

Tema 2

Análisis gráfico

Contenido

2.1. Introducción	1
2.2. Análisis exploratorio	2
2.2.1. Análisis exploratorio para variables con pocas modalidades	2
2.2.2. Análisis exploratorio para variables con muchas modalidades	3
2.3. Representaciones gráficas	5

2.1. Introducción

El objetivo de los gráficos en Estadística es facilitar la interpretación de las tablas de frecuencias. Se representa la información muestral mediante diagramas que ilustran las frecuencias de las distintas modalidades de forma clara y eficiente.

La finalidad de los gráficos es doble. Por un lado permiten **explorar** la distribución desde un punto de vista técnico, para detectar valores extraños, facilitar la elección de medidas resumen, etc. Por otro lado, permiten **transmitir visualmente** la información de las tablas de frecuencias de forma más atractiva y sencilla.

Cuando se localizan datos extraños en una muestra se debe valorar como proceder. Si está claro que se produjo un error, se debe subsanar. Si no está claro y puede ser un dato válido, aunque alejado del resto por cualquier circunstancia especial, no es conveniente eliminarlo. En este caso se debe tener en cuenta la existencia

de datos anómalos para determinar las medidas resumen más adecuadas (que entran dentro de las denominadas *medidas robustas*, ver Tema 7).

A continuación se verán las formas habituales de exploración y descripción gráfica de muestras para cada situación. Estos análisis se pueden realizar con la gran mayoría de los programas o paquetes estadísticos. Los gráficos de este libro se han realizado con el entorno estadístico de uso libre R (<http://www.r-project.org/>).

2.2. Análisis exploratorio

Al recoger o al almacenar los datos se pueden cometer errores que pasan desapercibidos, especialmente cuando se recogen grandes cantidades de datos. En ocasiones, esos errores son detectables estadísticamente y se pueden corregir. Por este motivo es importante realizar un análisis exploratorio previo antes de comenzar el estudio estadístico propiamente dicho. Además, el análisis exploratorio proporciona una primera idea de la forma de la distribución, que ayudará posteriormente a elegir las medidas de resumen más adecuadas a cada caso. La exploración preliminar depende del tipo de variables que se analice. Aunque cobra pleno sentido únicamente para variables continuas conviene realizar una exploración previa para todo tipo de variables.

2.2.1. Análisis exploratorio para variables con pocas modalidades

Si la variable de interés es nominal, ordinal o discreta (toma pocos valores distintos en la muestra) se puede realizar simplemente una tabla de frecuencias o un diagrama de barras y verificar si se observa algún valor extraño.

Un **diagrama de barras** es una representación gráfica sobre unos **ejes de coordenadas**, en el que se representan las **modalidades** de la variable en el eje de las equis y sobre cada modalidad se levanta una **barra de altura proporcional a la frecuencia** de esa modalidad.

El diagrama de barras debería utilizarse únicamente para variables cardinales, ya que la escala en el eje de las equis tiene que representar la escala de la variable. Actualmente, sin embargo, la mayor parte de los programas estadísticos no consideran la escala de la variable y con fines exploratorios se utiliza para representar cualquier distribución con pocas modalidades. Las barras están separadas, para indicar que se representan datos aislados (al contrario de lo que se hará posteriormente en el histograma).

Problema propuesto: Apartado a) del Problema 2.1.

2.2.2. Análisis exploratorio para variables con muchas modalidades

Si la variable de interés toma **muchas modalidades** en la muestra (variable continua), la tabla de frecuencias sería muy grande, por lo que sería muy difícil localizar valores extraños en ella. En estos casos se realizan los llamados **análisis exploratorios gráficos**.

Los gráficos más usuales para explorar las variables continuas son el gráfico de tallos y hojas y el gráfico de cajas.

Gráfico de tallos y hojas

En este tipo de gráfico aparece una primera columna titulada *Frequency* donde se indica el número de datos que se representan en cada fila.

A continuación aparece una columna titulada *Stem* (en español, *tallo*) que indica el “número” por el que empiezan los valores de la muestra representados en esa fila. Antes de interpretar cuáles son esos valores, se debe observar la penúltima fila del gráfico, titulada *Stem width*, que indica si el stem son unidades (1), decenas (10), centenas (100), etc.

Para identificar cuál es el siguiente dígito se observan los valores que aparecen bajo *leaf* (en español, hojas).

Si el número de cifras significativas es una o dos, el gráfico de tallos y hojas contiene toda la información. En caso contrario, contiene información de las dos cifras mayores.

El número de datos que se corresponden con cada hoja (leaf) aparece en la última fila, titulada *Each leaf*.

En el gráfico de tallos y hojas se etiquetan como *extremos* aquellos valores que están muy alejados del resto (tanto por encima como por debajo).

Los valores extremos son a menudo datos erróneos que se deben eliminar del estudio o corregir y, en cualquier caso, son datos influyentes cuyo efecto hay que valorar posteriormente en el estudio estadístico.

La magnitud de los valores extremos se identifica mejor en el siguiente gráfico (Gráfico de cajas), por ahora simplemente se observa que existe alguno.

Hasta ahora se ha identificado únicamente la información numérica. Sin embargo, este gráfico contiene información visual muy útil, ya que la longitud de cada fila indica la “densidad” o cantidad de datos que hay en las distintas zonas. Filas largas frente a filas cortas indica que los datos se concentran en ciertas zonas, habiendo “lagunas” (o falta de datos) en otras, mientras que filas de longitud similar indican que los datos se reparten de forma más uniforme.

En conclusión, se puede decir que el gráfico de **tallos y hojas** contiene la mayor parte de la **información muestral numérica ordenada** de tal forma que permite **identificar** visualmente **cómo se distribuyen los valores en su rango**.

Gráfico de cajas

El gráfico de cajas se basa en lo que se llama ‘medidas robustas’, en las que interviene sólo el orden de los valores y no su magnitud (de ahí que los valores extremos no influyan demasiado en los resultados, como se comprobará posteriormente). Es aparentemente muy simple, sin embargo, contiene información muy relevante.

La escala de datos se encuentra a la izquierda, en vertical.

En el gráfico de cajas se marcan con un círculo (o) los valores extremos (algunos programas marcan con un asterisco (*) los valores muy extremos). El valor límite a partir del cuál un dato se considera extremo se determina por criterios estadísticos. La mayor parte de los programas estadísticos utilizan los llamados *bigotes de Tukey*.

Si hay valores muy extremos es posible que el resto del gráfico no se aprecie bien por problemas de escala. Lo primero que debe hacerse es verificar si esos valores son errores claros.

Además de los datos extremos, el gráfico consta de una caja (habitualmente roja) de la que parten un segmento superior y otro inferior. Si no hay datos extremos, los segmentos marcan los valores máximo y mínimo respectivamente. En caso contrario marcan el último valor “normal” de la muestra, es decir, el valor más alto (o más bajo) de entre los que no son extremos.

Dentro de la caja roja aparece una raya negra que indica el “centro” de la distribución. En concreto, indica el valor que deja el 50 % de los datos muestrales por debajo y el otro 50 % por encima. Posteriormente se verá que este valor se denomina *Mediana*.

La caja roja marca el rango de valores “moderados”. El borde inferior de la caja marca el valor que deja el 25 % de los datos muestrales por debajo y el borde superior el valor que deja el 25 % de los datos muestrales por encima, así que en la caja están el 50 % de los valores “centrales”.

Si la caja es estrecha (en vertical), significa que los datos centrales están muy juntos (poco dispersos). En cambio, si es muy ancha, significa que están muy separados (muy dispersos). En realidad, juzgar si una sola caja es ancha o estrecha es difícil, en cambio, es sencillo comparar la dispersión de dos o más grupos en función de la anchura de las cajas.

En conclusión, el gráfico de **cajas** es útil para visualizar los **datos extremos** (en caso de que los haya), así como el **rango de valores**, el “**centro**” de la distribución y lo **agrupados o dispersos** que están los valores moderados entorno a ese centro.

Estos métodos de exploración son **univariantes**, es decir, involucran cada variable del análisis de forma independiente. En muchas ocasiones es útil ‘cruzar’ la información de dos o más variables que deberían guardar cierta relación, como la altura y el diámetro, para verificar si hay algún dato que se salga de lo común en esa relación. Esto se comentará posteriormente cuando se introduzcan las formas de determinar relaciones estadísticas (ver Tema 8).

En este sentido también es interesante hacer gráficos de cajas por *factores* (datos de empleo dependiendo del sexo, la comunidad autónoma, etc.), ya que de esta forma se pueden resaltar comportamientos anómalos que quedan ocultos al manejar la información general (por ejemplo, podría ocurrir que un dato de una comunidad autónoma dada comparado con todos no sea anómalo, pero sí si se compara sólo con el resto de datos de esa comunidad).

Problema propuesto: Apartado a) del Problema 2.2.

2.3. Representaciones gráficas

Como ya se ha señalado los gráficos estadísticos se utilizan para mostrar visualmente de forma sencilla la forma general de la distribución (dónde hay muchos datos, donde hay pocos, etc.). Según el tipo de variable que se necesite representar, se puede elegir entre los siguientes gráficos.

Gráfico de sectores

Es un **círculo dividido** en tantos **sectores** como modalidades tenga la variable, de forma que el **área** de cada sector es **proporcional a la frecuencia** de la modalidad que representa.

Es el gráfico más adecuado para representar variables nominales, porque al representar los valores sobre un círculo, no se considera ningún orden natural. De

todas formas, se utiliza incluso para representar variable ordinales o cardinales con pocos valores.

Diagrama de barras

Este gráfico, descrito e ilustrado en la Sección 2.2.1, está pensado para representar variables nominales, ordinales o cardinales discretas y, en la práctica, se suele utilizar con variables que no tienen un número excesivo de modalidades en la muestra.

Histograma

Un gráfico sobre unos **ejes de coordenadas** en el que se representa la variable escalada en el eje de las equis, los datos agrupados en clases y sobre cada clase se levanta una **barra de área proporcional a la frecuencia** (absolutas o relativas) de esa clase.

La agrupación en clases la suelen realizar automáticamente los programas estadísticos, pero se puede cambiar el número de clases para apreciar mejor la distribución. Es recomendable hacer pruebas hasta conseguir una agrupación fácil de interpretar. Al igual que al realizar agrupaciones (ver Sección 1.4) un número grande de intervalos dificulta la interpretación, mientras que un número muy bajo supone una pérdida de información, por lo que hay que buscar un número moderado.

Este gráfico es similar al diagrama de tallos y hojas, en el sentido en el que la longitud (vertical) de las barras representa, la ‘densidad’. Sin embargo es más fácil de interpretar, porque no incluye la información de todos los datos individuales, lo que hace la referencia visual más agradable.

Es el gráfico más adecuado para representar variables continuas y, en la práctica, se utiliza con variables con un número elevado de modalidades. Las barras del histograma están juntas, lo que refleja el hecho de que los datos no están aislados, si no que forman ‘un continuo’.

Existen otros tipos de gráficos que pueden ser interesantes según nuestros objetivos, como los gráficos de líneas o poligonales, etc. y que se pueden realizar con ayuda de la mayor parte de los programas estadísticos.

Problemas propuestos: Apartado b) de los Problemas 2.1 y 2.2.