Diplomado en Análisis de Información Geoespacial

Recordando las medidas de tendencia central, de dispersión y de la forma

Autor:
M. en G. Alberto Porras Velázquez

Contoy 137 Esq. Chemax, Col. Lomas de Padierna, Delegación Tlalpan, CP 14240, México, CDMX.

Introducción

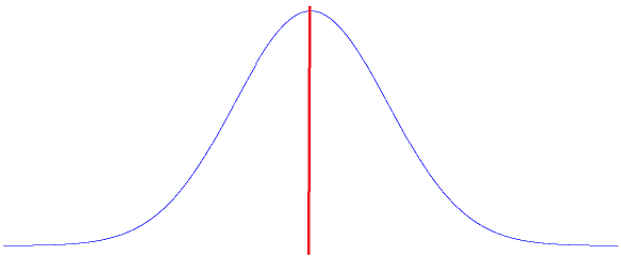
Una de las principales tareas de la estadística radica en la descripción de información numérica a través un conjunto de medidas que cuantifican diversos aspectos de la distribución de los valores de una variable en un conjunto de datos. La descripción básica involucra tres aspectos.

En primer lugar, suele ser de interés la identificación de un valor en torno al cual un conjunto de datos está centrado. A las medidas que describen esta característica se les llama *medidas de tendencia central*.

Otro aspecto de interés en la descripción radica en la caracterización de la variabilidad de los valores observados, cuantificados a través de las *medidas de dispersión*.
Finalmente, las *medidas de la forma* indican cómo los datos se agrupan de acuerdo con la frecuencia con que ocurren.

Por otro lado, la presentación de la información a través de gráficas es una herramienta de gran ayuda que permite identificar de manera cualitativa, a través de un análisis visual, las principales características de la distribución de los valores de una variable.

Medidas de tendencia central



La media, la moda y la mediana son las medidas de tendencia central más utilizadas. Cada una de estas medidas proporciona un valor de referencia para establecer cómo se centra un conjunto de datos.

La ***media*** se calcula mediante la expresión:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Es decir, la suma de todos los valores medidos de una variable, dividida entre el número total de observaciones. En la ecuación, la media de la variable X se representa con $\bar{X}$ y los valores observados de la variable son representados por las X_i, comenzando con la primera observación, cuando i=1 y terminando en la última, correspondiente a i=n.

La ***median****a* de una variable X corresponde con el valor central de un conjunto de n observaciones de la variable X ordenadas según su magnitud. La expresión matemática para el cálculo de la mediana depende del número total de observaciones.

Si n es par, entonces se calcula como:

$$M = {}^{1}/_{2}\,(X_{\frac{n}{2}} + X_{({}^{n}/_{2})+1}$$

En otras palabras, cuando el número de observaciones es par, la mediana M se calcula como el promedio de los dos valores del centro, con subíndices n/2 y (n/2) +1, con la condición de que las observaciones estén ordenadas según su magnitud.

Si n es impar, entonces la mediana corresponde con el valor observado que está en el centro cuando los datos están ordenados según su magnitud.

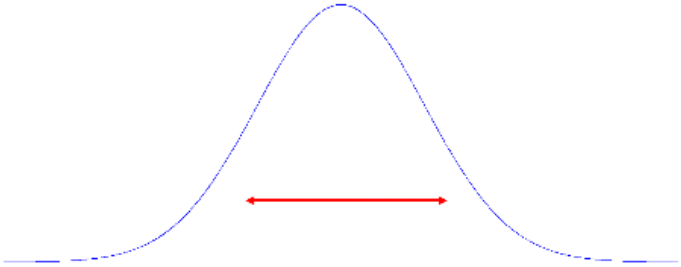
$M = X_{({}^{(n+1)}/_{2})}$ si n es impar

En la distribución de una variable la mediana es equivalente con el valor del segundo cuartil.

Contoy 137 Esq. Chemax, Col. Lomas de Padierna, Delegación Tlalpan, CP 14240, México, CDMX.

Finalmente, la ***moda*** se define como el valor de la variable cuya frecuencia de ocurrencia es más alta dentro de un conjunto de observaciones.

Medidas de dispersión



Las medidas de dispersión son utilizadas para cuantificar la variabilidad de un conjunto de datos. La forma más simple de describir la variación de un conjunto de datos es con el rango, calculado como la diferencia entre el máximo valor observado y el mínimo valor observado de la variable.

$$rango = X_{max} - X_{min}$$

El rango no proporciona suficiente información sobre la variabilidad de los datos y puede ser una medida engañosa; es por eso que la varianza y la desviación estándar son las medidas de dispersión más utilizadas.

La varianza, denotada por s^2, es una medida de dispersión de los datos con respecto a la media, de manera más específica, es el promedio de las desviaciones de los datos con respecto a la media elevadas al cuadrado.

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

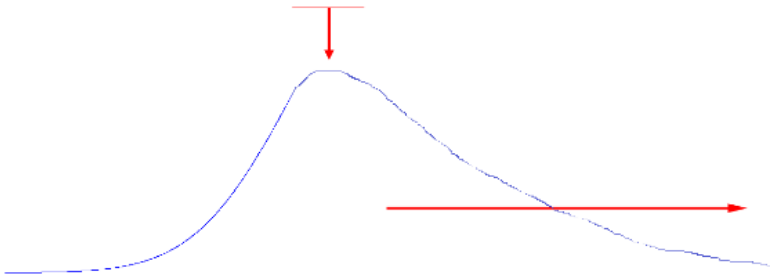
Contoy 137 Esq. Chemax, Col. Lomas de Padierna, Delegación Tlalpan, CP 14240, México, CDMX.

Cuando se calcula la varianza hay que contemplar si se calcula para una muestra o para una población. Cuando se trata de una muestra y se aplica la ecuación anterior, se suele obtener como resultado un valor que subestima la varianza real de la población. Es así que si se desea estimar la varianza de la población a partir de la muestra se debe realizar la división entre (n-1) en lugar de n en la ecuación anterior.

La varianza es una medida de dispersión que se expresa en las unidades de la variable al cuadrado, hecho que imposibilita su visualización en una gráfica como el histograma. Es por eso que en muchas ocasiones es mejor utilizar como medida de dispersión a la desviación estándar (s), calculada como la raíz cuadrada de la varianza (s^2) y que, en consecuencia, se expresa en las mismas unidades que la variable. La desviación estándar se entiende como la magnitud de la desviación promedio entre las observaciones y la media.

$$s = \sqrt{s^2}$$

Medidas de la forma



Las medidas de la forma son utilizadas para describir características tales como la simetría (o asimetría) que presenta la distribución de los datos, o qué tan aplanada o picuda es la forma de la distribución.

Para cuantificar la simetría de una distribución de datos se utiliza el ***coeficiente de sesgo (cs),*** cuya ecuación es:

Contoy 137 Esq. Chemax, Col. Lomas de Padierna, Delegación Tlalpan, CP 14240, México, CDMX.

$$cs = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}$$

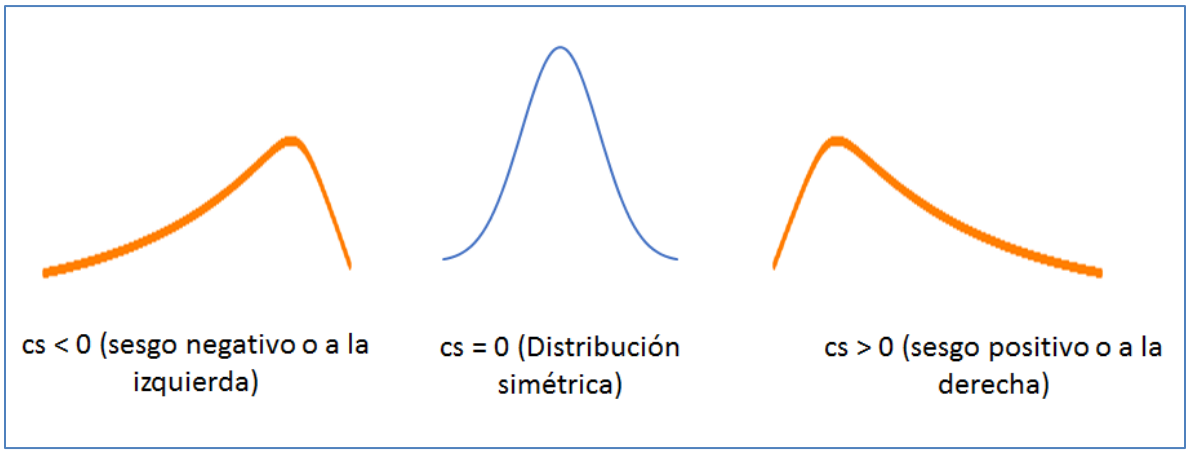
En términos generales, se trata de un promedio de las diferencias de las mediciones de la variable con respecto a la media $(X_i - \bar{X})^3$. Dado que las diferencias están elevadas al cubo, el coeficiente de sesgo puede tener un valor positivo, un valor negativo o un valor igual a cero.

Un valor positivo del coeficiente de sesgo nos indica que, en el promedio de diferencias elevadas al cubo, tienen mayor predominancia los términos en donde las X_i son mayores que $\bar{X}$, es decir, tienen mayor peso las diferencias con signo positivo. Este hecho se manifiesta como una cola de valores a la derecha de la media. Un sesgo positivo implica la existencia de observaciones con valores altos de la variable en comparación con la mayoría de las observaciones.

Por el contrario, un sesgo negativo implica que, en el promedio, pesan más los términos en donde X_i son menores que $\bar{X}$, lo que implica una cola de valores a la izquierda de la media. Así, un coeficiente de sesgo negativo implica la existencia de observaciones con valores bajos de la variable en comparación con la mayoría de las observaciones.

Un coeficiente de sesgo igual a cero implica que hay una compensación entre los términos de diferencias al cubo con valores positivos y con valores negativos que contribuyen al promedio. En este punto hay que mencionar que una distribución con coeficiente de sesgo igual a cero puede tomar diferentes formas. Una distribución simétrica forzosamente tendrá un coeficiente de sesgo igual a cero, pero si una distribución tiene coeficiente de sesgo igual a cero, no necesariamente tendrá una forma simétrica. Por otra parte, es pertinente mencionar que en la práctica será difícil encontrar una distribución de los datos en donde el coeficiente de sesgo cs sea exactamente igual a cero, pero sí se podrán encontrar coeficientes con valores

"cercanos" a cero, en donde cualitativamente se podría decir que la distribución exhibe simetría al analizar el histograma correspondiente.



Distribuciones y coeficiente de sesgo

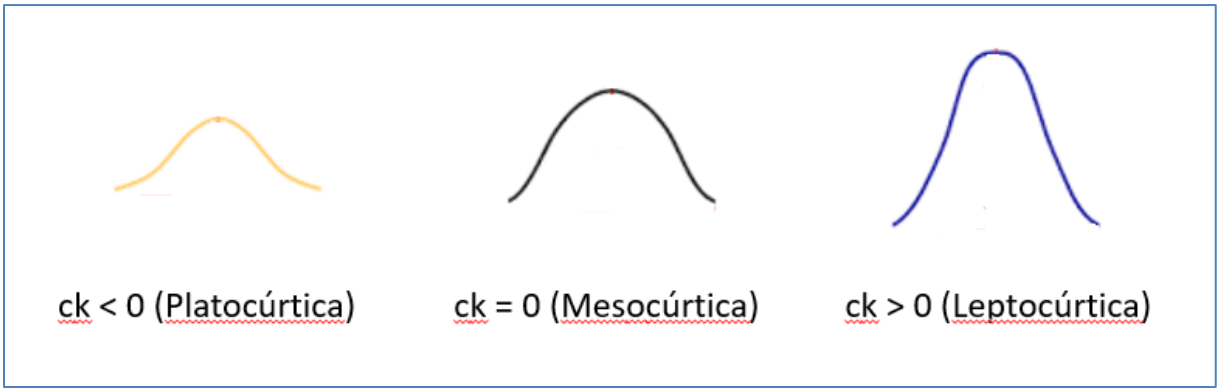
Por otra parte, el ***coeficiente de curtosis (ck)*** es la medida que define qué tan pronunciada es la punta (o pico) en una distribución. Su significado se relaciona con la distribución normal, cuyo coeficiente de curtosis es cero. La expresión matemática para el ck es:

$$ck = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{s^4} - 3$$

El -3 se agrega para hacer que la curtosis de la distribución normal sea igual a cero.

Para una distribución dada, la curva normal con una media y desviación estándar iguales a las de la distribución para la que se calcula la curtosis, sirve como patrón de referencia. A una distribución con ck = 0 (distribución normal) se le denomina mesocúrtica. A una distribución en donde el ck > 0 se le denomina como leptocúrtica, lo que implica que es más puntiaguda y con colas más anchas que la distribución normal de referencia. Finalmente, a una distribución más aplanada y con colas menos anchas

que las de las de la distribución normal de referencia se le denomina platocúrtica y tiene un ck < 0.



Distribuciones y coeficiente de curtosis

Relación entre el sesgo y las medidas de tendencia central

El sesgo de una distribución tiene impacto en algunas de las medidas de tendencia central. En una distribución simétrica los valores de la media, la moda y la mediana coinciden. Ahora imaginemos qué tendríamos que hacer para sesgar una distribución simétrica. Se tendrían que agregar algunos datos con valores muy altos (o muy bajos) en comparación con los datos de la distribución original. Podemos preguntarnos sobre el impacto de este sesgo en las medidas de tendencia central.

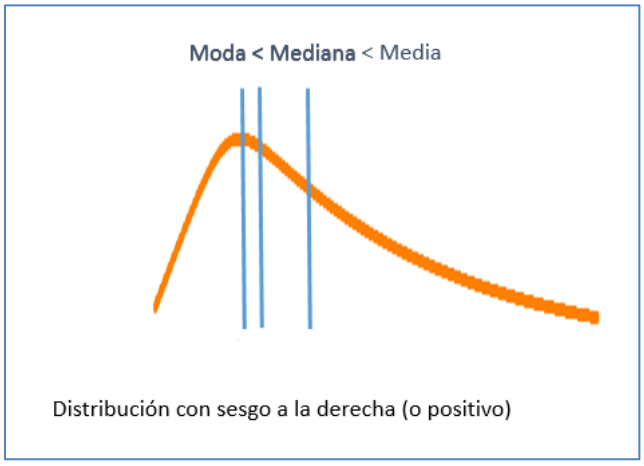
La moda es una medida de tendencia central que no es afectada por el sesgo, es decir, si agregáramos unas pocas observaciones con valores muy altos (o muy bajos) no se vería afectada debido a que el rango de valores con mayor frecuencia (en donde estaría la moda) permanecería inalterado.

El valor de la mediana podría cambiar un poco. Por ejemplo, si agregáramos unos pocos datos con valores altos de la variable, entonces la mediana correspondería con un valor similar al de la distribución no sesgada. Es decir, el valor central (o los valores

centrales cuando n es par) se recorrería hacia un dato con un valor un poco mayor en el caso de que agregáramos datos con valores altos. Recordemos que para el cálculo de la mediana los valores tienen que estar ordenados, por lo que, valores similares estarían cercanos entre sí en la ordenación, lo que implicaría que, si bien la mediana podría aumentar, este valor en general sería similar al original (el de la distribución no sesgada). El caso en que se agregaran datos con valores muy pequeños sería análogo, pero en la dirección opuesta. La mediana es una medida de tendencia central poco sensible al sesgo.

Finalmente, la media es la medida de tendencia central más sensible al sesgo. Si agregáramos datos con valores muy altos (o muy bajos) de la variable, estos tendrían un fuerte impacto en la sumatoria de la ecuación de la media y, en consecuencia, en el resultado final. Datos con valores muy altos incrementarían el valor de la media y en el caso opuesto (si se agregaran datos con valores muy bajos), la media disminuiría en comparación con el caso no sesgado.

Por lo general, en una distribución con sesgo positivo (o a la derecha) la media tendrá un valor mayor que la mediana, y la mediana tendrá un valor mayor que la moda. En una distribución con sesgo negativo la media tendrá un valor menor que la mediana, y la mediana tendrá un valor menor que la moda.



Relación entre las medidas de tendencia central en una distribución sesgada