



UNIVERSIDAD DE VALLADOLID

Dpto. Estadística e I.O.

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

ANEXO 3:A131 3º 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.....	43
Capítulo 6. Bibliografía.	45

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.

3º ESO. Académicas

- 3EAC1.- Fases y tareas de un estudio estadístico.
- 3EAC2.- Población, individuo, muestra. Variables estadísticas.
- 3EAC3.- Variables estadísticas: cualitativas, cuantitativas, discretas y continuas.
- 3EAC4.- Métodos de selección de una muestra estadística. Representatividad de una muestra.
- 3EAC5.- Frecuencias absolutas, relativas y acumuladas. Agrupación de datos en intervalos.
- 3EAC6.- Gráficas estadísticas.
- 3EAC7.- Parámetros de posición central (media, moda y mediana) y no central (primer y tercer cuartil).
- 3EAC8.- Parámetros de dispersión (rango, recorrido intercuartílico, varianza, desviación típica y coeficiente de variación).
- 3EAC9.- Diagrama de caja y bigotes.
- 3EAC10.- Interpretación conjunta de la media y la desviación típica.
- 3EAC11.- Utilización de los medios tecnológicos adecuados, para el análisis y la producción de información estadística.

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 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	44cddd285bc31c41e4eaeed74b8eebb

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

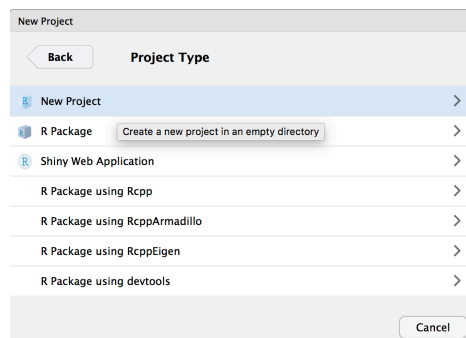
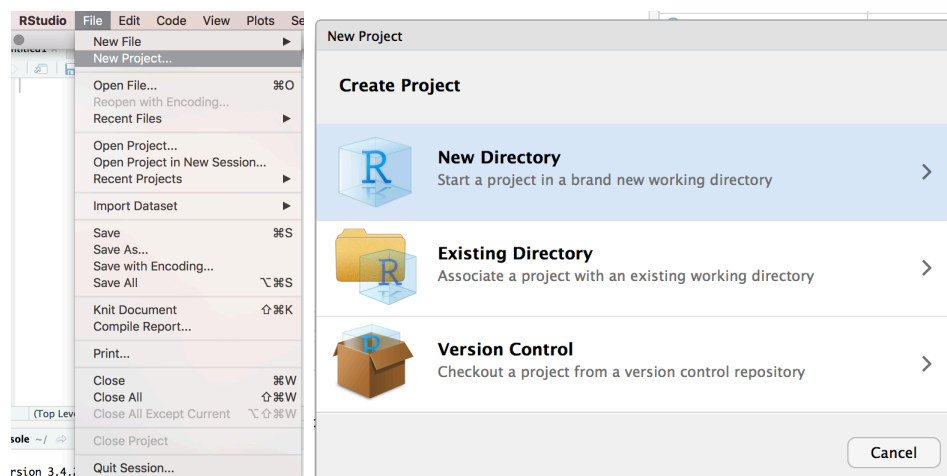
Una vez descargado, se ejecuta el programa instalador, y se van siguiendo los pasos del asistente de instalación, como en cualquier otro programa.

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:

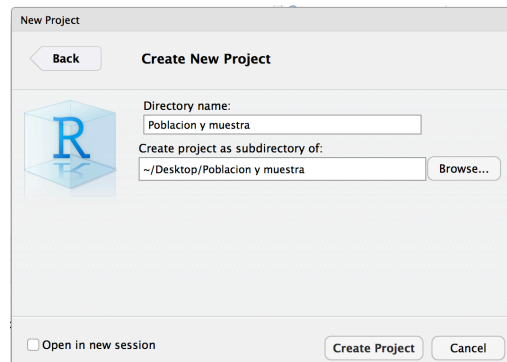
File/New project.../New directory/New Project



En la siguiente ventana, escribimos:

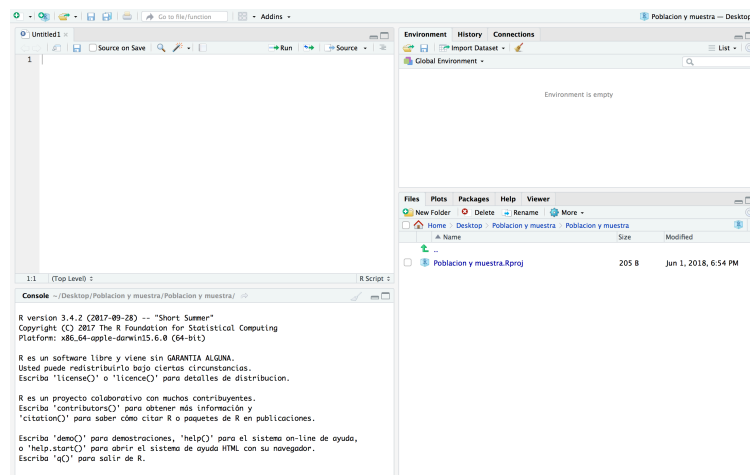
Directory name: Población y muestra.

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.

3EAC1.- 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.

3EAC2.- Población, individuo, muestra. Variables estadísticas.

Estos conceptos se han visto en el curso anterior. Veamos un repaso:

Población: Es el conjunto total de individuos sobre los que se quieren estudiar unos datos determinados.

Individuo: Cada uno de los componentes de la población. Pueden ser personas, animales, plantas, u objetos.

Cuando la población o colectivo sea muy grande, se hará difícil el estudio de la misma. Estos inconvenientes pueden ser superados mediante la elección de muestras.

Muestra: Es una parte de la población representativa de la misma. Tiene por tanto características similares. Ha de elegirse al azar. Se utiliza cuando la población es muy grande, o difícil de estudiar.

Variable estadística: Es el dato o característica que se quiere estudiar. Por ejemplo: la estatura, la nota de Matemáticas, el sexo de una persona, o su peso.

3EAC3.- Variables estadísticas: cualitativas, cuantitativas, discretas y continuas.

Variable cualitativa: Describen cualidades que no puede expresarse por números. Por ejemplo, provincias españolas, colores favoritos, qué libro lectura prefieren los adolescentes, o el coche más vendido.

Variable cuantitativa: Las variables cuantitativas toman valores numéricos. Por ejemplo, la estatura, o la nota de una asignatura. Pueden ser:

Variable cuantitativa discreta: Los valores de la variable son números enteros 1, 2, 3, 4, 5. Por ejemplo el número de compras de un producto en un mes, no puede ser 4,8.

Variable cuantitativa continua: Pueden tomar todos los valores dentro de un intervalo. La temperatura, la humedad, o la estatura, son ejemplos de variables cuantitativas continuas. Los valores razonables de la variable temperatura en una persona pueden valer desde 34,5°C hasta 42°C, pudiendo tomar cualquier valor intermedio.

3EAC4.- Métodos de selección de una muestra estadística. Representatividad de una muestra.

Para recoger los datos y determinar los valores de la variable se puede utilizar a toda la población, todo el universo sobre el que se realiza el estudio, o seleccionar una muestra.

En muchas ocasiones no es conveniente recoger valores de toda la población, porque es complicado o demasiado costoso, o incluso porque es imposible.

Métodos de selección de una muestra

Hay varios métodos para seleccionar una muestra. Veamos tres de ellos:

1.- Muestreo aleatorio simple

Cada individuo de la población tienen idéntica probabilidad de ser elegidos en la muestra.

2.- Muestreo aleatorio sistemático

Se colocan por orden los individuos de la población. Se selecciona un primer individuo de manera aleatoria, y a partir de él, se seleccionan los demás a intervalos fijos.

3.- Muestreo aleatorio estratificado

Se divide la población en grupos homogéneos de una determinada característica, por ejemplo su edad, su sexo o su nacionalidad. Estos grupos se denominan estratos. A continuación, se toma una muestra aleatoria simple en cada estrato.

Representatividad de una muestra

Cuando se elige una muestra los dos aspectos que hay que tener en cuenta son el tamaño y la representatividad de la misma.

Si la muestra no tiene el tamaño suficiente, el resultado no será fiable.

A medida que la muestra crece, los resultados serán más fiables. Sin embargo, cuanto mayor sea la muestra, mayor será el gasto para conseguirla. No siempre muestras grandes nos proporcionan mejores resultados. Por esto, debemos aprender a encontrar el tamaño adecuado para poder afirmar, con una confianza alta, que una población tiene cierta característica.

Con el tamaño adecuado de la muestra, y si ha sido elegida de forma aleatoria, podremos decir que es una muestra representativa.

Si el muestreo no se ha realizado de forma aleatoria, presentará un sesgo, y podremos decir que la muestra es sesgada.

Por ejemplo, si quisiéramos estudiar la estatura de una población, no sería razonable sacar las muestras de los equipos de baloncesto.

Ejemplo:

Aunque no lo hemos estudiado aún, vamos a calcular la longitud media de los pétalos de la tabla iris que contiene la muestra de 150 plantas de 3 especies diferentes, para ver las consecuencias de una mala elección de una muestra:

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

```
> iris
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
96	5.7	3.0	4.2	1.2	versicolor
97	5.7	2.9	4.2	1.3	versicolor
98	6.2	2.9	4.3	1.3	versicolor
99	5.1	2.5	3.0	1.1	versicolor
100	5.7	2.8	4.1	1.3	versicolor
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
103	7.1	3.0	5.9	2.1	virginica
150	5.9	3.0	5.1	1.8	virginica

Realmente, la tabla iris es una tabla de 150 plantas, de la que solo mostramos unas pocas. En esta tabla aparecen la longitud de los sépalos

(Sepal.Length), la anchura de los sépalos (Sepal.Width), la longitud de los pétalos (Petal.Length), la anchura de los pétalos (Petal.Width), y la especie a la que pertenece la planta estudiada (Species). Tenemos tres tipos de especies, que pueden verse en las siguientes fotos:



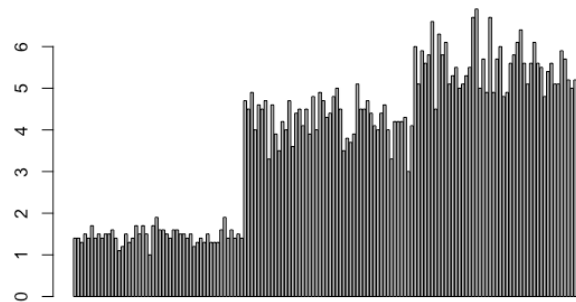
Iris Versicolor

Iris Setosa

Iris Virginica

Dibujamos el diagrama de barras de la longitud de los pétalos, con la instrucción Barplot:

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



Si nos fijamos en el diagrama, vemos que la longitud de los pétalos puede dividirse en tres grandes grupos. Estos grupos coinciden precisamente con la especie a la que pertenece.

Si para estudiar la longitud de toda la población, tomáramos una muestra de la primera parte de la tabla, seguramente obtendríamos resultados erróneos.

Calculamos la media de la longitud de los pétalos de esta población:

```
> mean(iris$Petal.Length)
[1] 3.758
```

Si tomásemos como muestra los 50 primeros individuos, la media sería:

```
> mean(iris[1:50,]$Petal.Length)
[1] 1.462
```

Los 50 siguientes:

```
> mean(iris[50:100,]$Petal.Length)
[1] 4.203922
```

Los últimos 50:

```
> mean(iris[100:150,]$Petal.Length)
[1] 5.523529
```

Si lo que queremos es estudiar la longitud media de los pétalos de esas 150 plantas, tendremos que hacer una muestra aleatoria.

Vamos a tomar una muestra al azar de 50 individuos.

```
> muestra<- sample(iris$Petal.Length,size=50).
```

Calculamos la media de esta muestra:

```
> mean (muestra)
[1] 3.502
```

Observa que aumentando la muestra ganamos precisión:

Muestra de 100 individuos:

```
> muestra100<- sample(iris$Petal.Length,size=100)
> mean (muestra100)
[1] 3.634
```

Vemos que no difiere mucho de la primera que tomamos, pero contiene un error.

```
> mean(iris$Petal.Length) =      [1] 3.758: Media población total.
> mean(iris[1:50,]$Petal.Length)= [1] 1.462: Media 50 primeros.
> mean(iris[50:100,]$Petal.Length)= [1] 4.203922: Media 50 siguientes.
> mean(iris[100:150,]$Petal.Length)= [1] 5.523529: Media últimos 50.
> mean (muestra)=                [1] 3.502: Media muestra aleatoria 50 ind.
> mean (muestra100)=              [1] 3.634: Media muestra aleat de 100 ind.
```

¿Qué muestra te parece más representativa?

3EAC5.- Frecuencias absolutas, relativas y acumuladas. Agrupación de datos en intervalos.

Frecuencia absoluta (n_i): Es el número de veces que aparece cada valor (x_i) de la variable.

La suma de las frecuencias absolutas es el número total de datos (N).

En "R", usaremos la función `table` para crear tablas de frecuencias absolutas:

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

Frecuencia relativa (f_i): es el resultado de dividir la frecuencia absoluta entre el número total de datos (N):

$$f_i = \frac{n_i}{N}$$

En "R", usaremos la tabla de frecuencias absolutas, y la dividiremos por el número de individuos, que se obtiene con la función `margin.table`:

```
> FRLongpetal<-(FALongpetal)/margin.table(FALongpetal)
```

¿Qué pasaría si quisieramos escribir la frecuencia relativa de todos los habitantes rubios de Valladolid? Habría que contar todos los habitantes rubios, y dividir por el número de habitantes.

Estos ejercicios se realizan mucho más rápido utilizando el ordenador, como vamos a ver a continuación.

Frecuencias acumuladas

La frecuencia absoluta acumulada (N_i) de un valor X_i del conjunto (X_1, X_2, \dots, X_N) es la suma de las frecuencias absolutas de los valores menores o iguales a X_i , es decir: $N_i = n_1 + n_2 + \dots + n_i$.

3EAC6.- Gráficas estadísticas.

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 vamos a cargar la tabla `mtcars` que recoge información de 32 tipos de coches. Para verla, escribimos:

```
> mtcars
```

A continuación, creamos la tabla de frecuencias de la variable número de cilindros (`mtcars$cyl`), con la función `table`:

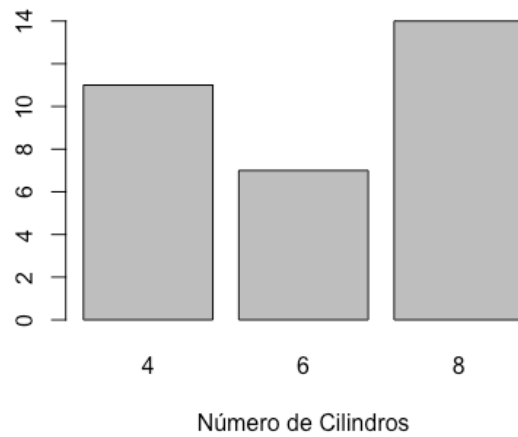
```
> Cyl_cars<-table(mtcars$cyl)
```

```
> Cyl_cars
```

```
 4  6  8  
11  7 14
```

Ahora dibujamos el diagrama de barras de la variable `Cyl_cars`

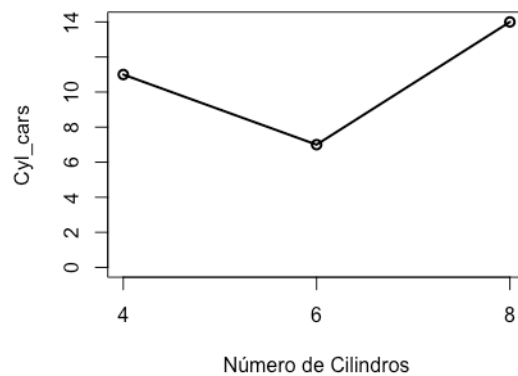
```
> barplot(Cyl_cars, xlab="Número de Cilindros")
```



Vemos que hay aproximadamente el doble de coches con 8 cilindros que con 6.

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

```
> plot(Cyl_cars, type='o', xlab="Número de Cilindros")
```



Estos diagramas nos sugieren que los coches con 6 cilindros son menos frecuentes.

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:

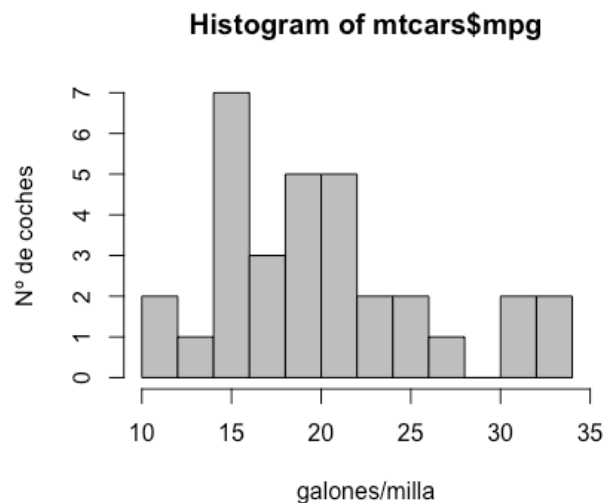
Tiene como eje horizontal una escala de valores de la variable que se mide, sobre la que se marcan los límites de las clases sobre la escala.

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.

Vamos a crear el histograma de la variable consumo en millas por galón:

```
> hist(mtcars$mpg, xlab="galones/milla",ylab="Nº de coches", xlim=c(10,35),  
breaks=12,col="gray")
```

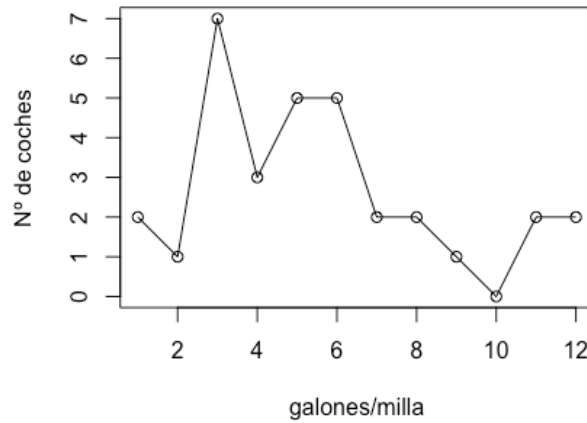


El polígono de frecuencias se construye uniendo los puntos medios de los lados superiores de los rectángulos. Podemos guardar en una variable todos los datos del histograma que acabamos de crear:

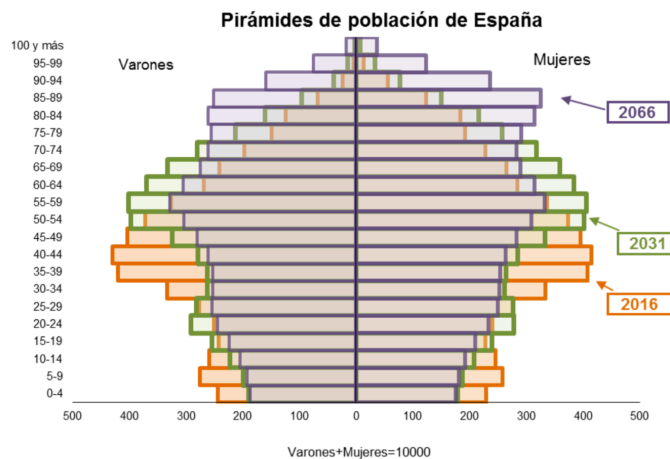
```
> Barras_Cons<- (hist(mtcars$mpg, xlab="galones/milla",ylab="Nº de  
coches", xlim=c(10,35), breaks=12,col="gray"))
```

Y a continuación, trazar el polígono de frecuencias:

```
> plot (Barras_Cons$counts,xlim=c(1, 12),ylim=c(0, 7), col="black",
xlab="galones/milla",ylab="Nº de coches",type = "o")
```



Las pirámides de población son histogramas dobles:

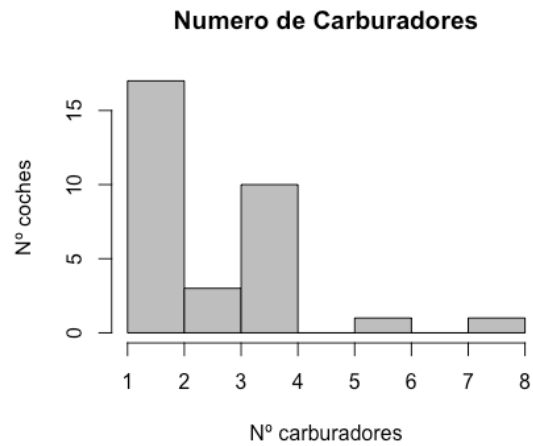


Fuente: <http://www.ine.es/>

Ejemplo:

Representa mediante un histograma el número de carburadores de los vehículos de la tabla mtcars.

```
> hist(mtcars$carb, col = "gray", main = "Numero de Carburadores",xlab =
"Nº carburadores", ylab = "Nº coches")
```



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=8`.

3EAC7.- Parámetros de posición central (media, moda y mediana) y no central (primer y tercer cuartil). Cálculo, interpretación y propiedades.

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)`

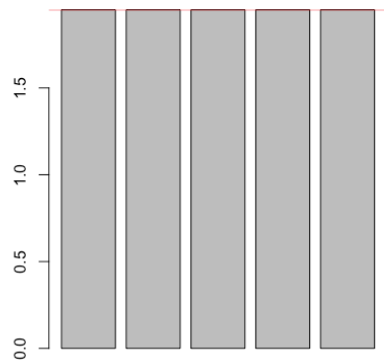
Interpretación y Propiedades:

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.

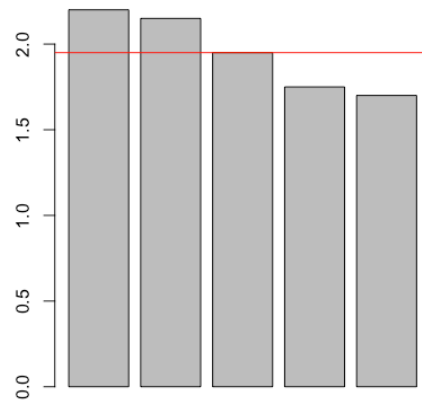
Si medimos, un segundo equipo de jugadores y obtenemos que miden 2.20m, 2.15m, 1.95m, 1.75m y 1.70m, éste segundo equipo tendría una estatura media de 1.95 m. Pero en este caso, este valor representa ninguno de sus componentes.

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 del consumo de los vehículos de la tabla mtcars, en millas por galón:

Media:

```
> mean (mtcars$mpg)
[1] 20.09062
```

Mediana:

```
> median (mtcars$mpg)
[1] 19.2
```

Moda:

Calculamos las frecuencias absolutas y la ordenamos:

```
> sort (table (mtcars$mpg))
13.3 14.3 14.7 15 15.5 15.8 16.4 17.3 17.8 18.1 18.7 19.7 21.5 24.4 26
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

27.3	32.4	33.9	10.4	15.2	19.2	21	21.4	22.8	30.4
1	1	1	2	2	2	2	2	2	2

En este caso, hay varias modas, que aparecen 2 veces: (10.4, 15.2, 19.2, 21, 21.4, 22.8, y 30.4) mpg

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.

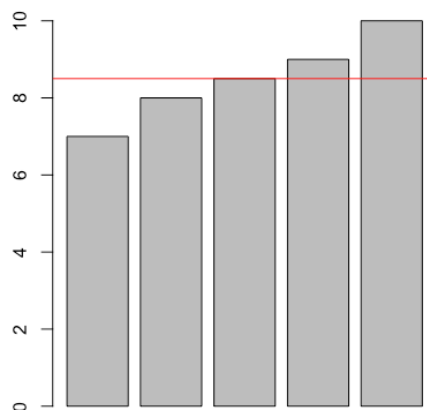
Media Geométrica: La media geométrica de una muestra se obtiene al multiplicarlos entre si y aplicarles la n-ésima raíz.

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

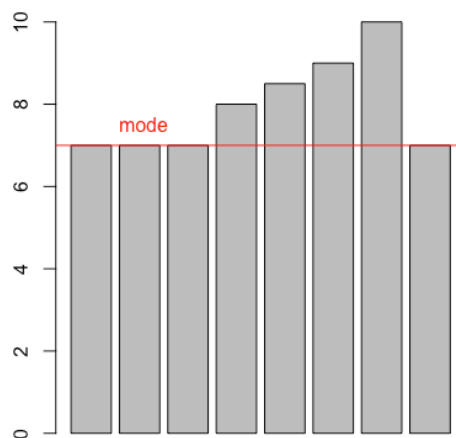
Su uso es adecuado en distribuciones asimétricas, puesto que no se ve afectada por las observaciones extremas, al no depender de los valores que toma la variable.

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 los consumos, y buscamos el mayor valor, ordenando la tabla con el comando sort:

```
sort (table (mtcars$mpg))
```

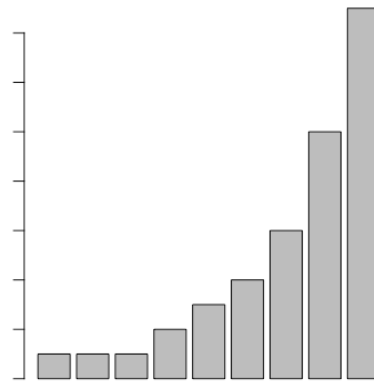


```
> Ej2<-c(7,7,7,8,8.5,9,10,7)
> barplot (Ej2)
> abline (h=7, col="red")
> text(2, 7.5, "mode", col = "red")
```

Interpretación y Propiedades

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

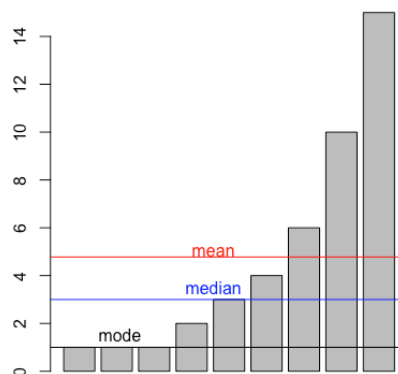
Ejemplo: Gráfico de la altura de una población de 9 árboles.



```
> 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 (Q_i): 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.

25%	25%	25%	25%
Q1	Q2	Q3	

3EAC8.- Parámetros de dispersión (rango, recorrido intercuartílico, varianza, desviación típica y coeficiente de variación).

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.

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

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

En "R": range (x)

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 S².

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

En "R": var (x)

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)

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.

```
> p1<-c(16, 0, 16, 0)      > p2<-c(16, 6, 10, 0)      > p3<-c(9, 7, 7, 9)

> mean (p1)                > mean (p2)                > mean (p3)
[1] 8                      [1] 8                      [1] 8

sd(p1)                     > sd(p2)                     > sd(p3)
[1] 9.237604               [1] 6.733003               [1] 1.154701
```

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:

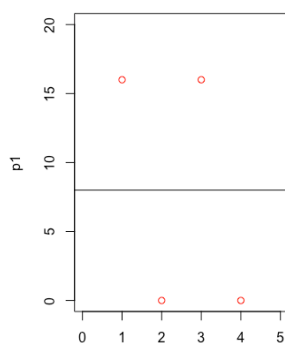
```
> plot( p1, col="red",xlim=c(0, 5),ylim=c(0, 20))
> abline(h=8)
```

Nube p2:

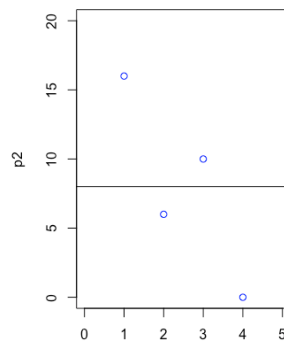
```
> plot( p2, col="blue",xlim=c(0, 5),ylim=c(0, 20))
> abline(h=8)
```

Nube p3:

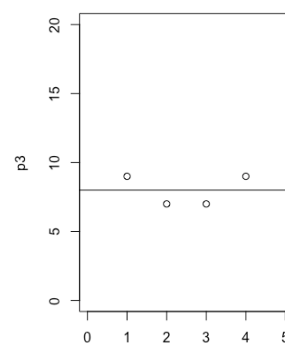
```
> plot( p3, col="black",xlim=c(0, 5),ylim=c(0, 20))
> abline(h=8)
```



Nube p1



Nube p2



Nube p3

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

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=mín, Q1=25%, Q2=50%, Q3=75%, y Q4=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
```

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 del consumo de los vehículos de la tabla `mtcars`.

```
> sd (mtcars$mpg)
[1] 6.026948
> range (mtcars$mpg)
[1] 10.4 33.9
> var (mtcars$mpg)
[1] 36.3241
```

En este ejemplo los datos aparecen muy dispersos. Hay mucha variabilidad en el consumo. Por eso la desviación media es alta.

El rango nos da los extremos de los consumos mínimo y máximo.

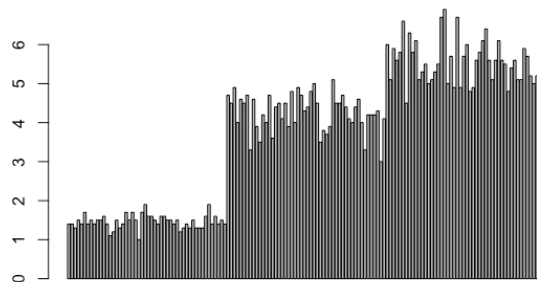
2.- Calcular las medias de las longitudes de los pétalos de la tabla `iris` que contiene la muestra de 150 plantas de 3 especies diferentes:

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)
```



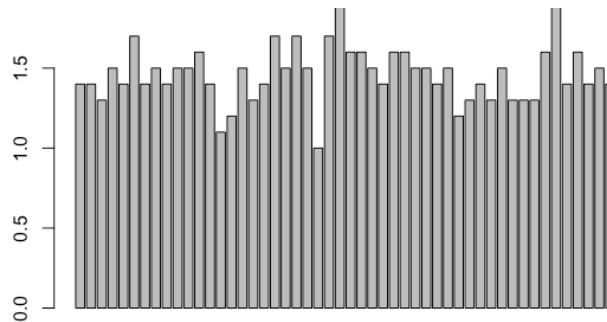
En el gráfico se aprecian tres escalones. Si nos fijamos en la tabla, las 50 primeras muestras son plantas de la especie Setosa, las 50 siguientes de la especie versicolor y las 50 últimas de la especie virgínica.

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
```

3EAC9.- Diagrama de caja y bigotes.

Este tipo de diagramas son representaciones semigráficas, que nos permiten observar las características principales de la muestra, y nos ayudan a detectar posibles valores atípicos.

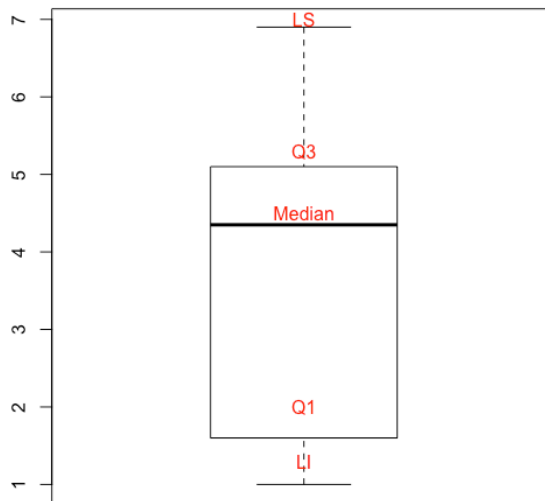
Pasos para construir un Box-Plot:

Primero se ordenan los datos de menor a mayor, para obtener el máximo, el mínimo, y los cuartiles.

Se forma un rectángulo, o caja, cuyos lados son los cuartiles Q_1 y Q_3 . En el centro, se señala la mediana Me .

Se añaden dos brazos, o bigotes, donde se señalan los valores máximo Máx. y mínimo, Mín.

Se pueden calcular además, unos límites superior e inferior. El inferior, Li , es $Q_1 - 1.5$ por el intervalo intercuartil, y el superior Ls es $Q_3 + 1.5$ por el intervalo intercuartil.

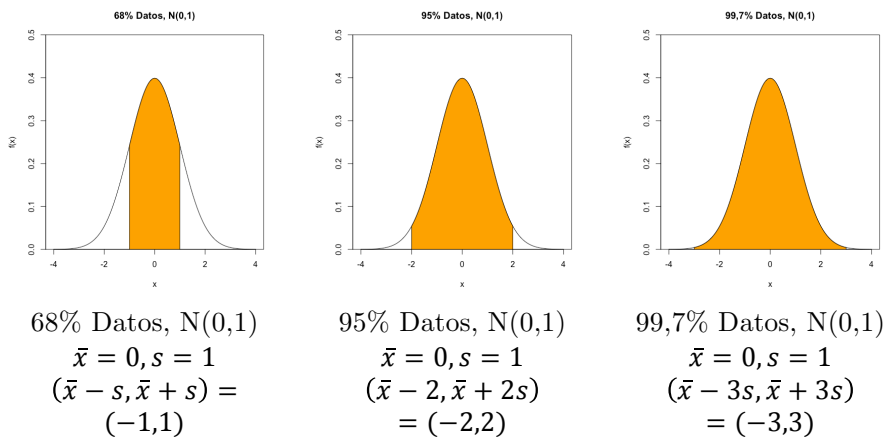


```
> boxplot(iris$Petal.Length)
> text(1.3, "LI", col = "red")
> text(7, "LS", col = "red")
> text(4.5, "Median", col = "red")
> text(5.3, "Q3", col = "red")
> text(2, "Q1", col = "red")
```

3EAC10.- Interpretación conjunta de la media y la desviación típica.

La desviación típica mide la distancia de los datos respecto de la media, dando una idea sobre cómo se agrupan los datos alrededor de la media.

Si los datos recogidos responden a lo que se conoce como una distribución normal, que por el momento no sabemos lo que esto significa, podríamos asegurar que en el intervalo generado entre la media menos una desviación típica, y la media más una desviación típica, están más del 68 % de los datos; en el intervalo entre la media menos 2 desviaciones típicas, y la media más 2 desviaciones típicas están más del 95 % de los datos, y entre la media menos 3 desviaciones típicas y la media más 3 desviaciones típicas están más del 99,7 % de los datos.



Código de las ilustraciones:

```
> radio=2
> regionX=seq((0-radio),(0+radio),0.01)
> xP <- c((0-radio),regionX,(0+radio))
> yP <- c(0,dnorm(regionX,0,1),0)
> curve(dnorm(x,0,1),xlim=c(-4,4),yaxs="i",ylim=c(0,0.5),ylab="f(x)",
+       main='95% datos, N(0,1)')
> polygon(xP,yP,col="orange")
```

Por ejemplo, si la altura de una planta está dentro de ese intervalo $(\bar{x} - s, \bar{x} + s)$, podríamos decir que es normal. Si se encuentra por encima del intervalo $(\bar{x} - 2s, \bar{x} + 2s)$ podríamos decir que es superior a la media, y si se

encuentra por encima del intervalo $(\bar{x} - 3s, \bar{x} + 3s)$, podríamos decir que es una planta gigante, o un caso atípico para su especie.

A la vista de las ilustraciones, se puede observar que prácticamente todos los datos distan de la media menos de 3 desviaciones típicas y que más del 68% distan menos de una desviación típica.

Con estos datos, es posible tomar decisiones, inferir o predecir con una cierta probabilidad lo que va a ocurrir.

Capítulo 5. Estándares de aprendizaje evaluables.

Estándares 3º ESO. Académicas

1.1. Distingue población y muestra justificando las diferencias en problemas contextualizados.

1.2. Valora la representatividad de una muestra a través del procedimiento de selección, en casos sencillos.

1.3. Distingue entre variable cualitativa, cuantitativa discreta y cuantitativa continua y pone ejemplos.

1.4. Elabora tablas de frecuencias, relaciona los distintos tipos de frecuencias y obtiene información de la tabla elaborada.

1.5. Construye, con la ayuda de herramientas tecnológicas si fuese necesario, gráficos estadísticos adecuados a distintas situaciones relacionadas con variables asociadas a problemas sociales, económicos y de la vida cotidiana.

2.1. Calcula e interpreta las medidas de posición (media, moda, mediana y cuartiles) de una variable estadística para proporcionar un resumen de los datos.

2.2. Calcula e interpreta los parámetros de dispersión (rango, recorrido intercuartílico y desviación típica) de una variable estadística (con calculadora y con hoja de cálculo) para comparar la representatividad de la media y describir los datos.

3.1. Utiliza un vocabulario adecuado para describir, analizar e interpretar información estadística de los medios de comunicación.

.

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

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>

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

Libros de texto IES Cristo Rey, curso 2017-12018

Matemáticas 1º ESO:

Autores: Francisco Javier García Crespo; Ruth Martín Escanilla

Editorial: EDITEX S.A

1ª ed. (2015)

ISBN: 8490784949 ISBN-13: 9788490784945

Matemáticas 2º ESO

Autores: Fernando ... [et al.] Alcaide Guindo

Pelorroto; Juan Antonio Rocafort (il.)

Editorial: EDICIONES SM

1ª ed. (01/05/2016)

ISBN: 8467586885 ISBN-13: 9788467586886

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

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>