Unidad 1. Tema 1.4 Medidas de resumen para variables cuantitativas
BIBLIOGRAFÍA BÁSICA

Las distribuciones de frecuencias son de gran utilidad en los análisis. El número de consultas médicas realizadas por especialidades por un centro de salud puede ser una información muy importante para la planificación de recursos.

Sin embargo, en ocasiones manejar un gran número de datos no es lo más aconsejable para determinados análisis, y si es muy beneficioso tener toda la distribución de una variable cuantitativa resumida en un solo valor o en muy pocos valores que representan con suficiente aproximación todo la distribución.

Las medidas de resumen para las variables cuantitativas se dividen en dos grandes grupos: las medidas de tendencia central y las denominadas medidas de posición.

Medidas de tendencia central.

Este grupo de medidas van a resumir toda una distribución, generalmente en un solo valor, que tiende a ocupar una posición central entre el menor valor y el mayor valor de la serie de datos y al rededor del cual se agrupan los valores que asume la variable. Ellas son la **media aritmética, la mediana y la moda.**

Media aritmética. También denominada promedio o simplemente media, es la más importante de todas las medidas de resumen. La media para su cálculo, tiene en cuenta todos los valores de la variable, este elemento hace que ella represente muy bien el comportamiento de la variable.

En una serie simple de datos, la media se obtiene sumando todos los datos y dividiéndolos entre el total de observaciones.

Ej.
Peso en kg. de 5 niños.
3,4,6,8,9,

La media se obtendría sumando los 5 pesos y dividiéndolos entre 5.

3+4+6+8+9/5=30/5

$\overline{X}$ =6

La media o el peso promedio de los 5 niños es 6kg.Su expresión de cálculo es:

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

Para una serie simple pero con las frecuencias absolutas previamente calculadas, la media se obtendrá multiplicando la frecuencia por el valor que asume la variable, sumar esos productos y dividirlos entre el total de observaciones.

Peso en kg. de 30 niños:

Peso (Kg)	No.
5	3
7	8
6	10
9	6
11	3
Total	30

$\overline{X}$ = (5x3)+ (7x8)+ (6x10)+ (9x6)+ (11x3) /30

$\overline{X}$ = 15+56+60+54+33/30

$\overline{X}$ = 218/30

$\overline{X}$ = 7.3kg.

La media de los pesos de los 30 niños es 7.3kg.

La expresión de cálculo es:

$$\overline{X} = \frac{\sum X.f_i}{n}$$

Veamos por último el cálculo de la media cuando se tienen los datos agrupados, cuando se tiene una escala con intervalos de clase.

Peso en kg. de 25 niños de una escuela primaria.

Peso en kg	**No.**
20–24	**4**
25–29	**10**
30–34	**8**
35–39	**3**
Total	**25**

Para el cálculo de la media tengo ahora la dificultad de no saber el peso exacto de los niños, los 4 niños que clasificaron en el primer intervalo de clase pueden haber tenido cualquier peso entre 20 y 24 kg. Para resolver este inconveniente asumimos que el peso de los niños es el valor central del intervalo de clase, la marca de clase.

Se procede de igual forma que para el cálculo de una serie simple con frecuencia. Multiplicamos la marca de clase por la frecuencia del intervalo, sumamos esos resultados y lo dividimos entre el total de observaciones.

Marca de clase	No	Mc. fr
22	4	88
27	10	270
32	8	256
37	3	111
Total	25	725

$\overline{X}$ = 725/25

$\overline{X}$ = 29kg

EL peso medio de los 25 escolares es de 29kg.

La expresión de cálculo es:

$$\overline{X} = \frac{\sum mc.f_i}{n}$$

Medias o promedios ponderados.

En ocasiones puede ocurrir que tengamos la información resumida por más de una media y después nos percatamos de la necesidad de calcular una media general.

Supongamos que le maestría en gerencia tiene una matrícula de 60 alumnos, 40 son varones y 20 son hembras. La edad promedio de los hombres es de 34.8 años y la de las mujeres es de 29.7 años. Si quisiéramos conocer la edad promedio general del grupo es lógico pensar que esas medias parciales que ya están calculadas por sexo me pueden servir para abreviar los cálculos.

Una tendencia generalizada es asumir que el promedio de las medias parciales sería la forma de cálculo de esa media general. Sin embargo, esa no es la forma correcta de operar, es necesario tener en cuenta el tamaño de las muestras de donde se obtuvieron esas medias, y para ello se multiplica el valor de la media por su **n** correspondiente. Este proceso, que se usa mucho en estadística se conoce con el nombre de **ponderación.**

La expresión de cálculo de la media ponderada sería entonces:

$$\overline{X} = \frac{n_1\overline{x_1} + n_2\overline{x_2} + n_3\overline{x_3}\ldots n_i\overline{x_i}}{n_1 + n_2 + n_3\ldots n_i}$$

La media, a pesar de ser la mejor medida de tendencia central y precisamente su precisión en los cálculos está dado en buena medida por el hecho de tener en cuenta todos los valores que asume la variable, esto puede convertirse en determinadas circunstancias en una limitación para su uso.

Supongamos que las edades de los estudiantes de la maestría se mueven en un rango de 25 a 35 años y al calcular la edad promedio fue de 30.5 años. Un profesional de cierto prestigio y que a dedicado buena parte de su vida a trabajar dentro del Sector Salud, matricula el curso a pesar de tener 60 años.

Al calcular la edad promedio de los cursistas ya no tendrá un valor de 30.5 años, valor que reflejaba muy bien el comportamiento del grupo, sino un valor mucho mayor que se aleja mucho del comportamiento real. En este caso, cuando existen valores extremos o aberrantes la media se ve afectada en su cálculo por los mismos y deja de ser una buena medida de resumen. Entonces preferimos para resumir la información la **mediana**, que como veremos a continuación, por su forma de cálculo no se ve afectada ante esta situación.

Mediana. Es la observación que en una serie **ordenada**, ocupa la posición central, por tanto divide a la serie en dos parte iguales. Por encima de ella se encuentra el 50% de las observaciones y a su vez su valor supera al 50% restante.

Por Ejemplo.

Peso en kg. de 5 niños.

5, 6, 7, 9, 11,

Ya se encuentra ordenada de menor a mayor, aunque a simple vista se aprecia que la mediana es 7kg, apliquemos la expresión para encontrarla:

5+1/2=3

En efecto la tercera posición corresponde al niño que peso 7kg.

Cuando la serie es par no hay un único valor central, la mediana sería en este caso la semisuma de los dos valores centrales.

3, 5, 6 ,7, 9, 11

El 6 y el 7 ocupan la posición central, si aplicamos la expresión matemática tendremos:
6+1/2=3.5

La mediana está entre la tercera y la cuarta posición.

Me = 6+7/2
Me = 6.5kg

La mediana por su propia forma de cálculo, no se verá afectada por valores extremos pues al realizar el ordenamiento de los datos estos serían la primera o la última observación.

Es una medida muy utilizada en epidemiología, sobre todo en los estudios de series cronológicas.

Puede ser calculada para datos agrupados, pero se usa con poca frecuencia. En el caso de tener escalas abiertas, la media no puede ser calculada, porque los intervalos abiertos no tienen marca de clase, elemento presente en la expresión de cálculo. En este caso se prefiere utilizar la mediana que no se ve afectada por este hecho.

La moda es la menos utilizada de las medidas de tendencia central, y como su nombre lo indica es el valor más frecuente.

Ej. Edad en años de 10 pacientes con SIDA.

20, 27, 23, 35, 23, 40,32, 23, 19, 17

La moda en este caso es 23 años

En este caso:

20, 24, 15, 34, 20, 38, 24, 34, 24, 20

Hay dos modas, 20 y 24 años, se dice que esa distribución es bimodal.

Por último, puede ocurrir que la moda no exista:

20, 35, 25, 18, 36, 24, 19, 33, 27, 15,

Puede ser calculada para datos agrupados, en cualquier libro de Estadística Ud. puede encontrar la fórmula.

Medidas de dispersión.

Una distribución de frecuencias para una variable cuantitativa no puede ser resumida con solo utilizar una medida de tendencia central. El hecho estriba en que si bien el valor que esa medida asuma nos indica que el resto de las observaciones están ubicadas a su alrededor, no sabemos cuán lejos o cerca pueden estar situadas y es fácil entender que mientras más próximas estén las observaciones de esa medida, mejor reflejará está el comportamiento de la distribución.

Veamos un ejemplo.

Días de hospitalización de 5 niños ingresados en un servicio de E.D.A.

3, 5, 7, 9, 11
$\overline{X}$ = 7 días
Me= 7 días

Días de hospitalización de 5 niños en un servicio de I.R.A.

1, 2, 7, 12, 13

$\overline{X}$ = 7 días

Me= 7 días

Aunque en los dos ejemplos coinciden los valores de las medias y las medianas, los valores de las observaciones no se distribuyen alrededor de las mismas de igual manera.

En el caso de las E.D.A. la dispersión es menor, los valores se concentran más alrededor de la media y la mediana, lo que nos hace pensar que en este caso las medidas de tendencia central reflejan mejor el comportamiento de la variable.

Por consiguiente una medida de la dispersión, de la variabilidad, debe ser añadida a la medida de resumen para completar el análisis.

El **rango, recorrido o amplitud** se obtiene por la diferencia entre el mayor y el menor valor observado.

En el ejemplo de la E.D.A. sería: 11-3=8

Para las I.R.A. tendríamos: 13- 1=12

Aunque logra cuantificar la variabilidad de los datos, no es una buena medida porque no tiene en cuenta el resto de las observaciones, solo los dos valores extremos.

Por consiguiente, una buena medida, debe tener en cuenta todos los valore que asume la variable, e ir midiendo cuan cerca o lejos están de esa medida de tendencia central. Además para obtener un valor único, lo lógico es hacer un promedio de esas diferencias. Apliquemos estas ideas al ejemplo de las E.D.A.

$\overline{X}$ =7

(3-7)+ (5-7)+ (7-7)+ (9-7)+ (11-7)/5

(-5)+(-2)+(0)+(2)+(5)/5

Como puede apreciarse, al realizar la suma algebraica de las diferencias de los valores con respecto a la media, estos se anulan y es algo que va a ocurrir siempre, invariablemente.

Esta situación se da por que los valores se alejan de la media en dos sentidos, por defecto y por exceso, dicho de otra manera, unos serán menores que la media y otros la superarán.

Sin embargo, nuestro interés es tener una medida de la dispersión, pero no tiene que precisar en qué sentido se alejan de la media, lo que se traduce en que no nos interesa el signo que se obtenga al encontrar las diferencias.

Así surge la **desviación media**, y para ello lo que se calcula es la diferencia modular.

$$D.M.= \frac{\sum \|x_i - \bar{x}\|}{n}$$

Con el mismo ejemplo de las E.D.A. tendremos:

D.M.= (|3-7|)+ (| 5-7|)+ (|7-7|)+(|9-7|)+(|11-7|)/5
D.M.= 4+ 2+ 0+ 2+ 4/5
D.M.=2.4 días.

Los niños permanecieron ingresados en el servicio de E.D.A. en promedio 7 días con una dispersión o variabilidad promedio de 2.4 días.

Varianza. Otra forma de lograr que la suma de las diferencias de los valores de la variable con respecto a la media no se anulen, es elevando esta diferencia al cuadrado. De esta forma el promedio de las desviaciones al cuadrado es la medida que se conoce bajo el nombre de **varianza**.

Su expresión de cálculo para una serie simple de datos es:

$$S^2 = \frac{\sum (x - \bar{x})^2}{n}$$

Con el mismo ejemplo anterior tendremos:

$S^2 = (3-7)^2 + (5-7)^2 + (7-7)^2 + (9-7)^2 + (11-7)^2 /5$

$S^2 = 16+4+0+4+16/5$

$S^2 = 8$

Como el resultado sería 8 días al cuadrado, algo realmente difícil de entender y analizar, se le saca la raíz cuadrada a ese resultado y obtendremos un valor de fácil interpretación: 2.8 días.

Esta medida que hemos obtenido al calcular la raíz cuadrada a la varianza es la denominada **Desviación estándar**, la más conocida y utilizada de las medidas de dispersión.
Su expresión de cálculo para una serie simple es:

$S = \sqrt{(x_1 - \bar{x})^2} /n$

Para el cálculo con datos agrupados utilizaremos el mismo ejemplo que se trabajo para la media, los pesos de 25 niños en edad escolar, que fue de 29 kg.

Al igual que en el caso de la media, se asume que la marca de clase es el valor de la variable para el número de observaciones que clasificaron en el intervalo.

M.C	Fi	(M.C.– $\bar{x}$)^2	(M.C.– $\bar{x}$)^2 fi
22	4	7= 49	196
27	10	2= 4	40
32	8	3= 9	72
37	3	8= 64	192
Total	25		Σ500

S^2 =500/25=20 kg.

S= $\sqrt{}$ 20

S=4.5 kg.

Su expresión de cálculo es:

$$S^2 = \frac{\sum (mc_i - \bar{x})^2 f_i}{n}$$

La media de los pesos de los niños fue de 29 kg., con una desviación estándar de 4.5kg.Una forma muy intuitiva de interpretar esta medida es la siguiente, si a la media se le resta y se le suma la desviación estándar, se forma un intervalo donde se encuentra una buena parte de las observaciones. Para este ej. Entre 15.5 y 24.5kg se encuentra los pesos de una parte importante de los niños estudiados.

Todas las medidas de dispersión que hemos visto hasta ahora se consideran medidas de **dispersión absoluta**.

Cuando queremos saber entre 2 o más variables cuales presentan una mayor variabilidad o dispersión, no lo podemos conocer comparando algunas de las medidas de dispersión analizadas, pues los valores que alcancen dependerán de la unidad de medida propia de cada variable.

Para realizar estas comparaciones se utilizan las denominadas **medidas de dispersión relativas** concretamente el **coeficiente de variación.**

El coeficiente de variación indica el tanto por ciento de la media que representa la desviación estándar.

Cv = S/ $\bar{X}$ x100

Supongamos que el peso promedio de un grupo de estudiantes fue de 62kg, con una DE. de 6kg, y la media de la talla de 162cm con una desviación estándar de 8 cm. Si comparamos las DE. Parece ser la talla la más variable de las dos. Veamos que ocurre cuando calculamos el CV.

CV_{peso} = 6/62 x 100=9.6%

CV_{talla} = 8/162 x 100=4.9%

El peso tiene una mayor dispersión que la talla.