

Conceptos básicos de estadística unidimensional

VARIABLE DISCRETA

Caso 1

Supongamos que un profesor tiene anotadas las calificaciones de 30 alumnos de un grupo, obtenidas en la primera evaluación en la materia de matemáticas:

5	3	4	1	2	8	9	8	7	6
6	7	9	8	7	7	1	0	1	6
9	9	8	0	8	8	8	9	5	7

Haciendo un recuento sencillo podríamos construir una tabla de frecuencias como sigue:

x_i	f_i	F_i	h_i	H_i	p_i (%)	P_i (%)
0	2	2	$2/30 = 0,07$	$2/30 = 0,07$	7	7
1	3	5	$3/30 = 0,10$	$5/30 = 0,17$	10	17
2	1	6	$1/30 = 0,03$	$6/30 = 0,20$	3	20
3	1	7	$1/30 = 0,03$	$7/30 = 0,23$	3	23
4	1	8	$1/30 = 0,03$	$8/30 = 0,27$	3	26
5	3	11	$3/30 = 0,10$	$11/30 = 0,37$	10	36
6	2	13	$2/30 = 0,07$	$13/30 = 0,43$	7	43
7	5	18	$5/30 = 0,17$	$18/30 = 0,60$	17	60
8	7	25	$7/30 = 0,23$	$25/30 = 0,83$	23	83
9	5	30	$5/30 = 0,17$	$30/30 = 1$	17	100
10	0	$F_{11} = 30$	0	$H_{11} = 30 / 30 = 1$	0	100
	$\sum_{i=1}^{11} f_i = 30$		$\sum_{i=1}^{11} h_i = 1$		100	

Observa que:

- x_i son los valores que toma la **variable**. En nuestro caso x_1 es la calificación 0, x_2 es la calificación 1, y así sucesivamente hasta x_{11} que es la que corresponde a la calificación 10. Decimos que hay un total de 11 **modalidades** de la variable (las posibles calificaciones enteras, desde 0 hasta 10). La variable es **discreta** porque sólo toma un número finito de valores. Además, es **cuantitativa** porque se representa mediante números. Hay variables estadísticas que no son cuantitativas, sino **cualitativas** porque no se pueden representar mediante números (el color de los ojos o el sexo de un ser humano, la raza de un perro, etcétera).
- f_i es la **frecuencia absoluta** de cada uno de los valores que toma la variable estadística. Así f_1 es 2 porque 2 alumnos obtuvieron la calificación 0, f_2 es 3 porque 3 alumnos obtuvieron la calificación 1, y así sucesivamente. Haz tú el recuento y ya verás (es de lo más tedioso de la estadística).
- F_i es la **frecuencia absoluta acumulada**. Se obtiene así:

$$\begin{aligned}F_1 &= f_1 \\F_2 &= f_1 + f_2 \\F_3 &= f_1 + f_2 + f_3 \\&\dots\dots\dots \\F_k &= f_1 + f_2 + f_3 + \dots + f_k\end{aligned}$$

En nuestro caso:

$$\begin{aligned}F_1 &= f_1 \Rightarrow F_1 = 2 \\F_2 &= f_1 + f_2 \Rightarrow F_2 = 2 + 3 = 5 \\F_3 &= f_1 + f_2 + f_3 \Rightarrow F_3 = 2 + 3 + 1 = 6 \\&\dots\dots\dots \\F_k &= f_1 + f_2 + f_3 + \dots + f_k \Rightarrow F_{11} = f_1 + f_2 + f_3 + \dots + f_{11} = 2 + 3 + 1 + \dots + 0 = 30\end{aligned}$$

En la práctica para obtener un valor de F_i se suman todos los valores de f_i que están a la misma altura o por debajo de la fila correspondiente al valor x_i de la variable.

F_i indica el número de individuos de la población que toman valores de la variable menores o iguales que el valor x_i . Así, por ejemplo, en nuestro caso, la frecuencia absoluta acumulada $F_5 = 8$ indica que hay 8 alumnos que han obtenido un 4 o menos, es decir, que hay 8 alumnos que han suspendido las matemáticas en la primera evaluación.

Para abreviar, muchas veces se utiliza el símbolo de sumatorio (Σ), muy habitual en estadística y probabilidad. De este modo en vez de escribir $F_k = f_1 + f_2 + f_3 + \dots + f_k$, escribiremos $F_k = \sum_{i=1}^k f_i$ (se lee “suma de todas las frecuencias absolutas, desde la primera hasta la k -ésima”)

- h_i es la **frecuencia relativa** correspondiente al valor x_i . Se obtiene mediante la fórmula $h_i = \frac{f_i}{n}$, donde n es el número de individuos objeto de estudio o **tamaño de la población** (en nuestro caso $n = 30$). Así pues, en el ejemplo es $h_1 = \frac{f_1}{n} = \frac{2}{30} = 0,07$, $h_2 = \frac{f_2}{n} = \frac{3}{30} = 0,10$, y así sucesivamente. La frecuencia relativa tiene mucho que ver con p_i , o **porcentaje**, que se calcula multiplicando la frecuencia relativa por 100: $p_i = h_i \cdot 100$. Entonces $p_1 = h_1 \cdot 100 = 0,07 \cdot 100 = 7$, $p_2 = h_2 \cdot 100 = 0,10 \cdot 100 = 10$, y así sucesivamente. El porcentaje indica el tanto por ciento de individuos que presentan una determinada modalidad de la variable. Así $p_8 = 17$ nos muestra que el 17% del alumnado de la clase ha obtenido una calificación igual a un 7.

A veces se cometen errores al tomar valores de la frecuencia relativa y del porcentaje. Lo normal es redondear a dos o tres decimales. En el caso de h_1 en realidad es $h_1 = \frac{f_1}{n} = \frac{2}{30} = 0,06666\dots$, que se ha redondeado a 0,07.

Pero también se podría haber tomado el valor 0,067, y en este caso el porcentaje sería $p_1 = 6,7$. Pero no hay que preocuparse, solamente hay que saber que cuantas más cifras significativas se tomen en el redondeo, menor será el error que se cometa en los resultados. Todo dependerá de lo precisos que queramos ser.

- H_i es la **frecuencia relativa acumulada** correspondiente al valor x_i . Se obtiene mediante la fórmula $H_i = \frac{F_i}{n}$. En nuestro ejemplo es $H_1 = \frac{F_1}{n} = \frac{2}{30} = 0,07$, $H_2 = \frac{F_2}{n} = \frac{5}{30} = 0,10$, y así sucesivamente. La frecuencia relativa tiene mucho que ver con P_i , o **porcentaje acumulado**, que se calcula multiplicando la frecuencia relativa acumulada por 100: $P_i = H_i \cdot 100$. $P_1 = H_1 \cdot 100 = 0,07 \cdot 100 = 7$, $P_2 = H_2 \cdot 100 = 0,10 \cdot 100 = 10$, y así sucesivamente. El porcentaje acumulado P_i indica el porcentaje de individuos de la población que toman valores de la variable menores o iguales que el valor x_i . Así, por ejemplo, en nuestro caso, el porcentaje acumulado $H_5 = 26$ indica que el 26% de los alumnos de la clase han obtenido un 4 o menos, es decir, que el 26% del alumnado ha suspendido las matemáticas en la primera evaluación.

Suponiendo que en el estudio de una variable estadística que presenta r modalidades (en nuestro caso $r = 11$, ya que hay 11 calificaciones posibles), hay algunas consecuencias inmediatas que siempre se cumplen. Son las siguientes:

- La suma de todas las frecuencias absolutas ordinarias ha de coincidir con el número total de individuos de la población, es decir, con el tamaño n : $f_1 + f_2 + f_3 + \dots + f_k = n$, o lo que es lo mismo: $\sum_{i=1}^k f_i = n$. En nuestro caso $\sum_{i=1}^{11} f_i = f_1 + f_2 + f_3 + \dots + f_{11} = 2 + 3 + 1 + \dots + 0 = 30 = n$.
- La suma de todas las frecuencias relativas ordinarias ha de ser 1, ya que representa la suma de todas las proporciones $\sum_{i=1}^k h_i = 1$ (compruébalo tú sumando todas las frecuencias relativas en nuestro ejemplo). Es tanto como decir que la suma de todos los porcentajes p_i ha de ser igual a 100 (lógico y natural).
- La última frecuencia absoluta acumulada ha de ser igual al tamaño n : $F_k = n$. En nuestro caso $F_{11} = 30$.
- La última frecuencia relativa acumulada ha de ser igual a uno: $H_k = 1$. Es tanto como decir que el último porcentaje acumulado P_k ha de ser 100 (también lógico y natural).

VARIABLE CONTINUA

Caso 2

Consideremos un parque automovilístico de 2478 coches clasificados según la distancia recorrida en un año:

kilometraje anual (en miles de kilómetros)	número de vehículos
$(e_{i-1}, e_i]$	f_i
(0, 4]	228
(4, 8]	634
(8, 12]	821
(12, 16]	475
(16, 20]	233
(20, 24]	87
	$n = 2478$

Observa que ahora la variable no es discreta porque la distancia recorrida por un coche en un año puede tomar cualquier valor de un intervalo y no se limita a valores numéricos aislados. En este caso se dice que la variable es **continua** y se suele agrupar en intervalos o **clases**, cuya amplitud (**amplitud de clase**) es la distancia entre los dos **extremos de clase** y suele ser la misma para todos los intervalos. En el ejemplo vemos que hay 6 clases y que la amplitud de cada clase es de 4 unidades. Para hacer determinados cálculos hemos de tomar un valor representativo de cada clase y que tomaremos como valor representativo de la variable. Recibe el nombre de **marca de clase** x_i y es

el punto medio de los extremos de clase o del intervalo: $x_i = \frac{e_{i-1} + e_i}{2}$. En el caso que nos ocupa tenemos que

$x_1 = \frac{e_0 + e_1}{2} = \frac{0 + 4}{2} = 2$, $x_2 = \frac{e_1 + e_2}{2} = \frac{4 + 8}{2} = 6$, y así sucesivamente. En los casos de variable continua es muy común que no nos den la marca de clase y tengamos que calcularla, escribiéndola en una columna justo a la derecha de los intervalos de clase (ver tabla siguiente). Es muy común también que no nos den ni los intervalos, sino solamente los datos y que nos pidan agruparlos en intervalos de una determinada amplitud empezando a partir de un determinado valor. Lo veremos en alguno de los ejercicios.

kilometraje anual (en miles de kilómetros)	marca de clase	número de vehículos
$(e_{i-1}, e_i]$	x_i	f_i
(0, 4]	2	228
(4, 8]	6	634
(8, 12]	10	821
(12, 16]	14	475
(16, 20]	18	233
(20, 24]	22	87
		$n = 2478$

A partir de esta tabla se procede como en el caso de variable discreta y se pueden completar las columnas de frecuencias absolutas acumuladas, frecuencias relativas, frecuencias relativas acumuladas, porcentajes y porcentajes acumulados (comprueba tú que los valores de estas columnas son los que se ofrecen a continuación).

kilometraje anual (en miles de kilómetros)	marca de clase	número de vehículos					
$(e_{i-1}, e_i]$	x_i	f_i	F_i	h_i	H_i	p_i	P_i
(0, 4]	2	228	228	0,092	0,092	9,2	9,2
(4, 8]	6	634	862	0,256	0,348	25,6	34,8
(8, 12]	10	821	1683	0,331	0,679	33,1	67,9
(12, 16]	14	475	2158	0,192	0,871	19,2	87,1
(16, 20]	18	233	2391	0,094	0,965	9,4	96,5
(20, 24]	22	87	2478	0,035	1	3,5	100
		$n = 2478$		1			

Basta un vistazo a la tabla anterior para darse cuenta, por ejemplo, de que el 9,4% de los coches han realizado entre 16000 y 20000 kilómetros. O de que hay 1683 vehículos que han hecho 12000 kilómetros o menos.

MEDIDAS DE CENTRALIZACIÓN

Hay determinadas medidas (también llamadas *estadísticos*) que dan una idea general de la distribución estadística, ya sea la variable discreta o continua. Algunas de ellas nos indican los puntos en torno a los cuales se encuentran los valores de la variable estadística en estudio; es decir, nos indican los puntos centrales de una distribución. Representan el conjunto de los datos mediante un solo valor numérico, tratando de resumir y sintetizar la distribución de frecuencias. Las medidas de posición central más utilizadas son la mediana, la moda y la media aritmética.

La media aritmética

La **media aritmética**, o simplemente **media**, de una distribución estadística asociada a una variable objeto de estudio, se representa mediante el símbolo $\bar{x}$ y su valor se calcula mediante la siguiente fórmula:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \dots + x_k f_k}{n} = \frac{\sum_{i=1}^k x_i f_i}{n}$$

Abreviadamente podemos escribir: $\bar{x} = \frac{\sum x_i f_i}{n}$

En el caso de variable discreta x_i son los valores que toma la variable y en el caso de variable continua son las marcas de clase.

En el **Caso 1** la media es:

$$\begin{aligned}\bar{x} &= \frac{0 \cdot 2 + 1 \cdot 3 + 2 \cdot 1 + 3 \cdot 1 + 4 \cdot 1 + 5 \cdot 3 + 6 \cdot 2 + 7 \cdot 5 + 8 \cdot 7 + 9 \cdot 5 + 10 \cdot 0}{30} = \\ &= \frac{0 + 3 + 2 + 3 + 4 + 15 + 12 + 35 + 56 + 45 + 0}{30} = \frac{175}{30} = 5,83\end{aligned}$$

Esta medida indica que el valor promedio, o calificación media de la clase en la materia de matemáticas, es 5,83.

En la práctica lo que se hace es añadir una columna a la derecha de la columna de las frecuencias absolutas f_i con los productos $x_i f_i$, y escribir la suma de todos estos valores en una última celda correspondiente a esta columna:

x_i	f_i	$x_i f_i$
0	2	0
1	3	3
2	1	2
3	1	3
4	1	4
5	3	15
6	2	12
7	5	35
8	7	56
9	5	45
10	0	0
	$n = 30$	$\sum x_i f_i = 175$

De este modo es más sencillo e intuitivo su cálculo: $\bar{x} = \frac{\sum x_i f_i}{n} = \frac{175}{30} = 5,83$.

En el **Caso 2** las columnas que nos interesan para el cálculo de la media serían, añadiendo la de los productos $x_i f_i$:

kilometraje anual (en miles de kilómetros)	marca de clase	número de vehículos	
$(e_{i-1}, e_i]$	x_i	f_i	$x_i f_i$
(0, 4]	2	228	456
(4, 8]	6	634	3804
(8, 12]	10	821	8210
(12, 16]	14	475	6650
(16, 20]	18	233	4194
(20, 24]	22	87	1914
		$n = 2478$	25228

Entonces $\bar{x} = \frac{\sum x_i f_i}{n} = \frac{25228}{2478} = 10,18$, valor que viene a decir que la distancia recorrida en un año por el parque automovilístico formado por 2478 coches es, por término medio, de 10180 kilómetros (el kilometraje anual viene dado en miles de kilómetros, por eso la media se multiplica por mil).

MEDIDAS DE DISPERSIÓN

Las medidas de centralización sintetizan la información: representan la totalidad del conjunto de datos mediante unos valores centrales. Sin embargo, un promedio no es suficiente. Es preciso añadir también una medida de cómo de representativo es dicho promedio.

Consideremos las siguientes distribuciones:

grupo A:	20	22	24	26	28
grupo B:	10	10	20	35	45

que podrían representar los pesos de dos grupos de niños. Observamos que los dos grupos tienen el mismo peso medio: $\bar{x} = 24$, siendo, no obstante, muy diferentes en cuanto a concentración-dispersión de sus valores. En el grupo A los valores se encuentran próximos a la media, luego tienen poca dispersión. En el grupo B, los valores están alejados de la media, estando formado por valores más dispersos.

Al grado en que los datos numéricos tienden a extenderse alrededor de la media se le llama variación o dispersión de los datos. Se utilizan distintas medidas de dispersión. Las más empleadas son: rango o recorrido, recorrido intercuartílico, desviación absoluta media, varianza, desviación típica y coeficiente de variación. Nosotros nos ceñiremos a las más importantes: varianza y desviación típica.

La varianza y la desviación típica

Antes de nada, es conveniente decir que el cálculo de estas medidas es igual para variable discreta que para variable continua. Solamente hay que tener en cuenta, como en el caso anterior de la media, que en el caso discreto x_i hace referencia a los valores que toma la variable y, en el caso continuo, a las marcas de clase.

La más empleada de las medidas de dispersión es la **varianza**, que se simboliza por $\text{Var}(X)$, $\sigma^2(X)$ o, simplemente, σ^2 (esta es la forma que nosotros utilizaremos). Se define como *la media de los cuadrados de las desviaciones respecto a la media*; esto es, la media de la variable $(X - \bar{X})^2$:

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

Desarrollando la expresión anterior queda una fórmula más cómoda a la hora de realizar los cálculos:

$$\sigma^2 = \frac{\sum x_i^2 f_i}{n} - \bar{x}^2$$

En la práctica se calcula la media y se añade, a la derecha de la columna de los productos $x_i f_i$, otra con los productos $x_i^2 f_i$ (basta multiplicar cada uno de los anteriores por x_i), y se escribe la suma de todos estos valores en una última celda correspondiente a esta columna.

Veámoslo para el **Caso 1**:

x_i	f_i	$x_i f_i$	$x_i^2 f_i$
0	2	0	0
1	3	3	3
2	1	2	4
3	1	3	9
4	1	4	16
5	3	15	75
6	2	12	72
7	5	35	245
8	7	56	448
9	5	45	405
10	0	0	0
	$n = 30$	$\sum x_i f_i = 175$	$\sum x_i^2 f_i = 1277$

De este modo tenemos que la varianza es $\sigma^2 = \frac{\sum x_i^2 f_i}{n} - \bar{x}^2 = \frac{1277}{30} - 5,83^2 = 42,57 - 33,99 = 8,58$. Observa que se ha utilizado en la fórmula el valor de la media calculado anteriormente.

Puesto que la varianza no viene dada en las mismas unidades que la variable objeto de estudio (por ejemplo, si la variable viene dada en metros, la varianza resulta en metros cuadrados), en su lugar se emplea la **desviación típica**, σ , definida como la raíz cuadrada positiva de la varianza:

$$\sigma = +\sqrt{\sigma^2}$$

En el **Caso 1** la desviación típica es $\sigma = +\sqrt{8,58} = 2,93$

Cuanto mayores sean la varianza o la desviación típica mayor será el grado de dispersión o alejamiento de los datos respecto de la media. En el caso trivial de que todos los valores de la variable estén concentrados en un punto (que coincidirá con la media), la varianza y la desviación típica se anularán.

La desviación típica anterior indica que el grado de dispersión de las calificaciones de los alumnos en la materia de matemáticas es, por término medio, de 2,93 puntos. Como la media era 5,83 podemos decir que “la mayoría” de los alumnos de la clase obtienen calificaciones entre $5,83 - 2,93 = 2,9$ y $5,83 + 2,93 = 8,76$. Se puede asumir que la dispersión es alta pues estos dos valores están cerca de los valores extremos: 0 y 10. Eso quiere decir que hay muchos alumnos con notas bajas y muchos alumnos con notas altas, lejos de la media.

Halla tú ahora la varianza y la desviación típica del **Caso 2**.

Ejercicios y problemas

1. Completar los datos que faltan en la siguiente tabla estadística:

x_i	f_i	F_i	h_i	H_i	$x_i f_i$	$x_i^2 f_i$
1	4		0,08			
2	4					
3		16	0,16			
4	7		0,14			
5	5	28				
6		38				
7	7	45	0,14			
8						
	$n =$					

Calcular la media, la varianza y la desviación típica.

2. Las calificaciones de dos grupos de diez alumnos en la Primera Evaluación en una cierta asignatura se recogen en la siguiente tabla:

Grupo A	0	1	1	3	5	5	6	8	8	9
Grupo B	2	2	4	4	4	5	5	6	6	8

¿Cuál de los dos grupos obtuvo mejores resultados? ¿Qué grupo es más homogéneo?

3. La siguiente tabla recoge los minutos de retraso en la incorporación al trabajo de los empleados de una empresa:

Retraso en minutos	(0, 4]	(4, 8]	(8, 12]	(12, 16]	(16, 20]
Número de empleados	5	15	18	10	4

Calcular el retraso medio y la desviación típica.

4. En un estudio sobre el sueldo en euros de 50 personas se han obtenido los siguientes datos:

Sueldo	(500, 700]	(700, 900]	(900, 1300]	(1300, 1500]	(1500, 2100]
Nº de personas	10	10	20	9	1

Calcular la media, la varianza y la desviación típica y explicar el significado de estos parámetros

5. Los pesos en kilogramos de 50 personas vienen dados por la tabla:

Peso	(50, 60]	(60, 70]	(70, 80]	(80, 90]	(90, 100]
Número de empleados	10	15	20	4	1

Calcular el peso medio y la desviación típica. Interpreta los resultados. ¿Se puede decir que es un grupo homogéneo?

6. En la sección de maternidad de un hospital se han tomado los pesos (en kilogramos) de 50 recién nacidos:

2,8	3,2	3,8	2,5	2,7	3,7	1,9	2,6	3,5	2,3
3,0	2,6	1,8	3,3	2,9	2,1	3,4	2,8	3,1	3,9
2,9	3,5	3,0	3,1	2,2	3,4	2,5	1,9	3,0	2,9
2,4	3,4	2,0	2,6	3,1	2,3	3,5	2,9	3,0	2,7
2,9	2,8	2,7	3,1	3,0	3,1	2,8	2,6	2,9	3,3

Construye, comenzando por el peso menor, una tabla de frecuencias absolutas con los datos agrupados en seis intervalos de amplitud 0,4. Calcula la media y la desviación típica.

7. Las puntuaciones obtenidas por 50 alumnos en una cierta prueba son las siguientes:

6	5	8	2	6	4	6	5	7	5
5	5	4	7	3	6	7	5	6	6
4	7	6	8	5	5	6	7	7	4
5	5	5	3	4	6	5	4	2	6
5	4	6	7	8	6	6	9	6	8

- a) Ordenar los datos anteriores en una tabla de frecuencias absolutas comenzando por la puntuación menor y terminando en la mayor. Complétala con las columnas correspondientes a los porcentajes absolutos, frecuencias absolutas acumuladas y porcentajes absolutos acumulados.
- b) A la vista de la tabla construida en el apartado a), ¿cuántos alumnos obtienen un 5 o más?, ¿qué porcentaje de alumnos obtienen menos de un 7?
- c) Calcular la media, la varianza y la desviación típica de la puntuación obtenida.
8. Las urgencias atendidas durante los treinta días de un mes en un centro de salud son:

1	5	3	2	1	6	4	2	2	3
4	3	5	1	0	1	5	3	3	6
2	4	6	3	2	4	3	2	1	5

- a) Realiza una tabla de frecuencias absolutas.
- b) Calcula la media de urgencias atendidas por día en el centro de salud. Calcula también la varianza y la desviación típica. Interpreta los resultados.