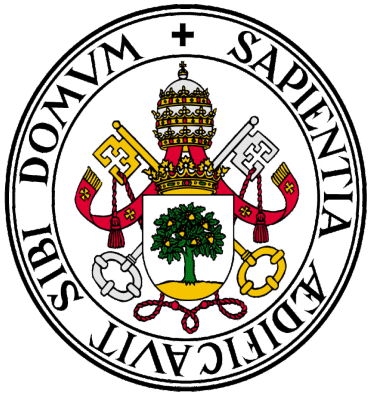


Universidad de Valladolid

Facultad de Ciencias Económicas y Empresariales

Trabajo de Fin de Grado

Grado en Administración y Dirección de empresas

Análisis exploratorio de datos. Una variable

Presentado por:

Ernesto Grande Atienza

Tutelado por:

José Luis Rojo

Valladolid, 09 de diciembre de 2021

RESUMEN

El objetivo principal de este trabajo es la descripción de una parte de la estadística que es el análisis exploratorio de datos, también conocido como AED, y distinguiremos esta rama del análisis estadístico tradicional. Explicaremos las características y el porqué es tan importante la realización de un buen gráfico, y detallaremos los diferentes tipos que hay para el estudio de una variable con diferentes tipos de ejemplos basados en la economía y población actual para así tener una visión más real.

Finalmente, terminaremos el trabajo con una breve introducción al estudio multidimensional y una de las técnicas más populares que es la recta de Tukey para ver la posible relación lineal entre variables.

Palabras clave: AED, gráfico, unidimensional, Tukey.

JEL: C10, C81, C82

ABSTRACT

The main objetive of this assignment is the description of a part of the statistcs that is the exploratory analysis of data, also known as EAD, and we will distinguish this branch from the traditional statistical analysis. We will explain the characteristics and why is so important to make a good graph, and we will detail the different types there are for the study of a variable with different types of examples based on the current economy and population to have a more realistic vision.

Finally, we will finish the work with a brief introduction to multidimensional study and one of the most popular techniques is the Tukey line to see the possible linear relationship between variables.

Keywords: EAD, graphic, one-dimensional, Tukey.

JEL: C10, C81, C82

ÍNDICE:

Índice de Tablas y Gráficos:

1. INTRODUCCIÓN. QUÉ ES EL ANÁLISIS EXPLORATORIO DE DATOS (AED). DIFERENCIAS CON EL ANÁLISIS ESTADÍSTICO TRADICIONAL

Comenzaremos dando una pequeña definición de lo que es la estadística. Es la ciencia que trata de estudiar cómo usar la información, explica cómo actuar en ciertas situaciones prácticas y ofrece diferentes maneras o métodos de investigación aplicable a otros campos.

La estadística ha tenido muchas definiciones a lo largo de la historia, muchas más precisas y extensas que otras, y todo dependiendo del autor que lo haya escrito. De todas las definiciones habidas, podemos extraer una nueva y probablemente más precisa: La Estadística es la ciencia que se ocupa del estudio de los fenómenos aleatorios.

Según Fernández-Abascal et al. (1994), la estadística presenta 2 planeamientos claramente diferenciados:

- La Estadística Descriptiva, la cual no saca conclusiones que vayan más allá de los propios datos que maneja. Una vez tenemos las observaciones, en mayor o menor cantidad, las ordena, resume y analiza, obteniendo así la información primaria contenida en las mismas. Esta rama de la Estadística se compone de un conjunto de técnicas y métodos para la recogida, descripción y análisis de datos, los cuales están sujetos a los fenómenos aleatorios que se estudian una vez han ocurrido; es decir, se trata de un estudio empírico, a posteriori, basado en la ocurrencia observada de los datos (frecuencias).
- El segundo planteamiento es la Estadística Teórica, que también podemos calificar como Matemática o Inferencial. Su función es el estudio de los fenómenos aleatorios antes de que éstos se produzcan, para encontrar pautas de comportamiento y poder preverlos. Es por tanto un estudio teórico, a priori, que se basa en la ocurrencia esperada de los posibles resultados (probabilidades).

Estos planteamientos no son antagónicos ya que la Estadística Teórica se apoya en la Descriptiva porque para poder conocer todas las realidades aleatorias, ya sea en un futuro o en unas condiciones distintas a las normales,

hay que conocer previamente su historia (en observaciones pasadas de la misma).

Antes de seguir vamos a describir algunos conceptos que son importantes para entender qué es la estadística:

- A la hora de realizar la observación y experimentación, podemos encontrarnos con 2 tipos de fenómenos:
    1. Fenómenos causales o deterministas: Son aquellos que, al repetirlos en las mismas condiciones, se obtiene el mismo resultado. Es decir, dadas unas causas puede deducirse el resultado final.
    2. Fenómenos inciertos o aleatorios: No es posible predecir el resultado ya que influye el azar encontrándonos resultados diferentes en iguales condiciones.
- Una variable estadística puede ser definida como esa característica o cualidad propensa de adquirir diferentes valores de un conjunto determinado o dominio de la variable. Existen dos tipos de variables: cuantitativas y cualitativas.
- Un dato es cada uno de los valores obtenido en el estudio estadístico.
- Los metadatos son un grupo de datos que denominan el contenido informativo de un objeto, es decir, dan una descripción entera de los datos que vamos a estudiar. La información que contiene puede ser en términos de su integridad, consistencia y precisión general. Son de gran importancia y si están incompletos, son inconsistentes o son inexactos, nos puede causar grandes problemas en el análisis posterior. .

El AED se puede definir como el arte de observar uno o varios conjuntos de datos y tratar de comprender la estructura subyacente de los datos ahí contenidos. Es decir, es una forma de analizar datos mirando números y gráficos y obteniendo un patrón a través del tratamiento de las muestras recogidas en el proceso de investigación de un campo científico.

Este análisis se realiza a través de gráficos y estadísticos cuyo objetivo es explorar la distribución descubriendo valores atípicos o outliers, evaluación de

datos ausentes, la forma de la distribución, discontinuidades o saltos, concentraciones de valores, etc...

El análisis se realizará de todos los casos a la vez o de forma separada por grupos. Si es de forma separada, los gráficos y estadísticos nos revelan si los resultados obtenidos derivan de una o varias poblaciones, señalando a la variable que determina esos grupos como agente diferenciador de las muestras. Además, nos posibilita comprobar, a través de técnicas gráficas y contrastes no paramétricos, si la búsqueda de datos ha sido a través de una población con distribución aproximadamente normal.

El examen que se hace antes de analizar los datos es necesario, aunque lleve tiempo, pero habitualmente se descuida por parte de quien realiza el análisis de datos. Las tareas en dicho examen previo parecerán ridículas o insignificantes, pero son parte esencial de este proceso.

Según C. Batanero et al. (1991), tenemos que saber los principios por los que se guía el AED. Hay que tener en cuenta que los datos están constituidos por dos partes: la “regularidad” y las “desviaciones”. La regularidad indica la estructura simplificada de un conjunto de observaciones (en una nube de puntos, por ejemplo, es la recta a la cual se ajusta). Las diferencias de los datos con respecto a esta estructura (diferencia en nuestro caso respecto a la recta), representan las desviaciones o residuos de los datos, que usualmente no tienen por qué presentar una estructura determinada. Tradicionalmente el estudio se ha concentrado en la búsqueda de un modelo que exprese la regularidad de las observaciones. Por el contrario, el análisis exploratorio de datos es básicamente el desglose de los mismos en las dos partes que hemos citado. En lugar de imponer, en hipótesis, un modelo a las observaciones, se genera dicho modelo desde las mismas.

Como ya hemos dicho, el objetivo es extraer la mayor cantidad de información posible y crear nuevas tesis con la ayuda de todas las herramientas con las que contamos. Por otro lado, al final del análisis, esta "tesis" no forma un contraste en el significado estadístico del término, por lo que es necesario obtener nuevos datos sobre el fenómeno y realizar análisis estadísticos tradicionales sobre el fenómeno para compararlos.

El AED además de toda la utilidad que nos proporciona, tiene la siguiente serie de características:

• Fuerte apoyo en representaciones gráficas: “Una idea fundamental del análisis exploratorio de datos es que al usar representaciones múltiples de los datos se convierte en un medio de desarrollar nuevos conocimientos y perspectivas. Esto puede ejemplificarse al pasar de tablas a gráficos, de lista de números a representaciones como la del ‘tallo’, reduciendo los números a una variedad discreta en un mapa estadístico para facilitar la exploración de la estructura total, construyendo gráficos, como el de la ‘caja’ que hace posible la comparación de varias muestras”. (Biehler, 1998,pg.2).

• Empleo preferente de los estadísticos de orden, porque son sensibles a la mayor parte de los datos, pero no lo son al valor de los datos individuales, para grupos pequeños de datos, y con ellos se disminuye el efecto producido por los valores atípicos, escasos y muy alejados de la norma.

• No necesita una teoría matemática compleja, utiliza nociones matemáticas muy elementales y procedimientos gráficos fáciles de realizar.

• Bastante parecida a la estadística descriptiva tradicional, pero se aleja de ella por su intención. Pues, al contrario que en ella, la representación o el cálculo no son en el análisis exploratorio de datos un fin, sino un medio de descubrir la información oculta en los mismos”. (Jullien y Nin 1989. págs. 30-31.)

• Uso de diferentes escalas o re-expresión: La escala en la que una de las variables es observada y registrada no es única. A veces, transformando los valores originales de la variable a una nueva escala se puede lograr que dichos valores sean más manejables.

En resumen, como indica Biehler (1998, pag. 5): “El currículum tradicional de Estadística Descriptiva debiera transformarse en dirección al análisis exploratorio de datos. Seria esencial, sin embargo, dar apoyo sustancial a la actitud investigadora, contra la tendencia de la mayor parte de las transposiciones didácticas de reducir el conocimiento a la técnica”.

PRINCIPIOS BASICOS DEL AED

De acuerdo con Hartwig y Dearing (1979), existen dos principios básicos que tienen que desarrollarse en el investigador para que lleve a cabo un buen AED, que son:

1. Escepticismo: El primer principio es ser escéptico con los resúmenes digitales del conjunto de datos, porque a veces oscurecen o no revelan cuál puede ser el aspecto más útil de los números. Por el contrario, suele haber demasiada confianza en el resumen numérico de estos.

   La estadística como concepto de análisis de datos parece enfatizar la importancia de los números, es decir, los datos estadísticos en sí mismos (resumen de números de datos) tienden a reducir la importancia de la representación gráfica de los datos en comparación con el método clásico. Por otro lado, es ampliamente utilizada por el AED.

   La visión habitual y clásica es que la información estadística es más fiable que la representación gráfica de los datos, sin embargo, esta información estadística puede incluso oscurecer o ignorar información que puede ser muy importante en AED (Análisis visual). Debe ser señalado que tiene que realizarse antes del análisis estadístico numérico, aunque este último sigue siendo el producto final deseado.

2. Actitud abierta: los analistas deben estar abiertos a patrones imprevistos en los datos, porque estos patrones pueden ser los aspectos más explicativos del análisis.

Para realizar un AED, hay que seguir un cierto número de etapas, estas son:

1. Organizar los datos de manera que sean aplicables para cualquier método estadístico.
2. Desarrollar un estudio descriptivo numérico que posibilite cuantificar algunos aspectos gráficos de los datos.
3. Realizar un examen gráfico sobre las relaciones entre variables examinadas y un análisis descriptivo numérico que mida el grado de interrelación que hay entre ambas.

4. Si es necesario, valorar algunos supuestos básicos que son los cimientos de la mayoría de las técnicas estadísticas, como la normalidad.
5. Observar y analizar los posibles datos atípicos y evaluar el impacto que tienen o puedan tener sobre los análisis estadísticos posteriores.
6. Determinar el impacto potencial (solo si fuera indispensable) que puedan tener los datos ausentes frente a la representatividad de los datos que estamos analizando.

2. ANÁLISIS EXPLORATORIO DE DATOS PARA UNA VARIABLE

2.1. Herramientas, Gráficas y Tablas

Una vez tenemos los datos recopilados y organizados, pasamos a realizar un análisis estadístico gráfico y numérico de las variables del problema para así obtener una idea de la información que hay en el conjunto de datos, así como descubrir posibles errores en la codificación de estos.

Dependiendo de la escala de medida de la variable analizada, haremos un tipo de análisis u otro. En la siguiente tabla vienen determinadas las representaciones gráficas y resúmenes descriptivos numéricos que mejores son para realizar el dicho análisis. En la tabla se sobreentiende que las escalas de razón suelen usar las medidas numéricas y representaciones gráficas de las escalas de intervalos además de las suyas propias.

Tabla 2.1: Tipos de gráficas y medidas del AED

Escala de medida	Representaciones graficas	Medidas de tendencia central	Medidas de dispersión
Intervalo	- Histogramas - Polígono de frecuencias - Boxplot - Gráfico de Cuantiles - Gráfico Cuantil-Cuantil	- Media aritmética - Mediana - Moda	- Desviación típica - Recorrido Intercuartílico
Razón	- Diagrama de Tallo y hojas	- Media geométrica	- Coeficiente de variación

Los gráficos estadísticos son una herramienta visual que ayuda a reflejar los datos complejos de forma sistematizada y simple. No existe una herramienta estadística única que sea tan poderosa como un gráfico bien elegido y esto se debe a que nuestra relación ojo-cerebro está demasiado desarrollada y, a través de pantallas gráficas, hace un buen uso de este sistema para obtener una visión profunda de la estructura de los datos. Somos capaces de resumir gran cantidad de información velozmente y sacar conclusiones destacadas, y a la vez enfocarnos en los detalles. Incluso para grupos pequeños de datos, hay muchos patrones y relaciones que son mucho más sencillos de entender en presentaciones gráficas que con cualquier otro método analítico de datos.

Las características de un buen gráfico estadístico son:

- Conseguir la atención del lector.
- Mostrar la información de una forma fácil, clara y exacta.
- No incitar al error.
- Favorecer la comparación de datos y recalcar las tendencias y las disparidades.
- Ilustrar el tema o mensaje del conjunto de datos existente.

Los gráficos que se pueden utilizar para el análisis exploratorio de datos son los siguientes:

1. Diagrama de tallo y hojas
2. Histograma
3. Diagrama de dispersión unidimensional
4. Gráfico de Cuantiles
5. Gráfico Cuantil-Cuantil
6. Gráfico de simetría
7. Diagrama de caja y bigotes

1. Diagrama de tallo y hojas

Es un híbrido entre una tabla y un gráfico ya que muestra la forma de un grupo de datos cuantitativos pero con un perfil muy parecido a un histograma. Para construir un diagrama de este tipo hay que escribir los primeros dígitos de cada

dato a la izquierda de una línea vertical (a este lado le llamaremos tallo). Después, se representa cada valor de datos escribiendo su dígito final en la fila correspondiente en el lado derecho de la línea (a cada dígito de este lado le llamaremos hoja). El siguiente paso es ordenar los datos de cada lado de la línea vertical de menor a mayor para que sea más fácil visualmente hablando.

El diagrama nos muestra información visual; la longitud cada fila nos presenta el número de valores en cada fila, por lo que la pantalla es esencialmente un histograma, acostado de lado. Aunque aparente mostrar o proporcionar la misma información que el histograma, tiene 2 ventajas importantes frente a este:

- Este diagrama es mucho más sencillo de componer a mano.
- Cada intervalo de datos nos brinda más información que un histograma ya que el tallo y la hoja facilitan el dato. Igual que para el histograma no hay un número de clases predeterminado, el diagrama de tallo y hojas tampoco tiene un número señalado de líneas de tallos. Es tan fácil como que, si creemos que se condensan mucho los datos, los expandamos o agrandemos empleando dos o más tallos para cada primer dígito.
- Este diagrama es útil sobre todo cuando queremos transmitir tanto los valores numéricos como la información gráfica sobre la distribución.

En algunas ocasiones, es necesario truncar los datos para alterar las unidades de medida multiplicando por una potencia de 10 y así conseguir valores adecuados para la observación del diagrama.

Es un método sencillo de construir a mano, pero también podemos hacer un diagrama de tallo y hojas a partir de datos desordenados. Las hojas no estarán ordenadas dentro de los tallos, pero podemos copiar fácilmente la visualización poniéndolas en orden. Cuando trabajamos con papel y lápiz, esta es una herramienta increíblemente sencilla para clasificar los datos de menor a mayor para calcular cuantiles de varios órdenes.

Por ejemplo, tomaremos el conjunto de datos de la tasa de fertilidad (nacimientos por cada mujer) a nivel mundial (N=199) del año 2019,

ordenándolos de menor a mayor siendo el primero la República de Corea con un 0,9 y el último Somalia con un 6. Como se puede observar, obtenemos un histograma con todos los valores ordenados, además de expresar la forma de la distribución.

La tasa de fertilidad nos da el número medio de hijos que tiene una mujer durante su vida fértil. Los países africanos son los que tienen una mayor tasa en comparación con el resto del mundo, destacando países como Níger, Angola, RP del Congo o Mali.

Por el contrario, los países con una tasa de fertilidad más baja son del sureste de Asia (Singapur, Macao, Taiwan o Hong Kong) y países europeos (Polonia, Rumanía, Eslovenia, etc...)

Como podemos observar, la tasa es más baja en países donde la mujer ha podido cambiar su rol dentro del matrimonio y dentro de la sociedad, ocupando así los números más bajos dentro de nuestro gráfico. Mientras, los países menos desarrollados son menos, por lo que son los que menor número de hojas tienen dentro del gráfico.

En este ejemplo no encontramos ningún dato atípico, dado que no hay valores aislados en los extremos, pero sí encontramos una cierta asimetría hacia la derecha (o positiva), ya que la mayoría de los datos se encuentran en la parte superior de la gráfica.

Gráfica 2.1: Tasa de fertilidad a nivel mundial (forma reducida)

0|9

1|011122233333444444444555555555555566666666666666777777777777778888888889999999

2|000000000000111111112222222333333344444444455677778888888999

3|0000133345555566778899

4|000111223333444455666666888 8

5|12334688

6|0

Otra forma de presentar el diagrama de tallo y hojas es extendiéndolo si estuviese bastante condensado. Para ello, separamos cada fila en dos subgrupos, uno del 0 al 4 y otro del 5 al 9, consiguiendo así una fácil

construcción ya que el tronco del diagrama de tallo y hojas retiene los valores numéricos de los datos.

Gráfica 2.2: Tasa de fertilidad a nivel mundial (forma extendida)

0|9
1 (0-4)|011122233333444444444
1 (5-9)|555555555555566666666666666777777777777778888888889999999
2 (0-4)|000000000000111111112222222333333344444444444
2 (5-9)|5567777888888999
3 (0-4)|000013334
3 (5-9)|5555566778899
4 (0-4)|0001112233334444
4 (5-9)|556666668888
5 (0-4)|12334
5 (5-9)|688
6|0

2. Histograma:

Es otra forma de resumir una distribución de datos cuantitativos. Consiste en dividir el rango de los datos en varios intervalos, contar el número de puntos en cada intervalo y trazar los recuentos como longitudes de barra. Las alturas relativas de las barras muestran la densidad relativa de observaciones en los intervalos. La longitud de los intervalos en aplicaciones (como Excel o Statgraphics) es la misma para todas las barras, pero en un principio, la longitud puede variar para una mejor representación de los datos, siendo una longitud mayor cuanto menor es la concentración de datos y una menor longitud cuando la concentración de datos es mayor.

Este tipo de gráfico se diferencia de un gráfico de barras en que en el histograma no hay separación natural entre los rectángulos adyacentes.

Una de las funciones más importantes del histograma es abastecer de información sobre la forma de la distribución ya que tiene una forma muy sencilla incluso para la mayoría de las personas sin conocimientos técnicos y sin dar una explicación muy extensa. Sin embargo, como dispositivo de análisis de datos tiene algunos inconvenientes ya que puede hacerse con más detalles, pero con una menor simplicidad o viceversa.

El gráfico puede ser asimétrico hacia la izquierda o hacia la derecha si su cola se extiende más hacia uno de esos lados, o ser simétrico si ambas colas son una imagen una de la otra.

En este ejemplo, hemos separado los países en dos grupos diferenciados por su grado de desarrollo. Para ello, se filtran los países en base a su PIB anual para diferenciarlos entre más y menos desarrollados y, en base a eso, analizar su Índice de Desarrollo Humano (IDH) en los diferentes histogramas que abarcan el primer grupo (más desarrollados) y el total de países.

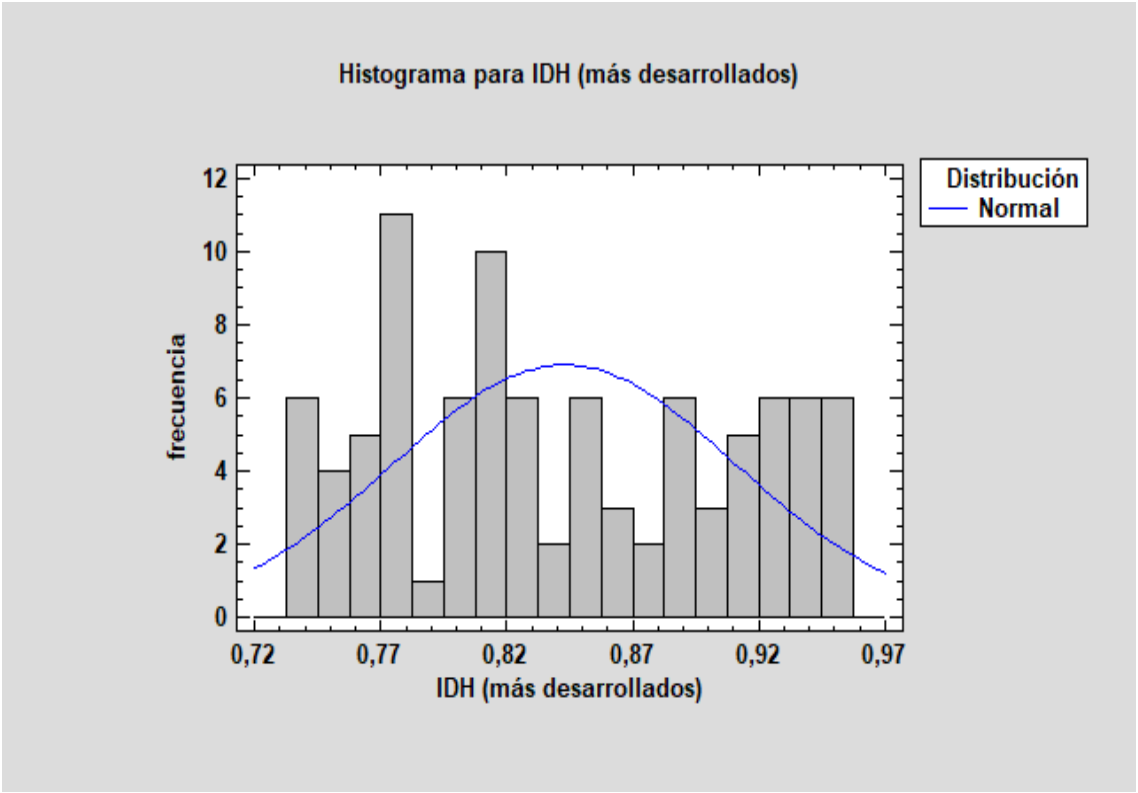
El IDH es un indicador elaborado por las Naciones Unidas que se utiliza para calificar a los países en cuatro niveles de desarrollo humano (esperanza de vida, educación y dos indicadores de ingreso per cápita). Por lo tanto, con estos cuatro indicadores, un país tiene un IDH elevado cuando todos estos factores son altos.

La dispersión de datos para el primer histograma es desde un IDH de 0,738 perteneciente a San Vicente y las Granadinas, hasta Noruega que es el valor más elevado con un 0,957. Según los datos obtenidos a través de la página de la Organización Mundial de la Salud, estos datos corresponden a la mayoría de los países del norte de Europa (Noruega, Alemania, Suecia...), EEUU y Canadá, el este de Asia y algunos países de Oceanía como Australia y Nueva Zelanda entre otros.

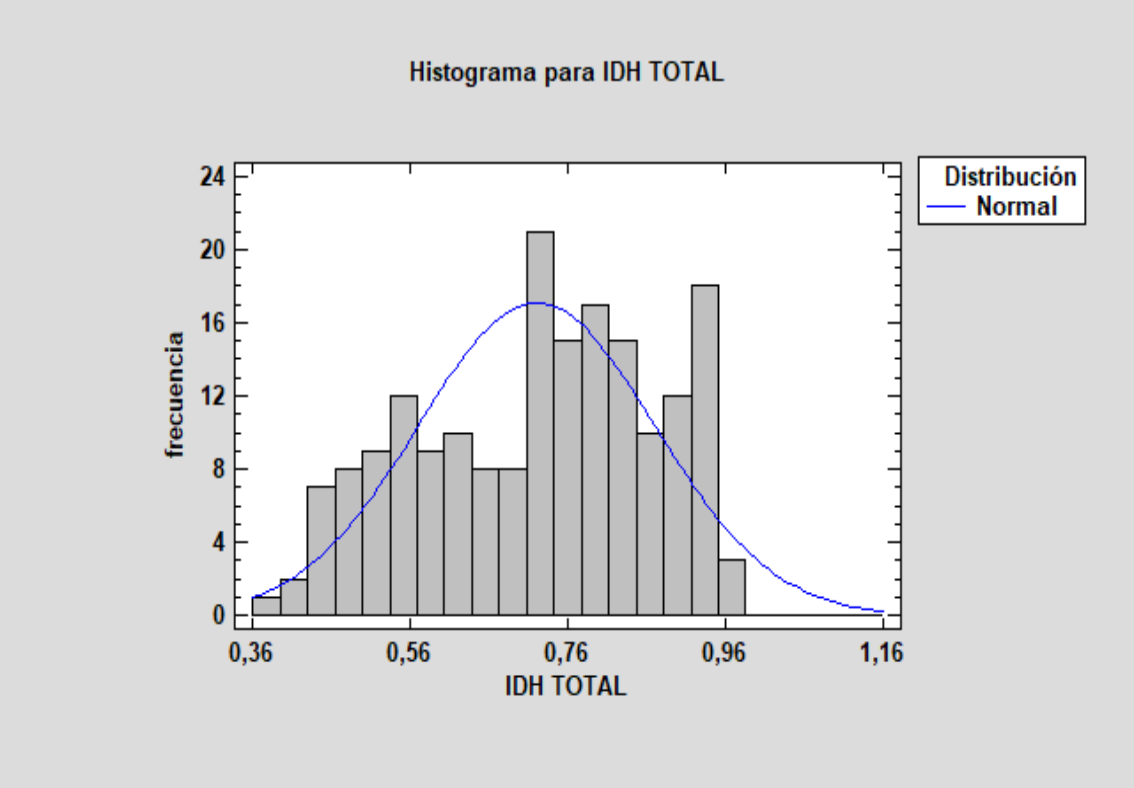
Por el contrario, entre los países con un IDH más bajo encontraríamos la gran mayoría de países del continente africano (Níger, República centro africana, R.D. del Congo...), y otros lugares del mundo serían Haití, Pakistán o Afganistán debido a las continuas guerras y tardía independencia de sus respectivos países colonizadores.

Es habitual comparar los histogramas con la función de densidad normal, para apreciar el grado de normalidad de los datos. Por ejemplo, estos datos no siguen una distribución normal, ya que no siguen una simetría natural si no que es más bien una distribución con alternación de alto y bajos debido a la concentración de datos en ciertos intervalos y lo contrario en otros intervalos. Para demostrarlo he ajustado, mediante el uso del programa Statgraphics, como sería la distribución normal de esta.

Gráfica 2.3: Índice de Desarrollo Humano para países más desarrollados.

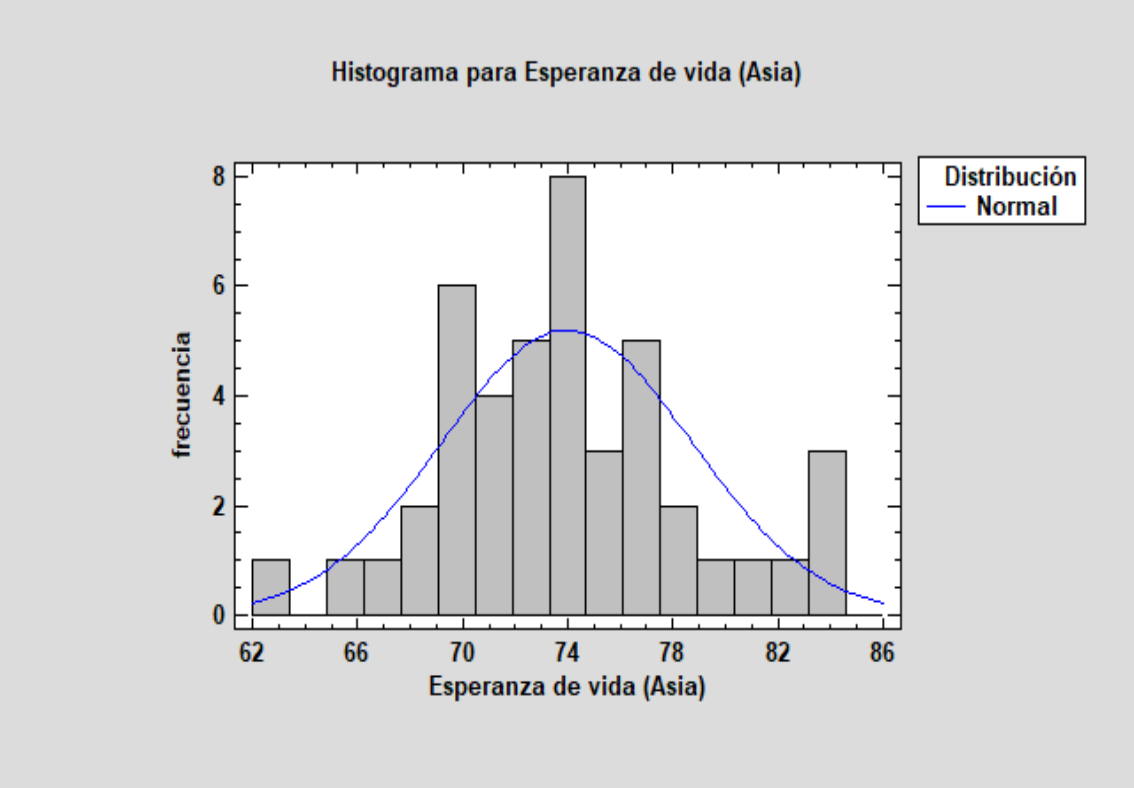


Gráfica 2.4: Índice de Desarrollo Humano para el total de países.

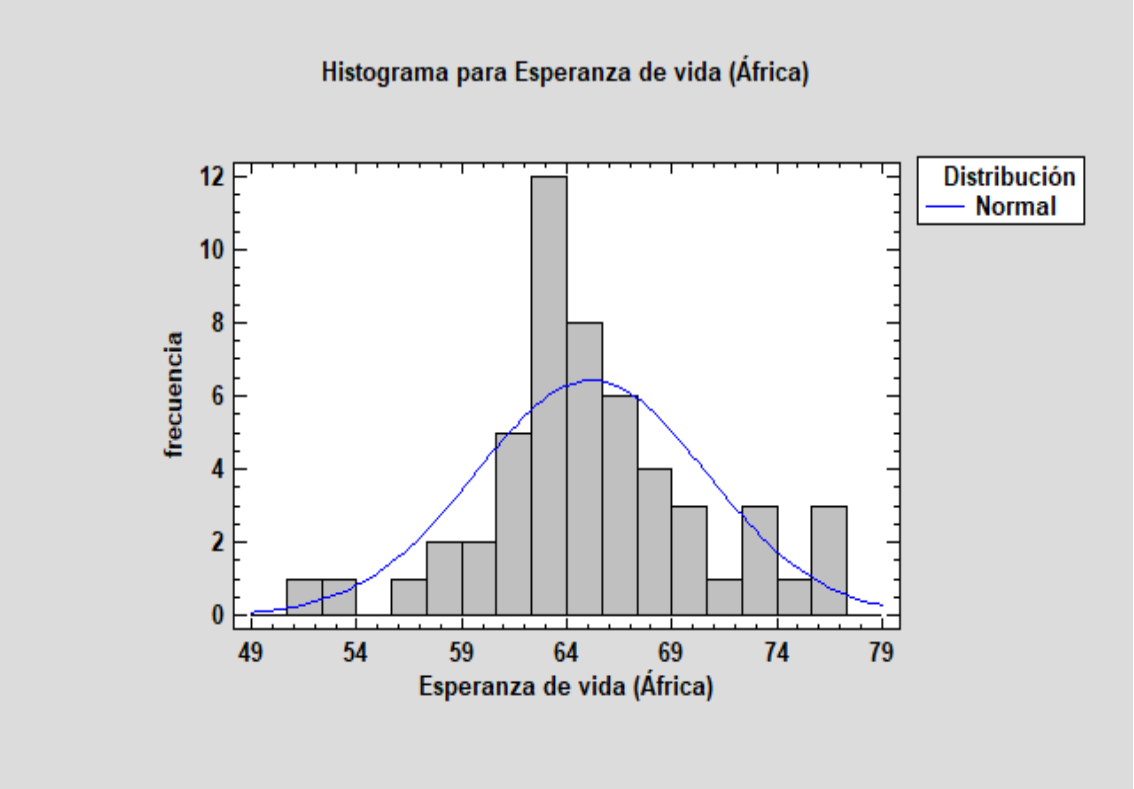


Como ejemplo de distribución que sigue una Normal ponemos el ejemplo de la Esperanza de vida de los continentes asiático y africano, en los que se puede observar que los valores más usuales coinciden con el máximo de altura de la campana de Gauss, aunque ninguno de ellos sea simétrico ya que ambos lados tienen una imagen diferente.

Gráfica 2.5: Esperanza de vida en Asia.



Gráfica 2.6: Esperanza de vida en África



3. Diagramas de dispersión unidimensional

Otra forma sencilla de representar el conjunto de datos disponible y su distribución es dibujando dichos datos a lo largo de una línea numérica o eje etiquetado de acuerdo con la escala de medición.

Su principal virtud es su gran compacidad, que permite utilizar en los márgenes de otras pantallas información que podamos agregar. En un diagrama de este tipo podemos ver claramente diferenciados los valores máximos y mínimo de los datos y podemos obtener impresiones muy aproximadas del centro de los datos, la extensión, la densidad local, simetría y valores atípicos.

Sin embargo, los cuantiles individuales ya no se encontrarán tan fácilmente, y es más probable que la resolución visual de los puntos sea un problema incluso para una cantidad moderada de puntos. Puede darse el caso de que haya varios valores repetidos en los datos y que no se muestren en el gráfico, pudiéndose paliar este problema apilando puntos, es decir, desplazándolos verticalmente cuando coinciden con otros. Esto sólo es una solución al problema de la superposición exacta y no nos ayuda cuando hay muchos puntos que se apilan entre sí.

En el siguiente ejemplo podemos observar la edad media para la maternidad del primer hijo en las provincias de España. Con esto podemos observar que mientras que Almería es la provincia donde se tiene antes al primer hijo con 29 años, en Vizcaya se espera casi a los 32 años y medio.

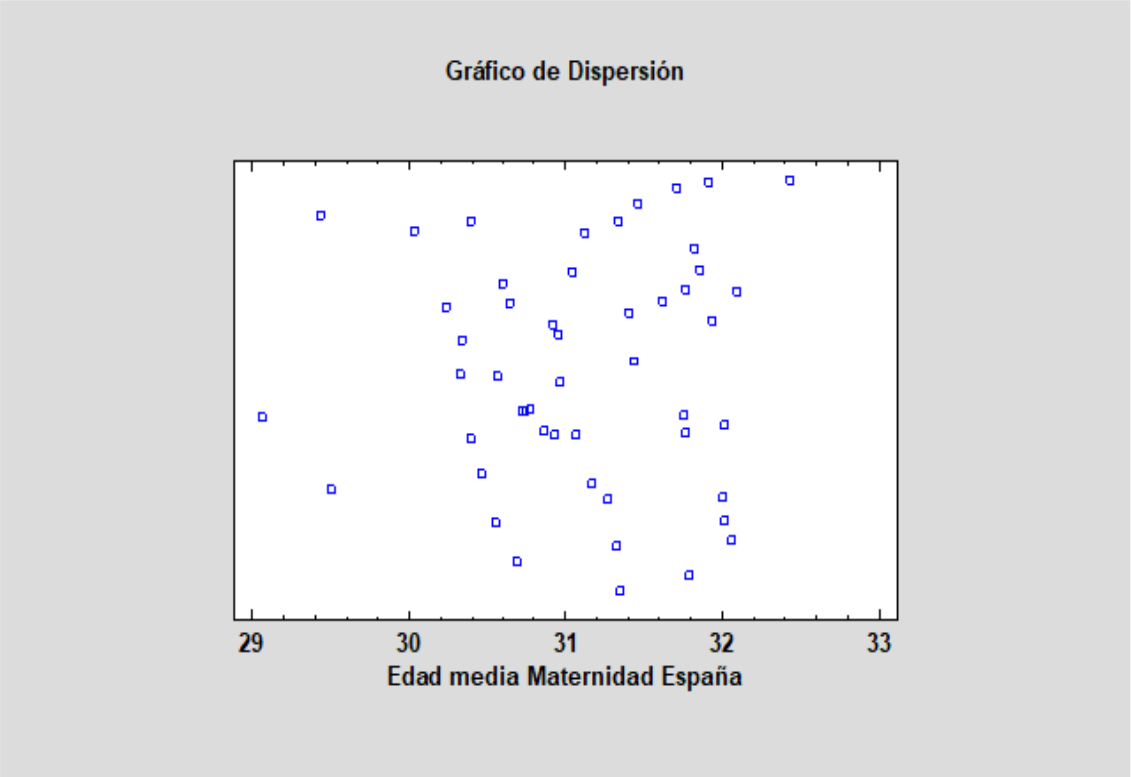
Podemos decir que pocos datos están superpuestos, lo que nos indica que entre los 4 años de diferencia que hay entre extremos, hay cierta dispersión de datos.

La evolución de la edad media a la que se es madre por primera vez en España no ha cesado de crecer en el periodo comprendido entre 1975 y 2019. De hecho, la edad media a la que las mujeres tienen a su primer hijo es casi 6 años más, que hace 46 años. Esto es debido a factores sociales y económicos como pueden ser el incremento de mujeres en el mercado laboral y al aumento de la calidad y esperanza de vida a parte de la precarización laboral de los

jóvenes, su dificultad para poder conciliar vida laboral y personal y el escaso apoyo público a las familias.

La zona norte con provincias de País Vasco, Galicia, Castilla y León, además de la capital (Madrid), son los que de media tienen más tarde a su primer hijo, mientras que zonas del sur y fuera de la península como pueden ser Ceuta, Melilla y Las Palmas tienden a tener a su primer hijo más jóvenes. Vemos así cierta asimetría hacia la izquierda.

Gráfico 2.7: Edad media de maternidad en España por CCAA.



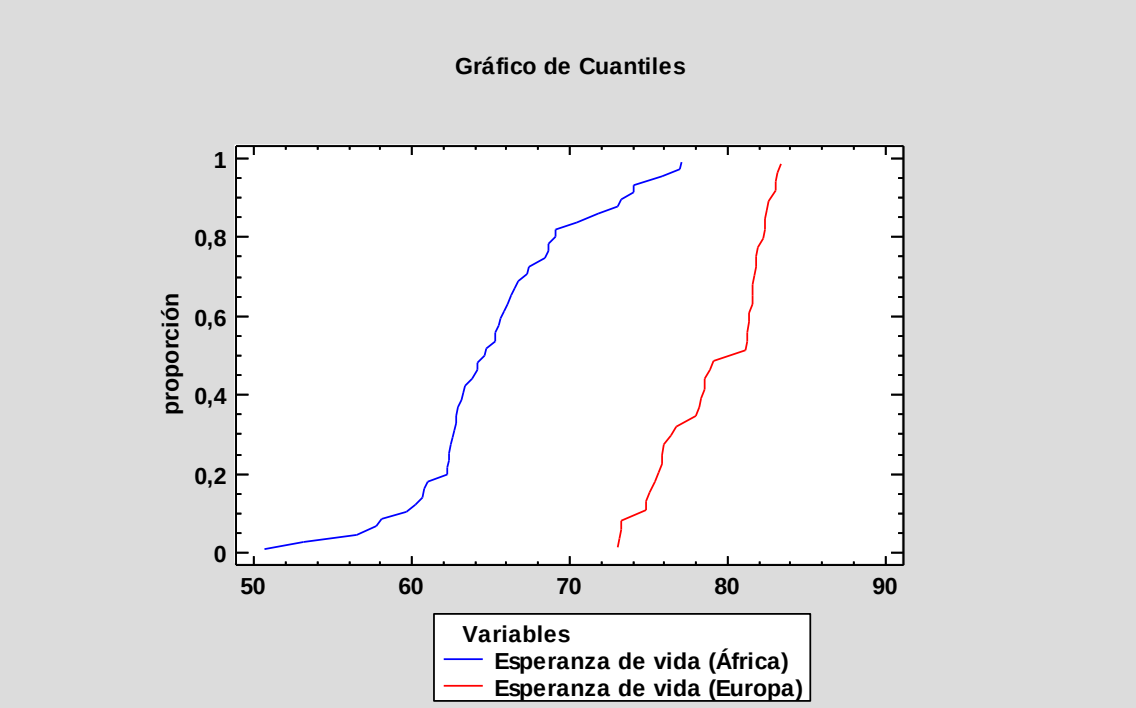
4. Gráfico de Cuantiles

Este tipo de gráfico muestra las frecuencias acumuladas del conjunto de valores mediante el trazado de una línea uniendo por segmentos los puntos sucesivos. Dado que lo que se representa son las frecuencias acumuladas, la curva es siempre creciente. Para representarse se construye sobre un sistema de ejes coordenados. En el eje de abscisas se marcan los intervalos de proporción que hayamos determinado previamente, y en el eje de ordenadas trazamos las frecuencias acumuladas.

Utilizamos este tipo de gráfico para representar datos de poblaciones tanto temporales como atemporales. A cada valor de la variable le corresponde su frecuencia más la de todos los datos anteriores a él. El último valor al que llega la línea es el total de datos que es el 100%.

Si a varios valores seguidos les pertenece el mismo valor en la curva, significa que la frecuencia del segundo valor es igual a cero.

Gráfico 2.8: Comparación de la esperanza de vida entre África y Europa.

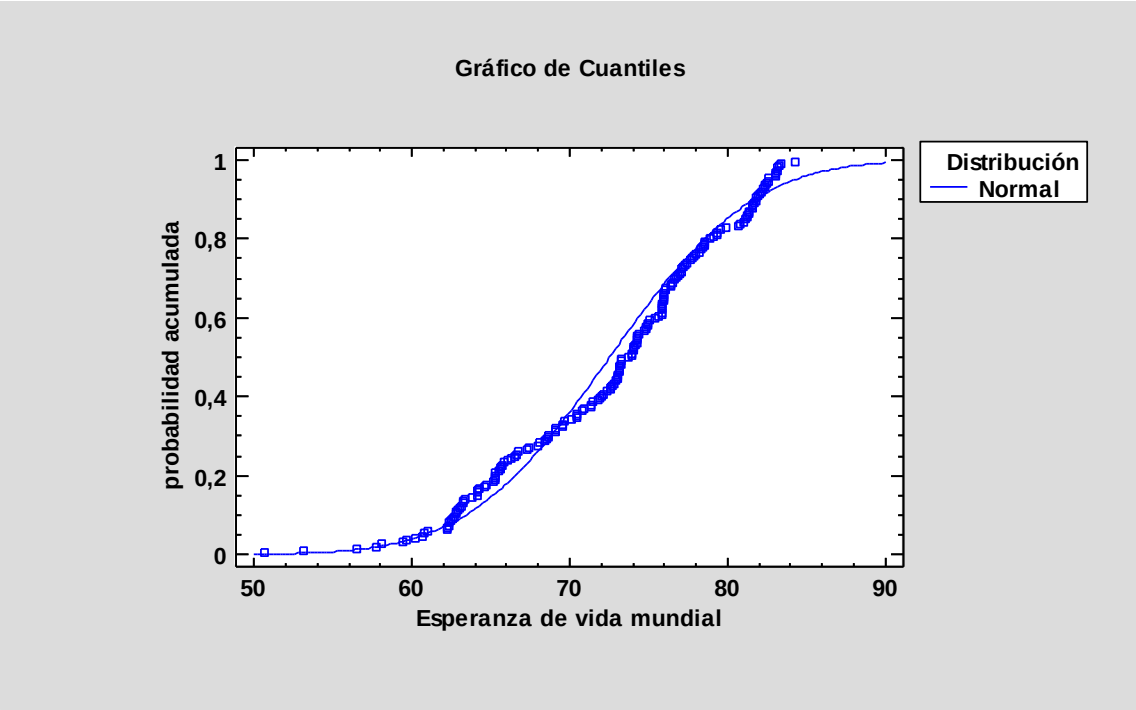


En este ejemplo comparamos la esperanza de vida de África y Europa, observándose a primera vista como la esperanza en Europa es mayor. Esto es fácil de observar ya que la curva de frecuencias acumuladas empieza más tarde en un continente que en otro. Podemos decir que mientras que el 50% de los países africanos vive como mucho 64 años aproximadamente, en los países europeos el 50% de los países tiene como mucho 78 años.

La línea que representa la esperanza de vida de Europa crece a un ritmo mucho más rápido porque los datos están menos dispersos que los de África, es decir, mientras que los datos europeos son 42 valores que van desde un mínimo de 73 años hasta un máximo de 83,4, los datos de la otra variable son 53 valores que se mueven entre un 50,7 y un 77,1.

Analizando el siguiente ejemplo en el que se compara la esperanza de vida de todos los países del mundo con una distribución normal para evaluar su grado de normalidad, no podemos rechazar la idea de que la variable proviene de una distribución normal con un 95% de confianza debido a que el P- valor más pequeño de las pruebas realizadas es mayor o igual a 0,05 (P- valor es 0,115195).

Gráfico 2.9: Esperanza de vida mundial.



5. Gráfico Cuantil-Cuantil

También llamados gráficos Q-Q, son una herramienta del AED para determinar las posibles similitudes o semejanzas entre la distribución de una variable numérica y una distribución normal, pero también se puede utilizar para dos distribuciones de dos variables numéricas.

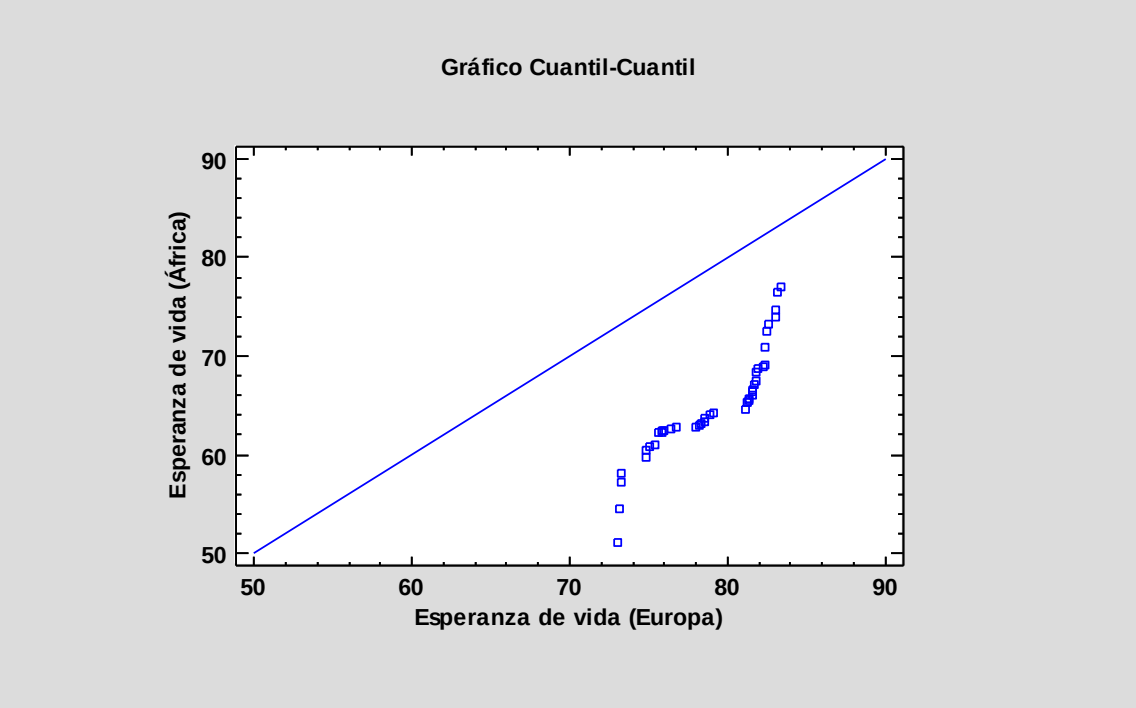
Existen 2 tipos de gráficos de cuantil-cuantil:

- Normales: Este tipo de gráfico se realiza trazando los cuantiles de una variable numérica respecto a los de una distribución normal.
- Generales: Estos en cambio trazan los cuantiles de una variable numérica respecto de los de una segunda variable numérica.

El siguiente gráfico ilustra los gráficos que denominamos generales. En él se representan los cuantiles para la esperanza de vida de Europa y África. En caso de que las distribuciones fuesen iguales, los puntos del diagrama construirían la bisectriz del primer cuadrante mientras que cuanto más se alejasen de dicha línea recta, más diferencias habría entre ambas distribuciones.

La densidad o concentración local de los datos es medible a través de la pendiente local del gráfico, es decir, cuanto menor sea la pendiente, mayor será la densidad de puntos de la variable en ordenadas y viceversa. La mayor densidad local de puntos se produce cuando hay muchas mediciones con exactamente el mismo valor, revelándose en el gráfico mediante una serie de puntos.

Gráfico 2.10: Comparación Esperanza de vida en Europa con Esperanza de vida en África.

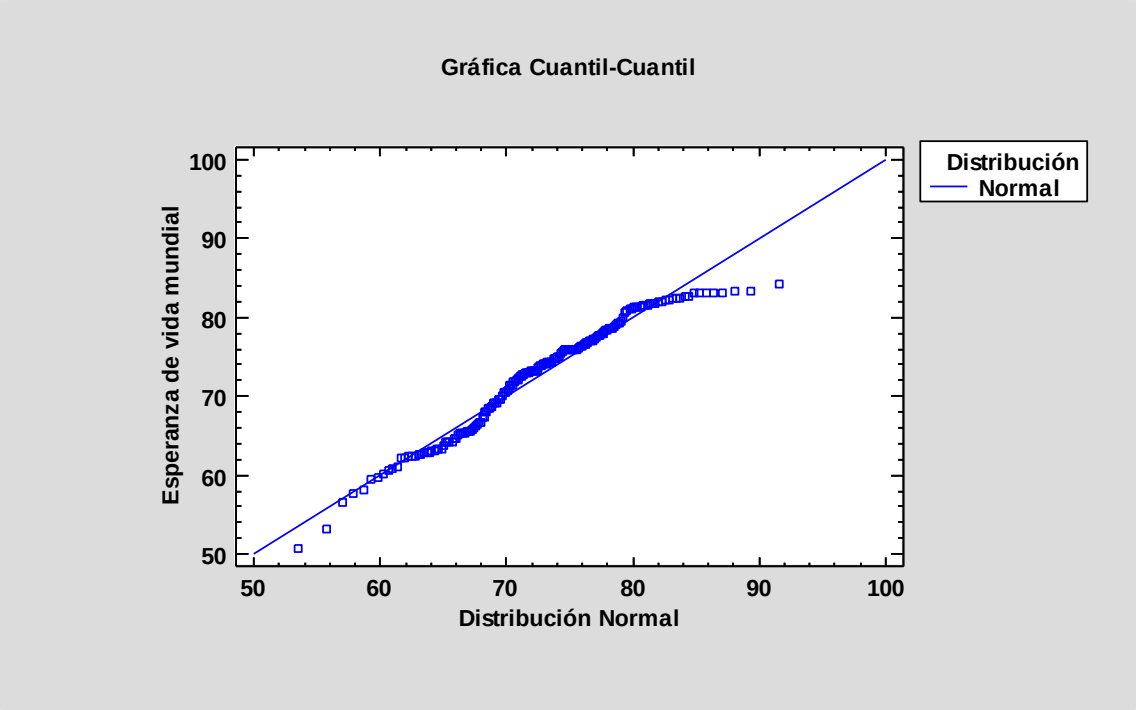


Si ambas muestras proviniesen de una misma población, los puntos podrían estar sobre la línea diagonal, en cambio, todos los puntos por debajo de la línea y en la zona derecha quiere decir que la media de la Esperanza de vida de Europa es más grande que la de África. En cambio la dispersión sería más grande en la variable africana que en la europea.

El gráfico de cuantil-cuantil nos permite investigar los efectos de las transformaciones logarítmicas y de raíz cuadrada en la distribución de datos y compararlos con la distribución normal.

En el siguiente gráfico podemos observar si existe normalidad en nuestra muestra ya que compara los datos de la variable en un eje, con la normal que mejor se ajuste a dicha variable.

Gráfico 2.11: Existencia de Normalidad en la variable Esperanza de vida a nivel mundial.



Puede ser que los datos sigan un tipo de distribución asimétrica, y haya que transformar dichos datos para que se distribuyan normalmente (algunos métodos de análisis lo requieren). Hay 2 tipos de transformaciones en este caso:

- La transformación logarítmica la usamos cuando los datos están sesgados de forma positiva y muchos de sus valores son muy grandes. Si en el conjunto de datos existen estos valores grandes, esta transformación le ayudará a que las varianzas sean más constantes y normalizará sus datos.

- En segundo lugar, la transformación de raíz cuadrada que es bastante similar a la anterior respecto a que reduce el sesgo derecho del conjunto de datos. Este tipo de transformación se puede aplicar a cero a diferencia de la logarítmica.

El gráfico de cuantil-cuantil es una buena visualización del conjunto, ya que es bastante fácil de construir y hace un buen trabajo al representar muchos aspectos de una distribución. Tiene 3 características fundamentales:

1. Al construirlo, no hacemos elecciones arbitrarias de valores de parámetros o límites de celda, y no se ajustan ni se asumen modelos para los datos.
2. Al igual que ocurre con una tabla, no es un resumen sino una visualización de todos los cuantiles.
3. Cada punto se representa en una ubicación distinta, incluso si hay duplicados exactos en los datos. El número de puntos que se pueden representar sin superposición está limitado solo por el programa utilizado para el trazado del gráfico.

6. Gráfico de simetría:

Este tipo de gráficos se utilizan para saber si la variable sigue una distribución de tipo normal, lo que implica, entre otras cosas, tener una apariencia simétrica. Una distribución es simétrica si ambos lados de la mediana tienen el mismo aspecto, es decir, las colas son estrictamente iguales. Cuanto más simétricos sean los datos, los puntos estarán entonces más cerca de la bisectriz. A veces, aunque la distribución siga una forma normal, pueden existir una serie de puntos que puedan estar por encima o debajo de la línea.

En cambio, unos puntos muy alejados de esta línea pueden ser existencia de datos atípicos.

En este ejemplo hemos creado 100 datos aleatorios mediante Excel y después lo hemos transformado en una Normal con media 100 y desviación típica 2 para conseguir el gráfico de Simetría y debajo un Histograma para observar sin dificultad la distribución.

Los puntos que divergen de la línea por debajo indican una cierta asimetría a la derecha que se puede ver después más abajo en el Histograma.

Gráfico 2.12: Comprobación de Normalidad a través de un gráfico de simetría a partir de datos aleatorios.

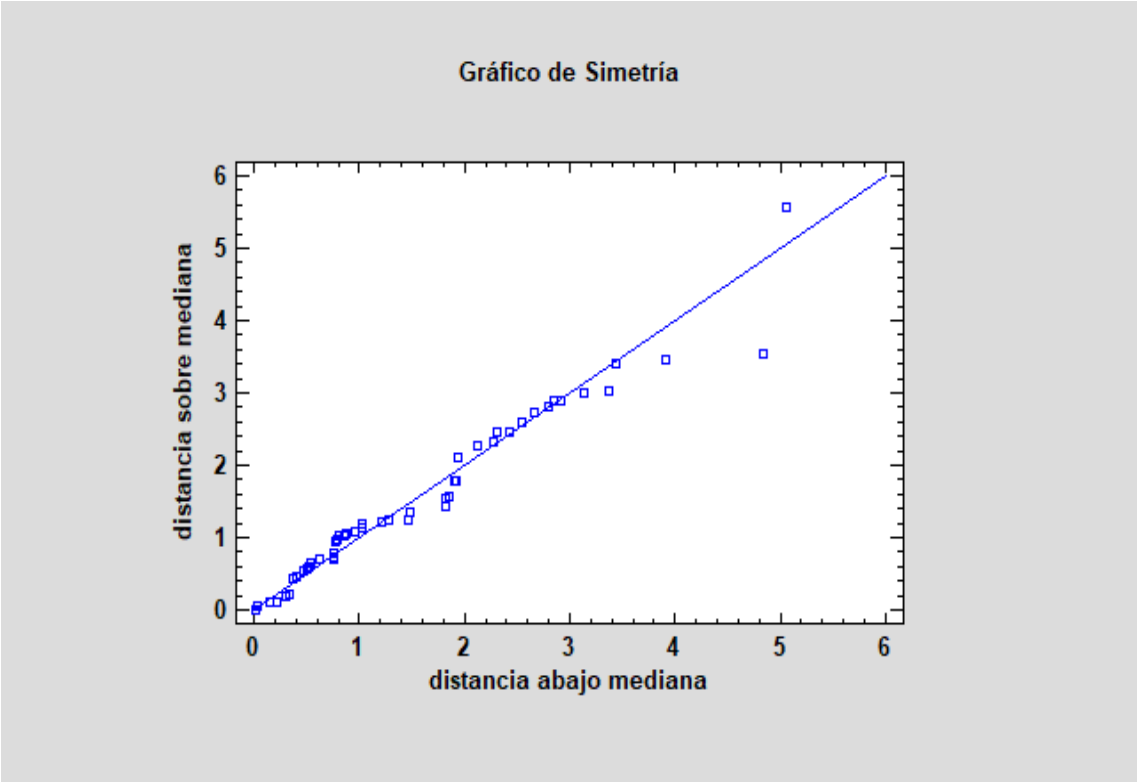
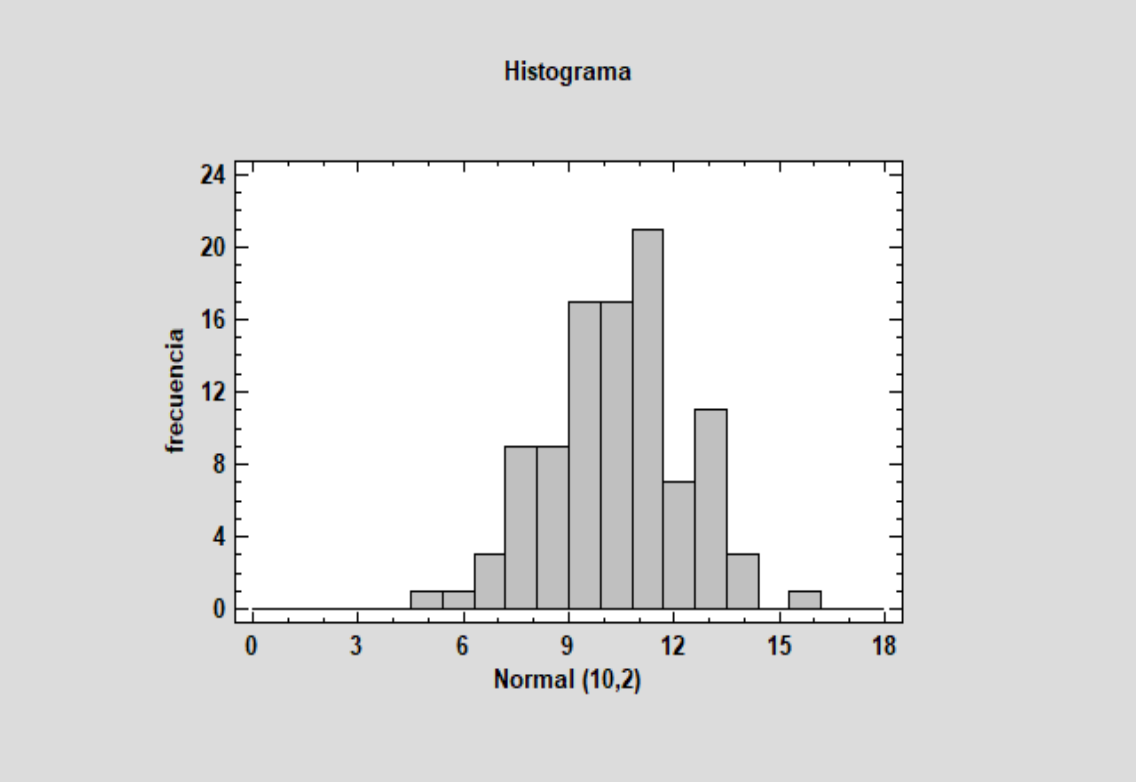


Gráfico 2.13: Histograma del Gráfico 2.12



7. Diagrama de caja y bigotes:

Es una presentación visual que muestra al lector varias características relevantes a la vez, como pueden ser la posición, la dispersión o la simetría.

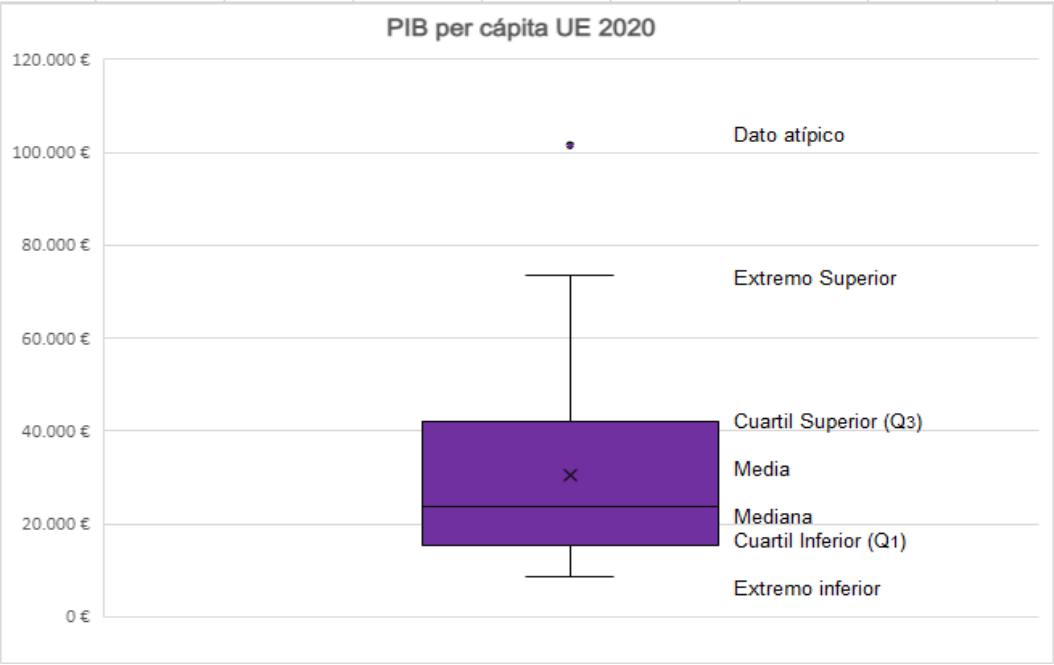
El largo del rectángulo o caja es el recorrido intercuartílico (Ri) que es la diferencia entre el tercer cuartil y el primer cuartil. Esta caja está dividida en dos partes por la mediana, que no tiene por qué estar en la mitad del rectángulo. Si no divide el rectángulo por la mitad, quiere decir que un lado, el más pequeño, tiene una concentración de datos mayor que el otro.

Para saber cuáles son estos datos atípicos hay que calcular el límite inferior y superior de los datos, que se extienden 1,5 veces el recorrido intercuartílico desde los extremos superior e inferior de la caja. Estos datos atípicos son aquellos que exceden de los extremos superior o inferior y brindan al lector la oportunidad de considerar las observaciones que parecen inusuales o inverosímiles.

En este caso utilizaremos los datos del PIB per cápita del 2020 en la Unión Europea, ya que los datos muestran muy bien todas las partes de un diagrama de este tipo. El cuartil superior (Q3) es la parte superior del rectángulo siendo esta cifra 41.090€ y el cuartil inferior (Q1) es la parte inferior, que es la cifra 16.085€. La mediana es 23.580€ y el Recorrido intercuartílico es de 25.005€. Sabiendo estos valores podemos descubrir la distancia de los bigotes tal y como hemos explicado antes. El extremo superior calculado como: (41.090 + 1,5*RI) da una cifra de 78.597,5€ y el extremo inferior calculado como (16.085 - 1,5*RI) dando la cifra de –21.422,5€. Como el valor más pequeño de la población es Bulgaria con 8750€, utilizamos ese como límite inferior porque no hay nada más por debajo, mientras que por encima del límite superior, siendo este 78597,5€, tenemos a Luxemburgo con un PIB per cápita de 101.640€ siendo este un valor atípico.

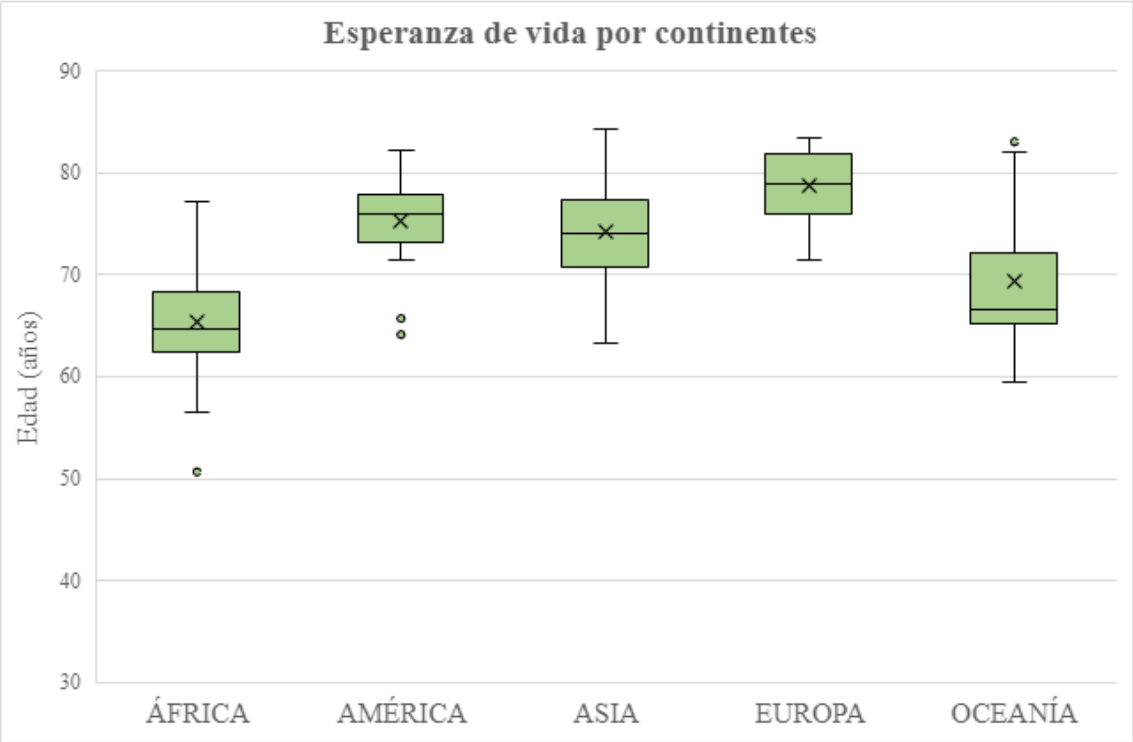
También podemos observar a través del gráfico que no sigue una distribución simétrica ya que ni la mediana corta por la mitad de la caja ni ambos bigotes tienen la misma longitud, esto quiere decir que la mayoría de los datos se acumulan en la zona inferior del gráfico y es asimétrica hacia la derecha.

Gráfico 2.14: PIB per capita en la Unión Europea (2020).



Los diagramas de caja y bigotes son útiles en situaciones en las que no es necesario o no es factible mostrar todos los detalles de la distribución.

Gráfico 2.15: Esperanza de vida por continentes.



En este gráfico de caja y bigotes analizaremos la esperanza de vida para los 5 continentes por separado (año 2019) y observaremos las diferencias existentes en los datos extremos, medias y datos atípicos.

A simple vista observamos que África tiene la esperanza de vida más baja de todos los continentes globalmente hablando, que América tiene el Recorrido Intercuartílico más pequeño con 4,6 años y Oceanía el más grande con 6,85 años. Los Ri de África y Europa coinciden con un 5,9, es decir la diferencia entre primer y tercer cuartil son la misma, mientras que la acumulación de datos en el continente africano es mayor en una parte que en otra ya que la mediana no divide la caja en dos partes iguales. En cambio, en el ejemplo europeo sí.

África tiene un valor atípico por debajo de su extremo inferior que sería Lesoto con un 50,7 de esperanza de vida, siendo además el valor más bajo de todo el mundo.

América en cambio tiene 2 datos atípicos que son Guyana y Haití con un 65,7 y 64,1 respectivamente. Aunque estos datos atípicos solo son para este continente ya que en Asia, Oceanía o África serían datos normales.

El último continente con algún dato atípico, pero en este caso por encima de su límite superior es Oceanía, en el que Australia tiene una esperanza de vida de 83 años.

El país que tiene una mayor esperanza de vida es Japón con un valor de 84,3 años que coincide a su vez con el límite superior de Asia.

En cuanto a la media podemos decir que Europa es el continente que más alta tiene la esperanza de vida de media con un 78,76 mientras que el continente con la menor media es África con un 65,36.

Para terminar, también podemos observar que ninguno sigue una distribución simétrica porque ninguna mediana se sitúa en la mitad de los rectángulos ni ninguno de los bigotes de ningún continente es de igual longitud. En el caso de África, América y Oceanía son asimétricas hacia la derecha, Europa claramente hacia la izquierda. Asia es asimétrica hacia la izquierda, pero levemente, ya que la mediana casi coincide con la mitad del rectángulo.

Podemos añadir que la dispersión de las variables es muy pequeña ya que el tamaño de la caja es relativamente pequeño.

3. ANÁLISIS EXPLORATORIO DE DATOS PARA DOS VARIABLES

Podemos realizar el análisis bidimensional para observar las posibles relaciones que hay entre dos variables a las cuales hayamos realizado el estudio unidimensional previamente. Nos encontraremos con 3 situaciones diferentes:

- Las dos variables son cualitativas
- Las dos variables con cuantitativas
- Una variable es cuantitativa y la otra cualitativa

Ambas variables son cualitativas: En este caso usaremos una tabla de contingencia, para estudiar si las dos variables son o no independientes, donde venga representada la frecuencia conjunta que representa el número de datos que pertenecen a la modalidad i-ésima de la primera variable y modalidad j-ésima de la segunda. Una vez analizadas, puede que no haya relación alguna entre las dos variables, pero en caso de que sí que la hubiese, comprobamos el tipo y el grado de su dependencia de forma gráfica y numérica.

Ambas variables son cuantitativas: En este caso podemos usar un diagrama de dispersión para la distribución conjunta que nos proporcionará la relación que pueda haber entre dichas variables estudiadas.

La estadística clásica, basada en la idea de normalidad indica que esta relación también puede ser expresada de forma numérica resumiendo la información del gráfico de dispersión y sin depender de las unidades de medida a través del coeficiente de correlación lineal. Este será igual a uno en valor absoluto si ambas variables están relacionadas de forma exacta y cuando no lo están, el coeficiente lineal es igual a 0. Aun así, siempre se recomienda observar el diagrama de dispersión para poder interpretar correctamente el coeficiente ya que así podremos verificar que los datos sean uniformes y no existan valores atípicos. Que exista una correlación no significa que haya una relación de causalidad entre las variables, y al revés, generalmente la ausencia de correlación no significa que no exista una relación de causalidad no lineal. Al estudiar la relación entre variables, debemos asegurarnos de que los individuos estudiados sean homogéneos respecto a dichas variables.

No obstante, cuando no puede presuponerse la idea de normalidad, hay alternativas basadas en estadísticos más robustos, como la recta de Tukey, que describimos más adelante.

Una variable cuantitativa y la otra cualitativa:

Cuando existe una variable de cada tipo, la investigación se convierte en comparar el comportamiento de la variable numérica en las diferentes subpoblaciones definida por la variable cualitativa. Podemos llegar a conclusiones erróneas si ignoramos la heterogeneidad debido a la existencia de subpoblaciones. Para realizar este análisis podemos usar el diagrama de caja y bigote y/o el diagrama Q-Q. El gráfico de esperanza de vida por continentes anterior muestra un diagrama de caja y bigotes múltiple que ilustra esta situación.

3.1 La recta de Tukey

Este método, llamado también “método de la línea resistente de tres grupos” (Emerson y Hoaglin, 1983) es aquella técnica que nos permite determinar la relación lineal entre dos variables. Para proceder a su cálculo tenemos que recurrir al estadístico de la mediana, que al no tener en cuenta todas las puntuaciones (no se ve afectada por observaciones extremas), solo las centrales, nos ayuda a construir la recta de Tukey de una forma resistente.

Para la correcta construcción de esta técnica debemos seguir una serie de pasos previos:

1. Organizar los valores existentes en orden ascendente en función de la variable X.
2. Dividir el total de observaciones previamente ordenadas de X en tres grupos, donde en cada uno tiene que haber aproximadamente un tercio del total. Diferenciándose así el grupo inferior, el grupo medio y el grupo superior.

La creación de los 3 grupos depende del número total de observaciones:

- Si el total es múltiplo de 3, es decir, al realizar la división no se dan decimales, cada grupo puede tener el mismo número de observaciones.

- Si la división nos da un resultado con decimales de 3 periódico, en los grupos inferior y superior habrá k observaciones (donde k es un número entero) y en el grupo medio k+1. Es decir, en el primer y último grupo habrá un número de observaciones igual a la parte entera de la división del total entre tres, y en el grupo medio, la parte entera más uno.
- Si la división nos da un resultado con decimales de 6 periódico, sería a la inversa que el caso anterior, habrá k observaciones en el grupo medio y k+1 observaciones en los demás.

3. Hallar las medianas de las variables X e Y de los tres grupos diferenciados.

Una vez tengamos estos pasos realizamos un gráfico en el que representamos un diagrama de dispersión de ambas variables. Tratando de ver si siguen una relación lineal.

Para un mejor trabajo, la relación de ambas variables tiene que seguir tres aspectos fundamentales:

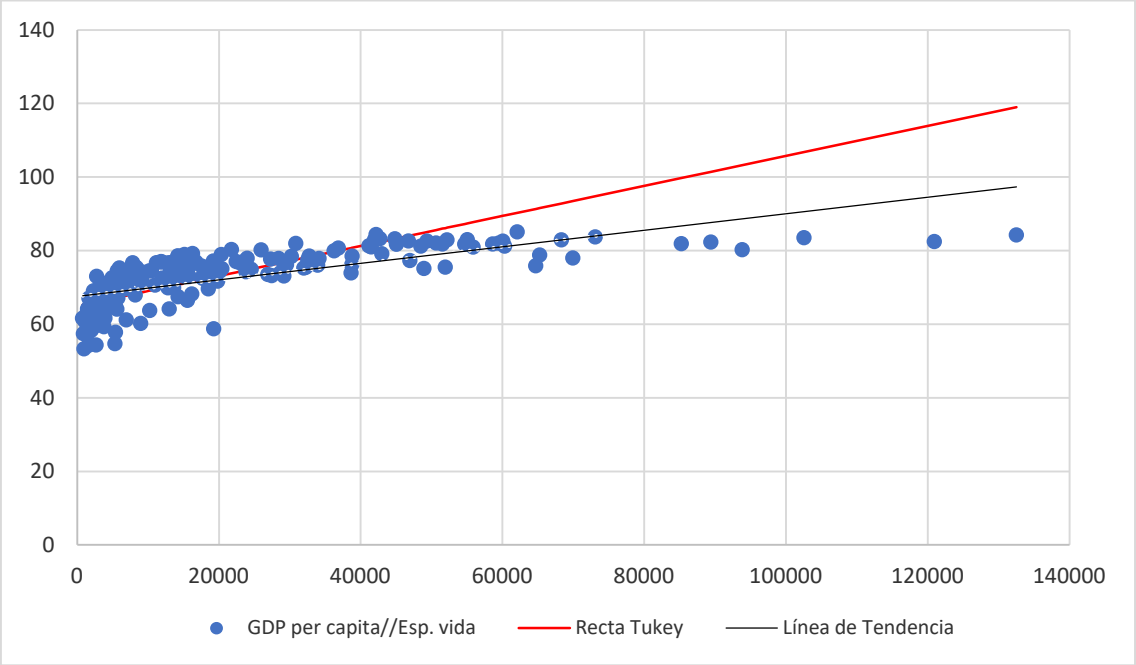
- Una fuerza de relación buena, es decir, el grado de ajuste de los datos a la hipotética recta que los enlazará. A mayor concentración de datos o más cerca de la recta imaginaria estén, una mejor relación existirá.
- Observar la dirección de la relación. En casos de relación lineal, a un valor mayor de la variable X, se le acredita otro mayor de la variable Y, esto dará lugar a una pendiente positiva y, por tanto, sería una relación positiva o directa. En el caso contrario, en el que a un menor valor de una variable se le acredita un menor valor de la otra variable, se dice que tiene una dirección negativa o inversa.
- Analizar la forma de la distribución, que puede ser lineal o curvilínea en base a como se relacionen ambas variables.

Como un ejemplo, hemos estudiado la relación entre el PIB per cápita y la esperanza de vida. Para facilitar la comparación, hemos calculado y representado adicionalmente la recta de regresión mínimo cuadrática.

Después pasamos a hacer los cálculos procedentes para hallar la recta de Tukey, que seguirá también una función lineal $Y = ax+b$. De la que obtendremos

los mismos datos que antes, es decir, pendiente y ordenada en el origen, pero con valores y fórmulas diferentes.

Gráfico 3.1: Comparación Línea de tendencia y Recta de Tukey.



En este gráfico podemos observar la nube de puntos de las variables PIB per cápita (nuestra variable X) y Esperanza de vida (nuestra variable Y) a nivel mundial (k=184 observaciones).

Para este caso, al tener 184 datos que es un número que al dividir entre los tres grupos tiene decimales con 3 periódico, los grupos primero y tercero tendrán 61 observaciones mientras que el grupo medio tendrá 62 como explicamos anteriormente.

Una vez diferenciados los grupos se hallan las medianas de todos los grupos, siendo estas:

MdX_1=	3503,61755	MdY_1=	64,57
MdX_2=	13933,2188	MdY_2=	74,2404146
MdX_3=	43005,5533	MdY_3=	80,6829268

Al obtener estos datos realizamos la línea de tendencia sobre el gráfico para obtener la recta de regresión.

Recta de regresión: $Y = 0{,}0002X + 67{,}603$

Pendiente: $b = 0{,}0002$

Ordenada en el origen: $a = 67{,}603$

Antes de seguir con la recta de Tukey, podemos añadir que la relación de ambas variables, como la mayoría de puntos están bastante concentrados en la parte izquierda de la gráfica, tiene un R^2 bastante pequeño (un 0,5016). Antes de hacer la línea de regresión se sabía que iba a tener una pendiente positiva ya que la dirección de los puntos es lineal y directa.

Seguidamente, pasamos a calcular la Recta de Tukey, cuyos procedimientos matemáticos son los siguientes:

Pendiente[1]: $b = \dfrac{MdY_3 - MdY_1}{MdX_3 - MdX_1}$ $\quad b = 0{,}0004079$

Ordenada en el origen: $a = \dfrac{\sum_{i=1}^{3}(MdY_i - b * MdX_1)}{3}$ $\quad a = 64{,}9462524$

Recta de Tukey: $Y = 0{,}0004079X + 64.9462524$

Una vez obtenida la Recta observamos en el gráfico que su pendiente es un poco más grande que la pendiente de la recta de regresión por lo que se ve que la línea roja crece mucho más rápido que la recta negra. Dicha diferencia entre ambas rectas es debida a la existencia de valores extremos o atípicos para el PIB per cápita.

[1] MdY es la Mediana de la Variable Y
MdX es la Mediana de la Variable X

CONCLUSIONES

En este trabajo hemos descrito las herramientas básicas del Análisis Exploratorio de datos, fundamentalmente para el análisis unidimensional.

En primer lugar, hemos descrito el objetivo, método y pasos para realizar el AED, distinguiéndolo de la E. descriptiva y de la inferencial.

Posteriormente hemos definido un total de siete herramientas fundamentales empleadas en el análisis con sus diferentes características, aplicándolas a datos obtenidos de las estadísticas económicas y sociales como pueden ser el PIB y el PIBpc por países, la esperanza de vida por países o la tasa de fertilidad, entre otros.

Finalizamos este trabajo con una pequeña descripción de las estrategias de AED bidimensional encontrándonos tres situaciones diferentes y haciendo hincapié en el caso de dos variables cuantitativas en el que usamos la recta de Tukey para comprobar si siguen una relación lineal.

BIBLIOGRAFÍA:

Ciberconta.unizar.es. 2021. Available at: <http://www.ciberconta.unizar.es/leccion/aed/ead.pdf>

Chambers, J., 1985. GRAPHICAL METHODS FOR DATA ANALYSIS. BELMONT, CALIF: WADSWORTH (U.A.).

UE - Unión Europea 2021. (2021). datosmacro.com.

https://datosmacro.expansion.com/paises/grupos/union-europea

PIB - Producto Interior Bruto 2021. (2021). datosmacro.com.

https://datosmacro.expansion.com/pib

INE - Instituto Nacional de Estadstica. (2021). *INE. Instituto Nacional de Estadística*. INE.

https://www.ine.es/

FMI - Fondo Monetario Internacional 2021. (2021). datosmacro.com.

https://datosmacro.expansion.com/paises/grupos/fmi

Statista. (2021, 21 septiembre). *Tasa de fertilidad mundial 2004–2019*.

https://es.statista.com/estadisticas/657184/tasa-de-fertilidad-a-nivel-mundial/

Indice de Desarrollo Humano - IDH 2019. (2019). datosmacro.com.

https://datosmacro.expansion.com/idh

Wikipedia contributors. (2010–2021). *Wikipedia*. Wikipedia. https://www.wikipedia.org/

Batanero, C., Estepa, A., & Godino, J. D. (1991). *Análisis exploratorio de datos: sus posibilidades en la enseñanza secundaria*.

https://www.ugr.es/~batanero/pages/ARTICULOS/anaexplora.pdf.

https://www.statgraphics.com/