



UNIVERSIDAD DE VALLADOLID

Dpto. Estadística e I.O.

La enseñanza de la estadística con herramientas didácticas como “R”

ANEXO 8:A212 1º Bachillerato Aplicadas.

**Trabajo Final del Máster Universitario de Profesor en Educación
Secundaria Obligatoria y Bachillerato. Especialidad de Matemáticas.**

Alumno: Julián Rodríguez Vaca.

Tutor: Dr. David Conde del Río.

Valladolid, Junio 2018.

Índice general.

Índice general.....	3
Capítulo 1. Introducción.....	5
Capítulo 2. Contenidos y estándares oficiales.....	7
Capítulo 3. Introducción a “R” Statistics.....	9
Capítulo 4. Contenidos y estándares oficiales con “R” Statistics.	13
Capítulo 5. Estándares de aprendizaje evaluables.....	35
Capítulo 6. Bibliografía	37

Capítulo 1. Introducción.

El objetivo del presente trabajo es plantear y ofrecer una propuesta para la mejora de la didáctica de la estadística mediante el empleo de un potente software, destinado hasta hoy a estudios superiores.

Por un motivo ético, se ha elegido Software Libre, con el que se pretende fomentar este tipo de herramientas en el aula.

Este software, además de ahorrar una gran cantidad de tiempo, permitirá hacer más dinámica esta parte, gracias al manejo de grandes volúmenes de información, la realización de gráficos estadísticos de manera automática, y permitiendo el análisis de los datos de una forma más adecuada.

Con esto el alumno comenzará a tomar contacto con un software de programación y un lenguaje de alto nivel, lo que le mostrará puertas por abrir, y le aportará una buena ventaja sobre todo si se plantea estudios superiores.

El trabajo se presenta en forma de memoria, donde se recopila cada punto del temario en su versión más extensa, y donde aparecen más de 50 ejemplos de cómo resolver los ejercicios de forma tradicional, y con “R”. Incorpora nueve anexos con el temario preparado para cada uno de los cursos, que el profesor puede proporcionar a sus alumnos. Tanto en la memoria como en los anexos, aparece todo el código utilizado en la elaboración del trabajo. Las versiones de introducción o de repaso de cada punto del temario se han dejado en cada uno esos anexos, para evitar la duplicación de los contenidos en la memoria.

Puesto que el bloque de estadística se presenta en todos los cursos en el último bloque de la asignatura de matemáticas, sufre los retrasos de todos los bloques precedentes, dejando en la mayoría de las ocasiones un tiempo muy reducido para el desarrollo del mismo. Con el uso de este método, no se trata de evitar que el alumno trabaje el tan necesario cálculo mental y manual. Sin embargo, si el grupo llega hasta este punto con retraso, uno de los motivos puede

ser precisamente el llevar trabajando cerca de ocho meses en esta línea. Por ello, se trata de optimizar el poco tiempo del que disponga el profesor, evitando pérdidas en la representación de gráficos a mano, nubes de puntos, o tablas de contingencia.

El BOCYL establece, en sus Ordenes EDU 362 y 363 del 4 de mayo de 2015, que el quinto bloque, «Estadística y probabilidad», es de suma importancia.

Esto no sólo es cierto, sino que además, en la era de la información y de la competitividad, el futuro de las empresas y de los países no dependerá tanto del volumen de información de que dispongan, sino de la mejor explotación que hagan de la misma.

Independientemente de su elección tras acabar la ESO o el Bachillerato, el alumno adquirirá los conceptos y el vocabulario necesarios para poder aplicarlos de manera prácticamente autónoma en su futura profesión.

Así, al finalizar sus estudios será capaz de realizar análisis críticos de una mayor cantidad de información mediante tablas y gráficas, con la ayuda de “R”.

Será capaz de recopilar datos por sí mismo, organizarlos, resumirlos, estudiarlos y explotarlos, lo que le será de gran utilidad en su ámbito profesional.

El contenido del trabajo está adaptado a la comunidad de Castilla y León, según las órdenes EDU 362 y 363 de 2015 por las que se establecen los currículos y se regulan la implantación, evaluación y desarrollo de la educación secundaria obligatoria y del bachillerato en la Comunidad de Castilla y León:

ORDEN EDU/362/2015, de 4 de mayo, por la que se establece el currículo y se regula la implantación, evaluación y desarrollo de la educación secundaria obligatoria en la Comunidad de Castilla y León.

ORDEN EDU/363/2015, de 4 de mayo, por la que se establece el currículo y se regula la implantación, evaluación y desarrollo del bachillerato en la Comunidad de Castilla y León.

Así, establecen los temas para el bloque de estadística que veremos a continuación.

Capítulo 2. Contenidos y estándares oficiales.

1º Bachillerato. Aplicadas

- 1BA1.- Estadística descriptiva bidimensional: Tablas de contingencia.
- 1BA2.- Distribución conjunta y distribuciones marginales.
- 1BA3.- Medias y desviaciones típicas marginales.
- 1BA4.- Distribuciones condicionadas.
- 1BA5.- Independencia de variables estadísticas.
- 1BA6.- Estudio de la dependencia de dos variables estadísticas.
 - Representación gráfica: Nube de puntos.
- 1BA7.- Dependencia lineal de dos variables estadísticas.
 - Covarianza y correlación: Cálculo e interpretación del coeficiente de correlación lineal.
- 1BA8.- Regresión lineal. Recta de regresión. Estimación. Predicciones estadísticas y fiabilidad de las mismas.

Capítulo 3. Introducción a “R” Statistics.

3.1 Sobre “R”:

R es un lenguaje y entorno para el procesamiento y representación de datos estadísticos. Es un proyecto de GNU es similar al lenguaje y al entorno S, que fue desarrollado en Bell Laboratories, por John Chambers y su equipo. Hay algunas diferencias importantes, pero gran parte del código escrito para S corre inalterado bajo R.

R proporciona una amplia variedad de técnicas estadísticas y gráficos, y es altamente extensible mediante la creación de librerías por los usuarios, al ser una herramienta de código abierto.

Uno de los puntos fuertes de R es la facilidad con la que se pueden crear gráficos de calidad, incluyendo símbolos matemáticos y fórmulas si es necesario.

R está disponible como Software Libre bajo los términos de la Licencia Pública General GNU de la Free Software Foundation en forma de código fuente. Se compila y se ejecuta en una amplia variedad de plataformas UNIX y sistemas similares (incluidos FreeBSD y Linux), Windows y MacOS.

Fuente: <https://www.r-project.org/about.html>

3.2 Descarga e instalación de RStudio:

Rstudio es un software gratuito que podemos descargar de forma totalmente legal y sin coste ni publicidad, de la siguiente página:

<https://www.rstudio.com/products/rstudio/download/>

Pulsando en la tecla download, que aparece en la columna Free (gratis), nos dirige a la zona para elegir nuestro sistema operativo:

Installers	Size	Date	MD5
RStudio 1.1.453 - Windows Vista/7/8/10	85.8 MB	2018-05-16	bf287e385aef53829204023087e98735
RStudio 1.1.453 - Mac OS X 10.6+ (64-bit)	74.5 MB	2018-05-16	00a0088424ed06ac434f7a966f602b9c
RStudio 1.1.453 - Ubuntu 12.04-15.10/Debian 8 (32-bit)	89.3 MB	2018-05-16	6cfd86770c7b6dbc13e66f4f59c299ce
RStudio 1.1.453 - Ubuntu 12.04-15.10/Debian 8 (64-bit)	97.4 MB	2018-05-16	63e36e8138e369d19f9aaf4b0e995bbc
RStudio 1.1.453 - Ubuntu 16.04+/Debian 9+ (64-bit)	64.4 MB	2018-05-16	85b3e76c9fad4613bc9cf0de1f34b183
RStudio 1.1.453 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	88.1 MB	2018-05-16	37cade7e162eab62483e6556e39dedee
RStudio 1.1.453 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	90.6 MB	2018-05-16	44cddd285bc31c41e4eaeed1d74b8eebb

Dependiendo del sistema operativo de nuestro PC, descargamos el que corresponda.

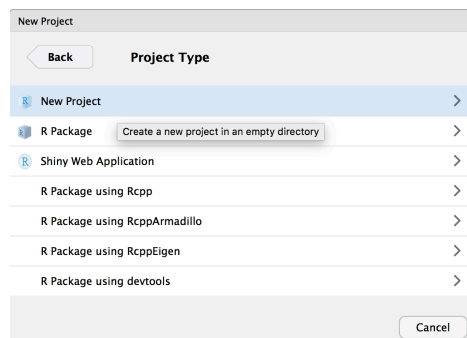
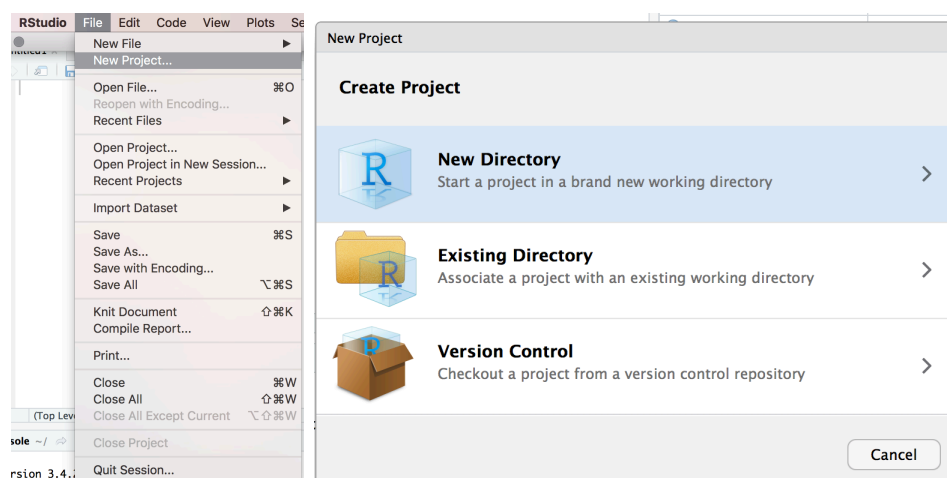
Una vez descargado, se ejecuta el programa instalador, y se van siguiendo los pasos del asistente de instalación, como en cualquier otro programa.

3.3.- Creación de un nuevo proyecto:

1.- Vamos a crear nuevo proyecto con el nombre población y muestra, ubicado en el escritorio del PC.

Para ello, abrir Rstudio y pulsar:

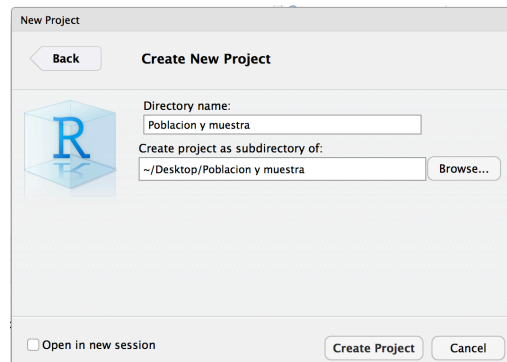
File/New project.../New directory/New Project



En la siguiente ventana, escribimos:

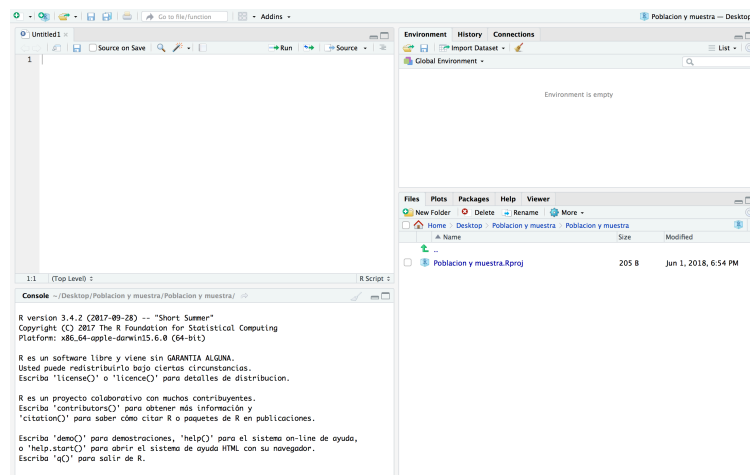
Directory name: Población y muestra.

Create project as subdirectory of: Click en browse... y creamos una carpeta en nuestro escritorio (desktop) que se llame Población y muestra.



Nota: las carpetas podemos crearlas tanto en el escritorio como en un USB o donde queramos, y luego localizarla usando la tecla browse.

Con esto, se nos abre el entrono de “R”, listo para empezar. Tendrá la siguietente pinta:



Los comandos se escriben en la zona inferior izquierda, y los gráficos se mostrarán en la ventana inferior derecha. Las ventanas superiores son para la selección y visualización de tablas y otras variables. Con esto el sistema está listo para comenzar a trabajar

Capítulo 4. Contenidos y estándares oficiales con “R” Statistics.

DeSeCo (2003) define competencia como «la capacidad de responder a demandas complejas y llevar a cabo tareas diversas de forma adecuada».

La competencia «supone una combinación de habilidades prácticas, conocimientos, motivación, valores éticos, actitudes, emociones, y otros componentes sociales y de comportamiento que se movilizan conjuntamente para lograr una acción eficaz».

Se contemplan, pues, como conocimiento en la práctica, es decir, un conocimiento adquirido a través de la participación activa en prácticas sociales y, como tales, se pueden desarrollar tanto en el contexto educativo formal, a través del currículo, como en los contextos educativos no formales e informales.

Fuente: <https://www.mecd.gob.es/>

En este trabajo se ha buscado contribuir a las competencias en:

Comunicación lingüística, mediante el fomento de un uso del vocabulario apropiado, de la lectura y sobre todo de la interpretación de los enunciados, que contribuyen finalmente a expresarse y comunicarse con propiedad.

Competencia matemática, mediante el análisis matemático del comportamiento de las variables de estudio de la población, extrayendo conclusiones en función de la regresión lineal y correlación de los datos de las variables, y bajo la interpretación conjunta de parámetros estadísticos.

Competencia digital, mediante el fomento de un uso ético, cívico y crítico de las nuevas tecnologías, y mediante el empleo de una herramienta Software de alto nivel.

Competencias sociales y cívicas, mediante el análisis de datos de nuestro entorno, como PIB, IPC, extrayendo conclusiones de posibles desigualdades salariales en poblaciones, o identificando las malas prácticas de las presentaciones de datos de forma interesada.

Competencia cultural, representando e interpretando la información con relación a ejemplos de plantas y otros datos del entorno.

La competencia aprender a aprender, mediante el ejemplo de la búsqueda de información para mejorar la utilización del software R de manera casi autodidacta.

Sentido de la iniciativa y espíritu emprendedor, mostrando al alumno el inicio de un camino, que con su propia iniciativa podrá recorrer hasta donde le lleve su curiosidad científica. Por su potencia y escasa inversión, el alumno será capaz de imaginar escenarios de emprendimiento, donde con un ordenador y este software como herramientas podrá realizar estudios de alto valor a nivel profesional.

1BA1.- Estadística descriptiva bidimensional: Tablas de contingencia.

Es muy común enfrentarse a la necesidad de estudiar dos características o variables estadísticas de una misma población.

Una variable estadística bidimensional es el conjunto de pares de valores de caracteres X e Y sobre una población, y se representa por (X,Y) .

Cada uno de los individuos de la población estará representado por una pareja (x_i, y_i) , donde x_i representa los datos, valores, o marcas de clase x_1, x_2, \dots, x_n , de la variable X , e y_i representa los datos, valores o marcas de clase, y_1, y_2, \dots, y_m de la variable Y .

Cada una de las variables estadísticas que forman la variable estadística bidimensional pueden ser:

Cualitativas.

Cuantitativas discretas.

Cuantitativas continuas.

Tablas de contingencia

Si el número de datos es grande y los pares se repiten, se utiliza una tabla de contingencias.

La tabla se contruye con las frecuencias marginales de todos los pares de valores.

$X \backslash Y$	y_1	y_2	...	y_m	Suma
x_1	f_{11}	f_{12}	...	f_{1m}	$f_{1 \cdot}$
x_2	f_{21}	f_{22}	...	f_{2m}	$f_{2 \cdot}$
...
x_n	f_{n1}	f_{n2}	...	f_{nm}	$f_{n \cdot}$
Suma	$f_{\cdot 1}$	$f_{\cdot 2}$...	$f_{\cdot m}$	N

x_1, x_2, \dots, x_n son los valores de la variable x .

y_1, y_2, \dots, y_n son los valores de la variable y .

f_{11} es la frecuencia del valor (x_1, y_1) .

En general f_{nm} , es la frecuencia del valor (x_m, y_n) .

$f_{1\cdot}$ es la suma de todas las frecuencias del valor $x_{1\cdot}$ (Fila 1)

$f_{\cdot 1}$ es la suma de todas las frecuencias del valor $y_{\cdot 1}$ (Columna 1)

Ejemplo:

Un grupo de alumnos lanzan a dos canastas, X e Y, con 4 lanzamientos cada uno, y el profesor de educación física anota los fallos la tabla. Cada casilla recoge el número de alumnos que han fallado en los dos lanzamientos. ¿Qué número de alumnos han fallado el primer lanzamiento en la canasta (X)? ¿Y al tercer lanzamiento de la canasta (Y)?

A mano, se sumarían las filas de cada valor de X y las columnas de cada valor de Y:

X\Y	1	2	3	4	Suma
1	12	6	4	2	$12+6+4+2=24$
2	8	7	3	0	$8+7+3+0=18$
3	6	5	2	1	$6+\dots+1=14$
4	4	4	1	0	8
Suma	30	21	10	3	64

Los alumnos que han fallado el primer lanzamiento de la canasta X han sido 24, y los que han fallado el tercer lanzamiento de la canasta Y, han sido 10.

Distribuciones bidimensionales.

Se denominan distribuciones bidimensionales al conjunto de parejas de valores (x_i, y_i) , que pueden presentarse mediante una tabla que las relaciona mediante las frecuencias absolutas de todos los posibles valores de la variable estadística bidimensional (X, Y) . Normalmente a la variable x se la llama independiente y a la variable y , variable dependiente.

Las tablas bidimensionales simples adoptan la forma siguiente:

Variable X	Variable Y	Frecuencia Absoluta
x_i	y_i	f_i
x_1	y_1	f_1
x_2	y_2	f_2
...
x_i	y_i	f_i
...
x_n	y_n	f_n
		$\sum_{i=1}^N f_i = N$

Las tablas bidimensionales de doble entrada, adoptan la forma siguiente:

$Y \backslash X$	x_1	x_2	...	x_i	...	x_n	Frecuencia absoluta de y
y_1	f_{11}	f_{21}	...	f_{i1}	...	f_{n1}	$\sum f_{i1}$
y_2	f_{12}	f_{22}	...	f_{i2}	...	f_{n2}	$\sum f_{i2}$
...
y_j	f_{1j}	f_{2j}	...	f_{ij}	...	f_{nj}	$\sum f_{ij}$
...
y_m	f_{1m}	f_{2m}	...	f_{im}	...	f_{nm}	$\sum f_{im}$
Frecuencia absoluta de x	$\sum f_{1j}$	$\sum f_{2j}$...	$\sum f_{ij}$...	$\sum f_{nj}$	N

En R se utiliza la función `table(Y,X)`

Se define f_{ij} a la frecuencia absoluta correspondiente al valor (x_i, y_j) multiplicada por N el número total de individuos. La última fila y la última

columna presentan las llamadas distribuciones marginales, y se corresponden con las distribuciones o tablas estadísticas correspondientes a las variables unidimensionales X e Y.

Ejemplo:

En una clase de 35 alumnos, hemos hecho una encuesta sobre el número de primos que tiene cada uno, con los resultados que figuran a continuación.

Y\X	0	1	2	3	Tot.
0	0	2	3	1	6
1	3	6	4	1	14
2	4	2	3	0	9
3	3	1	1	1	6
Tot.	10	11	11	3	35

La variable X indica el número de primos, y la variable Y el número de primas de los alumnos.

- Construye la tabla estadística bidimensional simple correspondiente.
- En las distribuciones marginales, calcula la media y la desviación típica.

a) La tabla bidimensional simple, es:

x_i	0	0	0	1	1	1	1	2	2	2	2	3	3	3
y_i	1	2	3	0	1	2	3	0	1	2	3	0	1	3
f_i	3	4	3	2	6	2	1	3	4	3	1	1	1	1

- Para el cálculo de la media y la desviación típica de las distribuciones marginales, nos ayudamos de las tablas siguientes:

x_i	$f_i x_i$	$f_i x_i^2$	$f_i x_i^3$
0	10	0	0
1	11	11	11
2	11	22	44
3	3	9	27
Total	35	42	82

$$\bar{x} = 1,2 \quad \sigma_x = 0,95$$

y_i	$f_i y_i$	$f_i y_i^2$	$f_i y_i^3$
0	6	0	0
1	14	14	14
2	9	18	36
3	6	18	54
Total	35	50	104

$$\bar{y} = 1,43 \quad \sigma_y = 0,965$$

1BA2.- Distribución conjunta y distribuciones marginales.

Para describir conjuntamente dos variables, nos ayudaremos de las tablas de frecuencias. En ellas, la frecuencia absoluta conjunta se representa por n_{ij} , e indica la cantidad de veces que se presenta la pareja (x_i, y_j) .

Tabla de frecuencias conjunta

Son tablas sobre las que se colocan las variables X e Y colocadas en orden creciente, la X en la primera columna, y la Y en la primera fila.

En la zona central aparecen las frecuencias conjuntas. Se pueden colocar las frecuencias absolutas, y las relativas separadas por una barra "/".

x_i/y_j	0	1	2	3	n_i	f_j
0	0/0	0/0	1/0'04	0/0	1	0'04
1	0/0	0/0	0/0	1/0'04	1	0'04
2	0/0	3/0'12	5/0'20	0/0	8	0'32
3	0/0	8/0'32	4/0'16	0/0	12	0'48
4	1/0'04	2/0'08	0/0	0/0	3	0'12
n_i	1	13	10	1	25	
f_j	0'04	0'52	0'04	0'04		1

Las frecuencias absolutas marginales serían las frecuencias de cada variable estudiada de forma independiente.

Para la X, (x_i) sería el número de veces que se repite el valor x_i sin tener en cuenta los valores de Y, la representamos por n_i .

Para la Y, (y_j) sería el número de veces que se repite el valor y_j sin tener en cuenta los valores de la X, la representamos por n_j .

A partir de las anteriores, y del mismo modo, se pueden obtener las frecuencias relativas marginales f_i y f_j .

En la tabla del ejemplo se han añadido una fila y una columna para recoger toda la información.

1BA3.- Medias y desviaciones típicas marginales.

Con las definiciones de media, desviación típica y varianza del apartado de distribuciones unidimensionales, utilizando para la X los valores y el número de veces que se repite n_i y N el número total de pares observados, y para la Y los valores y_j y el número de veces que se repite n_j y N el número total de pares observados, calcularemos las medias marginales, desviaciones típicas marginales y varianzas marginales.

1BA4.- Distribuciones condicionadas.

Las distribuciones condicionadas se obtienen a partir de la distribución de frecuencias conjuntas, al fijar el valor de una de las variables.

Frecuencia absoluta condicionada para X (x_i):

$$n_{i(j)} = n_{ij} \text{ para todo } i = 1, 2, \dots, k.$$

Frecuencia absoluta condicionada para Y (y_j) :

$$n_{(i)j} = n_{ij} \text{ para todo } j = 1, 2, \dots, h.$$

En las distribuciones condicionadas se suelen utilizar las frecuencias relativas condicionadas, que se definen mediante la expresión:

$$f_{i(j)} = \frac{n_{ij}}{n_j}$$

1BA5.- Independencia de variables estadísticas.

Estadísticamente, dos variables son independientes si su frecuencia relativa conjunta es igual al producto de sus frecuencias relativas marginales.

$$f_{ij} = \frac{n_{ij}}{n} = f_i \cdot f_j = \frac{n_i}{n} \cdot \frac{n_j}{n}$$

Además se cumplirá que todas sus frecuencias relativas condicionadas son iguales a sus correspondientes frecuencias marginales:

$$f_{i(j)} = f_i \text{ para todo } j \text{ y } f_{(i)j} = f_j \text{ para todo } i.$$

1BA6.- Estudio de la dependencia de dos variables estadísticas.
Representación gráfica: Nube de puntos.

Se pueden representar gráficamente las distribuciones bidimensionales en un diagrama de ejes X Y.

Considerando cada par de valores (x,y) como las coordenadas de un punto, se consigue una gráfica denominada diagrama de dispersión o nube de puntos.

Ejemplo:

Las siguientes parejas de valores (X,Y), muestran los resultados de una encuesta realizada a 30 alumnos, donde el primer valor son las horas de estudio, X y el segundo valor, el número de suspensos, Y:

(2,0)(2,2)(0,5)(2,1)(1,2)(2,1)(3,1)(4,0)(0,4)(2,2)(2,1)(2,1)(4,0)(3,1)(2,4)
 (2,1)(1,2)(2,1)(2,0)(3,0)(3,2)(2,2)(2,2)(2,1)(0,5)(1,3)(2,2)(2,1)(1,3)(1,4)

Construye la tabla estadística bidimensional de doble entrada y las tablas de distribuciones marginales.

- a) Realiza el diagrama de dispersión.
a) Las tablas estadísticas pedidas, son:

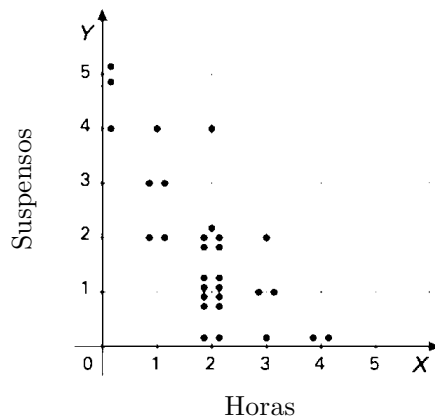
Y\X	0	1	2	3	4	Total
0	0	0	2	1	2	5
1	0	0	8	2	0	10
2	0	2	5	1	0	8
3	0	2	0	0	0	2
4	1	1	1	0	0	3
5	2	0	0	0	0	2
Total	3	5	4	4	2	30

Tablas de distribuc. marginales:

x_i	0	1	2	3	4	
f_i	3	5	16	4	2	30

y_i	0	1	2	3	4	5
f_i	5	10	8	2	3	2

Para crear el diagrama de dispersión, trazaríamos los ejes, e iríamos colocando cada punto.



Vamos a ver cómo hacer esto con "R".

Creamos una variable que se llame horas, y otra que se llame suspensos, con las horas y los suspensos que nos dice el enunciado. Para ello utilizamos la función concatenar:

```
> horas<-c(2,2,0,2,1,2,3,4,0,2,2,2,4,3,2,2,1,2,2,3,3,2,2,0,1,2,2,1,1)
> suspensos<-c(0,2,5,1,2,1,1,0,4,2,1,1,0,1,4,1,2,1,0,0,2,2,2,1,5,3,2,1,3,4)
```

Podemos consultar las variables que acabamos de crear:

```
> horas
[1] 2 2 0 2 1 2 3 4 0 2 2 2 4 3 2 2 1 2 2 3 3 2 2 2 0 1 2 2 1 1
> suspensos
[1] 0 2 5 1 2 1 1 0 4 2 1 1 0 1 4 1 2 1 0 0 2 2 2 1 5 3 2 1 3 4
```

Con esto, crear las tablas del enunciado y el diagrama de dispersión es muy sencillo:

Tabla bidimensional:

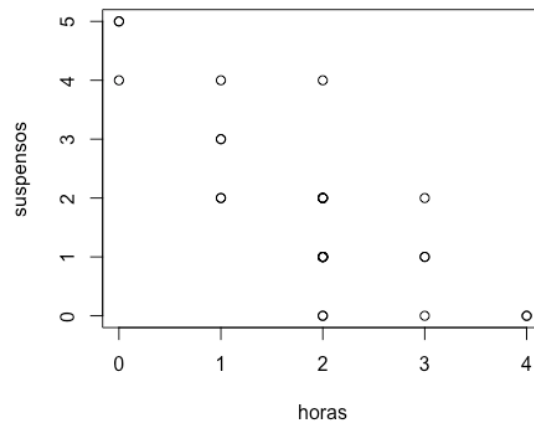
```
> tabla_bidim=table(suspensos,horas)
> tabla_bidim
      horas
suspensos 0 1 2 3 4
0      0 0 2 1 2
1      1 0 0 8 2 0
2      2 0 2 5 1 0
3      3 0 2 0 0 0
4      4 1 1 1 0 0
5      5 2 0 0 0 0
```

Las tablas de frecuencias también son muy sencillas de construir:

```
> fabshoras<-table(horas)
> fabshoras
horas
0 1 2 3 4
3 5 16 4 2
> fabSusp<-table(suspensos)
> fabSusp
suspensos
0 1 2 3 4 5
5 10 8 2 3 2
```

La nube de puntos se construye con la función plot:

```
> plot(horas,suspensos)
```



Dependencia lineal:

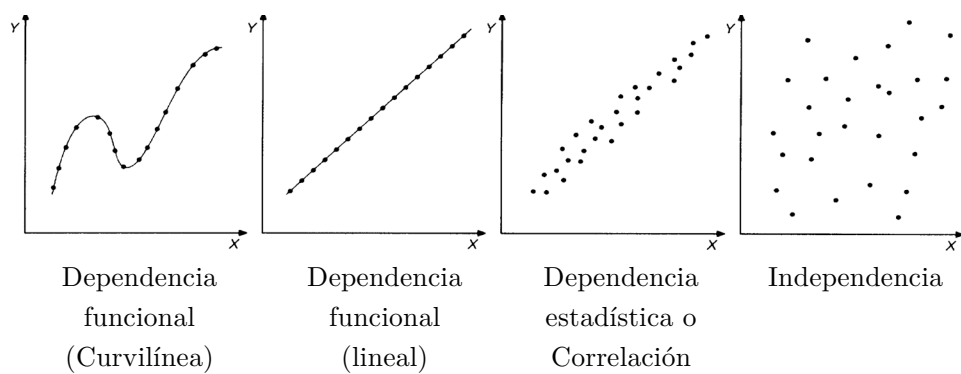
La forma de la nube de puntos, nos permite intuir si hay relación, o dependencia las dos variables. Esta dependencia, si existe, se llama correlación.

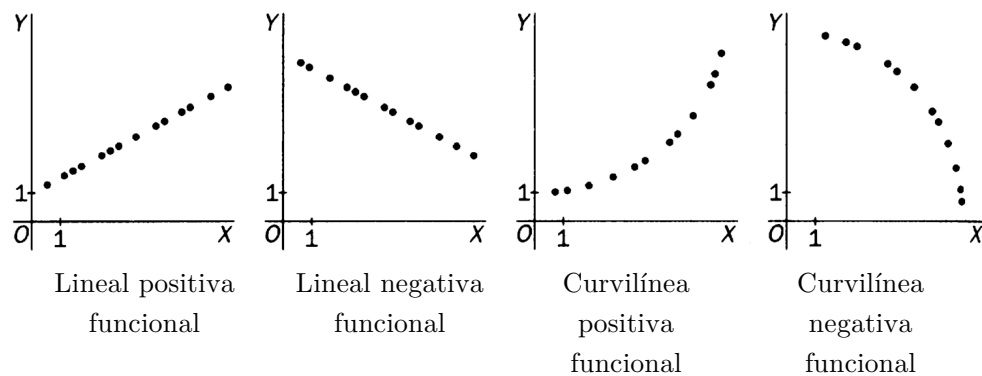
Existen varios tipos de dependencia:

Funcional, si la nube de puntos puede asemejar a la gráfica de una función.

Lineal, si la nube de puntos se asemeja a una recta.

Independencia o ausencia de correlación.





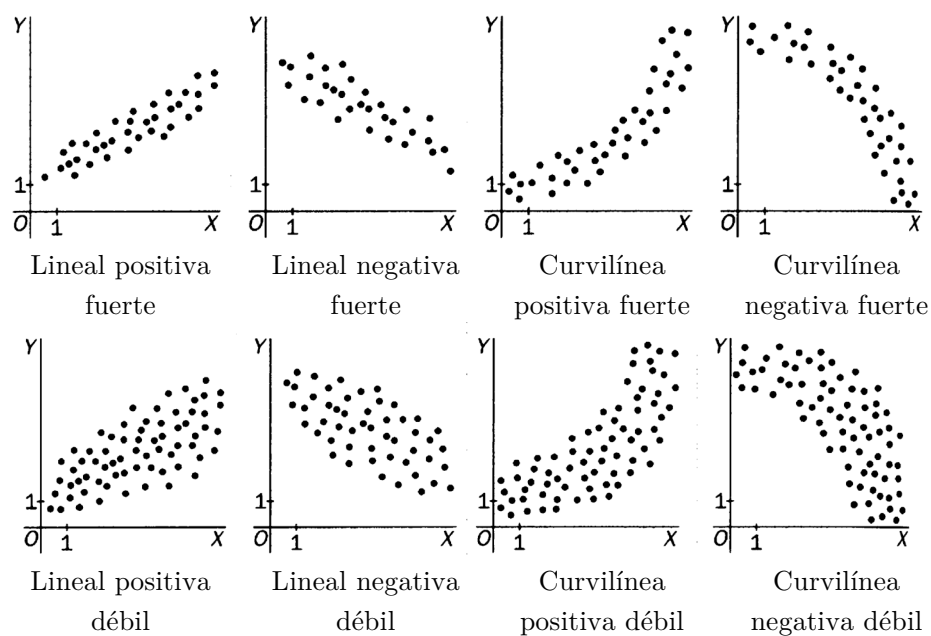
El grado de correlación, a su vez, puede ser:

Correlación fuerte, si la nube de puntos se aproxima a una recta o una curva.

Correlación débil, si la nube de puntos se aproxima poco a una recta o una curva.

Correlación positiva, si a medida que crece una variable, crece la otra.

Correlación negativa, si a medida que crece una variable, decrece la otra.



1BA7.- Dependencia lineal de dos variables estadísticas. Covarianza y correlación: Cálculo e interpretación del coeficiente de correlación lineal.

La correlación de tipo lineal se mide mediante el coeficiente de correlación lineal de Pearson, cuyo valor puede calcularse mediante la expresión:

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

Siendo

$$\sigma_{x,y} = \frac{\sum f_{ij} \cdot x_i \cdot y_j}{N} - \bar{x} \cdot \bar{y}$$

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^N x_i^2 \cdot f_i}{N} - \bar{x}^2} \quad \sigma_y = \sqrt{\frac{\sum_{j=1}^N y_j^2 \cdot f_j}{N} - \bar{y}^2}$$

Donde:

$\sigma_{x,y}$ es la covarianza o la varianza conjunta de las variables X e Y.

σ_x y σ_y son las desviaciones típicas de las variables X e Y, respectivamente.

Escala de valores del coeficiente de correlación lineal.

El coeficiente de correlación lineal de Pearson, r, siempre toma valores comprendidos entre -1 y 1. Nos permite analizar el grado de aproximación de la nube de puntos a una línea recta:

Si $-1 < r < 0$, existe correlación lineal negativa, y será más fuerte cuanto más se aproxime r a -1.

Si $0 < r < 1$, existe correlación lineal positiva, y será más fuerte cuanto más se aproxime r a 1.

Si $r = 1$ o $r = -1$, la correlación es una dependencia lineal.

Si $r = 0$, no existe correlación lineal, pero sí puede existir correlación curvilínea.

Ejemplo.

Se han recogido en una tabla bidimensional las notas de matemáticas (X) y física (Y) de 40 estudiantes.

Calcula el coeficiente de correlación lineal de Pearson y analiza si dependencia entre las calificaciones de ambas asignaturas.

Y\X	3	4	5	6	7	8	10	Total
2	4	0	0	0	0	0	0	4
5	0	7	11	0	0	0	0	18
6	0	0	0	5	3	0	0	8
7	0	0	0	5	2	0	0	7
9	0	0	0	0	0	1	0	1
10	0	0	0	0	0	0	2	2
Total	4	7	11	10	5	1	2	40

Para hacer este ejemplo a mano, necesitamos calcular las desviaciones típicas y la covarianza. Para calcularlas nos ayudaríamos de las siguientes tablas:

x_i	f_i	$f_i x_i$	$f_i x_i^2$
3	4	12	36
4	7	28	112
5	11	55	275
6	10	60	360
7	5	35	245
8	1	8	64
10	2	20	200
Tot	40	218	1292

y_i	f_i	$f_i y_i$	$f_i y_i^2$
2	4	8	16
5	18	90	450
6	8	48	288
7	7	49	343
9	1	9	81
10	2	20	200
Tot	40	224	1378

(x_i, y_i)	f_{ij}	$f_{ij} \cdot x_i \cdot y_j$
(3,2)	4	24
(4,5)	7	140
(5,5)	11	275
(6,6)	5	180
(6,7)	5	210
(7,6)	3	126
(7,7)	2	98
(8,9)	1	72
(10,10)	2	200
Total	40	1325

Las medias aritméticas, las desviaciones típicas y la covarianza, son:

$$\bar{x} = \frac{\sum_{i=1}^k x_i \cdot f_i}{N} = \frac{218}{40} = 5,45; \quad \sigma_x = \sqrt{\frac{\sum_{i=1}^N x_i^2 \cdot f_i}{N} - \bar{x}^2} = \sqrt{\frac{1292}{40} - 5,45^2} = 1,612$$

$$\bar{y} = \frac{\sum_{j=1}^k y_j \cdot f_j}{N} = \frac{224}{40} = 5,6; \quad \sigma_y = \sqrt{\frac{\sum_{j=1}^N y_j^2 \cdot f_j}{N} - \bar{y}^2} = \sqrt{\frac{1378}{40} - 5,6^2} = 1,758$$

Estos valores nos permiten calcular el coeficiente de Pearson.

$$\gamma = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{2,605}{1,612 \cdot 1,758} = 0,919$$

Observamos que el valor del coeficiente está próximo a 1, y por tanto, existe una correlación lineal positiva fuerte entre las dos variables del enunciado.

Vamos a ver cómo hacer esto con R.

De la tabla podemos obtener que los pares de notas son las siguientes:

(3,2);(3,2);(3,2);(4,5);(4,5);(4,5);(4,5);(4,5);(4,5);(4,5);(5,5);(5,5)...(10,10);

Definimos un vector con las notas de matemáticas y otro con las notas de física:

```
>M<-c(3,3,3,3,4,4,4,4,4,4,4,5,5,5,5,5,5,5,5,5,6,6,6,6,6,6,6,6,7,7,7,7,8,10,10)
```

```
>F<-c(2,2,2,2,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,6,6,6,6,6,6,7,7,7,7,7,6,6,6,7,9,10,10)
```

El coeficiente de correlación de las notas sería:

```
> cor(M,F)
```

```
[1] 0.9194978
```

Podemos comprobar que los datos introducidos son correctos, creando la tabla de frecuencias absolutas bidimensional, con la ayuda de la función table. Para ello creamos una tabla (data.frame) con las dos variables M, F, que llamaremos NotasMF:

```
> NotasMF<-data.frame (Fisica=F,Mate=M)
```

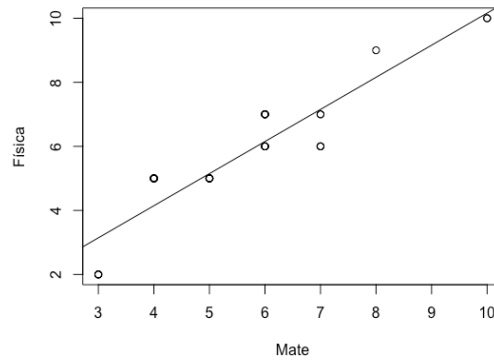
```
> table(NotasMF)
```

```
      Mate
Fisica  3  4  5  6  7  8 10
      2  4  0  0  0  0  0
      5  0  7 11  0  0  0
      6  0  0  0  5  3  0
      7  0  0  0  5  2  0
      9  0  0  0  0  0  1
     10  0  0  0  0  0  0  2
```

Vemos que la tabla coincide con la del enunciado, así que podemos decir que los datos introducidos son correctos.

Podemos ahora dibujar la nube de puntos y la recta de regresión, para ver que efectivamente, las parejas de puntos aparecen relativamente alineadas:

```
> plot(NotasMF$Mate, NotasMF$Fisica, xlab = "Mate", ylab = "Física")  
> regresion <- lm(Fisica ~ Mate, data = NotasMF)  
> abline(regresion)
```



1BA8.- Regresión lineal. Recta de regresión. Estimación.**Predicciones estadísticas y fiabilidad de las mismas.**

Las rectas de regresión lineal son las rectas que mejor se ajustan al diagrama de dispersión, o nube de puntos de una variable bidimensional.

Las ecuaciones de las rectas de regresión se calculan mediante la expresión:

$$\text{Recta de Y sobre X: } y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2} (x - \bar{x})$$

$$\text{Recta de X sobre Y: } x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2} (y - \bar{y})$$

Donde $\frac{\sigma_{xy}}{\sigma_x^2}$ y $\frac{\sigma_{xy}}{\sigma_y^2}$ son los coeficientes de regresión m y m' respectivamente.

Se cumple que $m \cdot m' = r^2$.

El punto de corte de ambas rectas es el (\bar{x}, \bar{y}) , y se denomina centro de gravedad de la distribución.

Estimaciones con las rectas de regresión.

Conociendo los valores de una variable, las rectas de regresión permiten hacer estimaciones o calcular de manera aproximada los valores esperados de la otra variable.

Estas estimaciones serán más fiables cuanto más se aproxime a 1 o -1 el coeficiente de correlación lineal.

Si el coeficiente de correlación lineal está próximo a 0, no tiene sentido tratar de hacer estimaciones mediante las rectas de regresión.

Ejemplo:

Una empresa ha invertido los diez últimos años en publicidad y ha obtenido las ventas que aparecen en la siguiente tabla, expresados en miles de euros.

7,5	8	8,5	10	10,5	12	13	14	15	18
200	205	230	240	250	270	280	300	310	325

Siendo X la variable "Inversión" e Y la variable "Beneficio", calcula:

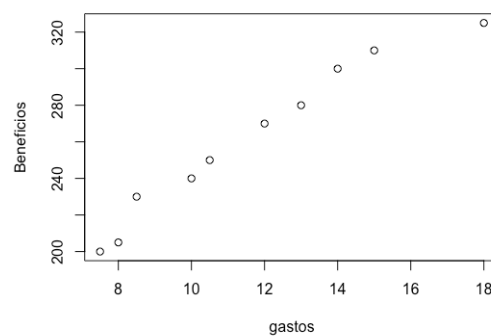
- El coeficiente de correlación lineal. Analiza la dependencia de ambas variables.
- La recta de regresión de Y sobre X.
- El volumen de ventas esperado, si la empresa decidiera invertir el próximo año 25000€ en publicidad.
- La inversión necesaria, si la empresa desea lograr 500000€ de beneficios.

Realizar este Ejemplo a mano, requiere los siguiente pasos:

Necesitamos calcular las desviaciones típicas y la covarianza. Para calcularlos, creamos la siguiente tabla bidimensional:

x_i	x_i^2	y_i	y_i^2	$x_i \cdot y_i$
7,5	56,25	200	40000	1500
8	64	205	42025	1640
8,5	72,25	230	52900	1955
10	100	240	57600	2400
10,5	110,25	250	62500	2625
12	144	270	72900	3240
13	169	280	78400	3640
14	196	300	90000	4200
15	225	310	96100	4650
18	324	325	105625	5850
116,5	1460,75	2610	698050	31700

Dibujamos también la nube de puntos colocando cada pareja de valores sobre un plano X,Y:



$$\bar{x} = \frac{116,5}{10} = 11,65; \sigma_x = \sqrt{\frac{1460,75}{10} - 11,65^2} = 3,22$$

$$\bar{y} = \frac{2610}{10} = 261; \sigma_y = \sqrt{\frac{698050}{10} - 261^2} = 41,04$$

$$\sigma_{xy} = \frac{31700}{10} - 11,65 \cdot 261 = 129,35$$

- a) $r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{129,35}{3,22 \cdot 41,04} = 0,98 \Rightarrow$ Es un coeficiente alto, lo que nos sugiere que existe relación entre ambas variables. Entonces tiene sentido hacer estimaciones.
- b) Recta de regresión Y sobre X: $y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$; $y = 12,49x + 115,44$.
- c) Para una inversión de 25000€, $X = 25$. Sustituyendo, obtenemos un beneficio estimado de 427690€.
- d) Si $y = 500$, lo sustituimos en la recta de regresión de X sobre Y, cuya ecuación es $x = 0,077y - 8,40$. Tras despejar la X, obtenemos una estimación de una inversión de 30100€.

Vamos a ver los pasos que habría que dar en R para hacer el mismo Ejemplo:

Creamos las variables gastos y beneficios, y la tabla Pub_GB con ambas variables:

```
> Gastos<-c(7.5,8,8.5,10,10.5,12,13,14,15,18)
> Beneficios<-c(200,205,230,240,250,270,280,300,310,325)
> Pub_GB<-(data.frame (Gas=Gastos,Ben=Beneficios))
```

Usaremos para estudiar la regresión lineal el comando `lm` (linear models). El primer argumento de este comando es una fórmula $y \sim x$ en la que se especifica cuál es la variable respuesta o dependiente "y" y cuál es la variable regresora o independiente "x".

El segundo argumento, llamado "data" especifica cuál es el fichero en el que se encuentran las variables.

El resultado lo guardamos en un objeto llamado “Reg_Gas_Ben” (regresión gastos beneficios). Este objeto es una lista que contiene toda la información relevante sobre el análisis.

Mediante el comando summary obtenemos un resumen de los principales parámetros estadísticos:

```
> Reg_Gas_Ben<-lm(Beneficios~Gastos, data=Pub_GB)
> summary(Reg_Gas_Ben)
```

Call:

```
lm(formula = Beneficios ~ gastos, data = Pub_GB)
```

Residuals:

Min	1Q	Median	3Q	Max
-15.341	-6.957	2.751	6.514	9.638

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	115.438	10.938	10.55	5.67e-06 ***
gastos	12.495	0.905	13.81	7.32e-07 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.208 on 8 degrees of freedom

Multiple R-squared: 0.9597, Adjusted R-squared: 0.9547

F-statistic: 190.6 on 1 and 8 DF, p-value: 7.315e-07

De este resumen, obtenemos las respuestas al apartado

a) Multiple R-squared: $0.9597 = r^2$

Otra forma:

```
> cor(Gastos, Beneficios)
[1] 0.9796541
```

b) Los coeficientes de la recta de regresión se obtienen de la columna Estimate Std. de la zona de coefficients, de la función summary:

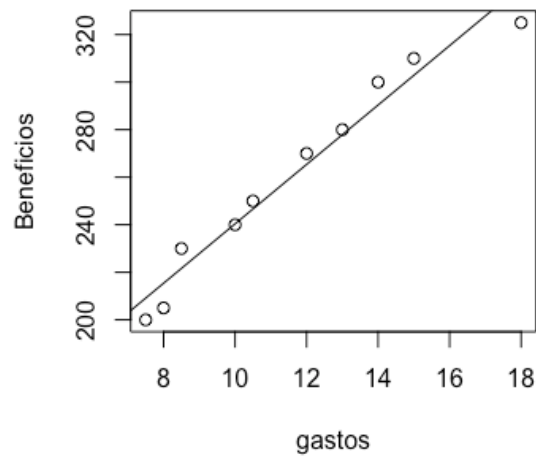
Coefficients:

	Estimate	Std.
(Intercept)	115.438	
gastos	12.495	

$y=12.495x+115.438$. Se puede dibujar mediante el comando:

```
> plot (gastos, Beneficios)
```

```
> abline(Reg_Gas_Ben)
```



c y d se calculan igual

Realmente, tras crear la tabla, y las dos variables, sólo hemos usado las funciones `lm` y `summary`.

```
> Gastos<-c(7.5,8,8.5,10,10.5,12,13,14,15,18);  
> Beneficios<-c(200,205,230,240,250,270,280,300,310,325);  
> Pub_GB<-(data.frame (Gas=Gastos,Ben=Beneficios));  
> Reg_Gas_Ben<-lm(Beneficios~Gastos, data=Pub_GB);  
> summary(Reg_Gas_Ben);
```

Capítulo 5. Estándares de aprendizaje evaluables.

Estándares 1º Bachillerato, Aplicadas

1.1. Elabora tablas bidimensionales de frecuencias a partir de los datos de un estudio estadístico, con variables discretas y continuas.

1.2. Calcula e interpreta los parámetros estadísticos más usuales en variables bidimensionales.

1.3. Calcula las distribuciones marginales y diferentes distribuciones condicionadas a partir de una tabla de contingencia, así como sus parámetros (media, varianza y desviación típica).

1.4. Decide si dos variables estadísticas son o no dependientes a partir de sus distribuciones condicionadas y marginales.

1.5. Usa adecuadamente medios tecnológicos para organizar y analizar datos desde el punto de vista estadístico, calcular parámetros y generar gráficos estadísticos.

2.1. Distingue la dependencia funcional de la dependencia estadística y estima si dos variables son o no estadísticamente dependientes mediante la representación de la nube de puntos.

2.2. Cuantifica el grado y sentido de la dependencia lineal entre dos variables mediante el cálculo e interpretación del coeficiente de correlación lineal.

2.3. Calcula las rectas de regresión de dos variables y obtiene predicciones a partir de ellas.

2.4. Evalúa la fiabilidad de las predicciones obtenidas a partir de la recta de regresión mediante el coeficiente de determinación lineal.

3.1. Describe situaciones relacionadas con la estadística utilizando un vocabulario adecuado..

Capítulo 6. Bibliografía

Software Rstudio:

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

Manual de R:

Título: R para profesionales de los datos: una introducción

Autor: Carlos J. Gil Bellosta

Fecha: 2018-04-22

https://www.datanalytics.com/libro_r/index.html

Tablas de contingencia en R:

https://rstudio-pubs-static.s3.amazonaws.com/34553_55e68158c79140be8d6ff5f60c77e0d1.html#6

Histogramas en R:

<https://www.cs.waikato.ac.nz/~fbravoma/teaching/explora.pdf>

Librerías en R:

<http://ggplot2.tidyverse.org/>

http://rstudio-pubs-static.s3.amazonaws.com/324830_8985f6dac8d34633b6cf23a92ff3e64c.html#sepal.length

Estructuras de datos en R:

http://www.dm.uba.ar/materias/analisis_de_datos/2009/2/practicas/TP2-2009.pdf

Creación de data.frames en R:

<http://r-econ.blogspot.com/2012/07/unir-varios-dataframes-en-un-solo-paso.html>

Correlación lineal en R:

<http://wpd.ugr.es/~bioestad/guia-r-studio/practica-3/>

Curso de introducción a la Estadística:

<http://www.postdata-statistics.com/IntroEstadistica/Tutoriales/Tutorial-00.pdf>

<http://www.postdata-statistics.com/IntroEstadistica/Tutoriales/Tutorial-02.pdf>
<http://www.postdata-statistics.com/IntroEstadistica/Tutoriales/Tutorial-04.pdf>
<http://www.postdata-statistics.com/IntroEstadistica/Tutoriales/Tutorial-05.pdf>
<http://www.postdata-statistics.com/IntroEstadistica/Tutoriales/Tutorial-06.pdf>
<http://www.postdata-statistics.com/IntroEstadistica/Tutoriales/Tutorial-07.pdf>
<http://www.postdata-statistics.com/IntroEstadistica/Tutoriales/Tutorial-08.pdf>
<http://www.postdata-statistics.com/IntroEstadistica/Tutoriales/Tutorial-09.pdf>
<http://www.postdata-statistics.com/IntroEstadistica/Tutoriales/Tutorial-10.pdf>
<http://www.postdata-statistics.com/IntroEstadistica/Tutoriales/Tutorial-11.pdf>

Contenidos y estándares oficiales:

Real Decreto 1105/2014, de 26 de diciembre, por el que se establece el currículo básico de la Educación Secundaria Obligatoria y del Bachillerato.

<https://www.boe.es/boe/dias/2015/01/03/pdfs/BOE-A-2015-37.pdf>

ORDEN EDU/363/2015, de 4 de mayo, por la que se establece el currículo y se regula la implantación, evaluación y desarrollo del bachillerato en la Comunidad de Castilla y León.

<http://bocyl.jcyl.es/boletines/2015/05/08/pdf/BOCYL-D-08052015-5.pdf>

Libros de texto:

Fundamentos y métodos de Estadística, 3ª Edición,:

Autores: M. López Cachero

Editorial: Ed Piramide

ISBN 84-368-0171-7

Estadística aplicada a las ciencias de la educación:

Autores: Joan Welkowitz, Robert B. Ewen, Jacob Cohen

Editorial: Ed Santillana

ISBN: 84-294-1903-9

Matemáticas 4º ESO

Autores: Fernando Alcalde, Joaquín Hernandez...

Editorial: EDICIONES SM

ISBN: 9788467586930

Matemáticas 1º Bachillerato:

Autores: Mª José Ruíz Jiménez, Jesús Llorente Medrano...

Editorial: EDITEX S.A

ISBN: 9788490785652

Matemáticas 2º Bachillerato:

Apuntes marea verde.

<http://apuntesmareaverde.org.es/grupos/mat/>

Apuntes de Complementos de Matemáticas del Máster en profesor de educación secundaria obligatoria y bachillerato, formación profesional y enseñanzas de idiomas

http://campusvirtual2017.uva.es/pluginfile.php/415026/mod_resource/content/1/CM2017-18-Material%20Estad%C3%ADstica-2.pdf

Apuntes de estadística para Ingenieros Técnicos Industriales. Curso 2005, Escuela Universitaria Politécnica de Valladolid.

Otros:

Definiciones de Medidas de centralización y de dispersión:

Wikipedia

Referencia censos bíblicos:

<https://www.bible.com/es/bible/149/NUM.1.RVR1960?parallel=149>

Definiciones. Encuesta, censos:

http://www.ine.es/explica/explica_pasos_primera_encuesta.htm

Imágenes flores iris:

<http://www.lac.inpe.br/~rafael.santos/Docs/R/CAP394/WholeStory-Iris.html>