Estadística descriptiva

Pedro O. Pérez M., PhD.

Herramientas computacionales: el arte de la analítica
Tecnológico de Monterrey

pperezm@tec.mx

09-2021

Medidas de tendencia central

- A primera vista, resumir los datos puede parecer bastante fácil: calculamos la media de los datos y listo. De hecho, si bien la media es fácil de calcular y conveniente de usar, es posible que no siempre sea la mejor medida para un valor central. Por esta razón, los estadísticos han desarrollado varias estimaciones alternativas a la media.

Media

- La estimación más básica de cómo están conformados los datos es el valor medio, media o promedio. La media es la suma de todos los valores dividida por el número de valores. Considera el siguiente conjunto de números: [3, 3, 1, 2]. La media es (3 + 5 + 1 + 2) / 4 = 11 / 4 = 2.75. El símbolo $\bar{x}$ representa la media de la muestra de una población (se pronuncia x-bar). La fórmula para calcula la media de una conjunto de N valores (x_1,x_2,...,x_N) es:

$$\text{Media} = \bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Media recortada

- Una variación de la media es la media recortada, la cual es calculada después de eliminar los extremos de un conjunto de valores ordenados y luego calculamos el promedio de los valores restantes. De un conjunto de valores ordenados (x_1, x_2,...,x_N) donde x_1 es el valor más pequeño y x_N es el valor más grande, la fórmula para calcular la media recortada con los p valores más pequeños y más grandes omitidos es:

$$\text{Media recortada} = \bar{x} = \frac{\sum_{i=p+1}^{N-p} x_i}{N-2p}$$

La media recortada elimina la influencia de valores extremos. Por ejemplo, la puntuación de los concursos internacionales de clavados se obtiene eliminando la puntuación máxima y mínima de los jueces y calculando el promedio de las puntuaciones restantes. Esto imposibilita que un solo juez manipule la puntuación, quizás para favorecer al competidor de su país. Las medias recortadas se utilizan ampliamente y, en muchos casos, es preferible utilizarlas en lugar de la media ordinaria.

Media ponderada

- Otro tipo de media es la media ponderada, que se calcula multiplicando cada valor x_i por un peso w_i y dividiendo la suma por la suma de los pesos. La fórmula para una media ponderada es:

$$\text{Media ponderada} = \bar{x_w} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$

- Hay dos razones principales para usar una media ponderada:
  1. Algunos valores son intrínsecamente más variables que otro, y las observaciones muy variables reciben un peso menor. Por ejemplo, si tomamos el promedio de varios sensores y uno de los sensores es menos preciso, entonces podríamos reducir el peso de ese sensor.
  2. Los datos recopiladores no representan igualmente a los diferentes grupos que nos interesa medir. Por ejemplo, debido a la forma en que se realizó un experimente en línea, es posible que no tengamos un conjunto de datos que refleje con precisión todos los grupos de usuarios. Para corregir eso, podemos dar un mayor peso a los valores de los grupos que estaban subrepresentados.

Moda

- La moda es el valor con mayor frecuencia en la distribución de datos. Si tomamos como ejemplo una muestra compuesta de los siguientes 5 números: 3, 8, 2, 8, 1; el valor modal es 8, ya que se es el que se repite la mayor cantidad de veces. La moda sirve para definir lo más común, lo que más se usa o lo que es más frecuente, en términos matemáticos, el valor de mayor frecuencia absoluta.

Mediana

- La mediana es el número del medio de una lista de datos ordenada. Si hay un número par de valores de datos, el valor medio es uno que no está realmente en el conjunto de datos, sino que es el promedio de los dos valores que dividen los datos ordenados en dos mitades. En comparación con la media, que usa todas las observaciones, la mediana solo depende de los valores en el centro de los datos ordenados. Si bien, esto puede parecer una desventaja, dado que la media es mucho más sensible a los datos, hay muchos casos en los que la mediana es una mejor métrica para la ubicación. Supongamos que queremos analizar los ingresos familiares típicos en los vecindarios de una zona. Al comparar un vecindario de ingresos altos con un vecindario de ingresos bajos, usar la media producirá resultados muy diferentes. Si usamos la mediana, no importa cuán rico sean los que vivan en un vecindario; la posición de la observación intermedia seguirá siendo la misma.

Por las mismas razones por las que usamos una media ponderada, también es posible calcular una mediana ponderada. Al igual que con la mediana, primero ordenamos los datos, aunque cada valor de datos tiene un peso asociado. En lugar de tomar el número del medio, la mediana pondera es el valor tal que la suma de los pesos es igual para ambas mitades de la lista ordenada. Como la mediana, la mediana ponderada es robusta a los valores atípicos (outliners).

Mediana

- La mediana se conoce como una estimación sólida de la ubicación, ya que no está influenciada por valores atípicos (casos extremos) que podrían sesgar los resultados. La exacta definición de un valor atípico es algo subjetiva, aunque se utilizan ciertas convenciones. Que un valor sea atípico, no indica que el dato sea inválido o erróneo. Aun así, los valores atípicos a menudo son el resultado de errores de datos, como la mezcla de datos de diferentes unidades (kilómetros y metros) o lecturas incorrectas de un sensor. Cuando los valores atípicos son el resultado de datos incorrectos, la media dará como resultado una estimación deficiente de la ubicación, mientras que la mediana seguirá siendo válida. En cualquier caso, los valores atípicos deben identificarse y, a menudo, requieren de una mayor investigación.

- La mediana no es la única estimación sólida de la ubicación. De hecho, una media recortada se usa ampliamente para evitar la influencia de valores atípicos. Por ejemplo, recortar el 10 % inferior y superior (una estrategia muy usada) de los datos proporcionará protección contra valores atípicos en todos los conjuntos de datos, excepto en los más pequeños. Se puede pensar en la media recortada como un compromiso entre la media y la mediana: es robusta a los valores extremos en los datos, pero utiliza más datos para calcular la estimación de la ubicación.

Desviación absoluta media

- La ubicación es solo una dimensión para resumir una característica. Una segunda dimensión, la variabilidad, también conocida como dispersión, mide que tan agrupados o dispersos están los datos. La variabilidad es un concepto muy importante a tener en cuenta: hay que medirla, reducirla, distinguir en la variabilidad aleatoria y la real, identificar las diversas fuentes de variabilidad real y tomar decisiones sobre ella. Asi como existen diferentes formas de medir la ubicación (media, mediana, etc.) también existen diferente formas de medir la variabilidad.

- Las estimaciones de variación más utilizadas se basan en las diferencia entre la media (una estimación de ubicación) y los datos observados. Para un conjunto de datos, [1, 4, 4], la media es 3 y la mediana es 4. Las desviaciones de la emdia son las diferencias (1 - 3 = -2, 4 - 3 = 1, 4 - 3 = 1). Estas desviaciones nos dicen qué tan dispersos están los datos alrededor del valor central.

- Una forma de mediar la variabilidad es estimar un valor típico para estas desviaciones. Promediar las desviaciones en sí no nos diría mucho: las desviaciones negativas compensan las positivas. De hecho, la suma de las desviaciones de la media es exactamente cero. En cambio, un enfoque simple es tomar el promedio de los valores absolutos de las desviaciones de la media. En el ejemplo anterior, el valor absoluto de las desviaciones es [2, 1, 1] y su promedio es (2 + 1+ 1) / 3 = 1.33. Esta se conoce como la desviación absoluta media y se calcula mediante la fórmula:

$$\text{Desviación absoluta media} = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N}$$

Desviación estándar

- Las estimaciones de variabilidad más conocidas son la varianza y la desviación estándar, que se basan en desviaciones cuadradas. La varianza es un promedio de las desviaciones cuadradas y la desviación estándar es la raíz cuadrada de la varianza.

$$\text{Varianza} = s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$

$$\text{Desviación estándar} = \sqrt{Varianza}$$

Desviación Absoluta Mediana

- La desviación estándar es mucho más fácil de interpretar que la varianza, ya que está en la misma escala que los datos originales. Aun así, con su fórmula más complicada y menos intuitiva, puede parecer extraño que en las estadísticas se prefiera la desviación estándar a la desviación media absoluta. Su preferencia se debe a que, matemáticamente, trabajar con valores cuadrados es mucho más conveniente que con valores absolutos, especialmente para modelos estadísticos.
- Aunque, ni la varianza, ni la desviación estándar o la desviación media absoluta son estimaciones robustas a valores atípicos y extremos. La varianza y la desviación estándar son especialmente sensibles a los valores atípicos, ya que se basan en las desviaciones al cuadrado.

- Una estimación más robusta de variabilidad es la Desviación Absoluta Mediana de la mediana, también llamada MAD:

$$\text{Desviación absoluta mediana} = \textit{Mediana}(|x_1 - m|, |x_2 - m|, ..., |x_N - m|)$$

donde m es la mediana. Al igual que la mediana, MAD no está influenciada por los valores extremos. También es posible calcular una desviación estándar recortada análoga a la media recortada.

Estimaciones basadas en percentiles

- Un enfoque diferente que se puede emplear para estimar la dispersión se basa en observar la separación de los datos clasificados. Las estadísticas basadas en datos ordenados, o clasificados, se denominan estadísticas de orden. La medida más básica es el rango: la diferencia entre el número más grande y el más pequeño. Es útil conocer los valores mínimos y máximos en sí mismos y para identificar los valores atípicos, pero el rango es extremadamente sensible a los valores atípicos y no es muy útil como medida general de dispersión.

- Para evitar la sensibilidad a valores atípicos, podemos mirar el rango de los datos después de eliminar los valores de cada extremo. Formalmente, este tipo de estimaciones se basan en diferencias entre percentiles. En un conjunto de datos, el n-ésimo percentil es un valor tal que, al menos, el n por ciento de los valores toman este valor o menos $(100 - n)$ por ciento de los valores toman este valor o más. Por ejemplo, si queremos conocer el percentil 80, ordenamos los dartos. Luego, comenzando con el valor más pequeño, continuamos el 80 por ciento del camino hasta el valor más grande. Ten en cuenta que la mediana es lo mismo que el percentil 50. El percentil es, esencialmente, lo mismo que un cuantil, con cuantiles indexados por fracciones (el cuantil .8 es el percentil 80).

- Una medida común de variabilidad es la diferencia entre el percentil 25 y el percentil 75, también llamado rango intercuartílico (o IQR). Por ejemplo, queremos conocer el IQR del conjunto de datos [3, 1, 5, 3, 6, 7, 2, 9]. Ordenamos los datos para obtener [1, 2, 3, 3, 5, 6, 7, 9]. El percentil 25 está en 2.5, el percentil 75 está en 6.5, por lo que el rango intercuartílico es 6.5 - 2.5 = 4.

Correlación

- El coeficiente de correlación da una estimación de la relación entre dos variables. El coeficiente de correlación de Pearson se calcular multiplicando las desviaciones de la media de la variable 1 por las de la variable 2, dividiendo por el producto de las desviaciones estándar:

$$r(x,y) = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{(N-1)s_x s_y}$$

- La correlación puede oscilar entre -1 y 1. Valores cercanos a 1, indican que existe una relación más cercana entre las dos variables; valores negativos cercanos a -1 indican una relación inversa. Por último, valores cercanos a 0, significan que no hay relación.
- Hay que recordar que "correlación no significa causalidad". En otras palabras, el hecho de que dos variables estén correlacionadas no significa que se afecten entre sí.

Covarianza

- La covarianza es el valor que refleja en qué cuantía dos variables aleatorias varían de forma conjunta respecto a sus medias. Nos permite saber cómo se comporta una variable en función de lo que hace otra variable. Es decir, cuando X sube ¿Cómo se comporta Y? Así pues, la covarianza puede tomar los siguiente valores:
  1. Si $Covarianza(X, Y)$ es menor que cero, entonces cuando X sube, Y baja. Hay una relación negativa.
  2. Si $Covarianza(X, Y)$ es mayor que cero, entonces cuando X sube, Y sube. Hay una relación positiva.
  3. Si $Covarianza(X, Y)$ es igual a cero, no hya relación entre X y Y.

- La fórmula de la covarianza se expresa como sigue:

$$Cov(x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N}$$