Tema 4:

El tipo de variabilidad condiciona el modo de analizarla

Introducción.....1

Criterios para la elección de herramientas estadísticas.....5

Del histograma a los polígonos de frecuencias 5

Análisis de la forma de la distribución..... 6

Recapitulación..... 13

Tema 4:

El tipo de variabilidad condiciona el modo de analizarla

INTRODUCCIÓN

Llegado este momento, hemos visto ya numerosos ejemplos de situaciones caracterizadas por su variabilidad. Hemos analizado también el procedimiento para descubrir y valorar el grado de variabilidad de una serie de datos. Una de las conclusiones fundamentales a la que hemos podido llegar es que, después de efectuar el análisis preliminar de los datos y de la distribución interna de los valores, el autor del análisis conoce bien las características principales de la distribución.

Después de realizar el análisis preliminar de los datos correspondientes a la renta personal disponible de los municipios de Euskadi en 2001, estamos preparados para elaborar una descripción de las características fundamentales de los datos y de su distribución:

- A la vista del amplitud del rango de la variable, se puede concluir que la variabilidad es alta, es decir, que hay importantes diferencias en la renta de los distintos municipios de la comunidad autónoma.

- La elevada variabilidad que muestra el rango de la variable se debe en gran medida a la existencia de un valor extremo, esto es, a la renta de un municipio (Laukiz), que es muy superior a la del resto de municipios.

- Pese a que la variabilidad total es muy elevada, existe un colectivo mayoritario de municipios que tuvieron valores de renta no tan diferentes entre sí. Se puede concluir que, en realidad, fueron pocos los municipios que tuvieron rentas bastante más altas o más bajas que las del colectivo mayoritario.

- La renta personal disponible de muchos de los municipios se situó en un rango comprendido entre los 8.500 y los 12.000 euros.

- Si miramos los municipios destacables por su diferencia en el valor de la

renta, vemos que fueron más los que tuvieron rentas bajas que los que las tuvieron altas.

Una vez comprobado que el análisis efectuado proporciona a quien lo realiza una imagen bastante clara de la situación, el paso siguiente es preguntarnos cómo nos la arreglaremos para transmitir la información de modo tal que otras personas puedan hacerse una idea lo suficientemente clara y precisa de la situación sin tener que hacer el análisis.

Está claro que la descripción que acabamos de realizar no resulta adecuada. El problema es que se trata de un conjunto de afirmaciones carentes de precisión; no hemos apoyado nuestras afirmaciones en referentes que permitan hacerse una idea inequívoca de la situación. Decir, por ejemplo, que la variabilidad de un colectivo mayoritario es baja es una afirmación muy ambigua. Necesitamos elaborar una descripción que evite las imprecisiones y los posibles equívocos. Una descripción como la que pretendemos debe cumplir las siguientes condiciones:

- Las afirmaciones deben fundamentarse en conclusiones obtenidas por métodos conocidos y aceptados, de modo que, cualquiera que utilizara dichos métodos pudiera llegar a las mismas conclusiones.

- Se deben explicitar los métodos utilizados y los resultados obtenidos. Cada afirmación debe ir acompañada del resultado que la avala. No sirve, por ejemplo, decir que la variabilidad es elevada si no se mencionan los valores extremos del rango. El concepto de rango es una herramienta estadística que conoce todo aquel que lo utiliza. Cuando lo usamos, de forma explícita, en una descripción, estamos informando al posible lector de la referencia que hemos usado para definir la variabilidad total de una serie de datos.

- Los procedimientos que utilizamos para analizar y describir los valores de una variable deben ser los adecuados a la naturaleza de los datos. Todos los procedimientos estadísticos no son aptos para todo tipo de datos y situaciones.

- La descripción debe mencionar todos los rasgos característicos de los datos y de su distribución. La ausencia de alguno de estos rasgos puede estar ocultando información esencial contenida en los datos.

La Estadística dispone de recursos abundantes para analizar y describir la variabilidad de distintos tipos de datos y variables. Ahora bien, como decíamos,

no todas las herramientas descriptivas son aptas para todo tipo de situaciones y, por ello, es preciso saber elegir, en cada caso, las adecuadas en función de las características de datos y variables. La cuestión es entonces adquirir el conocimiento necesario y desarrollar la capacidad para actuar en cada ocasión del modo acorde a la naturaleza de los datos que estamos manejando.

Es muy importante comprender e interiorizar la importancia que tienen los análisis y descripciones. Para empezar, debemos tener claro que las descripciones de la realidad o de un aspecto de ésta son, en muchas ocasiones, las únicas imágenes que tenemos de la propia realidad. ¿Qué sabe, por ejemplo, un ciudadano *corriente* sobre las estadísticas del desempleo?. Lo habitual es que la información que tenemos los ciudadanos proceda de las descripciones que proporcionan los medios de información.

En la mayoría de casos, las personas que leen o escuchan la descripción de una situación a través de los datos de los estudios que se publican, no tienen acceso a los valores que se han utilizado para su análisis. Esto significa que dichas personas obtendrán sus conclusiones a partir de la información proporcionada en la descripción o en los resultados del estudio que se divulgan. La imagen que proporcionan esos resultados, sea buena o mala, describa bien la situación o la manipule, es la responsable de las ideas que difunde.

No podemos olvidar que los análisis de los que se consideran expertos en distintas materias suelen tener una influencia considerable en la sociedad ya que crean opinión respecto a muchos temas. No ignoramos que los análisis que hacen los críticos de cine, por ejemplo, influyen en la imagen que se difunde de las películas y, en última instancia, en las decisiones de los posibles espectadores sobre la decisión de ir o no a verlas. Algo similar ocurre con los resultados de los estudios estadísticos que se difunden a través de los medios de comunicación. Las administraciones de los estados, entre otras instituciones, difunden constantemente informes estadísticos relacionados con la economía, la política o la sociedad del país. Esos informes, que suelen contener descripciones elaboradas mediante análisis estadístico de datos, tienen una influencia innegable en la opinión de los ciudadanos. Podemos concluir por ello que, en buena medida y en muchas ocasiones, los resultados de los estudios estadísticos constituyen una especie de intermediario entre la realidad y el conocimiento de ésta que tienen las personas.

Las reflexiones que hemos realizado en esta introducción han intentado poner en evidencia tres cuestiones innegables:

- Las descripciones estadísticas de la realidad o de algún aspecto de la misma constituyen, en muchos casos, la representación de una realidad cuyas características detalladas desconocemos.

- Las descripciones que hacemos de la realidad, sobre todo cuando éstas repercuten en la propia realidad, deben de estar apoyadas, en la mayor medida posible, en evidencias inequívocas y en referentes aceptados y comprendidos por los miembros de la sociedad.

- Una buena descripción estadística debe ser exhaustiva y, sobre todo, debe estar basada en herramientas analíticas acordes al tipo de información que ha utilizado.

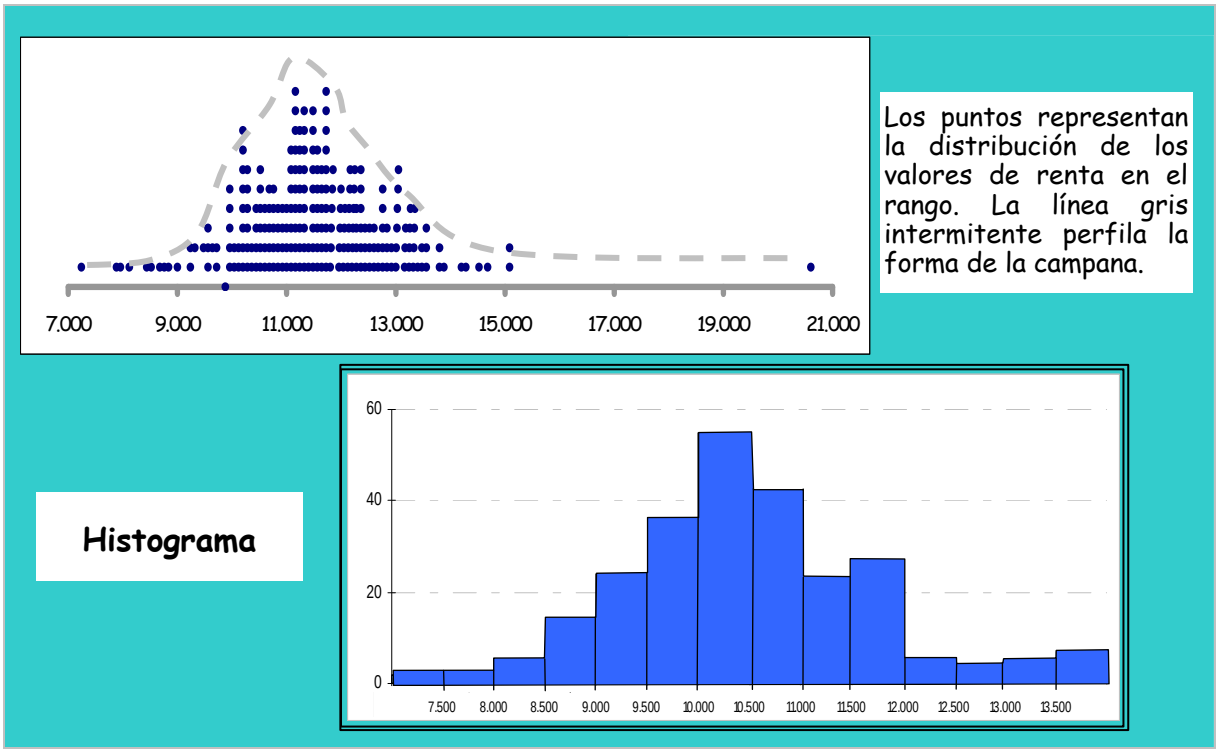
En las páginas que siguen presentaremos los fundamentos que nos permitan comprender cómo elegir los procedimientos estadísticos de análisis adecuados en función de la naturaleza de los datos que se estudian.

CRITERIOS PARA LA ELECCIÓN DE HERRAMIENTAS ESTADÍSTICAS

Los procedimientos estadísticos para el análisis de la variabilidad son distintos en función de la naturaleza de los datos y de las propiedades de la distribución de valores de la variable. El método adecuado para para saber qué procedimientos debemos utilizar en cada ocasión consiste en mirar las características de la distribución.

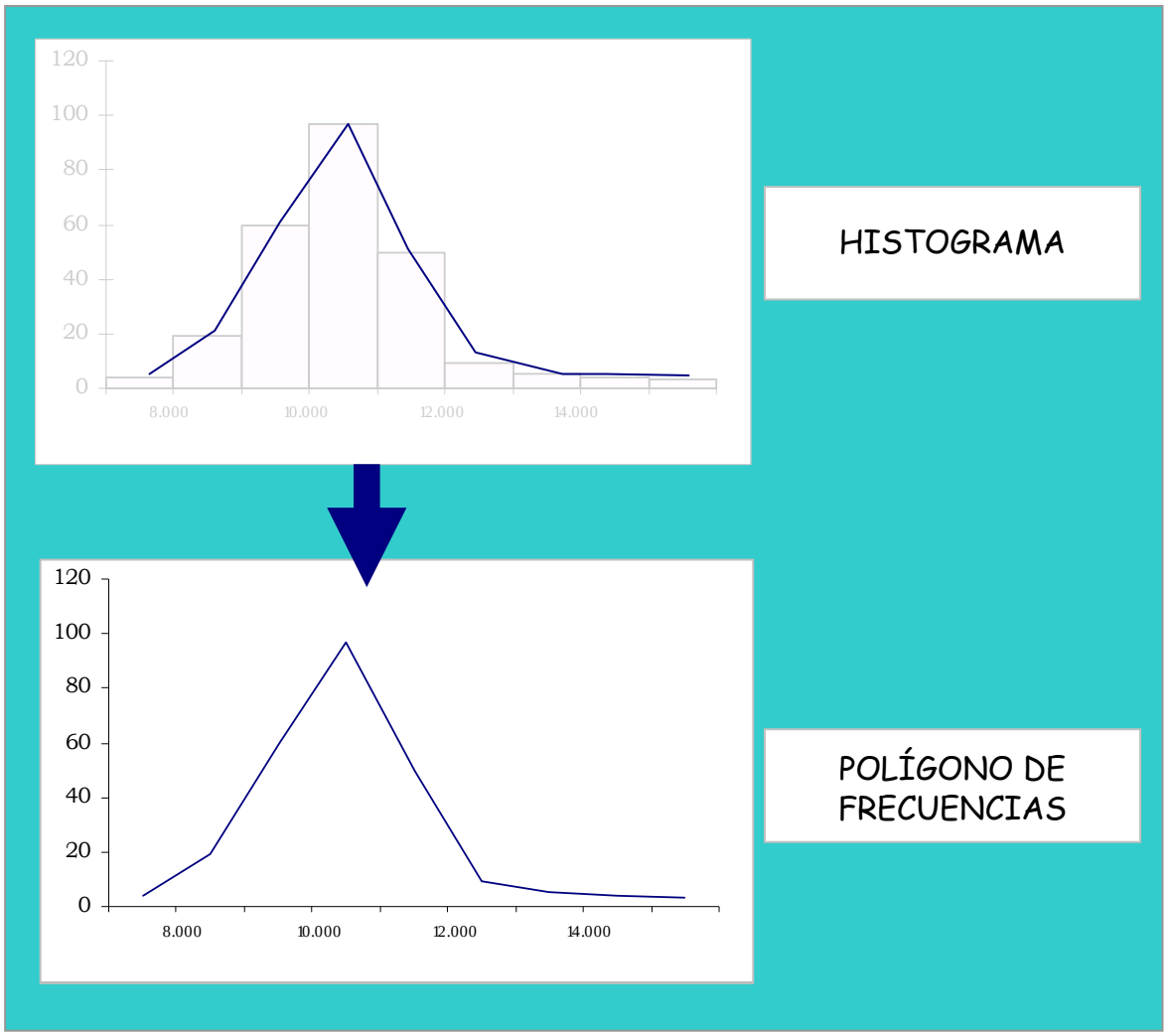
Del histograma a los polígonos de frecuencias

Para explicar el concepto de distribución de valores de la variable hemos utilizado los datos correspondientes a la renta personal disponible de los municipios de Euskadi en 2001. Con estos mismos valores hemos explicado posteriormente el histograma. En las dos ocasiones hemos tenido la oportunidad de ver que los datos de la variable se ajustaban a una distribución en forma de campana.



La forma perceptible de campana que dibujan los datos, se aprecia todavía de forma más clara cuando trazamos una línea que perfila el contorno de la distribución. De hecho, existe un tipo de gráfico pensado para este efecto. Nos referimos al *polígono de frecuencias*, para cuya realización el punto de partida es el propio histograma.

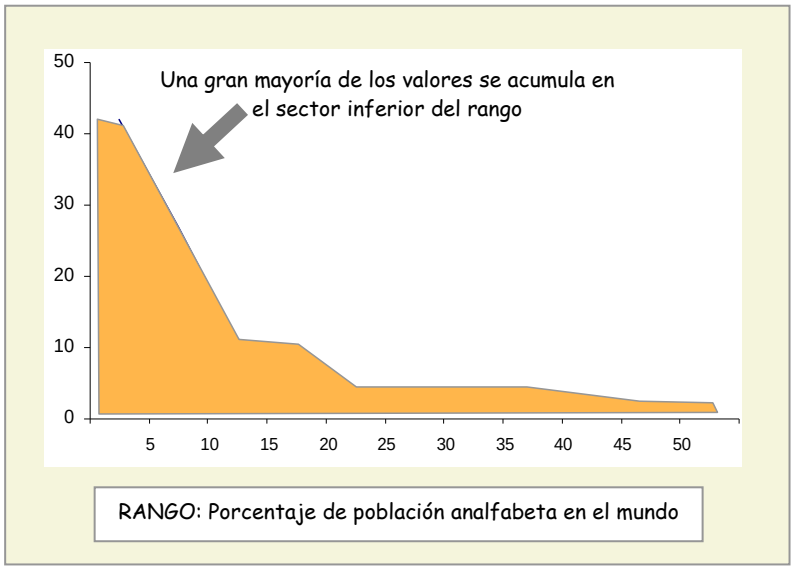
Como podemos ver en la ilustración, el procedimiento para elaborar el *polígono de frecuencias* consiste simplemente en trazar una línea que une el punto central de la parte superior de cada una de las barras que componen el histograma.



Análisis de la forma de la distribución

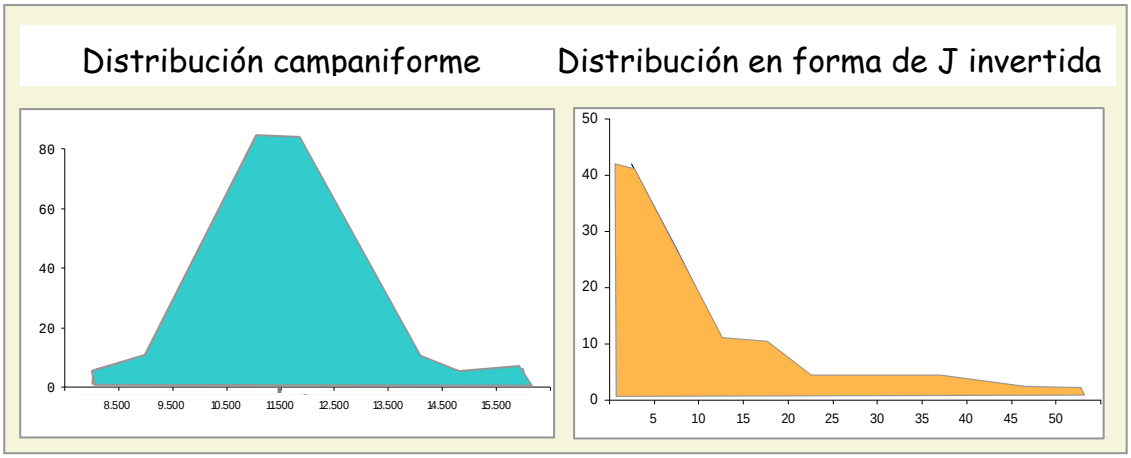
Aunque existe gran cantidad de variables cuyas distribuciones se ajustan a la forma de campana, existen también una buena cantidad de ellas que presentan otro tipo de formas. La forma de la distribución no es otra cosa que un reflejo del tipo de variabilidad de los datos; diferentes tipos de variabilidad producen distribuciones de formas distintas. En este sentido, es importante darse cuenta de que, aunque utilizamos la forma de la distribución como referente, lo importante es el significado de la misma, es decir, la información que nos proporciona la forma respecto a las características de los datos que estamos estudiando.

Para comprender la relación entre la forma y el fondo, es decir, entre la forma de la distribución y su significado con respecto al tipo de variabilidad, lo mejor es analizar y comparar distribuciones de diferente forma. Es lo que haremos a continuación.



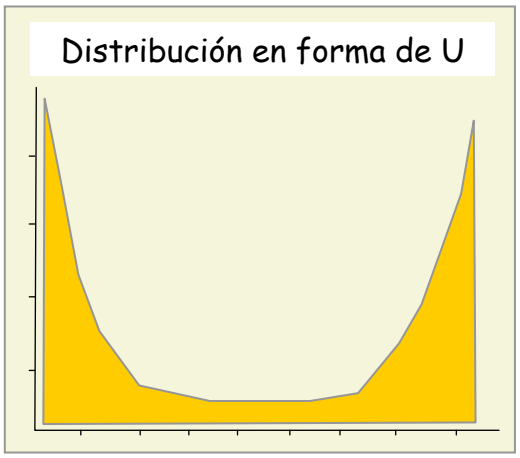
El gráfico que vemos en esta ocasión muestra el porcentaje de población analfabeta en distintos países del mundo. En el eje horizontal (abcisas) está representado el rango de la variable, que nos da una idea de la variabilidad de la población analfabeta: mientras en algunos países el porcentaje es inferior al 5%, en otros se supera el 50%. En el eje vertical (ordenadas) podemos leer el número de estados.

Si comparamos esta nueva distribución con las de tipo campaniforme, aunque la forma es completamente diferente, podemos observar que en esta también encontramos un número de valores (y por tanto de países) cuyos porcentajes de población analfabeta son similares entre sí. Sin embargo, los valores no se agrupan esta vez en el centro de la distribución sino en el extremo inferior de la misma, es decir, en el área izquierda del gráfico. A medida que nos desplazamos hacia el lado derecho de la distribución, es decir, hacia porcentajes de analfabetismo más elevados, disminuye de forma rápida el número de países. Esta curva de descenso, tan característica de este tipo de distribuciones, es la que explica el nombre que se les ha dado: distribuciones en forma de J invertida.



En ambos tipos de distribuciones existe un porcentaje de valores similares, agrupados en un sector del rango. En la distribución campaniforme el agrupamiento se sitúa en el sector central del rango. En la distribución en forma de J invertida el agrupamiento ocupa el extremo inferior del rango.

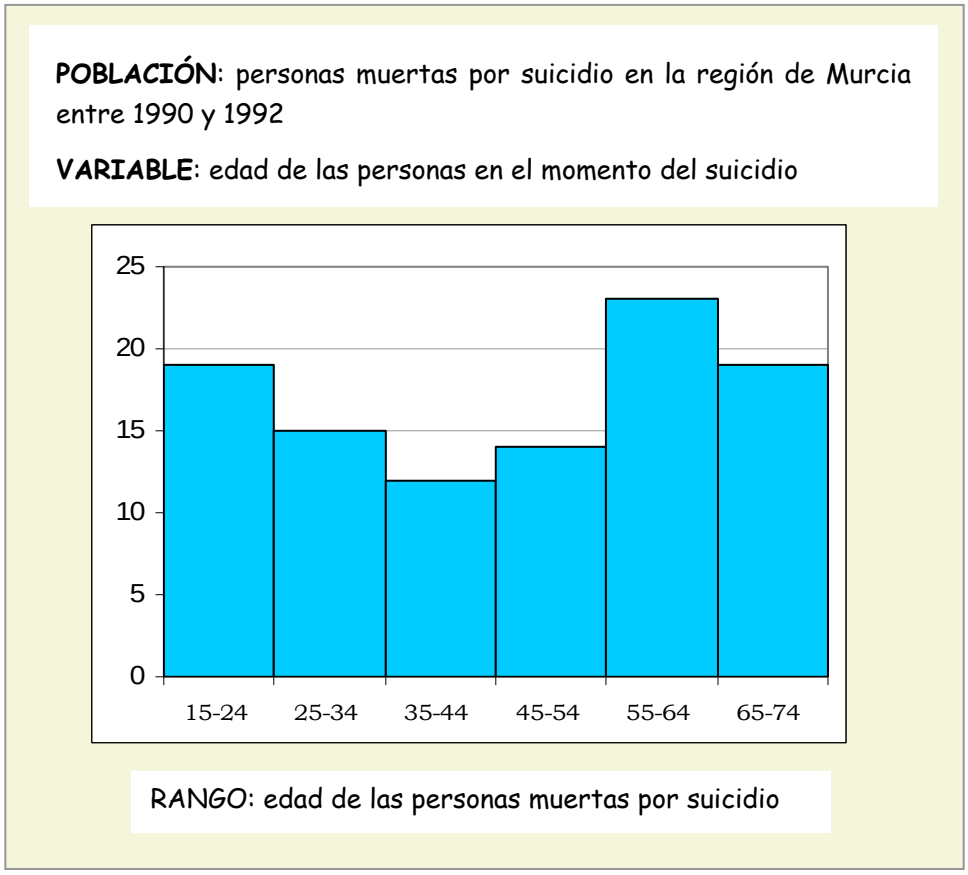
Otro tipo de distribución con una forma completamente diferente a las dos anteriores es la que conoce con el nombre de distribución en forma de U.



La distribución en forma de U se caracteriza por tener dos picos, uno a cada lado del rango, que se corresponden con las frecuencias de valores más elevadas. A diferencia de los dos tipos anteriores, en este tipo de distribución no existe una única mayoría de valores concentrados en un sector del rango.

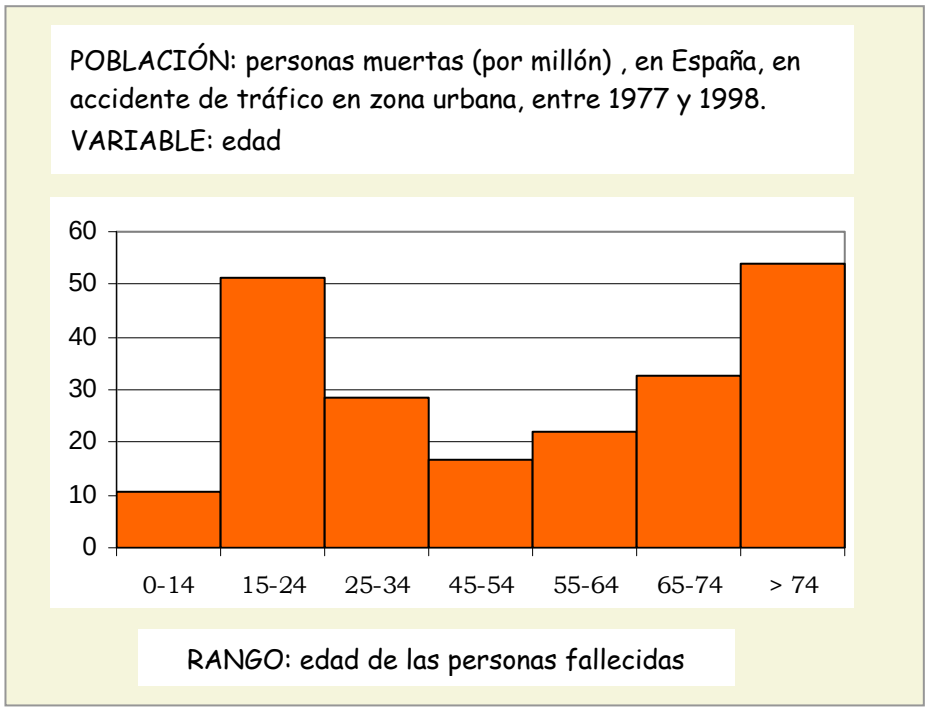
Distribuciones en forma de U se encuentran, por ejemplo, en variables que miden determinadas características o costumbres de los seres humanos cuya frecuencia de aparición es muy superior entre la población jóven y los ancianos. Un estudio dedicado a analizar las características de las personas que se habían suicidado en la región de Murcia, entre 1990 y 1992, permitió conocer que del total de muertos por esta causa, destacaban los individuos jóvenes y los

ancianos. (Gutiérrez García y Molina Clemente, 1996) El histograma que vemos a continuación, elaborado a partir de los datos que proporciona el estudio, muestra los dos picos típicos de estas distribuciones, que se corresponden con el número de suicidios en las poblaciones de jóvenes y de ancianos.



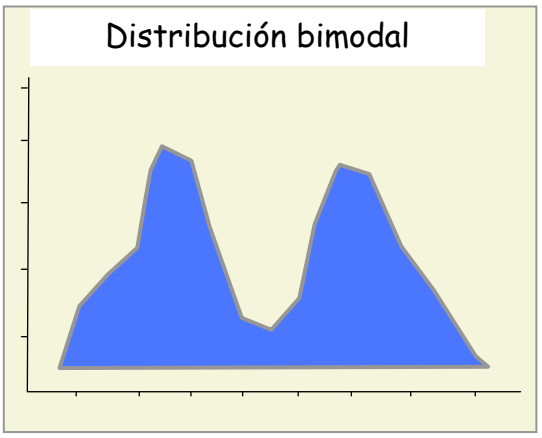
Histograma elaborado según datos de Gutiérrez García y Molina Clemente, 1996

Un caso similar al anterior es el estudio que analizó la sinistraliedad producida por los accidentes de tráfico en la población española, entre 1977 y 1988. (Monteagudo, et al., 2001). Decimos que es similar porque, nuevamente, los resultados mostraron que, entre los fallecidos en accidentes de tráfico ocurridos en zona urbana, destacaban claramente en número las personas jóvenes por un lado y la población anciana por otro. El histograma que mostramos a continuación, elaborado también a partir de los datos aportados por el estudio, muestra una distribución en forma de U más pronunciada y más abierta que la del ejemplo anterior. Esto significa que, en este segundo ejemplo, la tendencia a la concentración de elementos en los extremos del rango es superior. Dicho de otro modo, los jóvenes y adultos fallecidos por accidente de tráfico destacan mucho más dentro del total de la población de lo que lo hacen los fallecidos por suicidio.



> …si consideramos la distribución porcentual de la accidentalidad como peatones de los distintos grupos de edad, podemos observar como, especialmente en zonas urbanas, se obtiene una distribución en forma de J, donde los extremos que destacan son los niños y los ancianos. (Monteagudo, et al., 2001)

Para finalizar con la tipología de distribuciones, mencionaremos un último tipo, en cierta medida similar al anterior. Se trata de la distribución bimodal, conocida también como distribución en doble joroba de camello. Un vistazo rápido a su forma disipa cualquier duda sobre su denominación:



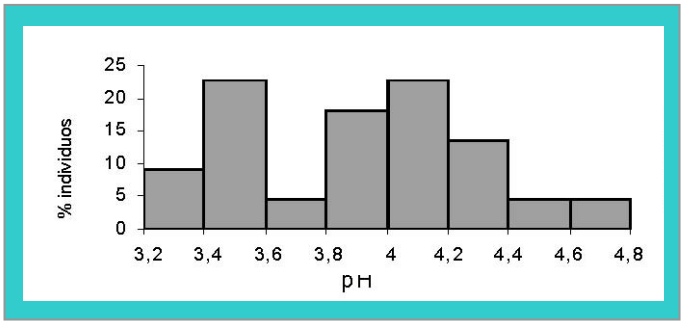
La similitud con el modelo de distribución en forma de U, se debe al hecho de que, en este caso también, existe una doble tendencia en el agrupamiento de los valores, es decir, en que hay dos sectores del rango en los que se acumulan más valores que en el resto. La diferencia entre los dos tipos de distribuciones estriba en que, en el caso de la bimodal, el agrupamiento no tiene por qué producirse en los extremos del rango sino en cualquier sector del mismo.

Una vez más, un ejemplo mostrará de forma más clara las propiedades y el significado de las distribuciones bimodales. En este caso, hemos escogido un estudio experimental cuyo objetivo es mejorar la selección de variedades de melocotonero adaptadas a las condiciones del Valle Medio del Ebro. Con este fin, los autores analizaron la variabilidad, en cuanto a criterios de calidad, de 29 poblaciones de melocotonero obtenidas a partir del cruzamiento de distintas especies. El análisis realizado lo describen como sigue:

> En cada una de las poblaciones en estudio se determinan parámetros agronómicos como las fechas de floración, el vigor y la carga productiva del árbol, y las fechas de maduración del fruto. En el momento de maduración fisiológica se evalúa la producción (kg/árbol) de cada descendencia y se eligen muestras de frutos representativos para los análisis de calidad del fruto. Además, se analizan caracteres pomológicos como la tipología del fruto (melocotón, nectarina o paraguayo; tipo sub-ácidos o normales; de carne amarilla, blanca o sanguina), forma, coloración de la epidermis, firmeza de la pulpa, sólidos solubles totales (SST), composición de azúcares y acidez. También se determina la composición de azúcares por HPLC a partir del zumo de los frutos. (Cantín et al., 2006)

De este estudio, lo que a nosotros nos interesa son los resultados que obtuvieron con respecto al Ph, es decir, con el grado de acidez de la fruta. El

interés reside en el hecho de que los análisis realizados mostraron una distribución bimodal de la variable Ph, tal como refleja el histograma:



Aunque el Ph de los frutos analizados presenta una variabilidad que oscila entre 3,2 y 4,8, el grado de acidez de una buena mayoría de los mismos se concentra en dos intervalos: 3,4- 3,6 y 3,8-4,2.

Aunque a los no entendidos en el tema pueda resultarnos totalmente ajeno el significado del Ph en este estudio, o la relación entre esta variable y la calidad de la fruta, el ejemplo nos sirve para visualizar un caso real de distribución bimodal y para comprender que el estudio de la forma de la distribución de las variables se utiliza en campos de conocimiento muy variados.

Recapitulación

A lo largo de este apartado hemos visto diferentes ejemplos sobre la forma en que se distribuyen las variables dentro del rango. Hemos visto también que existe una tipología de formas que permite clasificar las distintas distribuciones mediante nombres alusivos al perfil que dibujan. Llegados a este punto, es preciso recapitular y, sobre todo, definir cómo podemos utilizar el análisis de la forma de una distribución a la hora de decidir qué procedimientos utilizar para realizar el estudio de variabilidad.

Tenemos que decir, en primer lugar, que el "catálogo" de formas de la distribución que hemos presentado no incluye todos los tipos que se han identificado. No existe tampoco una única tipología comunmente aceptada que pueda servir como punto de referencia de todas las formas posibles. Al revisar los manuales de Estadística lo que encontramos es que cada autor, en función de su conocimiento o de sus preferencias, menciona unos tipos de formas u otros.

Las menciones al tipo de forma de una distribución pretenden ser más una ayuda para evocar una visualización rápida del modo en que se distribuyen los datos, que el resultado de comparar de forma precisa e inequívoca la distribución que se analiza con un patrón determinado de distribuciones.

Aunque la impresión que se pueda obtener al observar la variedad de tipos de distribuciones posibles es que en la práctica nos encontraremos con variedad de formas distintas, en realidad ocurre todo lo contrario. El tipo de forma, sin duda, más frecuente, cuando se trabaja con variables de razón, es el que corresponde a la distribución en forma de campana. De hecho, por razones cuya explicación excedería el conocimiento de un estudiante recién iniciado, la distribución campaniforme tiene una importancia excepcional en el mundo de la Estadística. Nos conformaremos con decir que muchas de las posibilidades que la Estadística ofrece en el ámbito de la predicción se fundamentan en las características de las distribuciones campaniformes.

Hemos dejado para el último lugar la conclusión fundamental de este tema, precisamente la que justifica su presencia. ¿Cuál es, definitivamente, el criterio para seleccionar los procedimientos de análisis adecuados al tipo de datos y de variables?.

Como decíamos en un principio, es la forma de la distribución la que debe orientar la ruta de análisis. A efectos de esta decisión, agruparemos los tipos de distribuciones en dos grupos: las que son campaniformes y las que no lo son.

Cada uno de estos grupos debe ser analizado mediante procedimientos diferentes. Solamente las distribuciones campaniformes pueden ser analizadas y descritas mediante las herramientas que, probablemente, a todos nos resulten más conocidas. Nos referimos a la media estadística, a la desviación estándar, al análisis de simetría y (esta última, menos conocida) a la curtosis.

En los capítulos que siguen se expondrán los fundamentos, los procedimientos de análisis y de interpretación de las herramientas mencionadas, es decir de aquellas que se pueden utilizar para el análisis de las distribuciones campaniformes.