HISTOGRAMA

¿Qué es?

Es una gráfica de la distribución de un conjunto de datos. Es un tipo especial de gráfica de barras, en la cual una barra va pegada a la otra, es decir no hay espacio entre las barras. Cada barra representa un subconjunto de los datos.

¿Qué muestra el histograma?

Un histograma muestra la acumulación ó tendencia, la variabilidad o dispersión y la forma de la distribución.

¿Para qué tipo de variable se usa?

Un histograma es una gráfica adecuada para representar variables continuas, aunque también se puede usar para variables discretas. Es decir, mediante un histograma se puede mostrar gráficamente la distribución de una variable cuantitativa o numérica.

Los datos se deben agrupar en intervalos de igual tamaño, llamados clases.

¿Se puede construir con los siguientes datos un histograma?

Tiempo en segundos de atención al cliente en una caja bancaria				
141	166	164	189	189
188	233	193	216	193
172	193	191	185	166
194	189	199	181	148
176	161	180	182	205
189	185	178	185	164
193	172	187	183	162
190	205	183	168	176
163	170	176	172	179
180	189	204	165	188

Los datos se refieren al tiempo en segundos de atención al cliente, son cuantitativos continuos, luego el histograma es una buena decisión de representación gráfica de estos datos.

¿Cómo se construye el histograma?

Utilizando software de aplicación estadística se puede obtener fácilmente el histograma de los datos, por lo que hoy en día nos debemos centrar más en su interpretación. Sin embargo, no está por demás hacer en forma manual el histograma de los datos.

Lo primero que se tiene que tener en cuenta es que los datos se deben agrupar en clases de igual tamaño. Teniendo en cuenta lo anterior, desarrollemos las ideas básicas de la agrupación de los datos.

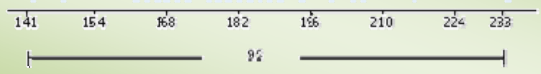
¿Cuántas clases?

Sugerencia 1:

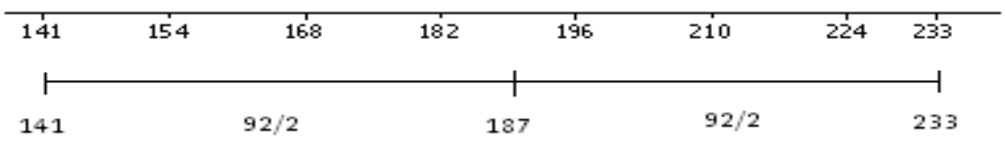
N	K
Número de datos	Número de Clientes
Menos de 50	5 – 7
50 – 100	6 – 12
100 – 200	7 – 12
Más de 250	10 - 20

Para los datos que se refieren a los tiempos de atención al cliente estos varían de 141 a 233 segundos. Si esto lo representaremos con una recta, la longitud sería de 92. A este valor de 92 se le conoce como rango y cómo puedes ver es igual a la diferencia entre el valor mayor y el valor menor. Lo podemos expresar de la siguiente manera:

$$R = y_{\max} - y_{\min}$$



Supongamos que deseamos clasificar los datos en 2 clases, lo que equivaldría a dividir la recta en 2 partes iguales; es decir, dividir 92/2 = 46. A este valor 46, se le conoce como amplitud o intervalo de clase. Entonces la primera clase comprendería los tiempos entre 141 y 187 y la segunda los tiempos entre 187 y 233, como se ve en la siguiente figura.



Ahora se nos ocurre clasificar los datos en cuatro clases, es decir, tenemos que dividir el rango entre 4; 92/4 = 23. Entonces la primera clase comprendería los tiempos entre 141 y 164, la segunda los tiempos entre 164 y 187, la tercera entre 187 y 210 y la cuarta entre 210 y 233.

¿Cuántas clases?
Sugerencia 2:

$$K = \sqrt{N}$$

Sugerencia 3:
Regla de Sturges:

$$K=1+3.322\log(N)$$

Recuerda que solo son sugerencias para realizar tu resumen. El mejor resumen es el que funcione.

Como puedes observar la amplitud de clase se obtiene dividiendo el rango entre el número de clases deseadas, entonces tenemos que:

$$amplitud\ \ de\ clase = \frac{Rango}{no.\ de\ clases\ deseadas}$$

¿Cuántas clases se deben utilizar?

Esta interrogante que es muy frecuente y que preocupa mucho al estudiante, se puede resolver fácilmente si se recuerda que el histograma es un resumen gráfico de los datos y como todo resumen no es único sino que depende de quién lo realice. Lo importante de un resumen es que resalte o ponga énfasis en lo más importante de la información.

En nuestro caso un buen resumen, es decir un buen histograma, debe de proporcionar una buena idea de la acumulación, dispersión y forma de la distribución de los datos. Por esta razón a veces es necesario hacer varios histogramas con diferente número de clases hasta obtener el que muestre eficientemente las características antes mencionadas. La sugerencia principal es que de ser posible, el número de clases se encuentre entre 5 y 20, tendiendo a un número mayor de clases según aumente el número de datos.

Atendiendo a la sugerencia anterior agrupemos ahora los datos en 6 clases, por lo tanto:

$$\text{amplitud de clase} = \frac{92}{6} = 15.33$$

Redondeando al entero mayor, tenemos que:

Amplitud de clase = 16

¿Por qué no utilizar la amplitud de clase de 15.333 que se obtuvo?

La idea es resumir la información de tal manera que podamos como se ha mencionado determinar tendencia, variabilidad y forma de la distribución de los datos.

El resumen debe ser fácil de obtener y no representar un problema adicional en el análisis de los datos. Esta es la razón fundamental por la que se redondea, para realizarlo fácilmente.

Ahora vamos a escribir las clases en una columna.

Clases
141 - 157
157 – 173
173 – 189
189 – 205
205 – 221
221 - 237

A continuación contemos el número de observaciones que pertenecen a cada clase. Antes de proceder al conteo el valor 189 ¿dónde se considera, en la tercera o cuarta clase?

¿En qué clase se consideran los datos que coinciden con los límites de clase?

Diferentes autores dan ideas diferentes de cómo solucionar esta cuestión, aquí recomendamos alguna de estas dos:

¿Por qué se redondea al valor mayor?

Por ninguna razón en especial. Mantenga en la mente que la amplitud de clase obtenida permitirá obtener el histograma y que una vez realizado este si muestra las características de la distribución será un buen instrumento; en caso contrario hay que modificar el número de clases y con esto la amplitud y volver a construir el histograma.

1. Considerar a los límites superiores en la clase siguiente. Es decir, contar el 189 en la cuarta clase. Esto equivale a leer la tercera clase, como el intervalo que incluye a los valores desde 173 a menos de 189. Entonces en la tabla se debe de indicar que el símbolo "-", se debe leer como "a menos de ", con lo que se resuelve el problema de datos coincidentes con los límites.

Si aparece un dato con valor de 189 como se menciona en el ejemplo, alguien propone lanzar una moneda. Sí cae águila lo cuentas en la tercera clase, sí cae sol lo cuentas en la cuarta clase. ¿Qué te parece esta sugerencia?

Seguramente no te gusta, pero por supuesto, que se puede utilizar ya que debes recordar que estás haciendo un resumen y que un dato pertenezca a una u otra clase no afecta mientras puedas mediante el resumen obtener ideas claras acerca de la distribución de los datos.

Clases
141 - 157
157 – 173
173 – 189
189 – 205
205 – 221
221 - 237

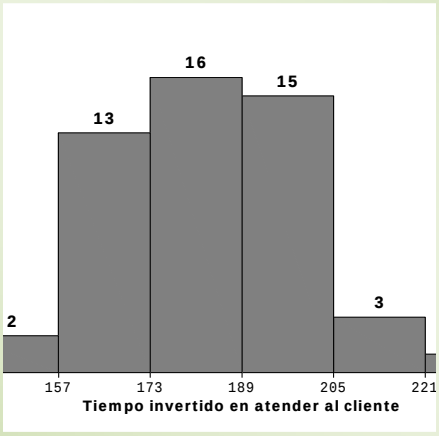
" – " Indica menos de:

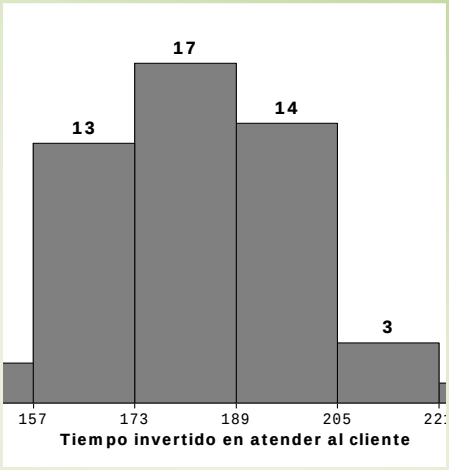
2. Se especifica un rango un poco más amplio que el rango de los datos y se introduce un decimal extra en los límites de las clases.

Para nuestro ejemplo el rango se incrementa de 92 a 93, es decir se incrementó en 1. Este aumento se reparte de forma igual entre la primera y la última clase. Es decir ½ = 0.5, entonces la primera clase iniciará en 141 - 0.5 = 140.5 y la última terminará en 237 + 0.5 = 237.5. Por lo tanto, las clases serían las siguientes:

Clases
140.5 – 157.5
157.5 – 173.5
173.5 – 189.5
189.5 – 205.5
205.5 – 221.5
221.5 – 237.5

Como se puede observar ya ningún dato coincide con los límites de clase.





Resultado: "sol", 189 se cuenta en la cuarta clase

Resultado: "águila", 189 se cuenta en la tercer clase

No hay un cambio importante de la distribución

Agrupemos ahora los datos en 6 clases siguiendo la primera recomendación. Entonces, tenemos:

Clases
141 - 157
157 – 173
173 – 189
189 – 205
205 – 221
221 - 237

" – " Indica menos de:

Ahora si podemos contar el número de observaciones que le corresponde a cada clase. A este número de observaciones se le conoce como ***frecuencia*** o ***frecuencia absoluta*** (f_i).

Clases	Frecuencia
141 - 157	2
157 – 173	13
173 – 189	17
189 – 205	14
205 – 221	3
221 - 237	1

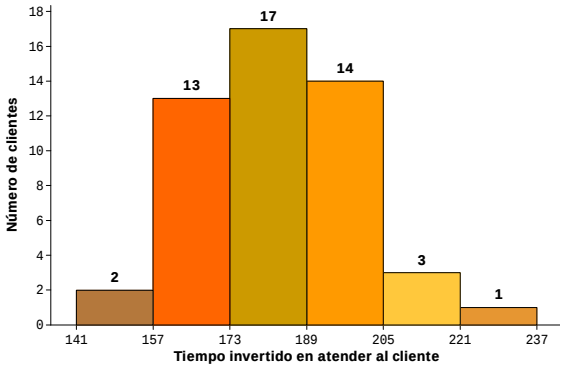
" – " Indica menos de:

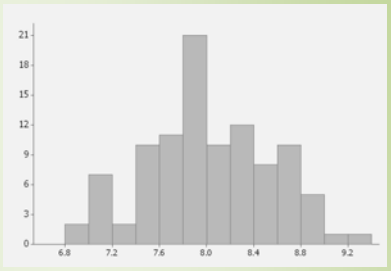
A la tabla de dos columnas, en que una de ellas indica las clases y la otra las frecuencias se le conoce como tabla de distribución de frecuencias, debido a que muestra con qué frecuencia se distribuyen los datos alrededor del valor de la variable.

Utilizar los nombres genéricos de clases y frecuencias no le indican al lector nada acerca de los datos representados en la Tabla. Por lo que en lugar de clases se debe escribir el nombre de los datos estudiados y en lugar de frecuencias el elemento donde se observaron o midieron éstos. En nuestro caso la variable es el tiempo invertido en la atención al cliente medido en segundos y las frecuencias son el número de clientes.

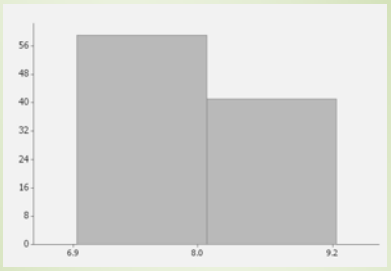
Tiempo invertido en atender al cliente	No. De clientes
141 - 157	2
157 – 173	13
173 – 189	17
189 – 205	14
205 – 221	3
221 - 237	1

" – " Indica menos de:

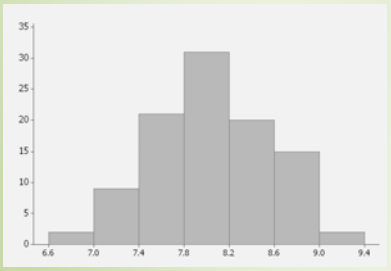




Poco resumen: muchas clases



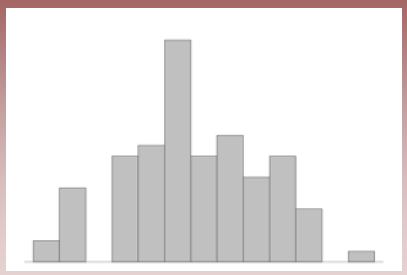
Demasiado resumen: pocas clases



Resumen Adecuado

Sí graficamos en el eje de las *X* a las clases y en eje de las *Y* a las frecuencias obtenemos el histograma de nuestros datos, que es la representación visual de la distribución de frecuencias.

¿Qué se puede hacer, si resultan espacios vacíos entre las barras?



¿Puede proporcionar la Tabla mayor información?

Se puede obtener mayor información a partir de los datos si se elaboran unas columnas adicionales en la Tabla de Distribución de Frecuencias. Si dividimos las frecuencias de cada clase entre el total de observaciones obtenemos la ***frecuencia relativa*** (***fr_i***), es decir la proporción de observaciones del total que pertenecen a cada clase.

Tiempo invertido en atender al cliente	No. De clientes	Proporción de clientes (fr_i)
141 - 157	2	2/50 = 0.04
157 – 173	13	13/50 = 0.26
173 – 189	17	17/50 = 0.34
189 – 205	14	14/50 = 0.28
205 – 221	3	3/50 = 0.06
221 - 237	1	1/50 = 0.02

" – " Indica menos de:

Entonces la primera clase indica que hay 2 clientes que tardaron de 141 a menos de 157 segundos en ser atendidos en la caja; 4% de los clientes observados tardaron de 141 a menos de 157 segundos en la caja.

Para cada clase es conveniente calcular un valor que la represente. Este valor se conoce como ***Marca de Clase*** (***X_i***) y es el punto medio de cada clase. Se obtiene de la siguiente manera:

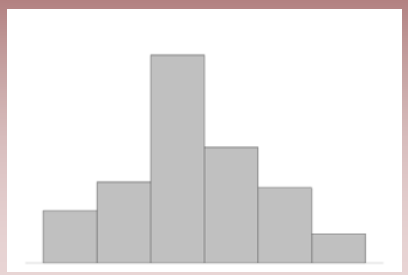
$$\frac{(Lim.\ sup. + Lim.\ inf.)_i}{2}$$

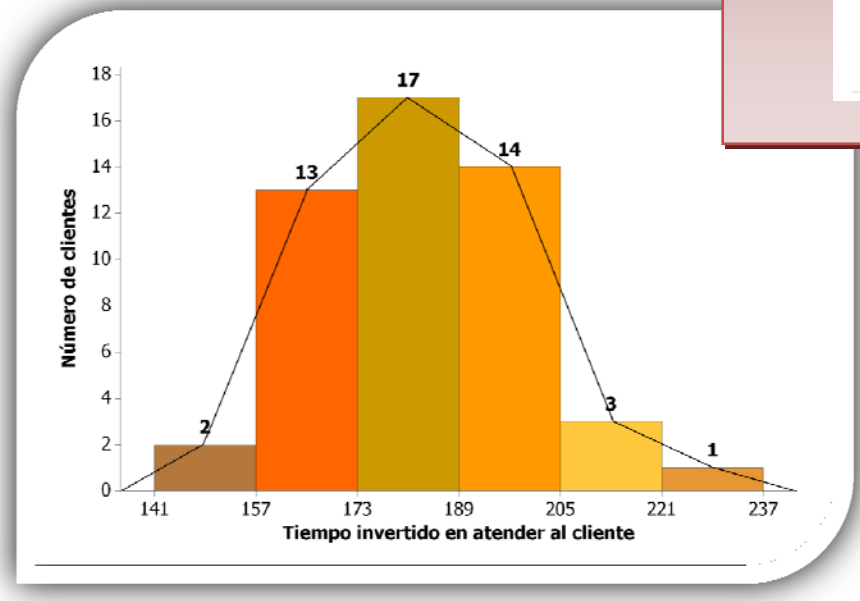
Tiempo invertido en atender al cliente	No. De clientes	Proporción de clientes (fr_i)	Marca de Clase (X_i)
141 - 157	2	2/50 = 0.04	149
157 – 173	13	13/50 = 0.26	165
173 – 189	17	17/50 = 0.34	181
189 – 205	14	14/50 = 0.28	197
205 – 221	3	3/50 = 0.06	213
221 - 237	1	1/50 = 0.02	229

" – " Indica menos de:

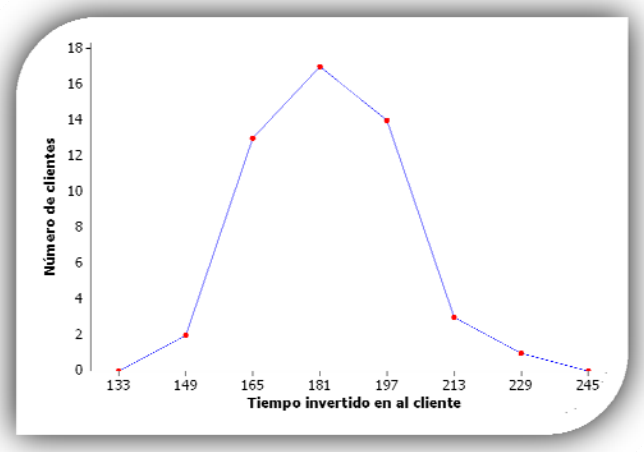
Si en el histograma colocamos las marcas de clase, estas serán el punto medio de cada barra y si unimos los puntos medios de la parte superior de cada barra obtenemos otra representación gráfica conocida como polígono de frecuencias.

Como un primer intento se debe de reducir el número de clases, con lo que se hacen más anchas las barras y se eliminan los espacios vacíos.





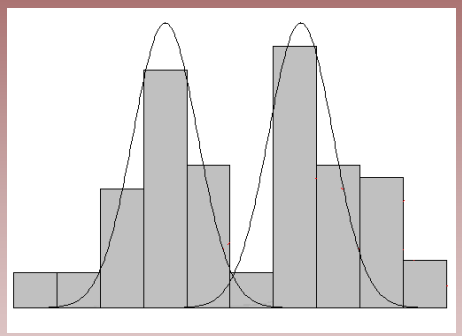
Observa que el inicio y terminación del polígono de frecuencias son la marca de clase de una clase anterior a la primera y la marca de clase de una posterior a la última, respectivamente. Es decir, el inicio es 133 y la terminación es 245 segundos.



Ahora que se ha construido el histograma y el polígono de frecuencias absolutas y relativas, procedamos a interpretarlos:

¿Y si el problema no se corrige?

Intenta aumentar el número de clases. Sí el problema no se corrige, es posible que se puedan observar dos distribuciones traslapadas o separadas



El polígono de frecuencia nos muestra de otra manera la forma de la distribución, que para nuestro ejemplo es aproximadamente simétrica.

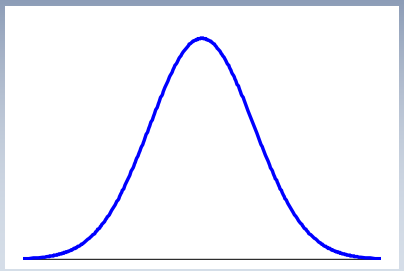
✓ La mayor acumulación o tendencia la encontramos en la tercera clase; 17 clientes, es decir el 34% de los clientes observados, tardaron entre 173 y casi 189 segundos en la caja. Es decir, el histograma muestra que la acumulación o tendencia del tiempo en que tardan los clientes en caja se encuentra entre 173 y 189 segundos.

✓ Sólo 2 clientes, es decir únicamente el 4% de los clientes observados fueron atendidos en caja en un tiempo menor a los 157 segundos.

✓ 4 clientes o sea el 8% de las personas tardaron 205 segundos o más en la caja.

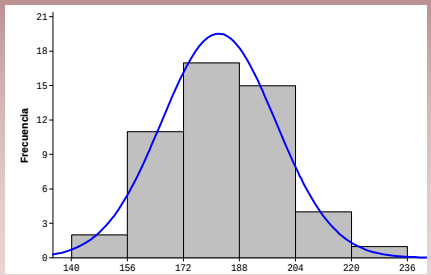
- ✓ 44 de los 50 clientes observados, es decir el 86% de los clientes estudiados tardaron entre 157 y 205 segundos. El histograma muestra que el rango o variabilidad total va de 141 a 237 segundos, y que el 88% de los clientes tardaron entre 157 y 205 segundos. Se puede decir, en otras palabras que el tiempo que tardaron el 88% de las personas varío entre 157 y 205 segundos

Una distribución simétrica con forma de campana se conoce en estadística como distribución normal



- ✓ La forma de la distribución es aproximadamente simétrica con respecto a la tercera clase y la curva suavizada nos muestra una distribución en forma aproximada a una campana que en estadística se conoce como distribución normal.

La distribución de nuestros datos es aproximadamente normal.



Como se puede observar el histograma nos muestra una fotografía reveladora de nuestros datos, que muy difícilmente podríamos apreciar a partir de ellos, si se encuentran sin agrupar.

Al observar el histograma al gerente del banco le gustaría en lugar de reducir el tiempo en caja, disminuir la variabilidad. Esto lo puede conseguir controlando variables que afectan al proceso, tal como tipo de operación realizada, número de operaciones aceptadas, hora del día en que se hizo la observación, etc. y el histograma obtenido reflejará la mejoría del proceso. Es en resumen esta gráfica sencilla un instrumento valioso para tener una buena idea acerca del comportamiento de nuestros datos.

Ojivas o Polígonos de Frecuencia Acumulada

En la tabla también se pueden adicionar columnas que indiquen el número de observaciones cuyo valor sea menor o igual que el límite superior de cada clase, lo que se conoce como ***frecuencia acumulada***.

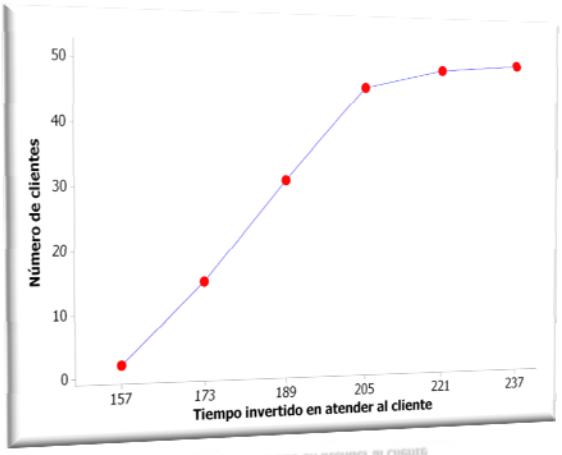
Así por ejemplo, para la tercera clase, el número de observaciones menores a 189 son 17 + 13 +2 = 32, que son las frecuencias de la tercera, segunda y primera clase respectivamente. Este valor es la *frecuencia acumulada* hasta la tercera clase.

Tiempo invertido en atender al cliente	No. De clientes	Proporción de clientes (fr_i)	Marca de Clase (X_i)	Frecuencia Acumulada (F_i)	Frecuencia Acumulada Relativa (Fr_i)
141 - 157	2	2/50 = 0.04	149	2	0.04
157 – 173	13	13/50 = 0.26	165	15	0.30
173 – 189	17	17/50 = 0.34	181	32	0.64
189 – 205	14	14/50 = 0.28	197	46	0.92
205 – 221	3	3/50 = 0.06	213	49	0.98
221 - 237	1	1/50 = 0.02	229	50	1.00

" – " Indica menos de:

La frecuencia acumulada puede ser absoluta (F_i) o relativa (Fr_i), según se utilice la frecuencia absoluta o relativa para obtenerla.

Si se grafican ahora sobre el eje de las *X* los límites superiores de clase y sobre el eje de las *Y* las frecuencias acumuladas absolutas o relativas obtenemos la gráfica conocida como Ojiva o Polígonos de frecuencia acumulada.



Algunas de las afirmaciones que podemos hacer al observar la gráfica, son las siguientes:

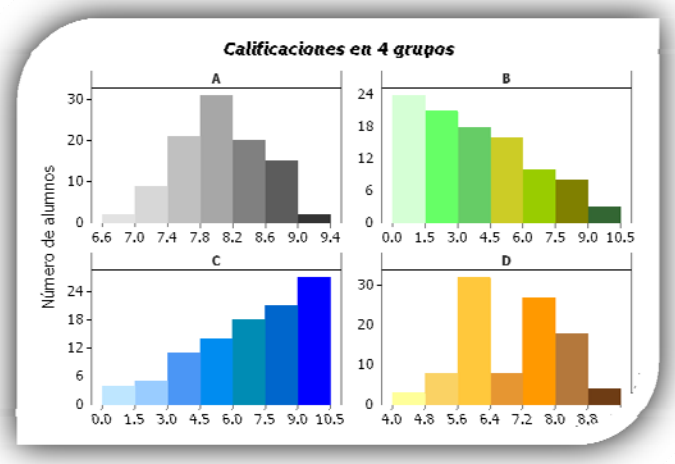
- ✓ 46 clientes tardaron menos de 205 segundos; es decir el 92% de los clientes tardaron menos de 205 segundos.
- ✓ El proceso es muy tardado sólo en el 2% de los casos; tardaron más de 221 segundos.
- ✓ El proceso es muy rápido sólo en el 4% de los casos; tardaron menos de 157 segundos
- ✓ El 64% de los clientes tardaron menos de 189 segundos.

Descripción sugerida:

- ***Indicar los valores alrededor de los cuales los datos se acumulan.***
- ***Indicar los valores extremos.***
- ***Indicar la variabilidad de los datos, (sin tomar en cuenta los valores extremos)***
- ***Indicar la forma de la distribución***

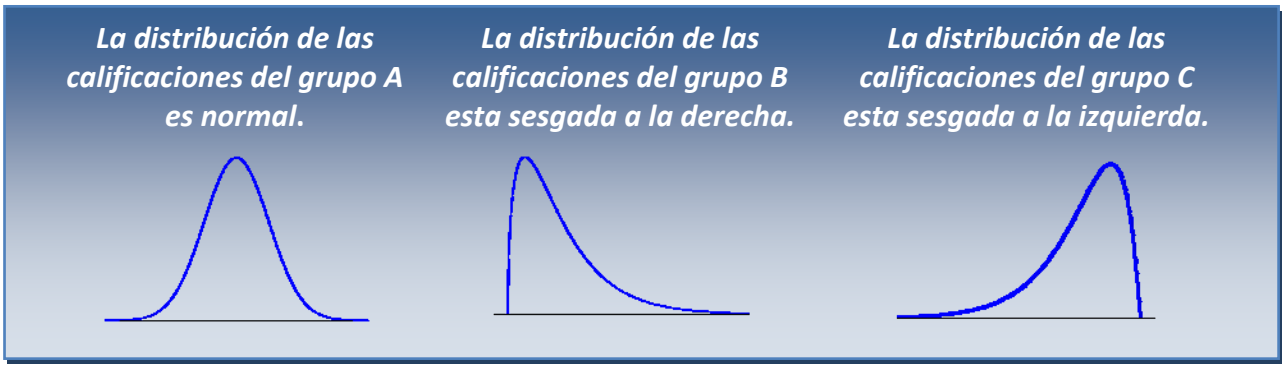
Interpretando unos histogramas

Hemos mencionado que hoy en día nos debemos centrar más en la interpretación de los histogramas que en su elaboración, ya que se cuenta con tecnología para elaborarlos. A continuación se presentan cuatro histogramas correspondientes a las calificaciones obtenidas en la asignatura de estadística en cuatro grupos diferentes. ¿Qué nos puede decir la forma de estos histogramas?



La sección A muestra que las calificaciones se distribuyen en forma aproximadamente simétrica, con respecto a la clase de mayor frecuencia, que comprenden las calificaciones de 7.8 a 8.2.; la mayoría de los estudiantes obtienen calificaciones entre 7.4 y 8.6, (acumulación o tendencia). A esta forma de la distribución se le conoce como *normal*.

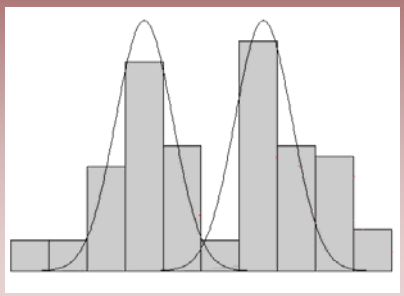
La sección B muestra una distribución con una cola larga a la derecha, es decir, muestra un *sesgo positivo*. La mayoría de los estudiantes obtuvieron calificaciones muy bajas, como lo muestra la acumulación de las calificaciones en la parte izquierda de la gráfica y muy baja densidad en la parte derecha. Esto se puede deber a varias razones, como por ejemplo, que el grupo este formado por muy malos estudiantes ó el profesor sea muy exigente ó el examen como instrumento de evaluación sea inadecuado, etc.



La sección C muestra una distribución con una cola larga a la izquierda, es decir, muestra un *sesgo negativo*.

La mayoría de los estudiantes obtuvieron calificaciones muy altas, como lo muestra la acumulación de las calificaciones en la parte derecha de la gráfica y muy baja densidad en la parte izquierda. Esto se puede deber a varias razones, como por ejemplo, que el grupo este formado por muy buenos estudiantes ó el profesor sea muy relajado ó el examen fue muy fácil, etc.

Para el grupo D se observan dos poblaciones translapadas.



La sección D muestra dos distribuciones normales traslapadas, una con acumulación entre 5.6 y 6.4 y la otra entre 7.2 y 8.0. Este se puede deber a que el grupo este conformado por estudiantes con distintos antecedentes en la asignatura (repetidores y regulares), con distintos hábitos de estudio, etc.

Examinemos ahora los histogramas que se muestran enseguida y que se refieren al peso en kilogramos de los estudiantes de un grupo de estadística. La gráfica con nombre total, muestra la distribución del peso de todos los estudiantes, mientras que las otras dos gráficas separan el peso de los hombres y de las mujeres, ¿Qué podemos apreciar en estas gráficas?

Un aspecto relevante que revelan estas gráficas es la acumulación o tendencia y la variabilidad.

La variabilidad total del peso se encuentra entre 35 y 105. La variabilidad para los hombres disminuye y se encuentra entre 55 y 105 y para las mujeres aun es menor y se encuentra entre 35 y 75 kilos. La variabilidad al formar grupos por el género disminuye, debido a que son grupos más homogéneos.

Para los hombres la tendencia se muestra entre 55 y 85 kilos, mientras que para las mujeres se encuentra entre 45 y 65.

En resumen una gráfica sencilla, como lo es el histograma, es un instrumento poderoso para obtener información del comportamiento de los datos y describir adecuadamente su distribución.