Universidad de Valladolid

Facultad de Ciencias Económicas y Empresariales

Trabajo de Fin de Grado

Grado en Marketing e Investigación de Mercados

UNA APROXIMACIÓN AL ANÁLISIS EXPLORATORIO DE DATOS

Presentado por:

Natalia Ruiz Fuentes

Tutelado por:

Jose Antonio Sanz Gómez

Valladolid, 1 de Abril de 2022

RESUMEN

Las técnicas de análisis gráfico del AED propuestas por Tukey (1977) permiten identificar la distribución de los datos y explorar si las variables objeto de estudio cumplen los requisitos de normalidad, simetría y unimodalidad. Se aplican seis técnicas principales a cuatro conjuntos de datos cuantitativos: PIB per cápita, PIB provincial, edad media y esperanza de vida de países de la OMS. Los resultados muestran que la utilización de las técnicas ha de hacerse siguiendo un orden: un análisis inicial haciendo uso del diagrama de tallo y hojas junto con el histograma para observar la normalidad, seguido del diagrama de caja y bigotes y ver tanto la simetría como los valores atípicos. Por último, el diagrama de dispersión ofrece una idea de las relaciones existentes entre las variables y su ajuste.

PALABRAS CLAVE: AED, Tukey, Análisis gráfico

ABSTRACT

EDA graphical tools developed by Tukey (1977) let identify and explore data distribution and the nature of variables in terms of normality, symmetry and unimodality. Graphical tools are applied to four datasets: GDP, provincial GDP, average age and life expectancy (OMS countries). Results show that in order to apply these techniques a proper order should be used; thus, an initial exploration making use of stem-and-leaf diagram and histogram to observe normality, followed by a box-and-whisker plot to determine symmetry and outliers; lastly, a dispersion analysis to conclude existential relations between variables and their adjustment.

KEY WORDS: *EDA, Tukey, Graphic analysis*

Códigos de clasificación del JEL: C10, 181, 182

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS Y FIGURAS PAGINADOS

ÍNDICE DE ANEXOS

1. INTRODUCCIÓN

El trabajo presente busca estudiar y extraer conclusiones de varios conjuntos de datos mediante el uso técnicas propias del Análisis Exploratorio de Datos,AED. Como su propio nombre indica, se considera exploratorio al ser uno de los primeros pasos que se realizan antes de lanzarse a trabajar con los datos. Dicha exploración se realiza esencialmente a través de técnicas que suponen herramientas visuales, gráficas y semi gráficas, y el uso de las mismas ayudan al investigador a:

- Conocer la estructura y distribución de los datos previamente a aplicar cualquier técnica estadística.
- Estudiar la relación entre variables explicativas.
- Dar con posibles errores y puntos extremos como anomalías.
- Refinar nuestras hipótesis.

De esta forma, se buscará dar respuesta a preguntas como:

- ¿Existe algún tipo de estructura (normalidad, multimodalidad, asimetría, homogeneidad entre grupos) en los datos que voy a analizar?
- ¿Cómo se sintetiza y presenta la información contenida en un conjunto de datos?
- ¿Existen datos atípicos? ¿Cuáles son? ¿Hay datos ausentes? ¿Tienen algún patrón sistemático?

Las técnicas propias del A.E.D., han sido utilizadas extensamente desde la publicación por primera vez de las mismas, en 1977, por John W. Tuckey, quién ensalzó la fase necesaria en cualquier investigación, la fase exploratoria de los datos. Sin embargo, una gran mayoría de los estudios e investigaciones comienza directamente el análisis de sus conjuntos de datos aplicando técnicas propias de la estadística descriptiva clásica, a través de los resúmenes descriptivos; Se tiene constancia de diversos autores que previamente han querido destacar la importancia de esta fase exploratoria y de su diferenciación de la fase de estadística descriptiva. Algunos trabajos destacables y de referencia son los de Monterde y Perea Lara, M (1991).

2. MARCO TEÓRICO

En ciencias sociales, en lo relativo a la ciencia estadística, existen multitud de técnicas estadísticas que se han ido empleado a lo largo del tiempo. Tukey (1977) propone un nuevo enfoque en su libro pionero Exploratory Data Analysis, lleno de nuevos procedimientos estadísticos y retoma otros. J. W. Tuckey nace en 1915 (Massachusetts) y fallece en 2000 (Nueva Jersey). Doctor en matemáticas con intereses en el ámbito estadístico, con fascinación por problemas y técnicas relativas al análisis de datos. Se le conoce por ser el pionero de este análisis pero también hizo contribuciones en el campo de la estadística en el análisis de la varianza, regresión y en un amplio rango de aplicaciones. En cierta manera él entendió que para enfrentarse a cualquier problema en el área de las matemáticas, tendría que entender de qué estaban compuestos los conjuntos de datos de los que partían sus investigaciones.

2.2. Estadística robusta en Análisis Exploratorio de Datos.

El A.E.D. presupone una actitud activa del investigador hacia el análisis como un medio para sugerir nuevas hipótesis de trabajo. Difiere de la Estadística Descriptiva clásica, con el fin de optimizar la cantidad de información de los datos recogidos, a través del uso de novedosas representaciones gráficas, a base de reducir la influencia de las puntuaciones extremas de los estadísticos a través del empleo de estadísticos resistentes. Depende de lo que vayamos encontrando, la experiencia y el conocimiento específico se utilizaran unas técnicas u otras para las que existen herramientas específicas, básicamente gráficas y no presupone normalidad (no considera de interés la media, varianza, regresión o correlación). En su lugar, además de los gráficos, usa mediana, recorrido intercuartílico. El uso de estadísticos robustos (resistentes) es muy aconsejable cuando los datos no se ajustan a una distribución normal. Estos estadísticos son los que se ven poco afectados por valores atípicos. Suelen estar basados en la mediana y en los cuartiles y son de fácil cálculo. Fruto del análisis exploratorio, a veces es necesario transformar las variables. (Pérez López, 2004).

3. ESTUDIO DE POBLACIONES INDIVIDUALES

El objetivo de este capítulo es estudiar por separado cada una las herramientas específicas del A.E.D. Para proceder con el análisis las variables objeto de estudio están compuestas por un conjunto de datos de carácter cuantitativo. Se abordarán todas ellas comenzando por una previa descripción de qué son y cómo se construyen, para, posteriormente, analizarlas utilizando el software STATGRAPHICS y ver particularidades pertinentes a cada una de ellas.

Las representaciones gráficas propuestas por Tukey (1977) proporcionan la información visual necesaria para obtener los estadísticos que ofrecen un resumen acerca de cómo se distribuyen los datos. Estos estadísticos se estudiarán en cada uno de los apartados siguientes para caracterizar las distribuciones y se pueden clasificar en cuanto a:

1. Características de posición o tendencia central: la mediana (en algún caso la moda). Es el valor alrededor del cual se agrupan los datos.
2. Características de dispersión: recorrido intercuartílico. Nos proporciona una medida de la desviación de los datos con respecto a la tendencia central.
3. Características de forma: simetría y asimetría. Nos proporcionan una medida de la forma gráfica de la distribución.

Las técnicas propias del AED suponen mirar y analizar el conjunto de datos de una forma peculiar que difiere de las técnicas de la estadística clásica convencional. Este enfoque permite mirar al conjunto de datos de tal manera que se pueda no sólo caracterizar la distribución en base a los estadísticos mencionados, sino que ofrece la ventaja de explorar esos datos en profundidad remarcando diferentes “problemas” que presenta dicha distribución para, posteriormente en la fase confirmatoria, aplicar las técnicas y estrategias pertinentes para solventarlos. Estos problemas que se observan en las representaciones visuales de la fase exploratoria son: puntos atípicos, asimetría de la distribución, agujeros y picos múltiples.

3.1. Diagrama de Tallo y Hojas (DTH).

Es la representación semi-gráfica de una variable en forma de Tabla (líneas y columnas), donde se muestra el rango y la distribución de los datos, la simetría y los candidatos a valores atípicos. Su uso es recomendable siempre que el número de datos no sea muy grande (alrededor de 50 observaciones). A diferencia de las otras técnicas, el DTH permite organizar gráfica y numéricamente a los valores para conseguir una inspección visual detallada de su distribución (Emerson, J. D., Hoaglin, D. C., 1983). Se podría decir que el DTH es la versión “mejorada” del histograma, ya que volcado en su costado no sólo conforma un histograma y mantiene las mismas características observables respecto al intervalo y la densidad, sino que además añade información relativa a cada una de las observaciones.

Tabla 1: Tabla que muestra la relación entre histograma y diagrama de tallo y hojas.

Histograma	Diagrama de tallo y hojas
No se pueden conocer los datos originales.	Preserva los valores de los datos: al observar el gráfico pueden ubicarse los diferentes valores de las observaciones y con ello reconstruir la Tabla de datos
Estimador de la densidad	Optimizar representación gráfica de los datos
Representación de intervalos.	Representación de: intervalos, gráfica de cada valor, profundidades, mediana, valores mínimos y máximos, recorrido intercuartílico.
Complejo de construir	Fácil de construir
No se puede calcular	Permite cálculo de la mediana y el recorrido intercuartílico

Fuente: realización propia.

Para construir el diagrama de tallo y hojas, con el fin de estudiar el PIB per cápita provincial en miles de euros, en España, se recogen las columnas correspondientes a las provincias y al PIB per cápita del Anexo 1: Provincias de España por PIB, y se ordenan las observaciones (52) de menor a mayor. A continuación hay que realizar los cálculos presentados en la siguiente Tabla:

Tabla 2: Cálculos para el diagrama de tallo y hojas.

Decisiones	Cálculo	Ejemplo: PIB per cápita (miles €)
Unidad a expresar[1]	Su cálculo no es una ecuación concreta. Dependerá del conjunto de datos y de la decisión del investigador.	En este caso, con valores expresados en miles de € sin decimales, se eligen los dos primeros dígitos: Valladolid, 26.901 (miles€). Se elige representarlo con dos dígitos para 26.000 (miles€), resultando una unidad de 26.000/26=1000,0.
Rango de los datos	$R = Máxx - Mínx$	$R = 36.404 - 18.050 = 18.354$
Número máximo de filas o intervalos	$Fmax = 10 * log10N$=10*log520=27,1600	
Amplitud máxima del intervalo	$a = (unidad/Fmax) * 10^{dígitos\ hojas}$	$\frac{1000}{27,160} \times 10^1 = 3681,885$

Fuente: realización propia.

A continuación se presenta el diagrama de tallo y hojas para el PIB per cápita provincial (Anexo 1) obtenido mediante STATGRAPHICS:

Figura 1. Diagrama de tallo y hojas. Método de análisis de una variable para el PIB per cápita provincial. Unidad = 1000,0.

Stem-and-Leaf Display for PIB: unit = 1000,0 1|2 represents 12000,0

```
 12     1|888888889999
 24     2|000000111111
 (9)    2|222223333
 19     2|5
 18     2|66777
 13     2|888889
  7     3|0011
  3     3|3
  2     3|5
  1     3|6
```

Fuente: obtenido mediante STATGRAPHICS (Versión 19).

El diagrama de la Figura 1 muestra que el rango de los datos ha sido dividido en 10 intervalos (tallos), cada uno de ellos representado por una línea de la Tabla. Los tallos se han decidido usando un dígito proveniente de los valores del conjunto de datos. En cada línea se representan los valores individuales del conjunto de datos con un dígito (hoja) a la derecha de la línea vertical (tronco). El DTH de la Figura 1 permite afirmar que la variable PIB per cápita provincial en miles de euros varía entre 18 y 36, y que el 50% central (mediana) de las provincias cuentan con un PIB

[1] Se divide cualquiera de los datos por el valor resultante de no tener en cuenta los decimales o bien la representación gráfica que vayamos a hacer de dicho valor en la gráfica (Monterde y Perea Lara, M.,1991). Una vez que tenemos varios datos a los que mirar, normalmente ayuda hacer el conjunto de datos lo más entendible posible respecto a de lo que partimos. Por ejemplo, una población descrita como 201 millones, o 201.2, proporciona más claridad a la hora de observar el dato que si se diese como 201,234,567 (Tukey, 1977).

de entre 22 y 23 (línea 3, entre paréntesis). La distribución del PIB per cápita provincial es asimétrica hacia la derecha, ya que los datos están concentrados a la izquierda en las tres primeras líneas, y bimodal, ya que presenta dos picos observables. Esto indica que los datos no se distribuyen normalmente, ya que la mayoría de las provincias cuentan con un PIB per cápita relativamente bajo. No se presentan valores extremos altos o bajos (separados por el resto de líneas). No sugiere casos extraordinarios.

Extraer los valores originales para reconstruir la Tabla

El PIB per cápita de la primera provincia es de 18000,0 (miles €), de los que llamaremos al primer dígito ("1") el tallo, y al segundo dígito ("8") la hoja. Así como, por ejemplo, la fila que muestra 1 | 888888889999 indica que hay 8 provincias con un PIB per cápita de 18.000 (miles €) y 4 provincias con un PIB de 19.000 (miles €).

Relación entre simetría y normalidad

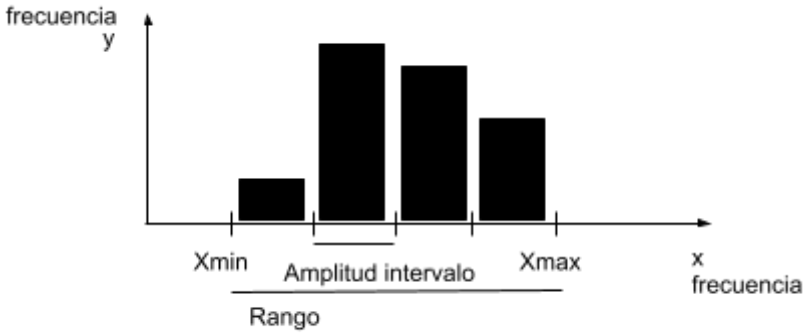
Los datos asimétricos indican que los datos podrían ser no normales. Cuando los datos son asimétricos, la mayoría de los datos se ubican en la parte superior o inferior de la gráfica.

3.2. Histograma (H).

Un histograma es una representación gráfica de una variable que se utiliza para examinar la forma y dispersión de los datos. Divide los valores de la muestra en intervalos y representa la densidad de los valores de datos en cada intervalo con una barra, es decir, muestra cómo se distribuyen los valores de una variable cuantitativa cuando ésta se divide en dichos intervalos uniformes. Los rectángulos[2] del histograma no se encuentran igualmente espaciados y la posición corresponde a la localización del intervalo sobre el eje X y la altura indica la frecuencia dentro de cada intervalo. El histograma permite detectar valores extremos, características de simetría de la distribución o presencia de varias modas.

[2] No se recomienda que las bases de los rectángulos tengan diferente tamaño ya que esto dificulta la interpretación.

Figura 2: Representación gráfica de un histograma.



Fuente: realización propia.

Este diagrama presenta algunas limitaciones; no se considera el tiempo en el que se obtuvieron los datos y la cantidad de clases influye en la forma del histograma. “La construcción de un histograma cumple generalmente dos propósitos. En primer lugar, es utilizado para representar datos, con el fin de buscar patrones, tendencias y eventualmente, algún número o números que resuma los datos. El segundo propósito de un histograma es servir como un estimador no paramétrico de la distribución subyacente. En ambos casos el, aparentemente, simple proceso de construir un histograma plantea, la mayoría de las veces, problemas de solución no inmediata.” (Capítulo 2. Gráficos Stem-and-Leaf frente a histogramas, Conde, J.E., Rull, V. y Vegas, T.)

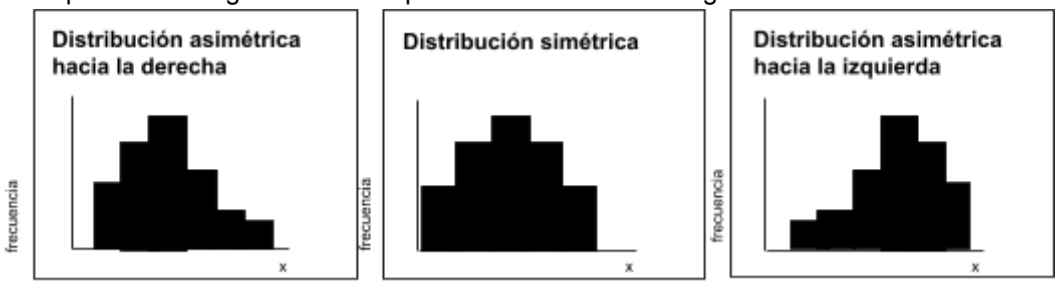
Debido a la estrecha relación que presenta con el DTH, se pueden extraer conclusiones comunes de ambos, para lo cual éstos han de estar construidos de la misma forma partiendo de un conjunto de datos idéntico, es decir, teniendo en cuenta que el ancho del intervalo del H sea igual a la unidad (escala) del DTH.

Simetría en un histograma

En un histograma la simetría hace referencia al grado en que los valores de la variable, equidistantes a un valor que se considere centro de la distribución (mediana), poseen frecuencias similares. A nivel visual se observa la representación gráfica del histograma dónde la distribución será simétrica si la mitad izquierda de la distribución de la imagen es el reflejo de la mitad derecha. En distribuciones

simétricas media y mediana coinciden. Si sólo hay una moda (distribución unimodal), el valor de ésta también será igual a las dos anteriores. El nivel de simetría se suele describir de acuerdo a tres categorías:

Figura 3: Representación gráfica de los tipos de asimetría en histogramas.



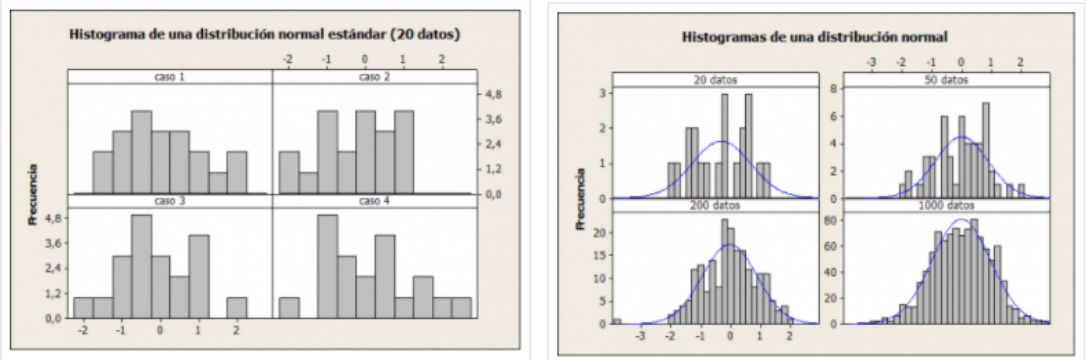
Fuente: realización propia.

- **Distribuciones simétricas:** dispersión es igual o muy similar a ambos lados.
- **Distribuciones asimétricas hacia la derecha:** está caracterizada por tener la cola más dispersa en el lado de los valores altos de la variable.
- **Distribuciones asimétricas hacia la izquierda:** más disperso al lado de los valores más bajos de la variable.

Normalidad en un histograma

Se observa la curva conocida como campana de Gauss para determinar si el conjunto de datos se ajusta a una distribución normal. Éste método sirve cuando hay muchos datos. Si es pequeño, puede llevar a extraer conclusiones erróneas ya que puede no ajustarse a la campana de Gauss y seguir una distribución normal. La Figura 4 afirma que al aumentar el número de datos, la forma del histograma se ajusta a la campana de Gauss:

Figura 4: Histogramas realizados con unos datos procedentes de una distribución normal estándar pero variando la cantidad de datos.



Fuente: https://www.caletec.com/6sigma/histogramas-y-normalidad-de-los-datos/.

Para construir el histograma, con el fin de estudiar la esperanza de vida general de los países de la OMS, se recogen los datos de la columna “Esperanza de vida” correspondiente al Anexo 2: países de la OMS por esperanza de vida. Para decidir las clases hay que realizar los cálculos de la Tabla 3. Su representación se lleva a cabo en un eje cartesiano de coordenadas (X,Y), siendo el eje de abscisas la representación del rango y el eje de ordenadas representa la frecuencia.

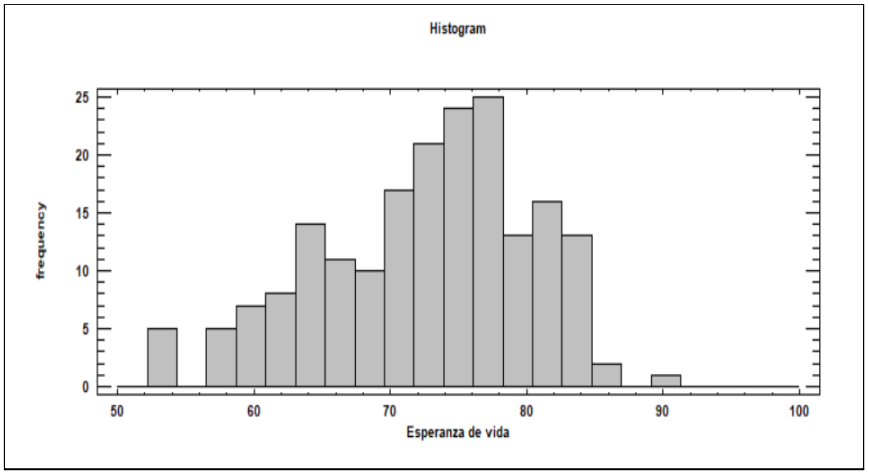
Tabla 3: Cálculos para decidir las clases de un histograma.

Decisiones	Cálculo	Ejemplo: países de la OMS por esperanza de vida
Rango de los datos	$R = Máxx - Mínx$	$R = 90 - 52,81 = 37,19$
Número de clases	$K = \sqrt{N}$	$K = \sqrt{192} = 13,86$
Amplitud o ancho del intervalo	$h = rango/K$	$h = 37,19/13,86 = 2,68$

Fuente: realización propia.

A continuación en la Figura 5 se observa el proceso de construcción del histograma mencionado anteriormente, de donde se puede detectar el rango, el número de clases, la amplitud del intervalo y la definición de dichas clases. En el eje X se observa el rango desde el valor máximo (90) al valor mínimo (52,81), con una amplitud del intervalo de dos unidades, y un número de clases de 14.

Figura 5: Histograma obtenido mediante el método de análisis de una variable en STATGRAPHICS para la variable esperanza de vida de los países de la OMS.



Fuente: obtenido mediante STATGRAPHICS (Versión 19).

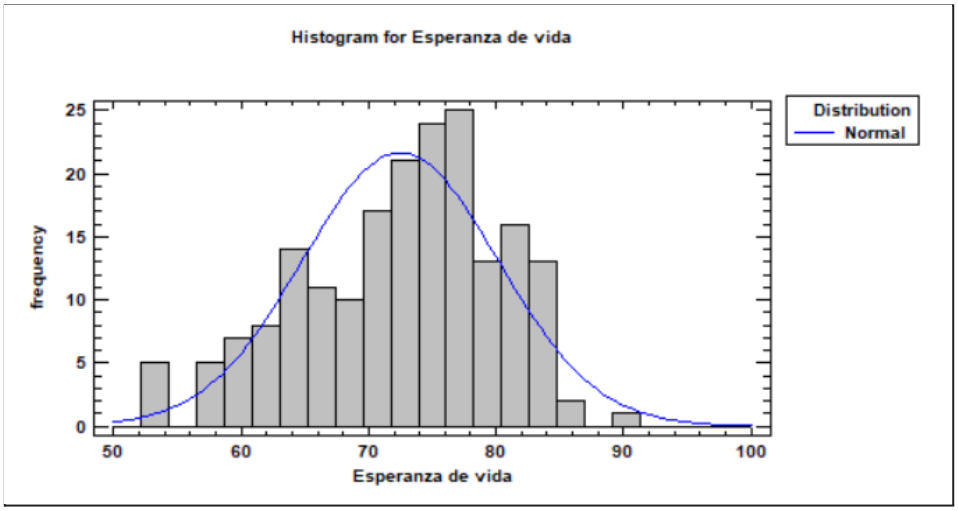
Tabla 4: Características de distribución para la Figura 5.

Agujeros en los intervalos	de 54,81 a 56,81 años y 86,81 a 88,81 años
Datos atípicos en el intervalo	de 52,81 a 54,81 años y de 88,81 a 90,81 años.
Distribución unimodal	la existencia de un sólo pico (moda).

Fuente: Realización propia.

El conjunto de datos analizado para la esperanza de vida general sigue una distribución normal (Figura 6), ya que se ajusta la campana de Gauss, comentada anteriormente.

Figura 6: Ajuste de la distribución para el histograma de esperanza de vida.

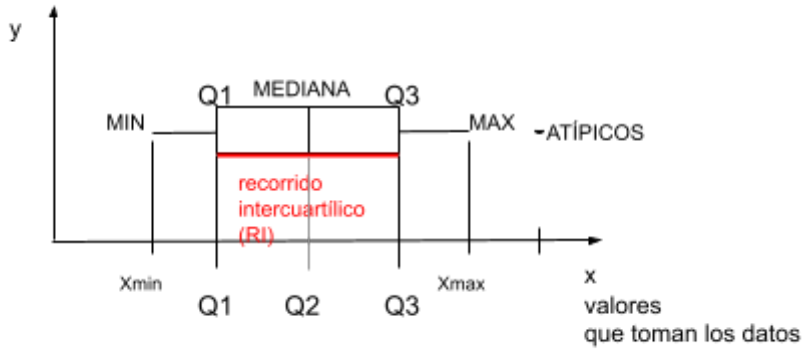


Fuente: obtenido mediante STATGRAPHICS (Versión 19).

3.3. Diagrama de Caja y Bigotes (DCB).

Este diagrama consiste en una representación gráfica de una variable que se utiliza para examinar la distribución, localización y dispersión de los datos. Compone un esquema que grafica los estadísticos de posición

Figura 7: Representación gráfica de los estadísticos de posición presentes en un diagrama de caja y bigotes para un eje de coordenadas (X,Y).



Fuente: Realización propia.

Para su construcción, se utilizan dichos estadísticos de posición, formados por tres estadísticos de la distribución de frecuencias: el primer cuartil Q1, la mediana y el tercer cuartil Q3. Una vez obtenidos, se sitúan en un eje cartesiano de coordenadas (X,Y). Con el objetivo de obtener los estadísticos, se observa la columna correspondiente al PIB per cápita provincial del conjunto de datos contenido en el Anexo 1: Provincias de España por PIB.

Tabla 5: Cálculos para el diagrama de caja y bigotes.

Decisiones	Cálculo	Ejemplo: PIB per cápita (miles €)
Primer cuartil (Q1)	$Q1 = N\ x\ ¼$	$Q1 = 52\ x\ ¼ = 13$ (Equivale a un PIB de 20.251)
Mediana[3] (Q2)	$m = (N + 1)/2$	$m = (52 + 1)/2 = 26,5$ (Equivale a un PIB de 22.339)
Tercer cuartil (Q3)	$Q3 = N\ x\ ¾$	$Q3 = 52\ x\ ¾ = 39$ (Equivale a un PIB de 27.870)

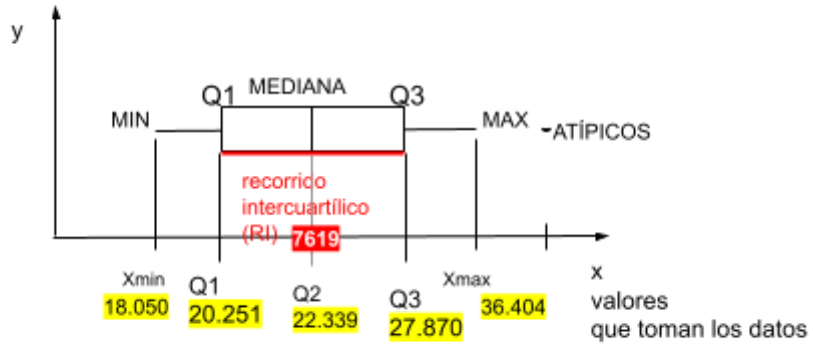
[3] En este caso, al ser un conjunto de observaciones pares, se cogen los dos centrales teniendo en cuenta m = 26,5, equivalente al id = 26,5, dónde en este caso sería el id = 26 y el id = 27, que se corresponden a Segovia y Pontevedra, con 22.212 y 22.586 (PIB per cápita en miles de euros). Por lo tanto, la mediana sería la suma de ambos valores entre 2, que sería igual a 22.339 millones de euros. Si N fuese impar, se coge el valor central directamente m = (N+1)/2.

Recorrido Intercuartílico	$RI = Q3 - Q1$	$RI = 27.870 - 20.251 = 7619$
Anchura de los bigotes	$a = 1,5 \times RI$	$a = 1,5 \times 7619 = 11.428,5$
Valores atípicos	$f1 = Q1 - 1,5RI$ $f3 = Q3 + 1,5RI$	$f1 = 20.251 - 1,5 \times 7619 = 8.822,5$ $f3 = 27870 + 1,5 \times 7619 = 16.441,5$

Fuente: realización propia.

Según los valores determinados, el gráfico de caja y bigotes se representa en un eje de coordenadas (X,Y) de la siguiente forma:

Figura 8: Representación del diagrama de caja y bigotes para el PIB per cápita provincial, siguiendo el esquema de la Figura 7.



Fuente: Realización propia.

En la Figura 8 se observan diversas características. La caja es rectangular y sus extremos equivalen a los cuartiles Q1 Y Q3, conteniendo la mitad del conjunto de datos. La línea central dentro de la caja muestra el valor que toma la mediana. Los bigotes se construyen tomando 1.5 veces RI (recorrido intercuartílico), a la izquierda y derecha de la caja, y se limitan por el último valor incluido en dichos intervalos. Los valores no incluidos ni en la caja ni en los bigotes se representan individualmente mediante un punto, y se consideran datos atípicos. Tanto la escala (el tamaño del diagrama) como la altura de la caja se deciden arbitrariamente.

Tabla 6: Cálculos para el diagrama de caja y bigotes.

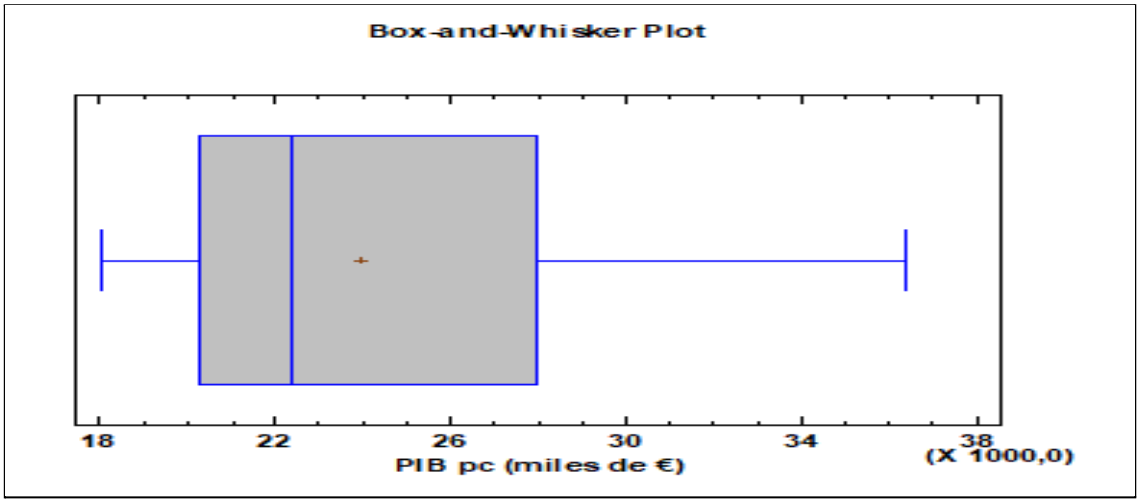
Tamaño de la caja (con respecto al resto del gráfico)	Nos informa de la variabilidad o dispersión
Caja	Contiene el 50% de los valores.
Punto mínimo y máximo	El valor más pequeño y el más grande del conjunto de datos sin

	tener en cuenta valores atípicos.
Puntos fuera	Valores atípicos.
Posición de la mediana	Dónde se localizan los valores centrales en el interior de la caja, hacia la izquierda o la derecha, y de la simetría de la distribución.
Posición de la media	El valor que identifica la media de la distribución

Fuente: realización propia.

En el siguiente ejemplo, se analiza el PIB per cápita provincial (columna PIBpc (miles de €), Anexo 1) obtenido mediante STATGRAPHICS:

Figura 9: Ventana de salida para el análisis de una variable para el PIB per cápita (miles de €).



Fuente: obtenido mediante STATGRAPHICS (Versión 19).

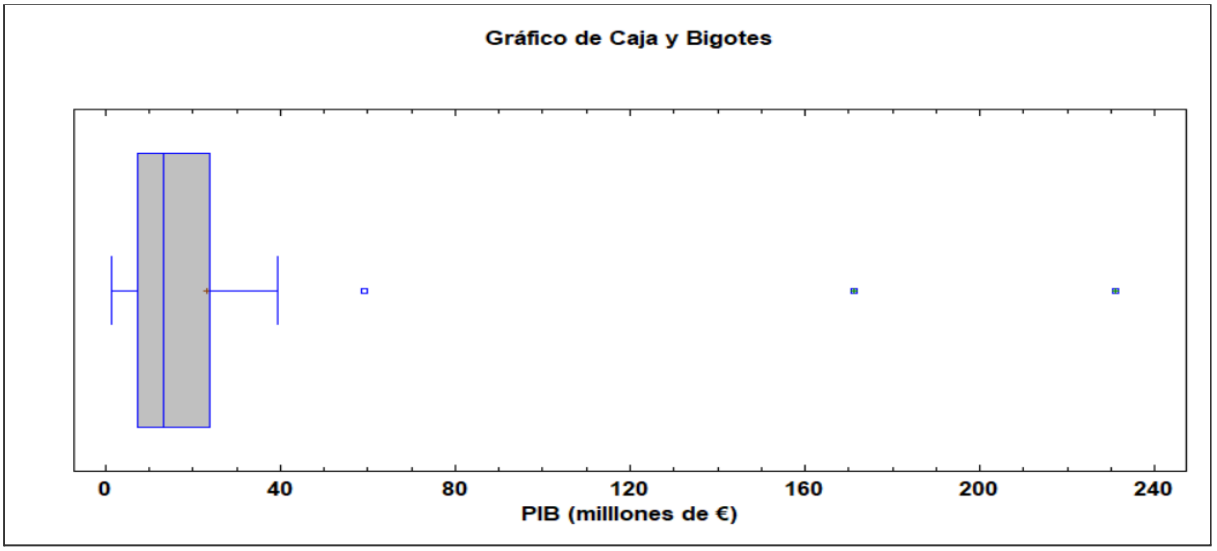
El diagrama de la Figura 9 muestra la variable representada en el eje de abscisas, conteniendo los dos dígitos principales del PIB per cápita provincial (unidad 1000,0). El rango de los datos ha sido calculado y representado de extremo a extremo de los bigotes, dónde no ha aparecido ningún valor atípico y por lo tanto no se observan puntos fuera de los mismos. La caja se ha representado de tal forma que contiene el 50% de los datos contenidos en la variable, y su división corresponde a la posición de la mediana.

Analizando el diagrama anterior se puede afirmar que la variable PIB per cápita en miles de euros para las provincias españolas varía entre 18 y 36, y que el 50% central de estas provincias tiene un PIB per cápita entre 20 (Q1) y 28(Q3), sin

presentar ningún valor atípico, ya que no aparecen puntos fuera de los bigotes. La distribución de X para las provincias es asimétrica hacia la derecha, ya que la zona de la derecha en el área central de la Figura es mayor que la de la izquierda, y la mediana corresponde aproximadamente al valor 22,5 de X, siendo la media 24 aproximadamente.

A continuación, se presenta el DCB para el PIB provincial (columna PIB (miles €), Anexo 1) obtenido mediante STATGRAPHICS:

Figura 10: Ventana de salida para el análisis de una variable para el PIB (miles de €).



Fuente: Obtenido mediante STATGRAPHICS (Versión 19).

La Figura 10 muestra la variable representada en el eje de abscisas, conteniendo los dos dígitos principales del PIB provincial en millones de euros. El rango de los datos ha sido calculado y representado de extremo a extremo de los bigotes (desde 1 millón hasta 231 millones), dónde han aparecido tres valores atípicos que se observan por los cuadrados fuera de los mismos (PIB provincial en millones de euros para Barcelona (171), Madrid (231) y Valencia (59). La caja se ha representado de tal forma que contiene el 50% de los datos contenidos en la variable (entre 7 y 23 millones de euros) que corresponden al cuartil 1 y cuartil 3, y su división corresponde a la posición de la mediana (13 millones de euros). La media aparece representada con una cruz roja, casi coincidente con el Q3 (23 millones de euros).

Se puede afirmar por tanto que el PIB provincial comprendido entre el 50% y el 75% cuentan con un PIB más disperso que el comprendido entre el 25% y el 50% ya que la caja es mayor a la derecha de la mediana. El bigote de la izquierda es más corto que el de la derecha, por ello el 25% de la población con un PIB provincial más bajo presenta mayor concentración que el 25% de los que tienen un PIB más alto.

3.4. Gráfico de Simetría (GS).

En estadística, un gráfico de simetría es una representación gráfica de una variable que se utiliza para visualizar si los datos de la muestra provienen de una distribución simétrica. Una distribución será simétrica si los datos en ambos lados de la mediana están distribuidos de la misma manera (si las colas de la distribución son imágenes exactamente iguales). La Figura 11 representa las tres posibilidades de simetría que puede seguir una distribución:

Figura 11: Representación gráfica de los tipos de simetría en un gráfico de simetría

Fuente: Realización propia.

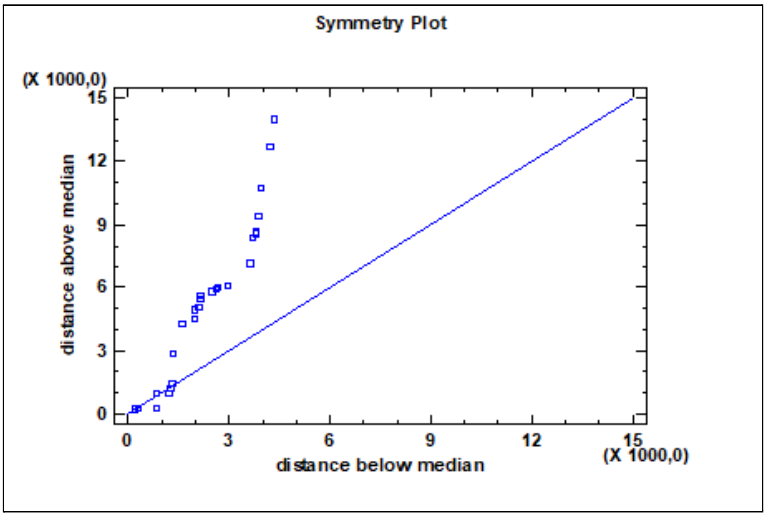
- **Datos asimétricos(derecha):** más valores distintos a la derecha de la media.
- **Datos simétricos:** los datos se distribuyen normalmente.
- **Datos asimétricos(izquierda):** más valores distintos a la izquierda.

Para su construcción, se ilustra la distancia superior de la mediana en el eje X versus la distancia inferior hasta la mediana en el eje Y, para cada uno de los puntos

de datos. Se coloca un punto en el gráfico que muestra su situación y respectiva distancia de la mediana. La línea de referencia en la gráfica es la bisectriz del primer cuadrante, que representa un conjunto de datos perfectamente simétricos.

La Figura 12 presenta el gráfico de simetría para el PIB per cápita provincial en miles de euros (Anexo 1) obtenido mediante STATGRAPHICS. El diagrama se ha construido cogiendo los datos correspondientes y se han expresado con una unidad de 1000,0. El diagrama muestra una distribución que no es simétrica, ya que los puntos no están situados cercanos a la línea recta de referencia (bisectriz del primer cuadrante) o sobre ella, que indicaría que los datos se distribuyen normalmente. Los puntos se encuentran alejándose de ésta hacia la izquierda, conformando una distribución asimétrica a la izquierda. Implica que hay asimetría negativa, es decir, hay más valores distintos a la izquierda de la mediana.

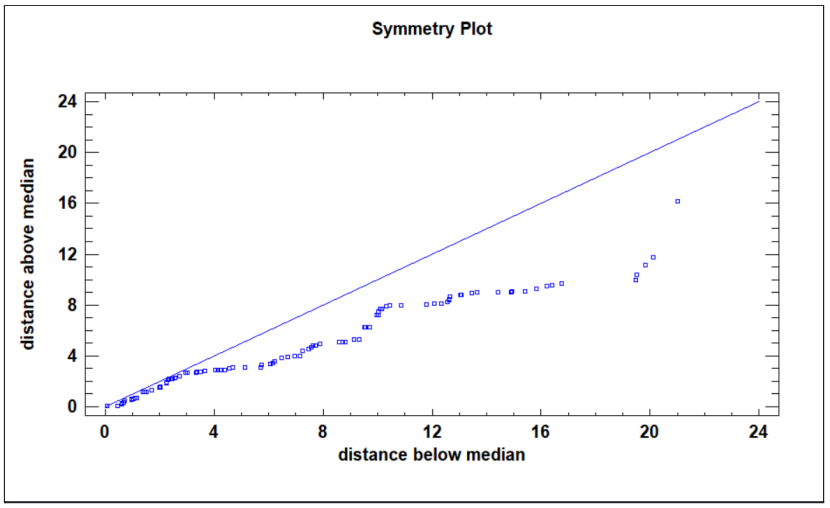
Figura 12: Gráfico de simetría. Método análisis de una variable para el PIB per cápita (miles de €).



Fuente: Obtenido mediante STATGRAPHICS (Versión 19).

La Figura 13 muestra un ejemplo de asimetría hacia la derecha. Se presenta el diagrama de simetría para la esperanza de vida de los países de la OMS (Anexo 2: países de la OMS por esperanza de vida) obtenido mediante STATGRAPHICS.

Figura 13: Gráfico de simetría. Método análisis de una variable para la esperanza de vida de los países de la OMS.



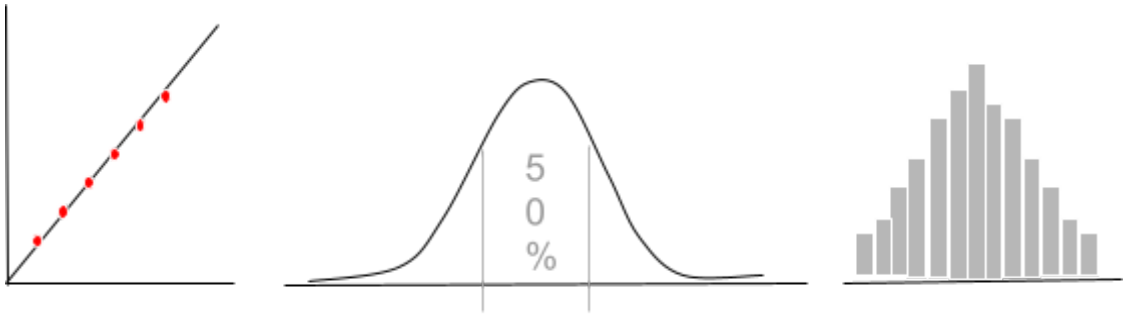
Fuente: Obtenido mediante STATGRAPHICS (Versión 19).

La distribución está caracterizada por estar compuesta por datos asimétricos hacia la derecha, lo cual implica asimetría positiva, es decir, se encuentran más valores distintos a la derecha de la mediana.

Simetría, probabilidad e histograma para una población normal

Una población normal se ajusta a la forma de una campana de Gauss, conteniendo el 50% de los datos a la izquierda y derecha de la media. En un gráfico de simetría se observa cuando los puntos se ajustan a la línea de referencia como se ha mencionado anteriormente.

Figura 14: Representación que muestra la equivalencia en diferentes tipos de gráficos de representar una población normal.



Fuente: Realización propia.

En la Figura 14 se muestra el gráfico de simetría con los puntos situados sobre el eje, dónde se observa el equivalente a representar una población que se distribuye

normalmente. A su vez, se muestra el gráfico que muestra la probabilidad bajo la cual el 50% de los datos se contienen entre las líneas respectivas, gráfico que emula la campana de Gauss. Además, a la derecha se presenta un histograma con una distribución normal.

3.5. Gráfico de Normalidad (GN).

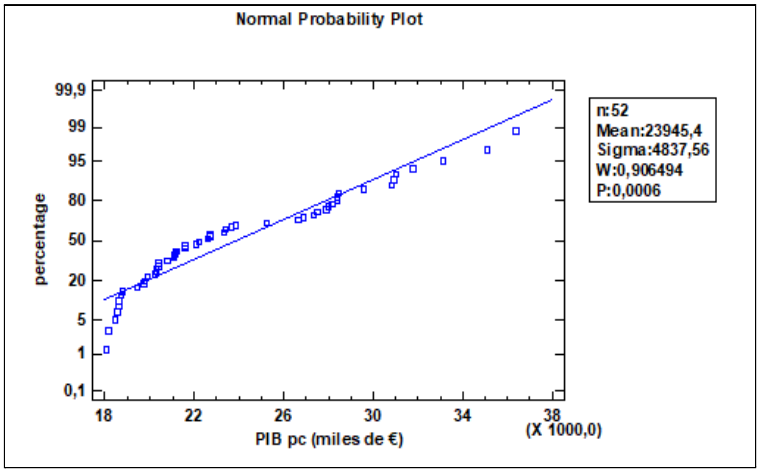
El Gráfico de Probabilidad Normal se usa para ayudar a juzgar si una muestra de datos numéricos proviene o no de una distribución normal. De no ser el caso, frecuentemente se puede determinar el tipo de alejamiento de la normalidad examinando la forma en la que los datos se desvían de la línea de referencia normal. Es un caso particular de gráficos Cuantil-Cuantil ("*Q-Q Plots*").

El gráfico probabilístico normal nos permite comparar la distribución empírica de un conjunto de datos con la distribución normal. Por tanto, dicho gráfico se puede considerar como una técnica gráfica para la prueba de normalidad de un conjunto de datos. La construcción del gráfico de probabilidad normal se realizará a través de los cuantiles de la normal estándar, de forma que aceptaremos la hipótesis de normalidad de los datos, siempre que los puntos en el gráfico tengan un comportamiento "suficientemente rectilíneo" (Castillo, S. y Lozano, E 2007).

Para construir el gráfico de probabilidad normal para un conjunto de datos se representan en el eje vertical los valores ordenados de los datos y en el eje horizontal el valor esperado del i-ésimo estadístico de orden de una distribución normal.

A continuación se presenta el gráfico de normalidad para el PIB per cápita en miles de euros de las provincias españolas (Anexo 1: provincias por PIB per cápita) obtenido mediante STATGRAPHICS:

Figura 15: Gráfico de probabilidad normal. Método de análisis de una variable para el PIB per cápita en miles de euros.



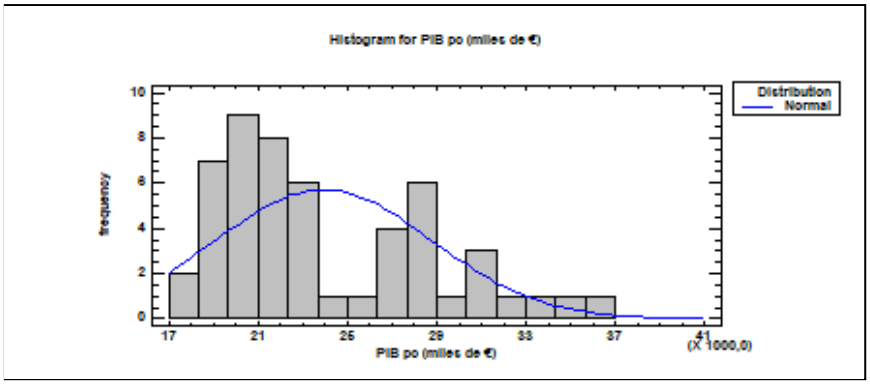
Fuente. Obtenido mediante STATGRAPHICS (Versión 19).

En la Figura 15 se observan las siguientes características de nuestra distribución:

- Si los datos vienen de una distribución normal, los puntos deberían de recaer aproximadamente a lo largo de la línea recta. Ésta recta se sitúa para ayudar a juzgar cómo de cerca se sitúan estos puntos, la cual ha sido determinada por la media y la desviación estándar de los residuos.
- Del gráfico se deduce que sigue una distribución asimétrica hacia la izquierda, la distribución en esta parte está más alejada de la línea, se desvía de ésta para abajo.
- Un extra, para comprobar la normalidad de la distribución, es seleccionar en STATGRAPHICS el ajuste de la distribución. Este panel muestra los resultados del test llevado a cabo para determinar si el PIB per cápita en miles de euros puede ser adecuadamente modelo por una distribución normal.

Para juzgar si un conjunto de datos se distribuye normalmente, se puede utilizar la representación visual del histograma y graficar en él tanto el diagrama de densidad cómo la curva de distribución normal para ver el ajuste que éste último tiene sobre el Histograma. A continuación, en la figura 16 se muestra el histograma con la correspondiente curva de distribución normal en azul (el histograma se describe en el siguiente apartado).

Figura 16: Ajuste de la distribución para el histograma para el análisis del PIB per cápita en miles de euros.



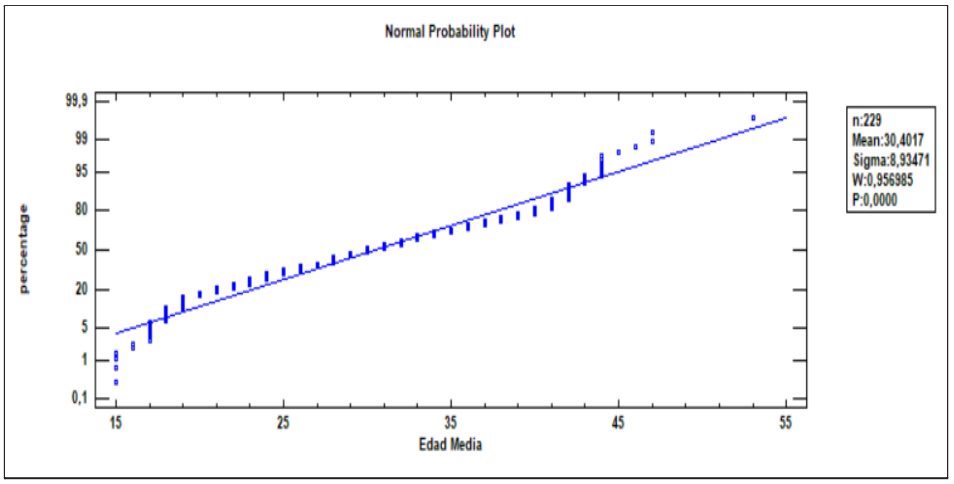
Fuente: Obtenido mediante STATGRAPHICS (Versión 19).

En la Figura 16 se observa que la distribución del PIB per cápita en miles de euros no sigue una distribución normal. De este gráfico se puede observar que los datos no se distribuyen normalmente, ya que la gráfica de la distribución normal no sigue la forma de una campana, también conocida como la campana de Gauss. Las características de este gráfico son las contrarias:

- Presenta asimetría hacia la izquierda.
- No es asintótica, los valores no tienden a infinito
- En el centro de la curva no se encuentran ni la media, ni la mediana ni la moda.
- Los elementos centrales del modelo no son ni la media ni la varianza.

A continuación se presenta el gráfico de normalidad para la edad media de los países de la OMS (Anexo 2: países OMS por edad media) obtenido mediante STATGRAPHICS:

Figura 17: Gráfico de probabilidad normal para el análisis de la variable países de la OMS por edad media.

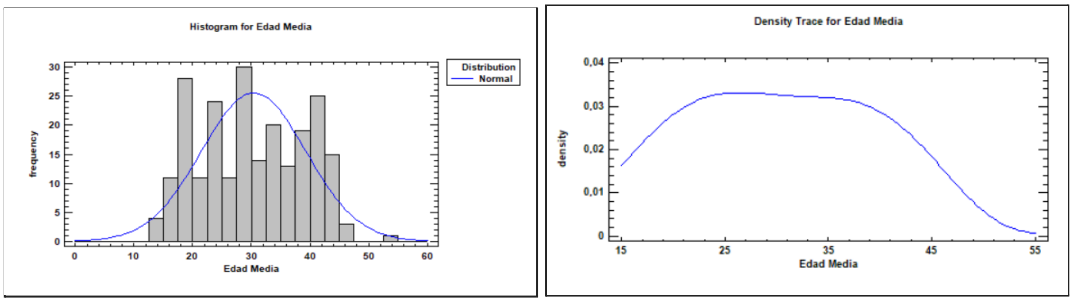


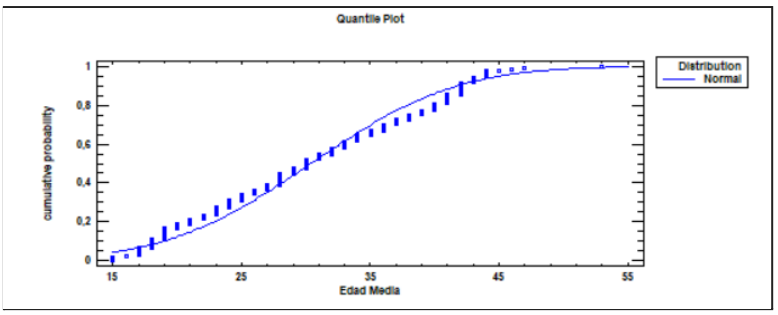
Fuente: Obtenido mediante STATGRAPHICS (Versión 19).

Del análisis para la variable edad media mediante el gráfico de probabilidad normal se puede observar que los datos se ajustan en cierta medida a la línea de referencia, lo que podría significar que la distribución se distribuye normalmente. A diferencia del anterior análisis para el PIB per cápita, hay un claro ajuste en la mayoría de los datos que antes no se percibía. Como se comentó antes, hay otras técnicas que permiten juzgar con mayor claridad si los datos se distribuyen de manera normal.

Con el objetivo de establecer un mejor juicio y a su vez comparar éstas otras técnicas con el gráfico de probabilidad normal, se presentan los gráficos obtenidos mediante STATGRAPHICS para el histograma con el ajuste de la distribución, la curva de densidad y el gráfico del cuantil del análisis de la variable edad media de los países de la OMS:

Figura 18: Curva de densidad, ajuste de la distribución y gráfico de cuantil para el análisis de la variable países de la OMS por edad media.





Fuente: obtenido mediante STATGRAPHICS (Versión 19).

Del histograma y la curva de ajuste se observa cierta normalidad, con una distribución simétrica, asintótica, dónde la media, la mediana y la moda se encuentran en el centro de la curva. Prácticamente casi todos los datos se encuentran bajo la curva, éste es, el área total que representa el 100% de los casos. Ante un aumento de datos, se podría tender a que se ajuste incluso más. Los elementos centrales del modelo son la media y la varianza. La curva de densidad: muestra el área bajo la cuál se encuentran el 100% de los casos. El gráfico del cuantil, al igual que el gráfico de probabilidad normal, muestra la distribución de los datos frente a la distribución normal esperada. Si los datos son normales, los puntos siguen la recta. Si no lo son, forman una curva que se desvía marcadamente de la línea. En este caso, se observa como el conjunto de los puntos se va ajustando a la línea prácticamente en todo su recorrido.

3.6. Gráfico de Dispersión (GD).

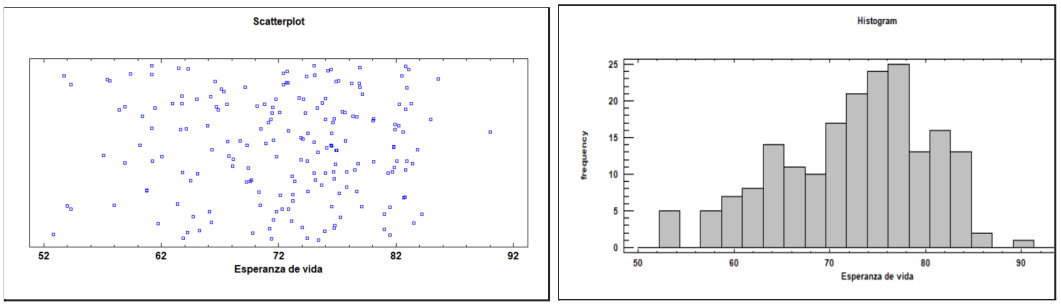
El gráfico de dispersión para análisis univariante es una representación gráfica que se utiliza para detectar valores extremos, explorar y describir la tendencia general de una variable (nubes de puntos en las cuales puede observarse la variable y la cercanía o lejanía de unos puntos a otros). Es un gráfico simple que muestra cada uno de los valores de una variable para cada observación usando puntos relativos al número observado.

Para construir un GD los valores de los datos se sitúan a lo largo del eje horizontal (abscisas). El eje vertical (ordenadas) se representa con el tiempo, a lo largo del cual los puntos se separan en base al tiempo determinado o aleatoriamente hacia

arriba o hacia abajo. Esto se hace para evitar que puntos con igual valor se superpongan. En STATGRAPHICS, la cantidad de separación se controla con el botón Separar en la barra de herramientas del análisis.

Tanto el diagrama de dispersión como el histograma se construye en base al eje X (abscisas), conocidos el punto mínimo y el punto máximo, se divide éste en intervalos y se representa la variable análisis. Ambos nos dan información de la distribución en cuanto a su concentración o dispersión, así como la localización de los valores atípicos. A continuación, se refleja en esta similitud en la Figura 19 graficando la variable esperanza de vida:

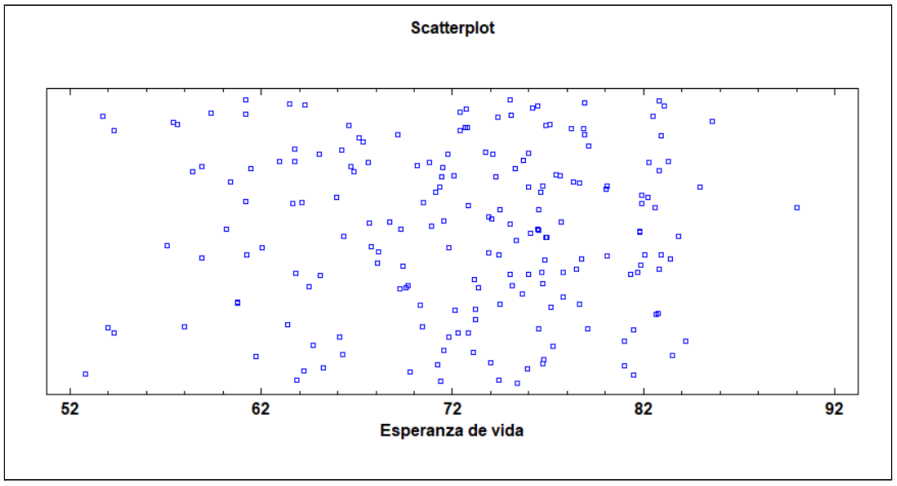
Figura 19: Gráfico de dispersión e histograma para el análisis de la variable esperanza de vida.



Fuente: Obtenido mediante STATGRAPHICS (Versión 19).

A continuación, se presenta el GD para la esperanza de vida de los países de la OMS (Anexo 2: países de la OMS por esperanza de vida):

Figura 20: Gráfico de dispersión para la variable esperanza de vida.



Fuente: obtenido mediante STATGRAPHICS (Versión 19).

La tendencia general de la variable esperanza de vida muestra que dicha esperanza se encuentra más concentrada entre los 72 y 82 años, dónde el gráfico presenta más valores y juntos entre ellos. A la izquierda de dicho intervalo los puntos aparecen más dispersos y a la derecha del intervalo apenas hay puntos. Además, muestra la presencia de valores atípicos, como el punto entre 88 y 92 años alejado del resto, o el cercano a 52, también ligeramente alejado.

4. CONCLUSIONES

John W. Tukey aporta en 1977 contribuciones a la estadística que se han establecido como la fase previa en cualquier análisis de datos. Esto ha marcado la diferencia en la manera en que se tratan los conjuntos de datos de los que parte cualquier investigación. Previamente, los estadísticos descriptivos eran de uso automático sin tener en cuenta si las variables objeto de estudio cumplían con los requisitos de normalidad, simetría y unimodalidad.

La incorporación del Análisis Exploratorio de Datos como fase exploratoria permite un análisis a fondo de la estructura de los datos cuantitativos a través de técnicas gráficas simples y al alcance de cualquiera, teniendo multitud de programas capaces de producir éstas representaciones gráficas fácilmente.

El Análisis Exploratorio de Datos se basa en diferentes técnicas que han de ser utilizadas siguiendo un orden, ya que no todas ellas ofrecen la misma información o no resulta tan fácil identificar ciertas características de la distribución. Para ello, se inicia el análisis con el Diagrama de Tallo y Hojas junto con el Histograma para observar y examinar la presencia de normalidad. A continuación, para observar la simetría y los valores atípicos se utiliza el Diagrama de Caja y Bigotes, ayuda a extender y apoyar el análisis anterior, ya que los primeros no detectan éstos con la misma exactitud. El Diagrama de Dispersión ofrece una idea de las relaciones existentes entre las variables y su ajuste.

El uso de estadísticos robustos (resistentes) es aconsejable cuando los datos no se ajustan a una distribución normal ya que se ven poco afectados por los atípicos y cómo se ha observado, son fáciles de calcular siguiendo las fórmulas.

5. REFERENCIAS BIBLIOGRÁFICAS

Libros:

Pérez López, C. (2004): *Técnicas de Análisis Multivariante de Datos.* PEARSON EDUCACIÓN, S.A., Madrid.

Tukey JW. (1977): *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.

Artículos académicos:

Castillo, S. y Lozano, E (2007). Q-Q Plot Normal. Los puntos de posición gráfica. Ini Inv, 2:a9. Departamento de Estadística e Investigación Operativa. Universidad de Jaén.

Conde, J.E., Rull, V. y Vegas, T. Análisis exploratorio de datos ecológicos y biométricos: gráficos stem-and-leaf (tallo-y-hoja) y boxplot (cajas gráficas). 153–162.

Emerson, J. D., & Hoaglin, D. C. (1983). Mathematical Aspects of Transformation. Resistant Lines for Y versus X. In D. C. Hoaglin, F. Mosteller, & J. L. Tukey (Eds.), Understanding Robust and Exploratory Data Analysis (pp. 129-163). New York: Wiley & Sons.

Monterde i Bort, H; Perea Lara, M. Capítulo 3. Sistemas de Representación Gráfica (Univariados). 1991.

Morgenthaler S. Exploratory Data Analysis, 1(1), 33–44. John Wiley & Sons, Inc. WIREs Comp Stat 2009 1 33–44.

Fuentes de datos:

INE. Instituto Nacional de Estadística. (s/f). INE. Recuperado el 14 de marzo de 2022, de https://www.ine.es/

Home - eurostat. (s/f). Europa.Eu. Recuperado el 14 de marzo de 2022, de https://ec.europa.eu/eurostat

Data at WHO. (s/f). Who.Int. Recuperado el 14 de marzo de 2022, de https://www.who.int/data/

Datos Mundial: El mundo en números. (s/f). DatosMundial.com. Recuperado el 14 de marzo de 2022, de http://www.datosmundial.com

Our World in Data. (s/f). Our World in Data. Recuperado el 14 de marzo de 2022, de http://www.ourworldindata.com

(S/f-a). Expansion.com. Recuperado el 14 de marzo de 2022, de http://www.datosmacro.expansion.com

(S/f-b). Inexmundi.com. Recuperado el 14 de marzo de 2022, de http://www.inexmundi.com

Wikipedia - The Free Encyclopedia. List of countries by median age -List of countries by median age. Recuperado el 14 de marzo de 2022, https://es.wikicore.net/wiki/List_of_countries_by_median_age

6. ANEXO

Los datos del Anexo 1 han sido recogidos de las siguientes fuentes:

- Instituto Nacional de Estadística: https://www.ine.es/
- Eurostat: https://ec.europa.eu/eurostat
- Data at WHO: https://www.who.int/data/
- Datos Mundial: www.datosmundial.com
- Nuestro mundo en datos: www.ourworldindata.com
- Datos Macro: www.datosmacro.expansion.com
- Inex Mundi: www.inexmundi.com

Anexo 1: Provincias de España por PIB

Lista de provincias de España ordenadas por producto interior bruto per cápita (PIBpc) *(según datos del INE del 2018).* [Consultado 18/10/2021 Anexo:Provincias de España por PIB - Wikipedia, la enciclopedia libre].

Id	Provincia	CCAA	PIB (miles de €)	PIB per cápita(miles de €)
1	Cádiz	Andalucía	22.535.246	18.050
2	Granada	Andalucía	16.687.601	18.181
3	Badajoz	Extremadura	12.423.261	18.453
4	Córdoba	Andalucía	14.534.325	18.525
5	Toledo	Castilla-La Mancha	12.814.575	18.617
6	Jaén	Andalucía	11.808.429	18.628
7	**Melilla**	*Melilla*	1.582.540	18.700
8	Málaga	Andalucía	31.023.255	18.801
9	Cáceres	Extremadura	7.664.977	19.464
10	Alicante	Comunidad Valenciana	36.521.398	19.757
11	Zamora	Castilla y León	3.459.100	19.813
12	Almería	Andalucía	13.979.829	19.919
13	**Ceuta**	*Ceuta*	1.720.295	20.251
14	Huelva	Andalucía	10.607.333	20.273
15	Sevilla	Andalucía	39.535.345	20.314
16	Guadalajara	Castilla-La Mancha	5.245.815	20.415
17	Ávila	Castilla y León	3.252.395	20.423
18	Las Palmas	Islas Canarias	23.553.372	20.813
19	Santa Cruz de Tenerife	Islas Canarias	22.269.949	21.076
20	Murcia	Región de Murcia	31.198.376	21.094
21	Albacete	Castilla-La Mancha	8.235.408	21.153
22	Salamanca	Castilla y León	7.048.640	21.187
23	Ciudad Real	Castilla-La Mancha	10.689.033	21.563
24	León	Castilla y León	10.006.588	21.579

25	Orense	Galicia	6.813.831	22.120
26	Segovia	Castilla y León	3.418.981	22.212
27	Pontevedra	Galicia	21.247.944	22.586
28	Cuenca	Castilla-La Mancha	4.536.392	22.691
29	**Asturias**	Asturias	23.258.673	22.709
30	Lugo	Galicia	7.692.177	23.320
31	Valencia	Comunidad Valenciana	59.123.107	23.363
32	**Cantabria**	Cantabria	13.737.756	23.646
33	Zaragoza	Aragón	27.348.811	23.816
34	Teruel	Aragón	3.367.236	25.262
35	Soria	Castilla y León	2.380.731	26.626
36	Valladolid	Castilla y León	13.998.460	26.901
37	Palencia	Castilla y León	4.407.310	27.346
38	**La Rioja**	La Rioja	8.593.185	27.482
39	Islas Baleares	Islas Baleares	32.767.619	27.870
40	Huesca	Aragón	6.134.249	28.015
41	Gerona	Cataluña	21.202.782	28.184
42	Castellón	Comunidad Valenciana	16.149.473	28.367
43	La Coruña	Galicia	26.682.181	28.386
44	Lérida	Cataluña	12.218.853	28.456
45	Burgos	Castilla y León	10.505.020	29.571
46	Tarragona	Cataluña	24.567.640	30.810
47	Barcelona	Cataluña	171.350.447	30.947
48	**Navarra**	Navarra	20.047.454	31.026
49	**Vizcaya**	País Vasco	36.085.689	31.792
50	Guipúzcoa	País Vasco	24.060.930	33.112
51	**Madrid**	Madrid	231.133.592	35.091
52	Álava	País Vasco	11.882.941	36.404

Anexo 2: Países de la Organización Mundial de la Salud por edad media y por Esperanza de Vida.

Lista de países por edad media y por esperanza de vida. Estimación del CIA World Factbook 2018.[4]

Id	Países	Edad Media	Esperanza de Vida
1	Afganistán	18	64,49
2	Albania	32	78,9
3	Argelia	38	76,69
4	Samoa Americana	25	-

[4] https://es.wikicore.net/wiki/List_of_countries_by_median_age

5	Andorra	44	-
6	Angola	15	60,78
7	Anguila	34	-
8	Antigua y Barbuda	31	76,89
9	Argentina	31	76,52
10	Armenia	35	76,7
11	Aruba	39	-
12	Australia	38	82,75
13	Austria	44	81,8
14	Azerbaiyán	32	76
15	Las Bahamas	32	73,75
16	Bahréin	32	77,16
17	Bangladesh	26	72,32
18	Barbados	38	79,08
19	Bielorrusia	40	74,5
20	Bélgica	41	81,7
21	Belice	22	74,5
22	Benin	18	61,47
23	islas Bermudas	43	-
24	Bután	27	71,46
25	Bolivia	24	71,24
26	Bosnia y Herzegovina	42	77,26
27	Botswana	24	69,28
28	Brasil	34	75,67
29	Islas Vírgenes Británicas	36	
30	Brunei	30	75,72
31	Bulgaria	42	75
32	Burkina Faso	17	61,17
33	Birmania	28	66,87
34	Burundi	17	61,25
35	Cabo Verde	25	72,78
36	Camboya	25	69,57
37	Camerún	18	58,92
38	Canadá	42	82,05
39	Islas Caimán	40	-
40	República Centroafricana	19	52,81
41	Chad	17	53,98
42	Chile	34	80,04
43	porcelana	37	-
44	Colombia	30	77,11

45	Comoras	19	64,12
46	República Democrática del Congo	18	60,37
47	República del Congo	19	64,29
48	Islas Cook	36	-
49	Costa Rica	31	80,1
50	Costa de Marfil	20	57,42
51	Croacia	43	78,2
52	Cuba	41	78,73
53	Curacao	36	-
54	Chipre	36	82,9
55	Republica checa	42	79,1
56	Dinamarca	42	81
57	Djibouti	23	66,58
58	Dominica	33	76,6
59	República Dominicana	28	73,89
60	Ecuador	27	76,8
61	Egipto	23	71,83
62	El Salvador	27	73,1
63	Guinea Ecuatorial	19	58,4
64	Eritrea	19	65,94
65	Estonia	42	78,5
66	Eswatini (Swazilandia)	21	-
67	Etiopía	17	66,24
68	UE	42	-
69	Islas Faroe	37	-
70	Fiyi	28	67,34
71	Finlandia	42	81,8
72	Francia	41	82,8
73	Polinesia francés	31	-
74	Gabón	18	66,19
75	Gambia	21	61,74
76	Palestina (Franja de Gaza)	17	-
77	Georgia	38	74
78	Alemania	47	81
79	Ghana	21	63,78
80	Gibraltar	34	-
81	Grecia	44	81,9
82	Granada	31	72,38

83	Guam	29	-
84	Guatemala	22	74,06
85	Guernsey	43	-
86	Guinea-Bissau	20	58
87	Guinea	18	61,19
88	Guayana	26	69,77
89	Haití	23	63,66
90	Honduras	23	75,09
91	Hong Kong	44	84,93
92	Hungría	42	76,2
93	Islandia	36	-
94	India	28	69,42
95	Indonesia	30	71,51
96	Iran	30	76,48
97	Irak	20	70,45
98	Irlanda	36	82,2
99	Isla del hombre	44	-
100	Israel	29	82,8
101	Italia	45	83,4
102	Jamaica	26	74,37
103	Japón	47	84,21
104	Jersey	38	-
105	Jordán	33	-
106	Kazajstán	30	73,15
107	Kenia	19	66,34
108	Kiribati	24	68,12
109	Corea del Norte	34	72,1
110	Corea del Sur	41	82,63
111	Kosovo	29	-
112	Kuwait	29	75,4
113	Kirguistán	26	71,4
114	Laos	23	67,61
115	Letonia	43	75,1
116	Líbano	30	78,88
117	Lesoto	24	53,71
118	Liberia	17	63,73
119	Libia	28	72,72
120	Liechtenstein	43	83,1
121	Lituania	43	76
122	Luxemburgo	39	82,3
123	Macao	39	-

124	Macedonia del Norte	37	76,7
125	Madagascar	19	66,68
126	Malawi	16	63,8
127	Malasia	28	76
128	Maldivas	28	78,63
129	Mali	15	58,89
130	Malta	41	82,5
131	Islas Marshall	22	65,24
132	Mauritania	20	64,7
133	Mauricio	35	74,42
134	México	28	74,99
135	Estados Federados de Micronesia	25	67,76
136	Moldavia	36	71,81
137	Mónaco	53	
138	Mongolia	28	69,69
139	Montenegro	40	76,9
140	Montserrat	33	-
141	Marruecos	29	76,45
142	Mozambique	17	-
143	Namibia	21	63,37
144	Nauru	26	-
145	Nepal	24	70,48
146	Países Bajos	42	81,9
147	Nueva Caledonia	32	-
148	Nueva Zelanda	37	81,86
149	Nicaragua	25	74,28
150	Nigeria	18	54,33
151	Níger	15	62,02
152	Islas Marianas del Norte	33	-
153	Noruega	39	82,8
154	Omán	25	77,63
155	Pakistán	23	67,11
156	Palau	33	69,13
157	Panamá	29	78,33
158	Papúa Nueva Guinea	23	64,26
159	Paraguay	28	74,13
160	Perú	28	76,52
161	Filipinas	23	71,1
162	Polonia	40	77,7

163	Portugal	42	81,5
164	Puerto Rico	41	-
165	Katar	33	80,1
166	Rumania	41	75,3
167	Rusia	39	72,66
168	Ruanda	19	68,7
169	San Bartolomé	44	-
170	Santa Elena, Ascensión y Tristán da Cunha	41	-
171	Saint Kitts y Nevis	35	-
172	Santa Lucía	34	76,06
173	San Martín	32	-
174	San Pedro y Miquelón	46	-
175	San Vicente y las Granadinas	33	72,42
176	Samoa	24	73,19
177	San Marino	44	85,6
178	Santo Tomé y Príncipe	18	70,17
179	Arabia Saudita	27	75
180	Senegal	18	67,67
181	Serbia	42	75,9
182	Seychelles	35	72,84
183	Sierra Leona	19	54,31
184	Singapur	34	83,3
185	Sint Maarten	41	-
186	Eslovaquia	40	77,4
187	Eslovenia	44	81,5
188	Islas Salomón	22	72,84
189	Somalia	18	57,07
190	Sudáfrica	27	63,86
191	Sudán del Sur	17	57,6
192	España	42	83,5
193	Sri Lanka	32	76,81
194	Sudán	19	65,1
195	Surinam	29	71,57
196	Suecia	41	82,6
197	Suiza	42	83,8
198	Siria	24	71,78
199	Taiwán	40	-
200	Tayikistán	24	70,88

201	Tanzania	17	65,02
202	Tailandia	37	76,93
203	Timor Oriental	18	69,26
204	Para llevar	19	-
205	Tonga	23	70,8
206	Trinidad y Tobago	36	73,38
207	Túnez	31	76,51
208	pavo	30	-
209	Turkmenistán	27	68,07
210	Islas Turcas y Caicos	33	-
211	Tuvalu	25	-
212	Uganda	15	62,97
213	Ucrania	40	73,2
214	Emiratos Árabes Unidos	30	77,81
215	Reino Unido	40	81,3
216	Estados Unidos	38	78,64
217	Uruguay	35	77,77
218	Uzbekistán	28	71,57
219	Vanuatu	22	70,32
220	Venezuela	28	72,13
221	Vietnam	30	75,32
222	Islas Virgenes	41	-
223	Wallis y Futuna	32	-
224	Palestina (Cisjordania)	21	73,9
225	Sahara Occidental	21	-
227	Yemen	19	66,1
228	Zambia	16	63,51
229	Zimbabue	20	61,2
230	China [+]	-	76,7
231	Jordania [+]	-	74,41
232	San Cristóbal y Nieves [+]	-	71,34
233	Suazilandia [+]	-	59,4
234	Togo [+]	-	60,76