## Universidad Nacional de Córdoba

Facultad de Ciencias Económicas

Escuela de Graduados

Doctorado en Demografía

## Propuesta de curso-taller optativo:

# Uso de R para el Análisis de Datos Demográficos

## Fundamentación

El doctorado en Demografía de la UNC tiene una carga importante de análisis de datos provenientes de censos, encuestas y registros. Algunas de las técnicas usadas se comparten con otras disciplinas: Sociología, Economía, Psicología; otras son propias de la Demografía, las que tratan con las variables específicamente demográficas. Para realizar las operaciones de análisis de esos datos, existe una variedad de programas que se ocupan de los procesos cálculo, de manera tal que el investigador sólo deba decidir qué procedimiento aplicar para un problema dado (en base al tipo de dato y los supuestos de cada técnica), y realizar una lectura e interpretación correcta de las “salidas” del software. Aun cuando se supone que el analista conoce las operaciones que conducen al resultado, su cómputo se confía al software. Los llamados “paquetes estadísticos” ofrecen, en un entorno único, las operaciones de aplicación más frecuente y tienen la ventaja de no requerir conocimientos estadísticos detallados. Pero esto se da con el riesgo de rutinizar procedimientos y lecturas y perder flexibilidad, tanto en los procedimientos que pueden aplicarse, como en las interpretaciones de resultados.

En su mayoría, estos paquetes usan la mecánica de “point & click”; seleccionar los comandos y aplicarlos al subconjunto de datos elegido. Esta práctica, si bien es de fácil aprendizaje y aplicación, atenta contra la reproducibilidad de los resultados, ya que no deja registro de las operaciones realizadas, por lo que impide seguir la secuencia de transformaciones que llevaron a un resultado, tanto para observadores externos como para el mismo investigador. Esto es crítico en todas las etapas del análisis, y se nota especialmente en las operaciones de limpieza de una base de datos, en la que se toman decisiones sobre casos perdidos, categorización, recodificación de variables, etc., a las que luego es difícil recordar, a menos que se haya tomado nota precisa.

Por el contrario, escribir las instrucciones de cada procedimiento y disponer luego de los comandos escritos, permite la trazabilidad del recorrido que siguió el investigador. Cuando un análisis se acompaña de la sintaxis que da cuenta de las operaciones realizadas, se gana en transparencia: otros investigadores pueden correr esa sintaxis sobre los mismos datos y alcanzar iguales resultados o replicarla sobre otros datos para comparar resultados. Más aun, cuando se analizan bases que proveen información de manera regular, como las encuestas de hogares o los registros, una vez que se decidió un conjunto de operaciones analíticas, es suficiente correr los mismos comandos sobre la nueva base.

Aunque algunos paquetes, como SPSS, dispongan de la opción de trabajar sobre la sintaxis, no es esa la práctica habitual. Además hay otras razones para elegir R:

* Es software libre, de código abierto.
* Dispone un repositorio de rutinas de análisis llamadas “paquetes” de aplicación en distintos campos disciplinares (análisis de datos de encuestas, de pruebas psicométricas, demográficos, series de tiempo, inferencia bayesiana, para graficar, de formatos, etc.). Actualmente el repositorio (la CRAN) tiene poco más de 15000 paquetes.
* Cuenta con una comunidad de usuarios muy amplia que ofrece ayuda a quienes se inician.
* Está altamente actualizado, porque la mayoría de las nuevas técnicas estadísticas se programan en R y se suben al repositorio (como paquetes) antes que los software comerciales las incorporen a sus actualizaciones.
* Gradualmente se va integrando en programas comerciales.

Por su uso cada vez más difundido y por las ventajas que se mencionaron, la incorporación en el Doctorado en Demografía de la UNC de un taller sobre uso de este software constituye un aporte valioso a la formación de los futuros doctores.

## Objetivo

Introducir al uso de R para el análisis de datos demográficos.

Motivar a los asistentes para que exploren la gran cantidad de paquetes específicos, según sus intereses de investigación.

## Contenidos

1. El Entorno de Desarrollo Integrado (IDE) *RStudio*. Operaciones elementales del paquete *base*. Los objetos R.
2. Lectura desde diferentes formatos: paquetes *foreign* y *readxl*.
3. Exploración y limpieza de una base. Recodificación de variables
4. Análisis descriptivo: tablas, medidas resumen.
5. Visualización: el paquete *ggplot2*.
6. Datos demográficos: el paquete *demography*.
7. Confección de informes: *RMarkdown*.

## Modalidad de cursado

Se trata de un seminario taller, en el que los participantes trabajarán sobre sus propias computadoras. En cada encuentro se presentarán los contenidos y se aplicarán los procedimientos sobre los mismos conjuntos de datos. Luego, los cursantes reproducirán los análisis con otros datos, que pueden ser propios o provistos en el curso.

## Evaluación

El curso se aprueba con un documento de análisis de datos, presentado en formato RMarkdown que incluya: objetivos, fuente(s) de datos, metodología de análisis, conclusiones.

## Referencias

Allaire J., Xie Y., McPherson J., Luraschi J., Ushey K., Atkins A., Wickham H., Cheng J., Chang W. & Iannone R. (2019). rmarkdown: Dynamic Documents for R. R package version 1.16, https://github.com/rstudio/rmarkdown.

R Core Team (2017). foreign: Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', .... R package version 0.8-70. <https://CRAN.R-project.org/package=foreign>

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Hyndman, R., Booth, H., Tickle, L. & Maindonald, J. (2019). demography: Forecasting Mortality, Fertility, Migration and Population Data. R package version 1.22. <https://CRAN.R-project.org/package=demography>

RStudio Team (2018). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL <http://www.rstudio.com/>.

Wickham, H. & Bryan, J. (2018). readxl: Read Excel Files. R package version 1.1.0. https://CRAN.R-project.org/package=readxl

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Xie Y., Allaire J., Grolemund G. (2018). R Markdown: The Definitive Guide. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 9781138359338, <https://bookdown.org/yihui/rmarkdown>.