Medidas de tendencia central y dispersión

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

Introducción

Con los datos, se hacen estimaciones que fortalecen tanto la visualización gráfica como tabular de los datos.

Las estimaciones usuales son:

- Medidas de tendencia central.
- Medidas de dispersión.

- Mientras las medidas de tendencia central indican donde se "centran" los datos, las de dispersión indican la variación de los datos.

- Ambas medidas son útiles tanto en la interpretación "**numérica**" como en la implementación de análisis de datos formales, con sus correspondientes pruebas de hipótesis.

Revisión del operador $\sum$

Para un conjunto de datos $X_1, X_2, \ldots, X_n$, la suma de los términos se abrevia como:

$$\sum_{i=1}^{n} X_i$$

- El subíndice *i* identifica los valores inicial (1) y final (n) de los valores considerados en la sumatoria.
- Por ejemplo, si la suma no inicia en 1 sino en 3 (por ejemplo), y termina en 12 se escribe como:

$$\sum_{i=3}^{12} X_i$$

Si se quiere expresar la suma de cuadrados de los términos se escribe como:

$$\sum_{i=1}^{n} X_i^2$$

Puede aplicar a más de dos variables, por ejemplo:

$$\sum_{i=1}^{n} Y_i X_i = Y_1 X_1 + Y_2 X_2 + \ldots + Y_n X_n$$

- Las siguientes reglas aplican para el operador:

$$\sum_{i=1}^{n}(X_i \pm Y_i) = \sum_{i=1}^{n} X_i \pm \sum_{i=1}^{n} Y_i$$

$$\sum_{i=1}^{n} cX_i = c\sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} c = nc$$

Medidas de Tendencia Central.

Las medidas usuales de tendencia central son:

- Media (promedio)
- Mediana
- Moda

La Moda

Es el valor más frecuente, es decir, el que más se repite.

- Cuando la variable es categórica, ya sea nominal u ordinal, es una medida muy usada ya que es muy sencillo interpretarla.
- Para datos continuos, se recurre regularmente a datos agrupados en intervalos de clase; así, se identifica intervalo de clase más frecuente y la moda se identifica como el punto medio de clase.
- Es importante comentar que en la práctica, pueden ocurrir casos donde hay más de una moda; es decir, hay más de un dato que tienen frecuencias repetidas altas.
- Se puede determinar para datos agrupados y no-agrupados.

Ventajas:

- Simple de estimar.
- Útil en la descripción de datos categóricos, sobre todo nominales.
- Fácil de interpretar.
- No es sensible de datos extremos.

Desventajas:

- En variables continuas suele NO ser la medida de tendencia central más apropiada.
- No tiene propiedades estadísticas deseables (teóricas).
- No siempre está en el centro de la distribución (datos asimétricos).

La Mediana

Es el valor que divide a la población (ordenada) en dos partes iguales; por lo que, no aplica a datos nominales.

- Aplica a datos fundamentalmente continuos; aunque puede aplicarse en datos ordinales. NO DEL TODO RECOMENDABLE.
- Intuitivamente, es el punto medio de la distribución (de frecuencias).
- A diferencia de la moda, la mediana si es única.
- Cuando el conjunto de datos es pequeño, no es difícil determinar la mediana; sin embargo, conforme el número de datos crece, el ordenar los datos puede marcar una tarea tediosa más que difícil (si no se cuenta con una herramienta de cómputo).
- Afortunadamente, la mayoría de las herramientas de cómputo ya cuentan con una función que puede dar orden a los datos sin ninguna dificultad.

Para datos no agrupados, dado que es el punto (dato) que divide a la población en dos partes iguales, la medida usa los datos ordenados.

- Cuando el tamaño de la muestra es un número non, el dato central es la mediana; es decir, so $n = 2k + 1$, entonces:

$$Me = X_{k+1}$$

- Si es un número par; es decir, $n = 2k$, la mediana es el promedio de los números del centro:

$$Me = \frac{1}{2}(X_{k-1} + X_{k+1})$$

- Por ejemplo si la muestra ordenada es:

2, 3, 4, 4, 5, 5, 6, 6, 7, 8, 10, 12, 16

- El total de datos es 13, por lo que $n = 2k + 1 = 2(6) + 1 = 13$, por lo que $k = 6$; por lo tanto, el dato que divide a la muestra en dos partes iguales es el 6 (7º dato ordenado):

$$Me = 6$$

- Si la muestra fuese 2, 3, 4, 4, 5, 5, 6, 7, 8, 10, 12, 16, $2(6) = 12$ por lo que $k = 6$ y los dos datos centrales son 6 y 7 (6º y 7º datos ordenados); por lo tanto la mediana es:

$$Me = \frac{1}{2}(6 + 7) = 6.5$$

La Media

- Cuando se tienen datos continuos, el promedio en la gran mayoría de los casos es "la mejor" medida de tendencia central.
- Si bien es común usarla con datos nominales y ordinales, lo que cambia es la interpretación de esta. NO ES DEL TODO CORRECTO DICHO USO.
- Por ejemplo, si los datos son nominales (hombre mujer; preferencia política; religión; etcétera), se puede mostrar que la media corresponde a las proporciones de cada categoría.
- Cuando es ordinal, (por ejemplo: grado de avance en una enfermedad; nivel de marginación; número de hijos, etcétera) el promedio con decimales no tiene sentido.

La Media

- Para datos no agrupados, la media– también referida como promedio o media aritmética - para un conjunto de datos $X_1, X_2, \ldots, X_n$, se define como:

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

- Lo más importante en cuanto a la estimación de la media es la interpretación; es decir, ¿Que significa desde un punto de vista estadístico el promedio?

La Media

- Inicialmente podemos mostrar los pesos de los estudiantes de secundaria; no hay distinción de otros aspectos en los estudiantes como:
  - Sexo,
  - Estatura,
  - Origen social,
  - Por mencionar algunos

- Recuérdese que se cuenta con 50 datos de estudiantes.

La Media

- Los siguientes datos corresponden a los pesos (kg) de 50 niños de 1º 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

$$\bar{X} = \frac{1}{50}\sum_{i=1}^{50} X_i = \frac{1}{50}(69 + 68 + 71 + \cdots + 66 + 77 = \frac{1}{50}(3{,}349) = 66.98$$

Interpretación: El peso promedio de los estudiantes de secundaria es de aproximadamente 67 kilogramos, independientemente del sexo, altura, origen social, etcétera.

Interpretación.

- En promedio en **México** en cada hogar **familiar** habitan 2.4 **hijos** e hijas.

  Interpretación: “En cada hogar familiar en México en promedio viven entre 2 y 3 tres hijos e hijas”.

- El porcentaje de divorcios en México asciende al 41% de acuerdo con datos del INEGI

  Interpretación: 41 parejas deciden **divorciarse** por cada 100 que se casan, de acuerdo con datos poblacionales del Instituto Nacional de Estadística y Geografía (INEGI 2019).

Continuemos con …..Python/R