Capítulo
15

ESTADÍSTICA

Con la llegada de los grandes computadores, ahora podemos procesar una enorme cantidad de datos, lo que ha contribuido al avance en áreas muy diversas como la medicina, el deporte, robótica, o la sociología, la que nos ayuda a estudiar el comportamiento y el consumo humano.

Existen dos tipos de estadística, la descriptiva y la inferencial. En este capítulo veremos la primera, la que se dedica a analizar los datos obtenidos, graficándolos y calculando diversos parámetros, como la media, los percentiles, la desviación estándar, etc.

Por otra parte, la estadística inferencial se preocupa de sacar conclusiones de una población a partir de una muestra de ella, lo cual es muy importante en cualquier estudio científico.

CONCEPTOS CLAVES

- **Organización de datos**
- **Representación de datos**
- **Medidas de tendencia central**
- **Rango**
- **Medidas de posición**

ORGANIZACIÓN DE DATOS

Si tenemos un conjunto de datos, existen diversas formas de organizarlos, acá veremos solo los más frecuentes.

- **Diagrama de tallo y hojas**

Este diagrama permite ordenar los datos de tal forma que los de mayor frecuencia se destaquen sobre los demás, esto también se produce en un gráfico de barras o un histograma como veremos más adelante.
En este diagrama se coloca en el tallo la o las cifras de mayor valor posicional y en las hojas las cifras restantes.
Por ejemplo, las siguientes notas: 3,2 ; 3,5 ; 4,1 ; 4,7 ; 4,9, 5,1 ; 5,5 ; 5,8 ; 5,9 ; 6,0 ; 6,5 ; 7,0 en un diagrama de tallo y hojas quedarían de la siguiente forma:

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

- **Tabla de frecuencias**

En las tablas de frecuencias, al lado del dato aparece la frecuencia del dato, es decir la cantidad de veces que se repite.

Ejemplo:

Dato	Frecuencia
12	3
15	4
18	7
21	6

También podemos disponer de una tabla de frecuencias acumuladas, donde aparece la cantidad de datos que son menores o iguales que él:

Dato	Frecuencia acumulada
12	3
15	7
18	14
21	20

Por ejemplo, que el dato 18 tenga una frecuencia acumulada de 14 indica que hay 14 datos que son menores o iguales que él.
También podemos tener una tabla de frecuencias relativas, esta indica que parte es la frecuencia del dato con respecto al total, esta se expresa en números decimales.

Siguiendo con el mismo ejemplo, tenemos que el dato 12 se repite 3 veces de un total de 20, es decir en términos de fracciones es $\frac{3}{20}$, o bien 3 : 20 = 0,15 en decimal.

Dato	Frecuencia	Frecuencia relativa
12	3	0,15
15	4	0,2
18	7	0,35
21	6	0,3

Si la frecuencia relativa la expresamos en términos porcentuales, se denomina frecuencia relativa porcentual:

Dato	Frecuencia	Frecuencia relativa	Frecuencia relativa porcentual
12	3	0,15	15%
15	4	0,2	20%
18	7	0,35	35%
21	6	0,3	30%

También podemos tener tablas con frecuencias relativas acumuladas o frecuencias porcentuales acumuladas:

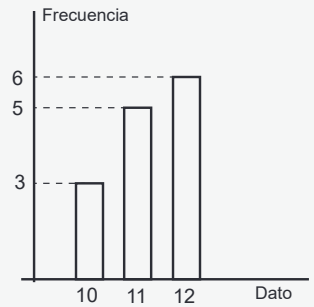
Dato	Frecuencia	Frecuencia relativa	Frecuencia relativa acumulada	Frecuencia relativa porcentual	Frecuencia acumulada porcentual
12	3	0,15	0,15	15%	15%
15	4	0,2	0,35	20%	35%
18	7	0,35	0,70	35%	70%
21	6	0,3	1,0	30%	100%

GRÁFICOS

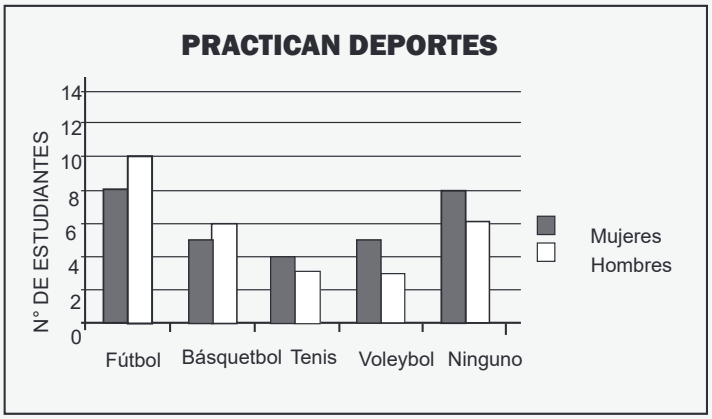
Existen diversos gráficos para representar datos, entre ellos los más importantes están el gráfico de barras y el histograma, en ambas las alturas de las columnas que se presentan están relacionadas con las frecuencias de los datos.

- **Gráfico de barras**

En el eje horizontal se colocan los datos y en el vertical las frecuencias de los datos, tal como se muestra en el siguiente ejemplo:

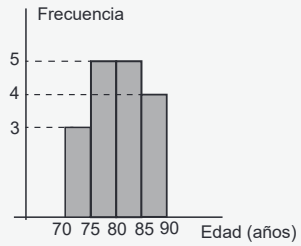


También existen los gráficos de barras dobles que nos permiten comparar dos variables:



- **Histograma**

Se utiliza frecuentemente cuando la variable es continua o discreta agrupada en intervalos, generalmente los intervalos son de la forma [a,b[y el último de la forma [a,b], a no ser que se indique lo contrario. El histograma está constituido por rectángulos contiguos donde su altura es proporcional a la frecuencia del intervalo:

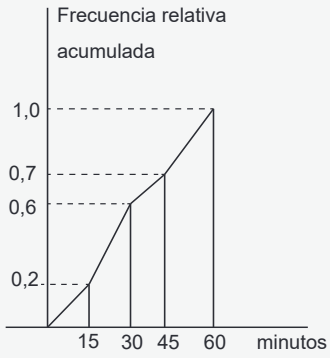


- **Ojiva (superior)**

La ojiva corresponde a un gráfico de frecuencias acumuladas, de frecuencias relativas o porcentuales acumuladas.

Este gráfico nos da información acerca de la cantidad de datos que son inferiores o superiores a un dato en particular.

Ejemplo de ojiva con frecuencia relativa acumulada

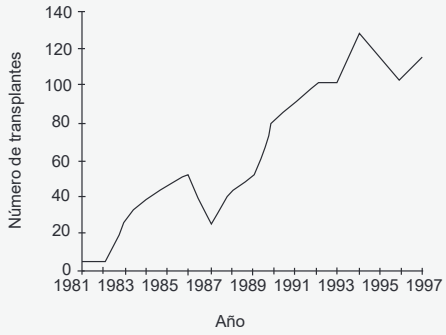


En este caso se observa que para el intervalo [30,45[, el gráfico aumenta de 0,6 a 0,7 es decir aumentó su frecuencia relativa en 0,1 en decir un 10%, por lo tanto un 10% de los datos está en este intervalo.
Por otro lado para el extremo derecho de este intervalo, es decir para el 45, presenta una frecuencia relativa acumulada de 0,7, esto significa que un 70% de los datos son inferiores a él.

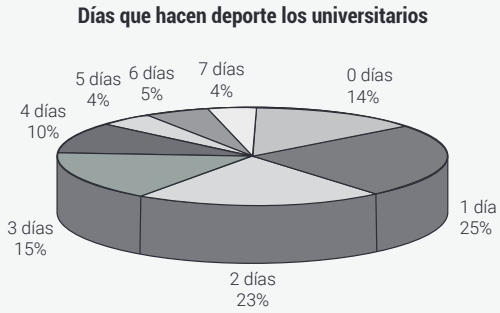
- **Otros gráficos**

También tenemos otros tipos de gráficos como los de línea, los circulares y los pictogramas.

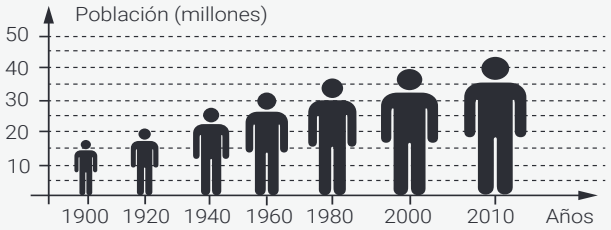
De línea:



Circulares:



Pictogramas:



MEDIDAS DE TENDENCIA CENTRAL

Las medidas de tendencia central son la media, la mediana y la moda .

- **Media , media aritmética o promedio $\overline{x}$**

A) Si se tiene datos dados sin frecuencia, la media se calcula sumando los datos y dividiendo esta suma por el total de datos.

Datos: $x_1, x_2, x_3,...,x_n \rightarrow$ $\overline{x} = \dfrac{x_1 + x_2 + x_3 + ... + x_n}{n}$

B) Si los datos vienen dados en una tabla de frecuencia, entonces la media se calcula multiplicando cada dato con su respectiva frecuencia, se suman estos productos y se divide por el total de datos.

Datos:

Dato	Frecuencia
x_1	f_1
x_2	f_2
...	...
x_n	f_n

$\rightarrow$ $\overline{x} = \dfrac{x_1 \cdot f_1 + x_2 \cdot f_2 + ... + x_n \cdot f_n}{f_1 + f_2 + ... + f_n}$

C) Si los datos vienen dados en una tabla de frecuencia relativas, entonces la media se calcula multiplicando cada dato con su respectiva frecuencia relativa y se suman todos estos productos.

Datos:

Dato	Frecuencia relativa
x_1	r_1
x_2	r_2
...	...
x_n	r_n

$\rightarrow$ $\overline{x} = x_1 \cdot r_1 + x_2 \cdot r_2 + ... + x_n \cdot r_n$

Nota: si en la tabla se indicaran las frecuencias relativas porcentuales, se efectúa la misma operación anterior pero se dividiría por 100.

D) Si los datos están agrupados en intervalos, entonces la media se calcula multiplicando cada marca de clase del intervalo con la frecuencia correspondiente, se suman estos productos y se divide por el total de datos. La marca de clase corresponde a la media de los datos extremos del intervalo.

Datos:

Datos	Marca de clase	Frecuencia
$[x_1, x_2[$	$\overline{x_1}$	f_1
$[x_2, x_3[$	$\overline{x_2}$	f_2
...		...
$[x_n, x_{n+1}]$	$\overline{x_n}$	f_n

La media tiene las siguientes propiedades:

1) Si a todos los datos se le suma una constante (o se le resta), entonces la media aumenta (o se disminuye) esa constante.

$x_1, x_2, x_3,...,x_n \rightarrow$ Media: $\overline{x}$

$x_1 + k, x_2 + k, x_3 + k,...,x_n + k \rightarrow$ Media: $\overline{x} + k$

2) Si todos los datos se multiplican (o dividen) por una constante, entonces la nueva media se obtiene multiplicando (o dividiendo) la media anterior por dicha constante.

$x_1, x_2, x_3,...,x_n \rightarrow$ Media: $\overline{x}$

$x_1 \cdot k, x_2 \cdot k, x_3 \cdot k,...,x_n \cdot k \rightarrow$ Media: $\overline{x} \cdot k$

- **Mediana**

Si se ordenan los datos en sentido creciente o decreciente, la mediana es el dato que se ubica al centro (en el caso de ser uno) o es la media de los dos datos centrales.

Si el número de datos es n y n es impar, la mediana es el dato de lugar $\frac{n+1}{2}$.

En el caso que n fuera par, la mediana es la media entre los datos de lugares $\frac{n}{2}$ y $\frac{n}{2}+1$.

- **Moda**

La moda es el dato que tiene mayor frecuencia.
Si todos los datos tienen la misma frecuencia diremos que no hay moda (muestra amodal).
Un conjunto de datos puede tener más de una moda.(muestra multimodal)

MEDIDAS DE DISPERSIÓN

Las medidas de dispersión nos indican cuan dispersos están los datos, una de esas medidas es el rango.

- **Rango**

Es la diferencia entre el dato mayor y el dato menor.
Este estadígrafo es cero en el caso en que todos los datos son iguales y es positivo en el resto de los casos.

MEDIDAS DE POSICIÓN O PERCENTILES

El percentil k o Pk, es un dato que es mayor o igual al k% de los datos.
Es decir, al hablar del percentil 60, es un dato que es mayor o igual al 60% de los datos.
Los cuartiles, son los percentiles 25, 50 y 75 y se designan como Q_1, Q_2 y Q_3 respectivamente, observa que Q_2 coincide con la mediana, con los cuartiles podemos construir el diagrama de caja o de cajón con bigotes como veremos más adelante.
Existen otros percentiles importantes, como los quintiles, que se ocupan bastante en Economía, como por ejemplo cuando hablamos de los quintiles de ingresos de un grupo familiar, los quintiles no son nada más que los percentiles: P_{20}, P_{40}, P_{60}, P_{80}.
Los deciles, son los percentiles, P_{10}, P_{20}, hasta el P_{90}.
Tenemos que distinguir si estamos calculando percentiles para datos discretos o percentiles para datos agrupados en intervalos, ya que su cálculo es diferente, como veremos en los siguientes ejemplos.

- **Percentil para datos discretos**

Nos referiremos a datos discretos a aquellos, que al ordenarlos de menor a mayor (o de mayor a menor) entre dos consecutivos, no existen datos entremedio, por ejemplo si las notas de Pablo en la asignatura de Física son: 4,6 ; 5,0 ; 5,2 ; 5,8 ; 6,0 ; 6,0 ; 6,8 y 7,0, entre dos notas consecutivas no hay otra entremedio (observa que han sido ordenadas en sentido creciente).
Para calcular percentiles para datos discretos, procederemos de la siguiente forma:
Supongamos que tenemos n datos y queremos calcular el percentil P_k:

1°) Se ordenan los datos en sentido creciente.

2°) Se calcula la expresión $\frac{kn}{100}$, si este valor te da decimal, el P_k es el dato de lugar siguiente al valor de esta expresión, si te da entero se calcula el promedio entre los datos de lugares $\frac{kn}{100}$ y $\frac{kn}{100}+1$.

Ejemplo:
Consideremos las notas de Pablo: 4,6 ; 5,0 ; 5,2 ; 5,8 ; 6,0 ; 6,0 ; 6,8 y 7,0

Determinamos el percentil 20, calculamos: $\frac{20 \cdot 8}{100}$, lo que nos da 1,6, como nos dio decimal aproximamos a

2, luego el percentil 20 corresponde al 2° dato, el cual es 5,0.

Si calculamos el tercer cuartil, esto corresponde al percentil 75, calculamos: $\frac{75 \cdot 8}{100}$, esto da 6, como es un entero, calculamos el promedio entre el 6° y el 7° dato, esto es $\frac{6 + 6{,}8}{2}$, por lo tanto el tercer cuartil es 6,4.

Observa que si calculamos el percentil 50 o la mediana, calculamos $\frac{50 \cdot 8}{100} = 4$, por lo que hay que calcular el promedio entre el 4° y el 5° dato: $\frac{5{,}8 + 6}{2} = 5{,}9$, lo que coincide con calcular la media entre los datos centrales tal como lo habíamos visto anteriormente.

Resumiendo:

Para datos discretos:	
1°) Se ordenan los datos en sentido creciente.	
2°) Se calcula: $\frac{kn}{100}$	
Si resulta decimal:	El percentil es el dato de lugar siguiente a $\frac{kn}{100}$.
Si resulta entero:	El percentil se calcula con el promedio de los datos de lugares: $\frac{kn}{100}$ y $\frac{kn}{100} + 1$.

- **Percentil para datos agrupados en intervalos**

Si tenemos n datos agrupados en intervalos, calcularemos la misma expresión anterior: $\frac{kn}{100}$, solo que ahora cambiaremos el criterio de cálculo del percentil:

Si $\frac{kn}{100}$ te da un entero, se busca el dato de lugar $\frac{kn}{100}$, si te da decimal, se calcula el promedio entre los datos más cercanos a este valor. Es importante considerar que cómo no sabemos cómo se distribuyen los datos en cada intervalo, solo podremos indicar en que intervalo se encuentra dicho percentil.

Ejemplo:

En la siguiente tabla se muestra la distribución de las notas en la última prueba de la asignatura de Lenguaje. Determina en qué intervalo se encuentra la mediana, el percentil 70 y el percentil 55.

Notas	N° de alumnos
[3 , 4[	5
[4 , 5[	8
[5 , 6[	11
[6 , 7]	6

Solución:

Si sumamos las frecuencias, obtenemos 5 + 8 + 11 + 6 = 30, luego n = 30.

La mediana corresponde al dato de lugar $\frac{50 \cdot 30}{100} = 15$, luego ubicamos el dato de lugar 15, para ello podemos ir sumando las frecuencias, hasta que sobrepasemos este valor, por ejemplo si sumamos las frecuencias de los dos primeros intervalos, obtenemos 5 + 8 = 13, aún no sobrepasamos los 15, por lo tanto tomamos un intervalo más: 5 + 8 + 11 = 24, por lo tanto la mediana está en el intervalo [5 , 6[. Un método equivalente es calcular la frecuencia acumulada y al primer intervalo que sobrepasemos el percentil, en ese intervalo se encontrará:

Notas	N° de alumnos	Frecuencia acumulada
[3 , 4[	5	5
[4 , 5[	8	13
[5 , 6[	11	24
[6 , 7]	6	

No sobrepasa el valor 15

Sobrepasa el valor 15, en este intervalo está el percentil.

El percentil 70 será el dato de lugar $\frac{70 \cdot 30}{100} = 21$, si ocupamos la técnica anterior, podemos darnos cuenta que también está en el intervalo [5 , 6[.

Si calculamos el percentil 55, tenemos: $\frac{55 \cdot 30}{100} = 16{,}5$, entonces tenemos que calcular el promedio de los datos de lugares 16 y 17, como ambos datos están en el intervalo [5 , 6[, el promedio de ambos sigue estando en ese intervalo, luego el percentil 55 está en el intervalo [5 , 6[.

Resumiendo:

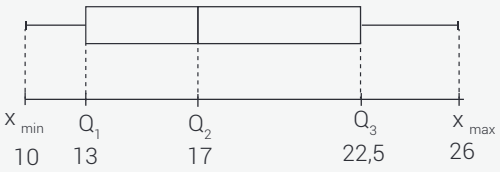
Para datos distribuidos en intervalos:	
1°) Se ordenan los datos en sentido creciente (generalmente los intervalos están ordenados en una tabla en este sentido)	
2°) Se calcula: $\frac{kn}{100}$	
Si resulta decimal:	El percentil se calcula con el promedio de los datos de lugares más cercanos a: $\frac{kn}{100}$ y se determina en qué intervalo está ocupando la frecuencia acumulada.
Si resulta entero:	El percentil es el dato de lugar $\frac{kn}{100}$ y se determina en qué intervalo está ocupando la frecuencia acumulada.

- **Diagrama de cajón y bigotes**

El diagrama de cajón y bigotes, es una representación visual de cuan dispersos están los datos entre los valores mínimo, los cuartiles y el valor máximo.

Ejemplo:

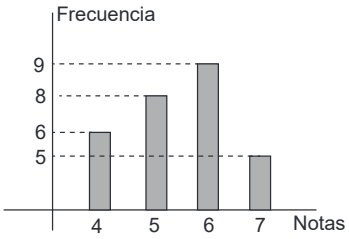
Supongamos que las edades de los nietos de una familia son: 10, 12, 12, 14, 16, 16, 18, 20, 22, 23, 24, 26.
Tenemos que el dato mínimo es 10, el primer cuartil es Q_1=13, el segundo cuartil o mediana es Q_2=17, tercer cuartil o Q_3 = 22,5 y el dato máximo igual a 26.
El diagrama de cajón y bigotes es un rectángulo donde en el extremo izquierdo se ubica Q_1, en el extremo derecho Q_3, y la mediana es una línea vertical que separa a este rectángulo.
Desde los extremos de este rectángulo o caja salen los bigotes que son segmentos que se extienden hasta el valor mínimo y al valor máximo por el otro:



Se llama valor intecuartílico a la diferencia entre Q_3 y Q_1, en este caso es 22,5 − 13 = 9,5, lo que indica que en un rango de 9,5 años está el 50% de los datos (pueden haber más). Por otro si consideramos la partición que produce la mediana, en este ejemplo el lado derecho es mayor que el lado izquierdo, lo que indica que los datos están más dispersos entre la mediana y el tercer cuartil.

EJERCICIOS RESUELTOS

1. En el siguiente gráfico se muestra la distribución de notas de un cierto curso en la última prueba de Física.



¿Cuál de las siguientes afirmaciones **NO** se puede deducir de la información dada en el gráfico?

A) La mediana es 5,5.
B) La moda es 6,0.
C) La media es inferior a la mediana.
D) El primer cuartil es 4.

Solución:

A) Para calcular la mediana, en primer lugar que determinar cuántos datos son, para ello sumamos las frecuencias: 6 + 8 + 9 + 5 = 28. Como es un número par de datos, los centrales son los de lugares 14 y 15, si sumamos las frecuencias del 4 y del 5 suman 14, por lo tanto el dato de lugar 14 es 5 y el dato 15 es 6, si calculamos la media entre 5 y 6 nos 5,5 por lo tanto la mediana es 5,5 , por lo tanto A es verdadera.

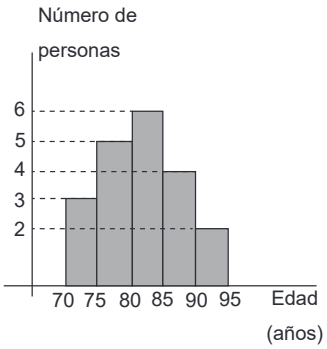
B) La moda es el dato con mayor frecuencia, en este caso es el 6, por lo tanto es correcta.

C) Para calcular la media tenemos que multiplicar cada dato con su respectiva frecuencia, sumar estos productos y el resultado dividirlo con la suma de las frecuencias, entonces

$\overline{x} = \frac{4\cdot6+5\cdot8+6\cdot9+7\cdot5}{6+8+9+5} = \frac{153}{28} \approx 5{,}46$, como la mediana era 5,5, se tiene que la afirmación es verdadera.

D) El primer cuartil es el dato de lugar $\frac{1}{4}\cdot 28 = 7$, luego debemos calcular la media entre el 7° y 8° dato, ambos datos son 5, por lo tanto su media es 5, luego E) es falsa.

2. En una casa de reposo, se ha consultado acerca de la edad de los residentes; con esta información se ha construido el siguiente histograma, donde los intervalos son de la forma [a, b[y el último de la forma [a , b].



¿Cuál de las siguientes afirmaciones es **FALSA**?

A) Por lo menos el 10% tiene a lo menos 90 años.
B) 85% tiene por lo menos 75 años.
C) El percentil 50 se encuentra en el intervalo [80 , 85[
D) El 30% tiene más de 85 años.

Solución:

A) En el intervalo [90, 95], se ubican 2 personas de un total de 20 (este total lo obtienes sumando las frecuencias), por lo tanto un 10% de los residentes tienen una edad mayor o igual que 90 años, luego A es verdadera

B) Las personas que tienen por lo menos 75 años son: 5 + 6 + 4 + 2 = 17, pero 17 de un total de 20 corresponde a un 85%, luego es verdadera.

C) El total de datos es 20, para hallar el percentil 50, calculamos $\frac{1}{2} \cdot 20 = 10$ y el 10° dato está en el intervalo [80, 85[, luego es verdadera.

D) Si sumamos las frecuencias de los dos últimos intervalos, tenemos 4 + 2 = 6, de un total de 20, luego un 30% de los datos son mayores o iguales que 85, pero no tienen más de 85, luego es falsa.

3. En una fábrica que produce ampolletas de bajo consumo, se ha tomado una muestra de 5000 ampolletas de un determinado modelo y se ha medido su vida útil, obteniéndose lo siguiente:

Nº de horas duración	Nº de ampolletas
10000 - 10100	1000
10101 - 10200	900
10201 - 10300	1100
10301 – 10400	1500
10401 - 10500	500

¿Cuál de las siguientes afirmaciones es **FALSA**?

A) El percentil 40 y el percentil 50 están en el tercer intervalo.
B) El percentil 60 está en el tercer intervalo.
C) El percentil 80 está en el cuarto intervalo.
D) El percentil 90 está en el último intervalo.

Solución:

En A), el total de datos es igual a la suma de las frecuencias, en este caso tenemos

1000 + 900 + 1100 + 1500 + 500 = 5000.

El percentil 40 es el dato que está en el lugar $\frac{40 \cdot 5000}{100} = 2000$, si sumamos las frecuencias de los dos primeros intervalos tenemos 1000 + 900 = 1900, con ello no alcanzamos los 2000, pero si sumamos los tres primeros intervalos tenemos 1000 + 900 + 1100 = 3000, luego el dato 2000 está en el tercer intervalo. El percentil 50 está en el lugar $\frac{50 \cdot 5000}{100} = 2500$, por lo visto anteriormente este dato está también en el tercer intervalo, luego la afirmación es verdadera.

En B), el percentil 60 es el dato que está en el lugar $\frac{60 \cdot 5000}{100} = 3000$, si sumamos los tres primeros intervalos, tal como vimos anteriormente, obtenemos 100 + 900 + 1100 = 3000, luego el dato está en el tercer intervalo, luego B) es verdadera.

En C), el percentil 80 está en el dato de lugar $\frac{80 \cdot 5000}{100} = 4000$, si sumamos las frecuencias de los tres primeros intervalos tenemos 3000, con lo que no alcanzamos aún el dato de lugar 4000, si sumamos los cuatro primeros intervalos, tenemos 4500, luego el percentil 80 se encuentra en el cuarto intervalo, la afirmación es verdadera.

En D), el percentil 90 está en el dato de lugar $\frac{90 \cdot 5000}{100} = 4500$ y tal como vimos anteriormente, alcanzamos este valor sumando los cuatro primeros intervalos, luego el percentil 90 está en el cuarto intervalo y no en el último, luego es falsa.

4. A un grupo de estudiantes de cuarto medio se le ha aplicado un facsímil de Matemática de 80 preguntas. Con la cantidad de preguntas buenas obtenidas se ha construido la siguiente tabla:

N° preguntas correctas	N° de estudiantes	Frecuencia relativa porcentual
[0 , 20[	4	
[20 , 40[		
[40 , 60[	36	45%
[60 , 80]	20	

¿Cuál de las siguientes afirmaciones **NO** se puede deducir de la tabla?

A) El total de alumnos que rindieron el facsímil es 80.
B) 20 alumnos obtuvieron a lo menos 20 y menos de 40 preguntas buenas.
C) El percentil 50 está en el intervalo [40, 60[.
D) El 30% obtuvo a lo sumo 40 preguntas buenas.

Solución:

Por la información dada, tenemos que el 45% de los datos está en el tercer intervalo, con ello podemos obtener el total de personas que rindieron el facsímil.

Si x es el total de personas, planteamos, $\frac{45}{100}x = 36 \rightarrow x = \frac{36 \cdot 100}{10} = 80$. Por lo tanto A es correcta.

Con el resultado anterior, podemos completar la tabla, con lo que se obtiene:

N° preguntas correctas	N° de estudiantes	Frecuencia relativa porcentual
[0 , 20[	4	5%
[20 , 40[	20	25%
[40 , 60[	36	45%
[60 , 80]	20	25%
Total	80	

B) Es correcta debido a que la frecuencia del intervalo [20 , 40[efectivamente es 20.

C) Para determinar en qué intervalo está el percentil 50 debemos, ir sumando la última columna, hasta sobrepasar el 50%, esto se obtiene sumando las tres primeras frecuencias relativas porcentuales: 5% + 25% + 45% = 75%, luego la mediana está en el tercer intervalo, C) es correcta.

D) Observemos que si sumamos las frecuencias relativas porcentuales de los dos primeros intervalos, obtenemos, 5% + 25% = 30%, por lo tanto un 30% de los que rindieron la prueba obtuvieron menos de 40 preguntas buenas, lo que se contradice con lo enunciado en D, ya que esta afirma que el 30% obtuvo 40 o menos preguntas buenas, luego es falsa.