LECCIÓN 3. DESCRIPCIÓN DE DATOS

SAAVEDRA, P.

Las ciencias experimentales se basan en la información aportada por conjuntos de datos, obtenidos como resultado de la observación de variables sobre los objetos del estudio en curso. En este capítulo se formalizan los conceptos de variable y de bases de datos y se presentan métodos de resumen de datos.

1. Variables estadísticas y bases de datos

Los **datos estadísticos** proceden de observar o medir atributos o magnitudes correspondientes a los elementos de una población. Tales atributos o magnitudes reciben el nombre de *variables estadísticas* o simplemente *variables*. Así, por ejemplo, el sexo o la edad son variables estadísticas. Las variables presentan modalidades o valores. Los de la variable *sexo*, por ejemplo, son: "varón" y "mujer". Los valores de una variable deben ser *incompatibles* y *exhaustivos*, esto es, la variable *debe asignar a cada elemento de la población un único valor* de un conjunto V de posibles valores. Cuando este conjunto es numérico ($V \subset \mathbb{R}$), la variable se dice *numérica* y en caso contrario, *categórica*. Los valores de las variables categóricas reciben el nombre de *categorías*. Si entre éstas existe una relación de orden, la variable se dice *ordinal*. Un caso especial de variables categóricas son las *binarias*, las cuales describen a menudo la presencia o ausencia de un carácter en cada elemento de la población.

Ejemplo 1. Las variables $Edad$, $Peso$ y $Talla$ son numéricas.

Ejemplo 2. De acuerdo con la forma de metabolismo de la glucosa, los individuos de una población Ω pueden clasificarse en los grupos NGT (normal), IFG (alteración de la glucosa en ayunas), IGT (intolerancia a la glucosa) y DM (diabetes mellitus). De esta forma la variable *Tolerancia a la glucosa*, que para cada sujeto indica el grupo al que pertenece, es categórica.

Ejemplo 3: La variable *hipertensión arterial* (criterio OMS), que expresa la presencia/ausencia de hipertensión arterial, es una variable categórica binaria.

La observación de una variable X sobre un elemento ω de la población de estudio Ω, proporciona un dato $x = X(\omega)$. El dato es categórico o numérico según la variable X sea categórica o numérica respectivamente. La observación de X sobre un conjunto de elementos $\{\omega_1, \ ... \ , \omega_N\}$ proporciona los datos $\{X(\omega_1), \cdots, X(\omega_N)\} = \{x_1, ..., x_N\}$. Un conjunto de variables $(X_1, ..., X_p)$ observadas sobre $\{\omega_1, ..., \omega_N\}$ proporciona un conjunto de datos que pueden organizarse en una tabla o matriz de datos como la que se muestra en la tabla 1.

Tabla 1. Estructura general de las bases de datos

	X_1	...	X_j	...	X_p
ω_1	$x_{1,1}$	...	$x_{1,j}$	...	$x_{1,p}$
...	...	...	...	...	...
ω_i	$x_{i,1}$	...	$x_{i,j}$	...	$x_{i,p}$
...	...	...	...	...	...
ω_N	$x_{N,1}$	...	$x_{N,j}$	...	$x_{N,p}$

Esta matriz recibe el nombre de tabla estadística de datos. Nótese que las columnas corresponden a las variables. Habitualmente cada fila corresponde a un objeto de estudio aunque en determinados estudios un elemento puede ocupar varias filas. El conjunto de datos lo representaremos habitualmente en la forma sintética $\{(x_{i,1}, ..., x_{i,p}) \ : \quad i = 1, ..., N\}$.

Ejemplo: La tabla 2 muestra algunas de las variables medidas en el estudio de Telde (Boronat *et al*, 2005). Se muestran solamente los datos correspondientes a 10 sujetos

de la muestra. En esta base de datos dada fila corresponde a un sujeto y cada columna es una variable.

TABLA 2. Fragmento de la base de datos del Estudio de Telde

ID	EDAD	SEXO	alcoh1	PESO	TALLA	IMC	COLESTEROL	HI
330	50	Mujer	No bebe	84.7	154	35.71429	196	4
761	52	Varón	< 30 g/día	104.0	169	36.41329	260	7
373	38	Mujer	No bebe	67.0	164	24.91077	214	4
354	42	Mujer	No bebe	51.0	158	20.42942	263	9
649	49	Mujer	No bebe	81.0	162	30.86420	191	4
1,345	53	Varón	No bebe	94.0	171	32.14664	214	4
489	59	Varón	< 30 g/día	76.0	166	27.58020	244	4
940	49	Mujer	No bebe	61.0	164	22.67995	205	4
1,154	33	Varón	No bebe	88.7	181	27.07488	188	4
562	64	Varón	< 30 g/día	100.0	175	32.65306	274	7

2. DATOS AGRUPADOS: DISTRIBUCIÓN DE FRECUENCIAS

En esta sección se considerarán datos correspondientes a variables categóricas observadas sobre un conjunto de sujetos. Se darán métodos de resumen de los mismos a través de tablas de frecuencias y representaciones gráficas.

Frecuencias absolutas y relativas. Considérese una variable estadística con categorías $C_1, \ldots, C_r$ observada sobre un conjunto de N sujetos. En tal contexto, la frecuencia absoluta de la categoría C_i se define como el es el número de sujetos n_i pertenecientes a C_i, mientras que la frecuencia relativa es la proporción f_i de que pertenecen C_i; esto

es: $f_i = n_i / N$, siendo $N = \sum_{i=1}^{p} n_i$ (número total de observaciones). La frecuencia relativa puede también expresarse en porcentaje obtenido como $f_i = 100 \times n_i / N$.

La variable *tolerancia a la glucosa* del ejemplo 2 se muestra resumida en la tabla 3. Nótese que el número de sujetos que pertenecen a la categoría DM es 128 (frecuencia absoluta) mientras que el porcentaje es 12,4%. Esta cantidad es lo que en términos epidemiológicos se llama la prevalencia de la enfermedad.

TABLA 3. Tabla de frecuencias para la variable Tolerancia a la glucosa

	Frecuencia	
Tolerancia a la glucosa	Absoluta	Relativa (%)
DM	128	12.4
IGT	107	10.4
IFG	132	12.8
NGT	663	64.4
Total	1,030	100.0

A menudo interesa convertir las variables numéricas en categóricas mediante la agrupación de sus valores en intervalos de clase. Este proceso recibe el nombre de categorización de la variable, siendo las nuevas categorías los intervalos de clase. La tabla 4 muestra la tabla de frecuencias de la variable *IMC* cuando sus valores se agrupan en intervalos de longitud 2.5. A modo de ejemplo, observemos que en esta tabla el intervalo 20-22.5 contiene a los 82 sujetos de la muestra cuyo valor de IMC es mayor que 20 y menor o igual que 22.5. Nótese que se han añadido dos columnas correspondientes a las frecuencias absolutas y relativas acumuladas, en las que se cuentan, respectivamente, el número total y el porcentaje de individuos cuyo valor de IMC es menor o igual que el valor superior de cada intervalo. Así, por ejemplo, la tabla nos indica que hay 698

sujetos con IMC menor o igual que 30, y que estas 698 personas componen el 67.8% de la muestra.

TABLA 4. Tabla de frecuencias para la variable IMC agrupada en intervalos

	Frecuencia		Frecuencia acumulada	
Valor de IMC	Absoluta	Relativa (%)	Absoluta	Relativa (%)
15 - 17.5	2	0.2	2	0.2
17.5 - 20	28	2.7	30	2.9
20 - 22.5	82	8.0	112	10.9
22.5 - 25	166	16.1	278	27.0
25 - 27.5	220	21.4	498	48.3
27.5 - 30	200	19.4	698	67.8
30 - 32.5	143	13.9	841	81.7
32.5 - 35	96	9.3	937	91.0
35 - 37.5	47	4.6	984	95.5
37.5 - 40	25	2.4	1,009	98.0
40 - 42.5	11	1.1	1,020	99.0
42.5 - 45	6	0.6	1,026	99.6
45 - 47.5	2	0.2	1,028	99.8
47.5 - 50	0	0.0	1,028	99.8
50 - 52.5	1	0.1	1,029	99.9
52.5 - 55	0	0.0	1,029	99.9
55 - 57.5	1	0.1	1,030	100.0
57.5 - 60	0	0.0	1,030	100.0

Representaciones gráficas para variables categóricas: diagramas de barras y circulares. Las variables categóricas pueden representarse mediante diagramas circulares y diagramas de barras. El diagrama circular se construye dividiendo un círculo en tantos sectores como categorías tenga la variable. Cada categoría de la variable tiene asignada un sector de área proporcional a su frecuencia. En el diagrama de barras se asigna a cada categoría un rectángulo levantado sobre un eje horizontal de área proporcional a su frecuencia. La figura 1 muestra los diagramas circular y de barras para la variable de grupos de tolerancia a la glucosa.

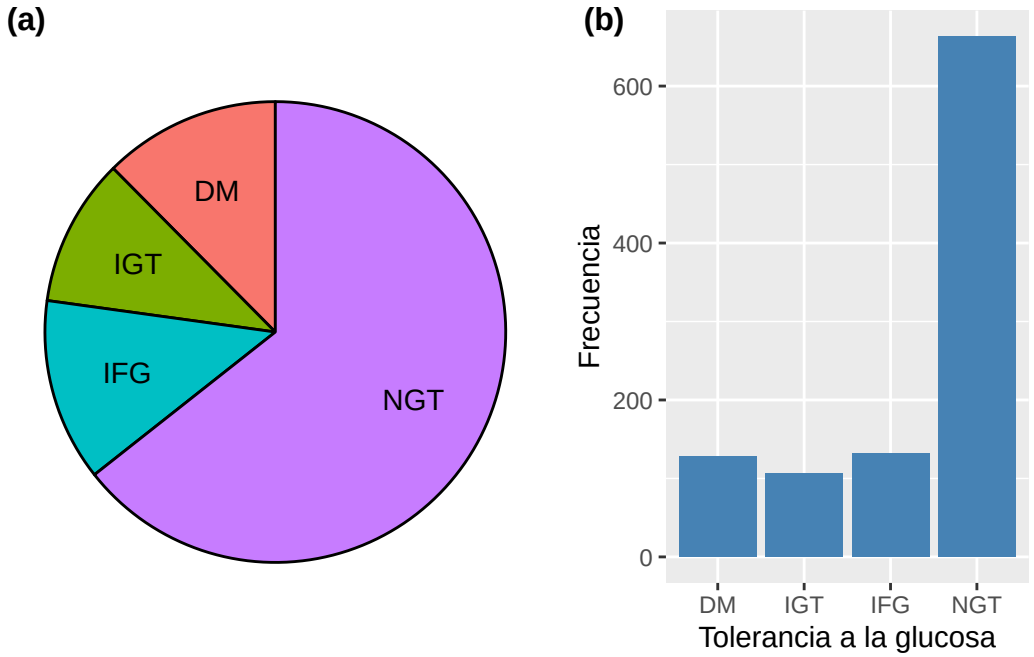

FIGURA 1. Diagramas: (a) circular; (b) barras

Histogramas. La distribución de frecuencias de una variable numérica puede representarse gráficamente mediante los *histogramas* de frecuencias absolutas y relativas. Para construir un histograma de frecuencias absolutas, los valores de la variable deben previamente agruparse en intervalos de la misma longitud. El histograma se obtiene levantando sobre cada intervalo un rectángulo *cuya altura coincide con la frecuencia absoluta* de ese intervalo. El histograma de frecuencias relativas requiere también agrupar los valores de la variable en intervalos, pero no necesariamente de la misma longitud. El histograma se completa levantando sobre cada intervalo un rectángulo *cuya área coincide con la frecuencia relativa de dicho intervalo*. La figura 2 corresponde a los histogramas de frecuencias absolutas (a) y frecuencias relativas (b) de la variable IMC cuyos valores se agruparon en intervalos de longitud 2.5.

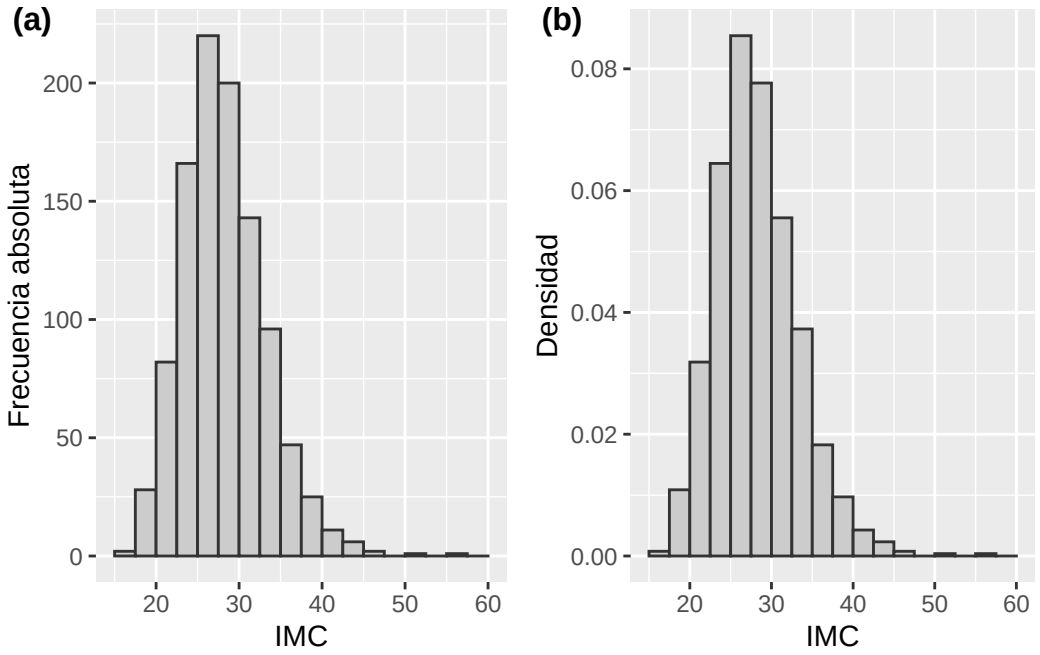

FIGURA 2. Histograma de frecuencias para el índice de masa corporal: (a) absolutas; (b) relativas

3. SÍNTESIS DE DATOS NUMÉRICOS

Las variables numéricas pueden resumirse a través de parámetros de posición (percentiles), localización (medias), dispersión (varianza, desviación estándar, coeficiente de variación, rango y rango intercuartílico) y de forma (coeficiente de asimetría). Se definen ahora estos parámetros para el conjunto de datos $\{x_i : i = 1, \dots, N\}$ correspondientes a una variable numérica X.

Percentiles. Para una variable numérica X, el percentil k es un valor P_k tal que el $k\%$ de las observaciones de X son inferiores a P_k. En la tabla 4 se puede observar que el 67.8% de la población tiene un valor de IMC inferior a 30 siendo, por tanto, 30 el percentil 67.8 de la variable IMC. El percentil 50 (P_{50}) de una distribución recibe el nombre de **mediana**. Los percentiles 25, 50 y 75 se denominan cuartiles 1, 2 y 3 respectivamente.

De la tabla de frecuencias para el IMC (tabla 4) se deduce también que el valor de 25 (kg/m^2) es el percentil 27, lo que supone que, de acuerdo con la OMS, el 73% de la población tiene sobrepeso. Se propone como ejercicio, obtener una aproximación a la mediana de esta variable.

Las curvas de percentiles son un instrumento de gran valor para el control del crecimiento en las poblaciones infantiles. La figura 3 muestra las curvas de percentiles 3, 10, 25, 50, 75, 90 y 97 para el índice de masa corporal correspondientes a una muestra de 484 niños de 0 a 12 años. Las curvas muestran la evolución de los percentiles desde el nacimiento hasta los 12 años.

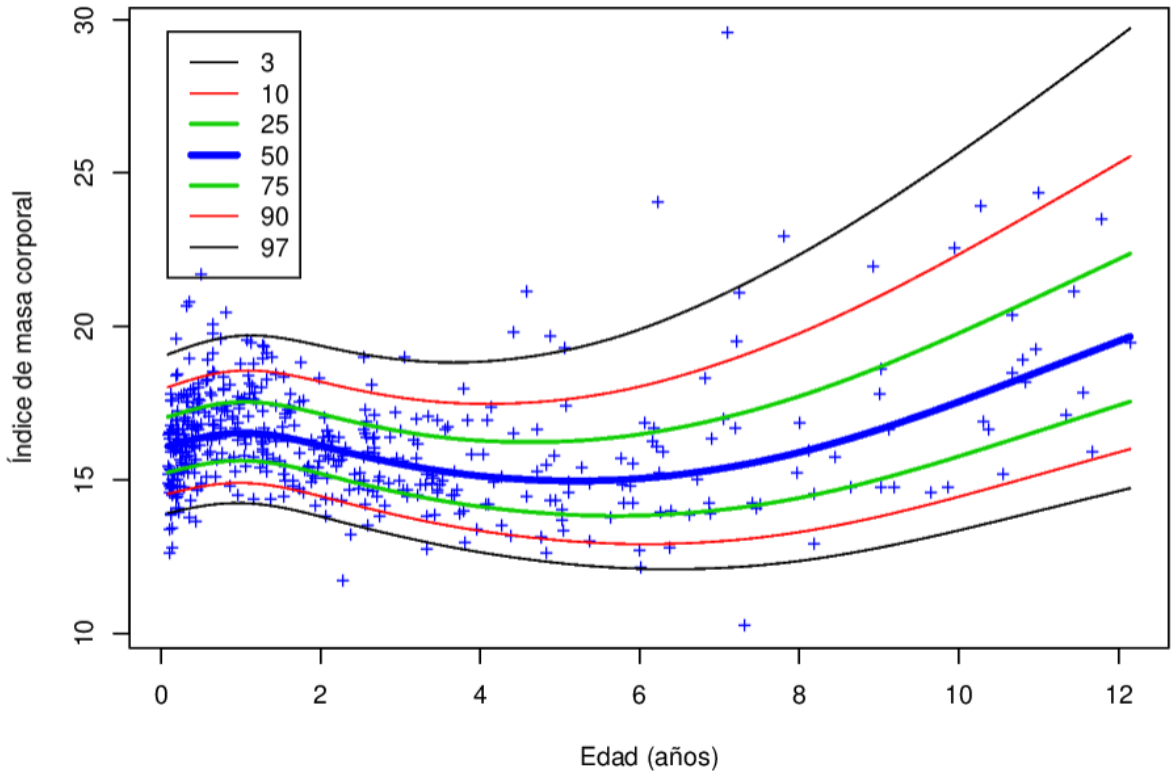

FIGURA 3. Curvas de crecimiento para el IMC

Media aritmética. La media aritmética representa el centro físico de gravedad del conjunto de datos. Se calcula así:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

La media de la variable *IMC* (índice de masa corporal) en el estudio de Telde fue 28.2. La figura 4 muestra la posición de la media en la base del histograma de frecuencias de dicha variable. Puede apreciarse que la media coincide efectivamente con el centro físico de masas del histograma, el punto en que dicha figura se mantendría en equilibrio sin "caer" a derecha ni a izquierda.

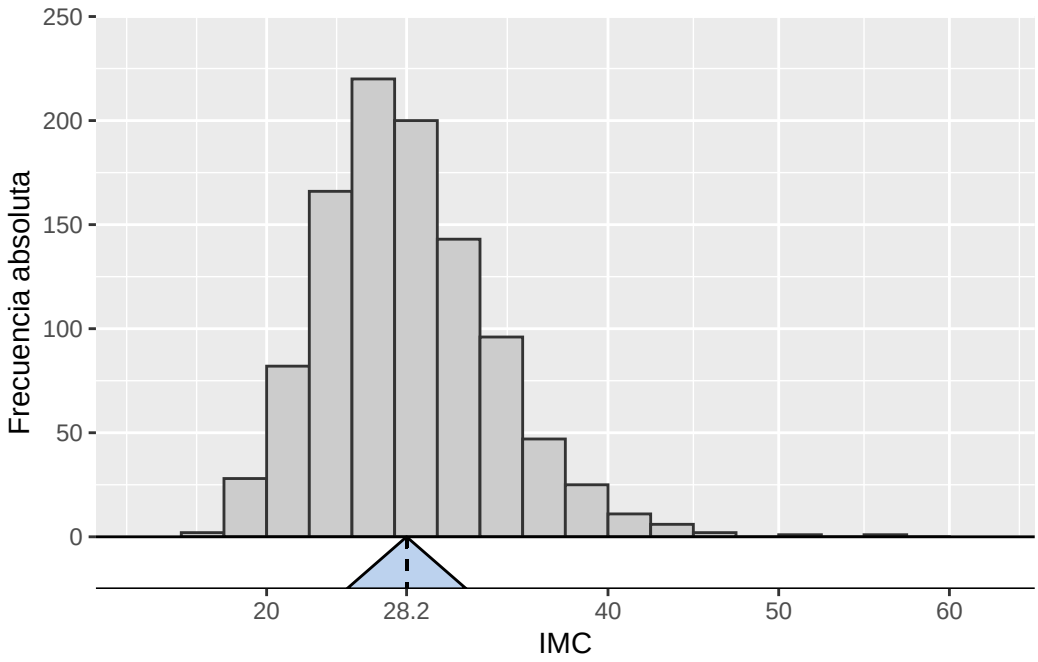

FIGURA 4. Posición de la media aritmética del IMC

Media geométrica. Posteriormente veremos que en determinados casos, un parámetro de resumen alternativo a la media aritmética es la media geométrica, la cual se define por:

$$\gamma = \{x_1 \times \cdots \times x_N\}^{1/N}$$

Varianza y desviación típica. La varianza mide la dispersión de los valores de una variable mediante:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

La raíz cuadrada de la varianza recibe el nombre de desviación típica o desviación estándar.

Coeficiente de variación. Varianza y desviación estándar son medidas de dispersión dependientes de las unidades en las que se mida la variable. Una medida de dispersión adimensional es el coeficiente de variación, el cual se define por:

$$\mathrm{cv}\,(X) = \frac{\sigma}{\mu}$$

Probaremos ahora que este parámetro es invariante para cambios de escala. En efecto, considérese el cambio de escala $Y = kX\ ,\ k > 0$. Si representamos por $\{x_1, \ldots, x_N\}$ los valores de X, los correspondientes a Y serán $\{kx_1, \ldots, kx_N\}$. La media aritmética μ_Y de la variable Y es entonces:

$$\mu_Y = \frac{1}{N}\sum_{i=1}^{N} kx_i = \frac{k}{N}\sum_{i=1}^{N} x_i = k\mu_X$$

donde μ_X es la media de la variable X. Análogamente para la varianza σ_Y^2:

$$\sigma_Y^2 = \frac{1}{N}\sum_{i=1}^{N}(kx_i - \mu_Y) = \frac{1}{N}\sum_{i=1}^{N}(kx_i - k\mu) = \frac{k^2}{N}\sum_{i=1}^{N}(x_i - \mu_X) = k^2\sigma_X^2$$

siendo σ_X^2 la varianza de la variable X. Por tanto:

$$\sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{k^2\sigma_X^2} = k\sigma_X$$

lo que supone:

$$\mathrm{cv}\,(Y) = \frac{\sigma_Y}{\mu_Y} = \frac{k\sigma_X}{k\mu} = \frac{\sigma_X}{\mu_X} = \mathrm{cv}\,(X)$$

Rango y rango intercuartílicos. El rango de una variable se define como la distancia entre los valores mínimo y máximo ($\text{rango}(X) = \max(X) - \min(X)$), mientras que el rango intercuartílico (IQR) es la distancia entre los cuartiles primero y tercero; esto es: $\text{IQR} = P_{75} - P_{25}$.

Ejemplo: La tabla 5 muestra las medidas de síntesis anteriores calculadas para la variable *triglicéridos*

TABLA 5. Medidas de síntesis para la variable Triglicéridos

	Valor
Media	122.61
Varianza	5,723.28
Desviación típica	75.65
Coeficiente de Variación	61.70
Percentil 25	75.00
Mediana	104.00
Percentil 75	149.00
Rango	816.00
IQR	74.00

Diagrama de cajas y barras (boxplot). Estos diagramas representan los percentiles de una variable y son especialmente útiles para una comparación gráfica de varias poblaciones. Su construcción se realiza de la siguiente forma. Sea $\{x_1, \ldots, x_N\}$ el conjunto de datos correspondientes a una variable numérica X y representemos por P_{25} , P_{50} y P_{75} los percentiles 25, 50 y 75 respectivamente. Los lados inferior y superior del

rectángulo corresponden a P_{25} (primer cuartil) y P_{75} (tercer cuartil) respectivamente y el segmento medio a P_{50} (mediana). Las barras (extremos de la distribución) se obtienen de la siguiente forma:

- Barra superior: $B = \min\{\max(X), P_{75} + 1.5(P_{75} - P_{25})\}$
- Barra inferior: $b = \max\{\min(X), P_{25} - 1.5(P_{75} - P_{25})\}$

Según los datos del estudio del Telde, los percentiles de la variable HDL en varones son: $P_{25} = 42$, $P_{50} = 49$ y $P_{75} = 56$. Además, $\min(HDL) = 21$ y $\max HDL = 90$. De esta forma, $P_{75} + 1.5(P_{75} - P_{25}) = 77$. Por tanto, $B = \min\{77\ ;\ 90\} = 77$. En la figura 5 se muestran conjuntamente los boxplots de la HDL para los subgrupos de hombres y mujeres. Nótese que la barra superior para los varones es $B = 77$. Las observaciones superiores a este valor se consideran como atípicas (outliers). Obsérvese también que los niveles de esta variable son *superiores* (distribución desplazada hacia arriba) en el grupo de mujeres.

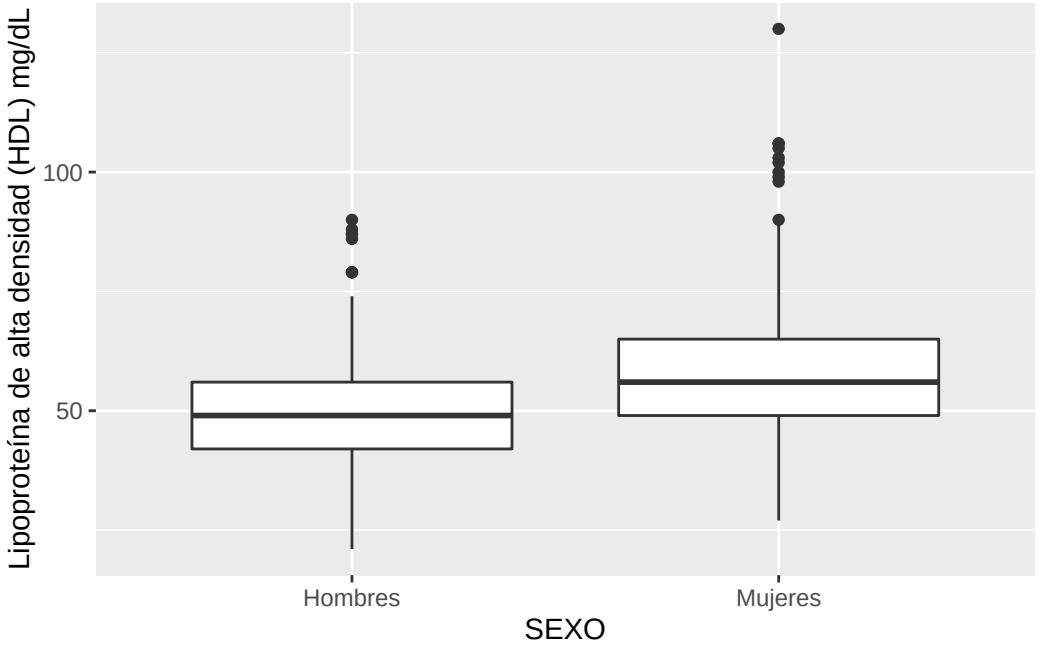

FIGURA 5. Distribución de la HDL según género

Coeficiente de asimetría. La figura 6 (a) muestra la distribución de los triglicéridos (TG) expresados en mg/dL. Nótese que la forma de la distribución presenta una notable *asimetría*. La primera implicación de este hecho suele ser una notable disparidad entre la media aritmética y la mediana del conjunto de datos. En el caso que nos ocupa, la

media es de 122.6 mg/dL mientras que la mediana es de 104 mg/dL. Otro efecto no deseable de la asimetría es la escasa fiabilidad del uso de numerosos procedimientos estadísticos. Para evaluar el grado de asimetría de una distribución usaremos el coeficiente de asimetría (*skewness*), el cual se define por:

$$a\left(X\right)=\frac{1}{\sigma^{3}N}\sum_{i=1}^{N}\left(x_{i}-\mu\right)^{3}$$

Cuando la distribución presenta una simetría perfecta, $a\left(X\right)=0$. Para la variable TG resumida en la figura 6 (a), $a\left(TG\right)\approx 3.029$.

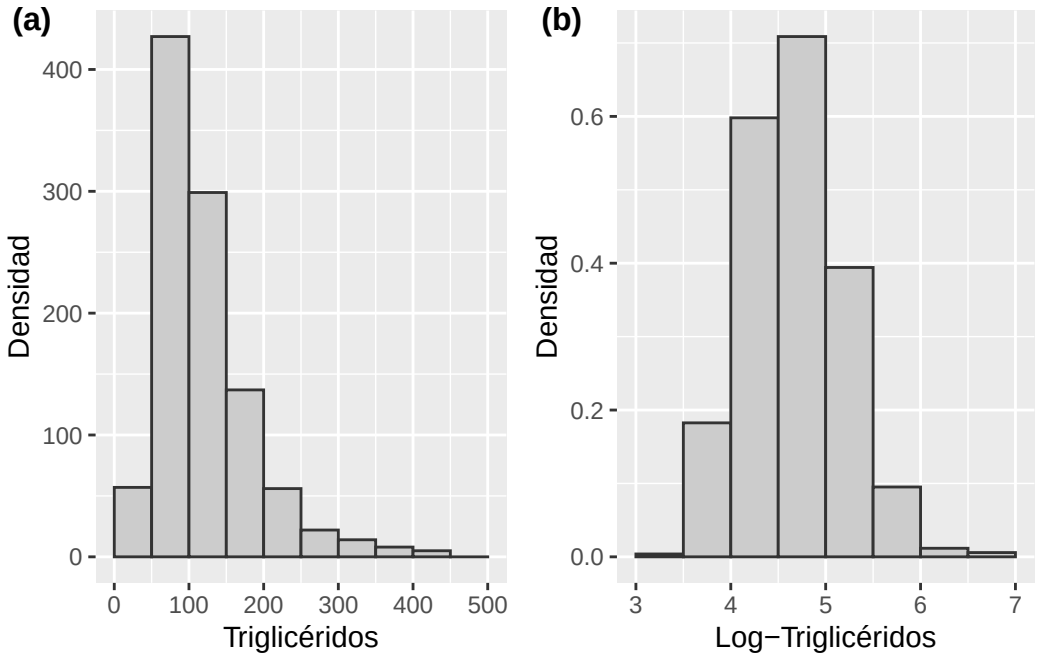

FIGURA 6. Distribución de los triglicéridos (TG) en: (a) en mg/dL; (b) en escala logarítmica

Cuando la distribución muestra una fuerte asimetría puede ser conveniente transformar la variable logarítmicamente. La figura 6 (b) muestra el histograma de la variable TG en escala logarítmica. Puede verse ahora que la distribución es prácticamente simétrica, siendo en este caso $a\left(\log TG\right)=0.397$. Una consecuencia inmediata es la similitud entre media y mediana (4.671 y 4.644 respectivamente).

La transformación logarítmica (en general, cualquier transformación de un marcador) conlleva perder la referencia de los valores de normalidad. Por tal motivo, una vez obtenida la media de los logaritmos conviene volver a la escala original mediante la

transformación inversa (para la logarítmica es la exponencial). Lo curioso, como veremos a continuación, es que la transformación inversa de la media de los logaritmos es la media geométrica de los datos originales. En efecto si $Y = \log X$ se tiene:

$$\mu_Y = \frac{1}{N}\sum_{i=1}^{N} \log x_i = \log \left\{ \prod_{i=1}^{N} x_i \right\}^{1/N}$$

El parámetro μ_Y resume los valores de la variable en la escala logarítmica. El resumen en la escala original puede hacerse mediante la transformación inversa (exponencial). De esta forma, la medida resumen es:

$$\exp\left(\mu_Y\right) = \exp\left(\log \left\{ \prod_{i=1}^{N} x_i \right\}^{1/N} \right) = \left\{ \prod_{i=1}^{N} x_i \right\}^{1/N}$$

lo que significa que $\exp\left(\mu_Y\right)$ es la media geométrica.

Para el caso de la variable TG (Triglicéridos) del estudio de Telde, se tiene que la mediana vale 104, la media de los valores en escala logarítmica vale 4.672, y:

$$\exp\left(4.672\right) = 106.9$$

Gráficos de dispersión. La relación entre dos variables numéricas X e Y puede explorarse mediante los diagramas de dispersión. Consideremos entonces el conjunto de datos correspondiente a la observación simultánea de las variables X e Y sobre un conjunto de N elementos. En el diagrama de dispersión, cada dato (x_i, y_i) se representa mediante un punto del plano XY con esas coordenadas.

La figura 7 corresponde al diagrama de dispersión para las variables glucemia basal y glucemia tras sobrecarga de glucosa (SOG) correspondientes al estudio de Telde. Nótese que los sujetos con tolerancia normal a la glucosa son aquellos que tienen la glucemia basal inferior a 100 mg/dL y la SOG inferior a 140 mg/dL. Cuando un individuo tiene un valor basal superior a 100 mg/dL e inferior a 126 mg/dL y la SOG se mantiene por debajo de 140 mg/dL, se dice que tiene alteración de la glucosa en ayunas (IFG). Si

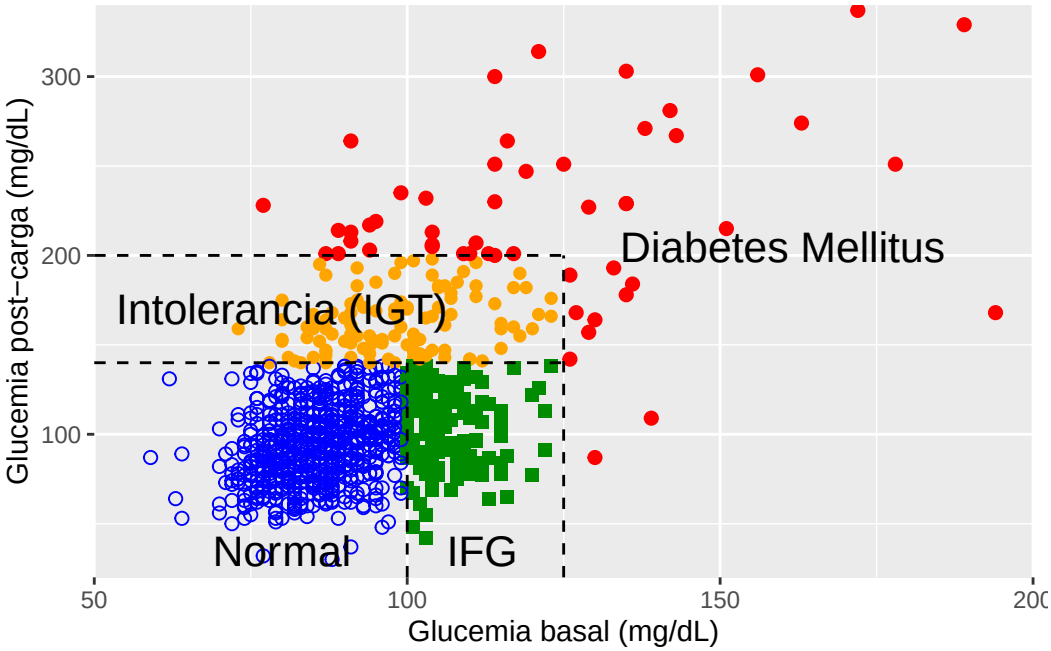

FIGURA 7. Tolerancia a la glucosa según valores de glucemia basal y post-carga

la basal es inferior a 126 mg/dL y la SOG está comprendida entre 140 mg/dL y 200 mg/dL, se dice que tiene intolerancia a la glucosa (IGT). Finalmente, cuando la basal supera los 126 mg/dL o la SOG los 200 mg/dL, se dice que tiene diabetes mellitus (DM).

4. BIBLIOGRAFÍA

Boronat, M., Varillas, M.V., Saavedra, P., Suárez, V., Bosch, E., Carrillo, A. and Nóvoa, F.J. (2005). Diabetes mellitus and impaired glucose regulation in the Canary Islands (Spain): prevalence and associated factors in the adult population of Telde, Gran Canaria. *Diabetes Care*, **28**, 2388:2393. pdf