ESTADÍSTICA DESCRIPTIVA, PROBABILIDAD E INFERENCIA CON R

Juan Antonio Espinosa Pulido, *I.E.S. Santa Catalina de Alejandría, Jaén, depmatiessantacatalina@gmail.com*
David Molina Muñoz, *Universidad de Granada, dmolinam@ugr.es*

RESUMEN.

En muchos campos de las matemáticas, y particularmente en la estadística, las situaciones prácticas involucran grandes volúmenes de datos, los cuales únicamente pueden ser analizados con la ayuda de programas informáticos. Dentro de la extensa variedad de software estadístico que existe destaca R, un programa de carácter gratuito que ha adquirido mucha popularidad en los últimos años.

En este taller introduciremos el programa R y haremos un repaso de las funciones básicas de que dispone para la resolución de problemas de estadística descriptiva, probabilidad e inferencia.

Nivel educativo: Bachillerato, Ciclo Formativo de Grado Superior, Universidad.

1. INTRODUCCIÓN.

La necesidad actual de profesionales capaces de enfrentarse a los problemas reales que se les presenten hace aconsejable complementar la formación teórica de los estudiantes con clases prácticas desde etapas muy tempranas de la educación.

En matemáticas, la complejidad de este tipo de problemas hace que, habitualmente, su resolución implique el manejo de grandes cantidades de información. El desarrollo que la tecnología ha experimentado en los últimos años ha simplificado mucho esta tarea, al proporcionar un gran número de programas informáticos para el tratamiento y el análisis de datos. Entre ellos se encuentra R, el cual se ha convertido en uno de los más utilizados por la comunidad matemática.

2. R: UN PROGRAMA PARA EL ANÁLISIS DE DATOS.

R es un entorno computacional para el tratamiento de datos y la generación de gráficos desarrollado inicialmente en 1993 por R. Gentleman y R. Ihaka del departamento de Estadística de la Universidad de Auckland. El término R también hace referencia al lenguaje de programación propio que dicho entorno incorpora.

Una de las principales características de R, que constituye su mayor ventaja, es su gratuidad. De hecho, R surgió como una alternativa gratuita al lenguaje de programación comercial S y, desde 1995, se incluye dentro del proyecto GNU de software libre.

Las implicaciones que el carácter gratuito de R conlleva son tremendamente importantes, ya que cualquier persona puede ejecutar, copiar, distribuir, modificar y mejorar el programa. Normalmente, las mejoras y nuevas funcionalidades que los usuarios proponen vienen en forma de paquete. En R, un paquete (también llamado librería) no es más que un conjunto de funciones con una temática común que complementa la instalación básica del programa. A día de hoy, se dispone de más de 5300 paquetes que hacen de R uno de los programas más completos dentro del panorama matemático.

2.1. ESTADÍSTICA DESCRIPTIVA CON R.

Como su propio nombre indica, la estadística descriptiva se encarga de la descripción y el resumen de conjuntos de datos. Son muchas las funciones que existen en R para el análisis descriptivo de una variable. Así, por ejemplo, la función *table* permite crear tablas de frecuencia para agrupar los datos, que pueden ser acumuladas si adicionalmente se emplea la orden *cumsum*. Dada una tabla de frecuencia, es posible crear distintos tipos de diagramas, como el de barras (mediante la función *barplot*) o el de sectores (usando la función *pie*).

Por otro lado, las funciones *max*, *min*, *mean*, *median*, *quantile*, *var* y *sd* calculan, respectivamente, el máximo, el mínimo, la media, la mediana, los cuantiles, la cuasi-varianza y la cuasi-desviación típica de una variable. A partir de estas medidas se pueden calcular otras como el rango, el rango intercuartílico o el coeficiente de variación.

Finalmente, R dispone de varias funciones que proporcionan un resumen con los principales indicadores descriptivos de una variable. Entre ellas destacan *fivenum*, que calcula el mínimo, el primer cuartil, la mediana, el tercer cuartil y el máximo de la variable, y *summary*, que además de las 5 medidas anteriores incluye la media entre sus resultados.

2.2. PROBABILIDAD CON R.

En múltiples ocasiones interesa obtener determinados valores de una distribución de probabilidad, habitualmente relacionados con su función de densidad o de distribución o con sus cuantiles. Con R es posible llevar a cabo este cometido para las distribuciones de probabilidad más relevantes, como la binomial, la de Poisson, la normal, la chi cuadrado, la t de Student o la F de Snedecor. Así, por ejemplo, para la distribución binomial las funciones *dbinom*, *pbinom* y *qbinom* calculan, respectivamente, valores de la función de densidad, de la función de distribución y de los cuantiles.

Otra tarea muy común que podemos realizar con R cuando se trabaja con distribuciones de probabilidad es la generación de números aleatorios, la cual en el caso de la distribución binomial se realiza mediante la función *rbinom*.

2.3. INFERENCIA CON R.

Uno de los principales objetivos de la inferencia estadística es estimar un determinado parámetro desconocido de la población bajo estudio. Esta estimación puede ser de carácter puntual, si únicamente se proporciona un valor para dicho parámetro, o confidencial, si lo que se calcula es un rango de valores entre los que figura el valor real del parámetro dado un cierto nivel de confianza. Otras veces lo que se pretende es comprobar alguna hipótesis inicial sobre la

población objeto de estudio a través de la información extraída de una muestra. Para ello, se emplean los contrastes de hipótesis.

Frecuentemente, se asume que las poblaciones de la que se extraen las muestras de observaciones que se emplean para la estimación o la resolución de contrastes de hipótesis están gobernadas por distribuciones normales. En estos casos, se emplean las técnicas de la rama de la inferencia conocida como inferencia paramétrica. Los parámetros sobre los que habitualmente se plantean los contrastes de hipótesis suelen ser la media o la varianza de la población, en el caso de que contar solo con una, o la diferencia de medias o el cociente de varianzas, cuando se dispone de dos poblaciones. Dependiendo del caso, el estadístico de prueba para la resolución del contraste sigue una distribución de probabilidad determinada.

La resolución de problemas de inferencia con R puede llevarse a cabo de forma sencilla calculando el estimador que corresponda (en el caso de problemas de estimación) o mediante la obtención de un estadístico de prueba y un valor crítico (si estamos ante un contraste de hipótesis). Para algunas situaciones concretas, R dispone de funciones que proporcionan la solución al problema de una forma más directa. Así, por ejemplo, la función *t.test* se emplea para resolver contrastes sobre la media de una población cuando su varianza se supone desconocida. Esta misma función puede utilizarse cuando se consideran dos poblaciones o grupos y se plantea un contraste sobre la diferencia de sus medias en el caso de varianzas desconocidas, pero iguales. La función *var.test,* por su parte, proporciona la solución a contrastes de hipótesis sobre el cociente de dos varianzas.

3. ORGANIZACIÓN DEL TALLER.

El taller se estructura en 4 partes. La primera de ellas tendrá un carácter introductorio y una duración aproximada de 15 minutos. En ella se darán indicaciones para la instalación del programa e instrucciones básicas para la carga y manipulación de datos. Cada una de las 3 secciones restantes, de unos 35 minutos de duración, estará dedicada a tratar cada uno de los apartados 2.1, 2.2 y 2.3.

Para maximizar la utilidad del taller, se pretende que los asistentes prueben *in situ* todas las funciones que se expongan y las utilicen para resolver pequeños problemas que se propondrán a lo largo del mismo. Para ello, sería necesario disponer de un puesto informático por cada asistente, a ser posible con la última versión del programa instalada. Igualmente, convendría disponer de un retroproyector para la proyección de la presentación de diapositivas que guiará el curso del taller.

REFERENCIAS.

PARADIS, E. (2002). *R para principiantes,* disponible en http://cran.r-project.org/.

R DEVELOPMENT CORE TEAM. (2000). *Introducción a R*, disponible en http://cran.r-project.org/.