Estadística Descriptiva – Análisis de Datos

8.1 INTRODUCCION

El concepto de Estadística

El origen de la Estadística se remonta a dos tipos de actividades humanas: los juegos de azar y las necesidades de los Estados: necesidades de describir numéricamente ciudades, provincias, etc. Los juegos de azar llevaron al estudio de la probabilidad, y éste condujo al tratamiento matemático de los errores de las mediciones y a la teoría que hoy constituye la base de la estadística, mientras que la segunda actividad condujo a la estadística descriptiva: presentación de datos en tablas y gráficos, aunque en nuestros días incluye también la síntesis de ellos mediante descripciones numéricas.

El método estadístico moderno se refiere a la Inferencia estadística: ésta se relaciona con el desarrollo de métodos y técnicas para obtener, analizar e interpretar datos cuantitativos de tal manera que la confiabilidad de las conclusiones basadas en los datos pueda ser evaluada objetivamente por medio del uso de la probabilidad. La teoría de la probabilidad permite pasar de datos específicos a conclusiones generales, por eso desempeña un papel fundamental en la teoría y aplicación de la estadística.

En épocas recientes la inferencia estadística ha adquirido la importancia que antes tenía la estadística descriptiva. La inferencia estadística trata de generalizaciones basadas en muestras de datos: se aplica a problemas como estimar, mediante pruebas, la emisión promedio de contaminantes en una turbina, verificar las especificaciones de un fabricante a partir de mediciones efectuadas sobre muestras de un producto, etc.

Cuando se hace una inferencia estadística, debe procederse con cautela: debe decidirse hasta qué punto pueden hacerse generalizaciones a partir de un conjunto de datos disponibles, si las generalizaciones son razonables, o si sería preferible disponer de otro conjunto de datos...

Algunos de los problemas importantes de la inferencia estadística se refieren precisamente a la evaluación de los riesgos y las consecuencias a las que uno se expone al hacer generalizaciones. Esto incluye una estimación de la probabilidad de tomar decisiones erróneas, las posibilidades de hacer decisiones incorrectas y de obtener estimaciones no comprendidas dentro de los límites permitidos. Todos estos problemas los aborda en los últimos años la teoría de la decisión.

Podemos sintetizar lo anterior, mediante las siguientes definiciones:

> El contenido de la **estadística moderna** incluye la recopilación, presentación y caracterización de la información a fin de que auxilie tanto en el análisis de datos como en el proceso de toma de decisiones.

> Se puede definir la **estadística descriptiva** como los métodos que implican recopilación, presentación y caracterización de un conjunto de datos con el objeto de describir en forma apropiada las diversas características de dicho conjunto.

Puede considerarse la **inferencia estadística** como los métodos que hacen posible la estimación de una característica de una población, o la toma de una decisión con respecto a una población, con base únicamente en resultados muestrales.

Para clarificar esta definición, se requieren algunas definiciones:

Población (o universo): es la totalidad de elementos que se consideran.
Muestra: es un subconjunto de una población que se selecciona para su estudio.
Parámetro: es una medida que se calcula para describir una característica poblacional.
Estadística: es una medida utilizada para describir una característica muestral.

La Bioestadística

La Bioestadística se define como "la aplicación de métodos estadísticos a la solución de problemas biológicos". También se la denomina biometría.

Antecedentes históricos

Como ya hemos dicho, a causa del interés por los juegos del azar, en el siglo XVII se desarrolló la teoría matemática de la probabilidad, gracias a los aportes, principalmente, de Pascal y Fermat. Jacques Bernoulli fundamentó la moderna teoría de la probabilidad en su obra *Ars Conjectandi*, y Abraham de Moivre fue el primero en combinar la estadística de su época con la teoría de probabilidad.

Un estímulo importante para el desarrollo de la estadística lo produjo la astronomía. Se cree que el primer personaje importante en bioestadística fue el astrónomo y matemático belga Adolphe Quetelet (1796-1874), que aplicaba los métodos teóricos y prácticos de la estadística a problemas de medicina, biología y psicología. Francis Dalton (1822-1911) es denominado padre de la bioestadística. Su mayor aporte lo constituye la aplicación del análisis estadístico al análisis de la variación biológica, así como el análisis de variabilidad y su estudio de regresión y correlación en medidas biológicas. Karl Pearson (1857-1936) continuó la tradición de Galton y sentó las bases para gran parte de la estadística descriptiva y de correlación. La figura dominante en el siglo XX en estadística y biometría ha sido Ronald Fisher (1890-1962).

8.2 RECOPILACION DE DATOS

Para el especialista, la información necesaria para toda investigación está constituida por **datos.** A fin de que un análisis estadístico resulte útil en la toma de decisiones, los datos deben ser apropiados. Hay, por lo menos, tres maneras de obtener datos: i) utilizar los datos publicados por fuentes gubernamentales, industriales o particulares; ii) a través de la experimentación; iii) realizando encuestas.

8.2.1 Tipos de datos

En una investigación estadística, se manejan diversas características, a las que se denomina **variables**. Los datos son los resultados que se observan para estas variables.

Básicamente existen dos tipos de variables, que producen dos tipos de datos: **cualitativos** y **cuantitativos**. Las primeras variables producen respuestas **categóricas**, en tanto que las segundas producen respuestas **numéricas**. Por otra parte, los datos cuantitativos pueden ser **discretos** o **continuos**.

Los datos cuantitativos discretos son respuestas numéricas que surgen de un proceso de conteo, mientras que los continuos son los que surgen de un proceso de medición.

Tipos de datos		Tipos de preguntas	Respuestas
Cualitativos		¿Posee vivienda propia?	Sí --- No ---
Cuantitativos	Discretos	¿Cuántos baños posee?	-------
	Continuos	¿Cuál es la superficie cubierta?	-------

8.2.2 Tipos de escalas de medición

Todos los datos son en última instancia resultado de un proceso de medición (hasta los datos discretos pueden considerarse resultado de una medición mediante conteo). Podemos distinguir cuatro niveles de medición: escala nominal, escala ordinal, escala de intervalo, escala de razón.

8.2.2.1 Escala nominal: corresponde a los datos cualitativos, cuando se clasifican en categorías que no implican orden.

¿Es propietario de automóvil?	Sí	No
¿Cuál es su afiliación política?		

8.2.2.2 Escala ordinal: cuando los datos cualitativos se clasifican en categorías distintas en las que existe algún orden.

Rango docente	Titular	Asociado	Adjunto	JTP
Auxiliar				

8.2.2.3 Escala de intervalo: es una escala ordenada en la cual la diferencia entre las mediciones es una cantidad que tiene significado preciso. Por ejemplo, si una persona mide 1,65 m, entonces tiene 5 cm más que otra que mide 1,70 m. Estos 5 cm representan la misma diferencia entre una persona que mide 1,82 m y otra que mide 1,77m.

8.2.2.4 Escala de razón: En este caso, además de que las diferencias son significativas e iguales en todos los puntos de la escala, existe un cero real, de modo que se pueden considerar cocientes de mediciones. Por ejemplo, una persona que mide 180 cm tiene el doble de altura de otra que mide 90 cm, mientras que una temperatura de 80ºC no significa precisamente el doble de otra de 40ºC.

Temperatura (en grados C)	de intervalo
Temperatura (en grados K)	de razón
Edad	de razón
Sueldo	de razón

8.3 LOS DATOS EN BIOLOGÍA

Haremos ahora referencia a ciertos aspectos específicos de la Bioestadística, donde podremos encontrar algunas diferencias con los conceptos generales antes estudiados. Lo que sigue es, por lo tanto, la adaptación, según la mayoría de los autores, de los conceptos antes estudiados.

8.3.1 Muestras y poblaciones

La Estadística trabaja con datos. En bioestadística, los datos se basan en observaciones individuales, es decir en medidas tomadas de la mínima unidad de muestreo. La propiedad medida por las observaciones individuales es el *carácter o variable*. En estadística se usa frecuentemente el término variable, pero en bioestadística es más común carácter. En cada unidad de muestreo puede medirse más de un carácter. Así en un grupo de 10 ratones podemos medir el pH de la sangre y el número de células rojas. De esta manera obtendríamos dos muestras de 10 observaciones o una única muestra bivariada de 10 observaciones.

La totalidad de observaciones individuales sobre las cuales se hacen inferencias se denomina *población* en Estadística, y a veces *universo.* Por ejemplo, las longitudes de la cola de todos los ratones blancos del mundo; los recuentos de leucocitos de todos los varones chinos de 20 años, o puede referirse a resultados de experimentos, como las

frecuencias de los latidos cardíacos producidos en ratone por inyecciones de adrenalina. En los primeros ejemplos, la población es finita, aunque sería imposible analizar cada uno de sus elementos. En el último ejemplo, al menos en teoría, podríamos repetir el experimento un número ilimitado de veces.

Aunque la mayoría de las veces las poblaciones son finitas, son tan superiores a las muestras extraídas de ellas que de hecho pueden ser consideradas infinitas.

8.3.2 Variables

Variable es una propiedad con respecto a la cual los individuos de una muestra difieren de algún modo verificable. Las variables biológicas pueden dividirse en:

Variables medibles
    Variables continuas
    Variables discontinuas
Variables clasificables en rangos
Atributos

8.3.2.1 Variables medibles: son aquellas cuyos diferentes valores pueden expresar de forma numéricamente ordenada. Pueden ser continuas: las que al menos en teoría pueden tomar infinitos valores entre dos determinados, o discontinuas -también llamadas discretas o merísticas-: son las que tienen valores numéricos fijos, sin posibles valores intermedios. En el primer caso, tenemos por ejemplo, longitudes, áreas, pesos, temperaturas, períodos de tiempo, velocidades. En el segundo, el número de crías, el número de colonias de microorganismos, el número de plantas en un cuadrado determinado.

8.3.2.2 Variables clasificables por rangos: son las que no pueden medirse, pero si pueden ordenarse por su magnitud.

8.3.2.3 Atributos: son las variables que no pueden expresarse cuantitativamente sino cualitativamente. Son propiedades como grávida e ingrávida, muerto o vivo, macho o hembra.

8.3.3 Observación sobre las variables continuas

La mayoría de las variables continuas son aproximadas. El valor exacto de la medida individual es desconocido. Por ejemplo, al decir que una medida es de 12,4 mm queremos dar a entender que la verdadera longitud está comprendida entre 12,35 mm y 12,45 mm. Si hubiésemos podido obtener una medida de 12,43 mm, esto significaría que la verdadera medida está entre 12,435 mm y 12,435 mm. En general, la última cifra de un número aproximado debería ser siempre significativa: debería implicar que la verdadera medida está en un intervalo desde media unidad del último orden por debajo hasta media unidad por encima de la medida registrada. Esto se aplica también al cero.

8.4 MANEJO DE DATOS

8.4.1 Propiedades de los datos cuantitativos

Ya vimos que el material con que cuenta el estadístico es un conjunto de datos. Pero, la recolección de datos es sólo uno de los aspectos de la estadística descriptiva ¿cómo se pueden utilizar esos datos?

A veces los datos estadísticos obtenidos de muestras, experimentos o cualquier colección de mediciones, son tan numerosos que carecen de utilidad a menos que sean condensados.

Veremos tres propiedades de los datos cuantitativos que permiten una mejor comprensión de la información por ellos aportada.

Estas propiedades pueden ser expresadas por diversas medidas, que agrupamos de la siguiente manera:

1. de tendencia central
2. de dispersión
3. de forma

Cuando se calculan a partir de los datos muestrales, reciben el nombre de **estadísticos**, y si se los calcula a partir de la población, se denominan **parámetros**.

8.4.2 Medidas de tendencia central

Con este nombre nos referimos a valores promedios que describen todo un conjunto de datos. Se utilizan cuatro promedios, frecuentemente, como medidas de tendencia central o de posición: la **media aritmética**, la **mediana**, la **moda** y el **rango medio**.

8.4.2.1 Media aritmética: es la medida de posición utilizada con más frecuencia. Si $X_{1,}\ X_2 \ldots. X_n$ constituyen una muestra de n observaciones, la media aritmética se define de la siguiente manera:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Si bien es una de las medidas más utilizadas posee la desventaja de ser muy afectada por los valores extremos, pues en su cálculo se utilizan todas las observaciones. Puede entonces dar una imagen distorsionada de la información contenida en los datos, por lo que no siempre es la mejor medida de posición.

8.4.2.2 Mediana: Es el valor que ocupa la posición central en un conjunto de datos, ordenados en forma creciente o decreciente. Así definida, la mitad de las observaciones es menor que la mediana, mientras que la otra mitad es mayor que la mediana. Resulta apropiada cuando el conjunto de datos posee observaciones extremas.

Para calcular la mediana, primero se deben ordenar los datos. Luego se debe determinar el dato que ocupa la posición $\frac{n+1}{2}$ (cuando n es impar) o la semisuma de los valores numéricos correspondientes a las dos observaciones centrales (cuando n es par).

Por ejemplo, si los datos son: 25 12 23 28 17 15, se obtiene el arreglo ordenado 12 15 17 23 25 28, y la mediana se obtiene promediando los valores 17 y 23, resultando igual a 20.

El cálculo de la mediana se ve afectado por el número de observaciones, y no por la magnitud de los valores extremos.

8.4.2.3 Moda: es el valor de un conjunto de datos que aparece con mayor frecuencia. Tampoco depende de los valores extremos, pero es más variable que las otras medidas de posición para las distintas muestras.

Cuando no hay ningún valor con frecuencia mayor, la distribución carece de moda. También se puede dar el caso de una distribución con más de una moda.

8.4.2.4 Rango medio: Es la media de las observaciones mayor y menor. Como intervienen solamente estas observaciones, si hay valores extremos, se distorsiona como medida de posición, pero frecuentemente ofrece una valor adecuado rápido y sencillo para resumir un conjunto de datos (cuando puede suponerse que no existen valores extremos).

8.4.3 Medidas de dispersión o de variabilidad

Las medidas de dispersión permiten conocer la variabilidad de un conjunto de datos. Estudiaremos las siguientes: **rango**, **varianza**, **desviación estándar** y **coeficiente de variación**.

8.4.3.1 Rango: Es la diferencia entre las observaciones mayor y menor. Si bien es una medida de dispersión simple, posee el inconveniente de que no toma en consideración la forma en que se distribuyen los datos entre los valores más pequeños y más grandes.

8.4.3.2 Varianza y desviación estándar: Una medida de variabilidad podría obtenerse a partir de la dispersión de cada una de las observaciones con respecto a algún valor particular, por ejemplo la media. Pero, como es fácil de comprobar, la suma de los desvíos de cada valor respecto a la media es siempre cero, es decir

$$\sum_{i=1}^{n} (X_i - \overline{X}) = 0$$

Por eso consideraremos una medida obtenida "promediando" los cuadrados de los desvíos, la **varianza muestral**:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n - 1} \qquad ([1])$$

El **desvío estándar muestral** es la raíz cuadrada de la varianza: $S = \sqrt{S^2}$

A partir de la fórmula puede observarse que ni la varianza ni el desvío estándar, pueden ser negativos, y hay un único caso en que pueden ser nulos: cuando todos los valores de la muestra son iguales.

La varianza y el desvío estándar miden la dispersión "promedio" en torno a la media, es decir cómo fluctúan las observaciones mayores por encima de la media y cómo se distribuyen las observaciones menores por debajo de ella.

A pesar de que la varianza posee ciertas propiedades matemáticas útiles, está expresada en unidades cuadradas (dólares cuadrados, etc.) lo que le hace perder significado. El desvío estándar no posee este inconveniente, y está expresado en las unidades originales.

8.4.3.3 Coeficiente de variación: Es una cantidad que mide la dispersión de los datos con respecto a la media: $CV = \frac{S}{\overline{X}}100$

El coeficiente de variación es una medida relativa. No se expresa en término de las unidades utilizadas, sino como porcentaje. Es útil cuando se compara la variabilidad de dos conjuntos de datos, o más, expresados en diferentes unidades. También es útil cuando se comparan dos o más distribuciones de datos expresados en la misma unidad, pero que difieren en tal forma que una comparación directa de los desvíos estándar no resulta útil.

8.4.4 Forma

Las medidas de **forma** describen la manera en que se distribuyen los datos. Una distribución de datos puede ser **simétrica** o no. Cuando no lo es, se denomina asimétrica o **sesgada**.

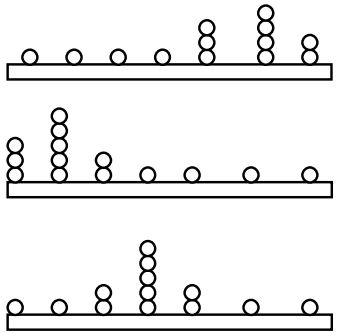
Para indicar la forma se comparan la media y la mediana de la distribución. Si las medidas son iguales se considera que los datos son simétricos, o que la distribución tiene sesgo cero. Cuando la media es mayor que la mediana, el sesgo es positivo o la asimetría es a la derecha, en cambio cuando la media es menor que la mediana, se dice que la distribución tiene sesgo negativo o asimetría a la izquierda.

El sesgo positivo ocurre cuando la media se ve aumentada por algunos valores extraordinariamente grandes; el sesgo negativo se da cuando la media se ve afectada por algunos valores extremadamente pequeños.

---

[1] La razón de utilizar (n - 1) en lugar de n se comprenderá más adelante, aunque si el tamaño de la muestra es grande, el uso de *n* o *(n - 1)* no produce diferencias significativas. En general, utilizaremos el denominador *(n - 1)* cuando se trata de la **varianza muestral**, y n para la **varianza de la población**.

El siguiente gráfico muestra lo que ocurre en cada situación: en cada una de las tres escalas se describe la distribución de un grupo de alumnos según las notas obtenidas (entre 40 y 100 puntos)



El primero de lo gráficos corresponde a un conjunto sesgado a la izquierda, donde la media es menor que la mediana, ya que hay pocas calificaciones bajas. En el segundo, los datos están sesgados a la derecha. La media es mayor que la mediana, ya que hay pocas calificaciones altas. El último gráfico muestra una distribución simétrica, con un desempeño que podríamos catalogar como "normal". La media y la mediana son iguales, al igual que la moda y el rango medio.

Cuantitativamente, la asimetría puede determinarse por medio de la siguiente ecuación:

$$As = \frac{3(\overline{X} - Med)}{S}$$

8.5 TRATAMIENTO DE DATOS AGRUPADOS

8.5.1 La distribución de frecuencias

Muchas veces es necesario manejar un gran número de datos, y en ese caso puede demandar mucho esfuerzo el cálculo de las medidas anteriores. Como regla práctica, cuando el conjunto contenga 20 o más observaciones, la mejor manera de examinar estos datos es presentarlos en forma resumida, elaborando tablas y gráficas.

La distribución de frecuencia es una tabla en la que se disponen los datos divididos en grupos y ordenados numéricamente, mostrando también el número de elementos de cada grupo o clase. Se sacrifica así parte de la información contenida en los datos: en lugar de conocer el valor exacto de cada uno, sólo sabemos que pertenece a una clase determinada. Pero lo que se pierde en información se compensa en legibilidad, ya que de esta forma se destacan características importantes de los datos.

El primer paso para construir una distribución de frecuencia es decidir el número de clases a utilizar y los límites de cada clase. En general el número de clases depende del número de observaciones, pero tiene poca utilidad una distribución con menos de 5 clases o con más de 15 clases. También depende del rango de los datos.

Es recomendable que todas las clases tengan la misma amplitud. Para determinar el tamaño de cada clase se divide el rango entre el número de clases que se desean.

Luego se necesita establecer límite para cada una de las clases, evitando que se superpongan.

Para ejemplificar, consideremos las siguientes 80 mediciones de la emisión (en toneladas) de óxido de azufre de una planta industrial:

31.8; 26.4; 17.3; 11.2; 23.9; 24.8; 13.9; 9.0; 13.2; 18.7; 25.9; 10.5; 22.7; 9.8; 6.2; 14.7; 26.1; 12.8; 17.6; 28.6; 23.7; 17.5; 15.9; 27.5; 26.8; 22.7; 18.0; 20.5; 11.0; 20.9; 15.5; 19.4; 16.7; 10.7; 18.1; 17.9; 19.1; 15.2; 22.9; 26.6; 20.4; 21.4; 19.2; 21.6; 16.9; 19.0; 9.4; 20.1; 18.5; 23.0; 24.6; 20.1; 16.2; 18.0; 7.7; 13.5; 23.5; 14.5; 28.5; 24.1; 14.4; 29.6; 19.4; 17.0; 20.8; 24.3; 22.5; 24.6; 18.4; 18.1

La observación más grande es 31.8, mientras que la más pequeña es 6.2, por lo tanto la amplitud de la distribución, o rango, es de 25.6. Podríamos entonces elegir 6 clases que tuvieran los límites: 5.0 - 9.9; 10.0 - 14.9, etc. O también las siete clases: 5.0 - 8.9; 9.0 - 12.9, etc. O las nueve clases: 5.0 - 7.9; 8.0 - 10.9, etc. Notemos que en todos los casos las clases no se traslapan, incluyen todos los casos y tienen la misma longitud.

Existe otra posibilidad: considerar los intervalos: 5.0 - 9.0; 9.0 - 13.0, etc. En este caso se presentan ambigüedades, ya que el valor 9 podría pertenecer a la primera o a la segunda clase. Para evitar esta dificultad, podemos hacer que la primera clase vaya de 4.95 a 8.95, la segunda de 8.95 a 12.95, etc. Estas son las fronteras de clase, y a pesar de las clases se traslapan, no hay ambigüedades, ya que estas fronteras son valores "imposibles" para los datos. En general, empleamos los fronteras de clase y no los límites para indicar que los datos son continuos.

Para nuestra ejemplo, tendremos:

límites de clase	etiqueta	frecuencia
4.95 - 8.95	///	3
8.95 - 12.95	~~////~~ ~~////~~	10
12.95 - 16.95	~~////~~ ~~////~~ ////	14
16.95 - 20.95	~~////~~ ~~////~~ ~~////~~ ~~////~~ ~~////~~	25
20.95 - 24.95	~~////~~ ~~////~~ ~~////~~ //	17
24.95 - 28.95	~~////~~ ////	9
28.95 - 32.95	//	2
Total		80

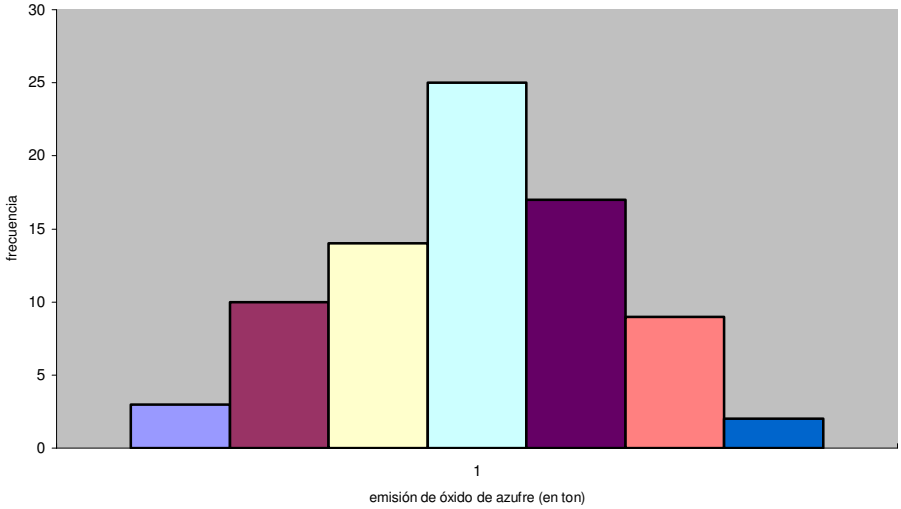
Una vez que los datos han sido ordenados, pierden su identidad, pues ya no se conoce su valor exacto. Esto puede evitarse de algún modo considerando el punto medio de cada intervalo, llamado **marca de clase.** La marca de clase de cada intervalo se obtiene mediante la semisuma de sus fronteras.

8.5.2 Gráficas de las distribuciones de frecuencias

8.5.2.1 Histogramas

Representar una distribución de frecuencias hace más evidente sus propiedades. La forma más común de representar una distribución de frecuencias es el **histograma**, que se construye con rectángulos adyacentes de alturas proporcionales a las frecuencias y cuyas bases se extienden entre las fronteras de clases sucesivas.

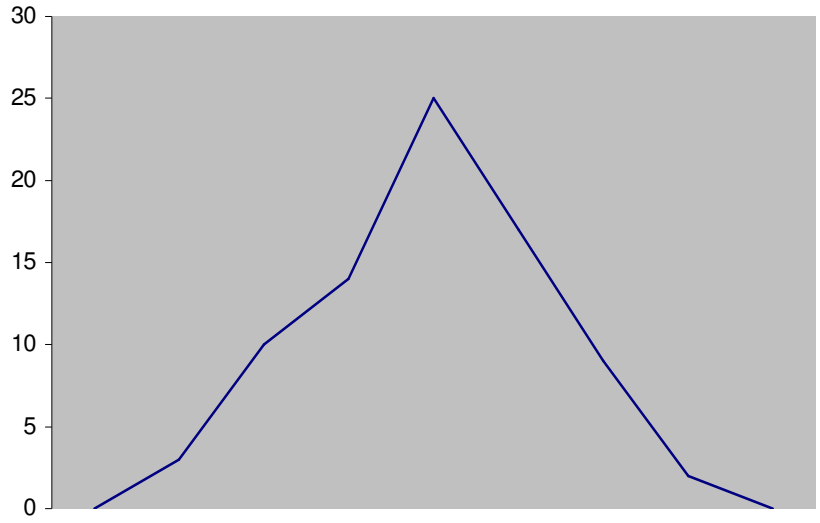
Para los datos anteriores, obtenemos el siguiente histograma:



Otras gráficas similares a los histogramas son los **diagramas de barras**, donde las alturas de los rectángulos representan las frecuencias, pero no se pretende fijar una escala horizontal continua.

8.5.2.2 Polígono de frecuencias

Una forma optativa de representar las distribuciones de frecuencias es el **polígono de frecuencias**. En él las frecuencias de cada clase son graficadas en sobre la marca de clase, y los puntos sucesivos se unen mediante segmentos de recta, después de haber agregado clases con frecuencia cero en los extremos de la distribución.



8.5.2.3 Distribuciones acumuladas

Existen formas alternativas de agrupar los datos: son las distribuciones **acumuladas "menor que" y "mayor que"**.

Para ello podríamos convertir la distribución de modo que muestre cuántas observaciones son menores que 4.95, menor que 8.95, etc.

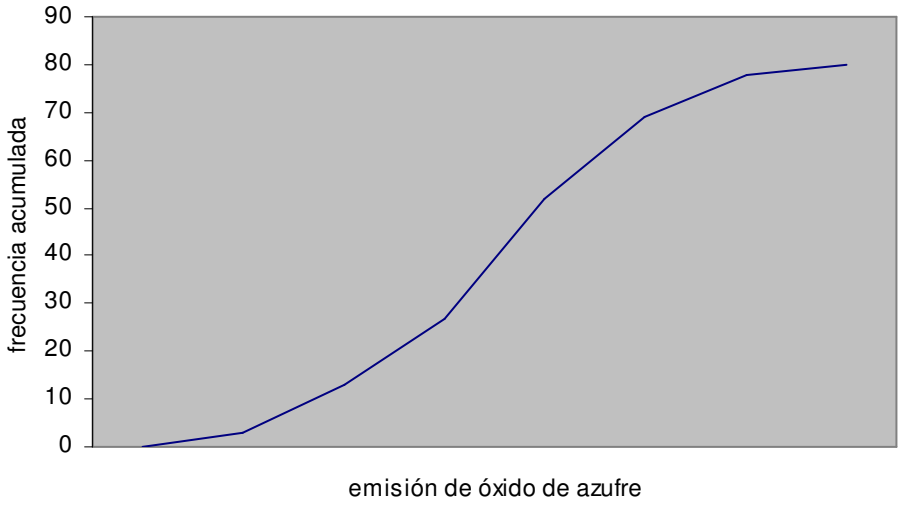
ton. de óx. de azufre	frec. acumulada
menos de 4.95	0
menos de 8.95	3
menos de 12.95	13
menos de 16.95	27
menos de 20.95	52
menos de 24.95	69
menos de 28,95	78
menos de 32.95	80

En lugar de "menos de 4.95", podríamos haber utilizado "menos de 5.0" o "menos de 4.9", etc.

Las distribuciones del tipo "mayor que" se construyen de la misma forma, pero en la práctica la más utilizada es la anterior. Con el fin de comparar distribuciones de frecuencias puede ser ventajoso convertirlas en distribuciones porcentuales. Puede hacerse lo mismo con las distribuciones acumuladas, obteniendo las **distribuciones porcentuales acumuladas**.

Las distribuciones acumuladas se presentan por lo general en forma de **ojivas**, las cuales son similares a los polígonos de frecuencias, excepto en que graficamos las

frecuencias sobre las fronteras en lugar de graficarlas sobre la marca de clase. Los puntos obtenidos se unen mediante segmentos de recta, obteniendo la gráfica de la distribución "menor que".



8.5.3 Cálculo de las medidas descriptivas para una distribución de frecuencias

Cuando los datos se presentan por medio de una distribución de frecuencias, perdemos la información acerca del valor de cada uno de ellos, ya que se encuentran reunidos en clases. En este caso sustituimos cada uno de los valores de un intervalo por la marca de clase.

Si llamamos X_i al punto medio de cada intervalo, y f_i a la frecuencia del intervalo, obtenemos las siguientes fórmulas para el cálculo de las diversas medidas descriptivas:

Media aritmética: $\overline{X} = \dfrac{\sum_{i=1}^{k} X_i f_i}{\sum_{i=1}^{k} f_i}$ **Varianza:** $S^2 = \dfrac{\sum_{i=1}^{k} (X_i - \overline{X})^2 f_i}{\sum_{i=1}^{k} f_i}$

(En estas fórmulas, k representa el número de intervalos o clases de la distribución)

Si bien es posible obtener expresiones para las demás medidas, sólo nos ocuparemos de las dos mencionadas más arriba.

En el caso de la mediana, su obtención puede hacerse en forma aproximada a partir del gráfico de la distribución acumulada, obteniendo del mismo el valor de la variable que corresponde a una frecuencia acumulada de $\frac{n}{2}$.