Características:

a) Que proporcionen la máxima información contenida en los datos en forma rápida y fácil de visualizar.
b) Que posean sencillez operativa.
c) Que permitan presentar los datos de una manera estética.

2.2.1 Histograma y tabla de frecuencias

Una de las características de los datos experimentales es su variabilidad. Independientemente de los esfuerzos que se hagan, no es posible duplicar con exactitud las condiciones experimentales o los resultados, en especial, en experimentos en los que existen variables que no se pueden controlar directamente por el investigador. Una de las razones por la que se recopilan datos, es determinar el promedio de la variable de estudio bajo una condición determinada, así como la variabilidad provocada por factores no-controlables entre experimentos, o por errores en las mediciones. Al revisar los datos obtenidos en las mediciones de diámetros de unas tuercas no se logra mucha información. Sin embargo, si se presentan de otra forma, por ejemplo en un histograma, se podrá entender e interpretar su comportamiento. Ahora bien, un histograma es un diagrama de barras sin espaciamiento entre ellas, construida colocando en el eje vertical a las frecuencias absolutas o relativas y en el eje horizontal a los límites de clase de una tabla de frecuencias con que se observan los datos (la frecuencia representa el número de veces que ocurre un evento, por ejemplo: nacimientos, días que llueve, etcétera), además identifica el centro y la variabilidad de los datos.

Pasos a seguir para construir un histograma:

1) Calcular el intervalo de los datos (Amplitud).

   Intervalo=Observación más grande menos la Observación más pequeña

2) Dividir el intervalo en clases de igual anchura. El número de clases es arbitrario, pero se obtiene una mejor descripción gráfica si se utiliza pocas clases cuando el número de datos es pequeño y un mayor número de clases cuando el conjunto de datos es grande. La mayoría de

los practicantes recomiendan entre 5 y 20 clases, un número menor sacrifica muchos detalles y uno más grande retiene demasiados. A continuación se presenta una guía para saber cuántas clases se deben formar:

Número de observaciones de un conjunto de datos	Número de clases,
Menos de 50	5 a 7
de 50 a 100	6 a 10
De 100 a 250	7 a 12
Más de 250	10 a 20

Algunos estadísticos gustan de usar la regla de Sturges para determinar el número de clases deseable.

$$k = 1 + 3.3 \log n$$

Donde k = Número de clases y n es igual al tamaño del conjunto de datos.

En muestras muy grandes se maneja calcular la raíz cuadrada del número total de observaciones. $\left(\sqrt{n}\right)$

La frontera de la clase más baja (o primera) deberá estar situada por debajo de la medición más pequeña. La anchura aproximada de las clases es igual al intervalo entre el número de clases.

$$\text{Anchura} = \frac{\text{intervalo}}{\text{No. de Clases}}$$

Al estar construyendo los límites de las clases, el problema que se presenta cuando los valores coinciden se puede evitar especificando un rango un poco más amplio que el rango de los datos e introduciendo un decimal extra en los limites de las clases.

3) Para cada clase, contar el número de observaciones que caen en dicha clase. Este número es la frecuencia de clase.

4) Calcular la frecuencia relativa de cada clase.

$$\text{Frecuencia Relativa} = \frac{\text{Frecuencia de Clase}}{\text{No. Total de Observaciones}}$$

5) El histograma es en esencia una gráfica de barras en las que las categorías son clases. En un histograma de frecuencia, la altura de las barras esta determinada por la frecuencia de clases. De forma similar las barras están determinadas por la frecuencia relativa de las clases.

Ejemplo 2.16: Se midió el diámetro de 84 tuercas y se registraron de la siguiente manera:

3.0	3.2	3.4	3.5	3.5	3.6	3.8	3.9	4.0	4.0	4.1	4.2	4.3	4.6
3.0	3.3	3.4	3.5	3.5	3.8	3.8	3.9	4.0	4.0	4.1	4.2	4.4	4.9
3.1	3.3	3.4	3.5	3.6	3.8	3.8	3.9	4.0	4.0	4.2	4.2	4.4	4.9
3.2	3.4	3.4	3.5	3.6	3.8	3.9	3.9	4.0	4.0	4.2	4.2	4.4	4.9
3.2	3.4	3.4	3.5	3.6	3.8	3.9	3.9	4.0	4.0	4.2	4.2	4.4	5.0
3.2	3.4	3.5	3.5	3.6	3.8	3.9	3.9	4.0	4.0	4.2	4.3	4.4	5.0

1) Intervalo (Amplitud) $5.0 - 3.0 = 2.0$

2) Debido a que existen 84 datos, se sabe que debemos formar de 6 a 10 clases según la guía para determinar el número de clases mostrada anteriormente. Elegimos arbitrariamente 7 clases.

Para determinar los límites de las clases, se amplía el rango, el cual se divide por el número de clases. En este caso, el rango se incrementa de 2.0 a 2.1, o sea, un aumento de 0.1, que se debe distribuir de manera equitativa entre la primera clase y la última, por lo que el límite inferior de la primera clase iniciará con 2.95 y el límite superior de la última será 5.05. Es recomendable que los tamaños de las clases sean iguales.

$$\text{La anchura de las clases } \frac{2.1}{7} = 0.3$$

En este ejemplo si se forman 7 clases de amplitud 0.3, se satisfacerá dicha recomendación. Con base en lo anterior, se proponen los siguientes límites para los intervalos de clase:

3) y 4) Cálculo de la Frecuencia de clases y Frecuencia relativa.

Clase	Intervalo de clase	Frecuencia de Clase	Frecuencia de clase relativa
1	$2.95 \le x < 3.25$	7	0.08333
2	$3.25 \le x < 3.55$	19	0.22619
3	$3.55 \le x < 3.85$	13	0.15476
4	$3.85 \le x < 4.15$	23	0.27380
5	$4.15 \le x < 4.45$	16	0.19047
6	$4.45 \le x < 4.75$	1	0.01190
7	$4.75 \le x < 5.05$	5	0.05952
	Total	84	0.99997

El histograma de Frecuencia de Clase quedaría:

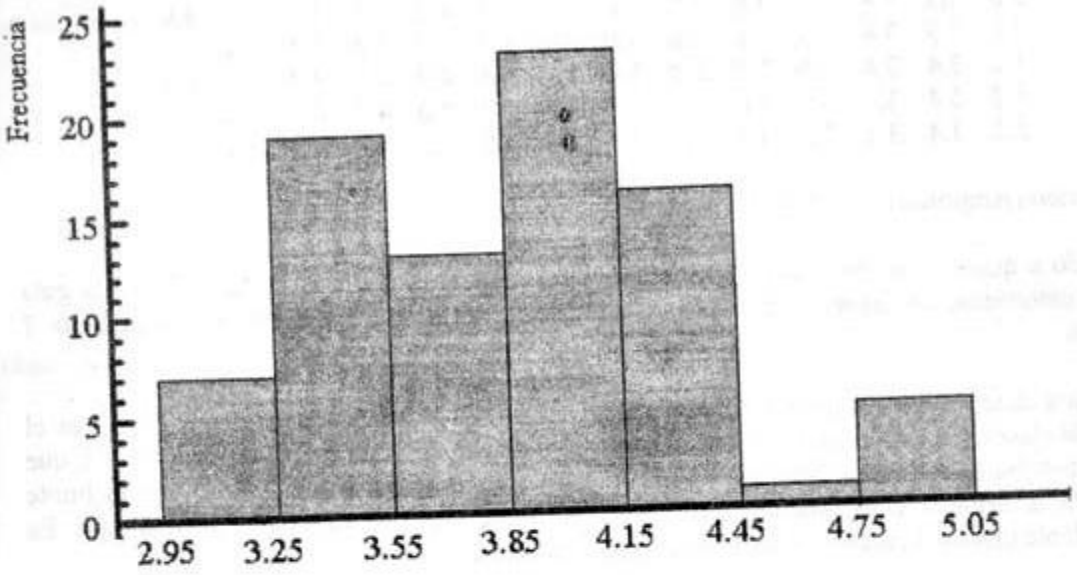

Fig. 2.14 Histograma para los datos del diámetro de las tuercas

Pasos para la construcción de una Tabla de Frecuencias:

1) Elección de número de clases.
2) Cálculo del intervalo de clases (Amplitud).
3) Elección del límite inferior de la primera o superior de la última clase y cálculo de los límites de las demás clases.
4) Cálculo de los valores medios de clase (V_i) sumando las fronteras de cada clase y dividiéndolas entre dos.
5) Cálculo de las Frecuencias Absolutas de clase (f_i).
6) Cálculo de las Frecuencias Relativas de clase (p_i).
7) Cálculo de las Frecuencias acumuladas relativas (F_i).
8) Cálculo de las frecuencias acumuladas absolutas (f_i acumulada).

Clase o intervalo	(V_i)	(f_i)	(p_i)	(F_i)	(f_i acumulada).
$2.95 \le x < 3.25$	0.78	7	0.08333	0.08333	7
$3.25 \le x < 3.55$	0.94	19	0.22619	0.30952	26
$3.55 \le x < 3.85$	1.10	13	0.15476	0.46428	39
$3.85 \le x < 4.15$	1.26	23	0.27380	0.73808	62
$4.15 \le x < 4.45$	1.42	16	0.19047	0.92855	78
$4.45 \le x < 4.75$	1.58	1	0.01190	0.94045	79
$4.75 \le x < 5.05$	1.74	5	0.05952	0.99997	84

Tabla 2.1 Tabla de frecuencias para los diámetros de las 84 tuercas

La tabla de frecuencias además de poseer sencillez operativa (son muy fáciles de construir), constituyen una forma de presentar los datos de manera tal que la información que contiene, es de fácil y rápida apreciación. Así en el ejemplo anterior se puede observar en forma inmediata:

a) Los valores numéricos de las mediciones de los diámetros de las tuercas que más frecuentemente se presentaron se encuentran entre $13.85 \leq x < 4.15$, 23 del total de las 84 observaciones pertenecen a este intervalo.

b) Aproximadamente el 92% de los diámetros de las tuercas son de menos de 4.45.

Sin embargo en las tablas de frecuencias existen las siguientes desventajas:

1) Pérdida de información al presentar las observaciones en intervalos sin especificar cuáles son los datos que pertenecen a ellos. Así, por ejemplo, no es posible saber directamente de la tabla cuales son los valores numéricos de los datos que pertenecen al intervalo $3.55 \leq x < 3.85$.

2) El hecho de haber escogido como valor representativo de la clase el valor medio implica que se supone que los datos que pertenecen a la clase tienen en promedio un valor cercano a este. Si ésta suposición no es correcta la información que nos proporciona (V_i) es poco confiable.

3) Dado que el número de clases y la anchura de las mismas se elige en forma arbitraria, no existe una representación única de los datos en las tablas de frecuencias.

2.2.2. Diagrama de Caja.

El diagrama de caja es una representación visual que describe al mismo tiempo varias características importantes de un conjunto de datos, tales como el centro, la dispersión, la desviación de la simetría y la identificación de observaciones que se alejan de manera poco usual del resto de los datos (este tipo de observaciones se conocen como "valores atípicos").

El diagrama de caja presenta los tres cuartiles, y los valores mínimo y máximo de los datos sobre un rectángulo, alineado horizontal o verticalmente.

El rectángulo delimita el rango intercuartílico (IQR) con la arista izquierda (o inferior) ubicada en el primer cuartil Q_1 y la arista derecha (o superior) en el tercer cuartil Q_3.

Se dibuja una línea a través del rectángulo en la posición que corresponde al segundo cuartil ($Q_2 = \tilde{X}$) (mediana).