Ejercicio de estadística descriptiva

El consumo diario de agua en una curtiembre responde a la siguiente distribución de frecuencias:

X (Miles de litros)	Días
20 - 30	1
30 - 40	15
40 - 50	39
50 - 60	32
60 - 70	11
70 - 80	2

a) Calcular la media, la varianza, el coeficiente de variación e indicar si la dispersión es alta o baja.
b) Calcular el porcentaje de los días que x>32 y x<51.
c) Calcular la mediana y el valor de la variable superado con 10%, 40% y 90% de probabilidad.
d) Calcular el recorrido intercuartílico y la desviación mediana.
e) Calcular el modo o moda e indicar, de acuerdo a los datos calculados, si la distribución es simétrica o no y, en caso de no serlo, qué tipo de asimetría presenta.
f) Dibujar el histograma y su curva correspondiente. Dibujar las curvas acumuladas.

Vamos a definir primer la variable aleatoria:

$X = Consumo\ diario\ de\ agua\ en\ una\ curtiembre\ (en\ miles\ de\ litros)$.

Antes de empezar a resolver el ejercicio, vamos a comentar algunas generalidades de los datos presentados en intervalos o clases.
La lectura del cuadro es la siguiente: hay 1 día en que se observó un consumo de agua en una curtiembre de entre 20 y 30 mil litros; hubo 15 días en que se observó un consumo de agua de entre 30 y 40 mil litros; etc.
Los “Días” corresponden a la frecuencia absoluta simple.

Los intervalos son cerrados en el límite inferior y abiertos en el límite superior. Esto quiere decir que si un día se observara un consumo de agua de exactamente 30 mil litros, esa observación debería contarse en el segundo intervalo y no en el primero. Y que, además, si un día se observara un consumo de exactamente 80 mil litros, esa observación quedaría fuera del último intervalo.
Por eso es importante que, antes de armar la tabla con intervalos veamos cuál es el valor mínimo de la variable y cuál es el valor máximo. El último intervalo debe tener un límite superior mayor al valor máximo, $X_{máx}$, de manera tal que dicho valor quede incluido en el intervalo.

Hay una regla que se debe seguir y es que no puede haber intervalos con frecuencia absoluta simple igual a cero. Es decir, intervalos que no contengan observaciones. Si esto ocurriera, por ejemplo, con el segundo intervalo, o con el anteúltimo, se pueden juntar el primer y segundo intervalo en un intervalo único del doble de ancho que los demás; o bien, en el segundo caso, juntar el anteúltimo y el último intervalo y hacer un único (y último intervalo) del doble de ancho que los demás. Exceptuando estos casos, siempre se intenta que los intervalos sean del mismo tamaño o ancho (el tamaño o ancho del intervalo es el límite superior menos el límite inferior del intervalo).

Cuando se tienen los datos sin agrupar y hay que agruparlos, se usa la regla de Sturges para saber cuál es la cantidad de intervalos óptima con que deben agruparse los datos. Es "óptima" en el sentido de que la pérdida de información por agrupar los datos en intervalos es mínima. La regla es la siguiente:

$$K = 1 + \frac{\ln(n)}{\ln(2)}$$

Donde n es la cantidad de observaciones de la muestra. La cantidad óptima de intervalos, que rara vez dará un número entero, se puede redondear para arriba o para abajo.

Para resolver el ejercicio vamos a agregar algunas columnas al cuadro.

X (Miles de litros)	f_i (Días)	X_i	F_i	F_i (%)	f_r	$f_i.X_i$	$f_i.(X_i-\bar{X})^2$
20 - 30	1	25	1	1%	0,01	25	590,49
30 - 40	15	35	16	16%	0,15	525	3067,35
40 - 50	39	45	55	55%	0,39	1755	721,11
50 - 60	32	55	87	87%	0,32	1760	1039,68
60 - 70	11	65	98	98%	0,11	715	2711,39
70 - 80	2	75	100	100%	0,02	150	1320,98

$$\boldsymbol{n = \sum = 100} \qquad \boldsymbol{\sum = 4930} \quad \boldsymbol{\sum = 9451}$$

Al límite inferior del intervalo lo vamos a denominar a y al límite superior lo vamos a denominar b. El subíndice i se refiere al número de intervalo. La cantidad de observaciones, n, es igual a 100.
En la tercera columna calculamos la *marca de clase* que es el valor representativo del intervalo y se calcula como $X_i = \frac{a_i+b_i}{2}$.
En la cuarta columna calculamos la frecuencia absoluta acumulada en la cual se van sumando las observaciones de los intervalos anteriores, de modo que, para el primer intervalo sólo se tiene la observación del primer intervalo; para el segundo intervalo se tiene la observación del primer intervalo más las 15 observaciones del segundo; y así sucesivamente, hasta sumar las 100 observaciones de la muestra en el último intervalo.

X (Miles de litros)	f_i (Días)	X_i	F_i	F_i (%)	f_r	$f_i . X_i$	$f_i . (X_i - \bar{X})^2$
20 - 30	1	25	1	1%	0,01	25	590,49
30 - 40	15	35	16	16%	0,15	525	3067,35
40 - 50	39	45	55	55%	0,39	1755	721,11
50 - 60	32	55	87	87%	0,32	1760	1039,68
60 - 70	11	65	98	98%	0,11	715	2711,39
70 - 80	2	75	100	100%	0,02	150	1320,98
	$\boldsymbol{n = \sum = 100}$					$\boldsymbol{\sum = 4930}$	$\boldsymbol{\sum = 9451}$

En la quinta columna tenemos la frecuencia acumulada porcentual, que se calcula como:

$$F_i\ (\%) = \frac{F_i}{n} . 100\%$$

En la sexta columna tenemos la frecuencia relativa simple, que se calcula como:

$$f_r = \frac{f_i}{n}$$

La séptima columna la vamos a utilizar para calcular la media y la octava columna la vamos a utilizar para calcular la varianza.

El punto (a) nos pide calcular la media, la varianza, el coeficiente de variación e indicar si la dispersión es alta o baja.

X (Miles de litros)	f_i (Días)	X_i	F_i	F_i (%)	f_r	$f_i.X_i$	$f_i.(X_i-\bar{X})^2$
20 - 30	1	25	1	1%	0,01	25	590,49
30 - 40	15	35	16	16%	0,15	525	3067,35
40 - 50	39	45	55	55%	0,39	1755	721,11
50 - 60	32	55	87	87%	0,32	1760	1039,68
60 - 70	11	65	98	98%	0,11	715	2711,39
70 - 80	2	75	100	100%	0,02	150	1320,98
	$\boldsymbol{n = \sum = 100}$					$\boldsymbol{\sum = 4930}$	$\boldsymbol{\sum = 9451}$

Para calcular la media utilizamos la siguiente fórmula:

$$\bar{X} = \frac{1}{n}.\sum_{i=1}^{k} f_i.X_i = \frac{1}{100}.4930 = 49{,}3\ miles\ de\ litros$$

Para calcular la varianza utilizamos la siguiente fórmula:

$$S^2 = \frac{1}{n}.\sum_{i=1}^{k} f_i.(X_i-\bar{X})^2 = \frac{1}{100}.9451 = 94,51\ (miles\ de\ litros)^2$$

Podemos dividir por n en lugar de (n-1) ya que el tamaño de muestra es grande (n>30).

Para calcular el coeficiente de variación:

$$C.V. = \gamma = \frac{S}{\bar{X}} = \frac{9{,}7216}{49{,}3} = 0{,}1972$$

El desvío *S* se calcula como la raíz cuadrada positiva de S^2. Recuerden que las varianzas y los desvíos siempre son positivos.

X (Miles de litros)	f_i (Días)	X_i	F_i	F_i (%)	f_r	$f_i.X_i$	$f_i.(X_i-\bar{X})^2$
20 - 30	1	25	1	1%	0,01	25	590,49
30 - 40	15	35	16	16%	0,15	525	3067,35
40 - 50	39	45	55	55%	0,39	1755	721,11
50 - 60	32	55	87	87%	0,32	1760	1039,68
60 - 70	11	65	98	98%	0,11	715	2711,39
70 - 80	2	75	100	100%	0,02	150	1320,98

La dispersión es baja, al ser menor a 0,20 o al 20%, podríamos decir que la variable que estamos analizando puede tener una distribución Normal. Si el coeficiente de variación fuera mayor al 20%, no podríamos hacer tal aseveración.

El punto (b) nos pide calcular el porcentaje de los días que x>32 y x<51. Para esto, vamos a utilizar la fórmula que nos da la frecuencia relativa acumulada hasta el valor X. Para utilizarla, hay que identificar en qué intervalo está contenido el valor de X para el cual queremos calcular la frecuencia relativa acumulada.

$$F(X) = \frac{1}{n}\left[F_{j-1} + \frac{(X - b_{j-1})}{h_j}.f_j\right]$$

Lo que necesitamos calcular es:

$$[F(51) - F(32)].100\%$$

Donde *n* es el tamaño de la muestra; F_{j-1} es la frecuencia acumulada hasta el intervalo anterior; b_{j-1} es el límite superior del intervalo anterior; h_j es el tamaño o ancho del intervalo que estamos considerando; y f_j es la frecuencia absoluta simple del intervalo que estamos considerando.

X (Miles de litros)	f_i (Días)	X_i	F_i	F_i (%)	f_r	$f_i . X_i$	$f_i . (X_i - \bar{X})^2$
20 - 30	1	25	1	1%	0,01	25	590,49
30 - 40	15	35	16	16%	0,15	525	3067,35
40 - 50	39	45	55	55%	0,39	1755	721,11
50 - 60	32	55	87	87%	0,32	1760	1039,68
60 - 70	11	65	98	98%	0,11	715	2711,39
70 - 80	2	75	100	100%	0,02	150	1320,98

El punto (c) nos pide calcular la mediana y el valor de la variable superado con 10%, 40% y 90% de probabilidad.

A continuación se encuentra la fórmula para calcular fractiles. Para saber cuál es el intervalo que debe utilizarse tenemos que analizar en qué intervalo está contenido el valor del fractil. Para eso utilizaremos la columna de frecuencia acumulada porcentual, que nos dice cuál es el porcentaje de observaciones que se van acumulando al final de cada intervalo.

$$X_w = a_j + \frac{(n \,.\, w \; - \; F_{j-1})}{f_j} \,.\, h_j \quad con\; 0 \leq w \leq 1$$

Donde a_j es el límite inferior del intervalo que estamos considerando; n es el tamaño de la muestra; w es el fractil que queremos calcular; F_{j-1} es la frecuencia acumulada hasta el intervalo anterior; h_j es el tamaño o ancho del intervalo que estamos considerando; y f_j es la frecuencia absoluta simple del intervalo que estamos considerando.

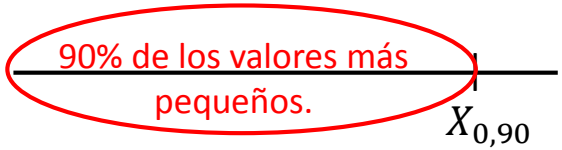
X (Miles de litros)	f_i (Días)	X_i	F_i	F_i (%)	f_r	$f_i . X_i$	$f_i . (X_i - \bar{X})^2$
20 - 30	1	25	1	1%	0,01	25	590,49
30 - 40	15	35	16	16%	0,15	525	3067,35
40 - 50	39	45	55	55%	0,39	1755	721,11
50 - 60	32	55	87	87%	0,32	1760	1039,68
60 - 70	11	65	98	98%	0,11	715	2711,39
70 - 80	2	75	100	100%	0,02	150	1320,98

Vamos a calcular primero la mediana, que es el fractil $X_{0,50}$. Tenemos que buscar en la columna amarilla aquel intervalo en donde hayamos acumulado el 50% de las observaciones más pequeñas. En el segundo intervalo, por ejemplo, acumulamos tan sólo el 16% de las observaciones más pequeñas, es decir, no hemos alcanzado todavía a acumular el 50% que corresponde a la mediana. En cambio, en el tercer intervalo ya hemos acumulado el 55% de las observaciones más pequeñas. Por lo tanto, el valor de la mediana será uno que esté comprendido entre 40 y 50 mil litros, es decir, estará incluido en el tercer intervalo, y ése es el que usaremos como referencia para hacer los cálculos.

$$\tilde{X} = X_{0,50} = 40 + \frac{(100 \,.\, 0{,}50 \; - 16)}{39} \,.\, 10 = 48{,}72 \; miles \; de \; litros$$

X (Miles de litros)	f_i (Días)	X_i	F_i	F_i (%)	f_r	$f_i . X_i$	$f_i . (X_i - \bar{X})^2$
20 - 30	1	25	1	1%	0,01	25	590,49
30 - 40	15	35	16	16%	0,15	525	3067,35
40 - 50	39	45	55	55%	0,39	1755	721,11
50 - 60	32	55	87	87%	0,32	1760	1039,68
60 - 70	11	65	98	98%	0,11	715	2711,39
70 - 80	2	75	100	100%	0,02	150	1320,98

El valor de la variable superado con 10% de probabilidad es aquel que deja a derecha el 90% de los valores más pequeños, por lo tanto, es el fractil $X_{0,90}$. Para calcularlo, debemos utilizar el quinto intervalo.



$$X_{0,90} = 60 + \frac{(100 \,.\, 0{,}90 \; - 87)}{11} \,.\, 10 = 62{,}73 \; miles \; de \; litros$$

Hagan el ejercicio de pensar de forma análoga los otros fractiles que les pide el ejercicio antes de avanzar la diapositiva siguiente.

X (Miles de litros)	f_i (Días)	X_i	F_i	F_i (%)	f_r	$f_i . X_i$	$f_i . (X_i - \bar{X})^2$
20 - 30	1	25	1	1%	0,01	25	590,49
30 - 40	15	35	16	16%	0,15	525	3067,35
40 - 50	39	45	55	55%	0,39	1755	721,11
50 - 60	32	55	87	87%	0,32	1760	1039,68
60 - 70	11	65	98	98%	0,11	715	2711,39
70 - 80	2	75	100	100%	0,02	150	1320,98

El valor de la variable superado con 40% de probabilidad es aquel que deja a derecha el 60% de los valores más pequeños, por lo tanto, es el fractil $X_{0,60}$. Para calcularlo, debemos utilizar el cuarto intervalo.

$$X_{0,60} = 50 + \frac{(100 \,.\, 0{,}60 \; - 55)}{32} \,.\, 10 = 51{,}56 \; miles \; de \; litros$$

Por último, para el fractil $X_{0,10}$ usamos el segundo intervalo:

$$X_{0,10} = 30 + \frac{(100 \,.\, 0{,}10 \; - 1)}{15} \,.\, 10 = 36 \; mil \; litros$$

X (Miles de litros)	f_i (Días)	X_i	F_i	F_i (%)	f_r	$f_i . X_i$	$f_i . (X_i - \bar{X})^2$
20 - 30	1	25	1	1%	0,01	25	590,49
30 - 40	15	35	16	16%	0,15	525	3067,35
40 - 50	39	45	55	55%	0,39	1755	721,11
50 - 60	32	55	87	87%	0,32	1760	1039,68
60 - 70	11	65	98	98%	0,11	715	2711,39
70 - 80	2	75	100	100%	0,02	150	1320,98

El punto (d) nos pide el recorrido intercuartílico y la desviación mediana. Para calcular el recorrido intercuartílico debemos calcular primero el primer cuartil y el tercer cuartil. En términos de fractiles:

$$Q_1 = X_{0,25} = 40 + \frac{(100 \,.\, 0{,}25 \; - 16)}{39} \,.\, 10 = 42{,}31 \; miles \; de \; litros$$

$$Q_3 = X_{0,75} = 50 + \frac{(100 \,.\, 0{,}75 \; - 55)}{32} \,.\, 10 = 56{,}25 \; miles \; de \; litros$$

$$f_s = Q_3 - Q_1 = 56{,}25 - 42{,}31 = 13{,}94 \; miles \; de \; litros$$

X (Miles de litros)	f_i (Días)	X_i	F_i	F_i (%)	f_r	$f_i.X_i$	$f_i.(X_i-\bar{X})^2$
20 - 30	1	25	1	1%	0,01	25	590,49
30 - 40	15	35	16	16%	0,15	525	3067,35
40 - 50	39	45	55	55%	0,39	1755	721,11
50 - 60	32	55	87	87%	0,32	1760	1039,68
60 - 70	11	65	98	98%	0,11	715	2711,39
70 - 80	2	75	100	100%	0,02	150	1320,98

Vamos a calcular la desviación mediana:

$$D.Me = \frac{\sum_{i=1}^{k}|X_i - \tilde{X}|}{n} = \frac{|25-48{,}72| + |35-48{,}72| + \cdots + |75-48{,}72|}{100}$$

Aquí podemos dividir por *n* en lugar de *(n-1)* porque el tamaño de muestra es grande y la diferencia es despreciable. En muestras pequeñas, *n*<30, se debe dividir siempre por *(n-1)*.

$$D.Me = 0{,}9\ miles\ de\ litros$$

El punto (e) nos pide el modo o moda e indicar, de acuerdo a los datos calculados, si la distribución es simétrica o no y, en caso de no serlo, qué tipo de asimetría presenta.

X (Miles de litros)	f_i (Días)	X_i	F_i	F_i (%)	f_r	$f_i . X_i$	$f_i . (X_i - \bar{X})^2$
20 - 30	1	25	1	1%	0,01	25	590,49
30 - 40	15	35	16	16%	0,15	525	3067,35
40 - 50	39	45	55	55%	0,39	1755	721,11
50 - 60	32	55	87	87%	0,32	1760	1039,68
60 - 70	11	65	98	98%	0,11	715	2711,39
70 - 80	2	75	100	100%	0,02	150	1320,98

Para calcular el modo se utiliza la fórmula que se detalla a continuación. Para identificar cuál es el intervalo modal debe considerarse que la marca de clase X_i es el valor representativo del intervalo, es decir, es 'como si' todas las observaciones del intervalo tomaran ese valor. En ese sentido, el intervalo modal es aquel que tiene la frecuencia absoluta más alta. En ese caso es el tercer intervalo.

$$Mo = a_{mo} + \frac{d_1}{d_1 + d_2} . h_j = 40 + \frac{24}{24 + 7} . 10 = 47{,}74 \ miles\ de\ litros$$

$$d_1 = f_j - f_{j-1}$$

$$d_2 = f_j - f_{j+1}$$

Donde a_{mo} es el límite inferior del intervalo modal; h_j es el tamaño o ancho del intervalo modal; f_j es la frecuencia absoluta simple del intervalo modal; f_{j-1} es la frecuencia absoluta simple del intervalo anterior al modal; y f_{j+1} es la frecuencia absoluta simple del intervalo posterior al modal.

X (Miles de litros)	f_i (Días)	X_i	F_i	F_i (%)	f_r	$f_i . X_i$	$f_i . (X_i - \bar{X})^2$
20 - 30	1	25	1	1%	0,01	25	590,49
30 - 40	15	35	16	16%	0,15	525	3067,35
40 - 50	39	45	55	55%	0,39	1755	721,11
50 - 60	32	55	87	87%	0,32	1760	1039,68
60 - 70	11	65	98	98%	0,11	715	2711,39
70 - 80	2	75	100	100%	0,02	150	1320,98

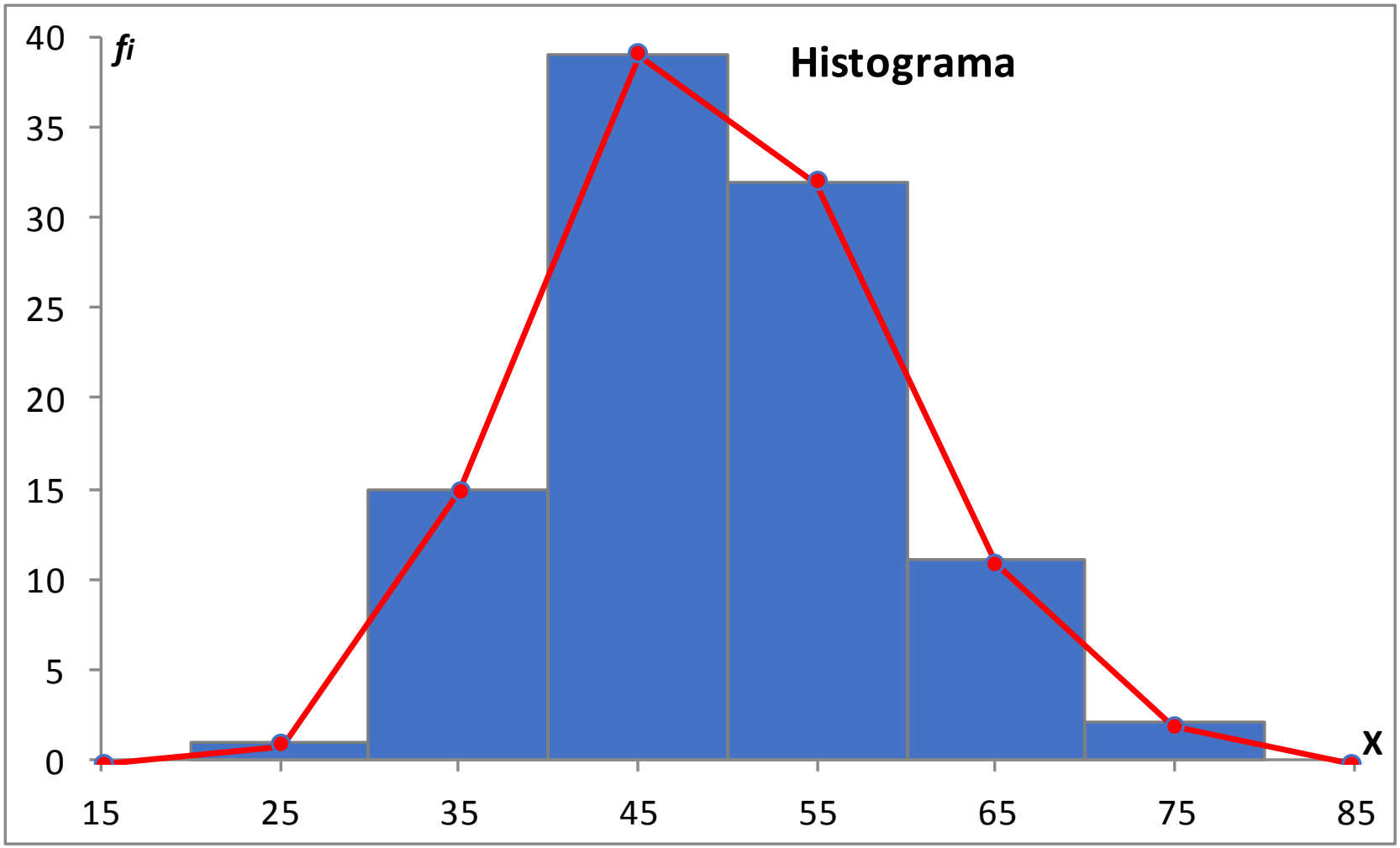
Respecto a la simetría,

$Mo = 47{,}74 \ miles\ de\ litros; \tilde{X} = 48{,}72\ miles\ de\ litros; \bar{X} = 49{,}3\ miles\ de\ litros$

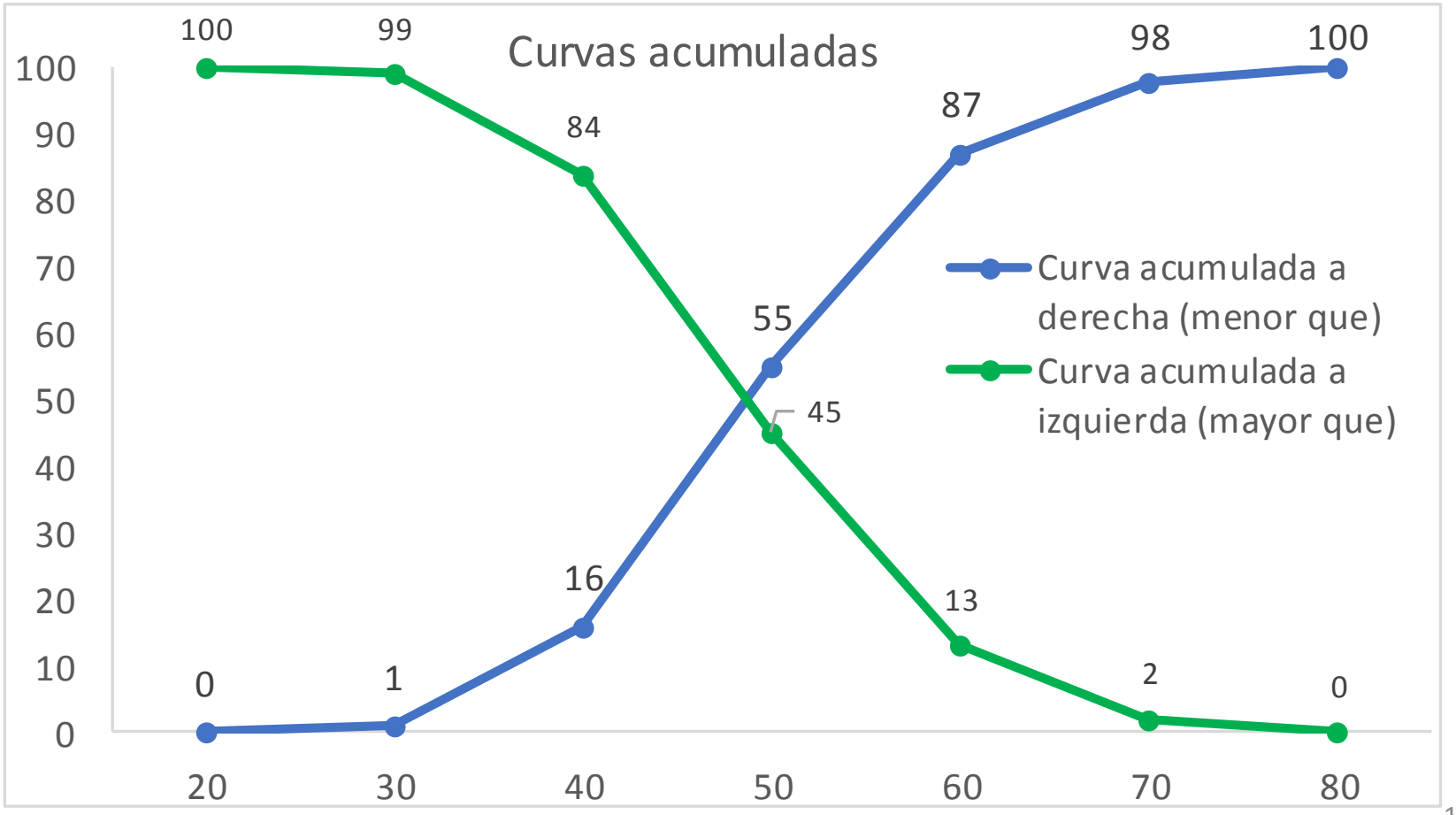
Es decir que $Mo < \tilde{X} < \bar{X}$, lo quiere decir que la forma de la distribución es asimétrica positiva.

El punto (f) nos pide el histograma y su curva correspondiente. Dibujar las curvas acumuladas.
Para dibujar el histograma usaremos las columnas sombreadas en verde.
Un histograma es una representación gráfica de una variable en forma de barras, donde la altura de cada barra (eje de ordenada) es proporcional a la frecuencia absoluta (puede ser simple o relativa) de los valores representados y el base de cada barra (eje de abscisas) es proporcional al tamaño o ancho del intervalo que representa. Sirven para obtener una "primera vista" general, o panorama, de la distribución de la población, o de la muestra, respecto a una característica, cuantitativa y continua. Las barras del histograma son todas del mismo color y no presentan espacios entre ellas.

En el eje de abscisas se muestran los límites de los intervalos o, como en este ejemplo, las marcas de clase. Si se muestran las marcas de clase, deben ubicarse en el medio de la base de su correspondiente barra. La curva u ojiva une con línea recta los puntos medios de cada intervalo en la cúspide de cada barra y, a su vez se une al eje de abscisas en los que serían los intervalos anterior y posterior a los de la tabla.



Las curvas acumuladas son dos. La curva acumulada a derecha, que se construye con la columna de frecuencia acumulada simple, sombreada en celeste; y la curva acumulada a izquierda, que no está en el cuadro, pero se puede calcular sumando las frecuencias absolutas simples desde los valores más grandes hacia los más pequeños de la variable. La información que brindan las curvas acumuladas es la cantidad de datos que se acumulan para valores menores que cierto valor *X* en el caso de la curva azul; y la cantidad de datos que se acumulan para valores mayores que cierto valor *X* en el caso de la curva verde.



Por ejemplo, si nos paramos en un valor de $X = 45$, y subimos por la línea amarilla hasta la curva acumulada derecha (azul), al movernos hacia el eje de ordenada por la línea de puntos, sabremos la cantidad de observaciones que se acumulan para valores de la variable menores que 45.
En cambio, si subimos por la línea amarilla hasta la curva acumulada a izquierda (verde), al movernos hacia el eje de ordenada por la línea de puntos, sabremos la cantidad de observaciones que se acumulan para valores de la variable mayores que 45.