Medidas de tendencia Central

Estadígrafos de tendencia central.

Cada vez que se observa un fenómeno cuantitativo, nos interesa saber si los datos recolectados se aglutinan en torno a ciertos valores representativos que son propios del fenómeno estudiado. Por ejemplo si pensamos en la Edad de los jugadores beisbol, la experiencia nos dice que sus edades varían entre los 17 y 35 años, siendo raro pero no imposible, encontrar jugadores con más de 35 años o menores de 17 años, además sabemos que la gran mayoría de estos jugadores tienen entre 23 y 30 años. Ahora la pregunta general se hace obvia, dada una colección de datos, ¿es posible saber a qué valores tienden dichos datos?, la respuesta la entregan los llamados estadísticos de tendencia central.

En consecuencia llamamos estadísticos de tendencia central a aquellos valores hacia los cuales tienden a aglomerarse los datos de una muestra. Los más utilizados son:

1) **La Media aritmética o Promedio aritmético:** es el estadígrafo de tendencia central más conocido, usado y abusado. Dada una colección de datos $X_1, X_2, ...X_n$, el promedio se define como LA SUMA DE LOS DATOS DIVIDIDA POR LA CANTIDAD DE DATOS y se denota por $\overline{X}$, en símbolos el promedio es:

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\sum X}{n}$$

Formalmente, representa el Centro de Masas de la muestra, en la práctica, esto significa que se puede considerar que cada dato tiene valor igual al promedio. Esta idea no es tan lejana, pues en el lenguaje corriente, muchas veces se habla del “hombre promedio” de la “familia promedio ”, etc. es decir hablamos de un sujeto TIPO al cual asimilamos a todos los sujetos estudiados. Obviamente esta asimilación podría resultar errónea, como veremos más adelante.

El promedio aritmético posee las siguientes propiedades frente a cambios de escala:

$$\overline{X \pm a} = \overline{X} \pm a$$
$$\overline{a \cdot X} = a \cdot \overline{X}$$
$$\overline{a} = a$$

Veamos un ejemplo numérico: si las calificaciones de un escolar en matemáticas durante un semestre son 50, 60, 30, 50 y 70, su calificación promedio es:

$$\bar{x} = \frac{50+60+30+50+70}{5} = 50.2$$

Es decir podemos asumir, que su rendimiento general en matemáticas es 50.2.

Si se dispone de una base de datos de gran tamaño, resulta trabajoso calcular $\overline{X}$. En este caso podemos obtener un valor aproximado para $\overline{X}$, a partir de la información que contiene una tabulación, esta aproximación se obtiene de

MULTIPLICAR LOS PUNTOS MEDIOS DE CADA INTERVALO POR LAS RESPECTIVAS FRECUENCIAS, SUMAR ESTOS PRODUCTOS Y LUEGO DIVIDIR POR LA CANTIDAD DE DATOS. Ejemplifiquemos usando la tabla de Edad de los consumidores de determinada bebida o producto.

Edad	frec	porcentaje	F
10-19	2	6.67	6.67
20-29	5	16.67	23.33
30-39	13	43.33	66.67
40-49	7	23.33	90.00
50-59	3	10.00	100.00
Total	30	100.00	

Así: $\bar{x} = \frac{15 \cdot 2 + 25 \cdot 5 + 35 \cdot 13 + 45 \cdot 7 + 55 \cdot 3}{30} = \frac{1090}{30} = 36.3$ años.

Cuando se usa el promedio como medida de centralización, debemos tener cuidado de que los datos sean homogéneos, es decir razonablemente parecidos, pues el promedio es muy sensible a valores extremos, esto es valores demasiado elevados o demasiado minimizados. En estos casos el promedio como resumen

“miente”. Por ejemplo, supongamos que preguntamos por el sueldo anual, en miles de pesos, a cinco personas vigilantes en edificios distintos, obteniendo 140, 150, 142, 160 y el sueldo del último encuestado sea 350, puesto que trabaja para una Empresa de Ensueños, al observar los datos vemos que los sueldos de los vigilantes, en general están alrededor de los $ 150 mil, sin embargo si los promediamos tenemos que dicho promedio es de $ 188,400 obviamente esta distorsión se produce por el astronómico sueldo de $ 350 mil. En estos casos, lo justo es no incluir en el promedio el sueldo astronómico, con lo que el promedio es de $ 148 mil o bien en vez del promedio usar el valor mediana que es $ 150 mil, lo que concuerda con la realidad que estamos estudiando.

2) **La Mediana (me):** es aquel valor que divide la muestra en dos partes iguales, de esta definición nos damos cuenta que la mediana no es otra cosa que el Percentil cincuenta o Cuartil 2, es decir Mediana=P_{50}=Q_2 . Notemos que la mediana es tanto un estadígrafo de posición y de centralización.

El procedimiento para calcular la mediana es el siguiente:

1- Ordenar los datos de menor a mayor
2- Calcular (n+1)/2
3- Contar los espacios con el valor indicado del paso 2, en el conjunto de datos ordenados

- **Percentil** de orden k = cuantil de orden k/100

  La mediana es el percentil 50

  El percentil de orden 15 deja por debajo al 15% de las observaciones. Por encima queda el 85%

  Cuartiles: Dividen a la muestra en 4 grupos con frecuencias similares.

  Primer cuartil = Percentil 25 = Cuantil 0,25

  Segundo cuartil = Percentil 50 = Cuantil 0,5 = mediana

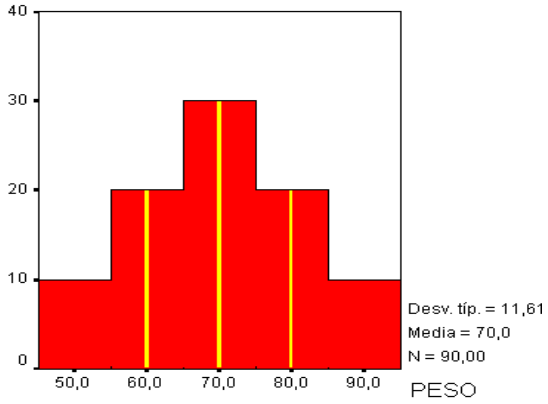
  Tercer cuartil = Percentil 75 = cuantil 0,75

Ejemplo

- ¿Qué peso no llega a alcanzar el 25% de los individuos?
  - Primer cuartil = percentil 25 = 60 Kg.

- ¿Qué peso es superado por el 25% de los individuos?
  - Tercer cuartil= percentil 75= 80 kg.
  - ¿Entre qué valores se encuentra el 50% de los individuos con un peso "más normal"?
  - Entre el primer y tercer cuartil = entre 60 y 80 kg.

Observar que indica cómo de dispersos están los individuos que ocupan la "parte central" de la muestra.



Estadísticos

PESO		
Percentiles	25	60,00
	50	70,00
	75	80,00

Número de años de escolarización

	Frecuencia	Porcentaje	Porcentaje acumulado
3	5	,3	,3
4	5	,3	,7
5	6	,4	1,1
6	12	,8	1,9
7	25	1,7	3,5
8	68	4,5	8,0
9	56	3,7	11,7
10	73	4,8	16,6
11	85	5,6	22,2
12	461	30,6	52,8
13	130	8,6	61,4
14	175	11,6	73,0
15	73	4,8	77,9
16	194	12,9	90,7
17	43	2,9	93,6
18	45	3,0	96,6
19	22	1,5	98,0
20	30	2,0	100,0
Total	1508	100,0	

Estadísticos

Número de años de escolarización		
N	Válidos	1508
	Perdidos	0
Media		12,90
Mediana		12,00
Moda		12
Percentiles	10	9,00
	20	11,00
	25	12,00
	30	12,00
	40	12,00
	50	12,00
	60	13,00
	70	14,00
	75	15,00
	80	16,00
	90	16,00

3) **La Moda (mo):** es aquel valor que más se repite en una muestra y se denota por m_o, por ejemplo si consideramos los datos 2,2,3,3,4,4,4,4,5,5 la moda en cuestión es 4. Sin embargo la definición dada cobra validez sólo si la variable es discreta.

Si la variable es continua, formalmente la moda no existe, pues es muy difícil que al sacar una muestra de números reales dos o más de ellos coincidan. Por ejemplo si se hilara muy fino y midiéramos el peso de las personas en milígramos, sería muy poco probable encontrar dos o más personas con igual peso en una muestra, pero generalmente el peso es medido en kilógramos enteros y en este caso, como se ha discretizado la variable es posible calcularla.

Cuando se dispone de un tallo y hoja, la moda corresponde al valor que más se repite dentro de la hoja más grande del tallo. Retomemos nuestro ejemplo de las edades de los bebedores de cerveza, el tallo y hoja se muestra a continuación, donde se ha destacado en negritas la hoja más larga:

```
1. | 69
2* | 23
2. | 677
3* | 112223
3. | 5555679
4* | 24
4. | 55568
5* | 1
5. | 78
```

Observamos que el valor más repetido en esta hoja es 35 años, que corresponde al valor de la moda.

Podemos utilizar la formula de la moda donde:

L_i : límite inferior del intervalo que contiene la frecuencia más alta

D_p : diferencia entre la frecuencia más alta y la del intervalo siguiente

D_a : diferencia entre la frecuencia más alta y la del intervalo anterior

A: ancho del intervalo

Veamos el ejemplo respectivo: Consideremos la tabulación de las edades, donde se ha ennegrecido el intervalo y la frecuencia modal:

Edad	frec	porcentaje	F
10-19	2	6.67	6.67
20-29	5	16.67	23.33
30-39	**13**	**43.33**	**66.67**
40-49	7	23.33	90.00
50-59	3	10.00	100.00
Total	30	100.00	

Aquí:

L_i : 30 años

Da : 13-5=6

Dp : 13-7=8

A: Ls-Li

con lo que:

$$M_o = Li + (\frac{8}{6+8})10 = 35.71 años$$

valor muy coincidente con el calculado a partir del tallo y hoja.

- Media ('mean') Es la media aritmética (promedio) de los valores de una variable. Suma de los valores dividido por el tamaño muestral.
  - Media de 2,2,3,7 es (2+2+3+7)/4=3,5
  - Conveniente cuando los datos se concentran simétricamente con respecto a ese valor. Muy sensible a valores extremos.
  - Centro de gravedad de los datos
  - Mediana ('median') Es un valor que divide a las observaciones en dos grupos con el mismo número de individuos (percentil 50). Si el número de datos es par, se elige la media de los dos datos centrales.
  - Mediana de 1,2,4,**5**,6,6,8 es 5
  - Mediana de 1,2,4,**5**,**6**,6,8,9 es (5+6)/2=5,5
  - Es conveniente cuando los datos son asimétricos. No es sensible a valores extremos.
    - Mediana de 1,2,4,**5**,6,6,800 es 5. ¡La media es 117,7!
- Moda ('mode') Es el/los valor/es donde la distribución de frecuencia alcanza un máximo.

Breve resumen

- Posición

  Dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos.

  - Cuantiles, percentiles, cuartiles, deciles,...
- Centralización

  Indican valores con respecto a los que los datos parecen agruparse.

  - Media, mediana y moda

- Dispersión
  - Indican la mayor o menor concentración de los datos con respecto a las medidas de centralización.
    - Desviación típica, coeficiente de variación, rango, varianza
- Forma
  - Asimetría
  - Apuntamiento o curtosis

Peso	**M. Clase**	**Fr.**	**Fr. ac.**
40 – 50	45	5	5
50 – 60	55	10	15
60 – 70	65	21	36
70 - 80	75	11	47
80 - 90	85	5	52
90 - 100	95	3	55
100 – 130	115	3	58
		58	

$$x = \frac{\sum_i x_i n_i}{n} = \frac{45 \cdot 5 + 55 \cdot 10 + \ldots + 115 \cdot 3}{58} = 69{,}3$$

$$me = Li + \frac{\left(\frac{n}{2} - Fa\right)}{f} A \qquad me = 60 + \frac{\left(\frac{58}{2} - 15\right)}{21} 10$$

$$mo = Li + \frac{(Da)}{Da+Dp} A \qquad me = 60 + \frac{(21-11)}{(21-11)+(21-10)} 10$$