

Maestría en Análisis y Visualización de Datos Masivos

Análisis e Interpretación de Datos

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

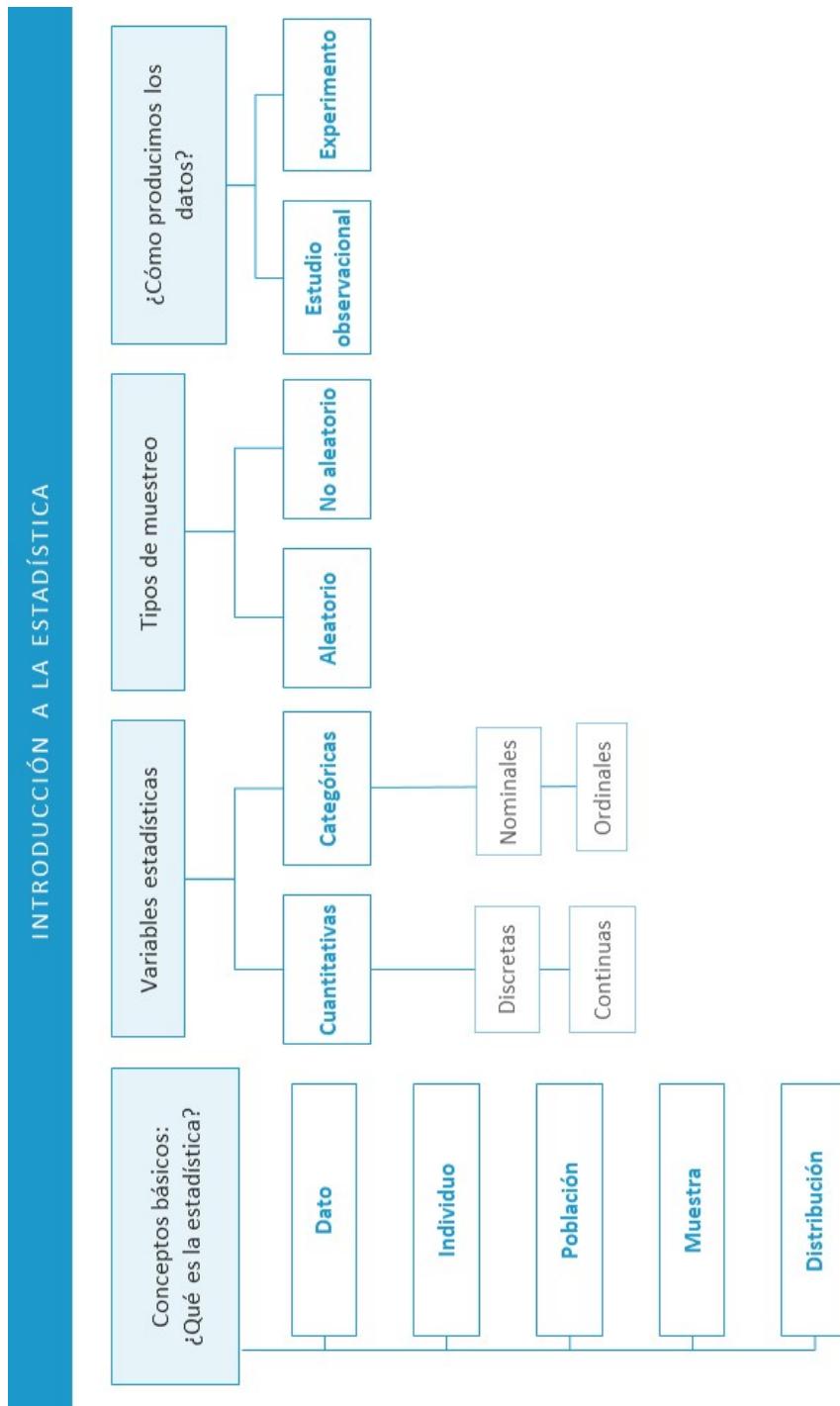
Series temporales

Estadística antes que cálculo

Técnicas de representación de datos

Bibliografía

Test



1.1. ¿Cómo estudiar este tema?

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

Ríus, 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).

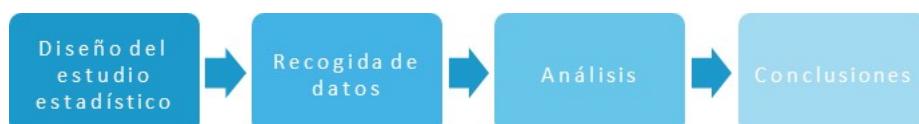
También será clave que practiques con los ejercicios que vienen al final del tema, los cuales están diseñados para que apuntes las ideas más importantes sobre tablas de frecuencias y gráficos estadísticos. Los dos esquemas que acompañan este tema te pueden ayudar a hacerte una buena idea de cómo está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

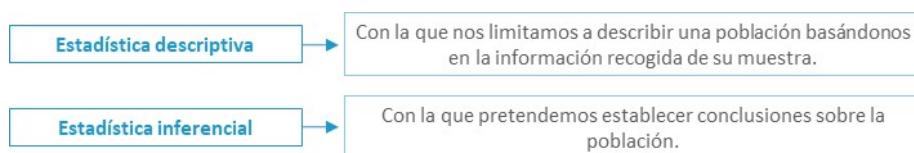


Ejemplo 1: De este modo en una misma empresa puede haber empleados y empleadas en diferentes puestos, encargándose uno de ellos del diseño del experimento para recoger los datos, otro de recogerlos, una tercera de analizarlos y un cuarto de exponerlos en una presentación delante del jefe de la empresa 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 compresión 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 extrapolalar 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 extrapolalar 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).

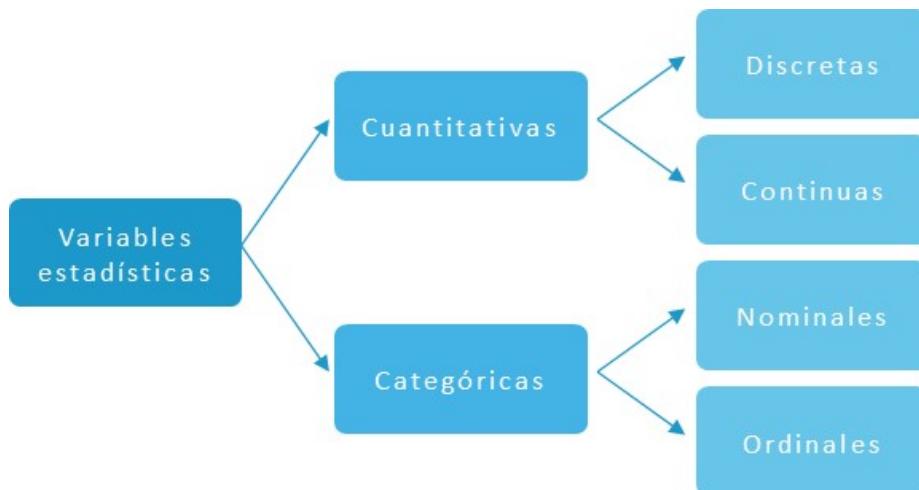
Ejemplo 2: En la Encuesta sobre Medios de Transporte que realizó el consorcio madrileño de transportes hace unos ocho años los encargados del estudio querían responder entre otras cuestiones a la siguiente pregunta concreta: «¿cuál es el uso que le están dando los madrileños al transporte público en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

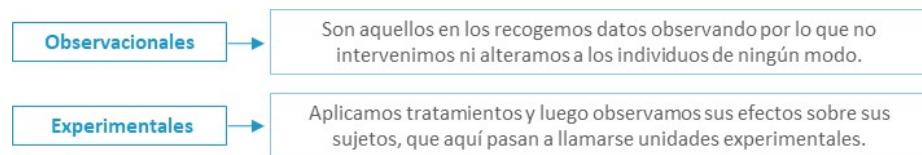
En la práctica una misma variable puede ser recodificada de diferentes modos, como por ejemplo la variable edad. En teoría se trata de una variable continua (la edad es el tiempo pasado desde el nacimiento, que es una magnitud continua), sin embargo, puede ser recogida en su dimensión puramente categórica 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 qué 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 suplantan 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

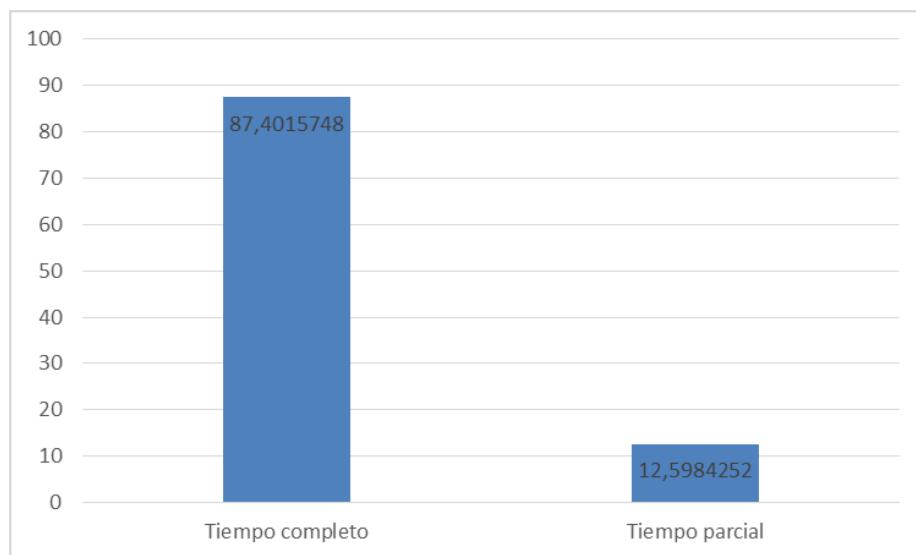
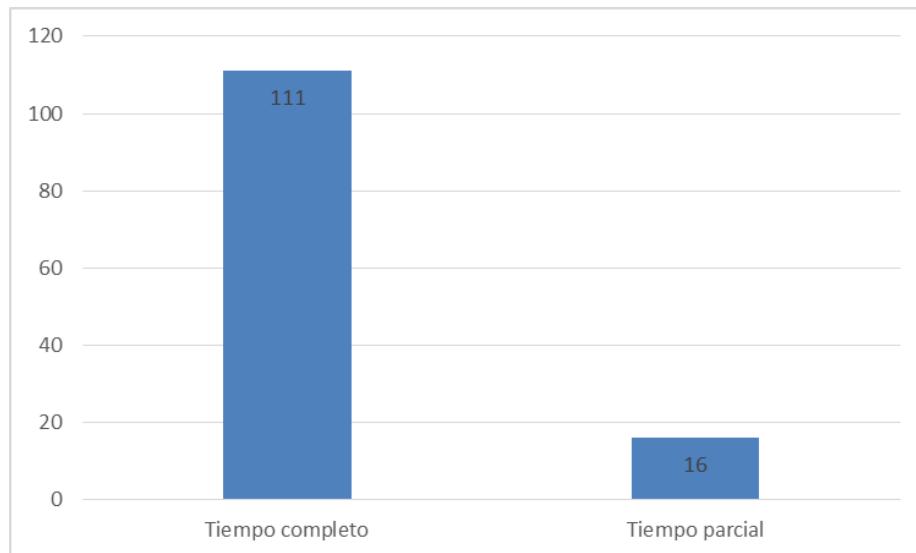
1.9. Gráficas básicas

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.

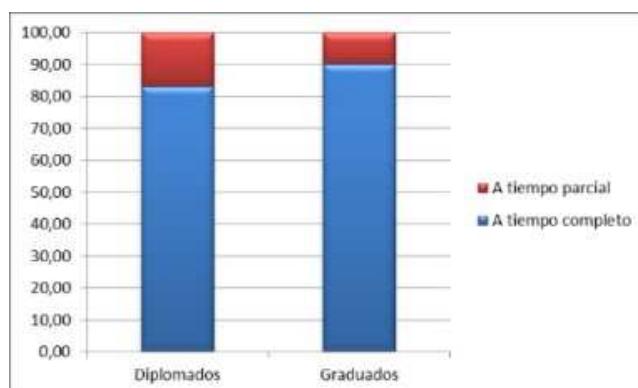
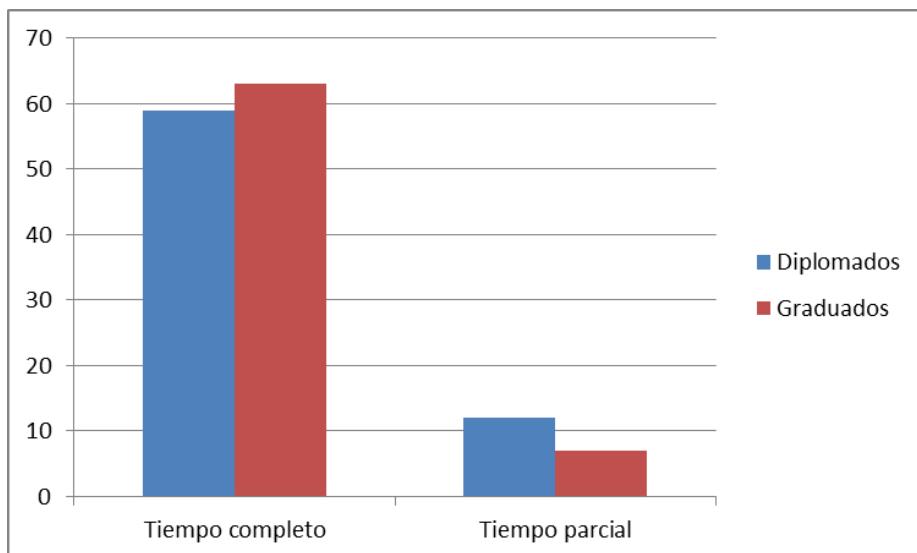
La pista esencial para saber que gráfico nos corresponde confeccionar es el tipo de variable que se pretende representar. El primer caso que se nos presenta es cuando tenemos variables de «tipo categórico» (en realidad no existe tal división pero a nivel práctico es útil manejarla), que pueden ser tanto 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:



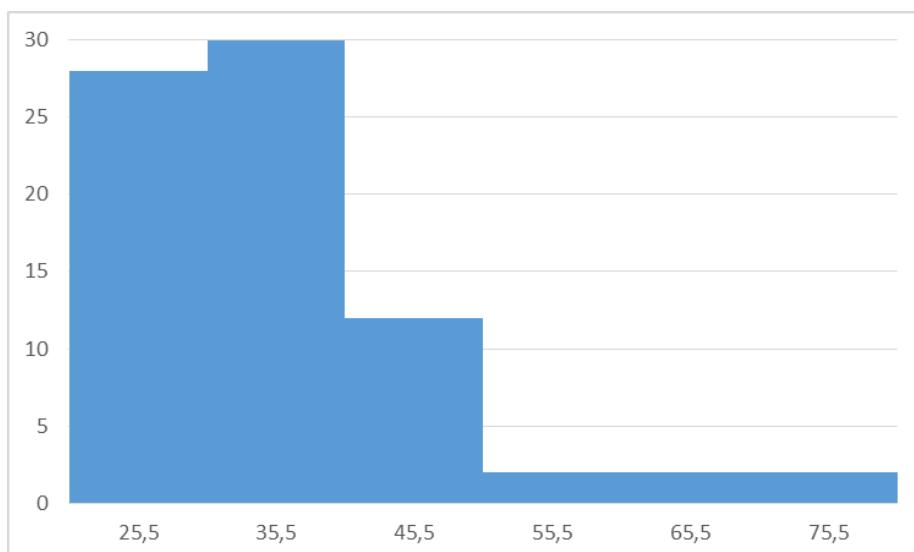
Uno de los **errores habituales** que se cometen en este tipo de gráficos es representar cada modalidad según su valor y dibujando cada elemento con esta escala. Esto no es correcto pues hay que considerar que las áreas de los dibujos tienen que ser proporcionales a las magnitudes que representan.

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 gráficas estadísticas como pueden ser el Excel y el SPSS.

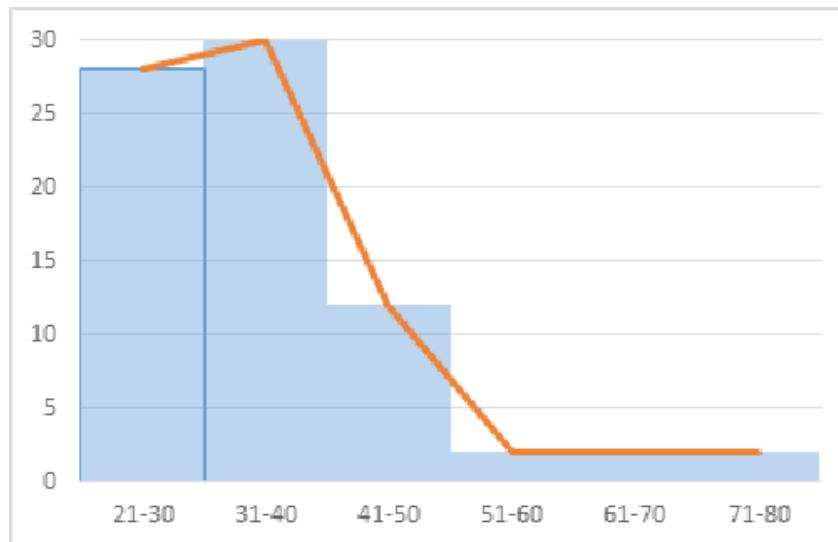
En el caso de las variables cuantitativas disponemos de otros gráficos básicos. El equivalente de algún modo al diagrama de barras en el caso cuantitativo continuo es el **histograma**. Este gráfico nos permite comunicar la continuidad a través de las **barras juntas**. Se suele emplear cuando disponemos 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.

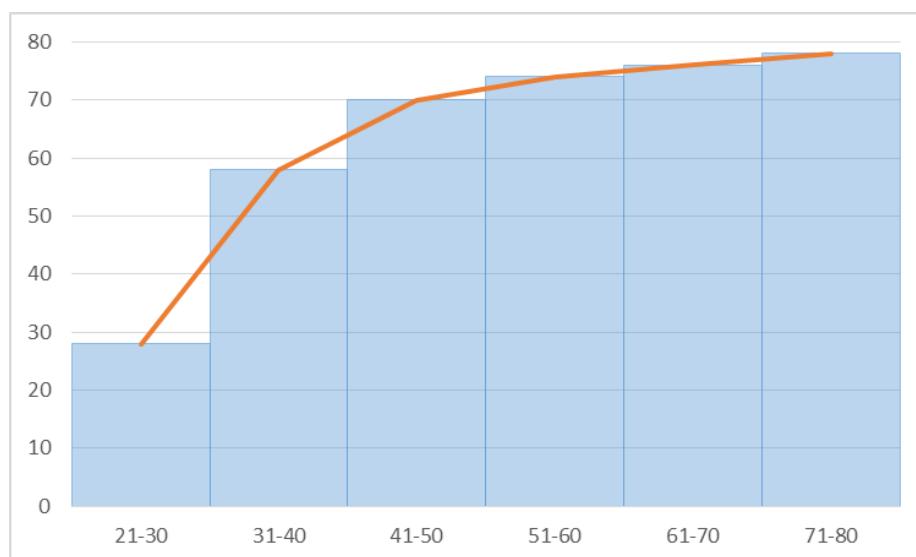


Un gráfico algo menos empleado que el histograma es el **polígono de frecuencias** que se obtiene al unir los puntos medios de las barras del histograma (muestro con el color de relleno rebajado el histograma asociado que no tendría por qué figurar acompañando al polígono de frecuencias).

Ejemplo 13:

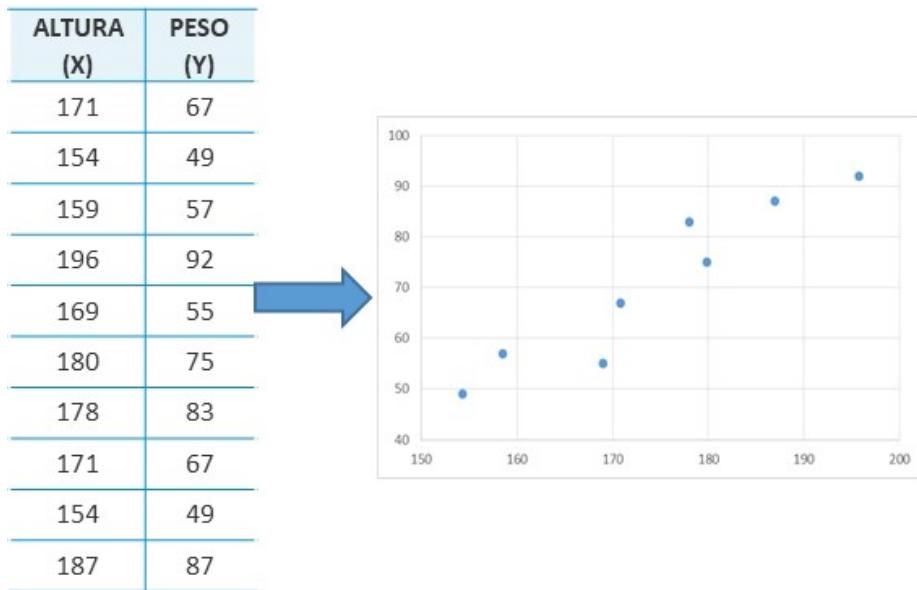


Este gráfico, al ser pura línea, acentúa las tendencias, por lo que viene bien para representar las frecuencias acumuladas, tal y como se ve en la siguiente versión:



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:

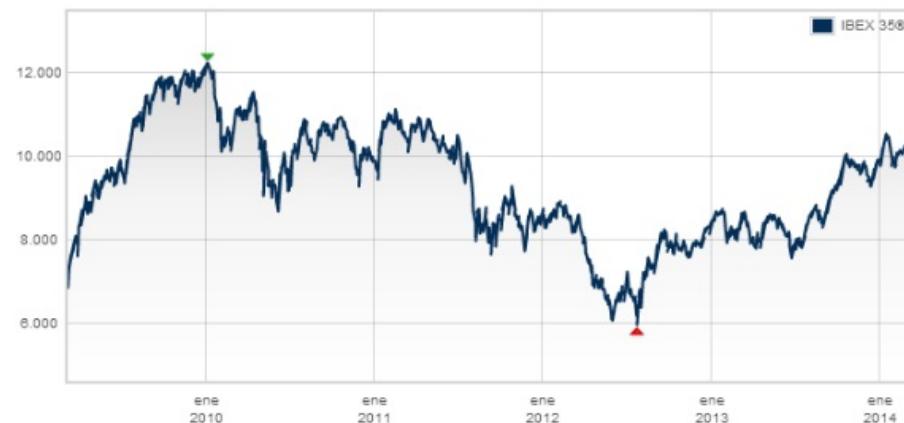


Cuando se dispone de frecuencia mayor que uno para algún par (x_i, y_j) lo que se hace es situarlos muy próximos entre sí indicando que todos esos puntos (n_{ij} puntos para ser más exactos) representan al mismo par.

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.

Ejemplo 15: Los índices económicos bursátiles como el IBEX 35 son un ejemplo muy común de gráficos de series temporales.



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

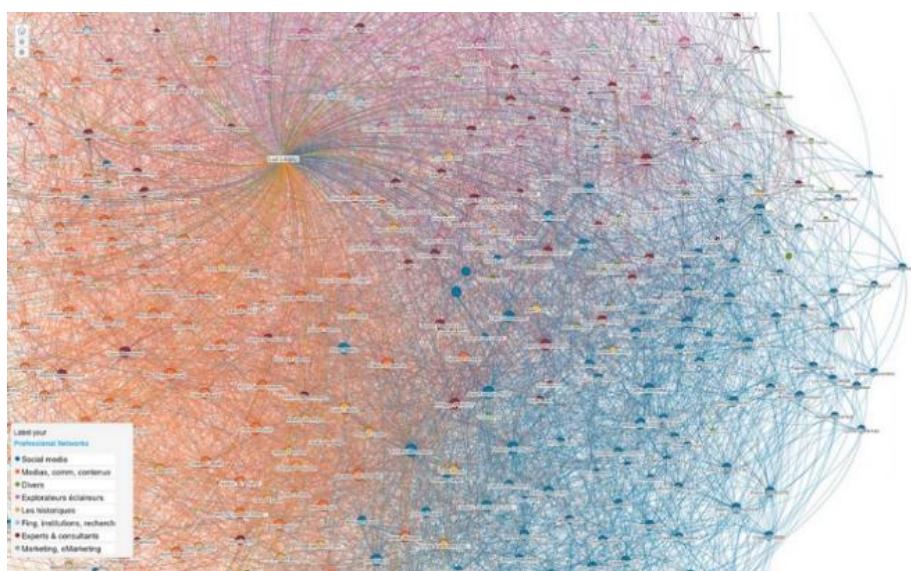
La **estadística es una disciplina clásica**. Su actual definición como «ciencia que recolecta y analiza los datos» proviene del **siglo XIX**. Está bastante claro que con la aparición de los computadores y, más recientemente, de Internet y el Big Data, **los entornos con los que actualmente trabaja 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**.



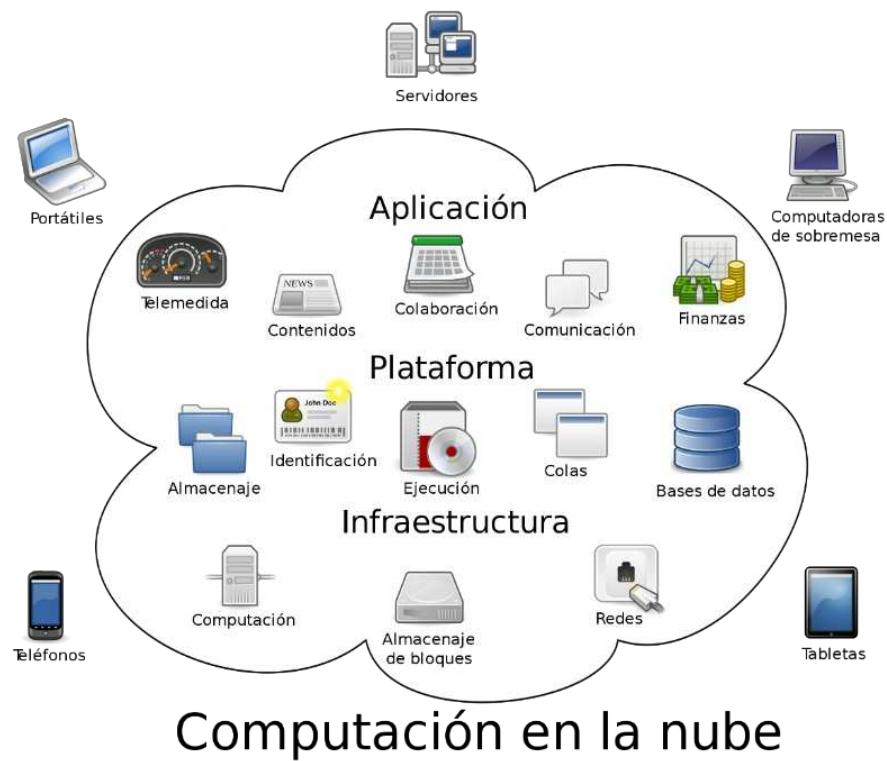
Red de usuarios. Fuente: https://c1.staticflickr.com/6/5217/5418037955_d361ba38ce_b.jpg

- **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.

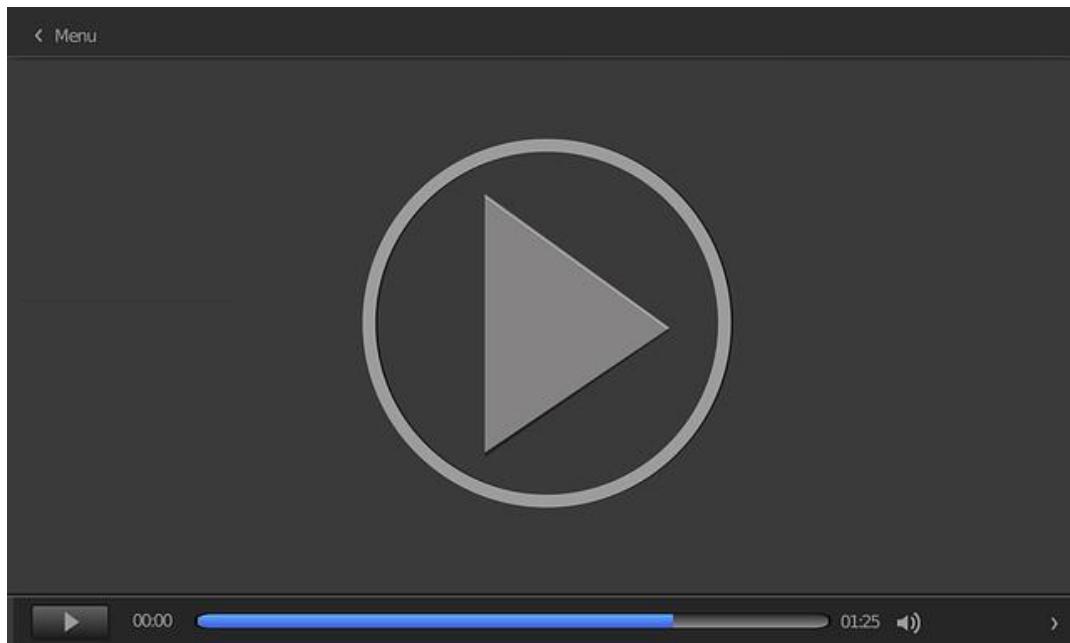


Privacidad. Fuente: https://pixabay.com/p-445153/?no_redirect

- ▶ **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.

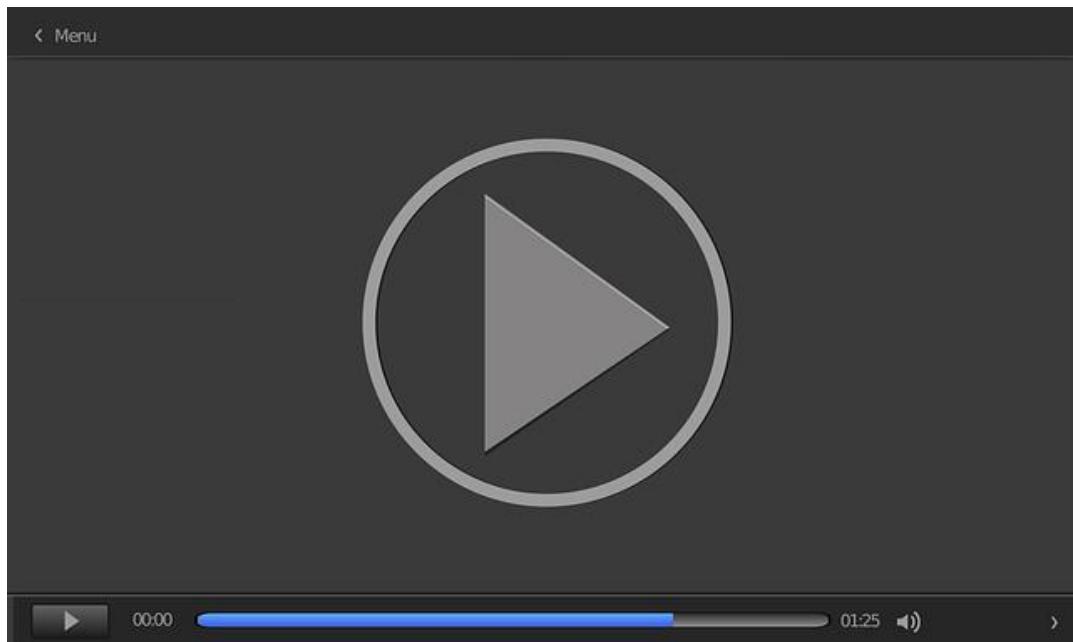
Ríus, F. (1998). *Bioestadística: Métodos y aplicaciones*. Málaga: Universidad de Málaga. Publicaciones. Versión electrónica:

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

Triola, M. F. (2009). *Estadística* (10^a 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 qué 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

En 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.

Ríus, 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 (10^a 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.

Análisis e Interpretación de Datos

Tema 2. Estadística computacional

Índice

[Esquema](#)

[Ideas clave](#)

[2.1. ¿Cómo estudiar este tema?](#)

[2.2. Principios básicos](#)

[2.3. Ámbitos de aplicación](#)

[2.4. Técnicas básicas de programación](#)

[2.5. Presentación del software «R»](#)

[A fondo](#)

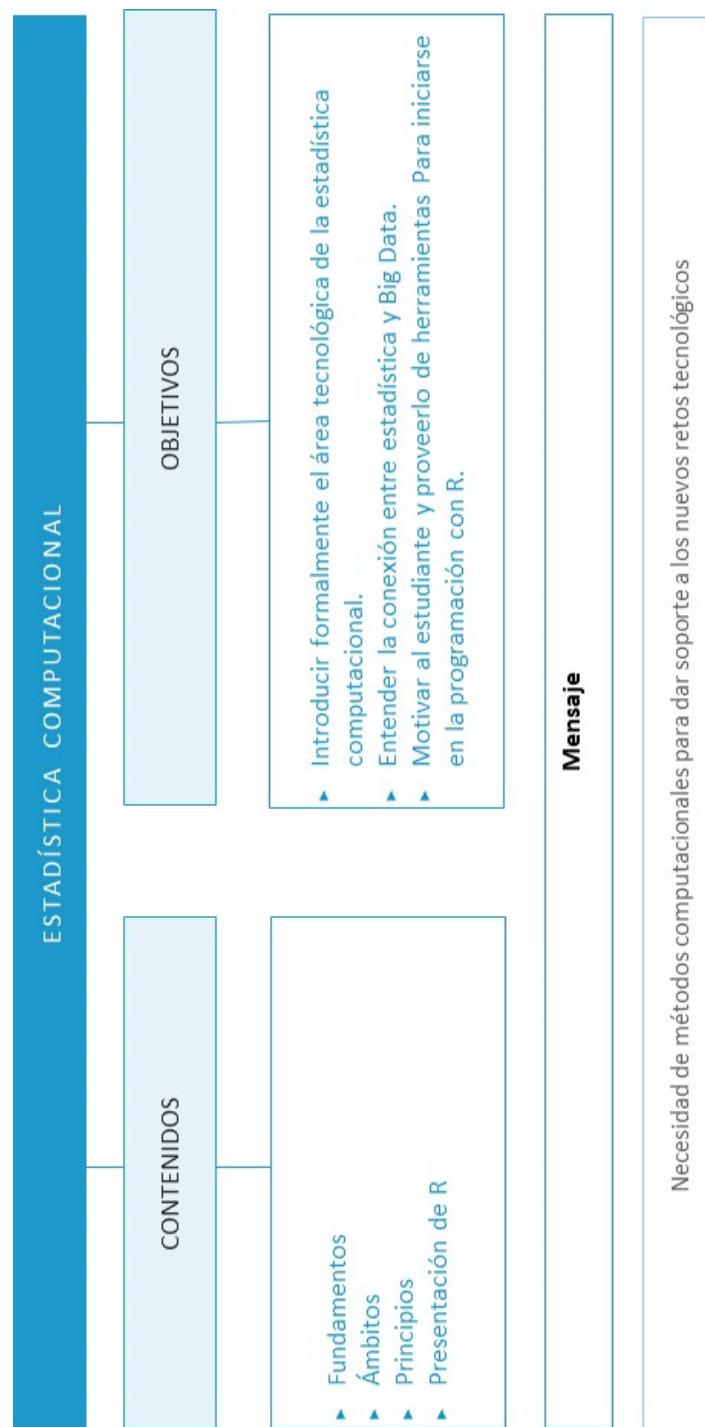
Sobre la relación entre Estadística, Informática y Big Data

Ejemplo de aplicación del análisis estadístico con métodos computacionales y usando software R a un problema real

[Bibliografía](#)

[Test](#)

Esquema



2.1. ¿Cómo estudiar este tema?

Para estudiar este tema lee las **Ideas clave** que encontrarás a continuación.

En este tema se introduce el área de estadística computacional. Se presenta una visión general del estado en esta ciencia emergente, tras la convergencia de áreas ya consolidadas como las matemáticas, estadística e informática. Tras presentar los **principios básicos** que definen la estadística computacional, se comparte una panorámica sobre las **áreas donde se está aplicando** esta nueva herramienta técnica, incluso podría decirse, áreas técnicas nuevas que se consolidan apoyadas en estrategias de lo que se conoce como **estadística computacional**. Algunas áreas son de reciente incorporación incluso al argot técnico, como inteligencia artificial, minería de datos y *Machine Learning*.

A continuación, se introduce el **lenguaje de programación estadística R** y se presentan sus elementos básicos para un uso a nivel principiante. Estos elementos no pretenden de ninguna manera ser un manual de uso para iniciarse, siquiera, en el *software*, algo imposible de abarcar en solo un apartado de este tema. Lo que sí se pretende es que sea un aliciente para motivarlos a iniciar los desarrollos computacionales con R y proveeros con las pautas sobre cómo enfrentar, por un lado, la formulación de un problema estadístico en términos de un algoritmo computacional y, por otro, ser capaz de escribir este algoritmo en el lenguaje R.

Los objetivos que intentamos alcanzar con este tema son:

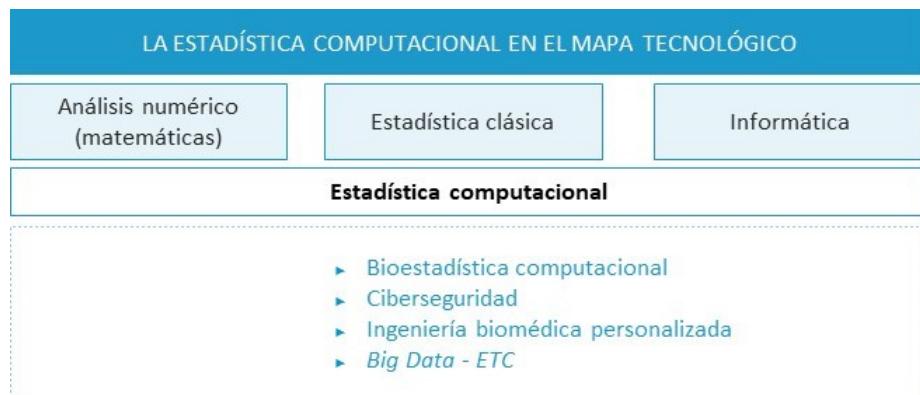
- ▶ Introducir formalmente el área tecnológica de estadística computacional.
- ▶ Entender y ser capaz de pensar en términos de la aplicación de los métodos de la estadística computacional a problemas técnicos que podemos tener hoy en día, haciendo énfasis en los problemas derivados del tratamiento de entornos *Big Data*.
- ▶ Presentar el *software* estadístico R.
- ▶ Motivar y proveer de herramientas para la programación con R.

La idea central que queremos recalcar es la necesidad que tenemos como profesionales de ser **capaces de controlar el uso de métodos computacionales para resolver problemas estadísticos** si pretendemos superar retos que en estos momentos conciernen a los científicos de datos.

Este conocimiento, la estadística computacional, puede ser entendida como el puente entre ciencias clásicas (matemáticas, estadística) y campos científicos aun en sus inicios (inteligencia artificial, bioinformática, ciberseguridad, medicina personalizada, *machine learning*). Los avances que podamos lograr en estas últimas áreas mencionadas serán dependientes en gran medida del rigor y alcance con que podamos abordar la estadística computacional.

2.2. Principios básicos

La estadística computacional se consolida como Ciencia que se sustenta en la implementación computacional de conceptos, reglas y fórmulas, derivadas de análisis matemáticos y estadísticos utilizados para describir la solución a un problema (ver Figura 1).



De arriba hacia abajo, vemos las bases tecnológicas que fundamentan los desarrollos en estadística computacional y conducen a las aplicaciones tangibles hoy día. Campos como el *Big Data* podrían dar soporte a nuevos retos tecnológicos incluso aun por definir.

Por otro parte, **análisis estadísticos** se refiere a la ciencia de recopilar, discutir, visualizar y analizar datos. Datos que pueden constituir una muestra finita o una porción de un espacio muestral infinito. Esta recopilación, discusión, visualización y análisis de datos se hace, esencial y complementariamente, basándose en métodos matemáticos.

Hoy en día, nos encontramos ante la necesidad/reto de analizar problemas que derivan del comportamiento de sistemas «grandes» (con muchos datos) incluso muchas interacciones ocultas entre estos datos, que generan información (nuevos datos) y que, en esencia, es lo que subyace al problema central que motiva el desarrollo de la estadística computacional: la necesidad de **entender sistemas Big Data** en un entorno intrínsecamente complejo.

Por tanto, los principios fundacionales de la estadística computacional subyacen en el conocimiento y control de **tres áreas técnicas**:

- ▶ **Programación:** nos centraremos en desarrollar, programar, aplicaciones en el software estadístico R. Con este contenido se pretenden dar pautas para el trabajo en la solución e implementación en R (es decir, implementación a nivel de principiante) de problemas matemáticos, estadísticos y de programación.

Presentaremos ejemplos prácticos que sirvan de guía en esta iniciación e incluso, y muy importante, atendiendo a lo que comentaremos más adelante como una de las reglas básicas de programación, presentaremos **problemas y soluciones numéricas sencillas** que deben servir de plantilla para la solución de problemas más complejos. Esta reutilización de código, concepto propio de la programación numérica y que sin definición estricta avanzamos, debe permitir al analista de datos poder modificar y ajustar ejemplos a sus necesidades derivadas del análisis de casos reales o modelos más complejos.

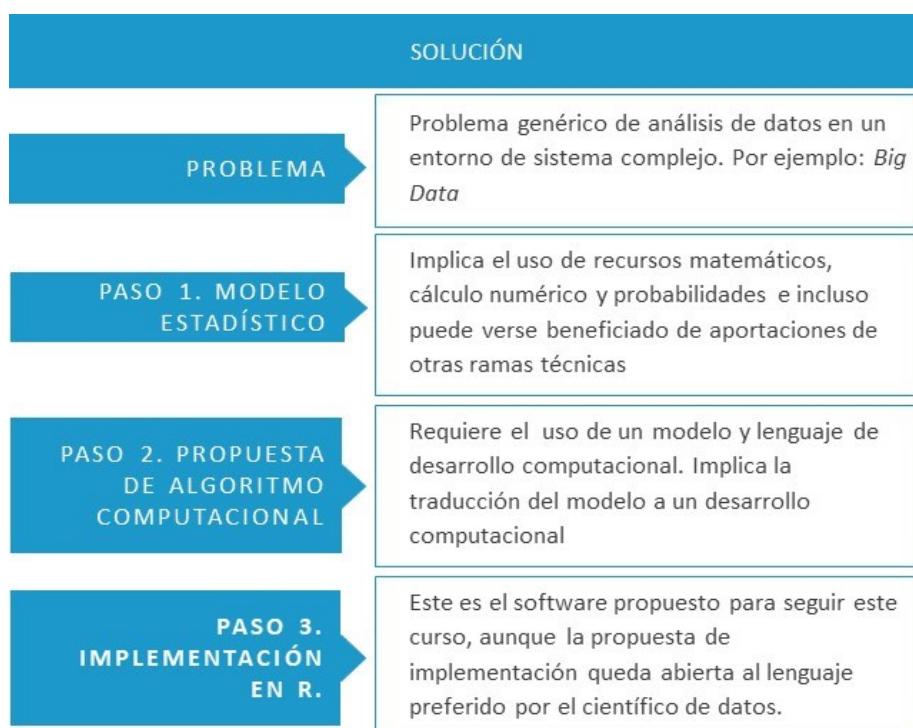
- ▶ **Análisis numérico:** durante la asignatura, veremos y motivaremos al alumno a enunciar problemas que conducen a soluciones en el marco de la estadística computacional. Estos problemas, pueden derivar de situaciones ya anunciadas dentro del entorno *Big Data* y los sistemas complejos. Estas situaciones, a su vez, constituyen la última etapa en el mapeo al análisis numérico de situaciones reales en áreas incluso emergentes hoy en día, como pueden ser la biología, la medicina, ciberseguridad, etc.

Frecuentemente, estos problemas no tienen solución analítica o el resultado exacto, —debido al elevado número de datos, requiere mucho tiempo para su evaluación—. En este sentido, se hace imprescindible el uso técnicas numéricas para aproximar el resultado. Distinción aparte merece el hecho de que los softwares estadísticos, por ejemplo, R, pueden ayudar en el cómputo de datos, en su arreglo y en su proceso de visualización. Aspectos que también trabajaremos en este curso.

En la siguiente figura se puede ver la propuesta de solución a un problema en el ámbito de análisis de datos para sistemas complejos que implica el uso de estadística computacional.

La solución, de manera global, se describe con **tres pasos fundamentales**:

- ▶ Modelado del problema.
- ▶ *Propuesta de solución numérica.*
- ▶ *Implementación numérica.*



- ▶ **Estadística clásica:** en este contexto presentaremos los métodos estadísticos básicos utilizados para describir y analizar **datos univariados**, temas propios de estadística descriptiva e inferencial de datos univariados, y se irán incorporando a los temas en la medida que se desarrolle el curso.

También retomaremos análisis propios de combinatoria y teoría de probabilidades, básicos para progresar en tratamientos estadísticos. Para análisis en sistemas de muchas variables utilizaremos **métodos de regresión lineal y multivariable**. Los modelos de regresión son extremadamente importantes. Por ejemplo, la regresión lineal, herramienta simple pero poderosa para investigar las dependencias lineales y basada en supuestos estrictos de distribución a los modelos de regresión no paramétricos imprescindibles en el análisis de entornos *Big Data*.

Con estos elementos podemos decir que vamos a trabajar con la conjunción de fuerzas que ofrecen los elementos de análisis matemático, estadística y programación, encaminados a resolver problemas de una alta dimensionalidad (es decir, número de datos y complejidad), o sea, estamos en la línea correcta para trabajar en el área de estadística computacional.

2.3. Ámbitos de aplicación

Áreas de aplicación para la estadística computacional... Parodiando mensajes propios del mundo de la programación, podríamos añadir: *running*.

¿Qué queremos decir con esto?

Estamos ante un **campo nuevo**, con aplicaciones con tendencia a resolver y simplificar problemas numéricos; por ejemplo, hacer funciones similares a calculadoras programables. Pero la estadística computacional, como hemos enunciado, es más que esto, lleva a plantear soluciones a problemas hasta ahora irresolubles y, por tanto, a generar nuevos problemas.

¿En qué áreas estamos viendo este impacto hoy en día? Véase la Figura 1 que resume el actual contexto de trabajo y entremos en más detalle en los siguientes párrafos.

- ▶ **Estadística computacional en Biología, bioestadística computacional:** es un área que une ramas ya consolidadas como bioinformática, genómica y biotecnología. Todas estas áreas tomadas como ejemplo muestran hoy en día un nicho de actividad importante para los científicos de datos. Son áreas que lucen por los prometedores avances que usan (y generan) grandes volúmenes y diversos tipos de datos. Por tanto, exigen el desarrollo de metodologías y herramientas eficientes de estadística computacional integradas con conocimiento biológico y algoritmos computacionales, o sea bioestadística computacional.
- ▶ **Big Data como base en el desarrollo de la medicina**, es decir, informática o ingeniería de datos para biomedicina, medicina personalizada: estas son algunas de las etiquetas que hoy en día se utilizan para sintetizar áreas emergentes de desarrollo. Tratan fundamentalmente del potencial que subyace en la utilización de «muchos datos» para lograr avances en estudios médicos. Por ejemplo, podemos citar los estudios derivados de análisis de genoma humano para la predicción de

comportamientos futuros de los individuos. Esto implica análisis de muestras de genotipos, comparación con datos de otros individuos (cuanto más, mejor). Todo esto es una aplicación más del análisis de datos e incluso lo que se conoce como *machine learning*, aplicado a problemas médicos.

- ▶ **Estadística computacional para facilitar trabajos de Ciberseguridad**. En estos momentos el tráfico de datos en el entorno virtual es una realidad y necesidad creciente. Este comportamiento genera acciones derivadas del uso que puedan tener estos datos, por un lado, la buena intención de sacar provecho de estos para fines legítimos u otras intenciones más oscuras encaminadas a utilizar estos datos para generar problemas de ámbito técnico, *malware*, o incluso para desestabilizar determinados entornos, *fakenews*. En cualquier caso, el objetivo es claro, controlar el tráfico de datos y poder proveerlo de seguridad. Para esto es necesario ser capaces de registrar, codificar, las transferencias que ocurren en red. Estos son procesos matemáticos que incluyen análisis dinámicos, contenidos avanzados para este curso, pero que en general ahora necesitamos entender como series de números que evolucionan en el tiempo. La posibilidad de desarrollar métodos seguros en este ámbito requiere de la participación de los estadísticos computacionales además de otros perfiles técnicos.

Con el panorama antes expuesto queremos dejar patente la **necesidad de trabajar en el ámbito de estadística computacional**, basándonos en el hecho de que es un campo reciente y, por tanto, abierto a **aportaciones y oportunidades** desde el punto de vista de contribuciones de desarrollo, además de que en estos momentos ya se ha consolidado como solución a muchos problemas existentes.

2.4. Técnicas básicas de programación

A continuación, queremos sintetizar lo que consideramos serán buenas prácticas en el ámbito de la programación de soluciones a problemas estadísticos.

Conocidas como «**buenas prácticas**», son extensibles a cualquier lenguaje de programación y en particular nos centraremos en **ponerlas en práctica en los códigos que se desarrollen con R**:

- ▶ **Prestar atención a la sintaxis de operaciones**, que se conoce como «expresividad» del lenguaje, lo que implica usar variables, denominaciones, etiquetas, cuadros de texto que ayuden a seguir el código sin necesidad de ser especialista en desarrollo informático.

Por ejemplo, si una variable guarda valores asociados a las notas de unos alumnos, lo más oportuno es denominarla «notas» o análogo. En caso de que los códigos sean extensos, o incluso como praxis general, se recomienda introducir al principio de los códigos cuadros de texto indicando qué significa cada variable utilizada.

- ▶ **Seccionar el programa**, esto se hace con la intención de facilitar el proceso de validación de código. Por ejemplo, si es un cálculo que como *input* utiliza las funciones f_1, f_2, f_3 , debemos dejar indicado en el código dónde se calculan cada una de estas funciones para poder estructurar, así, el proceso de verificación de código.
- ▶ **Facilitar lo que se conoce como «modularidad»**, similar a lo anterior, pero con más implicaciones. Se trata de dividir el programa tanto como sea posible en «módulos». Con estos módulos se gana en el proceso de validación de código, pero también se potencia un aspecto relevante en el ámbito de desarrollo computacional, la posibilidad de «reutilizar» código.

Un ejemplo es el hecho de que necesitemos generar alguna función como parte de la solución a nuestro problema. Es recomendable, en este caso, salvarla como función independiente, incluso si se trata de un diseño propio, y poder invocarla en futuras realizaciones.

Con esto sentamos unas bases para poder comenzar a programar a **nivel principiante**. Bases abiertas a su desarrollo y que, de forma general, constituyen una guía para perfilar buenas prácticas en la elaboración de *software*, particularmente *software* estadístico en el contexto que trabajamos.

2.5. Presentación del software «R»

R es un programa muy útil para el análisis, representación y visualización de datos. Es abierto (*open source*), gratuito y se puede descargar de Internet.

Accede a la página de descarga a través del aula virtual o desde la siguiente dirección:

<https://www.r-project.org/>

Aquí se pueden encontrar los ejecutables para los distintos sistemas operativos, pero, además, como corresponde a su definición también podemos tener acceso al código fuente (esto permite algunos estudios avanzados de implementación estadística prácticamente al alcance de cualquier usuario). Para este curso recomendamos inicialmente bajar los ejecutables para el sistema operativo que prefiera el alumno.

A modo de resumen presentamos algunas de las características de R que lo hacen ser el *software* de elección para conducir este curso y que, de hecho, justifican la tendencia al alza en su uso en el gremio de los científicos de datos.

- ▶ **Contiene implementaciones para el cálculo de «todas» las herramientas estadísticas**. Aquellas que no se encuentran, dada su especificidad o novedad, suelen ser añadidas por usuarios y agregadas como librerías de libre acceso.
- ▶ **Permite el acceso a otros programas de cálculo matemático**. Acceso entendido como compartición de librerías. Algunos programas que se pueden hibridar con R son: **C,C++, Fortran**.
- ▶ Es una potente herramienta de cálculo numérico que se basa en potenciar la **programación orientada a objetos** que, a su vez, le concede alta eficiencia para trabajar con distintos formatos de lectura/importación de datos externos.

Con estos elementos recomendamos proceder a la instalación de R y empezar a utilizarlo. Daremos algunas pautas que pueden servir de guía, pero como norma general recomendamos consultar manuales más extensos o ayudas *online* para poder gestionar los problemas de implementación que surjan durante el trabajo.

Estructuras básicas: bases de datos, operadores, funciones y librerías

El objetivo fundamental de un *software* diseñado para estadísticas es ser capaz de leer datos, manipularlos, operar con ellos y guardarlo adecuadamente. Todas estas funcionalidades las desarrolla R. En general R es muy versátil, en cuanto a que puede trabajar con distintos tipos de datos y cambiar de un tipo a otro según convenga.

Para la **lectura de datos**, veremos **cómo podemos leer datos de ficheros externos** y, por comodidad en este ejemplo, **guardarlos en una tabla**.

Primero, crearemos un fichero .txt (es el tipo más genérico de extensión que nos «conecta» con cualquier otra extensión de datos que trabajemos), que, por ejemplo, represente una lista finita de «edad» de un grupo de personas: 10,20,30,10,40,80 y encabezaremos una columna con el nombre: **edad**. Guardamos el fichero como edad.txt en nuestro directorio de trabajo, por ejemplo, PRÁCTICAS.

A continuación, abrimos la consola de R y vamos al directorio de trabajo PRACTICAS. Para esto usamos el comando:

```
>setwd("c:/PRACTICAS/")
```

Desde aquí ya podemos cargar el fichero edad.txt y guardarlo en una variable:

```
>TablaEdad<- read.table("edad.txt",dec=". ",header=T)
```

Otra manera muy recomendada de guardar datos en R es mediante la **creación de vectores**.

Por ejemplo, en el caso anterior, en lugar de importar una lista de edades pensemos en crear una lista que contga las edades de un grupo análogo al anterior. Creamos en este caso un vector «edad» para guardar esta información. En línea con el ejemplo anterior este vector se crearía así:

```
>edad<-c(10,20,30,10,40,80)
```

Con estas dos estructuras básicas de almacenamiento de datos podemos comenzar a operar en R.

En cuanto a operaciones, R puede funcionar como una **calculadora habitual**. Por ejemplo:

```
>4+5  
[1] 9
```

Y así, en esta línea, las operaciones algebraicas ya conocidas.

Para calcular algunas **funciones tipo exponenciales**, se pueden introducir directamente por pantalla, pues son de las que el programa carga por defecto. En línea con el desarrollo del código, se han implementado funciones más complicadas, incluso algunos estadísticos específicos para determinados contextos.

Todo esto lleva a que sea necesario, antes de escribir el código, investigar las librerías que contienen las funciones que necesitamos y dar al programa la orden de cargarlas antes de interpretar nuestro código.

Veamos un ejemplo:

En estadística suele ser usual el proceso de codificación de variables durante una recogida de datos. Por ejemplo, preguntar por «práctica de deporte» y tener respuesta «sí» o «no». Para trabajar, puede ser útil transformar estas variables a numéricas, ejemplo «1» y «0» respectivamente. Este proceso se designa con el término anglosajón *reencoding*.

¿Por qué es útil pensar en un término anglosajón? Porque con los programas que solemos trabajar hoy en día, los diseños de código suelen corresponder a estructuras anglosajonas y porque la manera de proceder para el investigador ante la duda de cómo introducir una funcionalidad, la opción inicial suele ser la búsqueda de bibliografía. En este caso, la búsqueda o respuesta a la pregunta: ¿existe alguna función que haga lo que necesito? Si buscamos sobre «R software recoding», inmediatamente nos dirigirá al uso de la función **recode** con ejemplos como el que mostramos a continuación.

Supongamos que declaramos una lista de datos sobre práctica deportiva:

```
>deportistas <- c("si","no","si","si")
```

Queremos recodificar estas variables como se indica en el párrafo anterior. Entonces, debemos cargar la librería «car» y utilizar la función «recode». Esta información se obtiene investigando en la documentación del programa.

```
>library(car)
>recode(deportistas,"'si'=1;else=0")
```

Esta es la **rutina de trabajo**: fijar el problema que necesitamos resolver, buscar qué hay ya implementado en el *software* que nos puede ayudar a escribir el algoritmo de solución (generalmente esto se refiera buscar funciones implementadas que estén implícitas en la solución al problema). Finalmente, como estrategia de programación para principiantes se recomienda acceder a ejemplos y mapearlos a el caso particular que se esté tratando.

Representación de datos: variables categóricas y variables numéricas

El análisis estadístico necesita manipular datos. Datos que, por su origen, pueden tener distinta forma o lo que en R denominamos «clase». Trabajaremos con dos clases de estadísticos: **categóricos o nominales** y **numéricos o cuantitativos**.

Mostraremos a modo de ejemplo cómo obtener **parámetros descriptivos** para cada uno de estos tipos de variables.

Supongamos que tenemos una variable categórica que guarda información sobre alumnos «aprobados» y «suspenso»:

```
>notaUnir<-c(rep("aprobados",20),rep("suspenso",10))
```

Esta variable «notaUnir» guarda una lista con la información de 20 «aprobados» y 10 «suspenso». En la definición, y para no repetir el valor categórico (20 y 10 veces respectivamente), utilizamos de forma recurrente la función propia de R **rep** (del término anglosajón *reproduce*). Utilizando otra función, **summary**, podemos obtener una descripción de esta variable:

```
>summary(notaUnir)
Length   Class  Mode
      30   character carácter
```

De manera análoga, con la función **summary**, podemos obtener los principales parámetros descriptivos para una variable numérica. Supongamos ahora que este es el caso de la nueva variable notaUnir,

```
> nota<-c(rep(8,20),rep(3,10))
> summary(nota)
Min.  1st Qu. Median    Mean 3rd Qu.    Max.
3.000   3.000   8.000   6.333   8.000   8.000
```

Notese que esta función provee de algunos estadísticos, pero no de todos, como es de esperar. El cálculo de aquellos que no se devuelven debe hacerse con funciones específicas que, como ya se comentó, hay que investigar para llegar a su forma a partir de la documentación del programa.

En casos más específicos o de interés técnico en los que puedan surgir situaciones en las que no tengamos implementadas las funciones necesarias, se desarrollará el algoritmo para «definir una nueva función», en caso de que necesitemos utilizarla de manera recurrente o dividir el problema, de forma tal que nos lleve a trabajar sobre funciones conocidas.

Tabulación de variables

En apartados anteriores hemos trabajado ya con tablas simples de una columna. A continuación, discutiremos aspectos más generales relacionados con las potencialidades que ofrece R en este contexto para el almacenamiento de datos.

Por medio de R podemos manipular **diferentes formatos de ficheros de bases de datos**, por ejemplo: .csv, .txt, ,dat, .xls, .sav .

Para **importar estos ficheros** se usarán comandos tipo `read` y `read.table` y comandos `write` y `write.table` **para guardarlos**, una vez que hayan sido modificados. Con las tablas se puede operar como con los vectores a nivel algebraico.

Existen **librerías especializadas** para la edición de tablas, colocación de etiquetas, títulos y otros aspectos estilísticos. A nivel de programación es necesario invocar esta librería al comenzar al escribir el *script* para poder usar las funciones que nos interesan y se comentan en este punto.

Gráficas básicas

Existen diversas maneras de representar datos gráficamente. En este curso nos centraremos en los **histogramas, diagramas de sectores y barras acumuladas**. A la hora de hacer un gráfico es imprescindible, desde el primer momento, poder acceder a las funciones para su correcto etiquetado, denominar ejes, regiones gráficas o lo que se necesite en este sentido. Un descuido en estos aspectos puede llevar a la incomprendición del gráfico.

Te recomendamos escribir un *script* para cada uno de los gráficos planteados, suponiendo que seguimos trabajando sobre la variable «nota» que contiene una lista resumida con información sobre resultados académicos:

```
>barplot(nota)
>pie(nota)
>hist(nota)
```

Al introducir estas instrucciones aparece un diagrama de barras, un diagrama de sectores y un histograma. Sin embargo, hay imprecisiones importantes en cuanto al estilo e incluso contenidos. Para esto se deben introducir parámetros asociados al etiquetado y estructura de los datos para cada caso. A continuación, en los temas siguientes se mostrarán ejemplos para trabajar con las estructuras introducidas en este tema.

Perspectivas

Como hemos intentado mostrar, R se puede usar en cualquier problema estadístico y para trabajar con cualquier tipología de datos.

En este curso nos limitamos a mostrar el uso de R para la realización de los análisis estadísticos que alcanzamos a estudiar a nivel principiante. Esto incluye análisis descriptivo de datos y algunos mínimos sobre visualización. A nivel superior, se puede indicar que R es útil, además, para realizar operaciones entre bases de datos de manera análoga a la forma en que trabajan algunos softwares funcionales que se ocupan del tema de la lógica relacional.

Con R podemos incluso desarrollar modelos de aprendizaje automático y modelado matemático avanzado, altamente demandados en temas de inteligencia artificial y *machine learning*.

Características del lenguaje R: flexible, reproducible, código abierto e interfaces controlables a través de línea de comandos.

- ▶ El hecho de ser **flexible**, como hemos venido comentando, está relacionado con la capacidad que tiene para ofrecer una solución numérica a cualquier problema estadístico que necesitemos. Su amplio número de funciones implementadas deben satisfacer las demandas a nivel de principiante. En términos de investigación computacional, se puede necesitar desarrollar nuevas funciones y esto es posible con la estructura de código que tenemos disponible.
- ▶ **Código reproducible.** Como anunciamos en el apartado de «buenas prácticas», si escribimos un código claro (legible) podemos reutilizar el código para distintas bases de datos.
- ▶ El hecho de ser un **código abierto** permite siempre identificar errores o introducir mejoras en los procesos de desarrollo ya implementados. Es importante destacar que esto es un esfuerzo comunitario, como otros códigos abiertos, y en este sentido todas las aportaciones son bienvenidas.
- ▶ El poder trabajar con **líneas de comandos** da un poder superior al usuario que no se limita a «activar funcionalidades», sino que puede mejorarlas y entrar a perfilar su codificación en función de sus especificidades.

Como **limitaciones al trabajo con R** podemos señalar que el método de trabajo no suele ser intuitivo y se presenta como un acto de investigación en cuanto al desarrollo.

El usuario encuentra un espacio vacío al comenzar un *script* y a partir de ahí, como hemos indicado, el trabajo se centra en identificar ¿qué necesito?, buscar referencias previas de estudios en R y, finalmente, con esta información, explorar en la documentación propia del programa detalles más finos, como pueden ser las sintaxis de las funciones a utilizar.

A nivel de **hardware**, R posee algunas **limitaciones de memoria**. De momento encontramos dos limitaciones fundamentales que pueden afectar el trabajo en entornos *Big Data*:

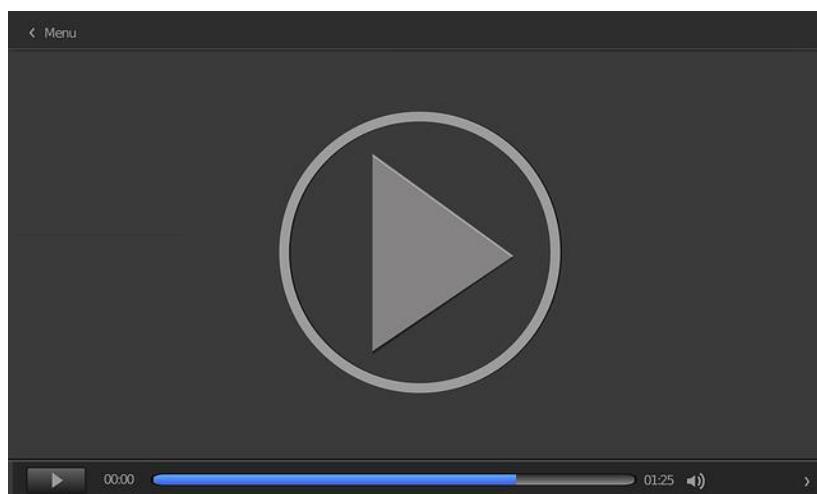
- ▶ **Acceso a la memoria operativa:** debido a que R trabaja en la memoria operativa, si por algo perdemos el cálculo, aquello que no hayamos guardado es susceptible de perderse. Por tanto, en cálculos extensos es recomendable ir guardando en el disco periódicamente.
- ▶ **Trabajo con extensas bases de datos:** como comentamos anteriormente, R tiene limitaciones de trabajo con algunas funciones que colapsan si las bases de datos superan determinadas dimensiones. Para resolver esto se usan paquetes auxiliares que en esencia se ocupan de dividir estas bases de datos tanto como sea necesario y utilizar conectores para operar de forma segura con ellas.

En general, este es el panorama de R en el entorno de estadística computacional.

Consideramos que la principal motivación para adentrarse en este contexto de desarrollo es la posibilidad de no ser un usuario de un código oscuro, sino ser capaz de usar y crear, incluso, aportaciones a un campo emergente como es el de la estadística computacional que hemos presentado en este tema.

Estadística computacional: inicio a la programación con R

En este vídeo vamos a introducir el concepto de estadística computacional y la programación con R como medio para llevarla a la práctica.



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=50489bf2-4dcf-41b3-a63d-acbd00aaf3eb>

Primeros pasos con R

Asumiendo que tienes [instalado R](#) y [RStudio](#), abre el IDE de RStudio y en un «R script» nuevo «Ctrl + Shift + N» escribe el siguiente código. En este ejercicio, es importante que leas los comentarios que hay en el propio código para comprender lo que hace cada instrucción.

Una vez copiado, prueba a ejecutar el *script* línea a línea. Para ello, posiciona el cursor en la primera línea y utiliza la opción «Run» o «Ctrl + Enter».

```
#####
##### Aprende a usar R desde RStudio #####
#####

# A) Para poner comentarios utiliza el símbolo '#' sin las comillas simples

#####
# B) Utiliza CTRL + SHIFT + C para comentar o descomentar un bloque de comentarios
# Esta línea es un comentario (Posiciona el cursor aquí y Utiliza CTRL + SHIFT + C)

#####
# C) Con el comando CTRL+ENTER se ejecutan las líneas de código seleccionadas
print('Esto es un primer mensaje por pantalla')
print('Esto es otro mensaje por pantalla')

#####
# D) Si no se selecciona nada, con el comando anterior se ejecuta 'línea por línea'
print('Esto es un segundo mensaje por pantalla')
print('Esto es otro mensaje por pantalla')

#####
# E) Con el comando CTRL+SHIFT+ENTER se ejecuta el script completo ::::: NO LO HAGAS :)

#####
##### Utiliza estructuras de datos en R #####
#####

# Es recomendable establecer un directorio de trabajo para R, en dicho directorio estarán todos los ficheros
# de tu proyecto.
#Indica aquí tu propia ruta
MAIN_DIR = 'C:/R-Proy/'
#Establece el directorio
setwd(MAIN_DIR)
#comprueba el directorio de trabajo actual
getwd()

#####
# Utiliza variables #####
# asigna el valor a una variable
y = 2
# Comprueba su tipo
class(y)

z = 2.3
class(z)

a = as.integer(2)
class(a)

a = as.integer(2.8)
class(a)
```

```

print(a)

b = 'str'
class(b)

#variables lógicas True/False
b = T
c = F
class(b)
class(c)

##### Utiliza vectores #####
vector1 = c(1,2,3,4,5,6,7,8,9,0)
class(vector1)
length(vector1)

# ¿Cuál es el tamaño de la variable a?
length(a)

vector1 = c(3:15)
print(vector1)

vector2 = c(2.1, 2.8, 3, 4.1, 55.2)
class(vector2)
# ¿Qué ocurre a continuación?
print(vector1 + vector2)

# ¿y ahora?
vector1 = c(1:5)
# ¿Funciona? ¿Qué ocurrió?
print(vector1 + vector2)

# Cadenas...
vector3 = c('str1', 'str2', 'str3')
class(vector3)

vector4 = c(8, 7.9, 'str2', 6, T, F)
class(vector4)
# Ten presente: todas las componentes del vector deben ser del mismo tipo -> cohercion.
# ¿Qué ocurre a continuación?
print(vector4)

vector5 = c(F, F, T, T, T)
class(vector5)

##### Utiliza matrices #####
# Debes crearlas a partir de vectores
vector1 = c(1:6)
print(vector1)
length(vector1)

m1 = matrix(vector1, ncol = 3, nrow = 2, byrow = F)
class(m1)
print(m1)

# Tamaño de matrix
length(m1)
dim(m1)
nrow(m1)
ncol(m1)

# Crea matrices a partir de vectores de distintas formas

```

```

m1 = matrix(vector1, ncol = 2, nrow = 2, byrow = F)
print(m1)

m1 = matrix(vector1, ncol = 2, nrow = 2, byrow = T)
print(m1)

m1 = matrix(vector1)
print(m1)
dim(m1)

m1 = matrix(vector1, ncol = 5, nrow = 6, byrow = T)
print(m1)

# Seguro habrá concluido que como las matrices se construyen a partir de vectores, estas solo aceptan un tipo
# de dato en sus componentes.
m2 = matrix(c('str',9, T, F, 8.1))
print(m2) #coercion

##### Utiliza factores #####
# Al igual que las matrices, se deben crear a partir de vectores.
# Los factores son útiles para manipular variables categóricas
vector1 = c(1:6)
f1 = factor(vector1)
class(f1)
print(f1)
levels(f1)
class(levels(f1))

f2 = factor(c(1,2,3,5,7))
print(f2)
levels(f2)

levels(f2) = c('a1', 'a2', 'a3', 'a5', 'a6')
print(f2)

##### Utiliza dataframes #####
# Es una de las principales estructuras de datos de R
# Puedes crear dataframe a partir de vectores, matrices y factores.
# Del dataframe, cada columna es un vector y por ello, los dataframes deben tener el mismo tipo de dato por
# COLUMNA.

df1 = data.frame(m1)
class(df1)
print(df1)

# La longitud no es el número de elementos, sino el número de columnas.
length(df1)
length(m1)

dim(df1)
nrow(df1)
ncol(df1)

# Utiliza nombres de filas y columnas (vectores)
colnames(df1) # == names(df1)
rownames(df1)

# Establece nombres de columnas o filas a tu dataframe
colnames(df1) = c('A1', 'A2', 'A3', 'A4', 'A5')
rownames(df1) = c('R1','R2','R3','R4','R5', 'R6')
print(df1)

```

```

# Más utilidades para los dataframes
mtcars
head(mtcars, 5)
tail(mtcars)

# Consulta para qué se utiliza la función str()
str(mtcars)

#####
##### Utiliza Listas #####
# En R, las listas son las estructuras de dato con mayor jerarquía.
# Una lista puede contener variables, vectores, matrices, factores, dataframes o incluso listas anidadas.
# Se podría afirmar que un dataframe es como una lista de vectores.

l1 = list(vector1, m1, f1, df1)
print(l1)
class(l1)
str(l1)
str(l1, max.level = 1)

# Establece nombres en tus listas
l2 = list('Object1' = vector1, 'Object2' = m1, 'Object3' = f1, 'Object4' = df1)
names(l2)
str(l2, max.level = 1)

l2 = list(vector1,m1,f1,df1)
str(l2, max.level = 1)
names(l2) = c('element1', 'element2', 'element3', 'element4')
str(l2, max.level = 1)

#####
##### Operaciones Lógicas #####
#####

# AND: &
T & F
# OR: |
T | F

#####
##### Utiliza variables en operaciones lógicas #####
a = 0
a > 3
a == 2
a != 4
!(a == 2)

#####
##### Utiliza vectores en operaciones lógicas #####
vector1 = c(1:4)
vector1 > 2
!(vector1 > 3)
vector1 %in% c(2,3)

#####
##### Utiliza matrices en operaciones lógicas #####
m1
m1 > 1
!(m1 > 2)
m1 %in% c(5,6)

#####
##### Selección de valores y porción de datos (Slice) #####
#####

#####
##### Selección de valores en vectores #####

```

```

vector1 = c(1:5)
vector1[5]

# Visualiza todos los elementos
vector1[-2]

# En R el índice comienza en 1 (en python que empieza en 0)
vector1[0]
vector1[c(1,6)]

# Es equivalente a v1[1:3]
vector1[c(1:3)]

# Visualiza todos los elementos
vector1[-c(1:3)]

# IMPORTANTE, APRENDER A USAR
# Comparación de valores en vectores
# Se pueden utilizar máscaras de booleanos
vector1[vector1>3]
vector1[!(vector1>3)]
vector1[vector1 %in% c(1,5)]
vector1[!(vector1 %in% c(2,3))]

# Utiliza nombre de componentes
names(vector1)
names(vector1) = paste('n', seq(1,length(vector1)), sep = '_')

# ¿Qué hace la instrucción anterior?
print(vector1)

vector1[c('n_1', 'n_2')]

# Utiliza una condición "x" para recuperar "y" valores de un vector.
# Tener en cuenta que "OJO" "x" e "y" deben tener el mismo tamaño.
names(vector1)[vector1>2]
vector1[names(vector1) == 'n_4']

##### Selección de valores en matrices #####
print(m1)

# Los elementos tienen posición i,j en la matriz (i = fila; j = columna)
m1[2,1]

# Recupera un vector de la matriz
m1[2:5, 4]

# Recupera una submatriz
m1[2:5, 1:4]

# Utiliza idx de 1 a length(m1)
m1[30]

# Lleva a cabo comparación de valores, pero ten presente que esto retorna un vector, no una matriz.
m1[m1>3]
m1[m1 %in% c(2,5)]

##### Selección de valores en dataframe #####
# La operativa es similar que en las matrices, pero además permite usar nombres de columnas
mtcars[1:5, c('cyl', 'wt')]

# Fila completa

```

```
mtcars[5,]

# Utiliza posición de columna
mtcars[4]
class(mtcars[4])
mtcars[,4]
class(mtcars[,4])

# Utiliza nombres de columna
mtcars['wt']
class(mtcars['wt'])

# Utiliza $ para referenciar las columnas
mtcars$wt
class(mtcars$wt)

# Utiliza condición en alguna(s) variable
mtcars[(mtcars$gear == 3),]
mtcars[(mtcars$carb == 3 & mtcars$mpg > 10),]
#obs: más adelante veremos como filtrar dataframes de mejor manera

##### Selección de valores en Listas #####
# Ten en cuenta que [[ ]] y $ son equivalentes

l2[1]
class(l2[1])

l2[[1]]
class(l2[[1]])

print(l2)

# ¿Recuerdas de donde salió el nombre "element4"?
l2$element4$A2
l2$element4$A3

l2[1:3]

# Utiliza la selección anidada
l2[[4]]
l2[[4]][['A1']]
l2[[4]][[['A1']]]
l2$element4$A4
```

Avanza un poco más con R

Asumiendo que tienes [instalado R](#) y [RStudio](#), abre el IDE de RStudio y en un «R script» nuevo «Ctrl + Shift + N» escribe el siguiente código. En este ejercicio, también es importante que leas los comentarios que hay en el propio código para comprender lo que hace cada instrucción.

Una vez copiado, prueba a ejecutar el *script* línea a línea. Para ello, posiciona el cursor en la primera línea y utiliza la opción «Run» o «Ctrl + Enter».

```
#####
##### Aprende a usar R desde RStudio #####
#####
```

```
#####
##### Sentencias condicionales #####
#####

##### Sentencia IF #####
condition = T
if (condition) {
print('...Se cumple la condición...')
}

##### Sentencia IF ELSE #####
condition = (6 == 2)

if (condition) {
print('Se cumple la condición')
} else {
print('NO se cumple la condición ')
}

##### Sentencia IF ELSE IF ELSE #####
q_flow = 35
if (q_flow > 90) {
print('nº mayor a 90')
} else if (q_flow < 90 & q_flow > 10) {
print('nº menor a 90 y mayor a 10')
} else {
print('nº menor a 10')
}

##### Condiciones y Vectores #####
vector1 = c(1:6)
condition = vector1 > 3
condition
if (condition) {
print('La condición se cumple')
}

# Alternativas para utilizar los vectores como parte de una condición
# or: a todos los componentes
any(condition)
# and: a todos los componentes
all(condition)

#####

##### Bucles #####
#####

##### While #####
contador = 0
while (contador < 6) {
print(contador)
contador = contador+1
}

##### While + break #####
contador = 0
while (T) {
print(contador)
contador = contador+1
if (contador == 2){
break
```

```
}

#####
# While + next #####
contador = 0
while (T) {
  if (contador == 2) {
    # Finaliza el bucle
    contador = contador+1
  next
}
print(contador)
contador = contador+1
if (contador == 5){
break
}
}

#####
# FOR #####
# Se utiliza principalmente para iterar vectores
for (var in 5:10) {
print(var)
}

var_values = c('cadena1', 'cadena2', 'cadena3')
for (var in var_values) {
print(paste('El valor de var es:', var))
}

# next y break se pueden utilizar en los bucles FOR, aunque es menos frecuente.
for (var in 1:10) {
if (var == 2){
next
}
print(var)
if (var == 5){
break
}
}

#####
# Instrucción REPEAT #####
# Es similar a while(True), se recomienda utilizar junto a 'break'
contador = 0
repeat {
if (contador == 2) {
#Finaliza el bucle
contador = contador+1
next
}
print(contador)
contador = contador+1
if (contador == 5){
break
}
}

# Es posible usar bucles anidados, incluso combinando los 3 tipos anteriores: while-for-repeat.

#####
##### Uso de Packages #####
#####
```

```
#importar paquetes: library() - require()
library(dplyr)
package_is_in = require(dplyr)
print(package_is_in)

library(MeInventoUnPaquete)
package_is_in = require(MeInventoUnPaquete)
print(package_is_in)

#instalar paquetes
package_name = 'dplyr'
install.packages(package_name)

#lista de paquetes instalados
installed.packages()
class(installed.packages())
rownames(installed.packages())
package_name %in% rownames(installed.packages())

#####
##### Utilities #####
#####

str(mtcars)
class(read.csv)
args(read.csv)
help(read.csv)
#?read.csv
?dplyr

# paste0 == paste(sep = '')
paste('Primera parte del str', 'segunda parte del str', sep = ', ')

u1 = 2
u2 = 4

# No se recomienda lo siguiente:
paste('El primer número es:', u1, 'y el segundo es:', u2)

# Esto es lo recomendado:
sprintf('El primer número es: %s y el segundo es: %s', u1, u2)
```

Practica con R los conceptos estudiados

Asumiendo que tienes [instalado R](#) y [RStudio](#), abre el IDE de RStudio y en un «R script» nuevo «Ctrl + Shift + N» escribe el siguiente código.

```
rm(list=ls())
#####
##### Tema 1#####

requiredPackages <- c("arsenal", "car", "corrplot", "gapminder","dplyr","DescTools", "foreign", "e1071",
"expss", "GGally", "ggplot2", "haven", "knitr","plotly", "remotes", "summarytools","ggridges","table1",
"tableone", "tidyverse", "SmartEDA")

session1 <- function(pkg){
new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
if (length(new.pkg))
install.packages(new.pkg, dependencies = TRUE)
```

```

sapply(pkg, require, character.only = TRUE)
}

session1(requiredPackages)
#####
##LOAD DATA
#Data Lending Club -https://www.kaggle.com/wordsforthewise/lending-club. Factores que determinan el Default en
los créditos. Modelo de riesgo
Datalc<-read.csv("https://raw.githubusercontent.com/millerjanny/Custom_UNIR/main/Data_LendingClub.csv")
# Exploración inicial
View(Datalc)
head(Datalc)
glimpse(Datalc)
names(Datalc)
str(Datalc)
dim(Datalc)
sapply(Datalc, function(x) sum(is.na(x)))

#####
#Tabla de frecuencias variable categórica
table(Datalc$home_ownership_n)
freq(Datalc$home_ownership_n, style = "rmarkdown")

#Tabla de frecuencias variable continua
Datalc$int_rate_cat <- factor(cut(Datalc$int_rate,
breaks=nclass.Sturges(Datalc$int_rate),include.lowest=TRUE))
freq(Datalc$int_rate_cat, style = "rmarkdown")
#####

#Pie chart
#Somos malos para juzgar el tamaño de los ángulos, que es lo que requieren los gráficos de tarta.
#Cambiar los colores de las porciones de la tarta hace que diferentes porciones parezcan más grandes o
pequeñas.
#Girar la tarta hace que las diferentes porciones parezcan más grandes o más pequeñas.
#Cuando hay pocas categorías, desperdician espacio.
#Cuando hay muchas categorías, son ilegibles.

df <- as.data.frame(table(Datalc$home_ownership_n)/length(Datalc$home_ownership_n))
colnames(df) <- c("class", "relative_freq")
pie <- ggplot(df, aes(x = "", y=relative_freq, fill = factor(class))) +
geom_bar(width = 1, stat = "identity") +
theme(axis.line = element_blank(),
plot.title = element_text(hjust=0.5)) +
labs(fill="class",
x=NULL,
y=NULL,
title="Pie Chart of Home ownership",
caption="Source: Lending Club")
pie + coord_polar(theta = "y", start=0)

#Barras
ggplot(data=Datalc, aes(home_ownership_n)) +
geom_bar(aes(y =(..count..)/sum(..count..))) +
ylab("Percent")+
geom_text(aes( label = scales::percent(..count..)/sum(..count..), accuracy = 0.01), y=
(..count..)/sum(..count..),accuracy = 0.01, stat= "count", vjust = -.5)+
scale_y_continuous(labels = scales::percent)

#comparación de Barras por variable cualitativa
ggplot(data=Datalc, aes(x=home_ownership_n)) +
geom_bar(aes(y =(..count..)/sum(..count..)*100),fill="steelblue") +
facet_wrap(~Default)+
```

```

ylab("Percent")

#Barras por variable cualitativa en el mismo gráfico
ggplot(Datalc, aes(x=home_ownership_n)) +
geom_bar(aes(y =(..count..)/sum(..count..)*100,accuracy = 0.01, fill = Default),
position = "dodge")+
ylab("Percent")

#Barras apiladas-variable cualitativa
ggplot(Datalc, aes(x = home_ownership_n, fill = Default)) +
geom_bar(aes(y =(..count..)/sum(..count..)*100,accuracy = 0.01))+ 
ylab("Percent")

#Histograma-Polígono de frecuencias-continua
ggplot(Datalc, aes(int_rate)) +
geom_histogram(mapping=aes(x=int_rate, y=..count../sum(..count..)*100),color="white",fill="darkblue",
bins=nclass.Sturges(Datalc$int_rate))+ 
geom_freqpoly(mapping=aes(x=int_rate, y=..count../sum(..count..)*100),
bins=nclass.Sturges(Datalc$int_rate))+ 
ggtitle("Distribution of Interest rate") +
xlab("Int Rate") +
ylab("Percent")

#Grafico de frecuencias acumuladas
ggplot(Datalc, aes(int_rate))+ 
geom_histogram(aes(y=cumsum(..count../sum(..count..)*100)),bins=nclass.Sturges(Datalc$int_rate),color='white',fill="darkblue")+

stat_bin(aes(y=cumsum(..count../sum(..count..)*100),bins=nclass.Sturges(Datalc$int_rate)),geom="line",color="green")+

xlab("Int Rate") +
ylab("Percent")

#comparar Grafico de frecuencias acumuladas por variable cualitativa
ggplot(Datalc, aes(int_rate, color = Default)) +
stat_ecdf(geom = "point")+
ylab("Relative frequency")

#Comparar distrubuciones de variable cuanti por variable cualitativa-crestas
ggplot(Datalc, aes(x = int_rate, y = Default)) +
geom_density_ridges(aes(fill = Default)) +
scale_fill_manual(values = c("#00AFBB", "#E7B800"))

ggplot(Datalc, aes(x = 'int_rate', y = 'home_ownership_n')) +
geom_density_ridges_gradient(aes(fill = ..x..), scale = 3, size = 0.3) +
scale_fill_gradientn(colours = c("#0D0887FF", "#CC4678FF", "#F0F921FF"),
name = "Interest rate")+
labs(title = 'Distribution of Interest rate by Home ownership')

#####
#El paquete gapminder contiene un fichero de datos de población, esperanza de vida y renta per cápita de los
países del mundo entre 1952 y 2007.
#La fundación Gapminder es una organización sin fines de lucro con sede en Suecia que promueve el desarrollo
global mediante el uso de estadísticas.
library(gapminder)
# Descripción de variables
# country: factor with 142 levels
# continent: factor with 5 levels
# year: 1952-2007
# lifeExp: life expectancy at birth
# pop: total population
# gdpPercap: per-capita GDP

```

```
#Gráfico de dispersión  
gap=data.frame(gapminder)  
ggplot(gap, aes(y=lifeExp, x=log(gdpPercap))) +  
geom_point()  
geom_smooth(method=lm)  
  
#Gráfico de linea-variable temporal  
spain <- gapminder %>%  
filter(country == "Spain")  
  
ggplot(spain, aes(x = year, y = lifeExp)) +  
geom_line(color = "#0099f9", size = 1) +  
geom_point(color = "#0099f9", size = 2) +  
labs(title = "Average life expectancy in Spain",  
subtitle = "Data from 1952 to 2007",  
caption = "Source: Gapminder dataset") +  
theme(plot.title = element_text(color = "#0099f9", size = 20, face = "bold", hjust = 0.5),  
plot.subtitle = element_text(size = 13, face = "bold", hjust = 0.5),  
plot.caption = element_text(face = "italic", hjust = 0))
```

Prueba a ejecutar el script anterior siguiendo estas indicaciones:

- ▶ Ejecuta cada línea de código, posiciona el cursor en la primera línea y utiliza la opción «Run» o «Ctrl + Enter».
- ▶ Observa la «Consola», la pestaña Environment y la pestaña Plot cuando ejecutes cada línea de código, especialmente cuando dibujes un gráfico.
- ▶ Encuentra la definición de cada función de R y comparte aquellas funciones que no conozcas en el foro de la asignatura.
- ▶ Repasa con R todos los conceptos vistos hasta ahora.

Sobre la relación entre Estadística, Informática y Big Data

Ferrero, R. y López, J. L. (s.f.). La estadística en la era del *Big Data*. *Data science* [Blog].

Recomendamos, especialmente, la lectura del apartado «El Universo del Big Data en Expansión», donde se muestra con datos cuantitativos la necesidad tangible hoy en día de desarrollar métodos computacionales para abordar problemas estadísticos.

Accede al artículo desde el aula virtual o a través de la siguiente dirección web:
<https://www.maximaformacion.es/blog-dat/la-estadistica-en-la-era-del-big-data/>

Ejemplo de aplicación del análisis estadístico con métodos computacionales y usando software R a un problema real

Recomendamos la revisión global a este trabajo: *Basic principles in Biostatistics: likelihood and statistical thinking*. Ejemplo práctico que muestra la elección de un problema real, desarrollo de un modelo estadístico, elección de un software de cómputo y cálculos numéricos para dar solución al problema.

Este trabajo es una guía, incluso una plantilla, de lo que debe ser el trabajo del analista de datos y puede servir incluso para mostrar posibles líneas de trabajo en las actividades que se propondrán durante el curso.

Accede al artículo desde el aula virtual o a través de la siguiente dirección web:

https://borishejblum.science/html/m2phds-basics/biostatistics_basics_mlepracticals#1_motivational_example

Bibliografía

Hey, T., Tansley, S. y Tolle, K. (2009). The Fourth Paradigm: Data-intensive Scientific Discovery. *Microsoft Research* [Web].

Disponible en: <https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/>

R Development Core Team (ed.). R manuals. *Cran* [Web].

Disponible en: <https://cran.r-project.org/manuals.html>

1. R soporta datos de tipo numérico en sus bases de datos:

 - A. Verdadero.
 - B. Falso.
 - C. Solo si se introducen como tipo .txt .
 - D. Ninguna de las propuestas es correcta.

2. R soporta datos de tipo categórico en sus bases de datos:

 - A. Verdadero.
 - B. Falso.
 - C. Solo si van acompañados de algún valor numérico.
 - D. Ninguna de las propuestas es correcta.

3. Histogram() es la etiqueta para desarrollar una función que elabore histogramas en un algoritmo desarrollado con R:

 - A. Incorrecto.
 - B. Correcto.
 - C. Falta colocar las etiquetas para completar el histograma.
 - D. Ninguna de las propuestas es correcta.

4. En estos momentos, R es un *software* que ofrece soporte ilimitado a la solución de problemas estadísticos en el entorno *Big Data*.

 - A. Sí, pero con limitaciones.
 - B. Nos impone la necesidad de trabajar para evitar problemas de asignación de memoria.
 - C. Posibilita el uso de funciones de código abierto para optimizar los recursos de memoria.
 - D. Todas las respuestas anteriores son correctas.

5. ¿Por qué puede ser relevante la irrupción del código R en temas de ciberseguridad?

- A. Facilita el tratamiento de muchos datos.
- B. Prima la lógica de los programas y la capacidad creativa del desarrollador a los mecanismos de control internos.
- C. Posibilidad de paralelización de procesos al tener las estructuras modularizadas.
- D. Todas las propuestas anteriores son correctas.

6. Uno de los objetivos básicos de la programación es la capacidad de desarrollar código que sea reutilizable:

- A. Verdadero, pero no aplicable al contexto estadístico donde cada código debe limitarse a un problema específico.
- B. Verdadero, extensible al área de la estadística donde se pretenden crear códigos generalistas que puedan ser utilizados sobre distintos escenarios.
- C. Falso, siempre se debe empezar el código de cero al implementar un problema.
- D. Ninguna de las anteriores.

7. Sobre el uso de la programación por módulos en R:

- A. Facilita la reutilización de código.
- B. Permite detectar errores (*bugs*) en un proceso de validación de código.
- C. Hace el código más expresivo.
- D. Todas las anteriores son correctas.

- 8.** R no permite compartir librerías con otros lenguajes:
- A. Verdadero, asociado a la seguridad propia del lenguaje.
 - B. Verdadero, en la línea de garantizar un uso matemáticamente correcto de los datos.
 - C. Falso, las librerías se pueden compartir con otros lenguajes de programación.
 - D. Ninguna de las anteriores.
- 9.** ¿Puede R trabajar con varios tipos de ficheros de datos?
- A. Sí, siempre que sean almacenables como .txt .
 - B. Sí, puede trabajar con varios tipos de ficheros, ejemplo .txt, .csv .
- 10.** ¿Puede un solo código R tratar simultáneamente variables categóricas y numéricas?
- A. Sí, es algo estándar.
 - B. No, una u otra, nunca simultáneamente. Puede dar errores en el proceso de compilación.
 - C. No, deben transformarse a uno u otro tipo y elegir un tipo para cada código.
 - D. Ninguna de las anteriores es correcta.

Análisis e Interpretación de Datos

Tema 3. Medidas que resumen la información

Índice

[Esquema](#)

[Ideas clave](#)

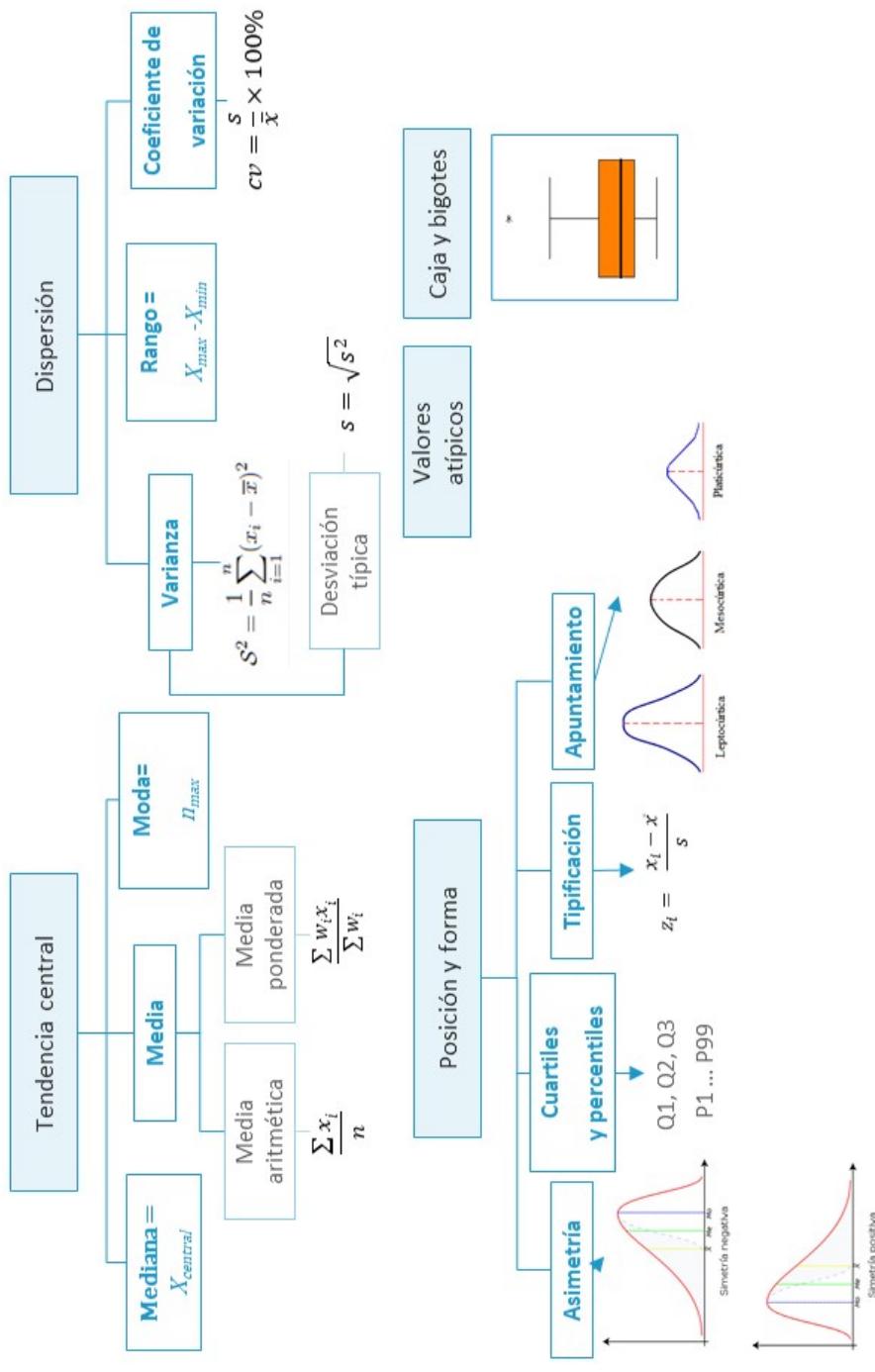
- [3.1. ¿Cómo estudiar este tema?](#)
- [3.2. Medidas de tendencia central](#)
- [3.3. Medidas de tendencia central robustas](#)
- [3.4. Medidas de dispersión](#)
- [3.5. Medidas de dispersión robustas](#)
- [3.6. Medidas de posición y forma](#)
- [3.7 Gráficos de caja](#)
- [3.8 Datos atípicos y análisis exploratorio de datos](#)
- [3.9. Referencias bibliográficas](#)

[A fondo](#)

- [Medidas de Tendencia Central con Excel](#)
- [Medidas estadísticas](#)
- [Estadísticas aplicadas al deporte](#)
- [Construir un diagrama de caja y bigotes en Excel](#)
- [Estadística y probabilidad](#)
- [Bibliografía](#)

[Test](#)

MEDIDAS RESUMEN DE LA INFORMACIÓN



3.1. ¿Cómo estudiar este tema?

Para estudiar este tema lee las **páginas 39-68** del siguiente libro:

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

Versión

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

Este tema versa sobre cómo resumir la información de una distribución estadística en números. Estos números son medidas que resumen la información y características de la muestra y/o la población. Para hacerse una idea global es importante que mires el esquema del tema, el cual te ayudará a hacerte una buena idea de cómo está estructurado el tema.

También será clave que practiques con las actividades propuestas en el tema, los cuales están diseñados para que apuntes las ideas más importantes sobre medidas resumen y algunos conceptos asociados también muy importantes, como la gráfica de caja y bigotes que es fundamental para hacerse una idea resumida de la distribución así como de la posible presencia de valores atípicos.

3.2. Medidas de tendencia central

Las primeras medidas que vamos a estudiar son las que giran alrededor de la idea de centro de la distribución de los datos. Son por tanto valores que se encuentran en el medio o la mitad de un conjunto o distribución de datos. Estas medidas o **estadísticos** también persiguen identificar valores que sean algo así como representantes de todos los datos.

La primera medida de tendencia central y más sencilla ya es conocida por nosotros aunque sea de forma informal, la **media aritmética** es la medida que consiste en la suma de todos los valores dividida por el número de estos valores.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Cuando el tipo de distribución de datos contenga de un dato para algún valor será necesario completar un poco la fórmula anterior para sumar tantas veces el valor repetido (x_i) como frecuencia presente este (n_i). De modo que nos queda esta nueva versión así:

$$\bar{x} = \frac{\sum n_i x_i}{\sum n_i} = \frac{n_1 x_1 + n_2 x_2 + \dots + n_k x_k}{n_1 + n_2 + \dots + n_k}$$

En algunas ocasiones —con mucha menos frecuencia que la media aritmética— surge una media que no está basada en una concepción frecuentista, sino que es ponderada estando cada valor multiplicado por un peso. Esta es la **media ponderada**:

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

donde los w_i son los pesos o ponderaciones de cada x_i .

En contadas ocasiones se presenta un tercer tipo: la **media armónica** (que curiosamente se denomina H a pesar de que se escribe sin «h») que se emplea por ejemplo con las velocidades. Ningún valor puede ser cero para poder calcularla, pues como sabemos no existen números reales que se obtengan al dividir por cero.

$$H = \frac{n}{\sum \frac{n_i}{x_i}}$$

Ejemplo 1: Una familia se va de puente de Madrid a Barcelona a una velocidad de 100km/h, y tres días después regresa el domingo por la noche a 120 km/h. ¿Qué velocidad media ha tenido la familia en los dos trayectos?

$$H = \frac{2}{\frac{1}{100} + \frac{1}{120}} = 109,1 \text{ km/h}$$

Si quieras profundizar sobre las clases de medias conviene que sepas que existen aún dos más: la **geométrica** y la **cuadrática**, las cuales puedes ver en el capítulo indicado en «Cómo estudiar este tema».

La **limitación de la media aritmética** (de ahora en adelante «media» a secas) consiste en lo mucho que le afectan las observaciones que presentan valores atípicos. Es por ello que la información que condensa no es suficiente para explicar cómo se distribuyen las observaciones.

Una medida de tendencia central que es más robusta que la media frente a valores extremos es la **mediana**, observación que ocupa el lugar central en un conjunto de datos. Al ocupar esta posición se nos presentan **dos casos**:

1. Cuando hay un número impar de observaciones la mediana ocupa justo el valor central.
2. Cuando el número de observaciones es par no existe posición central por lo que la mediana será el promedio entre las dos observaciones centrales.

En los casos en los que **tengamos grandes cantidades de datos**, la **mediana** nos proporciona información mucho más fiable sobre la tendencia general de los datos que la media. Si la cantidad (que no el porcentaje) **de outliers en nuestra muestra es alta**, es interesante separar dichos datos y realizar sobre ellos **un estudio aparte**. De esta forma, podremos constatar de forma mejor su naturaleza y **entender qué está ocurriendo**.

Ejemplo 2: Si tenemos los siguientes datos correspondientes a los puntos anotados por Gasol durante sus 13 años en la NBA:

PUNTOS	TEMPORADA
1441	2001-02
1555	2002-03
1381	2003-04
997	2004-05
1628	2005-06
1226	2006-07
1246	2007-08
1528	2008-09
1190	2009-10
1541	2010-11
1129	2011-12
673	2012-13
958	2013-14

Tabla 1: Puntos de Gasol en la NBA.

Lo primero que haríamos sería ordenarlos de menor a mayor (o de mayor a menor es indiferente).

POSICIÓN	PUNTOS	TEMPORADA
1	673	2012-13
2	958	2013-14
3	997	2004-05
4	1129	2011-12
5	1190	2009-10
6	1226	2006-07
7	1246	2007-08
8	1381	2003-04
9	1441	2001-02
10	1528	2008-09
11	1541	2010-11
12	1555	2002-03
13	1628	2005-06

Tabla 2: Puntos de Gasol ordenados.

Y de esta manera la mediana sería 1246 puntos, que es el valor que ocupa la posición siete que es la central al dejar tantos a la izquierda como a la derecha (6 menores y 6 mayores).

Para estudiar el caso par imaginemos que solo contamos con las temporadas completas por lo que llegamos hasta el 2012-13. De esta manera tendríamos dos puntuaciones anuales centrales.

POSICIÓN	PUNTOS	TEMPORADA
1	673	2012-13
2	997	2004-05
3	1129	2011-12
4	1190	2009-10
5	1226	2006-07
6	1246	2007-08
7	1381	2003-04
8	1441	2001-02
9	1528	2008-09
10	1541	2010-11
11	1555	2002-03
12	1628	2005-06

Tabla 3: Puntos de Gasol hasta el 2011/12 (número de temporadas par).

Y ahora la mediana sería el promedio de estas dos puntuaciones centrales:

$$Me = \frac{1246 + 1381}{2} = 1313,5$$

Un último caso de mediana surge cuando la distribución de datos se muestra en intervalos. En este caso es preciso emplear la siguiente fórmula de interpolación de la mediana:

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \times a_i$$

$$rango = x_{max} - x_{min}$$

Para aplicar esta fórmula hay que manejar el concepto de **intervalo mediano** que es aquel que contiene a la mediana. El límite inferior de este intervalo es L_{i-1} , su amplitud a_i , su frecuencia n_i , y por último N_{i-1} es la frecuencia acumulada hasta el intervalo anterior. Sabiendo esto ya puedes calcular la fórmula anterior. Como curiosidad la mediana para estas distribuciones con los **datos agrupados** (el otro nombre que reciben las distribuciones en intervalos) coincide con el punto cuya área acumulada es la mitad de la historia del histograma.

Una ventaja de la mediana frente a la media aritmética reside en que se emplea para variables cualitativas también, mientras que la media no puede ser empleada para estas variables sino tan solo para las cuantitativas.

Una última medida considerada tradicionalmente como de tendencia central es la **moda**, que no es más que el valor más frecuente del conjunto de datos. Si el conjunto contiene dos datos con la misma frecuencia diremos que la distribución es **bimodal** por tener dos modas. Y por otro lado nada impide que tenga tres modas, o cuatro, o las que sean, aunque no es habitual puede suceder.

3.3. Medidas de tendencia central robustas

Tal y como hemos comentado, la media es una medida que **se ve muy afectada por los valores extremos**. Por tanto, no proporciona un valor de tendencia central fiable cuando la aplicamos en conjuntos de datos que poseen datos outliers. Para solucionar este problema, podemos, como ya hemos comentado, usar la **mediana**. Dicha medida se centra en encontrar el valor central del conjunto de datos y, por tanto, **no tiene en cuenta los valores extremos**. Esto hace que, de manera natural, sea una medida muy robusta a la que los outliers no le afectan. Sin embargo, su significado, tal y como hemos estudiado es algo diferente del de la media aritmética.

Para solucionar esto, existen varias **versiones de la media aritmética que tratan de evitar que los outliers influyan en el resultado final obtenido**. Estudiaremos, en concreto, cómo funcionan la **media recortada** y la **media winsorizada**:

Media recortada: la media recortada realiza la media aritmética a un **subconjunto central del conjunto de datos**. De esta manera, los valores outliers quedan a los extremos y no influyen en el resultado final obtenido. Por lo general, hablamos de «**media recortada al y%**» donde y indica el porcentaje de datos que debemos dejar de lado por cada extremo. Por ejemplo, si tenemos 10 datos y calculamos una media recortada al 40 %, debemos obviar 4 datos a la izquierda y 4 a la derecha calculándose la media únicamente sobre los dos valores centrales. Otros datos a tener en cuenta es que una media recortada al 0 % es equivalente a calcular una media aritmética y que cuando hablamos de medias recortadas al 25 % el cálculo se denomina «**centrimedia**».

Veamos un ejemplo. Imaginad que tenemos los siguientes datos y nos piden calcular una **media recortada al 10 %**:

3	4	4	5	5	6	7	8	9	11
---	---	---	---	---	---	---	---	---	----

El 10 % de 10 valores corresponde a 1 valor. Por tanto, **eliminamos un valor por la derecha y otro por la izquierda** y realizamos la media aritmética sobre el siguiente subconjunto de datos:

	4	4	5	5	6	7	8	9	
--	---	---	---	---	---	---	---	---	--

Lo que nos da un valor de 6. Si directamente queremos recortar un número fijo de elementos en vez de trabajar con porcentajes podemos hablar de niveles. El ejemplo visto es de nivel 1 porque hemos recortado un valor a cada lado.

- **Media winsorizada:** la media winsorizada funciona de manera similar a la media recortada. La principal diferencia radica en que en la media winsorizada, en vez de eliminar los valores, **los sustituye por el menor y mayor valor que queda en el conjunto tras el proceso de eliminación**. La media winsorizada de nivel 2 del ejemplo visto consistiría en realizar una media aritmética sobre el siguiente conjunto de datos:

4	4	4	5	5	6	7	8	8	8
---	---	---	---	---	---	---	---	---	---

Como podemos ver, hemos eliminado los dos valores más extremos del conjunto y los hemos sustituido por los valores extremos del conjunto de valores restante. El resultado en este caso nos da un valor de 5,9.

Tal y como puede observarse, la idea principal de estas medias consiste en **eliminar los valores extremos y realizar los cálculos sobre valores situados en torno al centro de la distribución**. De esta manera los valores outliers no afectan al resultado final y podemos calcular un valor de tendencia central centrado únicamente en los valores más típicos encontrados en el modelo de datos.

3.4. Medidas de dispersión

Las medidas de dispersión nos indican cuánto se desvían los datos, aspecto que es fundamental para conocer cómo se distribuye el conjunto de datos.

La medida de dispersión más básica es el **rango**, que no es más que la diferencia entre la observación mínima y la máxima.

$$\text{rango} = x_{\max} - x_{\min}$$

El rango es útil hasta cierto punto pero son necesarias otras medidas que incluyan información sobre el resto de valores y no solo del máximo y el mínimo. Bajo esta filosofía surge la **varianza** que no es más que un promedio de las desviaciones de los datos a su media. Estas desviaciones son elevadas al cuadrado para que no le afecte el sentido de estas (si es negativo o positivo).

$$\text{Varianza} = s^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

Cuando nos refiramos a la varianza poblacional la designaremos con σ^2 .

Como manejando la varianza hemos elevado las desviaciones al cuadrado perdemos un tanto la referencia de magnitud respecto a los datos. Por este motivo nosotros manejamos su raíz, que es la conocida como **desviación típica** o desviación estándar.

$$\text{Desviación típica} = s = + \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

Donde he colocado el más delante de la raíz (lo cual estrictamente en notación del lenguaje matemático no es del todo correcto) para indicar que siempre es positivo su valor, cosa que es lógica puesto que mide el tamaño de una desviación o distancia que nunca puede ser negativa.

Conviene hacer una puntuación en cuanto al denominador empleado en esta fórmula. La tradición anglosajona de la estadística acostumbra a dividir por $n - 1$ en lugar de entre n . La razón tras ello será explicada cuando veamos los estimadores de la varianza, pues veremos que al dividir entre $n - 1$ este estadístico muestral cumplirá ciertos requisitos que nos interesarán. Se comenta esto para evitar futuras confusiones especialmente si se recurre a fuentes estadísticas americanas por ejemplo.

Ejemplo 3: Volviendo al ejemplo de Gasol anterior, ¿cuál sería la desviación típica de sus puntuaciones anuales? Para calcularla lo primero que hacemos es dibujar la tabla de frecuencias adecuada.

	PUNTOS	Puntos*Puntos		
1	673	452929		
2	958	917764		
3	997	994009		
4	1129	1274641		
5	1190	1416100		
6	1226	1503076		
7	1246	1552516		
8	1381	1907161		
9	1441	2076481		
10	1528	2334784		
11	1541	2374681		
12	1555	2418025		
13	1628	2650384		
SUMA	16493	21872551	s^2	72923,75
MEDIA	1268,692	1682503,923	s	270,044

Y entonces la fórmula para el cálculo es mejor realizarla a través de esta versión:

$$s^2 = \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{21872551}{13} - 1374,42^2 = 213132 \rightarrow s = \sqrt{213132} = 461,66$$

Un aspecto que no cubre la desviación típica es la comparación entre poblaciones, ya que esta refleja una magnitud que depende de la escala que tenga la población. Por ello, se maneja un estadístico llamado **coeficiente de variación**, el cual nos permite comparar la variación entre diferentes poblaciones ya que al dividir la dispersión por la media logramos que la medida carezca de unidades.

$$CV = \frac{s}{\bar{x}} \times 100\%$$

Ejemplo 4: Tenemos una población de mujeres en la que hemos medido dos variables: Peso y estatura. La estatura media resulta de 1,68m mientras que el peso medio es de 57Kg. No podemos comparar ambas magnitudes porque se encuentran en unidades diferentes, pero si también disponemos de las desviaciones típicas $s_{estatura} = 7,5\text{cm}$ y $s_{peso} = 12\text{kg}$ entonces podemos calcular los coeficientes de variación empleando la fórmula anterior:

$$CV_{estatura} = \frac{7,5}{1,68} \times 100\% = 4,46\%; CV_{peso} = \frac{12}{57} \times 100\% = 21,05\%$$

Ahora ya podríamos comparar las variables observando que la estatura tiene una variación casi cinco veces menor que el peso.

3.5. Medidas de dispersión robustas

De forma análoga a como calculábamos la media winsorizada en el apartado 2.3, es posible calcular la **varianza winsorizada** siguiendo la siguiente fórmula:

$$s_W^2 = \frac{1}{n} \sum_{i=1}^n (W_i - \bar{x}_\alpha^W)^2$$

Donde W_i hace referencia al conjunto de datos winsorizado según el mismo proceso visto en el apartado 2.3 con la media y \bar{x}_α^W hace referencia a la media winsorizada.

Siguiendo el mismo patrón que para definir la cuasivarianza, podemos definir la **cuasivarianza winsorizada** de la siguiente manera:

$$S_W^2 = \frac{1}{n-1} \sum_{i=1}^n (W_i - \bar{x}_\alpha^W)^2$$

$$z_i = \frac{x_i - \bar{x}}{s}$$

Donde la única diferencia con la fórmula anterior es que **se divide entre $n - 1$ en vez de entre n** . Es importante tener en cuenta que es posible, tal y como sucedía con la desviación típica y la varianza, que puede calcularse la **cuasideviación típica winsorizada** como la raíz de la cuasivarianza winsorizada.

3.6. Medidas de posición y forma

Las medidas de posición son necesarias para poder saber si un valor está alejado o no de su media, lo cual nos da idea de lo extremo que es comparado con la «mayoría» de los datos de su conjunto. Para esto una herramienta útil es la tipificación a través de la siguiente fórmula:

$$z_i = \frac{x_i - \bar{x}}{s}$$

Los valores z_i son llamados **puntuaciones tipificadas** y nos indican el número de desviaciones típicas que se aleja un valor de su media. Además, otra propiedad interesante que tienen es que nos permiten comparar diferentes variables que pueden ser de la misma o de diferentes poblaciones también, pues al tipificar un valor desaparecen sus unidades. Un elemento común del coeficiente de variación y de las puntuaciones típicas es precisamente esta propiedad. Sin embargo no conviene confundirlos, pues mientras que el coeficiente de variación lo es de toda la muestra o población la puntuación tipificada se calcula para cada puntuación (Rius et al., 1998).

Ejemplo 5: Regresando al ejemplo de Gasol, si quisiéramos saber cuándo fue comparativamente mejor su puntuación, si en 2002/2003 o en 2012/2013, tendríamos que disponer de las medias de puntuación de la NBA para ambas temporadas y sus desviaciones típicas. Si la media de puntos para los pívots fue de 707 puntos con una desviación típica de 451 puntos para el 2002/03 y de 729 puntos con una desviación típica de 411 puntos para el 2012/13. ¿Cuándo obtuvo Gasol mejores resultados en comparación con el resto de jugadores de la NBA en su misma posición? Para saberlo calcularíamos las puntuaciones tipificadas de Gasol para estas dos temporadas.

$$z_{2002/03} = \frac{1555 - 707}{451} = 1,88; z_{2012/13} = \frac{673 - 729}{411} = -0,14$$

Y de esta manera sabemos que fue en el 2002/03 cuando fue comparativamente mejor, alejándose casi dos desviaciones típicas frente a sus compañeros pívot. En la temporada 2012/13 fue sin embargo algo peor que la media de sus compañeros de esa misma temporada.

A partir de las puntuaciones típicas podemos saber precisamente si una puntuación es frecuente o no lo es dentro de su población. Para ello, basta con contemplar si z está comprendida entre -2 y 2. Cuando salen fuera de este rango se considera infrecuente. Esta afirmación se basa en la distribución normal, la cual veremos más adelante cuando abordemos las principales distribuciones de probabilidad.

Otra medida fundamental para tratar la posición de los datos es el **cuartil** que, como su nombre indica, proviene de dividir en cuartos iguales el conjunto de datos. De este modo tenemos tres cuartiles: Q_1 , Q_2 y Q_3 . Así el primer cuartil es el valor que hace que el 25% de los datos sean inferiores a él, el cuartil segundo deja el 50% de las observaciones y por tanto coincide con la mediana. El cuartil tercero por último será aquel que deja el 75% de las observaciones a su izquierda.

En realidad los cuartiles son a su vez casos específicos de **percentiles** y estos a su vez de **cuantiles** que son el nombre global que reciben estas medidas de posición. Los percentiles son aquellos valores que dejan un tanto por ciento de los datos a su izquierda. Así diremos que por ejemplo el percentil 40 es aquel que deja el 40% de los datos a su izquierda y el resto a su derecha (un 60%). Todavía hay un tipo de cuartil más que son los **deciles** que como su nombre indica habría nueve (el décimo no cuenta de modo análogo a los cuartiles que son tres).

Según esto los cuartiles 1º, 2º y 3º equivaldrán a los percentiles 25, 50 y 75 respectivamente. El cuartil segundo a su vez también equivaldría al decil quinto y a la mediana. Matemáticamente para que te familiarices con la notación quedaría:

$$P_{50} = D_5 = Q_2 = Me$$

Ejemplo 6: De nuevo nos basaremos en los datos de las puntuaciones de Gasol, ya que las hemos empleado anteriormente para ilustrar la mediana, que de hecho coincide con el cuartil segundo, como ya hemos comentado. Se ha generado esta tabla traspuesta (no es la habitual en columnas) con las frecuencias acumuladas, que son las que marcan las posiciones donde estarán los percentiles. Los primeros cálculos por tanto son los que identifican la posición de cada percentil del siguiente modo:

Para el $P_{25} = Q_1$ tenemos que el valor que ocupa la posición y por tanto sería el promedio entre 1129 y 997, que es 1063.

Para el $P_{50} = Q_2 = Me$ tenemos que el valor que ocupa la posición $(N + 1)/2 = 14/2 = 7$, luego es el que dato que ocupa el lugar 7 que es el 1246.

Para el $P_{75} = Q_3$ tenemos que el valor que ocupa la posición $3(N + 1)/4 = 42/4 = 10,5$; por lo que resulta el promedio o semisuma entre 1528 y 1541 que resulta 1534,5.

Puntuaciones anuales	673	958	997	1129	1190	1226	1246	1381	1441	1528	1541	1555	1628
N	1	2	3	4	5	6	7	8	9	10	11	12	13
Percentil				25			50			75			
Cuartil				1			2			3			

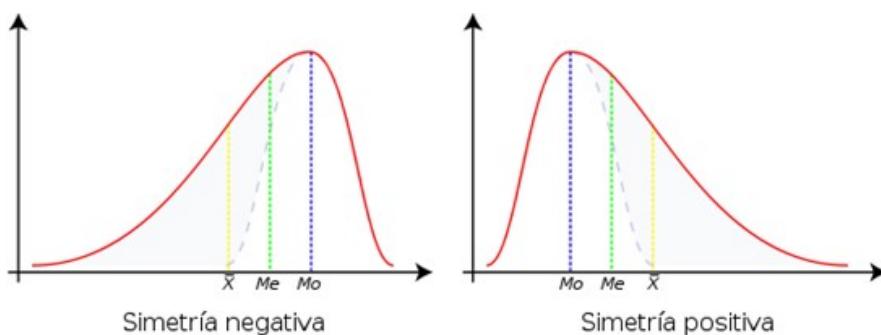
Una confusión habitual con los cuantiles es confundir la posición con el cuantil, algo para la que hay que estar atentos cuando los calculemos pues de otro modo acabamos dando por bueno un resultado que no lo es.

La fórmula vista anteriormente para el cálculo de la mediana con datos agrupados es extensible al resto de cuartiles de este modo:

$$Q_k = L_{i-1} + \frac{\frac{kN}{4} - N_{i-1}}{n_i} \times a_i$$

Ten en cuenta que estamos designando el inicio de un intervalo como L_{i-1} pero hay quien lo considera L_i . Esto te debe dar igual, pues escribiéndolo de un modo u otro siempre significa el inicio del intervalo mediano o del intervalo que contenga al cuartil correspondiente.

Pasaremos ahora al estudio de la forma de la distribución de datos, que engloba la simetría y el apuntamiento. En cuanto a la primera la distribución puede ser simétrica o bien presentar **asimetría positiva o negativa** (también designada simetría positiva o negativa), dependiendo hacia el lado que tenga la cola.

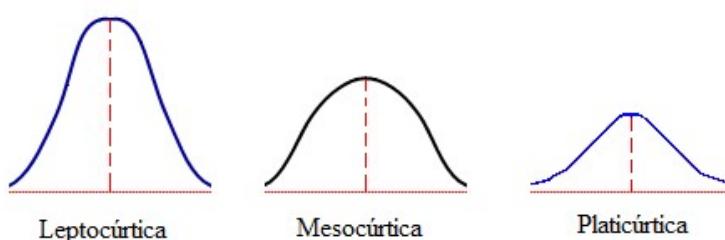


Gráfica 1: Tipos de asimetría y su relación con las medidas de tendencia central.

También podemos apreciar la relación entre la simetría o asimetría y la posición de la media, mediana y moda. En el caso de la simetría perfecta, que no aparece en la imagen coinciden media, mediana y moda. Para medir la simetría los programas estadísticos en ocasiones también confeccionan un coeficiente de simetría (As de Pearson) que resulta positivo cuando la asimetría es positiva resultando que la media sea mayor que la moda, nulo cuando es simétrica (Moda=Media) y negativo cuando la asimetría es negativa resultando que la moda sea mayor que media.

$$As = \frac{\bar{x} - Mo}{s}$$

En cuanto al **apuntamiento**, conviene que se conozcan los tres tipos de distribuciones según su forma sea más achitada (**platicúrtica**) o si es más puntaaguda (**leptocúrtica**). Siendo **mesocúrtica** en los casos intermedios.



Gráfica 2: Tipos de distribución según su apuntamiento.

3.7 Gráficos de caja

Con objeto de resumir la información del conjunto de datos haciendo énfasis en la distribución general de estos, se desarrolló el **diagrama de caja y bigotes**. (**boxplot** en inglés). Es probable que hayas visto alguna vez uno de ellos pero que no sepas con exactitud cómo está construido y por tanto como interpretarlo. Para confeccionarlo tenemos que contar con cinco medidas de las que ya hemos visto. Se dice que con estas **cinco medidas resumen** podemos condensar de manera rápida cualquier distribución estadística, reflejando algunas de las propiedades y facetas que no quedaban cubiertas por las gráficas tradicionales:

Mínimo, Q1, M, Q3, Máximo

A partir de los cuartiles primero y tercero se suele confeccionar otra medida que es muy útil: el **rango intercuartílico**.

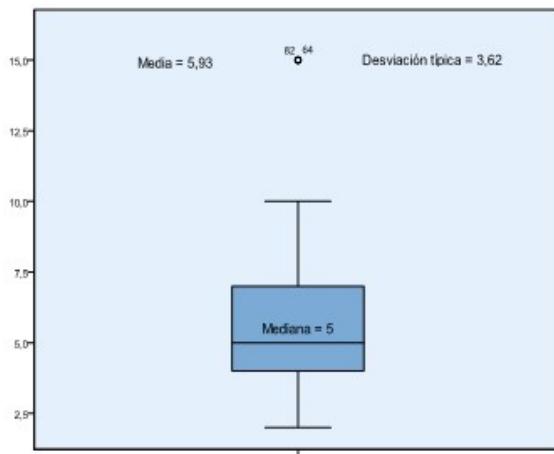
Rango IQ = Q3 – Q1

Este estadístico es muy útil pues nos marca el intervalo que ocupa el 50% central del conjunto de datos.

La manera de construir el diagrama de caja y bigotes es la siguiente:

1. Con los cuartiles 1 y 3 marcamos los límites de la caja.
2. La mediana establece la línea que parte la caja en dos.
3. Los bigotes tienen como principio y final el mínimo y el máximo, salvo que haya valores atípicos, en cuyo caso estos alcanzarán el tope de 1,5 veces el Rango Intercuartílico.

Ejemplo 7: Aquí podemos ver cómo es un diagrama de caja y bigotes correspondiente al número de horas semanales de estudio que dedican los estudiantes de cierta asignatura.



Gráfica 3: Diagrama de caja y bigotes de la variable «número de horas semanales de estudio».

La mediana está situada dentro de la caja. En este caso la he puesto título y su valor para ilustrarlo pero lo normal es que no figure explícitamente sino que se muestre simplemente a través de la línea. Los límites de la caja, es decir, los cuartiles, son aproximadamente 4 (dicho «a ojo» por estar algo más cerca del 5) y 7. De modo que la mitad de los datos están contenidos entre estos dos valores.

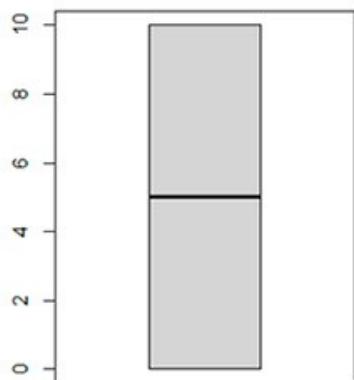
Los bigotes se confeccionan trazando una línea o bigote que une el dato mínimo con el máximo, salvo en el caso de que haya valores que excedan 1,5 veces el rango intercuartílico. En el caso del gráfico existirían dos datos atípicos, los cuales se remarcán con circulitos. Otras maneras de marcar estos datos extremos o atípicos es empleando asteriscos.

También existen programas estadísticos que establecen diferentes categorías de valores atípicos (por ejemplo, el programa del que proviene el gráfico anterior es el llamado SPSS, el cual emplea circulitos para los datos muy extremos y asteriscos para los **atípicos «normales»**). Precisamente, sobre este programa se desarrolló lo que es hoy el código R, y por esta razón histórica hemos querido incluirlo, al menos, en un ejemplo a lo largo de este curso.

Veamos la **realización de un diagrama de cajas** similar al anterior, pero utilizando el **código R**.

Ejemplo: supongamos tenemos la lista de notas de un grupo de 10 alumnos, que introducimos en la variable NotaAlumnos. Queremos ilustrar estas notas en un diagrama de cajas, usando los recursos de funciones ya instaladas en R para la realización de este gráfico. El código y los resultados quedarían como se muestra a continuación.

```
>NotaAlumnos <- c(10,10,10,10,10,0,0,0,0,0)  
  
>boxplot(NotaAlumnos)
```



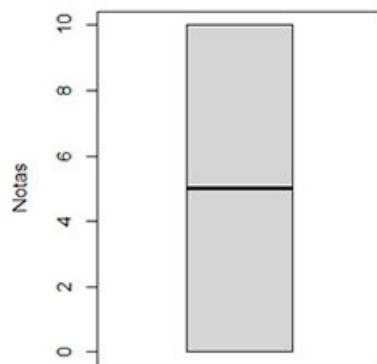
Nótese con este ejemplo el proceder recomendado ante un *software* con el cual nos estamos iniciando en la programación.

Primero, probaremos con un caso modelo, computable «a mano» y que sirva para comprobar que las funciones que utilizamos devuelven los valores correctos (es decir, un elemento más en el proceso de validación de código).

En este caso la función que nos interesa parece ser *boxplot*. No obstante, como vemos en la figura, el gráfico presumiblemente contiene los valores numéricos correctos. Se trata ahora de **mejorarlo**, usando los avances propios de la función que permiten los etiquetados de ejes u otras funcionalidades que hagan más legible nuestra figura. Para esto, debemos seguir el tercer paso recomendado a los principiantes, **visitar la función en la propia ayuda** del programa y seguir las instrucciones para mejorar la salida y/o presentación de resultados.

En este caso, por ejemplo, bastaría con añadir etiquetados de ejes:

```
>NotaAlumnos <- c(10,10,10,10,10,0,0,0,0,0)  
>boxplot(NotaAlumnos, ylab="Notas")
```



3.8 Datos atípicos y análisis exploratorio de datos

Los llamados **valores atípicos o extremos** (*outliers*) son aquellos valores que distan de la mayoría de los datos. El diagrama de caja y bigotes de hecho los marca con puntitos los simplemente atípicos (ver gráfico anterior) y con estrellitas los extremos.

Para establecer la diferencia entre unos y otros asume como límite la diferencia 3 veces el Rango IQ. De este modo entre 1,5 veces el Rango IQ y 3 veces el Rango IQ los datos se marcarán como simples atípicos, y a partir de 3 veces el Rango IQ en adelante serán marcados como valores extremos.

En realidad esto es una división (entre «simples atípicos» y «extremos») que hace el paquete SPSS pero lo normal es considerarlos todos en el mismo saco que es el de datos atípicos o extremos según los llamemos.

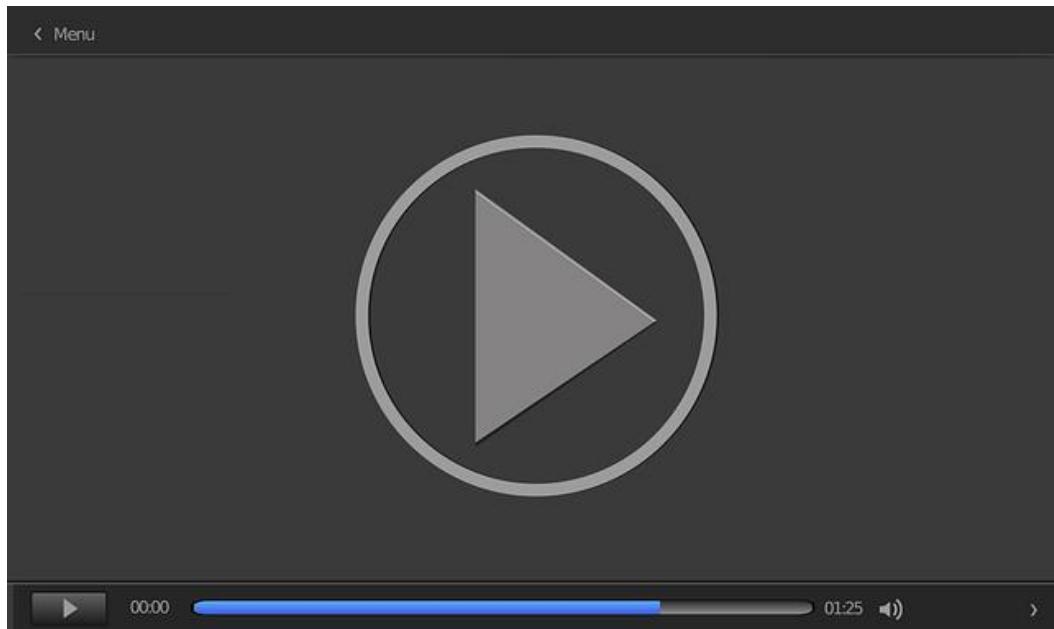
Aspectos que conviene saber de los valores atípicos:

- ▶ Afectan drásticamente a la media, pero por el contrario la mediana apenas se ve afectada por ellos.
- ▶ También se ve muy afectada la dispersión media que mide la desviación típica.
- ▶ Puede alterar la forma de la distribución de datos, especialmente en el histograma.

Por ello, cuando realizamos lo que se llama **análisis exploratorio de datos** cobra especial importancia la identificación y tratamiento de los *outliers* o valores atípicos. El análisis exploratorio de datos es el estudio de las características principales de un conjunto de datos sirviéndose de las medidas estadísticas (tendencia central, dispersión, posición, etc.) y de gráficas que ayuden a identificarlas.

Análisis exploratorio de datos con R

En este vídeo vamos a trabajar con R en el análisis de datos.



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=dffa76a4-137e-4893-b2bb-acbd00b21cc9>

Practica con R los conceptos estudiados

Asumiendo que tienes instalado R y RStudio, abre el IDE de RStudio y en un «R script» nuevo «Ctrl + Shift + N» escribe el siguiente código.

```
rm(list=ls())
#####
requiredPackages <- c("arsenal", "car", "chemometrics", "corrplot",
"gapminder", "dplyr", "DescTools", "foreign", "e1071", "expss", "GGally",
"ggplot2", "haven", "knitr", "plotly", "psych", "remotes",
"summarytools", "ggridges", "table1", "tableone", "tidyverse", "SmartEDA")

sesion1 <- function(pkg){
new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
if (length(new.pkg))
install.packages(new.pkg, dependencies = TRUE)
sapply(pkg, require, character.only = TRUE)
}

sesion1(requiredPackages)
#####
##LOAD DATA
#Data Lending Club -https://www.kaggle.com/wordsforthewise/lending-club.
Factores que determinan el Default en los créditos. Modelo de riesgo
Datalc<-
read.csv("https://raw.githubusercontent.com/millerjanny/Custom_UNIR/main/Data_LendingClub.csv")

Datalc$Default=recode_factor(Datalc$Default, '1' = "Default", '0' = "Non-default")
# Exploración inicial
#####
# Sample mean
mean(Datalc$dti_n)
mean(Datalc$dti_n,trim=0.05)
winsor.mean(Datalc$dti_n, trim = 0.05)
mean(filter(Datalc, Datalc$Default == "Default")$dti_n)
mean(filter(Datalc, Datalc$Default == "Non-default")$dti_n)

# Dispersion
var(Datalc$annual_inc)
sd(Datalc$annual_inc)
var(Datalc$dti_n)
sd(Datalc$dti_n)
sd_trim(Datalc$dti_n,trim=0.05)
winsor.sd(Datalc$dti_n, trim = 0.05)

# Coefficient of variation
```

```
sd(Datalc$annual_inc)/mean(Datalc$annual_inc)
sd(Datalc$dti_n)/mean(Datalc$dti_n)

# Resumen de todas las variables
summary(Datalc$dti_n)
summary(Datalc)

# descriptivo de variables cuantitativas by default
by(select(Datalc, dti_n), factor(Datalc$Default), summary)
(by(select(Datalc, annual_inc, loan_amnt, int_rate, fico_n,
dti_n), factor(Datalc$Default), summary))

# descriptivo de variables cuantitativas by default (mean, sd)
(cuanti_summary<-Datalc %>% tab_cols(total(label = "Total"), Default) %>%
tab_cells(annual_inc, dti_n, loan_amnt, int_rate, fico_n)
%>%tab_stat_fun(Mean = w_mean, "Std. dev." = w_sd, "Valid N" = w_n, method =
list) %>%
tab_pivot%>% tab_caption("Resumen de variables cuantitativas"))
```

Prueba a ejecutar el *script* siguiendo estas indicaciones:

- ▶ Ejecuta cada línea de código, posiciona el cursor en la primera línea y utiliza la opción «Run» o «Ctrl + Enter».
- ▶ Observa la «Consola» y «Environment» cuando ejecutas cada línea.
- ▶ Comprende por qué las líneas que empiezan con # no se ejecutan.
- ▶ Encuentra la definición de cada función de R.
- ▶ Comparte aquellas funciones que no conozcas en el foro de la asignatura.
- ▶ Repasa con R todos los conceptos vistos hasta ahora.

Practica la creación de gráficos con R

Asumiendo que tienes [instalado R](#) y [RStudio](#), abre el IDE de RStudio y en un «R script» nuevo «Ctrl + Shift + N» escribe el siguiente código.

```
rm(list=ls())
#####
requiredPackages <- c("arsenal", "car", "chemometrics", "corrplot",
"gapminder", "dplyr", "DescTools", "foreign", "e1071", "expss", "GGally",
"ggplot2", "haven", "knitr", "plotly", "psych", "remotes",
"summarytools", "ggridges", "table1", "tableone", "tidyverse", "SmartEDA")

sesion1 <- function(pkg){
new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
if (length(new.pkg))
install.packages(new.pkg, dependencies = TRUE)
sapply(pkg, require, character.only = TRUE)
}

sesion1(requiredPackages)
#####
##LOAD DATA
#Data Lending Club -https://www.kaggle.com/wordsforthewise/lending-club.
Factores que determinan el Default en los créditos. Modelo de riesgo
Datalc<-
read.csv("https://raw.githubusercontent.com/millerjanny/Custom_UNIR/main/Data_LendingClub.csv")

Datalc$Default=recode_factor(Datalc$Default, '1' = "Default", '0' = "Non-default")
# Exploración inicial
#####
#####Tablas de frecuencia de variables categóricas

# Frecuencias tablas cualitativas
table(Datalc$purpose, Datalc$Default)

#Totales filas y columnas
margin.table(table(Datalc$purpose, Datalc$Default), margin = 2)
margin.table(table(Datalc$purpose, Datalc$Default), margin = 1)
addmargins(table(Datalc$purpose, Datalc$Default))

## relativo total
prop.table(table(Datalc$purpose, Datalc$Default))

## relativo fila
prop.table(table(Datalc$purpose, Datalc$Default), margin=1)
```

```

## relativo columna
prop.table(table(Datalc$purpose, Datalc$Default), margin=2)

#Resumen de todas las variables cualitativas
(scu=by(select(Datalc,
emp_length,home_ownership_n,purpose),factor(Datalc$Default),summary))

(cuali_summary<-Datalc %>% tab_cols(Default) %>%
tab_cells(emp_length,home_ownership_n,purpose, total()) %>% tab_stat_rpct() %>%
tab_pivot%>% tab_caption("Resumen de variables cualitativas"))
#####
#algunos gráficos
ggplot(Datalc, aes(x = term)) +
geom_bar(position = "dodge") + #position = "dodge", to have a side-by-side (i.e. not stacked) barchart
theme_bw()

ggplot(Datalc, aes(x = term, fill = Default)) +
geom_bar(position = "dodge") + #position = "dodge", to have a side-by-side (i.e. not stacked) barchart
theme_bw()

Datalc %>%
count(term = factor(term), Default = factor(Default)) %>%
mutate(pct = prop.table(n)) %>%
ggplot(aes(x = term, y = pct, fill = Default, label = scales::percent(pct))) +
geom_col(position = 'dodge') +
geom_text(position = position_dodge(width = .9), # move to center of bars vjust = -0.5, # nudge above top of bar size = 3) +
scale_y_continuous(labels = scales::percent)

ggplot(Datalc, aes(x= Default, group=term)) +
geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count") +
geom_text(aes( label = scales::percent(..prop..), y= ..prop.. ), stat= "count", vjust = -.5) +
labs(y = "Percent", fill="Default") +
facet_grid(~term) +
scale_y_continuous(labels = scales::percent)

#####Medidas de localización###
# Resumen de todas las variables
summary(Datalc$dti_n)

# Sample median
median(Datalc$dti_n)

```

```

# First quartile
(Q1=quantile(Datalc$dti_n,probs = 0.25))

# Second quartile (= Median)
quantile(Datalc$dti_n,probs = 0.50)

# Third quartile
(Q3=quantile(Datalc$dti_n,probs = 0.75))

# quantiles
quantile(Datalc$dti_n,probs=c(0.25,0.5,0.75))

# Interquartile range
(IQR = Q3 - Q1)

# Lower and upper limits
Q1-1.5*IQR
Q3+1.5*IQR

# Boxplots

boxplot(Datalc$int_rate)
Q1 = quantile(Datalc$int_rate,probs = 0.25)
Q2 = quantile(Datalc$int_rate,probs = 0.50)
Q3 = quantile(Datalc$int_rate,probs = 0.75)
Q1-1.5*(Q3-Q1)
Q3+1.5*(Q3-Q1)

#Quitar Outliers

Q1 <- quantile(Datalc$int_rate, .25)
Q3 <- quantile(Datalc$int_rate, .75)
IQR <- IQR(Datalc$int_rate)
no_outliers_int_rate <- subset(Datalc, Datalc$int_rate> (Q1 - 1.5*IQR) &
Datalc$int_rate< (Q3 + 1.5*IQR))
dim(Datalc)
dim(no_outliers_int_rate)
boxplot(Datalc$int_rate)
boxplot(no_outliers_int_rate$int_rate)
#####
# Box-plot discriminando por variable categórica agregando la media
Datalc%>%
ggplot(aes(Default,int_rate, fill=Default)) +
geom_boxplot() +
stat_summary(fun.y="mean")+
theme(legend.position = "none")

#####Medidas de forma

```

```
skewness(Datalc$dti_n)
skewness(Datalc$annual_inc)
skewness(Datalc$int_rate)
kurtosis(Datalc$annual_inc, type = 1)
#https://search.r-project.org/CRAN/refmans/datawizard/html/skewness.html
####Tipificación de variables
Datalc$int_rate_z <- (Datalc$int_rate - mean(Datalc$int_rate)) /
sd(Datalc$int_rate)
df_z>Datalc %>% mutate(across(where(is.numeric), scale))
```

Prueba a ejecutar el *script* siguiendo estas indicaciones:

- ▶ Ejecuta cada línea de código, posiciona el cursor en la primera línea y utiliza la opción «Run» o «Ctrl + Enter».
- ▶ Observa la «Consola» y «Environment» cuando ejecutas cada línea.
- ▶ Comprende qué hacen las funciones de R.
- ▶ Repasa con R todos los conceptos vistos hasta ahora.

Practica la exploración de datos y las técnicas de imputación con ayuda de R

Asumiendo que tienes [instalado R](#) y [RStudio](#), abre el IDE de RStudio y en un «R script» nuevo «Ctrl + Shift + N» escribe el siguiente código.

```
rm(list=ls())
#####
requiredPackages <- c("arsenal", "car", "chemometrics", "corrplot",
"gapminder", "dplyr", "DescTools", "foreign", "e1071", "expss", "GGally",
"ggplot2", "haven", "knitr", "plotly", "psych", "remotes",
"summarytools", "ggridges", "table1", "tableone", "tidyverse", "SmartEDA",
"scales", "caret", "imputeMissing", "mice")
sesion1 <- function(pkg){
new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
if (length(new.pkg))
install.packages(new.pkg, dependencies = TRUE)
sapply(pkg, require, character.only = TRUE)
}
sesion1(requiredPackages)
#####
##LOAD DATA
#Data Lending Club -https://www.kaggle.com/wordsforthewise/lending-club.
Factores que determinan el Default en los créditos. Modelo de riesgo
Datalc<-
read.csv("https://raw.githubusercontent.com/millerjanny/Custom_UNIR/main/Data_LendingClub.csv")

Datalc$Default=recode_factor(Datalc$Default, '1' = "Default", '0' = "Non-
default")
# Exploración inicial
#####
####Identificación y tratamiento de outliers#####
#####Con boxplot identificar y eliminar atípicos#####
Q1 <- quantile(Datalc$int_rate, .25)
Q3 <- quantile(Datalc$int_rate, .75)
Q2 <- quantile(Datalc$int_rate, .5)
Q1 - 1.5*IQR
Q3 + 1.5*IQR
IQR <- IQR(Datalc$int_rate)
boxplot(Datalc$int_rate)$stats
no_outliers_int_rate <- subset(Datalc, Datalc$int_rate > (Q1 - 1.5*IQR) &
Datalc$int_rate < (Q3 + 1.5*IQR))

#otra manera
boxplot(Datalc$int_rate)
```

```

out<-boxplot(Datalc$int_rate)$out
no_outliers_int_rate_1<-Datalc[-which(Datalc$int_rate %in% out),]

#####identificar y truncar atípicos al P5 y P95#####
Datalc$int_rate_p5_95<- squish(Datalc$int_rate, quantile(Datalc$int_rate,
c(.05, .95)))
boxplot(Datalc$int_rate,Datalc$int_rate_p5_95)

#####limitando a un num sd#####
mean=mean(Datalc$int_rate)
std=sd(Datalc$int_rate)
Datalc$int_rate_3sd<-squish(Datalc$int_rate,c(mean-(3*std),mean+(3*std)))

#####
#####identificar e imputar NANs#####
#####
anyNA(Datalc)
sapply(Datalc, function(x) anyNA(x))
sapply(Datalc, function(x) sum(is.na(x)))
#####
#Adicional, no está en el temario
###
#mtcars, iris, ..https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html
data=airquality
anyNA(data)
sapply(data, function(x) anyNA(x))
sapply(data, function(x) sum(is.na(x)))
str(data)

#Imputar con median/mode(imputeMissings). Método rápido pero trata a cada
variable independientemente
DataIpmputed <- impute(data, method = "median/mode")
sapply(DataIpmputed , function(x) sum(is.na(x)))

#Imputar con KNN####(caret)
impknn <- preProcess(data, method = "knnImpute", k = 5)
data_knn <- predict(impknn, data)
sapply(data_knn, function(x) sum(is.na(x)))

#Imputar con Bagging (Bootstrap aggregating)
impbag <- preProcess(data, method = "bagImpute")
data_bag <- predict(impbag, data)
sapply(data_bag, function(x) sum(is.na(x)))

# Imputar con MICE (mice)...cart,rf,mean,https://cran.r-project.org/web/packages/mice/mice.pdf

```

```
#Multivariate Imputation via Chained Equations  
mice <- mice(data, method="cart")  
data_mice <- complete(mice)#Creates imputed data  
sapply(data_mice, function(x) sum(is.na(x)))
```

Recuerda ejecutar línea a línea el *script* anterior. Para ello, posiciona el cursor en la primera línea y utiliza la opción «Run» o «Ctrl + Enter». Analiza lo que muestra la «Consola» y el «Environment» conforme vayas ejecutando cada línea.

3.9. Referencias bibliográficas

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

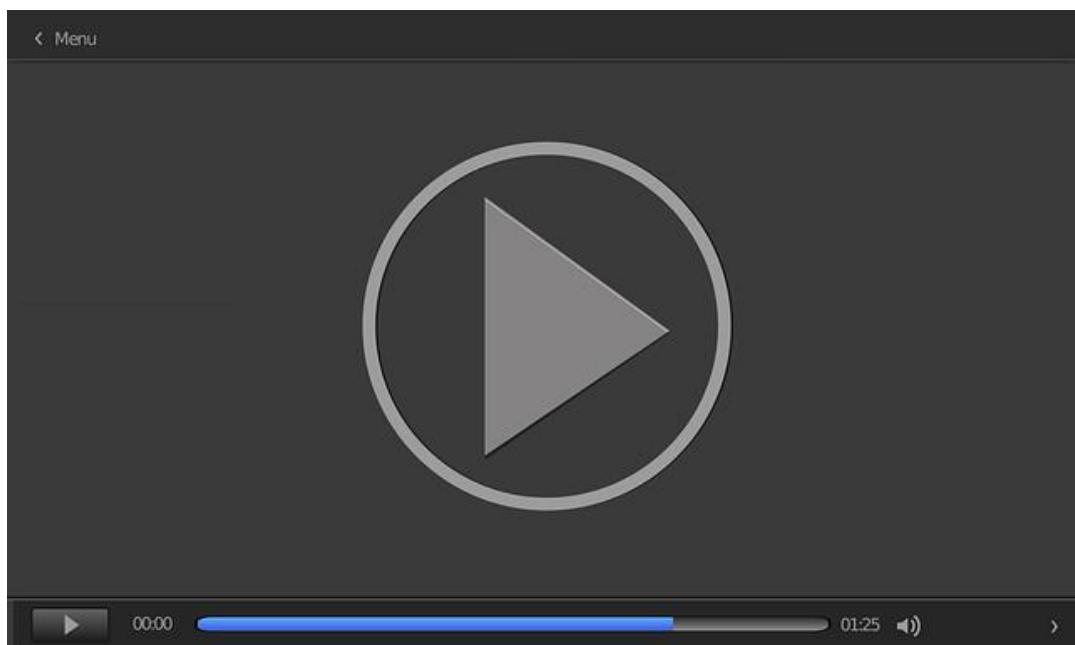
Versión electrónica:

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

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

Medidas de Tendencia Central con Excel

En este vídeo os ilustro sobre las operaciones y pasos necesarios para estudiar las medidas de tendencia central con Excel, haciendo énfasis en la media aritmética y la desviación típica y en cómo se complementan.



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=d2d7f1d6-e6a8-4a25-a59d-abdc00f2aaa1>

Medidas estadísticas

Triola, M. F. (2009). *Estadística* (10^a ed., pp. 74-136). México D.F.: Pearson Educación.

Es recomendable que le eches un vistazo al libro *Estadística* de M.F. Triola, cuyo primer tema contiene prácticamente todo lo que se trata en este tema.

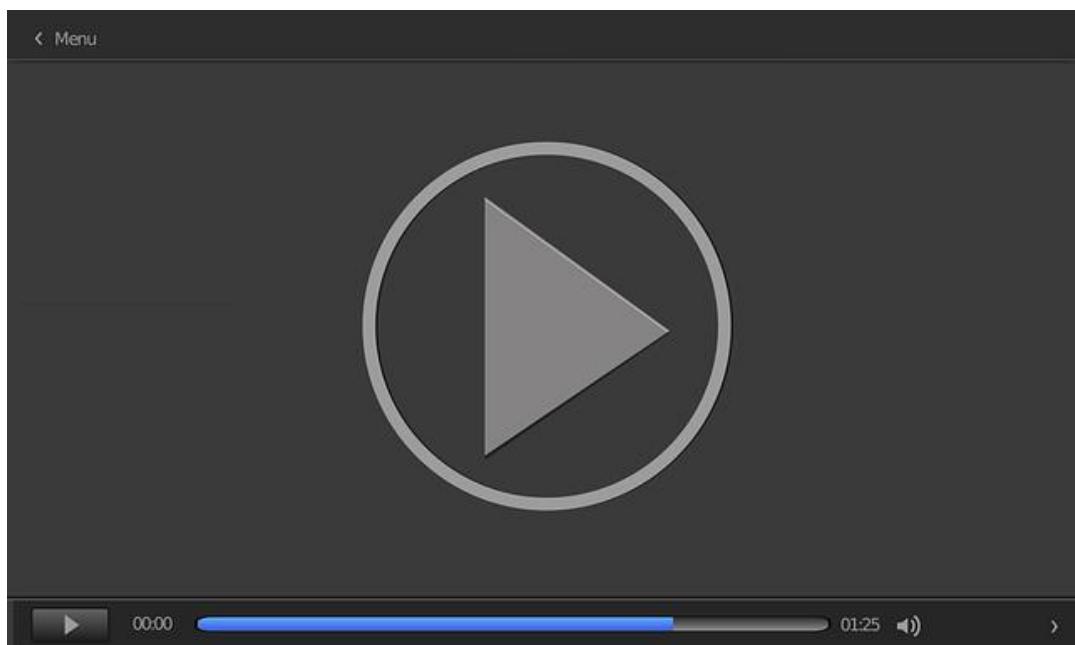
Estadísticas aplicadas al deporte

Como curiosidad, se recomienda el artículo *Estadísticas aplicadas al deporte*, de Paul Shirley publicado en El País.

Accede al artículo desde el aula virtual o a través de la siguiente dirección web: http://deportes.elpais.com/deportes/2014/02/09/actualidad/1391968023_590198.html

Construir un diagrama de caja y bigotes en Excel

Interesante vídeo de Mr. Reive (en inglés, aunque podrás encontrar otros similares en español) sobre cómo construir en Excel un diagrama de caja y bigotes realizando las modificaciones oportunas, ya que por defecto no las facilita.



Accede al vídeo:

<https://www.youtube.com/embed/ZFbPnwKwVWk>

Estadística y probabilidad

También se recomienda que visites esta web donde están alojados mini-vídeos sobre las medidas estadísticas y otros temas de estadística y probabilidad. Se trata de un proyecto abierto realizado por la Universidad Carlos III de Madrid. También tiene la posibilidad de consulta de los materiales en formato pdf.

Accede a la página desde el aula virtual o a través de la siguiente dirección web:

<http://163.117.132.198/minivideos/>

Bibliografía

Ríus, 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* (10^a ed.). México D.F.: Pearson Educación.

- 1.** ¿Cuántos cuartiles hay en una distribución de datos?

 - A. Paradójicamente hay dos, puesto que son tres pero como la mediana es el segundo se quedan en dos.
 - B. 4.
 - C. 3.
 - D. Depende si el conjunto de datos presenta frecuencias repetidas.

- 2.** ¿Qué cuantiles equivalen a la mediana?

 - A. El quinto decil.
 - B. El segundo percentil.
 - C. El segundo cuartil.
 - D. Las respuestas A y C son correctas.

- 3.** La mediana...

 - A. Es el valor central pero solo si el conjunto de datos es par.
 - B. Es el valor central pero solo si el conjunto de datos es impar.
 - C. Es el valor central siempre.
 - D. Depende si el conjunto de datos presenta frecuencias repetidas.

- 4.** La media...

 - A. Se ve afectada drásticamente por los valores extremos.
 - B. Es una medida con una representatividad mayor que la mediana.
 - C. Es más útil que la mediana para las variables cualitativas.
 - D. No es útil ni calculable para las variables cualitativas.
 - E. Las respuestas A y D son correctas.

5. La medida estadística que menos se ve afectada por los valores atípicos es:
- La desviación estadística.
 - La mediana.
 - La media aritmética.
 - La media armónica.
6. En la fórmula de la mediana para datos agrupados: ¿Qué representan las letras y símbolos?

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i} \times a_i$$

- L_{i-1} es el límite inferior del intervalo mediano.
 - $\frac{N}{2}$ corresponde con la posición que debería ocupar la mediana dentro del conjunto de datos.
 - a_i es la altura de cada intervalo.
 - L_{i-1} es el límite inferior del intervalo anterior al mediano.
 - Las repuestas A y B son correctas.
7. La varianza...:
- Es parecida a la desviación típica.
 - Aporta la misma información sobre la dispersión que la desviación típica.
 - junto con la desviación típica y la desviación estándar conforman las medidas de dispersión más conocidas.
 - Es el cuadrado de la desviación típica.
 - Las repuestas B y D son correctas.

- 8.** El diagrama de cajas se construye con:
- A. Cuatro valores: La mediana, el cuartil 1, el cuartil 3 y la desviación típica.
 - B. Cuatro valores: La mediana, el cuartil 1, el cuartil 3 y la varianza.
 - C. Cinco valores: La mediana, el cuartil 1, el cuartil 3, el mínimo y el máximo.
 - D. Los cinco valores de C. más los valores atípicos sin los cuales no se puede construir.
- 9.** Una medida estadística que nos permite comparar entre diferentes poblaciones es:
- A. El coeficiente de variación.
 - B. La desviación estándar.
 - C. La puntuación tipificada.
 - D. Las respuestas A y C son correctas.
- 10.** En cuanto a la asimetría...
- A. Es positiva cuando la cola está a la derecha y la Moda es mayor que la media.
 - B. Es negativa cuando la cola está a la izquierda y la Moda es mayor que la media.
 - C. Es negativa cuando la cola está a la derecha y la Moda es menor que la media.
 - D. Es positiva cuando la cola está a la derecha y la Moda es menor que la media.
 - E. Las respuestas B y D son correctas.

Análisis e Interpretación de Datos

Tema 4. Regresión y correlación

Índice

Esquema

Ideas clave

- 4.1. ¿Cómo estudiar este tema?
- 4.2. Introducción
- 4.3. Correlación
- 4.4. Regresión lineal
- 4.5. Gráfico de residuos
- 4.6. Regresión lineal multivariante
- 4.7 Regresión no lineal
- 4.8 LTS (Least Trimmed Squares)
- 4.9. Referencias bibliográficas

A fondo

Detectando puntos influyentes en nuestro modelo de regresión

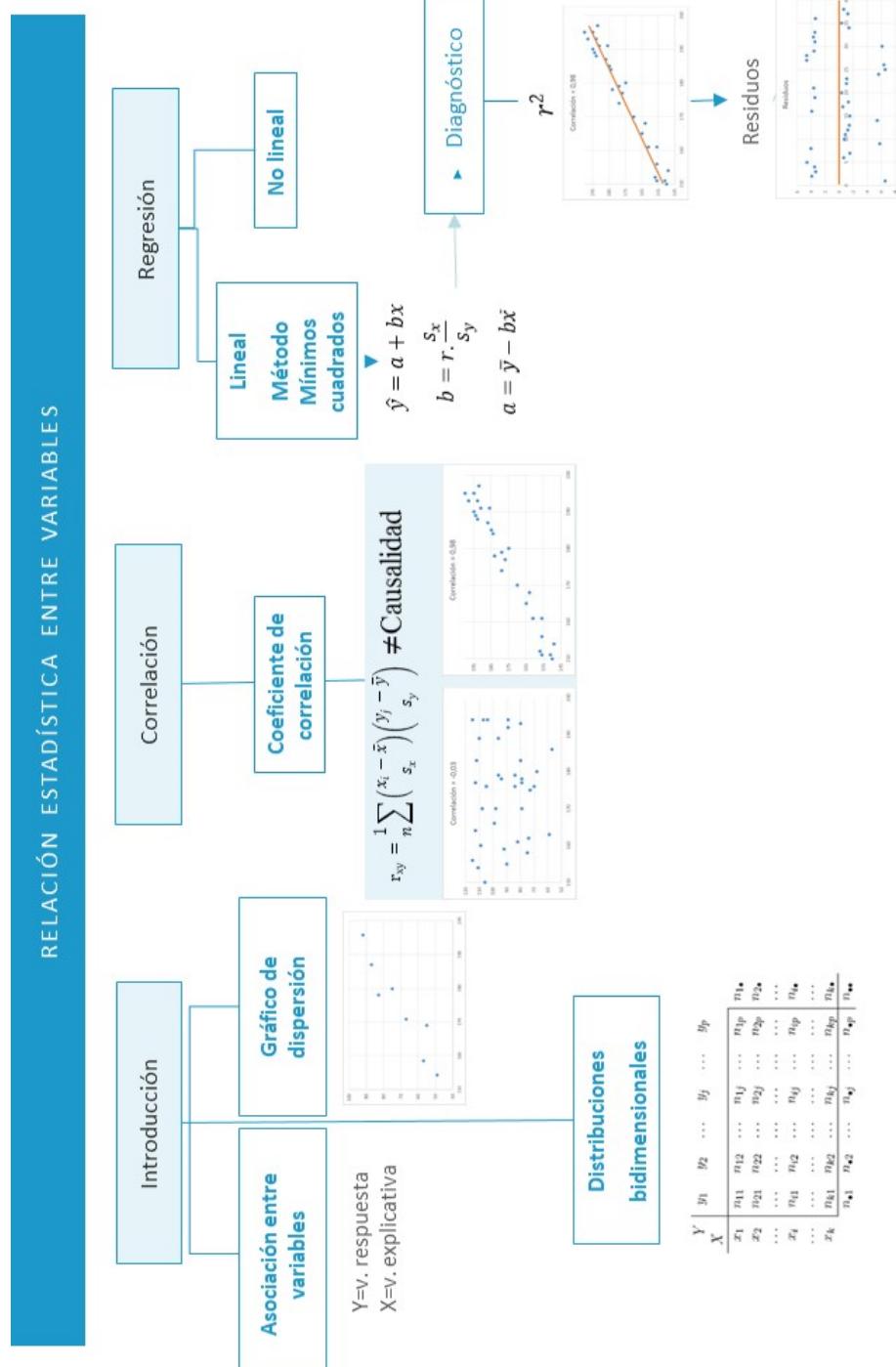
Tratamiento estadístico bidimensional

Método de mínimos cuadrados

Applets sobre correlación y regresión

Bibliografía

Test



4.1. ¿Cómo estudiar este tema?

Para estudiar este tema lee las **páginas 77-93** del siguiente libro:

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

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

Este tema trata sobre la relación estadística entre variables. Para hacerse una idea global es importante que revises el esquema inicial, el cual te ayudará a hacerte una buena idea de cómo está estructurado el tema.

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 relación estadística entre variables y algunos conceptos asociados también muy importantes. Del mismo modo presta atención a los ejemplos pues encierran muchas de las claves que te facilitarán la comprensión del tema.

4.2. Introducción

Es habitual querer estudiar la relación estadística entre dos variables. Para ello ya hemos visto alguna herramienta gráfica como el diagrama de dispersión. Para resumir la información en tablas de frecuencias bidimensionales empleamos un formato en el cual situamos cada variable en un eje y las frecuencias en el interior para cada combinación de pares de categorías de ambas variables.

X	y_1	y_2	\dots	y_j	\dots	y_p	
	n_{11}	n_{12}	\dots	n_{1j}	\dots	n_{1p}	$n_{1\bullet}$
x_1	n_{21}	n_{22}	\dots	n_{2j}	\dots	n_{2p}	$n_{2\bullet}$
\dots	\dots	\dots	\dots	\dots	\dots	\dots	\dots
x_i	n_{i1}	n_{i2}	\dots	n_{ij}	\dots	n_{ip}	$n_{i\bullet}$
\dots	\dots	\dots	\dots	\dots	\dots	\dots	\dots
x_k	n_{k1}	n_{k2}	\dots	n_{kj}	\dots	n_{kp}	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	\dots	$n_{\bullet j}$	\dots	$n_{\bullet p}$	$n_{\bullet \bullet}$

Tabla 1: La tabla de frecuencias bidimensional.

Donde los puntos nos sirven para indicar que se trata de **frecuencias marginales**, es decir, frecuencias de la distribución unidimensional de la variable que es la que resulta de no tener en cuenta la otra variable (de ahí lo de «marginal»).

Del mismo modo que para las variables unidimensionales, en el caso bidimensional tenemos estadísticos, pero en este caso no van a medir las propiedades individuales sino las conjuntas de ambas variables en el conjunto de datos bivariados.

El estadístico más importante en este caso es la **covarianza** que mide lo que covarian las dos variables. Técnicamente la definimos como la media aritmética de los productos de las desviaciones de cada variable respecto a su media.

$$S_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Donde los n_{ij} son las frecuencias que corresponden a la modalidad i de x y la modalidad j de y . La covarianza nos será muy útil para medir la fuerza de la relación entre las variables aunque a través de otro estadístico llamado coeficiente de correlación, pues del mismo modo que ocurría con la varianza, el hecho que la covarianza tenga un valor concreto no nos dice mucho sobre si es alta o no ya que ésta expresa una magnitud en términos absolutos y esto depende de la escala en la que se encuentre la variable.

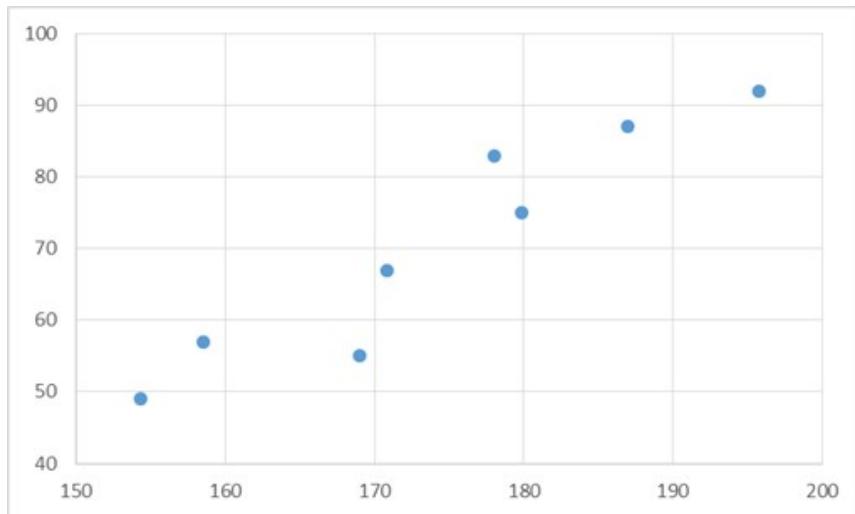
A diferencia de la varianza, la covarianza puede resultar negativa, pues al no haber cuadrado en la fórmula no será positiva necesariamente como lo es la varianza. Cuando esto ocurre será porque ambas variables actúan inversamente en el sentido de que a valores por encima de la media de una de ellas le corresponden valores por debajo de la media en la otra.

Del mismo modo que con la varianza, también contamos con una fórmula que nos facilita el cálculo:

$$S_{xy} = \frac{\sum x_i y_i}{N} - \bar{x}\bar{y}$$

Según refleja esta fórmula la covarianza es igual a la media de los productos menos el producto de las medias.

En cuanto a herramientas gráficas para representar las distribuciones conjuntas de dos variables lo idóneo es emplear gráficos de dispersión que nos muestran las nubes de puntos. Si bien, lo hemos visto anteriormente, ahora profundizaremos un poco más en esta herramienta gráfica, absolutamente fundamental para el estudio de la relación entre variables.



Gráfica 1: Ejemplo de gráfico de dispersión, una herramienta fundamental en el análisis de regresión.

Ejemplo 1: Vamos a confeccionar una tabla de frecuencias bidimensional y a calcular su covarianza, finalmente representaré la nube de puntos para que os hagáis una buena idea de la relación entre ambas.

ALTURA (X)	Peso (Y)	ALTURA (X)	Peso (Y)
171	110,5	183	66,5
158	79	155	77,5
185	67,5	192	71
172	111	159	79,5
167	108,5	189	94,5
164	107	182	91
178	89	161	105,5
166	58	195	97,5
177	63,5	180	90
184	92	162	56
188	119	176	88
178	64	175	87,5
161	55,5	191	70,5
187	93,5	155	77,5

Tabla 2: Un ejemplo de distribución de dos variables: peso y altura.

Para confeccionar la tabla de frecuencias vamos a agrupar los datos. Esto persigue crear categorías en las cuales ya tengamos frecuencias mayores que 1 y nos permitirá poner en práctica la confección de dichas tablas.

El primer paso de cara a agrupar los datos es ver cuántos intervalos establecemos y el tamaño de cada uno. Para ello calculamos los mínimos y máximos a través de su rango (recordemos que el $rango = valor_{max} - valor_{min}$).

	ALTURA	PESO
Min	155	55,5
Max	195	119

Con un rango de 40cm la variable altura la podemos agrupar en 4 intervalos de 10cm cada uno, mientras que la variable rango de peso = $119 - 55,5 = 63,5$ Kg. Por lo que podemos establecer por ejemplo 5 intervalos desde 55 a 120, lo cual hace que sean de 13 kg cada uno si todos tienen la misma amplitud.

A nivel teórico se han formulado ecuaciones para el cálculo de los intervalos, pero en la práctica resulta eficiente hacerlo empleando un buen criterio a partir del rango y con la idea de no manejar numerosos intervalos y que todos ellos tengan una frecuencia mínima aceptable.

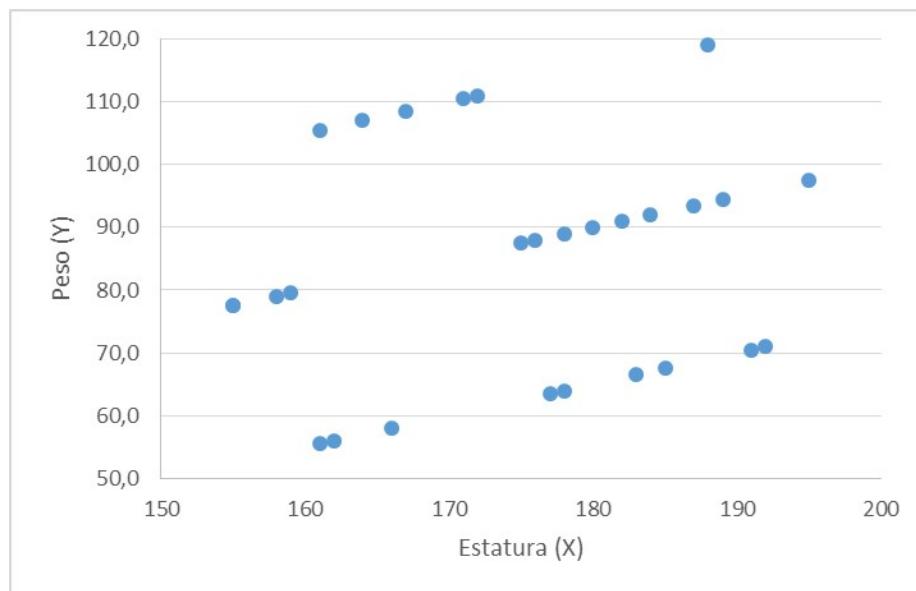
Así, la tabla bidimensional de frecuencias queda como sigue.

Altura (X)	Peso (Y)					Totales Altura
	55-68	68-81	81-94	94-107	107-120	
155-165	2	4	1	0	1	8
165-175	1	0	0	0	3	4
175-185	3	6	0	0	0	9
185-195	1	2	1	2	1	7
Totales Peso	7	12	2	2	5	28

Tabla 3: Un ejemplo de tabla de frecuencias bidimensional.

La fila donde figura «Totales peso» y la columna «Totales Altura» son las frecuencias marginales de la X e Y respectivamente.

El gráfico de dispersión correspondiente a este conjunto de datos resulta el siguiente:



Sin necesidad de avanzar muy deprisa en el tema, ya podemos adelantar a la vista del gráfico dos cosas:

- ▶ Lo primero sería que los datos podrían estar proviniendo de tres poblaciones distintas pues su comportamiento es claramente diferente situándose las observaciones sobre tres rectas.
- ▶ Lo segundo es que estas funciones son tan perfectamente lineales que parece matemática la relación entre estas variables. De hecho los datos en los que se basa el ejemplo han sido confeccionados por simulación imponiendo que el peso se generara a partir de una combinación lineal de la estatura.

Para **examinar convenientemente un diagrama de dispersión** y extraer toda la información valiosa conviene fijarse en:

- ▶ Los patrones que se manifiestan (en el caso del ejemplo anterior claramente tres líneas).
- ▶ La forma, dirección y fuerza de tal patrón (en el caso anterior la pendiente era positiva con una inclinación no muy alta...).
- ▶ Fijarse en los individuos que se alejan del patrón de la mayoría: los valores atípicos (en el caso anterior no parece haber ninguno).
- ▶ También podemos identificar los grupos aislados, que en estadística son llamados **clúster** (podríamos considerar a cada uno de los tres grupos como un conglomerado).

Un concepto que conviene aclarar desde el principio es el de **asociación entre variables**. Diremos que esta ocurre cuando determinados valores de una de ellas sean más propensos a su aparición según que valores tome la otra; es decir, que los valores de ambas variables están relacionados, sin que por esto tenga que haber necesariamente **causalidad**, sino simplemente asociación que podría estar dándose de forma indirecta.

Un ejemplo de esto bastante clásico es la asociación entre tener un mechero y padecer cáncer. Obviamente el tener el mechero no puede provocar cáncer pero sí que indirectamente puede afectar, pues si tenemos un mechero en el bolsillo es porque fumamos y el hecho de fumar sí que ha sido demostrado que es un factor importante para la aparición de cáncer.

Una pregunta que tenemos que hacernos de cara a estudiar la relación que existe entre dos o más variables es la siguiente: **¿Qué motivación tenemos para el estudio de la asociación entre variables?** Las respuestas principales son dos:

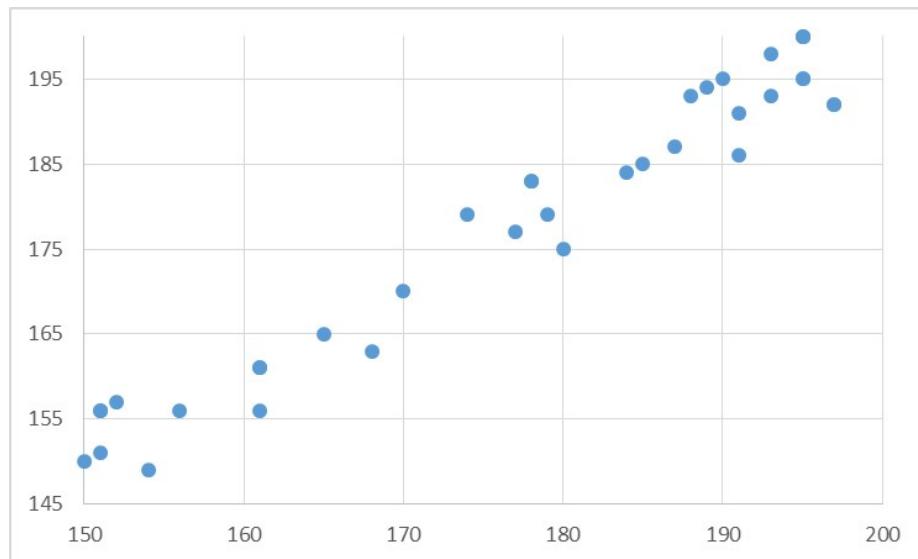
- ▶ **Estudiamos la asociación para describirla**, para conocer más sobre la manera en que las variables se relacionan (o correlacionan).
- ▶ También estudiamos la asociación **con idea de predecir o explicar una variable a partir de la otra**.

Así suele haber una variable que queremos conocer a partir de otra, bien porque podemos medir una más fácilmente que la otra o para sentar simplemente las bases de su asociación. Desde este punto de vista tenemos por tanto una **variable explicativa** (o predictora) y otra que es la **variable respuesta**, o bien la **variable independiente y la dependiente**. Siendo la variable dependiente la que pretendemos explicar a partir de la independiente. Ambas dualidades se pueden considerar a mi modo de ver equivalentes, aunque en algunos textos suelen diferenciarlas o preferir una a otra.

Ejemplo 2: En España la tasa de alcohol permitida es 0,03 g de alcohol por cada 100ml de sangre. Así que, por ejemplo, el número de cervezas que nos tomamos será proporcional al volumen de alcohol que generará en sangre. El estudio de esta relación es claramente estadístico, ya que tal proporción no es perfecta. No todos tenemos la misma capacidad de asimilación del alcohol. Entre estas dos variables la variable dependiente o respuesta sería volumen de alcohol en sangre y la independiente o explicativa sería el número de cervezas consumidas (o los litros de cerveza, según la midamos).

4.3. Correlación

Una de las relaciones más comunes entre dos variables es la lineal, quizás debido a su sencillez. Podemos fijarnos en los gráficos de dispersión en busca del patrón lineal. ¿Pero cómo dictaminar si la magnitud de este supuesto patrón es importante? Tenemos el gráfico que sigue, así a ojo de buen cubero diríamos que alta.



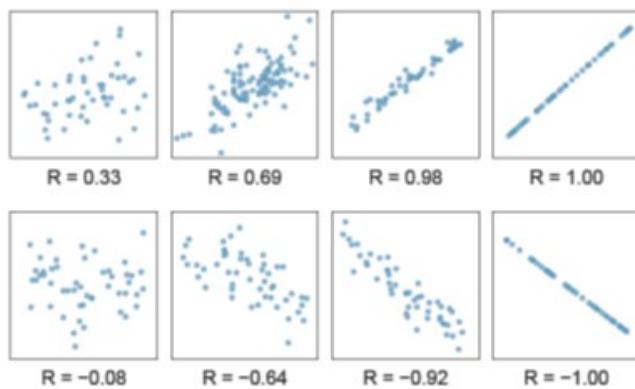
Gráfica 3: ¿Esta asociación lineal es alta?

Pero quizás eso no nos valga, así que disponemos de un estadístico que nos va a medir tal magnitud.

El **coeficiente de correlación de Pearson** es una medida de la fuerza de la relación lineal entre dos variables cuantitativas. Su fórmula es:

$$r_{xy} = \frac{1}{n} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = \frac{S_{xy}}{S_x S_y}$$

Ejemplos de coeficientes de correlación acompañados de su dispersión son los siguientes:



Gráfica 4: Diferentes magnitudes de correlación.

El coeficiente de correlación tiene las siguientes propiedades determinadas por su fórmula:

- ▶ Está limitado entre -1 y 1.

$$-1 \leq r \leq 1$$

- ▶ No importa si es calculado para x sobre y o para y sobre x , es decir, no distingue que una sea la dependiente o que lo sea la otra variable. La magnitud de la relación es por tanto la misma independientemente de la dirección de asociación que establezcamos.
- ▶ Cuando $r = 0$ indica que no existe asociación lineal.
- ▶ Cuando al contrario es -1 o +1 la relación es máxima y se dice que es «**perfecta**».

Conviene puntualizar un aspecto. La correlación mide la intensidad de la relación lineal pero no dice nada de otras relaciones no lineales que puedan existir. De esta manera si la correlación es nula, podría estar existiendo otro tipo de correlación.

Asimismo, tal y como sucedía con la media y la desviación típica la correlación es muy sensible a la presencia de *outliers*.

4.4. Regresión lineal

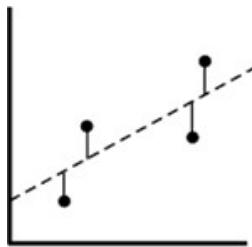
Del mismo modo que queríamos saber la intensidad de la relación lineal nos va a interesar conocer la forma explícita de tal relación lineal, es decir, la ecuación lineal que describe mejor la asociación encontrada. El estudio de esta cuestión se trata en lo que se conoce como **regresión lineal**.

El conocer la ecuación de regresión nos va a permitir calcular **predicciones** de una de las variables a partir de los valores que tome la otra. Una imposición que existe ahora por tanto es que la ecuación de regresión lineal queda restringida al caso de plantearnos la asociación en términos de que existe una variable que la explica: la **variable explicativa**, y otra que es explicada: la **variable respuesta**.

El procedimiento matemático que nos va a permitir calcular las ecuaciones de su regresión y sus componentes es el **método de los mínimos cuadrados (MMC)**. El MMC es clave en la historia de la estadística pues se debe a su descubrimiento un gran impulso que a la postre fue clave para su evolución. Personajes clave de las matemáticas como Legendre y Gauss fueron algunos de los «culpables» de dicha evolución (Stigler, 1998).

Este método se basa en la idea de minimizar los errores de predicción (e). Estos errores están definidos como la diferencia entre los valores predichos (\hat{y}_i) y las observaciones, es decir, son las distancias entre el valor que le correspondería según la recta de regresión y el que presenta la observación.

$$e = y_i - \hat{y}_i$$



Según el criterio de mínimos cuadrados se obtienen las siguientes **soluciones para la recta de regresión de y sobre x** (la recta que explica y a partir de x), las cuales garantizan que las distancias de la recta de regresión a las observaciones es mínima:

$$\hat{y} = a + b \times$$

$$b = r \frac{s_y}{s_x}$$

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

Donde a es el término independiente o punto de corte con el eje y y b es la pendiente de dicha recta. Para obtener las ecuaciones del caso de la recta de regresión de x sobre y simplemente tenemos que sustituir la y por la x en las ecuaciones anteriores.

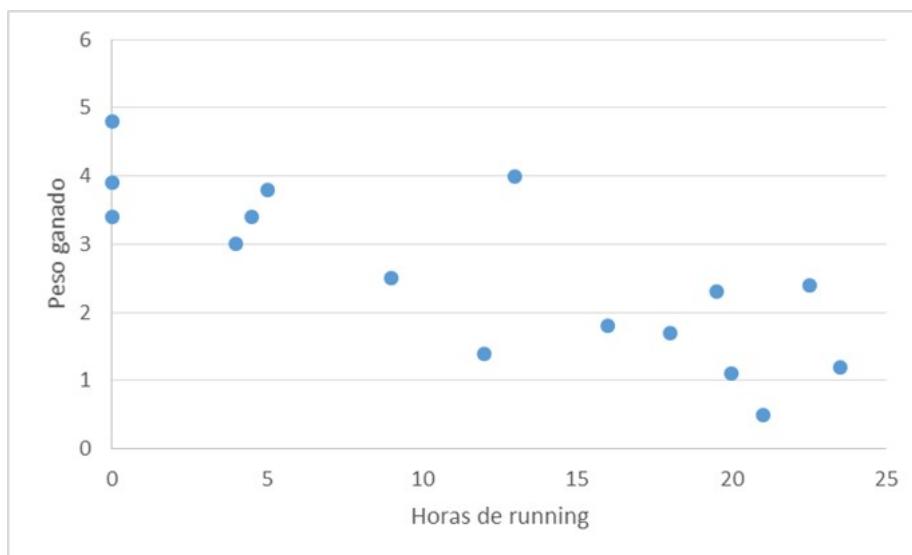
Ejemplo 3: En un estudio sobre ganancia de peso y ejercicio se sometió a 16 personas elegidas aleatoriamente de entre los pacientes que llegaban a una consulta de medicina general mientras cumplieran la condición de ser corredores ocasionales o aficionados. A los participantes se les pidió que se sometieran a la misma dieta hipercalórica (la cual incluía tres hamburguesas semanales, por ejemplo) durante un mes y que luego se pesaran de nuevo por la consulta y anotaran los minutos que habían corrido durante el mes.

El resultado fue el siguiente:

Horas de running/mes	0	0	0	4	4,5	5	9	12
Peso ganado (Kg)	4,8	3,4	3,9	3	3,4	3,8	2,5	1,4
Horas de running/mes	13	16	18	19,5	20	21	22,5	23,5
Peso ganado (Kg)	4	1,8	1,7	2,3	1,1	0,5	2,4	1,2

Tabla 4: Un estudio hipercalórico.

Dibujaríamos en primer lugar la nube de puntos:



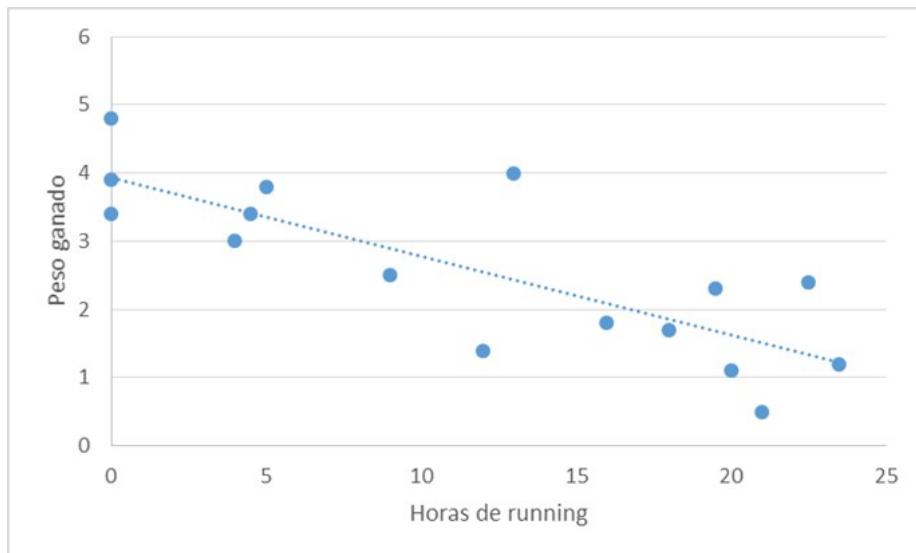
Gráfica 5: Gráfico de dispersión del estudio hipercalórico.

Según se observa, ya sabemos que tiene una pendiente negativa.

Calcularíamos los estadísticos necesarios de acuerdo a las fórmulas anteriores y resulta la siguiente ecuación:

$$\text{peso ganado} = 3,93 - 0,12 \times \text{horas running}$$

Si trazamos la recta sobre la gráfica resulta así:



Gráfica 6: Recta de regresión dibujada sobre el gráfico de dispersión.

Un aspecto interesante de construir la recta de regresión es que nos permite calcular las predicciones de y para un valor de x dado. Por ejemplo, a partir del estudio de la dieta hipercalórico que acabamos de ver podríamos querer saber cuál es el valor pronosticado si corremos 7 horas ese mes. Bastaría entonces con sustituir la x por su valor:

$$\hat{y}_{7\text{horas}} = 3,93 - 0,12 \times 7 = 3,09\text{kg}$$

En este caso no tenemos modo alguno de ver el error que estamos cometiendo pues ningún individuo de la muestra corrió tal cantidad de horas. Sin embargo si elegimos una cantidad que corresponda a las horas que corrió alguien dentro de nuestra muestra entonces podremos calcular el error cometido por la predicción. Por ejemplo, veamos qué error comete la recta prediciendo el peso ganado por la persona que corrió 4 horas ese mes.

$$\hat{y}_{4\text{horas}} = 3,93 - 0,12 \times 4 = 3,45\text{ kg} \rightarrow e = 3 - 3,45 = -0,45\text{ kg}$$

En cuanto a los errores es interesante saber que la suma de los errores de todas las observaciones es igual a cero, lo cual es una imposición por el hecho de exigir que la suma de los cuadrados de estos sean también cero.

Otro aspecto a tener en cuenta es que la recta de regresión siempre pasa por el punto de las medias (\bar{x}, \bar{y}).

Vamos a considerar ahora la relación entre la correlación y la regresión. Ya hemos visto que en la propia pendiente está presente el coeficiente de correlación, pero además para medir lo bueno que es el ajuste de la recta a los puntos, la llamada **bondad de ajuste**, empleamos el r^2 que mide el porcentaje de varianza de y explicada por x . El r^2 también es llamado **coeficiente de determinación**.

Ejemplo 4: Retomemos el ejemplo anterior y veamos lo bueno que era el ajuste logrado. Para ello tenemos que rescatar el coeficiente de correlación $r = 0,8$. De este modo $r^2 = 0,64$, lo cual quiere decir que este ajuste lineal hace que casi las dos terceras partes de la variabilidad del peso ganado queden explicadas por las horas de *running mensual*.

Otra manera de explicar el significado de r^2 es a través de esta otra relación:

$$r^2 = \frac{s^2\hat{y}}{s^2y} = \frac{\text{Varianza de las predicciones}}{\text{Varianza de las observaciones}}$$

Ejemplo 5: veamos cómo podemos calcular con el *software R* las cantidades aquí introducidas para caracterizar la relación entre variables. Tomemos como ejemplo un análisis del crecimiento de dos niños donde registramos su crecimiento durante siete años. (Nótese, en línea con lo propuesto en apartados anteriores, que tomamos ejemplos simples para introducir funciones y desarrollar código, estos ejemplos son extrapolables de forma inmediata a casos de mayor complejidad numérica).

Edad (años)/Altura individuo (m)	1	2	3	4	5	6	7
Niño 1	0,76	0,95	1,05	1,20	1,31	1,32	1,34
Niño 2	0,78	0,97	1,10	1,23	1,32	1,33	1,38

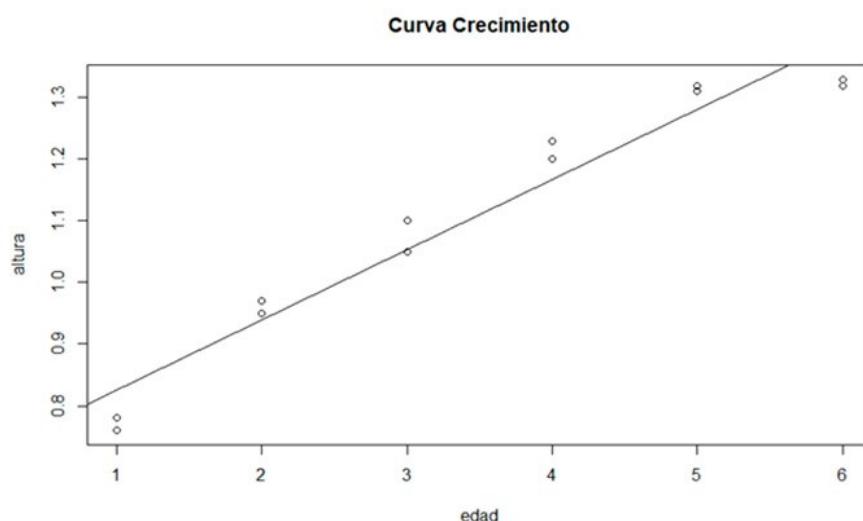
Queremos visualizar y cuantificar la relación entre las dos variables. Con este objetivo faremos el gráfico de dispersión (para «visualizar» la relación entre variables) y calcularemos la mejor aproximación a una regresión lineal a través de la recta de regresión.

Utilizamos un código como el que se muestra a continuación, se visualiza el gráfico de dispersión y finalmente se calcula la recta de regresión.

```
>edad<-c(1,1,2,2,3,3,4,4,5,5,6,6)
>altura<-c(0.76,0.78,0.95,0.97,1.05,1.10,1.20,1.23,1.31,1.32,1.32,1.33)
>scatter.smooth(x=edad, y=altura, main="Curva Crecimiento")
>plot(x=edad, y=altura, main="Curva Crecimiento")
>regresion<-lm(altura~edad)
>summary(regresion)
>abline(regresion)
```

Las líneas de código que se introducen responden a los siguientes comandos de este algoritmo (simple) para obtener los valores numéricos del modelo de regresión lineal aplicado a las dos variables de nuestro problema. El código describe:

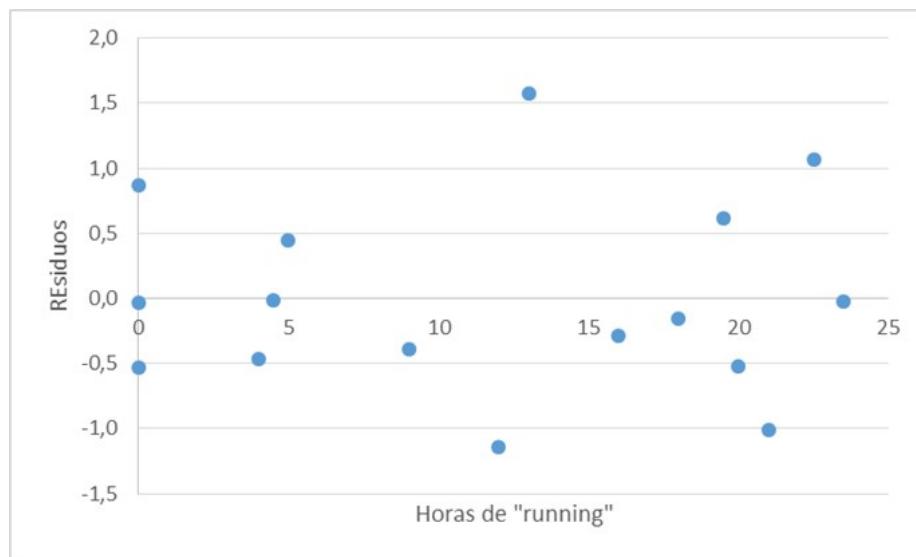
- ▶ Valor de la variable edad.
- ▶ Valor de la variable altura.
- ▶ Función para hacer un gráfico de puntos de las dos variables y una interpolación que visualmente transmite una idea del comportamiento de la función. No tiene más valor que esto, orientativo.
- ▶ El gráfico de los puntos individuales que surgen de la relación entre las variables en la muestra.
- ▶ Definición del modelo de regresión, lineal en este caso.
- ▶ Uso de la función summary para calcular los estadísticos principales del modelo. De aquí podemos extraer, pendiente de la recta, intercepto.
- ▶ Un esbozo visual de la recta de regresión calculada (ver figura abajo incluida).



Para valorar cuán buena es esta aproximación lineal se deben **revisar los valores de correlación obtenidos** y ya quedará a decisión del analista el elegir uno u otro modelo. Incluso, como se verá en epígrafes siguientes, la posibilidad de estudiar modelos de regresión no lineales.

4.5. Gráfico de residuos

Tal y como hemos visto la media de los residuos es siempre cero y esto propicia que podamos graficarlos para poder estudiarlos. El modo de hacerlo es empleando un **gráfico de residuos** que es un diagrama de dispersión con la variable explicativa en el eje de las X y los residuos en el ejes de las Y .



Gráfica 4: Gráfico de los residuos.

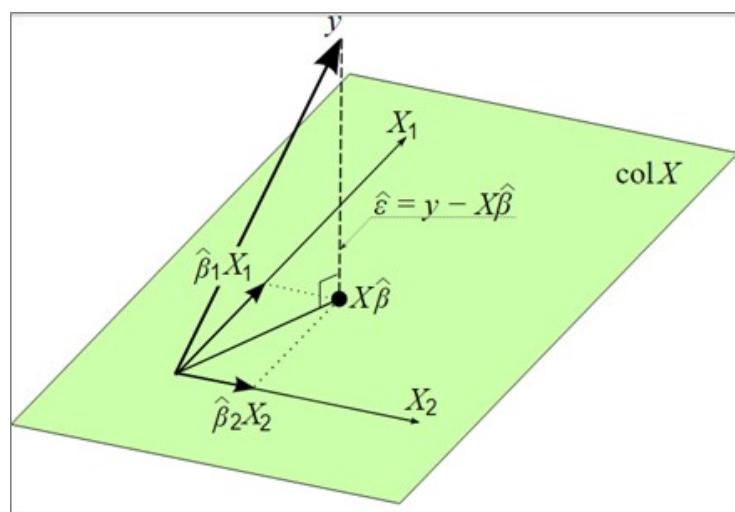
Cuando la recta de regresión está captando y ajustándose bien a las observaciones entonces no deberíamos apreciar ningún patrón en los residuos. Cuando tratemos la regresión desde la concepción de la probabilidad veremos que se impondrá precisamente en muchos modelos estadísticos este mismo requisito.

Otro aspecto que es posible ver con los residuos y de enorme interés para todo tipo de modelos estadísticos es lo que se conoce como **heterocedasticidad**, que es la diferente varianza que muestran estos a lo largo de la variable explicativa. Esto es en sí un tipo de patrón, uno que hace que los errores vayan creciendo o decreciendo según la magnitud de los valores de la variable explicativa y por tanto teóricamente no será deseable.

Por último, comentar **lo perjudicial que puede ser un valor extremo para el ajuste correcto de la recta de regresión**. Al haber una observación extrema de más podemos dar al traste con un buen ajuste de la recta de regresión pues esta tratará de «recoger» la observación extrema a base de alterar su pendiente y punto de corte. Al construir una ecuación de la recta diferente obtendremos predicciones muy alteradas por el valor extremo lo cual tendrá consecuencias nefastas para los errores.

4.6. Regresión lineal multivariante

Tal y como hemos visto en la regresión simple, es posible estimar el valor de una variable en función de otra. Esto, además, nos permitía estudiar si existía una relación de tipo lineal entre ellas. Sin embargo, **¿qué podemos hacer si una variable está relacionada con varias variables en vez de con una sola?**, ¿cómo podemos modelar esta situación? Una posibilidad es llevar a cabo una **regresión lineal multivariante o regresión múltiple**. Una regresión lineal múltiple consiste en la **generación de un plano que contenga a todos los puntos**. Dicho plano trata de definir la relación que existe entre una determinada variable con un determinado conjunto de variables.



Generación del plano de regresión. Fuente:

https://upload.wikimedia.org/wikipedia/commons/8/87/OLS_geometric_interpretation.svg

Pasemos a definir formalmente una relación lineal multivariante. Denominaremos $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ como el conjunto de variables cuya relación lineal con la variable Y quiere ser estudiada. Definimos **la relación lineal** de la siguiente forma:

$$Y = a + b_1X_1 + \dots + b_nX_n + \varepsilon$$

Por tanto, el método de regresión lineal multivariante debe ser capaz de calcular el valor a junto con todos los valores $\mathbf{B} = \{b_1, \dots, b_n\}$. Al igual que ocurre en la regresión lineal simple, hay un **error asociado a cada predicción**. Hay, por tanto, un residuo para cada posible combinación de valores utilizado en la generación del modelo lineal.

Para acabar, veremos **cómo aplicar un análisis de regresión lineal múltiple en R**. Primero, debemos definir nuestro conjunto de datos. En este caso, utilizaremos uno de los paquetes de datos que incluye R por defecto: el conjunto de datos mtcars. A continuación **aplicaremos el método de regresión lineal múltiple** incluido en R y, por último, **analizaremos los resultados y dibujaremos el gráfico de residuos**.

Todas estas acciones pueden realizarse mediante el **siguiente código**:

```
data(mtcars)
attach(mtcars) #Cargamos los datos
print(mtcars[1:5,]) #imprimimos las 5 primeras filas
cars.lm = lm(mpg~hp+wt) # realizamos la regresion
print(summary(cars.lm)) #mostramos los resultados
plot(cars.lm$residuals) #dibujamos el gráfico de residuos
abline(h=0) #pintamos linea horizontal en 0|
```

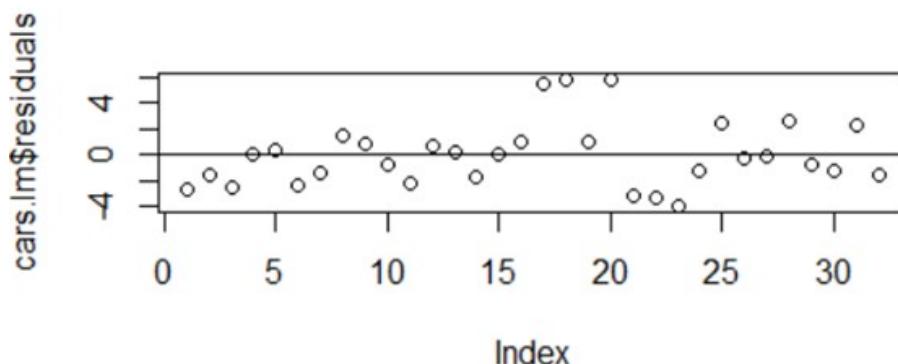
Como puede intuirse, hemos decidido **estimar la variable millas recorridas por galón (mpg) a partir de la potencia (hp) y peso (wt) del vehículo**.

Los datos referentes a los valores B y el término independiente, a, vienen especificados en el apartado coeficientes que nos da la función **summary**. Para nuestro ejemplo, mostramos los resultados a continuación:

Con lo cual, la relación lineal entre las distintas variables viene dada por la siguiente fórmula:

$$Mpg = 37.22727 - 0.03177 \cdot hp - 3.87783 \cdot wt$$

Asociado a cada coeficiente hay un ***p – value***. Veremos más acerca de los ***p – values*** en los temas siguientes. Por ahora basta con saber que **cuanto más cercano a 0, mejor explica el modelo la relación entre las variables**. Los asteriscos (*) van asociados al ***p – value*** y son un indicativo del error. **Cuantos más asteriscos, mejor es la aproximación**. En caso de que no aparezca nada en esa columna, podemos afirmar que el modelo de regresión calculado **no se ajusta a los datos** con lo cual la relación que hay entre las variables **no es lineal** y no puede ser expresada por el modelo. Por último, siempre es bueno echar un vistazo al **gráfico de residuos** asociado:



Tal y como podemos observar, **todo parece indicar que hay una relación lineal entre las variables**. No parece haber signos de **heterocedasticidad** ni de ningún **patrón** que pudiera hacernos sospechar que la relación entre las variables fuera no lineal.

4.7 Regresión no lineal

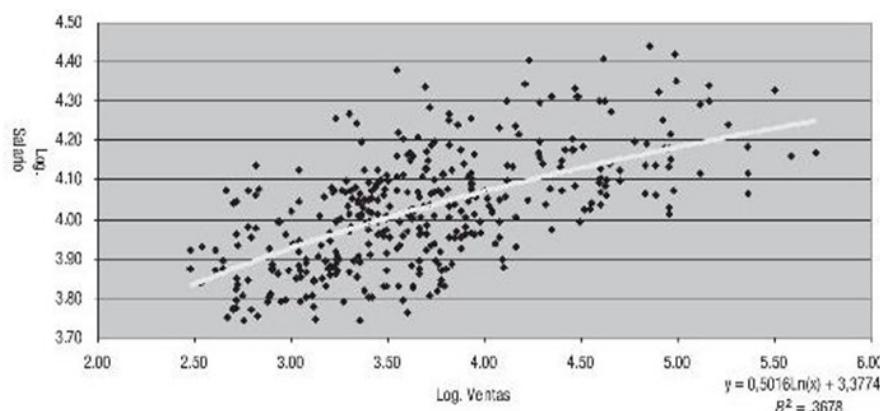
Cuando representamos gráficamente la relación entre dos variables cuantitativas a través de un diagrama de dispersión puede ocurrir que observemos que la relación lineal no es la que mejor describiría su asociación. En este caso puede que nos tengamos que plantear realizar **transformaciones en las variables** para que las relaciones ganen en linealidad.

Por otro lado, como **las correlaciones representan solo la asociación lineal** entre variables, las asociaciones no lineales no estarán siendo representadas en el valor de la correlación (Salvador Figueras y Gargallo, 2003).

Además del propio gráfico de dispersión podemos sospechar la existencia de relaciones no lineales a través de los errores que nos facilita el gráfico de residuos.

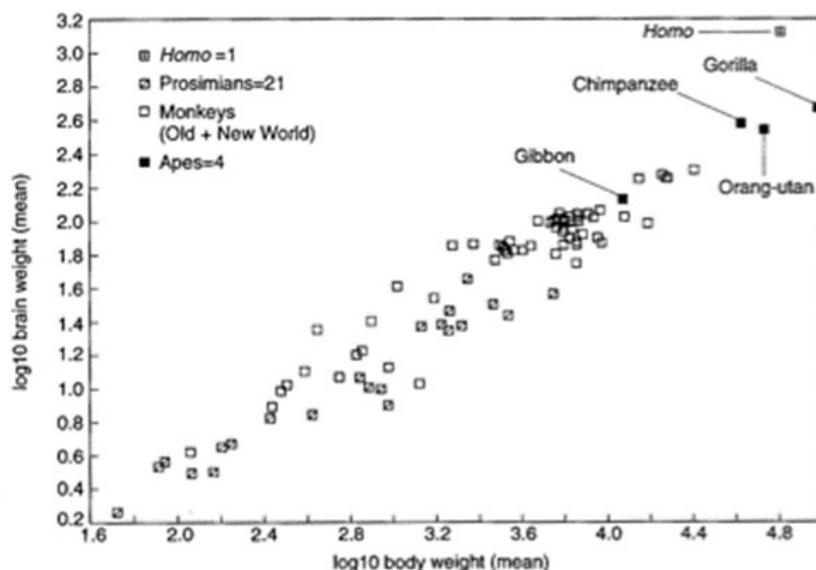
Existen numerosos casos en la naturaleza y la dimensión social de variables cuya relación se expresa mejor con funciones logarítmicas o exponenciales. Un ejemplo clásico son las variables que tienen que ver con salarios. De hecho es probable que la mayor importancia que se le ha dado en la teoría a la linealidad provenga del hecho de que es más sencilla matemáticamente.

Ejemplo 6: Podemos encontrar numerosos casos de relaciones económicas donde encaja mejor una relación logarítmica entre las variables.



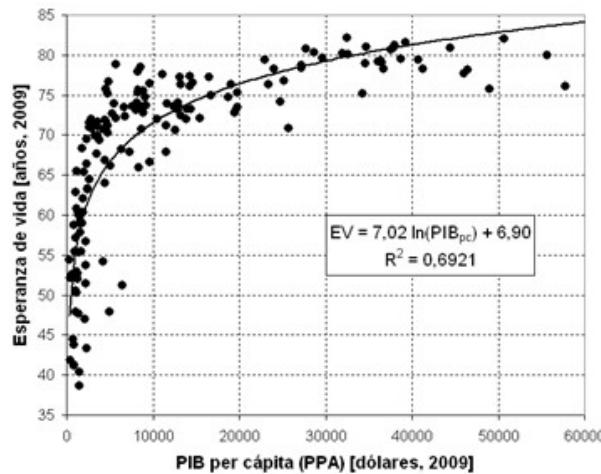
Gráfica 5: Diagrama de dispersión de una relación logarítmica en Economía.

En las ciencias naturales también son habituales este tipo de relaciones. Un ejemplo clásico es el de los tamaños de cerebros de animales.



Gráfica 6: Diagrama de dispersión de una relación logarítmica en las Ciencias Naturales.

En ambos casos se logró mostrar la linealidad previa transformación logarítmica de las variables. Otro ejemplo de relación logarítmica, esta vez sin transformar, lo podemos apreciar en la relación que existe entre la Renta per Cápita de un país y su esperanza de vida.



Gráfica 7: Relación logarítmica entre la Renta Per Cápita y la Esperanza de Vida. Fuente:http://es.wikipedia.org/wiki/Renta_per_c%C3%A1pita

La manera de leer esta ecuación o relación logarítmica es: «un crecimiento de un 100% en la renta per cápita predice un aumento de 7 años de vida».

4.8 LTS (Least Trimmed Squares)

La **regresión de mínimos cuadrados recortados** o Least Trimmed Squares es una variación del método de regresión por mínimos cuadrados visto que **trata de reducir la influencia de los outliers**. La idea que tiene debajo este método es muy simple. Simplemente se trata de, iterativamente, ir **cogiendo subconjuntos de puntos a los que se les va aplicando el método de mínimos cuadrados normales**. Al final, nos quedamos con la versión que minimice los residuos. Los pasos que sigue este método se detallan a continuación:

- ▶ **Selección del número de puntos para la realización de la regresión**: los procesos de regresión se realizan sobre **un conjunto fijo de puntos**. Por tanto, el tamaño de cada conjunto es un parámetro de gran importancia dentro de este algoritmo. Si el tamaño es pequeño, entonces podremos crear muchos conjuntos pero cada resultado estará formado por muy pocos elementos lo que seguramente nos dé resultados poco fiables. Por el contrario, si escogemos un tamaño muy grande, entonces podremos hacer pocos subconjuntos lo que nos da menos opciones de entre las que elegir.
- ▶ **Formación de los subconjuntos**: una vez seleccionado el tamaño de los subconjuntos, se generan todos los subconjuntos de puntos posibles que tengan ese tamaño.
- ▶ **Aplicación de la regresión de mínimos cuadrados**: se aplica el proceso de regresión de mínimos cuadrados **sobre cada subconjunto de puntos**. Para cada resultado se obtienen los residuos, es decir, el **error cometido en cada punto**.
- ▶ **Selección de la opción con menor error**: se comparan todas las soluciones calculadas y, **aquella que tiene la menor cantidad de residuos** es la seleccionada como solución final.

Este método nos proporciona la principal ventaja de ser **mucho menos sensible a los outliers que la regresión de mínimos cuadrados original**. Por ello, este método se enmarca dentro de lo que conocemos como **estadística robusta**. Al ir escogiendo subconjuntos de puntos dejamos fuera, para algunas soluciones, aquellos puntos problemáticos que hacen que el ajuste sea peor.

Veamos un ejemplo en R sobre cómo poder trabajar con este método. Primero, **debemos definir nuestro conjunto de datos**. Para ello, haremos uso de las siguientes sentencias:

```
library(robustbase)
par(new = FALSE)
x = 1:30
y = x*5
y[2] = 11
y[3] = 17
y[4] = 0
y[9] = 47
y[13] = 73
y[21] = 110
```

Como puede verse, primero hemos definido las librerías que vamos a utilizar en este ejemplo. En concreto, utilizaremos la librería **robustbase** que contiene el **método de mínimos cuadrados recortados** que estamos utilizando. A continuación, hemos definido unos datos de ejemplo. Para comprobar la efectividad hemos definido un ejemplo muy sencillo en donde hemos introducido variaciones a datos generados por la recta $y = 5x$.

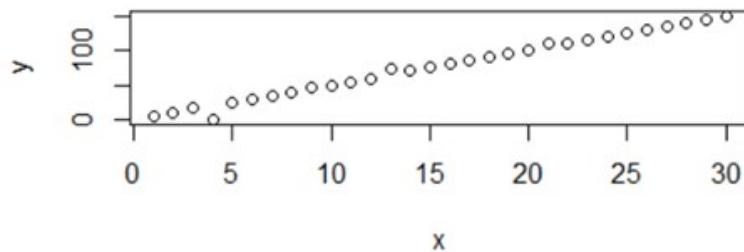
Podemos ver los datos con los que vamos a trabajar en este ejemplo en la siguiente tabla:

1	5
2	11
3	17
4	0
5	25
6	30
7	35
8	40
9	47
10	50
11	55
12	60
13	73
14	70
15	75
16	80
17	85
18	90
19	95
20	100
21	110
22	110
23	115
24	120
25	125
26	130
27	135
28	140
29	145
30	150

Podemos **generar la gráfica de dispersión** de los datos mediante el siguiente comando:

```
plot(x,y)
```

El resultado de la ejecución de dicha sentencia es el siguiente:



Una vez definidos los datos, aplicamos la función del método e imprimimos los resultados. A continuación, **mostramos el código de la función y el resultado obtenido tras su aplicación**. La función print imprime el contenido de la variable por pantalla.

```

resultado = ltsReg(x,y)
print(resultado)

Call:
ltsReg.default(x = x, y = y)

Coefficients:
Intercept           x
2.446e-14  5.000e+00

Scale estimate 0

```

De toda la información dada en el resultado, la parte importante viene determinada por el apartado **Coefficients**. Dichos coeficientes son los que nos proporcionan los **datos necesarios para construir la recta**. El valor Intercept hace referencia al término independiente mientras que la columna de x nos da el valor de la pendiente. Por tanto, la recta resultante es:

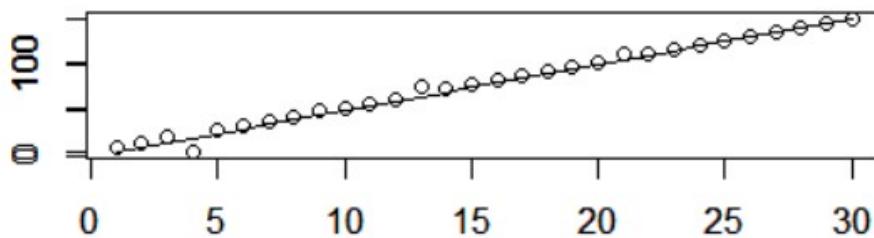
$$y = 5x + 2.446e-14$$

Lo cual es virtualmente $y = 5x$, la función principal que hemos usado para generar el ejemplo. Como podemos ver, a pesar de haber introducido inexactitudes en los datos, el **método es robusto y hace caso omiso de los valores outliers**.

Por último, podemos obtener una gráfica del resultado obtenido si ejecutamos las siguientes sentencias:

```
plot(x,y)
par(new = TRUE)
plot(x,yestimada,type="l")
```

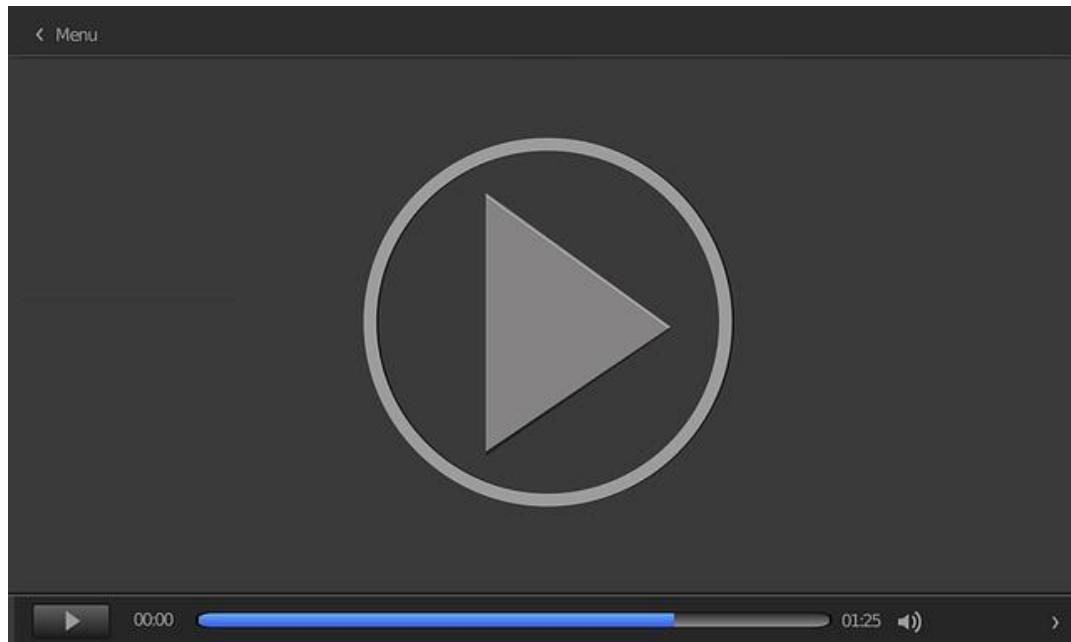
Dichas sentencias producen el siguiente resultado gráfico:



Podemos ver, por tanto, como la recta **se adapta perfectamente a los datos** con los que estamos trabajando.

Minería de datos: modelización y cálculos numéricos con R

En este vídeo vamos a introducir el concepto estadístico: minería de datos.



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=6e2736fe-3985-458b-a405-acbd00b21c41>

Practica con R los conceptos estudiados

Asumiendo que tienes [instalado R](#) y [RStudio](#), abre el IDE de RStudio y en un «R script» nuevo «Ctrl + Shift + N» escribe el siguiente código.

```
rm(list=ls())
#####
requiredPackages <- c("arsenal",
"car","chemometrics","corrplot","datarium","gapminder","dplyr","DescTools",
"foreign", "e1071", "expss", "GGally", "ggplot2", "haven","knitr","plotly",
"psych","remotes", "summarytools","ggridges","table1", "tableone",
"tidyverse", "SmartEDA", "scales", "caret", "imputeMissings", "mice")

sesion1 <- function(pkg){
new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
if (length(new.pkg))
install.packages(new.pkg, dependencies = TRUE)
sapply(pkg, require, character.only = TRUE)
}

sesion1(requiredPackages)
#####
#####Ejemplo 1#####
#El paquete gapminder contiene un fichero de datos de población, esperanza
de vida y renta per cápita de los países del mundo entre 1952 y 2007.
#La fundación Gapminder es una organización sin fines de lucro con sede en
Suecia que promueve el desarrollo global mediante el uso de estadísticas.
library(gapminder)
# Descripción de variables
# country: factor with 142 levels
# continent: factor with 5 levels
# year: 1952-2007
# lifeExp: life expectancy at birth
# pop: total population
# gdpPercap: per-capita GDP

#Gráfico de dispersión
gap=data.frame(gapminder)
ggplot(gap, aes(y=lifeExp, x=gdpPercap)) +
geom_point()+
geom_smooth(method=lm)

#Gráfico de dispersión
gap=data.frame(gapminder)
ggplot(gap, aes(y=lifeExp, x=log(gdpPercap))) +
geom_point()+
```

```

geom_smooth(method=lm)

#Correlación
cor(gap$lifeExp, gap$gdpPercap)
cor(gap$lifeExp, log(gap$gdpPercap))
cor(gap$lifeExp, log(gap$gdpPercap), method = "spearman")
cor(gap$lifeExp, log(gap$gdpPercap), method = "kendall")

#Modelo de regresión simple
model=lm(formula=lifeExp~log(gdpPercap), data=gap)
#resumen del modelo
summary(model)
#####Gráfico de residuos#####
ggplot(data.frame(x = seq(model$residuals), y = model$residuals)) +
  geom_point(aes(x, y)) +
  labs(x = "Index", y = "Residuals",
       title = paste("Residuals of", format(model$call)))
#####Gráfico de residuos#####
ggplot(data.frame(x = log(gap$gdpPercap), y = model$residuals)) +
  geom_point(aes(x, y)) +
  labs(x = "log(gap$gdpPercap)", y = "Residuals",
       title = paste("Residuals of", format(model$call)))

#####
#####Ejemplo 2#####
data("marketing", package = "datarium")
head(marketing, 4)
#Gráfico de dispersión
ggplot(marketing, aes(x = youtube, y = sales)) +
  geom_point() +
  stat_smooth(method=lm)
#correlación
cor(marketing$sales, marketing$youtube)
#Modelo de regresión simple
model1=lm(formula=sales~youtube, marketing)
#resumen del modelo
summary(model1)
#####Gráfico de residuos#####
ggplot(data.frame(x = seq(model1$residuals), y = model1$residuals)) +
  geom_point(aes(x, y)) +
  labs(x = "Index", y = "Residuals",
       title = paste("Residuals of", format(model1$call)))
#####Gráfico de residuos#####
ggplot(data.frame(x = marketing$youtube, y = model1$residuals)) +
  geom_point(aes(x, y)) +
  labs(x = "youtube", y = "Residuals",
       title = paste("Residuals of", format(model1$call)))

```

Prueba a ejecutar el *script* anterior siguiendo estas indicaciones:

- ▶ Ejecuta cada línea de código, posiciona el cursor en la primera línea y utiliza la opción «Run» o «Ctrl + Enter».
- ▶ Observa la «Consola» y «Environment» cuando ejecutas cada línea.

Profundiza utilizando R en el análisis de matriz de dispersión y correlaciones

Asumiendo que tienes [instalado R](#) y [RStudio](#), abre el IDE de RStudio y en un «R script» nuevo «Ctrl + Shift + N» escribe el siguiente código.

```
rm(list=ls())
#####
requiredPackages <- c("arsenal",
"car", "chemometrics", "corrplot", "datarium", "gapminder", "dplyr", "DescTools",
"foreign", "e1071", "expss", "GGally", "ggplot2", "haven",
"knitr", "plotly", "psych", "remotes", "summarytools", "ggridges", "table1",
"tableone", "tidyverse", "SmartEDA", "scales", "caret", "imputeMissings",
"mice", "robustbase", "rstatix")

sesion1 <- function(pkg){
new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
if (length(new.pkg))
install.packages(new.pkg, dependencies = TRUE)
sapply(pkg, require, character.only = TRUE)
}

sesion1(requiredPackages)
#####
#####Ejemplo 1#####
#####data=mcars#####
data(mtcars)
attach(mtcars)
print(mtcars[1:5,])
cars.lm=lm(mpg~hp+wt)
print(summary(cars.lm))
plot(cars.lm$residuals)
abline(h=0)

#####Ejemplo 2#####
#####data=marketing#####
data("marketing", package = "datarium")
head(marketing, 4)

#descriptivos#
describe(marketing)
summary(marketing)

#nans
na_count <- sapply(marketing, function(y) sum(length(which(is.na(y)))))
na_count <- data.frame(na_count)
```

```

print(na_count)
#matriz de dispersión y correlación
library(plotly)
library(GGally)
p <- ggpairs(marketing, title="correlación por pares")
ggplotly(p)

#matriz correlaciones
(corr <- round(cor(marketing, method='pearson'), 2))

#otra manera de calcular las correlaciones
#library(rstatix)
cor_mat(marketing,method = "pearson",
alternative = "two.sided",
conf.level = 0.95)

#Corr p-values
cor_pmat(marketing,method = "pearson",
alternative = "two.sided",
conf.level = 0.95)

#mapa de calor correlaciones
ggcorr(marketing, method = c("everything", "pearson"),
,label = TRUE, label_size = 3, label_color = "black",
label_alpha = TRUE) #https://briatte.github.io/ggcorr/, https://search.r-project.org/CRAN/refmans/GGally/html/ggcorr.html

#modelo de regresión
model=lm(formula=sales~youtube+facebook+newspaper, marketing)
summary(model)

#Gráfico de residuales
ggplot(data.frame(x = seq(model$residuals), y = model$residuals)) +
  geom_point(aes(x, y)) +
  labs(x = "Index", y = "Residuals",
       title = paste("Residuals of", format(model$call)))+
  geom_hline(yintercept=0)

#####Ejemplo 3#####
#####data=simulated#####
#rnorm(30, mean=10, sd=2)
#rlnorm(30, log(10), log(2))
x=1:30
y=5*x
y[2]=11
y[3]=17
y[4]=0

```

```
y[9]=47  
y[13]=73  
y[21]=110  
datos<-data.frame(x,y)  
attach(datos)  
plot(x,y)  
resultado=ltsReg(y ~ x, data=datos)  
print(resultado)  
  
datos$yestimada=resultado$fitted.values  
  
plot(x,y)  
par(new=TRUE)  
plot(x,yestimada, type="l")
```

Recuerda, ejecuta cada línea de código por separado. Posiciona el cursor en la primera línea y utiliza la opción «Run» o «Ctrl + Enter». Analiza la salida que se muestra en la «Consola» y en «Environment» cuando ejecutes cada línea.

4.9. Referencias bibliográficas

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

Versión electrónica:

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

Salvador, M. y Gargallo, P. (2003). *Análisis Exploratorio de Datos*.

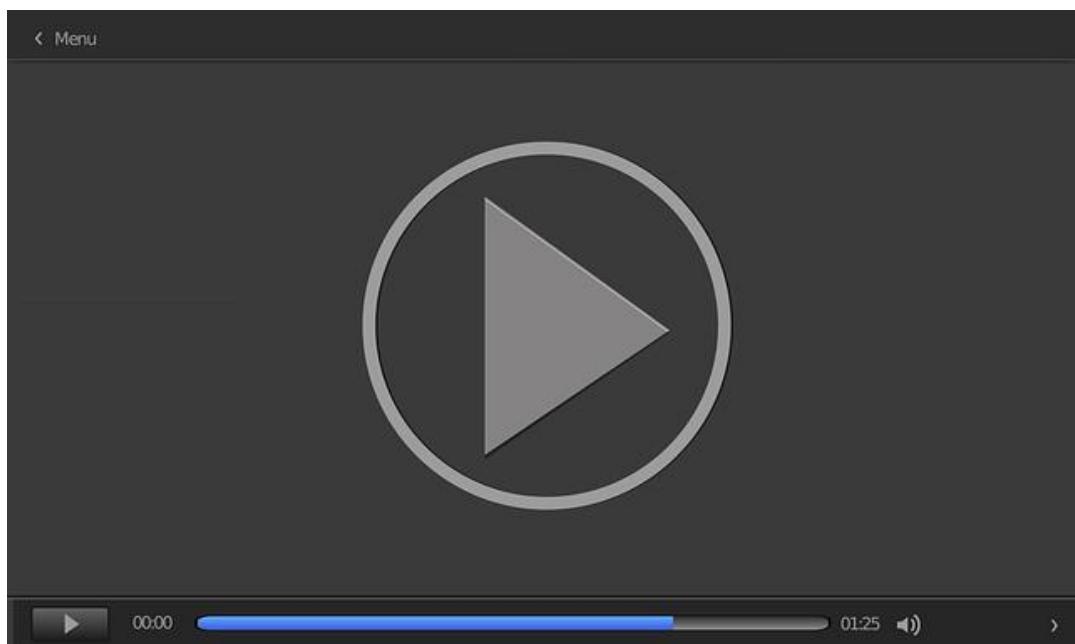
Recuperado de:

<http://www.5campus.com/leccion/aed>.

Stigler, S. M. (1998). *The history of statistics: The measurement of uncertainty before 1900* (7 ed.). London: Belknap Press of Harvard University Press.

Detectando puntos influyentes en nuestro modelo de regresión

En esta lección magistral se aborda el estudio de las implicaciones que pueden tener observaciones puntuales sobre el modelo de regresión elegido. Basándonos en un conjunto de datos sobre diabéticos trataremos aspectos como la representación gráfica de la asociación lineal entre variables cuantitativas, la búsqueda de patrones en el gráfico de dispersión, identificación de observaciones atípicas, valorando hasta qué punto ciertos puntos pueden ser influyentes, posibles cambios en la bondad de ajuste de la recta de regresión, etc.



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=cc38afb8-4434-4603-884a-abdc00f2aadb>

Tratamiento estadístico bidimensional

Un libro que está en castellano y está distribuido libremente en Internet es el de Gargallo (2003). En este libro el capítulo 6 está dedicado al tratamiento estadístico bidimensional y trata algunos temas más que no han sido tratados en este capítulo como la relación entre variables cualitativas.

Accede al libro desde el aula virtual o a través de la siguiente dirección web:

<http://www.5campus.com/leccion/aed>

Método de mínimos cuadrados

Si estás interesado en conocer más sobre el método de mínimos cuadrados y el origen de los primeros grandes resultados de la estadística, el libro de Stigler (1998) es realmente bueno, eso sí, de un nivel técnico en cuanto a las matemáticas alto.

Accede a una parte del libro desde el aula virtual o a través de la siguiente dirección web:

<http://books.google.es/books?id=M7yvkERHIIMC&printsec=frontcover>

Applets sobre correlación y regresión

Una web muy interesante sobre estadística donde encontrarás *applets* sobre correlación y regresión además de otras muchas más cosas. Se recomienda que no escatimes en investigar por los diferentes enlaces, pues se encuentran cosas realmente interesantes sobre estadística.

Accede a la página desde el aula virtual o a través de la siguiente dirección web:

<http://www.uvm.edu/~dhowell/SeeingStatisticsApplets/Applets.html>

Bibliografía

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

Versión electrónica:

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

Salvador, M. y Gargallo, P. (2003). *Análisis Exploratorio de Datos*.

Recuperado de:

<http://www.5campus.com/leccion/aed>.

Stigler, S. M. (1998). *The history of statistics: The measurement of uncertainty before 1900* (7 ed.). London: Belknap Press of Harvard University Press.

Martín Andrés, A. (2004). *Bioestadística para las ciencias de la salud*. Madrid: Norma-Capitel.

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

- 1.** Las frecuencias marginales son:
 - A. Aquellas que se marginan dejándose fuera de los cálculos.
 - B. Aquellas que alcanzan valores atípicos.
 - C. Aquellas que hacen referencia a una sola variable marginando a la otra.
 - D. Aquellas con las que se construye un diagrama de frecuencias bidimensional.

- 2.** La covarianza del mismo modo que la varianza...
 - A. No puede tomar valores negativos.
 - B. Su magnitud no nos dice mucho sobre si covaría entre sí mucho o poco las dos variables.
 - C. Mide la dispersión entre variables.
 - D. A y B son correctas.

- 3.** Un diagrama de dispersión nos permite ver:
 - A. Los patrones que se «esconden» en los datos.
 - B. Si se da algún tipo de relación lineal entre las variables
 - C. Si existe alguna observación atípica en los datos.
 - D. Las tres anteriores son correctas.

- 4.** La causalidad entre variables...
 - A. Se da siempre que tengamos algún tipo de asociación entre ellas.
 - B. No tiene por qué deducirse de la asociación que exista entre ellas.
 - C. Es equivalente a la asociación que pueda haber entre ellas.
 - D. A y C son correctas.

5. Cuando existe un asociación aproximadamente lineal entre variables que hemos modelado con una regresión lineal...
 - A. Una de las variables es la dependiente y la otra la independiente
 - B. Una de las variables es la explicativa y la otra la predictora.
 - C. La variable respuesta puede ser dependiente o independiente
 - D. Una de las variables es la variable explicativa y la otra la variable respuesta
 - E. A y D son correctas.
6. Si el coeficiente de correlación es nulo:
 - A. Puede estar existiendo otro tipo de relación entre las variables.
 - B. No existe causalidad entre ellas.
 - C. La asociación entre ellas es inexistente.
 - D. B y c son correctas.
7. El Método de los Mínimos Cuadrados sirve para...
 - A. Buscar la mejor asociación entre las variables.
 - B. Maximizar el ajuste entre los datos.
 - C. Hallar la ecuación de la recta que minimiza las desviaciones respecto a las observaciones
 - D. Complicar el análisis estadístico, pues ya está en desuso con la aparición de los ordenadores.
8. El coeficiente de determinación...
 - A. Es igual al cuadrado del coeficiente de correlación lineal.
 - B. Es igual a la raíz del coeficiente de correlación lineal.
 - C. Refleja el porcentaje de varianza explicado por las predicciones de la recta respecto al del total de observaciones.
 - D. A y C son correctas

- 9.** Los modelos lineales se emplean más en estadística porque...
- A. En la vida real son más comunes los modelos donde encajan adecuadamente.
 - B. Suponen una aproximación matemática sencilla a relaciones entre variables que pueden ser un tanto más complejas.
 - C. Otros modelos como el logístico no acaban de resultar manejables y se dan en casos raros.
 - D. A y C son correctas.
- 10.** Un *scatterplot* es:
- A. Un diagrama de residuos
 - B. Un gráfico que muestra la relación entre dos variables cuantitativas.
 - C. Una herramienta muy útil del Excel.
 - D. A y C son correctas.

Análisis e Interpretación de Datos

Tema 5. Probabilidad condicional y variables aleatorias

Índice

[Esquema](#)

[Ideas clave](#)

- [5.1. ¿Cómo estudiar este tema?](#)
- [5.2. Introducción a la teoría de la probabilidad](#)
- [5.3. Principios de la teoría de probabilidad](#)
- [5.4. Probabilidad condicional e independencia](#)
- [5.5. Variable aleatoria](#)
- [5.6. Modelos discretos](#)
- [5.7. Modelos continuos](#)
- [5.8. Referencias bibliográficas](#)

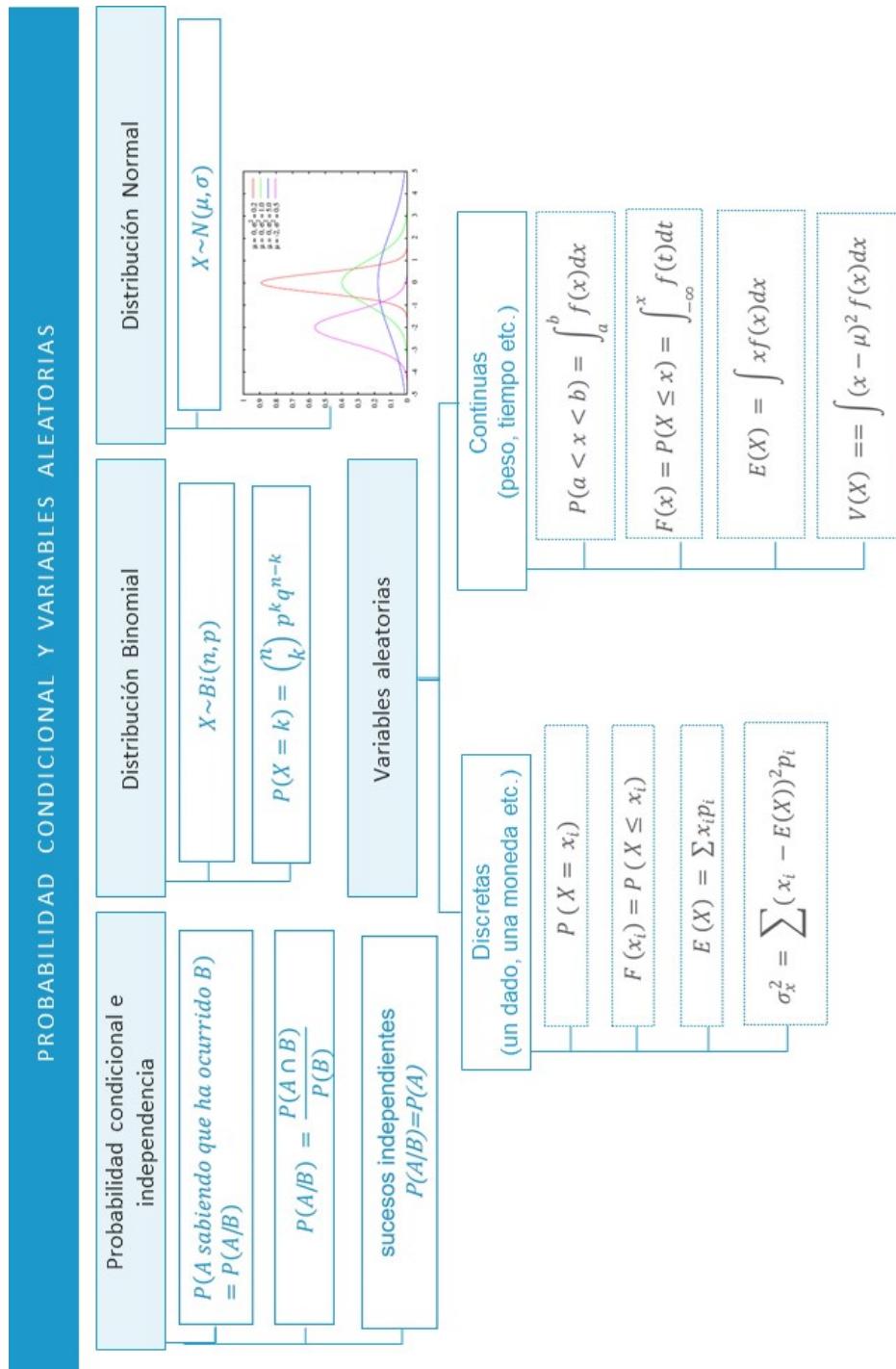
[A fondo](#)

[Aproximación de una distribución Binomial a una Normal](#)

[Modelos de probabilidad](#)

[Bibliografía](#)

[Test](#)



5.1. ¿Cómo estudiar este tema?

Para estudiar este tema lee las **páginas 106-115 y 123-159** del siguiente libro:

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

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

De los dos intervalos de páginas especificados, **el segundo no será necesario que lo estudies íntegro** pues algunas de las distribuciones de probabilidad discretas y continuas que tratan en el texto no las impartiremos por resumir esta parte y ser prácticos.

Nos limitaremos a las distribuciones que figuran en este tema, de modo que guíate por este criterio para saber si es necesario estudiarlas. **Para profundizar** de manera opcional puedes consultar el apartado **A fondo**, donde tienes un enlace a una página en donde se explican otros modelos teóricos de variables aleatorias que te podrían interesar.

Para hacerte una idea global de este tema es importante que mires el esquema del tema, el cual te ayudará a hacerte una buena idea de cómo está estructurado.

También será clave que practiques con los **ejercicios que vienen al final del tema** además de poder practicar con los que incluyen el libro de Ríus (1998) a partir de la página 159-162. Del mismo modo presta atención a los ejemplos que acompañan a los diferentes apartados a lo largo del capítulo, pues encierran muchas de las claves que te facilitarán la comprensión del tema.

5.2. Introducción a la teoría de la probabilidad

Definimos la probabilidad como una medida que se asocia a la ocurrencia de un suceso aleatorio, donde este es un evento sujeto a incertidumbre (como lanzar un dado, una moneda, etc.).

Cuando lanzamos un dado, estamos realizando un **experimento aleatorio**, pues no sabemos a ciencia cierta cuál será el resultado hasta que el dado se detenga. Lo que sí sabemos es que, si repetimos dicha tirada, tarde o temprano obtendremos, por ejemplo, un cinco.

La probabilidad es una función que nos va asociar la realización de un experimento aleatorio a un resultado.

Este resultado tendrá que estar definido previamente y formará parte del conjunto de resultados posibles del experimento aleatorio que denominamos **espacio muestral** y cuya notación es Ω .

Ejemplo 1: En el caso del lanzamiento de un dado y apuntar su resultado, el espacio muestral, designado por la letra Ω («Omega») es $\{1, 2, 3, 4, 5, 6\}$.

De modo que si tiráramos un dado repetidamente obtendríamos, por ejemplo, cinco en una proporción que sabemos se aproximará tanto como queramos a su valor teórico. La probabilidad la contemplaríamos como la frecuencia relativa del suceso «sacar un cinco» (número de veces que sale cinco al lanar el dado respecto al número total de lanzamientos). Esta visión de la probabilidad es la llamada **concepción frecuentista de la probabilidad**.

$$\lim_{n \rightarrow \infty} \text{freq. relativa}(A) = \frac{n_A}{n} = \text{Prob}(A)$$

Ejemplo 2:

si el suceso aleatorio A es «sacar un 5».

$$Prob(Sacar\ un\ 5) = \lim_{n \rightarrow \infty} freq.\ relativa(Sacar\ un\ 5) = \frac{n_5}{n}$$

Un enfoque sencillo de ver la probabilidad es el **enfoque clásico**, que tiene su origen en los albores de la disciplina probabilística, cuando se generó un interés fuerte por ella a raíz de los juegos de azar. La máxima expresión de este enfoque se da a través de la conocida como **regla de Laplace**:

$$p(A) = \frac{\text{casos favorables}}{\text{casos posibles}}$$

Donde el número de casos favorables y el de posibles podríamos ahora calcularlo con las técnicas de conteo vistas en este tema. La limitación que presenta este enfoque es que tenemos que considerar **sucesos equiprobables**, es decir con la misma probabilidad.

Ejemplo 3

Si A es «sacar un par» cuando tiramos un dado. Entonces tenemos que $A = \{2, 4, 6\}$, luego tenemos tres casos favorables mientras que los casos posibles son 6, los correspondientes a los seis resultados posibles.

De este modo:

$$p(A) = \frac{3}{6} = 1/2$$

Por otro lado, la noción frequentista la podemos ver aquí si nos imaginamos la repetición del experimento, esto es, del lanzamiento del dado muchas veces y luego observamos que la frecuencia relativa de A se aproximaría a $\frac{1}{2}$ (matemáticamente hablando diríamos que cuando el número de lanzamientos tiende a infinito la frecuencia relativa converge hacia la probabilidad).

5.3. Principios de la teoría de probabilidad

Los principios matemáticos sobre los que descansa una visión más moderna y formal de la probabilidad tienen su origen en los trabajos de **Kolmogorov**, en los que establece una serie de axiomas que constituyen la base matemática del lenguaje de la probabilidad que manejamos hoy en día.

El punto de partida es la **función de probabilidad P** , que será tal si se cumplen los siguientes cuatro axiomas, y entonces podremos decir que se ha definido correctamente un **espacio probabilístico**:

1. $P(\Omega) = 1$ La probabilidad del **sucedido seguro** es siempre 1. Ocurre siempre.
2. $0 \leq p(A) \leq 1$ La probabilidad de un suceso está entre 0 y 1. Será cero cuando no puede ocurrir nunca, también denominado, **sucedido imposible**.
3. Si A y B son dos **sucedidos disjuntos**, es decir que $A \cap B = \emptyset$, entonces:

$$P(A \cup B) = P(A) + P(B)$$

4. Si el espacio muestral está conformado por infinitos (en determinados casos es superfluo este 4º axioma, concretamente cuando es finito el espacio muestral) sucedidos disjuntos A_i entonces:

$$P(A_1 \cup A_2 \cup \dots) = \sum_{i=1}^{\infty} P(A_i)$$

De los axiomas podemos deducir **otros resultados útiles**:

1. $P(\emptyset) = 0$
2. Si \bar{A} es el complementario de A , entonces $P(\bar{A}) = 1 - P(A)$
3. Si $A \subset B$ entonces $P(A) \leq P(B)$
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

5.4. Probabilidad condicional e independencia

Vamos a comenzar con un caso concreto para explicar la probabilidad condicional. Más adelante la formalizaremos.

Ejemplo 4

Durante el año 2012-13 el número de estudiantes matriculados en cierta universidad española fue el siguiente:

	Derecho	Ingeniería Industrial	CC. Económicas	Total
Hombre	389	283	1156	1828
Mujer	483	52	728	1263
Total	872	335	1884	3091

Si elegimos al azar un estudiante, la probabilidad de que sea hombre y estudie Derecho es:

$$P(\text{Derecho} \cap \text{Hombre}) = \frac{389}{3091} = 0,126$$

Pero si ya supiéramos que es hombre, entonces la probabilidad de que estudie Derecho se ve modificada, pues la población de referencia ya no es la de todos los estudiantes sino la de los hombres:

$$P(\text{Derecho sabiendo que es hombre}) = \frac{389}{1828} = \frac{389/3091}{1828/3091} = \frac{P(D \cap H)}{P(H)}$$

Por tanto, cuando trabajemos con probabilidades como la anterior hablaremos de **probabilidad del suceso A condicionado a que ha ocurrido el suceso B**.

En el ejemplo anterior ocurría que sabíamos que era hombre y por tanto estaba condicionando la probabilidad de que estudiara Derecho. La probabilidad condicionada la especificamos así:

$$P(A \text{ sabiendo que ha ocurrido } B) = P(A|B)$$

Que se puede plantear por supuesto a la inversa, porque puede ser que ahora sepamos que ha ocurrido el otro evento y queramos entonces calcular la probabilidad de B:

$$P(B \text{ sabiendo que ha ocurrido } A) = P(B|A)$$

La expresión que nos permite el cálculo de la probabilidad condicionada es:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Si sabiendo que ha ocurrido uno de los eventos no modifica la probabilidad del otro diremos que son **independientes**. Por tanto, a partir de la fórmula anterior deducimos lo siguiente:

Si $P(B|A) = P(B)$ por no verse afectado por el hecho de que haya ocurrido A, entonces

$$P(B) = \frac{P(A \cap B)}{P(A)} \leftrightarrow P(A \cap B) = P(A) \times P(B)$$

Luego si dos sucesos son independientes entonces $P(A \cap B) = P(A) \times P(B)$

Ejemplo 5

Si tiramos una moneda repetidas veces, en principio, cada lanzamiento es un suceso independiente del anterior, pues la probabilidad de obtener una cara en el segundo lanzamiento es $\frac{1}{2}$ y no «importa» que haya sido cara o cruz el resultado del primer lanzamiento. Del mismo modo que si tenemos ya dos hijos varones y vamos a tener un tercero, a priori la probabilidad de que el tercero sea varón es la misma de que sea hembra.

Ten en cuenta que la igualdad de que la probabilidad de la intersección de dos eventos sea su producto de probabilidades también implica que son independientes. Por tanto siempre que apliques dicha igualdad cerciórate de que efectivamente los eventos son independientes.

5.5. Variable aleatoria

Hasta ahora hemos manejado los resultados de los experimentos aleatorios como «obtener una cara», «lograr un cinco lanzando un dado», etc. Claramente estos resultados son manejables de forma puramente numérica.

Así la serie de lanzamientos de monedas que resultan $\{CCXXXC\}$ podemos resumirla como 3 caras e incluso podemos asociar una variable matemática X al número de caras de modo que decir que $X = 3$ sea equivalente a decir que se han obtenido tres caras, pero con la ventaja de permitir un mejor manejo matemático.

Variable aleatoria

Variable cuyo valor representa el resultado de un experimento aleatorio.

Las variables aleatorias pueden ser de dos clases dependiendo de los valores que puedan tomar:

- ▶ La **variable aleatoria discreta** que solo puede tomar un número finito (o infinito numerable, puesto que la variable discreta puede tomar cualquier valor que se puede hacer corresponder a los números naturales N que son el 1, 2, 3, ...) de posibles resultados.
- ▶ La **variable aleatoria continua** que puede tomar todos los valores dentro de un intervalo dado.

Ejemplo 6

Si X = Número de caras obtenidas en n lanzamientos estamos ante una variable aleatoria discreta pues el número de caras será finito y además un valor entero positivo (un número natural).

Si X = Estatura, entonces X es una variable aleatoria continua pues en un intervalo dado puede tomar infinitos valores.

Ahora estudiaremos las principales características de estos dos tipos de variables además de los principales **modelos de probabilidad** discretos y continuos.

5.6. Modelos discretos

En un modelo de probabilidad discreto tenemos una probabilidad mayor o igual a cero para cada valor posible de la variable X .

Valor de X	X1	X2	X3	...	Xk
Probabilidad de X=x _i	P1	P2	P3	...	Pk

Estas probabilidades deben cumplir que la suma de todas ellas sea igual a uno y que cada probabilidad este contenida entre 0 y 1.

$$0 \leq p_i \leq 1$$

$$\sum p_i = 1$$

A la función que asigna una probabilidad a un valor discreto se le denomina **función de probabilidad**.

$$P(X = x_i)$$

Ejemplo 7

Si lanzamos una moneda dos veces, podemos obtener de 0 a 2 caras con unas respectivas probabilidades que apreciamos en la tabla siguiente.

Número de caras	0	1	2
$P(X = x_i)$	0,25	0,5	0,25

Observamos pues que se trata de un modelo discreto bien definido, ya que todas las probabilidades son mayores o iguales que 0 y, además, $0,25 + 0,5 + 0,25 = 1$.

Para describir una variable aleatoria discreta se emplea también el concepto de **función de distribución**, que nos indica la acumulación de probabilidad en un rango de valores discretos hasta uno dado (el x_i).

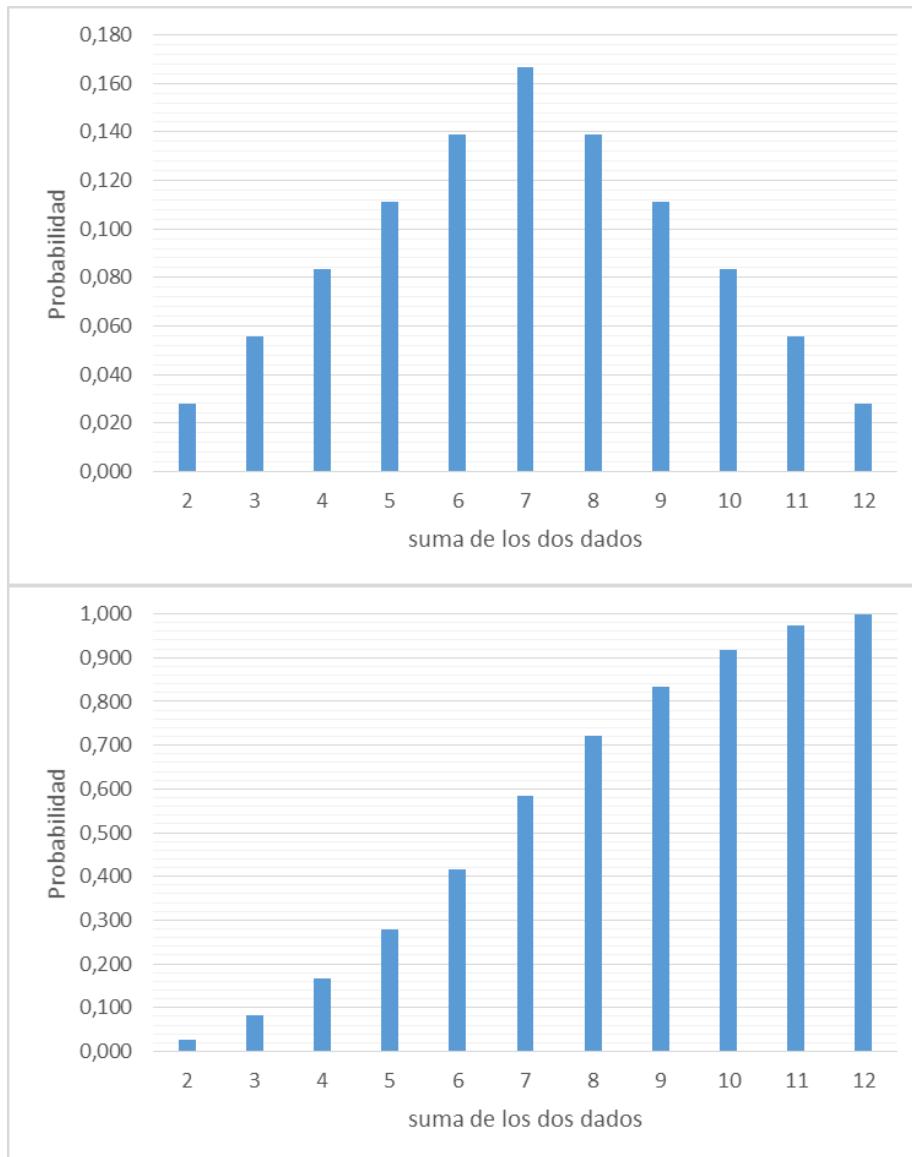
$$F(x_i) = P(X \leq x_i) = P(X = x_1) + P(X = x_2) + \cdots + P(X = x_i)$$

Ejemplo 8

Lanzamos dos dados y anotamos las puntuaciones. Los resultados posibles acompañados de la puntuación y sus respectivas probabilidades son las siguientes:

Xi	P(Xi)
2.....(1,1)	1/36
3.....(1,2); (2,1)	2/36
4.....(1,3); (2,2); (3,1)	3/36
5.....(1,4); (2,3); (3,2); (4,1)	4/36
6.....(1,5); (2,4); (3,3); (4,2); (5,1)	5/36
7.....(1,6); (2,5); (3,4); (4,3); (5,2); (6,1)	6/36
8.....(2,6); (3,5); (4,4); (5,3); (6,2)	5/36
9.....(3,6); (4,5); (5,4); (6,3)	4/36
10....(4,6); (5,5); (6,4)	3/36
11....(5,6); (6,5)	2/36
12....(6,6)	1/36

Si representamos gráficamente esta distribución de probabilidad a través de un diagrama de barras nos haremos una mejor idea de lo que es la función de distribución:



En los gráficos de barras anteriores observamos la función de probabilidad de la variable para todos los valores para a continuación mostrar la función de distribución.

Cuando manejamos variables aleatorias también tenemos el interés de resumir su información a través del **valor esperado** o **esperanza matemática** y se calcula como sigue:

$$E(X) = \sum x_i p_i$$

A la esperanza también la podemos designar como ***ux***.

Ejemplo 9

En el ejemplo de la moneda tenemos que la esperanza matemática (o «esperanza» a secas) o valor esperado es:

$$E(X) = \sum x_i p_i = 0 \times 0,25 + 1 \times 0,5 + 2 \times 0,25 = 1$$

Lo que se interpreta como que el valor medio del número de caras esperado cuando lanzamos dos veces una moneda normal es que obtengamos una.

Ejemplo 10

Y en el ejemplo de la suma de puntuaciones al lanzar dos dados la esperanza resulta:

$$E(X) = \sum x_i p_i = 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + 4 \times \frac{3}{36} + \dots + 10 \times \frac{3}{36} + 11 \times \frac{2}{36} + 12 \times \frac{1}{36} = 7$$

De modo que el valor esperado lanzando dos dados es 7.

Tenemos **estas propiedades sobre el valor medio esperado:**

- ▶ La esperanza de la suma es la suma de las esperanzas:

$$E(X + Y) = E(X) + E(Y)$$

$$E(a + bX) = a + bE(X)$$

También definimos la varianza de una variable aleatoria X como:

$$\sigma_x^2 = \sum (x_i - E(X))^2 p_i = (x_1 - E(X))^2 p_1 + (x_2 - E(X))^2 p_2 + \dots + (x_k - E(X))^2 p_k$$

Es habitual designar σ_x^2 como $V(X)$.

Ejemplo 11

En el ejemplo de las suma de los dos dados la varianza resulta:

$$V(X) = \frac{(2-7)^2 1}{36} + \frac{(3-7)^2 2}{36} + \frac{(4-7)^2 3}{36} + \dots + \frac{(10-7)^2 3}{36} + \frac{(11-7)^2 2}{36} + \frac{(12-7)^2 1}{36}$$

De todos modos la expresión anterior se suele calcular con una variante de la fórmula de la varianza que facilita el cálculo.

$$V(X) = \sum x_i^2 p_i - E(X)^2$$

De modo que el cálculo anterior se simplifica:

$$\sigma_x^2 = \frac{2^2 1}{36} + \frac{3^2 2}{36} + \frac{4^2 3}{36} + \dots + \frac{10^2 3}{36} + \frac{11^2 2}{36} + \frac{12^2 1}{36} - 7^2 = 54,83 - 49 = 5,83$$

También tenemos una serie de **reglas para la varianza**:

- ▶ $V(X + Y) = V(X) + V(Y) + 2Cov(X, Y)$ que también podemos expresar como
 $V(X + Y) = V(X) + V(Y) + 2\rho\sigma_x\sigma_y$ donde ρ es la correlación entre ambas variables, y σ_x, σ_y sus correspondientes desviaciones típicas.
- ▶ $V(a + bX) = b^2 V(X)$

Y si son independientes X e Y entonces la correlación es nula por lo que:

$$V(X + Y) = V(X) + V(Y) = V(X - Y)$$

Ejemplo 12

Si quieras estudiar en EEUU (e incluso a nivel laboral) te pueden exigir realizar la prueba conocida como SAT, la cual a su vez consta de varias partes. Dos de ellas miden la aptitud matemática y la lingüística. Si las puntuaciones obtenidas para matemáticas tienen una media de 419 con una desviación típica de 105 y la de lingüística alcanza una media de 407 con una desviación típica de 91 podemos construir la media y varianza de la puntuación obtenida por la suma de ambas pruebas.

Para las medias es fácil pues: (siendo M =Matemáticas y L = lingüística)

$$E(M + L) = E(M) + E(L) = 826$$

Sin embargo, para hallar la varianza necesitaríamos saber la correlación ρ entre ambas pruebas. Si además logramos saberla y vale 0,81 entonces podemos calcularla como:

$$V(M + L) = V(M) + V(L) + 2\rho\sigma_x\sigma_y \rightarrow V(M + L) = 105^2 + 91^2 + 2 \times 0.81 \times 105 \times 91 = 34781.1$$

Tras ver las propiedades de las variables aleatorias discretas vamos a pasar al estudio de uno de los modelos discretos más empleados. Este ocurre cuando realizamos un experimento aleatorio en el que queremos contar el número de «éxitos» de una determinada prueba que puede presentar dos resultados posibles (es decir, tendrás que ser dicotómica), entonces esta variable aleatoria diremos que se distribuye como una **distribución binomial** de parámetros n y p , donde n es el número de realizaciones y p la probabilidad de éxito en cualquiera de ellas. Para abreviar escribimos:

$$X \sim Bi(n, p)$$

Al ser la probabilidad de éxito p diremos que la del fracaso es $1 - p$ puesto que han de sumar uno entre ambas. La probabilidad del fracaso también se suele indicar con una q .

La probabilidad de que una variable aleatoria binomial (n, p) tome un valor concreto, es decir, alcance « k » éxitos es:

$$P(X = k) = \binom{n}{k} p^k q^{n-k}$$

Ejemplo 13

Precisamente en el ejemplo visto anteriormente la variable aleatoria número de lanzamientos en los que se obtiene «cara» se trataría de una distribución binomial donde « n » es 2 por ser dos lanzamientos y « p » es $\frac{1}{2}$ pues esa es la probabilidad de obtener cara (que es lo que consideramos éxito).

De modo que tendríamos
 $X = \text{«Número de caras al lanzar una moneda»}$ siendo:

$$X \sim Bi(2, \frac{1}{2})$$

Aquí observamos que el significado de «éxito» es figurado y connota cualquier evento que pueda ocurrir o no, siendo éxito cuando ocurre y fracaso cuando no lo hace.

Si quisieramos calcular, por ejemplo, la probabilidad de obtener dos caras en 6 lanzamientos ahora sería:

$$X \sim Bi(6, \frac{1}{2})$$

Y procederíamos del siguiente modo:

$$P(X = 2) = \binom{6}{2} \times 0,5^2 \times 0,5^4 = 15 \times 0,25 \times 0,0625 = 0,234$$

En el caso concreto de la distribución binomial la esperanza y varianza resultan:

- ▶ Si $X \sim Bi(n, p) \rightarrow E(X) = np$
- ▶ Si $X \sim Bi(n, p) \rightarrow V(X) = npq$

Ejemplo 14

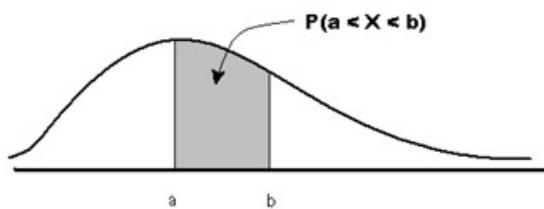
De modo que en el caso de la moneda lanzada tres veces tenemos que:

$$E(X) = 2 \times \frac{1}{2} = 1 \text{ cara} \quad y \quad V(X) = 2 \times \frac{1}{2} \times \frac{1}{2} = 1/2$$

5.7. Modelos continuos

En el caso continuo encontramos una diferencia importante, ya que todas las probabilidades siguen sumando uno, pero ahora ese valor unitario ha de repartirse entre infinitos valores que pueden tomar las x_i , por lo que la función de probabilidad será nula en un punto concreto y además en el caso continuo pasa a denominarse **función de densidad**, y se designa $f(x)$.

Al ser nulas estas probabilidades cobra importancia la función de distribución, ya que al acumular estas densidades alcanzará a tomar valores que serán proporcionales al área que encierre la función de densidad (generalmente una curva) entre ella misma y el eje de las x .



De modo que ahora la probabilidad es un área:

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

Siendo $f(x)$ la función de densidad. Y según apreciamos en el dibujo el área sombreada correspondería precisamente a la probabilidad acumulada en el intervalo (a, b) .

Si esta probabilidad la acumulamos hasta un punto dado entonces estamos en el caso de la **función de distribución acumulada de una variable aleatoria continua**:

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

Hay que tener en cuenta que el área seguirá sumando uno de modo que:

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

Ejemplo 15

Si tenemos que X es una variable aleatoria continua con la siguiente distribución:

$$f(x) = \begin{cases} \frac{x}{2} & \text{si } 0 \leq x \leq 2 \\ 0 & \text{en cualquier otra parte} \end{cases}$$

Y entonces,

$$\int_0^2 \frac{x}{2} dx = \left[\frac{x^2}{4} \right]_0^2 = \frac{4}{4} - 0 = 1$$

Además la función de distribución acumulada resulta:

$$F(x) = P(X \leq x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{x^2}{4} & \text{si } 0 \leq x \leq 2 \\ 1 & \text{si } x > 2 \end{cases}$$

En el caso continuo la esperanza será análogamente una integral y no una suma como lo era el caso discreto.

$$E(X) = \int xf(x)dx$$

Y la varianza de modo similar será:

$$V(X) = \int (x - E(X))^2 f(x)dx$$

Ejemplo 16

Siguiendo con el ejemplo anterior tendremos:

$$E(X) = \int_0^2 x \left(\frac{x}{2}\right) dx = \left[\frac{x^3}{6}\right]_0^2 = \frac{8}{6} = 4/3$$

$$V(X) = \int_0^2 (x - \frac{4}{3})^2 \left(\frac{x}{2}\right) dx = \int_0^2 x^2 \left(\frac{x}{2}\right) dx - \left(\frac{4}{3}\right)^2 = 2 - \frac{16}{9} = 2/9$$

La distribución de probabilidad teórica más conocida es sin duda la **distribución normal**. Durante un tiempo se creyó que todas las variables aleatorias eran continuas. Aunque hoy en día sabemos que no es cierto lo anterior, sí lo es que muchas de las variables aleatorias de la naturaleza se distribuyen como una normal (Martín Andrés, 2004).

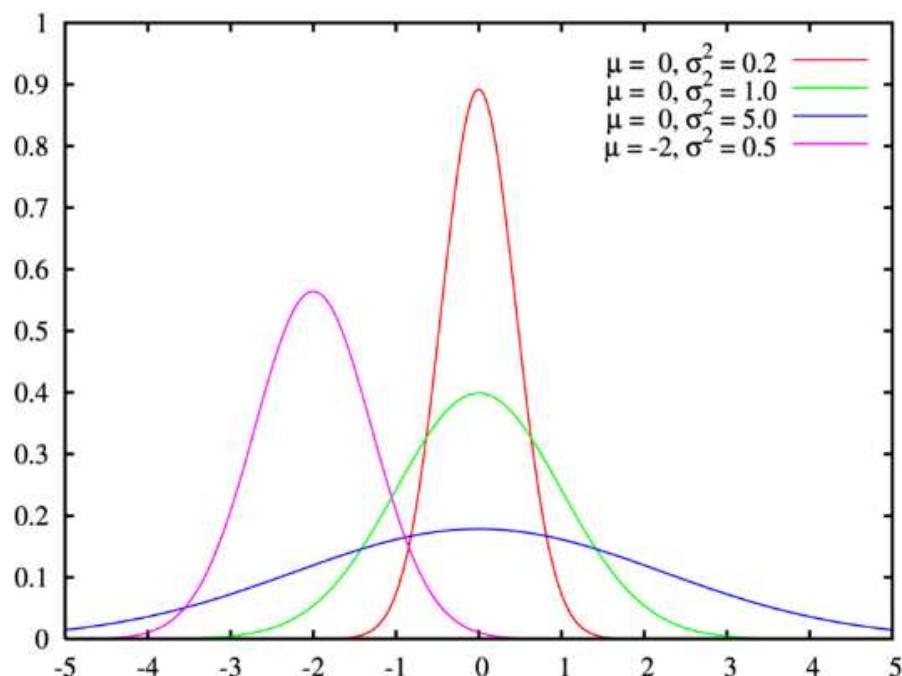
Sin embargo, la función matemática que las describe es compleja y no es práctica para manejarla.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Y tal y como se observa la **normal queda caracterizada por dos parámetros: su media μ y su desviación típica σ** . De modo que si X se distribuye como una normal lo anotamos como:

$$X \sim N(\mu, \sigma)$$

La forma de la normal es acampanada y como fue descubierta por Gauss, en ocasiones recibe el nombre de «Campana de Gauss».



Como se observa, la normal es simétrica respecto a su media y, de este modo, esta también es la mediana y, por la forma, también la moda. Luego **media, moda y mediana coinciden en la normal**.

Debido a la forma que tiene acampanada la mayor parte de **masa de probabilidad** (el área encerrada entre la curva y el eje de abscisas) se acumula en torno a la media, y cuánto más se aleja de ésta se hace más improbable —a un ritmo exponencial— que tome tal valor.

Es muy usada como referencia y por fines comparativos entre distribuciones una versión de la normal que se denomina **normal estándar**: $Z \sim N(0,1)$, que es entonces una normal con media 0 y desviación típica 1. En el caso de la gráfica anterior sería la que tiene color verde.

Podemos observar que cuanto mayor sea σ la distribución es más achata. De hecho σ está presente explícitamente en la gráfica pues es la distancia entre la media y el punto de la curva donde se produce una inflexión.

Para transformar una $N(\mu, \sigma)$ en una $N(0,1)$ empleamos las reglas vistas para la media y la varianza.

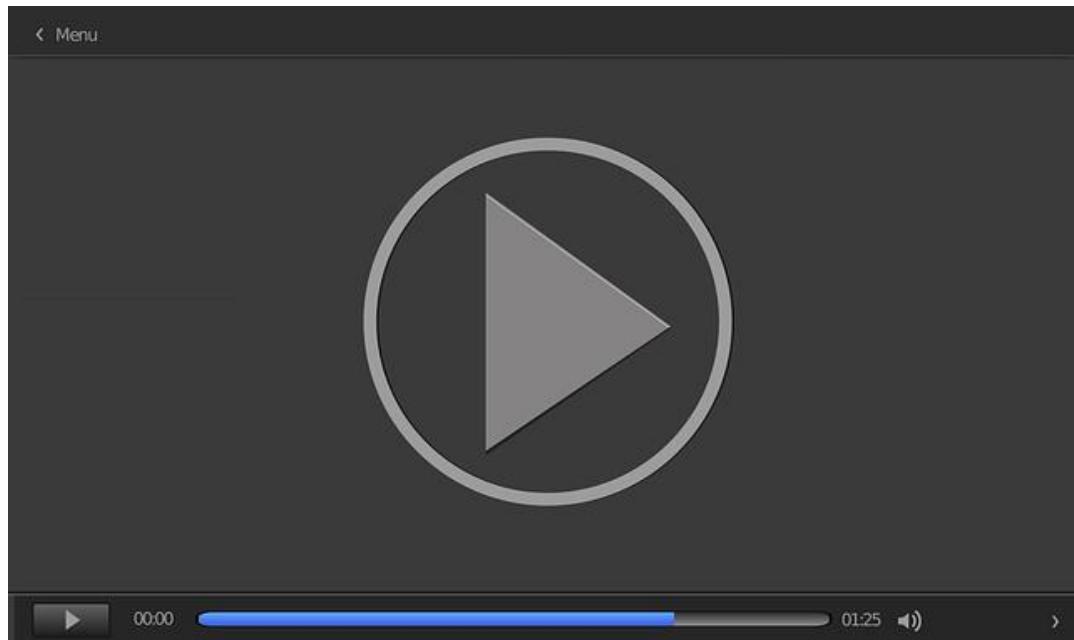
$$\frac{X-\mu}{\sigma} = z \leftrightarrow E\left(\frac{X-\mu}{\sigma}\right) = \frac{E(X)-\mu}{\sigma} = 0; V\left(\frac{X-\mu}{\sigma}\right) = \frac{V(X)}{\sigma^2} = 1$$

A esta transformación se le denomina **tipificación** y una de sus grandes utilidades concretas es la que permite trasladar cualquier valor de cualquier $N(\mu, \sigma)$ en su valor equivalente en una $N(0,1)$ de modo que se pueda saber su probabilidad al estar tabulada.

Esto sobre todo es importante cuando se consulta en las tablas de la $N(0,1)$ las relaciones entre los valores y sus probabilidades asociadas, aunque hoy debido a los ordenadores quizás está algo en desuso el empleo de estas tablas.

Apuntes sobre el concepto de «probabilidad»

En este vídeo veremos algunos apuntes sobre el concepto de «probabilidad» en el ámbito de la estadística computacional.



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=3a3c3a1f-6c54-43e4-aee8-acbd00c77afe>

Probabilidad con R

Asumiendo que tienes [instalado R](#) y [RStudio](#), abre el IDE de RStudio y en un «R script» nuevo «Ctrl + Shift + N» escribe el siguiente código. En este ejercicio, es importante que leas los comentarios que hay en el propio código para comprender lo que hace cada instrucción.

Una vez copiado, prueba a ejecutar el *script* línea a línea. Para ello, posiciona el cursor en la primera línea y utiliza la opción «Run» o «Ctrl + Enter».

```
#####
##### Repaso de instrucciones básicas en R #####
#####

#-----
# Cálculos estadísticos básicos
#-----
x<-c(1,1,2,2,2,3,3)

# tamaño de muestra
n=length(x)

# Tabla de frecuencias
table(x)
cumsum(table(x))
prop.table(x)
prop.table(table(x))

# Cálculo de la media muestral
suma<-sum(x)
media=suma/n
media
mean(x)
summary(x)

# Cálculo de la varianza muestral
y<-x^2
y
varianza=(sum(y)/n)-media^2
varianza

z<-x-media
z<-z^2
z
```

```

sum(z)

varianza=sum(z)/n
varianza

#Es la cuasivarianza muestral (n-1)
var(x)
((n-1)/n)*var(x)

summary(x)
#Cálculo de cuantiles
quantile(x,c(0.05,0.95))

library(e1071)
skewness(x) #simetria
kurtosis(x) #curtosis
plot(x)
hist(x)

#####
##### Ejemplos básicos de probabilidades #####
#####

# Cálculo de probabilidades, caso discreto y caso continuo
# -----
# CASO DISCRETO:
# -----
# Se sabe que una sucursal bancaria atiende un promedio de 6 clientes por
dia
# Sea la variable aleatoria X= número de clientes atendidos en un día
# X es discreta y sigue una distribución Poisson de parámetro lambda=6
# Calcular la probabilidad de que, en un día dado, la sucursal reciba más
de 6 y 8 o menos clientes
# Nos piden:
#  $P[6 < X \leq 8] = P[X \leq 8] - P[X \leq 6] = F(8) - F(6) = P[X = 7] + P[X = 8]$ 

ppois(c(6, 8), 6)
0.8472375-0.6063028

# Por tanto, también se puede calcular como:
dpois(7, 6)+dpois(8, 6)

# -----
# CASO CONTINUO:
# -----
# Se ha estudiado el nivel de glucosa en sangre en ayunas en un grupo de
diabéticos.

```

```

# Esta variable se supone que sigue una distribución Normal,
# con media 106 mg/100 ml y desviación típica 8 mg/100 ml.
# Se pide:
# ¿qué porcentaje de diabéticos tienen niveles de glucosa en sangre
comprendidos entre 90 y 130 mg/100 ml?
#  $P[\text{Niveles comprendidos entre } 90 \text{ y } 130] = P[90 \leq X \leq 130] = P[X \leq 130]$ 
-  $P[X \leq 90] = F(130) - F(90)$ 

pnorm(c(130, 90), mean = 106, sd = 8)
0.99865010 - 0.02275013

#####
##### No pierdas de vista las variables en R #####
#####

#-----
# Variables aleatorias discretas y continuas
# funciones de probabilidad y funciones de distribución
#-----

# Reto 1: Utiliza CTRL + SHIFT + ENTER y observa los gráficos

# Reto 2: Utiliza CTRL + ENTER desde el principio y observa paso a paso los resultados

par(mfrow=c(2,3))
# variable aleatoria discreta
Y <- rpois(1000,1)
hist(Y, main='Histograma - Poisson(1)')
# Funciones de probabilidad y distribución

x <- seq(-0.01, 5, 0.01)
plot(x, dpois(x, 1), type = "s", ylab = "P(X=x)", main = "Poisson(1)")
plot(x, ppois(x, 1), type = "s", ylab = "F(x)", main = "Poisson(1)")

# variable aleatoria continua
Y <- rnorm(1000,0,1)
hist(Y, main='Histograma - Normal(0,1)')
# Funciones de probabilidad y distribución
x <- seq(-5, 5, 0.01)
curve(dnorm(x,0,1), xlim = c(-5, 5), ylab = "f(x)", col = "blue",
main='Función de probabilidad')
curve(pnorm(x,0,1), xlim = c(-5, 5), ylab = "F(x)", col = "blue",
main='Función distribución')

```

5.8. Referencias bibliográficas

Amón, J. (1984). *Estadística para Psicólogos. Vol. 2: Probabilidad y Estadística Inferencial*. Madrid: Pirámide.

Lipschutz, S. (1971). *Teoría y problemas de probabilidad*. México: McGraw-Hill.

Martín, A. (2004). *Bioestadística para las ciencias de la salud* (1^a ed.). Madrid: Norma-Capitel.

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

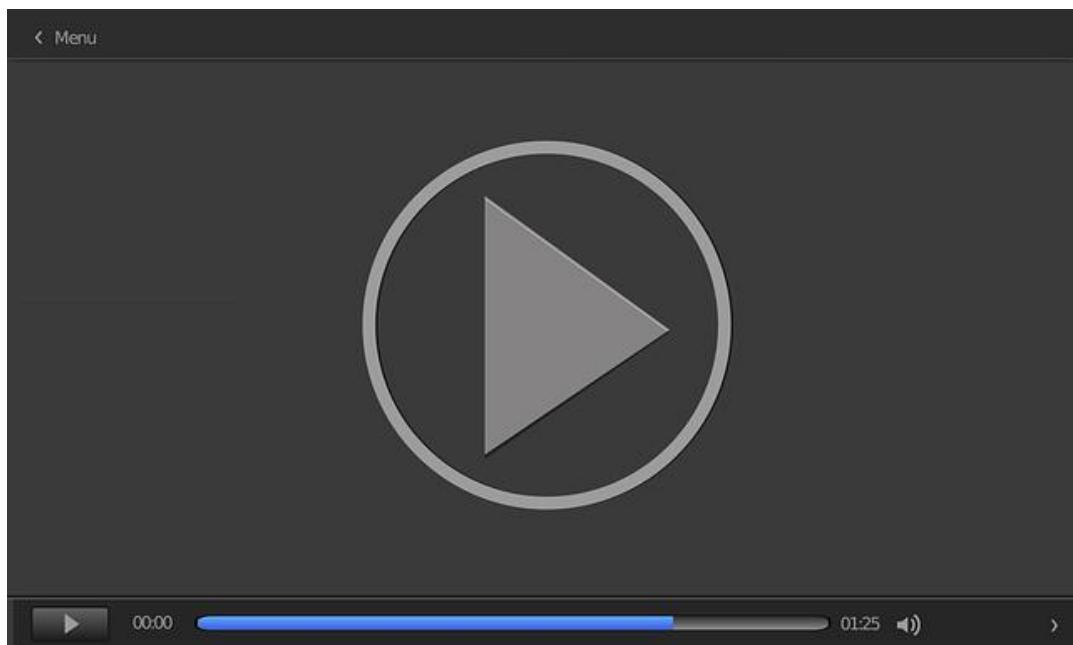
Ríus, F. (1998). *Bioestadística: Métodos y aplicaciones*. Málaga: Universidad de Málaga. Publicaciones.

Versión electrónica:

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

Aproximación de una distribución Binomial a una Normal

En esta lección magistral veremos cómo la distribución binomial y la normal están conectadas, ya que a partir de cierto tamaño del «n» la primera puede aproximarse a la segunda con las consiguientes ventajas que esto tiene.



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=d2a05549-c481-463a-a31a-abdc00f56343>

Modelos de probabilidad

Se recomienda que leas algo más sobre modelos de probabilidad, ya que solo hemos tratado los principales en este capítulo. En esta web de CEACES (un interesante proyecto de la Universidad de Valencia para la enseñanza de la Estadística) tienes información sobre más modelos tanto discretos como continuos. Además puedes descargar el material en formato PDF y también emplear una serie de programas *online* que te permiten calcular las probabilidades de diferentes modelos: binomial, binomial negativo, Poisson y normal.

Accede a la página desde el aula virtual o a través de la siguiente dirección web: <http://www.uv.es/ceaces/base/modelos%20de%20probabilidad/simple.htm>

Bibliografía

Amón, J. (1984). *Estadística para Psicólogos. Vol. 2: Probabilidad y Estadística Inferencial*. Madrid: Pirámide.

Lipschutz, S. (1971). *Teoría y problemas de probabilidad*. México: McGraw-Hill.

Martín, A. (2004). *Bioestadística para las ciencias de la salud* (1^a ed.). Madrid: Norma-Capitel.

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

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

Versión electrónica:

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

1. Si los tres primeros lanzamientos de una moneda han resultado cara, la probabilidad de que obtengamos cara en el cuarto lanzamiento es:

 - A. 1/16
 - B. 1/4
 - C. 1/2
 - D. Otro valor entre cero y uno.

2. Dos eventos se dice que son independientes cuando:

 - A. Siempre que ocurra A no tiene por qué ocurrir B.
 - B. La probabilidad de su intersección es nula.
 - C. El hecho de que ocurra uno no afecta a la probabilidad de que ocurra el otro.
 - D. Las respuestas B y C son correctas.

3. $F(x_i)$ es

 - A. La función de redistribución.
 - B. $1 - P(X > x_i)$
 - C. $P(X > x_i)$
 - D. Las respuestas A y B son correctas.

4. Si tiramos dos dados y sumamos sus puntuaciones. La probabilidad de obtener un 6,7 o un 8 es:

 - A. La mitad.
 - B. Más de la mitad.
 - C. 0,44.
 - D. $F(8) - F(6)$

5. La $V(X+Y)$ es igual a:
- $V(X)+V(Y)+2\rho\sigma_X\sigma_Y$
 - $V(X)+V(Y)$
 - $V(X)+V(Y)+2\text{Cov}(X,Y)$
 - Las respuestas A y C son correctas.
6. En una distribución binomial el parámetro «q» es:
- $1-P(\text{éxito})$
 - La probabilidad del fracaso.
 - $1/p$
 - Las respuestas A y B son correctas.
7. La distribución normal está caracterizada por dos parámetros que son:
- μ y ρ
 - σ^2 y ρ
 - μ y p
 - La media y la desviación típica.
8. Si $X \sim N(5,2)$ entonces la variable tipificada Z la obtenemos como...
- $\frac{X-5}{4}$
 - $\frac{X-2}{5}$
 - $\frac{X-5}{2}$
 - $\frac{Z-E(X)}{\sigma}$

9. Si tenemos un variable aleatoria X que se distribuye como una $Bi(10; 0,5)$:
- A. Su varianza es 2,5.
 - B. $\sigma = 5$
 - C. $E(X) = 2,5$
 - D. $\sigma = 2,5$
10. En una distribución normal a mayor sigma:
- A. Mayor altura de la función de densidad.
 - B. Más probable es encontrarse datos más dispersos que sigan tal distribución.
 - C. La forma de la campana será más achataada.
 - D. Las respuestas B y C son correctas.

Análisis e Interpretación de Datos

Tema 6. Distribución en el muestreo

Índice

[Esquema](#)

[Ideas clave](#)

[6.1. ¿Cómo estudiar este tema?](#)

[6.2. Distribución en el muestreo del conteo y la proporción muestral](#)

[6.3. Teorema Central del Límite y distribución de la media muestral](#)

[6.4. Aplicabilidad del Teorema Central del Límite en ámbitos Big Data](#)

[6.5. Estimulación puntual vs estimulación por intervalos](#)

[6.6. Propiedades de los estimadores](#)

[6.7. Referencias bibliográficas](#)

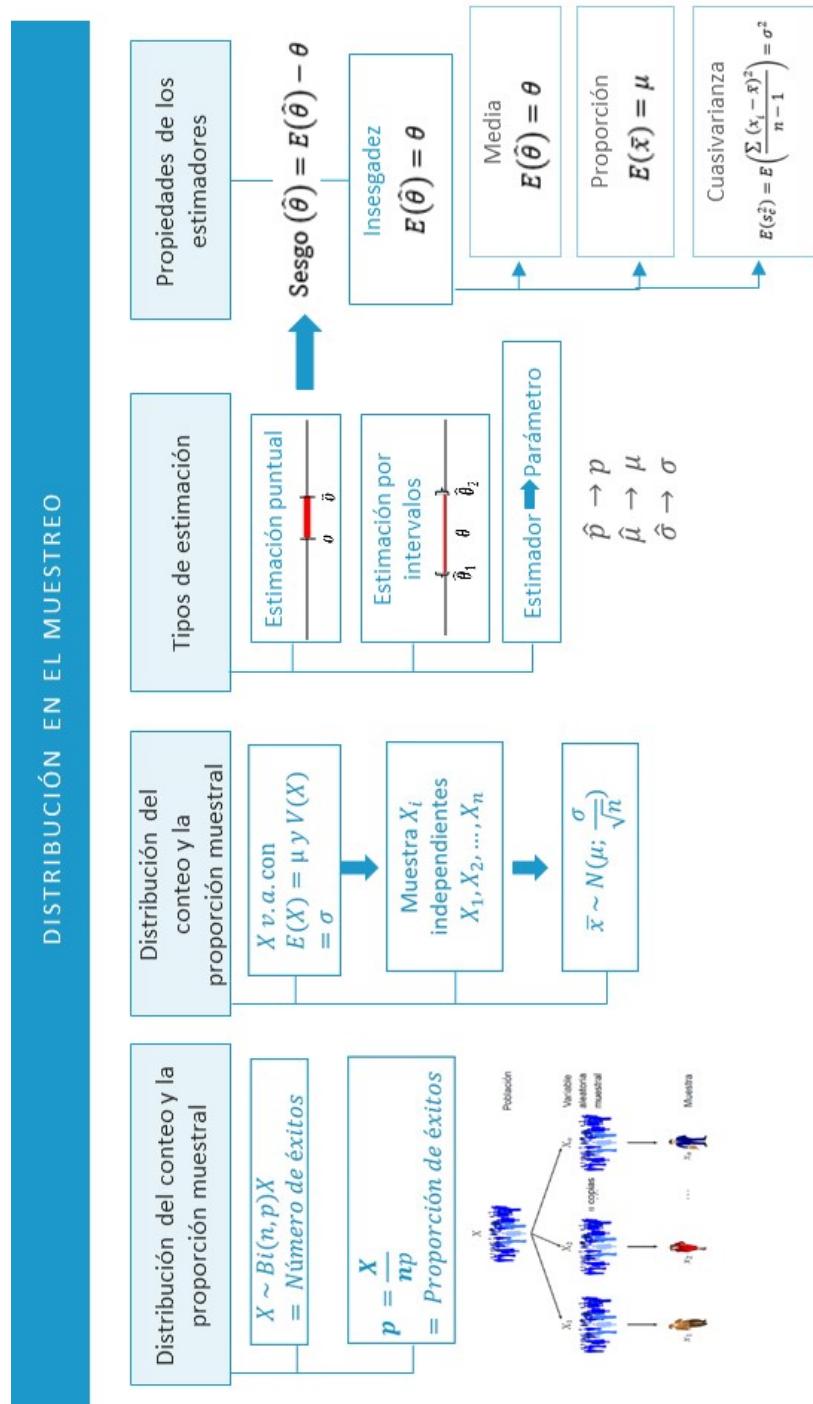
[A fondo](#)

[La distribución de los estadísticos muestrales](#)

[Profundizando sobre estimación puntual](#)

[Bibliografía](#)

[Test](#)



6.1. ¿Cómo estudiar este tema?

Para estudiar este tema lee las **Ideas clave**, además de los intervalos que se indican a continuación: **Páginas 269-277, 280-286 y 291-298** del libro: Triola, M. F. (2009). *Estadística* (10^a ed). México: Pearson. Estos tres fragmentos corresponden aproximadamente a diferentes apartados o aspectos vistos en este tema. **Páginas 169-173** del libro: Ríus, F. (1998). *Bioestadística: Métodos y aplicaciones*. Málaga: Universidad de Málaga Publicaciones.
<https://www.bioestadistica.uma.es/baron/apuntes/clase/apuntes/pdf/>

Para hacerte una idea global de este tema es importante que mires el esquema del tema, el cual te ayudará a hacerte una buena idea de cómo está estructurado y las relaciones que puedan existir entre algunos conceptos clave.

También **será clave que practiques con los ejercicios que vienen al final del tema**. Del mismo modo presta atención a los ejemplos que acompañan a los diferentes apartados a lo largo del tema, pues encierran muchas de las claves que te facilitarán la comprensión del tema.

6.2. Distribución en el muestreo del conteo y la proporción muestral

Hasta ahora nosotros hemos estudiado los modelos teóricos de probabilidad que pueden seguir las distribuciones, pero claro, nosotros generalmente no tendremos acceso a conocer con exactitud tales parámetros.

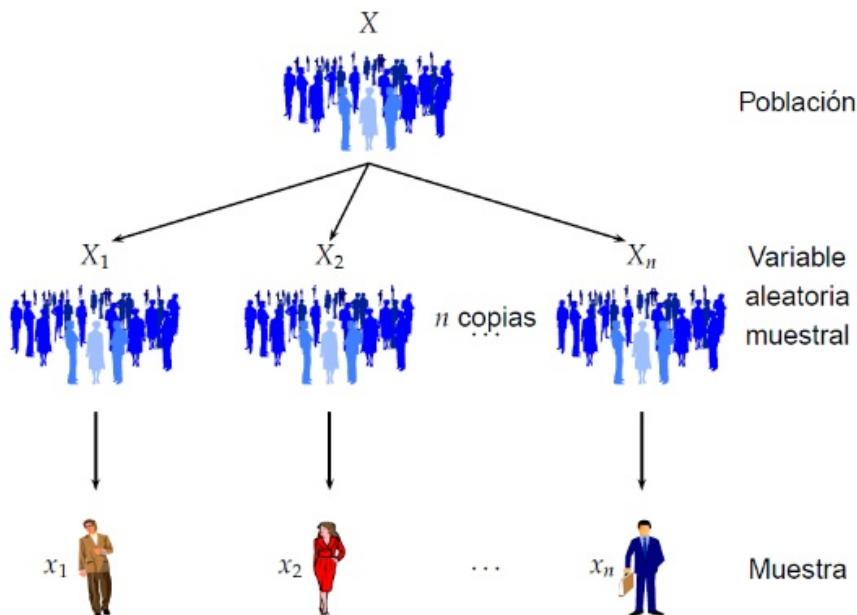
Resulta sencillo modelar el lanzamiento de un dado y saber con certeza que el parámetro asociado a que salga por ejemplo el «6», la proporción de éxitos « p » valga $1/6$, pero lo que resulta imposible de saber a ciencia cierta es la estatura del español medio, por ejemplo. Lo más que podremos hacer será recoger una muestra de españoles, medirlos e inferir que la media de la población será aproximadamente de por ejemplo 1,77 para los hombres y de 1,64 para las mujeres.

Por ello, resultará tan importante desarrollar técnicas inferenciales que nos permitan conocer la verdadera distribución de la población —o lo que es equivalente, conocer los parámetros que la definen— a partir de la distribución de probabilidad de un estadístico que construyamos con su muestra.

Aquí surge la primera definición que manejamos para desarrollar la técnica inferencial, la de **estadístico**, que es una función de la muestra obtenida de una población y entonces hablaremos de **distribución en el muestreo** (o muestral) para referirnos a la **distribución de dicho estadístico** cuando se cumplan dos requisitos:

- ▶ Cada observación X_1, X_2, \dots, X_n de la muestra sigue la misma distribución que la población de donde proviene.
- ▶ Todas las observaciones son independientes entre sí.

La lógica anterior la podemos observar con claridad en este gráfico, en el cual se intenta dar una idea sobre que es la distribución de un estadístico de la muestra, que claro, también será una variable aleatoria en sí.



Ejemplo 1

En el caso binomial, que vimos en el tema anterior, tenemos que para contar el número de seises obtenidos tras « n » lanzamientos lo modelábamos como una $Bi(n, p)$ donde « p » es la probabilidad de éxito en la población, y es la misma que tiene cada observación, es decir, cada lanzamiento. Por otro lado cada lanzamiento es independiente del anterior.

Siguiendo con el caso de la Binomial $Bi(n, p)$ observamos que nos sirve como modelo para la **distribución del conteo** (de «éxitos») **en una muestra** y también para la **distribución de la proporción muestral** (de los «éxitos»).

Así, cuando una población posea una proporción « p » de éxitos para determinado suceso, siempre y cuando la población sea mucho más grande y extraigamos la muestra siguiendo una m.a.s (retrotraerse al capítulo 1) entonces **el número de éxitos de dicha población será X y seguirá aproximadamente una $Bi(n, p)$.**

Ejemplo 2: Los tiros libres de Marc Gasol

En un partido de la NBA Marc Gasol recibió numerosas faltas y llegó a lanzar hasta 13 tiros libres de los cuales falló 5. Los aficionados del Memphis Grizzlies pensaron que no tuvo un buen día y los periodistas deportivos alimentaron esta idea. Para estudiar desde el punto de vista estadístico este asunto tengamos en cuenta que la probabilidad de que un jugador de la NBA falle un tiro libre es de un 25% aproximadamente, de acuerdo a los registros acumulados durante miles y miles de partidos. Teniendo en cuenta esto, el número de fallos cometidos al lanzar 13 tiros libres lo podemos modelar a través de una $Bi(13; 0,25)$. De modo que la probabilidad de que Marc Gasol cometiera 6 o más fallos en un partido resulta:

$$P(X \geq 5) = P(X = 5) + P(X = 6) + \dots + P(X = 13) = 0,206$$

Número de fallos	5	6	7	8	9	10	11	12	13	Total
Probabilidad	0,126	0,056	0,019	0,005	0,001	0,000	0,000	0,000	0,000	0,206

Redondeando las cantidades obtenidas en Excel tenemos que esta probabilidad es cercana a un 21% luego, en absoluto tuvo que ser un mal día pues se encuentra dentro de un porcentaje razonable de fallos si cada cinco partidos aproximadamente resultará uno con esa cantidad o más de fallos.

Del mismo modo que en el ejemplo anterior nos hemos interesado por contar los «éxitos» (que curiosamente en este caso eran fracasos pues recordemos que la concepción de «éxito» hace referencia a la ocurrencia o no de un suceso, que no tiene por tanto que ser un éxito tal y como lo solemos contemplar) podríamos haber querido estimar la **proporción muestral** de éxitos:

$$\hat{p} = \frac{X}{n} = \frac{\text{número de éxitos en una muestra}}{\text{tamaño de la muestra}}$$

Ejemplo 3: Una votación independentista

Imaginemos (aunque se trata de un tema perfectamente posible) que un conocido periódico catalán desea interrogar a sus lectores por medio de una encuesta sobre si están a favor o no de ser independientes de España. Supongamos ahora que en este periódico se asume que el 70% de sus lectores tiene tendencia independentista y por tanto votarían a favor de esta. ¿Cuál sería entonces la probabilidad qué con una muestra aleatoria de 500 lectores se alcance al menos una cifra del 65% a favor de la independencia o mayor?

Para resolver este ejemplo recurriremos en primer lugar a la distribución muestral del conteo que sabemos que es una X que se distribuye como una $Bi(500; 0,7)$, ya que en sí la proporción muestral \hat{P} no sigue una distribución binomial. Como el 65% de 500 son 325 tenemos:

$$P(\hat{p} \geq 0,65) = P(X \geq 325) = P(X = 325) + P(X = 326) + \dots + P(X = 500) = \dots$$

Y como este es un valor tremadamente grande, podemos hacer dos cosas:

Empleamos el Excel para calcularlo de manera cómoda y fácil.

Usamos la aproximación de la normal para una distribución binomial.

Desde luego que podemos hacer lo primero, bastaría con emplear la fórmula del Excel de la binomial y sustraer dicha cantidad a 1 pues:

$$P(X \geq 325) = 1 - P(X \leq 324) = 1 - F(324) = 1 - \text{BINOM.DIST}(324; 500; 0,7; 1)$$

Sin embargo, lo más correcto sería emplear el otro método, pues posiblemente lograremos, entre otras cosas, una mejor aproximación al valor real. El otro método se basa en el Teorema Central del Límite.

6.3. Teorema Central del Límite y distribución de la media muestral

Este teorema es un pilar fundamental para la estadística inferencial. Sin entrar en detalles excesivamente teóricos (y matemáticas complejas que las hay en la base de este teorema) conviene saber que en la naturaleza existen infinidad de variables que pueden considerarse normales.

El **Teorema Central del Límite** (TCL de aquí en adelante) afirma que cuando tenemos n variables independientes X_1, X_2, \dots, X_n (con n suficientemente grande) su suma $X_1 + X_2 + \dots + X_n$ es una variable aleatoria que se distribuye aproximadamente como una normal. Esta aproximación será mejor cuanto mayor sea n .

La explicación de por qué hay tantas variables en la naturaleza que se distribuyen aproximadamente como una normal se debe entonces a este teorema porque serían variables que están compuestas de muchas otras variables independientes entre sí, de manera que la combinación de estas variables resulta en una variables normal.

Ejemplo 4: Del porqué muchas variables médicas se comportan normalmente

Esto lo observamos en variables fisiológicas tales como el «nivel de ácido úrico en sangre» las cuales dependen de una combinación de factores y causas tales como la herencia, el ambiente, la alimentación, etc. Al actuar de modo aditivo e independiente, estas variables harían que el «nivel de ácido úrico en sangre» se comportara normalmente tal y como predice el TCL (Martín Andrés, 2004).

Lo más interesante del TCL va a llegar, no porque permita situar la distribución de la suma de ciertas variables, sino porque esto implica que podemos saber algo sobre un estadístico que suele ser mucho más útil para nosotros: la media muestral \bar{x} , que, de hecho, es función de la suma anterior ya que,

$$\bar{x} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

de modo que,

$$\bar{x} = \frac{X_1}{n} + \frac{X_2}{n} + \dots + \frac{X_n}{n}$$

y, por tanto, también es suma de k cantidades independientes y, entonces, por el TCL y siempre que $n \geq 30$, entonces:

Otra manera de enfocar lo que dice el Teorema es que si extraemos una muestra aleatoria de un tamaño « n » suficientemente grande de cualquier población con media μ y desviación estándar σ , entonces \bar{x} será aproximadamente una.

$$N(\mu; \frac{\sigma}{\sqrt{n}})$$

Antes de ver ningún ejemplo, vamos a aclarar —a modo de comprobación—, que no es fortuito que la media y desviación típica de

$$\bar{X}$$

resulten de este modo, pues:

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{\sum X}{n}\right) = \frac{1}{n} (\mu_{X_1} + \mu_{X_2} + \dots + \mu_{X_n}) = \\ &= \frac{1}{n} (\mu + \mu + \dots + \mu) = \frac{n\mu}{n} = \mu \\ \sigma_{\bar{X}}^2 &= \left(\frac{1}{n}\right)^2 (\sigma_{x_1}^2 + \sigma_{x_2}^2 + \dots + \sigma_{x_n}^2) = \\ &= \left(\frac{1}{n}\right)^2 (\sigma^2 + \sigma^2 + \dots + \sigma^2) = \frac{\sigma^2}{n} \end{aligned}$$

Ejemplo 5

Un analista de datos que trabaja en el departamento del Banco Santander y está encargado de la atención telefónica cogía muestras de 25 llamadas y calculaba con esto la duración media de las llamadas de los clientes. El problema que surgía con esto y del que tardaron en darse cuenta es que la dispersión reflejada por la desviación típica sería muy elevada, pues la media de duración de las llamadas era de 3 minutos y 20 segundos.

Veamos:

$$s_{\bar{x}} = \frac{200}{\sqrt{25}} = 40 \text{ segundos}$$

Así que el analista cuantitativo encargado se dio cuenta de que era razonable tomar muestra al menos 4 veces mayores (de 100 llamadas) pues así la desviación típica de la media se vería considerablemente reducida (a la mitad de hecho):

$$s_{\bar{x}} = \frac{200}{\sqrt{100}} = 20 \text{ segundos}$$

Lo cual supondría una estimación mucho mejor para el tiempo de llamada medio que estaría proporcionando cada muestra de 100 llamadas.

Ejemplo 6

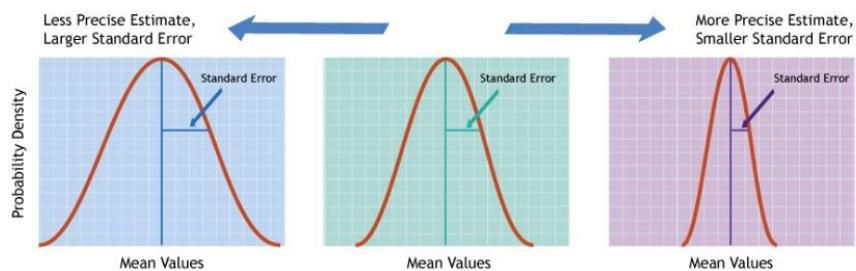
Continuando con el ejemplo anterior ocurrió que un segundo analista observó que tales medias seguían sin ser especialmente precisas, pues por el TCL sabía que la media muestral tenía que comportarse como una normal y, por tanto, —tal y como vimos anteriormente— el 95% de las observaciones aproximadamente se encuentran entre la media \pm 2 veces la desviación típica de modo que el 95% de las muestras de llamadas podían tener la media real entre 200 segundos \pm 40 segundos.

Esto hacía entonces que las medias verdaderas estuvieran contenidas en el 95% de los casos entre 160 y 240 segundos, esto es, llamadas medias de 2 minutos y 40 segundos a 4 minutos, lo cual es un rango muy amplio y, por tanto, poco preciso. De esta manera acordaron que se tomaran muestras lo suficientemente altas como para reducir este 95% a un rango de amplitud de unos 16 segundos. Para ello calcularon que la muestra debía ser igual a 2500 pues así:

$$s_{\bar{x}} = \frac{200}{\sqrt{2500}} = 4 \text{ segundos}$$

Y así para muestras de tamaño 2500 el 95% de las muestras resultarían con una media entre 192 y 208 segundos que ahora sí era un rango que consideraron bastante más preciso.

Gráficamente podemos ver más clara la distribución de la media muestral y como su ancho depende del **error estándar (standard error)** que cuanto menor es hace más precisa la estimación de \bar{x} . La manera de reducir este error es aumentando el tamaño de la muestra pues la varianza de la población siempre será la misma.



6.4. Aplicabilidad del Teorema Central del Límite en ámbitos Big Data

Tal y como hemos podido ver, **el teorema central del límite** asume que **la información con la que estamos trabajando sigue una distribución normal**. Hacer este tipo de asunciones **puede no ser siempre útil** cuando trabajamos con Big Data.

Debido a la complejidad de los datos con los que estamos trabajando, es posible que nos interese estudiar la naturaleza de aquellos datos que no cumplen la normalidad. Cuando trabajamos con conjuntos reducidos de datos, **los outliers vienen determinados por un conjunto muy reducido de información**. Generalmente, su presencia se relaciona con **errores en el proceso de muestreo o extracción de la información**. Los aparatos aplicados para realizar mediciones siempre tienen asociado un pequeño porcentaje de fallo que genera **mediciones erróneas** que se añaden al modelo malogrando el posterior proceso de análisis que debe llevarse a cabo.

Sin embargo, la naturaleza de los outliers en Big Data es distinto. Eso es debido, principalmente a que, debido a la alta cantidad de información disponible, **ya no estamos hablando de un número pequeño de mediciones sino de un conjunto de datos que posee un tamaño considerable**. Incluso formando parte de un porcentaje pequeño de la información global, es interesante estudiar dichos conjuntos de outliers y tratar de determinar su procedencia. Generalmente, bajo este tipo de contextos Big Data, ya no estamos hablando de fallos en las mediciones o de datos erróneos sino de **subconjuntos en las poblaciones de datos que cumplen una serie de propiedades diferentes a las de la media**. Por tanto, desechar por completo estos outliers implicaría eliminar una parte de la población lo que claramente produciría una **pérdida importante de la información** asociada a un sector de la población.

Esto, por supuesto, **no quiere decir que asumiendo normalidad en los datos estemos realizando análisis erróneos**. Únicamente estamos afirmando que dichos análisis **no permitirían obtener un análisis 100% completo** de la población estudiada. Es, por tanto, necesario un análisis exhaustivo de los datos con los que estamos trabajando antes de diseñar un proceso concreto de análisis de la información. De esta forma, podremos **sacar el máximo partido a nuestros datos** y aumentar la precisión de los resultados de los análisis.

Uno de los métodos que más se están empezando a utilizar en este tipo de análisis de datos complejos es lo que llamamos la **estadística robusta**. La estadística robusta es un nuevo campo dentro de la estadística cuyo objetivo es el **desarrollo de métodos de análisis estadístico que no se vean influenciados por los outliers**. De esta forma, mediante el uso de estos métodos, podemos llevar a cabo un análisis mucho más fiable de conjuntos de datos complejos como los que nos encontramos en entornos Big Data. De esta forma, utilizaríamos estos métodos para realizar un **análisis global que incluyera a toda la información** para luego complementar dicha información mediante la realización de **análisis exhaustivos** de diversos subconjuntos de la población (incluyendo outliers).

En resumen, si ya de por sí era muy importante en la estadística clásica **conocer la estructura, naturaleza y procedencia de los datos** antes de realizar cualquier análisis, con la aparición de los entornos Big Data este proceso es aún más importante y crítico si queremos obtener buenos resultados. Por desgracia, **no existen técnicas globales ni metodologías mágicas que puedan aplicarse a todos los conjuntos de datos ni a todos los posibles casos**. El único secreto consiste en **aprender a conocer y estudiar nuestros datos**: de qué población provienen, qué representan, qué puede provocar la aparición de distintas subpoblaciones, etc. Si sabemos de dónde provienen nuestros datos y, mediante **análisis preliminares**, identificamos su estructura general y específica, es posible determinar qué provoca los outliers y cuál serían los procesos de análisis más adecuados para cumplir el objetivo u objetivos propuestos.

6.5. Estimación puntual vs estimación por intervalos

Por lo general, podemos realizar estimaciones de dos formas distintas:

- ▶ **Estimador puntual** que el valor que da como estimación es único.
- ▶ **Estimación por intervalos** (o confidencial) que estima a través de un intervalo de confianza.

El tipo de estimación que estamos tratando hasta ahora sería la puntual, ya que será en el tema siguiente cuando veamos la estimación mediante intervalos.

Podemos observar en el siguiente gráfico cómo en la estimación puntual (caso a)) se estima mediante un solo valor, mientras que en la estimación por intervalos (caso b)) se requieren dos: el límite inferior y el superior de dicho intervalo.

Estimación puntual



Estimación por intervalos

6.6. Propiedades de los estimadores

Vamos a responder ahora a la cuestión: ¿cuál es el estimador más deseable para estimar un parámetro? Para ello, veamos las propiedades que puede tener un estimador que luego nos permita discernir cuál es ese estimador «deseado».

Si observamos el gráfico del punto anterior, al tratar de estimar puntualmente cometemos un error, pues no logramos acertar absolutamente con el parámetro, sino que cometemos un **sesgo en la estimación del parámetro**.

Este concepto de sesgo nos va a conducir a la primera propiedad deseable para un buen estimador:

Diremos que un estimador es **insesgado** para un parámetro cuando
«tienda» a producir estimaciones sin sesgo para dicho parámetro.

Ese «tienda», a nivel matemático nos va a obligar a que su valor esperado sea el parámetro que pretende estimar, es decir:

$$\text{Sesgo}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

$$E(\hat{\theta}) = \theta$$

Ejemplo 7

Haciendo ya referencia a estimadores concretos, podemos afirmar que la media muestral es un buen estimador entre otras cosas porque es un estimador insesgado de la media poblacional.

$$E(\bar{x}) = \mu$$

Del mismo modo el estimador de la proporción muestral es también un estimador insesgado de la proporción poblacional.

$$E(\hat{p}) = p$$

Vamos a aprovechar aquí para explicar **un estimador que es insesgado para la varianza poblacional**. Se podría pensar a priori que la varianza que vimos en la parte descriptiva nos puede servir como buen estimador, esto es, como estimador insesgado, pero no es así:

$$E(s^2) = E\left(\frac{\sum (x_i - \bar{x})^2}{n}\right) \neq \sigma^2$$

Si no que es otro estadístico que llamamos **cuasivarianza muestral** (s_c^2 el que es insesgado para la varianza poblacional):

$$E(s_c^2) = E\left(\frac{\sum (x_i - \bar{x})^2}{n - 1}\right) = \sigma^2$$

En realidad a nivel inferencial se emplea más la cuasivarianza que la varianza.

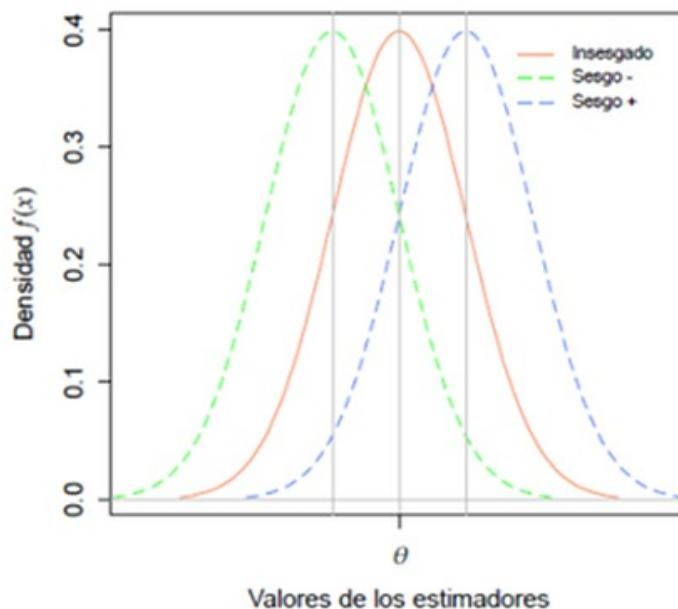
Para calcular la cuasivarianza a partir de la varianza hacemos un sencillo cálculo:

$$s_c^2 = \frac{n}{n - 1} s^2$$

Observamos por tanto que con la cuasivarianza hacemos estimaciones un tanto más grandes que con la varianza ya que:

$$\frac{n}{n-1} > 1$$

Distribuciones de estimadores sesgados e insesgados



Alguien podría plantearse la siguiente cuestión: ¿Y si encontramos dos estimadores insesgados para un mismo parámetro, cuál elegimos?

Esto sucede, por ejemplo, con la media muestral y la mediana, ya que la mediana también es un estimador insesgado de la media poblacional.

Para poder salir de este embrollo parece razonable exigir que aparte de ser insesgado nos produzca valores con poca dispersión, que no varíe mucho el valor de dicho estimador, que no se aleje en exceso del valor del parámetro.

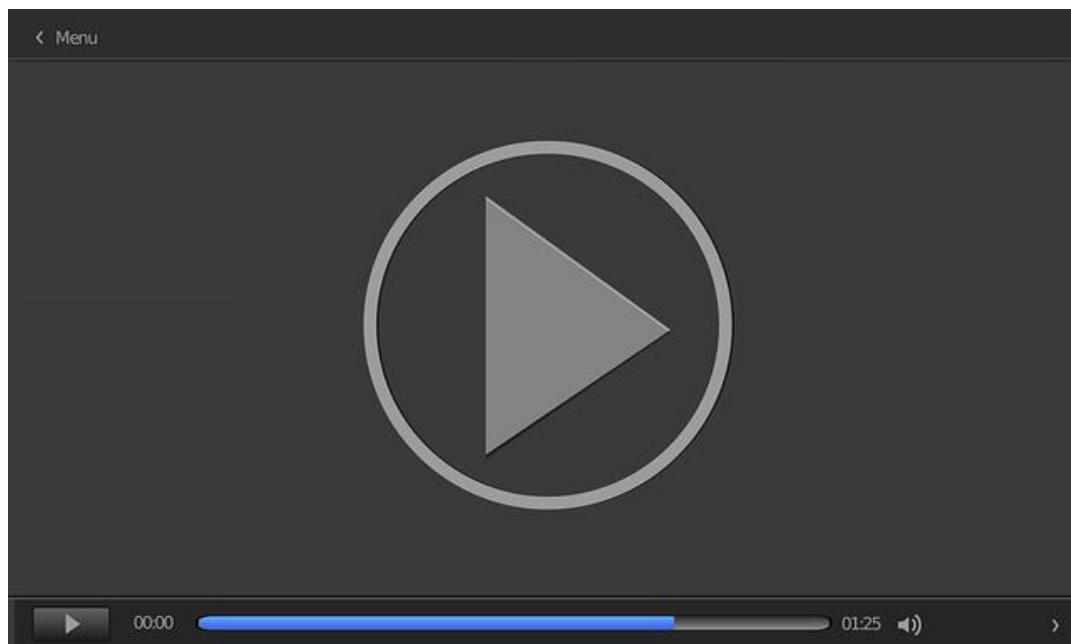
Luego, la siguiente propiedad deseada es la de encontrar el **estimador de varianza mínima**. Lo bueno de esta propiedad es que nos asegura (hay un teorema que lo garantiza) que de existir tal estimador de varianza mínima es único.

En el caso del estimador de mínima varianza para la media poblacional encontramos que es la media muestral el que presenta mínima varianza.

Existen otras propiedades deseables en un estimador como son la consistencia, eficiencia y suficiencia. Sin embargo, con las propiedades que hemos manejado por sí solas es suficiente para hacerse una buena idea de que criterio emplear para elegir un estimador y no otro.

Apuntes sobre el Teorema Central del Límite

En este vídeo veremos algunos apuntes sobre este teorema.



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=46a8177b-2def-4781-96e7-acbd00c77a18>

6.7. Referencias bibliográficas

Martín, A. (2004). *Bioestadística para las ciencias de la salud* (1^a ed.). Madrid: Norma-Capitel.

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

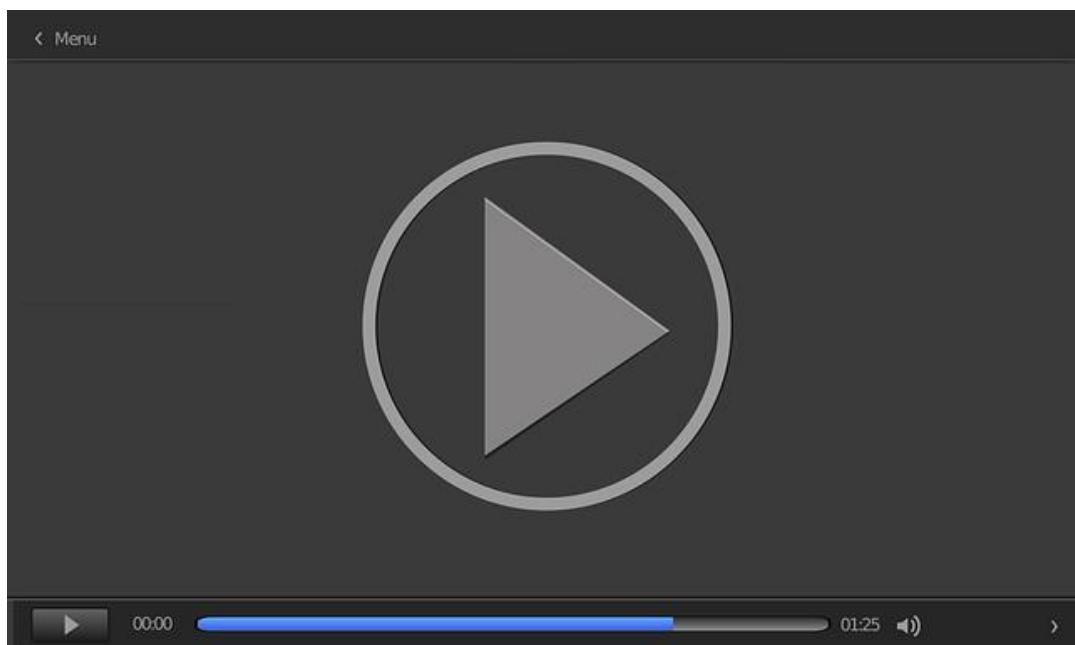
Ríus, F. (1998). *Bioestadística: Métodos y aplicaciones*. Málaga: Universidad de Málaga. Publicaciones. Versión electrónica:

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

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

La distribución de los estadísticos muestrales

En esta lección magistral veremos sirviéndonos de un *applet* bastante sofisticado cómo interpretar el concepto de distribución de un estadístico muestral, como es la media muestral de una población normal.



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=45d18d0c-0d31-4826-956d-abdc00f5af54>

Profundizando sobre estimación puntual

En este enlace de la página de CEACES sobre Estadística encontrarás más propiedades de los estimadores además de la técnica de generación de estimadores conocida como de máxima verosimilitud.

Accede al artículo desde el aula virtual o a través de la siguiente dirección web:

<http://www.uv.es/ceaces/tex1t/4%20estimacion/simple.htm>

Bibliografía

Martín, A. (2004). *Bioestadística para las ciencias de la salud* (1^a ed.). Madrid: Norma-Capitel.

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

Ríus, 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* (10^a ed). México: Pearson.

1. Para poder hablar propiamente de distribución muestral de un determinado estadístico...

 - A. Tendremos que recoger una muestra de una población cualquiera.
 - B. Las observaciones que se recojan habrán de ser independientes.
 - C. Las observaciones recogidas deben seguir la misma distribución que es la de la población de donde provienen.
 - D. Las respuestas B y C son correctas.
2. El arco circunflejo lo empleamos en estadística inferencial para...

 - A. Designar un estadístico cualquiera que provenga de la muestra.
 - B. Diferenciar el estadístico del parámetro poblacional.
 - C. Indicar que es un estimador del parámetro.
 - D. Las respuestas B y C son correctas.
3. \hat{p}

 - A. El estimador de la probabilidad.
 - B. El estimador de la proporción muestral.
 - C. El estimador de la proporción poblacional.
 - D. Un parámetro de cierto tipo de variable aleatoria.

4. ¿Por qué hay tantas variables en la naturaleza que se distribuyen normalmente?
- A. Porque según el TCL la suma de muchos efectos aleatorios independientes se comporta normalmente.
 - B. En realidad no hay tantas, es más la visión alterada del matemático que trata de que la realidad se parezca a las matemáticas y no a la inversa.
 - C. En realidad es por las binomiales que son las que abundan más y como su suma es normal acaba apareciendo que hay más normales.
 - D. Se debe a un misterio estadístico todavía por resolver.

5. La desviación típica de la media muestral es:

A. $\frac{\sigma}{\sqrt{n}}$

B. $\frac{\mu}{\sqrt{n}}$

C. $\frac{\sigma}{n}$

D. σ

6. Si multiplicamos por cuatro el tamaño de una muestra:

$$\sigma_{\bar{x}}$$

- A. Se reduce a la mitad.
- B. Se duplica.
- C. Se multiplica por 4.
- D. Se mantiene igual pues no le afectan cambios de n.

7. El error estándar es...

- A. Un error típico que se comete cuando estimamos.
- B. La desviación típica de la media.
- C. Proporcional al ancho de la curva de la distribución muestral de la media.
- D. Las respuestas B y C son correctas.

8. Señala la frase correcta.

- A. Todo estadístico es un estimador.
- B. Todo estimador es una variable aleatoria función de la muestra.
- C. Todo estimador es un parámetro de la muestra.
- D. Todo estadístico es un parámetro de la muestra.

9. ¿Cuál es un estimador insesgado de la media poblacional?

- A. \bar{x}
- B. $\hat{\mu}$
- C. Las respuestas A y B son correctas.
- D. μ

10. ¿En qué se diferencian los dos tipos principales de estimación?

- A. Una estima con un solo valor mientras que la otra con varios.
- B. Un tipo de estimación da intervalos posibles para el parámetro mientras que la otra solo da un valor.
- C. Una es confiente y la otra puntual.
- D. Depende del parámetro a estimar tendremos que usar una u otra.

Análisis e Interpretación de Datos

Tema 7. Intervalos de confianza

Índice

Esquema

Ideas clave

- 7.1. ¿Cómo estudiar este tema?
- 7.2. Introducción a los intervalos de confianza
- 7.3. Intervalo de confianza para la media de una población normal: varianza conocida y desconocida
- 7.4. Calculando el tamaño de la muestra
- 7.5. Intervalo de confianza para la proporción
- 7.6. Intervalo de confianza para la varianza de una población normal
- 7.7. Intervalo de confianza para la diferencia de medias y proporciones
- 7.8. Intervalos de confianza robustos
- 7.9 Referencias bibliográficas

A fondo

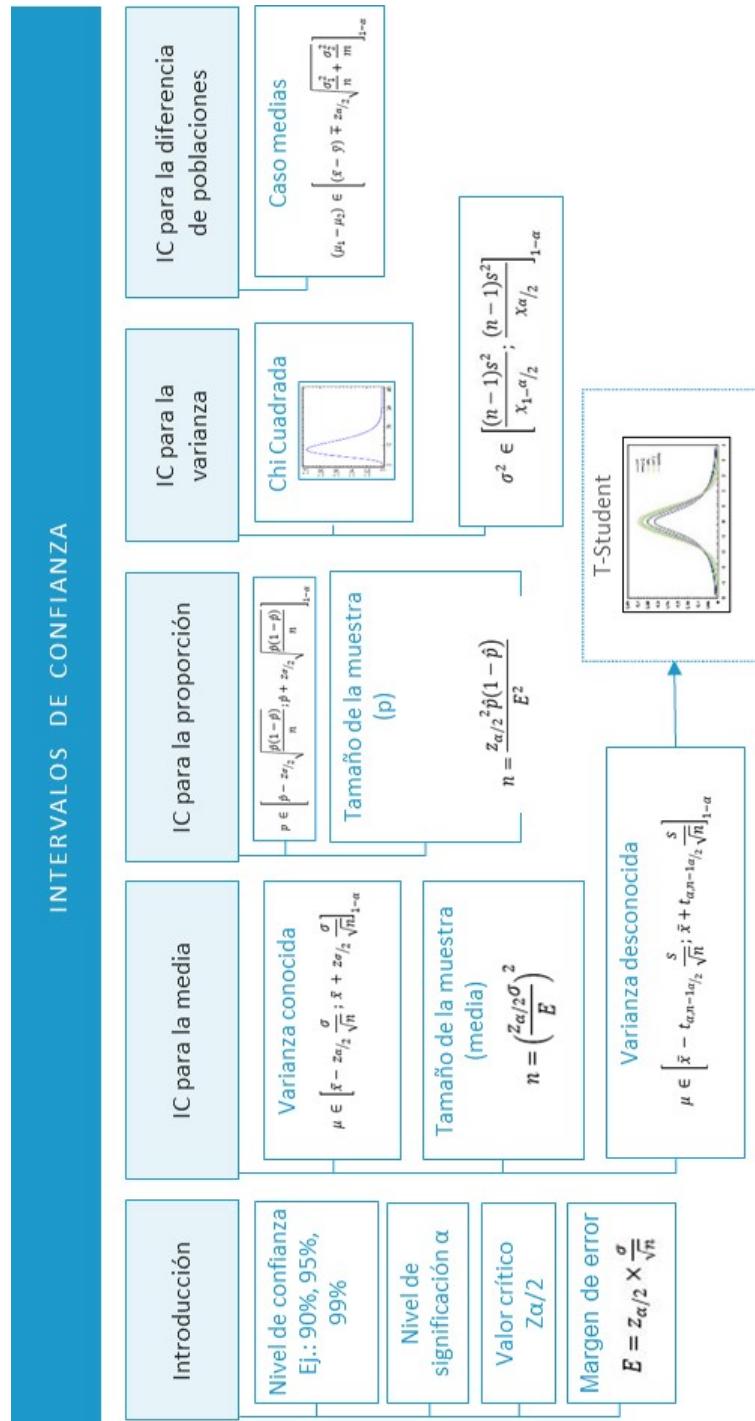
Buscando los valores críticos en las tablas de diferentes distribuciones

Un estadístico entre cervezas negras

Calculadoras online de las principales distribuciones de probabilidad

Bibliografía

Test



7.1. ¿Cómo estudiar este tema?

Para estudiar este tema lee las **Ideas clave**, además de los intervalos que se indican a continuación: **Páginas 320-331, 338-345 y 349-354** del libro: Triola, M. F. (2009). *Estadística* (10^a ed). México: Pearson. Estos tres fragmentos corresponden aproximadamente a diferentes apartados o aspectos vistos en este tema. **Páginas 175-199** del libro: Ríus, F. (1998). *Bioestadística: Métodos y aplicaciones*. Málaga: Universidad de Málaga. Publicaciones. <https://www.bioestadistica.uma.es/baron/apuntes/clase/apuntes/pdf/>

Para hacerte una idea global de este tema es importante que mires el esquema del tema, el cual te ayudará a hacerte una buena idea de cómo está estructurado y las relaciones que puedan existir entre algunos conceptos clave.

También **será clave que practiques con los ejercicios que vienen al final del tema**. Del mismo modo presta atención a los ejemplos que acompañan a los diferentes apartados a lo largo del tema, pues encierran muchas de las claves que te facilitarán la comprensión del capítulo.

7.2. Introducción a los intervalos de confianza

En el tema anterior planteamos el método de estimación puntual y, si bien se puede tener cierta utilidad, tiene una limitación seria. Pongamos que queríamos calcular la verdadera proporción de una población, p . Para ello, cogíamos una muestra y calculábamos su estimador de la proporción \hat{p} resultando que valdría pongamos 0,4.

¿Y esta es una buena estimación del intervalo? la respuesta es que no tenemos ni idea, **no es posible saber si una estimación puntual es buena o mala**, si se aleja poco o mucho del parámetro poblacional que pretendemos estimar, pues podríamos habernos topado con una «mala» muestra, de modo que ese valor no reflejase en absoluto al parámetro. Y no lo va a reflejar sobretodo porque no tenemos en cuenta la distribución probabilística del propio estimador, a través de la cual podríamos otorgarle el valor necesario y preciso a dicha estimación.

Por lo dicho anteriormente era necesario otro enfoque donde se diera un margen para situar al parámetro con cierta seguridad, teniendo en cuenta precisamente las desviaciones naturales del estimador. Los márgenes no serán otros que los que marquen los límites inferior y superior de un **intervalo de confianza**, denominado así porque su amplitud dependerá de la confianza que deseemos tener en que contenga realmente al parámetro.

Gracias al uso de los intervalos de confianza podremos asegurar con ciertas garantías que nuestra estimación no estará muy lejos del valor real. De este modo, ahora podemos ser más realistas en nuestras estimaciones afirmando, por ejemplo: «No sé con total seguridad, a partir de las muestras que manejo, cuál es la proporción de la población, pero estoy "casi seguro" de que rondará entre 0,35 y 0,51». Date cuenta de la diferencia conceptual en la afirmación basada en un intervalo frente a la puntual anterior.

Ese «casi seguro» del párrafo anterior se refiere precisamente a la idea de manejar un cierto nivel de seguridad que denominaremos **nivel de confianza** del intervalo y que está relacionado con la probabilidad de que efectivamente el parámetro este contenido en dicho intervalo de confianza.

Ejemplo 1

Los niveles de confianza más habituales que se manejan son altos, del estilo de 0,9; 0,95; 0,99 o incluso mayores, porque no se precisará manejar el mismo nivel de confianza si estamos realizando una encuesta política (sin ánimo de restarle importancia), que si estamos construyendo un puente o detectando los niveles a partir de los cuales se estable que un testo de detección de cáncer de positivo.

Los intervalos de confianza se suelen abreviar como IC acompañados de su respectivo nivel de confianza, el cual se expresa en ocasiones en modo porcentual. Así hablamos de **IC al 95%**, por ejemplo. La cantidad de probabilidad complementaria a nuestro nivel de confianza es llamada **nivel de significación** y la anotamos como α . De este modo el IC se anota genéricamente IC al $(1 - \alpha) \times 100\%$. Aquí vemos que el nivel de significación y el de confianza son complementarios.

El nivel de significación puede interpretarse también como el nivel de error que estamos dispuestos a asumir, en el sentido de que éste es precisamente la probabilidad de que el parámetro no esté contenido en el intervalo que construimos.

Los IC se construyen como apuntábamos en el tema anterior con dos valores estimados que configuran los límites a y b de un intervalo. De esta manera el $1 - \alpha$ será la probabilidad de que el parámetro este contenido entre a y b :

$$P[a \leq \theta \leq b] = 1 - \alpha$$

Ejemplo 2: Interpretando un intervalo de confianza

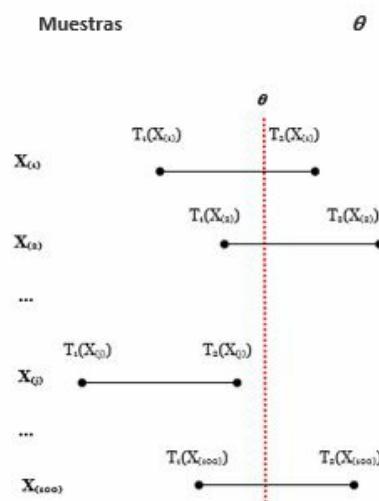
También es importante saber cómo interpretar un IC. Así, diremos que si el IC al 95% para la media de estatura de los españoles es el siguiente:

$$P[167\text{cm} \leq \mu \leq 192\text{cm}] = 0,95$$

Que lo interpretamos como sigue: «**de cada 100 muestras podemos afirmar que al menos 95 contendrán a dicho parámetro, mientras que 5 no lo contendrán**».

También podemos comentarlo así: «Tenemos un nivel de confianza del 95% de que el intervalo (167; 192) contenga a la media poblacional».

Esto lo podemos visualizar del siguiente modo:



Donde los $T_1(X_{(i)})$ y $T_2(X_{(i)})$ hacen referencia a los estadísticos empleados para construir los límites inferior o superior pero en formato de variables aleatorias, que son función de la muestra $X_{(i)}$, ya que es lo que realmente son; cuando se tiene una muestra concreta es cuando se convierten en límites concretos de un intervalo.

Tras caracterizar y definir los componentes de un IC, ya podemos pasar a estudiar la construcción de los principales IC para los parámetros más usuales, que ya manejamos en la estimación puntual: μ, p y α

7.3. Intervalo de confianza para la media de una población normal: varianza conocida y desconocida

Para hallar el **IC para la media μ de una población $N(\mu; \sigma)$ con σ conocida**, a un nivel de confianza $1 - \alpha$ razonamos como sigue. Vamos a ilustrar el proceso completo por tratarse del primer caso que se muestra.

$$X \text{ se distribuye con una } N(\mu; \sigma) \Rightarrow X: (x_1, \dots, x_i, \dots, x_n) \Rightarrow \bar{x} = \frac{\sum x_i}{n}$$

Por otro lado sabemos que \bar{x} será un estimador puntual de μ . Sabemos por el Teorema Central del Límite visto anteriormente, que si X es normal y cogemos una muestra aleatoria, entonces:

$$\bar{X} \sim N\left(\mu; \frac{\sigma}{\sqrt{n}}\right)$$

Comenzaríamos por definir el estadístico que va a conformar ambos límites del intervalo a construir:

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \rightarrow N(0; 1)$$

El hecho de que se use este estadístico y no otro es porque conocemos su distribución, que proviene de la tipificación de una normal cualquiera a una normal estándar. Además, ya vimos que cuando tratamos las variables normales, el 95% de las observaciones estaba comprendido entre $\mu + / - 2\sigma$ aproximadamente, lo cual está relacionado con el área que encierra la distribución normal entre dos desviaciones a un lado y a otro de la media, así que en cierto modo este era un concepto que ya hacía uso de la lógica de los IC a través del siguiente intervalo $(\mu - \sigma; \mu + \sigma)$.

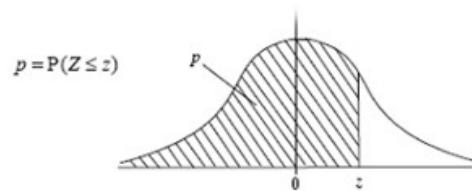
A continuación tendríamos que fijar el nivel de confianza, $1 - \alpha$, que hará que tomen un valor u otro los límites del IC que son los valores $-Z_{\alpha/2}$ y $Z_{\alpha/2}$ de la Normal estándar $N(0, 1)$ tales que:

$$P \left[-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \right] = 1 - \alpha$$

Conviene recordar, llegados a este punto, que la normal estándar, en las llamadas **tablas de la normal estándar**, simboliza z_α como el valor que deja a su derecha una probabilidad de α y que, por tanto, deja a su izquierda el complementario, esto es, $1 - \alpha$.

La cuestión de por qué entonces nuestros límites del intervalo $-z_{\alpha/2}$ y $z_{\alpha/2}$ presentan el $\alpha/2$ como subíndice, es porque el IC de confianza lo construimos de modo simétrico para que deje la misma probabilidad a ambos lados o colas de la función de probabilidad (ya que $\alpha/2$ por la izquierda + $\alpha/2$ por la derecha = α). El valor $z_{\alpha/2}$ que marca el límite del IC recibe el nombre de **valor crítico**.

A continuación podemos apreciar un fragmento de la tabla de la $N(0, 1)$ que contiene los valores de la variable que acumulan para esta distribución una determinada probabilidad, que en nuestro caso es $1 - \alpha/2$ (ya que deja a su derecha un $\alpha/2$) y que resulta muy útil para la confección de IC (y también en su momento para los Contrastes de Hipótesis).



z	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879

Ejemplo 3

Si nos piden hallar el IC al 99% de confianza de un determinado parámetro, lo primero que tenemos que hacer es descubrir el valor crítico que corresponde a $\alpha/2 = 0,005$, para lo cual inspeccionamos en la tabla dicho valor para descubrir que es 2,57 (lo cual no puedes ver en la tabla de arriba pues es un fragmento y no está completa).

Los valores críticos más comunes que se emplean son:

Nivel de confianza	α	$\alpha/2$	valor crítico
90%	0,1	0,05	1,65
95%	0,05	0,025	1,96
99%	0,01	0,005	2,57

Una vez ya sabemos hallar su nivel de significación, siguiendo con los pasos para construir el IC, se sustituye la variable aleatoria Z por su expresión como estadístico:

$$P \left[-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \right] = 1 - \alpha$$

Entonces operamos en las dos desigualdades para despejar el parámetro μ en la parte central ya que al que nos interesa situar dentro del intervalo. Finalmente obtenemos:

$$\mu \in \left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]_{1-\alpha}$$

Ejemplo 4

Cuando el IC sea al 95% de confianza, acudiendo a las tablas de la normal (0,1) observaremos que $z\alpha/2 = 1,96$, por lo que el intervalo resulta:

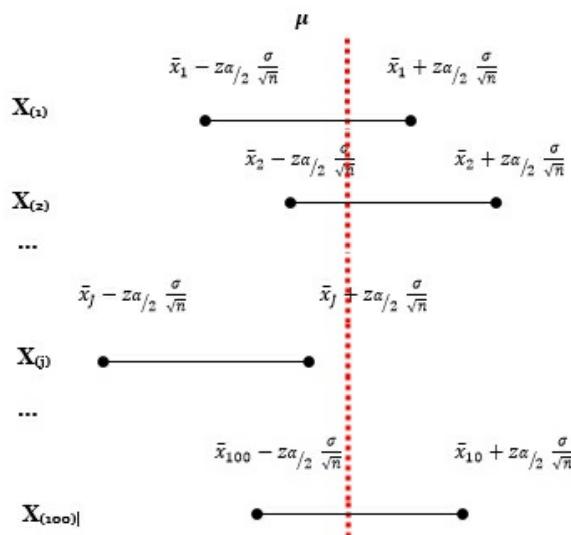
$$\mu \in \left[\bar{x} - 1,96 \frac{\sigma}{\sqrt{n}} ; \bar{x} + 1,96 \frac{\sigma}{\sqrt{n}} \right]_{0,95}$$

O lo que es equivalente:

$$P \left[\bar{x} - 1,96 \frac{\sigma}{\sqrt{n}} ; \bar{x} + 1,96 \frac{\sigma}{\sqrt{n}} \right] = 0,95$$

De nuevo recurro a la representación gráfica anterior para ilustrar la idea que hay detrás de este IC del cual acabamos de mostrar su construcción.

Muestras μ



Esta gráfica la interpretamos como que en un porcentaje del $\alpha\%$ de muestras el intervalo no contendrá al parámetro, ya que precisamente **α es** en este sentido **el error que estamos dispuestos a asumir en nuestro IC**.

$$E = z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

Es el llamado **margen de error** del parámetro. Así podemos expresar abreviadamente el IC del modo siguiente:

$$\mu \in [\bar{x} - E; \bar{x} + E]_{0.95}$$

O directamente:

$$(\bar{x} - E; \bar{x} + E)$$

Por último, conviene aclarar que no son correctas expresiones como la siguiente: «el parámetro μ ha caído dentro del intervalo...», pues el parámetro siempre es el que es (no lo conocemos, pero eso no quiere decir que no exista y tenga un valor concreto) y las variables, que dependen de las muestras concretas, son en realidad las que configuran el IC por medio precisamente de los estadísticos. Por tanto, no sería correcto decir algo así. Serían en todo caso los límites del IC los que caen donde caen, precisamente por tomar un valor concreto debido a una muestra particular.

En el primer caso que hemos tratado antes suponíamos conocida σ pero ahora vamos a partir de que **σ no es conocida**, lo que supone un caso más realista, ya que si desconocíamos μ sería extraño que conociéramos la desviación típica, más si cabe porque la obtenemos a partir de la media.

Ahora el estadístico a emplear para el contraste ya no se distribuye normalmente, sino que **sigue una distribución similar a la normal** que es la llamada T-Student.

La diferencia fundamental de esta distribución es que depende de un parámetro que son los grados de libertad (sin considerar el parámetro α), los cuales provienen del número de variables independientes que la conforman, esto es, de « n » que es el tamaño de la muestra. Al número de grados de libertad (GL de aquí en adelante) se le resta uno, pues el hecho de que la media ya tenga un valor concreto restringe las $n - 1$ variables restantes.

Otra diferencia para confeccionar el IC cuando se desconoce la varianza es que tendremos que utilizar otro valor que la suplante, y qué mejor valor para esto que su estimador muestral que es como vimos en el tema anterior la cuasivarianza muestral.

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

De este modo el nuevo IC resulta ser el siguiente (donde ya anotamos la cuasivarianza con una simple «s» prescindiendo del subíndice «c» porque en realidad a nivel inferencial siempre hacemos ya referencia a esta variante de la varianza para la estimación y no empleamos la «s» de estadística descriptiva donde se dividía entre «n» y no entre «n-1» como es el caso):

$$\mu \in \left[\bar{x} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}; \bar{x} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \right]_{1-\alpha}$$

Para saber cuál es el valor de la $t_{\alpha/2, n-1}$ tendremos que acudir a la tabla que se confecciona para los valores de la T-Student para un α dato y según sus G.L.

TABLA N° 4 DISTRIBUCIÓN t DE STUDENT									
$\frac{\alpha}{n-1}$	0,25	0,2	0,15	0,1	0,05	0,025	0,01	0,005	0,0005
1	1,0000	1,3764	1,9626	3,0777	6,3138	12,7062	31,8205	63,6567	636,6192
2	0,8165	1,0607	1,3862	1,8856	2,9200	4,3027	6,9646	9,9248	31,5991
3	0,7649	0,9785	1,2498	1,6377	2,3534	3,1824	4,5407	5,8409	12,9240
4	0,7407	0,9410	1,1896	1,5332	2,1318	2,7764	3,7469	4,6041	8,6103
5	0,7267	0,9195	1,1558	1,4759	2,0150	2,5706	3,3649	4,0321	6,8688
6	0,7176	0,9057	1,1342	1,4398	1,9432	2,4469	3,1427	3,7074	5,9588
7	0,7111	0,8960	1,1192	1,4149	1,8946	2,3646	2,9980	3,4995	5,4079
8	0,7064	0,8889	1,1081	1,3968	1,8595	2,3060	2,8965	3,3554	5,0413
9	0,7027	0,8834	1,0997	1,3830	1,8331	2,2622	2,8214	3,2498	4,7809
10	0,6998	0,8791	1,0931	1,3722	1,8125	2,2281	2,7638	3,1693	4,5869
11	0,6974	0,8755	1,0877	1,3634	1,7959	2,2010	2,7181	3,1058	4,4370

Ejemplo 5: Un IC para la estatura media de las mujeres policía

Para determinar la estatura media de las policías del Cuerpo Nacional de cara establecer un intervalo para el examen de entrada al Cuerpo, se tomó una muestra aleatoria de 10 mujeres resultando: 152, 166, 159, 155, 161, 159, 162, 158, 157, y 165 cm de estatura.

Para hallar ahora el valor de la altura media de las mujeres con un nivel de confianza del 95% hacemos lo que sigue.

Lo primero que haríamos sería identificar el estadístico que vamos a emplear. Vamos a suponer que al tratarse de la estatura será un v.a. normal pero con la varianza desconocida, de modo que emplearemos la fórmula anterior. Así tendremos que hallar la media y la cuasivarianza muestral:

$$\bar{x} = \frac{166+159+\dots+165}{10} = 159,40 \text{ cm}$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \rightarrow s = 4,30$$

Luego acudiendo a la tabla de T-Student sabemos que para un $\alpha = 0,025$ tenemos $t_{0,025;9} = 2,26$ (podemos incluir los dos parámetros como subíndice o solo el «n-1» es indistinto siempre que indiquemos el α). Ya tenemos todos los elementos y podemos construir nuestro intervalo para la media de la estatura:

$$\mu \in \left[159,40 - 2,26 \frac{4,30}{\sqrt{10}}; 159,40 + 2,26 \frac{4,30}{\sqrt{10}} \right]_{0,95} \rightarrow \mu \in (156,32; 162,47)$$

Así, interpretaríamos que de cada 100 muestras en 95 de ellas la media de las mujeres policía se encontrará entre 156,32 y 162,47 cm o bien que tenemos una confianza del 95% de que la media de estatura de las mujeres se encuentra entre 156,32 y 162,47 cm.

7.4. Calculando el tamaño de la muestra

Antes de proseguir enumerando los diferentes casos de estimación por IC conviene hacer un parón y sacar otra gran utilidad que nos aporta el uso de IC, que es la estimación del tamaño de muestra adecuado para un nivel de significación dado.

La pregunta que nos hacemos entonces es: **¿qué tamaño de muestra debo tener para asegurar una precisión determinada en el intervalo?**

Un ejemplo de esto sería: ¿qué cantidad de encuestas tengo que hacer a los jóvenes para saber con una precisión de 10 minutos el tiempo que pasan pegados al WhatsApp diariamente? Este tipo de cuestiones es muy útil saber responderlas, si bien no para el ejemplo anterior, para otros casos donde el estudio pueda ser médico, por ejemplo, como sucede en la bioestadística.

Si **E** el **margen de error deseado** que fijamos nosotros, despejando la « n » de la fórmula del margen de error obtenemos:

$$n = \left(\frac{z_{\alpha/2}\sigma}{E} \right)^2$$

Ejemplo 6

En los vuelos es crucial estimar el peso medio de los pasajeros (por razones de seguridad sin ir más lejos). Entonces cabe preguntarse, ¿cuántos pasajeros seleccionamos al azar y pesamos? La respuesta vendría dada por la fórmula anterior, para lo cual tendríamos que fijar un nivel de confianza dado y una precisión deseada.

Ejemplo 7

Si tuviéramos que calcular el tamaño muestral necesario (con una confianza del 95%) para medir la población de mujeres policía que vimos anteriormente con una precisión de 1cm , procederíamos sustituyendo en la ecuación anterior.

$$n = \left(\frac{1,96 \times 4,30}{1} \right)^2 = 71,03 \cong 71 \text{ mujeres policía}$$

Observa que la desviación típica no la conocíamos y, por ello, se ha empleado su estimación, el cual es un recurso empleado en la práctica, pues lo normal es que no conozcamos el valor real de la desviación típica. En otras ocasiones lo que se hace es estimar la desviación típica cogiendo una **muestra piloto** (que es una muestra que se recoge previamente de cara a «tantear» las características de la población para tener en cuenta estos aspectos para cuando se recoja la muestra grande. En la práctica también es común la estrategia de las muestras piloto). Si se emplea « s » en lugar de la desviación de la población real tendremos que usar la distribución T' en lugar de la Z tal y como sucedía cuando en la creación del IC no conocíamos σ .

También es habitual el redondeo puesto que el tamaño de la muestra ha de ser siempre —lógicamente— entera, así en el caso del ejemplo anterior redondeamos a 71 mujeres.

7.5. Intervalo de confianza para la proporción

Habiendo visto cómo se razona y procede para construir un IC, proseguimos con los diferentes casos que se nos presentan, esta vez nos interesa hallar el IC para la proporción donde de nuevo hemos empleado el TCL para saber que si la muestra es suficientemente grande \hat{p} se distribuirá como una normal de media p y desviación típica:

$$\sqrt{p(1-p)/n}.$$

$$p \in \left[\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]_{1-\alpha}$$

Que también podemos expresar brevemente a través del margen de error como:

$$\hat{p} \pm E \text{ con } E = z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

De modo más o menos análogo al de la media (teniendo en cuenta las diferencias en cuanto a desviación típica, etc.) el tamaño muestral necesario fijando un error determinado es:

$$n = \left(\frac{z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})}}{E} \right)^2 = \frac{z_{\alpha/2}^2 \hat{p}(1-\hat{p})}{E^2}$$

Ejemplo 8

Se ha interrogado en un trabajo de estadístico escolar a 100 jóvenes sobre si fuman o no. 30 afirmaron fumar mientras que 70 se declararon no fumadores. ¿Qué porcentaje de fumadores habrá en este instituto con un nivel de confianza del 95%?

Tenemos que $\hat{p} = 30/100 = 0,3$ siendo $n = 100$. El valor crítico es 1,96 como ya hemos visto. Por tanto tenemos que el IC para p es:

$$p \in \left[0,3 - 1,96 \sqrt{\frac{0,3(0,7)}{100}}, 0,3 + 1,96 \sqrt{\frac{0,3(0,7)}{100}} \right]_{0,95} = p \in (0,21; 0,39)$$

7.6. Intervalo de confianza para la varianza de una población normal

Vamos a estudiar ahora el caso del intervalo de confianza para la varianza σ^2 de una población $N(\mu; \sigma)$ con μ desconocida, para un nivel de confianza $1 - \alpha$.

Ahora si $X \rightarrow N(\mu; \sigma)$ y tenemos una muestra $X : (x_1, \dots, x_i, \dots, x_n)$ que es aleatoria con varianza s^2 tenemos que:

$$\sigma^2 \in \left[\frac{(n-1)s^2}{\chi_{1-\alpha/2}}, \frac{(n-1)s^2}{\chi_{\alpha/2}} \right]_{1-\alpha}$$

Donde

$$\chi_{\alpha/2}$$

es el valor crítico que deja una probabilidad acumulada de:

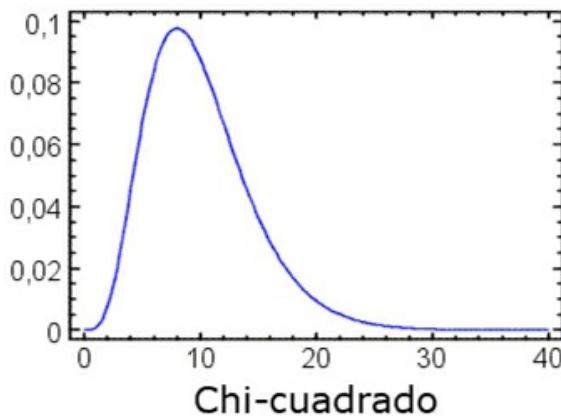
$$\alpha/2$$

en una distribución que no hemos visto todavía y que se denomina **Chi cuadrada de Pearson**.

$$\chi_{n-1}^2 = \frac{(n-1)s^2}{\sigma^2}$$

A pesar de que se escribe «Chi» se debe pronunciar «Ji». Los grados de libertad que tiene la Chi Cuadrado son $n - 1$, aspecto que hay que tener en cuenta cuando se localiza en las tablas.

Conviene saber que esta distribución **no es simétrica** como la normal o la T-Student tal y como podemos apreciar.



Además los valores de la X^2 han de ser positivos como su cuadrado indica (ya que un número al cuadrado no puede ser negativo). A medida que los G.L. de la X^2 aumentan se va acercando a la Normal. Del mismo modo que con la normal y la T-Student se emplea una tabla para localizar sus valores críticos.

DISTRIBUCION DE χ^2

Grados de libertad	Probabilidad											
	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,01	0,001	
1	0,004	0,02	0,06	0,15	0,46	1,07	1,64	2,71	3,84	6,64	10,83	
2	0,10	0,21	0,45	0,71	1,39	2,41	3,22	4,60	5,99	9,21	13,82	
3	0,35	0,58	1,01	1,42	2,37	3,66	4,64	6,25	7,82	11,34	16,27	
4	0,71	1,06	1,65	2,20	3,36	4,88	5,99	7,78	9,49	13,28	18,47	
5	1,14	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,07	15,09	20,52	
6	1,63	2,20	3,07	3,83	5,35	7,23	8,56	10,64	12,59	16,81	22,46	
7	2,17	2,83	3,82	4,67	6,35	8,38	9,80	12,02	14,07	18,48	24,32	
8	2,73	3,49	4,59	5,53	7,34	9,52	11,03	13,36	15,51	20,09	26,12	
9	3,32	4,17	5,38	6,39	8,34	10,66	12,24	14,68	16,92	21,67	27,88	
10	3,94	4,86	6,18	7,27	9,34	11,78	13,44	15,99	18,31	23,21	29,59	
No significativo										Significativo		

Un aspecto importante de cara a comprender la X^2 es que los valores críticos no son simétricos como ocurría anteriormente con la proporción (por ejemplo -1.96 y $+1.96$) sino que ahora tenemos que ambos son positivos y que presentan magnitudes diferentes.

Razón por la que en la tabla anterior está dividida en dos clases de valores críticos: los «no significativos», donde localizaremos al

$$x_{\alpha/2}$$

y los «significativos» donde localizaremos el

$$X_{1-\alpha/2}$$

Ejemplo 9:

Partiendo de los siguientes datos:

Peso en Kilos de 100 palomas mensajeras			
Peso en Kilos	Frecuencia	Peso en Kilos	Frecuencia
1,80	1	1,93	8
1,81	0	1,94	9
1,82	1	1,95	4
1,83	1	1,96	11
1,84	1	1,97	3
1,85	1	1,98	4
1,86	1	1,99	3
1,87	2	2,00	7
1,88	3	2,01	2
1,89	5	2,02	4
1,90	7	2,03	5
1,91	6	2,04	1
1,92	8	2,05	2

Calcula un intervalo de confianza al 95% para la varianza de la población correspondiente.

Lo primero que hacemos es establecer entonces la confianza = 0,95

Ya que $n = 100$, tenemos que buscando en las tablas de la Chi-Cuadrado (O ayudándonos del Excel) tenemos que $X_{0,025} = 73,4$ y $X_{0,975} = 128$.

Calculamos s^2 resultando 0,002822. De modo que $(n - 1)s^2 = 99 \bullet 0,002822 = 0,2794$. Ya contaríamos por tanto con todos los valores necesarios para sustituirlos en la fórmula del IC.

$$\sigma^2 \in \left[\frac{0,2794}{128}; \frac{0,2794}{73,4} \right]_{0,95} = [0,0021; 0,0039]$$

7.7. Intervalo de confianza para la diferencia de medias y proporciones

Ahora pasaremos a ver brevemente los casos en los que comparamos dos poblaciones y entonces nos va a interesar construir el intervalo de confianza para la diferencia de medias ($\mu_1 - \mu_2$) de dos poblaciones normales en diferentes casos. Como caso final veremos el caso de la diferencia de proporciones ($p_1 - p_2$) también.

- ▶ Supondremos primeramente que las **varianzas son distintas y conocidas**, al nivel de confianza $1 - \alpha$.

$$(\mu_1 - \mu_2) \in \left[(\bar{x} - \bar{y}) \mp z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \right]_{1-\alpha}$$

Donde «n» y «m» son los tamaños muestrales respectivos de la muestra de «las X » y de «las Y ».

- ▶ En este segundo caso las **varianzas serán iguales y conocidas**:

$$(\mu_1 - \mu_2) \in \left[(\bar{x} - \bar{y}) \mp z_{\alpha/2} \sigma \sqrt{\frac{1}{n} + \frac{1}{m}} \right]_{1-\alpha}$$

- ▶ Ahora tratamos el caso de **varianzas desconocidas** pero idénticas.

$$(\mu_1 - \mu_2) \in \left[(\bar{x} - \bar{y}) \mp t_{n+m-2; \alpha/2} \sqrt{\frac{1}{n} + \frac{1}{m}} \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}} \right]_{1-\alpha}$$

A pesar de que la fórmula es compleja, lo interesante es hacerse una idea de que la segunda raíz expresa algo así como la desviación típica combinada (o ponderada por sus respectivos tamaños muestrales) de ambas y, por ello, sirve de sustituta de la desviación típica σ de la fórmula de la varianza conocida, la cual ahora desconocemos.

Observa que al emplear las cuasivarianzas multiplicamos ahora por « $n - 1$ » y « $m - 1$ » estas y por ello dividimos entre « $n + m - 2$ » para obtener el promedio.

- ▶ Por último vamos a ver como calcular IC para la diferencia ($p_1 - p_2$) de proporciones poblacionales provenientes de dos poblaciones Binomiales con proporciones de éxito p_1 y p_2 respectivamente. De nuevo estas poblaciones pueden tener tamaños de muestra diferentes « n » y « m ». Además necesitamos que tales tamaños sean lo suficientemente altos. Bajo estas condiciones, tenemos:

$$(p_1 - p_2) \in \left[(\hat{p}_1 - \hat{p}_2) \mp z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n} + \frac{\hat{p}_2(1 - \hat{p}_2)}{m}} \right]$$

No vamos a ver ejemplos de aplicación de las fórmulas de diferencia de parámetros, puesto que las diferencias de parámetros se suelen plantear a nivel estadístico como contrastes de hipótesis que permitan observar si hay o no diferencias.

De todos modos puedes encontrar ejemplos en los textos propuestos en «Cómo estudiar este tema».

7.8. Intervalos de confianza robustos

Mediante el uso de las medidas robustas vistas en el tema 2, es posible redefinir todos los intervalos de confianza vistos para hacerlos **robustos frente a la presencia de outliers en las muestras**. Para ello, haremos uso de las **medidas robustas vistas en el tema 2** y trabajaremos, por tanto, con **conjuntos winsorizados**. Gracias a que estos conjuntos eliminan los valores más extremos de su conjunto, los outliers no tienen ninguna influencia en los análisis realizados sobre los datos. Los cambios a realizar sobre las fórmulas, en el caso del **intervalo de confianza para la media**, son los siguientes:

- ▶ **Uso de la media recortada:** dado que la media es una medida que, como ya vimos, es sensible a los outliers, la cambiaremos por la **media recortada**.
- ▶ **Uso de la cuasidesviación típica winsorizada:** en vez de usar la cuasidesviación típica normal, utilizaremos la cuasidesviación típica winsorizada. Dado que hemos eliminado en nuestras muestras los valores más extremos, es necesario que las medidas utilizadas en las fórmulas sean consecuentes con este hecho.
- ▶ **Modificación del tamaño de la población:** al eliminar los valores extremos de nuestra muestra, debemos reducir el tamaño n en consecuencia. Recordemos que utilizábamos un valor α para determinar el porcentaje de elementos que íbamos a no tener en cuenta en los cálculos. Dicho parámetro nos resultará útil para aplicarlo en las fórmulas y determinar el valor n real sobre el que estamos trabajando.

Tras realizar estos cambios y teniendo en cuenta que denominaremos como **β al valor de significación** para evitar **choques de nomenclatura**, podemos construir un intervalo de confianza robusto para la media recortada siguiendo la siguiente fórmula:

$$\left[\bar{x}_\alpha - t_{n-2k-1; \frac{\beta}{2}} \frac{S_w}{(1-2\alpha)\sqrt{n}}, \bar{x}_\alpha + t_{n-2k-1; \frac{\beta}{2}} \frac{S_w}{(1-2\alpha)\sqrt{n}} \right]$$

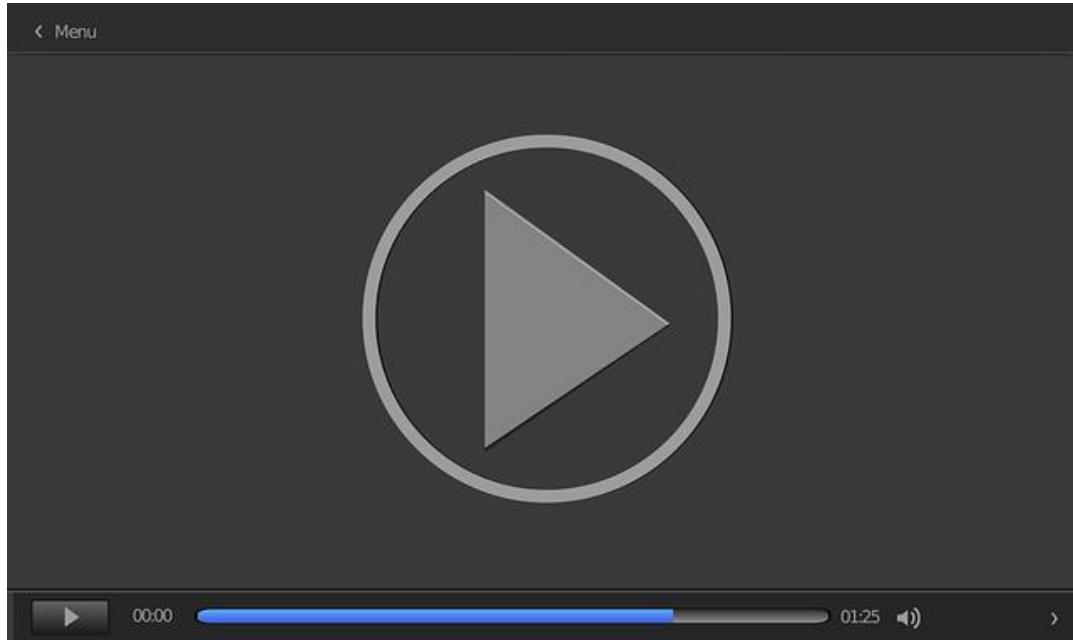
Donde **listamos qué es cada parámetro** a continuación:

- ▶ k : es un parámetro que se calcula como la parte entera de $n \cdot \alpha$.
- ▶ \bar{x}_α : es la media recortada de la muestra.
- ▶ $t_{n-2k-1; \frac{\beta}{2}}$: es una t-student con grados de libertad $n-2k-1$ con nivel de significación de $\beta/2$.
- ▶ S_w : es la cuasidesviación típica winsorizada.
- ▶ α : porcentaje de recorte usado en la media winsorizada utilizada.

Tal y como puede verse, **el intervalo resultante es muy similar al intervalo de confianza para la media normal** y todos los cambios realizados sobre él responden al proceso de winsorización realizado sobre la muestra extraída de la población con objeto de evitar trabajar con posibles outliers.

Diferenciación de conceptos: «confianza» vs. «intervalos de confianza»

En este vídeo vamos a ver estos dos conceptos y a aprender a diferenciarlos.



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=f60f9c06-961b-49df-a0eb-acbd00e37c03>

Practica con R los conceptos estudiados

Asumiendo que tienes [instalado R](#) y [RStudio](#), abre el IDE de RStudio y en un «R script» nuevo «Ctrl + Shift + N» escribe el siguiente código.

```
rm(list=ls())
#####
requiredPackages7 <- c('DescTools', 'dplyr', 'EnvStats', 'mosaic',
'rcompanion', 'samplingbook', 'TeachingDemos')

sesion7 <- function(pkg){
new.pkg <- pkg[!(pkg %in% installed.packages() [, "Package"])]
if (length(new.pkg))
install.packages(new.pkg, dependencies = TRUE)
sapply(pkg, require, character.only = TRUE)
}

sesion7(requiredPackages7)
#####
#####TLC#####
#https://opendatascience.com/central-limit-theorem-r
#Con una Bin(n=4,p=0.05)

n <- 4 # Number of trials (population size)
p<-0.05
s <- 2000 # Number of simulations
m <- c(20, 100, 500, 1000)
EX <- n*p
VarX <- n*p*(1-p)
Z_score <- matrix(NA, nrow = s, ncol = length(m))
for (i in 1:s){
for (j in 1:length(m)){ # loop over sample size
samp <- rbinom(n = m[j], size = n, prob = p)
sample_mean <- mean(samp) # sample mean
# Calculate Z score for mean of each sample size
Z_score[i,j] <- (sample_mean-EX)/sqrt(VarX/m[j])
}
}

par(mfrow=c(2,2))
for (j in 1:4){
hist(Z_score[,j], xlim=c(-5,5),
freq=FALSE, ylim=c(0, 0.5),
ylab="Probability", xlab="",
main=paste("Sample Size =", m[j]))}
```

```

# Density curve
x <- seq(-4, 4, by=0.01)
y <- dnorm(x)
lines(x, y, col="blue")}
#####
#####Intervalos de confianza#####
#####
Datalc<-
read.csv("https://raw.githubusercontent.com/millerjanny/UAM2022/main/Data_LendingClub.csv")

#####para la media con sigma conocida#####
library(TeachingDemos)
z.test(Datalc$int_rate, sd=sd(Datalc$int_rate))$conf.int

#####para la media con sigma desconocida#####
t.test(Datalc$int_rate,conf.level=0.95)$conf.int

library(DescTools)
MeanCI(Datalc$int_rate,conf.level=0.95)

library(rcompanion)
groupwiseMean(int_rate ~ 1,
data=Datalc,
conf=0.95,
digits=5)

groupwiseMean(int_rate ~ Default,
data=Datalc,
conf=0.95,
digits=3)

#####Para la proporción#####
library(dplyr)
library(mosaic)
Datalc$Default=recode_factor(Datalc$Default, '1' = "Default", '0' = "Non-default")
prop.test(~Default, data=Datalc, success='Default',
conf.level=0.95)$conf.int

#####para la varianza#####
library(EnvStats)
var(Datalc$int_rate)
varTest(Datalc$int_rate,conf.level=0.95)$conf.int

#####para la diferencia de proporciones#####

```

```
prop.table(table(Datalc$term, Datalc$Default), margin=1)
prop.test(~Default|term, data=Datalc, success='Default', conf.level=0.95)

#####para la diferencia de medias#####
#igualdad de varianzas int_rate por categoría Default
by(select(Datalc, int_rate), factor(Datalc$Default),summary)
var.test(int_rate~ Default, data = Datalc)$conf.int
t.test(int_rate ~ Default, data = Datalc, var.equal = FALSE)$conf.int

#####
#####Tamaño de muestra#####
#####
library(samplingbook)

#####media#####
#Ejemplo de la clase:tamaño muestral mínimo necesario (IC 95%) para estimar
la media de la estatura
#de mujeres policías con una precisión de 1cm
error=1
confianza=0.95
N=1000
s=4.30
sample.size.mean(error,s,N,confianza)

#####Proporción#####
data(election)
#DataFrame con el número de ciudadanos con derecho a voto y los resultados
de las elecciones en 2002 y 2005 para el Bundestag alemán, la primera
cámara del parlamento alemán.
#SPD_02:percentage for the Social Democrats SPD in 2002
# tamaño de muestra mínimo para estimar la P de votantes social demócratas
hoy
confianza=0.95
error=0.05
N=300
P=mean(election$SPD_02)
sample.size.prop(e=error, P=P, N=N,level=confianza)
#####
```

Prueba a ejecutar el *script* anterior siguiendo estas indicaciones:

- ▶ Ejecuta cada línea de código, posiciona el cursor en la primera línea y utiliza la opción «Run» o «Ctrl + Enter».
- ▶ Observa la «Consola» y «Environment» cuando ejecutas cada línea.

7.9 Referencias bibliográficas

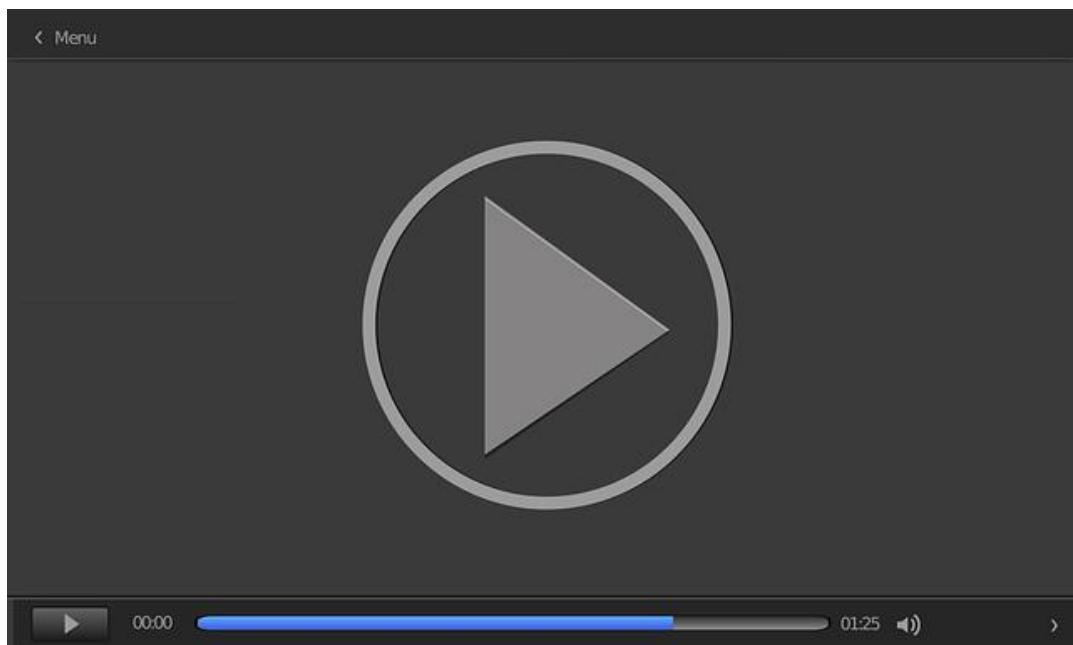
Ríus, F. (1998). *Bioestadística: Métodos y aplicaciones*. Málaga: Universidad de Málaga. Publicaciones. Versión electrónica:

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

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

Buscando los valores críticos en las tablas de diferentes distribuciones

En esta lección magistral veremos cómo manejar las tablas de distribuciones estadísticas. Veremos el manejo de la normal (0,1), la de la T-Sudent y la de Chi Cuadrada.



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=3d0a0024-ac4a-4ab5-ab93-abdc00f7fcd1>

Un estadístico entre cervezas negras

Te recomiendo que leas este artículo en inglés sobre el origen de la T-Student y el porqué de este nombre tan peculiar. Se trata de una interesante anécdota de la historia de la Estadística y, más aún si cabe, porque tiene relación con la cerveza negra, ¡una auténtica delicia!

Accede al artículo desde el aula virtual o a través de la siguiente dirección web:

<http://www.breweryhistory.com/journal/archive/121/bh-121-113.htm>

Calculadoras online de las principales distribuciones de probabilidad

Te recomiendo que investigues en la interesante web de Stat Trek en la cual puedes emplear diferentes *applet* para el cálculo de los valores de diferentes distribuciones: continuas como la normal, T-Student, chi cuadrada, etc. y discretas como la binomial, poisson, multinomial, etc.

Accede a la página desde el aula virtual o a través de la siguiente dirección web:

<http://stattrek.com/online-calculator/normal.aspx>

Bibliografía

Kreyszig, E. (1983) *Introducción a la Estadística Matemática*. México: Limusa. (Ver capítulo 11: Estimación de Parámetros).

Martín Andrés, A. (2004). *Bioestadística para las ciencias de la salud*. Madrid: Norma-Capitel.

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

1. ¿Qué es un valor crítico en términos de inferencia?

 - A. Un valor sumamente importante en los IC que depende de la muestra.
 - B. Es el valor de una distribución que acumula cierta probabilidad.
 - C. Se trata de un concepto fundamental cuando manejamos distribuciones normales.
 - D. Las respuestas A y C son correctas.

2. Empleamos intervalos de confianza entre otras razones porque...

 - A. La estimación puntual se queda corta en el sentido de que no sabemos cuán bueno es una estimación puntual.
 - B. Es una buena manera de aproximarnos al parámetro poblacional tanto como deseemos o podamos.
 - C. Si no podemos realizar la estimación puntual que es más exacta nos conformamos con el IC.
 - D. Las respuestas A y B son correctas.

3. ¿Qué es o a qué es igual « E »?

 - A. Al estimador, que también puede ser anotado como $\hat{\theta}$.
 - B. Al margen de error.
 - C. $z_{\alpha/2}$
 - D. Un parámetro de cierto tipo de variable aleatoria.

4. Si hemos calculado el IC para p y resulta: $0,325 < p < 0,375$. ¿Cuánto vale « E »?

 - A. 0,025
 - B. 0,25
 - C. 0,050
 - D. No se puede calcular con esta información.

5. Al IC $188\text{cm} < \mu < 209\text{cm}$ que marca con un 95% de confianza la estatura media de un equipo de la NBA lo interpretamos como...

- A. De cada 100 jugadores, 95 estarán contenidos en dicho intervalo.
- B. De cada 100 muestras de jugadores, 95 tendrán la media contenido en ese IC.
- C. Tenemos una confianza del 95% de que la media de los jugadores de un equipo de la NBA está contenida entre 188cm y 209cm.
- D. Las respuestas B y C son correctas.

6. ¿Qué fórmula es la correcta para hallar el IC de una media poblacional conocida su varianza?

A. $\mu \in \left[\bar{x} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}; \bar{x} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]_{1-\alpha}$

B. $\mu \in \left[\bar{x} - t_{\alpha/2} \frac{\sigma}{\sqrt{n}}; \bar{x} + t_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]_{\alpha}$

C. $\mu \in \left[\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}}; \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}} \right]_{1-\alpha}$

D. $\mu \in \left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}; \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]_{1-\alpha}$

7. Cuando σ no es conocida, en el cálculo de los valores críticos para construir los IC empleamos:

- A. La T-Student.
- B. La Chi Cuadrada.
- C. La Normal Z.
- D. La Normal (0,1).

8. La T-Student es...

- A. Una distribución similar a la Normal, de hecho se aproxima a ella a medida que aumenta el «n».
- B. Es diferente de la Normal y la Chi Cuadrada.
- C. Es simétrica.
- D. Las respuestas A y C son correctas.

9. ¿Qué es una muestra piloto?

- A. Un tipo de encuestas muy empleadas en aviación (por temas de seguridad).
- B. Una muestra que se recoge antes de hacer un estudio para tantear las características de la población.
- C. Todo estudio estadístico bien hecho debe constar de una muestra preliminar llamada «piloto» para posteriormente recoger la muestra definitiva.
- D. Es una muestra enorme que no siempre es posible recogerla, pero es lo ideal.

10. ¿En un IC qué porcentaje de las veces éste no contendrá al parámetro?

- A. $(1 - \alpha) \%$ de las veces.
- B. $(1 - \alpha) \times 100\%$ de las veces.
- C. $\alpha \times 100\%$ de las veces.
- D. Depende de la suerte que hayamos tenido con la muestra concreta que cojamos.

Análisis e Interpretación de Datos

Tema 8. Contrastes de hipótesis

Índice

Esquema

Ideas clave

- 8.1. ¿Cómo estudiar este tema?
- 8.2. Introducción a los contrastes de hipótesis
- 8.3. Dos tipos de error en la significancia estadística
- 8.4. Pasos a seguir en un contraste de hipótesis
- 8.5. Contrastados de hipótesis para una media
- 8.6. Contrastados de hipótesis para la proporción
- 8.7. Contrastados de hipótesis sobre la varianza
- 8.8. Contrastados paramétricos para dos muestras
- 8.9. Contrastados de hipótesis robustos
- 8.10 Referencias bibliográficas

A fondo

Aprendiendo a interpretar los resultados de un contraste de hipótesis efectuado con un programa estadístico y/o un artículo científico

A Close Look at Therapeutic Touch

Comparación de medias de dos poblaciones

Bibliografía

Test

CONTRASTES DE HIPÓTESIS			
Introducción	Dos tipos de error	Contrastes una población	Contrastes de dos poblaciones
<p>H_0 = Hipótesis nula H_1 = Hipótesis alternativa</p> <p>Contraste bilateral $H_0: \theta = \theta_0$ $H_1: \theta \neq \theta_0$</p> <p>Contrastes unilaterales</p> <p>$H_0: \theta \geq \theta_0$ $H_1: \theta < \theta_0$</p> <p>$H_0: \theta \leq \theta_0$ $H_1: \theta > \theta_0$</p>	<p>P(rechazar H_0 siendo H_0 verdadera) = α</p> <p>P(aceptar H_0 siendo H_0 falsa) = β</p> <p>p-valor</p>	<p>$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$</p> <p>Aceptamos si $-\bar{z}_{\alpha/2} \leq z_{exp} \leq \bar{z}_{\alpha/2}$</p> <p>Rechazamos si $z_{exp} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} < -\bar{z}_{\alpha/2}$</p> <p>$z_{exp} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} > \bar{z}_{\alpha/2}$</p>	<p>Ejemplo: caso bilateral con μ_1 y μ_2</p> <p>Caso σ desconocido o realista $H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$</p> <p>Aceptamos si $t_{exp} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \geq \bar{t}_{n+m-2; \alpha/2}$</p> <p>Rechazamos si $-\bar{t}_{n+m-2; \alpha/2} \leq t_{exp} \leq \bar{t}_{n+m-2; \alpha/2}$</p> <p>Si σ desconocida empleamos T-Student</p>
Se concluye que es cierto	<p>H_0</p> <p>H_1</p>	<p>H_0 Acierto $1-\alpha$</p> <p>H_1 Error de tipo I α</p>	<p>H_I Error de tipo II β</p> <p>Acierto (potencia = $1-\beta$)</p>

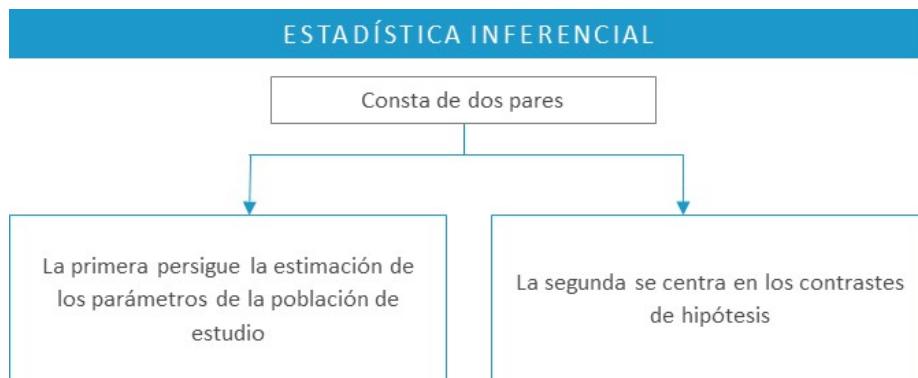
8.1. ¿Cómo estudiar este tema?

Para estudiar este tema lee las **Ideas clave**, además de los intervalos que se indican a continuación: **Páginas 389-398** del libro: Triola, M. F. (2009). *Estadística* (10^a ed). México: Pearson. Este fragmento corresponde aproximadamente a diferentes apartados o aspectos vistos en este tema. **Páginas 203-245** del libro: Ríus, F. (1998). *Bioestadística: Métodos y aplicaciones*. Málaga: Universidad de Málaga. Publicaciones. <https://www.bioestadistica.uma.es/baron/apuntes/clase/apuntes/pdf/>

Para hacerte una idea global es importante que mires el esquema del tema, el cual te ayudará a hacerte una buena idea de cómo está estructurado y las relaciones que puedan existir entre algunos conceptos clave.

También **será clave que practiques con las actividades que vienen al final del tema**. Del mismo modo presta atención a los ejemplos que acompañan a los diferentes apartados a lo largo del tema, pues encierran muchas de las claves que te facilitarán la comprensión del tema.

8.2. Introducción a los contrastes de hipótesis



Esta segunda parte de la inferencia es, en cierto modo, más amplia pues no está restringida a aseveraciones sobre los parámetros, sino que es posible también realizar contrastes de hipótesis sobre otras propiedades de una distribución estadística, como puede ser el contraste para ver si se distribuye normalmente.

Contraste de hipótesis

Es un procedimiento formal estadístico para decidir si una afirmación sobre una población parece manifestarse como verosímil o no a partir de los datos.

Se trata entonces de una herramienta muy poderosa, pues va enfocada directamente a poder decidir sobre cuestiones, lo cual le da un carácter fuertemente aplicado. Para poder resolver la veracidad o no de estas afirmaciones, o como decimos en lenguaje estadístico, contrastarlas, se establece que una afirmación de partida a contrastar, que es la llamada **hipótesis nula**: H_0 y otra de negación, que se da en caso de no ocurrir la primera o mejor dicho, de ser rechazada la H_0 , que es la **hipótesis alternativa**: H_1 .

La hipótesis nula se suele escoger porque es lo que se piensa, **es en principio lo que ya está establecido**, bien porque cierta teoría lo apoya, o bien porque empíricamente está consolidado, también puede establecerse porque tenemos una

fuerte intuición de que es cierto algo, etc.

Por otro lado, la hipótesis alternativa se plantea como lo novedoso, lo que «rompe» con algo establecido o conservador, aquello que se pretende que sea demostrado, podríamos decir.

Ejemplo 1



Si quisieramos contrastar si las antenas de repetición pueden provocar mayores probabilidades de tener cáncer a las personas que viven en su cercanía, nosotros no podríamos partir de esta afirmación como hipótesis nula, sino que partiríamos en principio de una afirmación neutral o «conservadora», que sería del tipo:

H_0 : La tasa de incidencia de cáncer en los bloques de viviendas con antenas de repetición instaladas es la misma que en aquellos donde no están instaladas.

Y una vez realizada esta afirmación que nos sirve de hipótesis nula, la alternativa quedaría fijada:

H_1 : La tasa de incidencia de cáncer en los bloques de viviendas con antenas de repetición instaladas no es la misma.

Hay diferentes maneras de plantear la alternativa, podemos plantear una negación total, en lo que se llamará un **contraste de dos colas o bilateral** por poder diferenciarse en dos direcciones, siendo mayor o menor.

Por el contrario, será de **una cola o unilateral** cuando se plantea solo una de las dos opciones. También conviene pensar que H_0 es la que suele plantearse en términos de igualdad y la H_1 en los de diferencia (que puede ser bilateral o unilateral). A continuación, podemos ver diferentes clases de contraste donde recordemos el parámetro es anotado como θ mientras que θ_0 se refiere a un valor concreto que toma este:

A	$H_0: \theta = \theta_0$ $H_1: \theta \neq \theta_0$
B	$H_0: \theta \geq \theta_0$ $H_1: \theta < \theta_0$
C	$H_0: \theta \leq \theta_0$ $H_1: \theta > \theta_0$
D	$H_0: \theta_1 \leq \theta \leq \theta_2$ $H_1: \theta < \theta_1 \text{ o } \theta > \theta_2$

Los casos A y el D son de dos colas o bilaterales, mientras que el B y el C son de una sola cola o unilaterales.

Ejemplo 2

El caso planteado antes sería bilateral, puesto que la tasa de cáncer de no ser igual puede ser mayor o menor. También podrá ser unilateral y, de hecho, parece más lógico en este caso, cuando se plantee que de rechazarse solo pueda ser superior.

Decimos que parece razonable pues lo que se espera es precisamente que pueda producir cáncer la cercanía de tales instalaciones y es tan inesperado el que rebaje la tasa de cáncer que no se plantearía en la H_1 .

En cuanto al lenguaje empleado cuando aceptamos la H_0 se suele decir algo del tipo: «no se han encontrado evidencias estadísticamente significativas de que la tasa de cáncer en las viviendas con antenas de repetición no se encuentre fuera de los valores normales...». Mientras, que de rechazar la H_0 y quedándonos, por tanto, con la H_1 decimos: «se han encontrado evidencias estadísticamente significativas de que existen diferencias en la tasa de...». También se suele omitir el término «evidencias» afirmando directamente: «...diferencias estadísticamente significativas».

El concepto de significancia es fundamental en los contrastes de hipótesis y está íntimamente relacionado con los niveles de significación que hemos visto en los intervalos de confianza.

La significancia estadística se da cuando los estadísticos que se emplean para el contraste de hipótesis toman valores a partir de los cuales rechazaremos H_0 . El nivel de significación α marcará el punto que de ser alcanzado por el estadístico rechazaremos H_0 .

Cualquier número basado en la muestra de los datos que nos ayude a decidirnos sobre H_0 y H_1 será el **estadístico de contraste**.

Ejemplo 3

Siguiendo con el ejemplo de las antenas de repetición, para contrastar la H_0 podríamos partir del estudio de la incidencia del cáncer en forma de tasa o proporción de presencia de cáncer, así sabríamos, por estudios anteriores, que la tasa normal de presencia de cáncer (su prevalencia) es de 7 % de la población, de modo que:

$$H_0: p \leq 0,07$$

$$H_1: p > 0,07$$

Así, nuestro estadístico de contraste viene determinado por que tiene que ser una variable que nos permita medir tal proporción y ya hemos visto anteriormente que tal v.a. es la proporción muestral \hat{P} que se distribuye como X/n donde X es una $Bi(n, p)$.

De esta manera recogeríamos una muestra aleatoria de vecinos en bloques de viviendas con antenas de repetición próximas y calcularíamos \hat{P} , si esta resulta diferente a 0,07 rechazaremos H_0 .

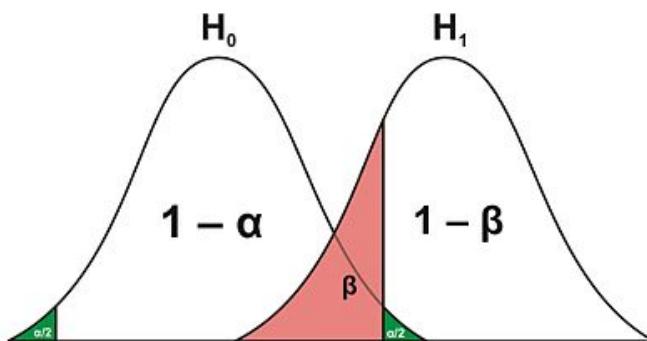
De todos modos, obviamente no seremos tan exigentes de rechazar si no toma el valor 0,07 exactamente, pues de lo contrario no aceptaríamos prácticamente nunca la H_0 . Lo que hacemos generalmente es **establecer un rango de valores que, de tomarlos el estadístico, aceptaremos H_0** , es la llamada **región de aceptación**. Por el contrario, de no tomar el estadístico un valor «razonable» para la H_0 , caerá en la región complementaria a esta, es decir, en la **región de rechazo o región de crítica**.

La distribución de probabilidad que supondremos que tiene el estadístico de contraste es precisamente aquella que resulta de suponer que H_0 es verdadera.

Esto es lógico pues se trata de partir que H_0 es cierta y, entonces, tras la recoger la muestra y calcular el estadístico, ver si de acuerdo al valor que ha tomado parece probable que H_0 sea cierta, o si por el contrario la deberemos rechazar y quedarnos con la H_1 .

8.3. Dos tipos de error en la significancia estadística

Como los datos de la muestra pueden ser los de una «mala muestra» (una de esas raras digamos) podríamos estar cometiendo un error. Al error de **rechazar la H_0 cuando realmente era verdadera** lo denominamos α o **error de Tipo I**. También es posible cometer una segunda clase de error que se da cuando **aceptamos H_0 siendo falsa** que es el llamado β o **error de Tipo II**.



Así, tenemos que la raya vertical del dibujo indica el valor a partir del cual, si es mayor el estadístico de contraste y cae en la región de rechazo, rechazaremos la H_0 , quedándonos con la H_1 . En el caso del dibujo se trata de un **contraste bilateral**, puesto que podemos rechazar la H_0 también por defecto cuando sea muy pequeño dicho valor. Conviene puntualizar que en estos casos también se dibuja otra función de probabilidad para H_1 a la izquierda, pues esos valores no pueden quedar en tierra de nadie (o es H_0 verdadera o lo es H_1). La siguiente tabla muestra los cuatro casos posibles que existen en un contraste de hipótesis, producto de las dos maneras de acertar en nuestra decisión y de las dos de errar:

		Cuando realmente es cierto	
		H_0	H_1
Se concluye que es cierto	H_0	Acierto $1-\alpha$	Error de tipo II β
	H_1	Error de tipo I α	Acierto (potencia = $1-\beta$)

Tal y como apreciamos en la gráfica, cobra cierta importancia, aparte de los dos tipos de error, un tipo de acierto, la posibilidad de realmente acertar cuando H_1 sea cierta $1 - \beta$, que se denomina **potencia del test o contraste**.

Ejemplo 4: Del porqué es deseable en algunos casos una potencia $1 - \beta$ alta

Imaginémonos que nuestro contraste de hipótesis viene asociado a un test estadístico para dictaminar si se tiene determinada enfermedad. En tal caso requeriremos de un contraste que presente una potencia elevada (o lo que es lo mismo un β pequeño), pues tal cosa hará que el contraste tienda a detectar a los enfermos con cierto grado de seguridad, ya que lo peligroso sería ignorar que están enfermos cuando realmente lo están (es decir aceptar H_0 cuando realmente es verdadera H_1). **Un test será mejor cuanta mayor potencia tenga.**

Las diferentes decisiones erróneas de la tabla anterior pueden ser traducidas en probabilidades condicionadas como vemos a continuación:

$$P(\text{aceptar } H_0 \mid \text{siendo } H_0 \text{ falsa}) = \beta$$

$$P(\text{rechazar } H_0 \mid \text{siendo } H_0 \text{ verdadera}) = \alpha$$

Se puede pensar que hacer un buen contraste de hipótesis será entonces fácil, sin más que reducir los tipos de error I y II. Sin embargo, tal cosa no es posible así sin más, puesto que **al reducir uno de los errores aumentamos el otro. La única vía para disminuir ambos a la vez es aumentando el tamaño de la muestra.**

Ejemplo 5



Antes del estreno del «ojo del halcón» para el Máster de tenis de Madrid tuvieron que probar su eficacia realizando múltiples tiros, que fotografiaron y luego compararon con las imágenes generadas informáticamente por el «ojo de halcón». Así se estableció:

H_0 = La proporción de aciertos del «ojo de halcón» es del 99% o mayor.

H_1 = La proporción de aciertos es menor del 99%.

De modo que si se aceptaba H_0 el aparato funcionaba bien y de rechazarlo tendrían que haberlo ajustado o incluso cambiarlo por otra.

Determinar el tamaño de la muestra de golpes necesario era clave aquí, pues requería un gran trabajo realizar tantas instantáneas y luego comprobarlas visualmente una a una. De esta manera se decidió qué valor tomaría cada tipo de error, $\alpha = 1\%$ o 0,01 y $\beta = 5\%$ o 0,05. Con estos errores deseados que se fijaron, obtuvieron que para tener tales errores tendrían que emplear un tamaño de muestra (es decir, una cantidad de golpes) de «n».

Nos queda un aspecto importante sobre los contrastes de hipótesis por tratar. Hemos hablado de que a partir de cierto valor del estadístico de contraste, rechazaríamos la H_0 por improbable, pero en realidad el proceso de decisión emplea una cuantía que es la **probabilidad de obtener un valor del estadístico de prueba que sea al menos tan extremo como el obtenido a partir de los datos muestrales**. A esta probabilidad se le denomina **p-valor**, que proviene de *p – value* del inglés.

Como regla práctica para decidir un contraste de hipótesis rechazaremos la H_0 cuando la probabilidad de que el estadístico tome cierto valor sea menor que el nivel de significación acordado, es decir, la **regla del p – valor** es:

Si p valor $\leq \alpha \rightarrow$ Rechazamos H_0

Si p valor $> \alpha \rightarrow$ Aceptamos H_0

La regla del p-valor no es la única posible, pues nosotros podríamos calcular el valor crítico del estadístico que hace que su probabilidad de valer tal cantidad o más sea $\alpha/2$ (o α si es de una cola el contraste) y decidir en base a ese valor. Cuando el estadístico tome un valor mayor que ese valor crítico rechazaremos.

Cuando decidimos en base al valor del estadístico estamos empleando la **regla tradicional** de decisión del contraste de hipótesis. ¿Por qué la denominamos «tradicional»? Pues porque para el uso del p-valor es conveniente el uso de programas informáticos que calculan la probabilidad asociada a cada valor concreto de la muestra, de modo que el método del p-valor no fue empleado hasta que se desarrollaron los paquetes estadísticos que lo calculaban y lo ofrecían sin problemas como resultado a cualquier tipo de test estadístico. Por supuesto que ambos métodos son equivalentes y llevan a la misma decisión de aceptación o rechazo.

Existe además una **relación entre los intervalos de confianza y los contrastes de hipótesis** que conviene tener clara. Al construir un IC para un cierto nivel de confianza nos bastará con ver si el valor del parámetro bajo H_0 está incluido en dicho intervalo en cuyo caso aceptaremos H_0 . Si el valor del parámetro por el contrario no está incluido rechazaremos H_0 .

Ejemplo 6: ¿Es cierto lo que dice la prensa sobre la tasa de jóvenes que fuman?

Retomando un ejemplo del tema anterior (ejemplo 8) vimos que en una muestra de cien estudiantes de Secundaria a los que se les interrogó sobre si fumaban o no los fines de semana, el IC resultante para la proporción al 95% fue el siguiente:

$$p \in (0,21; 0,39)$$

Imaginemos ahora que un artículo de prensa sostiene que la mitad de los jóvenes estudiantes de Secundaria fuma los fines de semana, de modo que esto traducido al lenguaje de los contrastes de hipótesis es como afirmar que:

H_0 : la proporción de jóvenes que fuman es de 0,5.

H_1 : la proporción de jóvenes que fuman es diferente de 0,5.

O en forma simplificada:

$$H_0: p = 0,5$$

$$H_1: p \neq 0,5$$

El nivel de significación α para el contraste de hipótesis sería el mismo que el empleado para el IC que fue del 5%; de modo que **como $p = 0,5$ no está incluido en el $IC(p) = (0,21; 0,39)$ rechazaríamos la H_0** , quedándonos con la H_1 , ya que **habríamos encontrado evidencias estadísticamente significativas de que la proporción de jóvenes que fuman es distinta de 0,5**.

Podemos afirmar por tanto que, de acuerdo a nuestro estudio estadístico, la afirmación del artículo aparecido en prensa no es cierta (eso sí, tendríamos que mencionar la cuantía de los errores que estamos manejando para indicar nuestro grado de seguridad al rechazar).

Ciertamente, tras ver este ejemplo, puede parecer que la capacidad de decidir si una afirmación es cierta o no recae sobre el estadístico, lo cual le da, porque no decirlo, un cierto «poder».

De hecho, este «poder» que nos otorgan los **contrastos de hipótesis para decidir** hace que sean **muy empleados en la investigación**, ya que suponen una **herramienta crucial** para desarrollar el método científico y en la que apoyar la validez de los resultados de cualquier investigación.

8.4. Pasos a seguir en un contraste de hipótesis

Existen multitud de formas que pueden ser adoptadas por un contraste de hipótesis, por ello, es fundamental tener claros los pasos que sigue el procedimiento del test:

- 1. Sintetizar la hipótesis que se desea probar** y ser capaz de expresarla en forma simbólica. Esta hipótesis será la nula.
- 2. Delimitar la hipótesis alternativa en base a la nula** que ya se ha determinado y, tener en cuenta que han de ser complementarias, y por tanto no puede existir ningún valor del parámetro que no esté contenido en una u otra hipótesis. Dicho de otra manera, cuando una de las hipótesis sea falsa, la otra deberá ser necesariamente verdadera y viceversa.
- 3. Fijar un error α** que se está dispuesto a cometer y para lo que se tendrá que tener en cuenta la naturaleza del estudio estadístico y la gravedad (las consecuencias) de cometer este error de tipo I. **Interesa que este error sea lo menor posible** para minimizar la probabilidad de rechazar H_0 cuando sea cierta.
- 4. Elegir el estadístico de prueba adecuado para contrastar las hipótesis planteadas.** Hay que tener en cuenta que deberá ser conocida la distribución muestral de este bajo H_0 (es decir, suponiendo H_0 cierta).
- 5. Se recoge una muestra aleatoria y se calcula el estadístico** y entonces se procede de dos maneras, dependiendo si se usa el método tradicional o el del p valor, aunque como hemos visto son equivalentes:
 - ▶ En el tradicional a través del estadístico resulta una **región de aceptación y otra crítica o de rechazo**.
 - ▶ **El cálculo del estadístico nos genera un p valor** o probabilidad de que tome ese valor bajo H_0 . Esta fase se realiza a través de un programa informático generalmente estadístico, ya que el cálculo del p valor es más sencillo sí.

1. En el caso de haber empleado el método de la región de aceptación/rechazo **aceptaremos la H_0 de caer el valor del estadístico dentro de la región de aceptación**, y por el contrario si cae fuera de la región de aceptación, cayendo por tanto dentro la región de rechazo entonces lo rechazaremos.
2. Tomaremos la decisión **de rechazar H_0 si el p valor es menor o igual que el α** fijado. Si por el contrario el p valor es mayor que el α fijado (usualmente mayor que 0,05 o 0,1 depende del que haya sido utilizado) entonces aceptamos la H_0 .

Conviene puntualizar que **hoy en día prácticamente solo se emplea el método del *p – valor***, sin embargo, es preferible no ignorar el método tradicional, pues en él reside la lógica de base de los contrastes de hipótesis. Además, tiene una mayor facilidad para imaginarnos visualmente o gráficamente el contraste que se plantea al situar directamente los valores críticos que toman los estadísticos de contraste sobre la gráfica de la distribución.

8.5. Contrastes de hipótesis para una media

Los estadísticos de prueba para diferentes contrastes de hipótesis para una población son los que siguen y son muy similares a los ya empleados para construir los intervalos de confianza que vimos en el tema anterior.

Empezaremos por el contraste para una media. Supongamos que partimos de una distribución $N(\mu; \sigma)$ de la cual cogemos una muestra aleatoria simple de tamaño n (x_1, x_2, \dots, x_n). Además, fijamos un nivel de significación α . Llegados a este punto podemos plantear como hemos visto diferentes tipos de contrastes:

$$\begin{aligned} H_0: \mu &= \mu_0 \\ H_1: \mu &\neq \mu_0 \end{aligned}$$

La regla de decisión será:

Rechazaremos H_0 si:

$$z_{exp} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha/2} \text{ o } z_{exp} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha/2}$$

Aceptaremos H_0 si se cumple que

$$-z_{\alpha/2} \leq z_{exp} \leq z_{\alpha/2}.$$

También podemos plantear el contraste según caiga en la región de aceptación o en la de rechazo.

$$\mu_0 \in \left[\bar{x} - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

aceptamos H_0

Mientras que si está fuera de dicho intervalo será equivalente a que caiga en la región crítica y por tanto rechazaremos H_0 .

$$\mu_0 \notin \left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}; \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

rechazamos H_0

Para el método del p-valor, tenemos que calcular la probabilidad de que el estadístico de prueba tome un valor al menos tan externo como

$$\bar{x}$$

Distinguiremos dos casos:

Si $\bar{x} < \mu_0$ ($Z_{exp} < 0$) entonces el p-valor es $2 \cdot P(Z < Z_{exp})$.

Si $\bar{x} > \mu_0$ ($Z_{exp} > 0$) entonces el p-valor es $2 \cdot P(Z > Z_{exp})$.

► $\begin{cases} H_0: \mu \leq \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$

Este contraste ya no es bilateral, sino que es unilateral, pues la hipótesis alternativa se plantea hacia un solo lado o cola (solo si es mayor). Otra diferencia respecto al caso anterior es que la región de aceptación ya no es un intervalo entre dos valores críticos, sino que va desde $-\infty$ hasta el valor crítico que esta vez es único.

Rechazaremos cuando:

$$Z_{exp} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} > z_\alpha$$

Mientras que aceptaremos si:

$$z_{exp} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \leq z_\alpha$$

Para el método del p-valor, tenemos que calcular la probabilidad de que el estadístico de prueba tome un valor al menos tan externo como

$$\bar{x}$$

Para este tipo de contraste unilateral, esta probabilidad viene dada por:

$$p - valor = P(Z > Z_{exp})$$

$$\begin{cases} H_0: \mu \geq \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$$

Es parecido al caso unilateral anterior solo que con la cola hacia el otro lado (se rechaza cuando es menor en lugar de cuando es mayor). Por tanto, las reglas de decisión serán contrarias a las de antes. Rechazaremos cuando:

$$z_{exp} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} < -z_\alpha$$

Mientras que aceptaremos si:

$$z_{exp} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \geq -z_\alpha$$

El cálculo del p-valor es similar al caso unilateral anterior. La única diferencia es que se modifica el sentido de la desigualdad para calcular la probabilidad:

$$p - valor = P(Z < Z_{exp})$$

Todos estos contrastes pueden llevarse a cabo con una pequeña adaptación en el caso de que σ no sea conocida, sin más que sustituir como referencia el valor crítico $Z_{\alpha/2}$ el equivalente en la distribución T-Student: $t_{\alpha/2}$ y también realizando otra sustitución usando la cuasidesviación típica s como estimador de la desconocida σ . En tal caso los grados de libertad que hay que manejar son iguales al tamaño de la muestra menos uno.

Ejemplo 7

Se dispone de una muestra de plasma de un adulto que contiene $2,4 \text{ mg/l}$ de hierro. Determinado aparato electrónico de medida estable diez determinaciones de la medida para la concentración de hierro, las cuales resultan:

2,4	2,2	2,5	3	3,2	3,3	3	3,4	3,2	3,3
-----	-----	-----	---	-----	-----	---	-----	-----	-----

La v.a. en este caso es X = «Concentración medida en una determinación en mg/l». La media de dicha v.a. es desconocida y la varianza también, pero partiremos suponiendo que la conocemos y vale 0,1. Partiremos suponiendo otra cosa más, y es que la variable aleatoria es normal (cosa habitual en este tipo de variables según quedó explicado en el capítulo donde se vio la normal). En este caso el contraste es de la forma:

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

Donde μ_0 es la media que resulta 2,4, luego resulta:

$$H_0: \mu = 2,4$$

$$H_1: \mu \neq 2,4$$

Para contrastarlo necesitamos hallar primeramente su estadístico de contraste que vale:

$$z_{exp} = \frac{2,95 - 2,4}{\sqrt{0,1/10}} = 5,50$$

Ahora elegiríamos el método que vamos a utilizar para contrastar dicho estadístico. Nosotros emplearemos los tres métodos para practicar, pero, en el futuro, cuando trabajemos con contrastes de hipótesis bastará con que empleemos el del *p – valor*. Una vez hecho esto fijamos el α que decidimos que valga 0,01 que es uno de los niveles de significación más habituales.

Para contrastar a la manera tradicional simplemente contrastaríamos Z_{exp} frente a $Z_{\alpha/2}$ entonces en este caso tenemos según nos indican las tablas de la normal que $Z_{0,005} = 2,57$ y entonces como $Z_{exp} = 5,50 > 2,57 = Z_{0,005}$ rechazamos H_0 , por lo que estaríamos encontrando evidencias estadísticamente significativas para un $\alpha = 0,01$ de que la media de concentración de hierro en sangre difiere de 2,4mg/l. Es decir, nos estaríamos quedando con H_1 .

Vayamos ahora al caso en que empleamos el método del p-valor, que es el más empleado actualmente, pues es más rápido al no tener que recurrir a ninguna tabla ni nada (cuando lo facilita el programa estadístico claro, porque hecho a mano al menos requiere su cálculo...).

Para ello, lo que hacemos es tratar de localizar la probabilidad de que el estadístico tome ese valor o uno más extremo:

$$p - valor = 2 \cdot p(Z \geq z_{exp}) = 2 \cdot p(Z \geq 5,50) = 0,000000038 < 0,01$$

El p-valor resultante es virtualmente igual a 0 (3,8 •10-8 para ser exactos) por lo que sabemos que vamos a rechazar la H_0 siempre, ya que este p-valor es menor que cualquier nivel de significación razonable que impongamos.

Es interesante que observes la diferencia conceptual del método del *p – valor* en la que las comparaciones se establecen en el nivel de las probabilidades, lo cual tiene, para empezar, una ventaja y es que no necesitamos saber el valor del estadístico de contraste al que equivale un nivel de significación dado, ya que comparamos directamente con las probabilidades y no tenemos que «retroceder» al nivel de los estadísticos para comparar entre ellos. Por el contrario, tendremos que «avanzar» un paso una vez dispongamos de nuestro valor crítico de contraste Z_{exp} para hallar la probabilidad asociada a él, que será nuestro *p – valor*, el cual compararemos entonces con el α .

Por último, según el método del intervalo de confianza calcularíamos el IC para la media para una confianza del 99% (o lo que es lo mismo para un $\alpha = 0,01$) el cual resulta:

$$\begin{aligned} P \left[-z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \right] &= 1 - \alpha \leftrightarrow P \left[-2,57 \leq \frac{2,95 - \mu}{\sqrt{0,1}/\sqrt{10}} \leq 2,57 \right] = 0,99 \leftrightarrow \\ &\leftrightarrow P \left[2,95 - 2,57 \frac{\sqrt{0,1}}{\sqrt{10}} \leq \mu \leq 2,95 + 2,57 \frac{\sqrt{0,1}}{\sqrt{10}} \right] = 0,99 \leftrightarrow \mu \\ &\in (2,69; 3,21) \end{aligned}$$

Y como 2,4 está fuera del intervalo de confianza, entonces rechazaríamos H_0 . Observa, por último, la diferencia entre el método tradicional y el del IC. En el IC comparamos directamente el valor hipotetizado del parámetro en H_0 que es 2,4 con los valores «aceptados» del parámetro en el IC que construimos, mientras que en el método de tradicional nos movemos en comparaciones de valores críticos de la distribución que estemos manejando.

8.6. Contrastes de hipótesis para la proporción

Pasaremos ahora a los **contrastes sobre la proporción**. De nuevo empleamos estadísticos muy similares a los que ya usamos en los intervalos de confianza.

Esta vez el estadístico de prueba es:

$$z_{exp} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

El primer **contraste** del que podemos partir es de nuevo el **bilateral**:

$$\begin{cases} H_0: p = p_0 \\ H_1: p \neq p_0 \end{cases}$$

Que aceptaremos si:

$$-z_{\alpha/2} \leq z_{exp} \leq z_{\alpha/2}$$

Ya que entonces estaría cayendo dentro de la región de aceptación. Por el contrario, si

$$z_{exp} < -z_{\alpha/2} \text{ o } z_{exp} > z_{\alpha/2}$$

rechazamos H_0 , porque estaría cayendo en la región crítica (que está conformada por dos partes).

El siguiente caso es el **contraste unilateral** de este tipo:

$$\begin{cases} H_0: p \leq p_0 \\ H_1: p > p_0 \end{cases}$$

Que aceptaremos si

$$z_{exp} \leq -z_\alpha$$

Ya que entonces estaría cayendo dentro de la región de aceptación. Por el contrario, si $Z_{exp} > Z_\alpha$ rechazamos H_0 , porque estaría cayendo en la región crítica.

El tercer caso lo tenemos cuando el **contraste unilateral** es para el otro lado.

$$\begin{cases} H_0: p \geq p_0 \\ H_1: p < p_0 \end{cases}$$

Que aceptamos si

$$z_{exp} \geq -z_\alpha$$

Por el contrario, si

$$z_{exp} < -z_\alpha$$

rechazamos H_0 , porque estaría cayendo en la región crítica.

El cálculo del *p – valor* para los tres casos se realizaría de forma análoga a lo visto para los contrastes de hipótesis para una media.

Ejemplo 8

Volvamos sobre el ejemplo 6. En él vimos que se rechazaba la H_0 en base al método del IC, que era el que queríamos ejemplificar en este momento. Veamos ahora qué ocurre con el método tradicional para contrastar la hipótesis. Lo primero que hacemos es plantear las hipótesis:

H_0 : la proporción de jóvenes que fuman es de 0,5.

H_1 : la proporción de jóvenes que fuman es distinta de 0,5.

Que como vimos de modo simplificado quedan:

$$H_0: p = 0,5$$

$$H_1: p \neq 0,5$$

Se trataba de un contraste bilateral:

$$\begin{cases} H_0: p = p_0 \\ H_1: p \neq p_0 \end{cases}$$

Por lo que el estadístico de contraste

retomando los datos iniciales del ejemplo con $n=100$ y $\hat{p} = 0,3$ resulta:

$$z_{exp} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leftrightarrow z_{exp} = \frac{0,3 - 0,5}{\sqrt{\frac{0,5(0,5)}{100}}} = -12,65$$

Que es un valor crítico extremadamente pequeño y mucho menor que $Z_{\alpha/2} = -Z_{0.025} = 1,96$ por lo que como $z_{exp} < z_\alpha$ rechazamos por estar cayendo en la región crítica.

El p-valor viene dado por $2 \bullet p(Z \leftarrow -12,65) \approx$. Esto nos indica que rechazaremos siempre la hipótesis nula: sin importar cuan pequeño sea α siempre tendremos que $p(Z < z_{exp}) < \alpha$.

8.7. Contrastes de hipótesis sobre la varianza

Por último, en cuanto a contrastes de parámetros de una población, veremos el contraste de hipótesis de la varianza para una población que, de nuevo, ha de distribuirse como una normal con media μ y varianza σ^2 desconocidas.

En los casos de la mediana y la proporción, también este requisito era importante pero ahora la normalidad lo será incluso más, pues los contrastes de hipótesis con el estadístico basado en la Chi Cuadrada no son robustos si falla la presunción de normalidad.

Emplearemos el estadístico que se distribuye como una Chi Cuadrada que ya empleamos para los intervalos de confianza.

El estadístico de contraste es entonces este:

$$\chi_{exp}^2 = \frac{(n - 1) s^2}{\sigma_0^2}$$

Tenemos tres casos posibles que vamos a estudiar como ya hicimos con la media y la proporción:

$$\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 \neq \sigma_0^2 \end{cases}$$

Se rechazará cuando

$$\chi_{exp}^2 < \chi_{n-1,\alpha/2}^2 \text{ o } \chi_{exp}^2 > \chi_{n-1,1-\alpha/2}^2$$

Mientras que aceptaremos cuando

$$\chi^2_{n-1,\alpha/2} \leq \chi^2_{exp} \leq \chi^2_{n-1,1-\alpha/2}$$

$$\begin{cases} H_0: \sigma^2 \leq \sigma_0^2 \\ H_1: \sigma^2 > \sigma_0^2 \end{cases}$$

Se rechazará cuando

$$\chi^2_{exp} > \chi^2_{n-1,1-\alpha}$$

Mientras que se aceptará cuando

$$\chi^2_{exp} \leq \chi^2_{n-1,1-\alpha}$$

$$\begin{cases} H_0: \sigma^2 \geq \sigma_0^2 \\ H_1: \sigma^2 < \sigma_0^2 \end{cases}$$

Que se rechazará cuando

$$\chi^2_{exp} < \chi^2_{n-1,\alpha}$$

Mientras que aceptaremos cuando

$$\chi^2_{exp} \geq \chi^2_{n-1,\alpha}$$

El cálculo del p-valor para los tres casos se realizaría de forma análoga a lo visto para los contrastes de hipótesis para una media, pero utilizando la variable Chi cuadrada con $n - 1$ grados de libertad para calcular las probabilidades.

Ejemplo 9: ¿Están cumpliendo las embotelladoras españolas de Coca-Cola?

La marca Coca Cola impone a sus empresas embotelladoras un riguroso control de calidad, para que sean capaces de embotellar el producto con una varianza mínima de líquido contenido en cada botella, cantidad que establecen en los contratos pertinentes. En este caso la marca registrada Coca Cola España® ha acordado con la embotelladora *Paco's bottle S.A.* (sita en Cáceres) embotellar el producto con una desviación típica $\sigma = 0,151$ cl.

Para testar el control de calidad se ha recogido la siguiente muestra de 24 botellas anotando los centilitros que contienen (se trata de botellas de 50cl):

47,26	47,26	47,29	47,26	47,02	47,38	47,29	47,11
47,38	47,38	47,38	47,20	47,35	47,32	47,29	47,17
47,17	47,20	47,20	47,38	47,29	47,53	47,11	47,47

Para ver si realmente está cumpliendo lo acordado vamos a contrastar:

$$\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 \neq \sigma_0^2 \end{cases}$$

La razón por la que elegimos este contraste unilateral es porque cuando no va a cumplir la empresa española es cuando la desviación típica sea mayor de lo acordado con Coca Cola.

$$H_0: \sigma^2 \leq 0,023$$

$$H_1: \sigma^2 > 0,023$$

Para aplicar el estadístico de contraste basado en la Chi Cuadrada debemos primero calcular $s = 0,119\text{cl}$.

$$\chi_{epx}^2 = \frac{(n - 1) \times s^2}{\sigma_0^2} \leftrightarrow \chi_{epx}^2 = \frac{(24 - 1) \times (0,119)^2}{0,023} = 14,161$$

Tras fijar nuestro $\alpha = 0,01$ (también lo exige Coca Cola) acudiríamos a las tablas para localizar nuestro valor crítico:

$$\chi^2_{n-1;1-\alpha} \rightarrow \chi^2_{23;0,01} = 41,06$$

Luego como:

$$\chi_{epx}^2 = 13,90 < 41,60 = \chi^2_{23;0,01}$$

Entonces aceptamos H_0 , lo cual equivale a afirmar que no hemos encontrado evidencias estadísticamente significativas de que la desviación típica de centilitros por botella sea mayor de lo acordado entre ambas empresas.

8.8. Contrastes paramétricos para dos muestras

Aunque los contrastes sobre el parámetro de una población son empleados, no cabe duda de que los que comparan dos poblaciones están más extendidos, seguramente, porque el interés suele girar en torno al estudio de si a partir de los datos podemos contrastar, si hay o no diferencia entre las distribuciones de dos o más grupos.

Ejemplo 10: Sobre qué tipo de cuestiones podemos tratar en los contrastes de hipótesis para dos muestras.

Este tipo de comparaciones entre grupos suelen ser del tipo de «¿Existen diferencias entre los CI de personas con trastornos psicóticos y las que presentan rasgos paranoides?» O «¿Existen diferencias entre las cantidades de aceite producidas en las diferentes cooperativas de aceite de la provincia de Castellón? ¿Son igual de longevas dos especies de hormigas? ¿Dos fármacos son igualmente efectivos para tratar determinada patología?» Son este tipo de cuestiones las que se pueden contrastar empleando contrastes de hipótesis entre dos poblaciones.

Todas estas comparaciones se establecen entre dos muestras que provendrán de dos poblaciones, aunque en ocasiones ambas muestras contienen datos relacionados, que denominamos entonces **datos apareados**, los cuales hacen que las muestras de donde procedan pasen a denominarse **muestras apareadas**.

Esta diferenciación es especialmente útil en ciencias de la salud, aunque esta presenta en otras áreas donde se aplica la estadística. El objetivo del estudio de las muestras apareadas es reducir las fuentes de variabilidad entre las muestras. Ejemplos de muestras apareadas son: cuando se toman análisis de sangre de los empleados de una empresa y un año después se vuelven a realizar análisis a los mismos empleados.

En realidad las muestras apareadas son un ejemplo concreto de **muestras dependientes o relacionadas**, ya que los datos de una muestra toman valores que no son libres de los valores que tome la otra sino que tienen una relación.

Si un empleado de la empresa anterior tenía hace un año 300 de índice de «colesterol del malo», seguramente seguirá teniendo una medida alta, aunque puede que se haya empezado a cuidar.

Por el contrario, las muestras se dice que son **independientes** cuando son independientes los valores que tomen las observaciones en una de ellas con los que tomen en la otra.

Tras esta breve introducción podemos pasar a explicitar algunos de los principales casos de comparación entre dos poblaciones con sus respectivos estadísticos y las condiciones previas que han de cumplir las poblaciones a comparar. Veremos solo el caso de comparación de medias y de proporciones en el caso de muestras independientes para simplificar.

A continuación, presentamos el contraste de hipótesis para comparar dos medias provenientes de dos poblaciones normales con varianzas **desconocidas e iguales**, de las que se recogen sendas muestras aleatorias e independientes. Nos basamos en que la distribución de las diferencias de medias de dos normales es también normal, por las propiedades que vimos cuando estudiamos la normal. Así, tenemos el siguiente estadístico de comparación:

$$t_{exp} = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\hat{s} \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

Donde

$$\hat{S} = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}}$$

También es válida la condición de que el tamaño de ambas muestras sea aproximadamente de 30 observaciones o más: n_x y $n_y \geq 30$.

Donde las dos poblaciones son x e y y $\mu_x - \mu_y$ es el valor que se establece como diferencia entre ambas en la H_0 . Además

$$s_x^2, s_y^2 \text{ y } n_x, n_y$$

hacen referencia a sus respectivas varianzas y tamaños de las muestras.

$$\begin{cases} H_0: \mu_x - \mu_y = d_0 \\ H_1: \mu_x - \mu_y \neq d_0 \end{cases}$$

Un caso más particular es cuando se parte de establecer que sean ambas medias iguales por lo que $d = 0$.

$$H_0: \mu_x - \mu_y = 0 \leftrightarrow H_0: \mu_x = \mu_y$$

$$H_1: \mu_x - \mu_y \neq 0 \leftrightarrow H_1: \mu_x \neq \mu_y$$

En este caso el estadístico anterior queda simplificado a:

$$t_{exp} = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\hat{S} \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$$

Que aceptaremos cuando

$$-t_{n_x+n_y-2; \alpha/2} \leq t_{exp} \leq t_{n_x+n_y-2; \alpha/2}$$

y rechazaremos en caso contrario, cuando caiga en la región de rechazo

$$t_{exp} > -t_{n_x+n_y-2; \alpha/2} \text{ o } t_{exp} > t_{n_x+n_y-2; \alpha/2}$$

También tenemos los casos de contraste unilateral:

$$\begin{cases} H_0: \mu_x - \mu_y \leq d_0 \\ H_1: \mu_x - \mu_y > d_0 \end{cases}$$

Este caso lo aceptamos cuando

$$t_{exp} \leq t_{n_x+n_y-2; \alpha}$$

y lo rechazamos cuando

$$t_{exp} > t_{n_x+n_y-2; \alpha}$$

Por último, tenemos este otro contraste unilateral:

$$\begin{cases} H_0: \mu_x - \mu_y \geq d_0 \\ H_1: \mu_x - \mu_y < d_0 \end{cases}$$

Este caso lo aceptamos cuando

$$t_{exp} \geq -t_{n_x+n_y-2; \alpha} \text{ y lo rechazamos cuando } t_{exp} < -t_{n_x+n_y-2; \alpha}$$

El cálculo del *p – valor* para los tres casos se realizaría de forma análoga a lo visto para los contrastes de hipótesis para una media, pero utilizando la variable T-Student $n_x + n_y - 2$ con grados de libertad para calcular las probabilidades.

El contraste de hipótesis para la comparación de dos medias con varianzas desconocidas y distintas es similar, pero se modifican las fórmulas para el estadístico de contraste. Podéis encontrar la fórmula para esta prueba en (Rius, 1998). A la hora de decidir qué prueba usar podemos realizar un contraste para determinar si las varianzas son iguales (ver, por ejemplo, Rius (1998)).

Ejemplo 11: ¿Hubo discriminación por la estatura?

Un grupo de aspirantes a un puesto de vigilante de seguridad que no fueron elegidos por una subcontrata del Metro de Madrid piensa en poner una denuncia, pues estima que hubo discriminación porque cogieron en el proceso de selección a los más altos en estatura, cuando la empresa en las condiciones no decía nada de ese punto.

Gracias a la estadística podemos resolver este asunto y descubrir si realmente hubo discriminación en este punto y, por consiguiente, su demanda podría tener un buen apoyo de cara a una resolución judicial favorable.

Veamos. Las estaturas en centímetros de los seleccionados fueron:

169	172	172	173	176	177	178	179	179	180
180	180	181	183	184	188	188	189	189	190
191	192	195							

Mientras que la de los no seleccionados fueron las siguientes:

162	168	171	172	173	173	174	177	177	178
178	179	179	179	180	180	180	180	181	181
182	182	183	183	184	184	186	186	187	189

El contraste es del tipo:

$$H_0: \mu_{seleccionados} \leq \mu_{no\ seleccionados}$$

$$H_1: \mu_{seleccionados} > \mu_{no\ seleccionados}$$



$$H_0: \mu_{seleccionados} \leq \mu_{no\ seleccionados} \leq 0$$

$$H_1: \mu_{seleccionados} \leq \mu_{no\ seleccionados} > 0$$

El cual se rechazará cuando la media del primer grupo —de los cogidos por la empresa— sea mayor que la de los no seleccionados.

Con estos datos resultan:

$$\bar{x} = 181,96\text{cm} ; \bar{y} = 178,93\text{cm} \text{ y } s_x = 7,22 ; s_y = 5,88 \text{ y } n_x = 23 , n_y = 30.$$

Donde la variable X corresponde a la altura de los seleccionados y la variable Y a la de los no seleccionados. De modo que el estadístico de contraste que se obtiene es:

$$\hat{s} = \sqrt{\frac{(23 - 1)7,22^2 + (30 - 1)5,88^2}{23 + 30 - 2}} = 6,492$$

$$t_{exp} = \frac{181,96 - 178,93}{6,492 \sqrt{\frac{1}{23} + \frac{1}{30}}} = 1,684$$

Supongamos que $\alpha = 0,05$ está bien para concluir tal diferencia. Ahora acudiríamos a las tablas correspondientes de la T-Student con $n_x + n_y - 2 = 51$ grados de libertad y vemos que:

$$t_{51;0,05} = 1,675$$

Como $t_{exp} = 1,684 > t_{51;0,05} = 1,675$, rechazaríamos la hipótesis nula de que son iguales las estaturas de ambos grupos y, por tanto, podemos afirmar que ha existido discriminación.

Empleemos ahora el método del *p – valor*. Para este contraste unilateral tenemos:

$$p\text{-valor} = P(T_{51} > 1,684) = 0,049 \blacksquare$$

Puesto que $0,049 < 0,05 = \alpha$, rechazamos la hipótesis nula. Sin embargo, observando el p-valor vemos que hemos estado muy cerca de no rechazar H_0 . De hecho, si hubiésemos tomado $\alpha = 0,01$ entonces aceptaríamos H_0 .

Nos quedaría por ver el contraste de proporciones entre dos poblaciones. Esta vez supondremos normalidad planteando como requisito que $np > 5$ y $n(1 - p) > 5$ para ambas poblaciones. Así tenemos que el estadístico que emplearíamos para el contraste es similar al anterior para las medias, pero sufriendo la lógica adaptación al caso de la proporción (cambio de parámetro y de la varianza que aquí pasa a ser $p(1 - p)/n$ como ya sabemos).

$$z_{exp} = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}}$$

Donde ahora los subíndices 1 y 2 hacen mención a sendas muestras. Fíjate que esta vez en que el estadístico de contraste se distribuye normalmente de modo que tenemos que recurrir a la tabla de la normal(0,1) para resolver los diferentes contrastes que se nos plantean que son los tres siguientes dependiendo si es bilateral o uno de los dos casos posibles de unilateral:

Caso bilateral:

$$H_0: p_1 = p_2$$

$$H_1: p_1 \neq p_2$$

Se acepta si:

$$-z_{\alpha/2} \leq z_{exp} \leq z_{\alpha/2}$$

Rechazándose en caso contrario.

El caso unilateral izquierdo:

$$H_0: p_1 \geq p_2$$

$$H_1: p_1 < p_2$$

Que se acepta si:

$$z_{exp} \geq -z_\alpha$$

Rechazándose en caso contrario.

Y por último tendríamos el caso unilateral derecho:

$$H_0: p_1 \leq p_2$$

$$H_1: p_1 > p_2$$

El cual aceptamos si:

$$z_{exp} \leq z_\alpha$$

Rechazándose en caso contrario.

El cálculo del *p – valor* para los tres casos se realizaría de forma análoga a lo visto para los contrastes de hipótesis para una media.

8.9. Contrastes de hipótesis robustos

De forma análoga a como se diseñaron los intervalos de confianza robustos a partir de los intervalos de confianza, es posible definir **contrastes de hipótesis robustos** que sean poco sensibles a la presencia de valores *outliers*. Para ello, en el caso de la media, utilizaremos la **media recortada** y la **cuasidesviación típica winsorizada**.

Por tanto, a la hora de realizar el siguiente **contraste bilateral** con un nivel de significación β :

$$H_0: \mu_\alpha = \mu_0$$

$$H_1: \mu_\alpha \neq \mu_0$$

Aceptaremos la hipótesis nula si:

$$\mu_0 \in \left[\bar{x}_\alpha - t_{n-2k-1; \frac{\beta}{2}} \frac{S_w}{(1-2\alpha)\sqrt{n}}, \bar{x}_\alpha + t_{n-2k-1; \frac{\beta}{2}} \frac{S_w}{(1-2\alpha)\sqrt{n}} \right]$$

Por el contrario, **rechazaremos la hipótesis nula** si se cumple lo siguiente:

$$\mu_0 \notin \left[\bar{x}_\alpha - t_{n-2k-1; \frac{\beta}{2}} \frac{S_w}{(1-2\alpha)\sqrt{n}}, \bar{x}_\alpha + t_{n-2k-1; \frac{\beta}{2}} \frac{S_w}{(1-2\alpha)\sqrt{n}} \right]$$

Con los **contrastes de hipótesis unilaterales** se sigue un proceso similar. Para el siguiente contraste:

$$H_0: \mu_\alpha \leq \mu_0$$

$$H_1: \mu_\alpha > \mu_0$$

Aceptamos la hipótesis nula si:

$$\mu_0 \geq \bar{x}_\alpha - t_{n-2k-1;\beta} \frac{S_w}{(1-2\alpha)\sqrt{n}}$$

Rechazamos la hipótesis nula si:

$$\mu_0 < \bar{x}_\alpha - t_{n-2k-1;\beta} \frac{S_w}{(1-2\alpha)\sqrt{n}}$$

Por último, para:

$$H_0: \mu_\alpha \geq \mu_0$$

$$H_1: \mu_\alpha < \mu_0$$

Aceptamos la hipótesis nula si:

$$\mu_0 \leq \bar{x}_\alpha + t_{n-2k-1;\beta} \frac{S_w}{(1-2\alpha)\sqrt{n}}$$

Rechazamos la hipótesis nula si:

$$\mu_0 > \bar{x}_\alpha + t_{n-2k-1;\beta} \frac{S_w}{(1-2\alpha)\sqrt{n}}$$

Tal y como hemos visto, los contrastes de hipótesis que acabamos de observar dan por hecho que el modelo de los datos sigue una distribución específica: la distribución normal. Aunque, por lo general, son aplicables en gran cantidad de casos, ¿qué opciones tenemos si nuestros datos no siguen dicha distribución? Una posible solución consiste en utilizar **contrastos de hipótesis robustos basados en el r – valor**. Dichos contrastes **no son dependientes de que los datos sigan una distribución normal** e incluyen en los cálculos una nueva variable, la variable F , consistente en la distribución específica que siguen los datos con los que estamos trabajando.

Como puede verse, este tipo de contrastes de hipótesis son **aplicables independientemente del tipo de distribución**, aunque, por supuesto, esta debe de ser **conocida**.

Principalmente, los contrastes de hipótesis robustos definen un nuevo concepto denominado **r-valor** (*robust value*) que, de forma similar al p-valor, ayuda a determinar **si se cumple o no la hipótesis alternativa**. Dicho valor se construye en función de la distribución que siguen los datos. Para el caso del siguiente contraste de hipótesis:

$$\begin{cases} H_0: \theta \leq \theta_0 \\ H_1: \theta > \theta_0 \end{cases}$$

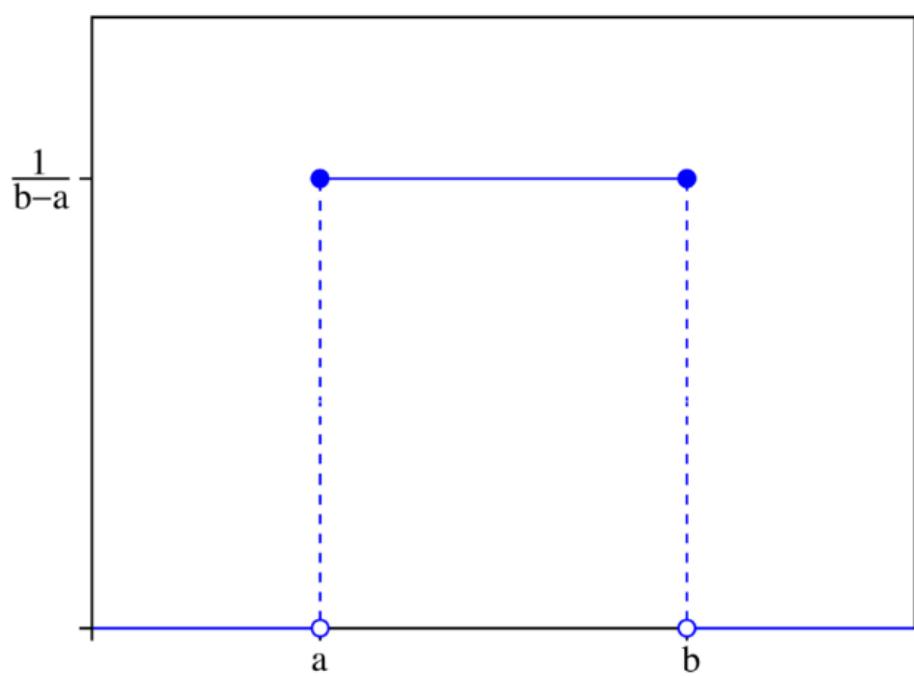
Donde θ hace referencia a la variable que estamos intentando comprobar, es posible definir el r-valor como:

$$R(F) = 1 - 8 c d (F, Uniforme)$$

Donde $R(F)$ es el r-valor asociado a la distribución F , F es la distribución que siguen los datos, c es la máxima frecuencia relativa de la muestra y la función d calcula la distancia entre las distribuciones. Dicha medida puede definirse tal y como se muestra a continuación:

$$d(F, G) = \int_{\theta}^{\infty} |F(x) - G(x)| dx$$

Por último, Uniforme hace referencia a una **distribución uniforme construida a partir de la variable que queremos contrastar**. La distribución uniforme tiene la siguiente forma:

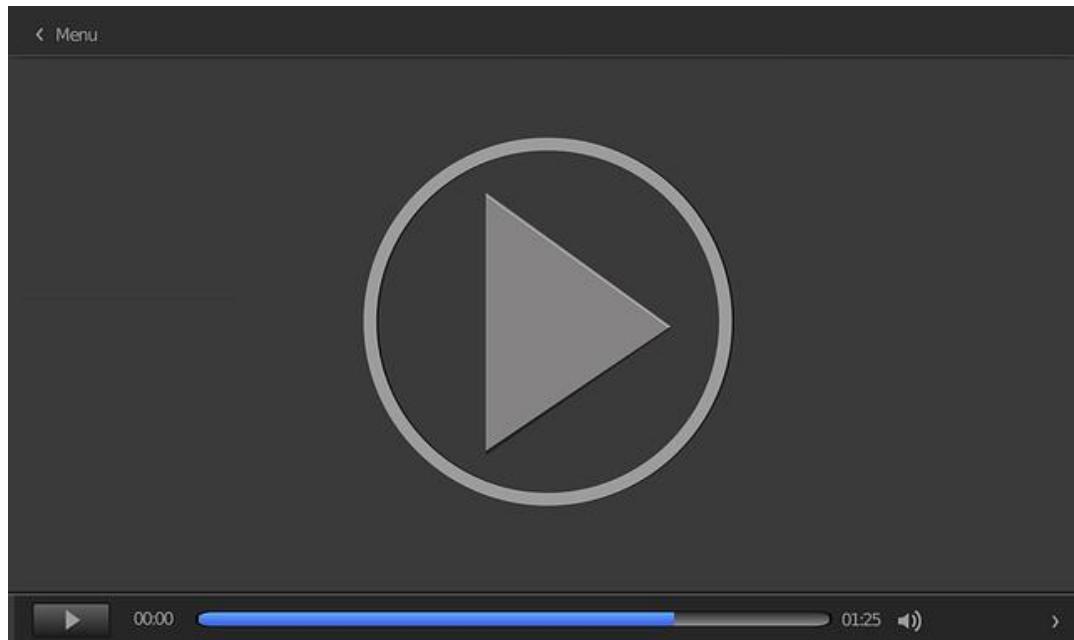


A pesar de la gran utilidad que poseen los contrastes de hipótesis robustos basados en r-valores, su utilización práctica es aún limitada siendo un **campo que está actualmente en pleno auge de desarrollo**. Se prevé que, en entornos Big Data, los contrastes de hipótesis robustos tendrán un gran uso debido a que son capaces de trabajar con distribuciones de datos complejas.

Debido a que su utilidad práctica no ha sido demostrada hasta hace poco, no hay disponible ningún paquete en R ni en otras plataformas que permitan automatizar su uso.

Infiriendo con herramientas estadísticas: contrastes de hipótesis

En este vídeo vamos a conocer dos herramientas fundamentales en la estadística inferencial.



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=bd35e27f-5e28-4937-9185-acbd00cf7efc>

8.10 Referencias bibliográficas

Contrastes de hipótesis de modelo variable. Recuperado de

http://www.ine.es/ss/Satellite?blobcol=urldata&blobheader=application%2Fpdf&blobheadername1=Content-Disposition&blobheadervalue1=attachment%3B+filename%3D525%2F945%2F115_3.pdf&blobkey=urldata&blobtable=MungoBlobs&blobwhere=525%2F945%2F115_3.pdf&ssbinary=true

García, A. (2005). *Métodos avanzados de estadística aplicada. Métodos robustos y de remuestreo*. Madrid: UNED.

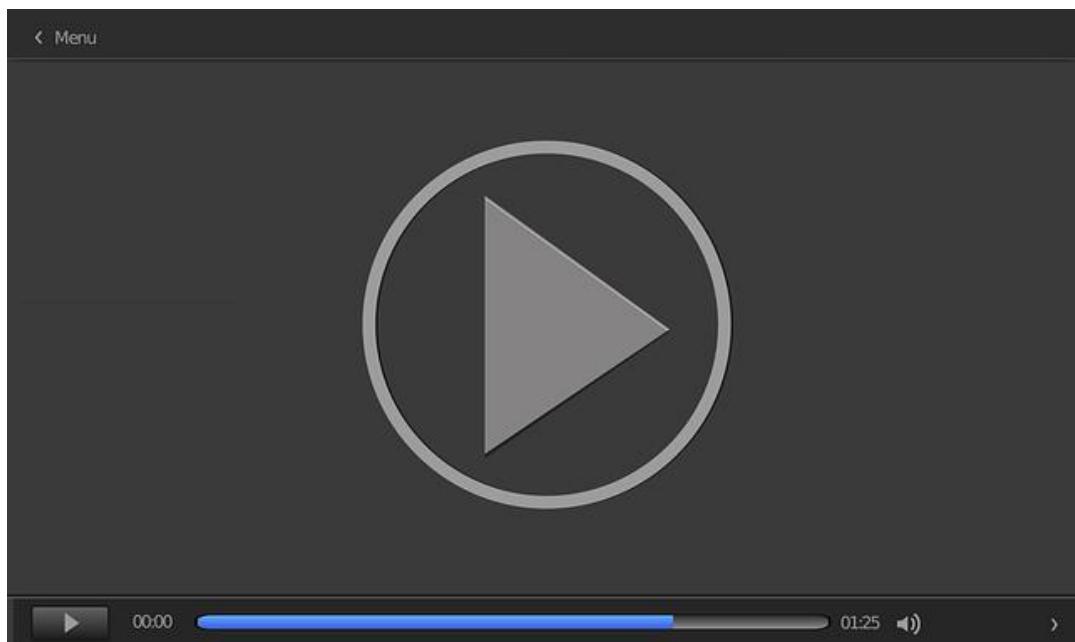
Ríus, F. (1998). *Bioestadística: Métodos y aplicaciones*. Málaga: Universidad de Málaga. Publicaciones. Versión electrónica:

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

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

Aprendiendo a interpretar los resultados de un contraste de hipótesis efectuado con un programa estadístico y/o un artículo científico

En esta lección magistral veremos cómo podemos manejarnos con las tablas y resultados que suelen ofrecer los programas estadísticos. También mostraremos cómo manejarnos con las tablas que suelen aparecer en los artículos científicos haciendo referencia a diferentes contrastes de hipótesis.



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=35822bff-bf2a-43e5-b38f-abdc00f9818e>

A Close Look at Therapeutic Touch

En este artículo se plantea una sencilla investigación estadística para dilucidar si ciertos paramédicos que afirmaban poder detectar y curar el «aura» de sus pacientes tenían alguna clase de fundamento o si, por el contrario, se trataba de una patraña con tintes estafadores. Cómo curiosidad la autora del artículo era menor de edad cuando realizó dicha investigación con la ayuda de, entre otros, su madre que era investigadora.

Accede al artículo desde el aula virtual o a través de la siguiente dirección web: <http://jama.jamanetwork.com/article.aspx?articleid=187390#Abstract>

Comparación de medias de dos poblaciones

Se recomienda que, para profundizar, le eches un vistazo a este enlace (en inglés) de una web australiana sobre estadística, donde se trata un caso de contraste de hipótesis de dos medias de diferentes poblaciones de un modo bastante completo y además usando un conocido programa estadístico (Minitab).

Accede a la página desde el aula virtual o a través de la siguiente dirección web:

<http://surfstat.anu.edu.au/surfstat-home/>

Bibliografía

- Kreyszig, E. (1983). *Introducción a la Estadística Matemática*. México: Limusa.
- Martín Andrés, A. (2004). *Bioestadística para las ciencias de la salud*. Madrid: Norma-Capitel.
- Moore, D. S. (2006). *Introduction to the practice of statistics* (5th. ed.). New York: Freeman and Company.

1. ¿Para qué sirve un contraste de hipótesis?

 - A. Para saber a ciencia cierta si una afirmación es verdadera.
 - B. Para saber con rigor matemático si una afirmación es falsa.
 - C. Para contrastar si una afirmación sobre una población parece cierta o no en base a los datos.
 - D. Las respuestas A y B son correctas.
2. Generalmente el orden que seguimos en un contraste de hipótesis es...

 - A. Establecer primero la H_0 para que quede delimitada aproximadamente la H_1 .
 - B. Plantearnos la hipótesis de rechazo para posteriormente fabricar su complementaria que es la hipótesis nula.
 - C. Negar la hipótesis alternativa para que surja la hipótesis nula.
 - D. Las respuestas B y C son correctas.
3. ¿Cuántos diferentes planteamientos tenemos para la H_1 ?

 - A. Solo tenemos una opción de H_1 , ya que está totalmente delimitada por la hipótesis nula.
 - B. Generalmente consideramos tres clases, donde una viene asociada al contraste bilateral y las otras dos a los laterales izquierdo y derecho respectivamente.
 - C. Cuatro, correspondiendo a los tres casos indicados en B) más un bilateral del tipo (que por cierto es un tanto atípico).
 - D. Las respuestas B y C son correctas.

4. ¿Con que tipo de frases interpretamos un contraste de hipótesis?
- A. «...tras realizar el contraste se demuestra la falsedad de la H_0 ...».
 - B. «...se han encontrado evidencias claras de que es la H_0 es verdadera...».
 - C. «... se han encontrado evidencia estadísticas aplastantes en favor de...».
 - D. «... no logramos encontrar evidencias estadísticamente significativas...»
5. ¿Quién tiene que caer en la región crítica para que rechacemos la H_0 ?
- A. El estadístico de contraste que estemos empleando.
 - B. El parámetro.
 - C. El valor que se contrasta del parámetro (por ejemplo μ).
 - D. Las respuestas B y C son correctas.
6. ¿Qué es $1 - \beta$?
- A. La probabilidad de que cometamos uno de los errores más graves que se pueden cometer.
 - B. La potencia del contraste.
 - C. El otro gran error en los contrastes de hipótesis junto con α .
 - D. Las respuestas B y C son correctas.
7. La $P(\text{rechazar } H_0 \mid \text{siendo } H_0 \text{ verdadera})$ es equivalente a:
- A. α
 - B. $1 - \alpha$
 - C. β
 - D. $1 - \beta$

8. La regla del p valor para decidir un contraste de hipótesis es:

- A. Si $p\text{ valor} < \alpha \rightarrow \text{Aceptamos } H_0$ mientras que si $p\text{ valor} \geq \alpha$ rechazamos.
- B. Si $p\text{ valor} > \alpha \rightarrow \text{Aceptamos } H_0$ mientras que si $p\text{ valor} \leq \alpha$ rechazamos.
- C. Depende del contraste que planteemos.
- D. Si $p\text{ valor} > \alpha/2 \rightarrow \text{Aceptamos } H_0$ mientras que si $p\text{ valor} \leq \alpha/2$ rechazamos.

9. ¿Cómo decidimos un contraste de hipótesis a través de un IC?

- A. Si μ_0 pertenece al IC construido para μ se acepta, de lo contrario rechazamos H_0 .
- B. Para empezar ha de ser bilateral y luego se procede como en A.
- C. Si μ_0 no pertenece al IC construido para μ se acepta, de lo contrario se rechaza H_0 .
- D. Para empezar ha de ser unilateral y luego se procede como en C.

10. Cuando en un contraste de hipótesis desconocemos la varianza de la población de la que provienen los datos empleamos

- A. La distribución T-Student como estadístico de contraste.
- B. La distribución normal si tenemos suficientes observaciones.
- C. La distribución Ji Cuadrada de Pearson.
- D. Si comparamos medias empleamos la Normal si son proporciones la T-Student.

Análisis e Interpretación de Datos

Tema 9. Regresión

Índice

[Esquema](#)

[Ideas clave](#)

[9.1. ¿Cómo estudiar este tema?](#)

[9.2. El modelo de regresión simple](#)

[9.3. Contrastando la regresión](#)

[9.4. Contrastando la regresión con el programa](#)

[9.5. La regresión como suma de cuadrados](#)

[9.6. Aplicación de las TIC](#)

[9.7 Referencias bibliográficas](#)

[A fondo](#)

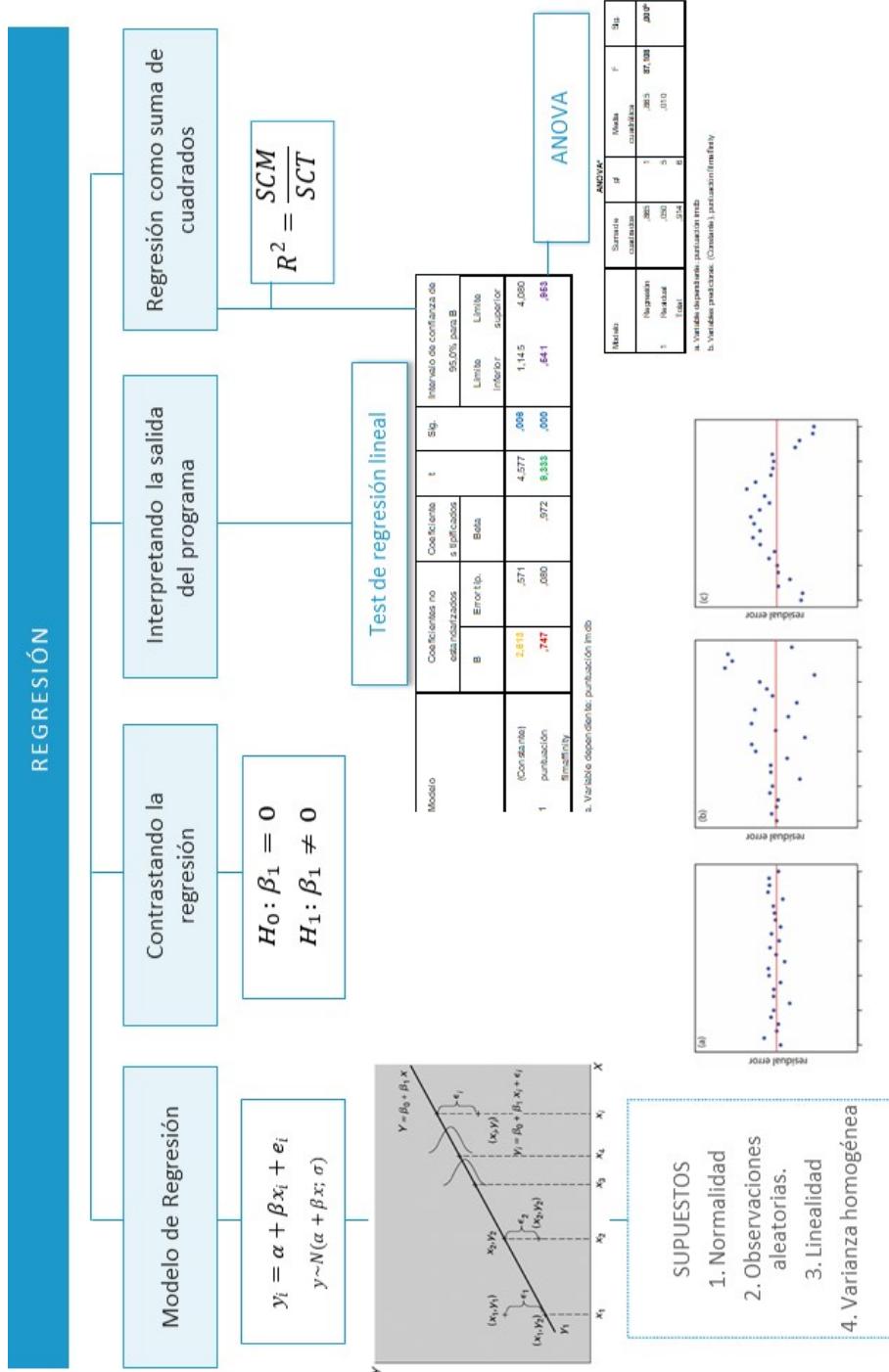
[Realizando un análisis de regresión completo](#)

[Dando el salto al caso de la regresión múltiple](#)

[Vídeos sobre regresión](#)

[Bibliografía](#)

[Test](#)



9.1. ¿Cómo estudiar este tema?

Para estudiar este tema lee las **Ideas clave**, además del intervalo que se indica a continuación:

Páginas 431-496 del libro: Newbold, P., Carlson, W. & Thorne, B. (2008). *Estadística para administración y Economía* (6^a Edición). Madrid. Pearson Educación.

Para hacerte una idea global es importante que mires el esquema del tema, el cual te ayudará a hacerte una buena idea de cómo está estructurado y las relaciones que puedan existir entre algunos conceptos clave.

También **será clave que practiques con las actividades que vienen al final del tema**. Del mismo modo presta atención a los ejemplos que acompañan a los diferentes apartados a lo largo del tema, pues encierran muchas de las claves que te facilitarán la comprensión del tema.

9.2. El modelo de regresión simple

Cuando vimos la parte descriptiva ya vimos una introducción de la Regresión. Sin embargo, era incompleta en el sentido de que se limitaba a un enfoque según lo descriptivo donde no se incorporaban todavía elementos inferenciales, los cuales se veían algo más adelante.

Ahora ya estamos en condiciones de revisitar la regresión apoyándonos en las herramientas inferenciales que ya conocemos, como puedan ser los contrastes de hipótesis. Lo que llamábamos valor estimado antes ahora pasa a ser la y poblacional que pretendemos estimar como un valor esperado de y para un valor de la x concreto.

Matemáticamente:

$$E(y/x) = \alpha + \beta x$$

Se puede observar que ahora empleamos la notación con las letras griegas **alfa y beta**, siendo la primera la **constante del modelo** y la segunda el **coeficiente de regresión o pendiente**. En ocasiones podemos encontrar otras notaciones donde en lugar de alfa se escribe β_0 y la pendiente es β_1 .

En realidad esta recta de regresión no sería más que una recta de medias «ideales» hacia las que tiende la variable y que se distribuye como una normal.

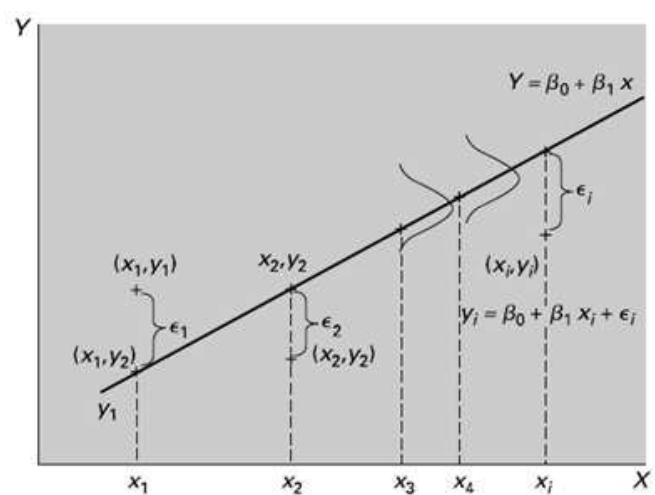
$$y \sim N(\alpha + \beta x; \sigma)$$

Esto lo interpretamos como que la variable y se desviaría σ ; es decir, que tendría una varianza σ^2 . De hecho, los errores son vistos como desviaciones de la recta con esta misma varianza y media cero, pues serían diferencia de normales con idénticas medias y varianzas, motivo por el cual se anularía la media de estos.

$$e_i = y_i - \hat{y}_i$$

Recordemos que el «gorrito» sobre la y se emplea para aludir al valor y estimado, mientras que la y sin el gorrito sería el valor real. Claro está que los errores solo podrán ser conocidos cuando se conozca el valor real de y , lo que permitiría calcularlos de acuerdo a la fórmula anterior. Los errores pueden ser contemplados como las desviaciones de la recta de regresión al valor real. Tenemos el siguiente modelo:

$$y_i = \alpha + \beta x_i + e_i$$



Se observa en el gráfico anterior como sobre la recta se sitúan las curvas normales que seguirán las distribuciones. También podemos apreciar que los errores no son más que desviaciones a la recta.

A parte de la Normalidad necesaria, el modelo presenta tres requisitos más:

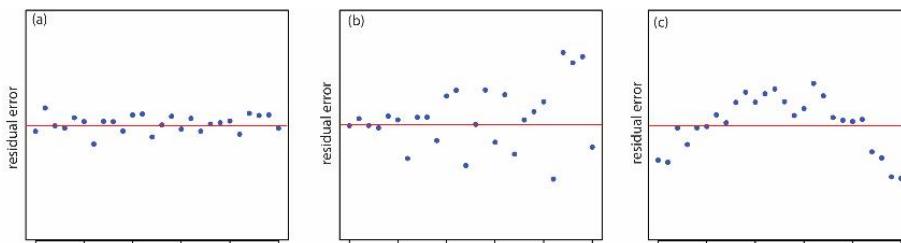
Las observaciones han de ser aleatorias (aleatoriedad de la muestra). Por tanto, no pueden mostrarse patrones al visualizar los errores. Unos subirán por encima de 0, otros estarán por encima pero sin pautas claras.

La relación es lineal.

La varianza σ^2 es homogénea a lo largo del rango de x.

Generalmente el punto 1 puede ser controlado por el estadístico al recoger los datos, mientras que las otras dos dependen de la naturaleza de las variables que se estudian y, por tanto, no dependen del que hace el estudio.

Visualmente, a través de los patrones que muestren los errores se pueden detectar las faltas al modelo de los tres puntos anteriores.



En el gráfico a) tenemos el caso deseable. Los errores están unos por encima otros por debajo y no varía su magnitud a lo largo del rango de la variable. En el caso b) tenemos que la varianza no sería constante (forma de embudo) lo que se llama «heterocedasticidad» (lo contrario de la «homocedasticidad»). Por último, en c) tenemos un claro caso de falta de linealidad.

De este modo vemos cómo podemos apreciar la aleatoriedad, la linealidad y la constancia de la varianza del modelo a través de la representación gráfica de los residuos. A su vez, existen diferentes maneras de representar los residuos, en el eje vertical siempre se colocan estos, pero en el eje horizontal podemos elegir entre colocar el orden de las observaciones, lo cual para detectar falta de aleatoriedad en la muestra es adecuado, sobre todo cuando el orden de las observaciones puede haber afectado a las medidas y por tanto a su independencia. También se suelen colocar en el eje de abscisas las y_i .

9.3. Contrastando la regresión

Para decidir si la regresión es significativa lo que haremos será contrastar sobre la pendiente. Si esta es nula, no hay relación lineal.

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Si los datos muestran con evidencia estadística suficientemente fuerte que el coeficiente debería ser distinto de 0, entonces rechazaríamos H_0 apoyando la regresión lineal. Como de costumbre (ya venimos haciéndolo desde que vimos los contrastes de hipótesis) apoyaremos el contraste en un estadístico que nos ayude a decidirnos.

Todo se reduce, por tanto, a una inferencia sobre la pendiente de la recta de regresión. Para saber cuál es el estadístico basta con darse cuenta (lo daremos por demostrado por no meternos de nuevo en las «tripas» matemáticas de la teoría) que el propio coeficiente beta se distribuye normalmente (b sería la estimación en la muestra del coeficiente poblacional β).

$$b \sim N(\beta; \frac{\sigma}{\sqrt{n} * s_x})$$

Esto nos acaba llevando a lo que nos interesa, que es:

$$t_{exp} = \frac{|b - \beta|}{\sqrt{n} * s_x}$$

Y como según el contraste si la hipótesis nula es cierta β sería igual a 0 tenemos que el estadístico será:

$$t_{exp} = \frac{|b|}{\sqrt{n} * s_x}$$

Que se distribuirá como una t con $n - 2$ grados de libertad. De todos modos, en la práctica lo resolveremos a través de la salida del ordenador, aspecto que trataremos en el siguiente apartado a través de un ejemplo.

9.4. Contrastando la regresión con el programa

Ejemplo

Partamos de los datos de uno de los ejercicios vistos en el tema de regresión:

Las dos bases de datos sobre películas más importantes que hay en la Web son IMDB y Filmaffinity. Las puntuaciones de las nominaciones a la mejor película en los últimos Oscar en ambas bases de datos son:

	Filmaffinity	IMBD
Bidrman	7.2	7.9
Boyhood	7.4	8.1
El gran Hotel Budapest	7.2	8.1
El Francotirador	6.3	7.4
La Teoría del Todo	7.1	7.8
Whiplash	7.9	8.6
Selma	6.7	7.6

El interés está en estudiar las puntuaciones de IMDB a partir de las puntuaciones en Filmaffinity. Para ello se introducen los datos en el programa SPSS (o el PSPP) y se obtiene como salida del test de regresión la siguiente:

Modelo	Coeficientes ^a							
	Coeficientes no estandarizados		Coeficientes tipificados		t	Sig.	Intervalo de confianza de 95,0% para B	
	B	Error típ.	Beta	Límite inferior	Límite superior			
1 (Constante) puntuación filmaffinity	2,613 ,747	,571 ,080		,972	4,577 9,333	,006 ,000	1,145 ,541	4,080 ,953

a. Variable dependiente: puntuación imbd

El interés a la hora de interpretarla está especialmente en las cifras resaltadas a color. En naranja y rojo, respectivamente, están las estimaciones de los coeficientes. Podemos ver que la estimación de la constante del modelo vale 2,613 y que la del coeficiente beta resulta 0,747. Recordemos que esta última nos indica la dirección de la relación lineal y que es la pendiente. Al ser positiva indica que a valores positivos altos de Filmaffinity le corresponden valores positivos y altos también en IMDB.

En verde tenemos el valor del estadístico t , que si bien no es tan importante nos ayuda también a confirmar que el contraste se rechaza pues es un valor muy elevado para una t que, recordemos, es una distribución que se aproxima a la normal a medida que crece n . En la columna de «sig.», en azul, tenemos los *p-valores* que nos indican que ambos coeficientes (el de la constante y el de regresión) son altamente significativos ($< 0,01$).

Especialmente importante es el 0,000 en azul pues es precisamente el *p – valor* asociado al contraste de la regresión que acabamos de ver en el apartado anterior. Nos está indicando que la regresión es significativa. Por último, es interesante echar un vistazo al IC del coeficiente de regresión beta en morado, aquí apreciamos que no contiene el 0 en el intervalo, pues de lo contrario no sería significativa la regresión

9.5. La regresión como suma de cuadrados

Es útil conocer otro enfoque para contrastar la regresión. Para ello debemos saber cómo descomponer en sumas de cuadrados la regresión. Veremos, además, como este enfoque está íntimamente relacionado con el coeficiente de correlación R visto en el tema de regresión de la parte descriptiva.

La desviación total del modelo de partida se considera como la **SCT o Suma de Cuadrados Totales** y no es más que la suma de las desviaciones a la media de todos los datos.

$$SCT = \sum (y_i - \bar{y})^2$$

El modelo que construimos debe ser capaz de contener parte de estas desviaciones. Si las contuviera o explicara todas sería un modelo perfecto, pero lo normal es que contenga una parte de estas desviaciones y, en función de cuanto de grande sea esta parte respecto la SCT, diremos que la SCM o Suma de Cuadrados del Modelo explica bien el modelo.

$$SCM = \sum (\hat{y}_i - \bar{y})^2$$

Aquí ya podemos rescatar lo dicho anteriormente sobre la relación entre esta descomposición y el coeficiente de determinación (recordemos que es el cuadrado del de correlación):

$$R^2 = \frac{SCM}{SCT}$$

La parte no explicada del modelo la forman los residuos a través de la SCR o Suma de Cuadrados de los Residuos:

$$SCR = \sum (\hat{y}_i - y_i)^2$$

La suma de cuadrados total de las desviaciones se puede por tanto descomponer en la suma de los cuadrados del modelo más la suma de los errores al cuadrado.

$$SCT = SCM + SCR$$

La salida de un contraste de regresión por el programa estadístico también puede ofrecer la salida de la tabla que descompone estas sumas de cuadrados y contrasta sus ratios a través de la distribución F , que es el cuadrado de la t vista anteriormente.

De este modo, se observa que en el fondo se pueden considerar equivalentes ambas formas de contrastar la regresión, o bien a través de la t o bien de la F .

La tabla que se muestra a continuación también es conocida como **tabla ANOVA** (*Analysis of Variance*) y es de gran importancia práctica en la estadística inferencial.

ANOVA ^a					
	Modelo	Suma de cuadrados	gl	Media cuadrática	F
1	Regresión	,865	1	,865	87,108
	Residual	,050	5	,010	
	Total	,914	6		

a. Variable dependiente: puntuación imdb

b. Variables predictoras: (Constante), puntuación filmaffinity

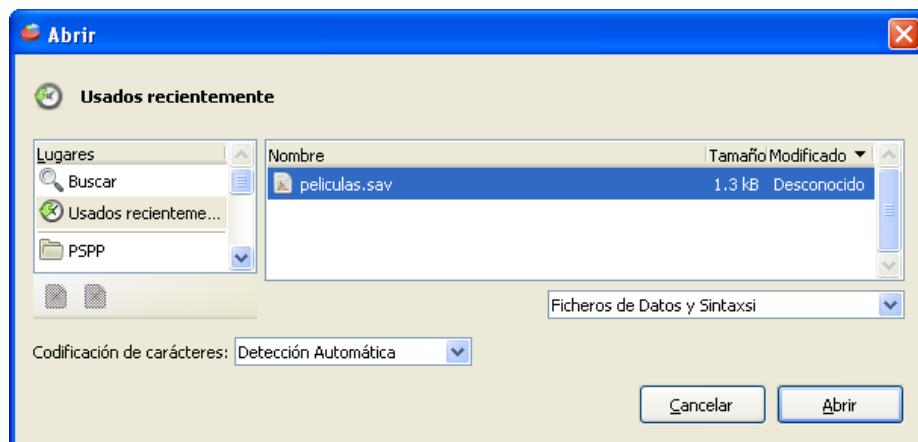
En negrita se resalta, por un lado, el resultado del test de la distribución F , el cual como dije anteriormente es equivalente al de la t al cuadrado ($87,108 = 9,332$). Por el otro lado, figura en negrita el p -valor asociado a la SCR o Suma de Cuadrados de la Regresión, que resulta significativo indicando que el modelo de regresión es significativo.

Podemos comprobar también que un buen modelo debe tener un F – *ratio* elevado, lo cual podremos saber más allá del programa si cotejamos con la tabla correspondiente de esta distribución del mismo modo que se ha venido haciendo con la Normal, la T y la Chi-Cuadrado.

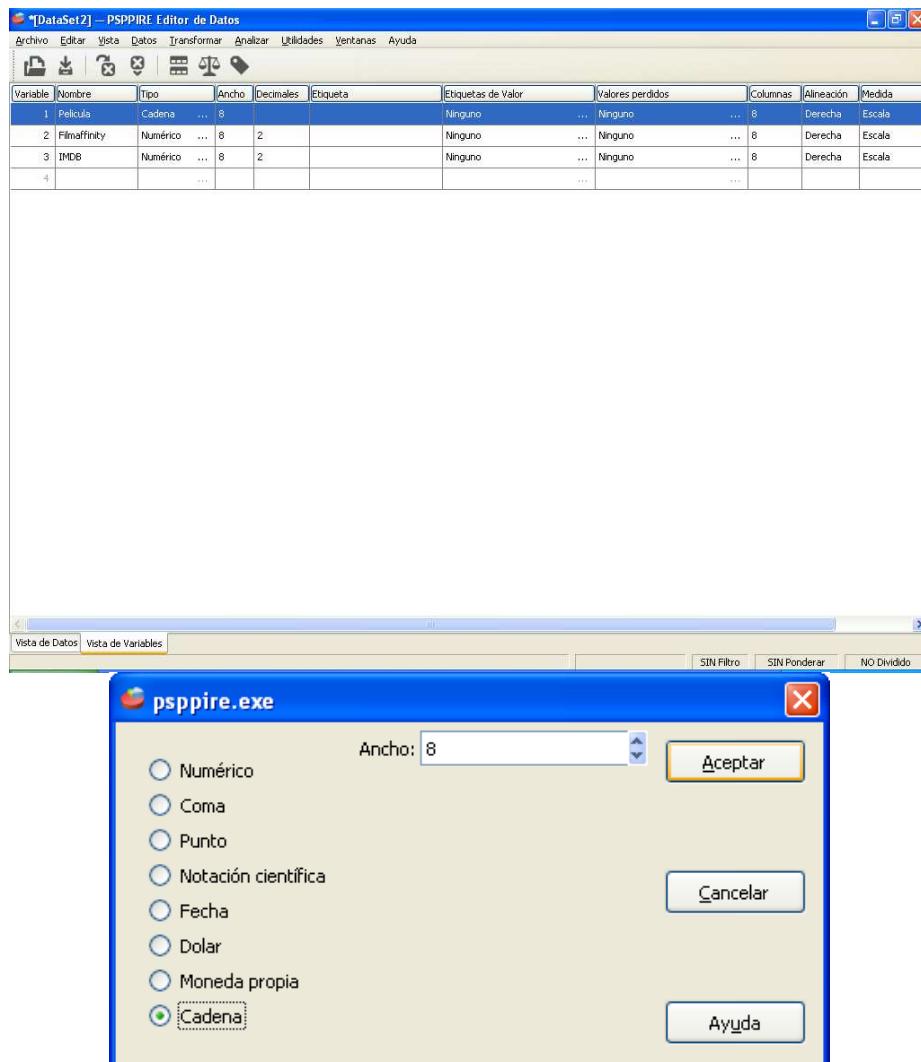
9.6. Aplicación de las TIC

Vamos a estudiar cómo llevar a cabo la regresión anterior con el PSPP (la versión libre que imita al SPSS).

- ▶ Una vez abierto el PSPP (instalarlo a través de <http://pspp.awardspace.com/>), abrimos el documento películas.sav (es recomendable tratar de introducir los datos en el programa para practicar).



Si decidimos introducir los datos desde el principio (opción recomendada) comenzamos definiendo las variables en «vista de variables»:



De este modo tenemos una variable cualitativa o categórica: «películas» y otras dos de escala, que son las puntuaciones. Tendremos que ajustar las características de las variables para poder introducir los datos que necesitan alojar. Por ejemplo, tendremos que aumentar el ancho de la variable película para que quepan los títulos, etc.

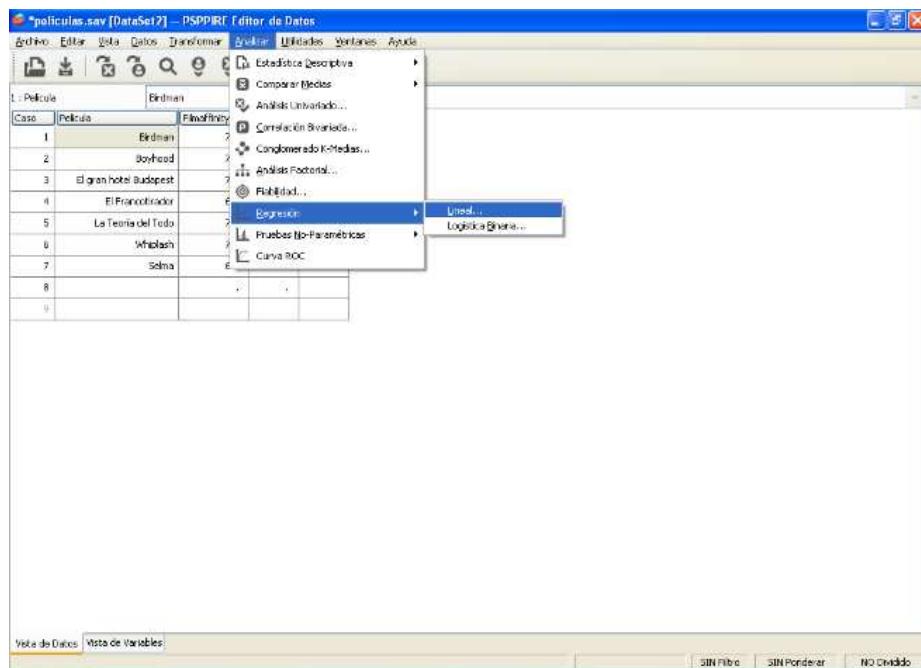
Cuando se acabe de introducir los datos deberíamos tener una hoja de datos como la siguiente:

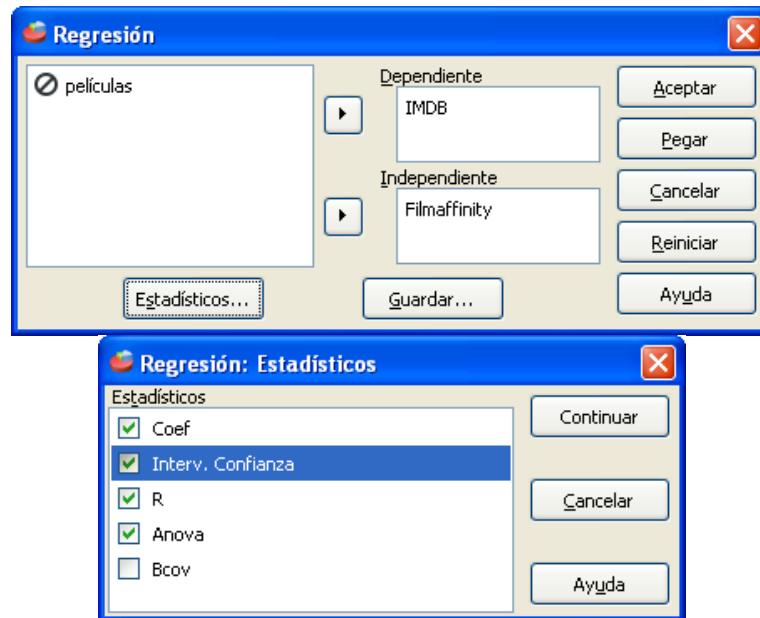
The screenshot shows the SPSS Data Editor window. The title bar reads "*peliculas.sav [DataSet2] – SPSS® Editor de Datos". The menu bar includes Archivo, Editor, Vista, Datos, Transformar, Analizar, Unidades, Ventanas, Ayuda. The toolbar has icons for file operations like Open, Save, and Print. The main area displays a table titled "L: Película" with the following data:

Caso	Película	FilmAffinity	IMBD
1	Birdman	7.20	7.30
2	Boyhood	7.40	6.10
3	El gran hotel Budapest	7.20	6.10
4	El Francotirador	6.30	7.10
5	La Teoría del Todo	7.10	7.30
6	Whiplash	7.50	8.50
7	Selma	6.70	7.50
8		-	-
9		-	-

At the bottom, there are tabs for "Vista de Datos" and "Vista de Variables", and buttons for SIN Filtro, SIN Ponderar, and NO Duplicado.

Para realizar el análisis de regresión de las puntuaciones de IMBD en función de las de Filmaffinity nos dirigimos a las opciones siguientes:





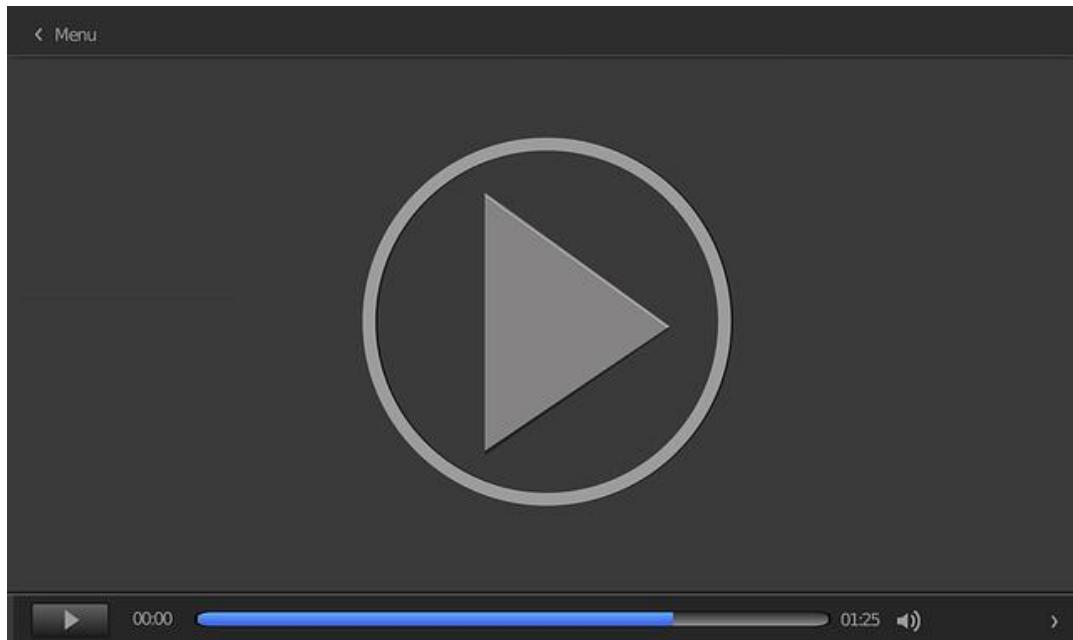
Ya podemos darle a aceptar y obtenemos el resultado de nuestro análisis de regresión:

Objetivo – PSPPIRE Visor de resultados																									
	Archivo Editar Ventanas Ayuda																								
REGRESSION																									
REGRESSION /VARIABLES= Filmaffinity /DEPENDENT= IMDB /STATISTICS=COEFF CI R ANOVA.																									
Resumen del modelo (puntuación en IMDB)																									
<table border="1"> <thead> <tr> <th>R</th> <th>R Cuadrada</th> <th>R Cuadrada Ajustada</th> <th>Error estándard del Estimador</th> </tr> </thead> <tbody> <tr> <td>.97</td> <td>.95</td> <td>.93</td> <td>.10</td> </tr> </tbody> </table>		R	R Cuadrada	R Cuadrada Ajustada	Error estándard del Estimador	.97	.95	.93	.10																
R	R Cuadrada	R Cuadrada Ajustada	Error estándard del Estimador																						
.97	.95	.93	.10																						
ANOVA (puntuación en IMDB)																									
<table border="1"> <thead> <tr> <th></th> <th>Suma de Cuadrados</th> <th>df</th> <th>Cuadrado medio</th> <th>F</th> <th>Sign.</th> </tr> </thead> <tbody> <tr> <td>Regresión</td> <td>.86</td> <td>1</td> <td>.86</td> <td>87.11</td> <td>.000</td> </tr> <tr> <td>Residual</td> <td>.05</td> <td>5</td> <td>.01</td> <td></td> <td></td> </tr> <tr> <td>Total</td> <td>.91</td> <td>6</td> <td></td> <td></td> <td></td> </tr> </tbody> </table>			Suma de Cuadrados	df	Cuadrado medio	F	Sign.	Regresión	.86	1	.86	87.11	.000	Residual	.05	5	.01			Total	.91	6			
	Suma de Cuadrados	df	Cuadrado medio	F	Sign.																				
Regresión	.86	1	.86	87.11	.000																				
Residual	.05	5	.01																						
Total	.91	6																							
Coeficientes (puntuación en IMDB)																									
<table border="1"> <thead> <tr> <th rowspan="2"></th> <th colspan="2">Coeficientes No Estandarizados</th> <th colspan="2">Coeficientes Estandarizados</th> <th rowspan="2">t</th> <th rowspan="2">Sign.</th> <th colspan="2">Intervalo de Confianza 95% para B</th> </tr> <tr> <th>B</th> <th>Error Estándar</th> <th>Beta</th> <th></th> <th>Límite Inferior</th> <th>Límite Superior</th> </tr> </thead> <tbody> <tr> <td>(Constant) puntuación en Filmaffinity</td> <td>2.61 .75</td> <td>.57 .08</td> <td></td> <td>.00 .97</td> <td>4.58 9.33</td> <td>.004 .000</td> <td>1.15 .54</td> <td>4.08 .95</td> </tr> </tbody> </table>			Coeficientes No Estandarizados		Coeficientes Estandarizados		t	Sign.	Intervalo de Confianza 95% para B		B	Error Estándar	Beta		Límite Inferior	Límite Superior	(Constant) puntuación en Filmaffinity	2.61 .75	.57 .08		.00 .97	4.58 9.33	.004 .000	1.15 .54	4.08 .95
	Coeficientes No Estandarizados		Coeficientes Estandarizados		t	Sign.			Intervalo de Confianza 95% para B																
	B	Error Estándar	Beta				Límite Inferior	Límite Superior																	
(Constant) puntuación en Filmaffinity	2.61 .75	.57 .08		.00 .97	4.58 9.33	.004 .000	1.15 .54	4.08 .95																	

Lo que figura antes de las tablas es la sintaxis que genera el programa. En realidad siempre es útil aprender a manejar la sintaxis con los programas estadísticos (da igual que el SPSS que el PSPP, que el R, que el SAS, etc.), pues existen ciertas opciones que solo serán accesibles desde el código fuente, especialmente a medida que se van haciendo complejos nuestros análisis. Otro motivo importante es el ahorro de tiempo que puede llegar a suponer.

Regresión lineal múltiple

En este vídeo vamos a conocer uno de los métodos más potentes que existen en estadística inferencial: la regresión lineal múltiple.



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=4b0064d9-9023-4c92-a20c-acbd00e37b71>

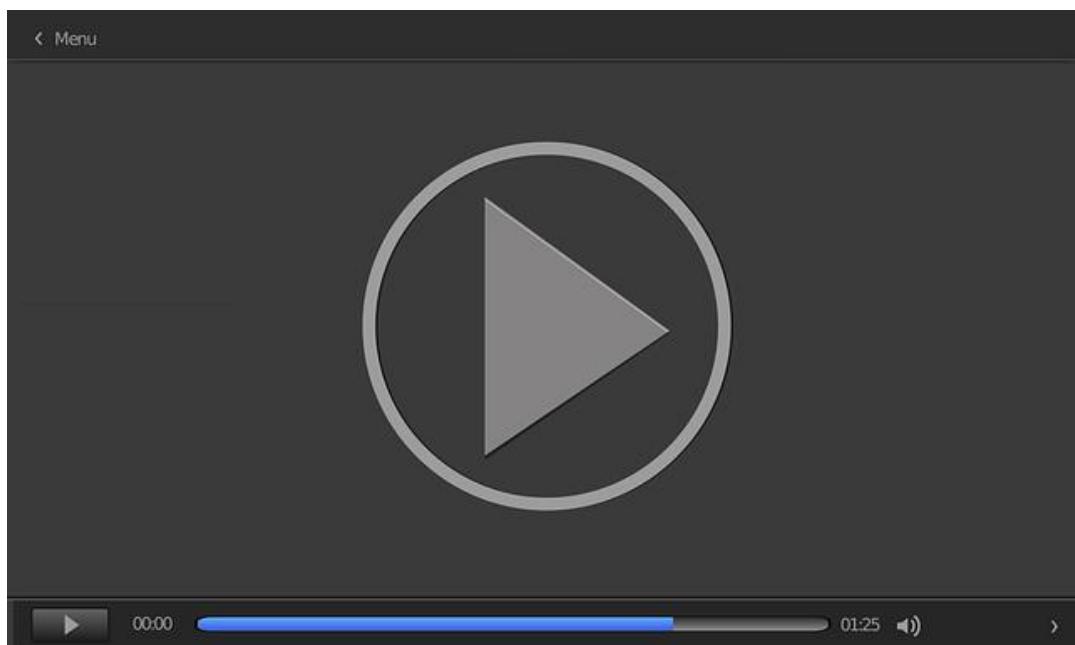
9.7 Referencias bibliográficas

Newbold, P. et al. (2008). *Estadística para administración y Economía* (6^a Edición). Madrid. Pearson Educación,

Martín, A. (2004). *Bioestadística para las ciencias de la salud* (1^a ed. (5^a)). Madrid: Norma-Capitel.

Realizando un análisis de regresión completo

En este vídeo voy a mostráros como realizar un análisis de regresión completo, teniendo en cuenta los supuestos teóricos que subyacen al modelo y tratando de dar una visión intuitiva y práctica del uso del programa estadístico.



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=beabd2df-9b3e-4c3f-a464-abdc00fae079>

Video. Análisis de regresión completo

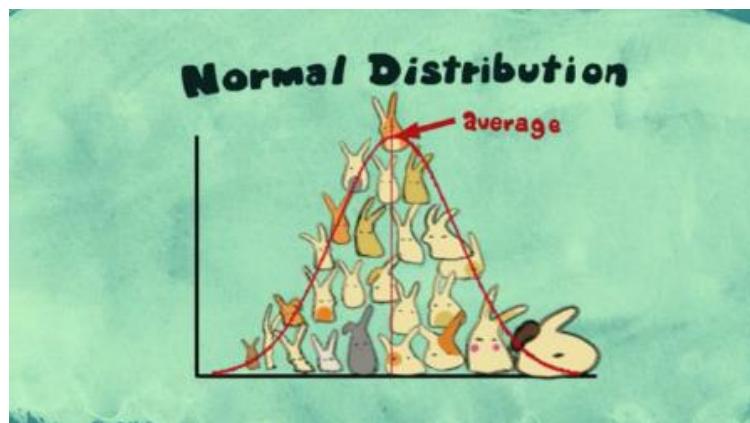
Dando el salto al caso de la regresión múltiple

Conceptualmente idéntica pero más compleja surge la regresión múltiple donde la diferencia es que contaremos con más de una variable predictora. Te recomiendo echarle un vistazo al capítulo 8 (pág. 354) del libro digital libre OpenIntroStatistics.

Accede al artículo desde el aula virtual o a través de la siguiente dirección web:

https://www.openintro.org/stat/textbook.php?stat_book=os

Vídeos sobre regresión



Dentro del portal de OpenIntro encontramos cuatro presentaciones animadas de diferentes aspectos relacionados con el tema. Te recomiendo especialmente el último de ellos.

Accede al vídeo a través del aula virtual o desde la siguiente dirección web: <https://www.openintro.org/stat/videos.php>

Bibliografía

Diez, D., Barr, C.; Çetinkaya-Rundel, M. (2015). *OpenIntro Statistics* (3^a Edición).

https://www.openintro.org/stat/textbook.php?stat_book=os

Newbold, P. et al. (2008). *Estadística para administración y Economía* (6^a Edición).

Madrid. Pearson Educación.

Martín, A. (2004). *Bioestadística para las ciencias de la salud* (1^a ed. (5^a)). Madrid:

Norma-Capitel.

- 1.** El modelo de regresión lineal se supone...
 - A. Aleatorio, curvilíneo y apolíneo.
 - B. Carente de errores en la medida de lo posible.
 - C. No heterocedástico.
 - D. Con una Tde student no muy elevada (de lo contrario se rechaza).

- 2.** ¿Qué es lo que se contrasta en un modelo de regresión lineal?
 - A. Si hay constante.
 - B. Si la pendiente es nula o no.
 - C. Si el error es mínimo respecto a la recta.
 - D. B y C son correctas.

- 3.** ¿Con que Test o pruebas podemos contrastar el modelo de regresión?
 - A. Con el F y la T de Student.
 - B. Con la T-de Student y la Normalidad de la variable.
 - C. Con el ANOVA.
 - D. A y C son correctas.

- 4.** R^2 equivale a:
 - A. SCM entre la SCR.
 - B. La SCM entre la SCT.
 - C. B y D son correctas.
 - D. 1 menos la SCR/SCT.

5. Si el IC para β_1 contiene al 0 esto querrá decir que: D B y C son correctas.
 - A. La regresión no es significativa.
 - B. La regresión es altamente significativa.
 - C. Con una probabilidad del 95% la regresión no será significativa.
 - D. B y C son correctas.
6. ¿Qué es alfa en el modelo de regresión?
 - A. El complementario de Beta.
 - B. La potencia del contraste.
 - C. La constante.
 - D. A y C son correctas.
7. En el ejemplo visto en el capítulo sobre las películas:
 - A. Las puntuaciones de IMDB eran la variable dependiente.
 - B. Las puntuaciones de Filmaffinity eran la variable predictoria.
 - C. Las películas eran las variables dependientes.
 - D. A y B son correctas.
8. Si al graficar los errores estos presentar una forma de embudo:
 - A. Es buena señal para el modelo.
 - B. Indicaría una falta de homocedasticidad.
 - C. Indicaría una falta de heterocedasticidad.
 - D. Ninguna de las anteriores.

- 9.** Los errores del modelo conviene que B sean casi todos elevados.
- A. Sean casi todos positivos.
 - B. Sean casi todos elevados.
 - C. Que se alternen lo mejor posible los positivos y los negativos sin grandes saltos ni patrones.
 - D. Que presenten claros patrones.
- 10.** El modelo de regresión lineal consta de tres parámetros.
- A. Alfa, gamma y el error.
 - B. Beta cero, beta uno y alfa.
 - C. Alfa, beta y sigma.
 - D. No es cierto, consta de dos.

Análisis e Interpretación de Datos

Tema 10. Análisis de componentes principales

Índice

[Esquema](#)

[Ideas clave](#)

[10.1. ¿Cómo estudiar este tema?](#)

[10.2. Motivación](#)

[10.3. Definición](#)

[10.4. Aplicaciones](#)

[10.5. Ejemplo de aplicación en R](#)

[10.6 Referencias bibliográficas](#)

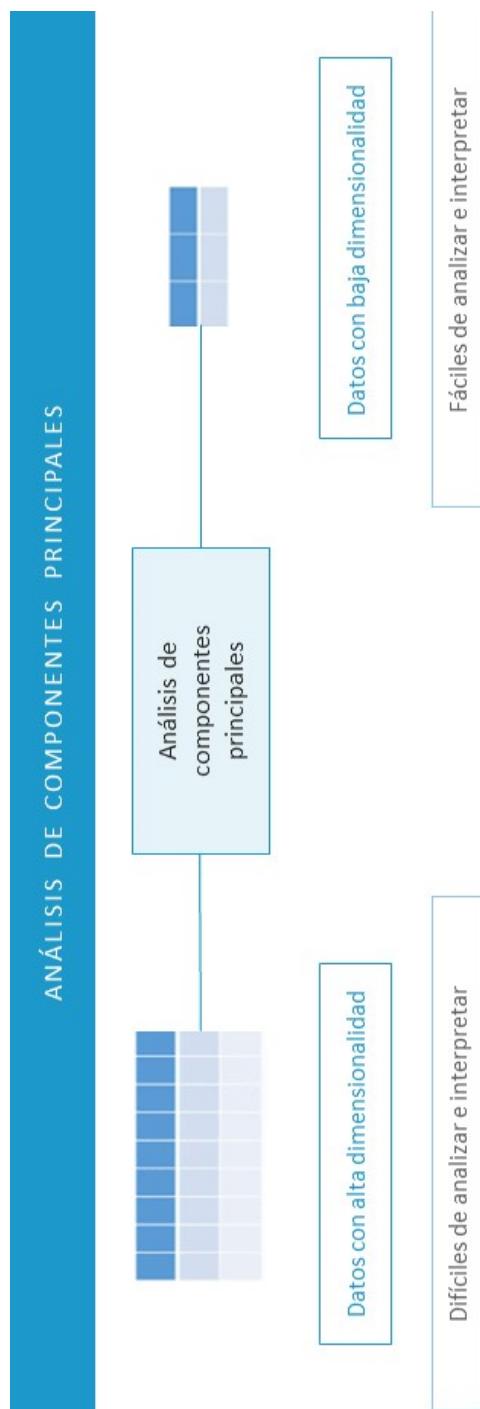
[A fondo](#)

[Análisis de componentes principales. Caso práctico en R](#)

[Explicación intuitiva del análisis de componentes principales](#)

[Aplicación del análisis de componentes principales en el diagnóstico socioambiental](#)

[Test](#)



10.1. ¿Cómo estudiar este tema?

Para estudiar este tema lee las **Ideas clave** que encontrarás a continuación.

Para hacerte una idea global es importante que mires el esquema del tema, el cual te ayudará a entender cómo está estructurado y las relaciones que puedan existir entre algunos conceptos clave.

Para estudiar este tema, repasa el contenido y **trata de reproducir en casa el ejemplo de R que se propone**. A lo largo del tema se plantea uno de los problemas típicos con los que solemos tener que tratar cuando trabajamos en entornos Big Data. Y dicho problema se soluciona mediante el uso de **técnicas de análisis de componentes principales**. Por último, veremos, de forma práctica sobre R, cómo aplicar las técnicas de análisis de componentes principales sobre un conjunto de datos concreto.

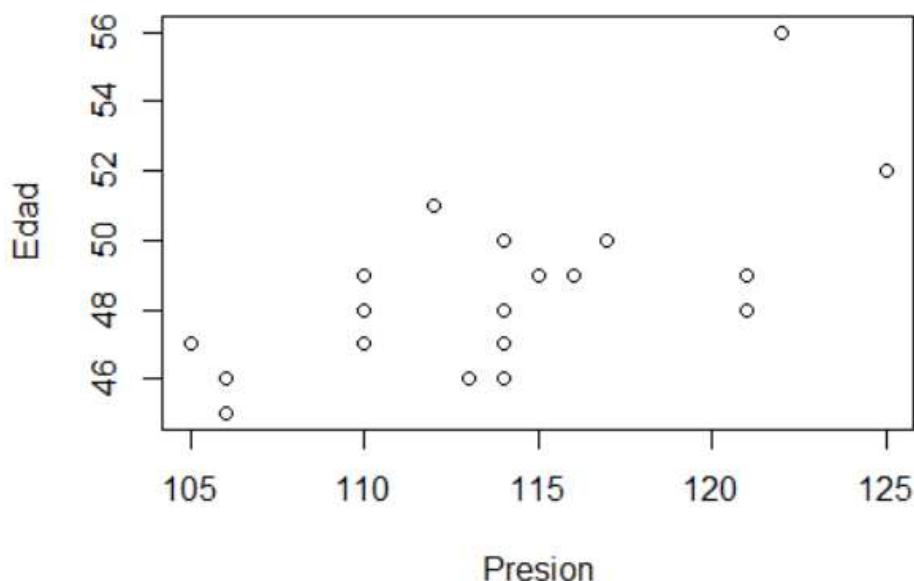
10.2. Motivación

Imaginemos que tenemos el siguiente conjunto de datos. Tenemos información acerca de 20 personas y de los factores influyentes en el padecimiento de una enfermedad coronaria.

	Presión arterial	Edad	Peso	Superficie corporal	Duración de la hipertensión	Pulso	Medida de estrés
I1	102	42	85,4	1,8	5,1	63	30
I2	117	51	94,2	2,1	3,8	70	14
I3	116	49	95,3	1,98	8,2	72	14
I4	117	50	94,7	2,01	5,8	73	97
I5	112	51	89,4	1,86	7	72	95
I6	120	38	99,5	2,25	9,3	71	10
I7	121	29	99,8	2,1	2,5	69	42
I8	110	47	90,9	1,9	6,2	66	8
I9	111	49	89,2	1,70	7,1	69	60
I10	117	48	92,7	2,09	5,6	64	35
I11	112	47	94,4	2,07	5,3	74	90
I12	115	49	94,1	1,92	5,6	71	21
I13	114	50	91,6	2,05	10,2	68	47
I14	110	47	87,1	1,92	5,6	67	80
I15	125	52	101,3	2,19	10	76	98
I16	112	46	94,5	1,98	7,4	69	98
I17	106	46	87	1,87	3,6	62	13
I18	109	46	94,5	1,9	4,3	70	15
I19	112	48	90,5	1,88	9	71	99
I20	120	56	95,7	2,09	7	75	94

Tabla 1. Conjunto de datos.

La dimensión del conjunto, esto es, el número de datos para cada individuo es de 7. ¿Qué opciones tenemos para realizar la representación gráfica de los mismos? Una primera aproximación consistiría en comparar 2 a 2 cada componente usando una **gráfica de dispersión**. Por ejemplo, comparando presión arterial y edad obtendríamos la siguiente gráfica:



Sin embargo, comparando solo estas dos características estamos lejos de dar un modelo representativo de todos los datos. Si hacemos la misma gráfica, pero utilizando 3 dimensiones esta vez (presión, peso y edad) obtenemos el siguiente resultado:

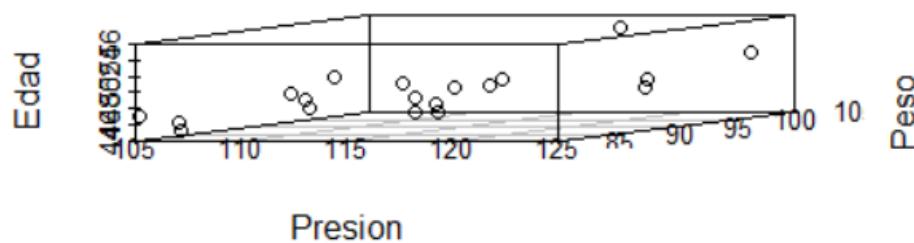


Figura 2: Gráfico de edad, presión y peso. (Generado por script en R mío).

Como podemos observar, este gráfico es **mucho más difícil de analizar** a simple vista que el anterior puesto que, a no ser que creemos un modelo 3D con el que el usuario pueda interactuar, es **imposible determinar a simple vista la distancia entre los puntos**. Aumentando las dimensiones del gráfico solo conseguimos que los modelos asociados sean aún más difíciles de entender.

Es bastante común, debido a la naturaleza y cantidad de datos, que cuando trabajamos en entornos Big Data, los conjuntos de datos tengan una alta dimensionalidad. Trabajar con datos que poseen estas características tiene los **siguientes inconvenientes**:

- ▶ La alta dimensionalidad hace que sea muy **complicado llevar a cabo representaciones coherentes de los datos**. Por tanto, sería de mucha utilidad aplicar métodos que nos permitan, de forma fácil e intuitiva, transformar los datos de forma que puedan representarse de forma fácil en una gráfica.
- ▶ Trabajar con datos que poseen una alta dimensionalidad hace que tengamos que tratar con cantidades ingentes de información. El número total de datos con los que trabajaríamos nosotros o cualquier sistema computacional se puede calcular como el **número de individuos por la dimensión o número de características de cada individuo**. De esta forma, cuanto mayor sea la dimensión, mayor será el número de datos totales con los que tenemos que trabajar. Sería interesante, por tanto, transformar los datos de forma que podamos reducir la dimensión del conjunto de datos sacrificando la menor cantidad de precisión e información posible en los datos.

Una manera simultánea de resolver los dos problemas planteados consiste en aplicar sobre los conjuntos de datos las llamadas **técnicas de análisis de componentes principales**.

10.3. Definición

Las técnicas de análisis de componentes principales nos permiten **reducir la dimensión de los datos con los que estamos tratando, sacrificando la menor cantidad de precisión de la información posible**. De forma más completa, podemos definir el proceso llevado a cabo por el análisis de componentes principales como:

El análisis de componentes principales consiste en encontrar transformaciones ortogonales de las variables originales para conseguir un nuevo conjunto de variables incorreladas, denominadas Componentes Principales, que se obtienen en orden decreciente de importancia.

Es decir, **creamos nuevas variables de dimensión como combinación de las variables ya existentes**. Por ejemplo, si tomamos como ejemplo el conjunto de datos presentado en la sección anterior, una posible solución que la técnica podría darnos sería la siguiente:

$$C_1 = 0,99 \cdot Edad - 0,9 \cdot Presion + 0,1 \cdot Peso$$

De forma que la nueva variable **C1 es una variable formada a partir de los valores de las dimensiones ya existentes**. Nótese que esto es solo un ejemplo de un posible resultado, no el resultado real que se obtiene tras aplicar la técnica y que veremos en la sección de ejemplos de aplicación. De hecho, todas las nuevas variables generadas se calculan siempre a partir de todas las variables del modelo, no solo a partir de algunas de ellas.

Por tanto, a partir de la fórmula anterior, podemos calcular el valor de la dimensión C1 para cada individuo, aplicando la fórmula a cada fila. Como resultado obtendríamos lo siguiente:

	Presión arterial	Edad	Peso	C1
I1	102	42	85,4	-39.43
I2	117	51	94,2	-45.57
I3	116	49	95,3	-46.36
I4	117	50	94,7	-46.33
I5	112	51	89,4	-41.37
I6	120	38	99,5	-51.43
I7	121	29	99,8	-50.41
I8	110	47	90,9	-43.38
I9	111	49	89,2	-41.57
I10	117	48	92,7	-45.81
I11	112	47	94,4	-46.63
I12	115	49	94,1	-45.58
I13	114	50	91,6	-43.94
I14	110	47	87,1	-42.14
I15	125	52	101,3	-50.89
I16	112	46	94,5	-47.61
I17	106	46	87	-41.16
I18	109	46	94,5	-46.71
I19	112	48	90,5	-42.43
I20	120	56	95,7	-44.79

Tabla 2. Resultado.

Pero ¿qué beneficios y características tiene la nueva variable calculada? Los enumeramos a continuación:

- ▶ **Es una variable formada a partir de las variables del modelo:** la nueva variable generada es una **variable en cuya generación han participado todas las dimensiones del modelo.** Por tanto, podemos considerarla como un **valor resumen** que agrega y aglutina varias dimensiones. La agregación realizada es una agregación ponderada donde **no todas las variables tienen el mismo peso en el cálculo del resultado final.** Aquellas variables que representen mejor los datos tendrán un mayor peso.
- ▶ **Es la variable que mejor recoge la variabilidad del modelo:** la variable calculada de la que estamos hablando, también llamada primera componente principal, es aquella que **mejor describe la variabilidad del modelo de datos. Por tanto, es aquella que minimiza la pérdida de información** resumiendo el modelo en una única variable. De esta manera, realizamos una reducción de dimensión de los datos óptima donde la información que se pierde, cuando utilizamos varias componentes principales, es muy reducida.

Los procesos de análisis de componentes principales no calculan una única variable sino **un conjunto de ellas.** Dichas variables están ordenadas en función de la variabilidad que aportan al modelo. Cuanto mayor sea la variabilidad, mejor representan estas variables al modelo de datos. Por tanto, cogiendo varias de ellas y maximizando la variabilidad, podemos obtener un **número reducido de dimensiones que nos permitan una correcta representación de la información.** Es importante entender que, a mayor número de variables generadas incluidas en el modelo de dimensionalidad reducida, mayor dimensionalidad, pero a la vez, **mayor fiabilidad a la hora de expresar la información.** Recordemos que cuanto mayor sea la dimensionalidad, más datos tendrá el modelo y más difícil será de entender y graficar.

Podemos definir el vector de componentes como el **vector de números que debemos multiplicar por los valores de las dimensiones de cada individuo**. En el ejemplo de C1, el vector asociado sería el siguiente:

$$C1 = \{0, 99 \ - 0, 9 \ 0, 1\}$$

El vector C1 nos ayuda a estudiar qué **influencia tienen las diferentes dimensiones implicadas sobre el valor agregado representado por la variable**. En el ejemplo centrado sobre presión, edad y peso con el que estamos trabajando, puede observarse lo siguiente:

- ▶ La edad tiene una fuerte influencia en el cálculo del valor agregado. Cuanto mayor es la edad, mayor es el valor agregado.
- ▶ El valor de presión tiene una fuerte influencia inversa en el cálculo de C1. A más presión menor valor agregado.
- ▶ El peso, al ser su valor asociado cercano a 0, tiene poca influencia sobre el valor agregado final.

Toda esta información, por supuesto, está asociada a una única componente. Al aplicar el análisis de componentes principales **calcularemos y trabajaremos con varias componentes a la vez donde, para cada una de ellas, el grado en que cada variable afecta a la componente variará**. El conjunto de nuevas componentes representará de forma fiable el conjunto de datos usando un número de dimensiones menor al original.

Todo esto resulta muy útil, pero **¿qué pasos deben seguirse exactamente para llevar a cabo un proceso de análisis de componentes?** Los detallamos a continuación:

- ▶ **1. Cálculo de las componentes principales:** se realiza el cálculo de las componentes principales. Para ello se hace uso de las matrices de varianzas y covarianzas ya que dichas **matrices explican la variabilidad del modelo**. En caso de estar interesados, en el siguiente enlace se explica, en detalle, el proceso de cálculo de las componentes principales. Dicho proceso no entrará en el examen final de la asignatura.
-

Podemos verlo junto a algunos ejemplos en el siguiente enlace:

https://www.mhe.es/universidad/ciencias_matematicas/pena/home/CAPITULO.PDF

- ▶ **2. Selección del número de componentes a incluir en el nuevo modelo:** en función de lo que necesitemos hacer y de la precisión que estemos dispuestos a sacrificar, debemos escoger el número de componentes que queremos que tenga nuestro nuevo conjunto de datos. **Cuantas más componentes incluyamos, tendremos mayor precisión, pero más información de la que hacernos cargo.**
- ▶ **3. Análisis de resultados y graficación del modelo:** tal y como hemos visto, una vez seleccionadas las componentes que queremos utilizar, podemos analizar **qué variables influyen en dichas componentes y qué gráficos utilizar para representar los datos**. Un número reducido de dimensiones permite representaciones mucho más claras del modelo de datos.

10.4. Aplicaciones

El análisis de componentes principales es una técnica muy utilizada en investigación y análisis de datos que tiene amplias aplicaciones en multitud de ámbitos. En esta sección proponemos algunos ejemplos:

- ▶ **Reducción del volumen de datos:** por supuesto, la aplicación principal del análisis de componentes principales es la de la **reducción del volumen de datos con el que estamos trabajando**. En entornos Big Data, donde la cantidad de información con la que trabajamos es enorme, es **esencial contar con herramientas de este tipo que nos permitan reducir la información lo suficiente como para poder trabajar con ella**.
- ▶ **Reducción del ruido en imágenes:** es normal, en multitud de medios, la obtención de imágenes que, debido a factores atmosféricos o interferencias de cualquier otro tipo, no se ven lo suficientemente claras. Generalmente, ha sido demostrado que **este ruido suele ir asociado a las últimas componentes de un proceso de análisis de componente principales**. Por tanto, la eliminación de dichas componentes, contribuirían a la eliminación y realce de dichas imágenes.

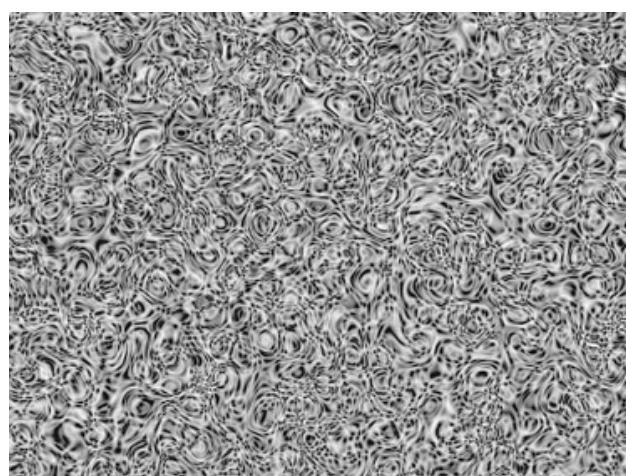


Figura 3. Ruido en imagen. Fuente: https://pixabay.com/p-643623/?no_redirect

- ▶ **Detección de cambios en los datos:** si poseemos datos de una misma población en dos momentos diferentes, es posible aplicar un análisis de componentes principales para estudiar los cambios producidos. De esta manera, **aquellas variables que participen más en los componentes principales serán las que han permanecido inamovibles.** Por el contrario, aquellas **variables que participen en componentes más lejanas serán aquellas que más han variado en el tiempo.**

10.5. Ejemplo de aplicación en R

En esta sección aprenderemos a llevar a cabo un **proceso de análisis de componentes principales usando la herramienta R**. Lo primero será definir nuestro conjunto de datos inicial. Mediante los siguientes comandos podemos introducir el ejemplo con el que hemos tratado en R:

```
Edad = c(47,49,49,50,51,48,49,47,49,48,47,49,50,45,52,46,46,46,48,56)
Presion = c(105,115,116,117,112,121,121,110,110,114,114,115,114,106,125,
114,106,113,110,122)
Peso = c(85.4,94.2,95.3,94.7,89.4,99.5,99.8,90.9,89.2,92.7,94.4,94.1,91.6,
87.1,101.3,94.5,87,94.5,90.5,95.7)
Superficiecorporal = c(1.8,2.1,1.98,2.01,1.86,2.25,2.1,1.9,1.7,2.09,2.07,
1.92,2.05,1.92,2.19,1.98,1.87,1.9,1.88,2.09)
Hipertension = c(5.1,3.8,8.2,5.8,7,9.3,2.5,6.2,7.1,5.6,5.3,5.6,10.2,5.6,
10,7.4,3.6,4.3,9,7)
Pulso =c(63,70,72,73,72,71,69,66,69,64,74,71,68,67,76,69,62,70,71,75)
Estres = c(30,14,14,97,95,10,42,8,60,35,90,21,47,80,98,98,13,15,99,94)
modelodatos = matrix(c(Edad,Presion,Peso,Superficiecorporal,Hipertension,
Pulso,Estres),nrow=20)
```

A continuación, utilizando la función **prcomp** y previa conversión logarítmica de los datos para mayor fiabilidad de los resultados, **calculamos los vectores de componentes**. El código a utilizar es el siguiente:

```
log.modelo <- log(modelodatos)
modelo.pca <- prcomp (log.modelo,center=TRUE,scale = TRUE)
```

Para ver los resultados de los cálculos podemos utilizar las dos siguientes líneas:

```
print(modelo.pca)
plot(modelo.pca,type="l")
```

La primera se encarga de imprimir por pantalla los valores de los vectores de componentes mientras que la segunda genera la siguiente gráfica de variabilidad haciendo uso de las desviaciones estándar calculadas en el modelo:

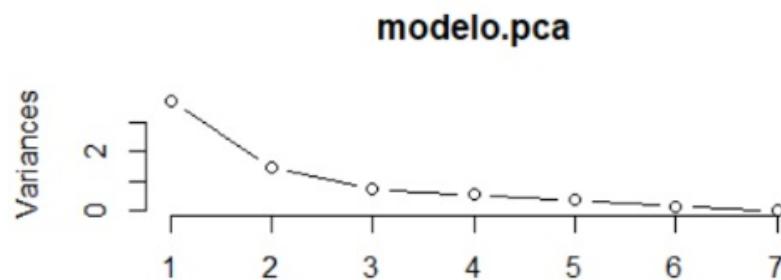


Figura 4. Gráfica de varianzas.

En la siguiente tabla 7x7 podemos ver las 7 componentes principales generadas.

	PC1	PC2	PC3	PC4
Presión arterial	0.3742470	-0.2582508	-0.142253834	-0.77310648
Edad	0.4977409	0.1796570	-0.041797493	-0.07303088
Peso	0.4561200	0.3228159	-0.007653248	0.16294087
Superficie corporal	0.3931690	0.3693694	0.102170140	0.34497471
Duración de la hipertensión	0.1885948	-0.4789927	0.835504126	0.12428394
Pulso	0.4355266	-0.2269764	-0.093178105	0.03853413
Medida del estrés	0.1559309	-0.6159819	-0.510656639	0.48423569

Tabla 3. Componentes principales generadas.

PC5	PC6	PC7
0.33483935	-0.06233825	0.24354160
0.01403677	0.38481971	-0.75140775
-0.23955235	0.48442220	0.60752441
0.55027670	-0.52350649	0.05171666
0.07220211	0.12629939	0.01787901
-0.67414565	-0.53973468	-0.05334177
0.26040412	0.17899374	0.03388748

Tabla 4. Componentes principales generadas.

Si analizamos, por ejemplo, la columna de PC2, podemos determinar qué variables son las que más han influido en el cálculo del resultado. Aquellas que tienen un valor absoluto más alto son la «Duración de la hipertensión y la medida del estrés».

Como podemos ver, ambas son negativas lo que implica una relación inversamente proporcional. Por tanto, **a valores altos de PC le corresponden valores pequeños de duración y estrés**. Por otra parte, podemos observar cómo los valores que menos han influido en la variable son los de pulso y edad puesto que son los valores más cercanos a 0.

Como podemos ver, en total se han creado **7 componentes principales** (una por dimensión). Es fácil observar que las primeras componentes aglutinan la mayor parte de la variabilidad mientras que las últimas apenas si tienen influencia. Para conseguir expresar toda la variabilidad habría que escoger las 7 componentes, pero entonces no reducimos la dimensión. El número de componentes a escoger dependerá de la fiabilidad del resultado y de la necesidad de exactitud en la información. En este ejemplo escogeremos las 2 primeras componentes principales ya que aglutinan gran parte de la información y nos permiten llevar a cabo una representación gráfica en 2 dimensiones de todos los datos. Por tanto, la versión reducida calculada del conjunto de datos queda tal y como puede verse a continuación:

	PC1	PC2
1	8.659.059	-2.321.267
2	8.696.830	-1.640.598
3	8.840.625	-2.031.832
4	9.098.019	-3.061.586
5	9.053.138	-3.195.898
6	8.889.027	-1.807.646
7	8.834.559	-2.085.719
8	8.582.932	-1.562.723
9	8.905.295	-2.936.821
10	8.852.548	-2.373.602
11	9.049.295	-2.954.202
12	8.803.646	-2.112.756
13	9.040.241	-2.877.702
14	8.879.258	-2.941.076
15	9.331.941	-3.282.783
16	9.070.010	-3.161.177
17	8.476.141	-1.608.344
18	8.660.625	-1.765.084
19	9.079.000	-3.318.188
20	9.223.736	-3.142.387

Tabla 5. Conjunto de datos.

Para generar la tabla anterior se han ejecutado las siguientes sentencias:

```
pc1 = modelo.pca$rotation[,1]
datoscompl = rep(0,20)
for(i in 1:7)
  datoscompl = datoscompl + log.modelo[,i] * pc1[i]

datoscomp2 = rep(0,20)
pc2 = modelo.pca$rotation[,2]
for(i in 1:7)
  datoscomp2 = datoscomp2 + log.modelo[,i] * pc2[i]

nuevosdatos = matrix(c(datoscompl,datoscomp2),nrow = 20)
```

Dichas sentencias únicamente multiplican cada columna del modelo de datos por el correspondiente valor del vector de componentes principales. Finalmente, suma el resultado generando una columna. La operación se repite para la primera y luego para la segunda componente. Finalmente, la sentencia matrix genera una matriz con el conjunto de datos deseado.

Estos datos sí que pueden graficarse fácilmente utilizando una gráfica de dispersión. Para generar la gráfica usaremos la siguiente sentencia sobre los cálculos realizados anteriormente:

```
plot(nuevosdatos)
```

Dicha sentencia dibuja una gráfica de dispersión utilizando las dos variables incluidas en la matriz nuevos datos. Vemos el resultado a continuación:

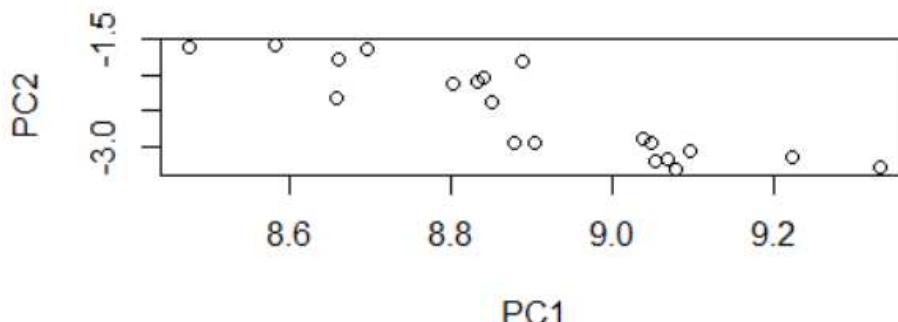
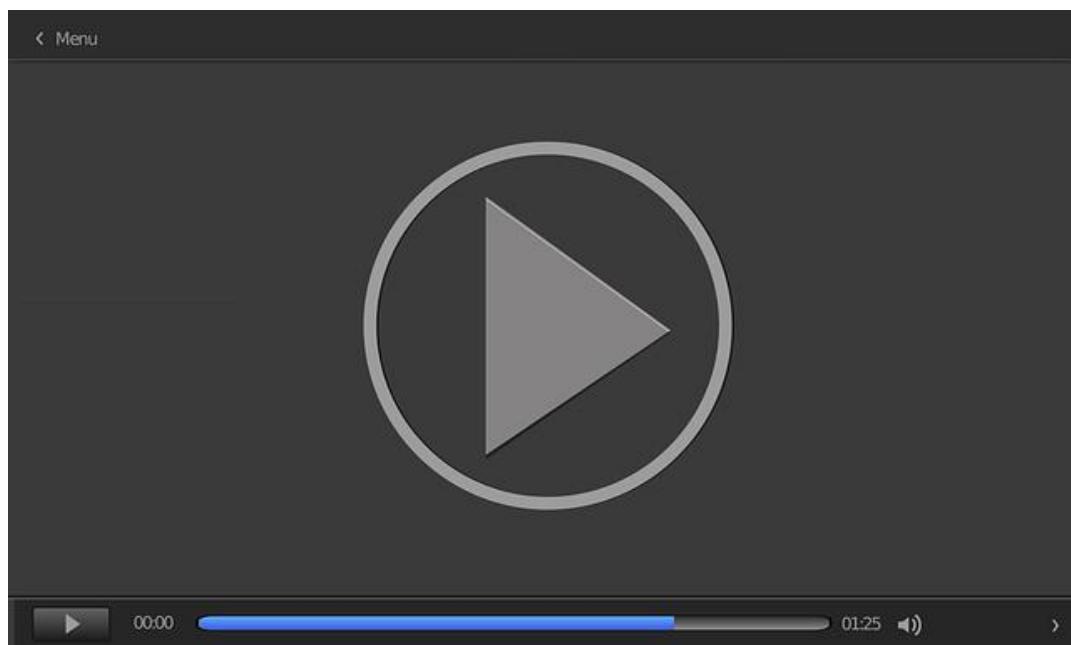


Figura 5. Gráfica de dispersión.

Al contrario de lo que ocurría con la gráfica de «Presión vs Edad» que vimos al principio del tema, en esta gráfica de dispersión sí se tienen en cuenta todas las variables del modelo de datos ya que PC1 y PC2 son **agregaciones de todos los datos del modelo original**.

Análisis de componentes principales usando R

En este vídeo vamos a conocer uno de los elementos estadísticos, como herramienta, más potentes para estudiar sistemas con muchas variables de las que no conocemos relaciones ni dependencias



Accede al vídeo:

<https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=4bc391f2-17e9-4109-bac4-acbd00fd6def>

10.6 Referencias bibliográficas

Componentes principales. Recuperado de
https://www.mhe.es/universidad/ciencias_matematicas/pena/home/CAPITULO.PDF

Computing and visualizing PCA in R. Recuperado de <https://www.r-bloggers.com/computing-and-visualizing-pca-in-r/>

Gurrea, M. (2000). *Análisis de componentes principales*. Madrid: MECD.

Análisis de componentes principales. Caso práctico en R

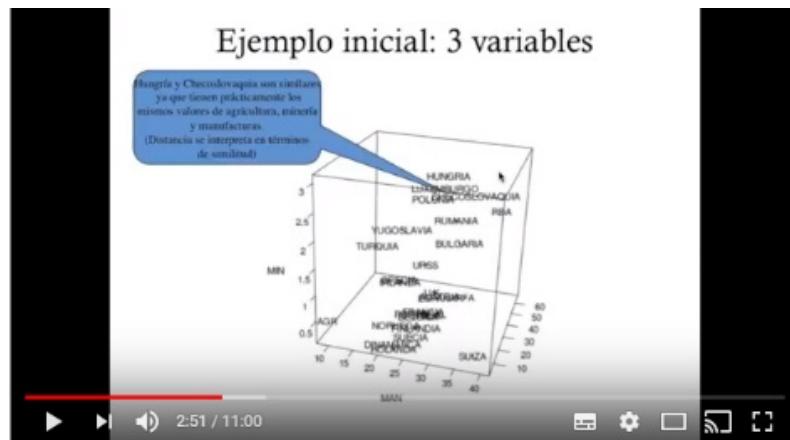
En este vídeo podemos encontrar una breve explicación de lo que es el análisis de componentes principales y como realizar un ejemplo práctico en R.

Accede al vídeo a través del aula virtual o desde la siguiente dirección web:

<https://www.youtube.com/watch?v=Hq0Ldt5S8Og>

Explicación intuitiva del análisis de componentes principales

Corta explicación acerca de qué es el análisis de componentes principales.



Accede al vídeo a través del aula virtual o desde la siguiente dirección web: https://www.youtube.com/watch?v=L1_iQAUa228

Aplicación del análisis de componentes principales en el diagnóstico socioambiental

Olivares, B. (2014). Aplicación del Análisis de Componentes Principales(ACP) en el diagnóstico socioambiental. Caso: sector Campo Alegre, municipio Simón Rodríguez de Anzoátegui. *Multiciencias*, 14(4), 364-374.

En este artículo puede verse una interesante aplicación del análisis de componentes principales a un caso real.

Accede al artículo desde el aula virtual o a través de la siguiente dirección web:

<http://www.redalyc.org/pdf/904/90433839011.pdf>

- 1.** ¿Para qué sirve el análisis de componentes principales?
 - A. Para obtener información de los datos.
 - B. Para reducir la dimensión de los datos.
 - C. Para aumentar la dimensión de los datos.
 - D. Todo lo anterior es falso.

- 2.** Los vectores de componentes principales:
 - A. Se obtienen agregando los valores de las variables del modelo.
 - B. Se obtienen utilizando datos externos al modelo.
 - C. Sirven para aumentar la dimensión de los datos.
 - D. B y C son correctas.

- 3.** ¿Cuándo es una variable representativa del vector de componentes?
 - A. Cuando el valor asociado es cercano a 1.
 - B. Cuando el valor asociado es cercano a -1.
 - C. Cuando el valor asociado es cercano a 0.
 - D. A y B son ciertas.

- 4.** ¿Qué componentes son las que representan el ruido en una imagen?
 - A. Las cercanas a PC1.
 - B. las cercanas al último PC.
 - C. Todas.
 - D. Ninguna.

5. Para detectar cambios en un modelo de datos...
 - A. Necesitamos distintos modelos de datos hechos en distintos tiempos o bajo distintas circunstancias.
 - B. Solo necesitamos un único modelo de datos.
 - C. Es necesario conocer la media.
 - D. Todo lo anterior es falso.
6. ¿Qué sucede si reducimos el número de componentes principales a utilizar en nuestro modelo reducido?
 - A. Perdemos precisión en los datos.
 - B. Reducimos la dimensión.
 - C. A y B son ciertas.
 - D. A y B son falsas.
7. ¿Cómo calculamos el modelo de datos reducido a partir del vector de componentes?:
 - A. Multiplicando cada individuo por la suma de los valores del vector de componentes principales.
 - B. Multiplicando cada valor de la variable de cada individuo por la derivada del vector de componentes.
 - C. Generamos el valor agregado asociado a cada individuo multiplicando cada valor del vector de componentes por la variable asociada y sumando el resultado.
 - D. Aplicando A y luego B.

- 8.** Si nos atenemos al número de variables a representar, ¿qué gráfica de dispersión es más fácil de entender e interpretar?:
- A. Una con 2 dimensiones.
 - B. Una con 3 dimensiones.
 - C. Una con 4 dimensiones.
 - D. Ninguna de las anteriores.
- 9.** ¿Qué función de R hemos utilizado para poder calcular el vector de análisis de componentes?
- A. matrix.
 - B. prcomp.
 - C. plot.
 - D. Ninguna de las anteriores.
- 10.** ¿Si tenemos un modelo de datos con 5 variables, ¿cuántos vectores de componente principal generará el algoritmo?
- A. 5.
 - B. 6.
 - C. 7.
 - D. 2.