

Tema 7

Otras medidas descriptivas usuales

Contenido

7.1. Introducción	1
7.2. Medidas robustas	2
7.2.1. Media recortada	2
7.2.2. Medidas de dispersión robustas	3
7.3. Medidas de forma	4
7.3.1. Asimetría	5
7.3.2. Curtosis	6

7.1. Introducción

Además de la medidas que se han visto en los temas anteriores la mayor parte de los programas estadísticos incluyen entre sus descriptivos algunas medidas robustas y algunas medidas de forma. En este tema se presentarán las opciones más habituales. Aunque se incluyen las fórmulas que permiten hacer los cálculos como en temas anteriores, se recomienda el uso de programas estadísticos, sobre todo en el contexto de las medidas robustas, que son especialmente adecuadas para trabajar con grandes volúmenes de datos.

7.2. Medidas robustas

Las **medidas robustas** son aquellas que no se ven afectadas por valores atípicos (o, más en general, por discrepancias entre los datos muestrales y el modelo teórico al que se asimilan los datos).

Ya se ha comprobado que la media no es una medida descriptiva robusta y, como alternativa, se había planteado la mediana. En esta sección se introducirá otra medida de localización central robusta así como algunas formas de cuantificar la dispersión en este contexto.

7.2.1. Media recortada

El inconveniente de la mediana como medida robusta es que sólo tiene en cuenta el orden de los valores, no las magnitudes, por lo que a veces puede resultar poco informativa. Una alternativa que se puede plantear es calcular la media eliminando la posibilidad de que esté afectada por los valores extremos excluyendo del estudio sistemáticamente los valores altos y los bajos.

La **media recortada al 10 %** es la media aritmética del 90 % central de los valores de la muestra. Es decir, se eliminan el 10 % de los valores extremos (el 5 % de los valores más altos y el 5 % de los más bajos) y se calcula la media de los que quedan. Como la media aritmética, es una medida de tendencia central para variables cardinales que tiene ventajas e inconvenientes. La mayor ventaja es que al eliminar el 10 % los valores extremos, en caso de que haya valores atípicos, éstos no influyen en su cálculo y en este sentido es mejor que la media. Sin embargo, eliminar individuos también supone un inconveniente, porque sabemos que las muestras son más informativas cuanto más grandes sean y eliminar datos supone una pérdida de conocimiento. Si se dispone de muchos datos, eliminar alguno no es problema, pero si la muestra es pequeña, sí puede serlo. Además, es menos operativa matemáticamente que la media.

Si el tamaño de muestra es suficientemente grande, comparando los valores de la media aritmética y de la media recortada se puede saber si hay valores muy extremos que hacen que la media aritmética no sea buena sin falta de hacer análisis exploratorios. Si se diferencian poco, significa que el 10 % de los valores que se excluyeron no aportaban mucho en el cálculo de la media, pero si se diferencian mucho significa que tenían mucho peso y que la media puede estar distorsionada por ellos.

En definitiva, se utilizará este descriptivo como medida de tendencia central únicamente si la muestra es suficientemente grande y hay valores atípicos que hacen que la media esté muy desvirtuada.

Si se sospecha que hay más o menos de un 10 % de valores atípicos se podría optar por hacer otro recorte (15 o 5 % o recortar sólo los altos y no los bajos, etc.), o elegir otras medidas robustas más avanzadas que sufran menor pérdida de información.

7.2.2. Medidas de dispersión robustas

Si hay valores atípicos, la media no es buena medida de centro. En este caso, la varianza, la desviación típica y el CV no sirven para cuantificar la dispersión, ya que miden la variabilidad en torno a un punto que no resulta representativo. Además, como se calculan a partir de las magnitudes de todos los datos, también están influenciadas por los datos extremos.

Las medidas de dispersión adecuadas a cada caso dependen de la medida de centro elegida.

Amplitud intercuartil

Cuando se elige la mediana, se puede cuantificar la dispersión mediante la amplitud intercuartil (la altura de la caja del gráfico de cajas, ver Sección 2.2.2), que es la diferencia entre el tercer cuartil y el primero, es decir,

$$AI = C_3 - C_1 = Q_{0,75} - Q_{0,25}.$$

Los cuartiles son, al igual que la mediana, ‘descriptivos de orden’ (sólo interviene en su cálculo el orden, y no la magnitud), por lo que son también medidas robustas. No obstante, debe tenerse en cuenta que si las muestras son pequeñas y hay un porcentaje relativamente alto de valores atípicos, la diferencia entre considerarlos o no puede ser tangible.

La amplitud intercuartil presenta el mismo inconveniente que la mediana: al estar basada en órdenes a veces no resulta suficientemente informativa.

La AI mide el rango en que se mueven “valores moderados” (eliminando el 25 % de los valores más bajos y el 25 % de los valores más altos). Si la amplitud intercuartil es muy pequeña, significa que los valores moderados están muy próximos entre sí, es decir, hay poca dispersión o variabilidad y, por lo tanto, el valor central (que es justo la mediana) resulta muy representativo. Si la amplitud intercuartil es muy grande ocurre justo lo contrario.

MEDA

En el Tema 4 se introducían las medidas de variabilidad calculando, en primer lugar, el *promedio de las desviaciones* absolutas. Como la media no es robusta, si hay una desviación muy grande (provocada por un dato atípico), esa medida de dispersión queda desvirtuada. Como alternativa a la media se tienen la mediana y la media recortada. La *mediana de las desviaciones absolutas* recibe, en general, el nombre de MEDA. Se suele utilizar, en particular, cuando se elige la mediana como medida robusta de centro. Su fórmula para una muestra de N datos $x_1, \dots, x_N$ con mediana Me es

$$\text{MEDA} = \text{Me}(|X - \text{Me}|) = \text{Me}\{|x_1 - \text{Me}|, \dots, |x_N - \text{Me}|\}.$$

Aunque en el Tema 4 se indicaba que el valor absoluto no resultaba muy operativo y se prefería sustituir por el cuadrado, esto ocurre principalmente cuando se trabaja con promedios como medidas de centro. Por motivos técnicos cuando se considera la mediana las distancias suelen medirse en valor absoluto, de ahí que se utilice más la MEDA que la ‘mediana de las desviaciones cuadráticas’, que sería otra posibilidad.

La ventaja de la MEDA respecto a la amplitud intercuartil es que se calculan diferencias de todos los valores al centro (en vez de la diferencia de únicamente 2 valores) así que aunque finalmente se utilice sólo el orden de las diferencias, puede resultar más informativa.

Desviación recortada

Si se elige la media recortada al 10% como medida de centro, ya se están seleccionando los valores que se consideran no atípicos para su cálculo, luego lo lógico será cuantificar la dispersión utilizando esos valores de la forma habitual.

Las tres medidas de dispersión robustas que se han introducido son absolutas (se miden en las mismas unidades que la variable), pero se pueden relativizar con respecto al valor central elegido como se hacía con el CV (dividiendo su valor por la medida de centro correspondiente).

Problema propuesto: Problema 7.1.

7.3. Medidas de forma

Las **medidas de forma** se suelen utilizar para comparar las distribuciones muestrales con la distribución más importante de la Estadística: la **distribución normal**.

La distribución normal, o **campana de Gauss**, es una abstracción de un tipo de distribuciones muy habituales en la práctica, con mucha densidad de valores agrupados alrededor de la media, pocos valores bajos o altos y simétrica (ver Figura 7.1).

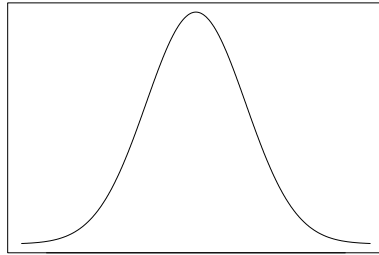


Figura 7.1: Distribución normal.

Muchas inferencias clásicas en Estadística están diseñadas para la distribución normal. De hecho, tanto la media como todas las medidas basadas en ella (como la varianza, la tipificación, etc.) son especialmente representativas en las distribuciones ‘tipo normal’. Si la forma de una distribución se aleja mucho de la normalidad suele ser conveniente un análisis más profundo, en busca de posibles subpoblaciones (por ejemplo, por sexos, especies, etc.) o bien optar por medidas mejor adaptadas a esas distribuciones, como suelen ser las medidas robustas.

Las características básicas de una distribución normal son la simetría y la forma de campana con la que los valores se agrupan en torno a la media.

7.3.1. Asimetría

El **coeficiente de asimetría** clásico o coeficiente de Fisher, γ_1 , cuantifica la falta de simetría de una variable cardinal respecto a su valor medio (también se podría analizar la asimetría respecto a la mediana u otro valor central). Su fórmula (implementada en la mayor parte de los paquetes estadísticos) es:

$$\gamma_1 = \frac{\sum_{i=1}^k (x_i - \bar{x})^3 f_i}{S_x^3}.$$

Se dice que una **distribución** es **simétrica** respecto a la media si lo que hay por debajo de la media se distribuye igual que lo que hay por encima, en cuyo caso el

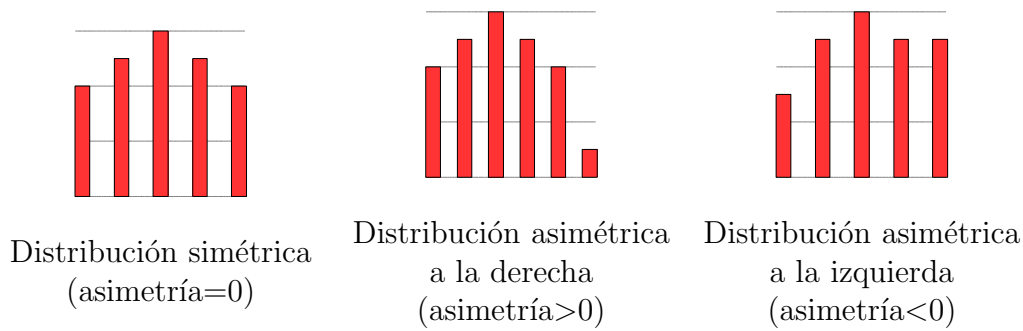


Figura 7.2: Distribuciones simétricas y asimétricas.

coeficiente de asimetría toma el valor 0. La simetría/asimetría se aprecia muy bien gráficamente en histogramas o diagramas de barras (ver Figura 7.2).

A grandes rasgos se puede decir que en las distribuciones asimétricas a la derecha hay valores altos con cierto peso (frecuencia) que tienden a alejarse más y los valores bajos tienden a estar más concentrados cerca de la media. En este caso el coeficiente de asimetría será un valor positivo. Si los valores bajos se alejan más de la media con cierta frecuencia y los altos están más concentrados será una distribución asimétrica a la izquierda y entonces el coeficiente de asimetría será un valor negativo.

El coeficiente de asimetría es un coeficiente adimensional (no tiene unidades), por lo que se puede utilizar para realizar comparaciones. Si no hay valores muy extremos suele variar pocas unidades alrededor de 0 (lo normal es entre -2 y 2, más raramente entre -4 y 4, y si toma valores más alejados de 0, suele ser por la existencia de valores atípicos). Si una distribución es muy asimétrica, habrá valores altos o bajos alejados de la media que pueden desvirtuar su valor, por lo que en este caso suele ser recomendable elegir una medida robusta.

Dada la estrecha relación de las dos medidas de forma consideradas, se ilustrarán ambos conceptos mediante un ejemplo después de introducir la curtosis.

7.3.2. Curtosis

El **coeficiente de curtosis** mide el grado de apuntamiento (forma campanoide alrededor de la media) de una variable en relación con la distribución normal. Su fórmula (también implementada en la mayor parte de los paquetes estadísticos) es:

$$\gamma_2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^4 f_i}{S_x^4} - 3.$$

Si una distribución es igual de apuntada que la normal se llama **mesocúrtica** y tiene un coeficiente de curtosis igual a 0. Si es más apuntada que la normal, se llama **leptocúrtica**, su curtosis será positiva y eso significa que los valores centrales están más concentrados que los de la normal y los extremos son menos frecuentes. El coeficiente de curtosis será negativo si la distribución es menos apuntada de lo normal (**platicúrtica**), y eso significa que los valores centrales están menos concentrados que los de la normal y los extremos son más frecuentes (ver Figura 7.3).

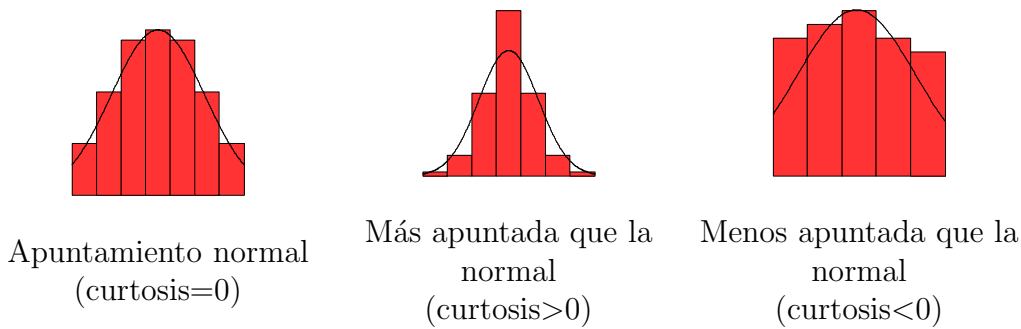


Figura 7.3: Distribuciones mesocúrticas, leptocúrticas y platicúrticas.

El coeficiente de curtosis también es un coeficiente adimensional (no tiene unidades), por lo que se puede utilizar para realizar comparaciones. Si no hay valores muy extremos suele variar pocas unidades alrededor de 0.

Problema propuesto: Problema 7.2.