

EJERCICIOS PRÁCTICOS MÓDULO 3:

- Estadística descriptiva

El conjunto de datos utilizado procede del portal del Departamento de Bioestadística de la Universidad de Vanderbilt y puede descargarse desde la dirección: <https://hbiostat.org/data/repo/diabetes.xls> (existen varios formatos) . Este archivo fue elaborado por el Dr. John Schorling, perteneciente al Departamento de Medicina de la Universidad de Virginia.

La base contiene información de 403 participantes y recoge 19 variables relacionadas con la salud de personas afroamericanas residentes en los condados de Buckingham y Louisa, en el estado de Virginia (EE. UU.). Estos individuos fueron evaluados con el objetivo de detectar la presencia de diabetes mellitus. Los registros forman parte de un estudio más amplio que incluyó a 1046 personas y que, además de la diabetes, analizó la prevalencia de la obesidad y otros factores de riesgo cardiovascular en esta comunidad durante el año 1997.

Aunque el archivo contiene numerosas variables, en este trabajo se analizarán principalmente el colesterol total, la glucosa estabilizada (HDL), el ratio colesterol/HDL —que permite identificar la condición diabética cuando los valores son iguales o superiores a 7—, la edad y el peso expresado en libras.

Para más información sobre el contexto de esta investigación, pueden consultarse los siguientes trabajos:

- Willems JP, Saunders JT, DE Hunt, JB Schorling: *Prevalence of coronary heart disease risk factors among rural blacks: A community-based study*. Southern Medical Journal 90:814-820; 1997.
- Schorling JB, Roach J, Siegel M, Baturka N, Hunt DE, Guterbock TM, Stewart HL: *A trial of church-based smoking cessation interventions for rural African Americans*. Preventive Medicine 26:92-101; 1997.

Este conjunto de datos se emplea por ser de acceso público, con fines exclusivamente académicos y educativos, y se utiliza únicamente para la realización de ejercicios prácticos del curso.

En la siguiente tabla se realiza la descripción de las variables recogidas en ambos ficheros de datos:

403 observations and 19 variables, maximum # NAs:262

| Name | Labels | Units | Levels | Storage | NAs |
|----------|--|---------|--------|---------|-----|
| id | Subject ID | | | double | 0 |
| chol | Total Cholesterol | | | double | 1 |
| stab.glu | Stabilized Glucose | | | double | 0 |
| hdl | High Density Lipoprotein | | | double | 1 |
| ratio | Cholesterol/HDL Ratio | | | double | 1 |
| glyhb | Glycosolated Hemoglobin | | | double | 13 |
| location | | | 2 | integer | 0 |
| age | | years | | double | 0 |
| gender | | | 2 | integer | 0 |
| height | | inches | | double | 5 |
| weight | | pounds | | double | 1 |
| frame | | | 3 | integer | 12 |
| bp.1s | First Systolic Blood Pressure | | | double | 5 |
| bp.1d | First Diastolic Blood Pressure | | | double | 5 |
| bp.2s | Second Systolic Blood Pressure | | | double | 262 |
| bp.2d | Second Diastolic Blood Pressure | | | double | 262 |
| waist | | inches | | double | 2 |
| hip | | inches | | double | 2 |
| time.ppn | Postprandial Time when Labs were Drawn | minutes | | double | 3 |

| Variable | Levels |
|----------|------------|
| location | Buckingham |
| | Louisa |
| gender | male |
| | female |
| frame | small |
| | medium |
| | large |

1. Explora el *dataset*: Carga los datos descargados en formato CSV. Usa `str()` y `summary()` para inspeccionar la estructura y obtener un resumen de las variables.
2. Crea una tabla de frecuencia simple con la variable LOCATION, y una tabla cruzada entre las variables GENDER y FRAME (Frame: pequeño, mediano, grande, constitución corporal).

3. Haz un histograma del colesterol total (chol): Usa `hist()`. Personaliza con color `skyblue`, título “Distribución del colesterol total” y etiqueta del eje X “Colesterol (mg/dl)”.
 - a. Guarda el gráfico en archivo PNG: Usa `png()` y `dev.off()` con nombre `hist_colesterol.png`
4. Muestra varios gráficos en una misma ventana: Usa `par(mfrow=c(2,2))`. Incluye histograma (copia el anterior), boxplot, barplot y scatterplot juntos.
 - a. Haz un boxplot del HDL: Usa `boxplot()`. Color `lightgreen`, título “Boxplot de HDL” y etiqueta del eje Y “HDL (mg/dl)”.
 - b. Haz un barplot de hombres vs mujeres: Usa `barplot()`. Colores contrastantes: `tomato` para “Hombres/Male” y `lightblue` para “Mujeres/Female”. Título “Frecuencia inclusión en el estudio según sexo”.
 - c. Haz un scatterplot de peso vs colesterol: Usa `plot()`. Puntos en color `darkorange`, título “Relación entre peso y colesterol”. Añade una línea de tendencia con `abline()` en color azul.