A. Medidas de Tendencia Central

I. Media aritmética (o promedio): este concepto, por todos conocido, es lo que usualmente se entiende como mejor representación de un conjunto de datos. Es decir, si se tiene $x_1, x_2, ..., x_n$ datos (u observaciones), entonces $\bar{x}$ será el valor que mejor los representa (en una gran cantidad de situaciones). La media siempre trata, en virtud de representar lo mejor posible al conjunto, de buscar el equilibrio o simetría con respecto a ellos (en Física se la entiende como una analogía del centro de gravedad del conjunto).

La forma de encontrarla es sencilla:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{(media muestral)}$$

si hemos tomado una muestra de n datos de una población de tamaño N, o

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \quad \text{(media poblacional)}$$

si hemos medido todos los datos de la población de tamaño N.

En Estadística esta distinción es importante porque la media poblacional es el verdadero promedio (notar que hay una única media poblacional) y en el caso de la media muestral (se lee "equis raya") esta es una estimación de la media poblacional (y existen tantas medias muestrales como muestras se puedan tomar)

Ejemplo 1.

A continuación se presentan los datos de las ocho calificaciones. La segunda tabla difiere de la primera en que los valores repetidos se le asigna una frecuencia distinta de uno. El cálculo de la media muestral se muestra para la segunda tabla.

x_1	9
x_2	5
x_3	7
x_4	6
x_5	4
x_6	4
x_7	10
x_8	9
$\bar{x}$	6,75

x_i	f
9	2
5	1
7	1
6	1
4	2
10	1
$\bar{x}$	6,75

$$\bar{x} = \frac{2 \cdot 9 + 5 \cdot 1 + 7 \cdot 1 + 6 \cdot 1 + 4 \cdot 2 + 10 \cdot 1}{8} = 6{,}75$$

Así, la calificación promedio es 6,75 y nos da una idea de cómo se comporta la variable calificaciones mediante un único valor (en este caso 6,75 representa a todas las calificaciones)

La gran ventaja de la media es que es una medida que representa a todas las observaciones; es decir, toma en cuenta a todo el conjunto y es su punto de equilibrio.

<u>Desventajas de la media</u>

Ejemplo 2.

X: "Calidad obtenida en 7 muestras de superficies a las que se les aplicó una capa de cierta pintura" (Dato: los resultados se categorizan en Mala, Regular, Buena, Muy Buena, Excelente)

x_1	Mala
x_2	Regular
x_3	Buena
x_4	Buena
x_5	Muy Buena
x_6	Excelente
x_7	Excelente

¿Qué notamos en este caso con respecto a la variable? Esta es una variable cualitativa (o categórica), es decir no numérica, por lo que no tiene sentido (ni es posible) calcular la media. Entonces no podemos utilizar la media para datos cualitativos.

Un error muy común es pensar que se le puede asignar un código numérico a los resultados de la variable cualitativa, por ejemplo Mala=0, Regular=1, Buena=2, Muy Buena=3, Excelente=4 y luego promediar estos resultados. Es decir

$$\bar{x} = \frac{0 + 1 + 2 \cdot 2 + 3 + 2 \cdot 4}{7} \cong 2{,}23$$

¿Qué significado le daríamos este resultado? La variable no tiene resultados numéricos y aunque quisiéramos relacionarlo con los valores asignados, ¿diríamos, "es bastante de 2 (Buena) pero también un poco de 3 (Muy Buena)"? No tendría mucho sentido. Entonces, las variables cualitativas no son adecuadas para calcular en ellas la media aritmética

Ejemplo 3.

X: "Salario de una muestra de 11 trabajadores"

x_1	16000
x_2	16100
x_3	16500
x_4	16800
x_5	16900
x_6	17000
x_7	17100
x_8	17200
x_9	17300
x_{10}	17400
x_{11}	17500
$\bar{x}$	16890,91

La media representa adecuadamente al conjunto de salarios y podríamos decir que en promedio las personas ganan $16.890,91 (ninguna persona lo discutiría por ganar un poco menos o un poco más que ese monto)

Sin embargo, si por alguna razón incluyéramos en la lista un sueldo muy grande en comparación con los dados, por ejemplo $100.000,

x_1	16000
x_2	16100
x_3	16500
x_4	16800
x_5	16900
x_6	17000
x_7	17100
x_8	17200
x_9	17300
x_{10}	17400
x_{11}	100000
$\bar{x}$	24390,91

¿Qué pensará ahora cualquiera de los trabajadores que ganan entre \$16.000 y \$17.400 si les decimos que el salario promedio es de \$24.390,91? No estarán de acuerdo en absoluto pues éste es un valor muy por encima del salario más alto (dejando de lado el de \$100.000 por supuesto)

Aquí ha ocurrido que el salario de \$100.000 ha distorsionado a la media, pues ésta, como punto de equilibrio, debe compensar la perturbación producida por este valor y situarse por encima del anterior punto de equilibrio.

Haciendo una analogía con el clásico juego para niños de plaza sube-baja, es como si todos los chicos en conjunto pesaran lo mismo de un lado y del otro. En ese caso la tabla sobre la cual se sientan permanecerá en equilibrio. Pero si de pronto se agregara uno más (o si uno se reemplazara por otro de mayor peso) en la realidad el sistema perdería el equilibrio y un lado bajaría más que el otro. Sin embargo, si el punto de apoyo fuera la media este punto se desplazaría horizontalmente hacia el lado de mayor peso hasta lograr el equilibrio.

Resumiendo. Las desventajas de la media son:

- No puede utilizarse en datos cualitativos
- Se distorsiona ante la presencia de valores extremos (tanto pequeños como grandes) y deja de ser una buena representación del conjunto.

II. <u>Mediana</u>

La mediana, cuyo símbolo es $\tilde{x}$ o Me, es sencillamente aquella observación que se encuentra justo en la mitad del conjunto de datos, <u>habiendo primero ordenado al mismo de menor a mayor</u>.

- Si la cantidad *n* de datos es impar, entonces $\frac{n+1}{2}$ da la posición en el conjunto de datos donde encontrar la mediana.
- Si la cantidad n de datos es par, entonces la mediana es el promedio entre los valore que dan las posiciones $\frac{n}{2}$ y $\frac{n+2}{2}$

Así, en el ejemplo de los salarios, la cantidad de datos (11) es impar, por lo que la mediana se hallará en la $\frac{n+1}{2}=\frac{11+1}{2}=6 \text{ (sexta posición)}$

Vemos que en los datos, previamente ordenados de mayor a menor, en la sexta posición se halla el valor $17.000 que es la mediana.

x_1	16000	
x_2	16100	
x_3	16500	
x_4	16800	
x_5	16900	
x_6	17000	
x_7	17100	
x_8	17200	
x_9	17300	
x_{10}	17400	
x_{11}	17500	

En el caso de las calificaciones, la cantidad de datos es par (8 calificaciones)

x_1	4
x_2	4
x_3	5
x_4	6
x_5	7
x_6	9
x_7	9
x_8	10

En la tabla <u>ahora ordenada desde la menor a la mayor de las calificaciones</u> aparecen dos observaciones en la mitad; las calificaciones 6 y 7 que corresponden a las posiciones $\frac{n}{2}=\frac{8}{2}=4 \text{ y } \frac{n+2}{2}=\frac{8+2}{2}=5$. En ellas aparecen las calificaciones 6 y 7 y la mediana es el promedio aritmético entre ellas, es decir 6,5.

Ventajas de la mediana.

¿Qué ocurriría en los casos de la variable cualitativa analizada antes, o del sueldo de $100.000 que distorsionaba a la media?

x_1	16000	
x_2	16100	
x_3	16500	
x_4	16800	
x_5	16900	
x_6	17000	$\tilde{x}$
x_7	17100	
x_8	17200	
x_9	17300	
x_{10}	17400	
x_{11}	100000	
$\bar{x}$	24390,91	

Mientras que la media se ve distorsionada por el sueldo de $100.000 la mediana, que es el valor de la observación situada en la mitad del conjunto habiendo ordenado previamente los datos, sigue siendo $17.000 como antes, sin importar la existencia de este dato extremo. En este caso la mediana (el sueldo mediano $17.000) es una mejor representación de los salarios de este conjunto de personas que la media.

En el caso de la variable "Calidad de los ensayos de pinturas…"

x_1	Mala	
x_2	Regular	
x_3	Buena	
x_4	Buena	$\tilde{x}$
x_5	Muy Buena	
x_6	Excelente	
x_7	Excelente	

Vemos que ahora sí es posible (y tiene sentido) encontrar la mediana en la calidad de los ensayos de pintura. La mediana es "Buena" y es realmente un valor o respuesta de la variable que funciona como representativo. Por supuesto, tuvo que ordenarse los datos previamente desde *peor* calidad hasta *mejor* calidad y la cantidad de datos es impar (no podría promediarse dos observaciones centrales si fueran par porque no son resultados numéricos)

Resumiendo. Las ventajas de la mediana son

- Sí puede utilizarse en datos cualitativos (mientras la cantidad de datos sea impar)
- No se ve distorsionada ante la presencia de valores extremos (tanto pequeños como grandes) y es una mejor representación del conjunto frente a la media distorsionada.

III. <u>Media pesada o ponderada</u>

No siempre las observaciones de la variable (es decir los datos) tienen todas la misma importancia o peso dentro del conjunto.

Por ejemplo, si una evaluación tuviera tres instancias, Escrita, Oral y Antecedentes y para aprobarla debe obtenerse un promedio de 7, entonces podría ocurrir la siguiente situación

				Promedio
Instancia	Escrita	Oral	Antecedentes	
Calificaciones	8	8	5	7

Aquí el promedio es sencillamente $\frac{8+8+5}{3} = 7$

Pero quien evalúa bien podría desear que los antecedentes pesen más en el promedio y para no tener que imponer un mínimo en dicha instancia le asigna un peso o ponderamiento a cada una en función de la importancia que para dicha persona tienen. Así

				Promedio
Instancia	Escrita (0,2)	Oral (0,3)	Antecedentes (0,5)	
Calificaciones	8x0,2=1,6	8x0,3=2,4	5x0,5=2,5	6,5

Ahora a pesar de ser las mismas calificaciones, los pesos o pondermientos de cada instancia de evaluación son distintos y el promedio (la suma de las calificaciones ponderadas) ya no es 7 como antes sino 6,5. Esto es porque la instancia Antecedentes representa el 50% de la nota final y es necesario, en consecuencia, esforzarse más en esta evaluación para llegar al promedio mínimo de 7.

La media ponderada es

$$\bar{x} = \frac{\sum(w \cdot x)}{\sum w}$$

donde w es el peso de cada observación.

Nótese que siempre Σw=1

En el caso original, las tres instancias tienen el mismo peso. Es decir, un tercio de la nota final. Por eso

$$\bar{x} = \frac{8 \cdot \frac{1}{3} + 8 \cdot \frac{1}{3} + 5 \cdot \frac{1}{3}}{1} = \frac{8 + 8 + 5}{3} = 7$$

Puede verse así que la media aritmética o promedio es un caso particular de la media pesada o ponderada cuando todos los datos tienen el mismo peso o ponderamiento.

IV. Moda

- La **moda** de un conjunto de observaciones es el valor de las observaciones que ocurre con *mayor frecuencia* en el conjunto.
- La moda es la única medida de tendencia central que puede ser calculada para variables cualitativas nominales (no es necesario que tengan un orden)
- El valor de la moda no se ve afectada por la existencia de valores extremos.
- Puede suceder que en una serie de datos haya más de una moda. En tal caso se denomina bimodal o multimodal, según el número de modas que presente.

Ventajas de la moda

- La moda, al igual que la mediana, se puede utilizar como una posición central para datos tanto cualitativos como cuantitativos.
- La moda no se ve mayormente afectada por los valores extremos. Incluso si los valores extremos son muy altos o muy bajos, nosotros escogemos el valor más frecuente del conjunto de datos como el valor modal. Podemos utilizar la moda sin importar qué tan grandes o qué tan pequeños sean los valores del conjunto de datos, e independientemente de cuál sea su dispersión.
- Podemos calcular la moda aun cuando una o más clases sean de extremo abierto.

Desventajas de la moda

- A menudo, no existe un valor modal debido a que el conjunto de datos no contiene valores que se presenten más de una vez.
- Cuando los conjuntos de datos contienen muchas modas, resultan difíciles de interpretar y comparar.