TEMA 1

MÓDULO:
FUNDAMENTOS DE ESTADÍSTICA

INTRODUCCIÓN A LA ESTADÍSTICA

DOLORES LORENTE

Diplomada en Estadística y Graduada en Estadística aplicada por la UCM. Responsable científica de datos en Big Data Analytics e Innovación.

Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

© de esta edición: Fundació IL3-UB, 2020

ÍNDICE

OBJETIVOS ESPECÍFICOS

- Asentar los principios básicos de estadística asimilando la terminología usada y de forma que teniendo unos datos, se puedan extraer unas conclusiones.
- Entender las diferencias que hay entre una variable discreta y una continua. Así como dominar los distintos tipos de variables estadística que hay.

1. INTRODUCCIÓN A LA ESTADÍSTICA

1.1. CONCEPTO DE ESTADÍSTICA

La primera cuestión para entender estadística es preguntarnos: ¿Qué es la estadística?
Si miramos en la RAE, estadística es el:

CITA

«Estudio de los datos cuantitativos de la población, de los recursos naturales e industriales, del tráfico o de cualquier otra manifestación de las sociedades humanas».

También encontramos la definición de:

CITA

«Rama de la matemática que utiliza grandes conjuntos de datos numéricos para obtener inferencias basadas en el cálculo de probabilidades».

De una forma un poco más tosca, se pude decir que los datos cuantitativos **son aquellos datos que se pueden contar.**

Un dicho que se puede usar para describir la necesidad de aplicar estadística en un estudio científico es que **"lo que no son cuentas, son cuentos".** Esto se podría interpretar como que lo que no se puede cuantificar o medir se convierte en una idea o intuición, por lo que no es lo mismo que tener una justificación o estudio científico de un fenómeno o suceso.

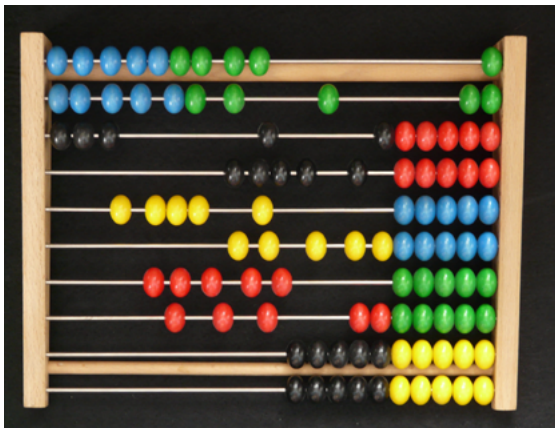

Imagen 1. Ábaco
Fuente de la imagen: *www.p1.pvxfuel.com*

Si nos vamos a *Wikipedia* y buscamos historia de estadística, se podría definir como la «disciplina ocupada de recolectar, resumir y analizar los datos ». Por tanto, la estadística se basa simplemente en tomar nota para cuantificar y **buscar para entender** cualquier objeto de estudio o fenómeno.

En definitiva, el nombre de Estadística se refiere al interés de esta rama de las Matemáticas en asuntos del Estado: censos de poblaciones, empadronamiento, índice de natalidad, de mortalidad, etc.

Actualmente, la Estadística interviene en los campos más diversos y su introducción tanto en el mundo científico como empresarial, se debe a la importancia indiscutible para el desarrollo de todas las ciencias (Psicología, Medicina, Economía, Historia, Física, Química, etc.).

Como curiosidad, comentar que uno de los primeros métodos que se usó para recolectar, resumir y analizar los datos en la historia de las matemáticas fue el ábaco, el cual permite realizar operaciones aritméticas sencillas y otras más complejas y cuyo origen real sigue siendo un misterio.

PIENSA UN MINUTO

¿Por qué crees que es tan importante la estadística?

RECUERDA

La estadística es el área de las matemáticas que **recuenta, agrupa y extrae información** de los datos. Esa información extraída ayuda a entender la realidad y a poder tomar decisiones en consecuencia.

Dicho en clave de humor: si sabes contar, contar literalmente 1, 2, 3, etc. y **si sabes sumar,** agrupar o resumir datos **entonces, sabes estadística.** Así de sencilla y fácil es la estadística.

- **Todo aquello que es susceptible de ser medido** (contado) o que se mide, entonces, **se puede analizar con estadística, por eso es tan importante y abarca tantas disciplinas.**

- Quizás por eso es preocupante que haya datos, que aun siendo conocidos (a través de medios de comunicación como periódicos o televisión) no se tomen los datos para analizar el grado de impacto real en la sociedad.

1.2 CONCEPTOS BÁSICOS

Para entender los conceptos de los siguientes temas, primero hay que comprender toda la terminología usada en estadística.

1.2.1. POBLACIÓN

Es el conjunto de todos los elementos cuyas propiedades se van a estudiar. En algunos sitios recibe el nombre de Universo. Existen dos categorías:

1. **Población finita.** Cuando se puede determinar a priori la cantidad exacta de elementos u observaciones. Por ejemplo:

    - Número exacto de universidades que hay en España.

- Número exacto de trabajadores en una empresa o compañía.
- Número de hijos que tiene una familia.

2. **Población infinita.** Cuando no se puede determinar a priori una cantidad exacta o es imposible de determinar. Por ejemplo:

- Número de estrellas en el universo.
- Número de granos de arena en el mar.
- Número de insectos a nivel mundial.

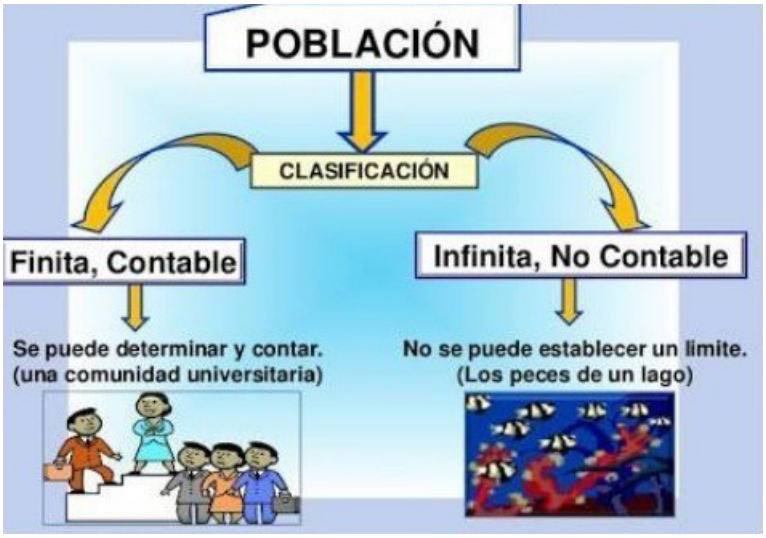

Imagen 2. Población finita e infinita
Fuente de la imagen: *www.portafoliosegundoparcialestadistica.blogspot.com*

1.2.2. MUESTRA

En muchas ocasiones, no se puede trabajar con toda la población, por lo que se usa un **subconjunto de la población.** Esto es, **una muestra representativa de la misma** y, para lograrlo hay que usar criterios y técnicas de muestreo.

Como norma general, **una muestra representativa debe reflejar todas las características de la población.**

Por ejemplo, si se desea conocer el resultado de las próximas elecciones de un país, resulta poco viable preguntar a toda la población de ese país. Por lo que se opta por obtener un subconjunto de esta población, asumiendo un margen de error, a partir del cual se podrán sacar conclusiones sobre la votación de toda la población.

Existen **dos clasificaciones** para la muestra estadística:

1. **Probabilística.** Todos los elementos de la población tienen la misma posibilidad de formar parte de ella. El azar interviene de alguna manera. Por Ejemplo:

- Todas las personas enfermas de un hospital, sin importar la enfermedad o las patologías que presenten.
- Todo el alumnado de una Universidad, sin importar la nota.

2. **No probabilística.** No todos los elementos de la población tienen la misma posibilidad de formar parte de la muestra. No se usa el azar, sino el criterio del investigador, es decir, se decide si la muestra es o no representativa. Por ejemplo:

- Todas las personas enfermas de un hospital que presentan una misma enfermedad (Covid) o patología (diabetes). Las personas que no presentan esa patología no son incorporadas en el estudio.
- Todo el alumnado de una Universidad que suspende o abandona. El estudio quiere determinar las causas de suspenso o abandono y no se incorporan los alumnos/as con aprobados.

$$\text{Tipos de Muestreo} = \begin{cases} \text{Probabilístico} & \text{Misma probabilidad} \\ \text{No Probabilístico} & \text{Distinta probabilidad} \end{cases}$$

NOTA

Hay distintas tipologías dentro del muestreo que se desarrollarán más adelante.

Imagen 3. Población y muestra
Fuente de la imagen: *www.economipedia.com*

En consecuencia, **muestra y población son conceptos relativos.**

Una **población** es ***un todo*** y una **muestra** es una ***fracción o segmento*** de ese todo.

1.2.3. EL DATO

Definición de dato

Cada uno de los individuos, cosas, entes abstractos que **integran una población o universo determinado.** Dicho de otra forma, cada valor observado de la variable.

Por ejemplo:

Se pueden estudiar personas, animales, enfermedades, temperaturas, bombillas, ordenadores, etc.

Cualquier estudio estadístico comienza con **la recogida de datos.** Esta recogida **puede ser física y directa** (aparatos de medición, encuestas, etc.) **o virtual** mediante la consulta de información a través de un programa.

Los datos pueden proceder de distintas instituciones u organismos públicos, de la navegación por Internet, de la propia compañía en que trabajemos o de datos recogidos a través de encuestas.

El objetivo de la recogida de datos, es poder analizarlos posteriormente. Por tanto, hay que procurar, en la medida de lo posible, hacer **que la presentación de estos datos sea de forma sencilla, coherente y atractiva para el consumidor final.**

En este sentido, la estadística dispone los datos, generalmente, en tablas y, en muchas ocasiones, se ayuda a su vez de gráficos que pretenden aclarar aspectos reseñables de los datos recogidos.

1.2.4 VARIABLE ESTADÍSTICA

La variable estadística es una **característica o cualidad de un individuo que se pueden medir, y puede adquirir diferentes valores.**

Por ejemplo, el color de los ojos, la estatura de una persona, el grupo sanguíneo o las notas obtenidas de una prueba.

Existen dos clasificaciones para las variables estadísticas:

1. **Cuantitativas o Numéricas.** Como su propio nombre indica, son variables que se pueden cuantificar, es decir, se pueden medir. Son de tipo numérico y, por tanto, se representan numéricamente. Dentro de esta clasificación hay dos tipologías:

   1.1 **Variables Discretas:** utilizan valores enteros y finitos. Son valores numéricos indivisibles. Por ejemplo: número de hijos, número de coches, número de viviendas, etc.

   Su representación matemática es el conjunto de los números naturales:

   $$N = \{1, 2, 3, 4,...\}$$

   1.2 **Variables Continuas:** utilizan valores objetivos y pueden tomar un número de valores infinitos. Se caracterizan por utilizar valores decimales (divisibles). Ejemplos: la altura, el peso, el dinero, etc.

Su representación matemática es el conjunto de los números racionales:

$$Q = \{a \div b \text{ donde } a, b \in Z \text{ y } b \neq 0\} = \{1.3, 2, 2.578, 3.275, 4,. . .\}$$

Variables Cuantitativas

Tipos	Definición	Ejemplos
Discreta	La variable solo puede tomar valores en número determinado de valores. En cada intervalo de valores la variable solo puede tomar un valor.	– Canastas en un partido (20; 21; 22; pero no 21,5). – Hijos por familia (0, 1, 2, 3,...).
Continua	La variable puede adquirir cualquier valor dentro de un intervalo de valores determinado.	– Peso (53,53 kg; 89,4 kg,...).

2. **Cualitativas o Categóricas.** Como su propio nombre indica, son variables que están vinculadas a una cualidad. Es decir, las variables cualitativas no pueden ser calculadas con números, sino que son clasificadas con palabras. Dentro de esta clasificación hay tres tipologías:

2.1 **Cualitativas ordinales:** son aquellas que siguen un orden o una jerarquía. Por ejemplo, el nivel socioeconómico (alto, medio o bajo), la nota de un examen (suspenso, aprobado, bien, notable, sobresaliente), potencia de un aparato o pulsaciones de un corazón (leve, moderado, fuerte).

2.2 **Cualitativas nominales:** se trata de aquellas variables que no siguen ningún orden en específico. Por ejemplo, el grupo sanguíneo, el color de pelo, el color de los ojos, etc.

2.3 **Cualitativa binaria o dicotómica:** es un caso particular de variable nominal con solo dos categorías o dos resultados posibles.

Si las dos categorías determinan dos estados cualesquiera se denomina ***binaria simétrica*** (por ejemplo: sexo). Si el 1 determina la presencia de una característica y el 0 su ausencia (como por ejemplo: enfermedad, trabajo, etc.), la variable se denomina ***binaria asimétrica.***

Variables Cualitativas

Tipos	Definición	Ejemplos
Nominal	Son variables cualitativas cuyas categorías no siguen ningún orden.	– Color (blanco, rojo, azul,...). – Lateralidad (zurdo, diestro).
Ordinal	Son las variables categóricas con orden o jerarquía.	– Nota examen. (suspenso, aprobado, notable, sobresaliente). – Nivel económico. (pobre, clase media, rico). – Medalla deportiva. (Oro, plata, bronce).
Binaria o dicotómica	Es un caso particular de la variable nominal. Son sólo dos categorías.	– Sexo (mujer, hombre). Simétrica. – Enfermo (si, no). Asimétrica.

Esquema resumen tipos de variables

$$Variables = \begin{cases} \text{Cuantitativas o Numéricas} \begin{cases} \text{Discreta} \\ \text{Continua} \end{cases} \\ \text{Cualitativas o Categóricas} \begin{cases} \text{Nominal} \\ \text{Ordinal} \\ \text{Binaria o Dicotómica} \end{cases} \end{cases}$$

Tipos de variables en el análisis de datos

En función del análisis que se realice, **las variables pueden recogerse de manera numérica pero ser categóricas.**

Por ejemplo, la variable fecha es una variable con la que se toma el dato (y se almacena) de manera numérica pero no es una variable cuantitativa ya que no tiene sentido sumar literalmente los años (por ejemplo: 2020+2021=4041).

La variable como tal, es una variable recogida o expresada de forma numérica (2020) cuyo uso es categórico, y sirve, por ejemplo, para clasificar la información sobre el año en que ocurrió un suceso.

Otro ejemplo podrían ser los códigos postales, que se almacena en forma de código numérico para hacer una consulta ágil de la información, aunque desde el punto de vista analítico sirve para clasificar ya que es el equivalente a los nombres de las poblaciones a las que haga referencia. Por ejemplo: el 28020 (corresponde a Madrid) y el 08300 (equivaldría a poner Barcelona).

En función del análisis de datos, existen **dos tipos de variables** que son:

- **Variables categóricas o de clasificación:** también se denominan variables cualitativas o variables de atributos. Los valores de una variable categórica son categorías o grupos mutuamente excluyentes. Los datos categóricos pueden tener o no tener un orden lógico. Por ejemplo:
  - Zonas geográficas (regiones, provincias, etc.) de una ciudad.
  - Clasificación del tipo de cliente.
  - Tipos de vehículo (gasolina, diésel, eléctrico, etc.).
- **Variables analíticas o cuantitativas:** son números que suelen representar una medición (frecuencia o conteo, sumas, etc.). Por ejemplo:
  - Importe de unas facturas.
  - Número de veces que ocurre una enfermedad.
  - Coste de mantenimiento de un hogar.

Este tipo de variables, a veces, necesita un tratamiento y no se puede trabajar con el dato "en bruto". Un ejemplo suele ser "la edad" cuyo valor en sí no aporta mucha información de forma individual. Sin embargo, aporta mucha información si se hace por rangos o mediante una clasificación. Por ejemplo: niños [0-18 años), adultos [18-65 años) y mayores [>65 años).

Resumen tipos de variables

$$Variables = \begin{cases} \text{Categórica o de Clasificación} \\ \text{Analítica o Cuantitativa} \end{cases}$$

IMPORTANTE

Las variables **binarias o dicotómicas** pueden ser tratadas, a su vez, de forma **categórica y analítica** (teniendo en cuenta que sólo se puede medir la frecuencia en este apartado).

RECUERDA

- La **variable estadística** es la que después **se relacionará con algún problema o fenómeno, que se desea investigar,** entender su causa o fenómenos asociados y/o buscar posibles soluciones.
- **El primer paso de cualquier tipo de análisis** radica en **estudiar el tipo de variables** de cualquier dataset.
- Una **variable estadística es una cualidad o cantidad** que poseen los individuos u objetos de estudio de una población de análisis.

1.3. MANIPULACIÓN BÁSICA DE LOS DATOS: EXPLORAR EL DATASET

Para poder describir el dataset, se introducirá el uso de la librería ***pandas,*** que es un paquete de ***Python*** que provee de una serie de funciones, métodos y estructuras de datos para trabajar los dataset de forma fácil e intuitiva.

Pandas está construido sobre la librería de numpy. Esto produce que algunos métodos o funciones de pandas no retornen tipos de datos creados en la propia librería, como los *DataFrame* o *Series,* sino que devuelva estructuras de numpy como el *numpy.ndarray.* Al mismo tiempo, pandas puede cargar, fácilmente, matrices que provengan de numpy.

Para empezar a trabajar, lo primero que necesitamos es poder cargar el dataset con pandas y, luego, lograr obtener alguna información sobre sus columnas. Con esta finalidad, se abrirá un fichero en formato .csv para, posteriormente, analizar de qué tipo son las variables.

La sentencia de comandos para importar las librerías con las que trabajaremos es ***import*** seguida del paquete, al que se le asocia un alias para mayor comodidad. Esto es:

```
In [ ]:
```

```
# Importing the libraries
import pandas as pd
import numpy as np
```

ACTIVIDAD

Titanic

El **objetivo** de esta actividad consiste en identificar el tipo y el subtipo de variables que corresponda.

Antes de empezar a procesar, analizar y visualizar los datos, es necesario **entender el dataset.** La fuente del *dataset* es de *Kaggle,* donde se facilita un archivo en formato CSV, con información del famoso accidente del Titanic.

También, se proporciona la información del diccionario de datos que se muestra a continuación:

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	Nivel Socieconómico: 1 = 1st (upper), 2 = 2nd (middle), 3 = 3rd (lower)
sex	Sex	
Age	Age in years	Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5
sibsp	# of siblings / número de hermanos, hermanas o marido/mujer	
parch	# of parents / número de parientes (padre, madre, hijos).	A algunos de los niños que viajaban sólo con la nanny se les puso la notación 0 para ellos.
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Solución

Abrir un archivo CSV

Para realizar tal acción se utiliza la función ***read_csv*** que retorna un DataFrame con la información del archivo. A continuación, vamos a cargar el archivo y aguardarlo en un *dataframe.*

IMPORTANTE

Las columnas están separadas por comas (",").

In [ ]:

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

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

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

Out[ ]:

PassengerId	Survived	Pclass	Name	Sex	Age	Sib sp	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/02. 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

In [ ]:

```
# Resumen información del fichero
print(df.info())
```

Out[ ]:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
None
```

Lo primero que debemos hacer al trabajar con un *dataset* es analizar el tipo de variables con el que se trabajará:

1. **PassengerId:** es la identificación de cada uno de los pasajeros del Titanic. El código es único, consecutivo y entero. Es una variable **numérica de tipo discreto.**

2. **Survived:** es la identificación de si sobrevivió o no al accidente. Es una variable **cualitativa binaria o dicotómica asimétrica.**

3. **Pclass:** es una variable que mide el nivel socioeconómico. Se clasifica en 3 niveles: bajo, medio y alto. Por tanto, es una variable **cualitativa ordinal.**

4. **Name:** es una variable que recoge el nombre del pasajero. Es una variable **cualitativa nominal.**

5. **Sex:** es una variable que identifica el género del pasajero: hombre o mujer. Por tanto, es una variable **cualitativa binaria o dicotómica simétrica.**

6. **Age:** es la edad del pasajero. Cuando tiene decimales es una edad estimada (el decimal 0.5 indica que es una edad estimada). La variable edad, en general, es una variable cuantitativa discreta, pero tal y como se han tomado los datos es una variable **cuantitativa continua.** Nótese: que tiene decimales.

7. **SibSp:** es el número de familiares: hermanos, hermanas o marido/esposa. Por tanto, se trata de una variable **cuantitativa discreta.**

8. **Parch:** es el número de familiares ascendientes o descendientes. Al igual que antes, se trata de una variable **cuantitativa discreta.**

9. **Ticket:** es el número del ticket del barco. Dado que el billete contiene números y letras, para identificar si tiene un orden establecido o no, nos fijamos en los datos que tienen una clasificación (empiezan por CA o PC seguido de un número secuencial). Por tanto, se trata de una variable **cualitativa ordinal.**

10. **Fare:** es la tarifa que paga cada pasajero. Es, por tanto, una variable **cuantitativa continua.**

11. **Cabin:** es el número de cabina, el cual tiene un orden de letra y número para clasificar las habitaciones dentro del barco. Es una variable **cualitativa ordinal.**

12. **Embarked:** es el puerto de embarque, en el que hay tres lugares donde se embarcaron los pasajeros: Cherbourg, Queenstown y Southampton. No tiene un orden, por lo que se trata de una variable **cualitativa nominal.**

Un paso más allá...

De las variables analizadas, realizaremos una parada para analizar 3 variables:

- Ticket.
- Cabin.
- Age.

Análisis de la variable Ticket

In [ ]:

```
# Análisis de la variable
Ticket billetes =
pd.unique(df['Ticket'])
billetes.sort()
print(billetes)
```

Out[ ]:

```
['110152' '110413' '110465' '110564' '110813' '111240' '111320' '111361'
 '111369' '111426' '111427' '111428' '112050' '112052' '112053' '112058'
 'A./5. 2152' 'A./5. 3235' 'A.5. 11206' 'A.5. 18509' 'A/4 45380'
 'A/5. 3336' 'A/5. 3337' 'A/5. 851' 'A/S 2816' 'A4. 54510' 'C 17369'
 'C 4001' 'C 7075' 'C 7076' 'C 7077' 'C.A. 17248' 'C.A. 18723' 'C.A. 2315'
```

Hay mucha tipología distinta. Puedes observar que hay clasificación formada principalmente por un prefijo y seguida por una numeración.

Los prefijos son, entre otros:

- A/5 o A/4.
- C + número C.A.
- PC.
- SOTON o STON.

Filtraremos solo por el grupo de los que empiezan por PC (Private Cabin) para observar si la numeración sigue un orden.

In [ ]:

```
# Filtro para ver los que empiezan por PC
filtro_billetes =[col for col in billetes if col.startswith('PC')!=0]

print(filtro_billetes)
```

Out[ ]:

```
['PC 17318', 'PC 17473', 'PC 17474', 'PC 17475', 'PC 17476', 'PC 17477', 'PC 17482', 'PC 1
7483', 'PC 17485', 'PC 17558', 'PC 17569', 'PC 17572', 'PC 17582', 'PC 17585', 'PC 17590',
'PC 17612', 'PC 17754', 'PC 17755', 'PC 17756', 'PC 17757', 'PC 17760', 'PC 17761']
```

Comprueba que hay saltos en la numeración, es decir no es una numeración consecutiva aunque a veces si se mantiene. Esto es, el primer elemento es PC 17318 y salta al PC 17473, luego sigue un orden hasta el PC 17477 donde se produce otro salto al 17482.

Claramente, sigue un "orden". No sigue un orden convencional seguramente debido a la clasificación del coste del billete asociado al tipo de cabina asociado.

Análisis de la variable Cabin

In [ ]:

```
# Ordenamos y quitamos valores nulos a Cabin
cabina = pd.unique(df['Cabin'].dropna())
cabina.sort()
print(cabina)
```

Out[ ]:

```
['A10' 'A14' 'A16' 'A19' 'A20' 'A23' 'A24' 'A26' 'A31' 'A32' 'A34' 'A36'
 'A5' 'A6' 'A7' 'B101' 'B102' 'B18' 'B19' 'B20' 'B22' 'B28' 'B3' 'B30'
 'B82 B84' 'B86' 'B94' 'B96 B98' 'C101' 'C103' 'C104' 'C106' 'C110' 'C111'
 'C90' 'C91' 'C92' 'C93' 'C95' 'C99' 'D' 'D10 D12' 'D11' 'D15' 'D17' 'D19'
 'E58' 'E63' 'E67' 'E68' 'E77' 'E8' 'F E69' 'F38' 'F4' 'G6' 'T']
```

SABÍAS QUE...

La función de cabina es un poco complicada y necesita un poco más de exploración. Falta la gran parte de la función *Cabin* y la función en sí no puede ignorarse por completo porque algunas de las cabinas pueden tener tasas de supervivencia más altas.

La primera letra de los valores de *Cabin* son las cubiertas en las que se encuentran las cabinas. Esas cubiertas se separaron, principalmente, para una clase de pasajeros, pero algunas de ellas fueron utilizadas por múltiples clases de pasajeros.

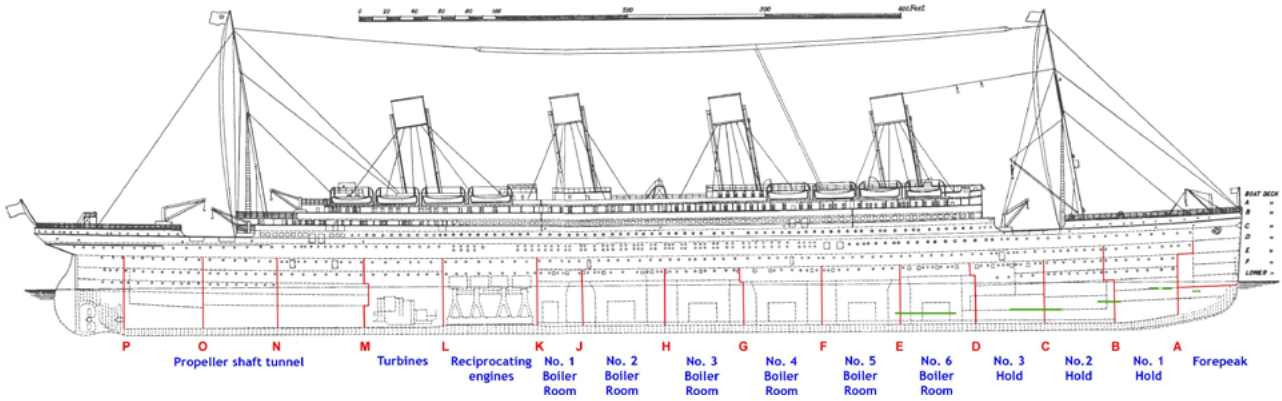

Imagen 4. Estructura de la cabina del Titanic
Fuente de la imagen: *www.vignette.wikia.nocookie.net*

- On the Boat Deck, there were 6 rooms labeled as T, U, W, X, Y, Z, but only the T cabin is present in the dataset.
- A, B and C decks were only for 1st class passengers.
- D and E decks were for all classes.
- F and G decks were for both 2nd and 3rd class passengers.
- From going A to G, distance to the staircase increases which might be a factor of survival.

Análisis de la variable Age

In [ ]:

```
edad = pd.unique(df['Age'])
edad.sort()
print(edad)
```

Out[ ]:

```
[ 0.42  0.67  0.75  0.83  0.92  1.    2.    3.    4.    5.    6.    7.
  8.    9.   10.   11.   12.   13.   14.   14.5  15.   16.   17.   18.
 19.   20.   20.5  21.   22.   23.   23.5  24.   24.5  25.   26.   27.
 28.   28.5  29.   30.   30.5  31.   32.   32.5  33.   34.   34.5  35.
 36.   36.5  37.   38.   39.   40.   40.5  41.   42.   43.   44.   45.
 45.5  46.   47.   48.   49.   50.   51.   52.   53.   54.   55.   55.5
 56.   57.   58.   59.   60.   61.   62.   63.   64.   65.   66.   70.
 70.5  71.   74.   80.     nan]
```

La edad, en general, se suele representar como variable cuantitativa discreta, esto es, con números enteros (1, 2,3,...) sin valorar los meses (días, horas, minutos o segundos).

Aunque en algunos estudios resulta relevante, como en las compañías aseguradoras cuando trabajan seguros de vida que miden milimétricamente el riesgo a asumir, y, por tanto, en ellas no se considera igual a una persona de 70 años que a otra con 70 años y 6 meses.

Como decíamos, las compañías aseguradoras miden el riesgo, luego una persona con 70 años y 6 meses tiene más probabilidades de ocurrencia que una persona recién cumplidos los 70 años que presenta menor riesgo. Cada vez el mercado es más competitivo, por lo que las compañías "ajustan" y miden mejor.

Lo normal es que en este tipo de mediciones, se busque mayor precisión y, por ende, la edad será una variable cuantitativa de tipo continuo (contemplando unidades decimales que podrán representar los meses o días).

Por tanto, no sólo hay que analizar la naturaleza de la información, si no también cómo está recogida dicha información. En general, será más fácil que las variables de tipo discreto estén alineadas.

Por ejemplo, nadie pondría un hijo y medio. Normalmente, en estadística, se estima al alza, por tanto, si sale en promedio un hijo y medio se pondrá como resultado 2 hijos.

Aunque la naturaleza de la información influye en el tipo de variable, no es determinante. Así pues, la manera en la que se recoge el dato influye en el tipo de variable. Un caso en el que se ve claramente son las herramientas de medición (sonidos, temperaturas, balanzas, etc.) puesto que pueden ser más o menos precisas.

Por ejemplo, la temperatura es una variable de tipo continuo, pero puede ocurrir que el aparato de medición no tenga suficiente precisión, por lo que al almacenar la información sólo contenga unidades enteras (30º, 35º, etc.) en lugar de unidades decimales (30.5º, 35.33º, etc.). En este caso, la variable temperatura en origen es de tipo cuantitativa continua aunque la toma del dato ha transformado la variable a una cuantitativa de tipo discreto.

IMPORTANTE

Como norma general, ***se tratan los datos como se han recogido,*** aunque siempre conviene hacer una mención y explicar cuando ocurran ese tipo de "incongruencias".

Desde el punto de vista del análisis de datos, el **tipo de variables** sería:

$$Variables = \begin{cases} \text{Clasificación: Pclass, Name, Sex, Age, Ticket, Cabin, Embarked, Survived.} \\ \text{Analíticas: SibSp, Parch, Fare, Survived (sólo conteo).} \end{cases}$$

OBSERVACIONES

- ***PassengerId:*** dato que es usado para agilizar el cruce de información y que, al mismo tiempo, no aporta información para el análisis de los datos.

- ***Survived:*** al ser una variable dicotómica (sobrevive si o no), su uso principal es de tipo clasificatorio, pero también se puede usar de manera analítica (como por ejemplo para realizar un conteo simple que sirva para saber la proporción de supervivientes que hubo dentro de toda la población).

IDEAS CLAVE

- **La estadística** es el estudio que **recuenta, clasifica y analiza todos los hechos** que tienen una determinada característica en común, para poder **llegar a conclusiones a partir de los datos** numéricos extraídos.

- La estadística **es una herramienta que utilizamos para la toma de decisiones** en cada uno de los campos de nuestro entorno social, por tanto es una materia importante dentro de nuestro proceso de aprendizaje.

- Es imprescindible analizar la información **antes de profundizar en cualquier tipo de análisis** y, para ello, el **primer paso consiste en analizar el tipo de variables** con el que trabajaremos.