UNIVERSIDAD NACIONAL AUTÓNOMA DE NICARAGUA-LEÓN

FACULTAD DE CIENCIAS Y TECNOLOGÍA

DEPARTAMENTO DE MATEMÁTICA, ESTADÍSTICA Y CIENCIAS ACTUARIALES

MONOGRAFÍA PARA OPTAR AL TÍTULO DE LICENCIADO EN CIENCIAS ACTUARIALES Y FINANCIERAS:

MODELO DE DISTRIBUCIÓN DE LA CUANTÍA DE LOS SINIESTROS: UNA INTRODUCCIÓN AL *R*

AUTORES:

- BR. TORRES COREA DALILA MERCEDES
- BR. VÍLCHEZ NARVÁEZ KAREN WALQUIRIA

TUTOR:

- ACTUARIO ÁLVARO ARAUZ

¨A LA LIBERTAD POR LA UNIVERSIDAD¨

TEMA

“MODELO DE DISTRIBUCIÓN DE LA CUANTÍA DE LOS SINIESTROS: UNA INTRODUCCIÓN AL R”

DEDICATORIA

A DIOS:

Por habernos dado sabiduría, y habernos permitido alcanzar nuestras metas y, así demostrar que no importan las adversidades para salir adelante.

A **Nuestros padres:**

Por darnos todo lo que tenemos y lo que somos, por sus consejos, sus valores, por su motivación constante, y sobre todo por su amor. Por brindarnos su apoyo, y estar siempre pendiente de nosotros.

AGRADECIMIENTO

Es nuestro deseo mencionar a quienes hicieron posible llevar a cabo este trabajo tan importante para la finalización exitosa de esta etapa académica. Para todas aquellas personas que hicieron posible esto, nuestro agradecimiento infinito por ser el soporte de nuestros logros.

A Dios:

En primer lugar agradecemos a Dios, nuestro padre celestial al cual le debemos todo y que sin su voluntad no hubiéramos concluido este trabajo. Por estar con nosotros siempre y en cada momento de nuestras vidas, el cual fue el motor que impulsó la idea de ser cada día mejores.

A nuestras familias:

Por estar ahí siempre con su apoyo, sus alientos y sus comentarios que servían de estímulo para auto exigirnos y elevar nuestras exigencias.

A nuestros maestros:

Queremos también de manera muy especial, **agradecer a nuestros profesores** de cada uno de los años recorridos en este proceso, que con sus constantes exigencias y su interés en proporcionar las armas y herramientas para hacer de cada uno de nosotros, un elemento social preparado académico y moralmente para enfrentar las fases de la vida secular laboral. En especial al MSc. Milton Carvajal, por brindarnos de su tiempo y conocimiento en la realización de esta tesis.

A nuestro tutor:

Queremos hacer una mención especial a la persona que con fe en nosotras tomo la responsabilidad de guiarnos en este camino, y desde el primer día que le propusimos este tema tuvo la constancia e interés profesional para sabernos guiar, por su tolerancia y fomento a amar cada vez más la rama que hemos decidido tomar, las gracias no son suficiente por su tiempo brindado, por la enseñanza que traspasan las teorías y a motivarnos a dar siempre lo mejor de nosotros para con este trabajo. A nuestro tutor Licenciado Álvaro Arauz.

Gracias a todos ellos por permitirnos alcanzar esta meta.

ÍNDICE

I. INTRODUCCIÓN

Las ciencias actuariales tiene por objeto la construcción de modelos que expliquen los fenómenos aleatorios denominados actuariales, que se enfrentan a una realidad que necesita de soluciones ante problemas suscitados en relación con los fenómenos que la integran, siniestralidad, mortalidad, ruina, supervivencia, por ejemplo.

Hoy en día existen en el mercado gran cantidad de programas, sin embargo, no satisfacen 100% los requerimientos de los usuarios, por eso es frecuente que se utilicen en las empresas aplicaciones informáticas. La gran mayoría de las computadoras cuentan con el sistema operativo de Microsoft que utiliza imágenes y símbolos. Este sistema contiene varias aplicaciones informáticas en Microsoft Office, como Word, Excel, PowerPoint, Access, entre otras. Las áreas administrativas de las empresas necesitan entregar cada día más información que les ayude a la toma de decisiones, pero los programas existentes no satisfacen sus necesidades y por ello utilizan las aplicaciones, sobre todo los actuarios. Dentro de las aplicaciones más usadas por los actuarios se encuentran Excel y SPSS, del cual se utiliza únicamente Excel (llamado sistema actuarial informatizado), ya que este es utilizado para cálculos, manejos de grandes volúmenes de datos numéricos, como clientes, proveedores, pasivos, nómina, inventarios, préstamos a empleados, bitácoras de consumo y entre otros.

El desarrollo de la ciencia informática afecta muy directamente al trabajo cotidiano del actuario. Calculadoras, ordenadores, etc., han revolucionado nuestro trabajo en cantidad y calidad. Los avances nos permiten en ocasiones un manejo cada vez más fácil.

Dentro de los muchos programas de uso habitual en seguros en otros países, no por actuarios nicaragüenses, hemos elegido uno para modelo de distribución de las cuantías de los siniestros de seguros el programa R.

El objetivo de este documento es proporcionar un punto de partida para personas interesadas en comenzar a utilizar R de manera que el actuario y otras

profesiones afines puedan usarlo de una manera básica. Dado que R ofrece una amplia gama de posibilidades, es útil para el principiante adquirir algunas nociones y conceptos y así avanzar progresivamente.
R es un sistema para análisis estadísticos y gráficos creado por Ross Ihaka y Robert Gentle- man1. R tiene una naturaleza doble de programa y lenguaje de programación y es considerado como un dialecto del lenguaje S creado por los Laboratorios AT&T Bell. S está disponible como el programa S-PLUS comercializado por Insightfu.

La importancia de la realización de esta monografía radica en que permite en primer instancia desarrollar la teoría y la programación en los seguros generales para facilitar el cálculo de las diferentes funciones estadísticas y distribuciones de probabilidad ya que no existe una documentación sobre la descripción de las funciones que tiene el programa R aplicada a dicho ámbito (seguros), la cual hoy en día es una herramienta muy útil y eficaz para el análisis de siniestralidad de las carteras de seguros.

Esto no significa que el programa propuesto sea el único capaz de realizar estas funciones aplicadas a los seguros generales. Ni siquiera nos atrevemos a decir que sea el mejor. Pero en cualquier caso, si se puede afirmar que R es uno de los programas más usados por los actuarios. Esta es la razón por lo cual se ha elaborado este trabajo. Además de brindar una información necesaria y que haya una documentación disponible para el estudio de este tema y realización de nuevos trabajos monográficos.
No sólo a los actuarios que ya conocen del tema puede servirle esta lectura, sino también a aquellos profesionales o estudiantes del seguro que de vez en cuando echan de menos un programa especializado.

Los aspectos más importantes a destacar en cada una de las partes que integran este trabajo son los siguientes:
En el capítulo I se presenta una breve descripción y conceptos básicos sobre la Distribución Normal, Distribución Log-normal, Distribución Exponencial y Distribución Gamma, que miden los costes de los siniestros que puede sufrir una cartera de seguro, dando sus características y propiedades fundamentales.

En el segundo capítulo se presenta los conceptos básicos, descripción, ambiente, características, propiedades y funcionamiento del programa R.

Y por último se presenta una estructura siguiendo algunos de los temas que se explican en la asignatura de la estadística actuarial I, incluyendo para cada tema una breve introducción y la resolución de casos con uso de R.

II. OBJETIVOS

Objetivo General:

- ☞ Aplicar la teoría del modelo de distribución de la cuantía de los siniestros al estudio de las situaciones más habituales en las ciencias actuariales y financieras utilizando la programación en R.

Objetivos Específicos:

- ➢ Identificar los elementos estadísticos que caracterizan a las distribuciones de probabilidad continuas que describen el comportamiento de los costes de reclamación de cada siniestro, habituales en el ámbito de la empresa aseguradora: función de probabilidad, función de distribución, momentos, funciones, características y propiedades.

- ➢ Describir los aspectos generales en cuanto a concepto, al ambiente, características, propiedades y funcionamiento del programa R.

- ➢ Plantear casos que den soluciones al tratamiento del riesgo, mediante la programación en R basadas en las distribuciones estadísticas de probabilidad, en cuanto al coste de reclamación de cada siniestro, Valor en Riesgo, etc.

III. MARCO TEÓRICO

CAPÍTULO I: MODELOS DE DISTRIBUCIONES CONTINUAS: MODELOS RELACIONADOS CON LA DISTRIBUCIÓN DE CUANTÍA DE CADA SINIESTRO.

1.1. Distribución Normal

En esta sección se dedica el estudio de la distribución de probabilidad más importante de la estadística, la distribución normal o gaussiana, que corresponde a una variable aleatoria continua. La distribución normal es una de las tantas distribuciones llamadas funciones de densidad de probabilidad continua, las que surgen debido a un proceso de medición de varios fenómenos de interés.

Cuando se dispone de una expresión matemática para representar un fenómeno continuo se puede calcular la probabilidad de que varios valores de la variable aleatoria ocurran dentro de ciertos intervalos. Sin embargo la probabilidad exacta de un valor en particular de una distribución continua es cero. Estos es lo que distingue a los fenómenos continuos que se miden, de los fenómenos discretos, que se cuenta. Por ejemplo el tiempo (en segundo) se mide, no se cuenta. Por lo tanto, es posible calcular la probabilidad de terminar un trabajo en un tiempo entre 74 y 80 segundos. Al reducir este intervalo, se puede calcular la probabilidad de terminar un trabajo en un tiempo entre 74 y 76 segundos. Con instrumentos de medición más finos o precisos, es posible calcular también la probabilidad de terminar un trabajo en un tiempo entre 74.99 y 75.01 segundos. Sin embargo la probabilidad de poder terminar un trabajo exactamente 75 segundos es cero.

Los modelos continuos tienen aplicaciones importantes en la ingeniería y las ciencias físicas, así como en los seguros, la administración y en las ciencias sociales. Algunos ejemplos de fenómenos aleatorios continuos son: la estatura; el peso; los cambios diarios en los precios de las acciones al cierre; en la distribución del coste del siniestro de una compañía de seguros; el tiempo entre llegada de clientes a un banco y los tiempos de servicio a clientes, etc.

La distribución normal es de importancia vital en la estadística por tres razones principales:

1. Parece que muchos fenómenos continuos siguen esta distribución o pueden aproximarse por ella.
2. Se puede usar para aproximar varias distribuciones de probabilidad discreta.
3. Proporciona la base para la inferencia estadística clásica debido a su relación con el teorema central del límite.

La distribución normal tiene varias aplicaciones teóricas importantes como se ilustra en el cuadro siguiente:

1.1.1. Propiedades de la distribución normal

Hay cuatro propiedades clave que están asociadas con la distribución normal.

- 1. Tiene forma de campana (y por lo tanto es simétrica)
- 2. Todas sus medidas de tendencia central (media, mediana, moda) son idénticas.
- 3. El intervalo medio es igual a 1.33 desviaciones estándar. Estos significa que el rango intercuartil está dentro de un intervalo de dos tercios de desviación estándar debajo de la media hasta dos tercios de desviación estándar arriba de la media.
- 4. La variable aleatoria asociada tiene un intervalo infinito $(-\infty < x < +\infty)$

La distribución normal es continua y tiene dos parámetros, su media μ y su desviación típica σ. Estos parámetros nos definen la tendencia y la dispersión. La función de densidad de probabilidad es una campana simétrica en la relación a la media. Para la distribución normal, la función de densidad de probabilidad normal es:

Distribución normal

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{1}{2}\right)\left[\left(\frac{x-\mu}{\sigma}\right)\right]^2}$$

Donde:

e=constante matemática con valor aproximado de 2.71828

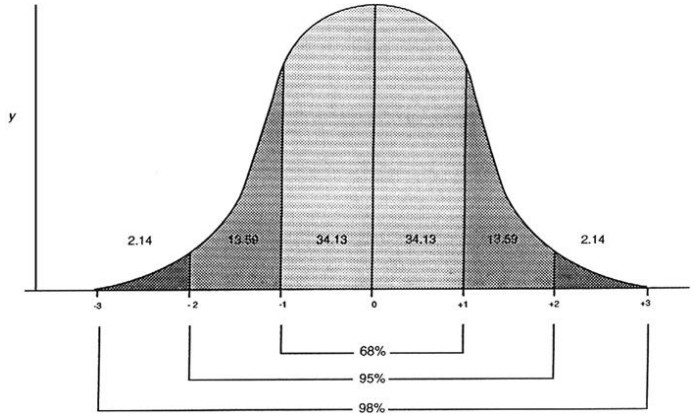
π=constante matemática con valor aproximado de 3.14159

μ=media de la población

σ=desviación estándar poblacional

X=cualquier valor de la variable aleatoria continua, donde $-\infty < x < +\infty$

Función de densidad de probabilidad de la distribución normal con media μ y varianza σ^2



Aunque el rango de una variable aleatoria normal X va de $-\infty < x < +\infty$, la probabilidad de que X tome valores muy pequeños o muy grandes es pequeña. Dicho de otra forma, hay una probabilidad del 95% de que X tome un valor en el intervalo de magnitud dos desviaciones típicas alrededor de la media.

Por desgracia, los cálculos de la expresión matemática de la función de densidad son tediosos. Para evitarlo es útil contar con un conjunto de tablas que proporcionan las probabilidades deseadas. Sin embargo, puesto que existe un número infinito de combinaciones de los parámetros μ y σ, se necesitaría un número infinito de tablas.

Si los datos se estandarizan, solo se requiere una tabla. Cualquier variable aleatoria normal X se puede convertir en una variable aleatoria normal estándar Z mediante la fórmula de transformación.

Formula de transformación:

El valor Z es igual a la diferencia entre X y la media de la población μ, dividida entre la desviación estándar σ

$$Z = \frac{X - \mu}{\sigma}$$

Una distribución normal estándar es aquella cuya variable aleatoria Z siempre tiene una media μ = 0 y desviación estándar σ = 1

Al sustituir la ecuación de la función de densidad por la fórmula de transformación, vemos que la función de densidad de probabilidad para la variable aleatoria normal estándar Z es:

Distribución Normal Estándar:

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(\frac{1}{2})z^2}$$

De esta manera, un conjunto de datos con distribución normal siempre se puede convertir en su forma estandarizada y después determinar cualquier probabilidad deseada, a partir de la tabla de distribución normal estándar.

1.2. Distribución Exponencial

Como los procesos Poisson son estacionarios, y se tiene una probabilidad igual de que el evento ocurra a todo lo largo del periodo relevante de tiempo, la distribución exponencial se aplica: si lo que interesa es el tiempo (o espacio) hasta la ocurrencia del primer evento, o el tiempo entre dos eventos sucesivos, o el tiempo que transcurre hasta que se presenta el primer evento, después de cualquier punto en el tiempo elegido al azar.

1.2.1. Características:

- Función de densidad de probabilidad:

A pesar de lo dicho sobre que la distribución exponencial puede derivarse de un proceso de Poisson, vamos a definirla a partir de la especificación de su función de densidad:

Dada una variable aleatoria X que tome valores reales no negativos {x > 0} diremos que tiene una distribución exponencial de parámetro λ (que para fines didácticos la sustituiremos con la letra b) con λ > 0, si y sólo si su función de densidad tiene la expresión:

$$f(x) = \lambda e^{-\lambda x}$$

- Función de distribución:

En la principal aplicación de esta distribución, que es la Teoría de la Fiabilidad, resulta más interesante que la función de distribución la llamada Función de Supervivencia o Función de Fiabilidad

$$F(x) = P(X \leq x) = 1 - \lambda . e^{-\lambda . x}$$

El valor esperado y la varianza de una distribución exponencial de probabilidad, en donde la variable se designa como tiempo X, son:

- Media

$$E(X) = \frac{1}{\lambda}$$

- Varianza

$$V(X) = \frac{1}{\lambda^2}$$

1.3. La Distribución Log-Normal

Esta distribución es frecuentemente utilizada como modelo de distribución del coste de un siniestro debido que es asimétrica positiva, que es una característica general de las distribuciones del coste del siniestro. Su rango va desde cero a infinito.

Una variable aleatoria X tiene una distribución log-normal con parámetros μ y σ si Y=ln X tiene una distribución normal con media μ y desviación típica σ.

La variable aleatoria tipificada viene dada por:

Log N = $\frac{\ln x - \mu}{\sigma}$ ~ N (μ, σ)

Y su función de densidad de probabilidad por:

$$f(x) = \frac{1}{\sigma_x \sqrt{2\pi}} e^{\left\{-\frac{1}{2}\left[\frac{\ln x - \mu}{\sigma}\right]^2\right\}}$$

1.3.1. Características:

- Función de distribución:

La función de distribución se obtiene tipificando y teniendo en cuenta que la variable aleatoria $\frac{Y-\mu}{\sigma} = Z$, es N(0,1), se tiene:

$$F(X) = F(\ln x) = P(Y \leq \ln x)$$

$$= P\left(\frac{Y-\mu}{\sigma} \leq \frac{\ln x - \mu}{\sigma}\right) \cong P\left(Z \leq \frac{\ln x - \mu}{\sigma}\right) = F\left(\frac{\ln x - \mu}{\sigma}\right)$$

- Media

Media = $e^{\mu+\frac{1}{2\sigma^2}}$

- Varianza

Varianza= $e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$

1.4. Distribución Gamma

Es una distribución de probabilidad continua adecuada para modelizar el comportamiento de variables aleatorias con asimetría positiva y/o los experimentos en donde está involucrado el tiempo. La distribución gamma es una de las distribuciones más utilizadas en estadística actuarial cuando se dispone de un conjunto de datos positivos, unimodales y con asimetría positiva. Una variable aleatoria tiene una distribución gamma si su función de densidad está dada por:

$$f(x,\alpha,\beta)=\begin{cases}\dfrac{1}{\beta*\Gamma(\alpha)}.x^{\alpha-1}.e^{-x/\beta} & ,\text{para } x>0;\ \alpha,\beta>0;\\ 0\ , \quad \text{de otra manera}\end{cases}$$

1.4.1. Características:

- Función de distribución

La función de distribución acumulativa de gamma $F(x)$, la cual permite determinar la probabilidad de que una variable aleatoria de gamma X sea menor a un valor específico x, se determina de la siguiente expresión:

$$F(x)=P(X\leq x)=\frac{1}{\beta*\Gamma(\alpha)}\int_0^x t^{\alpha-1}.e^{-t/\beta}dt,\ \ x>0.$$

Como se mencionó anteriormente, es una distribución adecuada para modelizar el comportamiento de variables aleatorias continuas con asimetría positiva. Es decir, variables que presentan una mayor densidad de sucesos a la izquierda de la media que a la derecha. En su expresión se encuentran dos parámetros, siempre positivos, α y β de los que depende su forma y alcance por la derecha, y también la función gamma Γ(α), responsable de la convergencia de la distribución.

- Media

  E(X)=αβ

- Varianza

  $\mathrm{Var}(X) = \alpha\beta^2$

- Los factores de forma de la distribución gamma:

- Coeficiente de asimetría: $\frac{2}{\sqrt{\alpha}}$.

- Curtosis relativa: $3(1+\frac{2}{\alpha})$.

- Con lo anterior se puede observar que la distribución gamma es leptocúrtica y tiene un sesgo positivo. También observamos que conforme el parámetro crece, el sesgo se hace menos pronunciado y la curtosis relativa tiende a 3.

1.5. Distribución Inversa Gaussiana

La distribución Gaussiana Inversa tiene aplicaciones en las áreas de análisis de supervivencia, confiabilidad, finanzas, etc. La distribución inversa Gaussiana o distribución de Wald en una de las distribuciones más utilizadas para ajustar datos relativos al coste de los siniestros.

1.5.1. Características:

- Función de densidad

Una variable aleatoria X sigue una distribución inversa Gaussiana con parámetros μ y λ, y representaremos X~$\mathcal{IG}$ (μ, λ), si su función de densidad viene dada por,

$$f(x) = \sqrt{\frac{\lambda}{2\pi x^3}\, exp\left\{\frac{\lambda}{2\mu^2 x}(x-\mu)^2\right\}}\ , x > 0$$

donde λ, μ>0.

- Función de distribución

Se trata de una distribución asimétrica positiva y unimodal. La función de distribución puede expresarse en términos de la función de distribución Φ de una normal estándar y viene dada por,

$$F(x) = \Phi\left(\sqrt{\frac{\lambda}{x}\left(\frac{x}{\mu} - 1\right)}\right) + e^{2\lambda/\mu}\Phi\left(-\sqrt{\frac{\lambda}{x}\left(\frac{x}{\mu} + 1\right)}\right),\ x > 0.$$

- Función generatriz de momentos

La distribución inversa Gaussiana posee momentos positivos y negativos de todos los órdenes. La función generatriz de momentos viene dada por,

$$M_x(t) = exp\left\{\frac{\lambda}{\mu}\left(1 - \sqrt{1 - \frac{2\mu^2 t}{\lambda}}\right)\right\},$$

y su función característica,

$$\varphi x(t) = \mathrm{E}\left(e^{itX}\right) = M_x(it) = exp\left\{\frac{\lambda}{\mu}\left(1 - \sqrt{1 - \frac{2i\mu^2 t}{\lambda}}\right)\right\}.$$

- Media
  E(X)=μ

- Varianza
  $\mathrm{Var(X)} = {\mu^3}/{\lambda}$

1.6. Contraste De Bondad De Ajuste

Un contraste de bondad de ajuste se emplea para verificar si un conjunto de datos (muestra aleatoria) procede de una población con una cierta distribución de probabilidad. Existen diferentes tests de bondad de ajustes, aquí veremos los más usuales que son:

- El test X^2de bondad de ajuste.
- El test de Kolmogorov-Smirnov.
- El test de normalidad de Shapiro-Wilks.
- El test de normalidad de Lilliefors.
- El test de Kolmogorov-Smirnov para dos muestras.

1.6.1. Contraste X^2 de Pearson de bondad de ajuste:

Este test es el más antiguo y el más conocido de los tests de bondad de ajuste, y fue introducido por Pearson en 1900, para utilizarlo con datos nominales. Esta prueba es aplicable para variables aleatorias discretas o continuas.

Aquí se pueden considerar dos casos:

1. *Cuando los parámetros de la distribución de la población son todos conocidos* bien porque se conocen previamente o porque se pueden estimar de una muestra distinta a la que se utiliza para realizar el contraste de la bondad de ajuste.
2. *Cuando uno o varios de los parámetros de la distribución no son conocidos* y hay que estimarlo a partir de la misma muestra que se utiliza para realizar el contraste de bondad de ajuste.

Para hacer un test X^2 de bondad de ajuste que permita comprobar si los datos pueden considerarse procedentes de una población o se ajustan a una distribución debemos realizar el siguiente procedimiento:

1. Obtener las medidas de tendencia central y de dispersión.
2. Agrupar los intervalos para que ninguna de las casillas halla una frecuencia menor que 5.
3. Obtener las columnas de las frecuencias teóricas ya sea de la distribución exponencial, log-normal, gamma e inversa gaussiana.
4. Se plantean las hipótesis: una nula (H_0) y otra alternativa (H_a).
5. Se obtiene el valor del estadístico X^2 mediante la siguiente formula:

$$\mathrm{X}^2 = \sum_{\mathrm{i}=1}^{\mathrm{k}} \frac{(\mathrm{o_i} - \mathrm{e_i})^2}{\mathrm{e_1}}$$

Donde:

$\mathrm{o_i}$: Frecuencia observada (corresponde a los datos de la muestra)

$\mathrm{e_i}$: Frecuencia esperada (corresponde al modelo propuesto)

6. Se obtiene el valor $X^2_{1-\alpha}$procedente de la tabla *ji-cuadrada* de Pearson.

Donde α representa el nivel de significancia.

7. Se toma la decisión de aceptar o rechazar la H_0. Esto es

   $\mathrm{X}^2 = \sum_{\mathrm{i}=1}^{\mathrm{k}} \frac{(\mathrm{o_i}-\mathrm{e_i})^2}{\mathrm{e_1}} < X^2_{1-\alpha}$ se acepta H_0

   $\mathrm{X}^2 = \sum_{\mathrm{i}=1}^{\mathrm{k}} \frac{(\mathrm{o_i}-\mathrm{e_i})^2}{\mathrm{e_1}} > X^2_{1-\alpha}$ se rechaza H_0

CAPÍTULO II: GENERALIDADES DEL LENGUAJE DE PROGRAMACIÓN R

2.1. Definición:

R es un conjunto integrado de programas para manipulación de datos, cálculo y gráficos.

2.2. Características de que dispone de R:

a) Almacenamiento y manipulación efectiva de datos.
b) Operadores para cálculo sobre variables indexadas (Arrays), en particular matrices.
c) Una amplia, coherente e integrada colección de herramientas para análisis de datos.
d) Posibilidades gráficos para análisis de datos, que funcionan directamente sobre pantalla o impresora.
e) Un lenguaje de programación bien desarrollado, simple y efectivo, que incluye condicionales, ciclos, funciones recursivas y posibilidad de entradas y salidas. (Debe destacarse que muchas de las funciones suministradas con el sistema están escritas en el lenguaje R)

R realmente es un lenguaje y conjunto de módulos estadísticos que, mediante cualquiera de los interfaces de que dispone, permite realizar análisis de datos y representación de los mismos. Es un software para el análisis estadístico de datos considerado como uno de los más interesantes; apoyan esta opinión la vasta variedad de métodos estadísticos que cubre, las capacidades gráficas que ofrece y, también muy importante, el hecho de ser un software libre, es decir, gratuito.

R posee muchas funciones para análisis estadísticos y gráficos; los resultados de análisis estadísticos se muestran en la pantalla, y algunos resultados intermedios (como valores P-, coeficientes de regresión, residuales,...) se pueden guardar, exportar a un archivo, o ser utilizados en análisis posteriores, los últimos pueden ser visualizados de manera inmediata en su propia ventana y ser guardados en varios formatos (jpg, png, bmp, ps, pdf, emf, pictex, xfig; los formatos disponibles dependen del sistema operativo).

El lenguaje R permite al usuario, por ejemplo, programar bucles (’loops’ en inglés) para analizar conjuntos sucesivos de datos.

2.3. Cómo funciona R

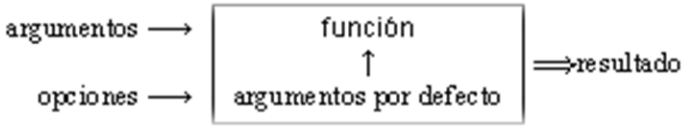
R es un lenguaje Orientado a Objetos: bajo este complejo término se esconde la simplicidad y flexibilidad de R.

R es un lenguaje interpretado (como Java) y no compilado (como C, C++, Fortran, Pascal,. . .), lo cual significa que los comandos escritos en el teclado son ejecutados directamente sin necesidad de construir ejecutables. La sintaxis de R es muy simple e intuitiva. Por ejemplo, una regresión lineal se puede ejecutar con el comando lm(y ˜x). Para que una función sea ejecutada en R debe estar siempre acompañada de paréntesis, inclusive en el caso que no haya nada dentro de los mismos (por ej., ls()). Si se escribe el nombre de la función sin los paréntesis, R mostrará el contenido (código) mismo de la función.

En este documento, se escribirán los nombres de las funciones con paréntesis para distinguirlas de otros objetos, a menos que se indique lo contrario en el texto.

Orientado a Objetos significa que las variables, datos, funciones, resultados, etc., se guardan en la memoria activa del computador en forma de objetos con un nombre específico. El usuario puede modificar o manipular estos objetos con operadores (aritméticos, lógicos, y comparativos) y funciones (que a su vez son objetos).

El uso y funcionamiento de los operadores es relativamente intuitivo. Una función en R se puede delinear de la siguiente manera:



Los argumentos pueden ser objetos (“datos”, fórmulas, expresiones,...), algunos de los cuales pueden ser definidos por defecto en la función; sin embargo estos argumentos pueden ser modificados por el usuario con opciones. Una función en R puede carecer totalmente de argumentos, ya sea porque todos están definidos por defecto (y sus valores modificados con opciones), o porque la función realmente no tiene argumentos.

Veremos más tarde en detalle cómo usar y construir funciones. Por ahora esta corta descripción es suficiente para entender el funcionamiento básico de R.

Todas las acciones en R se realizan con objetos que son guardados en la memoria activa del ordenador, sin usar archivos temporales. La lectura y

escritura de archivos solo se realiza para la entrada y salida de datos y resultados (graficas,. . .). El usuario ejecuta las funciones con la ayuda de comandos definidos. Los resultados se pueden visualizar directamente en la pantalla, guardar en un objeto o escribir directamente en el disco (particularmente para gráficos). Debido a que los resultados mismos son objetos, pueden ser considerados como datos y analizados como tal. Archivos que contengan datos pueden ser leídos directamente desde el disco local o en un servido remoto a través de la red.

Las funciones disponibles están guardadas en una librería localizada en el directorio R HOME/library (R HOME es el directorio donde R está instalado). Este directorio contiene paquetes de funciones, las cuales a su vez están estructuradas en directorios. El paquete denominado base constituye el núcleo de R y contiene las funciones básicas del lenguaje para leer y manipular datos, algunas funciones gráficas y algunas funciones estadísticas (regresión lineal y análisis de varianza). Cada paquete contiene un directorio denominado R con un archivo con el mismo nombre del paquete (por ejemplo, para el paquete base, existe el archivo R HOME/library/base/R/base). Este archivo está en formato ASCII y contiene todas las funciones del paquete.

El comando más simple es escribir el nombre de un objeto para visualizar su contenido. Por ejemplo, si un objeto n contiene el valor 10:

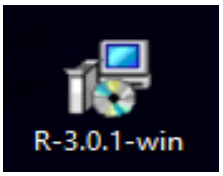
```
> n
[1] 10
```

El dígito 1 indica que la visualización del objeto comienza con el primer elemento de n. Este comando constituye un uso implícito de la función *print*, y el ejemplo anterior es similar a print(n) (en algunas situaciones la función print debe ser usada explícitamente, como por ejemplo dentro de una función o un bucle).

El nombre de un objeto debe comenzar con una letra (A-Z and a-z) y puede incluir letras, dígitos (0-9), y puntos (.). R discrimina entre letras mayúsculas y minúsculas para el nombre de un objeto, de tal manera que x y X se refiere a objetos diferentes (inclusive bajo Windows).

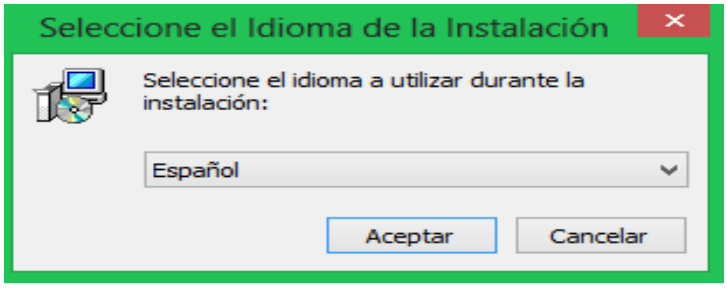
2.4. Instalar R.

Para ello hacemos doble clic sobre el icono del archivo compilado de R.

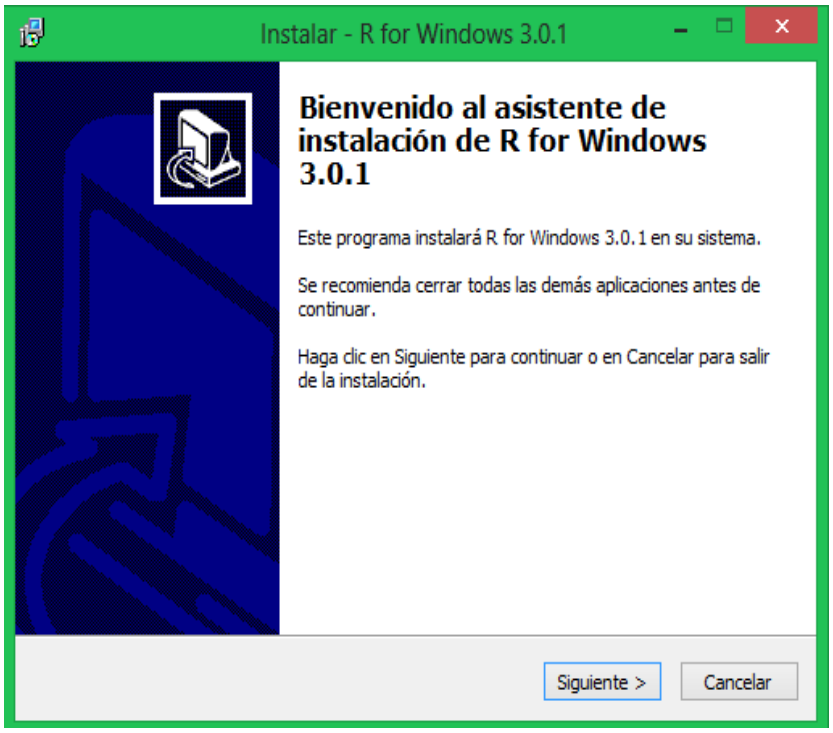


Al abrirse la ventana “Abrir archivo-Advertencia de seguridad”, hacemos clic sobre el botón “Ejecutar”

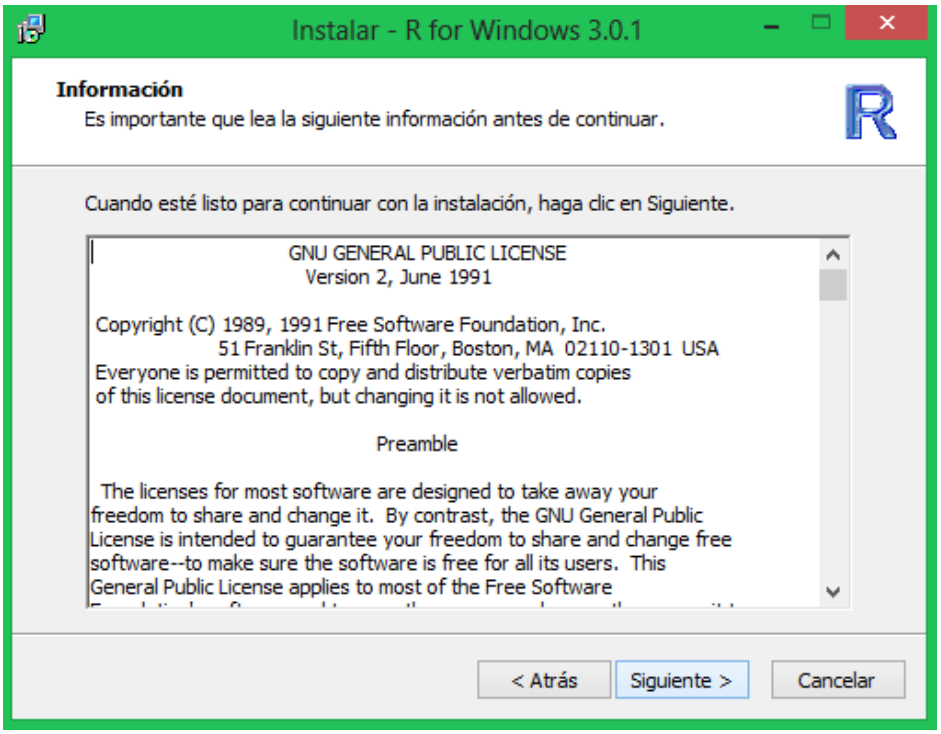
-Seleccionamos el idioma de instalación.



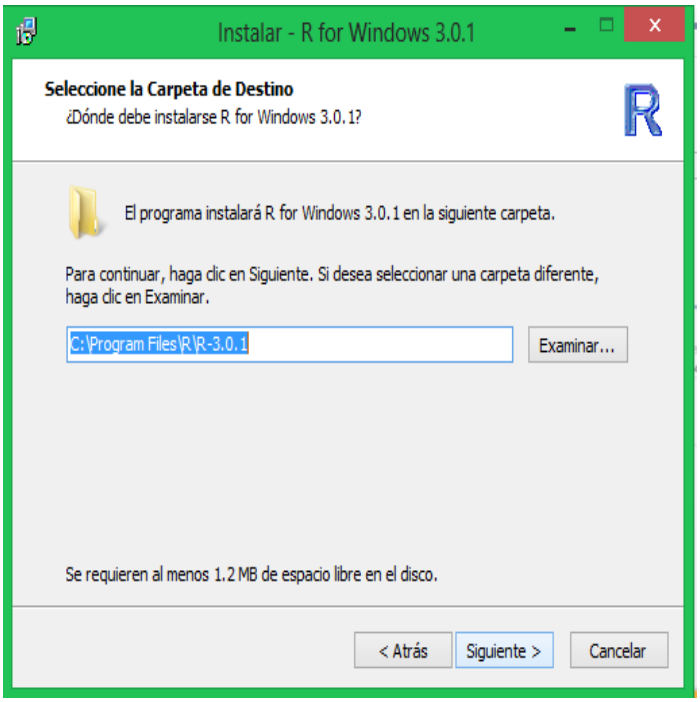
-Seguimos las instrucciones del asistente de instalación de R



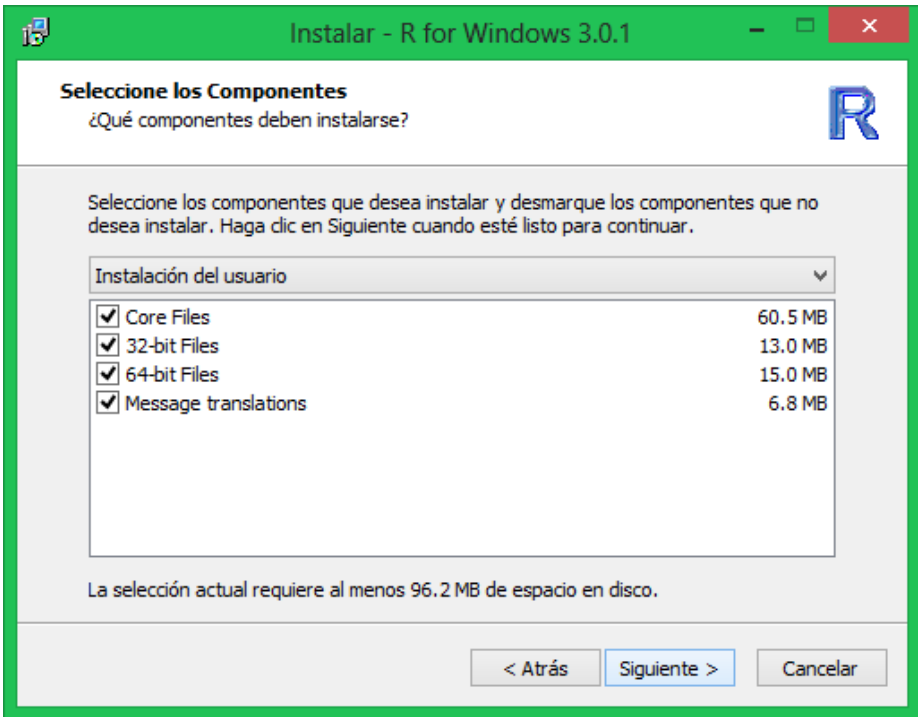
-Leemos las condiciones de licencia de R



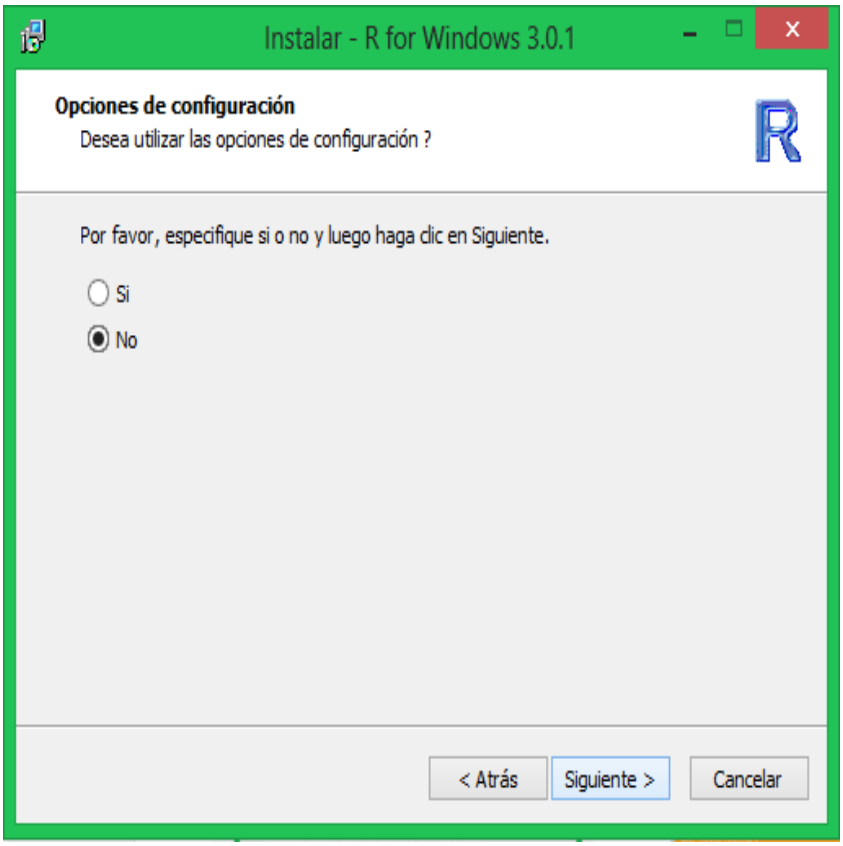
-Seleccionamos la carpeta donde instalaremos R



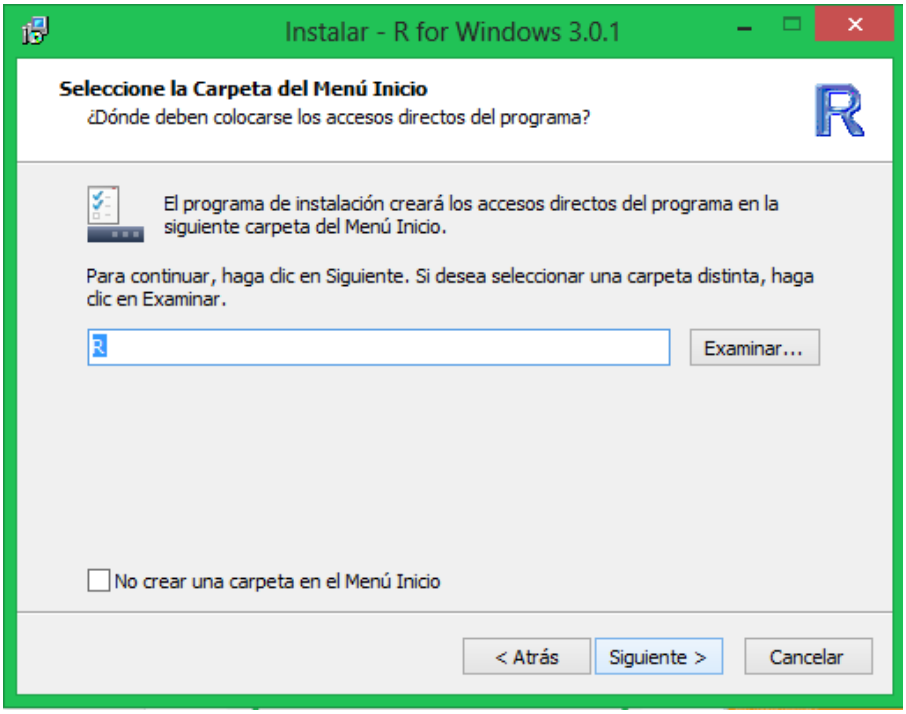
-Seleccionamos los componentes a instalar.



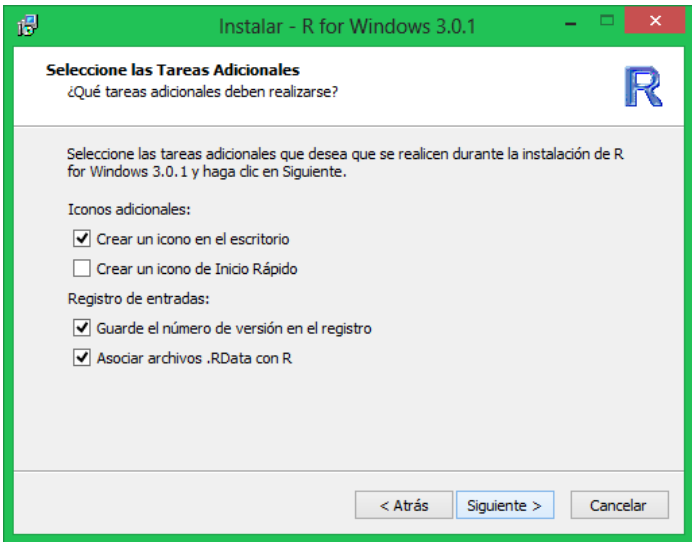
-Especificamos si utilizaremos opciones de configuración



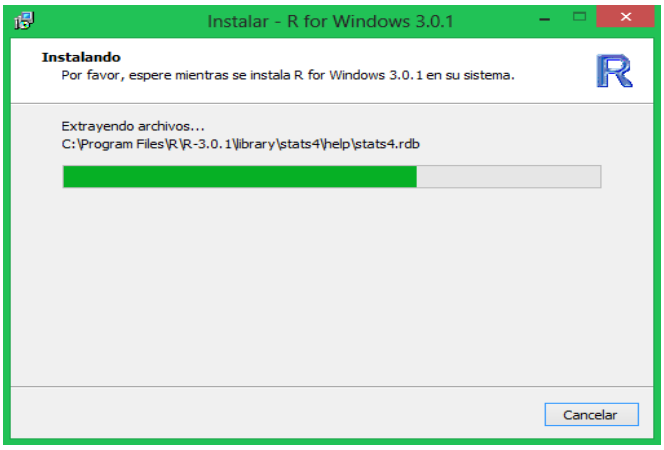
-Seleccionamos dónde se crearán accesos directos al programa



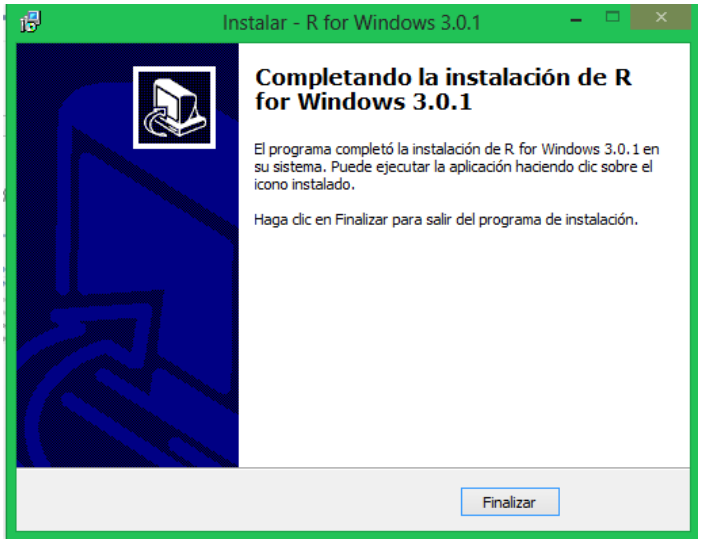
Seleccionamos tareas adicionales como la de “Crear un icono en el escritorio”



Una vez ejecutadas las acciones anteriores, R se instalará automáticamente.



-Para terminar el proceso hacemos clic sobre el botón “Finalizar”



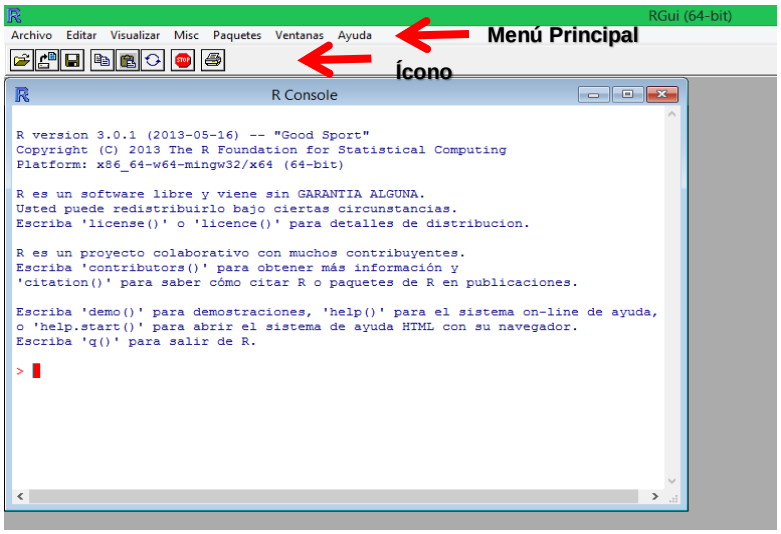
2.5. El ambiente de trabajo en R

2.5.1. Iniciar una sesión de trabajo en R.

Hacemos doble clic sobre el ícono de R que aparece en el escritorio.

2.5.2. El ambiente de trabajo en R

Al abrir R se mostrará la siguiente imagen:



En la imagen podemos identificar los siguientes elementos:

- El menú principal:

Compuesto por los menús: Archivo, Editar, Visualizar, Misc, Paquetes, Ventanas, y Ayuda. Al desplegar estos menús, podemos realizar procedimientos complementarios a la escritura de programas en R. Las funciones específicas a las que podemos acceder a través de cada menú, las daremos a conocer a lo largo del manual.

- Los íconos de funciones:

Constituyen accesos abreviados o rápidos a las funciones más usadas de R, como: abrir archivos de programas (documentos con extensión *.txt, *.R); cargar espacios de trabajo (archivos con extensión *.RData); copiar; pegar; copiar y pegar consecutivamente en la Consola, interrumpir la ejecución de instrucciones, e imprimir.

- La consola:

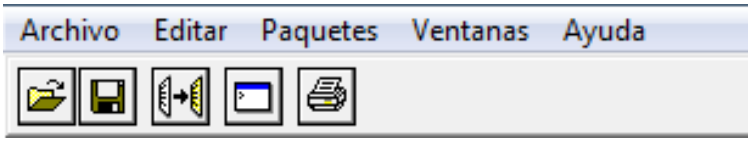
La consola es el espacio en donde: en letras rojas, aparecen las instrucciones dadas a R y en letras azules, sus resultados. Las instrucciones pudieron ser escritas en la ventana Script y luego ejecutadas, apareciendo automáticamente en letras rojas en la Consola; o pudieron escribirse directamente en ella; en este último caso, si las instrucciones están completas, la ejecución se realizará al presionar la tecla ENTER, de no ser el caso, aparecerá el signo +, indicando que nos falta terminar de escribirlas.

Otros elementos de R, son visibles al realizar procedimientos previos:

- La ventana Script

Para obtener esta ventana, desplegamos el menú Archivo → Nuevo Script.

Como podemos observar, al activar la ventana Script, los menús Archivo y Editar muestran opciones sólo pertinentes a esta ventana. También se reduce el número de íconos disponibles y se presentan los íconos: Ejecutar (correr línea o seleccionar), y cambiar a Consola (retornar foco a la consola).



La ventana Script es un espacio en donde podemos escribir instrucciones. Para que R las ejecute primero debemos seleccionarlas y luego realizar cualquiera de las acciones siguientes: desplegar el menú: Editar → correr línea o seleccionar; presionar la tecla F5; presionar simultáneamente las teclas CTRL+R; o presionar el siguiente ícono:

Aunque las instrucciones para R, también pueden escribirse en la Consola, una de las ventajas de escribirlas en la ventana Script es que podemos introducir modificaciones, fácilmente y que podemos guardarlas en un archivo, para uso futuro.

Para grabar un Script desplegamos el menú Archivo → Guardar o Archivo → Guardar como

Para abrir una Script existente desplegamos el menú Archivo → Abrir Script…

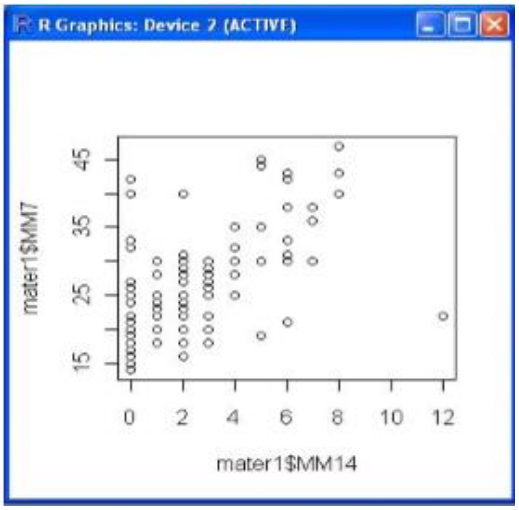
- El editor de datos

Para acceder a esta ventana, previamente debemos haber pedido a R que haga la lectura de un archivo de datos. Luego desplegamos el menú Editar →Editor de datos, y escribimos el nombre del conjunto de datos que deseamos editar, por ejemplo: mater1.

	CASEID	MMIDX	MM1	MM2
1	000804001 03	7	2	0
2	001009101 02	2	2	0
3	001202101 02	1	2	0
4	001607501 03	3	2	0
5	002106101 02	2	2	0
6	005700501 02	3	2	0
7	007703301 02	3	2	0
8	008101201 02	2	2	0
9	008106601 02	1	2	0
10	010700701 02	1	2	0
11	010804701 02	5	2	0
12	012002901 02	1	2	0
13	012002901 02	10	2	0

- La ventana de gráficos

Se activa automáticamente al dar instrucciones a R, para realizar un gráfico.



- Organizar ventanas.

Para trabajar con mayor comodidad, se puede organizar la presentación de las ventanas de Script, Consola y/o gráficos, en forma paralela, ya sea de manera vertical u horizontal. Para ello desplegamos el menú:
Ventanas→ Mosaico vertical o ventanas→ Mosaico Horizontal.

R organizará todas las ventanas que, se encuentren abiertas. En el siguiente gráfico se observa cómo se organizan tres ventanas abiertas en una sesión, desplegando el menú Ventana→ Mosaico Vertical.

2.6. Elementos De Programación En Lenguaje R

"R trabaja con objetos" .Estos objetos pueden ser: estructuras de datos como los vectores, los factores, las matrices, los marcos de datos, entre otros; o de funciones como las funciones matemáticas, las funciones estadísticas, las funciones para realizar gráficos, o inclusive funciones para realizar otras funciones.

2.6.1. Los objetos de R

Los objetos que más usaremos en este manual son los siguientes: vectores, factores, matrices, marcos de datos y funciones. Aquí describiremos sus principales características:

- Vector:

El vector es la estructura de datos básica y puede asumir diversos modos, entre ellos: numéricos, caracteres y lógicos.

Ejemplo: el vector numérico x es una secuencia de números consecutivos del 1 al 4.

```
> x
[1] 1 2 3 4
```

- Factor:

Un factor es un objeto que tiene como base un vector, al cual se le ha identificado sus niveles. Estos niveles describen grupos en el vector.

Ejemplo: el factor yf tiene 8 elementos y dos grupos o niveles, el grupo de hombres y el grupo de mujeres.

```
> yf
[1] hombre mujer  hombre mujer  mujer  mujer  hombre hombre
Levels: hombre mujer
```

- Matriz:

Es una tabla o arreglo de dos dimensiones (filas y columnas). Una matriz tiene todos sus elementos de un mismo modo, factores o vectores, pero no ambos .

Ejemplo: la matriz m1 tiene 20 elementos, 5 filas y cuatro 4 columnas.

```
> m1
     [,1] [,2] [,3] [,4]
[1,]    1    6   11   16
[2,]    2    7   12   17
[3,]    3    8   13   18
[4,]    4    9   14   19
[5,]    5   10   15   20
```

- Marco de datos (data frame):

Es una estructura de datos compleja, de vectores y factores de la misma longitud. Ejemplo: el marco de datos grupo 1, está compuesto por los factores resi y sexo y los vectores, edad y nota.

```
> grupo1
    resi sexo edad nota
1 urbana    h   15   15
2  rural    m   16   15
3  rural    h   17   14
4  rural    m   17   13
5 urbana    h   18   17
6 urbana    m   19   18
7 urbana    h   18   10
8 urbana    h   16   17
9  rural    m   15   12
```

- Funciones:

Son objetos que nos permiten realizar diversas tareas con otros objetos. Estas tareas pueden ser tratamiento de archivos, tratamientos de variables, operaciones matemáticas simples y complejas; análisis estadísticos; y gráficos. Ejemplo 1: calcular la media de la variable edad del marco de datos grupo1

```
> mean(grupo1$edad)
[1] 16.77778
```

Donde:

- Mean es la función que calcula la media de la variable cuantitativa edad
- grupo1 es el marco de datos al que pertenece la variable edad
- $ es la notación usada para vincular una variable a su marco de datos correspondiente

Ejemplo 2: usar la función attach() para vincular las variables con su marco de datos, luego calcular la media de las variables edad y nota, y presentar la distribución de frecuencias de las variables resi y sexo que se encuentran en el marco de datos grupo1.

```
> attach(grupo1)
> mean(edad)
[1] 16.77778
> mean(nota)
[1] 14.55556
> table(resi)
resi
 rural urbana
     4       5
> table(sexo)
sexo
h m
5 4
```

Donde:

- Attach es la función para vincular variables con su marco de datos.
- Mean es la función usada para estimar la media de los vectores edad y nota.
- Table es la función usada para elaborar las tablas de distribución de frecuencia de los factores resi y sexo

Como podemos observar, al usar la función attach(), evitamos escribir el nombre del marco de datos seguido de la notación $, cada vez que invocamos el nombre de la variable sobre la cual deseamos se aplique una función dada.

Para desvincular las variables de su marco de datos usamos la función detach(). En esta situación, para ejecutar una función sobre las variables necesitaremos escribirlas precedidas del nombre del marco de datos y la notación $, como en el siguiente ejemplo:

```
> detach(grupo1)
> mean(edad)
Error in mean(edad) : object 'edad' not found
> mean(grupo1$edad)
[1] 16.77778
```

2.6.2. Atributos intrínsecos de los objetos

Entre los atributos intrínsecos de los objetos tenemos: el modo (numérico, carácter, lógico…) y la longitud (número de elementos que contiene un objeto). Es importante tomar en consideración estos atributos de los objetos, ya que depende de ellos la aplicabilidad de una función.

2.6.3. Creación de objetos

En R podemos crear un objeto usando el operador asignar (<- o ->). Al crearse el objeto, R lo guarda en memoria, y solo podremos visualizarlo cuando lo "invoquemos", tal como observaremos a continuación.

```
> z <- c(7,9,5,8,9,3,NA,4)
> z
[1]  7  9  5  8  9  3 NA  4
> w <- z < 8
> w
[1]   TRUE FALSE  TRUE FALSE FALSE  TRUE    NA  TRUE
```

2.6.4. Creación de factores

Para crear factores, se usa la función factor().

Ejemplo 1: a partir del vector x = 1, 2, 1, 2, 2, 2, 1,1, crear el factor xf.

```
> x<-c(1,2,1,2,2,2,1,1)
> x
[1] 1 2 1 2 2 2 1 1
> xf<-factor(x)
> xf
[1] 1 2 1 2 2 2 1 1
Levels: 1 2
```

Un ejemplo de funciones que se pueden ejecutar con un factor es table(), cuyo resultado es la distribución de frecuencia de los niveles del factor.

```
> table(xf)
xf
1 2
4 4
```

2.6.5. Creación de una matriz

Para crear una matriz usamos la función matrix(), detallamos sus componentes, e indicamos el número de filas (nrow) y/o columnas (ncol) de las que estará compuesta.

Ejemplo: crear la matriz m1 de 5 filas y 4 columnas con la secuencia de números consecutivos del 1 al 20.

```
> m1 <- matrix(1:20,nrow=5)
> m1
     [,1] [,2] [,3] [,4]
[1,]    1    6   11   16
[2,]    2    7   12   17
[3,]    3    8   13   18
[4,]    4    9   14   19
[5,]    5   10   15   20
```

Además con la función dim(), podemos reportar el número de filas y columnas que tiene la matriz.

```
> dim(m1)
[1] 5 4
```

2.6.6. *Creación de un marco de datos* (data frame)

Para crear un marco de datos, utilizamos la función data.frame().

Ejemplo 1. Crear el marco de datos grupo1, con los factores resi y sexo y los vectores edad y nota.

- Creación de los vectores de tipo caracter x y w, y de los vectores numéricos y y z:

```
> x<-c("urbana","rural","rural","rural","urbana","urbana","urbana","urbana","rural")
> x
[1] "urbana" "rural"  "rural"  "rural"  "urbana" "urbana" "urbana"
[8] "urbana" "rural"
> w<-c("h","m","h","m","h","m","h","h","m")
> w
[1] "h" "m" "h" "m" "h" "m" "h" "h" "m"
> y<-c(15,16,17,17,18,19,18,16,15)
> y
[1] 15 16 17 17 18 19 18 16 15
> z<-c(15,15,14,13,17,18,10,17,12)
> z
[1] 15 15 14 13 17 18 10 17 12
```

- Conversión de los vectores x y w en los factores xf y wf:

```
> xf<-factor(x)
> xf
[1] urbana rural  rural  rural  urbana urbana urbana urbana rural
Levels: rural urbana
> wf<-factor(w)
> wf
[1] h m h m h m h h m
Levels: h m
```

- Creación del marco de datos grupo1 con los factores resi y sexo y los vectores edad y nota.

```
> grupo1<-data.frame(resi=xf, sexo=wf, edad=y, nota=z)
> grupo1
    resi sexo edad nota
1 urbana    h   15   15
2  rural    m   16   15
3  rural    h   17   14
4  rural    m   17   13
5 urbana    h   18   17
6 urbana    m   19   18
7 urbana    h   18   10
8 urbana    h   16   17
9  rural    m   15   12
```

Una vez creado el marco de datos, podemos llamar variables, a los vectores y factores incluidos en él.

Bases de datos externas, pueden ser leídas por R como marcos de datos usando la función read.table (ver procedimientos en el capítulo 4).

2.6.7. *Solicitar ayuda*

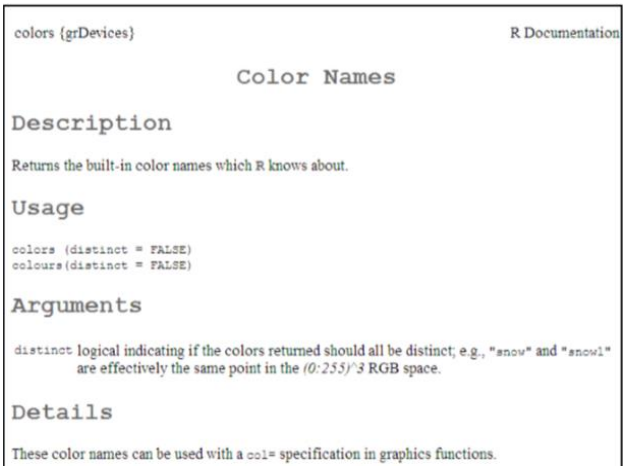
Hay diversas maneras de solicitar ayuda para escribir en lenguaje R, aquí describiremos tres:

- Si conocemos el nombre de la función:

Escribimos en la consola el nombre de esta, precedido por el signo de cierre de interrogación (?). Como resultado R nos mostrará una descripción de la función,

la forma de su uso, los argumentos, detalles, valores, funciones asociadas y ejemplos.

```
> ?colors
starting httpd help server ... done
```



colors {grDevices} R Documentation

Color Names

Description

Returns the built-in color names which R knows about.

Usage

```
colors (distinct = FALSE)
colours(distinct = FALSE)
```

Arguments

distinct logical indicating if the colors returned should all be distinct; e.g., "snow" and "snow1" are effectively the same point in the *(0:255)^3* RGB space.

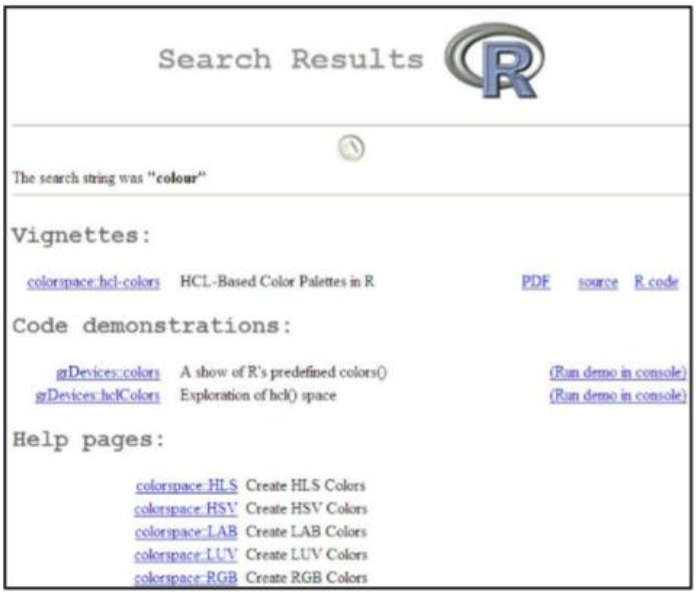
Details

These color names can be used with a col= specification in graphics functions.

- Si no conocemos el nombre de la función:

Escribimos doble signo de interrogación antes de la palabra (en inglés) asociada a la función. Luego de ello R nos mostrará una página con diversos recursos asociados a la palabra sobre la cual buscamos información.

```
> ??colour
```



Search Results

The search string was "**colour**"

Vignettes:

colorspace:hcl-colors HCL-Based Color Palettes in R PDF source R code

Code demonstrations:

grDevices:colors A show of R's predefined colors() (Run demo in console)
grDevices:hclColors Exploration of hcl() space (Run demo in console)

Help pages:

colorspace:HLS Create HLS Colors
colorspace:HSV Create HSV Colors
colorspace:LAB Create LAB Colors
colorspace:LUV Create LUV Colors
colorspace:RGB Create RGB Colors

- Solicitar ejemplos de la manera en que se emplea una función:

```
> example(mean)

mean> x <- c(0:10, 50)

mean> xm <- mean(x)

mean> c(xm, mean(x, trim = 0.10))
[1] 8.75 5.50
```

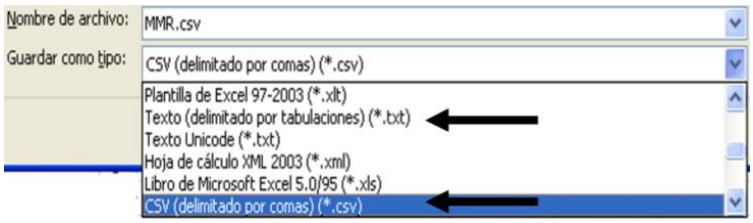
2.7. Tratamiento Y Exploración De Archivos

Como señaláramos en el punto 3.3.4 del capítulo anterior, R puede leer archivos externos como marcos de datos (data frame). En esta parte leeremos archivos con extensión *.txt, *.csv, y *.dat, usando el paquete básico instalado en R.

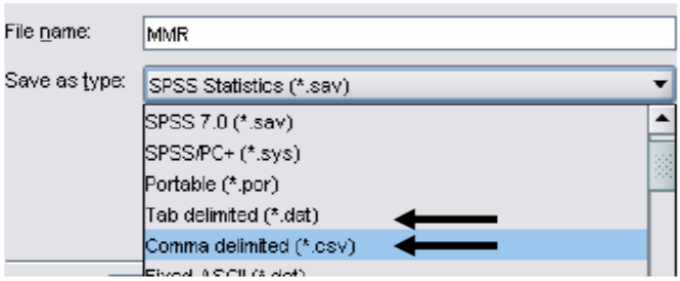
Dado que la mayoría de veces las bases de datos se encuentra archivados en formato SPSS o Excel, lo primero que haremos será preparar estos archivos para que puedan ser leídos por el paquete básico de R. Aunque debemos mencionar que R cuenta con paquetes que se pueden descargar para leer directamente los archivos con extensión *.sav y *.xls o *.xlsx.

2.7.1. Preparar archivos externos que puedan ser leídos por el paquete básico de R

Para transportar archivos creados en Excel o en SPSS que puedan ser leídos por el paquete básico de R, primero seleccionamos la opción Guardar como/Save as; luego abrimos la pestaña Guardar como tipo/Save as type; y finalmente seleccionamos un formato. Los formatos disponibles desde Excel son: *.csv o *.txt:



Y los formatos disponibles desde SPSS son: *.csv o *.dat.



Los archivos resultantes serán:

2.8. Lo anexo del R para el ámbito Actuarial.

2.8.1. El paquete Actuar:

Una de las mayores ventajas de R, es que cuenta con paquetes que extienden su configuración básica, los cuales son agrupados en un repertorio oficial y están organizados por temas según su naturaleza y función.
Es este inmenso repertorio el que hace posible el uso de este lenguaje en distintas profesiones, ya que cuenta con paquetes relacionados con econometría, ecología, finanzas, agricultura y muchas otras áreas de estudio. Dentro de este extenso repertorio encontramos un paquete especialmente útil para las ciencias actuariales. Dicho paquete lleva por nombre: actuar.
El paquete "actuar" es una biblioteca de funciones de ciencias actuariales y sus funcionalidades se pueden dividir en 3 grupos:

i) Modelización de distribuciones de pérdidas.
ii) Teoría del riesgo.
iii) Teoría de la credibilidad.

Además, este paquete permite un proceso más dinámico de modelización-estimación-diagnóstico-predicción, comparado con otros de su tipo.
Actuar contiene funciones de distribuciones de pérdidas, permite manipular y almacenar datos de la forma intervalo-frecuencial con facilidad. Dichas funciones convierten a R en una plataforma para modelizar y calcular variables de teoría del riesgo, calcular coeficientes de ajuste para cualquier frecuencia y severidad de siniestros y permite la simulación de portafolios y modelos de credibilidad. Estas herramientas facilitan el trabajo en el ámbito de los seguros patrimoniales y ayudan a mejorar los tiempos de respuesta de los análisis hechos por los actuarios.

Por otra parte, el paquete contiene funciones en su mayoría orientadas a las distribuciones de pérdidas, de igual forma está enfocado principalmente en seguros patrimoniales y no en los de vida. Ambas implican fuertes limitaciones en el software para el desarrollo del extenso campo de la actuaría. En el futuro, se espera que, con la cooperación de los usuarios, pueda abarcar todas las áreas de las ciencias actuariales.

IV. DISEÑO METODOLÓGICO.

4.1. Tipo de estudio: es de tipo cuantitativo, descriptivo y de corte transversal.

4.2. Cuantitativo: se realizó el cálculo de la frecuencia ajustada siguiendo el procedimiento que el programa R permite para obtener estimaciones del coste de los siniestros.

4.3. Descriptivo: se describieron todos los datos obtenidos en la operación del caculo del coste de los siniestros, con el fin de analizar su resultado.

4.4. De corte transversal: se realizó en el período comprendido julio – diciembre 2016.

4.5. Área de estudio: Seguro privado

4.6. Unidad de análisis: cartera de seguros

4.7. Población: 275,485 siniestros ocurridos en el ramo de hogar de una compañía de seguro privado.

4.8. Criterios de inclusión:

- Cartera de seguros de hogar.

4.9. Criterios de exclusión:

- Cartera de seguros de automóvil
- Cartera de seguros de vida individual y colectivo

4.10. Tipo de variable

4.10.1. Dependiente:

- Cálculo de la frecuencia ajustada de póliza

4.10.2. Independiente:

- Tipo de distribución.
- Los parámetros de las distribuciones
- Número de pólizas.
- Frecuencia de probabilidad.

4.11. Fuente de la recolección de la información:

- Secundaria: Bases de datos estadísticas de una compañía de seguro, libros, informes y documentos de Internet.

4.12. Procesamiento de la información:

Se introdujeron los datos de la información utilizando el método electrónico computarizado para el procesamiento de datos, mediante el uso del programa de R, Excel, Word y PowerPoint. Año 2013, en la cual se hizo por cada variable de estudio.

4.13. Análisis de los datos: Los datos se analizaron por cada tipo de distribución de probabilidad del coste de los siniestros, presentados en tablas.

V. RESULTADOS.

Caso: La tabla siguiente abreviada presenta el coste medio de 290,608 siniestros que ocurrieron en ramo de hogar de una compañía de seguro de Nicaragua. Usando R:

a) Realice un ajuste a una distribución exponencial, log-normal, gamma, inversa gaussiana y calculas las frecuencias absolutas teóricas que se obtendrían ajustando a los datos las distribuciones continuas conocidas.

b) Mediante el contraste chi-cuadrado, encontrar la distribución que se ajusta mejor a los datos observados.

INTÉRVALOS DE COSTE	COSTE MEDIO	FRECUENCIA ABSOLUTA
0- 5,000	$3,420.00	4,114.00
5,001-10,000	$7,829.00	10,116.00
10,001-15,000	$12,595.00	15,796.00
15,001-20,000	$17,503.00	15,611.00
20,001-25,000	$22,532.00	15,749.00
25,001-30,000	$27,577.00	16,514.00
30,001-35,000	$32,530.00	16,891.00
35,001-40,000	$37,501.00	16,540.00
40,001-45,000	$42,477.00	14,817.00
45,001-50,000	$47,836.00	15,824.00
50,001-65,000	$57,511.00	35,342.00
65,001-80,000	$73,004.00	29,496.00
80,001-95,000	$86,908.00	15,447.00
95,001-120,000	$105,503.00	29,457.00
120,001-145,000	$131,445.00	9,553.00
⋮	⋮	⋮
⋮	⋮	⋮
2,500,001-3,000,000	$2753,731.00	72.00
3,000,001-3,500,000	$3227,466.00	37.00
3,000,001-4,000,000	$3709,932.00	10.00
Más de 4,000,000	$5320,783.00	32.00

5.1. Procedimientos General:

> # Se abre un nuevo script en R y se introduce el vector de puntos medios de los intervalos

```
>cuantia=c(3420,7829,12595,17503,22532,27577,32530,37501,42477,47836,57511,73004,86908,105503,131445,156548,184365,222845,273772,323500,375027,425010,476258,548943,650104,749889,851063,954198,1097989,1301022,1495463,1699401,1925553,2087044,2358041,2753731,3227466,3709932,5320783) ; cuantía
[1]   3420   7829  12595  17503  22532  27577  32530  37501  42477
[10] 47836  57511  73004  86908 105503 131445 156548 184365 222845
[19] 273772 323500 375027 425010 476258 548943 650104 749889 851063
[28]  954198 1097989 1301022 1495463 1699401 1925553 2087044 2358041
2753731
[37] 3227466 3709932 5320783
```

> # Añadimos las frecuencias observadas y calculamos la media de la muestra

```
>siniestros=c(4114,10116,15796,15611,15749,16514,16891,16540,14817,15824,35342,29496,15447,29457,9553,6437,5069,4976,2898,1876,1417,973,818,1220,738,611,407,313,467,337,229,143,107,75,79,72,37,10,32);
media=weighted.mean(cuantia,siniestros); media
[1] 84212.6
```

Para hacer la prueba de bondad de ajuste es necesario crear la función tal como sigue:

> # FUNCION CONCHI (Prueba Chi-cuadrado)

```
> conchi <- function(valfcont, fteocont, par) { D <- sum( (valfcont-fteocont)**2/fteocont)
+ grad <- length(valfcont)-par-1 ; pvalue <- 1-pchisq(D, grad) ; print(D) ; print(pvalue) }
```

5.2. Frecuencia Teórica Exponencial:

> # 1. Empezamos con la distribución exponencial, hay que calcular el parámetro b de la distribución estadística exponencial

```
> b=1/media; b
[1] 1.187471e-05
```

```
> # 2. Se calcula las probabilidades sobre los límites inferiores de cada intervalo del importe de cuantía
>x=c(0,5001,10001,15001,20001,25001,30001,35001,40001,45001,50001,65001,80001,95001,120001,145001,170001,200001,250001,300001,350001,400001,450001,500001,600001,700001,800001,900001,1000001,1200001,1400001,1600001,1800001,2000001,2200001,2500001,3000001,3500001,4000001) ;
prexp=pexp(x,b); prexp
[1] 0.00000000 0.05765649 0.11197817 0.16316846 0.21140787 0.25686650
[7] 0.29970466 0.34007340 0.37811507 0.41396381 0.44774605 0.53785099
[13] 0.61325454 0.67635536 0.75948597 0.82126384 0.86717360 0.90698118
[19] 0.94862938 0.97163004 0.98433239 0.99134740 0.99522151 0.99736103
[25] 0.99919514 0.99975452 0.99992513 0.99997717 0.99999304 0.99999935
[31] 0.99999994 0.99999999 1.00000000 1.00000000 1.00000000 1.00000000
[37] 1.00000000 1.00000000 1.00000000
> # 3. Hay que crear el vector que incluye la probabilidad acumulada total = 1
>  prexp=c(prexp,1)
> # 4. Hay que obtener la probabilidad de cada intervalo que será la diferencia de uno respecto al otro.
> # La función "diff" nos da la diferencia entre cada componente del vector
>  prexp=diff(prexp); prexp
[1] 5.765649e-02 5.432168e-02 5.119029e-02 4.823941e-02 4.545863e-02
[6] 4.283816e-02 4.036874e-02 3.804167e-02 3.584875e-02 3.378223e-02
[11] 9.010494e-02 7.540355e-02 6.310082e-02 8.313061e-02 6.177788e-02
[16] 4.590975e-02 3.980759e-02 4.164820e-02 2.300065e-02 1.270235e-02
[21] 7.015008e-03 3.874112e-03 2.139519e-03 1.834107e-03 5.593875e-04
[26] 1.706085e-04 5.203418e-05 1.586999e-05 6.316440e-06 5.875548e-07
[31] 5.465430e-08 5.083938e-09 4.729075e-10 4.398981e-11 4.383605e-12
[36] 1.276756e-13 3.330669e-16 0.000000e+00 0.000000e+00
> # 5. Ahora ya se puede calcular la probabilidad teórica según la distribución estadística  exponencial de parámetro b,
>  fteoexp=prexp*sum(siniestros); fteoexp
[1] 1.675544e+04 1.578631e+04 1.487631e+04 1.401876e+04 1.321064e+04
[6] 1.244911e+04 1.173148e+04 1.105521e+04 1.041793e+04 9.817388e+03
[11] 2.618522e+04 2.191288e+04 1.833760e+04 2.415842e+04 1.795314e+04
[16] 1.334174e+04 1.156840e+04 1.210330e+04 6.684174e+03 3.691405e+03
[21] 2.038617e+03 1.125848e+03 6.217613e+02 5.330063e+02 1.625625e+02
```

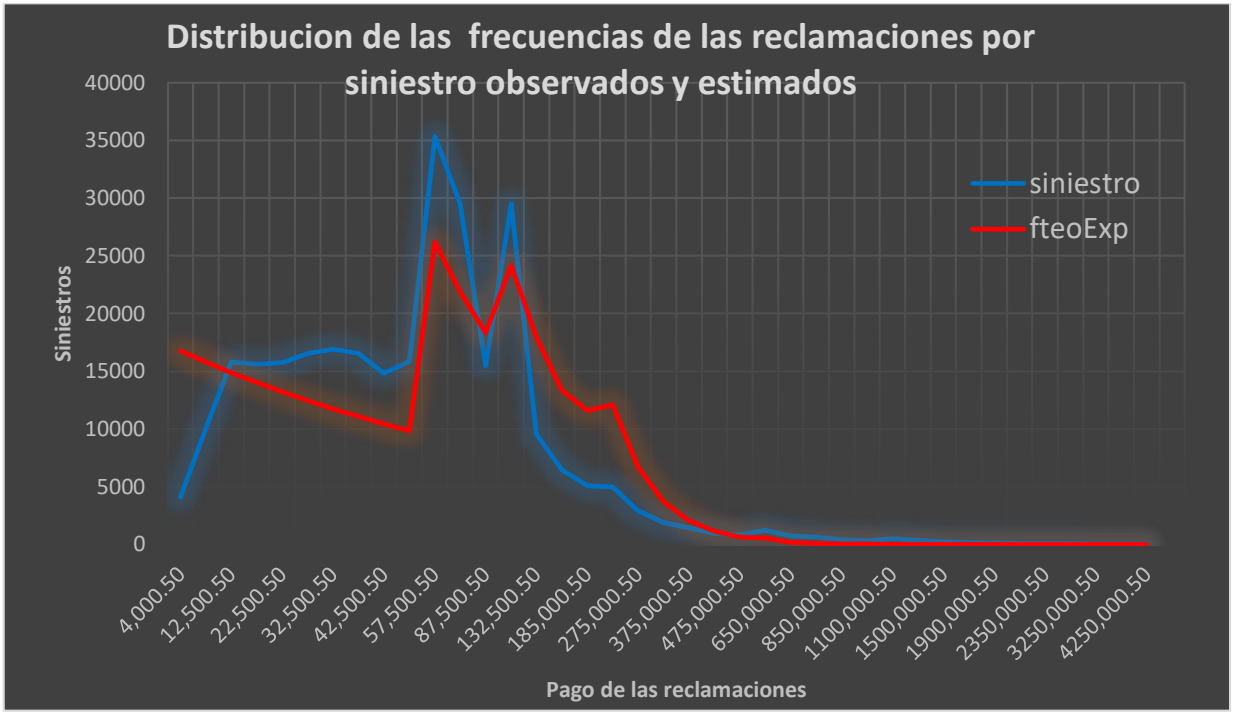
[26] 4.958020e+01 1.512155e+01 4.611946e+00 1.835608e+00 1.707481e-01
[31] 1.588298e-02 1.477433e-03 1.374307e-04 1.278379e-05 1.273911e-06
[36] 3.710356e-08 9.679191e-11 0.000000e+00 0.000000e+00

Cuadro comparativo Frecuencias observadas y frecuencias teóricas

Cuadro comparativo	Frecuencias observada	Frecuencias teóricas exponencial
[1,]	4,114	16,755.44
[2,]	10,116	15,786.31
[3,]	15,796	14,876.31
[4,]	15,611	14,018.76
[5,]	15,749	13,210.64
[6,]	16,514	12,449.11
[7,]	16,891	11,731.48
[8,]	16,540	11,055.21
[9,]	14,817	10,417.93
[10,]	15,824	9,817.40
[11,]	35,342	26,185.22
[12,]	29,496	21,912.88
[13,]	15,447	18,337.60
[14,]	29,457	24,158.42
[15,]	9,553	17,953.14
[16,]	6,437	13,341.74
[17,]	5,069	11,568.40
[18,]	4,976	12,103.30
[19,]	2,898	6,684.17
[20,]	1,876	3,691.41

Cuadro comparativo	Frecuencias observadas	Frecuencias teóricas Exponencial
[21,]	1,417	2,038.62
[22,]	973	1,125.85
[23,]	818	621.76
[24,]	1,220	533.01
[25,]	738	162.56
[26,]	611	49.58
[27,]	407	15.12
[28,]	313	4.61
[29,]	467	1.84
[30,]	337	0.17
[31,]	229	0.02
[32,]	143	0.00
[33,]	107	0.00
[34,]	75	0.00
[35,]	79	0.00
[36,]	72	0.00
[37,]	37	0.00
[38,]	10	0.00
[39,]	32	0.00

Representación gráfica de los siniestros observados y frecuencia teórica exponencial



En los datos de los pagos de las reclamaciones y el número de siniestros observados que se consideraron en este caso para la distribución exponencial tiene como media 84675.28, mientras que la mediana es de 58699.27; debido a que la media es mayor que la mediana, la distribución de los valores observados tiene un sesgo positivo, es decir, hacia la derecha. Como el coeficiente de sesgo es mayor que cero (esto es $P = 0.91973229 > 0$) no tiene una distribución normal. Lo que significa que la ocurrencia de pocos siniestros tiene reclamaciones muy altas, lo que halan la cola de la distribución hacia la derecha.

5.3. Frecuencia teórica Log-normal

```
> # 1. Volvemos a empezar para calcular la frecuencia teórica de la distribución estadística log-normal.
> #  Para esto es necesario encontrar la varianza para poder calcular los parámetros, mu(μ) y sigma(σ)
>  varianza=sum((cuantia-media)^2*siniestros)/sum(siniestros); varianza
[1] 25139552344
>  a=sqrt(varianza)/media; a ; sigma=sqrt(log(1+a^2)); sigma
[1] 1.882789
[1] 1.230449
>  mu=log(media/sqrt(1+a^2)) ; mu
[1] 10.5841
> # 2. Se calcula la probabilidad para cada límite inferior del importe de la cuantía
>  praln=plnorm(x,mu,sigma); praln
[1] 0.00000000 0.04651425 0.13212829 0.21567378 0.29009787 0.35503740
[7] 0.41153964 0.46084304 0.50407294 0.54218141 0.57595530 0.65718019
[13] 0.71685639 0.76213624 0.81675053 0.85470969 0.88221287 0.90628197
[19] 0.93313481 0.95029562 0.96188617 0.97004997 0.97599342 0.98043796
[25] 0.98648395 0.99026308 0.99275450 0.99446656 0.99568289 0.99723467
[31] 0.99813217 0.99868593 0.99904509 0.99928749 0.99945652 0.99962537
[37] 0.99978346 0.99986596 0.99991259
> # 3. Hay que crear el vector que incluye la probabilidad acumulada total = 1
>  praln=c(praln,1); praln
[1] 0.00000000 0.04651425 0.13212829 0.21567378 0.29009787 0.35503740
[7] 0.41153964 0.46084304 0.50407294 0.54218141 0.57595530 0.65718019
[13] 0.71685639 0.76213624 0.81675053 0.85470969 0.88221287 0.90628197
[19] 0.93313481 0.95029562 0.96188617 0.97004997 0.97599342 0.98043796
[25] 0.98648395 0.99026308 0.99275450 0.99446656 0.99568289 0.99723467
[31] 0.99813217 0.99868593 0.99904509 0.99928749 0.99945652 0.99962537
[37] 0.99978346 0.99986596 0.99991259 1.00000000
> # 4. Hay que obtener la probabilidad de cada intervalo que será la diferencia de uno respecto al otro.
> # La función "diff" nos da la diferencia entre cada componente del vector
```

```
>  prln=diff(praln); prln
[1] 4.651425e-02 8.561404e-02 8.354549e-02 7.442410e-02 6.493953e-02
[6] 5.650224e-02 4.930340e-02 4.322990e-02 3.810847e-02 3.377389e-02
[11] 8.122489e-02 5.967619e-02 4.527986e-02 5.461429e-02 3.795916e-02
[16] 2.750318e-02 2.406910e-02 2.685284e-02 1.716081e-02 1.159055e-02
[21] 8.163800e-03 5.943449e-03 4.444538e-03 6.045992e-03 3.779124e-03
[26] 2.491429e-03 1.712052e-03 1.216333e-03 1.551776e-03 8.975080e-04
[31] 5.537584e-04 3.591545e-04 2.424058e-04 1.690304e-04 1.688444e-04
[36] 1.580951e-04 8.249948e-05 4.662576e-05 8.741285e-05
> # 5. Ya se puede calcular las frecuencias teóricas de la distribución log-normal
>  fteoln=prln*sum(siniestros); fteoln
[1]  13517.41306  24880.12506  24278.98667  21628.23759  18871.94715
16420.00281
[7]  14327.96233  12562.95416  11074.62548   9814.96374  23604.60323
17342.37916
[13]  13158.68861  15871.34834  11031.23642   7992.64269   6994.67349
7803.65094
[19]   4987.06770   3368.30797   2372.46561   1727.21371   1291.61844
1757.01377
[25]  1098.24369   724.02920   497.53607   353.47601   450.95850   260.82302
[31]   160.92662   104.37317    70.44507    49.12159    49.06754    45.94370
[37]    23.97501    13.54982    25.40287
```

Cuadro comparativo frecuencias observadas y frecuencias teóricas

Cuadro comparativo	Frecuencias observada	Frecuencias teóricas Log-normal
[1,]	4,114	13,517.41
[2,]	10,116	24,880.13
[3,]	15,796	24,278.99
[4,]	15,611	21,628.24
[5,]	15,749	18,871.95
[6,]	16,514	16,420.00
[7,]	16,891	14,327.96
[8,]	16,540	12,562.95
[9,]	14,817	11,074.63
[10,]	15,824	9,814.96
[11,]	35,342	23,604.60
[12,]	29,496	17,342.38
[13,]	15,447	13,158.69
[14,]	29,457	15,871.35
[15,]	9,553	11,031.24
[16,]	6,437	7,992.64
[17,]	5,069	6,994.67
[18,]	4,976	7,803.65
[19,]	2,898	4,987.07
[20,]	1,876	3,368.31

Cuadro comparativo	Frecuencias observada	Frecuencias teóricas Log-normal
[21,]	1,417	2,372.47
[22,]	973	1,727.21
[23,]	818	1,291.62
[24,]	1,220	1,757.01
[25,]	738	1,098.24
[26,]	611	724.03
[27,]	407	497.54
[28,]	313	353.48
[29,]	467	450.96
[30,]	337	260.82
[31,]	229	160.93
[32,]	143	104.37
[33,]	107	70.45
[34,]	75	49.12
[35,]	79	49.07
[36,]	72	45.94
[37,]	37	23.98
[38,]	10	13.55
[39,]	32	25.40

Representación gráfica de los siniestros observados y frecuencia teórica Log-normal

En los datos de los pagos de las reclamaciones y el número de siniestros observados que se consideraron en este caso para la distribución Log-normal tiene como media 84476.15, mientras que la mediana es de 37287.37; debido a que la media es mayor que la mediana, la distribución de los valores observados tiene un sesgo positivo, es decir, hacia la derecha. Como el coeficiente de sesgo es mayor que cero (esto es $P = 0.91819379 > 0$) no tiene una distribución normal. Lo que significa que la ocurrencia de pocos siniestros tiene reclamaciones muy altas, lo que halan la cola de la distribución hacia la derecha.

5.4. Frecuencia teórica Gamma

> # 1. Iniciamos el procedimiento para la distribución gamma, empezamos calculando los parámetros (a=α, b=β)

> bgamma=media/varianza; bgamma ; agamma=media*bgamma; agamma

[1] 3.349805e-06

[1] 0.2820958

> # 2. Se calcula la probabilidad para cada límite inferior del importe de la cuantía

> pragamma=pgamma(x,agamma,bgamma); pragamma

[1] 0.0000000 0.3491467 0.4229881 0.4725208 0.5106165 0.5418409 0.5684035

[8] 0.5915610 0.6121039 0.6305656 0.6473247 0.6898725 0.7240817 0.7524833

```
[15] 0.7907513 0.8210039 0.8455626 0.8695157 0.8999511 0.9222114
0.9388981
[22] 0.9516271 0.9614644 0.9691440 0.9799756 0.9868411 0.9912696
0.9941639
[29] 0.9960746 0.9981984 0.9991608 0.9996047 0.9998121 0.9999101
0.9999567
[36] 0.9999854 0.9999976 0.9999996 0.9999999
> # 3. Hay que crear el vector que incluye la probabilidad acumulada total = 1
> # pragamma=c(pragamma,1); pragamma
> # 4. Hay que obtener la probabilidad de cada intervalo que será la diferencia
de uno respecto al otro.
> # La función "diff" nos da la diferencia entre cada componente del vector
>  prgamma=diff(pragamma); prgamma
[1] 3.491467e-01 7.384143e-02 4.953270e-02 3.809570e-02 3.122435e-02
[6] 2.656265e-02 2.315750e-02 2.054290e-02 1.846164e-02 1.675914e-02
[11] 4.254779e-02 3.420925e-02 2.840159e-02 3.826799e-02 3.025259e-02
[16] 2.455865e-02 2.395315e-02 3.043543e-02 2.226023e-02 1.668677e-02
[21] 1.272900e-02 9.837263e-03 7.679631e-03 1.083155e-02 6.865479e-03
[26] 4.428570e-03 2.894243e-03 1.910764e-03 2.123721e-03 9.624009e-04
[31] 4.439067e-04 2.074664e-04 9.795646e-05 4.662835e-05 2.869463e-05
[36] 1.215905e-05 2.013879e-06 3.396468e-07
> # 5. Ya se pueden calcular las frecuencias teóricas
>  fteogamma=prgamma*sum(siniestros); fteogamma
[1] 1.014648e+05 2.145891e+04 1.439460e+04 1.107092e+04 9.074046e+03
[6] 7.719319e+03 6.729756e+03 5.969932e+03 5.365099e+03 4.870339e+03
[11] 1.236473e+04 9.941482e+03 8.253730e+03 1.112098e+04 8.791646e+03
[16] 7.136941e+03 6.960976e+03 8.844781e+03 6.469001e+03 4.849308e+03
[21] 3.699151e+03 2.858787e+03 2.231762e+03 3.147734e+03 1.995163e+03
[26] 1.286978e+03 8.410902e+02 5.552834e+02 6.171702e+02 2.796814e+02
[31] 1.290028e+02 6.029140e+01 2.846693e+01 1.355057e+01 8.338889e+00
[36] 3.533517e+00 5.852495e-01 9.870408e-02
```

Cuadro comparativo frecuencias observadas y frecuencias teóricas

Cuadro comparativo	Frecuencias observada	Frecuencias teóricas Gamma
[1,]	4,114	101,464.80
[2,]	10,116	21,458.91
[3,]	15,796	14,394.60
[4,]	15,611	11,070.92
[5,]	15,749	9,074.05
[6,]	16,514	7,719.32
[7,]	16,891	6,729.76
[8,]	16,540	5,969.93
[9,]	14,817	5,365.10
[10,]	15,824	4,870.34
[11,]	35,342	12,364.73
[12,]	29,496	9,941.48
[13,]	15,447	8,253.73
[14,]	29,457	11,120.98
[15,]	9,553	8,791.65
[16,]	6,437	7,136.94
[17,]	5,069	6,960.98
[18,]	4,976	8,844.78
[19,]	2,898	6,469.00
[20,]	1,876	4,849.31

Cuadro comparativo	Frecuencias observada	Frecuencias teóricas Gamma
[21,]	1,417	3,699.15
[22,]	973	2,858.80
[23,]	818	2,231.76
[24,]	1,220	3,147.73
[25,]	738	1,995.16
[26,]	611	1,286.98
[27,]	407	841.10
[28,]	313	555.28
[29,]	467	617.17
[30,]	337	279.68
[31,]	229	129.00
[32,]	143	60.30
[33,]	107	28.47
[34,]	75	13.55
[35,]	79	8.34
[36,]	72	3.53
[37,]	37	0.59
[38,]	10	0.10
[39,]	32	0.00

Representación gráfica de los siniestros observados y frecuencia teórica gamma

En los datos de los pagos de las reclamaciones y el número de siniestros observados que se consideraron en este caso para la distribución Gamma tiene como media 85092.97, mientras que la mediana es de 15394.44; debido a que la media es mayor que la mediana, la distribución de los valores observados tiene un sesgo positivo, es decir, hacia la derecha. Como el coeficiente de sesgo es mayor que cero (esto es P=1.31492593>0) no tiene una distribución normal. Lo que significa que la ocurrencia de pocos siniestros tiene reclamaciones muy altas, lo que halan la cola de la distribución hacia la derecha.

5.5. Frecuencia teórica Inversa gaussiana

> # 1. Para la distribución estadística inversa gaussiana se debe calcular el parámetro beta y usar la distribución pinvgauss del paquete statmod;

> library(statmod); beta=varianza/media

> library(statmod); beta=varianza/media

> # 2. Se calcula la probabilidad para cada límite inferior del importe de la cuantía

> praig=pinvgauss(x,media,media^2/beta); praig

[1] 0.00000000 0.03859115 0.16162080 0.27197314 0.35896120 0.42777202

[7] 0.48333674 0.52915187 0.56762506 0.60043797 0.62879296 0.69487527

[13] 0.74225134 0.77803219 0.82164654 0.85277682 0.87610737 0.89735565

[19] 0.92245421 0.93965458 0.95202876 0.96124550 0.96829192 0.97378970

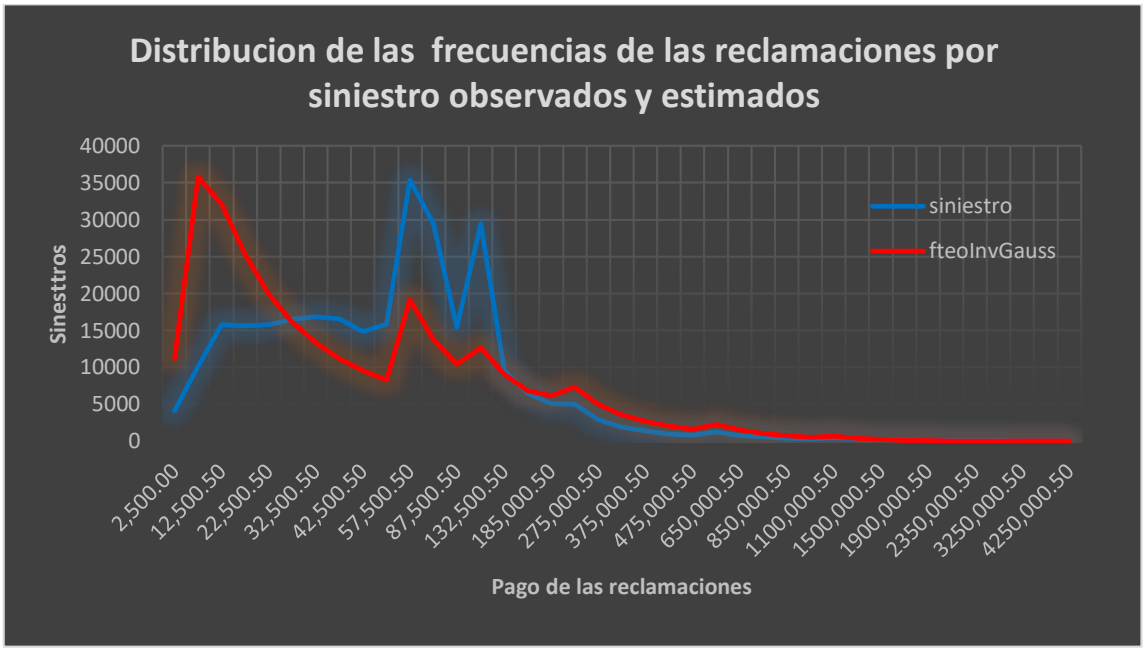
```
[25] 0.98165363 0.98683748 0.99037890 0.99286381 0.99464397 0.99689793
[31] 0.99815173 0.99887499 0.99930367 0.99956316 0.99972288 0.99985758
[37] 0.99995134 0.99998282 0.99999378
> # 3. Hay que crear el vector que incluye la probabilidad acumulada total = 1
>  praig=c(praig,1); praig
[1] 0.00000000 0.03859115 0.16162080 0.27197314 0.35896120 0.42777202
[7] 0.48333674 0.52915187 0.56762506 0.60043797 0.62879296 0.69487527
[13] 0.74225134 0.77803219 0.82164654 0.85277682 0.87610737 0.89735565
[19] 0.92245421 0.93965458 0.95202876 0.96124550 0.96829192 0.97378970
[25] 0.98165363 0.98683748 0.99037890 0.99286381 0.99464397 0.99689793
[31] 0.99815173 0.99887499 0.99930367 0.99956316 0.99972288 0.99985758
[37] 0.99995134 0.99998282 0.99999378 1.00000000
> # 4. Se calcula la probabilidad de cada tramo
>  prig=diff(praig); prig
[1] 3.859115e-02 1.230297e-01 1.103523e-01 8.698806e-02 6.881082e-02
[6] 5.556472e-02 4.581513e-02 3.847319e-02 3.281291e-02 2.835499e-02
[11] 6.608231e-02 4.737607e-02 3.578085e-02 4.361434e-02 3.113028e-02
[16] 2.333055e-02 2.124828e-02 2.509856e-02 1.720037e-02 1.237418e-02
[21] 9.216743e-03 7.046422e-03 5.497776e-03 7.863929e-03 5.183849e-03
[26] 3.541424e-03 2.484909e-03 1.780158e-03 2.253958e-03 1.253806e-03
[31] 7.232587e-04 4.286820e-04 2.594832e-04 1.597252e-04 1.346962e-04
[36] 9.376634e-05 3.147190e-05 1.096217e-05 6.222161e-06
> # 5. Y ya se pueden calcular las frecuencias teóricas
>  fteoig=prig*sum(siniestros); fteoig
[1] 11214.897410 35753.401350 32069.271841 25279.425642 19996.973551
[6] 16147.552547 13314.244033 11180.615659  9535.694471  8240.188180
[11] 19204.047285 13767.865151 10398.202677 12674.677546  9046.708089
[16]  6780.044853  6174.920035  7293.841622  4998.564806  3596.036537
[21]  2678.459335  2047.746582  1597.697603  2285.320669  1506.467923
[26]  1029.166225   722.134421   517.328268   655.018323   364.365942
[31]   210.184774   124.578424    75.407880    46.417411    39.143796
[36]    27.249248     9.145985     3.185694     1.808210
```

Cuadro comparativo frecuencias observadas y frecuencias teóricas

Cuadro comparativo	Frecuencias observada	Frecuencias teórica Inversa gaussiana
[1,]	4,114	11,214.90
[2,]	10,116	35,753.40
[3,]	15,796	32,069.27
[4,]	15,611	25,279.43
[5,]	15,749	19,996.97
[6,]	16,514	16,147.55
[7,]	16,891	13,314.24
[8,]	16,540	11,180.62
[9,]	14,817	9,535.69
[10,]	15,824	8,240.19
[11,]	35,342	19,204.05
[12,]	29,496	13,767.87
[13,]	15,447	10,398.20
[14,]	29,457	12,674.68
[15,]	9,553	9,046.71
[16,]	6,437	6,780.04
[17,]	5,069	6,174.92
[18,]	4,976	7,293.84
[19,]	2,898	4,998.56
[20,]	1,876	3,596.04

Cuadro comparativo	Frecuencias observada	Frecuencias teórica Inversa gaussiana
[21,]	1,417	2,678.46
[22,]	973	2,047.75
[23,]	818	1,597.70
[24,]	1,220	2,285.32
[25,]	738	1,506.47
[26,]	611	1,029.17
[27,]	407	722.13
[28,]	313	517.33
[29,]	467	655.02
[30,]	337	364.37
[31,]	229	210.18
[32,]	143	124.58
[33,]	107	75.41
[34,]	75	46.42
[35,]	79	39.14
[36,]	72	27.25
[37,]	37	9.15
[38,]	10	3.19
[39,]	32	1.81

Representación gráfica de los siniestros observados y frecuencia teórica Inversa gaussiana



En los datos de los pagos de las reclamaciones y el número de siniestros observados que se consideraron en este caso para la distribución inversa gaussiana tiene como media 84566.21, mientras que la mediana es de 30678.07; debido a que la media es mayor que la mediana, la distribución de los valores observados tiene un sesgo positivo, es decir, hacia la derecha. Como el coeficiente de sesgo es mayor que cero (esto es $P = 1.01522347 > 0$) no tiene una distribución normal. Lo que significa que la ocurrencia de pocos siniestros tiene reclamaciones muy altas, lo que halan la cola de la distribución hacia la derecha.

5.6. Prueba de bondad de ajuste

> # 1. Para hacer el contraste de bondad de ajuste, se puede hacer primero un cuadro para comparar los resultados observados (variables siniestros) con las frecuencias teóricas obtenidas;

```
> fteogamma[39]=0
> Ajustes = cbind(siniestros,fteoexp,fteoln,fteogamma,fteoig)
> Ajustes;   matplot(Ajustes, type="l")
```

	fteoexp	fteoln	fteogamma	fteoig
[1,]	4114	1.675544e+04	13517.41306	1.014648e+05	11214.897410
[2,]	10116	1.578631e+04	24880.12506	2.145891e+04	35753.401350

[3,]	15796	1.487631e+04	24278.98667	1.439460e+04	32069.271841
[4,]	15611	1.401876e+04	21628.23759	1.107092e+04	25279.425642
[5,]	15749	1.321064e+04	18871.94715	9.074046e+03	19996.973551
[6,]	16514	1.244911e+04	16420.00281	7.719319e+03	16147.552547
[7,]	16891	1.173148e+04	14327.96233	6.729756e+03	13314.244033
[8,]	16540	1.105521e+04	12562.95416	5.969932e+03	11180.615659
[9,]	14817	1.041793e+04	11074.62548	5.365099e+03	9535.694471
[10,]	15824	9.817388e+03	9814.96374	4.870339e+03	8240.188180
[11,]	35342	2.618522e+04	23604.60323	1.236473e+04	19204.047285
[12,]	29496	2.191288e+04	17342.37916	9.941482e+03	13767.865151
[13,]	15447	1.833760e+04	13158.68861	8.253730e+03	10398.202677
[14,]	29457	2.415842e+04	15871.34834	1.112098e+04	12674.677546
[15,]	9553	1.795314e+04	11031.23642	8.791646e+03	9046.708089
[16,]	6437	1.334174e+04	7992.64269	7.136941e+03	6780.044853
[17,]	5069	1.156840e+04	6994.67349	6.960976e+03	6174.920035
[18,]	4976	1.210330e+04	7803.65094	8.844781e+03	7293.841622
[19,]	2898	6.684174e+03	4987.06770	6.469001e+03	4998.564806
[20,]	1876	3.691405e+03	3368.30797	4.849308e+03	3596.036537
[21,]	1417	2.038617e+03	2372.46561	3.699151e+03	2678.459335
[22,]	973	1.125848e+03	1727.21371	2.858787e+03	2047.746582
[23,]	818	6.217613e+02	1291.61844	2.231762e+03	1597.697603
[24,]	1220	5.330063e+02	1757.01377	3.147734e+03	2285.320669
[25,]	738	1.625625e+02	1098.24369	1.995163e+03	1506.467923
[26,]	611	4.958020e+01	724.02920	1.286978e+03	1029.166225
[27,]	407	1.512155e+01	497.53607	8.410902e+02	722.134421
[28,]	313	4.611946e+00	353.47601	5.552834e+02	517.328268
[29,]	467	1.835608e+00	450.95850	6.171702e+02	655.018323
[30,]	337	1.707481e-01	260.82302	2.796814e+02	364.365942
[31,]	229	1.588298e-02	160.92662	1.290028e+02	210.184774
[32,]	143	1.477433e-03	104.37317	6.029140e+01	124.578424
[33,]	107	1.374307e-04	70.44507	2.846693e+01	75.407880
[34,]	75	1.278379e-05	49.12159	1.355057e+01	46.417411
[35,]	79	1.273911e-06	49.06754	8.338889e+00	39.143796

```
[36,]  72  3.710356e-08  45.94370  3.533517e+00  27.249248
[37,]  37  9.679191e-11  23.97501  5.852495e-01   9.145985
[38,]  10  0.000000e+00  13.54982  9.870408e-02   3.185694
[39,]  32  0.000000e+00  25.40287  1.014648e+05   1.808210
> # 2. Ya podemos aplicar la función "conchi" para hacer el contraste de bondad de ajuste, aunque hay que tener en cuenta que hay que agrupar las frecuencias inferiores a 5.
> # Para la exponencial:
> #H0: Los siniestros tienen una distribución exponencial
> #Ha: Los siniestros no tienen una distribución exponencial
>  siniestros1=c(siniestros[1:27],sum(siniestros[28:39]))
>  fteoexp1=c(fteoexp[1:27],sum(fteoexp[28:39]))
>  conchi(siniestros1,fteoexp1,1)
[1] 610486.3
[1] 0
> # Para la log-normal
> #H0: Los siniestros tienen una distribución log-normal
> #Ha: Los siniestros no tienen una distribución log-normal
>  siniestros1=c(siniestros[1:39])
>  fteoln1=c(fteoln[1:39])
>  conchi(siniestros1,fteoln1,2)
[1] 58441.88
[1] 0
> # Como esto es laborioso, el cuadro de resumen nos queda de la siguiente manera:
```

Valores	Fteoexp	Fteoln	fteogamma	Fteoig
D(x^2)	610486.30	58441.88	326175.10	111386.07
P(probabilidad)	0	0	0	0

Por lo tanto se rechaza la Hipótesis nula (H0) de que los siniestros siguen una distribución: exponencial, log-normal, gamma e inversa gaussiana; dado que la chi-cuadrado es mayor que el chi-teórico a un nivel de significancia de 0.05, es decir, no existe evidencia estadística de que los datos esperados se ajusten a los datos observados(datos reales).

VI. CONCLUSIONES

Después de haber finalizado nuestro trabajo monográfico llegamos a las siguientes conclusiones que:

a) Las distribuciones continuas descritas en este documento son muy útiles en los seguros generales, ya que muchos de los cálculos que hemos desarrollado podrían ser llevado a cabo tan sólo disponer de datos empíricos de la frecuencia de siniestralidad y del montante de los siniestros. Además, algo que le otorga vital importancia es especialmente que sus propiedades son bien conocidas, se encuentran totalmente definidas a través de un pequeño número de parámetros (uno en el caso de la exponencial y la inversa gaussiana, dos en la log-normal y gamma); sin omitir que dichas distribuciones nos permiten efectuar inferencias acerca del comportamiento de la cartera de seguros.

b) El lenguaje de programación R, que consta de un sistema base y de librerías adicionales, cubre las necesidades descritas como herramienta de cálculo gratuita y de libre acceso, siendo además un software muy dinámico, ya que las funciones y las librerías se van ampliando gracias a las aportaciones de usuarios que pasan un estricto proceso de elaboración y contrastación. Estamos hablando, además, de un software ampliamente utilizado en el campo actuarial, siendo constante la aparición de librerías especializadas en temas actuariales como el paquete actuar que contiene funciones de distribuciones de pérdidas, permite manipular y almacenar datos de la forma intervalo-frecuencial con facilidad. Dichas funciones convierten a R en una plataforma para modelizar y calcular variables de teoría del riesgo, calcular coeficientes de ajuste para cualquier frecuencia y severidad de siniestros y permite la simulación de portafolios y modelos de credibilidad.

c) En cuanto a la estimación de las frecuencias teóricas y la prueba de bondad de ajuste que se realizó en R para cada una de las distribuciones de probabilidad continua, se puede apreciar de que los datos reales (siniestros) de esta compañía no se ajustan a ninguno de estos modelos

estadísticos, dado que no existe homogeneidad en cada uno de los siniestros y tampoco existe uniformidad en el pago de las indemnizaciones. Sin embargo, a la hora de poder decidir según los resultados podríamos optar por la distribución log-normal dado que un mejor ajuste tiene un menor valor D. Sin embargo, el valor de D (De la prueba de bondad de ajuste) por sí solo no permite rechazar o no la hipótesis de que la distribución real del número de siniestros es la considerada; por lo que es necesario calcular la probabilidad *p* ya que a mayor valor de este, mejor es el ajuste. Además, esta distribución es frecuentemente utilizada como modelo de distribución del coste de un siniestro, debido a que es asimétrica positiva que es una característica general de la distribución del costo de los siniestros que no tienen una distribución normal.

d) El presente trabajo recoge los frutos de la investigación realizada sobre el sistema informático R. Entre las razones de porque utilizarlo estaba el hecho de que empezamos a encontrar libros y artículos internacionales que incluían ejemplos actuariales solucionados con R o que presentaban nuevas librerías especializadas en dicha temática, pensadas tanto para la práctica profesional como para la enseñanza y la investigación. Cabe mencionar que la apuesta para este lenguaje de programación fue arriesgada dado que aparte de ser desconocido para nosotros, nunca habíamos programado en una plataforma como esta, pero el resultado ha sido muy satisfactorio.

e) Los actuarios y la tecnología podemos adaptarnos para generar mayor aprovechamiento del tiempo y reducción de las cargas de trabajo. Que la ausencia algunos conocimientos no es limitación cuando se tiene interés y dedicación, y que podemos dirigir o colaborar en trabajos tan pequeños como este o grandes con el fin de brindar soluciones.

VII. RECOMENDACIONES

- A las compañías de seguros:

Actualizar y capacitar a su área técnica en el manejo del programa R, ya que este evita realizar cálculos repetitivos y complejos, dado que los actuarios y la tecnología podemos adaptarnos para generar mayor aprovechamiento del tiempo y reducción de las cargas de trabajo. A pesar de la existencia de programas como Excel, sqlite, Access, no son las únicas herramientas a las que se les pueden sacar provecho, que también podemos utilizar otros programas como el R que nos brinden mayor desempeño. Teniendo además como ventaja que es un programa de libre acceso, por lo que no le generaría ningún costo.

- A estudiantes y profesores:

Romper la zona de confort y no limitarse a lo ya aprendido e incentivarse en la adquisición de nuevos conocimientos para generar nuevas soluciones. Es necesario que al momento de programar cuiden la ortografía, dado que R distingue entre mayúsculas y minúsculas; así como en la aplicación de tildes, ya que al hacer caso omiso a estas indicaciones le produciría errores a la hora de correr el programa.

- A la institución Universitaria:

Incentivar a que brinden apoyo a los estudiantes impartiendo cursos de informática y programación en R. Ampliar el pensum académico de la carrera de Ciencias Actuariales y Financieras y carrera afines con la extensión de un ***Sistemas Actuariales Informatizados III (SAI III).*** Así también facilitando documentación referente a nuestro campo y réplicas de los documentos ya existentes.

VIII. BIBLIOGRAFÍA

8.1. Referencias Bibliográficas:

1. Allen L. Wester; *Estadística aplicada a los negocios y la economía*. Tercera edición, Colombia 2001.
2. Rosario Collatón Chicana; *Introducción al uso de R y R Commander para el análisis estadístico de datos en ciencias sociales*; 2014.
3. Gudelia Figueroa Preciado, José A. Montoya laos; *Introducción al Software Estadístico R.* Agosto, 2015
4. Emmanuel Paradis; *R para principiantes*. Instituto de Ciencias de la evolución.
5. Julio Sergio Santana, Efraín Mateo Farfán; *El arte de programar en R: Un lenguaje para la estadística*. Noviembre de 2014.
6. Francesc Carmona; *Curso básico de R*. Febrero, 2007.
7. F.Tusell; *Lectura, manipulación y análisis de datos en R*. Curso 2004-2005
8. Arthur Chapentier; *Computational Actuarial Science with R*. University of Québec at Montreal Canada.
9. Shyamal Kumar; *Using R for Actuarial Science*.
10. Emilio Gómez Déniz; *Economía Financiera Cuantitativa y Actuarial. Modelización de Riesgo e Incertidumbre en Seguros y Auditoría Contable*. Curso Doctorado. Bienio 2008-2010.
11. *Introducción a R; Notas sobre R: Un entorno de programación para Análisis de Datos y Gráficos*. Versión 1.0.1 (2000-05-16)
12. Maikol Solís; Taller: *Introducción a la Estadística para Actuarios* (agosto, 2016. León, Nicaragua)

8.2. Referencias Electrónicas:

1. http://www.tutorialr.es/es/index.html
2. http://osluz.unizar.es/proyectos/r
3. http://itamactuaria.blogspot.com/2009/03/paquete-actuar-en-r-statistics.html
4. https://www.r-project.org/about.html
5. http://www.econanalytics.org/actuar-paquete-de-funciones-para-ciencias-actuariales-en-r/
6. http://www.holamundo.mx/proyecto-r-una-herramienta-actuarial/

IX. ANEXOS

Anexo 1: Tabla del coste medio de 290,608 siniestros que ocurrieron en ramo de hogar de una compañía de seguro de Nicaragua.

INTERVALOS DE COSTE	COSTE MEDIO	FRECUENCIA ABSOLUTA
0- 5,000	3420	4114
5,001-10,000	7829	10116
10,001-15,000	12595	15796
15,001-20,000	17503	15611
20,001-25,000	22532	15749
25,001-30,000	27577	16514
30,001-35,000	32530	16891
35,001-40,000	37501	16540
40,001-45,000	42477	14817
45,001-50,000	47836	15824
50,001-65,000	57511	35342
65,001-80,000	73004	29496
80,001-95,000	86908	15447
95,001-120,000	105503	29457
120,001-145,000	131445	9553
145,001-170,000	156548	6437
170,001-200,0000	184365	5069
200,001-250,000	222845	4976
250,001-300,000	273772	2898
300,001-350,000	323500	1876

INTERVALOS DE COSTE	COSTE MEDIO	FRECUENCIA ABSOLUTA
350,001-400,000	375027	1417
400,001-450,000	425010	973
450,001-500,000	476258	818
500,001-600,000	548943	1220
600,001-700,000	650104	738
700,001-800,000	749889	611
800,001-900,000	851063	407
900,001-1,000,000	954198	313
1,000,001-1,200,000	1097989	467
1,200,001-1,400,000	1301022	337
1,400,001-1,600,000	1495463	229
1,600,000-1,800,000	1699401	143
1,800,001-2,000,000	1925553	107
2,000,001-2,200,000	2087044	75
2,200,001-2,500,000	2358041	79
2,500,001-3,000,000	2753731	72
3,000,001-3,500,000	3227466	37
3,000,001-4,000,000	3709932	10
Más de 4,000,000	5320783	32

Anexo 2: Algoritmo de programación en R.

```
# Introducir el vector de puntos medios de los
intervalos(ecosdelaeconomia.wordpress.com)
cuantia=c(3420,7829,12595,17503,22532,27577,32530,37501,42477,47836,57
511,73004,86908,105503,131445,156548,184365,222845,273772,323500,375
027,425010,476258,548943,650104,749889,851063,954198,1097989,1301022
,1495463,1699401,1925553,2087044,2358041,2753731,3227466,3709932,532
0783) ; cuantia
# Añadimos las frecuencias observadas y calculamos la media de la muestra
siniestros=c(4114,10116,15796,15611,15749,16514,16891,16540,14817,15824
,35342,29496,15447,29457,9553,6437,5069,4976,2898,1876,1417,973,818,12
20,738,611,407,313,467,337,229,143,107,75,79,72,37,10,32);
media=weighted.mean(cuantia,siniestros); media

# FUNCION CONCHI (Prueba Chi-cuadrado)
Conchi <- function(valfcont, fteocont, par) { D <- sum( (valfcont-
fteocont)**2/fteocont)
+ grad <- length(valfcont)-par-1 ; pvalue <- 1-pchisq(D, grad) ; print(D) ;
print(pvalue) }

# 1. Empezamos con la distribución exponencial, hay que calcular el parámetro
b de la exponencial
b=1/media; b
# 2. se calcula las probabilidades sobre los limites inferiores de cada intervalo
# del importe de cuantía

x=c(0,5001,10001,15001,20001,25001,30001,35001,40001,45001,50001,6500
1,80001,95001,120001,145001,170001,200001,250001,300001,350001,40000
1,450001,500001,600001,700001,800001,900001,1000001,1200001,1400001,
1600001,1800001,2000001,2200001,2500001,3000001,3500001,4000001) ;
prexp=pexp(x,b); prexp
# 3. creamos el vector que incluye la probabilidad acumulada total = 1
 prexp=c(prexp,1)
```

```
# 4. la probabilidad de cada intervalo será la diferencia de uno respecto a otro.
# La función "diff" nos da la diferencia entre cada componente del vector
 prexp=diff(prexp); prexp
# 5. ahora ya se puede calcular la probabilidad teórica según la exponencial de parámetro b,
 fteoexp=prexp*sum(siniestros); fteoexp
# 1. volvemos a empezar para calcular la frecuencia teorica de la distribución log-normal,
# es necesarioencontrar la varianza para poder calcular los parámetros, mu y sigma
 varianza=sum((cuantia-media)^2*siniestros)/sum(siniestros); varianza
 a=sqrt(varianza)/media; a ; sigma=sqrt(log(1+a^2)); sigma
 mu=log(media/sqrt(1+a^2)) ; mu
# 2. se calcula la probabilidad para cada limite inferior del importe de la cuantía
 praln=plnorm(x,mu,sigma); praln
# 3. se completa la probabilidad acumulada hasta 1
 praln=c(praln,1); praln
# 4. se calcula la probabilidad de cada intervalo
 prln=diff(praln); prln
# 5. ya se puede calcular las frecuencias teoricas de la distribución log-normal
 fteoln=prln*sum(siniestros); fteoln
# 1. iniciamos el procedimiento para la distribución gamma, empezamos calculando
# los parámetros (a,b)
 bgamma=media/varianza; bgamma  ; agamma=media*bgamma; agamma
 pragamma=pgamma(x,agamma,bgamma); pragamma
# 5. se completa la probabilidad acumulada hasta el total
# pragamma=c(pragamma,1); pragamma
# 6. se calcula la probabilidad de cada intervalo
 prgamma=diff(pragamma); prgamma
# 7. ya se pueden calcular las frecuencias teoricas
 fteogamma=prgamma*sum(siniestros); fteogamma
# 1. para la distribución inversa gaussiana los parámetros hay que calcular
```

```
#  el parámetro beta y usar la distribución pinvgauss del paquete statmod;
library(statmod); beta=varianza/media ;
praig=pinvgauss(x,media,media^2/beta); praig
# 2. se complementa la probabilidad acumulada hasta 1
 praig=c(praig,1); praig
# 3. se calcula la probabilidad de cada tramo
 prig=diff(praig); prig
# 4. y ya se pueden calcular las frecuencias teoricas
 fteoig=prig*sum(siniestros); fteoig
# b) Para hacer el contraste de bondad de ajuste, se puede hacer primero un
cuadro
# para comparar los resultados observados (variables siniestros) con las
frecuencias
# teoricas obtenidas;
Ajustes = cbind(c(siniestros,0),fteoexp,fteoln,fteogamma,fteoig)
Ajustes;   matplot(Ajustes, type="l")
 siniestros=c(siniestros,0); siniestros
# Ya podemos aplicar la función "conchi" para hacer el contraste de bondad de
ajuste,
# aunque hay que tener en cuenta que hay que agrupar las frecuencias
inferiores a 5.
# Para la exponencial:
#H0: Los siniestros tienen una distribución exponencial
#Ha: Los siniestros no tienen una distribución exponencial
 siniestros1=c(siniestros[1:27],sum(siniestros[28:39]))
 fteoexp1=c(fteoexp[1:27],sum(fteoexp[28:39]))
 conchi(siniestros1,fteoexp1,1)
# Para la log-normal
#H0: Los siniestros tienen una distribución log-normal
#Ha: Los siniestros no tienen una distribución log-normal
 siniestros1=c(siniestros[1:39])
 fteoln1=c(fteoln[1:39])
 conchi(siniestros1,fteoln1,2)
```

Anexo 3: Glosario.

Mean: Es la función que calcula la media de la variable cuantitativa edad.

Attach: Para vincular las variables con su marco de datos.

Detach: Para desvincular las variables de su marco de datos.

Getwd: Nos dice cuál es el actual directorio de trabajo.

Setwd: Si queremos volver al directorio que se encuentra en el nivel inmediatamente superior al actual podemos hacer *setwd("...").*

Dir: Nos da un listado de los ficheros en el actual directorio de trabajo.

Seg: Crear vectores.

Length: Nos da la longitud de un vector:

Rep: Repite una secuencia de números.

Matrix: Para crear una matriz.

Nrow: Número de filas.

Ncol: Columnas.

Dim: Reportar el número de filas y columnas que tiene la matriz.

Data.Frame: Para crear un marco de datos.

Read.table: Es la función que nos permite leer el archivo MMR.csv desde R.

Header: Es un argumento de la función *read table* que lee la primera fila del archivo MMR.csv, como una fila que contiene los nombres de las variables de la base de datos.

Sep: Indica el elemento que actúa como separador de los datos en este caso, la coma.

***Caseid*:** Variable de tipo carácter y no tipo numérico.

***Función Labels*:** Permite conocer los casos (filas) que han sido seleccionadas para integrar el nuevo marco de datos, y además nos lista los nombres de las variables consideradas en él.

Subset: Función que selecciona todos los elementos del objeto.