INTRODUCCIÓN A LA ESTADÍSTICA DESCRIPTIVA

Raúl David Katz

Introducción

Los primeros usos de la estadística significaron la recolección de datos para describir diferentes aspectos de un "estado" o país: tamaño de poblaciones, tasas de natalidad y de mortalidad, ingresos etc. Hoy en día los medios de difusión publican datos del INDEC (Instituto Nacional de Estadísticas y Censos) sobre el valor de la canasta básica para una familia tipo o la variación mensual del empleo en el país.
En estos contextos la palabra estadística hace referencia a la información expresada en forma numérica.
Desde una perspectiva más amplia, la "Estadística" como disciplina se relaciona con las técnicas y los métodos que se han desarrollado para planear experiencias, recopilar, organizar, resumir, analizar, interpretar y comunicar la información proveniente de datos tanto cuantitativos como cualitativos.
Es por ello que la estadística desempeña una función importante en problemas prácticos de diferentes disciplinas.

- ✓ Se realizan encuestas para recabar información previa al día de las elecciones y predecir el resultado de las mismas.
- ✓ Se diseñan experiencias para evaluar los efectos de nuevos tratamientos.
- ✓ Se consideran índices económicos durante un determinado período y se utiliza la información para predecir la situación económica futura.
- ✓ Se observa el consumo de combustible de un vehículo cuando viaja a diferentes velocidades para estudiar la existencia de alguna relación entre ambas variables.
- ✓ Se selecciona al azar una muestra de un lote suministrado por un nuevo proveedor para estimar la proporción de artículos defectuosos, con el objeto de evaluar su calidad.

Una revisión superficial sobre qué es la estadística sugiere una carencia de uniformidad.

Kendall y Stuart afirman: "la estadística es la rama del método científico que se ocupa de los datos obtenidos al observar o medir características o propiedades de alguna población".

Fraser dice: " la estadística trata con métodos para obtener conclusiones a partir de los resultados de experimentos o procesos.

Freund considera a la estadística como algo que abarca el conocimiento relacionado con la toma de decisiones en situaciones de incertidumbre.

Todas estas consideraciones tienen algunos elementos en común. Cada definición implica la recopilación de datos teniendo como objetivo la inferencia. A partir de los datos de una muestra se busca realizar estimaciones, predicciones u otras generalizaciones sobre un conjunto mayor de datos (población).
En los procedimientos de esta naturaleza siempre existe la posibilidad de tomar decisiones erróneas. Nunca podrá tenerse un 100% de confianza cuando se realizan generalizaciones de una muestra a una población. La cuantificación de la confiabilidad de las conclusiones en una población a partir de los datos de una muestra se realiza en términos de probabilidad. De ahí la importancia por comprender los conceptos probabilísticos.

En esta introducción nos hemos referido en forma implícita a tres ejes temáticos: Estadística Descriptiva, Estadística Inferencial y Probabilidad, que serán objeto de nuestro estudio, con diferente intensidad.
Al finalizar el cursado de la asignatura Probabilidad y Estadística, no encontrará todas las respuestas a las situaciones prácticas que le hemos presentado, pero esperamos haber logrado familiarizarlo con un lenguaje y un tipo de pensamiento diferente al habitual, muy ligado al tratamiento de situaciones determinísticas. No es lo mismo preguntarse: ¿durante cuánto tiempo funcionará cierto mecanismo?, que, ¿cuál es la probabilidad de que un mecanismo funcione al cabo de 100 horas?
Le recordamos que uno de los objetivos de la estadística es hacer inferencias con respecto a una población a partir de la información contenida en una muestra y proporcionar una medida de la bondad de dichas inferencias.
Para aproximarnos a ese objetivo iniciaremos el estudio de la Estadística Descriptiva, pero le proponemos previamente indagar acerca de los significados de: Métodos, Técnicas y Método Científico, mencionados anteriormente.

Tres situaciones para empezar

Situación 1

El gerente de operaciones de una planta empacadora desea estudiar las fallas en las cajas de cartón que se utilizan en el proceso de empaque.
Los datos sin procesar que se muestran a continuación corresponden a 50 cajas de cartón falladas, las cuales se tomaron de la producción de una semana.
Notamos con A cartón roto, con B cartón abultado, con C cartón sucio, con D cartón agrietado, con E error de impresión y con F etiqueta ilegible.

Los siguientes datos corresponden a las fallas observadas.

C	B	C	A	E	C	B	D	F	B	D	C	B
C	F	A	B	C	C	D	B	E	F	C	B	B
A	C	D	C	B	C	B	D	F	C	B	E	B
B	B	C	B	E	C	B	D	B	E	B		

Situación 2

Un negocio de artículos para el hogar ha registrado la cantidad de televisores, de cierta marca, vendidos por semana. Los siguientes datos corresponden a las ventas semanales del último año.

6	5	4	6	7	7	6	8	5	7	4	6	6
7	6	5	6	6	5	7	7	4	7	6	5	4
6	7	7	6	5	8	4	7	4	5	5	6	5
6	6	6	4	8	6	5	5	4	6	5	4	6

Situación 3

Un bar de la ciudad tiene una forma específica para preparar un trago muy solicitado.
La fórmula contempla agregar 500 gramos de azúcar. Resulta de suma importancia agregar esa cantidad, ya que de lo contrario, el trago resulta muy dulce o desabrido.
El dueño del bar comprobó que en ocasiones los tragos resultan excesivamente dulces y en otras muy desabridas. Como el azúcar que se utiliza tiene buenos antecedentes de calidad decidió controlar el peso de los contenidos de las bolsas. Los siguientes datos corresponden a los pesos en gramo de 50 bolsas que había en existencia.

470	528	531	518	468	547	499	488	500	512	497	499
457	532	484	508	511	516	502	507	473	489	516	474
540	492	497	519	526	488	471	485	509	478	513	530
503	514	535	530	554	508	469	511	478	494	503	530
486	520										

En cada una de las situaciones presentadas se realizan observaciones de una característica que varía y que resulta de interés.
Interesa conocer:

- ✓ cuáles son las fallas y en particular las más frecuentes en la producción de cajas de cartón, para actuar sobre esas fallas y mejorar consecuentemente el proceso de fabricación,
- ✓ la cantidad de televisores de una cierta marca que se venden por semana para decidir cuántos de esos televisores conviene tener en existencia, con el objeto de satisfacer la demanda en forma inmediata,
- ✓ el peso del contenido de bolsas de azúcar que se utilizan para preparar un trago, pues una variación muy grande con respecto a los 500 gramos generaría tragos muy dulces o desabridos.

Las observaciones de cada una de esas características generan un conjunto de datos. Para que estos datos resulten comprensibles es necesario organizarlos, representarlos gráficamente y definir medidas descriptivas que sinteticen la información.
La parte de la estadística que se relaciona con estos procedimientos se conoce como **estadística descriptiva.**

Como señaláramos en cada una de las situaciones introducidas, existe una característica que varía.
En la situación 1 varía el tipo de falla que puede observarse en una caja de cartón.
En la situación 2 varía la cantidad de televisores que se venden por semana en un negocio.
En la situación 3 varía el peso del contenido de azúcar.

Llamaremos **variable** a toda característica que varía.

En relación a los ejemplos introducidos, las cajas con fallas, las semanas y las bolsas de azúcar constituyen respectivamente las **unidades elementales** sobre las cuales se realizan las observaciones.

Clasificación de variables

- Una variable es **cualitativa** cuando expresa un atributo o cualidad de la unidad elemental que se observa.
- Una variable es **cuantitativa** cuando se expresa numéricamente.

Propuesta

Clasifica las variables de las situaciones introducidas.

Las variables cuantitativas se clasifican en discretas y continuas

❖ Una variable cuantitativa es **discreta** cuando el conjunto de los valores que puede asumir es finito o infinito numerable.

Si observamos la cantidad de azulejos fallados que hay en una caja que contiene cien, entonces la variable cantidad de azulejos fallados en la caja puede tomar los valores de cero a cien. El conjunto { 0, 1, 2, ...,100 } es finito y por lo tanto la variable es discreta.
Si observamos la cantidad de veces que lanzamos simultáneamente los cinco dados de la generala hasta obtener una generala servida, entonces la variable número de lanzamientos hasta obtener una generala servida puede tomar cualquier valor entero no negativo. El conjunto { 1, 2, 3, ...n,...} = N es infinito numerable y por lo tanto la variable es discreta.

En general un conjunto se dice infinito numerable cuando puede ponerse en correspondencia biunívoca con los números naturales.

El conjunto de los números naturales pares es infinito numerable. ¿ Por qué ?

❖ Una variable cuantitativa es **continua** cuando puede tomar cualquier valor real o de un intervalo real.

La variable tiempo que transcurre hasta la falla de una lámpara, desde un punto de vista teórico puede ser cualquier valor real no negativo. Por lo tanto es una variable continua.

Propuesta

1) Clasifica las variables cuantitativas de las situaciones introductorias.
2) Los turistas de un vuelo proveniente de Europa deben completar una ficha con los siguientes datos: nacionalidad, ocupación, grupo sanguíneo, días de permanencia en el país, peso del equipaje, estado civil. Clasifique las variables en cuestión.

3) ¿ Cuáles de las siguientes variables son continuas y cuáles son discretas?

- Número de personas que se atienden en un período de 5 minutos en la ventanilla de un banco.
- Tiempo de atención a un cliente, en la ventanilla de un banco.
- Cantidad de llamadas que se reciben por hora en una central de emergencia.
- Número de autos que llegan a una estación de servicios en el período de una hora para cargar combustible.
- Cantidad de combustible en litros que carga un auto.
- Distancia recorrida por un auto con un litro de nafta.

Organización y representación de datos

En muchas situaciones, la primera tarea que debe emprenderse en el tratamiento estadístico de un conjunto de datos consiste en organizar los mismos en forma de una tabla, a fin conocer la distribución de esos datos. Pero también las representaciones gráficas son fundamentales para visualizar esa distribución y encontrar patrones y/o relaciones.

Cómo ordenar datos en una tabla

Para ordenar datos una de las técnicas más usuales consiste en construir una tabla de frecuencias. Para construir dicha tabla se distribuyen los datos en un número finito de clases y luego se registra la cantidad de datos que aparece en cada una de ellas.
En relación a la situación 1 podemos considerar cada tipo de falla como una clase y constatar, por ejemplo, que 4 de las 50 fallas observadas corresponden a cajas con etiqueta ilegible En este caso decimos que 4 es la frecuencia absoluta de cajas con etiqueta ilegible y $\frac{4}{50}$ es la frecuencia relativa o proporción de cajas con etiqueta ilegible, sobre el total de cajas fallas observadas.

Llamemos con:

C_1 : la clase formada por las cajas con cartón sucio,

C_2 : la clase formada por las cajas con cartón abultado,

C_3 : la clase formada por las cajas con cartón agrietado,

C_4 : la clase formada por las cajas con error de impresión,

C_5 : la clase formada por las cajas con etiquetas ilegibles,

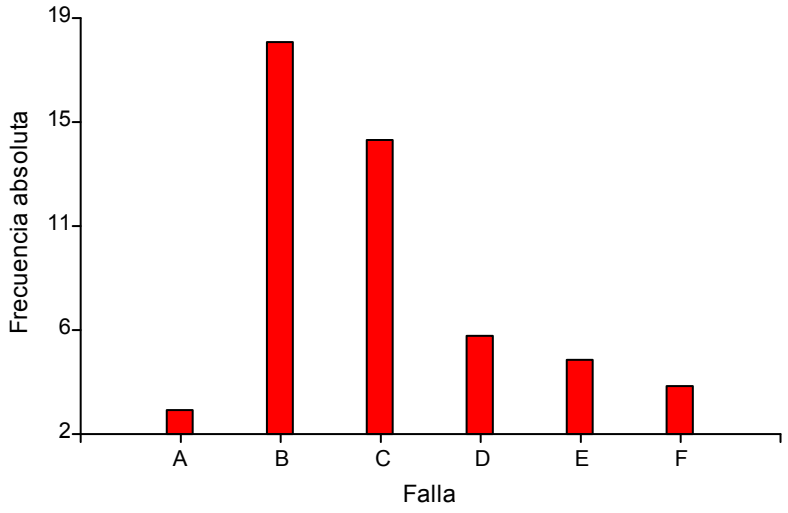
C_6 : la clase formada por las cajas con cartón roto.

Si realizamos el cómputo de cuántas veces se presenta cada tipo de falla obtenemos la siguiente tabla.

Clase	Cómputo de frecuencias	Frecuencias absolutas	Frecuencias relativas	Porcentajes
C_1	xxxxxxxxxxxxxx	14	$\frac{14}{50}$	28%
C_2	xxxxxxxxxxxxxxxxxx	18	$\frac{18}{50}$	36%
C_3	xxxxxx	6	$\frac{6}{50}$	12%
C_4	xxxxx	5	$\frac{5}{50}$	10%
C_5	xxxx	4	$\frac{4}{50}$	8%
C_6	xxx	3	$\frac{3}{50}$	6%

Un primer análisis de los datos, facilitado por la construcción de la tabla, permite observar que las fallas más frecuentes son B y C, es decir, cajas con cartón abultado, que representan un 36% y cajas con cartón sucio que representan un 28%. Entre ambas fallas suman un 64%, de modo que si consideramos que lo observado esa semana, es reflejo de un proceso estable, actuando sobre esas dos causas se resuelve alrededor de la dos tercera parte de las fallas.

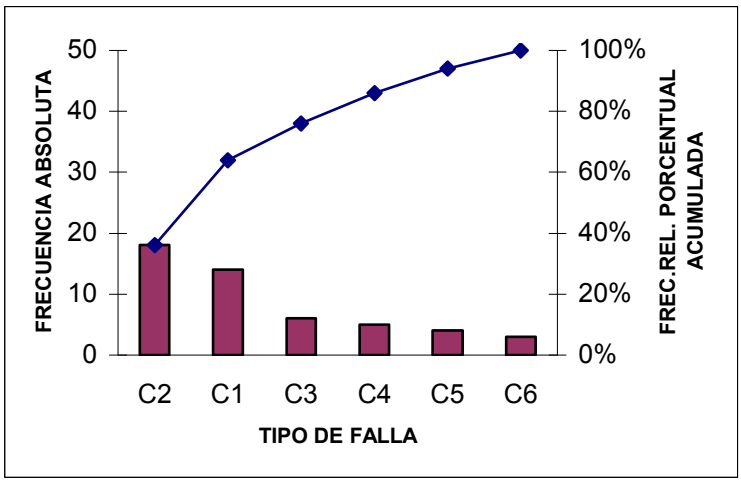
Para representar gráficamente se utiliza un diagrama de barras. Las categorías de la variable (distintas fallas) se representan sobre el eje horizontal, y sobre cada una de ellas se levantan barras de altura proporcional a la frecuencia (absoluta o relativa) o porcentaje correspondiente.



A: cartón roto
B: cartón abultado
C: cartón sucio
D: cartón agrietado
E: error de impresión
F: etiqueta ilegible

Una alternativa es el **diagrama de Pareto** que consiste en un diagrama de barras en que las categorías se ordenan de modo tal que las frecuencias o porcentajes se representan por

orden decreciente. Es útil acompañar el diagrama con una poligonal que muestra las frecuencias o porcentajes acumulados.



El nombre de Pareto fue dado por el Doctor J. Juran en honor al economista italiano Vilfredo Pareto (1848 – 1923) quien realizó un estudio sobre la distribución de la riqueza, encontrando que la minoría de la población poseía la mayor parte de la riqueza. Hoy en día un 20% de la población tiene un 80 % de la riqueza. El Dr. Juran aplicó este concepto a la calidad. Si se tiene un problema con muchas causas, alrededor del 20% de las causas resuelven el 80% del problema. En relación a nuestro ejemplo el 33% de las causas (cartón abultado y cartón sucio) representan el 64% de las fallas.

En relación a la situación 2 la variable discreta cantidad de televisores vendidos por semana asume valores enteros comprendidos entre 4 y 8 (en total 5 valores diferentes). En este caso podemos considerar que cada valor de la variable define una clase. De este modo la clase C_1 queda definida por el valor 4, la clase C_2 por el valor 5 y así sucesivamente.

Si realizamos el cómputo de frecuencias obtenemos la siguiente tabla.

Clase C_k	Valor de la variable x_k	frecuencia absoluta n_k	frecuencia relativa f_k	frecuencia acumulada F_k
C_1	4	9	$\frac{9}{52}$	$\frac{9}{52}$
C_2	5	12	$\frac{12}{52}$	$\frac{21}{52}$
C_3	6	18	$\frac{18}{52}$	$\frac{39}{52}$
C_4	7	10	$\frac{10}{52}$	$\frac{49}{52}$
C_5	8	3	$\frac{3}{52}$	$\frac{52}{52}$
Suma		52	1	

Hemos notado con,
x_k : valor de la variable que define la clase C_k , para k=1,2,....5.
n_k : frecuencia absoluta de la clase C_k

f_k : frecuencia relativa de la clase C_k , donde $f_k = \frac{n_k}{n}$ y n el total de las observaciones

F_k : frecuencia relativa acumulada hasta *la clase* C_k . ($F_k = f_1 + f_2 + \cdots\cdots f_k$)

Por ejemplo, $F_3 = \frac{39}{52}$, significa que en el $\frac{39}{52}$. *100 = 75%* de las semanas se vendieron a lo sumo 6 televisores.

Asimismo observemos las siguientes propiedades de las frecuencias:

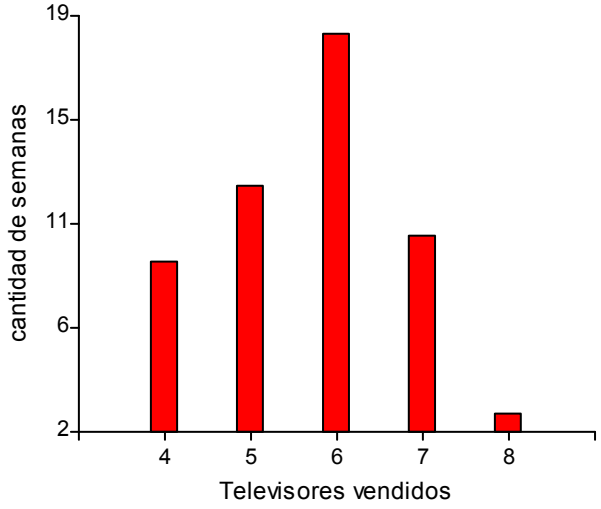
Si se tienen r clases entonces,

- $n_1+n_2+ \cdots\cdots n_r = n$, $\left(\sum_i^r n_i = n \right)$. La suma de las frecuencias absolutas de las r clases es igual al total de datos.
- $f_1 + f_2 + \cdots\cdots f_r = 1$, $\left(\sum_i^r f_i = 1 \right)$. La suma de las frecuencias relativas de las r clases es igual a 1.

Para la representación se utiliza una gráfica de bastones.
En el eje horizontal se representan los valores de la variable y en el eje vertical las correspondientes frecuencias absolutas o relativas. Sobre cada valor de la variable se traza un bastón cuya longitud es proporcional a la frecuencia de dicho valor. Se obtiene de este modo la **gráfica de la distribución de frecuencias absolutas o relativas.**

Los conjuntos de pares ordenados $\{(x_k , n_k)\}$ y $\{(x_k , f_k)\}$, con $k = 1, 2 , \cdots\cdots, r$ constituyen las **distribuciones de frecuencias absolutas y relativas** respectivamente.

La siguiente gráfica corresponde a la distribución de frecuencias absolutas



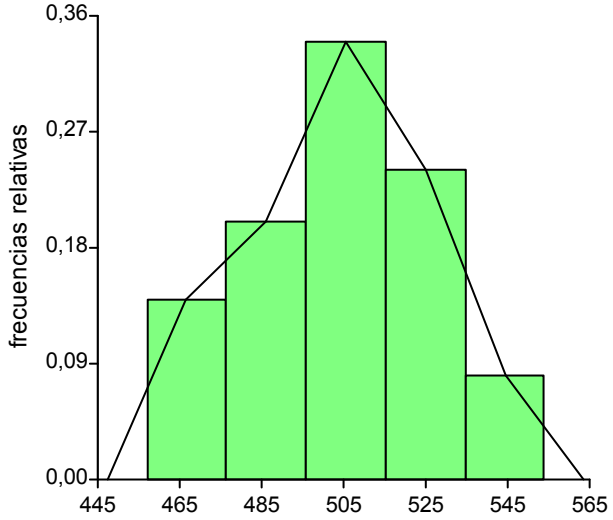
Para ordenar en tabla los datos correspondientes a una variable continua se procede de la siguiente manera. Se busca el mínimo, x_m , y el máximo, x_M, de los valores. Para la situación 3, $x_m = 457$ gramos y $x_M = 554$ gramos. Conocidos el mínimo y el máximo sabemos que los restantes valores de la variable se encuentran en el intervalo $[x_m , x_M]$. Interesa conocer cómo se distribuyen esos valores en dicho intervalo o en un intervalo que lo contenga. A tal fin agruparemos los datos en intervalos adyacentes, de modo que cada dato pertenezca a uno y solo uno de esos intervalos.
Por comodidad particionamos el intervalo [455 , 555) en cinco intervalos de igual amplitud. A cada uno de esos intervalos los llamaremos **intervalo de clase.**

Una vez definidos los intervalos procedemos a realizar el cómputo de frecuencias, es decir, contamos la cantidad de datos que pertenecen a cada intervalo y confeccionamos la tabla con las frecuencias absolutas, relativas y acumuladas. Asimismo destacamos el punto medio de cada intervalo.

Intervalo de clase I_k	Punto medio x_k	Frecuencias absolutas n_k	Frecuencias relativas f_k	Frecuencias acumuladas F_k
[455 ; 475)	465	7	0.14	0.14
[475 ; 495)	485	10	0.20	0.34
[495 ; 515)	505	17	0.34	0.68
[515 ; 535)	525	12	0.24	0.92
[535 ; 555)	545	4	0.08	1.00

Para la representación gráfica de la distribución de los datos utilizaremos un **histograma de áreas y el polígono de frecuencias relativas.**

El punto de partida para graficar el histograma es la tabla de frecuencias. Sobre el eje horizontal se representan los extremos de los intervalos de clase y sobre cada uno de ellos se construye un rectángulo de área igual a la frecuencia relativa de cada clase. Si los intervalos tienen igual amplitud entonces las alturas de los rectángulos son proporcionales a las frecuencias.

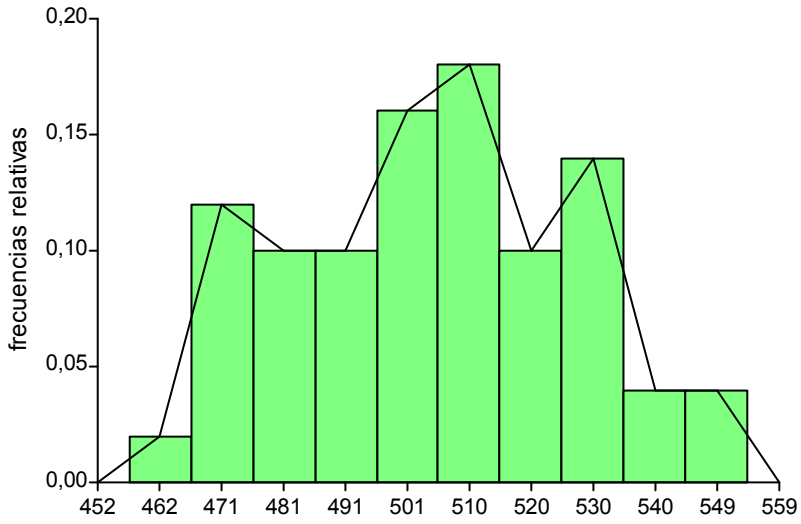


Para recordar:

- Lo importante de un histograma son las áreas de los rectángulos.
- El área de cada rectángulo representa la proporción de datos de cada intervalo de clase.
- El área total que encierra el histograma es igual a uno.
- El área comprendido entre dos valores cualesquiera de la variable es indicador de la proporción de datos que se encuentran en el intervalo delimitado por esos valores.

La forma de un histograma depende del número de intervalos de clase que se consideren. Cuando se emplean "pocos" intervalos o "demasiados" intervalos de clase la visualización del histograma no ofrece buena información. En el primer caso no se discrimina convenientemente la distribución de los datos y en el segundo de los casos no se alcanza a lograr un patrón de la distribución de los mismos. En la práctica se acostumbra seleccionar el número de intervalos aproximadamente igual a la raíz cuadrada del número de observaciones.

Cuando se consideran 10 intervalos de clase el histograma (para los datos de la situación 3) toma la siguiente forma:



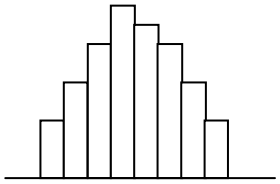
Cada histograma se acompaña con el **polígono de frecuencias relativas** que se obtiene uniendo los puntos medios de las bases superiores de los rectángulos y se completa como lo muestra la figura.
El polígono se construye de modo que el área que encierra es igual al área del histograma y constituye una alternativa para visualizar la distribución de los datos de una variable continua.

También suele ser útil el **polígono de frecuencias relativas acumuladas**.
Mostramos su construcción para el caso de variables continuas.

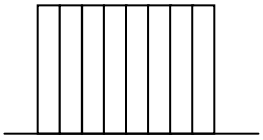
Sobre el eje horizontal se marcan los puntos extremos de los intervalos de clase y sobre el eje vertical las frecuencias relativas acumuladas. El origen del polígono coincide con el extremo inferior del primer intervalo de clase. La ordenada del polígono de frecuencias acumuladas correspondiente a un valor cualquiera de la variable es igual al área encerrada por el histograma hasta ese valor de la variable. De este modo cada ordenada mide la proporción de los datos que son menores o iguales a ese valor.

Las figuras que siguen son algunas formas de histogramas que pueden presentarse respondiendo a diferentes comportamiento de los datos

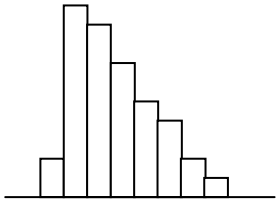
Teniendo en cuenta que los histogramas muestran información, es interesante observar las distintas formas que pueden tomar de acuerdo al grupo de datos que representan



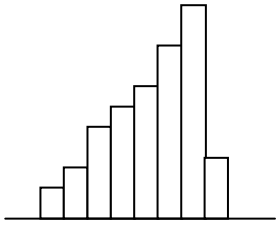
FORMA NORMAL O SIMÉTRICA



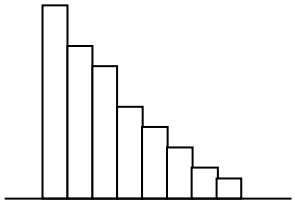
UNIFORME



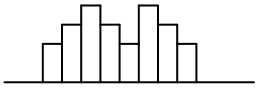
SESGADO HACIA LA DERECHA



SESGADO HACIA LA IZQUIERDA



FORMA DE "JOTA"



BIMODAL

El histograma "bimodal", con dos máximos diferenciados, se presenta cuando se mezclan datos de distinto origen "centrados en valores distintos".

Propuesta

Asocia un histograma con:

a) la distribución de ingresos en un país donde hay muchos pobres y pocos ricos.
b) la distribución de ingresos en un país donde hay muchos ricos y pocos pobres.
c) la distribución de las alturas de los alumnos que cursan el séptimo año del tercer ciclo de la EGB. con los alumnos que cursan el tercer año de la Educación Polimodal, correspondientes a una escuela.

Para la interpretación gráfica de la información, también suele ser útil el **polígono de frecuencias relativas acumuladas**.

Mostramos su construcción para el caso de una variable continua.

El polígono de frecuencias relativas acumuladas se obtiene teniendo en cuenta las frecuencias acumuladas de cada clase, que podemos visualizar mediante rectángulos. Sobre el eje horizontal se marcan los puntos extremos de los intervalos de clase y sobre el eje vertical las frecuencias relativas acumuladas. El origen del polígono coincide con el extremo inferior del primer intervalo de clase. Los restantes vértices tienen por abscisa los extremos de cada uno de los intervalos y por ordenada la frecuencia acumulada hasta dicho valor.

Observemos que la ordenada del polígono de frecuencias acumuladas correspondiente a un valor cualquiera de la variable es igual al área encerrada por el histograma hasta ese valor de la variable. De este modo cada ordenada mide la proporción de los datos que son menores o iguales a ese valor.

Para la situación 3 el polígono de frecuencias acumuladas resulta:

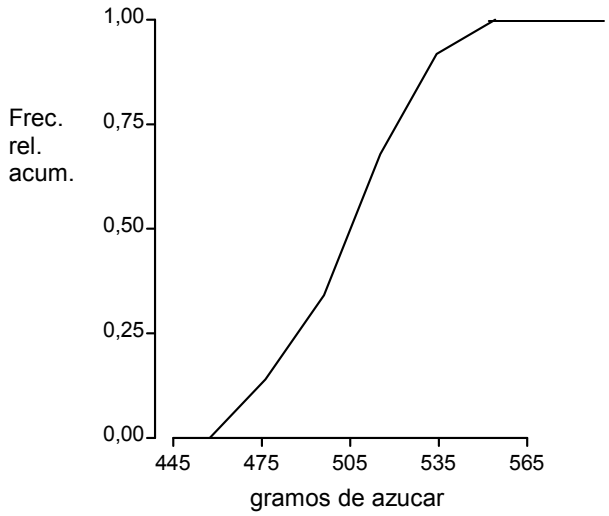


Diagrama de tallo y hoja.

Desde el enfoque del análisis exploratorio de datos, se han ideado una serie de gráficas apropiadas para estudiar la estructura de los datos.
Uno de estos gráficos exploratorios, alternativo del histograma, es el diagrama de tallo y hoja.
Explicamos su construcción utilizando los datos correspondientes a la situación 3.
Se construye una columna (el tallo) con las centenas y decenas de los datos. Cada renglón se completa con las unidades correspondientes (las hojas).

45	7
46	8 9
47	0 3 4 1 8 8
48	8 4 9 8 5 6
49	9 7 9 2 7 4
50	0 8 2 7 9 3 8 3
51	8 2 1 6 6 9 3 4 1
52	8 6 0
53	1 2 0 5 0 0
54	7 0
55	4

El diagrama de tallo y hoja resulta más informativo que el histograma ya que conserva los datos originales y al mismo tiempo permite visualizar la forma en que se distribuyen los datos.

Los conceptos de población y muestra

En el párrafo introductorio decimos que a partir de los datos de una muestra se busca realizar estimaciones, predicciones u otras generalizaciones sobre un conjunto mayor de datos (población). En lo que sigue definimos los conceptos de población y muestra.

Llamamos **población estadística** al conjunto formado por todos los resultados de las observaciones posibles en relación a un objetivo prefijado.

Llamamos **muestra** a un subconjunto finito de la población.

Para comprender mejor veamos los siguientes ejemplos.
Si se desea estudiar a qué distancia, medida en cuadras, viven los alumnos que concurren a la Facultad Regional Rosario, los datos que se obtienen al considerar a todos los alumnos constituyen la población estadística. Cabe destacar que cada alumno es la unidad elemental sobre la cual se realiza la observación y el conjunto de todos los alumnos conforman la población física.
Si solo se consideran los datos de los alumnos de una especialidad, por ejemplo los que cursan Ingeniería Mecánica, teniendo en cuenta el objetivo, estos datos constituyen una muestra de la población recién definida.
Si el objetivo fuera estudiar a qué distancia viven los alumnos de esa especialidad, entonces los datos que se obtendrían con los alumnos de esa especialidad, constituirían mi población en estudio.

De este modo, un conjunto de datos constituye una población o una muestra según el objetivo que se plantea.

La cantidad de datos que conforman una muestra o una población se denomina tamaño de la muestra o población respectivamente.

Propuesta

En relación a las situaciones introducidas, le proponemos plantear diferentes posibilidades de modo que los datos dados correspondan a una muestra o a una población. En cada caso explicite el tamaño.

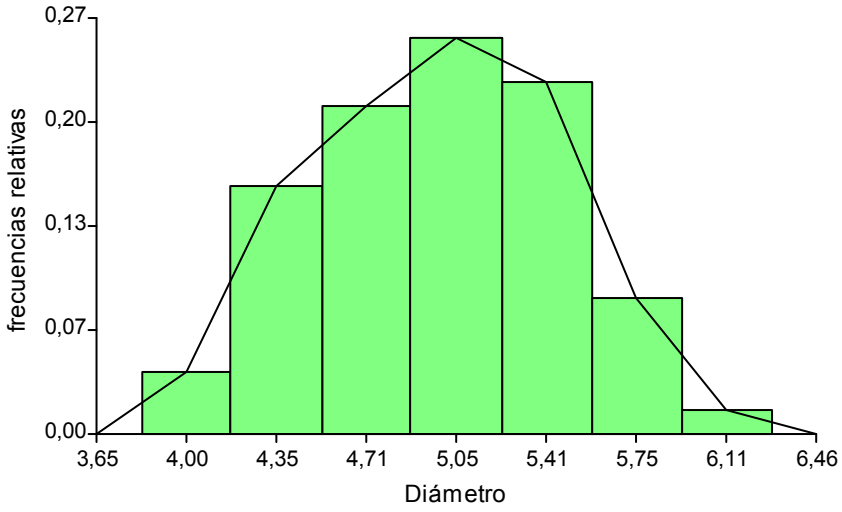
Existen poblaciones que no son finitas. Si consideramos el conjunto de los resultados de las observaciones que teóricamente podrían realizarse si se observara indefinidamente el diámetro de las tuercas producidas por un proceso, obtendríamos una población infinita.

Los siguientes datos constituyen una muestra de 400 observaciones de esa población teórica

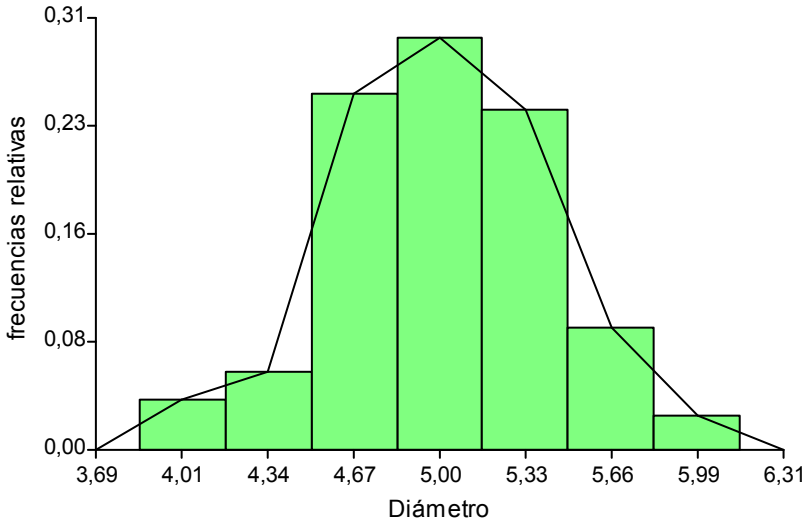
5,16	5,37	4,75	5,16	5,61	5,40	5,30	4,81	4,49	4,82
4,97	4,59	5,42	4,75	4,54	5,50	5,23	4,85	5,20	5,41
5,34	4,28	4,47	5,19	5,47	5,28	5,34	4,92	5,07	5,41
5,38	4,78	5,60	5,15	4,19	4,90	5,70	4,73	5,05	5,63
5,26	4,64	4,84	5,27	4,72	4,40	4,78	4,66	4,72	5,11
5,04	5,16	5,45	4,81	4,45	4,67	4,36	5,40	4,62	5,01
4,62	5,60	5,34	5,14	4,57	5,10	5,35	5,49	4,07	4,89
4,21	4,77	4,94	5,57	4,22	5,29	4,72	5,40	5,28	5,21
4,86	5,02	5,46	4,75	5,29	5,27	4,91	5,23	5,95	4,59
4,69	5,58	5,12	5,49	5,70	5,31	4,52	5,05	5,05	5,29
5,07	4,98	5,09	4,91	4,56	4,59	4,86	4,22	4,53	5,22
4,45	4,93	4,15	4,76	4,66	4,96	4,58	5,21	5,08	4,79
4,73	5,01	4,39	4,80	5,14	5,44	5,00	5,03	4,73	5,08
5,21	4,34	3,91	5,16	5,45	4,96	4,64	4,72	5,10	4,11
4,76	5,01	4,93	5,40	4,44	5,15	4,29	4,21	5,45	4,80
5,97	4,15	4,47	5,70	5,86	5,69	4,33	4,56	4,89	4,10
4,31	4,52	5,14	4,49	5,66	4,76	4,88	4,66	5,03	4,84
4,96	5,60	4,76	4,64	4,73	5,33	5,23	5,63	4,14	5,25
4,86	5,27	5,29	5,53	6,28	4,44	5,21	4,97	4,91	4,83
4,37	4,72	5,07	4,27	4,34	5,16	5,96	4,54	4,85	5,03
4,78	5,05	4,79	5,51	5,39	5,02	4,28	4,85	4,70	5,10
5,63	5,26	5,25	5,25	4,81	6,12	5,68	4,97	4,93	4,14
5,53	4,58	5,22	5,48	5,24	4,85	5,64	4,80	5,08	5,53
5,41	4,60	5,16	4,22	3,83	4,82	5,02	4,55	5,33	5,34
5,87	5,29	4,53	4,60	4,40	6,15	5,94	5,42	5,05	4,73
5,26	4,68	4,72	4,52	4,36	5,16	5,69	5,02	5,21	5,53
4,01	4,89	5,13	5,08	5,34	5,34	4,77	5,53	5,19	5,25
5,39	4,97	4,90	4,67	4,74	5,96	4,62	5,32	4,70	5,18
4,95	4,51	5,04	4,20	5,37	5,14	5,07	5,15	5,67	4,84
5,52	5,36	4,22	5,42	4,95	5,41	5,07	4,81	5,74	3,85
4,94	5,46	4,10	4,31	5,99	5,04	4,79	3,90	5,14	4,99
5,27	4,29	5,61	5,43	4,46	4,72	5,30	5,18	5,40	5,16
5,53	5,17	5,15	4,15	4,29	5,15	5,51	5,31	4,81	5,26
4,45	4,89	4,81	4,78	5,27	5,37	4,46	5,37	4,77	5,40
4,99	5,55	4,85	5,66	5,31	4,69	4,81	5,08	5,14	5,70

5,71	4,91	5,41	5,65	5,70	5,48	5,07	4,58	5,27	5,43
5,29	4,66	4,44	5,06	5,00	4,79	5,66	4,70	5,48	5,15
4,91	4,80	5,28	4,49	5,10	4,84	5,12	5,42	4,79	4,48
3,91	5,63	4,72	4,91	4,78	4,78	5,05	5,26	5,00	4,74
5,58	5,00	5,08	5,08	5,40	4,66	5,49	5,58	4,94	5,20
5,70	6,00	5,12	4,02	5,11	4,81	4,54	4,63	5,08	4,96
5,68	5,08	5,41	4,76	4,69	4,93	4,78	5,05	4,25	3,85
4,56	5,40	5,23	5,45	4,83	5,65	5,29	5,15	4,89	4,74
5,18	4,91	5,32	5,06	5,27	4,55	5,21	5,01	5,39	4,51
5,22	5,38	5,73	5,50	4,95	4,99	4,26	4,73	4,69	3,99
4,72	4,70	4,32	4,93	4,38	4,78	5,73	5,10	4,45	5,80
5,83	5,26	5,31	4,94	4,81	5,10	4,99	5,12	4,78	4,80
3,99	5,22	4,45	4,46	5,76	4,57	5,61	4,65	5,64	4,97
4,54	4,21	4,75	5,34	4,52	4,81	4,60	5,21	4,63	5,28
4,29	4,59	5,13	4,40	5,08	5,17	5,00	5,29	5,32	4,63

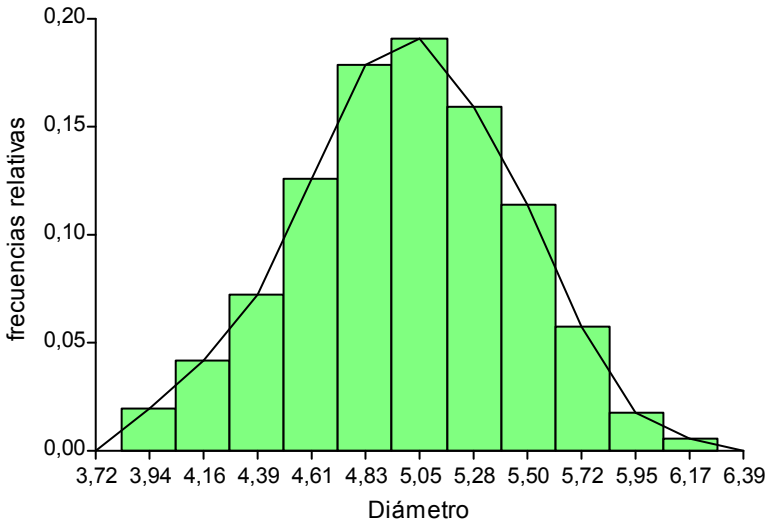
La distribución de frecuencias de las primeras 200 observaciones se visualiza a través del siguiente histograma



La distribución de frecuencias de las siguientes 200 observaciones se visualiza a través del siguiente histograma



El siguiente histograma y polígono de frecuencias relativas corresponde a las 400 observaciones.



El polígono de frecuencias relativas en la medida que se aumenta el número de observaciones y consecuentemente el número de intervalos de clase, tiende a una "curva suave" que denominamos curva de densidad y que describe la distribución de los diámetros de las tuercas en la "población infinita". En el caso particular del ejemplo que venimos estudiando esa curva toma la forma de una campana simétrica. El área que encierra una curva de densidad, a igual que el área encerrada por un histograma o un polígono de frecuencias relativas, es igual a uno. Asimismo el área encerrada por la curva de densidad sobre un intervalo (a , b), que se expresa mediante una integral definida, representa la probabilidad de que una tuerca elegida al azar tenga un diámetro comprendido entre a y b. Por ejemplo, si ese área es igual a 0,20, esto significa que en un "número grande" de observaciones de diámetros, alrededor del 20% de los mismos se encuentran en ese intervalo.
En las siguientes unidades estudiaremos diferentes curvas de densidad, las más usuales en las aplicaciones de la ingeniería.

Valores característicos

Cuando se tiene un número finito de datos, ya sea de una muestra o de una población, no sólo interesa tabular y representar gráficamente la información, también importa resumirla a través de valores numéricos (caso de las variables cuantitativas) que pudieran caracterizar al conjunto de datos y revelar algunas de sus particularidades esenciales.

Llamamos **parámetros** a las características numéricas de una población.

Llamamos **estadísticos** a las características numéricas de una muestra.
Se acostumbra notar con letras griegas a los parámetros y con letras latinas a los estadísticos.

Los valores que se utilizan con mayor frecuencia para resumir la información de un conjunto de datos son los que se refieren a la tendencia central o localización y los de variabilidad o dispersión. Hay diferentes formas de medir estas características.
La siguiente tabla muestra los valores más usuales y que pasamos a considerar.

Valores característicos	
De tendencia central	De variabilidad
Media	Desviación estándar
Mediana	Variancia
Moda	Recorrido o rango
	Recorrido intercuartílico
	Coeficiente de variación

Convengamos en que si tenemos un conjunto finito de n datos, escribimos $x_1, x_2, ...x_n$ cuando corresponden a una muestra de tamaño n, y $x_1, x_2, ...x_N$ cuando corresponden a una población finita de tamaño N.
(x_1 denota el primer dato, x_2 el segundo, y así sucesivamente.)

La media

Los siguientes datos corresponden a la antigüedad (en años) de todos docentes que se desempeñan en una división: 10, 9, 9, 4, 9, 4, 15, 11, 19.

La antigüedad media de los docentes es $\frac{(10+9+9+4+9+4+15+11+19)}{9} = 10$ años

Si el objetivo es evaluar la edad media de los docentes de esa división, los datos que se tienen corresponden a una población finita de tamaño N = 9.
En este caso la media calculada se denomina media poblacional y se nota con la letra griega μ (se lee mu) , $\mu = 10$ años.
En cambio si se utilizan estos datos para estimar la antigüedad media de todos los docentes que trabajan en la facultad, la media calculada correspondería a una muestra de tamaño n=9. En este caso haremos referencia a la media muestral, que se nota con $\bar{x}$. Desde esta perspectiva $\bar{x} = 10$ años.

Le proponemos que reflexione sobre cuál es la población estadística correspondiente en cada caso del ítem anterior.

Hemos hecho referencia a la media poblacional (parámetro) y a la media muestral (estadístico) incurriendo en ambos casos en un abuso del lenguaje.

La media a la que nos estamos refiriendo es la media aritmética. Existen otras medias. Más adelante haremos referencia a la media geométrica.

En general:

Si $x_1, x_2, ...x_n$ es una muestra de tamaño n entonces la media muestral o media aritmética es $\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$

Si $x_1, x_2, ...x_N$ es una población finita de tamaño N entonces la media poblacional

$$\mu = \frac{x_1 + x_2 + ... + x_N}{N}$$

Cuando los datos se presentan en forma de una distribución de frecuencias ya sean absolutas o relativas: (x_k, n_k) o (x_k, f_k) con k = 1,2,.....r, entonces según corresponda a una muestra o a una población resulta

$$\bar{x} = \frac{1}{n} \sum_{1}^{r} x_k.n_k = \sum_{1}^{r} x_k.f_k \quad o \quad \mu = \frac{1}{N} \sum_{1}^{r} x_k.n_k = \sum_{1}^{r} x_k.f_k$$

donde $\sum_{1}^{r} n_k = n$ o N y $\sum_{1}^{r} f_k = 1$

De ahora en más convengamos en considerar, salvo que se enuncie lo contrario, que los datos corresponden a una muestra.

Le proponemos verificar en relación a las situaciones introducidas que:

- el promedio (media aritmética) de televisores vendidos por semana es $\bar{x}$ = 5,73 televisores
- el peso promedio de las bolsas de azúcar, calculado a partir de los datos agrupados en intervalos de clase es $\bar{x}$ = 503,4 kilogramos. (En este caso se considera x_k punto medio del intervalo de clase C_k y n_k la frecuencia absoluta de dicho intervalo.)
  Este promedio así calculado difiere ligeramente del promedio que se obtendría de considerar los datos inicialmente dados. El agrupamiento de los datos en intervalos de clases favorece el análisis de la distribución de los mismos, pero genera pérdida de información cuando se calculan los valores característicos.

La mediana

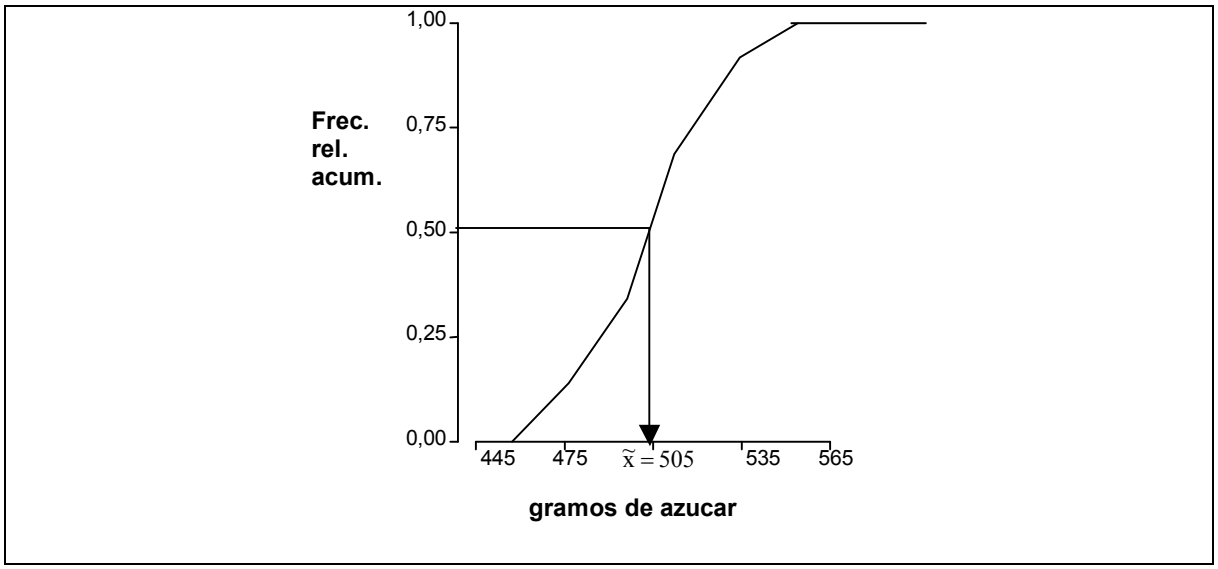
Si consideramos nuevamente los datos correspondientes a las antigüedades de los docentes y los ordenamos de menor a mayor (4, 4, 9, 9, 9, 10, 11, 15, 19) observamos que el valor central del ordenamiento (el quinto) es igual a 9. Decimos que el valor 9 es la

mediana del conjunto de datos y escribimos $\tilde{x}$ = 9 años (mediana muestral) o $\tilde{\mu}$ = 9 años (mediana poblacional).
Si hay un número par de datos, la mediana se calcula promediando los dos valores centrales.

Le proponemos que verifique que la mediana para el número de televisores vendidos por semana es igual a seis.

Observe además que: $F_2 = \frac{21}{52} = 0.40 < 0.50$, $F_3 = \frac{39}{52} = 0.75 > 0.50$. A partir de estas consideraciones puede determinarse que $\tilde{x}$ = 6 televisores. ¿Por qué?

Cuando los datos corresponden a una variable continua y se encuentran agrupados por intervalos de clase el valor de la mediana se obtiene en forma aproximada a partir del polígono de frecuencias acumuladas en la forma que se indica.



Propuesta

Compare el valor 5.05 con el que se obtiene promediando los valores centrales de los datos originales.
El uso del diagrama de tallo y hoja (con los números de cada renglón ordenados en forma creciente) le facilitará la tarea.

La moda

Llamaremos moda al valor de la variable que se presenta con mayor frecuencia

En relación a la antigüedad de los docentes la moda es 9 años y notaremos $\hat{x} = 9$ años o $\hat{\mu} = 9$ años según se consideren los datos correspondientes a una muestra o a una población.
En relación a la situación 1 la característica que se da con mayor frecuencia es A (cartón roto) y por lo tanto constituye la moda.

¿Cuál es el valor de la moda en relación a la cantidad de televisores vendidos por semana?

Cuando los datos se encuentran agrupados en intervalos de clase de igual amplitud, llamaremos intervalo modal al intervalo de mayor frecuencia. En relación al peso de las bolsas de azúcar el intervalo [495;515) es el intervalo modal. En ese intervalo la "densidad de frecuencia" es máxima; interesa considerar la cantidad de datos que hay en el intervalo en relación a su amplitud.

Algunas observaciones:

Si por ejemplo, consideráramos el ingreso medio de los grupos familiares de los alumnos de un curso, obtendríamos un valor que se modificaría significativamente si incorporáramos a los datos el ingreso de Bill Gate. En este caso la media aritmética dejaría de ser un valor apropiado para caracterizar la tendencia central, resultando la mediana más adecuada a tal fin.
Si bien la media aritmética es el valor más usual para caracterizar la tendencia central tiene la desventaja de ser "sensible" a los valores extremos (valores muy grandes o pequeños en relación a los restantes datos). Por otra parte, la media aritmética se determina involucrando en su cálculo todos los datos. En cambio el valor de la mediana depende únicamente del valor central, constituyendo este aspecto una desventaja respecto de la media aritmética.
Asimismo la media aritmética y la mediana no están definidas para datos correspondientes a una variable cualitativa. De ahí la importancia de la moda.

Propuesta

Si se reemplazara en los datos correspondientes a las antigüedades de los docentes, el valor máximo 19 por 30, ¿cuál de los valores característicos: media, moda, mediana, se modificaría?

Valores característicos de la variabilidad

Por lo general los valores característicos de tendencia central no proporcionan suficiente información para una adecuada descripción de los datos.
Consideremos por ejemplo, las calificaciones trimestrales en Matemática de tres alumnos: Andrés, Ignacio, Gabriela.

Andrés:	10	7	4
Ignacio:	5	9	7
Gabriela:	7	6	8

La media aritmética de las calificaciones en los tres casos es 7. Sin embargo las calificaciones de Andrés presentan mayor variación con respecto a la media que las calificaciones de Ignacio y estas a su vez tienen mayor variación, con respecto a la media, que las calificaciones de Gabriela.
¿Cómo medir esa variación con respecto a la media?
En un primer intento parecería razonable promediar las desviaciones con respecto a 7.
Si realizamos los cálculos obtenemos cero en los tres caso.
En relación a las calificaciones de Andrés tendríamos: $\frac{(4-7)+(7-7)+(10-7)}{3}=0$
Realice los cálculos con las calificaciones de Ignacio y de Gabriela.

Le proponemos probar en general que si $x_1, x_2, ...x_N$ es una población finita de tamaño N con media ($\mu = \frac{1}{N}\sum_{1}^{N} x_i$) entonces $\sum_{1}^{N} (x_i - \mu) = \frac{1}{N}\sum_{1}^{N} (x_i - \mu) = 0$
Igual resultado se verifica cuando los datos corresponden a una muestra.

Si calculáramos la media de las desviaciones absolutas, $\frac{1}{N}\sum_{1}^{N} |x_i - \mu|$, o las medias de las desviaciones al cuadrado, $\frac{1}{N}\sum_{1}^{N} (x_i - \mu)^2$, obtendríamos valores que describen de diferente manera la mayor o menor variación respecto de la media.
Le proponemos verificar que los valores $(2, \frac{4}{3}, \frac{2}{3})$ y $(6, \frac{8}{3}, 1)$ son respectivamente las desviaciones medias absolutas y al cuadrado de las calificaciones de Andrés, Ignacio y Gabriela.
Para cuantificar la variabilidad de los datos con respecto a su media priorizaremos las desviaciones medias al cuadrado.

Variancia y desviación estándar

Si $x_1, x_2, ...x_N$ es una población finita de tamaño N se define:

- la **variancia poblacional** y se nota con σ^2 (se lee sigma al cuadrado) al número positivo:

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

- la **desviación estándar poblacional** como σ (raíz cuadrada de la variancia).

Si $x_1, x_2, ...x_n$ es una muestra de tamaño n se define:

- la **variancia muestral** y se nota con s^2 al número positivo:

$$s^2 = \frac{1}{n-1} \sum_1^n (x_i - \bar{x})^2 \text{ donde } \bar{x} = \frac{1}{n} \sum_1^n x_i$$

- **la desviación estándar muestral** como s (raíz cuadrada de la variancia muestral).

Cuando los datos se presentan en forma de una distribución de frecuencias ya sean absolutas o relativas : (x_k , n_k) o (x_k, f_k) con k = 1, 2,.....r, entonces

$$s^2 = \frac{1}{n-1} \sum (x_k - \bar{x})^2 \, n_k \quad \text{o} \quad s^2 = \frac{n}{n-1} \sum (x_k - \bar{x})^2 \, f_k$$

Tanto la variancia como la desviación estándar, ya sea poblacional o muestral, caracterizan la variación de los valores de la variable respecto de su media. Si una variable asume "frecuentemente" valores alejados de su media, tanto la variancia como la desviación estándar resultan grande. La ventaja de la desviación estándar radica en que se expresa en las mismas unidades que la variable.

Le proponemos que reflexione sobre el siguiente argumento: si la desviación estándar en una población es pequeña, bastan unos pocos datos de la misma para estimar con buena precisión la media poblacional a través de la media muestral.

Algunos cálculos

La variancia muestral del número de televisores vendidos por semana es :

$$s^2 = \frac{1}{52-1}[(4 - 5.73)^2.9 + (5 - 5.73)^2.12 + (6 - 5.73)^2.18 + (7 - 5.73)^2.10 +$$
$$+ (8\text{-}5.73)^2.3] = 1.298 \text{ televisores}^2 \quad \text{y} \quad s = 1.14 \text{ televisores.}$$

Para calcular la variancia muestral del peso de las 50 bolsas de azúcar, a partir de los datos agrupados en intervalos de clase debe tomarse como x_k el punto medio del intervalo. De este modo se obtendrá $s^2 = 536$ (grs.)^2 s = 23.15 grs.
Estos valores difieren ligeramente de los que se obtendría tomando lo 50 datos.

El coeficiente de variación

En general es difícil hacer una interpretación de los valores de la variancia y la desviación estándar en razón de que los mismos dependen de las unidades de medida.

Consideremos los siguientes datos correspondientes a las alturas y pesos de los jugadores titulares de un equipo de basquetball.

Alturas: 1.98, 2.10, 2.05, 1.85, 1.90
Pesos: 92, 96, 98, 88, 92
Si calculamos el promedio y la desviación estándar poblacional de las alturas y de los pesos obtenemos:
$\mu_A = 1.976$ m $\quad \mu_P = 93.2$ kg.
$\sigma_A = 0.092$ m $\quad \sigma_P = 3.487$ kg.

Un primer análisis nos permite observar que la desviación estándar de los pesos es mayor a la desviación estándar de las alturas. Sin embargo, si expresamos esas desviaciones como una fracción de sus respectivas medias obtenemos:

$\frac{\sigma_A}{\mu_A} \cong 0.046 \qquad \frac{\sigma_P}{\mu_P} \cong 0.037$

Esto significa que σ_A representa el 4.6 % de μ_A mientras que σ_P representa solamente un 3.7% de μ_P.
Desde esta perspectiva la desviación estándar de los pesos es relativamente menor que la desviación estándar de las alturas. En este caso diremos que los datos correspondientes a los pesos de los jugadores presentan mayor homogeneidad que las alturas.

En general los cocientes $\frac{\sigma}{\mu}$ o $\frac{s}{\bar{x}}$ se denominan coeficientes de variación poblacional y muestral respectivamente.

Cabe destacar que el coeficiente de variación es adimensional, es decir, no depende de las unidades consideradas.

La desigualdad de Tchebyshev

Hemos visto que dado un conjunto de datos: $x_1, x_2, ...x_N$, a partir de los mismos podemos calcular la media $\mu = \frac{1}{N}\sum_1^N x_i$ y la desviación estándar $\sigma = \sqrt{\frac{1}{N}\sum_1^N (x_i - \mu)^2}$.
Estos dos valores resumen la información, pero a partir de los mismos no es posible reconstruir el conjunto de datos. Sin embargo estos valores, μ y σ contienen suficiente información para acotar el porcentaje de los datos que se encuentran en los intervalos de la forma $(\mu - k.\sigma, \ \mu + k.\sigma)$ con $k > 1$.
Este resultado se debe al matemático ruso Chebyshev quien probó que para cualquier conjunto de datos por lo menos el $100[1 - (\frac{1}{k})^2]$ % de los mismos se encuentran en el intervalo $(\mu - k.\sigma, \ \mu + k.\sigma)$.

De este modo, para k=2 se tiene que por lo menos el 75% de los datos se encuentran en el intervalo ($\mu - 2.\sigma$, $\mu + 2.\sigma$), y para k=3, por lo menos el 88% de los datos se encuentran en el intervalo ($\mu - 3.\sigma$, $\mu + 3.\sigma$).

En el año 1996 la revista Clarín Fútbol 96 publica las siguientes edades del plantel profesional de Rosario Central :

23	20	26	18	21	21	28	23	20	21	18	20	21
24	20	21	22	21	21	20	37	23	21	26	19	20
19	27	19	21	24								

Le proponemos calcular la edad media y la desviación estándar de las edades y responder a las siguientes preguntas.
¿Qué porcentaje de las edades se encuentran en :
ι) ($\mu - 2.\sigma$, $\mu + 2.\sigma$)
ii) ($\mu - 3.\sigma$, $\mu + 3.\sigma$) ?

Como observará el valor 37 no queda comprendido en el intervalo ($\mu - 3.\sigma$, $\mu + 3.\sigma$).
Ese valor corresponde a la edad de Omar Palma.

El recorrido o rango

Otro valor característico de la dispersión es el recorrido o rango, que se define como la diferencia entre el valor máximo y el valor mínimo de los datos.
Si notamos con x_m el valor mínimo y con x_M el valor máximo entonces $R = x_M - x_m$.

En relación a los datos correspondientes a la antigüedad de los docentes:

10 9 9 4 9 4 15 11 19,

$x_m = 4$ y $x_M = 19$, por lo tanto, R = 15

El recorrido o rango tiene la ventaja de la facilidad de su cálculo y la desventaja de que en su determinación sólo se consideran dos valores del conjunto de datos. A igual que la media aritmética es sensible a valores extremos.

¿ Cómo se modifica el recorrido de los datos anteriores si se agrega el valor 35?

El recorrido o rango se utiliza en las aplicaciones estadísticas al control de calidad cuando el número de datos no supera los diez.

Otros valores característicos

Ya hemos visto que la mediana divide los datos ordenados en dos partes iguales. Cuando se divide un conjunto ordenado en cuatro partes iguales, los puntos de división se conocen como cuartiles. De este modo el primer cuartil, Q_1, es el valor que tiene (aproximadamente)

el 25% de las observaciones menores que él. El segundo cuartil, Q_2 , coincide con la mediana y el tercer cuartil,Q_3 , tiene (aproximadamente el 75% de las observaciones menores que él.
Para calcular los cuartiles utilizaremos un procedimiento similar al empleado para determinar la mediana.

Consideremos las distancias medidas en cuadras, entre la facultad y las viviendas de 10 de sus alumnos elegidos al azar:

3 8 10 14 16 | 20 25 30 35 40

Los datos se presentan ordenados en forma creciente, de modo que la mediana o segundo cuartil es $Q_2 = \frac{16+20}{2} = 18$ cuadras (promedio de los dos valores centrales).
El primer cuartil es la mediana de las primeras 5 observaciones. En consecuencia Q_1=10 cuadras.
El tercer cuartil es la mediana de las segundas 5 observaciones, de modo que Q_3=30 cuadras.

Tal vez las definiciones dadas no le resultan suficientemente precisas. Inclusive algunos programas estadísticos utilizan una regla diferente para calcular los cuartiles, pero las diferencias serán pequeñas para considerarlas importantes.

De forma análoga se definen los percentiles o deciles que dividen al conjunto de datos ordenados en 100 o 10 partes iguales respectivamente.

Cuando se dice que la inteligencia de un alumno está en el percentil 90 significa que su inteligencia es superior al 90% de la población e inferior al 10% restante.

El recorrido intercuartílico

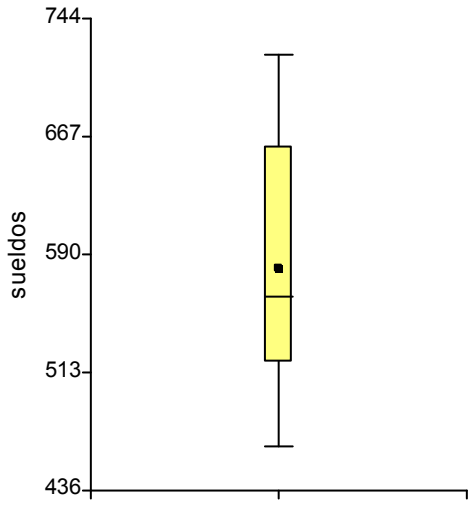
La diferencia RI = $Q_3 - Q_1$ se denomina recorrido intercuartílico y suele emplearse como medida de variabilidad. Un valor pequeño para RI significa que en un intervalo de amplitud reducida se encuentra el 50% de los datos (aproximadamente).
El RI es menos sensible a valores extremos que el rango o recorrido. ¿Por qué?

Diagrama de caja

Otro diagrama desarrollado por Tukey desde el enfoque del análisis exploratorio de datos es el diagrama de caja.
Este diagrama describe al mismo tiempo varias características importantes de un conjunto de datos tales como la tendencia central, la dispersión , la desviación de la simetría y la identificación de observaciones que se alejan de manera poco usual del resto de los datos (valores atípicos)-

Los siguientes datos corresponden a los sueldos de 10 operarios de una sección, de una fábrica:
450, 520, 730, 480, 575, 660, 520, 610, 710, 550.

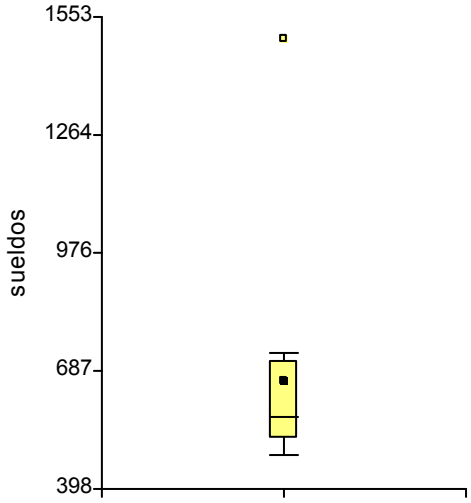
La figura muestra el diagra



El lado inferior y superior de la caja se corresponden con el primer y tercer cuartil respectivamente. El segmento interior de la caja indica la mediana. Cuando los datos tienden a distribuirse simétricamente, el primer y tercer cuartil están aproximadamente a la misma distancia de la mediana (Q_3-Q_2 ≈ Q_2-Q_1). En el ejemplo Q_3-Q_2 > Q_2-Q_1 lo que implica que los datos tienden a distribuirse con asimetría hacia la derecha. El punto interior a la caja indica el valor de la media aritmética.
Fuera de la caja aparecen dos líneas (bigotes) que se extienden hasta un máximo de 1.5 veces el recorrido intercuartílico si no se alcanza antes el los valores mínimos y máximos.
El bigote inferior comienza en: máximo { x_m , Q_1 - 1.5 (Q_3-Q_1)}.
El bigote superior termina en: mínimo { x_M , Q_3 + 1.5 (Q_3-Q_1)}, donde x_m y x_M simbolizan el valor mínimo y máximo de los datos.
Cuando aparecen valores más allá de los bigotes se consideran atípicos y se marcan con cuadraditos.

Si a los datos se incorpora el salario del jefe de la sección, que es de $1.500, el diagrama se visualiza de la siguiente manera:

.



El valor $1500 aparece como un valor atípico.

La media geométrica

Al comenzar el año el precio de un artículo era de $ 100. Al cabo del primer, segundo, tercer y cuarto trimestre los precios eran de $110, $132, $165, y $188,1 respectivamente. En consecuencia el aumento al cabo del primer trimestre fue del 10% y al cabo de todo el del 88.1%.
Los cocientes:
$\frac{110}{100} = 1.10$, $\frac{132}{110} = 1.2$, $\frac{165}{132} = 1.25$, $\frac{188,1}{165} = 1.14$ nos indican que el aumento porcentual de cada trimestre con respecto al trimestre anterior fueron del 10%, 20%, 25% y 14% respectivamente.

Si solo dispusiéramos de los aumentos porcentuales por trimestre, podríamos determinar el aumento porcentual durante todo el año realizando el producto 1.10 x 1.20 x 1.25 x 1.14 cuyo resultado 1.881 nos indica que el aumento durante el año fue del 88.1%.

Observe que:

- $188.1 = 100 \times 1.10 \times 1.20 \times 1.25 \times 1.14 = 100 \times 1.881 = 100 \left(1 + \frac{88.1}{100} \right) = 100 + \frac{88.1}{100} \times 100$
- el aumento porcentual durante el año no se obtiene sumando los aumentos porcentuales de cada trimestre.
- Los aumentos porcentuales por trimestre fueron variando.

Si quisiéramos determinar un aumento porcentual constante por trimestre, que produjera al cabo del año un aumento del 88.1% deberíamos realizar la operación:

$\sqrt[4]{(1.10)(1.20)(1.25)(1.14)}$ cuyo resultado aproximado, 1.1711, nos indica que aumentando cada trimestre el precio un 17.11% sobre el trimestre anterior, al cabo del año el aumento porcentual es del 88.1%.

El número $\sqrt[4]{(1.10)(1.20)(1.25)(1.14)}$ se denomina la media geométrica de los números 1.10, 1.20, 1.25 y 1.14.

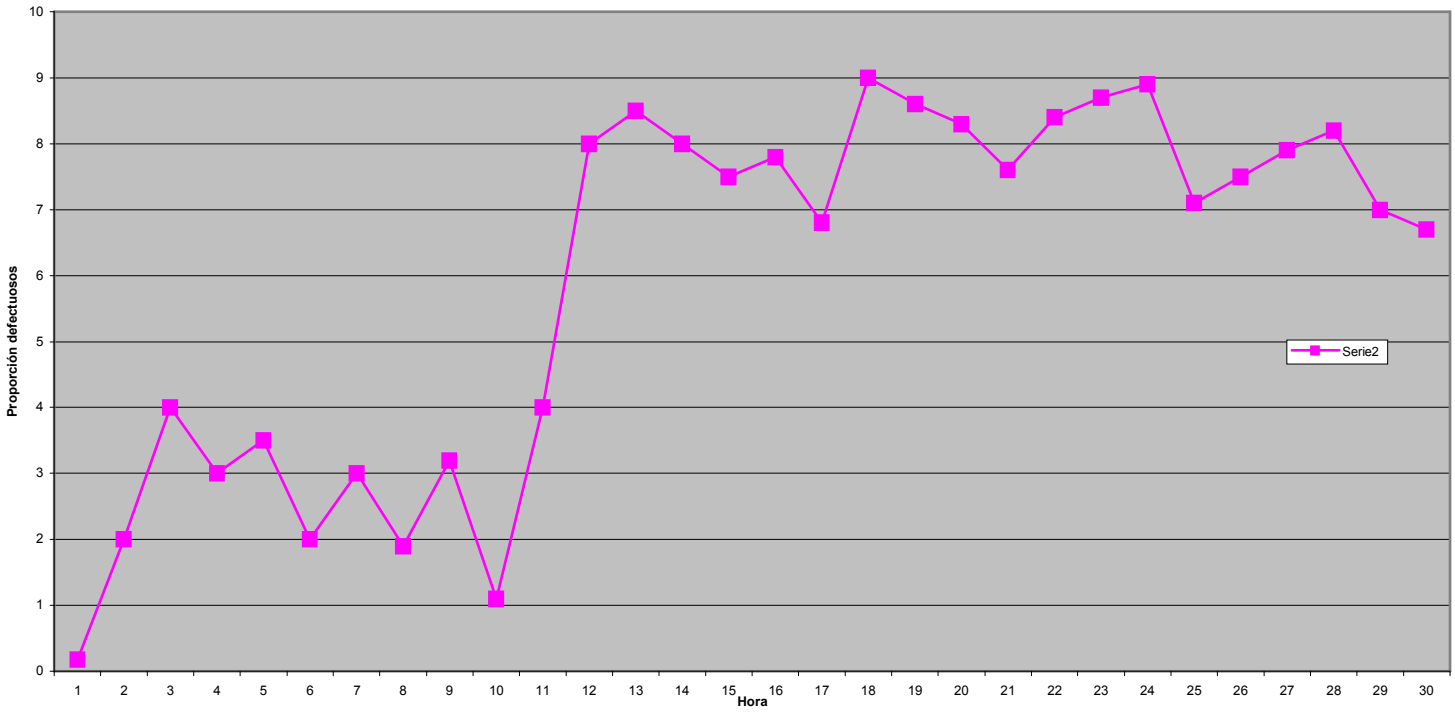
En general, la media geométrica , m_g , de las n observaciones $x_1, x_2, ...x_n$ es:

$m_g = \sqrt{x_1 x_2 .....xn}$

Le proponemos comparar la media aritmética con la media geométrica en diferentes conjuntos de datos. ¿Cuáles son sus observaciones?

Gráficas de series de tiempo

Los histogramas, los diagramas de tallo y hoja, y diagramas de caja son representaciones visuales muy útiles para mostrar la variabilidad presente en un conjunto de datos pero que no toman en cuenta los cambios en el tiempo.
Al registrar las observaciones de una variable en función del tiempo se obtiene un conjunto de números que se denomina una serie de tiempo o serie cronológica.
Para graficar una serie cronológica , sobre el eje horizontal se representa la variable tiempo (en minutos, horas, días, años, etc.), mientras que en el eje vertical se representan los correspondientes valores observados.



La gráfica nos muestra que después de la décima hora hubo un sostenido aumento de la proporción de unidades defectuosas. En casos como éste, la causa se encuentra mirando

atentamente si tuvo lugar algún cambio alrededor del momento del aumento. Estos cambios pueden estar relacionados con la materia prima, maquinaria, operario, etc.

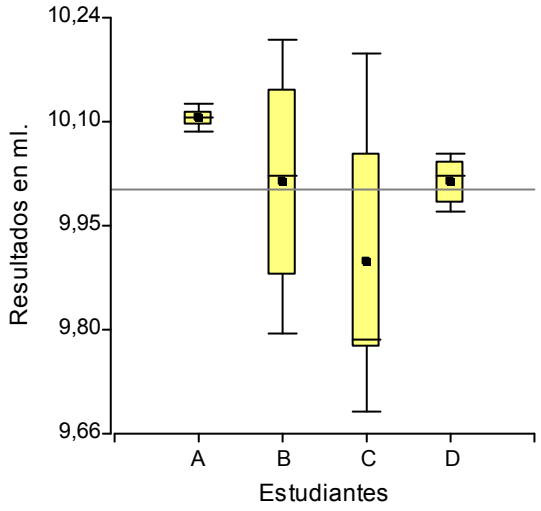
Exactitud y precisión de las mediciones

Cuatro estudiantes que denotamos con A, B, C y D realizan cinco mediciones de una magnitud cuyo verdadero valor se conoce y es igual a10 ml.

Lo resultados que obtienen son:

A:	10,08	10,11	10,09	10,10	10,12
B:	9,88	10,14	10,02	9,80	10,21
C:	10,19	9,79	9,69	10,05	9,78
D:	10,04	9,98	10,02	9,97	10,04

El siguiente diagrama de cajas representa las mediciones realizadas por cada estudiante.
La línea marcada corresponde al valor y = 10.



Los resultados obtenidos por A tienen dos particularidades. Todos los valores están muy próximos, se encuentran comprendidos entre 10,08 y 10,12. Por otra parte todos los resultados son superiores al verdadero valor, que es de 10 ml.

Resulta evidente que han surgido dos tipos de errores en las mediciones de este estudiante. En primer lugar existen **errores aleatorios**, los cuales provocan que haya resultados menores y mayores al valor medio, que es de 10,10 ml. Los errores aleatorios son pequeños, ya que la desviación estándar de las mediciones es pequeña. En este caso decimos que las mediciones son **precisas**.

Por otra parte las mediciones de A tienen **error sistemático**, que se refleja a través de resultados, que en su totalidad superan al verdadero valor. En este caso decimos que las mediciones no son exactas, ya que el promedio de las mismas se encuentra demasiado alejado del valor verdadero. Estos errores suelen reducirse a través de la calibración.

Un análisis para las mediciones de los restantes estudiantes permite concluir que: las mediciones de B son exactas pero no precisas, las mediciones de C no son exactas ni precisas y las de D son exactas y precisas.

Observe que la noción de precisión está asociado a la de desviación estándar de las mediciones y la noción de exactitud con el valor medio de las mismas.

Si las mediciones se repiten en una sucesión rápida (no se tarda más de "una hora" en realizarlas), lo que daría lugar a que permanezcan invariables las condiciones de temperatura, presión y otras condiciones del laboratorio, la precisión medida a través de la desviación estándar es **la precisión dentro de rachas** y se denomina **repetitividad**. En cambio cuando las mediciones se realizan en circunstancias diferentes, podrían cambiar las condiciones del laboratorio. En este caso no sería sorprendente encontrar una mayor desviación estándar, ocasionada por esas condiciones diferentes. Esa desviación estándar calculada en tales condiciones refleja **la precisión entre rachas** y se denomina **reproducibilidad.**

Propuesta

Calcule la media aritmética, la desviación estándar y el coeficiente de variación correspondiente a cada grupo de cinco mediciones.

Propuestas para la revisión

1) El siguiente diagrama de tallo y hojas, "espalda con espalda", corresponde a las calificaciones que obtuvieron en un examen de Probabilidad y Estadística los alumnos de las divisiones A y B.

8 5	1	
8 5 5	2	
6 5 5 2	3	0 5
9 8 5 5	4	2 2
8 5 4 0 0	5	0 5 5
9 8 8 5 2	6	0 2 3 8
5 5 2 0	7	2 5 5 8 8
5	8	0 0 1 5 8
	9	3 3 5 8
	10	0

a) A partir del diagrama describa algunas características de las distribuciones de las calificaciones de ambas divisiones.
b) Calcula la media aritmética, la mediana, la desviación estándar y el coeficiente de variación de las calificaciones para:
   i) los alumnos de la división A.
   ii) los alumnos de la división B.
   iii) para los alumnos de ambas divisiones consideradas conjuntamente.

c) ¿Puede a partir de las medias aritméticas de ambas divisiones obtener la media aritmética conjunta? En caso afirmativo explique cómo.
d) Idem c), pero para la mediana.
e) ¿Qué otro recurso gráfico conoce para comparar el rendimiento de ambas divisiones?

f) ¿Considera que existen diferencias en el rendimiento de ambas divisiones? En caso afirmativo enuncie posibles causas que expliquen esa diferencia y cómo procedería para indagar acerca de esas posibles causas.

2) El tiempo promedio y la desviación estándar para la limpieza de un equipo es de 50 horas y 4 horas respectivamente.
Analiza cuáles de las siguientes afirmaciones son verdaderas y cuáles son falsas. Fundamente.

a) Más del 50% de los equipos requieren un tiempo de limpieza superior a 62 horas.
b) Menos del 15% de los equipos que requieren a lo sumo 38 horas para su limpieza.
c) Por lo menos el 75% de los equipos requieren más de 42 y menos de 58 horas para su limpieza.

3) La edad (en años) de los operarios de una sección es: 23, 21, 22, 26.

a) Calcule la edad media y la desviación estándar de las edades.
b) Si se mantienen los operarios durante los próximos dos años ¿cuál es la edad media y la desviación estándar dentro de 2 años?
c) ¿Cuáles son sus observaciones en relación a los valores calculados en a) y en b)?

4) Los siguientes datos corresponden a los sueldos (en pesos) de los cinco empleados de una sección: 380 420 400 450 410

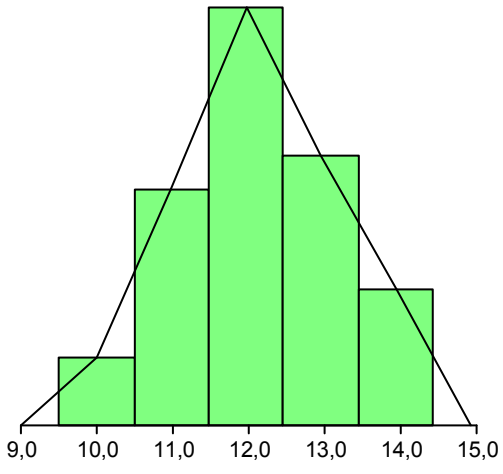
a) Calcule la media y desviación estándar de los sueldos.
b) Si se aumentan los sueldos en un 10%, ¿cuál es la nueva media y desviación estándar ?
c) ¿Cuáles son sus ob servaciones?

5) Sea $y_i = a + b\, x_i \quad i = 1,2, ...N$, donde a y b son constantes. Encuentra la relación entre la media y la desviación estándar de los valores x_i con la media y desviación estándar de los y_i .

6) Sean μ y σ la media y desviación estándar de $x_1, x_2, ...x_N$ y sea $z_i = \frac{x_i - \mu}{\sigma}$.

¿ Cuáles son los valores de la media y la desviación estándar de los z_i ?

7) El siguiente histograma y polígono de frecuencias relativas corresponden a 100 observaciones de los tiempos, en minutos, que una persona tarda para viajar con su auto desde su casa al trabajo.

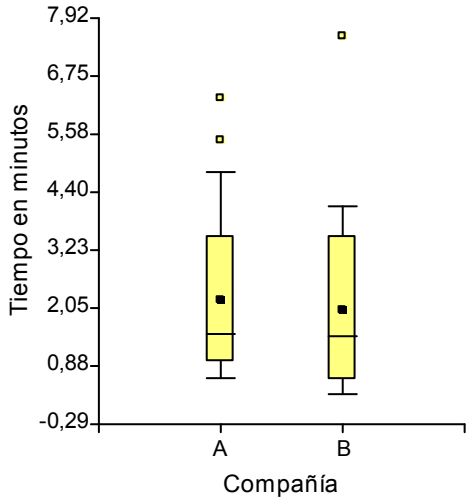


Analice cuáles de las siguientes afirmaciones son verdaderas y cuáles son falsas. En cada caso justifique.

a) El tiempo medio supera los 12 minutos.
b) El valor de la mediana es superior a12.
c) El valor del primer cuartil se encuentra en el primer intervalo de clase.
d) El valor del tercer cuartil se encuentra en el cuarto intervalo de clase.
e) La proporción de tiempos observados entre 11.5 y 12.5 es menor que la proporción de tiempos inferiores a 11.5.

.

8) a) El siguiente diagrama de cajas corresponde a dos muestras, cada una de tamaño 20, del tiempo en minutos en que dos compañías A y B demoraron en resolver problemas en la línea, que impide a un cliente recibir o hacer llamadas.



Ante la evidencia gráfica redacte un informe comparando las dos compañías. ¿Considera que existen diferencias en cuanto al tiempo en que las dos compañías resuelven los problemas a sus clientes?

9) Una característica de calidad que interesa en el proceso de llenado de bolsas de té es el peso de las mismas. Llenar las bolsas con la cantidad exacta es difícil debido a las variaciones de temperatura y humedad en el interior de la fábrica, las diferencias en la densidad del té y la alta velocidad con que trabaja la máquina de llenado (cerca de 170 bolsas por minuto).

Si la cantidad promedio de té en la bolsa excede el valor que figura en la etiqueta, la compañía tendría pérdidas porque estaría regalando parte del producto y si la cantidad promedio es inferior al valor declarado, la compañía estaría violando las leyes de veracidad de las etiquetas y podría generar desconfianza entre sus clientes.

La siguiente tabla proporciona el peso en gramos de una muestra de 50 bolsas de té producidas en una hora por una misma máquina.

5.65	5.44	5.42	5.40	5.53	5.34	5.54	5.45	5.52	5.41
5.57	5.40	5.53	5.54	5.55	5.62	5.56	5.46	5.44	5.51
5.47	5.40	5.47	5.61	5.53	5.32	5.67	5.29	5.49	5.55

5.77	5.57	5.42	5.58	5.58	5.50	5.32	5.50	5.53	5.58
5.61	5.45	5.44	5.25	5.56	5.63	5.50	5.57	5.67	5.36

Represente gráficamente la información y calcule valores que resuman la información de modo que le faciliten el análisis de decisión respecto a la característica de calidad del envasado, cuando las bolsas indican en su etiqueta un peso promedio de 5.50 gramos.

10) Los valores del contenido de humedad, en porcentajes, de 42 muestras de arcilla resultaron:

10.3	10.7	9.2	11.4	11.5	8.4	10.3	10.2	9.8	10.2	11.4	10.4
11.6	10.3	11.4	10.3	7.8	8.7	11.7	9.3	9.8	11.4	10.4	11.2
10.6	9.8	10.1	13.1	10.5	12.2	9.6	7.0	12.3	12.3	12.3	9.7
10.7	11.5	10.6	13.0	13.0	9.0						

a) Construya un diagrama de tallo y hoja.
b) Represente gráficamente la información a través de un histograma y de un polígono de frecuencias relativas.
c) Calcule los valores de la media aritmética, mediana desviación estándar y variancia a partir de los datos dados. Compare luego con los valores que obtiene cuando los calcula a partir de los datos agrupados en intervalos de clase.

11) Un aspecto clave de la calidad de un producto es su peso. La norma establece que su peso mínimo sea de 2 kg. Un ingeniero de producción informa que se está cumpliendo con tal norma ya que el peso promedio del producto es de 2.5 kg. ¿Está usted de acuerdo con el ingeniero?

12) En un barrio residencial existen 50 viviendas construidas sobre terrenos cuyas superficies, en m^2, varían. También varían los m^2 cubiertos de cada terreno. Sean: (x_i , y_i) los m^2 de los terrenos y superficies cubiertas respectivamente.
No se conocen los valores de x_i ni de y_i , pero sí sus respectivas medias aritméticas y medianas.

a) ¿Permite dicha información conocer la media aritmética y la mediana de las superficies no cubiertas? De ser posible explique cómo procedería para hallarlos, de lo contrario aclare porqué no es posible hallarlos.
b) Siendo $x_i, > y_i$, ¿puede afirmarse que el desvío estándar de los valores x_i es mayor que el desvío estándar de los valores y_i ? Explique

13) En una empresa se llevan los registros del número de fallas de equipos por mes. Si la media es igual 10 y la mediana es igual a 5, ¿qué valor adoptaría para caracterizar la tendencia central? Justifique.

14) Se observan 100 cajas de 200 azulejos cada una y se registra la cantidad de azulejos fallados por caja.

La siguiente tabla resume la información obtenida.

Cantidad de azulejos fallados:	0	1	2	3	4
Cantidad de cajas:	40	30	x	10	5

a) ¿Cuánto vale x ? (No existen cajas que contienen más de 4 azulejos fallados)
b) ¿ Cuál es la media aritmética y la desviación estándar de azulejos fallados por caja?. ¿ Qué mide la desviación estándar?
c) Determine el valor de la moda y de la mediana. Explique qué representan esos valores.
d) Analice si las siguientes afirmaciones son falsas o verdaderas. Justifique.
   d-1 El 60% de las cajas contienen azulejos fallados.
   d-2 El 15 % de las cajas contienen más de dos azulejos fallados.
   d-3 El total de azulejos fallados en las 100 cajas es igual a 10.

Le sugerimos recurrir a la siguiente bibliografía para ampliar la lectura y ejercitación propuesta:

- Berenson, Mark y otros. "Estadística para Administración". México. Pearson Educación. 2001.
- Canavos, George. "Probabilidad y Estadística. Aplicaciones y Métodos". México. McGraw Hill 1988.
- Devore Jay L. "Probabilidad y Estadística para Ingeniería y Ciencias". México. Thomson Editores 1998.
- Miller Irwin y Freund J. "Probabilidad y Estadística para Ingenieros". Prentice Hall 1993.
- Montgomery Douglas, Runger George. "Probabilidad y Estadística Aplicadas a la Ingeniería". México. McGraw Hill 1996.
- Scheaffer R., MaClave James. "Probabilidad y Estadística para Ingeniería". México. Grupo Editorial Iberoamericana1993.
- Walpole Ronald, Myers Raymond. "Probabilidad y Estadística". México. Pearson Educación 1999.