

Análisis e Interpretación de Datos

Tema 1. Introducción a la estadística

Índice

Esquema

Ideas clave

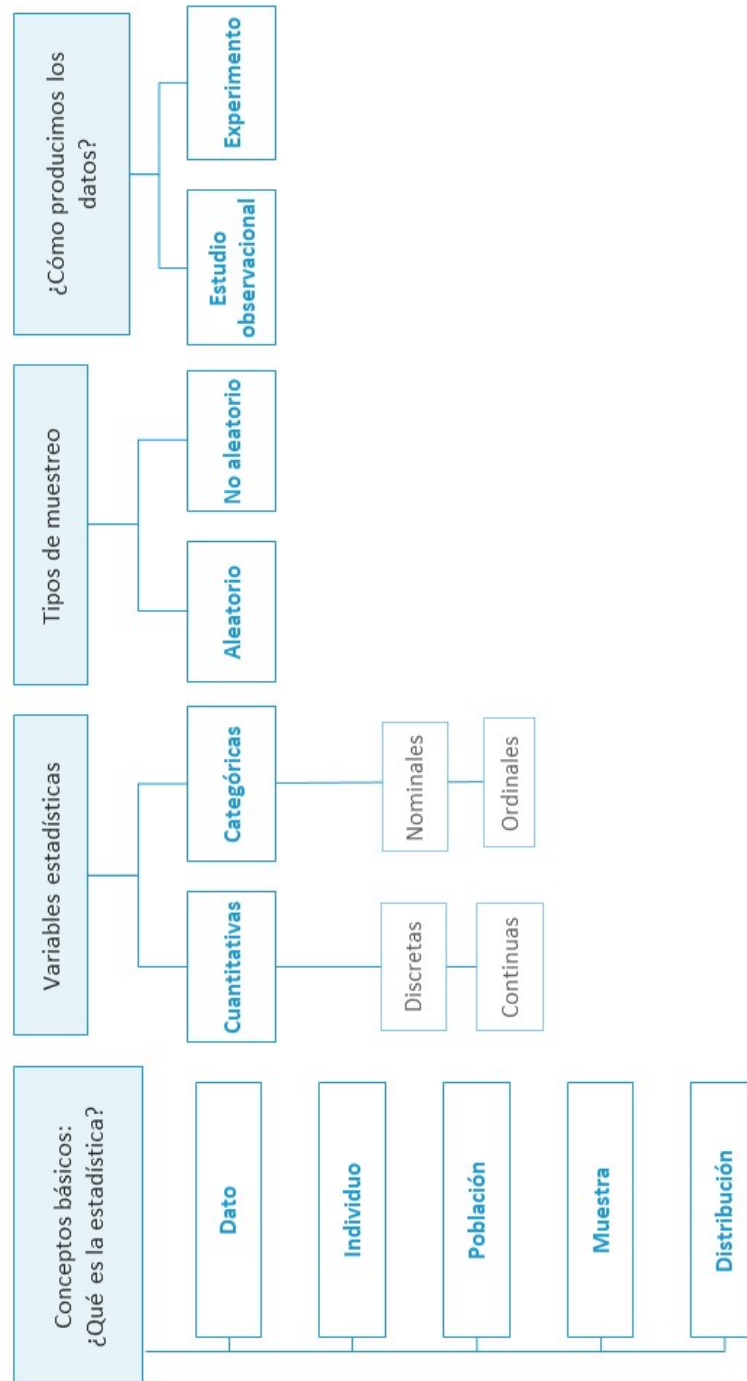
- 1.1. ¿Cómo estudiar este tema?
- 1.2. ¿Qué es la estadística?
- 1.3. Población, muestra y muestreo
- 1.4. Tipos de variables estadísticas
- 1.5. Diseño de experimentos
- 1.6. Razonamiento estadístico
- 1.7. Representando los datos: distribución de frecuencias
- 1.8. Tabulación de variables
- 1.9. Gráficas básicas
- 1.10. El arte de elegir el gráfico adecuado
- 1.11. Retos de la estadística en el Big Data
- 1.12. Referencias bibliográficas

A fondo

- Realizando un informe Analytics
- Efecto Hawthorne
- Series temporales
- Estadística antes que cálculo
- Técnicas de representación de datos
- Bibliografía

Test

INTRODUCCIÓN A LA ESTADÍSTICA



1.1. ¿Cómo estudiar este tema?

Para estudiar este tema lee las **páginas 13-37** del siguiente libro:

Ríos, F. (1998). *Bioestadística: Métodos y aplicaciones*. Málaga: Universidad de Málaga. Publicaciones.

Versión electrónica:

<https://www.bioestadistica.uma.es/baron/apuntes/clase/apuntes/pdf/>

Este primer tema consta de una parte introductoria para repasar los conceptos y técnicas clave sobre los que trabaja la ciencia estadística y también aborda una primera necesidad que surge a partir de los datos, sobre cómo organizarlos y presentarlos. O dicho de otro modo, este capítulo trata de responder a esta cuestión: ¿Cómo organizamos los datos para poder comprender la información que contienen? (O como diría Moore, para «aprender» de ellos.

También será clave que practiques con los ejercicios que vienen al final del tema, los cuales están diseñados para que apuntes las ideas más importantes sobre tablas de frecuencias y gráficos estadísticos. Los dos esquemas que acompañan este tema te pueden ayudar a hacerte una buena idea de cómo está organizado.

1.2. ¿Qué es la estadística?

Podemos pensar en un primer lugar que la estadística es simplemente una colección de datos cualquiera. Así decimos informalmente estadísticas del paro, de intención de voto, etc. Pero esta definición no es la que nos interesa, ya que hace mención a estudios concretos, pero no expresa una visión de esta disciplina como ciencia que estudia los datos de manera más amplia.

Una definición un tanto exhaustiva de la estadística diría que es la ciencia que maneja los datos a través de un proceso que va desde el diseño del estudio, recogida de los datos, análisis, para finalmente organizar, resumir y mostrar la información contenida en ellos para sacar conclusiones. De manera resumida podemos dar otra definición: la **estadística** es la ciencia que nos permite aprender de los datos (Moore, 2006).

Conviene aclarar que el hecho que no se desarrolle el proceso estadístico completo con todas sus fases no quiere decir que no se «haga estadística». Podemos realizar estadísticas partiendo de datos ya producidos (habiéndose hecho previamente el diseño y la recogida de datos) de modo que comencemos nuestra labor estadística en la fase de análisis de datos.

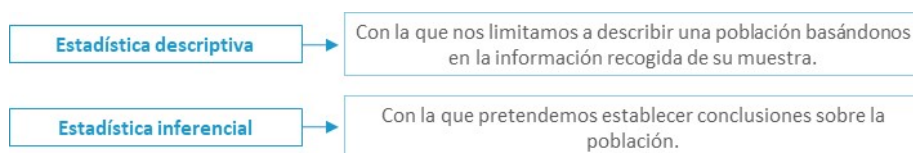


Ejemplo 1: De este modo en una misma empresa puede haber empleados y empleadas en diferentes puestos, encargándose uno de ellos del diseño del experimento para recoger los datos, otro de recogerlos, una tercera de analizarlos y un cuarto de exponerlos en una presentación delante del jefe de la empresa para que este pueda tomar las decisiones oportunas. Cada uno de los cuatro empleados está trabajando a su manera como estadístico pero en una fase diferente.

Todas las fases de un estudio estadístico son **igualmente importantes**, pero, de hecho, se suele decir que no hay buen análisis posible si los datos han sido recogidos de cualquier manera sin seguir unos criterios estadísticos mínimos, y es por ello que la etapa de recogida de datos es sumamente delicada y de suma importancia. Luego veremos cómo garantizar la recogida de unos «buenos» datos. Además, las fases explicadas anteriormente tampoco son únicas, pues otros autores afirman que el identificar una cuestión o problema de estudio también es en sí una fase previa.

Alguien podría preguntarnos alguna vez «¿para qué sirve la estadística?». Entonces, podríamos responderle, no sin razón, que el objetivo de la estadística es «ganar en comprensión de un fenómeno a partir de los datos que se manejan sobre este» (Moore, 2006).

La estadística de acuerdo al nivel de uso pretendido que le demos como herramienta puede ser de dos clases:



Los primeros temas de este curso se centran en la que tradicionalmente se llama **estadística descriptiva**, mientras que más adelante, con la probabilidad veremos la parte inferencial, aquella que descansa sobre un aparato matemático mayor y que nos permitirá fundamentar gran parte de las técnicas estadísticas conocidas.

1.3. Población, muestra y muestreo

La definición de estadística emplea primeramente el concepto de **dato**, que no solo es un número, sino un número en un contexto, con lo cual es **información** recolectada sobre algo. Pero ese «algo» es lo que llamaremos **individuo** el cual conforma un colectivo que llamamos **población**, que es finalmente sobre lo que nos interesa estudiar y sacar conclusiones. Por lo tanto, la estadística no se encarga de cualquier fenómeno, sino de aquellos que son colectivos y que no atienden a leyes deterministas (de las cuales se encargan las ciencias exactas), es decir, de aquellos que contienen algún elemento de **incertidumbre**.

El proceso mediante el cual seleccionamos a los individuos que van a formar parte de la muestra se denomina **muestreo** y es clave para garantizar un mínimo de calidad en los datos obtenidos (es decir, una información importante sobre la población), que ayude a validar futuros análisis y conclusiones. Lo deseable al recoger la muestra es que los individuos seleccionados configuren una **muestra representativa** de su población, es decir, que contenga una diversidad muy similar a la de la población de origen.

Siempre que obtengamos una muestra estamos expuestos al **error de muestreo**, producto de inferir o extrapolar a partir de un trozo de realidad (la muestra), el cómo será la realidad entera (la población). La clave será reducir este error, inherente al propio proceso de **muestreo**, al mínimo.

El proceso de extrapolar las características y propiedades de la muestra a las de la población se conoce como **inferencia estadística** y, dada su importancia, ha devenido en una rama de la estadística (generalmente se habla de estadística descriptiva y de la inferencial).

Ejemplo 2: En la Encuesta sobre Medios de Transporte que realizó el consorcio madrileño de transportes hace unos ocho años los encargados del estudio querían responder entre otras cuestiones a la siguiente pregunta concreta: «¿cuál es el uso que le están dando los madrileños al transporte público en la zona de la sierra de Madrid?».

Para ello los encuestadores fueron debidamente formados y realizaron encuestas en pueblos de la serranía. Lo que ocurre es que no les pudieron preguntar a todos los habitantes de todos los pueblos, ya que ello hubiera excedido los costes presupuestados.

De modo que se seleccionó una muestra aleatoria de viviendas para que sus inquilinos fueran encuestados y posteriormente se respondió a la pregunta a partir de los datos de la muestra recogida extrapolándolos a toda la población de Madrid.

Como el estudio anterior son en realidad todos los estudios que se llevan hoy en día en España, pues **los censos** o registros exhaustivos a toda la población ya no se practican desde el año 2000 cuando tuvo lugar el último censo de la población española.

1.4. Tipos de variables estadísticas



Tal y como observamos en el esquema existen dos tipos de variables estadísticas: las **categóricas** y **cuantitativas**. La primera de ellas está dividida a su vez en dos clases, dependiendo de si las categorías son meramente cualitativas, son las llamadas **nominales**, o si además poseen orden, las **ordinales**. Las cuantitativas pueden ser **discretas** cuando toman un número finito de valores o **continuas** cuando pueden tomar infinitos valores como por ejemplo las magnitudes físicas (altura, peso, etc.)

La clasificación anterior de los tipos de variables no es única. Otros autores las subdividen de otro modo, aunque este es probablemente el más común. También podemos **clasificar las variables según su enfoque metodológico**:

- ▶ Variables dependientes.
- ▶ Variables independientes.

Las dependientes son las que sus valores dependen de los que tomen otros de acuerdo a un determinado rol hipotético que asumimos que juega cada variable y que hará que planteemos un modelo estadístico u otro en nuestros análisis estadísticos (como cuando planteamos una regresión lineal).

Ejemplo: aprobado en Lengua en el 1er Cuatrimestre será variable dependiente de otra independiente como puede ser el número de horas de estudio de Lengua. Se supone que pretendemos explicar el hecho de aprobar Lengua a partir del número de horas estudiadas para la asignatura, lo cual parece razonable (aunque existirán otros factores).

Es por ello que también recibe el nombre de **variable explicada o respuesta**, mientras que la independiente también recibe el nombre de **variable explicativa o predictora**. Depende del gusto de los autores el emplear una terminología u otra, porque en el fondo, variable dependiente, respuesta y explicada por un lado, e independiente, explicativa y predictora por el otro, no son más que sinónimos de un mismo rol que desempeña la variable. En economía u otras disciplinas pueden emplearse otros términos equivalente como variables endógenas y exógenas, etc.

Otro tipo de variable al que conviene ponerle nombre es el de las **variables intermediarias u omitidas**, variables que no son contempladas por el estudio o el modelo planteado en cuestión, pero que en el fondo estarían actuando de variables explicativas de nuestra variable dependiente, pero de un modo digamos oculto, o mejor dicho «desde la sombra». Conviene identificarlas para no establecer asociaciones y presuponer causalidades infundadas.

Ejemplos en el terreno educativo son la renta familiar sobre el rendimiento escolar, el profesor sobre la motivación del alumno y el ambiente familiar sobre la integración de los estudiantes. La variable nivel de estudios de los padres es un ejemplo clásico de este tipo de variables. En ocasiones los análisis estadísticos se realizan «controlando» el efecto de dichas variables para eliminar determinado influjo sobre la variable respuesta en el cual no estamos interesados (El análisis de covarianza o ANCOVA permite este tipo de controles, aunque son técnicas que se ven en cursos más avanzados de estadística).

Otro tipo de variable muy empleado en estadística es el de las **variables dicotómicas**, ya que son muy útiles para describir el hecho de que ocurra algo (1) o no ocurra (0).

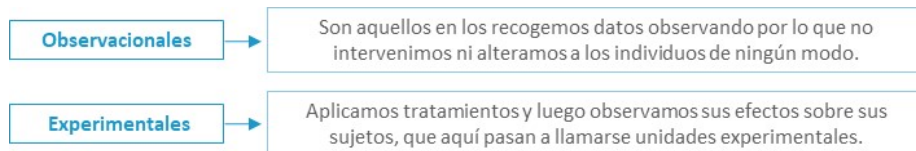
En la práctica una misma variable puede ser recodificada de diferentes modos, como por ejemplo la variable edad. En teoría se trata de una variable continua (la edad es el tiempo pasado desde el nacimiento, que es una magnitud continua), sin embargo, puede ser recogida en su dimensión puramente categórica ordinal si solo apuntamos o codificamos los intervalos de edad, tal y como ocurre en numerosas encuestas. (Ejemplo: Menor de edad- De 18 a 25 años- Mayor de 25).

Ejemplos de cada tipo de variable son:

- ▶ Categórica nominal es el género, el grupo al que pertenecen los alumnos, etc.
- ▶ Categórica ordinal es el curso al que pertenecen los alumnos (Ejemplo: 1ºESO, 2ºESO,..., 2ºBachillerato).
- ▶ Cuantitativa discreta es el número de asignaturas suspensas en un cuatrimestre.
- ▶ Cuantitativa continua es el tiempo empleado en hacer el examen.

1.5. Diseño de experimentos

Los estudios estadísticos pueden ser de dos clases:



Un estudio observacional es cualquier encuesta de las vistas anteriormente, ya que en ellas no apliquemos cambios ni sometamos a ningún tratamiento a los encuestados. Los diseños experimentales se emplean muy a menudo en la rama bioestadística, ya que es habitual aplicar tratamientos médicos y luego querer observar las diferencias entre ellos.

1.6. Razonamiento estadístico

Para aprender a pensar estadísticamente debemos desarrollar un pensamiento crítico basado en varias preguntas (adaptadas de *Estadística* de Triola, 2009):

1. ¿Cuál es el objetivo del estudio?
2. ¿Quién es la fuente de los datos?
3. ¿Con que tipo de muestreo han sido obtenidos los datos?
4. ¿Existen variables que influyan en los resultados y que se hayan omitido?
5. ¿Las gráficas resumen adecuadamente los datos?
6. ¿Las conclusiones se extraen directa y naturalmente de los datos?
7. ¿Se ha cumplido el objetivo marcado al principio del estudio y tienen sentido y utilidad práctica las conclusiones obtenidas?

El hecho de plantearnos quién es la fuente es importante porque esta puede, en un momento dado, no ser neutral con el resultado de los objetivos del estudio y este interés propio puede alterarlos. A esto muchas veces se le llama el «cocinado» de datos que viene a ser esa pequeña o grande manipulación y preparación que sufren las conclusiones de los datos para beneficio de quien presenta los resultados del estudio.

Diremos entonces que el estudio estadístico tiene un **sesgo**. Este concepto es fundamental para el pensamiento estadístico, y todas las preguntas anteriores deben ir enfocadas a plantearnos si existe o no sesgo. Por supuesto, existen muchas fuentes de sesgo donde la anterior es tan solo la más coloquial. Es donde solemos decir: «tal o cual estudio o investigación están sesgados...». Cuando veamos los estimadores y sus propiedades en temas posteriores aprenderemos otras variaciones del concepto de sesgo.

Ejemplo 3: Los grandes medios de comunicación suelen colaborar asiduamente con una misma agencia de estudios de opinión, la cual se encarga por ejemplo de sondear los votos a los partidos en un momento coyuntural concreto. Este tipo de estudio se puede prestar a sesgo por diferentes motivos.

Entre ellos, diríamos que el momento en el que se realiza el estudio, el momento en que se publica, la ideología predominante en los dueños de la agencia de comunicación en cuestión, el uso de cuestionarios un tanto restringidos o con preguntas dirigidas que pudiera haber producido un **efecto de redacción en la pregunta**, etc.

Ejemplo 4: Imagínate que eres un analista de datos y tienes que empezar a trabajar los análisis sobre un archivo Excel cuya tabla de datos es la siguiente:

| Y | X |
|------|------|
| 0,50 | 9,89 |
| 7,62 | 1,03 |
| 5,73 | 7,43 |
| 1,90 | 7,92 |
| 4,65 | 6,20 |
| 7,68 | 5,29 |
| 2,96 | 9,45 |
| 2,31 | 8,46 |
| 1,27 | 3,42 |
| 3,19 | 7,05 |

Si no te dan ninguna información extra a partir de aquí no podrías realizar estadísticas con sentido pues desconoces el contexto en que se ha producido estos datos, a las variables que hacen referencia X e Y, cómo han sido recogidos, etc.

Si se te facilita más información y puedes saber que estas variables pertenecen a unas actas de una asignatura de un grado universitario y que son una m.a.s. de 10 alumnos por cada uno de los grupos del curso, mañana y tarde, los cuales corresponden respectivamente a las columnas X e Y.

1.7. Representando los datos: distribución de frecuencias

Ahora vamos a pasar la fase de organización y representación de datos. Lo primero que se nos ocurre hacer con los datos es contarlos. Anotar sus repeticiones, es decir, el número de veces que se repite un valor o una categoría de una variable. A estas magnitudes las llamamos **frecuencias**.

Clasificamos las frecuencias de la siguiente manera:

- ▶ Las **absolutas**, que denotamos n_i donde la i hace referencia a la categoría o valor i -ésimo de la variable (también llamado **modalidad**).
- ▶ Las **relativas** que se obtienen como las absolutas en relación al N total o suma de todas las frecuencias absolutas de todas las modalidades, que en realidad no es más que el tamaño de la muestra:

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

- ▶ Las **absolutas acumuladas** que resultan de ir sumando las frecuencias de las modalidades de la variable hasta una dada. Para diferenciarlas de las anteriores se las distingue con letras mayúsculas: N_1, N_2, \dots, N_k . Dándose entonces la circunstancia que N_k , que es la última frecuencia absoluta acumulada (que a veces simplemente se dice «frecuencia acumulada» por abreviar) coincide con el tamaño de la muestra N . Matemáticamente: $N_i = n_1 + \dots + n_i$, para $i > 1$.
- ▶ Las **relativas acumuladas** que por analogía con las anteriores son las sumas de las frecuencias relativas hasta determinada modalidad de la variable.

$$F_i = \frac{N_i}{N}, \text{ y donde } F_k = 1.$$

1.8. Tabulación de variables

Las clases de frecuencias anteriores las organizamos y presentamos mediante una **tabla de frecuencias**, la cual consta de k filas, correspondientes a cada una de las k modalidades de que consta la variable.

| Modalidades | Frecuencias (absolutas) | Frecuencias relativas | Frecuencias absolutas acumuladas | Frecuencias relativas acumuladas |
|-------------|----------------------------|--------------------------|--|--|
| 1 | n_1 | f_1 | N_1 | F_1 |
| 2 | n_2 | f_2 | N_2 | F_2 |
| ... | ... | ... | ... | ... |
| k | n_k | f_k | N | 1 |
| SUMA | N | 1 | | |

La forma más empleada de tabla de frecuencias consiste en la columna de los valores y sus frecuencias normales, es cuando se pretende registrar más información cuando se incorporan el resto de columnas. En la práctica se suelen incluir las columnas de frecuencias «normales» y la de relativas pero en forma de porcentajes.

Ejemplo 5:

| | Frecuencia | % | % válido |
|-----------------|------------|-------|----------|
| Tiempo completo | 111 | 74,49 | 87,40 |
| Tiempo parcial | 16 | 10,73 | 12,60 |
| No aplicable | 22 | 14,76 | |
| TOTAL | 149 | | |

En esta tabla se aprecia que en el lugar que tendría que figurar la columna de frecuencias relativas la suplantán los porcentajes. El motivo es claro si se tiene en cuenta que se trata de conceptos equivalentes, las frecuencias relativas son al tanto por uno lo que los porcentajes al tanto por cien.

No todos los individuos tienen que tener asociado obligatoriamente un valor para cada variable, cuando esto sucede diremos que el individuo presenta un **valor perdido** (o *missing*) en dicha variable. Cuando existen valores perdidos es habitual colocar otra columna en la tabla de frecuencias con la coletilla «válidos», dando a entender que en esa columna no se contabilizan los valores perdidos. Esto sucede en la tabla anterior tal y como se puede apreciar, ocurriendo que en este caso se considera la modalidad o categoría «No aplicable», que a efectos prácticos se trata de un caso especial de perdidos cuando no procede su respuesta por parte del individuo.

Ejemplo 6: Cuando en una encuesta se pregunta primero si se tienen hijos y a continuación en otra pregunta cuántos hijos se tienen, esta segunda pregunta dará lugar a valores «no procede» o «no aplicables» para los individuos que hayan contestado que no tienen hijos en la primera.

Un caso aparte dentro de las tablas de frecuencias es aquel en el que las modalidades de la variable continua se muestran por **intervalos**. En este caso tenemos que considerar los conceptos de límite inferior y superior del intervalo, y el valor que representará a dicho intervalo que se denomina **marca de clase** del intervalo. Esta marca de clase tendrá su utilidad como valor promedio o representante de dicho intervalo, aspecto que trataremos en el tema siguiente cuando veamos las medidas resumen estadísticas. Al ser el valor o punto medio del intervalo se calcula así:

$$x_i = \text{marca de clase} = \frac{L_{i-1} + L_i}{2}$$

| Modalidades | Marcas de clase | Frecuencias |
|---------------|-----------------|-------------|
| L_0-L_1 | x_1 | n_1 |
| L_1-L_2 | x_2 | n_2 |
| ... | ... | ... |
| $L_{k-1}-L_k$ | x_k | n_k |

Ejemplo 7:

| Modalidades | Marcas de clase | Frecuencias |
|-------------|-----------------|-------------|
| 15-19 | 17 | 3575 |
| 20-24 | 22 | 4985 |
| ... | ... | ... |
| 60-64 | 62 | 1257 |

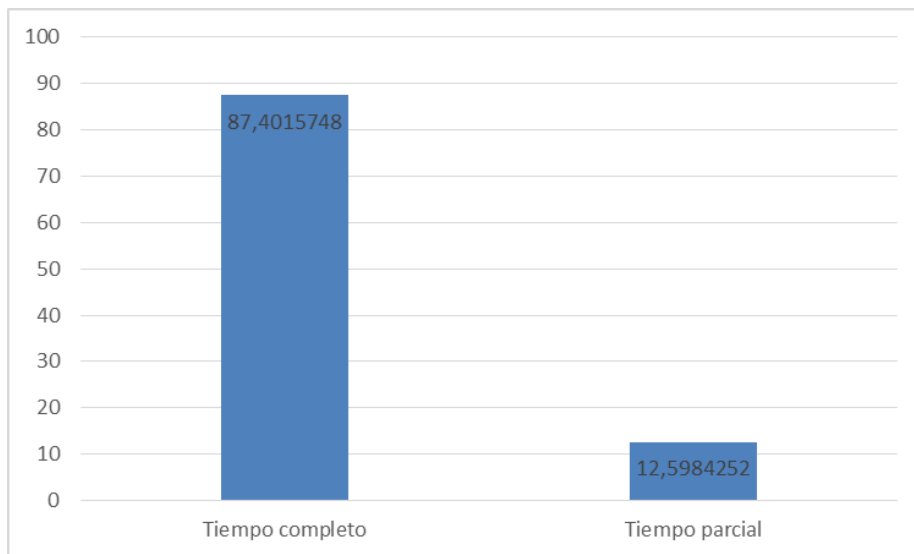
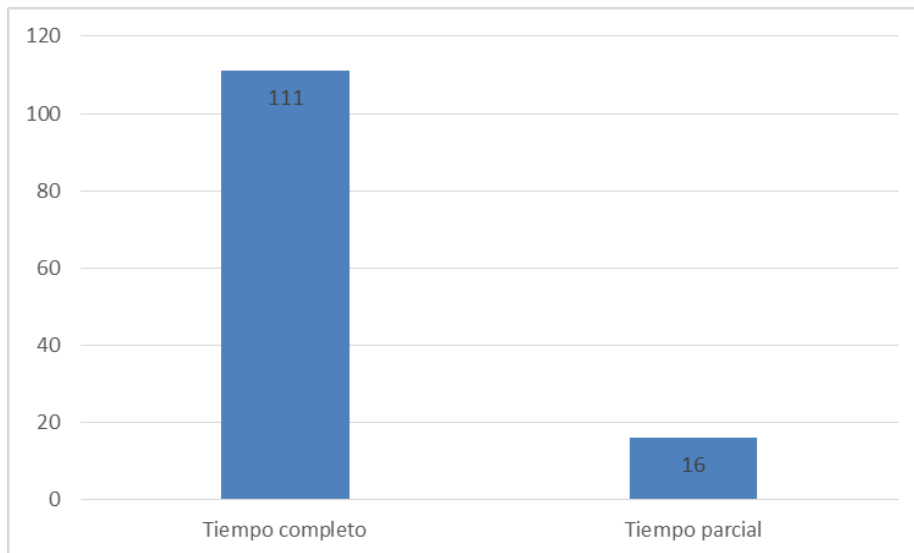
1.9. Gráficas básicas

Existe un dicho en estadística: «Más vale un buen gráfico que mil tablas de frecuencias». Si bien puede que sea una exageración, en muchos casos es cierto. Visualmente somos capaces de asimilar cosas más rápidamente y con mayor claridad que codificadas de un modo más complejo y analítico.

Uno de los dilemas clave cuando tenemos una base o conjunto de datos es el siguiente: ¿Cómo describir visualmente tales o cuales variables? O dicho de otro modo, ¿cuál es el gráfico idóneo para representarlos? Antes de responder a estas cuestiones es necesario saber la «oferta» de gráficos disponible para saber elegir el adecuado. Es en esta cuestión en la que nos centraremos en este apartado.

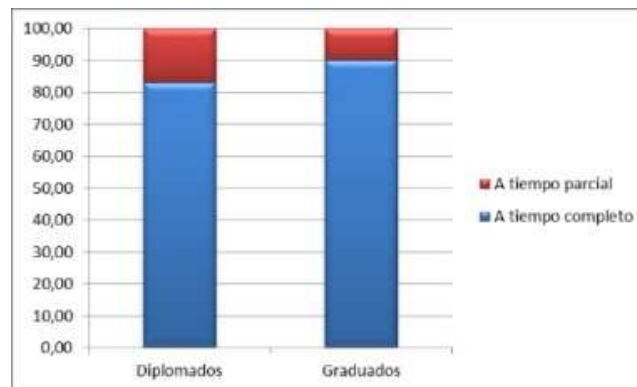
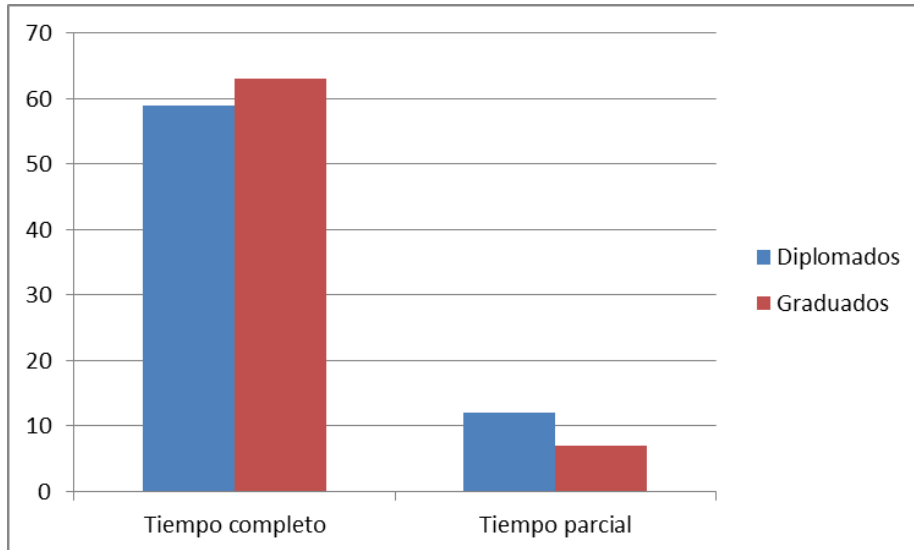
La pista esencial para saber que gráfico nos corresponde confeccionar es el tipo de variable que se pretende representar. El primer caso que se nos presenta es cuando tenemos variables de «tipo categórico» (en realidad no existe tal división pero a nivel práctico es útil manejarla), que pueden ser tanto cualitativas (de ambos tipos: nominales y ordinales) como cuantitativas discretas, donde cada valor discreto sería una de las categorías. En estos casos utilizaremos **diagramas de barras**. Lo anterior equivale a decir que todas las variables pueden ser representadas con diagramas de barras excepto las continuas.

Ejemplo 8:



En ocasiones los diagramas de barras pueden ser un poco más complejos, esto ocurre cuando «cruzamos» dos variables categóricas.

Ejemplo 9:

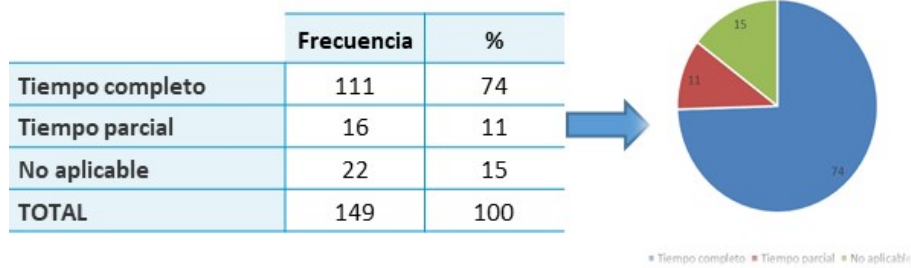


De los dos gráficos anteriores es más habitual el diagrama de barras de la izquierda, siendo el de la derecha un caso especial menos frecuente — pero con sus «adeptos» — denominado **diagrama de barras apiladas**.

Para representar gráficamente variables cualitativas tenemos el **gráfico de sectores**, también llamado gráfico circular, de porciones, de tarta, o *pie chart* en inglés (*pie* = tarta).

Se trata de un gráfico muy habitual que estamos más o menos acostumbrados a ver por doquier. El único requisito que hay que tener en cuenta es el de representar los porcentajes de las modalidades y que estos siempre sumen el 100%. El área o sector circular que ocupa cada modalidad es proporcional a su porcentaje en relación con el total. Es preferible usarlo cuando el número de categorías no es excesivo. Cuando hay muy pocas diferencias entre las categorías o porciones podríamos plantearnos realizar el gráfico de barras en su lugar.

Ejemplo 10:



Otro gráfico de uso habitual y exclusivo para las variables cualitativas es el **pictograma**, el cual como su propio nombre apunta se trata de un gráfico que se basa en un dibujo. La elección de este gráfico puede reportar ventajas cuando queremos acentuar ciertas diferencias o porque se trata de un elemento que visual o simbólicamente tiene cierta potencia.

Ejemplo 11: Para resumir información de carácter militar el pictograma puede ser muy apropiado, sobre todo de cara a acentuar ciertas diferencias a la hora de comparar. Un ejemplo clásico es el de comparar el gasto militar entre países o bien el de las armas militares como en el gráfico siguiente:



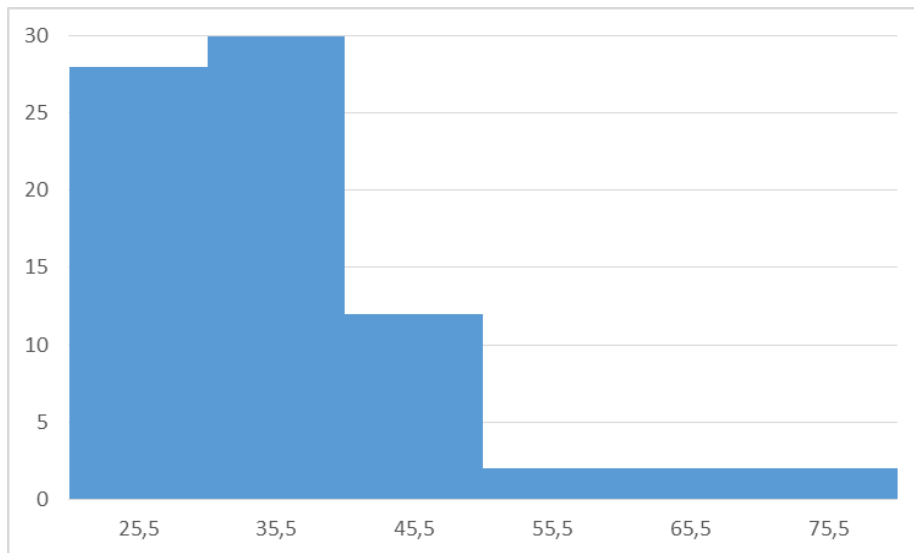
Uno de los **errores habituales** que se cometen en este tipo de gráficos es representar cada modalidad según su valor y dibujando cada elemento con esta escala. Esto no es correcto pues hay que considerar que las áreas de los dibujos tienen que ser proporcionales a las magnitudes que representan.

En el ejemplo anterior si se fija uno bien el valor en millones de euros del segundo misil, el *Meteor* es el doble aproximadamente que el del *Sparrow* y, sin embargo, no es el doble de alto el primero que el segundo sino que es su área la que es aproximadamente el doble. El criterio para comparar en los pictogramas será, por tanto, el área, tal y como apuntan algunos autores (Ríus et al., 2006, 25). Según lo dicho las frecuencias serán proporcionales al tamaño de estas áreas.

Uno de los motivos que hace que el uso de los pictogramas sea limitado se debe al hecho de que no estén disponibles en los principales programas que se emplean para la elaboración de gráficos estadísticos como pueden ser el Excel y el SPSS.

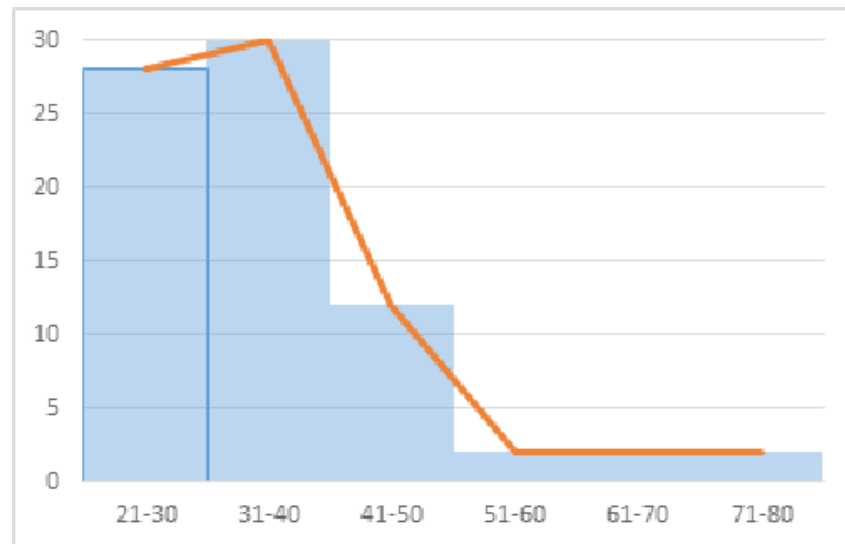
En el caso de las variables cuantitativas disponemos de otros gráficos básicos. El equivalente de algún modo al diagrama de barras en el caso cuantitativo continuo es el **histograma**. Este gráfico nos permite comunicar la continuidad a través de las **barras juntas**. Se suele emplear cuando disponemos de la información agrupada en intervalos, que es la manera más común en la que se manejan las variables cuantitativas continuas.

Ejemplo 12: En el siguiente caso representamos las estatuillas de Oscar ganadas por actrices dependiendo de su edad (Triola, 2009). La variable «edad» es continua de modo que parece apropiado mostrar su distribución con un histograma. El valor que figura en el eje de abscisas es la marca de clase de cada intervalo.

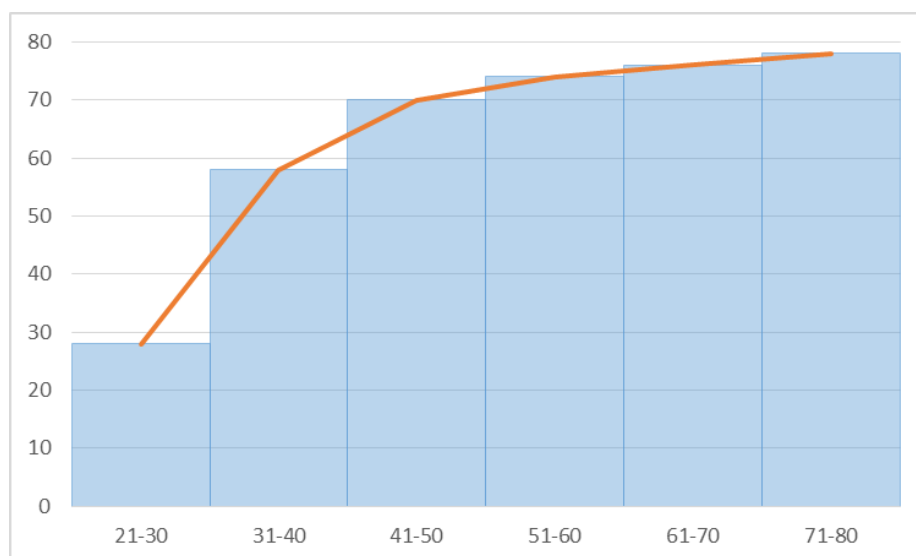


Un gráfico algo menos empleado que el histograma es el **polígono de frecuencias** que se obtiene al unir los puntos medios de las barras del histograma (muestro con el color de relleno rebajado el histograma asociado que no tendría por qué figurar acompañando al polígono de frecuencias).

Ejemplo 13:

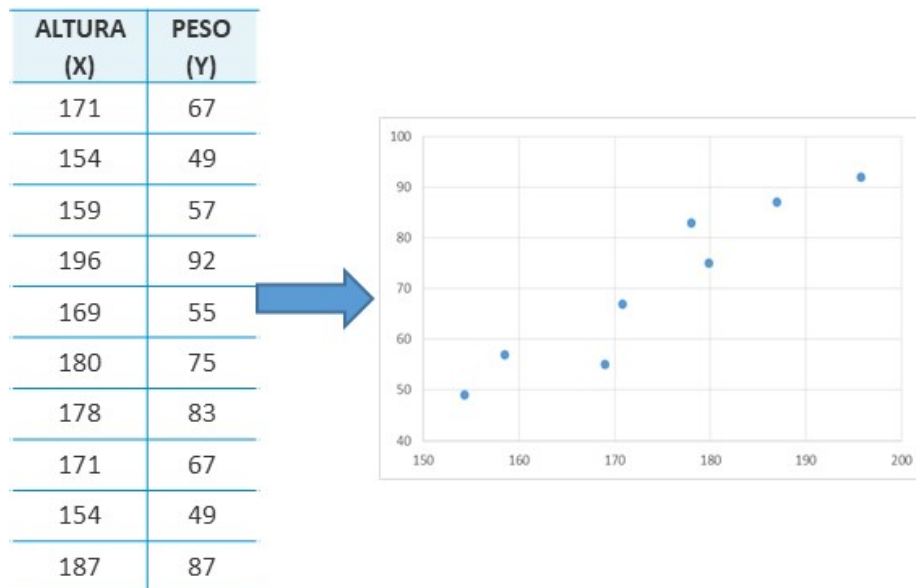


Este gráfico, al ser pura línea, acentúa las tendencias, por lo que viene bien para representar las frecuencias acumuladas, tal y como se ve en la siguiente versión:



Otro gráfico muy empleado en el caso cuantitativo es el de **dispersión** (también llamado **nube de puntos**) el cual nos sirve para representar los valores de un individuo en dos variables continuas.

Ejemplo 14:



Cuando se dispone de frecuencia mayor que uno para algún par (x_i, y_j) lo que se hace es situarlos muy próximos entre sí indicando que todos esos puntos (n_{ij} puntos para ser más exactos) representan al mismo par.

También es posible mostrar la información de una variable cualitativa con este gráfico diferenciando los puntos por colores o usando un símbolo. Por ejemplo «H» o «M» para indicar género (Hombre y Mujer).

Otra gráfica muy común en nuestro día a día (sobre todo en las secciones de economía de los periódicos) es la llamada **serie temporal** (*time plot* en inglés), en la que se muestran una línea que recorre diferentes valores o frecuencias a lo largo del tiempo. La variable temporal se sitúa siempre en el eje horizontal.

Para aprender más sobre series temporales consulta el apartado A fondo.

1.10. El arte de elegir el gráfico adecuado

Uno de los problemas habituales cuando tenemos un conjunto de datos y nos disponemos a representarlos gráficamente es que **no sabemos por dónde empezar**. Es raro encontrar un libro que aborde esta cuestión explícitamente, pero lo cierto es que es un momento en el cual llegamos a dudar de que el gráfico que vamos a emplear sea realmente el más adecuado o que no pareciendo que sea erróneo sospechamos que tiene que haber algún otro gráfico que sea realmente bueno para describir los datos.

Y entonces, **¿cuál es el gráfico más adecuado para mis datos?** Lo primero que tenemos que tener en mente para responder con seguridad a esta pregunta es la siguiente tabla, que aunque al principio quizás tengamos que acudir a ella con cierta frecuencia, acabaremos por interiorizarla a nuestra manera.

| Tipo de variable | | Opciones gráficas | |
|------------------|----------|--|-------------------------|
| Cualitativa | Nominal | Diagrama de barras, sectores, pictograma | |
| | Ordinal | | Diag. Barras acumuladas |
| Cuantitativa | Discreta | Diagrama de barras ("normales" y acumuladas) | |
| | Continua | Histograma, dispersión (dos continuas), serie temporal | Polígono de frecuencias |

1.11. Retos de la estadística en el Big Data

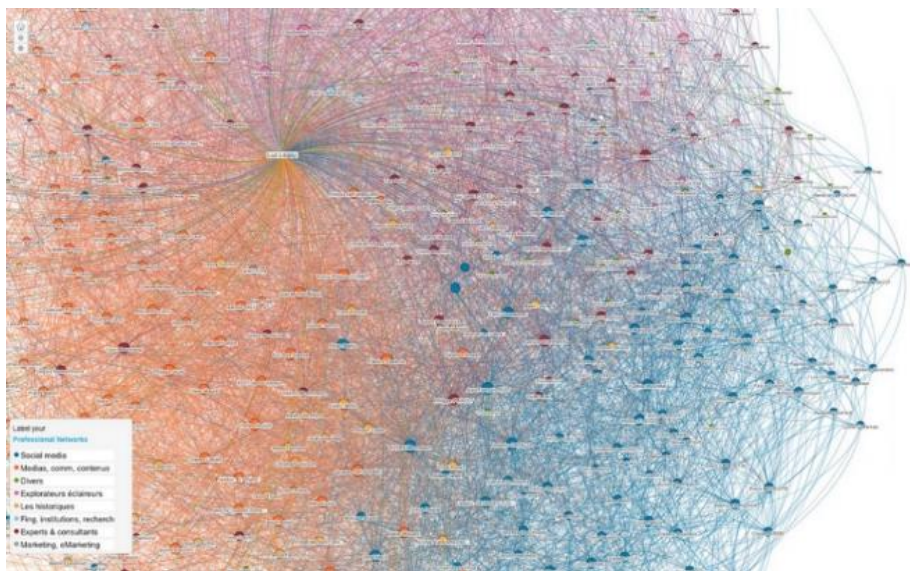
La **estadística es una disciplina clásica**. Su actual definición como «ciencia que recolecta y analiza los datos» proviene del **siglo XIX**. Está bastante claro que con la aparición de los computadores y, más recientemente, de Internet y el Big Data, **los entornos con los que actualmente trabaja la estadística han variado enormemente**. Mientras que antes se trabajaban con conjuntos relativamente pequeños de datos, actualmente la cantidad de información que hay disponible para llevar a cabo todo tipo de análisis está más allá casi de nuestro propio entendimiento. Esto genera un problema que hasta ahora nunca había sucedido: «**tenemos tantos datos que no hay manera de analizarlos**». La consecuencia de esto es que a pesar de que nunca habíamos tenido tantos datos, **somos incapaces de aprender nada de ellos**. Y ¿de qué sirve realmente entonces tener los datos? La respuesta es: para nada. Para solucionar esto, la estadística debe adaptarse a este nuevo entorno y desarrollar nuevos métodos y prácticas que nos permitan analizar y aprender de los datos que tenemos a nuestra disposición.

De manera más específica, estos son **los problemas a los que tiene que enfrentarse la estadística clásica**, al ser aplicada a entornos Big Data:

- ▶ **1. Excesiva cantidad de información y datos:** generalmente, los métodos estadísticos no están pensados para manejar grandes cantidades de datos por lo que, en general, **no están diseñados para ser especialmente eficientes**. Esto puede provocar problemas al aplicar estos métodos a grandes cantidades de datos debido a que el **tiempo necesario para llegar a cabo los cálculos necesarios puede ser inviable**. Por tanto, se hace necesaria la creación de códigos eficientes que nos permitan:
 - Aplicar los métodos estadísticos clásicos necesarios.

- Desarrollar nuevos métodos estadísticos que sean capaces de trabajar con altas cantidades de información.

Otro problema importante asociado a la gran cantidad de información disponible es el que generan en este tipo de conjuntos de datos **los outliers**. La tendencia de los métodos estadísticos clásicos es la de **la eliminación y supresión de los outliers**. Cuando trabajamos con conjuntos reducidos de datos, este enfoque puede resultar adecuado debido a que la cantidad de outliers es reducida. Sin embargo, cuando trabajamos en entornos Big Data, **los outliers pueden estar formado por una cantidad muy grande de datos**. Por ello, **eliminarlos u obviarlos puede no ser la solución más adecuada**.



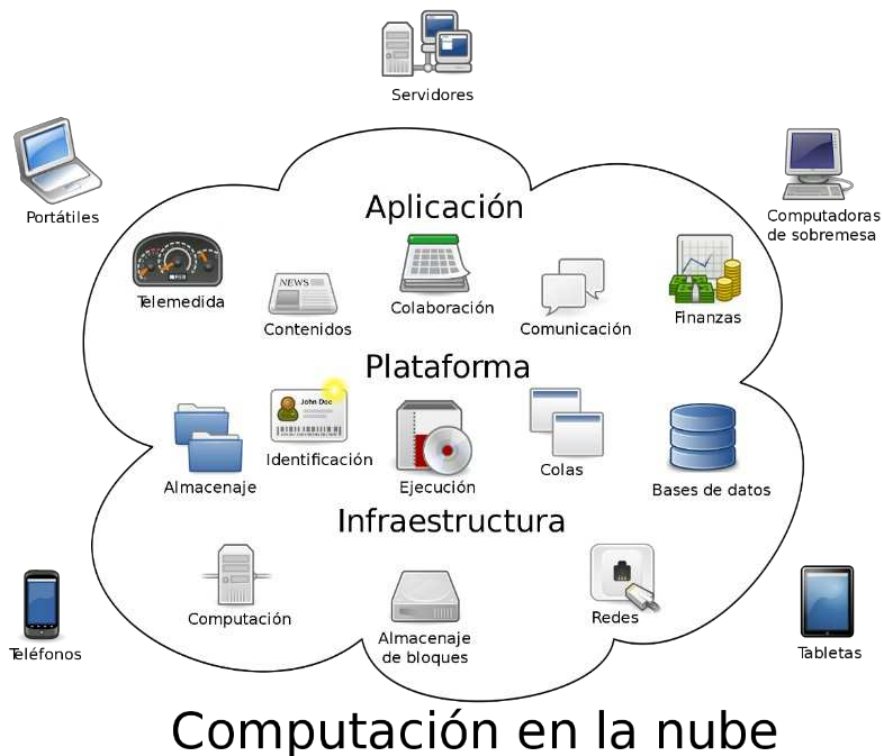
Red de usuarios. Fuente: https://c1.staticflickr.com/6/5217/5418037955_d361ba38ce_b.jpg

- ▶ **2. Complejidad de los datos:** la gran cantidad de datos en entornos Big Data no es el único problema que la estadística debe resolver para poder trabajar de forma adecuada con los entornos actuales. La **complejidad inherente a la información disponible es otro gran reto**. Disponemos de muchos datos pero, además, **dichos datos son extremadamente complejos y difíciles de interpretar**. Esto es debido, sobre todo, a su procedencia. Por lo general, los datos con lo que se suele trabajar

en Big Data, son **datos extraídos de usuarios de Internet**. Es lo que se conoce como «**la huella digital**». Multitud de páginas web almacenan de forma automática datos relativos a todos los usuarios que pasan por ellas. Este compendio de información contiene, por lo general, **datos referentes a todo tipo de actuaciones que los usuarios llevan a cabo en la web**. La heterogeneidad de dicha información hace necesaria, por parte de los métodos de análisis estadístico, de la aplicación de procesos que permitan **transformar los datos de forma que puedan ser fácilmente interpretados y analizados**.

- ▶ **3. Necesidad de infraestructuras potentes de análisis:** la gran cantidad de datos disponibles hace necesaria la utilización de entornos de computación extremadamente eficientes que permitan proporcionar los resultados de los análisis en tiempos adecuados. Por suerte, gracias a los clústeres y a las recientes tecnologías de computación en la nube, la capacidad de procesamiento de información y de cómputo de los ordenadores actuales ha aumentado exponencialmente. Por tanto, es posible crear una red de procesadores o pagar un módico precio para la utilización de un clúster en la nube y tener, de esta manera, acceso a un entorno de computación que nos proporcione suficiente capacidad de cómputo para los análisis que queramos realizar.

Para aprovechar al máximo las infraestructuras de cómputo, es interesante hacer uso de **métodos que sean fácilmente paralelizables**. De esta manera, la capacidad de cómputo puede aprovecharse al máximo y la **generación de resultados es mucho más rápida y eficiente**. Esto es debido a que, si paralelizamos los métodos, todos los ordenadores de la red pueden estar trabajando al mismo tiempo.



Computación en la nube. Fuente:

https://upload.wikimedia.org/wikipedia/commons/thumb/f/ff/Cloud_computing-es.svg/2000px-Cloud_computing-es.svg.png

- **4. Políticas de privacidad:** los datos de la mencionada «huella digital» que dejan los usuarios en Internet son una fuente fiable y extensa de información cuya utilización requiere de la autorización de los usuarios y de la web en concreto que haya obtenido esta información. Por tanto, no son datos que estén al alcance de todo el mundo sino que, cuando se necesite llevar a cabo un estudio estadístico, es necesario pedir los datos (o comprarlos) a la empresa en cuestión que posea la información que necesitamos.

Puede que incluso necesitemos **cruzar datos que posean varias empresas** a la hora de llevar a cabo nuestro análisis. Por tanto, aunque pueda parecer que hay una alta cantidad de información disponible, es necesario tener en cuenta que dicha información, **por lo general, es privativa y, por tanto, no todo el mundo puede acceder ni hacerse con dichos datos**. Generalmente, las empresas almacenarán los datos y tratan de monetizarlos y sacarles rendimiento como puedan.

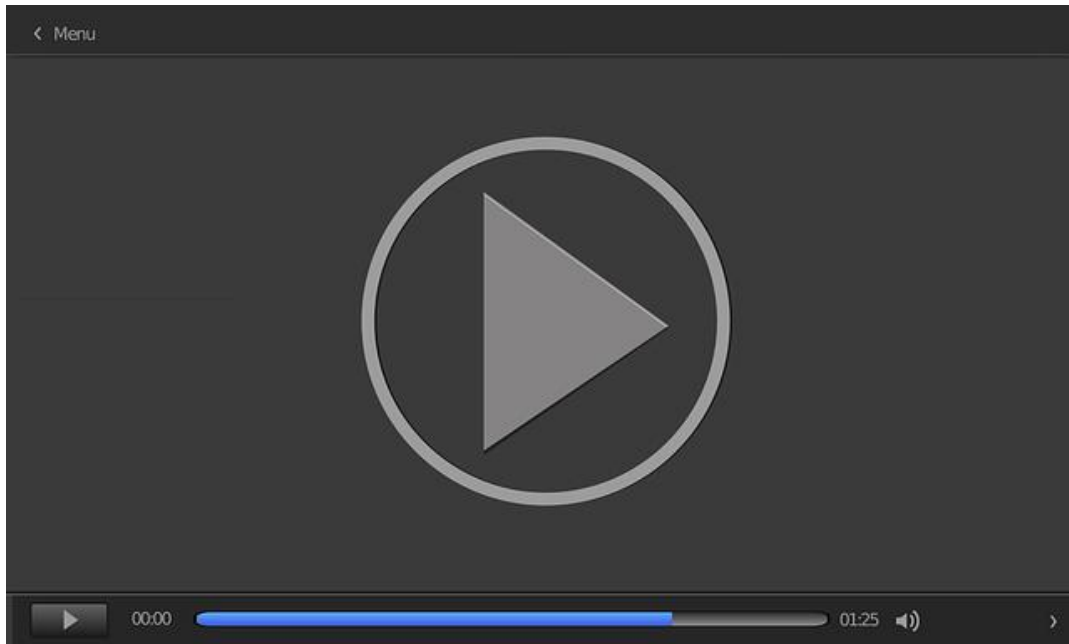


Privacidad. Fuente: https://pixabay.com/p-445153/?no_redirect

- **5. Recogida de datos sin previa especificación del problema**: en la estadística clásica, tal y como hemos visto, se diseña el estudio y luego se recoge la información. Por lo general, para ello, se utilizan encuestas o algún método de extracción de información que nos permita obtener la información necesaria. Como podemos observar, en la estadística clásica primero se diseña el problema y el modelo de datos y luego se extraen.

Directrices generales para la elaboración de un informe estadístico

En este vídeo vamos a establecer las directrices generales para la elaboración de un informe estadístico.



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=77b5a672-ce90-4d20-b4bd-acbc00c99a8e>

1.12. Referencias bibliográficas

Moore, D. S. (2006). *Introduction to the practice of statistics* (5th. ed.). New York: Freeman and Company.

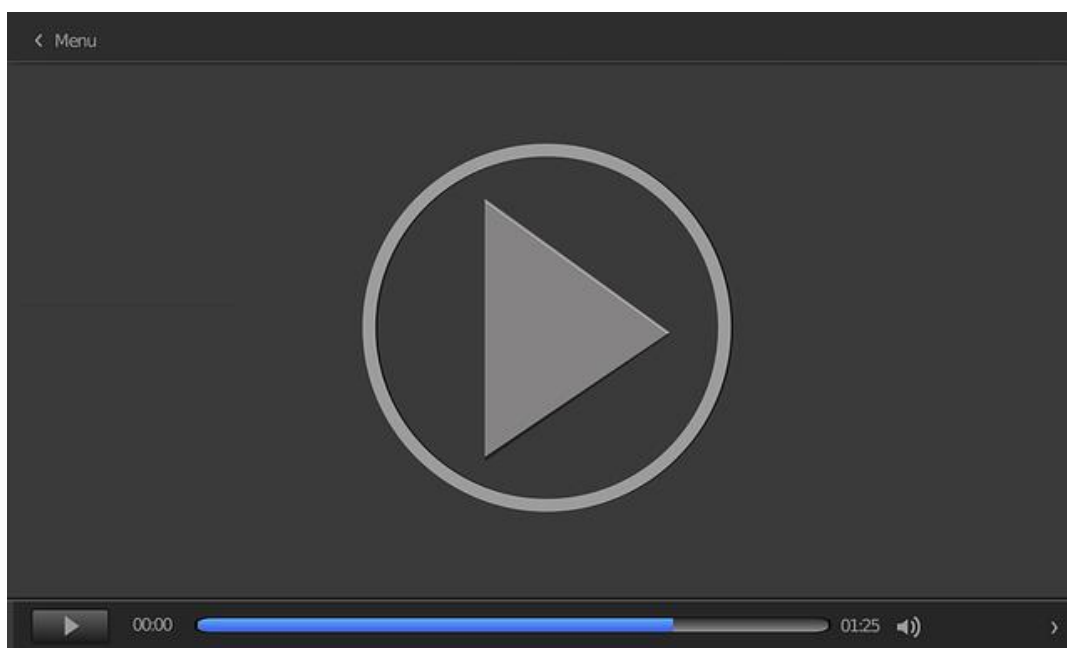
Ríos, F. (1998). *Bioestadística: Métodos y aplicaciones*. Málaga: Universidad de Málaga. Publicaciones. Versión electrónica:

<https://www.bioestadistica.uma.es/baron/apuntes/clase/apuntes/pdf/>

Triola, M. F. (2009). *Estadística* (10ª ed.). México D.F.: Pearson Educación.

Realizando un informe Analytics

En esta lección magistral aprenderemos a realizar un informe con Google Analytics.



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=e1b1a4b3-1803-41fe-b7c5-abdc00f2aa38>

Efecto Hawthorne

¿Has oído hablar del efecto Hawthorne? Te animo a que investigues por tu cuenta un poco de este efecto y sus orígenes en la industria americana de los años 50 del pasado siglo. También puedes aprovechar para reflexionar que implicaciones puede tener su existencia en los estudios estadísticos.

Accede al artículo a través de la siguiente dirección web:

http://es.wikipedia.org/wiki/Efecto_Hawthorne

Series temporales

Para profundizar y saber más sobre series temporales (lo cual excede en cierto modo el carácter introductorio y general de esta asignatura) te recomiendo al menos indagar sobre las componentes de una serie temporal, lo cual te servirá para desarrollar un «buen ojo» para juzgar y analizar las series temporales con las que trates de aquí en adelante.

Puedes consultar por ejemplo este breve resumen en Wikipedia.

Accede al artículo a través de la siguiente dirección web:

http://es.wikipedia.org/wiki/Serie_temporal#Componentes

Estadística antes que cálculo

El *speech* breve de Arthur Benjamin nos muestra de un modo elocuente la importancia que debería tener la estadística en nuestros currículos acorde con lo útil que resulta en nuestro día a día; todo ello en detrimento de las matemáticas clásicas y el cálculo los cuales ya no serían en general tan necesarios... (nota: puedes además poner los subtítulos en español o inglés para facilitar su seguimiento).

Accede al vídeo a través de la siguiente dirección web:

http://www.ted.com/talks/arthur_benjamin_s_formula_for_changing_math_education

Técnicas de representación de datos

Vídeo de TED para profundizar en técnicas de representación de datos aplicado a estudios demográficos realizado por Hans Rosling. Nota: puedes además poner los subtítulos en español o inglés para facilitar su seguimiento.

Accede al vídeo a través de la siguiente dirección web:

http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html

Bibliografía

Moore, D. S. (2006). Introduction to the practice of statistics (5th. ed.). New York: Freeman and Company.

Ríos, F. (1998). Bioestadística: Métodos y aplicaciones. Málaga: Universidad de Málaga. Publicaciones. Versión electrónica:

<https://www.bioestadistica.uma.es/baron/apuntes/clase/apuntes/pdf/>

Triola, M. F. (2009). Estadística (10ª ed.). México D.F.: Pearson Educación.

1. ¿De qué clase son cada una de las siguientes variables?

| | | | |
|--|---|---|----------|
| Tipo de madera (pino, cedro, roble) | 1 | A | Ordinal |
| Tipo de pintura (metálica, plástica, seca) | 2 | B | Continua |
| Grosor de la pintura (en milímetros) | 3 | C | Discreta |
| Grosor de la pintura (gruesa, normal, fina, ultrafina) | 4 | D | Nominal |
| Color de la pintura (rojo, violeta, azul, verde) | 5 | | |
| Meses del año (Enero, Febrero...) | 6 | | |
| Número de hijos | 7 | | |

2. La estadística ha sido definida como:

- A. El arte de manejar con rigor los números.
- B. La ciencia que analiza la información y la convierte en números.
- C. La ciencia del aprendizaje a partir de los datos.
- D. La ciencia que produce, analiza y extrae conclusiones de los datos.
- E. Las respuestas C y D son correctas.

3. Con la estadística manejamos:

- A. Información en forma de datos.
- B. Números contextualizados.
- C. Individuos de una población.
- D. Las respuestas A y B son correctas.

4. Hoy en día en España los censos...
 - A. Los llevaba a cabo el INE todos los años para temas muy importantes como la Encuesta de Población Activa, El Censo de Población y Viviendas, etc...
 - B. Ya no existen como tal.
 - C. Solo existe uno, el Censo de Población y Vivienda, que se lleva a cabo cada diez años.
 - D. Las respuestas A y B son correctas.

5. Decimos que una muestra es representativa cuando:
 - A. Ha sido obtenida mediante métodos aleatorios.
 - B. Es de un tamaño cercano al de la población de la que proviene.
 - C. Posee una diversidad muy parecida a la de la población.
 - D. Las respuestas A y C son correctas.

6. Decimos que los estudios experimentales:
 - A. Son superiores a las observaciones, pues permiten manipular a los individuos con la libertad que eso presupone.
 - B. Son junto con los observacionales los dos grandes tipos de estudios estadísticos.
 - C. Son más cuestionados que los observacionales pues interfieren en exceso.
 - D. Las respuestas B y C son correctas.

7. Un pictograma representa la información:
 - A. En el área del dibujo.
 - B. En la altura del dibujo.
 - C. En la anchura del dibujo.
 - D. Todo lo anterior es falso.

8. Referente a la infraestructura requerida para llevar a cabo análisis de datos en Big Data:

- A. Es necesario poseer un clúster propio.
- B. No hace falta usar infraestructuras de computación potentes.
- C. La computación en la nube no es una opción.
- D. Todo lo anterior es falso.

9. La aplicación de la estadística en Big Data:

- A. No plantea ningún problema.
- B. Se produce falta de información.
- C. La información es, a veces, demasiado compleja.
- D. Todo lo anterior es cierto.

10. La estadística:

- A. Es una disciplina clásica.
- B. Es una disciplina reciente.
- C. Engloba únicamente el apartado de extracción de información.
- D. A y C son ciertas.