Ejercicio de estadística para 3º de la ESO

Unibelia

La estadística es una disciplina técnica que se apoya en las matemáticas y que tiene como objetivo la interpretación de la realidad de una población o muestra sobre la que se hace el estudio. La utilidad que ésta tiene es decisiva a la hora de tomar decisiones en una infinidad de asuntos relacionados con las distintas áreas de planificación del estado, empresas y hasta en los hogares. Comprender el significado de las medidas de centralización, dispersión y las gráficas son fundamentales para tener un conocimiento básico y útil que nos ayude a comprender la importancia de su uso y sus limitaciones.

Estadística

Esta clase de matemática está especialmente diseñada para dar apoyo a alumnos de 3º de la ESO como así también a quienes quieren acceder a los ciclos formativos.

El ejercicio consta de varios pasos que deberemos realizar para analizar y sacar conclusiones de los datos obtenidos de las encuestas realizadas a una muestra poblacional (a un subconjunto de la población) sobre una sola variable.

Ejercicio

Sacar conclusiones estadísticas de una variable (edad de madres primerizas) de una muestra poblacional encuestada en un hospital de la isla de Gran Canaria.

Hay que recordar que la muestra debe ser representativa por lo que en nuestro caso deberíamos haber tomado una mayor cantidad de datos (edades de madres primerizas) pero a fin de simplificar el ejemplo, lo haremos con tan solo 30.

Estas son las edades de un grupo de madres:

28, 34, 43, 30, 47, 38, 34, 40, 31, 33

42, 33, 42, 39, 30, 32, 47, 37, 32, 35

41, 35, 37, 33, 39, 34, 32, 43, 40, 38

¿CÓMO EMPEZAMOS?

Preparamos los datos

Antes de comenzar a realizar cualquier análisis, debemos preparar los datos y decidir una serie de cuestiones relacionadas con su agrupamiento.

1º Ordenar los valores de manera ascendente

28 - 30 – 30 - 31 - 32- 32 - 32 - 33 - 33 - 33 – 34 – 34 - 34 – 35 – 35 – 37 - 37 - 38 - 38– 39 - 39 40 - 40 - 41 -42 - 42 - 43 - 43 - 47 - 47

2º Decidir si realizar el análisis en intervalos o no

Lo primero que deberemos pensar es en cómo agruparemos los datos. Tendremos dos posibilidades:

No agruparlos y considerar a cada valor como una observación que acumulará frecuencias (esto se utiliza cuando los datos son pocos y/o **la variable no asume demasiados valores** sobre todo cuando es una **variable discreta**)

Agruparlos y considerar intervalos de valores (esto se utiliza cuando debemos trabajar con un gran volumen de datos y/o la **variable asume una gran cantidad de valores**, sobre todo cuando la **variable es continua**)

No existe una regla fija para determinar en cuántos intervalos dividiremos la muestra pero por lo general podremos utilizar una tabla que relaciona la cantidad de datos con respecto a la cantidad de intervalos "razonables".

Número de Datos	Número de Intervalos
< 20	Sin intervalos
20 - 50	7
50 - 75	10
75 - 100	12

En nuestro caso hemos decidido optar por agrupar los datos porque la cantidad de supera los 30 y existe una gran variedad de valores de la variable.

3º Calcular en Rango

El rango es una medida de dispersión como lo es la **desviación típica** o la **varianza** y se define como la diferencia entre el Valor Máximo y el Valor Mínimo de los datos, es decir, nos indica la amplitud de la variación de fenómeno entre su límite mayor y su límite menor.

= Max( ) – Min( ) (En nuestro caso) 47 – 28 = 19

R= 19

3º Calcular el número de intervalos

Un intervalo es un conjunto de valores que oscilan entre dos límites.

Para calcular el intervalo tomaremos el rango y, en caso de ser conveniente, sumaremos un número de forma tal que pueda ser dividido de forma exacta.

En nuestro caso:

Rango: 19

Debemos recordar que el número de intervalos "razonables" era 7 para una distribución que estaba comprendida entre 20 y 50 valores.

Sumo a nuestro rango un número conveniente = 2

Total 21 y lo divido por 7 (el siete es el número de intervalos razonables que habríamos visto en la tabla y para llegar a él tuve que sumarle 2).

21 : 7 = 3

Al número 3 (el cociente), se le llama **amplitud el intervalo** y significa que en el intervalo sólo podrán entrar 3 números distintos entre los límites.

4º Fabricar los intervalos

Para fabricar el primer intervalo debemos tomar el valor más pequeño de la distribución al que llamaremos primer límite inferior. A partir de ahí se considerará la amplitud de intervalos para formar el primer límite superior.

Cómo dijimos anteriormente, cada intervalo tiene una amplitud de 3, quiere decir que 3 son los valores que podrán entrar dentro del intervalo, en el primer caso serán: 28, 29 y 30.

Para fabricar el segundo intervalo, tomaremos el valor siguiente al límite superior del intervalo anterior y pensaremos nuevamente en la amplitud 3 del intervalo por lo que obtendremos que en el segundo intervalo los valore serán 31, 32 y 33 y así sucesivamente.

L_i	Amplitud	L_s
28	3	30
31	3	33
34	3	36
37	3	39
40	3	42
43	3	45
46	3	48

6º Marca de Clase X_j o C_i

La marca de clase es el punto medio del intervalo

Habiendo encontrado los límites, hallaremos la marca de clase que es el valor que tomaremos para sacar los cálculos de las medidas de centralización y dispersión.

La fórmula es simple

$$X_j o\ C_i = \frac{(L_s - L_i)}{2} + L_i$$

En nuestro ejercicio

$$\frac{(30-28)}{2} + 28 = \frac{2}{2} + 28 = 29$$

Ahora sacamos la marca de clase para cada uno de los intervalos y comenzamos a armar nuestra tabla de distribución de frecuencias.

L_i - L_s	C_i
28 - 30	29
31 - 33	32
34 - 36	35
37 - 39	36
40 - 42	41
43 - 45	44
46 - 48	47

7º Obtenemos la frecuencia de cada intervalo

Para obtener la frecuencia de un intervalo sólo debemos contar cuántos valores se encuentran dentro del intervalo que hemos pensado.

Ej.: Primer intervalo 28 – 30.

Dentro de este intervalo entrarán los valores de la distribución de frecuencias subrayados y en negrita. Así la frecuencia con que se repite este intervalo es 3

28 - 30 – 30 - 31 - 32- 32 - 32 - 33 - 33 - 33 – 34 – 34 - 34 – 35 – 35 – 37 - 37 - 38 - 38– 39 - 39 40 - 40 - 41 -42 - 42 - 43 - 43 - 47 - 47

Ahora calcularemos la frecuencia para cada uno de los intervalos y calcularemos una serie de columnas que luego nos servirán para sacar algunas conclusiones y gráfica:

L_i - L_s	C_i	f_i	$C_i \cdot f_i$	F_i	h_i	H_i	Sector
28 - 30	29	3	87	3	0,10	0,10	36º
31 - 33	32	7	224	10	0,23	0,33	84º
34 - 36	35	5	175	15	0,17	0,50	60º
37 - 39	38	6	228	21	0,20	0,70	72º
40 - 42	41	5	205	26	0,17	0,87	60º
43 - 45	44	2	88	28	0,07	0,93	24º
46 - 48	47	2	94	30	0,07	1,00	24º
Total		30	1101		1,00		360º

C_i= Marca de Clase

f_i = Frecuencia

$F_i = Frecuencia\ acumulada$

h_i =Frecuencia Relativa

H_i =Frecuencia Relativa Acumulada

$Sector = h_i \times 360º$

Calcular las medidas de centralización

Media aritmética

Media: Es el promedio que surge de la muestra. Es el resultado de la suma del producto de cada valor o marcas de clases por su respectiva frecuencia y al resultado dividirlo por el total de observaciones.

$$\bar{x} = \frac{\sum c_i \cdot f_i}{n}$$

c_i= Marca de Clase

f_i = Frecuencia

En nuestro caso:

$$\text{media} = \frac{29 \cdot 3 + 32 \cdot 7 + 35 \cdot 5 + 38 \cdot 6 + 41 \cdot 5 + 44 \cdot 2 + 47 \cdot 2}{30}$$

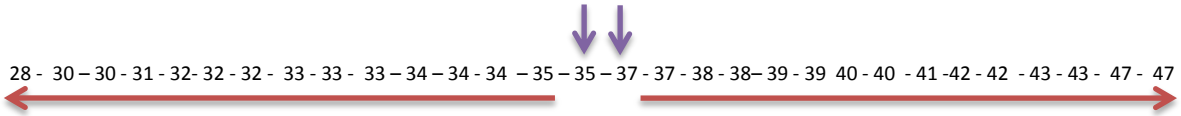
$$\text{media} = \frac{87 + 224 + 175 + 228 + 205 + 88 + 94}{30}$$

$$\text{media} = \frac{1101}{30}$$

$$\text{media} = 36{,}7$$

Mediana

La Mediana representa el valor de la variable que ocupa la posición central en un conjunto de datos ordenados de menor a mayor.



Cuando calculamos la media en una distribución dejamos a ambos lados una misma cantidad de datos, en nuestro caso, 14. La posición 15 y 16 son el número 35 y 37 sobre los cuales aplicaremos un promedio a fin de calcular la mediana.

$$mediana = \frac{35+37}{2}$$

$$mediana = \frac{72}{2} = 36$$

$$mediana = 36$$

Recuerda que si la cantidad de datos de la distribución es impar, habrá solo un número en el medio y no habrá que hacer el promedio.

Moda

La moda es el valor que más se repite en la frecuencia. Cuando tenemos intervalos, será la marca de clase del intervalo que mayor frecuencia tenga, en nuestro caso:

$$Moda = 32$$

Calcular medidas de dispersión

Desviación media

La desviación media es la media aritmética (promedio) de los valores absolutos de las desviaciones (diferencia entre un valor) respecto a la media.

$$D_{\bar{x}} = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + |x_3 - \bar{x}| + \cdots |x_n - \bar{x}|}{N}$$

Resumiendo

$$D_{\bar{x}} = \sum_{i=1}^{n} \frac{|x_i - \bar{x}|}{N}$$

Veamos nuestro caso:

En nuestro caso, como estamos tomando intervalos, deberemos tomar la desviación de la marca de clase c_i con respecto a la media que hemos encontrado y multiplicarlos por su frecuencia.

L_i - L_s	C_i	f_i	$C_i \cdot f_i$	F	h	H	Sector
28 - 30	29	3	87	3	0,10	0,10	36º
31 - 33	32	7	224	10	0,23	0,33	84º
34 - 36	35	5	175	15	0,17	0,50	60º
37 - 39	38	6	228	21	0,20	0,70	72º
40 - 42	41	5	205	26	0,17	0,87	60º
43 - 45	44	2	88	28	0,07	0,93	24º
46 - 48	47	2	94	30	0,07	1,00	24º
Total		30	1101		1,00		360º

$$D_{\bar{x}} = \frac{|x_1 - \bar{x}|f_1 + |x_2 - \bar{x}|f_2 + |x_3 - \bar{x}|f_3 + \cdots |x_n - \bar{x}|f_n}{N}$$

$$D_{\bar{x}} = \frac{|29 - 36{,}7|3 + |32 - 36{,}7|7 + |35 - 36{,}7|5 + |38 - 36{,}7|6 + |41 - 36{,}7|5 + |44 - 36{,}7|2 + |47 - 36{,}7|2}{30}$$

$$D_{\bar{x}} = \frac{|7{,}7|3 + |4{,}7|7 + |1{,}7|5 + |1{,}3|6 + |4{,}3|5 + |7{,}3|2 + |10{,}3|2}{30}$$

$$D_{\bar{x}} = \frac{23{,}1 + 32{,}9 + 8{,}5 + 7{,}8 + 21{,}5 + 14{,}6 + 20{,}6}{30}$$

$$D_{\bar{x}} = \frac{129}{30}$$

$$D_{\bar{x}} = 4{,}3$$

Este valor indica que, en promedio, las variaciones que existen entre los datos de la muestra con respecto a la media da como resultado 4,3. Luego, podremos traspolarlo a futuras muestras. De todas maneras, con el fin de eliminar el problema que surge a raíz de los valores absolutos en otros cálculos algebraicos, se opta por tomar otras dos medidas de dispersión.

Varianza

Si se desea una medida de la dispersión sin los inconvenientes para el cálculo que tiene la desviación media, una solución es elevar al cuadrado las desviaciones antes de calcular el promedio. Así, se define la **varianza** como:

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

En nuestro caso:

$$\sigma^2 = \frac{(29-36{,}7)^2 \cdot 3 + (32-36{,}7)^2 \cdot 7 + (35-36{,}7)^2 \cdot 5 + (38-36{,}7)^2 \cdot 6 + (41-36{,}7)^2 \cdot 5 + (44-36{,}7)^2 \cdot 2 + (47-36{,}7)^2 \cdot 2}{30}$$

$$\sigma^2 = \frac{(7{,}7)^2 3 + (4{,}7)^2 7 + (1{,}7)^2 5 + (1{,}3)^2 6 + (4{,}3)^2 5 + (7{,}3)^2 2 + |(10{,}3)|^2 2}{30}$$

$$\sigma^2 = \frac{177{,}87 + 154{,}63 + 14{,}45 + 10{,}14 + 92{,}45 + 106{,}58 + 212{,}18}{30}$$

$$\sigma^2 = \frac{177{,}87 + 154{,}63 + 14{,}45 + 10{,}14 + 92{,}45 + 106{,}58 + 212{,}18}{30}$$

$$\sigma^2 = \frac{768{,}3}{30}$$

$$\sigma^2 = 25{,}61$$

El inconveniente con la varianza es que los datos que nos proporciona son cuadrados. Es decir, en nuestro caso serían años al cuadrado. Para quitar este inconveniente es que se utiliza otra medida de dispersión que se llama desviación típica.

Desviación típica muestral

Se denomina desviación típica a la raíz cuadrada de la varianza

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

$$\sigma = \sqrt{\frac{(29-36{,}7)^2 3 + (32-36{,}7)^2 7 + (35-36{,}7)^2 5 + (38-36{,}7)^2 6 + (41-36{,}7)^2 5 + (44-36{,}7)^2 2 + (47-36{,}7)^2 2}{30}}$$

$$\sigma = \sqrt{\frac{(7{,}7)^2 3 + (4{,}7)^2 7 + (1{,}7)^2 5 + (1{,}3)^2 6 + (4{,}3)^2 5 + (7{,}3)^2 2 + |(10{,}3)|^2 2}{30}}$$

$$\sigma = \sqrt{\frac{177{,}87 + 154{,}63 + 14{,}45 + 10{,}14 + 92{,}45 + 106{,}58 + 212{,}18}{30}}$$

$$\sigma = \sqrt{\frac{768{,}3}{30}}$$

$$\sigma = \sqrt{25{,}61}$$

$$\sigma = 5{,}06$$

Si observamos este valor, veremos que se acerca a nuestra desviación media. La desviación típica nos dice que los datos tienen un promedio geométrico de desviación con respecto a la media de 5,06 años, es decir, que en promedio los datos se alejan de la media en 5,06 años.

Graficar

Una vez que hemos encontrado nuestras medidas de centralización y dispersión nos queda el último paso, la gráfica.

Existirá una serie de gráficos que serán los más adecuados para cada una de las variables a analizar y los tipos de datos que asuman dichas variables.

Grafico de Sector

En nuestro ejemplo, para determinar a simple vista cual qué rango de edad es el más frecuente entre las embarazadas por primera vez, podemos hacer un gráfico de sectores en el que cada sector es proporcional a la frecuencia absoluta del intervalo.

L_i - L_s	Sector
28 - 30	36º
31 - 33	84º
34 - 36	60º
37 - 39	72º
40 - 42	60º
43 - 45	24º
46 - 48	24º
Total	360º

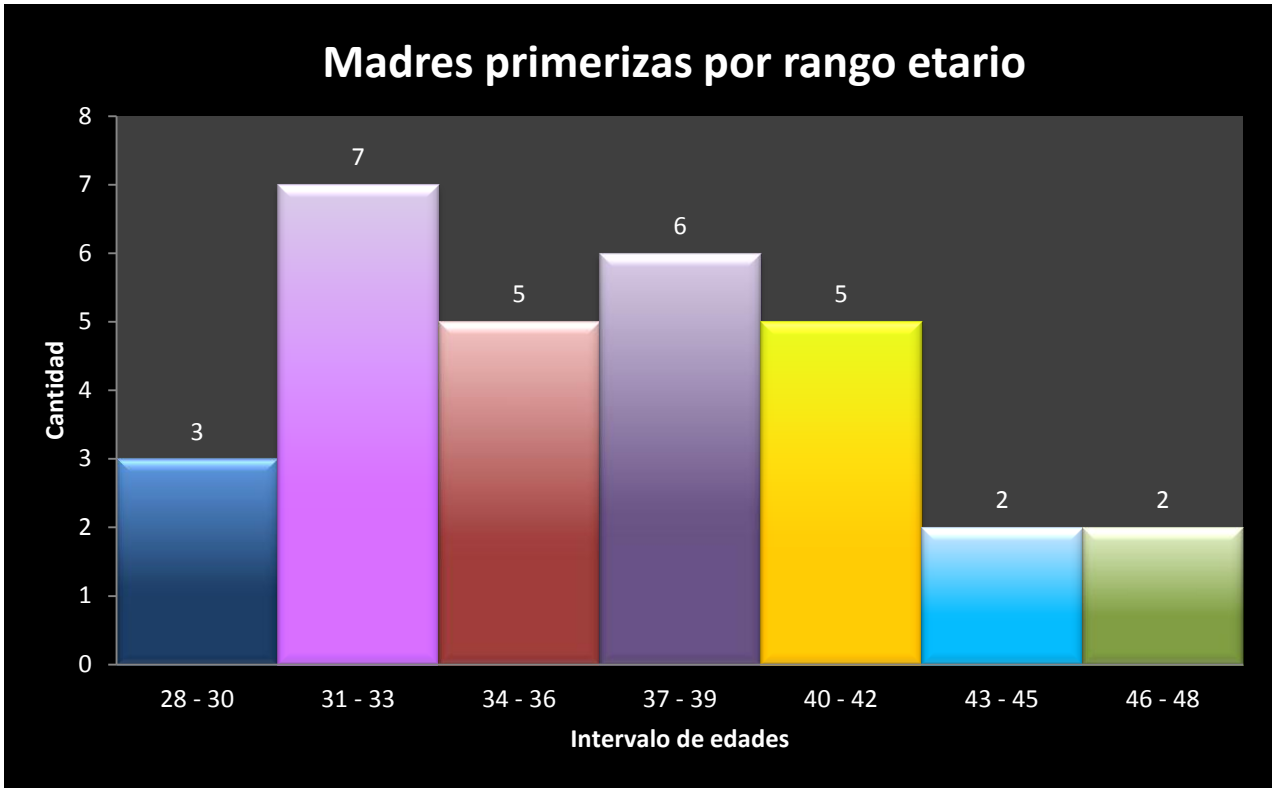
En el gráfico se ve el porcentaje de madres primerizas por grupo de edades. Podemos observar rápidamente cuales son los grupos más numerosos y menos numerosos porcentualmente, (también podríamos haber puesto sus respectivas cantidades). Este tipo de gráfico es muy utilizado ya que es muy fácil de comprender y resume de manera muy práctica la información.

Histograma

En un gráfico de barras aparecerán en el eje inferior los intervalos y en el eje vertical las frecuencias de dichos intervalos. La altura de cada intervalo es proporcional a su frecuencia.

L_i - L_s	C_i	f_i
28 - 30	29	3
31 - 33	32	7
34 - 36	35	5
37 - 39	38	6
40 - 42	41	5
43 - 45	44	2
46 - 48	47	2
Total		30

Para hacer el gráfico tomaremos las frecuencias absolutas.



Como vemos, el gráfico nos permite sacar conclusiones rápidas. El intervalo de edades donde hay un mayor número de madres es de 31-33 años. Los menores, de 43-45 y de 46-48. También nos permite apreciar un grupo de valores dónde más se acumulan las observaciones, de 31 hasta los 42 años y se puede decir que en los extremos se registran las menores frecuencias.

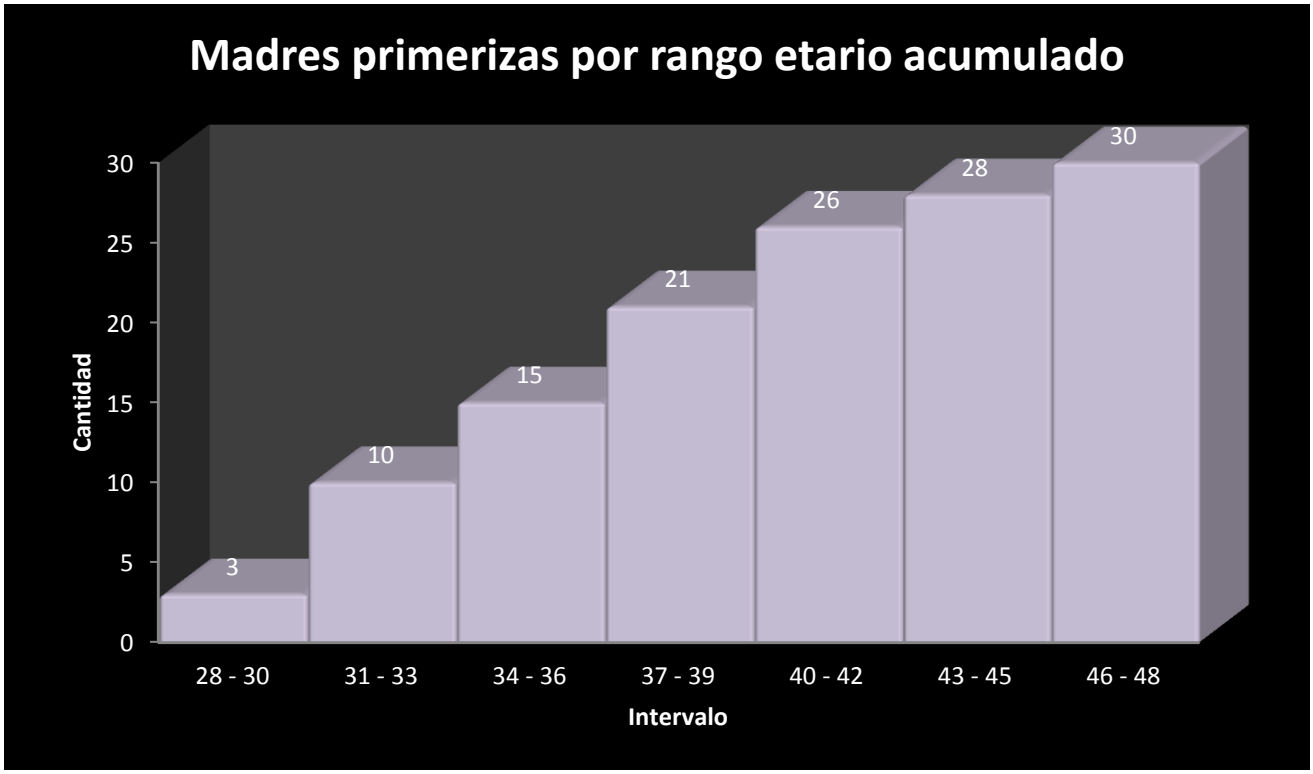
Histograma de Frecuencias Acumuladas

Un histograma de frecuencias acumuladas nos permite ver en un solo pantallazo cómo se van acumulando las frecuencias hasta llegar a la totalidad de las observaciones. Es muy conveniente su uso para gráficas económicas.

Para hacer el gráfico tomaremos las frecuencias acumuladas (F)

L_i - L_s	F
28 - 30	3
31 - 33	10
34 - 36	15
37 - 39	21
40 - 42	26
43 - 45	28
46 - 48	30
Total	

Haremos en nuestro caso también este gráfico para que se vea cómo nos puede ayudar la exposición de información mediante esta herramienta estadística.



El gráfico nos muestra de manera simple y rápida como se van acumulando los datos hasta llegar a la totalidad de la muestra. En nuestro caso podemos decir, por ejemplo, que la mitad de las madres primerizas tienen menos de 36 años o que el 70% (21 de 30) de las madres tienen menos de 40 años.

Conclusión

Más allá de la exactitud con la que debemos realizar cada cálculo estadístico, de nada servirán sin un correcto y objetivo análisis de las variables ya que, como dijimos al principio, la estadística nos ayuda a comprender la realidad que observamos.