1. Introducción

1.1 Estadística en ingeniería

Entre las tareas de las cuales son responsables los ingenieros hoy en día se pueden mencionar el diseño y desarrollo de productos de todo tipo, generación y distribución de energía, desarrollo de nuevos materiales y equipos, construcción y mantenimiento de obras de infraestructura, avances en comunicaciones y en informática, etc.

En cualquiera de los procesos en los que ellos intervienen se presentan fuentes de variabilidad que afectan el comportamiento de características de interés, ya sean del proceso en sí o de alguna de sus salidas. Esta variabilidad provoca incertidumbre, por lo que no se pueden predecir con exactitud los valores de estas características en una unidad o en una ejecución del proceso en particular. En consecuencia, la mayoría de las veces su trabajo está envuelto en una nube de variabilidad e incertidumbre, en medio de la cual, deben tomar decisiones con riesgos mínimos. Con respecto a esta toma de decisiones y a los problemas que pueden surgir, la Estadística aporta conceptos y procedimientos para su resolución y, además, para realizar investigaciones que involucren datos sobre las características de interés.

1.2 Los procesos y la variabilidad

En general, a un **proceso** se lo puede definir como un conjunto de fases sucesivas de un fenómeno natural o de una operación artificial o también, como una secuencia de pasos (etapas), ordenados con cierta lógica, para obtener un resultado.

De estas definiciones se desprende que en todo proceso hay un estado inicial (Entrada) y uno final (Salida) y que entre ambos se produce alguna transformación o cambio (Figura 1.1)

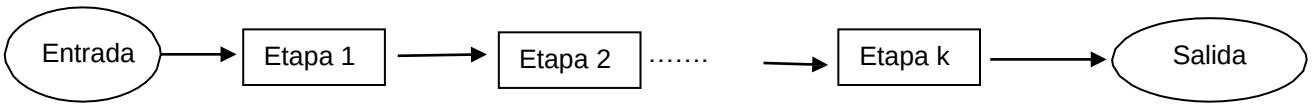


Figura 1.1. El concepto de proceso

Considere, a modo de ejemplo, al proceso que ocurre desde que se hace click en un botón de una aplicación hasta que se obtiene la respuesta a esa acción (se ingresa a una sección, se abre una ventana, etc). En estos casos se desarrolla un conjunto definido de etapas que culminan brindando la respuesta correspondiente. Por parte de quien desarrolla la aplicación, se busca que el producto cumpla ciertos requerimientos, por ejemplo, que el tiempo desde que se hace click hasta que se

brinda la respuesta no supere cierto valor. Sin embargo, la evidencia indica que los tiempos tienen mucha variabilidad y en ocasiones puede suceder que no se cumpla el requerimiento.

¿Por qué se presenta variabilidad en los procesos? En todos ellos actúan numerosos factores que varían permanentemente. En el ejemplo considerado, se pueden mencionar al dispositivo utilizado, la red local, la cantidad de memoria y/cloud que tiene ocupada el usuario, etc. Estos factores pueden variar en un momento dado o a través del tiempo y todo influye en la característica de interés que es el tiempo de respuesta al hacer click.

En síntesis, el concepto de "proceso" incluye a la "variabilidad", que se manifiesta en una o varias características del producto o servicio (o de cualquier otra salida) y esta variabilidad genera incertidumbre ya que no se pueden predecir con exactitud los valores que asumirán esas características para una salida en particular.

1.3 Conceptos fundamentales que intervienen en un problema "estadístico"

Como se menciona anteriormente, los ingenieros están envueltos en determinadas tareas, a partir de las cuales surgen diferentes situaciones o problemas que los mismos deben enfrentar.

Algunos ejemplos de ellos son:

1. Quejas porque una aplicación resulta muy lenta en dar respuesta.
2. Análisis de las fallas en una red informática.
3. Control de cambios de un producto de software.
4. Evaluación de la calidad de una ruta.
5. Controlar la calidad de los dispositivos de almacenamiento digital producidos.

En estas situaciones pueden surgir preguntas a responder. Por ejemplo: "¿Cada cuánto tiempo ocurren las fallas?", "¿El dispositivo de almacenamiento tiene la capacidad especificada?", "¿Cuál dispositivo responde mejor en determinada situación?", etc.

El primer paso para resolver estos problemas es plantearlos en forma precisa, lo que implica definir entre otras cosas, **la/las característica/s de interés** y el **conjunto para el cual se quieren obtener conclusiones.**

En el caso del control de calidad de los dispositivos de almacenamiento producidos, una característica de interés (que define la calidad del producto) es la capacidad real del dispositivo. Por ejemplo, un dispositivo puede tener una capacidad nominal de 32 Gb pero la capacidad real puede diferir de ese valor. El conjunto podría ser el total de

los dispositivos (de ese modelo determinado) que se pueden producir (un conjunto muy grande e indeterminado y que difícilmente se pueda evaluar en su totalidad).
En el caso del proceso al hacer click en un botón de una aplicación, la característica de interés podría ser la que ya se mencionó, que es el tiempo desde que se hace click hasta que se brinda la respuesta. En esta situación, el conjunto no está conformado por elementos concretos como son los dispositivos de almacenamiento, aquí es más abstracto, podría ser el total de ocasiones que se hace click en el botón y se ejecuta el proceso (también es un conjunto muy grande e indeterminado).
En muchas situaciones se requiere el estudio del comportamiento de la característica de interés (conocer como varía la misma de unidad a unidad) y entonces es necesario llevar adelante una investigación empírica para obtener la información pertinente. La Estadística cumple un rol fundamental en dicha resolución desde el planteo mismo de la pregunta inicial y proporciona métodos para obtener, organizar y resumir datos que se convierten luego en información de utilidad; así como herramientas para la toma de decisiones en presencia de variabilidad e incertidumbre.

En todo problema que involucre datos sobre características o el estudio de la variabilidad de las mismas se debe comenzar realizando un adecuado planteo. En los párrafos siguientes se definen y describen los conceptos involucrados en dicho planteo. Estos conceptos serán recurrentes a lo largo de toda la materia.

Definiciones:

- La **población** es el conjunto de todos los elementos bajo estudio, es decir, el universo respecto del cual se quiere obtener conclusiones o tomar decisiones.
- A cada uno de los elementos que componen la población se lo denomina ***unidad elemental*** (también se le puede llamar unidad de análisis).

Según la cantidad de elementos o unidades que la componen, una población puede clasificarse en ***finita o infinita***. En el caso de poblaciones finitas, el total de unidades elementales, o tamaño de la población, se simboliza con N. En el caso de poblaciones infinitas, el número de unidades elementales no está claramente determinado o limitado. Por ejemplo, todas los dispositivos que se podrían producir manteniendo las condiciones de operación o todos los “clicks” que se podrían hacer correspondientes a un botón de una aplicación.
Reconsiderando la situación de la calidad de los dispositivos de almacenamiento, si la población está compuesta por la totalidad de los dispositivos, se entiende que se trata de los que se vienen produciendo (y de los que se seguirán produciendo si la producción continúa con un comportamiento estable).

Pero, considere la siguiente situación:
Un cliente compró 100 dispositivos que serán utilizados con un determinado fin y se necesita que no tenga menos capacidad que la nominal (mas exigencias). Entonces interesa conocer el comportamiento de los dispositivos de almacenamiento (en cuanto a su capacidad real) que serán utilizadas con dicho fin. En este caso, la población está formada por los 500 dispositivos que adquirió el cliente.
Observe la importancia de definir a la población de acuerdo al objetivo planteado.

Definiciones:

- Una ***variable*** es cualquier característica que pueda tomar diferentes valores o categorías en las unidades elementales.

Las variables se clasifican, según los valores o categorías que pueden asumir, en cuantitativas o cualitativas.

- Las **variables cualitativas o categóricas** clasifican a las unidades elementales en categorías o niveles. Por ejemplo, “el tipo de problema de soporte técnico encontrado durante la instalación y configuración de un software”, “la condición de defectuosa o no de piezas plásticas”, “el estado de una máquina en un momento determinado (ocupada, inactiva o fallada)” y “la preferencia del consumidor en relación a una familia de productos”. A este tipo de variables también se les llama **atributo**.

- Las **variables cuantitativas** le asignan a cada unidad elemental un número. Estas, a su vez, se pueden clasificar en **discretas y continuas**.

Para las variables discretas, el conjunto de valores posibles es finito o infinito numerable y se asocia, generalmente a un subconjunto de los números enteros. Son variables discretas “cantidad de pruebas realizadas a un software hasta su implementación”, “el número de piezas defectuosas en cajas de 20 unidades” y “la cantidad de llamadas que ingresan a un helpdesk en períodos de 10 minutos”.
Para las variables continuas, el conjunto de valores posibles es un intervalo o conjunto de intervalos de números reales. Son variables continuas “la longitud de piezas metálicas”, “tiempo de llamada de los clientes de una helpdesk”, “la capacidad real de almacenamiento de un dispositivo” y “la temperatura de salida de un polímero”.
Las variables se simbolizan con letras mayúsculas; mientras que los valores posibles

se simbolizan con letras minúsculas. Por ejemplo, para la población de dispositivos de almacenamiento, la variable de interés es X: capacidad real del dispositivo (en Gb). Uno de los dispositivos analizadas en el estudio presentó una capacidad real de 31.7 Gb, es decir x = 31,7 Gb.

Resumiendo, se puede decir que si la pregunta es "¿quiénes son los objetos bajo estudio?, surge en primer lugar el concepto de unidad elemental y luego el de población, como la totalidad de las unidades elementales, asociadas a un objetivo. Y si la pregunta es ¿qué características se van a observar en dichas unidades?, surge el concepto de variable.
En la situación tomada como ejemplo, para la misma población de dispositivos, podrían haberse observado otras variables. En la Tabla siguiente se mencionan y clasifican algunas variables posibles.

Variable	Clasificación	Valores posibles (*)
"cumplimiento de la especificación"	Cualitativa o categórica	Sí, No
"tipo de ficha de conexión"		USB, Mini USB, etc
"nivel de calidad"		Bajo, Medio, Alto
"nº de fallas"	Cuantitativa discreta	0, 1, 2, 3, ……
"peso"	Cuantitativa continua	(0, +∞)
"tensión máxima que soporta"		(0, +∞)
"longitud"		(0, +∞)

Definición:

- Cualquier medida que resuma información de la población se denomina ***parámetro***.

Generalmente, los parámetros se simbolizan con letras griegas. Por ejemplo, la proporción poblacional se simboliza con la letra **π**, el promedio poblacional se simboliza con la letra **μ** y la desviación estándar poblacional se simboliza con **σ**. (cada parámetro específico se define más adelante)

Cuando se analizan datos de una determinada característica, el estudio puede ser muestral o poblacional, según se estudie a un subconjunto o a toda la población.

Definiciones:

- Un **censo o estudio exhaustivo** es un estudio en el que se observan todas las unidades de una población.

- Un **estudio por muestreo** es un estudio en el que se observa a un subconjunto de unidades de una población.

Los estudios poblacionales no siempre son posibles de llevar a cabo, ya sea porque la población es infinita o porque los ensayos que deben hacerse para medir la característica en estudio son destructivos o muy costosos. También puede ser que el proceso sea muy lento y se demore mucho tiempo en obtener todas las observaciones.

Definición:

- Una ***muestra*** es un subconjunto de elementos de la población bajo estudio. Su tamaño o cantidad de elementos lo simbolizaremos con **"n".**

Existen diferentes maneras de seleccionar una muestra; pero es fundamental tener presente que sólo a partir de ***muestras aleatorias o probabilísticas*** se pueden extender los resultados a toda la población de manera válida.

¿Qué es una muestra aleatoria o probabilística? La idea básica de una muestra de este tipo es que cada unidad de la población tenga una posibilidad o chance conocida de ser seleccionada para la muestra.

En las ***muestras por conveniencia***, en cambio, las unidades se incorporan porque se consiguen fácilmente, por voluntad del participante, etc. Estas muestras pueden resultar sesgadas si no representan el patrón de variabilidad de la población en estudio.

Además de la manera, es también importante elegir adecuadamente el tamaño de la muestra ya que ambos influyen en la calidad de las conclusiones que se obtendrán.

Definición:

- Se denomina ***estadístico*** a cualquier medida que resuma información de una muestra (por ejemplo, el promedio de una muestra, la proporción muestral de unidades con cierta característica, el valor más frecuente en la muestra, etc.). Se usa diferente notación para estadísticos y parámetros.

Una vez obtenidos los datos provenientes de una muestra o población, se debe proceder al análisis de los mismos. Se organizan, resumen y presentan de manera de facilitar su análisis e interpretación. Esto se realiza aplicando herramientas de análisis descriptivo, es decir, se construyen cuadros y gráficos y se obtienen indicadores o medidas de resumen (estadísticos o parámetros, según corresponda). Este análisis permite poner de manifiesto el patrón de variabilidad de las características de interés y obtener información sobre el problema planteado.

Si se pudo estudiar a toda la población, con este análisis es suficiente para dar respuesta al problema. Si se estudió a un subconjunto de la población (estudios muestrales), se deben aplicar herramientas de análisis inferencial, para generalizar las conclusiones obtenidas a toda la población, corriendo cierto riesgo de obtener conclusiones erróneas.

Definición:

- **Análisis descriptivo de los datos**: consiste en la aplicación de herramientas (tablas, gráficos, indicadores) para resumir y/o presentar un conjunto de datos, sean estos de una muestra o de una población finita.

Si el estudio es exhaustivo (censo), con las herramientas de análisis descriptivo es suficiente para dar respuesta al problema planteado, ya que se cuenta con información de todas las unidades de la población y se pueden obtener los valores de los parámetros de interés. En cambio, en el caso de estudios por muestreo, las conclusiones obtenidas deben inferirse a la población a través de herramientas de análisis inferencial.

Definición:

- **Análisis inferencial de los datos:** consiste en la aplicación de herramientas (intervalos de confianza, pruebas de hipótesis) que permiten extender las conclusiones de una muestra a la población, con riesgos controlados. Estas herramientas se apoyan en la Teoría de la Probabilidad.

En el ejemplo de los dispositivos, si se desea evaluar la calidad en cuanto a la capacidad real de almacenamiento de todos los dispositivos producidos, sería razonable trabajar con muestras, es decir, seleccionar aleatoriamente un conjunto de dispositivos y medirles la capacidad real para luego analizar los valores obtenidos. Si a

partir de esos datos se obtiene la capacidad real promedio, esa medida sería un estadístico.

2. Análisis Descriptivo de datos

Considerando el conjunto de datos obtenidos, se presentan algunas de las técnicas más usadas para:

- *la presentación de los mismos en forma ordenada (tablas y gráficos)*
- *el cálculo de medidas resúmenes.*

Para lograr estos objetivos, se debe pensar primero en los distintos valores de la variable de interés que se observan, y la frecuencia con la que aparece cada uno. A partir de esta idea, surge el concepto de **distribución de frecuencias**

2.1 Distribuciones de frecuencias y gráficos

El primer paso para ordenar y presentar los datos es obtener la distribución de frecuencias. La misma queda definida al indicar todos los valores de la variable que se observan en el conjunto de datos, indicando para cada uno su frecuencia.

Antes de analizar los datos es importante determinar primero si se recogieron datos cualitativos o cuantitativos ya que se usan herramientas distintas para cada uno de ellos.

2.1.1 Datos correspondientes a un atributo (variable cualitativa)

Ejemplo 2.1:

Con respecto a una red de fibra óptica en una ciudad se obtuvieron datos correspondientes a la zona en la que se reportaron fallas durante un determinado año. A partir de los mismos se obtuvo la siguiente tabla que contiene la distribución de frecuencias

Zona	*Número de fallas* N_k	*Proporción de fallas* f_k
Norte	18	0,171
Oeste	11	0,105
Centro	13	0,124
Sur	34	0,324
Suroeste	29	0,276
Totales	105	1

Para definir la población bajo estudio se debe hacer referencia al objetivo definido previamente. En este ejemplo se supone que se quiere estudiar solo las fallas de un año porque fue el período de tiempo en el que se utilizó un nuevo método de

transmisión. Entonces queda:

- Población: 105 fallas ocurridas durante el año correspondiente.
- Unidad elemental: Cada falla
- Característica en estudio: Zona donde ocurrió la falla (variable cualitativa o atributo)

En este caso se estudió a la población completa, es un censo. Note que el elemento no es un objeto si no cada falla que ocurrió.

Para armar la distribución de frecuencias se particionó al conjunto de los 105 fallas en subconjuntos o **clases** según los niveles del atributo (en el ejemplo, 5 niveles que se corresponden con las 5 zonas).

- El número de elementos que pertenecen a cada clase recibe el nombre de **frecuencia absoluta (n_k).**

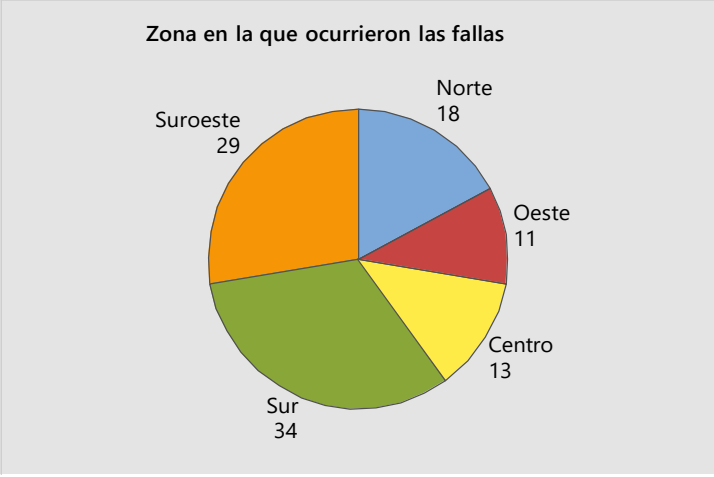
- El cociente entre la frecuencia absoluta y el número total de observaciones recibe el nombre de f**recuencia relativa (f_k)**.

La suma de las frecuencias absolutas es igual al número total de observaciones (N) y en consecuencia, la suma de las frecuencias relativas es siempre igual a 1.
Es muy frecuente expresar a las frecuencias relativas como porcentaje; así en el ejemplo diremos que el 32 % de las fallas ocurrieron en zona sur durante el año estudiado.

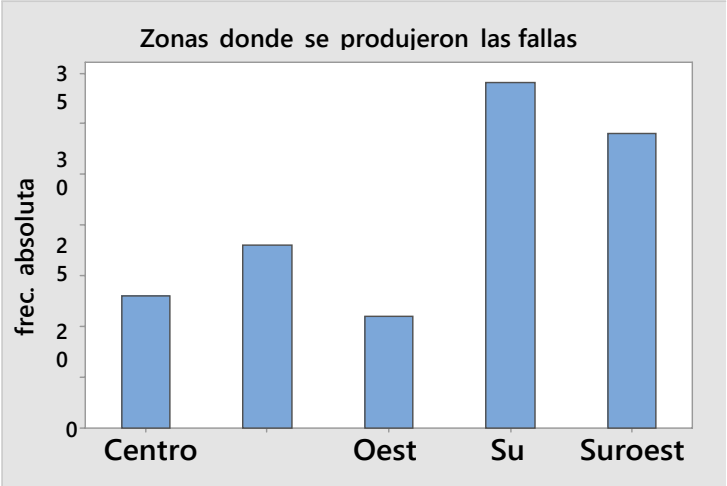
Representación gráfica de un atributo o variable cualitativa.

Los distintos tipos de gráficos representan a la distribución de frecuencias indicando la n_k o la f_k.

- **Gráfico circular o sectores:**



- **Gráfico de barras**



- **Diagrama de pareto**

Es un caso especial del diagrama de barras, que se usa con frecuencia en control de calidad. Las barras se grafican en orden descendente. Puede también incluir una segunda escala (del 0 al 100), encima de las barras de las clases, que muestre los porcentajes acumulados.

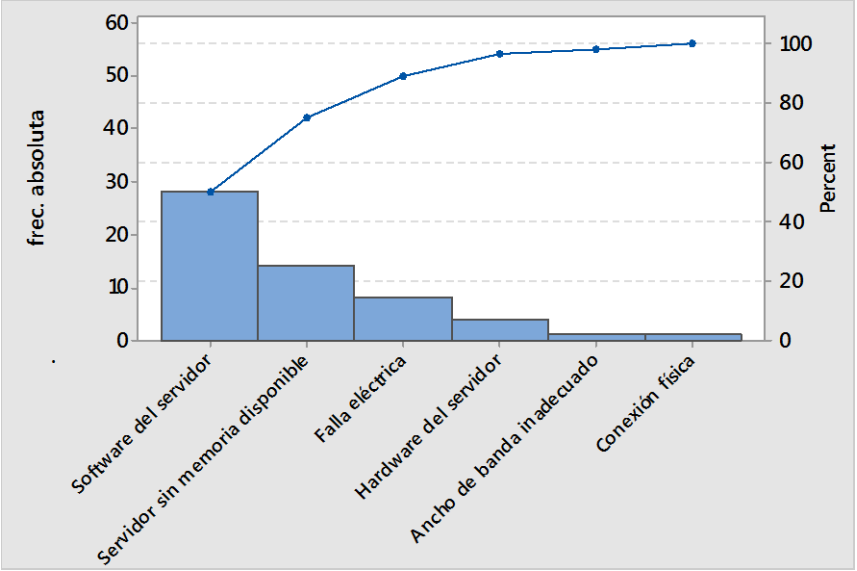
Este tipo de diagrama lleva el nombre del economista italiano V. Pareto y en general representa la “ley de Pareto”, esto es: la mayor parte de los defectos aparece sólo en unas pocas categorías.

Ejemplo 2.2:

Un analista de redes registró las causas principales que propiciaron fallas en los sistemas durante los últimos seis meses, obteniendo el siguiente resultado:

Razón de la falla	*Frecuencia*
Conexión física	1
Falla eléctrica	8
Software del servidor	28
Hardware del servidor	4
Servidor sin memoria disponible	14
Ancho de banda inadecuado	1

Este resultado se presenta en un diagrama de Pareto:



2.1.2 Datos correspondientes a una variable cuantitativa discreta

Ejemplo 2.3:

Una empresa registra la cantidad de veces en el día que se utiliza una determinada terminal durante un período de 50 días. Los resultados obtenidos fueron:

84	88	87	89	88	89	88	91	87	85
88	89	90	88	87	91	86	89	85	88
86	90	89	84	91	92	89	88	94	90
87	89	91	86	90	89	91	92	89	88
85	88	87	88	91	87	92	90	85	87

A fin de ordenar la información se particiona al conjunto de 50 días en clases, según la variable en estudio “número de veces que se utiliza una terminal por día” y se realiza el cómputo de frecuencias según se indica en la siguiente tabla:

Valor de la variable x_k	Frecuencia absoluta n_k	Frecuencia relativa f_k	Frecuencia absoluta acumulada N_k	Frecuencia relativa acumulada F_k
84	2	0,04	2	0,04
85	4	0,08	6	0,12
86	3	0,06	9	0,18
87	7	0,14	16	0,32
88	10	0,20	26	0,52
89	9	0,18	35	0,70
90	5	0,10	40	0,80
91	6	0,12	46	0,92
92	3	0,06	49	0,98
93	0	0,00	49	0,98
94	1	0,02	50	1,00
Totales	50	1		

- **La frecuencia absoluta acumulada (N_k)** es la cantidad de elementos correspondientes a valores de la variable menores o iguales a x_k.
- **La frecuencia relativa acumulada (F_k)** es la proporción de elementos cuyo valor de la variable es menor o igual que x_k.

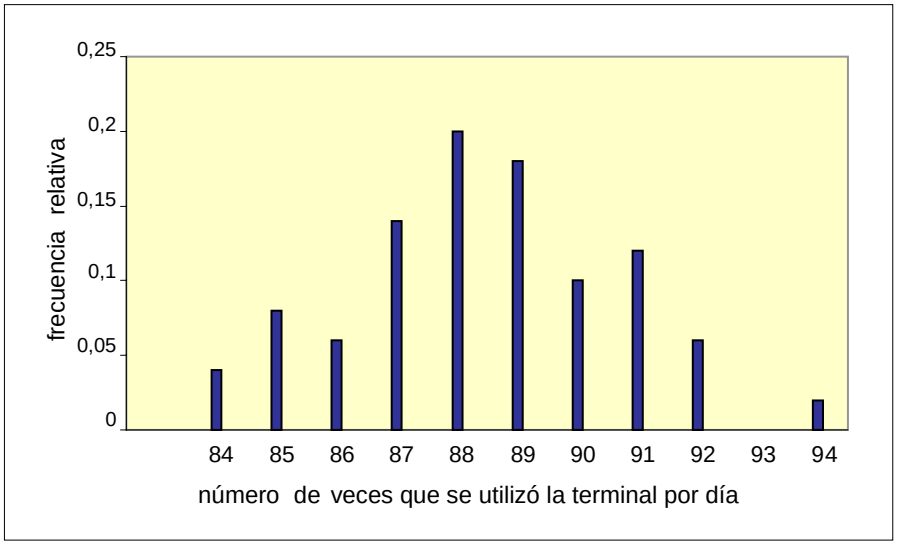
Como ejemplo se interpretan los valores correspondientes a la quinta fila:

En el 20% de los **días** se utilizó 88 veces la terminal (f_5).
En el 57% de los **días** se utilizó 88 veces o menos la terminal (F_5).
Reemplazando "el 20% de los" por "10" o "el 52% de los por "26" se obtienen las interpretaciones para n_6 y N_6.

Representación gráfica de una variable cuantitativa discreta.

- **Gráfico de bastones**

El mismo se obtiene representando la frecuencia (absoluta o relativa) correspondiente a cada valor de la variable mediante un segmento cuya longitud es proporcional a dicha frecuencia.



2.1.3 Datos correspondientes a una variable cuantitativa continua

Ejemplo 2.4:

En una gran empresa se realizan backup periódicos sobre ciertas base de datos y quiere analizar el espacio que ocupan los mismos. A tal fin se registra la espacio (en Gb??) que ocupan 30 backups elegidos al azar. Los resultados obtenidos fueron:

85 - 117 - 92 - 120 - 94 - 110 - 151 - 90 - 80 - 116 - 95 - 102 - 100 - 113 - 118 - 140 - 133 -108 - 115 - 148 - 110 - 130 - 100 - 120 - 108 - 125 - 105 - 130 - 112 – 150

- **Diagrama de tallo y hoja**

Como un paso previo a la construcción de la distribución de frecuencias, los datos

pueden organizarse en un diagrama de **tallo-hoja**. En este tipo de diagramas, cada valor observado se descompone en “dígitos tallo” y “dígitos hoja”.

En el ejemplo planteado, la decena y la centena de cada valor observado forman los “dígitos tallo” y la unidad el “dígito hoja”.

Así, para las dos primeras observaciones (85 y 117) resultan:

8|5

11| 7

Es conveniente presentar a los dígitos hoja ordenados en forma creciente para facilitar la posterior utilización del diagrama tallo-hoja, tanto en forma gráfica como tabular.

En el ejemplo, el diagrama de tallo-hoja queda de la siguiente manera:

8	**0 5**
9	**0 2 4 5**
10	**0 0 2 5 8 8**
11	**0 0 2 3 5 6 7 8**
12	**0 0 5**
13	**0 0 3**
14	**0 8**
15	**0 1**

Otra forma de organizar la información es construyendo la distribución de frecuencias. Para definir las clases, el intervalo total de variación de la variable se particiona en subintervalos de igual amplitud. Cada uno de ellos identifica a una clase y recibe el nombre de intervalo de clase.

Al intervalo total de variación lo determinan el valor mínimo (80 m^2) y el máximo (151 m^2) que asume la variable.

La diferencia entre ambos valores (en este caso 71 m^2) se llama **rango**. Los 30 valores observados pertenecen al intervalo [80, 151].

Una vez obtenidos los intervalos de clase se procede para obtener las frecuencias de la misma manera que en los casos vistos anteriormente.

Cuando se agrupan datos a través de intervalos de clase, se produce una pérdida de información por la no conservación de los valores individuales. Demasiados intervalos provoca pérdida de efectividad como medio de resumir datos y podría pasar que muchos intervalos tengas frecuencias muy bajas o nulas; en cambio, pocos intervalos condensan tanto la información que arrojan poca luz sobre el comportamiento de la característica.

La elección del número de subintervalos está estrechamente relacionada con la cantidad de datos que se consideran. Es común usar entre 5 y 20 subintervalos.

Hoy en día el armado de la distribución de frecuencias y los gráficos son realizados por diferentes programas de informática y toma relevancia la interpretación de la

información que da la distribución en lugar de la construcción en sí. Los intervalos cumplen que son semiabiertos para que cada valor de la variable pertenezca a uno y sólo uno de los intervalos.

En el ejemplo: $79 < x \leq 91$ ó $(79 , 91]$

Es preferible, por facilidad en el análisis, que los intervalos posean igual amplitud, pero podrían no serlo en ciertas situaciones particulares.

La tabla siguiente muestra la distribución de frecuencias del ejemplo:

Intervalo de clase	Punto medio	Frecuencia absoluta	Frecuencia relativa	Frecuencia absoluta acumulada	Frecuencia relativa acumulada
$79 < x \leq 91$	85	3	0,10	3	0,10
$91 < x \leq 103$	97	6	0,20	9	0,30
$103 < x \leq 115$	109	8	0,27	17	0,57
$115 < x \leq 127$	121	6	0,20	23	0,77
$127 < x \leq 139$	133	3	0,10	26	0,87
$139 < x \leq 151$	145	4	0,13	30	1,00
		30	1,00		

Como ejemplo se interpretan los valores correspondientes a la tercera fila:

El 27% de los backups ocupan un espacio entre 103 y 115 Gb (f_3).

El 57% de los backups ocupan un espacio menor a 115 Gb (F_3)

Reemplazando “el 27% de los” por “8” o “el 57% de los por “17” se obtienen las interpretaciones para n_3 y N_3

Representación gráfica de una variable cuantitativa continua.

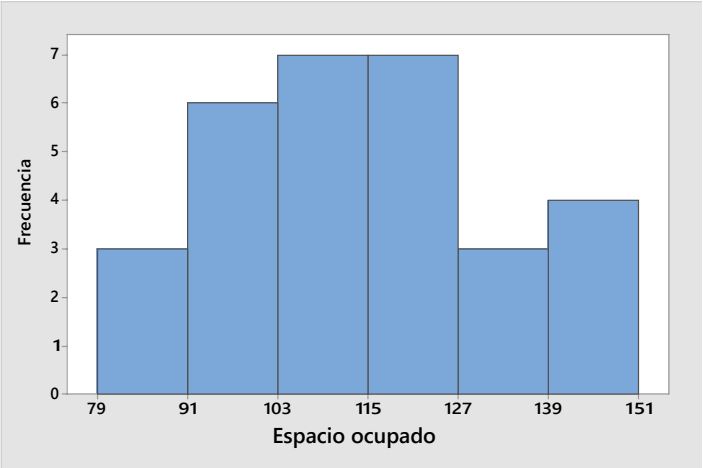
- **Histograma**

• Las bases de las barras tienen la longitud igual a la amplitud del intervalo de clase que representan y se ubican sobre el eje de la abscisa.

• El área de cada barra es proporcional a la frecuencia del intervalo de clase.

• Si los intervalos de clase son de igual amplitud, las alturas de las barras resultan proporcionales a las frecuencias de las clases. En caso de amplitudes diferentes, las alturas deben ser calculadas para que se verifique la condición anterior.

Para el ejemplo, se presenta a continuación el histograma correspondiente:



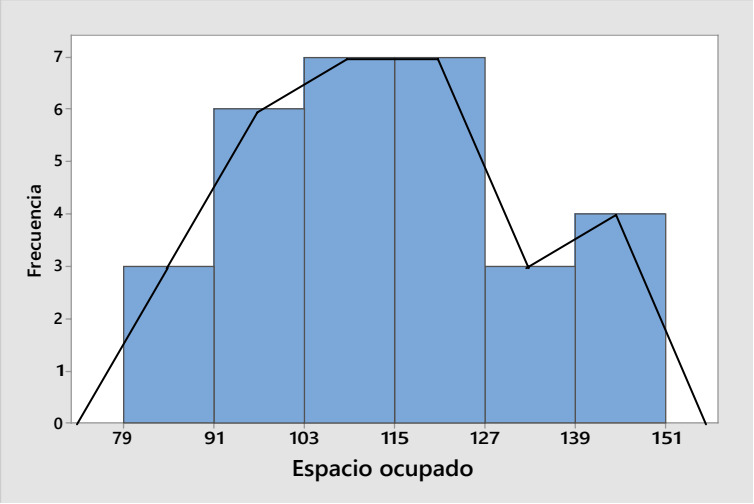
Los histogramas son más fáciles de interpretar si los intervalos de clase tienen la misma amplitud.

El histograma, al igual que el diagrama de tallo-hoja, proporciona una impresión visual del aspecto que tiene la distribución de las observaciones, así como información sobre la dispersión de los datos.

Al pasar de los datos originales o del diagrama de tallo-hoja a la distribución de frecuencias y al histograma, se pierde parte de la información debido a que ya no se tienen las observaciones originales. Sin embargo, esta pérdida en la información a menudo es pequeña si se le compara con la facilidad de interpretación ganada al utilizar la distribución de frecuencias y el histograma.

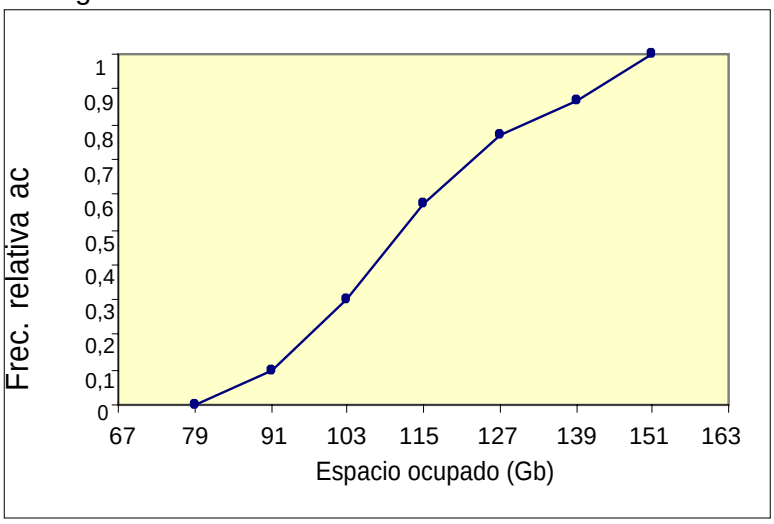
Para conjuntos de datos pequeños, los histogramas pueden cambiar claramente de apariencia si el número de clases o el ancho de éstas cambia. Esto no se da si el número de observaciones es grande.

También se puede agregarle al histograma un **polígono de frecuencias**. El mismo se conforma uniendo con líneas el punto medio de la parte más alta de cada barra como se muestra en la siguiente gráfica:



- **Polígono de frecuencias acumuladas**

De igual forma se puede construir el polígono de frecuencias acumuladas como se muestra en la siguiente figura:



Se presenta a continuación otro ejemplo sobre variable continua:

Ejemplo 2.5

Los datos de la siguiente tabla representan la resistencia a la tensión, en libras por pulgada cuadrada (psi), de 80 muestras de una nueva aleación de aluminio y litio, que está siendo evaluada como posible material para la fabricación de elementos estructurales de aeronaves.

Resistencia a la tensión de 80 muestras de aleación de aluminio-litio

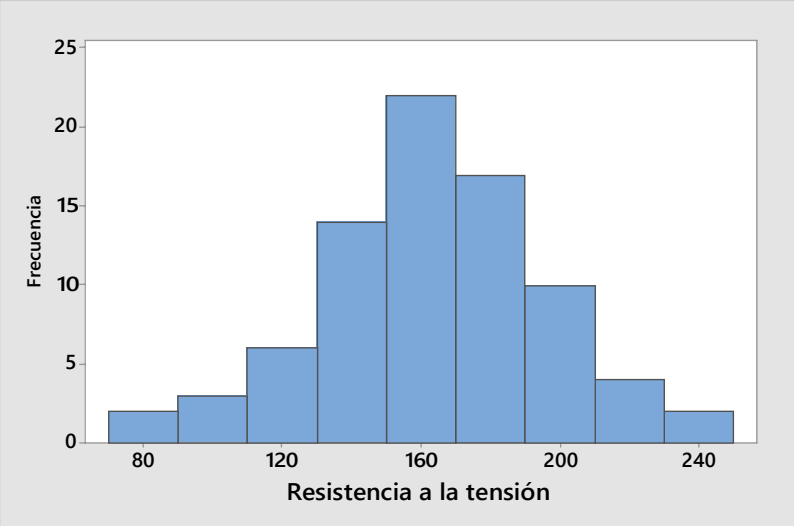
105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Los datos fueron registrados conforme se realizaba la prueba y en este formato no conllevan mucha información con respecto a la resistencia a la tensión. No es fácil responder a preguntas tales como “¿Qué porcentaje de las muestras fallaron debajo de los 120 psi?”.

Dado que se tienen muchas observaciones. Una de las formas de representación gráfica es mediante el ya visto diagrama de tallo y hoja:

Tallo	*Hoj*	*Frecuenci*
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

Otra gráfica apropiada es el histograma:



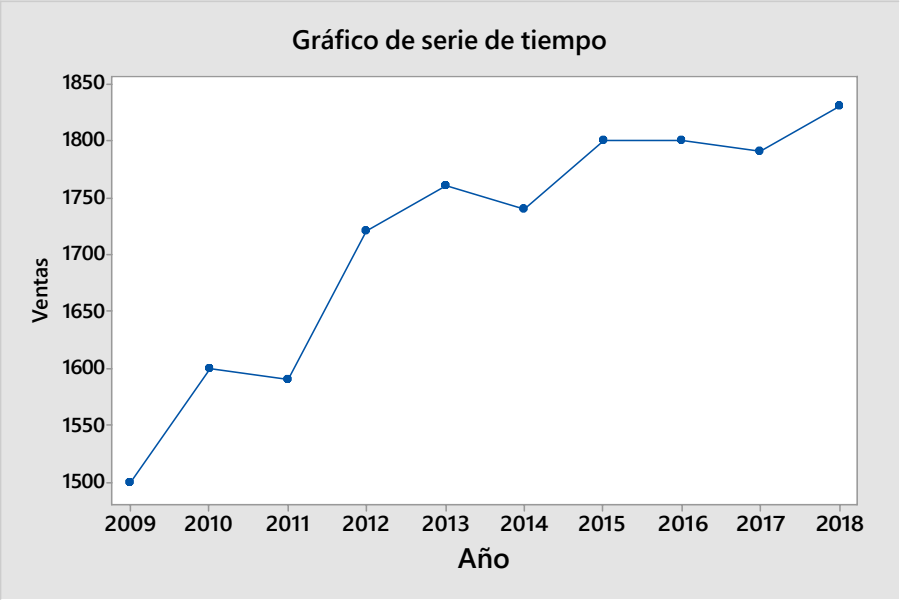
2.1.4 Gráficas de series de tiempo

Las gráficas consideradas hasta el momento (histogramas, diagramas de tallo y hoja) son métodos visuales muy útiles para mostrar la variabilidad presente en los datos. Sin embargo, con frecuencia el tiempo es un factor importante que contribuye a la variabilidad observada en los datos, y los métodos mencionados no lo toman en cuenta, es decir, se considera que los datos son tomados todos en el mismo momento o que no hay cambios importantes en el tiempo. Una serie de tiempo, o secuencia de tiempo, es un conjunto de datos en los que las observaciones se registran en el orden en que ocurren. La **gráfica de una serie de tiempo** es un diagrama en el que el eje vertical denota el valor observado (por ejemplo X), mientras que el eje horizontal denota el tiempo (que puede ser minutos, días, años, etc.). Cuando se grafican las

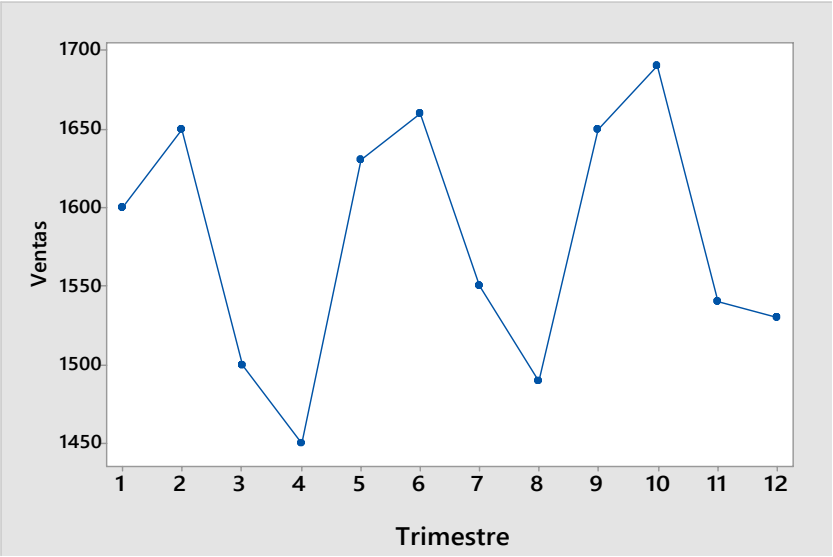
mediciones como una serie de tiempo, puede ocurrir que se observen tendencias, ciclos u otras características importantes de los datos que, de otra forma, pasarían inadvertidas.

Por ejemplo, considerando la figura 1.3, la parte (a) presenta una serie de tiempo de las ventas anuales de una compañía durante los últimos diez años. La impresión general que ofrece esta gráfica es que las ventas tienen una **tendencia** a crecer. Existe cierta variabilidad en esta tendencia, donde, las ventas en varios años aumentaron con respecto a las del año anterior, mientras que en algunos casos disminuyeron. La parte (b) presenta las ventas de los tres últimos años notificadas por trimestre. Esta gráfica muestra de manera clara que las ventas anuales de la empresa exhiben una variabilidad **cíclica** por trimestre, donde las ventas en los dos primeros trimestres son mayores que en los dos últimos.

Esta tendencia (a) y ciclos (b) no podrían observarse en cualquier gráfico que muestre la distribución de frecuencia de los datos. Es por esto que es de mucha importancia que el análisis en el tiempo, cuando se pueda, se realice antes de obtener las distribuciones de frecuencias. Además, si se observa alguna tendencia patrón o ciclo, carece de sentido estudiar las frecuencias obviando el tiempo ya que no puede dejarse de lado el “efecto” tiempo y trabajar suponiendo que los datos se tomaron en el mismo momento.



(a)



(b)

Figura 1.3. Ventas de una compañía por año (a) y por trimestre (b)

2.2 Medidas características de una muestra

Las medidas que resumen la información de una distribución de frecuencias o de un conjunto de datos reciben el nombre **estadísticos o parámetros**, según si se trata de una muestra o de una población.

Estas medidas pueden dar distintos tipos de información. Por ejemplo, una medida muy conocida es el promedio o media aritmética. Esta medida da información del centro de la distribución, alrededor de qué valor varían las observaciones, pero no da información sobre la variabilidad en sí, es decir, conociendo solamente un promedio no podemos saber si las distintas observaciones están cerca o lejos de ese promedio.

Entonces, las medidas se clasifican según el tipo de información que brindan.

2.2.1 Medidas de posición

Estas medidas están referidas a la posición de la distribución de frecuencias sobre el eje de las abscisas. Las tres primeras que se muestran en la tabla indican específicamente al centro de la distribución, por eso se llaman medidas de **tendencia central**. La simbología no es la misma según si se trata de un parámetro o de un estadístico. En la tabla se muestra la notación correspondiente a un estadístico.

NOMBRE	NOTACION	DEFINICION
Media aritmética	$\overline{x}$	Es el promedio de las observaciones
Moda	$\hat{x}$	Es el valor de la variable con mayor frecuencia
Mediana	$\widetilde{x}$	Es el mínimo valor de la variable que acumula, por lo menos, el 50 % de las observaciones ordenadas en forma creciente

También existen otras medidas de posición que, al igual que la mediana, se definen a partir de tener las observaciones ordenadas en orden creciente (medidas de orden):

Cuartiles	q_1 q_2 q_3	Son los mínimos valores de la variable que acumulan respectivamente, por lo menos : el 25% , el 50% y el 75% de las observaciones ordenadas en forma creciente.
Percentiles	p_1 p_2 ……p_{99}	el 1% , el 2% ..... el 99% de las observaciones ordenadas en forma creciente.

- **Media aritmética o promedio**

Es la más conocida y utilizada de las medidas de posición. No coincide necesariamente con un valor de la variable.

El cálculo del promedio de n observaciones de la variable X (x_i con i = 1, 2,.....n), resulta:

$$\bar{x} = \frac{\sum_{1=1}^{n} x_i}{n} \quad (1)$$

Si las n observaciones están agrupadas en r clases, la fórmula (1) resulta:

$$\bar{x} = \frac{\sum_{1=1}^{r} x_i . n_i}{n} = \sum_{1=1}^{r} x_i . f_i \quad (2)$$

En los casos en que las observaciones se encuentren agrupadas en intervalos de clase, se le da a x_i el valor del punto medio del intervalo de clase correspondiente.

Si los datos corresponden a una población, es decir, se trata de un parámetro, el símbolo del promedio es la letra griega **μ**

Características del promedio:

- Toma en consideración toda la información por lo tanto es muy sensible a la influencia de los valores extremos, lo que puede ser una ventaja o desventaja, según la situación.
- Es una medida de posición útil para comparar dos o más distribuciones, sólo si éstas tienen forma semejante.

Se dice que una medida es sensible a valores extremos cuando la misma toma un valor mayor o menor según si en el conjunto de datos se encuentran uno o más valores muy grandes o muy chicos respectivamente. Esto se muestra luego con un ejemplo

- **Moda**

Es el valor de la variable con mayor frecuencia.

Características de la moda:

- Algunos conjuntos de observaciones no poseen moda.
- Algunos conjuntos de observaciones tienen dos modas (son bimodales).
- Es la única medida que puede calcularse para cualquier tipo de variables.

- **Mediana**

Es el mínimo valor de la variable que acumula, por lo menos, el 50 % de las observaciones ordenadas en forma creciente, por tal razón es uno de los llamados estadísticos de orden.

$$\min(x / F(x) \geq 0{,}5)$$

Si se cuenta con n datos ordenados; si n es impar, la mediana es el que está en la posición central; pero si n es par, al ser dos los datos que se encuentran en la posición central, la mediana es el promedio de ambos.

Característica de la mediana:

- Al no tomar en cuenta toda la información ya que depende de qué magnitud tenga el o los valores centrales y no de la magnitud de ningún valor extremo, no es sensible a dichos valores extremos.

- **Cuartiles y Percentiles**

Se definen de forma similar a la mediana pero en vez de acumular el 50% acumulan el valor indicado (cuartil 1 el 25%, cuartil 2 el 50%,…, percentil 1 el 1%, etc).

De las definiciones se observa que el cuartil 2 (q_2) coincide con la mediana $(\tilde{x})$ y con el percentil 50 (p_{50}).

Si son medidas poblacionales se pueden utilizar letras mayúsculas para la simbología (Q_1, Q_2, etc)

Si se cuenta con una tabla de frecuencias, los datos están ordenados y tanto la mediana como los cuartiles se pueden obtener directamente observando la columna correspondiente a las frecuencias acumuladas, buscando el valor que acumula el porcentaje pretendido.

Todas las medidas definidas hasta aquí, a excepción de la Moda, solo pueden obtenerse para variables cuantitativas. La Moda también puede obtenerse para variables cualitativas

Ejemplo 3.1:

Tomando las observaciones del ejemplo 2.3 (página 12) se calculan las siguientes medidas:

- Media o Promedio:

$\bar{x} = \frac{(84 + 88 + ...)}{50} = 88{,}44$. Utilizando la fórmula (1), con los 50 datos "sueltos".

$\bar{x} = \frac{(84x2 + 85x4 + ... + 94x1)}{50} = 88{,}44$. Utilizando la fórmula (2), a partir de la distribución de frecuencias.

Interpretación: En promedio la terminal se utilizó 88,44 veces por día

- Mediana

$\tilde{x} = 88$. Ordenando los 50 datos, en las dos posiciones centrales (n es par), que son la 25 y 26, se encuentra el valor 88, o bien directamente se observa en la distribución de frecuencias (página 12), en la última columna, que el primer valor que supera el 50% de frecuencia acumulada es el 88. Note que como el 88 esta repetido 10 veces, incluyéndolo acumula más del 50%, pero esto no quita que en la posición central esté dicho valor.

Interpretación: En la mitad de los días se utilizó 88 veces o menos la terminal

- Moda

$\hat{x} = 88$. Se observa también directamente en la distribución de frecuencias que 88 (con una frecuencia de 10) es el más observado.

Interpretación: La cantidad de usos por día que más veces se repite es 88

- Cuartiles

$q_1 = 87$ y $q_3 = 90$. Se obtienen de igual forma que la mediana ya sea ordenando los datos o directamente de la distribución de frecuencias, pero se busca el primer valor que acumula 25% y 75% respectivamente

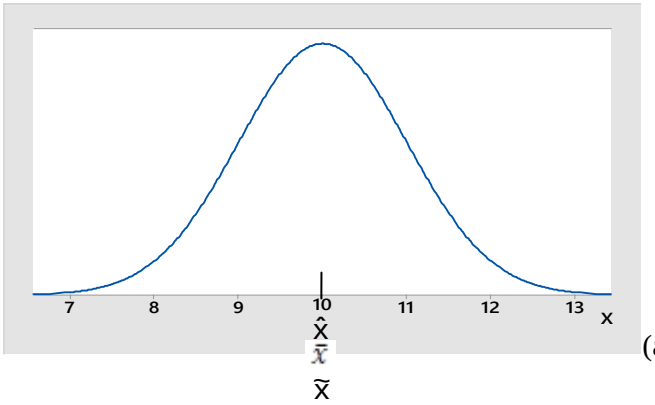
Interpretación: En el 25% de los días se utilizó 87 veces o menos la terminal

Interpretación: En el 75% de los días se utilizó 90 veces o menos la terminal

➢ Comparación de media aritmética, mediana y moda

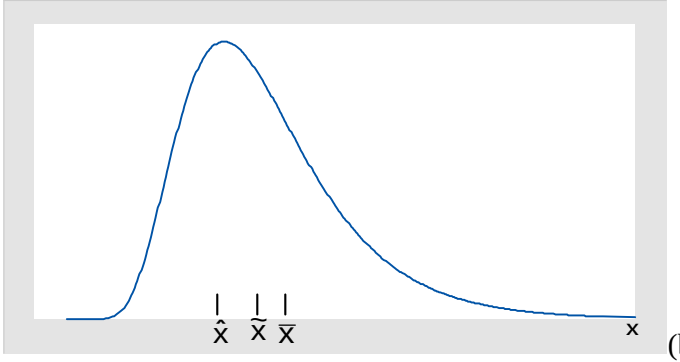
En el ejemplo 3.1 se observa que tanto la mediana como la moda coinciden y la media es levemente mayor. En general pueden ser diferentes las tres medidas y esto depende principalmente de la simetría de la distribución. Si la distribución es simétrica y con una sola moda, las tres medidas coinciden. Si la distribución es asimétrica, significa que de un lado puede tener valores extremos más alejados y esto hace que la media, la cual es sensible a estos valores, quede posicionada hacia ese lado. A continuación se presentan distribuciones simétricas (a) y asimétricas (b y c) y se ubican las medidas de tendencia central:

- Distribución Simétrica:



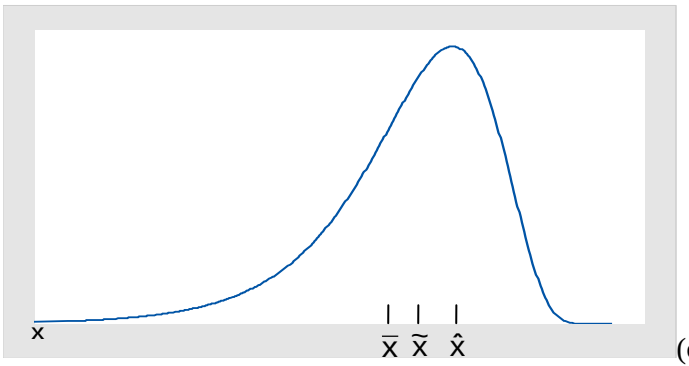
(a)

- Distribuciones Asimétricas o Sesgadas:
  Distribución Asimétrica a la derecha



(b)

Distribución Asimétrica a la izquierda



(c)

Note que el lado de la asimetría no lo marca la parte más elevada de la gráfica (valores más frecuentes) si no “la cola”, que es donde están ubicados los valores extremos alejados y menos frecuentes.

En el ejemplo, el motivo por el que la media es levemente mayor puede ser porque el valor 94 está un poco alejado hacia la derecha.

Pero si cambiamos el valor 94 por 104 (más alejado) queda:

$\bar{x} = 88{,}64$

$\tilde{x} = 88$

Se observa que la media aumenta un poco al aumentar este valor extremo pero la mediana no cambia porque el valor central sigue siendo 88.

2.2.2 Medidas de dispersión

Como se menciona al comienzo del punto 2, las medidas pueden dar distinto tipo de información. Analizando comparativamente las distribuciones que se ven en la figura 2.1 se observa que a pesar de que están igualmente centradas, los valores de la variable de cada una de ellas están alejados del promedio de manera distinta. Esta situación hace ver la necesidad de utilizar medidas que den información referida a la variabilidad o al alejamiento de los datos entre sí. Estas medidas son llamadas de **dispersión**.

Si para un conjunto de datos las medidas de dispersión son altas, entonces significa que existe mucha dispersión. Se dice también que el conjunto de datos es heterogéneo o que tiene mucha variabilidad.

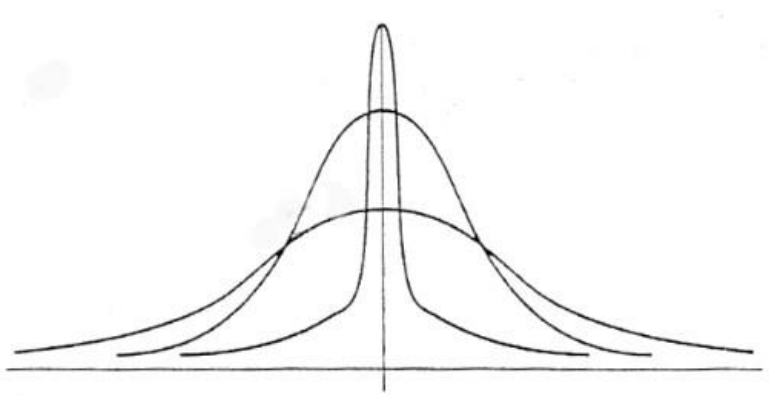


Figura 2.1

En la tabla se presentan medidas de dispersión con la simbología para muestra (estadísticos):

NOMBRE	NOTACION	DEFINICION
Rango	**r**	Es la diferencia entre el mayor y el menor valor de las observaciones
Varianza muestral	**s^2**	Es el promedio, aproximado, de los cuadrados de las diferencias entre los valores de las observaciones y su correspondiente media aritmética
Desvío estándar muestral	**s**	Es la raíz cuadrada positiva de la varianza.
Recorrido intercuartílico	**riq**	Es la diferencia entre el cuartil 3 y el cuartil 1.
Coeficiente de variación	**cv**	Es el cociente entre el desvío estándar y la media aritmética

- **Rango**

Es la diferencia entre el máximo valor de las observaciones (x_M) y el mínimo valor de las mismas (x_m)

$$r = x_M - x_m$$

Proporciona una primera información sobre la dispersión de los valores pero basta que al menos uno de los dos valores que intervienen en su cálculo esté excesivamente alejado para que pierda importancia la información que brinda.

- **Varianza**

También se le puede llamar Variancia. Es el promedio, aproximado, de los cuadrados de los desvíos de las observaciones con respecto a su media aritmética.

$$s^2 = \frac{\sum_{1=1}^{n} (x_i - \bar{x})^2}{n-1}$$

O, si se cuenta con los datos agrupados en r clases, $s^2 = \frac{\sum_{1=1}^{r} (x_i - \bar{x})^2 . n_i}{n-1}$

Notar que cada término positivo que se suma depende de la distancia de cada observación a la media. Por esto, **mayor será la varianza cuando más lejos este cada valor del promedio.**

Esta medida esta expresada en unidades al cuadrado, por lo que no puede interpretarse su valor en el contexto del problema. Se define debido a que muchas propiedades que se verán en otros capítulos se demuestran a partir de la varianza.

En el caso que los datos correspondan a una población, es decir, que se obtiene un parámetro, el símbolo de la variancia es la letra griega **σ^2**, y hay una pequeña modificación en la fórmula:

$$\sigma^2 = \frac{\sum_{1=1}^{N} (x_i - \mu)^2}{N}$$

Notar que como el cálculo es poblacional, aparecen las diferencias con la media poblacional. En el denominador no se resta el valor 1, esto es debido a que cuando se calcula a partir de una muestra, se debe hacer esa corrección por el hecho de que en el numerador se obtienen las diferencias con la media muestral. **En esta unidad, cuando no haya nada que indique si los datos pertenecen a una muestra o una población, se trabaja con la variancia muestral (y con todas las medidas muestrales).**

- **Desvío estándar**

Es la raíz cuadrada positiva de la varianza.

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{1=1}^{r} (x_i - \bar{x})^2 . n_i}{n-1}}$$

Este estadístico, esta expresado en la misma unidad de las observaciones. **Esto permite su interpretación en el contexto**. Para facilitar dicha interpretación, luego se presenta la regla empírica. Si es poblacional se simboliza **σ.**

Tanto para esta medida como la varianza, al igual que para la media aritmética, se utilizan todas las observaciones para el cálculo, y son muy sensibles a valores extremos.

- **Rango Intercuartílico**

Es la diferencia entre el tercer cuartil y el primer cuartil.

$$riq = q_3 - q_1$$

Esta medida no es sensible a valores extremos ya que se utilizan en el cálculo dos medidas de posición que no lo son.

Indica en qué rango varía el 50% central de los datos.

- **Coeficiente de variación**

Es una medida de variación relativa:

$cv = \frac{s}{\bar{x}}$ si es muestral, o $CV = \frac{\sigma}{\mu}$ si es poblacional

Es el desvío estándar expresado como porcentaje de la media aritmética, por lo tanto no viene expresado en unidades.

Es útil para la comparación de la variabilidad relativa entre distribuciones que no están expresadas en la misma unidad de medida o bien, entre distribuciones que si bien están expresadas en la misma unidad, poseen promedios muy dispares.

Ejemplo:

Siguiendo con los datos del ejemplo 2.3:

- Rango

$r = 194 - 84 = 110$

- Varianza:

$$s^2 = \frac{\left((84-88{,}44)^2 + (88-88{,}44)^2 + \ldots\right)}{49} = 227{,}68$$

- Desvío estándar:

$$s = \sqrt{227{,}68} = 15{,}09$$

- Rango intercuartílico

$$riq = 90 - 87 = 3$$

- Coeficiente de variación

Se supone ahora que además de observar la cantidad de aspiradoras vendidas por días, se tienen los datos de la cantidad de empleados que concurren a trabajar por día. Teniendo entonces los dos conjuntos de datos (X: cantidad de aspiradoras vendidas por día e Y: Cantidad de trabajadores que concurren a trabajar por día). ¿En cuál de los dos se presenta menor dispersión relativa?

Para la variable Y se tiene que $\bar{y} = 12{,}2$ y $s_y = 0{,}85$ (valores supuestos)

No se pueden comparar los desvíos estándar directamente ya que seguramente los datos de la variable Y son mucho menores a los de X.

$$c.v._X = \frac{15{,}09}{88{,}44}.100 = 17\%$$

$$c.v._Y = \frac{0{,}85}{12{,}2}.100 = 7\%$$

La menor dispersión relativa se presenta en la cantidad de empleados que concurren a trabajar (es lógico pensando que esa cantidad varía poco de día en día).

2.3 Regla empírica

Es posible que dos conjuntos de datos distintos tengan el mismo rango pero difieran considerablemente en el grado de variación de los datos. En consecuencia, el rango, si bien es muy sencillo de calcular e interpretar, es una medida que puede no llegar a dar la información necesaria sobre la dispersión de un conjunto de datos.

Las unidades de medición de la desviación estándar son las mismas unidades de la variable. Para poder entender e interpretar mejor su valor, se puede combinar con la media del conjunto de datos, teniendo en cuenta que el desvío indica “alejamiento” de dicha medida. En este sentido se ha estudiado qué porcentaje de datos quedan

cubiertos si se arman intervalos de la forma $\overline{x} \pm k.s$ (a la media se le suma k veces el desvío estándar).

Una regla práctica útil es la que se conoce como **regla empírica**, a saber:

Si un conjunto de datos tiene una **distribución aproximadamente simétrica y de forma campanular** (se le llama distribución Normal) verifica lo siguiente en cuanto al porcentaje de datos que pertenecen, aproximadamente, a cada intervalo:

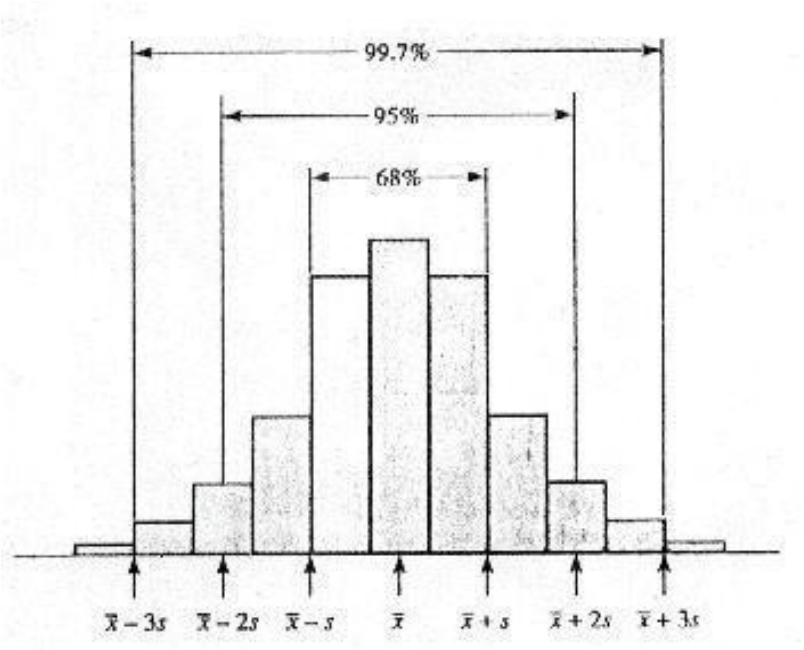


Figura 2.2

En la figura 2.2 se indican los porcentajes correspondientes a los tres intervalos cuando k toma los valores de 1, 2 y 3.

Con esta información, se pueden hacer interpretaciones teniendo solamente la media y el desvío y sabiendo que la forma de la distribución es campanular.

Notar que sería erróneo pensar que el rango de variación de una variable es $\overline{x} \pm s$. Este intervalo estaría cubriendo solo el 68% de los datos en muchos casos, y podría ser aún menor en algunos casos.

Si se quiere tener un intervalo aproximado donde varían el 100% de los datos, se puede tomar k=3, es decir $\overline{x} \pm 3.s$, ya que este intervalo cubriría prácticamente el total de los datos.

Se ha estudiado en diversos ámbitos que la forma campanular y simétrica es la más común, por eso se estudiaron los porcentajes en ese caso.

Ejemplo:

Con respecto al ejemplo 2.4 correspondiente al espacio ocupado por los backups de una empresa, la media es 113,8 Gb y la desviación estándar 18,267 Gb.

Contando solamente con estas dos medidas se pueden obtener los tres intervalos mencionados:

k=1 $\Rightarrow$(95,533 – 132,067)

k=2 $\Rightarrow$ (77,266 – 150,334)

k=3 $\Rightarrow$ (58,999 – 168,601)

Estos tres intervalos tendrían que contener aproximadamente el 68%, 95% y 100% de los datos.

Como en este caso se cuenta con la totalidad de los datos, se pueden obtener los porcentajes reales para este conjunto de observaciones y verificar la regla empírica.

k	$\bar{x} \pm k.s$	Proporción aproximada según la regla empírica	Proporción real de observaciones en el intervalo
1	95,533 – 132,067	68%	67%
2	77,266 – 150,334	95%	100%
3	58,999 – 168,601	100%	100%

Notar que los porcentajes reales son muy parecidos a los que indica la regla empírica. Eso depende principalmente de la forma de la distribución, **si la misma no es campanular y simétrica, los porcentajes serán diferentes.**

2.3.1 Detección de valores anómalos (muy extremos)

Hay ocasiones en que un conjunto de datos contiene observaciones muy alejadas del resto (muy grandes o muy chicas) y puede ocurrir que no se desee incluirlas para su análisis. Dichas observaciones se denominan **valores anómalos u outliers**.

Uno de los métodos para determinar si una observación es un valor anómalo es observar si se aleja demasiado de la media.

El valor “z” de un valor “x” de un conjunto de datos es la distancia a la que se encuentra x de la media, medida en unidades de la desviación estándar.

Valor $z = \frac{x - \bar{x}}{s}$ o $z = \frac{x - \mu}{\sigma}$ si es poblacional.

Luego, para determinar cuándo un valor de z es muy grande, se tiene en cuenta **la regla empírica**, que indica que el 100% (aproximadamente) se encuentra a menos de 3 desvíos de la media. **Esto significa que un valor de z puede considerarse grande si es mayor a 3**. Este criterio no debe considerarse siempre determinante, ya que puede encontrarse en algún caso algún valor de z mayor a 3 y que el mismo no sea anómalo.

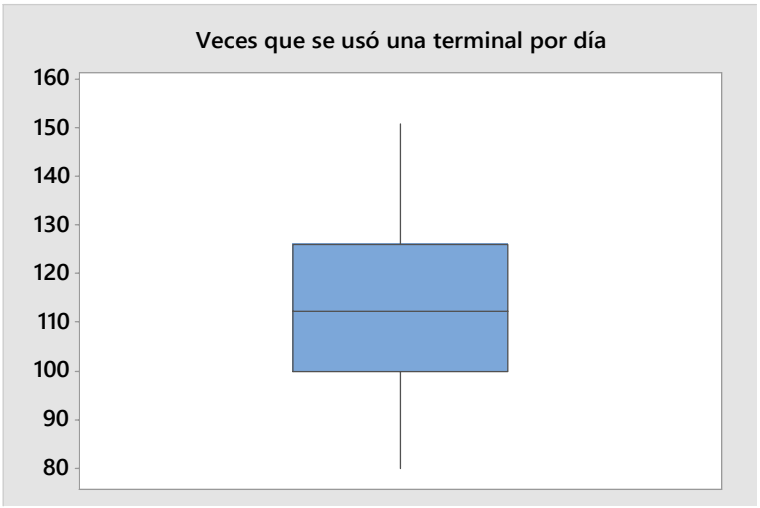
Estos valores por lo general son atribuibles a una de las siguientes causas:

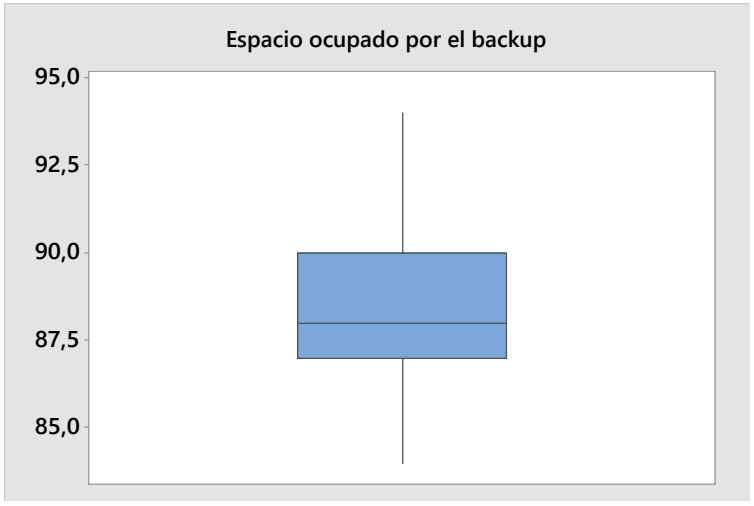
- La observación se registra incorrectamente.
- La observación proviene de una población distinta.
- La observación es correcta pero representa un suceso poco común (fortuito)

2.4 DIAGRAMAS DE CAJA O BOX – PLOT

Es otro gráfico que se utiliza para representar la distribución de los datos de variables cuantitativas. El mismo se realiza a partir de medidas de posición, específicamente las medidas de orden. En estos diagramas se representan los tres cuartiles junto con los valores máximo y mínimo de las observaciones.

Los diagramas de caja que se presentan a continuación, corresponden a los datos observados para las características “cantidad de veces que se utilizó una terminal” y “espacio ocupado por una backup” correspondientes a los ejemplos 2.3 y 2.4 respectivamente.

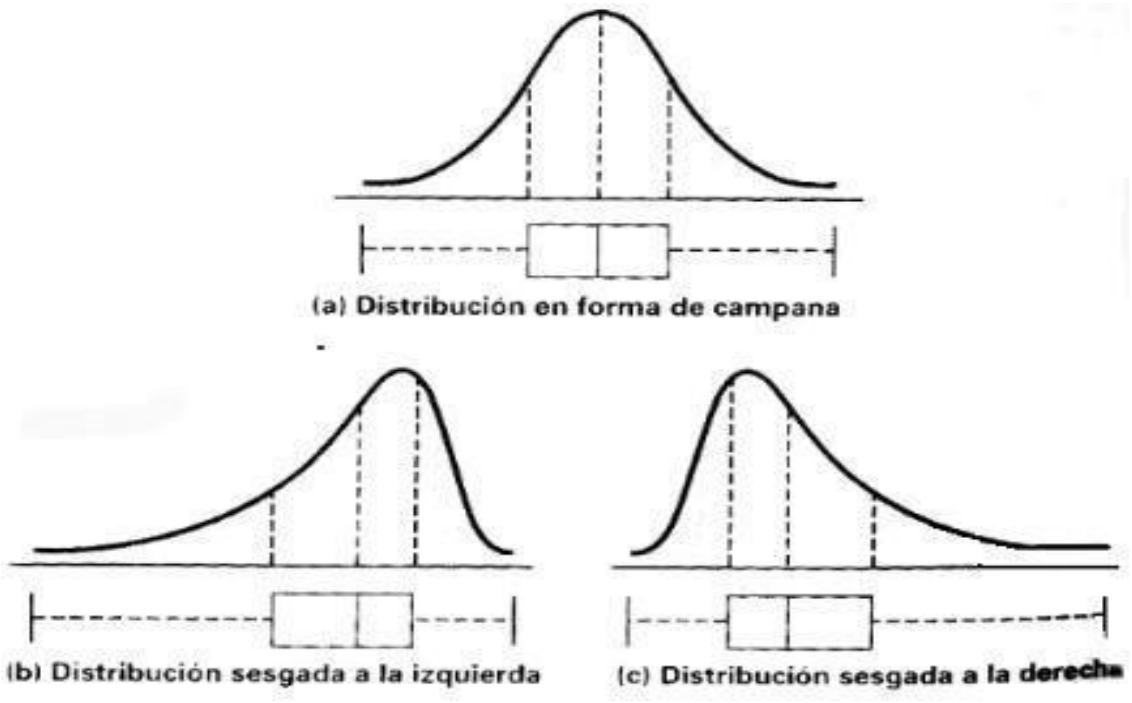




El lado inferior de las cajas (central) corresponde al primer cuartil, el lado superior al tercer cuartil y el segmento que divide a las cajas corresponde a la mediana (segundo cuartil). Un segmento de recta une el lado inferior de las cajas con el mínimo valor observado y otro segmento une el lado superior de las cajas con el máximo valor observado.

La longitud total del diagrama es el rango y la longitud de la caja es el rango intercuartílico.

A continuación se presentan tres casos de distribuciones con distinta simetría donde se visualiza cómo quedaría el diagrama de caja según la forma de la distribución:



Notar que, si dividimos los datos ordenados en cuatro partes cada una con el 25%, los intervalos de variación de cada parte se observan en el diagrama de caja. Estos gráficos constituyen una herramienta eficaz para el análisis de la simetría de una distribución de frecuencias y su estudio comparativo con otras distribuciones.