MEDwave

Estadística Aplicada a la Investigación en Salud
Medwave. Año XI, No. 3, Marzo 2011. Open Access, Creative Commons.

Medidas de tendencia central y dispersión

Autor: Fernando Quevedo Ricardi[(1)]
Filiación:
[(1)]Departamento de Educación en Ciencias de la Salud, Facultad de Medicina, Universidad de Chile
Correspondencia: fquevedo@med.uchile.cl
doi: 10.5867/medwave.2011.03.4934

Ficha del Artículo
Citación: Quevedo F. Medidas de tendencia central y dispersión. *Medwave* 2011 Mar;11(3). doi: 10.5867/medwave.2011.03.4934
Fecha de envío: 5/1/2011
Fecha de aceptación: 13/1/2011
Fecha de publicación: 2/3/2011
Origen: solicitado
Tipo de revisión: sin revisión por pares

Resumen

En la sección Series, Medwave publica artículos relacionados con el desarrollo y discusión de herramientas metodológicas para la investigación clínica, la gestión en salud, la gesión de la calidad y otros temas de interés. En esta edición se presentan dos artículos que forman parte del programa de formación en Medicina Basada en Evidencias que se dicta por e-Campus de Medwave. El artículo siguiente pertenece a la Serie **"Estadística Aplicada a la Investigación en Salud"**.

Las medidas de tendencia central son medidas estadísticas que pretenden resumir en un solo valor a un conjunto de valores. Representan un centro en torno al cual se encuentra ubicado el conjunto de los datos. Las medidas de tendencia central más utilizadas son: **media**, **mediana** y **moda**. Las medidas de dispersión en cambio miden el grado de dispersión de los valores de la variable. Dicho en otros términos las medidas de dispersión pretenden evaluar en qué medida los datos difieren entre sí. De esta forma, ambos tipos de medidas usadas en conjunto permiten describir un conjunto de datos entregando información acerca de su posición y su dispersión.

Los procedimientos para obtener las medidas estadísticas difieren levemente dependiendo de la forma en que se encuentren los datos. Si los datos se encuentran ordenados en una tabla estadística diremos que se encuentran "agrupados" y si los datos no están en una tabla hablaremos de datos "no agrupados".

Según este criterio, haremos primero el estudio de las medidas estadísticas para datos no agrupados y luego para datos agrupados.

Medidas estadísticas en datos *no agrupado*

Medidas de tendencia central

Promedio o media

La medida de tendencia central más conocida y utilizada es la media aritmética o promedio aritmético. Se representa por la letra griega μ cuando se trata del promedio del universo o población y por Ȳ (léase Y barra) cuando se trata del promedio de la muestra. Es importante destacar que μ es una cantidad fija mientras que el promedio de la muestra es variable puesto que diferentes muestras extraídas de la misma población tienden a tener diferentes medias. La media se expresa en la misma unidad que los datos originales: centímetros, horas, gramos, etc.

Si una muestra tiene cuatro observaciones: 3, 5, 2 y 2, por definición el estadígrafo será:

$$\bar{Y} = \frac{3+5+2+2}{4} = \frac{12}{4} = 3$$

Estos cálculos se pueden simbolizar:

$$\bar{Y} = \frac{Y_1 + Y_2 + Y_3 + Y_4}{4}$$

Donde Y1 es el valor de la variable en la primera observación, Y2 es el valor de la segunda observación y así sucesivamente. En general, con "n" observaciones, Yi representa el valor de la i-ésima observación. En este caso el promedio está dado por

$$\bar{Y} = \frac{Y_1 + Y_2 + Y_3 + \ldots\ldots\ldots + Yi + \ldots\ldots\ldots + Y_n}{n}$$

De aquí se desprende la fórmula definitiva del promedio:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

Desviaciones: Se define como la desviación de un dato a la diferencia entre el valor del dato y la media:

$$Desviación = (Y_i - \bar{Y})$$

Ejemplo de desviaciones:

Y_i	$\bar{Y}$	$(Y_i - \bar{Y})$
3	3	0
5	3	+2
2	3	-1
2	3	-1
Suma		0

Una propiedad interesante de la media aritmética es que la suma de las desviaciones es cero.

Mediana
Otra medida de tendencia central es la mediana. La mediana es el valor de la variable que ocupa la posición central, cuando los datos se disponen en orden de magnitud. Es decir, el 50% de las observaciones tiene valores iguales o inferiores a la mediana y el otro 50% tiene valores iguales o superiores a la mediana.

Si el número de observaciones es par, la mediana corresponde al promedio de los dos valores centrales. Por ejemplo, en la muestra 3, 9, 11, 15, la mediana es (9+11)/2=10.

Moda
La moda de una distribución se define como el valor de la variable que más se repite. En un polígono de frecuencia la moda corresponde al valor de la variable que está bajo el punto más alto del gráfico. Una muestra puede tener más de una moda.

Medidas de dispersión

Las medidas de dispersión entregan información sobre la variación de la variable. Pretenden resumir en un solo valor la dispersión que tiene un conjunto de datos. Las medidas de dispersión más utilizadas son: Rango de variación, Varianza, Desviación estándar, Coeficiente de variación.

Rango de variación
Se define como la diferencia entre el mayor valor de la variable y el menor valor de la variable.

$$\text{Rango de variación} = \text{Máximo} - \text{Mínimo}$$

La mejor medida de dispersión, y la más generalizada es la varianza, o su raíz cuadrada, la desviación estándar. La varianza se representa con el símbolo σ^2 (sigma cuadrado) para el universo o población y con el símbolo s2 (s cuadrado), cuando se trata de la muestra. La desviación estándar, que es la raíz cuadrada de la varianza, se representa por σ (sigma) cuando pertenece al universo o población y por "s", cuando pertenece a la muestra. σ^2 y σ son parámetros, constantes para una población particular; s2 y s son estadígrafos, valores que cambian de muestra en muestra dentro de una misma población. La varianza se expresa en unidades de variable al cuadrado y la desviación estándar simplemente en unidades de variable.

Fórmulas
Donde μ es el promedio de la población.

$$\sigma^2 = \frac{(Y_1 - \mu)^2 + (Y_2 - \mu)^2 - \ldots\ldots + (Y_N - \mu)^2}{N}$$

$$\sigma^2 = \frac{\sum_i (Y_i - \mu)^2}{N}$$

Donde Ȳ es el promedio de la muestra.

$$s^2 = \frac{(Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 + \ldots\ldots + (Y_N - \bar{Y})^2}{n \quad 1}$$

$$s^2 = \frac{\sum_i (Y_i - \bar{Y})^2}{n - 1}$$

Consideremos a modo de ejemplo una muestra de 4 observaciones.

Según la fórmula el promedio calculado es 7, veamos ahora el cálculo de las medidas de dispersión:

Y_i	$(Y_i - \bar{Y})$	$(Y_i \quad \bar{Y})^2$
3	-4	16
6	-1	1
8	+1	1
11	+4	16
		34

s2 = 34 / 3 = 11,33 Varianza de la muestra

La desviación estándar de la muestra (s) será la raíz cuadrada de 11,33 = 3,4.

Interpretación de la varianza (válida también para la desviación estándar): un alto valor de la varianza indica que los datos están alejados del promedio. Es difícil hacer una interpretación de la varianza teniendo un solo valor

de ella. La situación es más clara si se comparan las varianzas de dos muestras, por ejemplo varianza de la muestra igual 18 y varianza de la muestra b igual 25. En este caso diremos que los datos de la muestra b tienen mayor dispersión que los datos de la muestra a. esto significa que en la muestra a los datos están más cerca del promedio y en cambio en la muestra b los datos están más alejados del promedio.

Coeficiente de variación

Es una medida de la dispersión relativa de los datos. Se define como la desviación estándar de la muestra expresada como porcentaje de la media muestral.

$$CV = \frac{s \otimes 100}{\bar{y}}$$

Es de particular utilidad para comparar la dispersión entre variables con distintas unidades de medida. Esto porque el coeficiente de variación, a diferencia de la desviación estándar, es independiente de la unidad de medida de la variable de estudio.

Medidas de tendencia central y de dispersión en datos agrupados

Se identifica como datos agrupados a los datos dispuestos en una distribución de frecuencia. En tal caso las fórmulas para el cálculo de promedio, mediana, modo, varianza y desviación estándar deben incluir una leve modificación. A continuación se entregan los detalles para cada una de las medidas.

Promedio en datos agrupados

La fórmula es la siguiente:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i n_i}{n}$$

Donde *ni* representa cada una de las frecuencias correspondientes a los diferentes valores de *Yi*.

Consideremos como ejemplo una distribución de frecuencia de madres que asisten a un programa de lactancia materna, clasificadas según el número de partos. Por tratarse de una variable en escala discreta, las clases o categorías asumen sólo ciertos valores: 1, 2, 3, 4, 5.

Yi- nº de partos	ni	Yi ni	Ni (Frec acumulada)
1	4	4	4
2	13	26	17
3	16	48	33
4	6	24	39
5	3	15	42
Total	42	117	

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i n_i}{n} = \frac{117}{42} = 2{,}78$$

Entonces las 42 madres han tenido, en promedio, 2,78 partos.

Si la variable de interés es de tipo continuo será necesario determinar, para cada intervalo, un valor medio que lo represente. Este valor se llama marca de clase (Yc) y se calcula dividiendo por 2 la suma de los límites reales del intervalo de clase. De ahí en adelante se procede del mismo modo que en el ejercicio anterior, reemplazando, en la formula de promedio, Yi por Yc.

Mediana en datos agrupados

Si la variable es de tipo discreto la mediana será el valor de la variable que corresponda a la frecuencia acumulada que supere inmediatamente a n/2. En los datos de la tabla 1 Me=3, ya que 42/2 es igual a 21 y la frecuencia acumulada que supera inmediatamente a 21 es 33, que corresponde a un valor de variable (Yi) igual a 3.

Si la variable es de tipo continuo es necesario, primero, identificar la frecuencia acumulada que supere en forma inmediata a n/2, y luego aplicar la siguiente fórmula:

$$Me = LI + \frac{\left[\frac{n}{2} - N_{i-1}\right]}{n_i} A_i$$

Donde:

Li	=	Límite inferior del intervalo de clase que contiene a la mediana.
n	=	Tamaño de la muestra.
N_{i-1}	=	Frecuencia acumulada del intervalo anterior.
A_i	=	Amplitud del intervalo (diferencia entre los límites).

Moda en datos agrupados

Si la variable es de tipo discreto la moda o modo será al valor de la variable (Yi) que tenga la mayor frecuencia absoluta ( ). En los datos de la tabla 1 el valor de la moda es 3 ya que este valor de variable corresponde a la mayor frecuencia absoluta =16.

Más adelante se presenta un ejemplo integrado para promedio, mediana, varianza y desviación estándar en datos agrupados con intervalos.

Varianza en datos agrupados

Para el cálculo de varianza en datos agrupados se utiliza la fórmula

$$s^2 = \frac{\sum_i (y_i - \bar{y})^2 n_i}{n-1}$$

Con los datos del ejemplo y recordando que el promedio (Y) resultó ser 2,78 partos por madre,

Y_i	n_i	$Y_i n_i$	$(Y_i \ \bar{Y})^2$	$(y_i - \bar{y})^2 n_i$
1	4	4	3,1684	12,67
2	13	26	0,6084	7,9
3	16	48	0,0484	0,7744
4	6	24	1,4884	8,93
5	3	15	4,9284	14,7852
Total	42	117		45,06

$$s^2 = \frac{\sum_i (y_i - \bar{y})^2 n_i}{n-1} = 45,06 \ / \ 42\text{-}1 = 45,06 \ /41 = 1,1$$

Cuando los datos están agrupados en intervalos de clase, se trabaja con la marca de clase (Yc), de tal modo que la fórmula queda:

$$s^2 = \frac{\sum_i (y_c - \bar{y})^2 n_i}{n-1}$$

Donde Yc es el punto medio del intervalo y se llama marca de clase del intervalo:

Yc=(Límite inferior del intervalo + límite superior del intervalo)/2.

Percentiles

Los percentiles son valores de la variable que dividen la distribución en 100 partes iguales. De este modo si el percentil 80 (P80) es igual a 35 años de edad, significa que el 80% de los casos tiene edad igual o inferior a 35 años.

Su procedimiento de cálculo es relativamente simple en datos agrupados sin intervalos.

Retomemos el ejemplo de la variable número de partos:

Yi- nº de partos	ni	Ni
1	4	4
2	13	17
3	16	33
4	6	39
5	3	42
Total	42	

El percentil j (Pj) corresponde al valor de la variable (Yi) cuya frecuencia acumulada supera inmediatamente al "j" % de los casos (jxn/100).

El percentil 80, en los datos de la tabla, será el valor de la variable cuyo Ni sea inmediatamente superior a 33,6 ((80x42) /100).

El primer Ni que supera a 33,6 es 39. Por lo tanto al percentil 80 le corresponde el valor 4. Se dice entonces que el percentil 80 es 4 partos (P80=4). Este resultado significa que un 80% de las madres estudiadas han tenido 4 partos o menos.

Si los datos están agrupados en una tabla con intervalos, el procedimiento es levemente más complejo ya que se hace necesaria la aplicación de una fórmula.

$$P_j = LI + \frac{\left[\frac{jn}{100} - N_{i-1}\right]}{n_i} A_i$$

Se aplica a los datos del intervalo cuya frecuencia acumulada (Ni) sea inmediatamente superior al "j" % de los casos (jxn/100).

En la siguiente tabla se muestra la distribución de 40 familias según su ingreso mensual en miles de pesos. Nótese que para calcular el centro de clase se usaron los límites reales de cada intervalo.

Ingreso Familiar (miles)	n_i	Y_c	$y_c \times n_i$	$(y_c - \bar{y})$	$(y_c - \bar{y})^2$	$(y_c - \bar{y})^2 \times n_i$	A_i	N_i
80 - 99	8	90	720	-49,13	2.413,76	19.310,08	20	8
100 - 119	10	110	1100	-29,13	848,56	8.485,60	20	18
120 - 159	11	140	1540	0,87	0,76	8,36	40	29
160 - 199	6	180	1080	40,87	1.670,36	10.022,16	40	35
200 - 249	5	225	1125	85,87	7.373,66	36.868,30	50	40
Total	40		5565			74.694,50		

1. El ingreso mensual promedio será:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_c n_i}{n} = \frac{5565}{40} = 139,13$$

2. La mediana será:

$$Me = LI + \frac{\left[\frac{n}{2} - N_{i-1}\right]}{n_i} A_i = 120 + \frac{\left[\frac{40}{2} - 18\right]}{11} \times 40 = 120 + 7,27 = 127,27$$

Esto significa que un 50% de las familias tiene ingreso mensual igual o inferior a $127.270.

3. El percentil 78 será:

$$P_{78} = LI + \frac{\left[\frac{78n}{100} - N_{i-1}\right]}{n_i} A_i = 160 + \frac{[31,2 - 29]}{6} 40 = 160 + 14,66 = 174,66$$

Por lo tanto se puede decir que 78% de las familias tienen ingreso igual o inferior a $174.660.

4. Los percentiles 10 y 90 serán:

$$P10 = LI + \frac{\left[\frac{10n}{100} - N_{i-1}\right]}{n_i} A_i = 80 + \frac{[4-0]}{8} 20 = 80 + 10 = 90$$

$$P90 = LI + \frac{\left[\frac{90n}{100} - N_{i-1}\right]}{n_i} A_i = 200 + \frac{[36-35]}{5} 50 = 200 + 10 = 210$$

A base de los valores de los percentiles 10 y 90 se pueden hacer tres afirmaciones:

- El 10% de las familias tiene ingreso igual o inferior a $90.000.
- El 90% de las familias tiene ingreso igual o inferior a $210.000.
- El 80% central, de las familias, tiene ingreso entre $90.000 y $210.000

5. - La varianza será:

$$s^2 = \frac{\sum_i (y_i - \bar{y})^2 n_i}{n-1} = 74.694,5 \,/\, 39 = 1915,24$$

6. La desviación estándar es la raíz cuadrada de esta cifra, es decir: 43,76.