



**UNIVERSIDAD DE VALLADOLID**

**Dpto. Estadística e I.O.**

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

ANEXO 5:A141 4º ESO Académicas.

**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.....	51
Capítulo 6. Bibliografía .....	53



## 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.

### 4º ESO. Académicas

- 4EAC1.- Identificación de las fases y tareas de un estudio estadístico.
- 4EAC2.- Gráficas estadísticas: Distintos tipos de gráficas. Análisis crítico de tablas y gráficas estadísticas en los medios de comunicación. Detección de falacias.
- 4EAC3.- Medidas de centralización y dispersión: interpretación, análisis y utilización.
- 4EAC4.- Comparación de distribuciones mediante el uso conjunto de medidas de posición y dispersión.
- 4EAC5.- Introducción a la estadística bidimensional. Dependencia estadística y dependencia funcional.
- 4EAC6.- Construcción e interpretación de diagramas de dispersión. Introducción a la correlación.
- 4EAC7.- Utilización de medios informáticos para calcular parámetros, representar variables unidimensionales y representar nubes de puntos.





## 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/>

Choose Your Version of RStudio

RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace. [Learn More about RStudio features.](#)

Product	License	Price	Action
RStudio Desktop	Open Source License	FREE	DOWNLOAD
RStudio Desktop	Commercial License	\$995 per year	BUY
RStudio Server	Open Source License	FREE	DOWNLOAD
RStudio Server Pro	Commercial License	\$9,995 per year	DOWNLOAD
RStudio Server Pro + RStudio Connect	Commercial License	\$29,995 per year	TALK

Integrated Tools for R: ●

Priority Support: ●

Access via Web Browser: ●

Enterprise Security: ●

Project Sharing: ●

Try RStudio Server Pro for free!

FREE

DOWNLOAD

Learn More

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

**RStudio Desktop 1.1.453 — Release Notes**

RStudio requires R 3.0.1+. If you don't already have R, download it [here](#).

Linux users may need to import RStudio's public code-signing key prior to installation, depending on the operating system's security policy.

**Installers for Supported Platforms**

Installers	Size	Date	MD5
RStudio 1.1.453 - Windows Vista/7/8/10	85.8 MB	2018-05-16	b287e385aef53829204023087e98735
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	44cddd285bc31c41e4eac1d74b8eebb

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

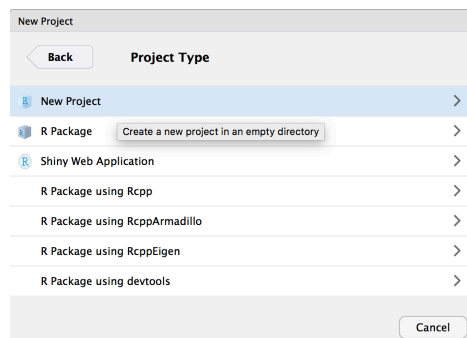
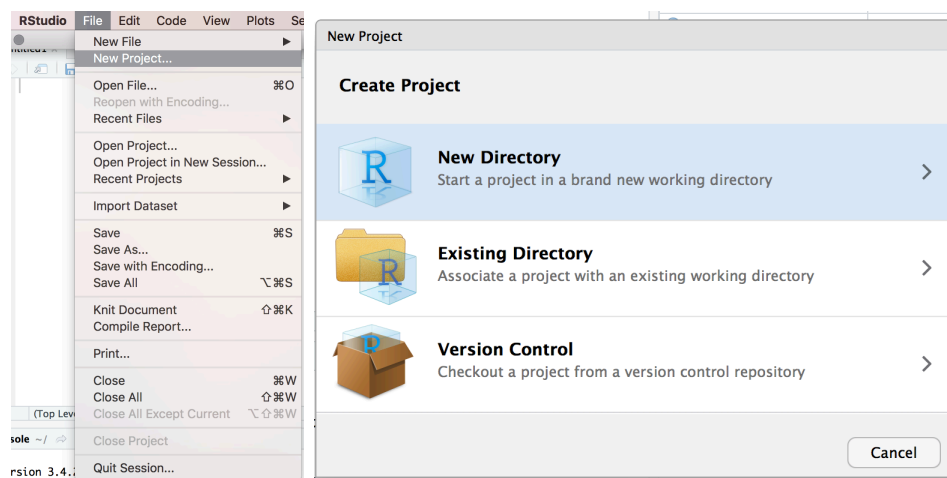
Una vez descargado, se ejecuta el programa instalador, y se van siguiendo los pasos del asistente de instalación, como en cualquier otro programa.

### 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:

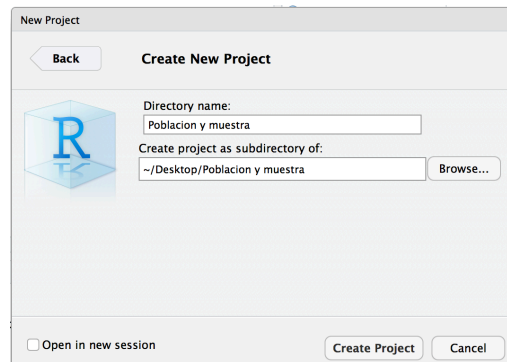
File/New project.../New directory/New Project



En la siguiente ventana, escribimos:

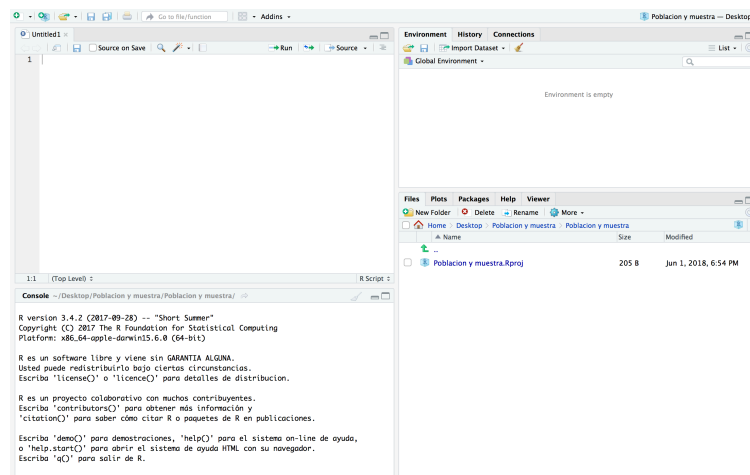
Directory name: Población y muestra.

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 siguiente 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.

#### **4EAC1.- Identificación de las fases y tareas de un estudio estadístico.**

El tratamiento estadístico de un problema comienza siempre con la elección de la magnitud o variable que se quiere estudiar de una determinada población.

Para ello, se elige el método de selección de la muestra, para pasar a la recogida de datos. Una vez obtenidos los datos, se ordenan y presentan en tablas o gráficas, de forma que sean más fáciles de interpretar. Por tanto, podemos decir que un estudio estadístico consta de las siguientes fases y tareas:

1.- Determinación del objeto de estudio. Localización del objeto de estudio. Definición de la población e identificación las características cuantitativas y cualitativas a estudiar, especificando la forma en la que los datos serán recogidos.

2.- Selección de las variables de estudio. Cálculo del tamaño de la muestra y de los recursos para conseguirla.

3.- Recogida de los datos: diseño del cuestionario y diseño muestral.

4.- Organización de los datos: estudio de cada variable, creación de tablas y representación gráfica de la forma más apropiada para favorecer su interpretación.

5.- Representación y tratamiento de los datos.

6.- Interpretación y análisis. Recomendaciones y toma de decisiones a partir de las conclusiones.

Muchas veces los tres primeros puntos nos los dan cuando nos plantean el problema.

**4EAC2.- Gráficas estadísticas: Distintos tipos de gráficas. Análisis crítico de tablas y gráficas estadísticas en los medios de comunicación. Detección de falacias.**

Las gráficas estadísticas permiten representar la información de un estudio estadístico de forma visual. El tipo de gráfico a utilizar se elegirá dependiendo del tipo de variable y de las características a estudiar.

**Diagrama de barras y polígono de frecuencias**

En un diagrama de barras cada valor se representa con una barra cuya altura es proporcional a su frecuencia.

Si se marcan los puntos medios de los extremos superiores de las barras y se unen mediante rectas, se obtiene el polígono de frecuencias.

El diagrama de barras muestra las frecuencias absolutas de los datos. Cuanto más alta es la barra más se da el valor al que corresponde. La altura indica la frecuencia absoluta de la variable.

En "R", se utiliza la función `barplot`.

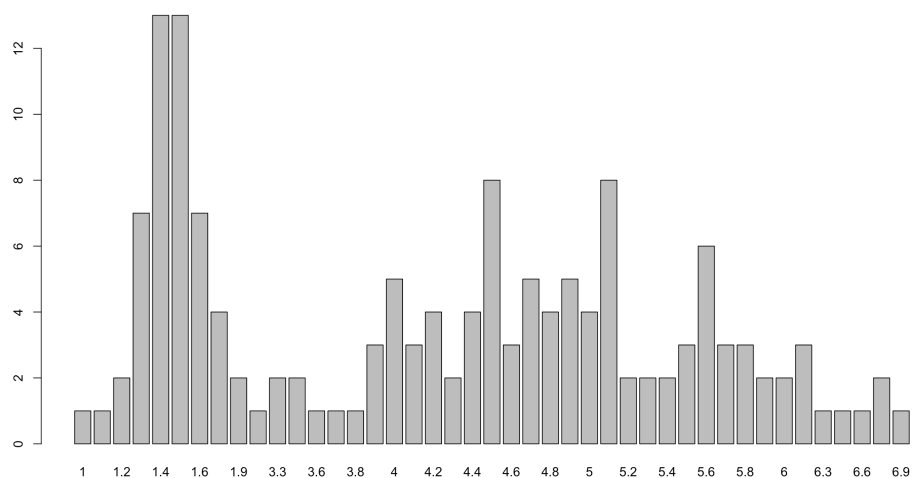
Para dibujar los diagramas de barras y los polígonos de frecuencias en "R", primero creamos la tabla de frecuencias de la variable `iris$Petal.Length`, con la función `table`:

```
> Longpetal<-table(iris$Petal.Length)
```

A continuación dibujamos el diagrama de barras de la variable `Longpetal`

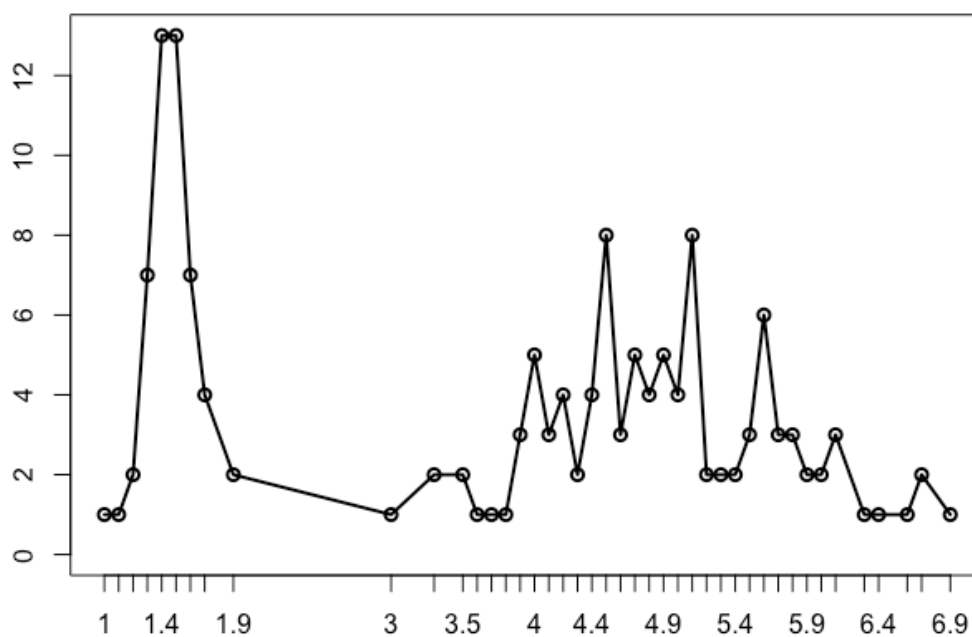
```
> barplot (Longpetal) #Diagrama de barras de la variable Longpetal
```





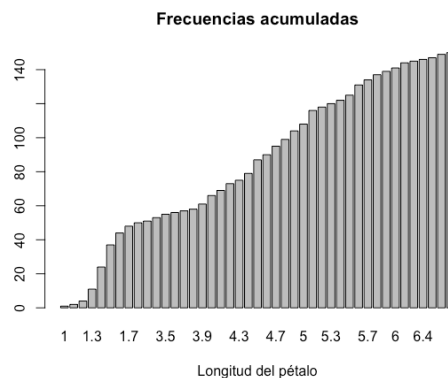
El polígono de frecuencias se dibuja uniendo los centros de las alturas de los rectángulos del diagrama de barras.

```
> plot (Longpetal, type="o") #Polígono de frecuencias de Longpetal
```



Estos diagramas nos sugieren que las longitudes más repetidas están en torno a 1.4, con cantidades de 13 individuos en cada una.

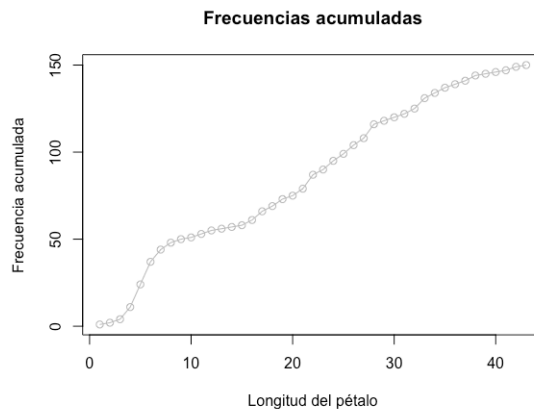
Si quisiéramos hacer lo mismo con las frecuencias acumuladas, usaríamos las funciones cumsum para calcular las frecuencias acumuladas de la variable Longpetal



```
> barplot (cumsum(Longpetal), col="gray", xlab="Longitud del pétalo",
main="Frecuencias acumuladas")
```

Con los comandos plot y lines, y la función cumsum, dibujamos el polígono de frecuencias acumuladas:

```
> plot (cumsum(Longpetal), col="gray", xlab="Longitud del pétalo",
ylab="Frecuencia acumulada", main="Frecuencias acumuladas")
> lines (cumsum(Longpetal), col="gray", xlab="Longitud del pétalo",
ylab="Frecuencia acumulada", main="Frecuencias acumuladas")
```



### Diagrama de sectores

En un diagrama de sectores, la amplitud de cada sector circular representa el valor de la variable:

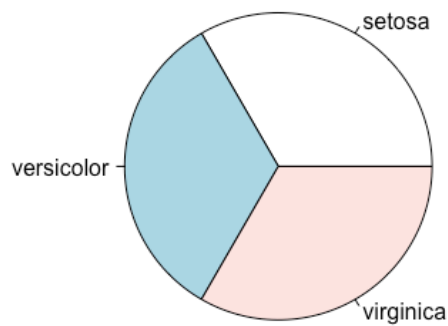
$$\text{Amplitud del sector} = 360^\circ \cdot (f_i/N) = 360^\circ \cdot h_i$$

En "R" se representa con la función pie.

**Ejemplo:**

Representa mediante un diagrama de sectores la distribución por especies de las 150 muestras de plantas de la tabla iris:

```
> pie(table(iris$Species)):
```



Viendo el diagrama podemos pensar que hay las mismas muestras de cada una de las tres especies.

Vamos a tomar una muestra de 20 individuos al azar:

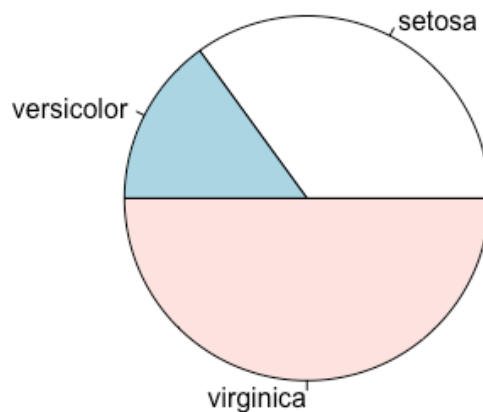
```
> muestra<- sample (1:nrow(iris), size=20,replace=FALSE)
> irismuestra<- iris[muestra, ]
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
97	5.7	2.9	4.2	1.3	versicolor
41	5.0	3.5	1.3	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
137	6.3	3.4	5.6	2.4	virginica
114	5.7	2.5	5.0	2.0	virginica
119	7.7	2.6	6.9	2.3	virginica
113	6.8	3.0	5.5	2.1	virginica
128	6.1	3.0	4.9	1.8	virginica
103	7.1	3.0	5.9	2.1	virginica
44	5.0	3.5	1.6	0.6	setosa
12	4.8	3.4	1.6	0.2	setosa
45	5.1	3.8	1.9	0.4	setosa
109	6.7	2.5	5.8	1.8	virginica
111	6.5	3.2	5.1	2.0	virginica
27	5.0	3.4	1.6	0.4	setosa

144	6.8	3.2	5.9	2.3	virginica
73	6.3	2.5	4.9	1.5	versicolor
77	6.8	2.8	4.8	1.4	versicolor
110	7.2	3.6	6.1	2.5	virginica
33	5.2	4.1	1.5	0.1	setosa

Vamos a ver si se ha conservado constante el número de plantas de cada especie:

`>pie (irismuestra$Species):`



Viendo el diagrama de sectores, vemos que ha salido muy beneficiada la especie virginica, en contra de versicolor.

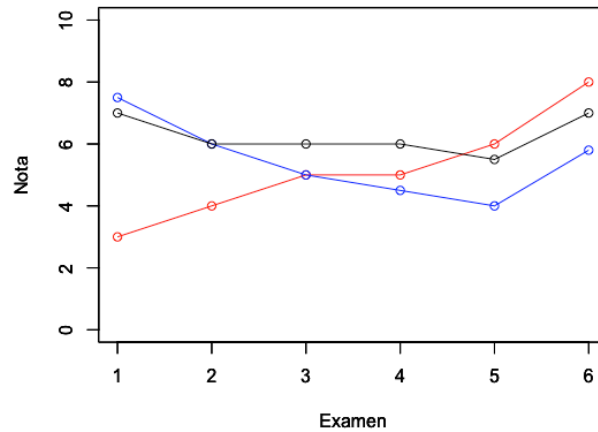
### Diagramas lineales

En los diagramas lineales, cada valor se representa mediante un punto cuya ordenada es proporcional a su frecuencia. A continuación, se unen los puntos mediante segmentos.

Sirven para ver la evolución de una o varias variables estadísticas en distintas fases.

**Ejemplo:**

El gráfico muestra la evolución de las notas de un alumno a lo largo de seis exámenes durante curso en tres asignaturas:



Rojo=Lengua; Azul=Matemáticas; Negro=Física

```
> L<-c(3,4,5,5,6,8)
> M<-c(7.5,6,5,4.5,4,5.8)
> F<-c(7,6,6,6,5.5,7)
> plot(L,xlim=c(1,6),ylim=c(0,10),col="red",xlab="Examen",ylab="Nota",type
="o")
> par(new=T)
> plot(M,xlim=c(1,6),ylim=c(0,10),col="blue",xlab="Examen",ylab="Nota",type
="o")
> par(new=T)
> plot(F,xlim=c(1,6),ylim=c(0,10),col="black",xlab="Examen",ylab="Nota",type
="o")
```

**Histograma y polígono de frecuencias.**

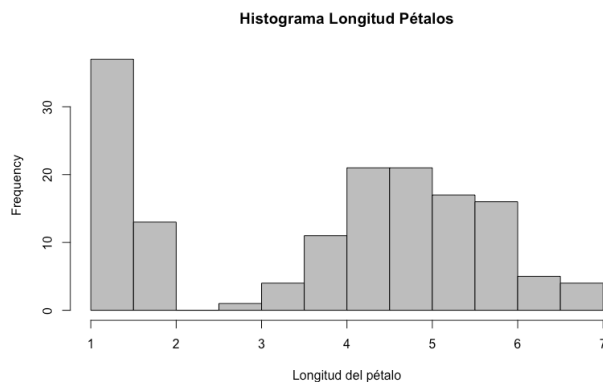
Es la representación gráfica más frecuente para datos agrupados en clases o intervalos. Consiste en un conjunto de rectángulos contruidos de la siguiente forma:

Tiene como eje horizontal una escala de valores de la variable que se mide, sobre la que se marcan los límites de las clases sobre la escala.

Como eje vertical, tiene una escala de frecuencias absolutas o relativas.

La base de los rectángulos es la amplitud del intervalo, y la altura es la frecuencia absoluta.

```
> hist(iris$Petal.Length, breaks=12, col="gray", xlab="Longitud del pétalo",
main="Histograma Longitud Pétalos"):
```

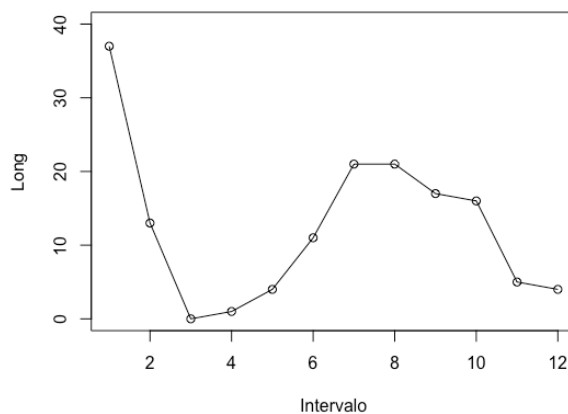


El porígon de frecuencias se construye uniendo los puntos medios de los lados superiores de los rectángulos. Podemos guardar en una variable todos los datos del histograma que acabamos de crear:

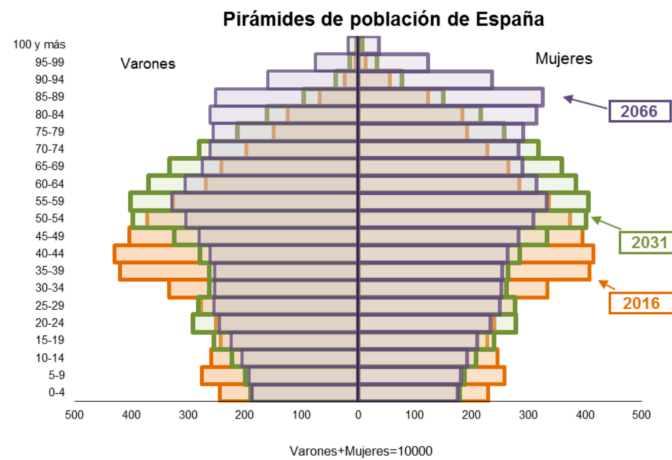
```
> Barras_LP<- (hist(iris$Petal.Length, breaks=12, col="gray",
xlab="Longitud del pétalo", main="Histograma Longitud Pétalos"))
```

Y a continuación, trazar el polígono de frecuencias:

```
> plot (Barras_LP$counts,xlim=c(1, 12),ylim=c(0, 40),
col="black",xlab="Intervalo",ylab="Long",type = "o")
```



Las pirámides de población son histogramas dobles:

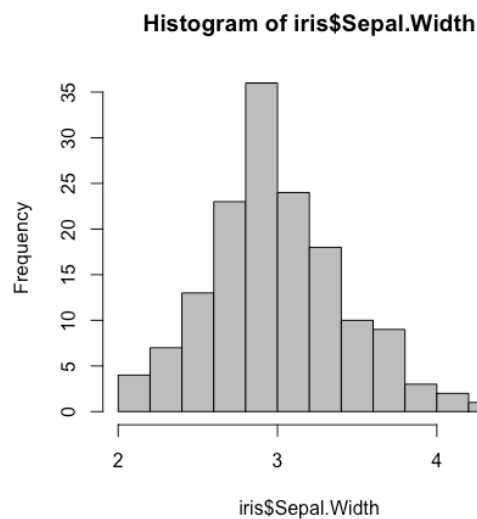


Fuente: <http://www.ine.es/>

### Ejemplo:

Representa mediante un histograma los datos de la anchura de los pétalos de las flores de la tabla iris.

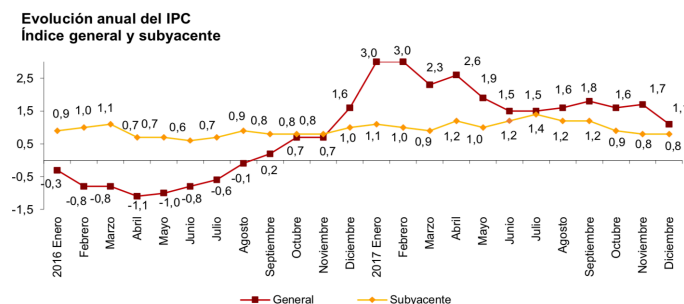
```
> hist(iris$Sepal.Width, col="gray")
```



En "R" podemos establecer el número de intervalos del histograma, con la el parámetro `breaks` dentro de la función `hist`. Si no ponemos nada, el programa elige el mejor valor posible. En este caso, ha elegido automáticamente `breaks=12`, utilizando la regla de  $K = \sqrt{n}$ .

### Análisis crítico de tablas y gráficas estadísticas en los medios de comunicación (4º ESO).

Los medios de comunicación nos muestran constantemente datos en forma de tablas y gráficas para facilitar la interpretación de los mismos por parte de los espectadores.



Fuente: <http://www.ine.es/>

Con cierta frecuencia ocurre que con unos mismos datos podemos encontrar opiniones distintas.

La asistencia a una manifestación, las inversiones en sanidad o en educación, parecerán distintas en función de quién nos muestre la información.

En este sentido, la estadística nos ayuda a desarrollar nuestro pensamiento crítico, y a detectar estas diferencias.

En multitud de ocasiones, los medios de comunicación recurren a malas prácticas para esconder la realidad. Estos son algunos ejemplos de malas prácticas:

Cuestionarios mal planteados:

Son numerosos los casos en los que los cuestionarios tienen las posibles respuestas dirigidas en un sentido u otro, impidiendo a la población responder libremente. Los resultados que se obtienen de estos estudios estarán sesgados.

Por ejemplo, si preguntamos por nuestra fruta preferida y sólo ofrecemos como repuestas: “naranja”, “pera” o “manzana”, no se verán representados los que les guste la sandía.

Errores en la obtención de datos:



Puesto que los instrumentos de medida no tienen precisión absoluta, en ocasiones los medios de comunicación redondean o truncan hacia arriba o hacia abajo según interese.

Delimitación imprecisa de la población:

Si se desea estudiar si los niños vallisoletanos hacen poco deporte, habrá que dejar claro qué edades en concreto se considerarán, si entendemos por vallisoletano a todos los nacidos en Valladolid, o sólo a los que están empadronados en Valladolid, o sólo a los que viven habitualmente en Valladolid...

Selección de la muestra inapropiada o no representativa:

La muestra en ocasiones no representa a la población, sobre todo cuando la elección de los individuos de la muestra no se hace de forma aleatoria. Por ejemplo, si queremos estudiar los grupos de música favoritos de los alumnos del instituto, tendremos que seleccionar muestras de edades variadas, y en las proporciones que aparecen en el instituto. No estaría bien sólo coger la muestra de los alumnos de 2º de la ESO.

Errores en las gráficas:

Muchas veces se presentan los diagramas de barras truncados, sin representar el origen, o con las escalas en los ejes distintas. Hay que dejar claras las variables que se miden.

### **Ejemplos de falacias en los medios de comunicación:**

*"Se calcula que en España hay un millón de cazadores que cada temporada realizan 250.000 millones de disparos". ¿Es posible?*

*"El día siguiente a la muerte de la cantante Lola Flores, una emisora afirmó que la capilla ardiente, instalada a las cuatro de la tarde del día anterior, había recibido más de 500.000 personas".* Según esta cifra, al haber transcurrido un total de 16 horas (57.600 segundos), los visitantes desfilaban ante el féretro de Lola Flores a una velocidad de nueve personas por segundo. ¿Improbable, no?.

Otro de los ejemplos más comunes son los grados. En ocasiones los periodistas aseguran que alguien ha dado un *"giro de 360º"*; y eso significa que esa persona ha completado una vuelta entera para quedarse en el mismo lugar. Lo correcto sería decir que alguien ha dado *"un giro de 180º"* en su vida. Por no

hablar de cuando utilizan en una noticia números larguísimos o cambian el sentido de la noticia al cambiar una cifra.

Fuente: [Apuntes Metodología y Evaluación, UVA 2018](#)

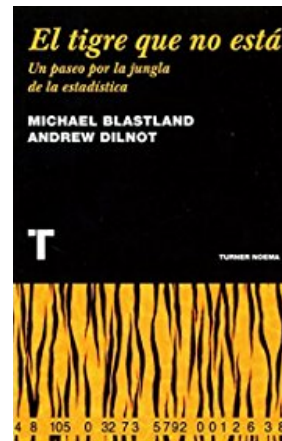
Máster Universitario de Profesor en Educación Secundaria Obligatoria y Bachillerato.

Especialidad de Matemáticas

Lecturas recomendados:



Darrell Huff



“How to lie with statistics (Darrell Huff)”: Lo que estas páginas, escritas con ingenio y humor, nos ofrecen es, en realidad, un curso de sentido común para aprender a descubrir los ardides con los que cada día pretenden engañarnos, manipulando cifras y gráficas, los medios de comunicación, los políticos, la publicidad... Lo que aquí se nos cuenta resulta divertido; pero es bueno tomarlo en serio, porque, como nos dice el autor, “los desaprensivos ya conocen estos trucos; los hombres honrados deben aprenderlos en defensa propia”

Fuente: [Biblioteca UVA](#)

“El tigre que no está. Un paseo por la jungla de la estadística”. En su contraportada, puede leerse: Los diarios, los políticos, la televisión y los guardianes de la salud nos bombardean diariamente con cifras y porcentajes: “Invertiremos dos mil billones en Educación”, “Se han mejorado los tiempos de espera en un catorce por ciento”, “Dos de cada tres adolescentes sufren depresión”, “Cada bebida alcohólica aumenta el riesgo de cáncer un diez por ciento”.

Blastland y Dilnot, armados con un incombustible humor, algo de conocimiento estadístico y una buena carga de sentido común, nos introducen en el arte de “ver más allá” de las cifras: en palabras de los autores, tras leerlo el

lector podrá distinguir entre un grupo de rayas y un tigre. Este libro breve, divertidísimo y muy esclarecedor, fue para la revista

The Economist

Uno de sus libros del año, y ha conocido un resonante éxito en sus versiones británica y norteamericana.

Fuente: [DivulgaMat](#)

**4EAC3.- Medidas de centralización y dispersión: interpretación, análisis y utilización.**

Normalmente interesa resumir la información de una muestra en un solo valor, para hacernos una idea de cómo se comporta la variable y así poder realizar comparaciones.

Las medidas de tendencia central más habituales son la media, la mediana y la moda.

Medidas de centralización:

**Media aritmética  $\bar{X}$**  : La media aritmética es el valor que se obtiene al sumar todos los individuos de la muestra, y al dividir esta suma entre el número de individuos. Es la medida de tendencia central que más se utiliza.

Tanto las empresas, como los países, como los medios de comunicación, constantemente hablan de medias de datos, como el gasto medio, el salario medio, o la altura media.

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

En "R": `mean(x)`

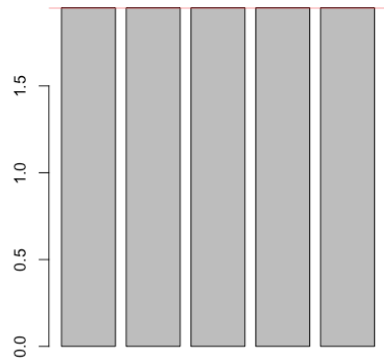
**Interpretación y Propiedades:**

Dos poblaciones totalmente distintas pueden tener la misma media.

Si medimos por ejemplo a los 5 jugadores de un equipo de baloncesto y obtenemos que todos miden 1.95 m, dicho equipo tendría una estatura media de 1.95 m. Este valor representa adecuadamente a esta población, porque todos los datos están muy próximos a la media.

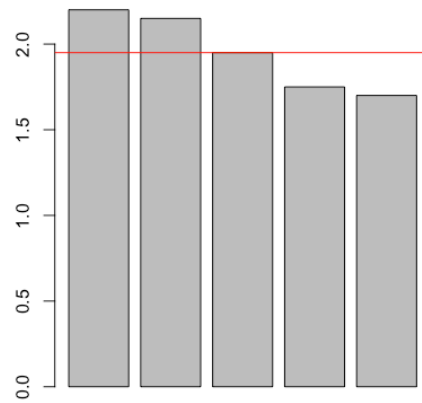
Si medimos, un segundo equipo de jugadores y obtenemos que miden 2.20m, 2.15m, 1.95m, 1.75m y 1.70m, éste segundo equipo tendría una estatura media de 1.95 m. Pero en este caso, este valor representa ninguno de sus componentes.

Altura jugadores Eq1:



```
> Eq1<-c(1.95,1.95,1.95,1.95,1.95)
> barplot (Eq1)
> mean(Eq1)
[1] 1.95
> abline (h=mean (Eq1), col="red")
```

Altura jugadores Eq 2:



```
> Eq2<-c(2.20,2.15,1.95,1.75,1.75)
> barplot (Eq2)
> mean(Eq2)
[1] 1.95
> abline (h=mean (Eq2), col="red")
```

### Ejemplos:

Calcular la media, la mediana y la moda de las notas de la primera evaluación de la tabla Población.xlsx

Media:

```
> mean(Población$EV_P)
[1] 4.554526
```

Mediana:

```
> median(Población$EV_P)
[1] 4.862917
```

Moda:

Calculamos las frecuencias absolutas de las notas de la primera evaluación:

```
> FA_EV_P<-table(Población$EV_P)
> FA_EV_P
```

```

0.7 0.95 1.06 1.75 1.97 1.99 3.5 3.58 4.71 4.8 4.81 4.86 5 5.03 5.28 5.52
6.4 6.62
1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1
6.8 7.53 7.72 9.18
1 1 1 1

```

Como aparecen desordenadas, podemos ordenarlas de mayor a menor para que nos resulte más cómodo encontrar la moda:

```

> FA_EV_P<-sort(FA_EV_P, decreasing = TRUE)
> FA_EV_P
5 0.7 0.95 1.06 1.75 1.97 1.99 3.5 3.58 4.71 4.8 4.81 4.86 5.03 5.28 5.52
2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
6.4 6.62 6.8 7.53 7.72 9.18
1 1 1 1 1 1

```

La moda sería el valor de 5, que se repite dos veces.

**Media Ponderada:** La media ponderada (MP) de una muestra, se calcula asignando a cada observación unos pesos que indicarán la importancia que tiene cada uno de los valores observados.

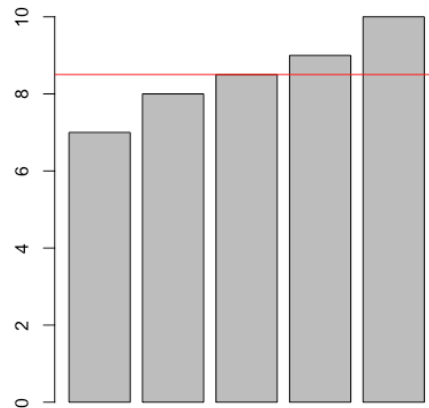
**Media Geométrica:** La media geométrica de una muestra se obtiene al multiplicarlos entre si y aplicarles la n-ésima raíz.

Para el cálculo de la media aritmética se suman los valores para luego dividirlos por el número de valores. En esta ocasión se multiplican para posteriormente aplicar la n-ésima raíz.

La media geométrica implica que no existan números negativos, o si existen que sean impares, puesto que las raíces de los números negativos pueden no existir en el conjunto de los números reales.

**Mediana, Me:** es el valor que se encuentra en el centro, una vez ordenados los datos.

En R: median(x)



```
> Mediana<-c(7,8,8.5,9,10)
> barplot (Mediana)
> median(Mediana)
[1] 8.5
> abline (h=median(Mediana),col="red")
```

### Interpretación y Propiedades

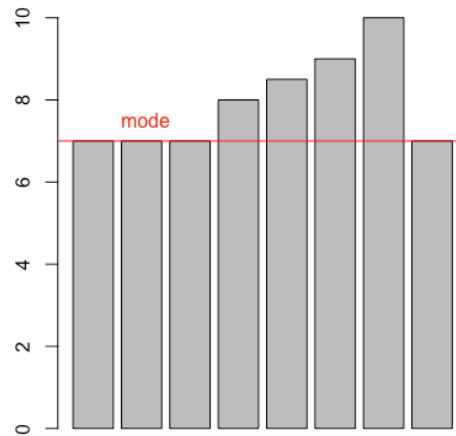
Su uso es adecuado en distribuciones asimétricas, puesto que no se ve afectada por las observaciones extremas, al no depender de los valores que toma la variable.

En el caso de la mediana, siempre tomará un valor de la variable que estudiamos, cosa que no ocurría con la media.

**Moda, Mo:** La moda es la modalidad que más se repite.

En "R" calculamos la tabla de frecuencias absolutas de las notas, y buscamos el mayor valor, ordenando la tabla con el comando sort:

```
FA_EV_P<-sort(FA_EV_P, decreasing = TRUE)
```

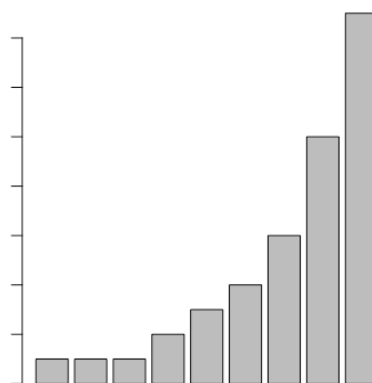


```
> Ej2<-c(7,7,7,8,8.5,9,10,7)
> barplot (Ej2)
> abline (h=7, col="red")
> text(2, 7.5, "mode", col = "red")
```

### Interpretación y Propiedades

En función de los datos a estudiar, unas medidas pueden ser más representativas que otras.

**Ejemplo:** Gráfico de la altura de una población de 9 árboles.



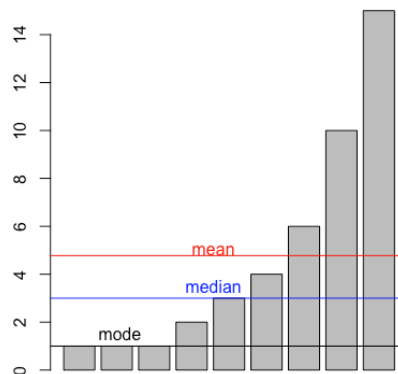
```
> Ej1<-c(1,1,1,2,3,4,6,10,15)
> barplot (Ej1)

> mean (Ej1)
```



```
[1] 4.777778
> median (Ej1)
[1] 3
> table(Ej1)
Ej1
 1  2  3  4  6 10 15
 3  1  1  1  1  1  1
```

La media sería 4,77778, la mediana 3, y la moda 1.



```
> abline (h=mean (Ej1), col="red")
> abline (h=median (Ej1), col="blue")
> abline (h=1, col="black")
> text(5, 5, "mean", col = "red")
> text(5, 3.5, "median", col = "blue")
> text(2, 1.5, "mode", col = "black")
```

¿Cuál de las tres medidas de posición te parece que representa mejor a esta población?

Para conocer la altura media de la población, la media sería más representativa que la moda.

Si nos fijamos en la mediana, podemos ver que hay un 50% de la población con unos valores muy superiores al otro 50%. Si en lugar de hablar de alturas, estuviéramos hablando de sueldos, podríamos pensar que hay mucha desigualdad en esta población, y que abunda la pobreza.

Si nos fijásemos en la moda y analizáramos el número de suspensos de una población, podríamos decir que en esa población suspenden pocas asignaturas. Sin embargo, estaríamos pasando por alto una cantidad muy alta de suspensos repartida en el resto de la población.

Es importante reflexionar qué medida nos viene mejor. Esta reflexión en ocasiones la utilizan los medios de comunicación para informar de manera inresada. Si por ejemplo, fuéramos el ministro de educación, y quisiéramos vender la imagen de que en nuestro país el nivel de suspensos es muy bajo, gracias a la buena gestión del gobierno, nos tendríamos que apoyar en la moda como medida de tendencia central.

Al contrario, si quisiéramos que el estado invirtiese más en educación, ofreceríamos un dato de suspensos más alto, como puede ser la media.

Los cuartiles (Qi): Los cuartiles son los valores que dividen los datos en 4 partes iguales, es decir, en cada tramo está el 25 % de los datos recogidos en el estudio.

$$\begin{array}{cccc} 25\% & 25\% & 25\% & 25\% \\ Q1 & Q2 & Q3 & \end{array}$$

### Medidas de dispersión:

Las medidas de dispersión son una serie de valores que nos informan cómo se encuentran los datos de agrupados o desagrupados.

**Desviación media:** mide la distancia media que hay entre todos los valores de la muestra y el valor medio.

$$S = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})$$

Interpretación: si dos muestras tienen desviaciones medias diferentes, podremos interpretar que la muestra con menor desviación media, contiene los datos más concentrados hacia su valor medio.

**Rango o recorrido:** mide la diferencia entre el valor mayor y el valor menor de la muestra.

En "R": range (x)

Interpretación: cuanto mayor sea el rango más dispersos estarán los valores.

**Varianza:** Se utiliza para medir la dispersión de una variable con respecto a su media. A mayor varianza, mayor dispersión.

Se calcula como suma de las diferencias al cuadrado de cada valor respecto a la media de la muestra. Esta suma se divide entre el número de datos. La varianza se suele representar con la letra V, o  $S^2$ .

$$V = S^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2 \cdot f_i}{N} = \frac{\sum_{i=1}^N x_i^2 \cdot f_i}{N} - \bar{x}^2$$

En "R": var (x)

Interpretación: cuanto mayor sea la varianza más dispersos estarán los valores, y por lo tanto, menos representativa será la media.

Cuando una variable se expresa en una unidad, su varianza se expresa en dicha unidad al cuadrado. Por ejemplo, si trabajamos con distancias en Km, la varianza se medirá en  $\text{Km}^2$

**Desviación típica o desviación estándar:** La desviación típica es otra medida de dispersión y se calcula como raíz cuadrada de la varianza. Es la medida de dispersión que más se utiliza:

$$S = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2 \cdot f_i}{N}} = \sqrt{\frac{\sum_{i=1}^N x_i^2 \cdot f_i}{N} - \bar{x}^2}$$

En "R": sd (x)

Interpretación: la desviación estándar de un grupo repetido de medidas da una idea de su variabilidad respecto de la media. Puede ser interpretada como una medida de incertidumbre.

**Coeficiente de variación:** El coeficiente de variación se define mediante la expresión:

$$g = \frac{S}{|\bar{x}|}; \quad \bar{x} \neq 0$$

Permite comparar la variabilidad de dos o más muestras, independientemente de sus unidades de medida, al cancelarse éstas en la división.

**Recorrido intercuartílico o intervalo intercuartil** es la distancia entre el tercer y el primer cuartil:

$$R = \text{Recorrido intercuartílico} = Q3 - Q1$$

En "R" la función `quantile(x)`, nos devuelve los cuartiles

$Q0=\text{mín}$ ,  $Q1=25\%$ ,  $Q2=50\%$ ,  $Q3=75\%$ , y  $Q4=\text{Máx}$

```
> notas=c(7.4,5.6,7.2,4.0,8.2,7.6,7.2,8.7,8.1,5.0, 6.5, 6.2)
```

```
> quantile(notas)
```

```
0%   25%   50%   75%  100%
```

```
4.000 6.050 7.200 7.725 8.700
```

```
Q1=6.050
```

```
Q3=7.725
```

#### 4EAC4.- Comparación de distribuciones mediante el uso conjunto de medidas de posición y dispersión.

Las medidas de posición y de dispersión nos aportan información sobre cómo se distribuyen los individuos dentro de nuestra muestra

Por ejemplo, tres poblaciones (16, 0, 16, 0), (16, 6, 10, 0) y (9, 7, 7, 9) tienen como media un valor de 8, pero sus desviaciones estándar poblacionales son 9.2, 6.7 y 1.15, respectivamente.

<code>&gt; p1&lt;-c(16, 0, 16, 0)</code>	<code>&gt; p2&lt;-c(16, 6, 10, 0)</code>	<code>&gt; p3&lt;-c(9, 7, 7, 9)</code>
<code>&gt; mean (p1)</code>	<code>&gt; mean (p2)</code>	<code>&gt; mean (p3)</code>
<code>[1] 8</code>	<code>[1] 8</code>	<code>[1] 8</code>
<code>sd(p1)</code>	<code>&gt; sd(p2)</code>	<code>&gt; sd(p3)</code>
<code>[1] 9.237604</code>	<code>[1] 6.733003</code>	<code>[1] 1.154701</code>

Si nos fijamos, vemos que la tercera población tiene una desviación mucho menor que las otras dos porque sus valores están más cerca de 8. Esto podemos observarlo dibujando las nubes de puntos:

Nube p1:

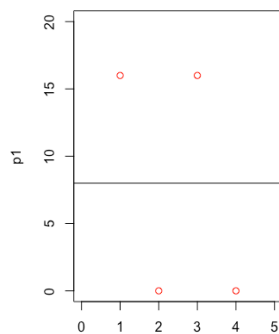
```
> plot( p1, col="red",xlim=c(0, 5),ylim=c(0, 20))
> abline(h=8)
```

Nube p2:

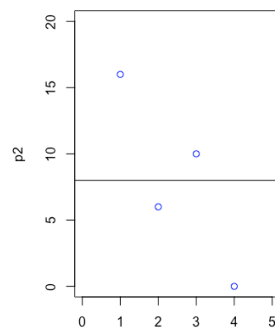
```
> plot( p2, col="blue",xlim=c(0, 5),ylim=c(0, 20))
> abline(h=8)
```

Nube p3:

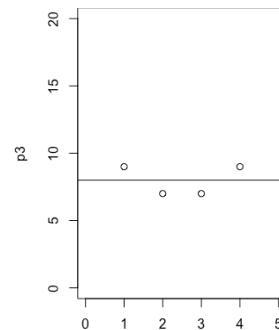
```
> plot( p3, col="black",xlim=c(0, 5),ylim=c(0, 20))
> abline(h=8)
```



Nube p1



Nube p2



Nube p3

Observa que la desviación típica representa la distancia de los valores de la variable a la media.

Utilizando la función `summary` obtenemos la información de las medidas de posición más interesantes:

```
> summary (notas)
```

Min	1st Qu.	Median	Mean	3rd Qu.	Max.
4.000	6.050	7.200	6.808	7.725	8.700

### Ejemplos:

**Desviación media:** se calcula con la función `sd(x)`

**Rango:** se calcula con la función `range(x)`

**Varianza:** se calcula con la función `var(x)`

**Desviación típica:** se calcula con la función `sd(x)`

1.- Calcular la Desviación media, el rango, la varianza y la desviación típica de las notas de la primera evaluación, de la tabla `Población.xls`

```
> sd(Población$EV_P)
[1] 2.321369
> range(Población$EV_P)
[1] 0.70 9.18
> var(Población$EV_P)
[1] 5.388753
```

En este ejemplo los datos aparecen muy dispersos. Hay muchos suspensos, muchos aprobados cerca del 5, y también hay buenas notas. Por eso la desviación media es alta.

El rango nos da los extremos de las notas mínima y máxima.

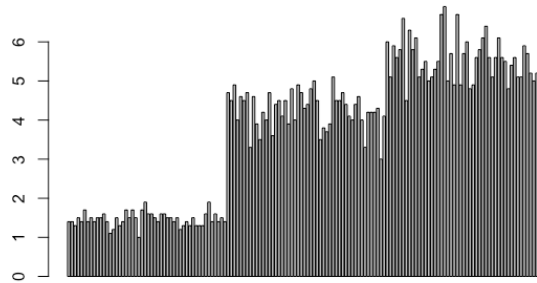
2.- Calcular las medias de las longitudes de los pétalos de la tabla `iris` que contiene la muestra de 150 plantas de 3 especies diferentes:

Cargar la tabla de datos interna de “R” que se llama `iris`

```
> iris
```

Dibujamos el diagrama de barras de la longitud de los pétalos

```
> barplot (iris$Petal.Length)
```



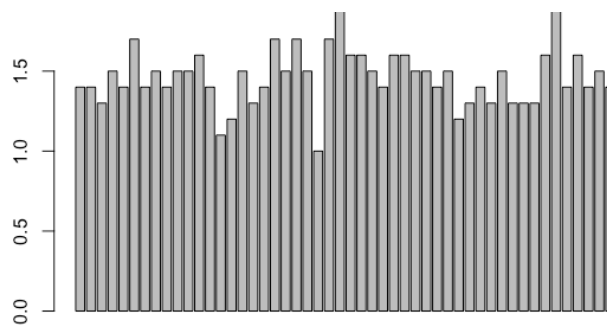
En el gráfico se aprecian tres escalones. Si nos fijamos en la tabla, las 50 primeras muestras son plantas de la especie Setosa, las 50 siguientes de la especie versicolor y las 50 últimas de la especie virgínica.

Podemos decir, por tanto, que la primera especie tiene de media los pétalos más cortos que las otras dos especies.

Para hacer un estudio más riguroso de estas plantas, podría interesar estudiarlas por separado.

Podemos analizar la longitud de las plantas por separado, limitando la tabla a la especie setosa, con la instrucción: `iris[iris$Species == "setosa",]`.

```
> barplot (iris[iris$Species == "setosa",]$Petal.Length)
```



Igualmente, podemos extraer la media de la longitud de los pétalos de todas estas plantas:

```
> mean (iris[iris$Species == "setosa",]$Petal.Length)
[1] 1.462
```

Podemos hacer lo mismo para las otras especies:

```
> mean (iris[iris$Species == "versicolor",]$Petal.Length)
[1] 4.26
> mean (iris[iris$Species == "virginica",]$Petal.Length)
[1] 5.552
```

O para la anchura de los pétalos:

```
> mean (iris[iris$Species == "setosa",]$Petal.Width)
[1] 0.246
> mean (iris[iris$Species == "versicolor",]$Petal.Width)
[1] 1.326
> mean (iris[iris$Species == "virginica",]$Petal.Width)
[1] 2.026
```

En este caso, la anchura de los pétalos se comporta de manera similar a las longitudes.

Si calculamos las medianas:

```
> median (iris[iris$Species == "setosa",]$Petal.Width)
[1] 0.2
> median (iris[iris$Species == "versicolor",]$Petal.Width)
[1] 1.3
> median (iris[iris$Species == "virginica",]$Petal.Width)
[1] 2
```

Vemos que hay algo de diferencia con respecto a las medias, pero no mucha.

Podemos obtener más información de las variables con la función `summary`, que nos devuelve los valores mínimo, máximo, la media, la mediana y los cuartiles.

```
> summary(iris[iris$Species == "setosa",]$Petal.Width)
  Min. 1st Qu.  Median    Mean 3rd Qu.   Max.
0.100  0.200   0.200   0.246  0.300   0.600
```



#### 4EAC5.- Introducción a la estadística bidimensional. Dependencia estadística y dependencia funcional.

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.

### Dependencia estadística y dependencia funcional. Dependencia lineal:

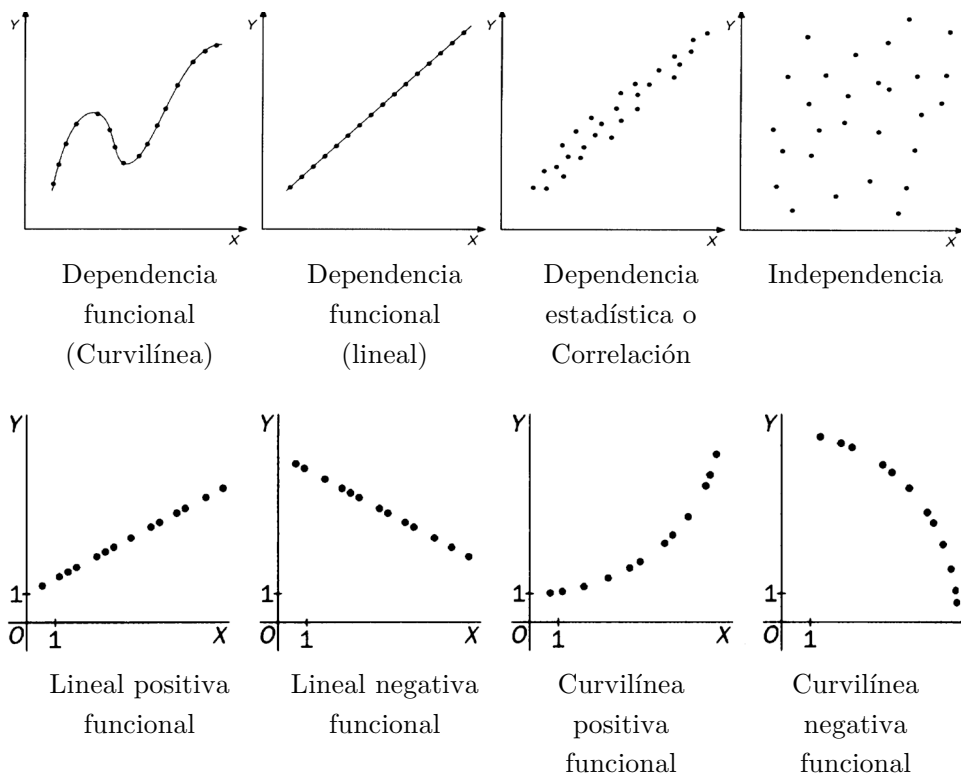
La forma de la nube de puntos, nos permite intuir si hay relación, o dependencia las dos variables. Esta dependencia, si existe, se llama correlación.

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.



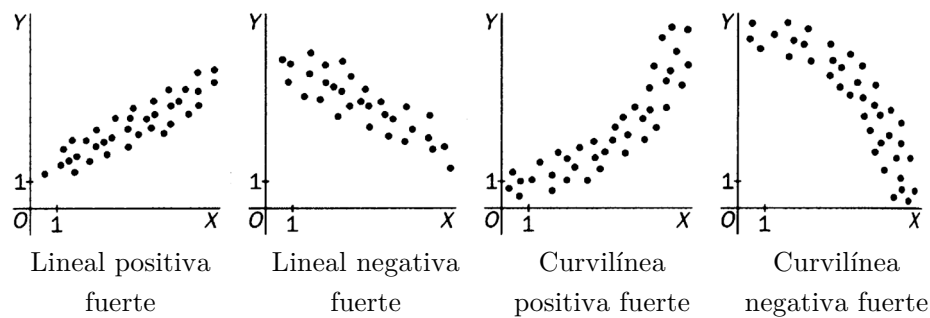
El grado de correlación, a su vez, puede ser:

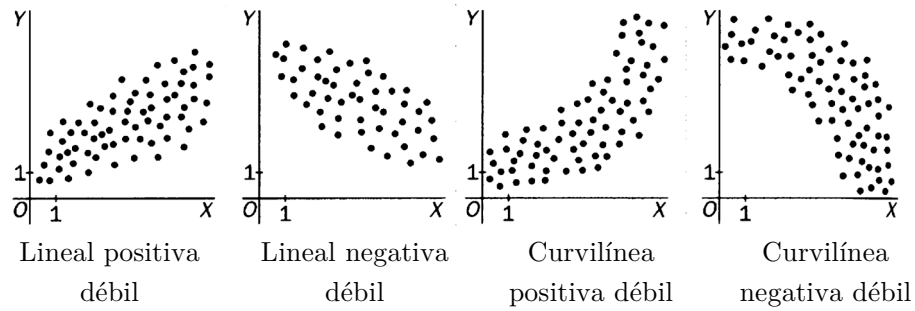
Correlación fuerte, si la nube de puntos se aproxima a una recta o una curva.

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.





### 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:

$Y \setminus X$	$x_1$	$x_2$	...	$x_i$	...	$x_n$	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.

$Y \setminus X$	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.

a) Construye la tabla estadística bidimensional simple correspondiente.

b) 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

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

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

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

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

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

#### 4EAC6.- Construcción e interpretación de diagramas de dispersión. Introducción a la correlación.

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:

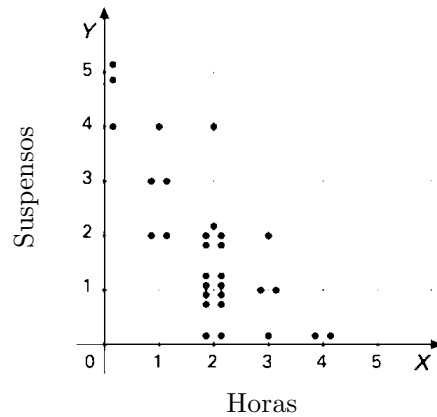
Y\X	0	1	2	3	4	Total
0	0	0	2	1	2	5
1	0	0	8	2	0	10
2	0	2	5	1	0	8
3	0	2	0	0	0	2
4	1	1	1	0	0	3
5	2	0	0	0	0	2
Total	3	5	4	4	2	30

Tablas de distribuc. marginales:

xi	0	1	2	3	4	
fi	3	5	16	4	2	30

yi	0	1	2	3	4	5
fi	5	10	8	2	3	2

Para crear el diagrama de dispersión, trazáramos 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 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
```

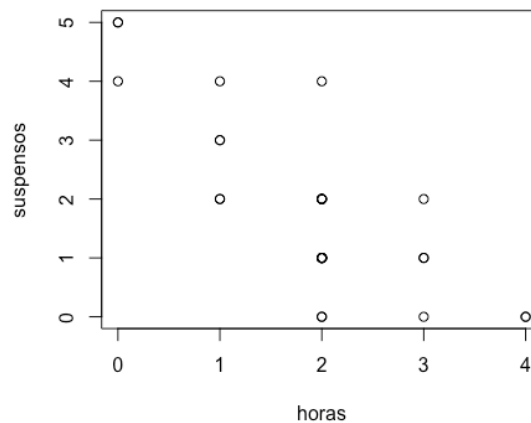


Las tablas de frecuencias también son muy sencillas de construir:

```
> 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)
```



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.

$\sigma_x$  y  $\sigma_y$  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.

## Capítulo 5. Estándares de aprendizaje evaluables.

### Estándares 4º ESO. Académicas

4.1. Interpreta críticamente datos de tablas y gráficos estadísticos.

4.2. Representa datos mediante tablas y gráficos estadísticos utilizando los medios tecnológicos más adecuados.

4.3. Calcula e interpreta los parámetros estadísticos de una distribución de datos utilizando los medios más adecuados (lápiz y papel, calculadora u ordenador).

4.4. Selecciona una muestra aleatoria y valora la representatividad de la misma en muestras muy pequeñas.

4.5. Representa diagramas de dispersión e interpreta la relación existente entre las variables.

.



## 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](https://www.datanalytics.com/libro_r/index.html)

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](http://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](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>

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/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.

<http://bocyl.jcyl.es/boletines/2015/05/08/pdf/BOCYL-D-08052015-4.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 3º ESO:

Apuntes marea verde.

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

Matemáticas 4º ESO

Autores: Fernando Alcalde, Joaquín Hernandez...

Editorial: EDICIONES SM

ISBN: 9788467586930

Matemáticas 1º Bachillerato:

Autores: M<sup>a</sup> José Ruíz Jiménez, Jesús Llorente Medrano...

Editorial: EDITEX S.A

ISBN: 9788490785652

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](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](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>

Gráfica de ejemplo padrón (pirámide de población):  
Instituto nacional de estadística  
<http://www.ine.es/prensa/np994.pdf>

Gráfica de ejemplo de Análisis crítico de tablas y gráficas estadísticas en los medios de comunicación (4º ESO). Inflación:  
Instituto nacional de estadística  
<http://www.ine.es/daco/daco42/daco421/ipc1217.pdf>