TAREA SESIÓN 1: Introducción a la Inferencia Estadística, Muestreo y Preproceso de datos

# **Descripción de la práctica y objetivos**

Esta tarea forma parte del trabajo no presencial del alumno relativo a la primera sesión de la asignatura Herramientas Estadsiticas para Big Data del Diploma de Especialización.

Mediante ésta se pretende que el alumno pueda reforzar los contenidos vistos durante las horas presenciales de forma práctica.

Adicionalmente, la realización de un informe con los resultados, se utilizará para evaluar esta parte de la asignatura.

Para realizar la práctica podéis ayudaros de [**Quick R**](http://www.statmethods.net/), de la **ayuda del RStudio** y de la **bibliografía sugerida** (o de cualquier otra fuente que consideréis necesaria).

Los ***data sets* usados** a lo largo de este trabajo práctico los tenéis disponibles en **PoliformaT**.

**El alumno deberá entregar a través de PoliformaT un proyecto con:**

1. **Fichero con el script** de todo el código de R escrito para la práctica denominado **Tarea HE Big Data-1.R**
2. **Ficheros de datos** de R creados en la práctica:

* **Andalucia.RData**
* **JaenIndicadores.RData**

1. **Documento informe** con las respuestas a las preguntas que se hagan en esta práctica, acompañadas de las salidas que sean necesarias.

**NOTA: si generas un informe con *Markdown*, el código y los resultados estarán todos juntos.**

# **Contenido**

**1. CÁLCULO DE NUEVAS VARIABLES, RECODIFICACIÓN Y FILTRADO**

**2. ANÁLISIS ESTADÍSTICO DESCRIPTIVO DE DATOS**

**3. DISTRIBUCIONES DE PROBABILIDAD**

**4. CONTRASTES DE HIPÓTESIS E INTERVALOS DE CONFIANZA**

# **Cálculo de nuevas variables, recodificación y filtrado**

# Descripción del dataset

El fichero JaenIndicadores.txt[[1]](#footnote-1) contiene datos sobre indicadores importantes de los municipios de la provincia de Jaén en el año 2001, e incluye las siguientes variables[[2]](#footnote-2):

* Código INE del municipio.
* Nombre del municipio.
* Consumo de energía eléctrica en megavatios por hora.
* Consumo medio de agua en invierno, en metros cúbicos por día.
* Consumo medio de agua en verano, en metros cúbicos por día.
* Destino de los residuos sólidos urbanos: las posibilidades son vertedero controlado, vertedero incontrolado, compostaje.
* Cantidad de residuos sólidos urbanos, en toneladas.

# Ejercicios

1. Importar el fichero **JaenIndicadores.txt** y denominar a la hoja de datos (*data frame*) ***Datos.Jaen[[3]](#footnote-3)***
2. Recodificar la variable *Poblacion* en una variable cualitativa tipo factor llamada *Tamaño* con tres categorías:

* Si la población es inferior a 2000 habitantes, Tamaño será “Pequeño”.
* Si la población está entre 2000 y 4500 habitantes, Tamaño será “Mediano”.
* Si la población es superior a 4500 habitantes, Tamaño será “Grande”.

1. Calcular los siguientes promedios que se especifican a continuación y añadirlos como nuevas variables al fichero *Datos.Jaen* obtenidas a partir de las variables existentes:

* Variable ***elec.hab*** que contendrá el consumo de energía eléctrica por habitante, obtenida como *Consumo.de.energia.electrica/Poblacion*
* Variable ***agua.hab***que contendrá el consumo medio de agua por habitante y día, obtenida como *(Consumo.de.agua..Invierno + Consumo.de.agua..V erano)/Poblacion*
* Variable ***res.hab*** que contendrá los residuos sólidos urbanos por habitante, obtenida como *Residuos.solidos.urbanos..Cantidad/Poblacion*

1. Crear una nueva hoja de datos con todas las variables que contiene actualmente el *data frame Datos.Jaen*, pero referida sólo a los municipios de tamaño mediano y denominarla ***Datos.Jaen.Medianos***
2. Guardar la hoja de datos *Datos.Jaen* con las nuevas variables creadas en los apartados anteriores y la hoja que contiene los datos de las poblaciones medianas (*Datos.Jaen.Medianos*) en un archivo de datos de R y llamadlo **JaenIndicadores.RData**

# **Análisis Estadístico Descriptivo de Datos**

# Descripción del dataset

El fichero **Andalucia.txt**1 contiene datos sobre diversos indicadores de los municipios andaluces obtenidos del [Instituto de Estadística y Cartografía de Andalucía](http://www.juntadeandalucia.es/institutodeestadisticaycartografia/temas/index-est.htm). Los valores de estos indicadores están contenidos en las siguientes variables[[4]](#footnote-4):

* Código INE.
* Municipio.
* Tasa de actividad en 2001.
* Nº de líneas ADSL en funcionamiento en 2007.
* Edad media del municipio en 2007.
* Renta familiar disponible por habitante. Aparece agrupado en varias categorías de renta. Hay numerosos datos faltantes que aparecen señalados como “..”.
* Crecimiento vegetativo en 2006. El crecimiento vegetativo es la diferencia entre el número de nacidos y el número de fallecidos.
* Número de parados en 2007.

Población total del municipio en 2001, 2003, 2006 y 2007.

# Ejercicios

1. Importar el fichero **Andalucia.txt** y denominar a la hoja de datos (*data frame*) ***Datos.Andalucia***. Comprobar si en el archivo .txt hay datos faltantes y cómo están codificados.
2. A partir de la variable código INE, construir una variable tipo factor que distinga la provincia de pertenencia de cada municipio, denominarla “Provincia” y añadirla al *data frame*.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Provincia** | Almería | Cádiz | Córdoba | Granada | Huelva | Jaén | Málaga | Sevilla |
| **Id[[5]](#footnote-5) cod INE** | 4 | 11 | 14 | 18 | 21 | 23 | 29 | 41 |

Obtener la distribución de frecuencias absolutas y relativas, un diagrama de barras con las frecuencias absolutas y un diagrama de sectores con las frecuencias relativas en porcentajes de esta variable tipo factor. ¿Qué provincia tiene más municipios? ¿Cuál tiene menos? ¿Qué porcentaje representa en cada caso?

1. Obtener un resumen descriptivo de la variable tasa de actividad de 2001 que incluya parámetros de posición, dispersión, asimetría y curtosis, histograma y diagrama de caja. En función de este resumen, contestar a las siguientes preguntas:
   1. ¿Cuál es la tasa media de actividad de los municipios andaluces? ¿Crees que este valor es adecuado para representar la Tasa de Actividad de los municipios andaluces durante 2001?
   2. ¿Cómo valoras la homogeneidad de los valores de la tasa de actividad en los municipios andaluces? ¿Qué parámetro elegirías para representar la dispersión de la Tasa de Actividad de 2001?
   3. ¿En ese sentido, qué municipios andaluces destacan significativamente del resto (como atípicos) por su alta tasa de actividad y por su baja tasa de actividad? ¿Se te ocurre alguna explicación al respecto?
   4. ¿Cómo valoras la simetría de la distribución de frecuencias?
2. Obtener un gráfico de caja de la Tasa de actividad en 2001 en función de la provincia y describe brevemente la información que contienen los datos a partir del gráfico.
3. Guardar la hoja de datos *Datos.Andalucia* con la nueva variable creada en los apartados anteriores junto con los parámetros obtenidos en un archivo de datos de R y llámalo **Andalucia.RData**.

# **Distribuciones de probabilidad**

1. Consideremos una variable aleatoria que sigue una distribución B (15, 0.33). Se pide:
   1. ¿Qué valor de la variable deja por debajo de sí el 75% de la probabilidad?
   2. Calcular el percentil 95% de la distribución.
   3. Obtener una muestra de tamaño 1000 de esta distribución, representarla gráficamente las frecuencias observadas de cada valor de la distribución mediante un diagrama de barras y comparar éste con las frecuencias esperadas según el modelo que genera los datos.
2. Consideremos una variable aleatoria W con distribución N (250, 13). Se pide:
   1. P [240 < W ≤ 245.5]
   2. P [W ≥ 256].
   3. Si queremos desechar el 5% de valores más altos de la distribución y el 5% de valores más bajos, ¿con qué intervalo de valores nos quedaremos?
   4. Obtener una muestra de tamaño 1000 de la distribución, representar la función de densidad de esta distribución y compararla con el histograma de la muestra obtenida.

# **Contrastes de Hipótesis e Intervalos de Confianza**

# Descripción del dataset

Mediante una red de sensores se han recogido datos sobre la temperatura media diaria (ºC) en dos estaciones A y B durante 52 días. Los valores recogidos de la temperatura se encuentran en la hoja de datos “Temper” incluida en el fichero ***Temperatura.RData***.

# Ejercicios

1. Cargar el fichero **Temperatura.RData**.
2. Crear dos nuevas variables, temp.A y temp.B, que contengan las temperaturas de las estaciones A y B, respectivamente.
3. Da un intervalo de confianza para la temperatura media diaria de la estación A, al 95%, y a partir de éste indica si se puede admitir, y por qué, que la temperatura media diaria en dicha estación sea de 19ºC, con ese mismo nivel de confianza.
4. Plantea un test de hipótesis que refleje la pregunta del apartado anterior y resuélvelo sin usar el intervalo de confianza (riesgo de 1ª especie 5%)
5. Determina si puede admitirse, con un riesgo de primera especie de 1%, que la temperatura media diaria es la misma en las dos estaciones. Plantea previamente el correspondiente contraste de hipótesis.
6. Obtén un intervalo de confianza (99%) para la diferencia de temperaturas entre estaciones. ¿Aporta alguna información adicional al resultado obtenido en el apartado anterior?
7. Se sabe que a lo largo de los 52 días, la estación A falló 5 días y la B 7 días. ¿Puede afirmarse con un nivel de confianza del 90% que la proporción de días fallados es la misma en las dos estaciones?

1. Sáez Castillo, A.J., 2010. *Métodos Estadísticos con R y R Commander*, Jaén: Universidad de Jaén. [↑](#footnote-ref-1)
2. Observad que en el archivo .txt faltan algunos datos de Linares y Jaén, y en su lugar aparece “-”. [↑](#footnote-ref-2)
3. Recuerda que para facilitar el la referencia a las variables durante su utilización te puede ser útil establecer cómo activa la hoja de datos Datos.Jaen (*attach()*). [↑](#footnote-ref-3)
4. [↑](#footnote-ref-4)
5. El id representa los 2 primeros dígitos del código INE [↑](#footnote-ref-5)