Desarrollo de contenido

IU Digital de Antioquia

INSTITUCIÓN UNIVERSITARIA DIGITAL DE ANTIOQUIA

Unidad 2

Estadística

ORD No 74 de 2017 · VIGILADA MinEducación

Estadística

Unidad 2. Medidas de tendencia central, dispersión, correlación y regresión

En esta unidad, se presentan los parámetros estadísticos que tienden al centro y se muestra su importancia en la descripción de un conjunto de datos. Estos tipos de medidas son: media aritmética, mediana y moda. Además, el cálculo de la variación del conjunto de datos con respecto a las medidas encontradas que tienden al centro, las cuales permiten mostrar el grado de homogeneidad de dichos datos a analizar. Finalmente, se establece la posible relación entre las variables y las predicciones, al encontrar una función matemática, como la recta de regresión lineal.

Para empezar, es clave entender que las medidas de tendencia central permiten identificar el parámetro estadístico que conduce al centro. Así que medir la variación de

los promedios en las medidas de tendencia central, ayuda a reconocer qué tan dispersos están los datos unos de otros, y en cuanto a los promedios.

Por otra parte, la regresión y la correlación, facilitan determinar la relación entre dos o más variables, y sus causas y efectos, cuantificando el grado de relación.

Tema 1: Medidas de tendencia central

Las medidas de tendencia central como su nombre lo indica, permiten identificar el parámetro estadístico que tiende al centro. Son valores de la variable alrededor de los cuales se agrupa una gran cantidad de valores; tenemos tres tipos de medidas de tendencia central: la media aritmética, la mediana y la moda.

Media aritmética

La media aritmética es la medida más conocida como “promedio” y también es llamada simplemente media; esta se escribe como x. Está en un conjunto n de datos, y es la suma de los datos divididos por la misma cantidad de datos que aparecen en el conjunto (n).

Datos sin agrupar	Datos agrupados
$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i \times f_i}{n}$

x_i para datos cuantitativos discretos hace referencia a la variable y cuando son cuantitativos continuos hace referencia a la marca de clase.

Ejemplo:

LIM INFERIOR	LIM SUPERIOR	EDAD	MARCA DE CLASE Xi	f_i	$x_i * f_i$
8	11	[8, 11)	9,5	12	9,5*12=114
11	14	[11, 14)	12,5	7	12,5*7=87,5
14	17	[14, 17)	15,5	2	15,5*2=31
17	20	[17, 20)	18,5	2	18,5*2=37
20	23	[20, 23)	21,5	4	21,5*4=86
23	26	[23, 26)	24,5	3	24,5*3=73,5
				n=30	429

Mediana

La mediana en un conjunto de n datos, es el valor que indica la mitad, donde el 50% de las observaciones están por encima o por debajo del dato. La mediana se denota como $\boldsymbol{Me}$.

Primero, se calcula la posición de la mediana con $\boldsymbol{X\_Me=(n+1)/2}$; si el resultado es un número entero, esa es la posición en donde se encuentra la mediana; de lo contrario, (cuando el resultado es un número decimal) se suman los números que se encuentran en las dos posiciones entre el número encontrado, y el resultado de la suma se divide en dos.
Para calcular la mediana con datos agrupados,hay que seguir dos pasos:

- Encontrar el intervalo en el que se encuentra la mediana usando la fórmula:

$$posición = \frac{n}{2},\ si\ n\ es\ par$$

$$posición = \frac{n+1}{2},\ si\ n\ es\ impar$$

- Usar la fórmula de la mediana:

$$M_e = L_i + \frac{\frac{n}{2} - F_{i-1}}{f_i} . A_i$$

Donde:

- L_i: límite inferior de intervalo en el cual se encuentra la mediana.
- n: número de datos del estudio. Es la sumatoria de las frecuencias absolutas.
- F_{i-1}: frecuencia acumulada de intervalo al que se encuentra la mediana.
- A_i: amplitud del intervalo en el que se encuentra la mediana.
- f_i: frecuencia absoluta del intervalo en el que se encuentra la mediana.

Ejemplo 1:

Para encontrar la posición de la Tabla 2. Cálculo de la mediana variables cuantitativas discretas, se usa la fórmula:

$$X_{Me} = \frac{40+1}{2} = 20,5$$

Por tanto, la posición de la mediana se encuentra entre X20 y X21, véase la frecuencia absoluta acumulada.

Cantidad de hermanos	*INCLU*	*INCLUD*	
1	6	6	Hasta acá hay 6 datos con 1 hermano
2	10	16	Entre la posición 7 y 16 hay 2 hermanos
3	8	24	Entre la posición 17 y 24 hay 3 hermanos
6	5	29	Entre la posición 25 y 29 hay 6 hermanos
7	6	35	Entre la posición 30 y 35 hay 7 hermanos
8	5	40	Entre la posición 36 y 40 se encuentran 8 hermanos
	n = 40		

Ejemplo 2:

La mediana de este conjunto de datos es:

LIM INFERIOR	LIM SUPERIOR	EDAD	MARCA DE CLASE Xi	f i	F i
8	11	[8, 11)	9,5	12	12
11	14	[11, 14)	12,5	7	19
14	17	[14, 17)	15,5	2	21
17	20	[17, 20)	18,5	2	23
20	23	[20, 23)	21,5	4	27
23	26	[23, 26)	24,5	3	30
				30	

Primero se encuentra la posición:

$$\frac{n}{2} = \frac{30}{2} = 25$$

Se ubica la posición 15 en la segunda fila. La fila se denota como j, la anterior fila como j-1.

Tema 2: Medidas de dispersión

Medir la variación respecto a los promedios encontrados en las medidas de tendencia central, es un cálculo importante en el tratamiento estadístico de datos pues permite identificar qué tan dispersos están los datos unos de otros y con respecto a los promedios. Para medir el grado de dispersión de una variable, se utilizan, principalmente, los indicadores a continuación.

Medidas de variabilidad absoluta

- Esta variable permite comparar distribuciones con la misma medida
- Rango
- Varianza
- Desviación típica estándar

Medidas de variabilidad relativa

En este apartado, estudiaremos la desviación típica y el coeficiente de variación.

- Desviación típica:

La varianza denotada como $INCLUDEPICTURE\ "https://iudigital.instructure.com/equation\_images/s\%255E2"\ \backslash *\ MERG$ da origen a otra mucho más significativa: la desviación típica o estandarizada. La varianza es la suma de las diferencias alrededor de la media, elevadas al cuadrado, dividida entre el tamaño de la muestra.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times f_i}{n}$$

Las unidades de la varianza son los cuadrados de las unidades de los datos: pesos cuadrados, alumnos cuadrados, etc. y, por lo tanto, son medidas difíciles de interpretar. Por ello, la varianza da origen a la desviación típica o estándar (simbolizado como s), obteniéndose extrayendo la raíz cuadrada de la varianza, tomando siempre el valor positivo. Esta es la medida de dispersión más conocida y más utilizada en el análisis de datos estadísticos.

LIM INFERIOR	LIM SUPERIOR	EDAD	MARCA DE CLASE Xi	fi	xi*fi	$(x_i - \bar{x})^2 \times f_i$
8	11	[8,11)	9,5	12	114	(9,5 - 14,3)² X 12 = 276,48
11	14	[11,14)	12,5	7	87,5	(12,5 - 14,3)² X 7 = 22,68
14	17	[14,17)	15,5	2	31	(15,5 - 14,3)² X 2 = 2,88
17	20	[17,20)	18,5	2	37	(18,5 - 14,3)² X 2 = 35,28
20	23	[20,23)	21,5	4	86	(21,5 - 14,3)² X 4 = 207,36
23	26	[23,26)	24,5	3	73,5	(24,5 - 14,3)² X 3 = 312,12
				n=30	429	856,8

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - x)^2 \times f_i}{n} = \frac{856.8}{30} = 28.56$$

Y la desviación típica entonces es la raíz de la varianza, es decir:

$$s = \sqrt{28.56 = 5.344}$$

- Coeficiente de variación:

Esta medida arroja el porcentaje de variación y se usa también cuando se desea comparar la variabilidad de dos o más distribuciones con diferentes unidades de medidas. El coeficiente de variación hace comparable el grado de dispersión entre dos o más variables, y se define como:

$$CV = \frac{s}{\bar{x}} x 100\%$$

s: desviación típica (o estándar)
x ¯: media

Por lo tanto, el coeficiente de variación de la distribución de las edades es igual a:

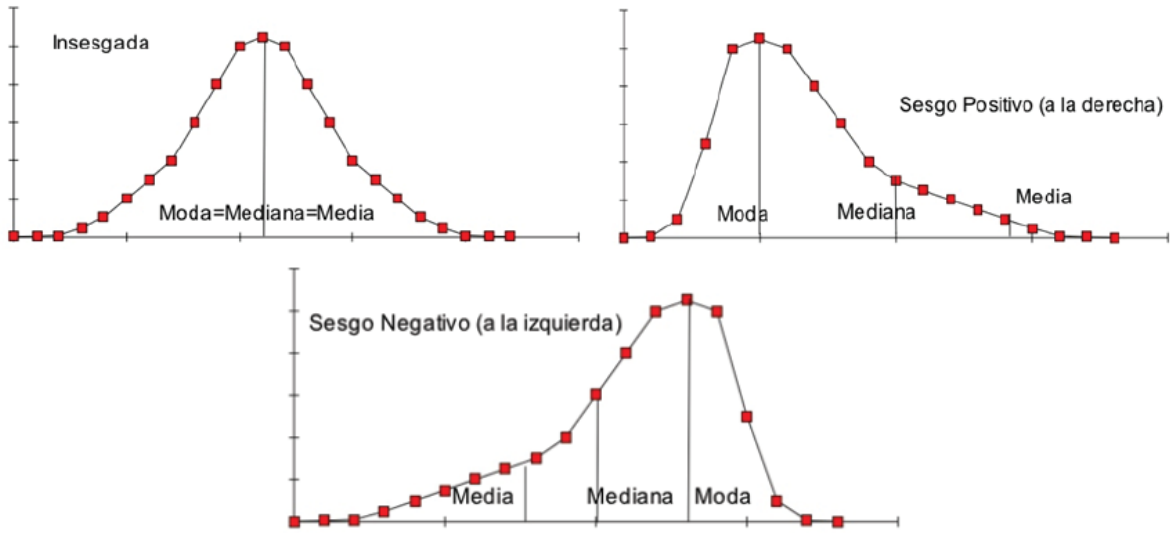
$$cv = 5.344/14.3 \times 100\% = 37.37\%$$

Cuando el coeficiente de variación es muy alto, se dice que la media aritmética no es lo suficientemente representativa en la distribución.

- Asimetría:

La simetría de una distribución indica homogeneidad en los datos, que los valores de la media aritmética, mediana y moda son iguales, y que la distribución tiene la forma de una campana de Gauss o normal.

Cuando las distribuciones son consideras altas, la distribución se convierte en asimétrica ya sea positiva o negativa.



Las fórmulas para calcular el grado de asimetría son:

$$A_s = \frac{\bar{x} - Mo}{s}$$

$$A_s = \frac{m_3}{s^3},\ donde\ m_3 = \frac{\sum (x_i - \bar{x})^3}{n};\ para\ datos\ \text{sin}\ agrupar$$

$$A_s = \frac{m_3}{s^3},\ donde\ m_3 = \frac{\sum (x_i - \bar{x})^3 * f_i}{n};$$

para datos agrupados, f_i, frecuencia absoluta

Su interpretación es:

Si A_s<0, entonces es una distribución negativa, es decir que es dispersa a la izquierda.

Si A_s=0, entonces es una distribución simétrica.

Si A_s>0, entonces es una distribución positiva, es decir que es dispersa a la derecha.

- Curtosis:

Esta medida es conocida como medida de apuntamiento y permite identificar el grado de variación en la cima de las distribuciones, con respecto a la moda comparada con la distribución normal o campana de Gauss.

La fórmula para hallarla es:

$$A_p = \frac{m_4}{s^4} - 3, \; con$$

$$m_4 = \frac{\sum (x_i - \bar{x})^4}{n}; \; para \; datos \; \text{sin} \; agrupar$$

$$m_4 = \frac{\sum (x_i - \bar{x})^4 * f_i}{n}; \; para \; datos \; agrupados, \; f_i, \; frecuencia \; absoluta$$

Su interpretación es:

Si A_p<0, entonces, es una distribución platicúrtica o achatada.

Si A_p=0, entonces, es una distribución normal o mesocúrtica.

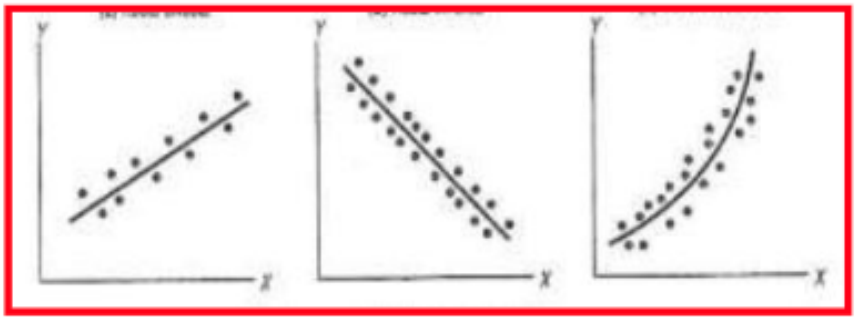
Si A_p>0, entonces, es una distribución leptocúrtica o apuntada.

Tema 3: Correlación y regresión

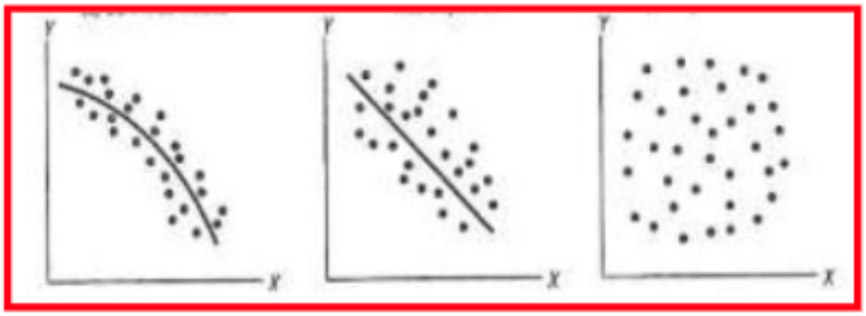
La regresión y la correlación permiten determinar la relación que existe entre dos o más variables, causas y efectos cuantificando el grado de relación, como por ejemplo: entre los salarios y el rendimiento, consumo de alcohol y los años, salario y horas de trabajo, ingresos y gastos, entre otros.

1.Diagrama de dispersión

Una relación entre la variable X (independiente) y la variable Y (dependiente),se representa en el plano cartesiano.Con el diagrama de punto de las observaciones de una distribución,se puedenidentificar las distintas formas que es posibleadoptar desde funciones matemáticas sencillas hasta unas complejas. La relación más sencilla es la lineal y es la que se presentará a continuación.



IM_IUD_DesSof_Est_CR_01



IM_IUD_DesSof_Est_CR_02

Ilustración 1. Diagramas de dispersión. Adecuado a partir de Moreno (2015, p.84).

2.Correlación

La correlación es una forma de medir el grado de dispersión entre dos variables; por tanto, la correlación es una prueba de hipótesis que debe ser sometida a contraste y el coeficiente de correlación cuantifica la correlación entre las dos variables, cuando esta exista.La relación se medirá a través del coeficiente de correlación de Pearson (rp), donde su fórmula es:

$$r_p = \frac{Cov}{s_x s_y}, donde\ Cov = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \overline{xy}$$

Este coeficiente tiene las siguientes características:

- El coeficiente de correlación varía entre –1 y 1

- El signo indica la dirección de la correlación:

a. Cuando el signo es positivo (+) la correlación es directa o positiva e indica que cuando aumenta xtambién aumenta y,ocuando disminuye xtambién disminuye y.
b. Cuando el signo es negativo (–), la correlación es inversa o negativa e indica que cuando aumenta xdisminuye y,ocuando disminuye xaumenta y.

- La interpretación se realiza a través de la siguiente Tabla que fue tomada y adaptada del libro de Estadística de Ciro Martínez.

Tabla 1. Interpretación coeficiente de correlación de Pearson.

Interpretación	Valores de r +	Valores de r –
Correlación perfecta	r = 1	r = –1
Correlación excelente	$0{,}90 \leq r < 1$	$-1 < r \leq 0{,}90$
Correlación aceptable	$0{,}80 \leq r < 0{,}90$	$-0{,}90 < r \leq -0{,}80$
Correlación regular	$0{,}60 \leq r < 0{,}80$	$-080 < r \leq -0{,}60$
Correlación mínima	$0{,}30 \leq r < 0{,}60$	$-0{,}60 < r \leq -0{,}30$
No hay correlación	$0 \leq r < 0{,}30$	$-0{,}30 < r \leq 0$

a.Regresión

El análisis de regresión,se realizapor medio de una ecuación matemática que estima la relación de las variables; en este caso,es una ecuación lineal de la forma y=mx+b, llamada recta de regresión,en estadística.

En estadística,$\mathbf{m}$ ya no se llama la pendiente de una ecuación (aunque su significado es el mismo), sino coeficiente de regresióny $\mathbf{b}$ que es el corte con el eje,se llama coeficiente de posición.

Existen varias fórmulas para encontrar estos tipos de coeficientes, y se puede utilizar cualquiera de ellas.

Coeficiente de regresión

$$m = \frac{Cov}{s_x^2} \quad (i)$$

$$m = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2} \quad (ii)$$

$$m = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \quad (iii)$$

$$m = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \quad (iv)$$

Coeficiente de Posición

$$b = \frac{\sum y_i - m\sum x_i}{n} \quad (i)$$

$$b = \bar{y} - m\bar{x} \quad (ii)$$

Ejemplo 1:

Edad x	Días de hospitalización y	$x_i y_i$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	3	3	49	1,14
2	3	6	36	1,14
3	4	12	25	0,00
4	4	16	16	0,00
5	5	25	9	0,87
6	7	42	4	8,60
7	5	35	1	0,87
8	3	24	0	1,14
9	2	18	1	4,27
10	6	60	4	3,74
11	6	66	9	3,74
12	3	36	16	1,14
13	5	65	25	0,87
14	3	42	36	1,14
15	2	30	49	4,27
$\sum 120$	$\sum 61$	$\sum 480$	$\sum 280$	$\sum 32{,}93$
$\bar{x} = \frac{120}{15} = 8$	$\bar{x} = \frac{61}{15}$ $\bar{x} = 4{,}07$		$s_x = 4{,}32$	$s_y = 1{,}48$

[1]Tomado de Moreno(2015, p.86).

Para encontrar el coeficiente de correlación de Pearson, se debe encontrar primero la covarianza, sumando la tercera columna y haciendo la diferencia con el producto entre las medias de las dos variables:

$$Cov = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y} = \frac{480}{15} - (8)(4{,}07) = -0{,}53$$

Como la varianza da negativa,se sabe que la relación entre las dos variables es inversa o negativa, y cuando aumentan los años,disminuyenlos días de hospitalización. Ahora,se encuentra el coeficiente de correlación de Pearson a través del nivel de relación con la covarianza y las desviaciones de la variable xy de la variable y:

$$r_p = \frac{Cov}{s_x s_y} = \frac{-0{,}53}{(4{,}32) \times (1{,}48)} = -0{,}08$$

Después de encontrar este resultado, se concluye que no hay relación entre las dos variables (ver la Tabla 2.Cálculo Covarianza);es decir,que los días de hospitalización de los niños depende en un 0,75%2de la edad. La recta de regresión se calcula así:

$$m = \frac{Cov}{{s_x}^2} = \frac{-0{,}53}{(4{,}32)^2} = -0{,}286$$

$$b = \bar{y} - m\bar{x} = 4{,}07 — 0{,}286 \times (8) = 4{,}2952$$

Entonces la recta de regresión de la forma $y = mx + b$, es
$y = -0{,}286x + 4{,}2952$

Y su diagrama de dispersión es: