*.
- Aumenta de manera escalada.

EJEMPLO

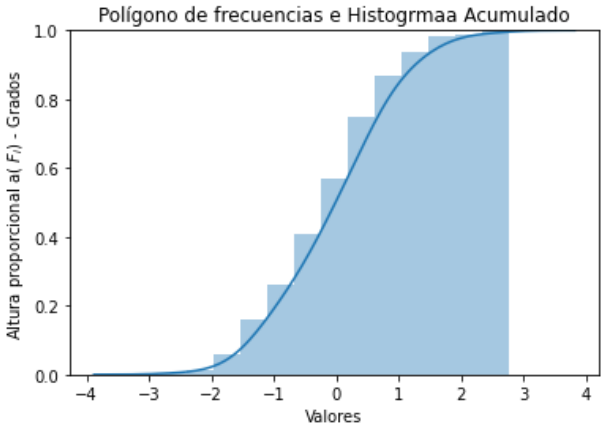
Basándonos en el ejemplo anterior, construimos el histograma acumulado y el polígono de frecuencias asociado.

In [ ]:

```
# Poligono de frecuencias
plt.title("Polígono de frecuencias e Histogrmaa Acumulado")
plt.ylabel("Altura proporcional a( $F_{i}$) - Grados ")
plt.xlabel("Valores")
```

```
# Axis - Tratamiento eje OY
plt.ylim(0,1)

# Data
x = np.random.randn(200)

# Plot
kwargs = {'cumulative': True}
sns.distplot(x, hist_kws=kwargs, kde_kws=kwargs)
plt.show()
```

Out [ ]:



2.1.5.5. DIAGRAMA DE TALLO Y HOJAS

El diagrama de tallos y hojas (o Stem-and-leaf plot) **sólo se puede utilizar con variables cuantitativas.**

Para construirlo, basta con ordenar a los individuos de menor a mayor. A la izquierda se coloca el tallo y a la derecha se colocará la hoja. El tallo está compuesto por la parte entera de las observaciones y las hojas por la parte decimal. Las hojas siempre deben estar ordenadas de menor a mayor.

IMPORTANTE

Para construir este gráfico, existen varias formas, aunque la más sencilla y visual es con el paquete stemgraphic de *python.* Por defecto en colab no está instalado, por tanto se necesita instalar dicho paquete.

In [ ]:

```
import sys

!{sys.executable} -m pip install stemgraphic

import stemgraphic
```

EJEMPLO

Las edades en una excursión de 10 personas han sido: 30, 22, 41, 16, 23, 15, 32, 22, 21, 23.

Luego el **tallo** corresponde con las **decenas:** 1, 2, 3 y 4. Y las **hojas** hacen referencia a las **unidades:** 1, 2, 3,...

In [ ]:

```
# Raw data
x = [ 30, 22, 41, 16, 23, 15, 32, 22, 21,  23]
y = pd.Series(x)

# Plot
fig, ax = stemgraphic.stem_graphic(y)
```

Out [ ]:

Esta representación de los datos es semejante a la de un histograma pero, además de ser fáciles de elaborar, presentan más información que estos. También se puede representar a través de una sencilla tabla:

Tallos	Hojas
1	5 6
2	1 2 2 3 3
3	0 2
4	1

2.1.5.6. DIAGRAMA DE SECTORES

El uso principal de un diagrama de sectores es para las **variables de tipo cualitativas o discretas.** Son gráficos que sólo son útiles si **las categorías son pocas,** en el momento que hay un excesivo volumen no resulta tan comprensible.

Los datos se representan en un círculo, de modo que el ángulo de cada porción o **sector es proporcional a la frecuencia absoluta** correspondiente.

Recibe multitud de nombres, también se llama *diagrama circular,* de tarta o de queso o quesito.

Sabiendo que 360º son los grados que tiene una circunferencia, su fórmula se puede expresar de la siguiente forma:

$$\text{Ángulo} = f_i \cdot 360^o$$

EJEMPLO

En una clase de 30 alumnos/as, 12 juegan a baloncesto, 6 practican la natación, 9 juegan al fútbol y el resto no practica ningún deporte. Según la tabla siguiente:

Deporte	n_i	Ángulo	f_i	Porcentaje
Baloncesto	12	144°	0,4	40
Natación	6	72°	0,2	20
Fútbol	9	108°	0,3	30
Sin deporte	3	36°	0,1	10
$\sum$	30	360°	1	100

In [ ]:

```
# Gráfico circular o de pastel
plt.title('Gráfico circular o de pastel')

# etiquetas de los sectores
categorias = ['Baloncesto', 'Natación', 'Fútbol', 'Sin deporte']

# porciones o sectores
porcentajes = [40, 20, 30, 10]

# Destacar algunas porciones
explode = [0, 0, 0, 0]  # hay que cambiar el valor 0 por 0.1

# Plot
plt.pie(porcentajes, labels=categorias, explode=explode,  autopct='%1.1f%%', shadow=True
, startangle=90)
plt.show()
```

Out [ ]:

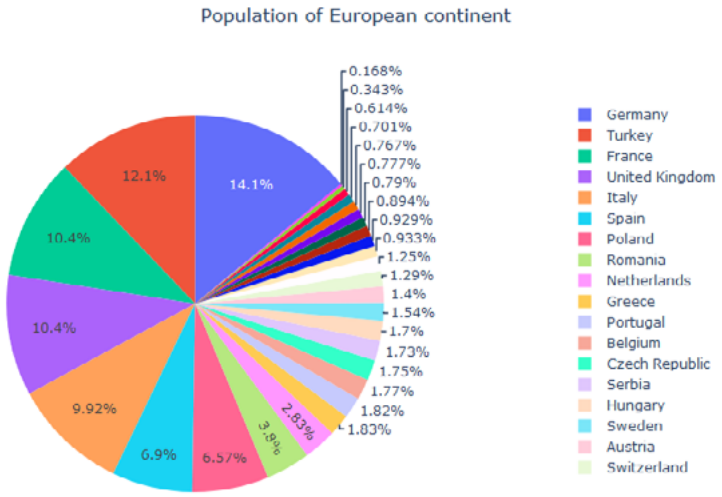
OBSERVACIÓN

Si tuviéramos 50 categorías resultaría menos legible e interpretable.

In [ ]:

```
import plotly.express as px
df = px.data.gapminder().query("year == 2007").query("continent == 'Europe'")
df.loc[df['pop'] < 2.e6, 'country'] = 'Other countries' # Represent only large countries
fig = px.pie(df, values='pop', names='country', title='Population of European continent'
)
fig.show()
```

Out [ ]:



ACTIVIDAD

Titanic

El **objetivo** de esta actividad consiste en crear una **tabla de frecuencias** para las siguientes variables:

- **PClass** (variable que clasifica la variable socioeconómica de cada pasajero/a, variable cualitativa ordinal).
- **Sex** (variable que identifica el género de los pasajeros/as, variable cualitativa binaria o dicotómica).
- **Age** (variable que tiene la edad de los pasajeros/as, variable cuantitativa continua).

Además, pedimos **añadir el gráfico correspondiente para cada tipo de variable.**

Solución

El primer paso consiste en analizar el tipo de variables, aunque te recordamos que este análisis ya se realizó en el tema 1:

Análisis del tipo de variables - Titanic

In [ ]:

```
# Se importan las librerías
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Carga del fichero desde el enlace web y creación del dataframe
url_data = 'https://raw.githubusercontent.com/md-lorente/data/master/titanic.csv'

# Creacion Dataframe
df = pd.read_csv(url_data, sep=',')

# Visualización del dataframe (la cabecera)
df.head()
```

Out[ ]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Para resolver la actividad, realiza los siguientes **pasos:**

1. Copia el dataframe con la variable que deseas analizar.
2. Añade un descriptivo para facilitar la visualización de los análisis y de los gráficos.
3. Construye la tabla de frecuencias:
   - Frecuencia Absoluta (*ni*).
   - Frecuencia Absoluta Acumulada (*Ni*). Frecuencia Relativa (*fi*).
   - Frecuencia Relativa Acumulada (*Fi*).
4. Realiza los gráficos asociados al tipo de variable.

Análisis de la variable Clase (pclass)

In [ ]:

```
# Copy df con el tipo de clase
data_class = df['Pclass'].copy(deep='True')

df_class = pd.DataFrame(data_class)

# Descriptivo de Pclass
conditions = [
        (df['Pclass'] == 1 ) ,
        (df['Pclass'] == 2 ) ,
        (df['Pclass'] == 3 ) ]
choices = ['1ª Clase', '2ª Clase', '3ª Clase']
df_class['Class'] = np.select(conditions, choices, default='-')
```

```
# Add Tabla de frecuencias del tipo de clase

## Frecuencia Absoluta
df_class['Frec Absoluta'] = df_class.groupby('Pclass')['Pclass'].transform('count')
df_class = df_class.drop_duplicates()
```

```
#df class.sort values('Pclass', ascending=False)

df_class = df_class.sort_values(by ='Pclass' )

## Frecuencia Absoluta Acumulada
df_class['Frec Absoluta Acum'] = df_class['Frec Absoluta'].cumsum()

## Frecuencia Relativa
df_class['Frec Relativa'] = df_class['Frec Absoluta'] / df_class['Frec Absoluta'].sum()

## Frecuencia Relativa Acumulada
df_class['Frec Relativa Acum'] = df_class['Frec Absoluta'].cumsum()/df_class['Frec Absolu
ta'].sum()


# View df
df_class
```

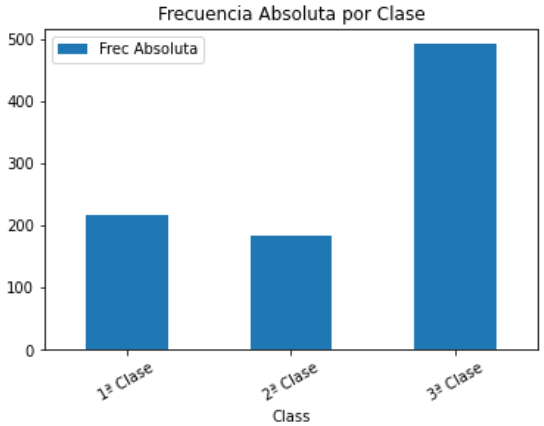
Out[ ]:

	Pclass	Class	Frec Absoluta	Frec Absoluta Acum	Frec Relativa	Frec Relativa Acum
1	1	1ª Clase	216	216	0.242424	0.242424
9	2	2ª Clase	184	400	0.206510	0.448934
0	3	3ª Clase	491	891	0.551066	1.000000

In [ ]:

```
# Gráfico de barras
df_class.plot.bar(x='Class', y='Frec Absoluta', rot=30,
                  title="Frecuencia Absoluta por Clase")
ax.invert_yaxis()
plt.show()
```

Out[ ]:

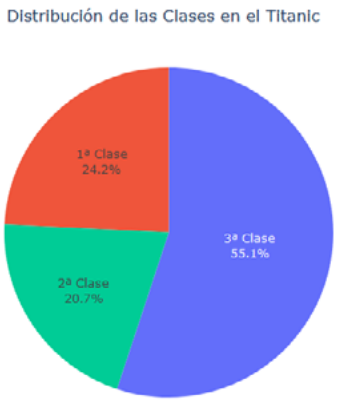


Al tratarse de una variable categórica con pocas variables, también se puede dibujar el gráfico de quesitos.

In [ ]:

```
import plotly.express as px

fig = px.pie(df_class, values='Frec Absoluta', names='Class', title='Distribución de las Clases en el Titanic')
fig.update_traces(textposition='inside', textinfo='percent+label')
fig.update_layout(showlegend=False)
fig.show()
```

Out[ ]:



Análisis de la variable Género (sex)

In [ ]:

```
# Copy df con la variablegénero
data_sex = df['Sex'].copy(deep='True')

df_sex = pd.DataFrame(data_sex)


#AddTabladefrecuenciasdeltipodeclase ## Frecuencia Absoluta
df_sex['Frec Absoluta'] = df_sex.groupby('Sex')['Sex'].transform('count')
df_sex = df_sex.drop_duplicates()

df_sex = df_sex.sort_values(by ='Sex' )

## Frecuencia Absoluta Acumulada
    df_sex['Frec Absoluta Acum'] = df_sex['Frec Absoluta'].cumsum()

## Frecuencia Relativa
    df_sex['Frec Relativa'] = df_sex['Frec Absoluta'] / df_sex['Frec Absoluta'].sum()

## Frecuencia Relativa Acumulada
    df_sex['Frec Relativa Acum'] = df_sex['Frec Absoluta'].cumsum()/df_sex['Frec Absoluta']. sum()


# View df
df_sex
```

Out[ ]:

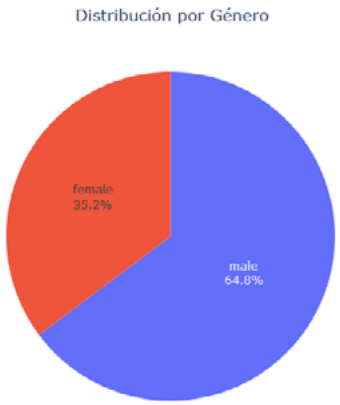
	Sex	Frec Absoluta	Frec Absoluta Acum	Frec Relativa	Frec Relativa Acum
1	female	314	314	0.352413	0.352413
0	male	577	891	0.647587	1.000000

In [ ]:

```
import plotly.express as px
```

```
fig = px.pie(df_sex, values='Frec Absoluta', names='Sex', title='Distribución por Género'
)
fig.update_traces(textposition='inside', textinfo='percent+label')
fig.update_layout(showlegend=False)
fig.show()
```

Out[ ]:



Análisis de la variable Edad (age)

In [ ]:

```
# Copy deep df con la variable edad (age)
data_age = df['Age'].copy(deep='True')

df_age = pd.DataFrame(data_age)

# Clasificación de la Edad
conditions = [
       ( (df_age['Age'] >= 0)  & (df_age['Age'] < 5)  ) ,
       ( (df_age['Age'] >= 5)  & (df_age['Age'] < 10)  ) ,
       ( (df_age['Age'] >= 10) & (df_age['Age'] < 19) )  ,
       ( (df_age['Age'] >= 19) & (df_age['Age'] < 30) )  ,
       ( (df_age['Age'] >= 30) & (df_age['Age'] < 50) )  ,
       ( (df_age['Age'] >= 50) & (df_age['Age'] < 65) ) ,
       (df_age['Age'] >= 65 )  ]
choices = ['1 Niños de 0-4', '2 Niños de 5-9', '3 Niños de 10-18',
           '4 Adultos de 19-29', '5 Adultos de 30-49', '6 Adultos de 50-64', '7 Adultos +
65']
df_age['Edad'] = np.select(conditions, choices, default='n/a')


# View df
df_age
```

Out[ ]:

	Age	Edad
0	22.0	4 Adultos de 19-29
1	38.0	5 Adultos de 30-49
2	26.0	4 Adultos de 19-29
3	35.0	5 Adultos de 30-49
4	35.0	5 Adultos de 30-49
...	...	...
886	27.0	4 Adultos de 19-29
887	19.0	4 Adultos de 19-29
888	NaN	n/a
889	26.0	4 Adultos de 19-29
890	32.0	5 Adultos de 30-49

891 rows × 2 columns

In [ ]:

```
# Se elimina age para hacer bien la agrupación
df_age.drop(['Age'], axis='columns', inplace=True)


# Add Tabla de frecuencias del tipo de clase

## Frecuencia Absoluta
df_age['Frec Absoluta'] = df_age.groupby('Edad')['Edad'].transform('count')
df_age = df_age.drop_duplicates()
df_age = df_age.sort_values(by ='Edad' )

## Frecuencia Absoluta Acumulada
df_age['Frec Absoluta Acum'] = df_age['Frec Absoluta'].cumsum()

## Frecuencia Relativa
df_age['Frec Relativa'] = df_age['Frec Absoluta'] / df_age['Frec Absoluta'].sum()

## Frecuencia Relativa Acumulada
df_age['Frec Relativa Acum'] = df_age['Frec Absoluta'].cumsum()/df_age['Frec Absoluta'].
sum()

# View df
df_age
```

Out[ ]:

	Edad	Frec Absoluta	Frec Absoluta Acum	Frec Relativa	Frec Relativa Acum
7	**1 Niños de 0-4**	**40**	**40**	0.044893	0.044893
24	**2 Niños de 5-9**	**22**	**62**	0.024691	0.069585
9	**3 Niños de 10-18**	**77**	**139**	0.086420	0.156004

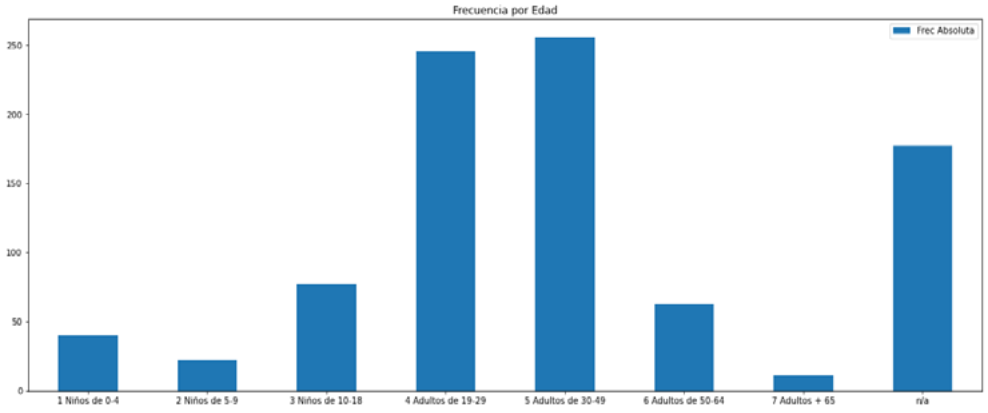
0	4 Adultos de 19-29	245	384	0.274972	0.430976
1	5 Adultos de 30-49	256	640	0.287318	0.718294
6	6 Adultos de 50-64	63	703	0.070707	0.789001
33	7 Adultos + 65	11	714	0.012346	0.801347
5	n/a	177	891	0.198653	1.000000

Al agrupar por "grupos de edad", creamos 7 categorías, por lo que la variable de tipo "continua" que se tenía en origen se ha transformado en variable cualitativa.

In [ ]:

```
# Frecuencia Absoluta
df_age.plot.bar(x='Edad', y='Frec Absoluta', title="Frecuencia por Edad",
                rot=0, width=0.5, figsize=(20,8)   )
plt.show()
```

Out[ ]:



In [ ]:

```
# Frecuencia Absoluta sin los n/a
#  print(df_age[:-1] )

df_age[:-1].plot.bar(x='Edad', y='Frec Absoluta', title="Frecuencia por Edad",
                rot=0, width=0.5, figsize=(20,8)   )
plt.show()
```

Out[ ]: