



UNIVERSIDAD DE VALLADOLID

Dpto. Estadística e I.O.

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

**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

Agradecimientos.

Este trabajo no habría sido posible sin la ayuda de mi esposa, Esther García, quien ha soportado la agotadora labor de atender a nuestra preciosa e incansable hija, Carla.

Agradecerle al equipo de profesores del Máster el haberme aportado unos conocimientos de semejante valor.

A mi tutor, David Conde, por haberse preocupado en todo momento por el avance de este trabajo, y por haber confiado en mi planificación y en mi criterio.

A mi segunda familia, Angelines, Alberto y María quienes me han mostrado su más sincero apoyo y en ocasiones, la admiración que me ha ayudado a no abandonar.

Y por supuesto, a mis padres y hermanos, quienes me lo han dado todo.

Es a ellos a quien dedico este trabajo.

“No importa lo mal que lo hayas hecho hasta ahora. A partir de ahora,
hazlo bien”

David Conde

Índice general.

Agradecimientos.....	3
Índice general.	7
Capítulo 1. Introducción.....	9
Capítulo 2. Contenidos y estándares oficiales.....	13
Capítulo 3. Introducción a “R” Statistics.....	19
Capítulo 4. Contenidos y estándares oficiales con “R” Statistics.	23
Capítulo 5. Estándares de aprendizaje evaluables.....	97
Capítulo 6. Bibliografía.	103

Capítulo 1. Introducción.

“Enseñar a los alumnos a utilizar software libre y a participar en la comunidad del software libre es una lección cívica llevada a la práctica.”

Richard Stallman.

<https://www.gnu.org/education/edu-schools.es.html>

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.

A principios de Septiembre de 2017 comenzaba la parte común de este Master, con la asignatura de Sociedad, Familia y Educación, donde se nos invitaba a hacer la reflexión sobre la situación familiar y el contexto en el que viven nuestros alumnos. En ella aprendimos a ponernos en el lugar de los que menos tienen, y la necesidad de acercar la enseñanza a todos los hogares, de una manera inclusiva. Por todo esto, y por un motivo de principios morales y éticos que he visto en casa desde pequeño, he elegido Software Libre, con el que pretendo fomentar este tipo de herramientas en el aula.

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

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. En ella 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. Estos anexos vienen a poner de manifiesto parte de lo aprendido en la asignatura de Diseño Curricular en Matemáticas, y son prácticamente nueve Unidades Didácticas, en las que se deja al usuario final la posibilidad de personalizar lo relativo a la temporalización, recursos, o atención a la diversidad, en función del contexto al que lo quiera aplicar.

Tanto en la memoria como en los anexos, aparece todo el código utilizado en la elaboración del trabajo, ya sea para la resolución de los problemas, como para las ilustraciones de ejemplo.

Se ha tratado de utilizar un lenguaje que sea comprensible por los alumnos buscando siempre mantener el rigor, como aprendimos en las asignaturas de Innovación Docente en Matemáticas e Iniciación a la Investigación educativa en Matemáticas.

Las versiones de introducción o de repaso de cada punto del temario se han dejado en cada uno de estos nueve anexos, para evitar la duplicación innecesaria 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 evadir la teoría, ni de evitar que el alumno trabaje el tan necesario cálculo mental y manual. El mismísimo Da Vinci, dijo que “los que se enamoran de la práctica sin la teoría son como los pilotos sin timón ni brújula, que nunca podrán saber a dónde van”, reflexión que nos invitaban a hacer en la asignatura de Didáctica de la matemática.

En parte por esto, se ha conservado todo el contenido teórico, y no se ha eludido el realizar los ejercicios de la forma tradicional. Un contenido teórico que el profesor debe conocer con solvencia, como aprendimos en Complementos de Matemáticas. El otro motivo de mantener la forma tradicional de resolución, es que al finalizar el Bachillerato, el alumno se enfrentará a una prueba donde no le permitirán hacer uso de este software. De este modo, el alumno dispone también del material necesario para preparar la prueba sin la ayuda del ordenador.

En cualquier caso, si el grupo llega hasta este último bloque con retraso, uno de los motivos puede ser precisamente el llevar trabajando cerca de ocho

meses en la línea tradicional. La metodología propuesta, trata de optimizar el poco tiempo del que disponga el profesor, evitando pérdidas en la representación de gráficos a mano, diagramas de barras, nubes de puntos, o tablas de contingencia. En Metodología y Evaluación en Matemáticas, nos planteábamos si resulta práctico competir contra un ordenador en esto, cosa que hemos podido constatar gracias a las Prácticas Externas Matemáticas, donde hemos podido ver el tiempo que le lleva al alumno el simple hecho de dibujar dos ejes y representar ocho o diez puntos a mano.

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.

Esta reflexión nos parece acertada, puesto que 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.

Y es que como aprendíamos en la asignatura de Modelos Matemáticos, la estadística está por todas las partes, al igual que lo están las matemáticas como nos mostraba con pasión, Encarnación Reyes. En la naturaleza, en la salud, en la industria, en el ocio, en la educación, etc. Por su parte, la tecnología es capaz de generar y recoger todos estos datos con una facilidad cada vez mayor, como vimos en Ideas y conceptos matemáticos a través de la historia, donde aprendimos también la importancia que va cobrando la estadística día tras día, y la proyección de futuro que tiene.

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

Así, al finalizar sus estudios será capaz de recopilar datos por sí mismo, organizarlos, resumirlos, estudiarlos y explotarlos, lo que le será de gran utilidad en su ámbito profesional. Será capaz de realizar análisis críticos de una mayor cantidad de información mediante tablas y gráficas, con la ayuda de “R”. No debemos olvidar, que en estas edades, el desarrollo del pensamiento crítico del alumno está en pleno apogeo, como sabemos de Aprendizaje y desarrollo de la personalidad. Para potenciar este desarrollo, y como fomento de la lectura, se proponen dos libros: “How to lie with statistics”, y “El tigre que no está. Un paseo por la jungla de la estadística.” Ambos, presentan con humor las falacias que podemos encontrarnos a diario en los medios de comunicación.

El contenido del trabajo está adaptado a la comunidad de Castilla y León y sacado de fuentes primarias, de la misma manera que aprendimos a hacerlo en la asignatura de Procesos y Contextos educativos. La fuente en cuestión, son 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.

De este contenido oficial, se ha seleccionado la parte de estadística, como indica el título del TFM. No obstante, el software tiene potencial para incorporar también la parte de Probabilidad. Esto se deja como puerta abierta para el futuro, para completar la totalidad del quinto bloque de la asignatura a todos sus niveles de ESO y Bachillerato.

Capítulo 2. Contenidos y estándares oficiales.

A continuación se detallan los contenidos oficiales de cada uno de los cursos de ESO y Bachillerato.

1º ESO.

- 1E1.- Población e individuo. Muestra. Variables estadísticas.
- 1E2.- Variables cualitativas y cuantitativas discretas.
- 1E3.- Frecuencias absolutas y relativas.
- 1E4.- Organización en tablas de datos recogidos en una experiencia.
- 1E5.- Diagramas de barras, y de sectores. Polígonos de frecuencias.
- 1E6.- Medidas de tendencia central.
- 1E7.- Medidas de dispersión.

2º ESO.

- 2E1.- Población e individuo. Muestra. Variables estadísticas.
- 2E2.- Variables cualitativas y cuantitativas discretas y continuas.
- 2E3.- Frecuencias absolutas y relativas.
- 2E4.- Organización en tablas de datos recogidos en una experiencia.
- 2E5.- Diagramas de sectores, de barras, histogramas y polígonos de frecuencias.
- 2E6.- Otros gráficos estadísticos provenientes de los medios de comunicación.

- 2E7.- Medidas de tendencia central.
- 2E8.- Medidas de dispersión.
- 2E9.- Iniciación en la hoja de cálculo.

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). Cálculo, interpretación y propiedades.
- 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.
- 3EAC12.- Uso de la calculadora científica, de la hoja de cálculo y de otros programas para hacer representaciones gráficas y calcular parámetros.

3º ESO. Aplicadas

- 3EAP1.- Fases y tareas de un estudio estadístico.
- 3EAP2.- Población, individuo, muestra. Variables estadísticas.
- 3EAP3.- Variables cualitativas, discretas y continuas.
- 3EAP4.- Métodos de selección de una muestra estadística. Representatividad de una muestra.
- 3EAP5.- Frecuencias absolutas, relativas y acumuladas. Agrupación de datos en intervalos.
- 3EAP6.- Gráficas estadísticas.

- 3EAP7.- Parámetros de posición: central (media, moda y mediana) y no central (primer y tercer cuartil).
- 3EAP8.- Cálculo, interpretación y propiedades.
- 3EAP9.- Parámetros de dispersión: rango, recorrido intercuartílico, varianza y desviación típica. Cálculo e interpretación.
- 3EAP10.- Diagrama de caja y bigotes.
- 3EAP11.- Interpretación conjunta de la media y la desviación típica.
- 3EAP12.- Uso de la calculadora científica, de la hoja de cálculo y de otros programas, para la representación gráfica, el cálculo de parámetros y su interpretación.

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.

4º ESO Aplicadas

- 4EAP1.- Identificación de las fases y tareas de un estudio estadístico. Población, individuo, muestra.
- 4EAP2.- Gráficas estadísticas: Distintos tipos de gráficas. Análisis crítico de tablas y gráficas estadísticas en los medios de comunicación.
- 4EAP3.- Interpretación, análisis y utilidad de las medidas de centralización y dispersión.

- 4EAP4.- Comparación de distribuciones mediante el uso conjunto de medidas de posición y dispersión.
- 4EAP5.- Construcción e interpretación de diagramas de dispersión. Introducción a la correlación.
- 4EAP6.- Introducción a la estadística bidimensional. Dependencia estadística y dependencia funcional
- 4EAP7.- Construcción e interpretación de diagramas de dispersión. Introducción a la correlación.
- 4EAP8.- Utilización de medios informáticos para el cálculo de parámetros, la representación de unidimensionales representación de nubes de puntos.

1º Bachillerato. Aplicadas.

- 1BA1.- Estadística descriptiva bidimensional: Tablas de contingencia.
- 1BA2.- Distribución conjunta y distribuciones marginales.
- 1BA3.- Medias y desviaciones típicas marginales.
- 1BA4.- Distribuciones condicionadas.
- 1BA5.- Independencia de variables estadísticas.
- 1BA6.- Estudio de la dependencia de dos variables estadísticas.
- 1BA7.- Representación gráfica: Nube de puntos.
- 1BA8.- Dependencia lineal de dos variables estadísticas.
- 1BA9.- Covarianza y correlación: Cálculo e interpretación del coeficiente de correlación lineal.
- 1BA10.- Regresión lineal. Recta de regresión. Estimación. Predicciones estadísticas y fiabilidad de las mismas.

1º Bachillerato. Ciencias sociales.

- 1BC1.- Estadística descriptiva bidimensional: Tablas de contingencia.
- 1BC2.- Distribución conjunta y distribuciones marginales.
- 1BC3.- Distribuciones condicionadas.
- 1BC4.- Medias y desviaciones típicas marginales y condicionadas.
- 1BC5.- Independencia de variables estadísticas.
- 1BC6.- Dependencia de dos variables estadísticas. Representación gráfica:
- 1BC7.- Diagrama de dispersión (o nube de puntos).
- 1BC8.- Dependencia lineal de dos variables estadísticas.

- 1BC9.- Covarianza y correlación: Cálculo e interpretación del coeficiente de correlación lineal.
- 1BC10.- Regresión lineal. Predicciones estadísticas y fiabilidad de las mismas.
- 1BC11.- Coeficiente de determinación.

2º Bachillerato. Ciencias Sociales

- 2BC1.- Población, individuo y muestra.
- 2BC2.- Métodos de selección de una muestra. Tamaño y representatividad de una muestra.
- 2BC3.- Estadística paramétrica. Parámetros de una población y estadísticos obtenidos a partir de una muestra.
- 2BC4.- Estimación puntual. Media y desviación típica de la media muestral y de la proporción muestral. Teorema central del límite. Distribución de probabilidad de la media muestral en una población normal. Distribución de la media muestral y de la proporción muestral en el caso de muestras grandes.
- 2BC5.- Estimación por intervalos de confianza. Relación entre nivel de confianza, error máximo admisible y tamaño muestral.
- 2BC6.- Intervalo de confianza para la media poblacional de una distribución normal con desviación típica conocida.
- 2BC7.- Intervalo de confianza para la media poblacional de una distribución de modelo desconocido y para la proporción en el caso de muestras grandes.

Capítulo 3. Introducción a “R” Statistics.

3.1 About “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/>

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

[Learn More](#)

[Try RStudio Server Pro for free!](#)

RStudio Desktop
Open Source License

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	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	37cade7e162eab62483e655e39dedee
RStudio 1.1.453 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	90.6 MB	2018-05-16	44cddd285bc31c41e4eae1d74b8eebb

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

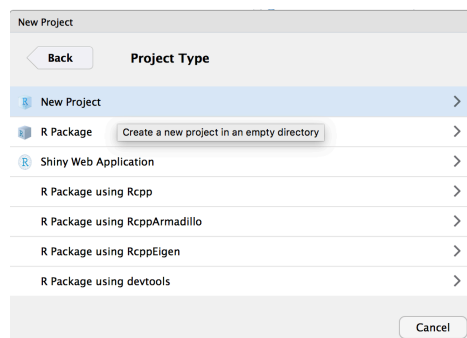
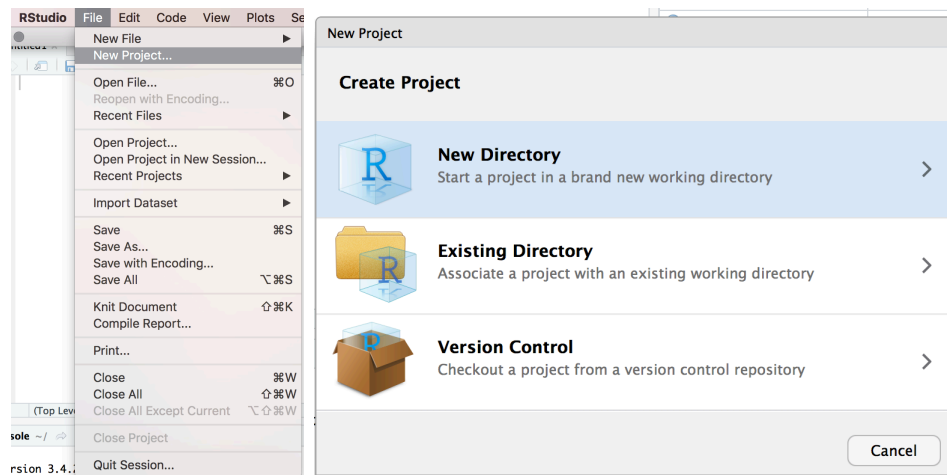
Una vez descargado, se ejecuta el programa instalador, y se van siguiendo los pasos del asistente de instalación, como en cualquier otro programa.

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:

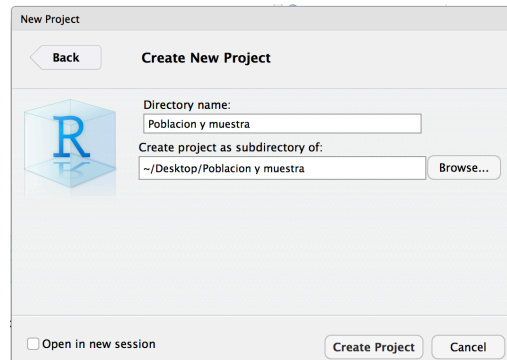
File/New project.../New directory/New Project



En la siguiente ventana, escribimos:

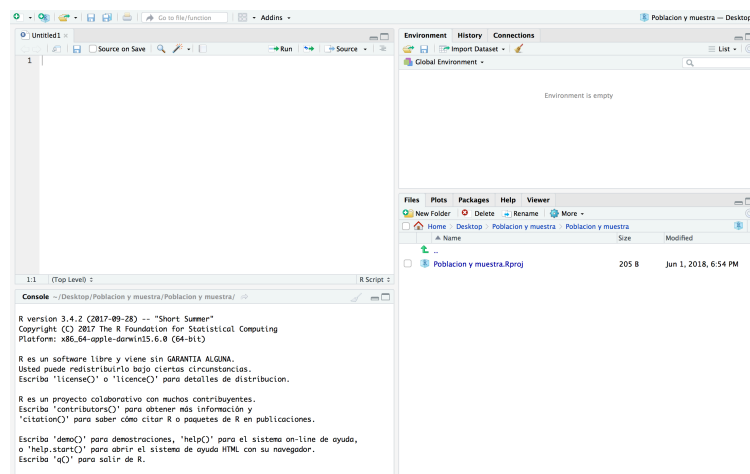
Directory name: Población y muestra.

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.

4.1.- Fases y tareas de un estudio estadístico (3EAC1)

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.

4.2.- Población e individuo. Muestra. Variables estadísticas (1E1).

La Estadística es una ciencia que se ocupa del estudio de los métodos y procedimientos para recoger, clasificar, resumir y analizar datos observados sobre una población de individuos.

El objetivo de cualquier estudio estadístico es obtener información acerca de las características de los individuos de cierto colectivo, llamado población estadística.

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.

Ejemplo. Población.

Vamos a crear un archivo de población con los alumnos de la clase, y su sexo. Para crear la población, nos vamos a ayudar de otro programa informático: LibreOffice Calc.

Lo abrimos, y vamos colocando en la primera fila, uno encima del otro, los nombres de toda la clase por orden de izquierda a derecha, y de adelante a atrás según estemos sentados en el aula. En la segunda columna, colocaremos el primer apellido de cada compañero. En la tercera columna, colocaremos el segundo apellido de cada compañero.

Marta	Hernández	Poza
Paula	Fernández	Mate
Mario	Rodríguez	Crespo
Alberto	Camino	Arranz
Elsa	Moral	Rodríguez
David	Castro	Martin
Álvaro	Martínez	Lobato
Alberto	Martín	Sueña
Sofía	González	Manuel
Claudia	Álvarez	Alonso
María	Lozano	Alonso
Manuel	Martínez	Lobato
Esteban	González	Aguado
Jesús	González	Ortega
Iván	Cano	Hernández
Luis	Mateo	Sánchez
Jimena	Sacristán	Blas
Marga	Domingo	Garzo
Eva	Viñas	Zamora
Ana	Merino	Valdezate
Francisco	Rodríguez	Álvarez
Andrea	Moreno	Jiménez
Carla	Rojo	Domínguez

Lo guardamos con el nombre de Población.xlsx en la carpeta del proyecto Población y muestra. Ya tenemos un archivo con la población de la clase. Esto se conoce como censo. En la antigüedad, los reyes pedían hacer censos constantemente. Incluso en la biblia aparecen referencias a censos.

En el libro de Números, en su capítulo 1, se puede leer lo siguiente:

Censo de las tribus de Israel:

El primer día del segundo mes del segundo año desde la salida de los israelitas de Egipto, el Señor dio las siguientes instrucciones a Moisés, que se encontraba en el santuario, en el desierto del Sinaí.

«Haz un censo de todos los hombres, mayores de veinte años, capaces de ir a la guerra. En la lista anota la tribu y familia a la que pertenezcan.»

Ejemplo 2. Muestra:

Tomar una muestra representativa de la población que acabamos de crear.

Vamos a tomar como muestra a 10 compañeros. Los 10 que queramos:

Marta	Hernández	Poza
Paula	Fernández	Mate
Mario	Rodríguez	Crespo
Alberto	Camino	Arranz
Elsa	Moral	Rodríguez
David	Castro	Martin
Álvaro	Martínez	Lobato
Alberto	Martín	Suaña
Sofía	González	Manuel
Claudia	Álvarez	Alonso

4.3.- Variables cualitativas y cuantitativas (1E2).

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.

Ejemplo 3. Variable cualitativa.

Añadir a la tabla anterior, una variable cualitativa a estudiar: color del pelo. La llamaremos Pelo, para no poner espacios ni caracteres especiales.

Nombre	Apellido1	Segundo2	Pelo
Marta	Hernández	Poza	castaño
Paula	Fernández	Mate	rubio
Mario	Rodríguez	Crespo	moreno
Alberto	Camino	Arranz	rubio
Elsa	Moral	Rodríguez	castaño
David	Castro	Martin	rubio
Álvaro	Martínez	Lobato	moreno
Alberto	Martín	Suaña	castaño
Sofía	González	Manuel	rubio
Claudia	Álvarez	Alonso	castaño
María	Lozano	Alonso	castaño
Manuel	Martínez	Lobato	castaño
Esteban	González	Aguado	moreno
Jesús	González	Ortega	castaño
Iván	Cano	Hernández	castaño
Luis	Mateo	Sánchez	moreno
Jimena	Sacristán	Blas	moreno
Marga	Domingo	Garzo	pelirrojo
Eva	Viñas	Zamora	castaño
Ana	Merino	Valdezate	moreno
Francisco	Rodríguez	Álvarez	rubio
Andrea	Moreno	Jiménez	moreno
Carla	Rojo	Domínguez	pelirrojo

Ejemplo. Variable cuantitativa:

Añadir a la tabla anterior, una columna con una variable cuantitativa discreta y otra continua.

Nombre	Apellido1	Segundo2	Pelo	Sexo	Hermanos	Nota_1EV
--------	-----------	----------	------	------	----------	----------

Marta	Hernández	Poza	castaño	Chica	0	4,71
Paula	Fernández	Mate	rubio	Chica	2	1,75
Mario	Rodríguez	Crespo	moreno	Chico	2	5,52
Alberto	Camino	Arranz	rubio	Chico	2	1,97
Elsa	Moral	Rodríguez	castaño	Chica	0	4,81
David	Castro	Martin	rubio	Chico	2	3,50
Álvaro	Martínez	Lobato	moreno	Chico	1	6,40
Alberto	Martín	Suaña	castaño	Chico	0	4,80
Sofía	González	Manuel	rubio	Chica	2	1,06
Claudia	Álvarez	Alonso	castaño	Chica	1	5,00
María	Lozano	Alonso	castaño	Chica	1	3,58
Manuel	Martínez	Lobato	castaño	Chico	2	5,03
Esteban	González	Aguado	moreno	Chico	1	6,62
Jesús	González	Ortega	castaño	Chico	2	5,00
Iván	Cano	Hernández	castaño	Chico	1	4,86
Luis	Mateo	Sánchez	moreno	Chico	0	9,18
Jimena	Sacristán	Blas	moreno	Chica	3	7,53
Marga	Domingo	Garzo	pelirrojo	Chica	1	0,70
Eva	Viñas	Zamora	castaño	Chica	1	5,28
Ana	Merino	Valdezate	moreno	Chica	1	6,80
Francisco	Rodríguez	Álvarez	rubio	Chico	2	1,99
Andrea	Moreno	Jiménez	moreno	Chica	1	7,72
Carla	Rojo	Domínguez	pelirrojo	Chica	3	0,95

4.4.- Métodos de selección de una muestra estadística. Representatividad de una muestra (3EAC4).

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 es 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:



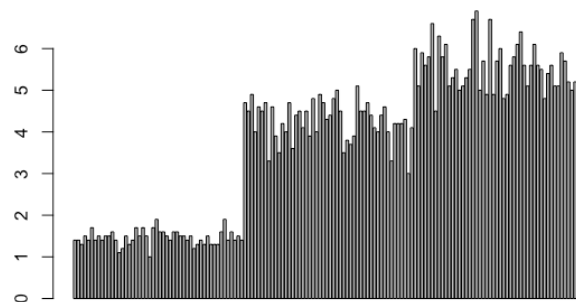
Iris Versicolor

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?

4.5.- Frecuencias absolutas y relativas (1E3).

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 aplicaremos la instrucción `prop.table`:

```
> FRLongpetal<-prop.table(FALongpetal)
```

También podemos usar la función `margin.table`, que nos devuelve el número de individuos, y dividir toda la tabla de frecuencias absolutas por dicha instrucción:

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

¿Qué pasaría si quisiéramos 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: $N_i = n_1 + n_2 + \dots + n_i$.

4.6.- Gráficas estadísticas (4EAP2)

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.

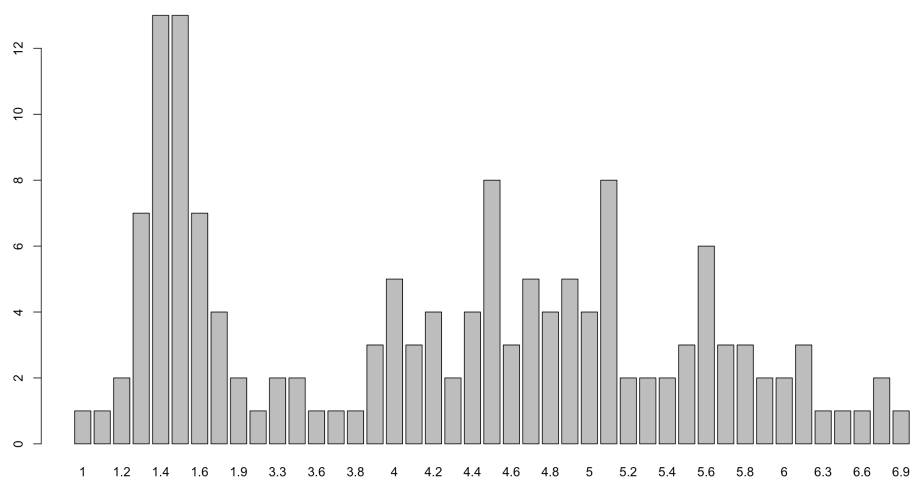
En “R”, se utiliza la función `barplot`.

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)
```

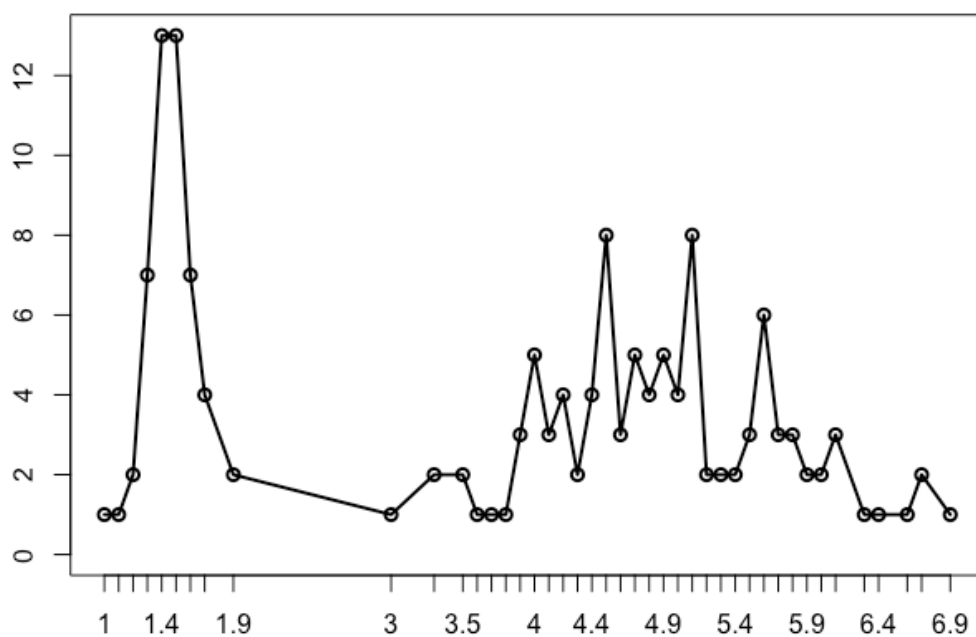
A continuación dibujamos el diagrama de barras de la variable `Longpetal`

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

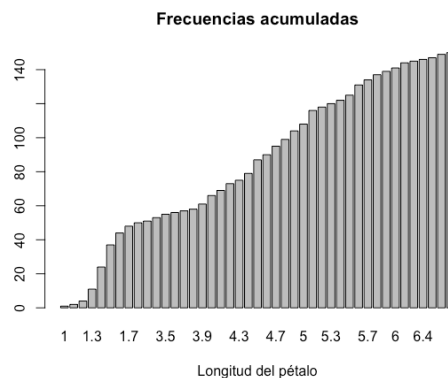
El polígono de frecuencias se dibuja uniendo los centros de las alturas de los rectángulos del diagrama de barras.

```
> 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")
```

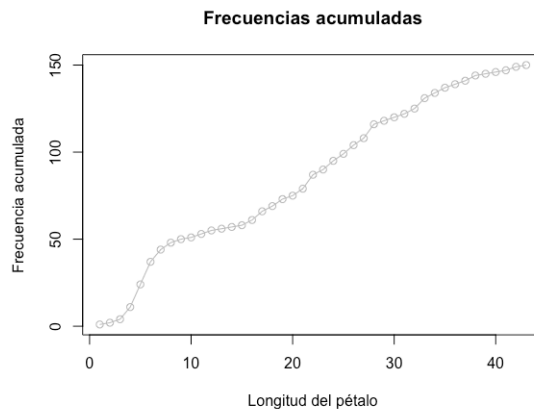


Diagrama de sectores

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

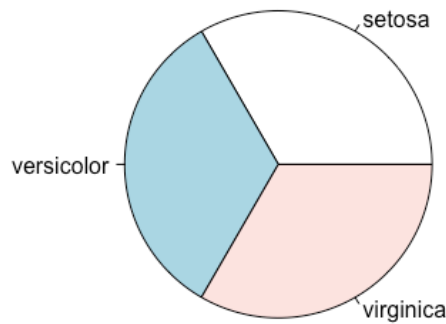
$$\text{Amplitud del sector} = 360^\circ \cdot (f_i/N) = 360^\circ \cdot h_i$$

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:

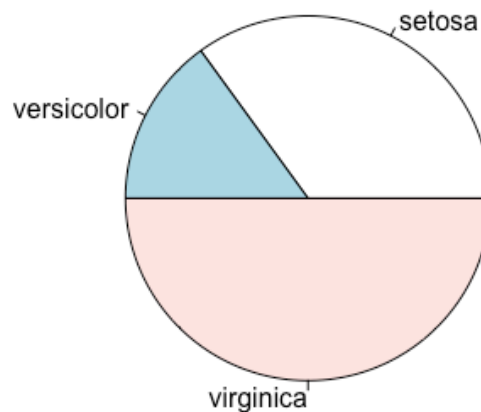
```
> muestra<- sample (1:nrow(iris), size=20,replace=FALSE)
> irismuestra<- iris[muestra, ]
```

	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.

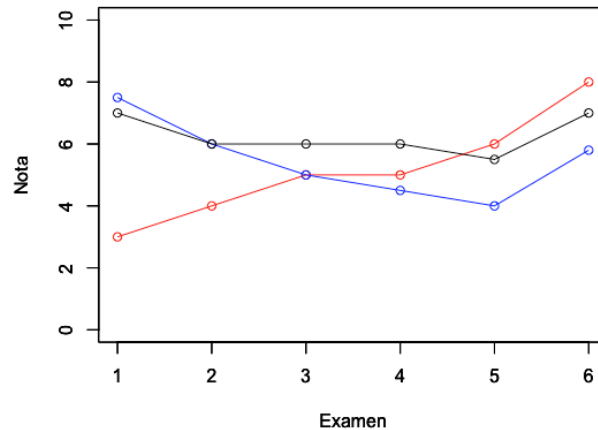
Diagramas lineales

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.

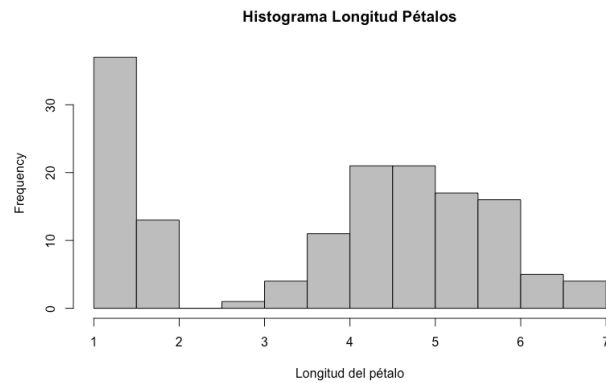
Es la representación gráfica más frecuente para datos agrupados en clases o intervalos. Consiste en un conjunto de rectángulos contruidos de la siguiente forma:

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"):
```

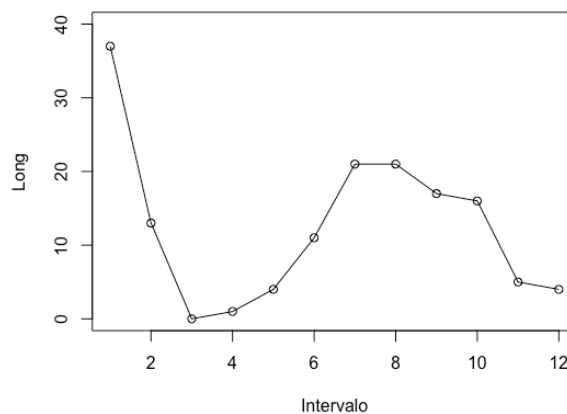


El porígon de frecuencias se construye uniendo los puntos medios de los lados superiores de los rectángulos. Podemos guardar en una variable todos los datos del histograma que acabamos de crear:

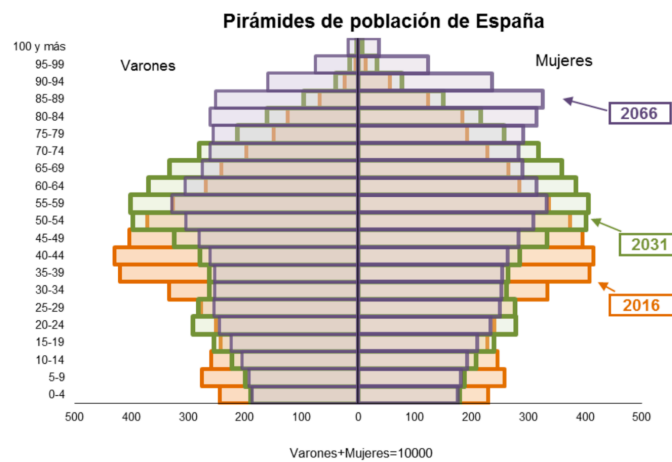
```
> Barras_LP<- (hist(iris$Petal.Length, breaks=12, col="gray",
xlab="Longitud del pétalo", main="Histograma Longitud Pétalos"))
```

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:

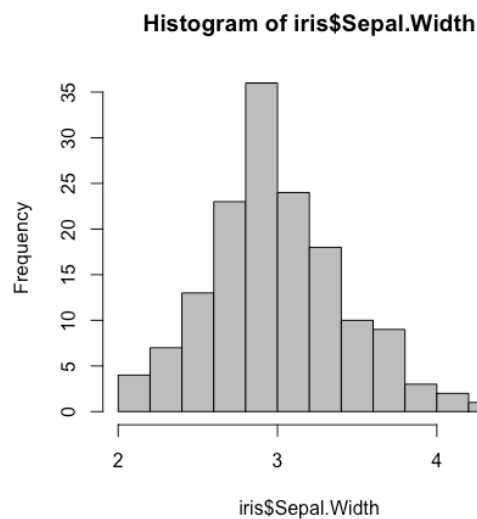


Fuente: <http://www.ine.es/>

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 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 $K = \sqrt{n}$.

4.7.- Organización en tablas de datos recogidos en una experiencia (1E4).

Es muy común encontrarnos los datos colocados en tablas y gráficas para ayudar a su interpretación. Estos datos pueden haber sido obtenidos mediante encuestas, de muy diversos tipos.

Encuesta: Una encuesta es un procedimiento que permite obtener los datos para hacer un estudio de ellos. Puede ser oral o escrita. En función de las preguntas realizadas, se construye una tabla donde se representan los datos.

La tabla Población.xlsx que hemos construido es una encuesta. Podríamos seguir insertando columnas, con las preguntas que se nos ocurriesen.

Hemos visto que si una encuesta se efectúa sobre una población muy grande, puede resultar muy complicado extraer conclusiones. Vamos a aprender a utilizar las herramientas informáticas para analizar estos datos.

Ejemplo (1º y 2º ESO). Tablas en "R":

Ya tenemos los datos de todos los alumnos de la clase en una tabla. Ahora vamos a empezar a estudiar los comportamientos de las variables con la ayuda del software "R".

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

2.- Importa la tabla con los alumnos de la clase. La llamaremos Población, escribiendo lo siguiente en la ventana inferior derecha:

```
>Población<-read_excel("~/Desktop/Población y muestra/Población.xlsx").
```

Vamos a estudiar las frecuencias relativas y absolutas de los alumnos rubios, por un lado, y las frecuencias de tener 2 hermanos por otro, sin la necesidad de hacer cuentas.

Nota: Los nombres de las Variables, escritos en cada cabecera, no deben contener espacios ni caracteres especiales. Para poner espacios, usaremos el guión bajo _.

Construimos una tabla con las frecuencias absolutas de la variable Pelo. Para ello usamos la función `table`, escribiendo lo siguiente:

```
> FA_Pelo<-table(Población$Pelo)
```

A la tabla de frecuencias absolutas la estaremos llamando `FA_Pelo`.

Mediante el símbolo `<-`, que representa una especie de flecha, estaremos diciéndole al programa que escriba dentro de la variable lo que nosotros queramos. En este caso, como queremos escribir la tabla de frecuencias absolutas de la variable Pelo de nuestra población, hemos escrito a continuación de la flecha, la función que queremos utilizar, "table". A continuación y entre paréntesis, hemos escrito la tabla de donde tiene que buscar los datos, (Población\$Pelo). El símbolo `dollar` indica la columna a la que queremos referirnos.

Para visualizar lo que acabamos de escribir en la tabla `FA_Pelo` escribimos:

```
> FA_Pelo
```

El programa nos muestra lo siguiente:

castaño	moreno	pelirrojo	rubio
9	7	2	5

Por lo tanto, podemos ver que hay 9 alumnos castaños, 7 morenos, 2 pelirrojos, y 5 rubios.

La **moda**, en este caso, sería tener el pelo castaño, por ser el valor más repetido.

Las frecuencias absolutas del número de hermanos serían:

```
> FA_Hermanos<-table(Población$Hermanos)
> FA_Hermanos
0 1 2 3
4 9 8 2
```

La **moda** de nuestra clase sería tener un único hermano.

Para las frecuencias relativas usamos la función `margin.table`, que nos devuelve el tamaño de la población o muestra:

```
> FR_Pelo<-(FA_Pelo)/margin.table(FA_Pelo)
> FR_Pelo
  castaño   moreno pelirrojo   rubio
0.39130435 0.30434783 0.08695652 0.21739130
```

Con esta instrucción, el programa divide cada miembro de la tabla de frecuencias absolutas, por el número de miembros, y coloca el resultado en la nueva tabla.

Hallamos las frecuencias relativas de la variable Hermanos. Tenemos dos formas:

Primera forma: Usando la función `margin.table` sobre la tabla de frecuencias absolutas

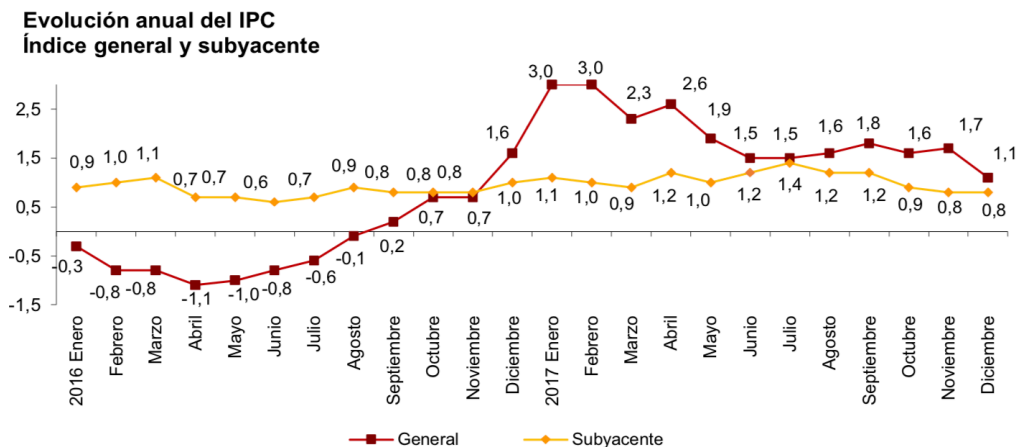
```
> FR_Hermanos<-(FA_Hermanos)/margin.table(FA_Hermanos)
> FR_Hermanos
      0      1      2      3
0.17391304 0.39130435 0.34782609 0.08695652
```

Segunda forma: Usando la función `prop.table` sobre la tabla de frecuencias absolutas:

```
> FR_Hermanos<-prop.table(FA_Hermanos)
> FR_Hermanos
      0      1      2      3
0.17391304 0.39130435 0.34782609 0.08695652
```

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?.

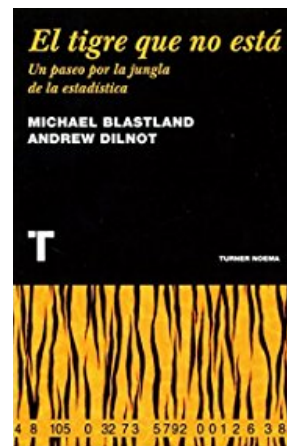
Otro de los ejemplos más comunes son los grados. En ocasiones los periodistas aseguran que alguien ha dado un "*giro de 360°*"; y eso significa que esa persona ha completado una vuelta entera para quedarse en el mismo lugar. Lo correcto sería decir que alguien ha dado "*un giro de 180°*" en su vida. 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 recomendadas:



"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](#)

4.8.- Parámetros de posición: central (media, moda y mediana) y no central (primer y tercer cuartil). Cálculo, interpretación y propiedades. (3EAP7)

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)`

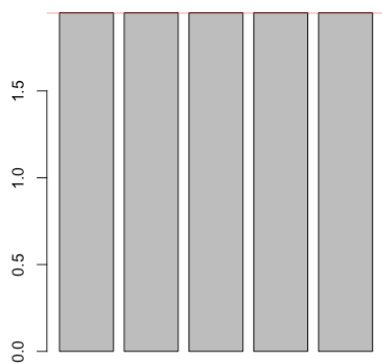
Interpretación y Propiedades:

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.

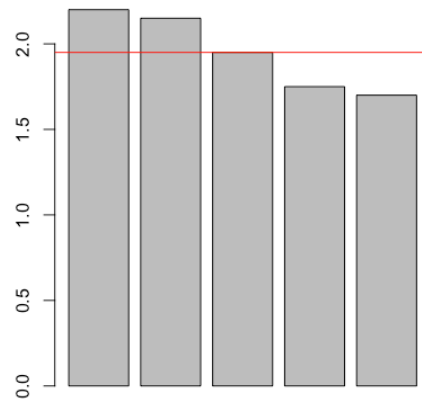
Si medimos, un segundo equipo de jugadores y obtenemos que miden 2.20m, 2.15m, 1.95m, 1.75m y 1.70m, éste segundo equipo tendría una estatura media de 1.95 m. Pero en este caso, este valor representa ninguno de sus componentes.

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")
```

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.

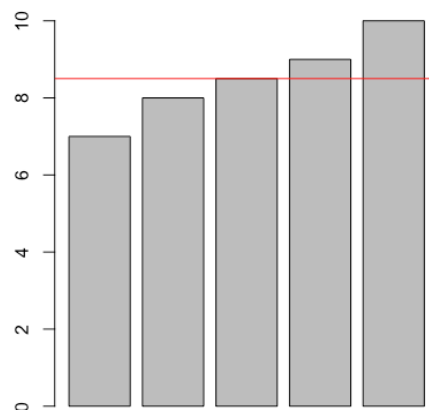
Media Geométrica: La media geométrica de una muestra se obtiene al multiplicarlos entre si y aplicarles la n-ésima raíz.

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

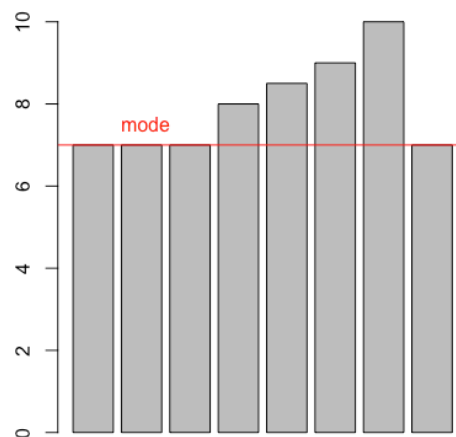
la variable.

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:

```
Moda<-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.

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:

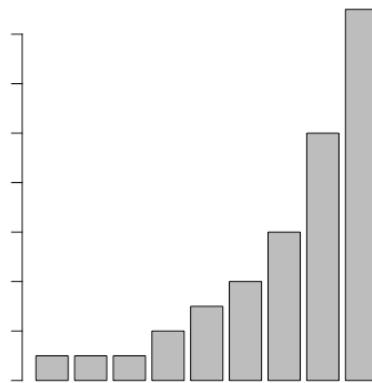
```
> FA_EV_P<-table(Población$EV_P)
> FA_EV_P
 0.7 0.95 1.06 1.75 1.97 1.99  3.5 3.58 4.71  4.8 4.81 4.86    5 5.03 5.28 5.52
6.4 6.62
  1   1   1   1   1   1   1   1   1   1   1   1   2   1   1   1   1
6.8 7.53 7.72 9.18
  1   1   1   1
```

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.

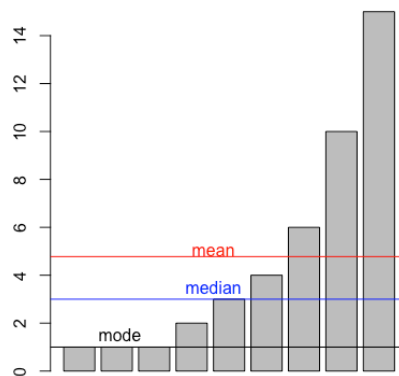
Ejemplo: Gráfico de la altura de una población de 9 árboles.



```
> 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	

4.9.- Parámetros de dispersión: rango, recorrido intercuartílico, varianza y desviación típica. Interpretación. Cálculo e interpretación. (3EAP8)

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 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.

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.

> 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:

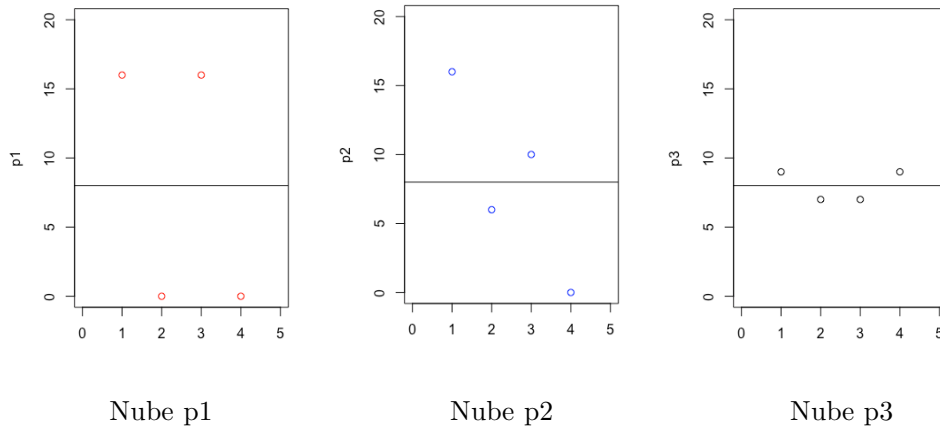
```
> plot( p1, col="red",xlim=c(0, 5),ylim=c(0, 20))
> abline(h=8)
```

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)
```



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

Coefficiente 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
```

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.xlsx`

```
> 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.

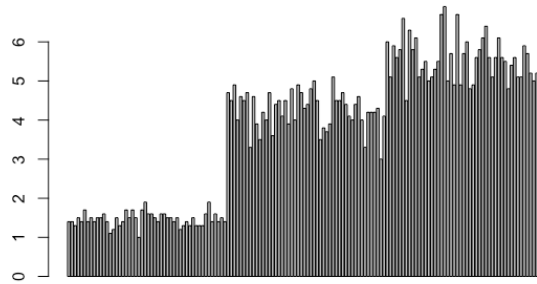
2.- Calcular las medias de las longitudes de los pétalos de la tabla `iris` que contiene la muestra de 150 plantas de 3 especies diferentes:

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)
```

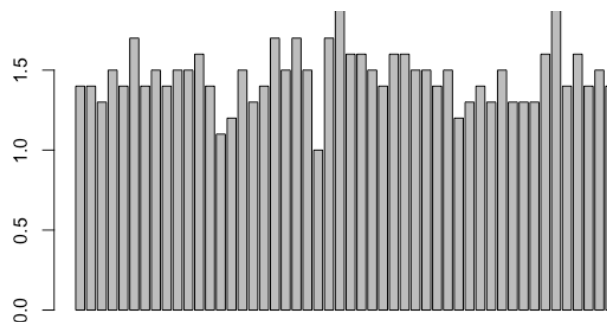
En el gráfico se aprecian tres escalones. Si nos fijamos en la tabla, las 50 primeras muestras son plantas de la especie Setosa, las 50 siguientes de la especie versicolor y las 50 últimas de la especie virgínica.

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
```

4.10.- Diagrama de caja y bigotes. Interpretación conjunta de la media y la desviación típica. (3EAP9)

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.

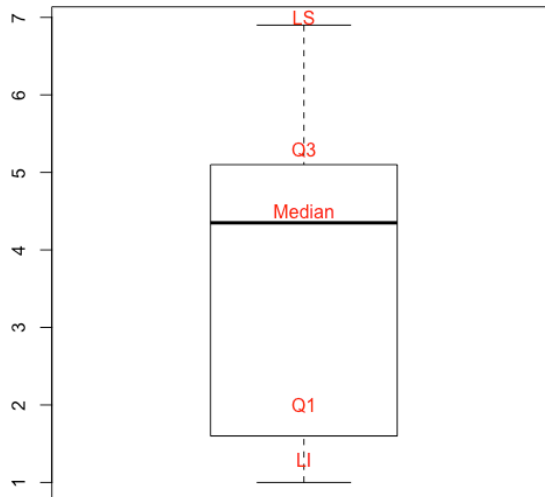
Pasos para construir un Box-Plot:

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.

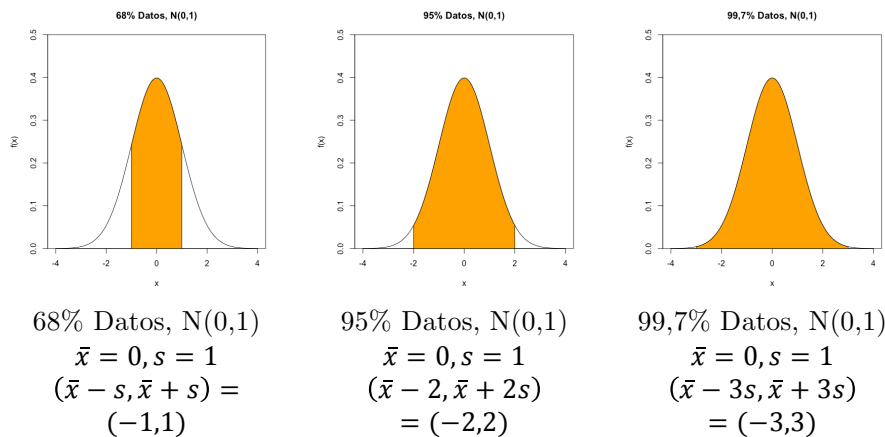


```
> boxplot(iris$Petal.Length)
> text(1.3, "LI", col = "red")
> text(7, "LS", col = "red")
> text(4.5, "Median", col = "red")
> text(5.3, "Q3", col = "red")
> text(2, "Q1", col = "red")
```

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.

4.11.- Estadística descriptiva bidimensional: Tablas de contingencia (1BC1).

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 . (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:

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

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

Distribuciones bidimensionales.

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

Las tablas bidimensionales simples adoptan la forma siguiente:

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

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

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

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

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

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

Ejemplo:

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

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

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

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

a) La tabla bidimensional simple, es:

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

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

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

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

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

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

4.12.- Distribución conjunta y distribuciones marginales. Distribución de frecuencias conjuntas (1BC2).

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

Tabla de frecuencias conjunta

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

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

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

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

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

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

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

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

4.13.- Distribuciones condicionadas (1BC4).

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

Frecuencia absoluta condicionada para X (x_i):

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

Frecuencia absoluta condicionada para Y (y_j) :

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

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

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

4.14.- Independencia de variables estadísticas (1BC5).

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

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

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

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

4.15.- Estudio de la dependencia de dos variables estadísticas. Representación gráfica: Nube de puntos (1BC6).

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:

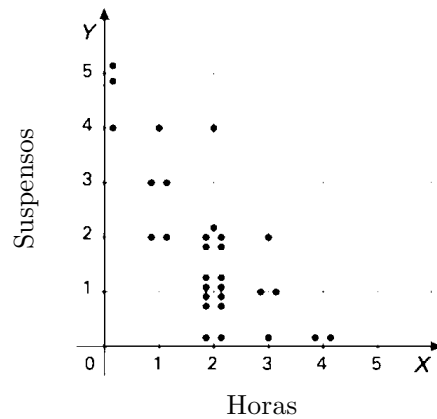
Y\X	0	1	2	3	4	Total
0	0	0	2	1	2	5
1	0	0	8	2	0	10
2	0	2	5	1	0	8
3	0	2	0	0	0	2
4	1	1	1	0	0	3
5	2	0	0	0	0	2
Total	3	5	4	4	2	30

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, trazariamos 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 0 2 0 0 0
4 1 1 1 0 0
5 2 0 0 0 0

```

Las tablas de frecuencias también son muy sencillas de construir:

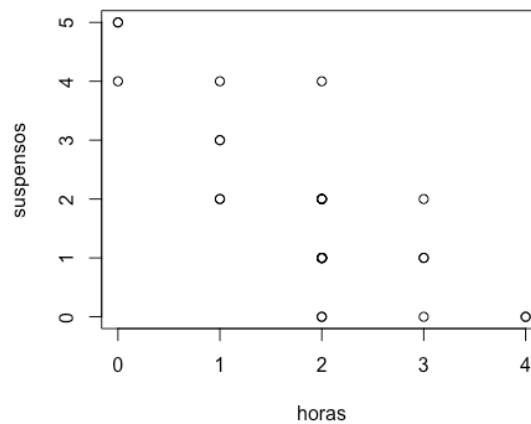
```

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

```

La nube de puntos se construye con la función plot:

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



Dependencia lineal:

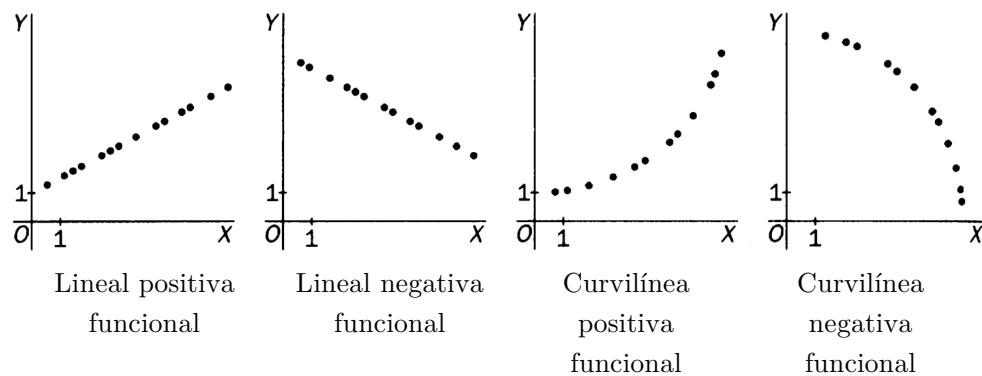
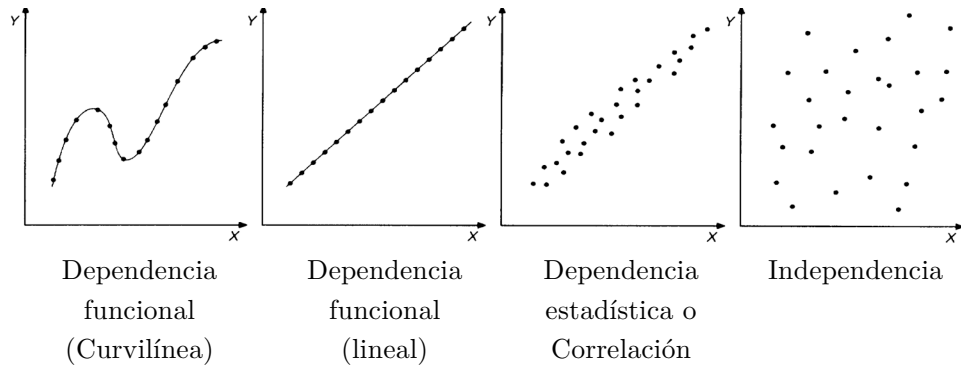
La forma de la nube de puntos, nos permite intuir si hay relación, o dependencia las dos variables. Esta dependencia, si existe, se llama correlación.

Existen varios tipos de dependencia:

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

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

Independencia o ausencia de correlación.



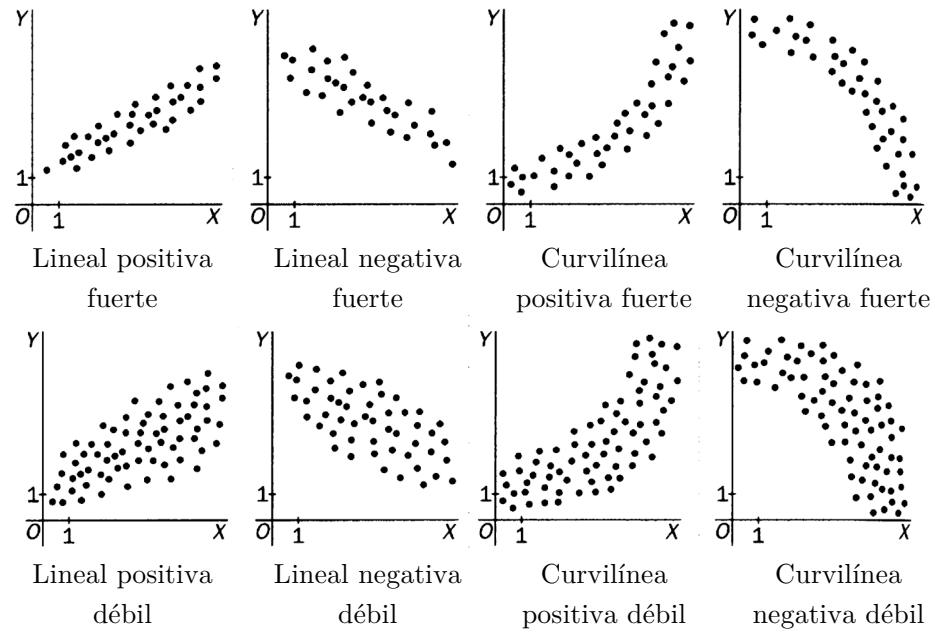
El grado de correlación, a su vez, puede ser:

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

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

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

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



4.16.- Dependencia lineal de dos variables estadísticas. Covarianza y correlación: Cálculo e interpretación del coeficiente de correlación lineal (1BC7).

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

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

Siendo

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

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

Donde:

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

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

Escala de valores del coeficiente de correlación lineal.

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

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

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

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

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

Ejemplo.

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

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

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

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

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

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

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

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

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

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

Estos valores nos permiten calcular el coeficiente de Pearson.

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

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

Vamos a ver cómo hacer esto con R.

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

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

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

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

[illegible]

El coeficiente de correlación de las notas sería:

> cor(M,F)
[1] 0.9194978

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

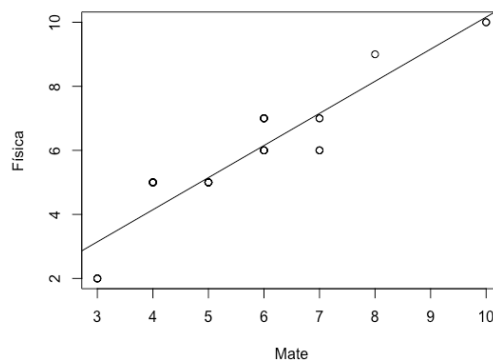
```
> NotasMF<-data.frame (Fisica=F,Mate=M)
> table(NotasMF)
```

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

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

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

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



4.17.- Regresión lineal. Predicciones estadísticas y fiabilidad de las mismas. Coeficiente de determinación (1BC8).

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

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

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

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

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

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

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

Estimaciones con las rectas de regresión.

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

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

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

Ejemplo:

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

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

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

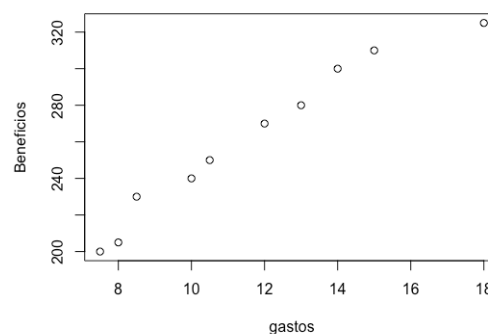
- El coeficiente de correlación lineal. Analiza la dependencia de ambas variables.
- La recta de regresión de Y sobre X.
- El volumen de ventas esperado, si la empresa decidiera invertir el próximo año 25000€ en publicidad.
- La inversión necesaria, si la empresa desea lograr 500000€ de beneficios.

Realizar este Ejemplo a mano, requiere los siguiente pasos:

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

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

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



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

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

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

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

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

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

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

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

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

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

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

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

Call:

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

Residuals:

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

Coefficients:

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

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

Residual standard error: 9.208 on 8 degrees of freedom

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

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

De este resumen, obtenemos las respuestas al apartado

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

Otra forma:

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

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

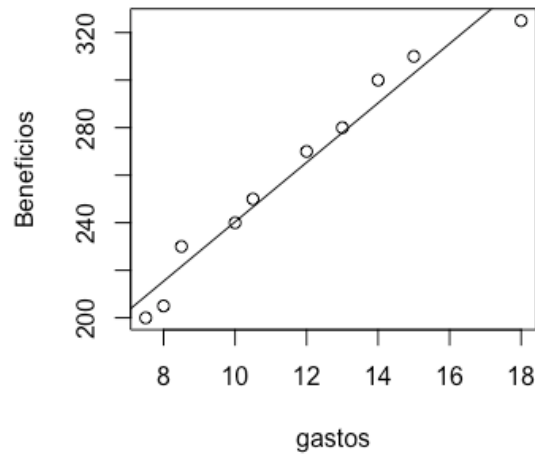
Coefficients:

	Estimate	Std.
(Intercept)	115.438	
gastos	12.495	

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

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

```
> abline(Reg_Gas_Ben)
```



c y d se calculan igual

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

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

4.18.- Estadística paramétrica. Parámetros de una población y estadísticos obtenidos a partir de una muestra. Estimación puntual (2BC2).

La estadística paramétrica es una rama de la estadística inferencial que comprende los procedimientos estadísticos y de decisión que están basados en las distribuciones de los datos reales. Estas son determinadas usando un número finito de parámetros. Esto es, por ejemplo, si conocemos que la altura de las

personas sigue una distribución normal, pero desconocemos cuál es la media y la desviación de dicha normal.

La media y la desviación típica de la distribución normal son los dos parámetros que queremos estimar.

Parámetro: Es un valor que representa a toda la población.

Estadístico: Es un valor que representa a toda la muestra. Cuando se utilizan para estimar parámetros, se les llama estimadores.

Lo normal es que nos interese conocer los parámetros. Esto implica estudiar al 100% de la población, lo que en ocasiones es imposible. Lo que haremos será tomar una muestra, y obtener un estimador. A partir de este estimador, obtendremos una aproximación al parámetro que buscamos.

La **media muestral** la representamos por \bar{x} , o por la letra m , y se define como:

$$\bar{x} = \frac{\sum_{i=1}^k x_i \cdot n_i}{n} = \sum_{i=1}^k x_i \cdot f_i$$

La **desviación típica muestral** la representamos por la letra s , y se define como:

$$s = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{n}}$$

Normalmente usaremos la media muestral y la desviación típica muestral como estadísticos.

La **media poblacional**, o la media de una distribución, la representamos por la letra griega μ y se define:

$$\mu = E(x) = \sum_{i=1}^N x_i \cdot p(x_i)$$

$$\mu = E(x) = \int_a^b x \cdot f(x) dx$$

La **desviación típica poblacional**, o de una distribución, la representamos por la letra griega σ y se define:

$$\sigma^2 = \sum_{i=1}^N (x_i - \mu)^2 \cdot p(x_i) = E(x^2) - E^2(x)$$

$$\sigma = \sqrt{E(x^2) - E^2(x)}$$

$$\sigma^2 = \int_a^b (x - \mu)^2 \cdot f(x) dx = 0$$

Normalmente usaremos la media poblacional y la desviación típica poblacional como **parámetros**.

Durante este proceso de obtención de los parámetros a través de los estadísticos, se pierde parte de la información de la población total. Por ejemplo, si en un grupo de tres personas una de ellas ingiere tres helados, el parámetro que con más frecuencia se utiliza para resumir datos estadísticos, la media aritmética del número de helados ingeridos por el grupo sería igual a 1, valor que no parece resumir fielmente la información.

Ninguna de las personas se sentiría identificada con la frase resumen: "He ingerido un helado de media".

4.19.- Media y desviación típica de la media muestral y de la proporción muestral. Distribución de la media muestral en una población normal. Distribución de la media muestral y de la proporción muestral en el caso de muestras grandes (2BC3).

El ajuste de los fenómenos a la distribución normal, se conoce como Teorema Central del Límite.

Si X es una variable aleatoria de una población de media μ finita y desviación típica σ finita, entonces:

Si tomamos un número n de muestras, y calculamos sus medias muestrales, éstas se distribuirán siguiendo una distribución normal de media μ y desviación típica $\frac{\sigma}{\sqrt{n}}$, a medida que crece el tamaño de la muestra (valores de $n > 30$).

Observa que tomando muestras, podemos conocer las medias y desviaciones típicas muestrales. El teorema nos dice que estas medias muestrales se irán "colocando" alrededor de la media poblacional desconocida, y con una desviación típica muestral de la que podemos extraer la desviación típica de la población original.

Las diferentes medias dan lugar a una variable aleatoria que vamos a representar por \bar{X} .

El Teorema Central del Límite nos garantiza que:

La media de la variable aleatoria \bar{X} es la media poblacional μ .

La desviación típica de la variable aleatoria \bar{X} es $\frac{\sigma}{\sqrt{n}}$, donde σ es la desviación típica poblacional y n es el tamaño de las muestras elegidas.

Para valores de n suficientemente grandes ($n \geq 30$), la distribución de \bar{X} se aproxima a una normal:

$$N(\mu, \frac{\sigma}{\sqrt{n}})$$

Ejemplo:

Los parámetros de una distribución normal son media $\mu = 10$ y desviación típica $\sigma = 2$. Se extrae una muestra de $n=100$ individuos. Calcula $P(8 < \bar{x} < 12)$.

Por el TCL, sabemos que la media muestral de una población normal se comporta como otra a distribución normal $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = N\left(10, \frac{2}{\sqrt{100}}\right) = N\left(10, \frac{2}{10}\right) = N(10, 0.2)$.

Para calcular la probabilidad que nos piden, tenemos que tipificar, para poder a continuación buscar en la tabla de la normal $N(0,1)$.

$$P(8 < \bar{x} < 12) = P\left(\frac{8-10}{0.2} < z < \frac{12-10}{0.2}\right) = P(-1 < z < 1) =$$

$$P(z < 1) - P(z < -1) =$$

$$2P(z < 1) - 1 = 2(0.8416) - 1 = 0.6832$$

Esto, que resulta laborioso con las tablas, en "R" se resume a una fórmula:

```
> pnorm(12,10,2)-pnorm(8,10,2)
[1] 0.6826895
```

Rstudio nos ahorra mucho tiempo, que podemos emplear para profundizar en los conceptos importantes. Por esto ampliaremos el Ejemplo y nos sobrará tiempo con respecto al uso de tablas y calculadora.

Ejemplo ampliado en RStudio:

Los parámetros de una distribución N $\mu = 10$ y desviación típica $\sigma = 2$. Se extrae una muestra de 100 individuos. Calcula la media y la desviación típica de la muestra que acabas de extraer.

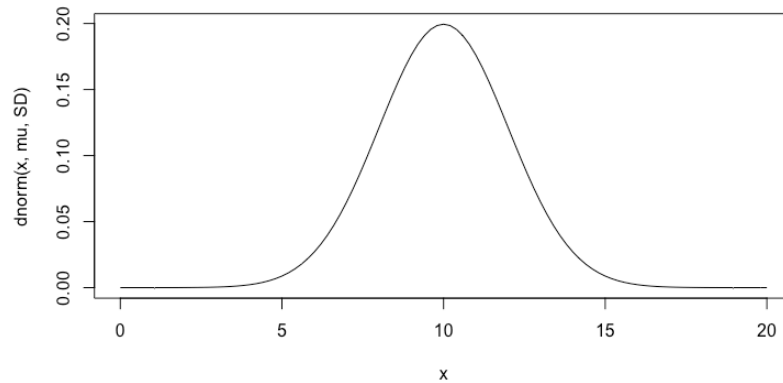
Haz lo mismo con muestras de 10, 100, 1000 y 10000 individuos, y compara las medias y desviaciones típicas poblacionales.

Finalmente, calcula $P(8 < \bar{x} < 12)$, para una muestra de 100 individuos.

Vamos a dibujar la curva:

```
> mu <- 10
> SD <- 2
```

```
> curve(dnorm(x, mu, SD), xlim = c(0,20))
```



Creamos la muestra de 100 individuos:

```
> muestra1 <- rnorm(100, mu, SD).
```

Nuestra muestra es la siguiente:

```
> muestra1
 [1] 11.080435  8.831985 10.123373 11.430427  9.155632  9.181577  9.827503 11.953020  8.891469
[10]  7.910513 12.579297 13.189111  6.734687  9.598388  7.913820 12.489522 15.102830  7.631180
[19]  9.368541  8.334997  9.357042  8.727217  7.854169  9.694912 12.528185 10.270649 11.610947
[28] 10.933908  9.417848 13.565228 11.208273  6.316510 11.951416 11.268456 10.024933 10.127910
[37]  3.627051 10.818820 11.828534 12.372750  9.533983 11.750733  9.202805 14.074441  7.803966
[46]  3.784245  9.875275 11.120781  6.168908  9.720031  8.384544  8.365658 12.025326 12.546405
[55]  7.447684 11.064643 11.386440  9.263883  8.330976  9.684476 10.824802 11.714280 12.109117
[64] 10.606221  7.879907 10.533375 10.031382  9.648328  5.757429 11.623302 10.484103  8.658699
[73]  9.260123  7.889299  6.237275  5.537658 12.642345  7.312772 10.300098  9.469288 12.273294
[82] 11.164070 12.476208 13.594020  7.942232  8.147640  9.497009  9.336401 10.808773  7.603004
[91] 10.134582  9.742037  8.094339  8.435306 13.117876  9.804331  5.713864  9.378496 10.916927
[100] 13.618820
```

Vamos a calcular la media y la desviación típica de nuestra muestra:

```
> mean(muestra1)
[1] 9.866873
> sd(muestra1)
[1] 2.189425
```

Vemos que no coinciden con la media y la desviación típica del enunciado, pero se aproximan.

Si tomáramos una muestra mucho mayor, la aproximación sería cada vez mejor.

Prueba con muestras de 10, 100, 1000 y 10000 individuos.

```
> muestra10 <- rnorm(10, mu, SD)
> muestra100 <- rnorm(100, mu, SD)
> muestra1000 <- rnorm(1000, mu, SD)
> muestra10000 <- rnorm(10000, mu, SD)
> mean(muestra10)
[1] 9.600026
> sd(muestra10)
[1] 2.038081
> mean(muestra100)
[1] 10.3521
> sd(muestra100)
[1] 2.031775
> mean(muestra1000)
[1] 9.939411
> sd(muestra1000)
[1] 1.96159
> mean(muestra10000)
[1] 9.988372
> sd(muestra10000)
[1] 2.003216
```

Calculamos por último la probabilidad $P(8 < \bar{x} < 12)$

La probabilidad de que un valor sea menor de 12, se calcularía:

```
> pnorm(12,10,2)
[1] 0.8413447
```

A esta probabilidad tenemos que restarle la probabilidad de que un valor sea menor que 8.

```
> pnorm(8,10,2)
[1] 0.1586553
```

Restando ambos valores se obtiene la solución al problema que nos plantean:

```
> 0.8413447-0.1586553
[1] 0.6826894
```

Directamente, podríamos haber escrito:

```
> pnorm(12,10,2)-pnorm(8,10,2)
[1] 0.6826895
```

Vamos ahora a crear un vector con las medias de muestras aleatorias de tamaño 100 de la siguiente manera:

```
> X<-c(mean(rnorm(100, mu, SD)), mean(rnorm(100, mu, SD)),
mean(rnorm(100, mu, SD)), mean(rnorm(100, mu, SD)), ... ,mean(rnorm(100,
mu, SD)))
```

Cada vez que añadamos otro “mean(rnorm(100, mu, SD))”, el vector X guardará una nueva media muestral de la muestra aleatoria de tamaño 100. Tras 62 elementos, el vector que contiene las medias muestrales tiene la siguiente pinta:

```
> sort (X)
[1] 9.538430 9.706938 9.709580 9.712088 9.824376 9.828127 9.833791 9.835903
[9] 9.842129 9.855327 9.865662 9.891405 9.896161 9.905310 9.923183 9.932834
[17] 9.937090 9.938889 9.947757 9.961318 9.961487 9.971978 9.973545 9.981441
[25] 9.998461 10.007454 10.013477 10.019264 10.036496 10.037037 10.048259 10.052356
[33] 10.053054 10.055682 10.057340 10.069083 10.076235 10.076599 10.078313 10.078521
[41] 10.085567 10.087228 10.118698 10.134970 10.157032 10.157227 10.178497 10.179751
[49] 10.187288 10.187952 10.199113 10.215878 10.235498 10.270368 10.305931 10.315877
[57] 10.316511 10.318490 10.325075 10.361024 10.375965 10.569114
```

Se ha ordenado para facilitar la visualización de cómo se van colocando las medias muestrales en torno a la media poblacional, que en el ejercicio valía 100

Tomando 10 muestras:

```
> mean(X)
[1] 10.06102
```

Tomando 20 muestras:

```
> mean(X)
[1] 10.01828
```

Tomando 30 muestras:

```
> mean(X)
[1] 9.995963
```

Tomando 40 muestras:

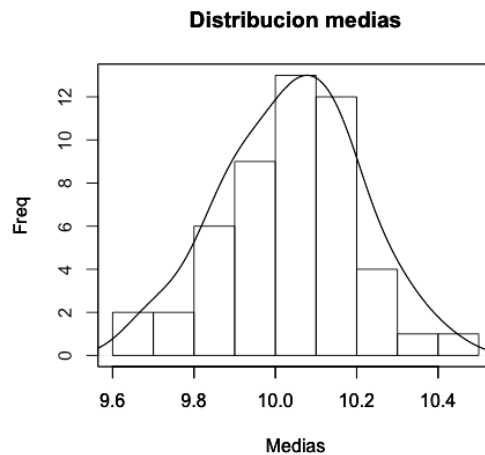
```
> mean(X)
[1] 10.03961
```

Tomando 50 muestras:

```
> mean(X)
[1] 10.03903
Tomando 145 muestras:
> mean(X)
[1] 10.05481
```

Se observa que a partir de 30 muestras no se consigue ganar mucha más precisión.

Con 50 muestras, la distribución que siguen las medias muestrales es la siguiente:



```
>plot (density (X),xlab='Medias',ylab='Freq',main='Distribucion medias',
xlim=c(9.6,10.5), yaxt="n"); par (new=TRUE)
>hist(X,xlab='Medias',ylab='Freq',main='Distribucion medias',
xlim=c(9.6,10.5) ,ylim=c(0,13))
```

Ejemplo:

Una fábrica produce bielas de bicicleta de 100 gramos, con una desviación típica de 2 gramos. Se crean lotes de 50 bielas. Calcula la probabilidad de que la media de las bielas de un lote sea menor que 99 gramos.

Tenemos la media poblacional $\mu = 100$, la desviación típica poblacional $\sigma = 2$, y el tamaño de la muestra, $n=50$.

Sabemos que la media se distribuye según una $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = N(100, 0.28)$. Para calcular estas probabilidades, tenemos que tipificar para pasar a una $N(0,1)$.

$$P(\bar{x} < 99) = P\left(Z < \frac{99 - 100}{0.28}\right) = N\left(\frac{99 - 100}{0.28}\right) = p(z < -3.54) = 1 - P(z < 3.54)$$

Nota: Como la distribución normal es simétrica, los valores negativos pueden calcularse usando los positivos.

Buscamos en la tabla 3.54 y obtenemos que

$$P(z < 3.54) = 0.9998.$$

$$P(\bar{x} < 99) = 1 - p(z < 3.54) = 1 - 0.9998 = 0.0002.$$

Una probabilidad muy baja.

Con "R"

> #Datos del enunciado. Declaración de variables:

> mu<-100

> sigma<-2

> n<-50

Sabemos por el TCL que la media se distribuye según una $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$:

> DT<-(sigma/sqrt(n))

> #Hallamos la probabilidad que nos piden:

> pnorm(99,mu,DT)

[1] 0.000203476

Ejemplo:

Calcula la probabilidad de que un lote de la fábrica anterior de 400 bielas pese más de 40100 gramos.

Como la media muestral es igual a $\bar{X} = \frac{\sum_{i=1}^N [x_i]}{N}$, entonces $\sum_{i=1}^N [x_i] = n\bar{x}$, por lo que su distribución es una normal de media $n\mu$ y desviación típica $n\frac{\sigma}{\sqrt{n}} = \sigma\sqrt{n}$:

$$N(n\mu, \sigma\sqrt{n}).$$

En nuestro caso, $N(n\mu, \sigma\sqrt{n}) = N(400 \cdot 100, 2\sqrt{400}) = N(40000, 40)$.

Queremos calcular $P(\sum_{i=1}^N [x_i] > 40100) = P(z > \frac{40100 - 40000}{40}) = 1 - P(z > 2.5) = 1 - 0.9938 = 0.0062$.

Unas 6 bielas de cada mil pesarán más de 40'1 kg.

Con "R"

Datos del enunciado. Declaración de variables:

```
> mu<-100
> sigma<-2
> n<-400
> Limite<-40100
```

Como la media muestral es igual a $\bar{X} = \frac{\sum_{i=1}^N [x_i]}{N}$, entonces $\sum_{i=1}^N [x_i] = n\bar{x}$, por lo que su distribución es una normal de media $n\mu$ y desviación típica $n\frac{\sigma}{\sqrt{n}} = \sigma\sqrt{n}$:

$$N(n\mu, \sigma\sqrt{n}).$$

```
> X_muest<-(n*mu)
> DT_muest<-(sigma*sqrt(n))
```

Cálculo de la probabilidad pedida:

```
> 1-pnorm(Limite,X_muest,DT_muest)
[1] 0.006209665
```

4.20.- Estimación por intervalos de confianza. Relación entre confianza, error y tamaño muestral (2BC4).

Supongamos una población de la que deseamos conocer, por ejemplo, su media. Para ello seleccionamos una muestra aleatoria de la que podemos calcular su media muestral. Este valor se conoce como estimador o estimador puntual. Con este estimador se puede inferir con una cierta probabilidad, el parámetro que buscamos. El proceso que hemos seguido, se conoce como estimación puntual. Al hacer esto, no podemos asegurar que no haya error. Al error cometido de le denomina sesgo.

Un estimador es insesgado o centrado, cuando su sesgo sea nulo por ser su esperanza igual al parámetro que se desea estimar. La media muestral y la proporción muestral son estimadores centrados.

La varianza de un estimador define la eficiencia del mismo. Concretamente, para medir la eficiencia de un estimador centrado se utiliza la inversa de la varianza.

Intervalos de confianza

Cuando queremos conocer un parámetro de una población, tomamos una muestra, partir de la cual calculamos un intervalo que contiene a dicho parámetro, con un nivel de confianza. Este intervalo se conoce como intervalo de confianza. A la probabilidad de que el estimador se encuentre dentro de dicho intervalo se denomina nivel de confianza.

Método:

Elegimos un estimador $t(X)$, que será el estadístico pivote. Tiene que cumplir:

- 1.- Debe estar relacionado con el parámetro θ que queremos estimar.
- 2.- Debemos conocer su distribución de probabilidad, e ésta no debe depender del valor de θ .

Utilizando la distribución de probabilidad de $t(X, \theta)$, y con el nivel de confianza que nos pidan " γ ", se calculan los valores críticos k_1 y k_2 .

Conceptos:

Intervalo de confianza: Si $P(a < X < b) = 0.95$, El intervalo de confianza es (a, b) .

Nivel de confianza o coeficiente de confianza: $1 - \alpha = \gamma$, en nuestro ejemplo, será 0.95.

Nivel de significación o de riesgo: α , en nuestro ejemplo, 0.05.

Valor crítico: k_1 y k_2 , que dejan a la derecha, o a la izquierda, un área $\alpha/2$. En la $N(0, 1)$ son -1.96 y 1.96 para $\alpha = 0.05$.

Margen de error: Diferencia entre los extremos del intervalo de confianza.

Máximo error admisible: Valor que no puede exceder el valor absoluto de la diferencia entre el estimador y el parámetro.

Ejemplo:

Determina un intervalo de confianza y el margen de error con un nivel de confianza del 0.95 de una $N(2, 0.1)$.

El margen de error es la distancia entre los extremos del intervalo de confianza. Lo calculamos para la $N(2, 0.1)$. Sabemos que en una $N(0, 1)$:

$$P(-1.96 < Z < 1.96) = 0.95 \Rightarrow$$

Para la $N(2, 0.1)$:

$$P(-1.96 < \frac{x - 2}{\sqrt{0.1}} < 1.96) = 1 - \alpha = \gamma = 0.95 \Rightarrow$$

Operando:

$$P\left((0.1(-1.96)) + 2 < X < (0.1 \cdot 1.96) + 2\right) = 0.95 \Rightarrow$$

$$P(1.8 < X < 2.2) = 0.95;$$

La variable aleatoria X estará en el intervalo $(1.8, 2.2)$ con un nivel o coeficiente de confianza de 0.95.

El margen de error viene dado por la amplitud del intervalo:

Margen de error: $2.2 - 1.8 = 0.4$.

Veamos como hacerlo en R.

Instalamos y cargamos la librería TeachingDemos:

```
>install.packages("TeachingDemos")
> library(TeachingDemos)
> z.test(dnorm(2,0.1), mu=2, stdev=0.1, conf.level=0.95)
One Sample z-test
data:  dnorm(2, 0.1)
z = -19.344, n = 1.0, Std. Dev. = 0.1, Std. Dev. of the sample mean =
0.1, p-value <
2.2e-16
alternative hypothesis: true mean is not equal to 2
95 percent confidence interval:
-0.1303806  0.2616122
sample estimates:
mean of dnorm(2, 0.1)
0.06561581
```

El intervalo de confianza al 95% al estar centrado en la media, se obtiene sumando y restando los valores que nos devuelve en “95 percent confidence interval”: -0.1303806 0.2616122:

```
> 2-0.1303806
[1] 1.869619
> 2+0.2616122
[1] 2.261612
El margen de error es:
> 2.261612-1.869619
[1] 0.391993
```

Capítulo 5. Estándares de aprendizaje evaluables.

Estándares 1º ESO. Aplicadas.

1.1. Define población, muestra e individuo desde el punto de vista de la estadística, y los aplica a casos concretos.

1.2. Reconoce y propone ejemplos de distintos tipos de variables estadísticas, tanto cualitativas como cuantitativas.

1.3. Organiza datos, obtenidos de una población, de variables cualitativas o cuantitativas en tablas, calcula sus frecuencias absolutas y relativas, y los representa gráficamente.

1.4. Calcula la media aritmética, la mediana y la moda y los emplea para resolver problemas.

2.1. Emplea la calculadora y herramientas tecnológicas para organizar datos, y calcular las medidas de tendencia central.

3.1. Identifica los experimentos aleatorios y los distingue de los deterministas.

3.2. Calcula la frecuencia relativa de un suceso mediante la experimentación.

3.3. Realiza predicciones sobre un fenómeno aleatorio a partir del cálculo exacto de su probabilidad o la aproximación de la misma mediante la experimentación.

4.1. Describe experimentos aleatorios sencillos y enumera todos los resultados posibles, apoyándose en tablas, recuentos o diagramas en árbol sencillos.

Estándares 2º ESO.

1.1. Define población, muestra e individuo desde el punto de vista de la estadística, y los aplica a casos concretos.

1.2. Reconoce y propone ejemplos de distintos tipos de variables estadísticas, tanto cualitativas como cuantitativas.

1.3. Organiza datos, obtenidos de una población, de variables cualitativas o cuantitativas en tablas, calcula sus frecuencias absolutas y relativas, y los representa gráficamente.

1.4. Calcula la media aritmética, la mediana (intervalo mediano), la moda (intervalo modal), y el rango, y los emplea para resolver problemas.

1.5. Interpreta gráficos estadísticos sencillos recogidos en medios de comunicación.

2.1. Emplea la calculadora y herramientas tecnológicas para organizar datos, generar gráficos estadísticos y calcular las medidas de tendencia central y el rango de variables estadísticas cuantitativas.

2.2. Utiliza las tecnologías de la información y de la comunicación para comunicar información resumida y relevante sobre una variable estadística analizada.

3.1. Identifica los experimentos aleatorios y los distingue de los deterministas.

3.2. Calcula la frecuencia relativa de un suceso mediante la experimentación.

3.3. Realiza predicciones sobre un fenómeno aleatorio a partir del cálculo exacto de su probabilidad o la aproximación de la misma mediante la experimentación.

4.1. Describe experimentos aleatorios sencillos y enumera todos los resultados posibles, apoyándose en tablas, recuentos o diagramas en árbol sencillos.

4.2. Distingue entre sucesos elementales equiprobables y no equiprobables.

4.3. Calcula la probabilidad de sucesos asociados a experimentos sencillos mediante la regla de Laplace, y la expresa en forma de fracción y como porcentaje.

Estándares 3º ESO.

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.

3.2. Emplea la calculadora y medios tecnológicos para organizar los datos, generar gráficos estadísticos y calcular parámetros de tendencia central y dispersión.

3.3. Emplea medios tecnológicos para comunicar información resumida y relevante sobre una variable estadística analizada.

Estándares 4º ESO.

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.

Estándares 1º BACH. Aplicadas.

- 1.1. Elabora tablas bidimensionales de frecuencias a partir de los datos de un estudio estadístico, con variables discretas y continuas.
- 1.2. Calcula e interpreta los parámetros estadísticos más usuales en variables bidimensionales.
- 1.3. Calcula las distribuciones marginales y diferentes distribuciones condicionadas a partir de una tabla de contingencia, así como sus parámetros (media, varianza y desviación típica).
- 1.4. Decide si dos variables estadísticas son o no dependientes a partir de sus distribuciones condicionadas y marginales.
- 1.5. Usa adecuadamente medios tecnológicos para organizar y analizar datos desde el punto de vista estadístico, calcular parámetros y generar gráficos estadísticos.
- 2.1. Distingue la dependencia funcional de la dependencia estadística y estima si dos variables son o no estadísticamente dependientes mediante la representación de la nube de puntos.
- 2.2. Cuantifica el grado y sentido de la dependencia lineal entre dos variables mediante el cálculo e interpretación del coeficiente de correlación lineal.
- 2.3. Calcula las rectas de regresión de dos variables y obtiene predicciones a partir de ellas.
- 2.4. Evalúa la fiabilidad de las predicciones obtenidas a partir de la recta de regresión mediante el coeficiente de determinación lineal.
- 3.1. Describe situaciones relacionadas con la estadística utilizando un vocabulario adecuado.

Estándares 2º BACH.

- 2.1. Valora la representatividad de una muestra a partir de su proceso de selección.
- 2.2. Calcula estimadores puntuales para la media, varianza, desviación típica y proporción poblacionales, y lo aplica a problemas reales.
- 2.3. Calcula probabilidades asociadas a la distribución de la media muestral y de la proporción muestral, aproximándolas por la distribución normal de parámetros adecuados a cada situación, y lo aplica a problemas de situaciones reales.
- 2.4. Construye, en contextos reales, un intervalo de confianza para la media poblacional de una distribución normal con desviación típica conocida.
- 2.5. Construye, en contextos reales, un intervalo de confianza para la media poblacional y para la proporción en el caso de muestras grandes.

2.6. Relaciona el error y la confianza de un intervalo de confianza con el tamaño muestral y calcula cada uno de estos tres elementos conocidos los otros dos y lo aplica en situaciones reales.

Capítulo 6. Bibliografía.

Software Rstudio:

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

Manual de R:

Título: R para profesionales de los datos: una introducción

Autor: Carlos J. Gil Bellosta

Fecha: 2018-04-22

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

Tablas de contingencia en R:

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

Histogramas en R:

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

Librerías en R:

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

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

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

Estructuras de datos en R:

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

Creación de data.frames en R:

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

Correlación lineal en R:

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

Curso de introducción a la Estadística:

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

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

Contenidos y estándares oficiales:

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

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

ORDEN EDU/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>

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

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

Libros de texto:

Fundamentos y métodos de Estadística, 3ª Edición,:

Autores: M. López Cachero

Editorial: Ed Piramide

ISBN 84-368-0171-7

Estadística aplicada a las ciencias de la educación:

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

Editorial: Ed Santillana

ISBN: 84-294-1903-9

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

Matemáticas 1º Bachillerato:

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

Editorial: EDITEX S.A

ISBN: 9788490785652

Matemáticas 2º Bachillerato:

Apuntes marea verde.

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

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

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

Apuntes Metodología y Evaluación del Máster Universitario de Profesor en Educación Secundaria Obligatoria y Bachillerato. Especialidad de Matemáticas

http://campusvirtual2017.uva.es/pluginfile.php/451950/mod_resource/content/1/6.-%20Recursos.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

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>

Referencias a Libros recomendados:

How to lie with Statistics:

http://almena.uva.es/search~S1*sp?/tcomo+mentir/tcomo+mentir/1%2C1%2C2%2CB/frameset&FF=tcomo+mentir+con+estadisticas&1%2C%2C2

El tigre que no está:

DivulgaMat

http://vps280516.ovh.net/divulgamat15/index.php?option=com_content&view=article&id=9810:el-tigre-que-no-estun-paseo-por-la-jungla-de-la-estadica&catid=53:libros-de-divulgaciatemca&Itemid=35

Inferencia estadística:

http://proyectodescartes.org/iCartesiLibri/materiales_didacticos/EstadisticaProbabilidadInferencia/Estadistica1D/1Introduccion.html

Imágenes flores iris:

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