# Análisis de datos en los estudios epidemiológicos II

**Introducción**
En este capitulo continuamos el análisis de los estudios epidemiológicos centrándonos en las medidas de tendencia central, posición y dispersión, índices fundamentales para conocer mejor la distribución de los datos de un estudio.

**Los índices de posición y centralidad**
Otro de los aspectos fundamental a conocer de cualquier distribución de datos es la tendencia central y la posición que ocupan los datos respecto a un determinado valor. Siendo cierto que las distribuciones de frecuencia son un importante medio para ordenar un conjunto de datos e informar sobre algunos patrones de grupo, también lo es el hecho de que ofrecen poca información. En muchas investigaciones interesa mucho más conocer el resumen global de las características del grupo en estudio que podemos conseguir utilizando las medidas de tendencia central y de posición. Las medidas de tendencia central más comunes son: la media, la mediana y la moda, cada una de las cuales puede utilizarse como índice para caracterizar una distribución de datos. Son índices estadísticos que nos indican el valor de la variable hacia el cual tienden a agruparse los datos.
Las medidas de posición más utilizadas son*:* los cuartiles y los percentiles, índices que nos informan del orden o de la posición que ocupa un dato dentro del conjunto de los datos observados en una distribución.

**La moda.**
La moda de un conjunto de datos es el valor de la variable que se repite con mayor frecuencia. Es la medida de tendencia central más sencilla de calcular, ya que en realidad no se calcula sino que se observa. Se utiliza tanto para variables cualitativas como cuantitativas o cuasicuantitativas. Cuando estamos trabajando con variables cualitativas y cuasicuantitativas la moda se corresponde con la modalidad de la variable que más se repite, es decir, la de mayor frecuencia.
Imaginemos que hemos realizado un estudio para valorar el grupo sanguíneo en un grupo de mujeres embarazadas y obtenemos los siguientes datos:

| Grupo sanguíneo | Numero de mujeres |
|---|---|
| A | 14 |
| B | 11 |
| AB | 5 |
| O | 10 |
| **TOTAL** | **40** |

Como podemos observar la moda se corresponde con el grupo sanguíneo A, ya que hay 14 mujeres con este grupo sanguíneo.
Cuando trabajamos con variables cuantitativas tenemos que tener en cuenta si los datos obtenidos están o no agrupados en intervalos.

**Cálculo de la moda con datos no agrupados en intervalos**
Procedemos de la misma forma que en la situación anterior. En el estudio anterior conocemos las edades de las mujeres, que es la siguiente:

| Edad | n | Edad | n |
|---|---|---|---|
| 24 | 0 | 31 | 2 |
| 25 | 1 | 32 | 3 |
| 26 | 3 | 33 | 2 |
| 27 | 1 | 34 | 10 |
| 28 | 1 | 35 | 1 |
| 29 | 3 | 36 | 7 |
| 30 | 5 | 37 | 1 |

Como podemos observar la moda es 34 años, ya que es la edad que más se repite. Hay 10 mujeres que tienen 34 años.

**Cálculo de la moda con datos agrupados en intervalos**
Cuando los datos están agrupados en intervalos la moda corresponde con el punto medio del intervalo de mayor frecuencia. Imaginemos que en un estudio sobre los valores de colesterol hemos obtenido unos datos, que hemos procedido a ordenar en una tabla de distribución de frecuencias en intervalos.

| Intervalos | n | Frecuencia acumulada |
|---|---|---|
| 235,5-256,5 | 3 | 30 |
| 214,5-235,5 | 6 | 27 |
| 193,5-214,5 | 9 | 21 |
| 172,5-193,5 | 7 | 12 |
| 151,5-172,5 | 5 | 5 |

El punto medio del intervalo de mayor frecuencia es el de 193,5-214,4, ya que hay 9 personas cuyos valores de colesterol están correspondidos en este intervalo, luego la moda será la media de estos dos valores, es decir 204.
Tenemos que recordar que las distribuciones de frecuencia con una sola moda se denominan unimodales, con dos modas se denominan bimodales. Una distribución que contenga más de dos modas se denomina multimodal. Imaginemos que en el ejemplo anterior también 9 personas tiene sus valores de colesterol comprendidos en el intervalo 151,5-172,5, estamos frente a una distribución bimodal.
La moda es una medida poco utilizada en investigación o, al menos, como medida única de tendencia central, ya que tiene poco peso y fluctúa mucho de una muestra a otra. Sin embargo es frecuente su utilización y descripción en estudios de tipo demográfico, social, etc. Por ejemplo "Los sujetos tipo (modal) de estudio fueron niñas, de colegios privados de Madrid, del área metropolitana, con antecedente de anorexia".

**La media aritmética**
La media aritmética es una de las medidas de tendencia central más utilizada, ya que en ella se basan muchas de las pruebas de la estadística inferencial.

Se representa por $\overline{X}$ y se describe como la suma de todos los valores obtenidos de una variable, divididos por el número total de sujetos en estudio.
El cálculo de la media a diferencia del de la moda, solo se puede aplicar a las variables cuantitativas, por lo que tenemos que tener en cuenta también si los datos están o no agrupados en intervalos.
La fórmula general para el cálculo de la media es:

$$\overline{X} = \frac{\sum Xj}{N} = \frac{X_1 + X_2 + ... + X_n}{N}$$

**Cálculo de la media con datos no agrupados en intervalos**
Cuando los datos no están agrupados en intervalos el cálculo es sencillo y se reduce a aplicar la fórmula anterior.
Imaginemos que estamos estudiando los niveles de obesidad en un grupo de adolescentes de un instituto de enseñanza superior y tomamos una muestra de 10 niños y niñas, obteniendo los siguientes pesos:

47, 52, 54,48, 40, 45, 50,52, 46, 47

La media sería 48.1, es decir, el peso medio de este grupo es de 48.1 Kilos.

**Cálculo de la media con datos agrupados en intervalos**
Si tenemos los datos agrupados en intervalos el procedimiento varía en algunos aspectos:

1. Hay que calcular el punto medio de cada intervalo.
2. Hay que multiplicar este punto medio por la frecuencia correspondiente, es decir, por el número de personas que tiene sus valores en ese intervalo
3. El resultado del producto anterior se divide por N.

$$\overline{X} = \frac{\sum n_j X_j}{N}$$

- Siendo $n_j$ la frecuencia del intervalo
- Siendo $X_j$ el punto medio de cada intervalo.

La media es una medida muy sensible a la variación de las puntuaciones, basta con que varíe una sola puntuación para que varíe la media. No es recomendable su uso cuando la distribución de frecuencias que estamos estudiando tiene puntuaciones muy extremas.

**La mediana**
Es el punto por encima y debajo del cual quedan contenidos el 50% de los datos de una distribución de frecuencias, es decir, la puntuación que ocupa el nivel central.
Tal y como hemos indicado para el cálculo de la media, también hay que tener en cuenta en la mediana si los datos están o no agrupados en intervalos.

**Cálculo de la mediana con datos no agrupados en intervalos**
En este caso hay que tener en cuenta si el número de datos es par o impar. Si es impar procedemos a ordenar los datos de menor a mayor y aquel que ocupa la posición central es la mediana. Por ejemplo, tenemos los valores 2, 5, 7, 6, 4. Procedemos al cálculo de la mediana ordenándolos de menor a mayor 2, 4, 5, 6, 7. Como podemos observar la mediana es 5, la posición central.
Si el número de datos es par procedemos de la misma manera que en el caso anterior, ordenamos los datos y con aquellos dos que ocupan en nivel central se realiza la media aritmética, es decir, se suman los dos datos y se divide por dos.
Ejemplo tenemos los valores 2, 4, 5, 6, 7, 8. La mediana será: (5+6)/2=5.5

**Cálculo de la mediana con datos agrupados en intervalos**
Cuando los datos están agrupados en intervalos se aplica la siguiente fórmula

$$Md = Li + \left( \frac{\frac{N}{2} - n_d}{n_c} \right) \cdot i$$

Siendo $L_i$ el limite exacto inferior del intervalo crítico.
N el número total de datos
$n_d$ el número de datos pro debajo del intervalo critico
$n_c$ la frecuencia del intervalo critico
i = La amplitud del intervalo crítico

La mediana es menos sensible que la media a las variaciones de las puntuaciones. Es más representativa que la media cuando la distribución de frecuencias tiene puntuaciones muy extremas, puesto que la mediana depende de los valores centrales de la distribución y no se ve afectada por los valores extremos.
También son muy utilizados otros índices para informar de la distribución, tales como los cuartiles, deciles y especialmente los percentiles. Éstas no son medidas de tendencia central sino de posición que, como ya indicamos, nos informan del orden o posición que ocupa un dato dentro del total de los datos observados.

**Cuartiles y percentiles**
Definimos el percentil como el valor de la variable por debajo del cual se encuentra un porcentaje determinado de observaciones. Por ejemplo, si hablamos del percentil 35, expresado como P 35, estamos informando de que es el valor de la variable por debajo del cual se encuentran el 35% de todas las puntuaciones.
Por su parte los cuartiles son los valores de la variable que dejan por debajo de sí el 25%, 50% y 75% de las puntuaciones y se representan por $Q_1$, $Q_2$ y $Q_3$ .

- El primer cuartil es el valor de la variable que deja por debajo de sí el 25% de los datos.
- El segundo cuartil es el valor de la variable que deja por debajo de sí el 50% de los datos. Se corresponde con la mediana de la distribución y el percentil 50.
- El tercer cuartil es el valor de la variable que deja por debajo de si el 75% de los datos.

El cálculo es similar al que realizamos para calcular la mediana en datos agrupados.
Como conclusión a este apartado de las medidas de tendencia central podemos decir que, en general, la utilización de uno u otro índice dependerá entre otros de los siguientes aspectos:

- Del objeto de nuestra investigación
- Del tipo de variables en estudio
- De la distribución de los datos obtenidos en el estudio

**La variabilidad**
Como hemos podido ver las medidas de tendencia central no informan de la totalidad de la distribución. Dos conjuntos de datos pueden tener una media idéntica y sin embargo diferir en cuanto a la variabilidad de sus datos. Por tanto para describir de manera correcta una distribución de datos, no sólo necesitamos conocer los índices de tendencia central sino en qué medida cada dato de esa distribución se aparta del punto central que hemos determinado.
Para valorar esta variabilidad utilizamos lo que se denominan medidas de dispersión, entre las que se encuentran el rango o amplitud, la amplitud semiintercuartil, la desviación media, la desviación típica y la varianza y el coeficiente de variación. De todas ellas vamos a describir las características y el cálculo de la desviación típica, la varianza y el coeficiente de variación.

**La Varianza**

Se representa por $S^2_X$ y se define como la media de los cuadrados de las diferencias entre cada valor de la variable en estudio y la media de esa distribución de datos de la variable.
La fórmula para su cálculo es la siguiente:

$$S_x^2 = \frac{\sum \left(X_j - \overline{X}\right)^2}{N}$$

Desarrollando la formula podemos obtener esta expresión:

$$S_x^2 = \frac{\sum X_j^2}{N} - \overline{X}^2$$

Siendo

- $X_j$: Los valores de la variable
- $\overline{X}$ : La media
- N: número de datos de la muestra del estudio

La desviación típica se representa por $S_X$ y es igual a la raíz cuadrada positiva de la varianza.

$$Sx = \sqrt{S^2}$$

Utilizamos la varianza y la desviación típica solamente cuando nuestras variables de estudio son cuantitativas y, tal como indicamos para el cálculo de la media y la mediana, tenemos que tener en cuenta si nuestros datos están o no agrupados en intervalos.

**Cálculo de la varianza con datos no agrupados en intervalos**

En este caso su cálculo se limita a aplicar la fórmula. Vamos a proceder a su cálculo a través de un ejemplo: Imaginemos que quiero conocer la varianza y la desviación típica de las edades de un grupo de 5 mujeres que acuden a mi consulta de enfermería a un programa de menopausia, cuyas edades son: 44, 58, 62, 50, 52.
En primer lugar debemos calcular la media:

$$\bar{X} = \frac{\sum Xj}{N} = \frac{44 + 58 + 62 + 50 + 52}{5} = 53.2$$

Para calcular la varianza aplicamos la fórmula:

$$S_x^2 = \frac{\sum \left( Xj - \bar{X}^2 \right)^2}{N} = =$$

$$\frac{(44 - 53.2)^2 + (58 - 53.2)^2 + (62 - 53.2)^2 + (50 - 53.2)^2 + (52 - 53.2)^2}{5} = 39.36 \text{ años}^2$$

Puesto que la varianza se obtiene como resultado de una suma de cuadrados, tiene como unidades de medida el cuadrado de las unidades de medida en que se mide la variable estudiada. Por ello nuestra varianza se expresará en años al cuadrado.

**Cálculo de la varianza con datos agrupados en intervalos**

En este caso la fórmula para calcular la varianza es:

$$S_x^2 = \frac{\sum n_j \left( X_j - \bar{X} \right)^2}{N}$$

Siendo:
nj= frecuencia de cada intervalo
Xj = punto medio de cada intervalo
N= número total de datos

$\bar{X}$ = media

**La desviación típica**

Se representa por $S_x$ y es la raíz cuadrada de la varianza:

$$S_x = \sqrt{\frac{\sum\left(X_j - \bar{X}\right)^2}{N}}$$

En el ejemplo anterior la desviación típica será $\sqrt{39.36}$ = 6,27 años, que al venir expresada en las mismas unidades que la variables resulta más fácil de expresar.

**Características de la varianza y la desviación típica**

- Siempre toma valores positivos
- Si los datos de una distribución son iguales entre sí los valores de varianza y desviación típica serán cero.
- Son índices muy sensibles a la variación de cualquier puntuación de la variable. Basta que varíe una puntuación para que varíen la varianza y la desviación típica.
- Sólo se utiliza para variables cuantitativas
- No se recomienda su cálculo cuando tampoco se recomienda el de la media.
- De la observación de fórmula se deduce fácilmente que cuando los datos se alejan mucho de la media (muy dispersos) el numerador de la fórmula tendrá un valor muy grande y por tanto una varianza y desviación típica grande.

**El coeficiente de variación**

Es un índice muy utilizado cuando pretendemos comparar la variabilidad de dos o más grupos en estudio. Se representa por CV y es igual a la desviación típica dividida por la media.

$$\mathrm{CV} = \frac{S_x}{\bar{X}}$$

Se puede utilizar tanto para comparar el comportamiento de la misma variable en dos grupos distintos, por ejemplo el valor de glucemia en un grupo de niños pequeños y en uno de adultos, como para comparar el comportamiento de dos variables distintas en un mismo grupo. Por ejemplo la altura y el valor de la presión arterial. Veámoslo en el siguiente ejemplo: Imaginemos que hemos realizado un estudio cuyas dos variables principales han sido la edad y el nivel de glucemia y hemos obtenido los siguientes datos:

Edad ( X)
Media 69,6 años
Desviación típica: 10,44 años

Nivel de glucemia ( Y)
Media 97 mg
Desviación típica: 10,44 mg.

Deseamos comparar la variabilidad de ambas variables y no podemos hacerlo a través del análisis de sus desviaciones típicas, ya que los años nada tienen que ver con los miligramos. Por ello para poder comparar la variabilidad de ambas variables utilizamos el coeficiente de variación.

En nuestro ejemplo: $\mathrm{CV} = \frac{S_x}{\bar{X}}$ $CVx = \frac{10,44}{69,6} = 0,15$ $CVy = \frac{21,30}{166} = 0,128$

Como el CV de la variable X es mayor que el CV de la variable Y podemos decir que la variable X, la edad, presenta mayor dispersión que la variable Y, el nivel de glucemia.

**Bibliografía**


- Carrasc JL. El método estadístico en la investigación médica. 6ª Edición. Editorial Ciencia 3; 1995
- Rodríguez Miñón P. Estadística Aplicada a la Biología. 3ª Edición. Editorial UNED; 1984.
- Polit Denise y Hungler Bernadette. Investigación científica en ciencias de la salud. 6ª edición. Edit McGraw-Hill Interamericana; 2000.
- Fernando Villar et al. Diseño y análisis Epidemiológico. Revista Rol de Enfermería. 1987. 112: 13-17.