

Valladolid, Junio 2018.

Índice general.

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

Capítulo 1. Introducción.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Capítulo 2. Contenidos y estándares oficiales.

2º Bachillerato Ciencias Sociales

- 2BC1.- Población y muestra. Métodos de selección de una muestra. Tamaño y representatividad de una muestra.
- 2BC2.- Estadística paramétrica. Parámetros de una población y estadísticos obtenidos a partir de una muestra.
- 2BC3.- 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.

Capítulo 3. Introducción a “R” Statistics.

3.1 Sobre “R”:

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

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

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

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

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

3.2 Descarga e instalación de RStudio:

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

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

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

Installers	Size	Date	MD5
RStudio 1.1.453 - Windows Vista/7/8/10	85.8 MB	2018-05-16	bf287e385aef53829204023087e98735
RStudio 1.1.453 - Mac OS X 10.6+ (64-bit)	74.5 MB	2018-05-16	00a0088424ed06ac434f7a966f602b9c
RStudio 1.1.453 - Ubuntu 12.04-15.10/Debian 8 (32-bit)	89.3 MB	2018-05-16	6cfd86770c7b6dbc13e66f4f59c299ce
RStudio 1.1.453 - Ubuntu 12.04-15.10/Debian 8 (64-bit)	97.4 MB	2018-05-16	63e36e8138e369d19f9aaf4b0e995bbc
RStudio 1.1.453 - Ubuntu 16.04+/Debian 9+ (64-bit)	64.4 MB	2018-05-16	85b3e76c9fad4613bc9cf0de1f34b183
RStudio 1.1.453 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	88.1 MB	2018-05-16	37cade7e162eab62483e6556e39dedee
RStudio 1.1.453 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	90.6 MB	2018-05-16	44cddd285bc31c41e4eaeed174b8eebb

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

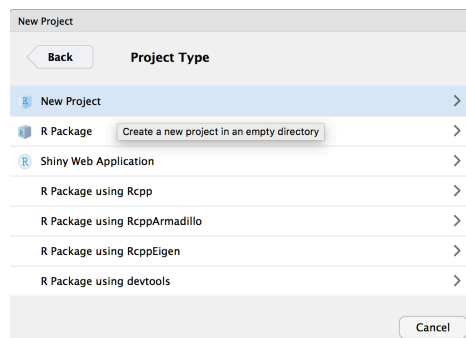
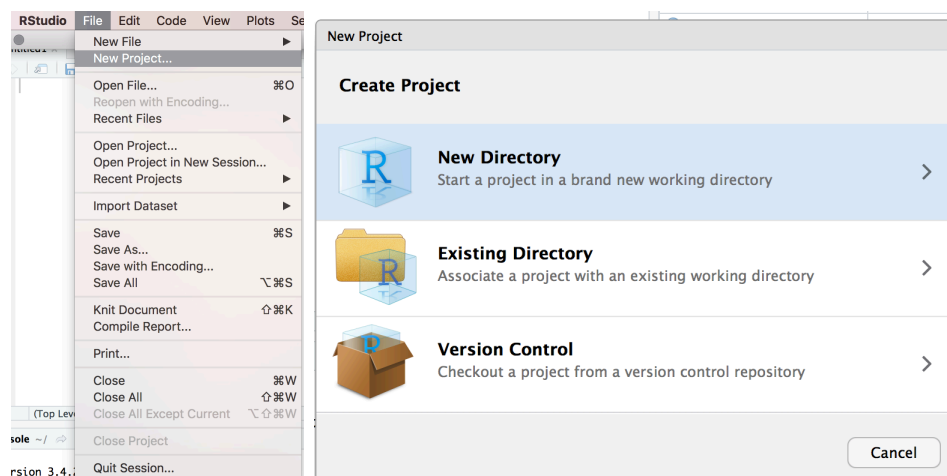
Una vez descargado, se ejecuta el programa instalador, y se van siguiendo los pasos del asistente de instalación, como en cualquier otro programa.

3.3.- Creación de un nuevo proyecto:

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

Para ello, abrir Rstudio y pulsar:

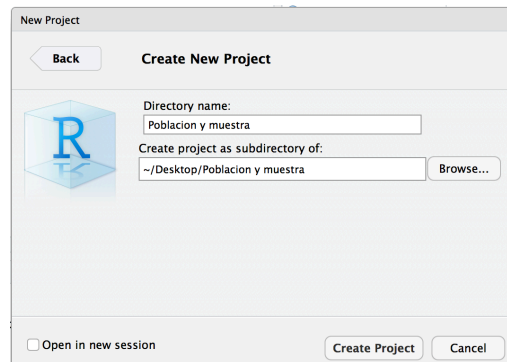
File/New project.../New directory/New Project



En la siguiente ventana, escribimos:

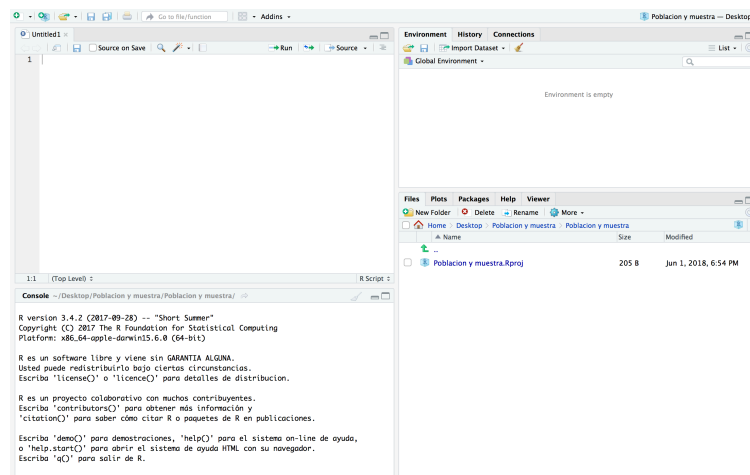
Directory name: Población y muestra.

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



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

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



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

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

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

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

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

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

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

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

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

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

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

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

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

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

2BC1.- Población y muestra. Métodos de selección de una muestral. Tamaño y representatividad de una muestra.

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 componenets 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.

Un **muestreo** es un proceso mediante el cual se selecciona la muestra de la población.

El **tamaño de la muestra** es el número de los elementos que contiene.

Cuando la muestra comprende a todos los elementos de la población, se denomina **censo**.

Métodos de selección de una muestra.

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

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

Si N es el tamaño de la población y n el de la muestra, entonces:

a) Fracción de muestreo: $f=n/N$. Indica el porcentaje de población encuestada.

b) Factor o coeficiente de elevación: $k=1/f =N/n$. Indica el número de unidades que hay en la población, por cada elemento de la muestra.

Ejemplo:

El número de viviendas de un Barrio es 10.000, si tomamos una muestra de 400 viviendas, el porcentaje de encuestados es el 4%, y cada vivienda entrevistada representa a 25 viviendas del barrio.

Los métodos de muestreo se dividen en dos bloques:

Muestreo no probabilístico: no se usa el azar, sino el criterio del investigador, suele presentar grandes sesgos y es poco fiable. Los métodos de muestreo no probabilístico más utilizados son: Muestreo por cuotas, Muestreo por rutas aleatorias, Muestreo "bola de nieve".

Muestreo probabilístico: se utilizan las leyes del azar mediante diferentes procedimientos de muestreo que se describen a continuación:

1. Muestreo aleatorio simple (m.a.s) (es el más importante y básico):

Consiste en seleccionar una muestra de tamaño n de una población de tamaño N con elementos homogéneos, de modo que cada elemento de la población tiene la misma probabilidad de ser elegido para la muestra, las observaciones se realizan con reemplazamiento, de manera que la población es idéntica en todas las extracciones, o sea, que la selección de un individuo no debe afectar a la probabilidad de que sea seleccionado otro cualquiera aunque ello comporte que algún individuo pueda ser elegido más de una vez. ("se hacen tantas papeletas numeradas como individuos hay, se coge una y se devuelve, se vuelve a coger otra y se devuelve, etc.").

También se pueden realizar sin reemplazamiento, de forma que al no ser devuelta la unidad otra vez a la población, la probabilidad de que salga un elemento depende de las extracciones anteriores, en el caso de ser una población finita. El número total de muestras posibles será igual al número de combinaciones que se pueden hacer con N elementos tomados de n en n .

Para elegir una muestra de una población finita se utilizan frecuentemente los números aleatorios (tabla elaborada para ese fin o generándolos con el ordenador).

2. Muestreo sistemático:

Se utiliza cuando el número de elementos de la población es elevado y están ordenados por listas. Se toma un individuo al azar y a continuación a intervalos constantes se eligen todos los demás hasta completar la muestra. Si el orden de los elementos es tal que los individuos próximos tienden a ser más semejantes que los alejados, el muestreo sistemático tiende a ser más preciso que el aleatorio simple, al cubrir más homogéneamente toda la población. Si se sospecha que pueda presentarse algún tipo de periodicidad, se procedería a cambiar, al azar, cada cierto tiempo el punto de partida o se utilizaría el muestreo aleatorio simple.

Los intervalos constantes se calculan tomando el valor recíproco de la fracción de muestro, que recibe el nombre de factor de elevación k .

Ejemplo: Se necesita tomar una muestra de 250 estudiantes de un colegio que cuenta con 1.000, entonces el intervalo de selección sería 4. Para iniciarla, se toma un número al azar entre 1 y 4, a partir de él, se aplica a la lista el intervalo 4. Si por ejemplo elegimos el 2, la muestra estaría compuesta por las personas 2, 6, 10, 14, 18, 22,.....

3. Muestreo estratificado:

Se utiliza sobre todo en encuestas de opinión, donde los elementos (personas) son heterogéneos en razón de su sexo, edad, etc.... Interesa que la muestra tenga la misma composición a la de la población la cual se divide en clases o estratos. Si por ejemplo en la población el 20% son mujeres y el 80% hombres, se mantendrá la misma proporción en la muestra. La muestra se toma asignando un número o cuota de miembros a cada estrato, esto se conoce en la literatura estadística como afijación de la muestra y escogiendo los elementos por m.a.s. dentro de cada estrato.

Existen diversos tipos de afiliación, cuya utilización depende de las características de la población a investigar, son entre otros afiliación simple (es el menos recomendable, consiste en repartir la muestra total en partes iguales para cada estrato), afiliación proporcional (es el más utilizado, consiste en dividir la muestra total en partes proporcionales a la población de cada estrato) , afiliación óptima, óptima por costes variables, valoral, etc.

Ejemplo: El número de viviendas de un barrio es de 10.000 y tenemos 2.000 del tipo A, 7.000 del tipo B y 1.000 del tipo C. Se realizan 400 entrevistas ¿Cómo se dividiría la muestra, utilizando la afiliación proporcional?

Si consideramos n_1 , n_2 y n_3 al tamaño de la muestra en cada uno de los estratos, tendríamos:

$$n_1/2000=n_2/7000=n_3/1000=400/10000$$

Es decir: $n_1=(2000*400/10000) = 8$, $n_2= 280$ y $n_3 = 40$

4. Muestreo por conglomerados:

La muestra se obtiene al seleccionar directamente cierto número de grupos en los cuales aparecen distribuidas las unidades de la población original – llamados conglomerados – . En cada etapa del muestreo, en lugar de seleccionar elementos al azar, seleccionamos conglomerados tan homogéneos entre sí como sea posible y heterogéneos internamente como la población de estudio.

Puede ser:

a) Monoetápico: En una sola etapa.

b) Bietápico: En dos etapas.

c) Polietápico: generalización para un número cualquiera de etapas. Se utiliza para investigar poblaciones complejas, donde se combinan las ideas de conglomerados y estratificados con el muestro aleatorio simple entre las unidades finales del muestreo. Consiste en muestrear conglomerados dentro de cada conglomerado.

Tamaño y 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 muestra. Si la muestra es demasiado pequeña, aunque esté bien elegida, el resultado no será fiable.

Queremos estudiar la estatura de la población española. Para ello elegimos a una persona al azar y la medimos.

Evidentemente este resultado no es fiable. La muestra es demasiado pequeña.

Si la muestra es demasiado grande los resultados serán muy fiables, pero el gasto puede ser demasiado elevado. Incluso, en ocasiones, muestras demasiado grandes no nos proporcionan mejores resultados. Vamos a aprender a encontrar cual es el tamaño adecuado para que podamos afirmar que la población tiene tal característica con una probabilidad dada, grande.

Cuando una muestra tenga el tamaño adecuado, y haya sido elegida de forma aleatoria diremos que es una muestra representativa.

Si la muestra no ha sido elegida de forma aleatoria diremos que la muestra es sesgada.

Ejercicios:

Indica si es población o muestra:

1) En una ganadería se mejora el pienso de todas las ovejas con un determinado tipo de grano.

2) En otra ganadería se seleccionan 100 ovejas para alimentarlas con ese tipo de grano y estudiar su eficacia.

En el primer caso, todas las ovejas, son la población. En el segundo se ha elegido una muestra.

En una serie de televisión tienen dudas sobre qué hacer con la protagonista, si que tenga un accidente o si debe casarse. Van a hacer una consulta. ¿A toda la población o seleccionado una muestra representativa?

Observa que no sabemos bien cual seria la población, ¿los que ven esa serie? o ¿toda la población española? Si son los que ven la serie, ¿cómo los conocemos? ¿Cómo preguntar a todos? Parece más operativo preguntar a una muestra.

El estudio de la vida media de unas bombillas, ¿se puede hacer sobre toda la población?

El estudio es destructivo. Si se hiciera sobre toda la población nos quedamos sin bombillas. Es imprescindible tomar una muestra.

3. Para estudiar el número de accidentes de una población de mil conductores, de los cuales la mitad tiene carnet de conducir entre 5 y 20 años, la cuarta parte lo tiene más de 20 años y la otra cuarta parte lo tiene menos de 5 años. Se quiere elegir por muestreo aleatorio estratificado proporcional, 50 conductores, ¿cuántos seleccionarías de cada grupo?

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

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:

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

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

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

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".

2BC3.- 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.

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.

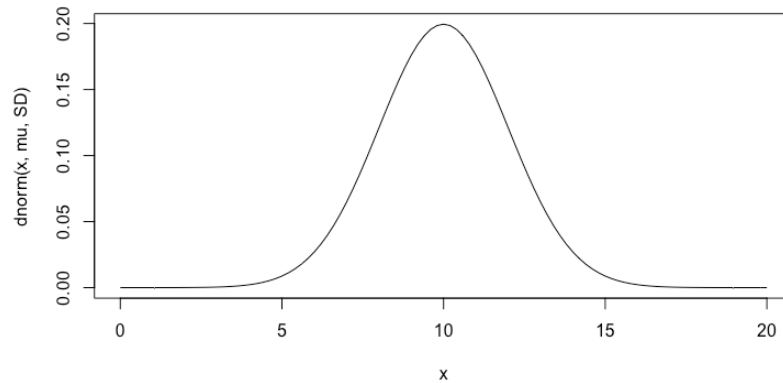
Haz lo mismo con muestras de 10, 100, 1000 y 10000 individuos, y compara las medias y desviaciones típicas poblacionales.

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:

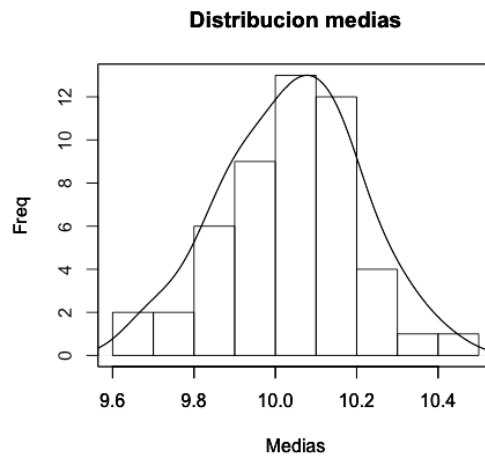
```
> mean(X)
[1] 10.03961
```

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
```

BC4.- Estimación por intervalos de confianza. Relación entre confianza, error y tamaño muestral.

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 $N(2, 0.1)$:

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

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 2º Bachillerato Ciencias Sociales

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.

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

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

Libros de texto:

Fundamentos y métodos de Estadística, 3ª Edición,:

Autores: M. López Cachero

Editorial: Ed Piramide

ISBN 84-368-0171-7

Estadística aplicada a las ciencias de la educación:

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

Editorial: Ed Santillana

ISBN: 84-294-1903-9

Matemáticas 4º ESO

Autores: Fernando Alcalde, Joaquín Hernandez...

Editorial: EDICIONES SM

ISBN: 9788467586930

Matemáticas 1º Bachillerato:

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

Editorial: EDITEX S.A

ISBN: 9788490785652

Matemáticas 2º Bachillerato:

Apuntes marea verde.

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

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

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

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

Otros:

Definiciones de Medidas de centralización y de dispersión:

Wikipedia

Referencia censos bíblicos:

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

Definiciones. Encuesta, censos:

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

Imágenes flores iris:

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