

# Fundamentos de Análisis de Datos

## Trabajo Práctico 1

Profesora: Dra. Andrea Alejandra Rey

**Ejercicio 1.** En el archivo `Dieta.xlsx` se encuentran los datos correspondientes a 173 personas que están siguiendo una dieta. Para cada una de ellas, se registró el sexo y el consumo de grasas saturadas y de alcohol, así como del total de calorías diarias.

1. Analizar si existen datos faltantes y, en caso afirmativo, eliminar tales registros.
2. Calcular las siguientes medidas estadísticas descriptivas clásicas del consumo de grasas: rango, media, mediana, desvío estándar y rango intercuartil.
3. Realizar gráficos boxplots de los datos sobre el consumo de calorías en función de la variable categórica. ¿Qué puede observarse?
4. Dividir la cantidad de calorías consumidas en dos categorías: MODERADA (menor o igual a 1700) o ALTA (mayor a 1700). Analizar el consumo de alcohol de acuerdo a la cantidad de calorías consumidas según las categorías definidas.

**Ejercicio 2.** El archivo `Sociodemograficos.xlsx` contiene datos sobre distintos indicadores socio-demográficos de varios países.

1. ¿Cuáles son las variables de interés? ¿Cuántos países fueron analizados?
2. ¿Cuáles son los países con menor y mayor tasa de natalidad?
3. Realizar un diagrama de dispersión con las tasas de natalidad y de mortalidad infantil. ¿Qué puede observarse? Justificar lo observado a partir del gráfico con una medida cuantitativa.
4. Calcular el vector de medias y medianas.
5. Calcular las matrices de covarianzas y de correlaciones. A partir de estas matrices dar un ejemplo de dos variables fuertemente correlacionadas positivamente, de dos variables fuertemente correlacionadas negativamente y de dos variables no correlacionadas.

**Ejercicio 3.** Vamos a considerar el conjunto de datos `swiss` disponible en R.

1. Cargar la base de datos y explorarla. ¿Cuántos registros y cuántas variables tiene? Describir las variables de estudio.

2. Se desea comparar las provincias entre sí. ¿Es adecuado utilizar la distancia Euclídea para realizar la comparación? Justificar la respuesta.
3. Buscar la presencia de datos atípicos mediante la distancia de Mahalanobis. Comentar los resultados obtenidos.

**Ejercicio 4.** El Departamento de Psicología de una universidad ubicada en una ciudad céntrica realizó un estudio sobre la asistencia a clases teóricas no obligatorias dependiendo de la localidad de residencia del estudiantado. Para tal fin, se seleccionaron 40 estudiantes en la Ciudad A, 40 estudiantes en la Ciudad B y 40 estudiantes en la Ciudad C, y se contabilizó la cantidad de clases a las que cada uno/a asistió. Los resultados obtenidos se muestran en la siguiente tabla.

Ciudad A	Ciudad B	Ciudad C
11	13	6
14	10	7
7	12	3
15	7	5
11	5	9
13	10	6
11	10	1
16	16	6
10	9	0
15	7	2
18	7	5
12	2	6
9	6	11
9	9	6
10	9	7
10	8	0
15	8	5
10	10	7
14	3	5
10	6	4
10	5	7
12	2	4
14	9	2
12	3	8
15	4	9
7	5	6
13	10	1
6	8	4
10	5	7
15	9	7
20	10	8
10	8	9
13	13	7
10	10	5
6	0	1
14	2	6
8	1	9
10	1	4
8	0	7
11	4	16

1. Armar un **data frame** en **R** con los datos de la tabla anterior, creando dos variables: una que represente la cantidad de asistencias a las clases teóricas no obligatorias y otra que represente la localidad de residencia. ¿Qué tipo de variable es cada una?
2. Analizar los datos de la muestra mediante gráficos y medidas estadísticas descriptivas. ¿Se observan diferencias en los valores promedios por localidad?
3. Realizar un test ANOVA para comparar las medias de las 3 poblaciones. Plantear las hipótesis nula y alternativa del test, informar los resultados obtenidos y la decisión tomada.
4. Si se han obtenido diferencias significativas entre las localidades, determinar cuáles son esas diferencias utilizando el test de Tukey.