Medidas de dispersión

Héctor de la Torre Gutiérrez
hdelatorreg@up.edu.mx

Medidas de dispersión

El paso a seguir, una vez que se han estimado los indicadores de tendencia central, son las medidas de dispersión.

Dispersión, se refiere al efecto de dispersar; y a su vez dispersar significa "separarse entre sí" en este caso las observaciones.

Entonces las medidas de dispersión indican la variación de los datos; cabe destacar que son regularmente complementarias a las medidas de tendencia central.

Las medidas usuales de dispersión son:

- Rango
- Rango interquartil
- Desviación absoluta
- Varianza
- Desviación estándar
- Coeficiente de variación

Medidas de dispersión

Las medidas de dispersión aplican fundamentalmente a variables continuas (razón e intervalo). En este caso, es importante entonces hacer las siguientes preguntas:

- ¿Para que sirve conocer la variación?
- ¿Qué es mejor, mayor variación o menor variación?
- ¿Cómo influye en la interpretación y/o análisis posteriores de los datos?

El Rango

Si se trata de ver como varían los datos, la medida "quizá" más simple y obvia es la diferencia entre la observación mayor y la menor; es decir, entre el máximo y el mínimo.

- El rango de un conjunto de datos (no agrupados) se define, precisamente como:

$$R = Max(X_1, X_2, \dots, X_n) - Min(X_1, X_2, \dots, X_n) = X^{(n)} - X^{(1)}$$

- Como puede verse, es un cálculo muy sencillo, solo requiere la identificación de los valores que determinan su estimación.
- Si bien da una idea de la variación de los datos, su interpretación es muy vaga en ese sentido.

El Rango

El estadístico del Rango, requiere algo más que solo su valor para que sea de utilidad; un rango de 20 unidades, no indica si se trata de miles o cientos.

- Si el promedio está más cerca del Máximo, hay mayor frecuencia de valores grandes (mayores al promedio) que valores chicos (menores al promedio).
- También es importante notar que al usar solo dos valores se pierde información valiosa de la muestra.
- Por estas razones, el Rango, no es la medida más apropiada para estimar la dispersión de la muestra.

El Rango Inter-quartil:

Una variación modificada del Rango y que también suele usarse es el llamado Rango inter-quartil.

- "Quartil" significa que los datos se dividen en "cuatro" partes iguales; es decir los puntos que corresponden al 0.25, 0.50 y 0.75 de probabilidades.
- Al punto de 0.25, se le denomina "primer quartil"- $X_{0.25}$; al de 0.5, corresponde al segundo quartil (o Mediana)- $X_{0.50}$; y el de 0.75 - $X_{0.75}$ al tercer quatil.
- El Rango interquartil, RI, es entonces :

$$RI = X_{0.75} - X_{0.25}$$

El Rango Inter-quartil:

- Es importante, en este caso identificar los puntos correspondientes a cada cuarta parte de los datos; 0.25, 0.50 y 0.75. ¿Cómo?
- El proceso es como sigue:
  - Se ordenan los datos de menor a mayor.
  - De la cantidad de datos, *n*, los puntos correspondientes a los cuartiles son:

    $\text{Observación}_p[n * p]$; donde $p = quantil\ deseado\ (0.25, 0.75, etc)$
  - El valor $[n * p]$ redondeado, indica la posición ordenada de la observación en los datos; por ejemplo, si el resultado de la operación es 12 y $p = 0.25$, la observación ordenada en el lugar 12 corresponde al $X_{0.25}$
- Aunque el cálculo de Rango Inter-quartil es un poco más elaborado, que el Rango, su principal objetivo es eliminar los datos extremos que suelen afectar mucho las estimaciones, en este caso el Rango.

Medidas de dispersión

Como ejemplo usemos los datos originales de los pesos de estudiantes de secundaria.

69	68	71	79	63
71	72	66	70	61
67	59	67	71	64
66	76	62	78	61
59	67	62	75	60
73	74	57	65	62
80	67	71	61	65
70	74	56	57	59
72	64	69	59	70
69	63	65	66	77

- El valor máximo corresponde a 80 ; y el mínimo a 56; por lo tanto el Rango, R, es
$$R = 86 - 56 = 24$$
- El 24 por si solo no dice mucho; sin embargo si se incluye que la mediana es 67, significa que los datos tienen una variación de 24 unidades centradas en el 67.
- De esta manera nos da mejor idea de la estructura de los datos (la forma de la distribución?)

Medidas de dispersión

- Con respecto al Rango inter-quartil, primero identificamos las observaciones que corresponden a cada quartil;

$$Observación_{0.25} = 50 * 0.25 = 12.5 \approx 13$$
$$Observación_{0.75} = 50 * 0.75 = 37.5 \approx 38$$

Entonces, $X_{0.25} = 62$ y $\boldsymbol{X_{0.75} = 71}$

- Por lo tanto el Rango inter-quartil es:

$$RI = 71 - 62 = 9$$

Medidas de dispersión

- La interpretación es diferente a la del rango, ya que usa información adicional; en este caso se diría que el 50% de las observaciones centrales varían en 9 unidades centradas en 67 (mediana) cambian de 62 hasta 71.

- Otra interpretación interesante sería que al seleccionar una observación de forma aleatoria; con el 50% de probabilidad estará entre 62 y 71.

Otras ideas de como medir la dispersión:

- Tomar el promedio de los valores absolutos de las diferencias.
- Tomar el “promedio” de las diferencia al cuadrado.
- Tomar raíz cuadrada de la opción anterior.

- Cada una de ellas, dio origen a medidas de dispersión; a saber: La Desviación Media Absoluta; La Varianza y la Desviación Estándar, respectivamente.
- Más adelante se darán detalles del porqué se usan más la Varianza y Desviación Estándar como medidas de dispersión que la Desviación Media Absoluta.

Desviación Media Absoluta (DMA):

Por lo expuesto, la expresión de cálculo de la DMA es:

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

donde:

| | indica la función valor absoluto.

- Dado que solo toma en cuenta la magnitud de los valores (no el signo), la interpretación es el promedio de las desviaciones (los errores) independientemente si son positivos o negativos. Si es cercana a cero, nos dice que en promedio, los errores respecto al promedio son “pequeños”.
- La principal desventaja de la medida es que, en general, la función valor absoluto no es fácil de manejar analíticamente.

Varianza (S^2)

Una alternativa usual en vez del valor absoluto, es elevar al cuadrado los errores y tomar el "promedio". Esta es la definición de varianza.

- La varianza, es una medida de dispersión muy usada en la descripción de datos, sobre todo cuando se hacen inferencias.
- Por definición entonces, la varianza de un conjunto de datos es:

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

Varianza (S^2)

- La principal desventaja es sobre la interpretación; ya que al estar elevados al cuadrado los errores, esta debe ser con unidades cuadradas.

- Por ejemplo:
  - Con kilogramos, la varianza tendría como unidades kilogramos cuadrados.
  - Con temperatura, serían gracias centígrados al cuadrado.
  - Años de escolaridad, años cuadrados.
  - Longitud, las unidades serían longitud al cuadrado (áreas), etcétera

Desviación Estándar (S)

- La forma “lógica” para evitar las unidades al cuadrado de la varianza, es tomar su raíz cuadrada; lo que da origen a la desviación estándar.
- Por lo anteriormente dicho, la expresión de la desviación estándar es:

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

- De aquí, la interpretación es similar a la de la DMA, con unidades iguales a los datos originales. Es decir, que tanto se “desvían” en promedio los datos de la media aritmética.

Medidas de dispersión.

- Como interpretación se nota que en promedio la desviación absoluta promedio las observaciones distan del promedio en **5.0208 Kg** sin distinguir si son menores o mayores al promedio..
- La varianza S^2 indica que hay una desviación de **37.5710 Kg^2** - como interpretación específica del problema, no tiene mucho sentido. La solución es usar la desviación estándar.
- En este caso S toma el valor de **6.1295 Kg; es decir, los datos varían en promedio 6.1295 kilogramos respecto al promedio.**
- Sobre el CV, el valor dice que los datos respecto al promedio varían en el **9.15%**. Es decir, los valores como porcentajes respecto al promedio tienen una desviación estándar de **9.15%**.

Coeficiente de Variación (CV)

- Una medida de variación muy usada en términos prácticos, sobretodo en el proceso de estadística descriptiva es el Coeficiente de Variación (CV).
- El CV toma en cuenta simultáneamente la medida de tendencia central (el Promedio) y la de dispersión (Desviación Estándar).
- Como ya se comentó, la medida de dispersión no dice mucho por sí sola; el CV en este caso es una buena alternativa ya que se resume en un solo número.
- En la teoría de muestreo, es muy útil para medir la precisión de los estimadores de los parámetros poblacionales (no se discute aquí).

Coeficiente de Variación (CV)

- También se usa para comparar la dispersión de poblaciones que quizá no es fácil de detectar.
- El CV se define como:

$$CV = \frac{S \times 100}{\bar{X}}$$

- La medida es adimensional (no tiene unidades), por lo que su interpretación no depende de que se este analizando.
- La expresión nos dice que tanta dispersión tiene un conjunto de datos con respecto a la media (por unidad).

Continuemos con …..R/Python