4A. ESTADÍSTICA DESCRIPTIVA Y REPRESENTACIONES GRÁFICAS.

La descripción y representación gráfica de los datos es la parte inicial de cualquier estudio de investigación. A veces es su único objetivo. Pretende sintetizar la información contenida en los datos y mostrarla visualmente para contribuir a los análisis posteriores e incluso sugerir hipótesis nuevas o complementarias para la investigación. La estadística descriptiva y gráficos de las variables categóricas se obtienen con el cuadro *Frecuencias*, y las de las variables cuantitativas con el cuadro *Explorar*, aunque también se pueden obtener parcialmente con el cuadro *Frecuencias*.

ESTADÍSTICA DESCRIPTIVA Y REPRESENTACIÓN GRÁFICA DE VARIABLES CATEGÓRICAS

La estadística descriptiva de ***variables categóricas*** consiste en clasificar a los sujetos según la categoría a la que pertenecen en una tabla con la distribución de frecuencias absolutas (recuentos) y relativas (proporciones o porcentajes). Es el caso de las variables *Sexo*, *Hospital* y *Nivel de masa corporal* de la Tabla 4A.1. Cuando hay *user missing* los porcentajes respecto al total (*user missing* incluidos) y los porcentajes respecto a los valores válidos (*user missing* excluidos) son diferentes. Es el caso de las variables *TabacoBi* y *TabacoOr* de la Tabla 4B.1. También se pueden indicar los porcentajes acumulados, de mayor interés en el caso de las variables ordinales.

Tabla 4A.1. Estadística descriptiva de las variables categóricas *Sexo*, *Hospital* y *Nivel de masa corporal*.

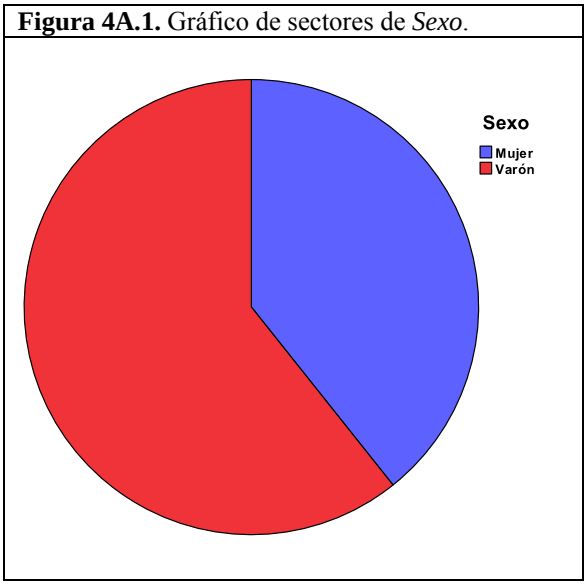
Sexo	n (%)
Mujer	44 (39,3)
Varón	68 (60,7)
Total	112 (100)

Hospital	n (%)
Príncipe de Asturias	28 (25,0)
Gregorio Marañón	28 (25,0)
Ramón y Cajal	23 (20,5)
12 de Octubre	17 (15,2)
Clínico San Carlos	16 (14,3)
Total	112 (100)

Nivel de masa corporal	n (%)
Imc Normal	18 (16,1)
Sobrepeso	51 (45,5)
Obesidad	43 (38,4)
Total	112 (100)

La representación gráfica de variables categóricas se hace mediante el ***diagrama de barras*** o con el ***de sectores***. En la Figura 4A.1 se muestra el diagrama de sectores de *Sexo*. El ángulo de cada sector es proporcional a la frecuencia de la categoría. En el diagrama de barras es recomendable ordenar las categorías de las variables nominales por frecuencias o recuentos descendentes (de mayor a menor frecuencia), y las de las variables ordinales por valores o códigos ascendentes (del valor o código menor al valor o código mayor). En la Figura 4A.2 se muestra el diagrama de barras de la variable nominal *Hospital* ordenado por frecuencias descendentes (de mayor a menor frecuencia). En la Figura 4A.3 se muestra el diagrama de barras de la variable ordinal *Nivel de masa corporal* por valores ascendentes (del valor o código menor –el 0 de *Imc Normal*– al mayor –el 2 de *Obesidad*–).

Figura 4A.1. Gráfico de sectores de *Sexo*.



Videotutorial 4A1Frecuencias1.avi

Se muestra como se obtienen la tabla de frecuencias y el gráfico de sectores de *Sexo*, la tabla de frecuencias y el diagrama de barras ordenado por recuentos descendentes de la variable nominal *Hospital* y la tabla de frecuencias y el diagrama de barras ordenado por valores ascendentes de la variable ordinal *ObesidadOr*, todos ellos con el cuadro *Frecuencias*.

Figura 4A.2. Diagrama de barras de *Hospital*.

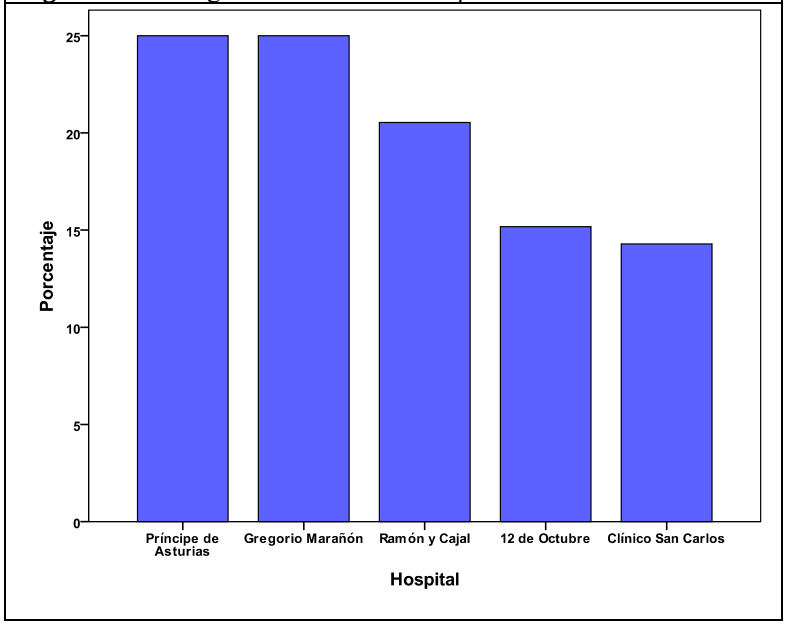
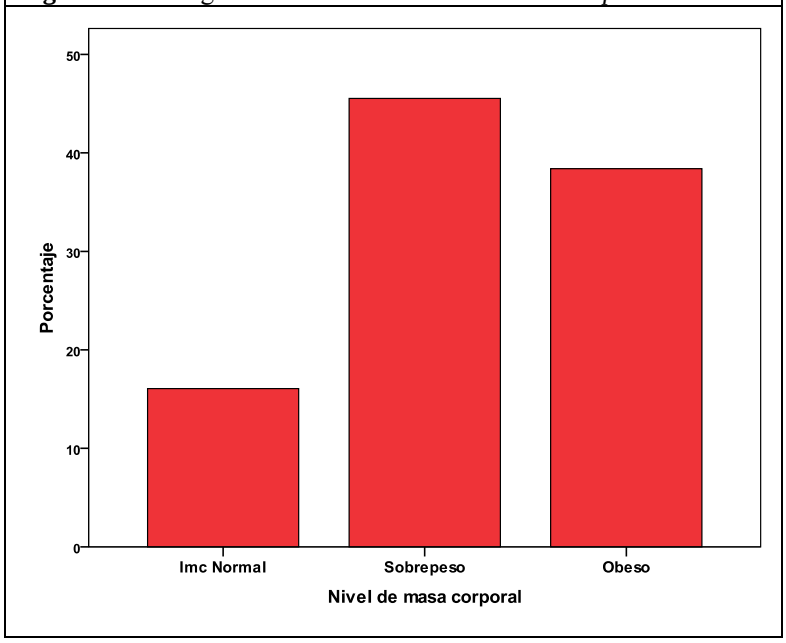


Figura 4A4. Diagrama de barras de *Nivel de masa corporal*.



ESTADÍSTICA DESCRIPTIVA DE VARIABLES CUANTITATIVAS

La descripción de *variables cuantitativas* consiste en sintetizar los datos con un índice de tendencia central (media, mediana o moda) y otro de dispersión (desviación estándar, amplitud intercuartil o rango). También se pueden dar índices de posición (percentiles, deciles, cuartiles) y de forma (asimetría y apuntamiento).

Medidas de tendencia central o de centralización

Media (M) aritmética.
Medida de tendencia central basada en valores: suma de los valores (S) dividida por el número de casos (n): M = S/n. Tiene las mismas unidades de medida que la variable. Representa el centro de gravedad de la distribución. Describe bien la tendencia central si la distribución de datos es simétrica y se ajusta a la ley normal, en cuyo caso es similar a la mediana. La media de *Imc* es 29,3 kg/m^2 y la de *Glucosa* 154,9 mg/dl (Tabla 4A.2).

Mediana (Md).
Medida de tendencia central basada en ordenaciones: corresponde al valor central de los valores ordenados. Es el valor que divide la distribución en dos partes iguales: el 50% de las observaciones presentan valores iguales o

inferiores a la mediana y el otro 50% presenta valores superiores. Si la distribución es simétrica la mediana es muy similar a la media. La mediana de *Imc* es 28,6 kg/m^2 y la de *Glucosa* 145 mg/dl (Tabla 4A.2).

Moda (Mo).
Medida de tendencia central más grosera: corresponde al valor que ocurre más frecuentemente. En caso de haber varios valores de la variable que tengan la misma frecuencia máxima, habrá varias modas. La moda de *Imc* es 28,4 kg/m^2 y la de *Glucosa* 113 mg/dl (Tabla 4A.2).

Tabla 4A.2. Estadística descriptiva de las variables las cuantitativas *Imc* y *Glucosa*.

	Imc (kg/m^2)	**Glucosa (mg/dl)**
n	112	112
Media	29,3	154,9
Mediana	28,6	145
Moda	28,4	113
Varianza	19,7	5117
Desviación Estándar	4,4	71,5
Primer cuartil	26,6	95,5
Tercer cuartil	32,2	218,7
Amplitud intercuartil	5,6	123
Mínimo	15,5	46
Máximo	40,2	298
Amplitud o Rango	24,7	252
Coeficiente de variación	15,2	46,2
Asimetría	0,140 (EE = 0,228)	0,380 (EE = 0,228)
Apuntamiento	0,160 (EE = 0,453)	−1,040 (EE = 0,453)

Medidas de dispersión y de posición

Son índices que indican el grado de proximidad o alejamiento de los valores respecto a un valor central.

Amplitud, **Rango** o **Recorrido** (R).
Es la medida de dispersión más grosera. Corresponde a la diferencia entre los valores **máximo** y **mínimo**. La amplitud de *Imc* es 24,7 kg/m^2 (mínimo 15,5 y máximo 40,2) y la de *Glucosa* 252 mg/dl (mínimo 46 y máximo 298) (Tabla 4A.2).

Varianza (V).
Para obtener una medida de dispersión más precisa que el Rango, lo más intuitivo es calcular las diferencias de cada valor (x_i) con la media (M) (variable centrada), y obtener su promedio: $\Sigma(M-x_i)/n$. Pero habría diferencias positivas y negativas, por lo que tenderían a anularse entre sí. Para evitarlo se elevan al cuadrado esas diferencias, quedando el numerador así: $\Sigma(M-x_i)^2$. Este numerador es la denominada *Suma de cuadrados* (*SC*) o *Sum of Squares* (*SS*). El denominador, *n*, también se corrige por *n−1*, debido a que para el cálculo de este parámetro con estos datos se utiliza otro parámetro (la media *M*) obtenido con los mismos datos, lo cual obliga a uno de los casos a tomar un valor determinado, "no es libre de tomar cualquier valor". Este es el concepto de ***grados de libertad*** (gl) ó *degree of freedom* df) de un parámetro, tan importante en estadística. En general es el número total de observaciones *n*, menos el número de parámetros estimados con estas mismas observaciones que intervienen en su cálculo, en este caso *n−1*, ya que en el cálculo de la varianza interviene la media que ha sido calculada con estos mismos valores. Todos los valores son "libres" de tomar cualquier valor excepto uno.

El parámetro resultante de dividir la *Suma de Cuadrados* ($SC= \Sigma(M-x_i)^2$) de la variable centrada, entre los grados de libertad ($gl=n-1$) es la ***Varianza***, $V= \Sigma(M-x_i)^2/n-1$. La varianza también se denomina *Media Cuadrática* (*MC*), y puede expresarse como $MC = SC/gl$. Esta expresión es de utilidad para entender el análisis de la varianza explicado en el Unidad 11. Tiene como unidades de medida el cuadrado de la unidad de medida de la variable, siendo este su principal inconveniente. Es la medida de dispersión basada en los valores más importante, aunque no es la más empleada. La varianza de *Imc* es 19,7 (kg/m^2)^2 y la de *Glucosa* 5117 (mg/dl)^2 (Tabla 4A.2).

Desviación Estándar (*DE*).
También llamada *Desviación Típica*, es la medida de dispersión basada en valores más utilizada. Es la raíz cuadrada positiva de la varianza, $DE=+\sqrt{V}$, en un intento de corregir el cuadrado de las diferencias. Tiene las mismas unidades de medida que la variable, lo que la hace más inteligible. La desviación estándar de *Imc* es 4,4 kg/m^2 y la de *Glucosa* 71,5 mg/dl (Tabla 4A.2). Mide la variabilidad de los datos alrededor de la media y solo tiene interpretación práctica cuando la distribución es normal, en cuyo caso se cumple:

- El intervalo M ± 1×DE contiene el 68% central de las observaciones, aproximadamente.
- El intervalo M ± 2×DE contiene el 95% central de las observaciones; es el Intervalo de Normalidad.

– El intervalo M ± 3×DE contiene el 99,7% central de las observaciones, aproximadamente.

Tabla 4A.3. Medidas de posición.

Nº Orden	Glucosa	Cuantiles
1	46	
2	51	
.	.	
10	61	
11	**62**	Percentil 10
12	64	
.	.	
21	90	
22	**91**	Percentil 20
23	92	
.	.	
27	95	
28	**95**	1er Cuartil
29	97	
.	.	
32	100	
33	**102**	Percentil 30
34	103	
.	.	
43	116	
44	**120**	Percentil 40
45	120	
.	.	
55	140	
56	**144**	**(145) Mediana**
57	**146**	
58	146	
.	.	
66	161	
67	**161**	Percentil 60
68	163	
.	.	
77	186	
78	**186**	Percentil 70
79	192	
.	.	
83	206	
84	**218**	3er Cuartil
85	219	
.	.	
88	232	
89	**242**	Percentil 80
90	242	
.	.	
99	259	
100	**260**	Percentil 90
101	263	
.	.	
112	298	

Coeficiente de variación (CV).

La varianza, la desviación estándar, la amplitud intercuartil y el rango son índices de dispersión absolutos. No sirven para comparar la dispersión de dos variables distintas. Cada índice tiene las unidades de medida de cada variable, por lo que comparan no solo la dispersión de las variables sino sus unidades de medida. El ***Coeficiente de variación*** es la medida de dispersión relativa más utilizada que carece de unidades de medida y representa el

valor de la *DE* en unidades de la "media" (*CV* = *DE*/*M*). En general unos datos con CV > 30% se consideran datos dispersos. Se puede ver que la dispersión de *Glucosa* (CV=46,2) es mayor que la *Imc* (CV=15,2%), pero no tanto como indican la comparación de sus respectivas varianzas.

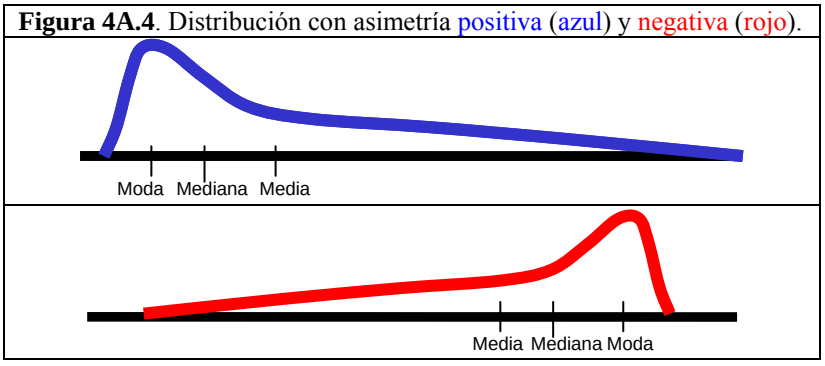
Percentiles (P), **Cuartiles** (Q) y **Amplitud intercuartil** (AIC).
Los *Cuantiles* o *n-tiles* son cada uno de los n−1 valores de la variable que dividen los datos ordenados en n partes de igual tamaño. Son medidas que informan de la posición respecto al resto de individuos o respecto a un grupo de referencia. En la Tabla 4A.3 se muestran los 112 valores de *Glucosa* ordenados de menor a mayor para facilitar el cálculo de las medidas de posición. Si n=100 los cuantiles son ***Percentiles*** (P), cada uno de los 99 valores de la variable que dividen los datos ordenados en 100 partes de igual tamaño. El percentil 90 (P90) de *Glucosa* es 260 mg/dl (Tabla 4A.3), valor que divide la muestra en un 90% de casos con glucosa menor o igual a 260, y un 10% de casos con glucosa mayor de 260 mg/dl. Si n=10 los cuantiles son ***Deciles*** (D). En caso de n=4 son ***Cuartiles*** (Q). El primer cuartil Q1 es 95 mg/dl (Tabla 4A.3). Indica que el 25% de los casos tienen glucosa igual o inferior a 95 mg/dl, y el 75% de los casos tienen glucosa mayor de 95 mg/dl. El tercer cuartil Q3 es 218 mg/dl (Tabla 4A.3). Indica que el 75% de los sujetos tiene glucosa igual o inferior a 218 mg/dl y el 25% de los casos tienen glucosa mayor de 218 mg/dl. El programa ofrece los cuartiles calculados por dos procedimientos: Promedio ponderado y Bisagras de Tukey. Si n=1 es la Mediana (glucosa de 145), que coincide con el segundo cuartil (Q2), el decil 5 (D5) y el percentil 50 (P50). El intervalo o **Amplitud intercuartil** (AIC) es la diferencia entre el tercer y el primer cuartil, AIC=Q3−Q1. Es el intervalo de valores que contiene el 50% central de los casos de la distribución, y corresponde a la "caja" del diagrama de caja, como se verá más adelante. En el caso de Glucosa es AIC=218–95=123 (Tabla 4A.3).

Medidas de forma

Son el índice de asimetría y el de apuntamiento.

Asimetría (AS).
Informa de la magnitud y del sentido de la desviación respecto a la simetría. Puede ser:
- **Asimetría Nula** (AS=0): distribución simétrica, el índice de asimetría esta dentro +/−2 veces su EE (Error Estándar). El error estándar es la desviación estándar de la distribución muestral. Es un concepto de estadística inferencial que se desarrolla en la próxima unidad. Es el caso de *Imc* (AS=0,14; EE=0,228) y *Glucosa* (AS=0,38; EE=0,228) (Tabla 4A.2).
- **Asimetría Positiva** (AS>0): la cola se aleja por la derecha. La media es mayor que la mediana (Figura 4A.4). El índice de asimetría es mayor que +2 veces su EE. Es frecuente en las variables utilizadas en biomedicina, sobre todo en las distribuciones acotadas por el 0 (Creatinina, Bilirrubina, etc.).
- **Asimetría Negativa** (AS<0): la cola se aleja por la izquierda. La media es menor que la mediana (Figura 4A.4). El índice de asimetría es menor de −2 veces su EE. Es mucho menos frecuente.

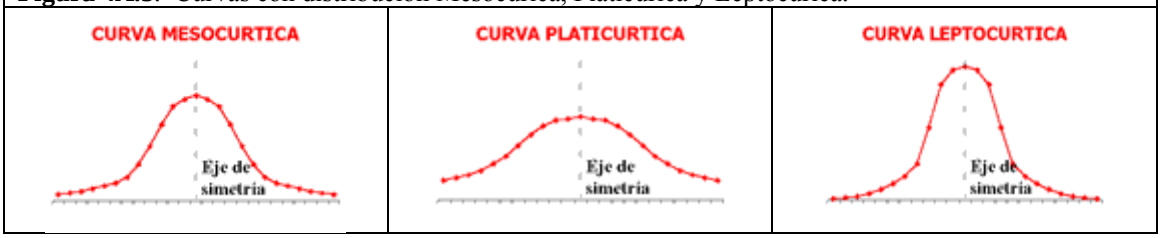
Figura 4A.4. Distribución con asimetría positiva (azul) y negativa (rojo).



Apuntamiento (AP) o **Curtosis**.
Informa de la magnitud y del sentido de la desviación respecto al apuntamiento. Una distribución puede ser:
- **Mesocúrica** (AP=0): con apuntamiento Nulo, como la ley normal (Figura 4A.5). El índice de apuntamiento esta comprendido entre +/− 2 veces su EE. Es el caso de Imc: AP=0,16 (EE=0,453) (Tabla 4A.2).
- **Leptocúrica** (AP>0): con apuntamiento Positivo., más apuntada que la ley normal, los valores se acumulan en la parte central (Figura 4A.5). El índice de apuntamiento es mayor que +2 veces su EE.
- **Platicúrica** (AP<0): con apuntamiento Negativo, más aplanada que la ley normal, los valores se acumulan en las colas (Figura 4A.5). El índice de apuntamiento es menor que −2 veces su EE. Es el caso de *Glucosa*: AP=−1,04 (EE=0,453) (Tabla 4A.2).

La distribución del *Imc* es simétrica y mesocúrica. Ambos índices se alejan de cero menos de dos veces su EE. Por ello la media es muy similar a la mediana y la desviación estándar tiene interpretación práctica. La distribución de de *Glucosa* es simétrica y platicúrica.

Figura 4A.5. Curvas con distribución Mesocúrica, Platicúrica y Leptocúrica.



¿Índices basados en valores o en ordenaciones?

Los ***índices basados en valores*** (*media*, *varianza* y *desviación estándar*) se calculan con los propios valores de la variable. Tradicionalmente han sido los más utilizados, pero se ven afectados por la presencia de valores extremos, por lo que solo deberían usarse en distribuciones simétricas que no presenten anomalías, es decir cuando siguen la ley normal.

Los ***índices basados en ordenaciones*** (*mediana* y *amplitud intercuartil*) se calculan con el orden que ocupan los valores cuando se ordenan de menor a mayor. Son índices generales que permiten describir correctamente variables cuantitativas. Su interpretación práctica es más fácil de entender sin necesidad de tener conocimientos estadísticos y utilizan las mismas unidades de medida que la variable. Para variables cuantitativas que no cumplen la normalidad, situación muy habitual en ciencias de la salud, y para variables ordinales, es necesario utilizar índices basados en ordenaciones, y no se deben utilizar índices basados en valores.

Comunicación de resultados

Los resultados de un estudio de investigación suelen comenzar con la tabla-resumen de estadística descriptiva y con los gráficos. Tanto las tablas como los gráficos deben ser simples, claros y autoexplicativos. En el caso de variables cuantitativas deben llevar las unidades de medida. En los estudios comparativos se presentan para diferentes grupos, como la que se muestra en la Tabla 4A.4, referida a un estudio sobre adolescentes diabéticos. Se presentan los grupos en columnas y las variables con los índices descriptivos en filas.

- Para ***variables categóricas*** se emplean el número absoluto (*n*) y su proporción en porcentaje (%), generalmente con un decimal. Es el caso de la variable *Sexo*.
- Para ***variables cuantitativas*** se emplea una medida de tendencia central y una de dispersión en función de la distribución de los datos.
  - Si se asume la normalidad y la simetría, se pueden utilizar la ***media*** y la ***desviación estándar***, como sucede con el Peso: M=56,3 kg, DE=8,2 kg en diabéticos y M=59,4 kg, DE=9,1 kg en controles (Tabla 4A.4). No se recomienda utilizar expresiones del tipo 56,3 ± 8,2 kg puesto que crean confusión.
  - Si no sigue la normalidad, se deben utilizar la ***mediana*** y la ***amplitud intercuartil***, como podría suceder en este ejemplo con la Glucemia: Md=130 mg/dl, AIC=112 a 131 mg/dl en diabéticos y Md=109 mg/dl, AIC=98 a 127 mg/dl en controles (Tabla 4A.4).
  - Incluso, en ocasiones excepcionales, en variables con poca dispersión, como la Edad en este estudio sobre adolescentes, sería incluso más explicativo utilizar las medidas más groseras, como la ***moda*** y el ***rango***: Mo=15 años, R=12 a 18 años en diabéticos y Mo=13 años, R=12 a 18 años en controles (Tabla 4A.4).

Tabla 4A.4. Resultados descriptivos de un estudio sobre adolescentes diabéticos.

	Diabéticos	**Controles**
Número de sujetos n	60	84
Sexo Masculino n (%)	26 (43%)	39 (46%)
Femenino n (%)	34 (57%)	45 (54%)
Peso (kg) Media (Desviación Estándar)	56,3 (8,2)	59,4 (9,1)
Edad (años) Moda (Rango)	15 (12 a 18)	13 (12 a 18)
Glucemia (mg/dl) Mediana (Amplitud Intercuartil)	130 (112 a 131)	109 (98 a 127)

REPRESENTACIÓN GRÁFICA DE VARIABLES CUANTITATIVAS

Son el Histograma, el diagrama de caja y el diagrama de tallo y hojas.

Histograma.
Es la representación gráfica de una variable cuantitativa basada en valores. En la Figura 4A.6 se muestran los histogramas de *Imc* y de *Glucosa*. Tiene la ventaja de que se pueden representar intervalos de clase de diferente

amplitud, aunque está muy extendido su uso con intervalos de igual amplitud. La superficie o área de la especie de "barra" definida por cada intervalo es proporcional a la frecuencia de los valores representados. Sus principales inconvenientes son la distorsión que produce la presencia de valores alejados y la diferente morfología que adopta al cambiar la amplitud de los intervalos.

Figura 4A.6. Histogramas de *Índice de masa corporal (kg/m^2)* y *Glucosa plasmática (mg/dl)*.

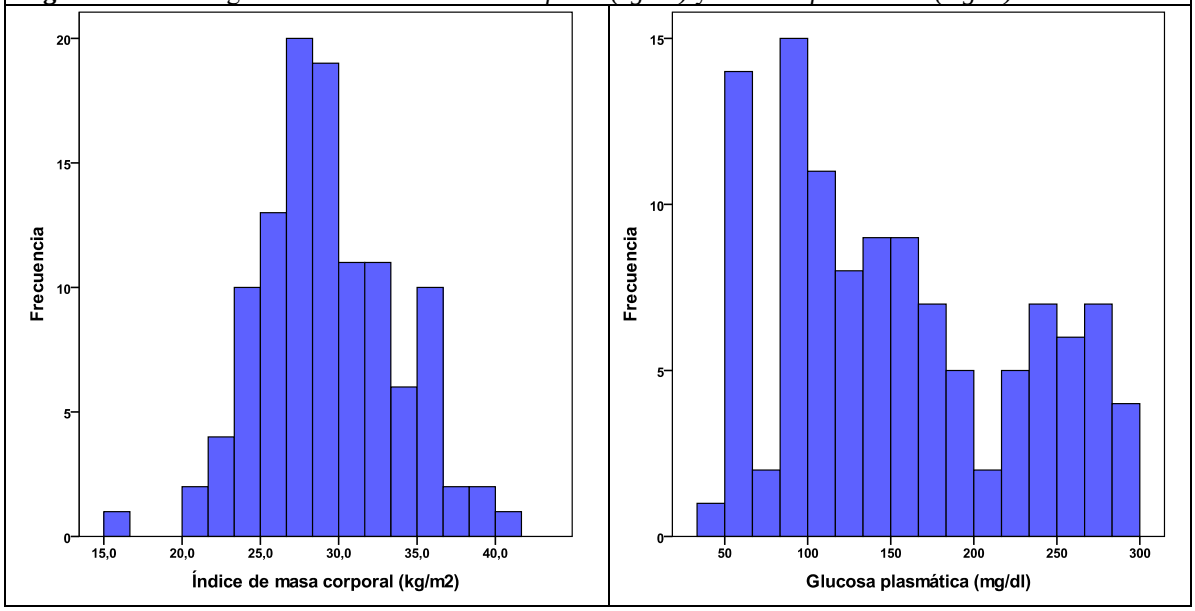


Diagrama de Caja o "**Boxplot**".

Es la representación gráfica de una variable cuantitativa basada en ordenaciones. Sintetiza muy bien la distribución de los valores y muestra claramente la tendencia central, dispersión, asimetría y valores alejados. Es muy útil para comparar varias variables o una misma variable entre diferentes grupos o a través del tiempo. En la Figura 4A.7 se muestran los diagramas de caja de *Imc* y de *Glucosa*. El trazo grueso horizontal central es la mediana y las dos bases de la caja son el primer y el tercer cuartil. La caja representa la amplitud intercuartil (AIC) y permite evaluar la asimetría central de la distribución. Los dos trazos horizontales más delgados son las patillas, bigotes o "whiskers". El superior es la cabeza y el inferior el pie. Corresponden a los valores máximo y mínimo de la distribución que no se consideran valores alejados y permiten evaluar la asimetría de las colas de la distribución. Los *valores alejados* o atípicos son los que se alejan del cuartil más próximo más de 1,5 veces la AIC. El programa distingue los valores alejados *exteriores* (que marca con un punto –o–) si se alejan del cuartil más próximo entre 1,5 y 3 veces la AIQ, de los valores alejados *extremos* (que marca con un asterisco –*–) si lo hacen más de 3 veces la AIQ). En el diagrama de caja de *Imc* hay un valor alejado exterior, el caso 92 con Imc de 15,5 kg/m^2, es el punto en la parte inferior del gráfico de la Figura 4A.7.

Figura 4A.7. Diagrama de caja de Índice de masa corporal (kg/m^2) y Glucosa plasmática (mg/dl).

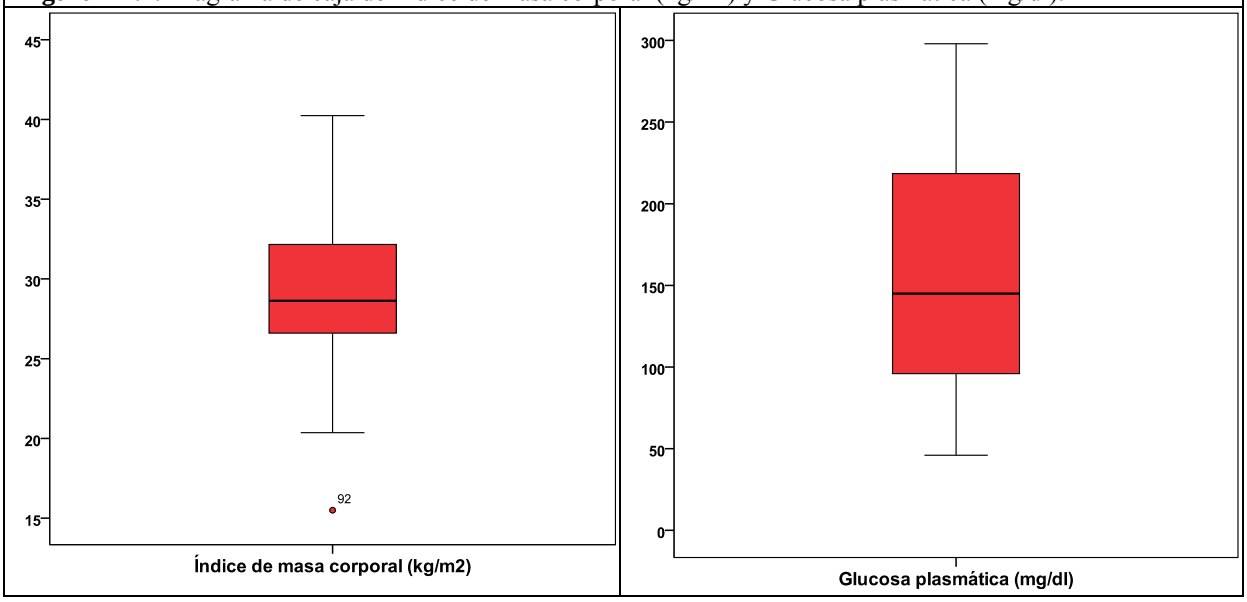


Diagrama de tallo y hojas (*stem and leaf plot*) de Tukey

Es una representación gráfica de una variable cuantitativa basada a la vez en valores y ordenaciones. Se forma descomponiendo cada valor en dos partes: el último dígito de la derecha (que es la hoja) y las cifras restantes de la izquierda (que es el tallo). En Figura 4A.8 se ofrecen los diagramas de Tallo y hojas de *Imc* y *Glucosa*. En el de *Imc* cada Tallo es la mitad de cada decena; la mitad inferior tiene las hojas 0-1-2-3-4, y la mitad superior las hojas 5-6-7-8-9. Hay un caso (Frecuencia 1) con valor alejado (Tallo) que es ≤ 15, hay 2 casos con valor 20 (Tallo=2 y Hola=0), 7 casos con *Imc* de 22 ó 23 (Tallo=2 y Hoja=2 en 3 casos −22− y Hoja=3 en 4 casos −23−) y así sucesivamente. Se van ordenando los datos originales en intervalos. Ofrece simultáneamente una tabla (se muestra la distribución de frecuencias) y una representación gráfica (una imagen de la forma general de la distribución similar a un histograma "tumbado" horizontal) de la variable. Permite localizar fácilmente las medidas de posición y de orden (mediana, cuartiles y percentiles), los posibles valores alejados, las concentraciones de datos y la existencia de lagunas. Es una representación gráfica (a la vez que tabla) bastante completa que, sin embargo, no se utiliza tanto como el clásico histograma y el novedoso diagrama de caja.

Figura 4A.8. Diagrama de Tallo y hojas de *Índice de masa corporal (kg/m^2)* y *Glucosa plasmática (mg/dl)*.

Índice de masa corporal (kg/m^2)			**Glucosa plasmática (mg/dl)**		
Frecuencia	Tallo	Hojas	Frecuencia	Tallo	Hojas
1	Alejados	≤ 15	9	0	455555555
2	2	00	6	0	666666
7	2	2223333	15	0	888889999999999
13	2	4444444555555	13	1	0000000111111
23	2	66666666666777777777777	11	1	22222333333
23	2	88888888888888999999999	11	1	44444455555
14	2	00000000111111	10	1	6666666777
9	2	222233333	6	1	888999
12	2	444445555555	4	2	0011
5	2	66667	3	2	223
2	2	89	11	2	44444455555
1	2	0	8	2	66667777
			5	2	88889

Videotutorial 4A2Explorar1.avi

Se muestra como se obtienen la estadística descriptiva y los gráficos (histogramas, diagrama de cajas y de tallo y hojas) de las variables *Imc* y *Glucosa* con el cuadro *Explorar*, que ofrece una estadística descriptiva bastante completa. La Moda, cualquier n−til, por ejemplo los Terciles y el Percentil 95 y los Histogramas con la curva normal, se obtienen con el cuadro *Frecuencias*.