Introducción al lenguaje de programación R: Manejo eficiente de datos, gráficos y automatización de procesos

Profesor titular Lic. Matias Poullain

# Módulo 0: Instalación de R y RStudio

Antes de comenzar, es necesario instalar el lenguaje R y su interfaz RStudio en la computadora por lo que se espera que para la primer clase, usted haya completado este módulo.

En él hay una guía con las instrucciones de cómo hacerlo. En caso de ya tenerlos instalados, no es necesario realizar el módulo 0.

## Paso 1: Descargar R

En primer lugar, es necesario descargar el lenguaje R. Para ello vamos a la página web oficial: <https://www.r-project.org> y clickeamos en **download R** (“Descargar R”):

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Seleccionamos el distribuidor de Argentina:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Y descargamos R para Windows. En el caso de que su computadora tenga un sistema operativo distinto (macOS o Linux) descargar el que corresponda. Dado Windows es el más común, las instrucciones y el curso en su totalidad se dictará en Windows. Sin embargo, la instalación y la utilización de R es muy similar en cualquiera de las tres plataformas.

Luego de clickear sobre **Download R for Windows**:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Clickear sobre **install R for the first time**:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Y descargamos R clickeando sobre **Download R 4.1.3 for Windows**. Puede que la versión no sea la misma, es decir que los números “4.1.3” hayan cambiado. No le preste atención e instálelo de todas formas. Cada tanto, R se actualiza y distribuyen una nueva versión. Sin embargo los cambios entre versiones no son muy importantes, a menos que haya pasado mucho tiempo (unos años) entre versiones, que no es nuestro caso.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Se descargará el archivo **R-4.1.3-win.exe** en su carpeta de descargas.

## Paso 2: Instalar R

En este paso se va a instalar R en la computadora. Esta acción es equivalente a enseñarle a la computadora un idioma nuevo.

Ubique el archivo **R-4.1.3-win.exe** recientemente descargado y ejecútelo con un doble click izquierdo. Al hacerlo, usualmente aparece una ventana con la pregunta **¿Quiere permitir que esta aplicación realice cambios en su dispositivo?**. Si aparece, seleccione **Sí** y empezarán las configuraciones para la instalación.

En la selección de idioma, seleccione **Español** y **Aceptar**:

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Continúe con la instalación clickeando siempre sobre siguiente. Los pasos que debería observar son los siguientes:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Y comenzará la instalación, espere hasta que termine y finalmente clickee sobre **Finalizar.**

## Paso 3: Descargar RStudio

¡Felicidades, su computadora ahora puede hablar en R fluído! Sin embargo, nosotros no… Para ayudar a entender y manejar el lenguaje se utilizará la interfaz RStudio. Para descargarla diríjase a la página web oficial <https://www.rstudio.com/>. Dentro de la pestaña **Products** clickee sobre **RStudio:**

**Interfaz de usuario gráfica, Texto, Correo electrónico, Sitio web

Descripción generada automáticamente**

Baje por la página hasta encontrar el botón **DOWNLOAD RSTUDIO DESKTOP** y clickeelo:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Baje por la página hasta encontrar el botón **DOWNLOAD** debajo de **RStudio Desktop Open Source License Free** y clickeelo. Asegúrese de clickear el **DOWNLOAD** de la izquierda (NO el de la derecha), tal como se muestra en la imagen siguiente:

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza baja

Y clickee sobre **DOWNLOAD RSTUDIO FOR WINDOWS**. En este caso, lo mismo que para R; si usted no tiene Windows debe seleccionar su instalador correspondiente (que se encuentran un poco más abajo). También pueden haber diferencias en las versiones que se descargan pero las diferencias son mínimas y no nos generarán problemas.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Se descargará el archivo **RStudio-2022.02.0-443.exe** en su carpeta de descargas.

## Paso 4: Instalar RStudio

Para poder instalar RStudio es necesario tener previamente instalado R (**Pasos 1** y **2**). Si usted no pudo completar los pasos 1 y 2, no continúe con el **Paso 4**. Guarde una captura de pantalla del error que obtuvo y solucionaremos el error en la primera clase.

Si pudo completar con los pasos 1 a 3, ubique el archivo **RStudio-2022.02.0-443.exe** y haga doble click izquierdo sobre él. Los pasos a seguir son similares al **Paso 2** que se realizó anteriormente. Haga click en **Sí**, **Siguiente**, **Aceptar** o **Finalizar** según la opción que aparezca para completar la instalación.

Al haber completado el Módulo 0, usted tiene instalado en su computadora a R y RStudio y estos dos están intercomunicados entre sí y listos para ser utilizados.

# Módulo 1: R y RStudio

En este módulo se introducen los conceptos básicos de R y de RStudio. Se introducirá un poco de terminología específica relacionada a la programación en general y a estos programas en particular. Entendiendo que el público al que está dirigido tiene escaso conocimiento de esta temática y que tampoco busca ser experto, los conceptos que se introducirán no presentarán una profundidad ni complejidad excesiva. De este módulo se espera que el alumno entienda qué es un lenguaje de programación, qué es R, qué es RStudio, cómo se usa y que ventajas tiene R por sobre Microsoft Excel.

## A. R como lenguaje de programación para datos

Los ordenadores permiten ensamblar, manipular y visualizar conjuntos de dato a velocidades asombrosas. Pero, para lograr hacerlo, es necesario que el usuario adquiera ciertos conocimientos básicos sobre la programación. R es un lenguaje de programación orientado a objetos, esto quiere decir que se pueden guardar objetos virtuales (como algún número, serie de números, texto, tablas, imágenes, etc…) bajo algún nombre para consultarlos, modificarlos y borrarlos a pedido. Existen muchos lenguajes orientado a objetos, todos ellos se diferencian en la sintaxis de los comandos, en los procesos que ocurren “detrás de escena”, en las capacidades, limitaciones y el “para qué” fueron creados. Algunos fueron creados para construir páginas webs, otros para crear software, otros para mantener sistemas interconectados . A su vez, R es de los pocos lenguajes creados para manipular datos, lo que lo vuelve extremadamente potente en este ámbito.

A continuación se presentan algunas definiciones básicas:

**Comando**: Línea de código que realiza una acción específica.

**Algoritmo**: Concepto abstracto que describe cómo solucionar un problema. Los algoritmos no son solo informáticos, se encuentran en nuestra vida cotidiana, aquí un ejemplo de algoritmo aplicado a la cocina:

1. Cortar la calabaza en trozos y salpimentar.
2. Hornear la calabaza a 180 °C hasta que esté blanda.
3. Caramelizar las cebollas a fuego lento con la mantequilla.
4. Freír los ajos previamente troceados.
5. Licuar la calabaza con el ajo y la cebolla. Añadir un poco de agua si es necesario, para que quede la consistencia de crema.
6. Decorar con las almendras laminadas.

Cada línea realiza una acción particular y en su conjunto, el algoritmo produce una sopa de calabaza.

**Función:** Materialización del algoritmo. Son las líneas de código necesarias para resolver el problema. En el ejemplo anterior, el conjunto y orden de las líneas de la receta hacen el código de la función. Estas, además, tienen parámetros de entrada y un resultado. Un parámetro de entrada para la receta de calabaza podría ser la cantidad de comensales y el resultado es la sopa de calabaza, cuyo volumen fue definido por la cantidad de comensales.

**Paquete**: Librería de funciones útiles. Son funciones redactadas por otros usuarios, agrupadas en un paquete y compartidas de forma gratuita para ser utilizadas por la comunidad. Cada paquete suele enfocarse en algún aspecto en específico. Por ejemplo, el paquete *ggplot2* tiene funciones relacionadas a la visualización de los datos mientras que el paquete *stringr* sirve para la manipulación de textos.

**Script**: Archivo de texto donde se desarrollan y se guardan los algoritmos.

**Objeto**: Información guardada bajo un nombre específico.

**Ambiente**: Espacio virtual donde se guardan los objetos mientras se usa el programa.

**Consola**: Ventana donde se ejecutan los comandos y se devuelven los resultados.

Para visualizar la consola de R puede buscarla en su buscador de archivos como **RGui** y clickear sobre el símbolo Imagen que contiene dibujo

Descripción generada automáticamente. Debería abrirse una ventana similar a la siguiente:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Hemos aquí R en todo su esplendor. En su forma más básica, R se visualiza como una consola (parte blanca de la imagen). En ella se pueden ejecutar todos los comandos necesarios para resolver cualquier problema con datos. Sin embargo, para facilitar la interacción con este lenguaje, existe un programa llamado RStudio que funciona de interfaz gráfica.

## B. RStudio como interfaz facilitadora para programar en R

Busque RStudio en su buscador de archivos y ábralo. Su símbolo distintivo es Icono

Descripción generada automáticamente. Se abrirá una ventana similar a la siguiente:

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Podrán observar 4 sub-ventanas dentro de la ventada de RStudio. Cada una de ellas tiene funciones distintas:

1. Scripts: Donde el usuario escribe el código de su programa. No es mas que un lector de archivos de texto con algunas modificaciones estéticas y funcionales: El texto tendrá colores distintos según si se refiere a funciones, a números, a texto, a funciones lógicas, etc. Una vez escrita una línea, esta se ejecuta clickeando sobre **Run** arriba a la derecha o bien presionando **Crtl + Enter**.
2. Environment (Ambiente) + otras: Se muestran los objetos guardados en el ambiente a los que se puede acceder clickeandolos. En la imagen, el ambiente está vacío. La solapa de ambiente será la mas utilizada de esta sub-ventana. Las demás
3. Consola: Es la misma consola que la presentada en la parte A. En ella se realizarán todas las ejecuciones de comandos y se muestran los resultados. Cuando uno corre una línea de código en el script, esta línea es ejecutada en la consola automáticamente.
4. Files (archivos), Plots (Imágenes), Packages (Paquetes), Help (Ayuda), etc:
   1. Archivos: Los archivos de la carpeta donde uno se encuentra ubicado. Puede abrir scripts, tablas y otros archivos desde aquí.
   2. Imágenes: Cuando se realicen gráficos, estos aparecerán acá.
   3. Paquetes: Lista de paquetes instalados.
   4. Ayuda: Se encuentra la documentación de las funciones de los paquetes.

## C. R vs. Excel

Si usted está realizando este curso, es porque pudo observar de primera mano algunas limitaciones que tiene Excel y siente curiosidad sobre cómo puede ir más allá y R es una opción para hacerlo. Frente a problemas grandes, complejos y repetitivos, R presenta muchas ventajas por sobre Excel. Algunas de ellas son:

* Programabilidad y automatización: La implementación de acciones sobre tablas de datos se realiza de forma escrita por lo que las acciones quedan guardadas, su secuencia queda explicita en el script y puede ser repetida todas las veces necesarias. La mayor parte del trabajo se encuentra en la escritura del primer script, luego solo basta aplicarlo y esta segunda parte suele tomar literalmente menos de 1 segundo de forma casi automática. Esto es extremadamente útil para procesos repetitivos y periódicos.
* Velocidad: La velocidad de cálculo es muy superior. Esto queda en evidencia especialmente en problemas grandes.
* Escalabilidad: La cantidad de líneas y columnas de una tabla empieza a ser menos importante. Para el usuario prácticamente no hay diferencias entre el manejo de una tabla de 100 filas frente a una de 1.000.000 de filas.
* Exactitud de cálculos: Se ha demostrado que Excel puede cometer errores en cálculos no tan complejos. Cosa que no ocurre en R.
* Encontrar errores cometidos: La programabilidad también permite buscar, encontrar y corregir errores, dado que cada paso se encuentra escrito en el script.
* Gratis y de código abierto: No es necesario pagar una licencia para usar R y todo dentro de él es modificable. R puede incorporar nuevas funciones y la cantidad disponible es enorme ya que toda la población tiene acceso a compartir sus funciones.
* Complejidad: R puede realizar operaciones mucho más complejas que las provistas en Excel.
* Gráficos destacables: El diseño de los gráficos es altamente modificable. Puede observar algunos ejemplos en <https://www.r-graph-gallery.com/>.
* Lectura de distintos formatos: Puede leer información de cualquier formato, incluyendo el formato Excel.

Sin embargo, para problemas pequeños, Excel tiene ventajas por sobre R:

* Intuitividad: Excel es mucho más intuitivo que R lo que facilita su aprendizaje en especial para aquellos que están iniciando en el mundo de las tablas de datos.
* Registro y modificación puntal de datos: Es muy sencillo construir una tabla de cero y modificar puntualmente algún valor cuando la cantidad de valores a modificar es pequeña.
* Cálculos básicos: Es más rápido para el usuario usar Excel cuando necesita realizar algún cálculo sencillo.
* Visualización de las tablas: La interfaz de Excel muestra todo el tiempo la/s tabla/s con la que el usuario esta trabajando lo que permite echar un vistazo rápido y de forma sencilla la tabla con la que se está trabajando.
* Reconocimiento: La difusión de Excel es muy extensa y es difícil encontrar gente que no lo haya utilizado alguna vez, por lo que la mayoría de las personas sabe cómo leer información y realizar operaciones básicas desde un archivo Excel.

En conclusión, R y Excel tienen ventajas uno por sobre el otro. Excel es el programa predilecto para iniciarse en el manejo de datos y para el abordaje de problemas chicos y puntuales. Por otro lado, presenta limitaciones que tarde o temprano hacen necesario pasar a usar procesadores de datos mucho más potentes como lo es R. A la (no tan) larga, descubrirán que usar R es una forma mucho más eficiente que Excel y podrán reducir enormemente mucho esfuerzo en la realización de sus trabajos.