ESTADÍSTICA DESCRIPTIVA

Tabla de contenidos

Tema	Página
1. Estadística. Definición	2
2. Cónceptos básicos	2
3. Ramas de la Estadística	3
Estadística. Diagrama conceptual	4
Estadística Descriptiva	5
1. Relevamiento de datos	5
2. Un concepto importante. Variable	8
2.1. Concepto	8
2.2. Clasificación de las variables	8
2.3. Escalas de medición de las variables	9
2.4. Simbología de las variables	10
3. Tabulación de datos	10
3.1. Serie simple y distribución de frecuencias	10
3.2. Representación gráfica	15
4. Otras distribuciones de frecuencias	19
4.1. Distribución de frecuencias relativas	19
4.2. Distribución de frecuencias acumuladas	19
4.3. Distribución de frecuencias relativas acumuladas	20
4.4. Representación gráfica	20
5. Medidas descriptivas	24
5.1. Medidas de tendencia central	24
5.2. Medidas de dispersión o variabilidad	40
5.3. Medidas de distribución, Asimetría y Kurtosis	44
5.4. Media geométrica	45
5.5. Media armónica	46
6. Un gráfico muy descriptivo. Diagrama de caja (boxplot)	47
Bibliografía	49

Autora: María Elena Marcoleri

2010

Estadística

1. DEFINICIÓN

La **Estadística** es una ciencia que estudia la aplicación del método científico en el análisis de datos, numéricos o no, con el fin de contribuir a tomar decisiones racionales.

La **Estadística** es una ciencia con base matemática referente a la recolección, análisis e interpretación de datos, que busca explicar condiciones regulares en fenómenos de tipo aleatorio. Es aplicable en una amplia variedad de disciplinas, desde la física hasta las ciencias sociales, desde las ciencias de la salud hasta el control de calidad, y es usada para la toma de decisiones en áreas de negocios e instituciones gubernamentales.

La palabra "estadística" procede del latín *statisticum collegium* ("consejo de Estado") y de su derivado italiano *statista* ("hombre de Estado" o "político"). El término alemán *Statistik*, que fue primeramente introducido por Gottfried Achenwall (1749), designaba originalmente el análisis de datos del Estado, es decir, "la ciencia del Estado" (también llamada "aritmética política" de su traducción directa del inglés). No fue hasta el siglo XIX cuando el término *estadística* adquirió el significado de recolectar y clasificar datos. Este concepto fue introducido por el inglés John Sinclair.

En su origen, la estadística estuvo asociada a datos, a ser utilizados por el gobierno y cuerpos administrativos (a menudo centralizados). La colección de datos acerca de estados y localidades continúa ampliamente a través de los servicios de estadística, nacionales e internacionales. En particular, los censos suministran información regular acerca de la población.

Desde los comienzos de la civilización han existido formas sencillas de estadística, pues ya se utilizaban representaciones gráficas y otros símbolos en pieles, rocas, palos de madera y paredes de cuevas para contar el número de personas, animales o ciertas cosas. Hacia el año 3000 a. C. los babilónicos usaban ya pequeñas tablillas de arcilla para recopilar datos en tablas sobre la producción agrícola y de los géneros vendidos o cambiados mediante trueque. Los egipcios analizaban los datos de la población y la renta del país mucho antes de construir las pirámides en el siglo XI a. C. Los libros bíblicos de Números y Crónicas incluyen, en algunas partes, trabajos de estadística. El primero contiene dos censos de la población de Israel y el segundo describe el bienestar material de las diversas tribus judías. En China existían registros numéricos similares con anterioridad al año 2000 a. C. Los griegos clásicos realizaban censos cuya información se utilizaba hacia el 594 a. C. para cobrar impuestos.

2. CONCEPTOS BÁSICOS

- En Estadística la **población**, también llamada **universo** o **colectivo** es el conjunto de elementos de referencia sobre el que se realizan las observaciones. Puede estar constituida por personas, animales, plantas, artículos o cosas. Es un conjunto generalmente inaccesible, que reúne unas características determinadas. Por ejemplo, la población de habitantes de San Salvador de Jujuy hoy, los estudiantes de la Facultad de Ciencias Económicas de la UNJu en el corriente año, los libros de la biblioteca de la Facultad cuando empiezan las clases de este año. Y así muchos ejemplos más.
- **Muestra estudiada:** es el grupo de elementos en el que se recogen los datos y se realizan las observaciones, siendo realmente un **subconjunto representativo de la población** y es accesible y limitado. El número de muestras que se puede obtener de una población es una o más. Por ejemplo, un conjunto de 100 estudiantes de la Facultad de Ciencias Económicas, en el cual estén representados todos los cursos.

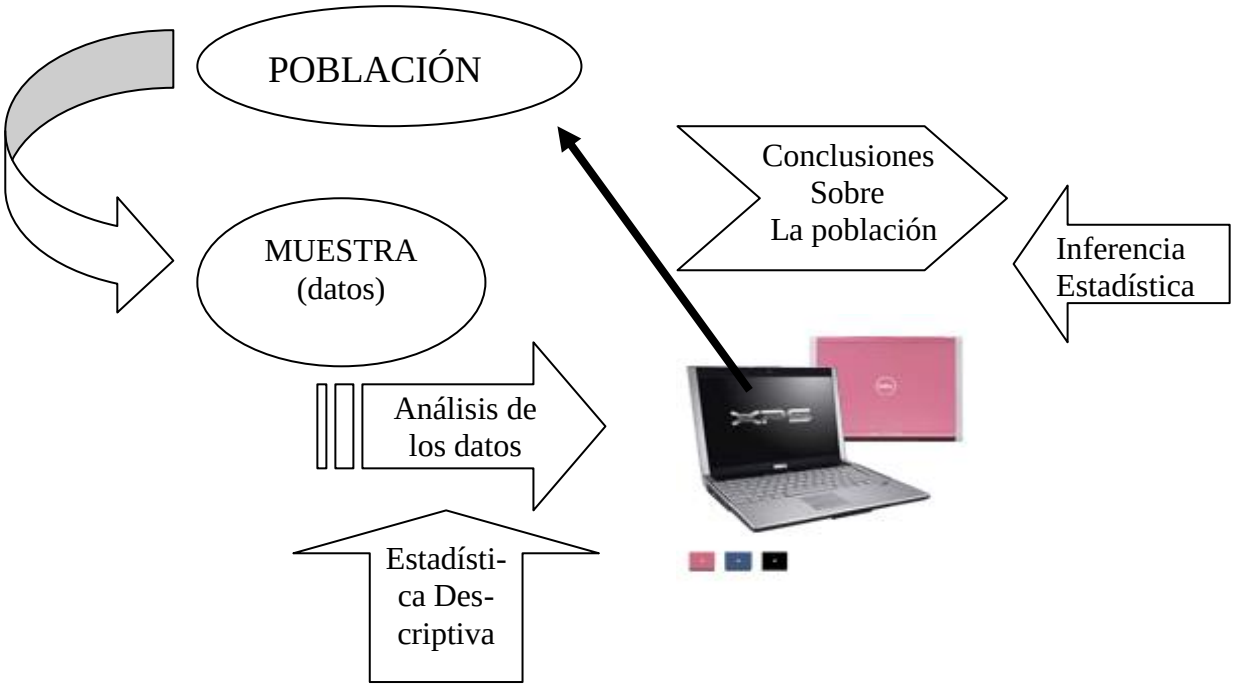
- En Estadística se llama **parámetro** a un valor representativo de una población. El parámetro es el cálculo de valores en la población. Es una medida descriptiva de alguna característica de una población. También se puede decir que es el resultado que generaliza las características de la población; se puede dar en porcentaje o en promedio. Por ejemplo, el ingreso familiar mensual promedio de los hogares de San Salvador de Jujuy en un momento determinado, la proporción de estudiantes de la Facultad que tienen quince o más materias aprobadas, la proporción de libros de la biblioteca de la Facultad que fueron adquiridos en los últimos cinco años. Generalmente se simbolizan con letras griegas: μ, σ, π, α, β, etc.
- En cambio, un **estadístico** o una **estadística,** es una medida descriptiva que resume una característica de una muestra extraída de la población. Por ejemplo, el ingreso familiar mensual promedio de 500 hogares de San Salvador de Jujuy (representativos de todos los hogares de la ciudad) en un momento determinado, la proporción de una muestra de 100 estudiantes de la Facultad que tienen quince o más materias aprobadas, en una muestra de 55 libros de la biblioteca de la Facultad que fueron adquiridos en los últimos cinco años, la proporción de libros que corresponde al Área Contable. La palabra ***estadísticas*** también se refiere al resultado de aplicar un algoritmo estadístico a un conjunto de datos, como en estadísticas económicas, estadísticas criminales, estadísticas demográficas, etc.

3. RAMAS DE LA ESTADÍSTICA

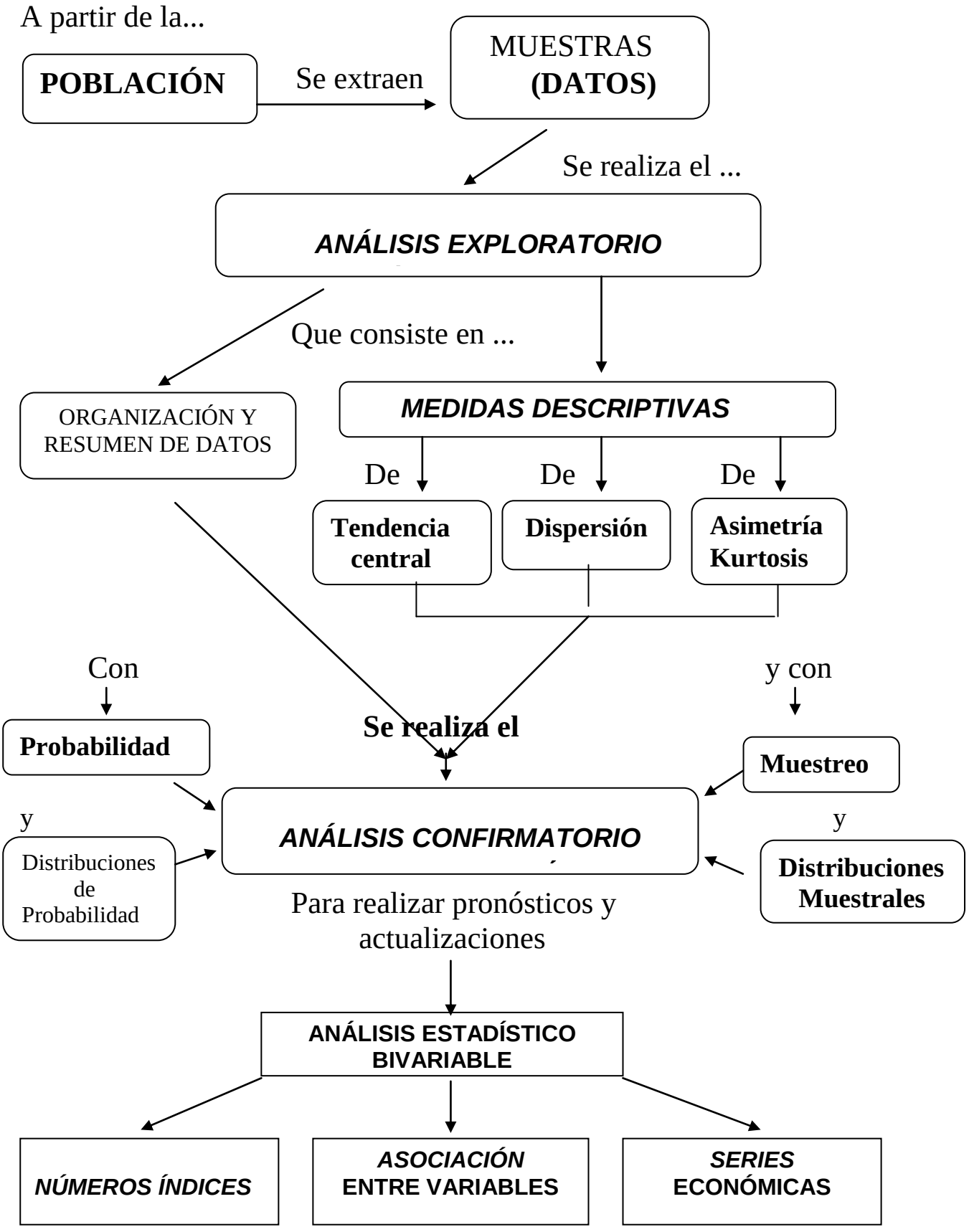
La Estadística se divide en dos ramas:

- La **Estadística Descriptiva**, que se dedica a los métodos de recolección, tabulación, análisis, presentación e interpretación de datos originados a partir de los fenómenos en estudio, a fin de describir en forma apropiada sus principales y diversas características. Los datos pueden ser resumidos numérica o gráficamente.
- La **Inferencia Estadística**, que se dedica a la generación de los modelos y predicciones asociadas a los fenómenos en cuestión teniendo en cuenta la aleatoriedad de las observaciones muestrales. Se usa para modelar patrones en los datos y extraer conclusiones acerca de la población bajo estudio, analizando sólo una muestra de esa población.

Ambas ramas (descriptiva e inferencial) comprenden la **estadística aplicada**. Hay también una disciplina llamada **estadística matemática**, la cual se refiere a las bases teóricas de la materia.

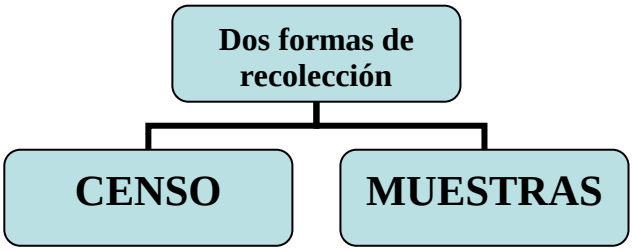


ESTADÍSTICA



ESTADÍSTICA DESCRIPTIVA

1. RELEVAMIENTO DE DATOS



Se denomina **Censo**, en **estadística descriptiva**, al recuento de individuos que conforman una población estadística, definida como un conjunto de elementos de referencia sobre el que se realizan las observaciones. El censo de una población estadística consiste, básicamente, en obtener el número total de individuos mediante las más diversas técnicas de recuento.

El censo es una de las operaciones estadísticas que no trabaja sobre una muestra, sino **sobre la población total.**

Uno de los casos particulares de censo pero, al mismo tiempo, uno de los más comunes, es el denominado **censo de población**, en el cual el objetivo es determinar el número de personas humanas que componen un grupo, normalmente un país o una nación. En este caso, la población estadística comprendería a los componentes o habitantes del grupo, país o nación.

En general, en un censo de población se pueden realizar algunas actividades extras que no se corresponden específicamente con la operación censal estadística. Se busca calcular el número de habitantes de un país de territorio delimitado, correspondiente a un momento o período dado, pero se aprovecha igualmente para obtener una serie de datos demográficos, económicos y sociales relativos a esos habitantes.

La **muestra es e**l grupo de sujetos (personas, animales, seres microscópicos u objetos inanimados) que se utilizarán como objeto de estudio en una investigación. Será a ellos a quienes se les aplique el procedimiento experimental (las pruebas, mediciones, entrevistas, encuestas, tratamientos médicos farmacológicos o no farmacológicos) y serán ellos los que, distribuidos o no en dos o más grupos, cada uno de éstos con una condición experimental específica, nos darán, después del análisis de los resultados, la respuesta positiva o negativa a la pregunta que generó el desarrollo de la investigación, respuesta que se expresará, por medio de una publicación científica, a través de una serie de conclusiones.

Existen varios tipos de muestras, de los cuales en el cuadro siguiente se mencionan los más comúnmente utilizados:

Muestras probabilísticas	Muestra aleatoria simple
	Muestra sistemática
	Muestra estratificada
	Muestra de conglomerados
Muestras no probabilísticas	Muestra de juicio
	Muestra de cuota
	Muestra bola de nieve

Un muestreo es probabilístico cuando se puede determinar de antemano la probabilidad de selección de cada uno de los elementos de la población, es decir que la selección de cada elemento debe ser realizada al azar con una probabilidad conocida a priori.

Las razones para estudiar muestras en lugar de poblaciones son diversas y entre ellas se pueden señalar:

a. Ahorrar tiempo. Estudiar a menos individuos es evidente que lleva menos tiempo.
b. Como consecuencia del punto anterior se ahorran costos.
c. Estudiar la totalidad de los pacientes o personas con una característica determinada en muchas ocasiones puede ser una tarea inaccesible o imposible de realizar.
d. Aumentar la calidad del estudio. Al disponer de más tiempo y recursos, las observaciones y mediciones realizadas a un reducido número de individuos pueden ser más exactas y plurales que si las tuviésemos que realizar a una población.
e. La selección de muestras específicas permitirá reducir la heterogeneidad de una población al indicar los criterios de inclusión y/o exclusión.

Muestreo aleatorio simple

La forma más común de obtener una muestra es la selección al azar. Es decir, cada uno de los elementos de una población tiene la misma posibilidad de ser elegido. Si no se cumple este requisito, se dice que la muestra es viciada. Para tener la seguridad de que la muestra aleatoria no es viciada, debe emplearse para su constitución algún método aleatorio (al azar). El procedimiento empleado es el siguiente: 1) se asigna un número a cada individuo elemento de la población y 2) a través de algún medio mecánico (bolillas dentro de una bolsa, tablas de números aleatorios, números aleatorios generados con una calculadora o computadora, etc.) se eligen tantos elementos como sea necesario para completar el tamaño de muestra requerido. Este procedimiento, atractivo por su simpleza, tiene poca o nula utilidad práctica cuando la población objetivo es muy grande y heterogénea.

Muestreo aleatorio sistemático

Una muestra sistemática es obtenida cuando los elementos son seleccionados en una manera ordenada. La manera de la selección depende del número de elementos incluidos en la población y el tamaño de la muestra. El número de elementos en la población es, primero, dividido por el número deseado en la muestra. El cociente indicará si cada décimo, cada onceavo, o cada centésimo elemento en la población tendrá que ser seleccionado. El primer elemento de la muestra se selecciona al azar. Por lo tanto, una muestra sistemática puede dar la misma precisión de estimación acerca de la población, que una muestra aleatoria simple cuando los elementos en la población están ordenados al azar.

Este procedimiento exige, como el anterior, numerar todos los elementos de la población, pero en lugar de extraer n números aleatorios sólo se extrae uno. Se parte de ese número aleatorio i, que es un número elegido al azar, y los elementos que integran la muestra son los que ocupan los lugares i, i+k, i+2k, i+3k,...,i+(n-1)k, es decir se toman los individuos de k en k, siendo k el resultado de dividir el tamaño de la población entre el tamaño de la muestra: $k = N/n$. El número i que empleamos como punto de partida será un número al azar entre 1 y k.

El riesgo se este tipo de muestreo está en los casos en que se dan periodicidades en la población, ya que al elegir a los miembros de la muestra con una periodicidad constante (k) puede ocurrir que se introduzca una homogeneidad que no se da en la población. Por ejemplo, si se debe seleccionar una muestra sobre listas de 10 individuos en los que los 5 primeros son varones y los 5 últimos mujeres, si se emplea un muestreo aleatorio sistemático con $k = 10$ siempre seleccionaríamos o sólo hombres o sólo mujeres, no podría haber una representación de los dos grupos.

Muestreo aleatorio estratificado

Una muestra es estratificada cuando los elementos de la muestra son proporcionales a su presencia en la población. La presencia de un elemento en un estrato excluye su presencia en otro. Para este tipo de muestreo, se divide a la población en varios grupos o estratos (formados por elementos homogéneos entre sí) con el fin de dar representatividad a los distintos factores que integran el universo de estudio. Para la selección de los elementos o unidades representantes, se utiliza el método de muestreo aleatorio. Las estimaciones de la población, basadas en la muestra estratificada, usualmente tienen mayor precisión (o menor error muestral) que si la población entera muestreada mediante muestreo aleatorio simple. Trata de obviar las dificultades que presentan los anteriores ya que simplifican los procesos y suelen reducir el error muestral para un tamaño dado de la muestra. Consiste en considerar categorías típicas diferentes entre sí (estratos) que poseen gran homogeneidad respecto a alguna característica (se puede estratificar, por ejemplo, según la profesión, el municipio de residencia, el sexo, el estado civil, etc). Lo que se pretende con este tipo de muestreo es asegurarse de que todos los estratos de interés estarán representados adecuadamente en la muestra. Cada estrato funciona independientemente, pudiendo aplicarse dentro de ellos el muestreo aleatorio simple o el sistemático para elegir los elementos concretos que formarán parte de la muestra. En ocasiones las dificultades que plantean son demasiado grandes, pues exige un conocimiento detallado de la población. (tamaño, geográfico, sexos, edades,...).

La distribución de la muestra en función de los diferentes estratos se denomina afijación, y puede ser de diferentes tipos:

Afijación Simple: a cada estrato le corresponde igual número de elementos muestrales.

Afijación Proporcional: la distribución se hace de acuerdo con el peso (tamaño) de la población en cada estrato.

Afijación Óptima: se tiene en cuenta la previsible dispersión de los resultados, de modo que se considera la proporción y la desviación típica. Tiene poca aplicación ya que no se suele conocer la desviación.

Muestreo de conglomerados

Para obtener una muestra de conglomerados, primero se divide la población en grupos que son convenientes para el muestreo. En seguida, seleccionar una porción de los grupos al azar o por un método sistemático. Finalmente, tomar todos los elementos o parte de ellos al azar o por un método sistemático de los grupos seleccionados para obtener una muestra. Bajo este método, aunque no todos los grupos son muestreados, cada grupo tiene una igual probabilidad de ser seleccionado. Por lo tanto la muestra es aleatoria.

Los métodos anteriores están estructurados para seleccionar directamente los elementos de la población, es decir, que las unidades muestrales son los elementos de la población. En el muestreo por conglomerados la unidad muestral es un grupo de elementos de la población que forman una unidad, a la que se llama conglomerado. Las unidades hospitalarias, los departamentos universitarios, una caja de determinado producto, etc, son conglomerados naturales. En otras ocasiones se pueden utilizar conglomerados no naturales como, por ejemplo, las urnas electorales. Cuando los conglomerados son áreas geográficas suele hablarse de "muestreo por áreas".

El muestreo por conglomerados consiste en seleccionar aleatoriamente un cierto numero de conglomerados (el necesario para alcanzar el tamaño muestral establecido) y en investigar después todos los elementos pertenecientes a los conglomerados elegidos.

Una muestra de conglomerados, usualmente produce un mayor error muestral (por lo tanto, da menor precisión de las estimaciones acerca de la población) que una muestra aleatoria simple del mismo tamaño. Los elementos individuales dentro de cada "conglomerado" tienden usualmente a ser iguales. Por ejemplo la gente rica puede vivir en el mismo barrio, mientras que la gente pobre puede vivir en otra área. No todas las áreas son muestreadas en un muestreo de áreas. La variación entre los elementos obtenidos de las áreas seleccionadas es, por lo tanto, frecuentemente mayor que la obtenida si la

población entera es muestreada mediante muestreo aleatorio simple. Esta debilidad puede ser reducida cuando se incrementa el tamaño de la muestra de área.

El incremento del tamaño de la muestra puede fácilmente nacerse en la muestra de área. Los entrevistadores no tienen que caminar demasiado lejos en una pequeña área para entrevistar más familias. Por lo tanto, una muestra grande de área puede ser obtenida dentro de un corto período de tiempo y a bajo costo. Por otra parte, una muestra de conglomerados puede producir la misma precisión en la estimación que una muestra aleatoria simple, si la variación de los elementos individuales dentro de cada conglomerado es tan grande como la de la población.

Muestreo intencionado o de juicio

También recibe el nombre de sesgado. El investigador selecciona los elementos que a su juicio son representativos, lo que exige un conocimiento previo de la población que se investiga. Es utilizado generalmente en los estudios de casos.

Muestreo por cuotas

También llamado muestreo accidental, se divide a la población en estratos o categorías, y se asigna una cuota para las diferentes categorías y, a juicio del investigador, se selecciona las unidades de muestreo. La muestra debe ser proporcional a la población, y en ella deberán tenerse en cuenta las diferentes categorías. El muestreo por cuotas se presta a distorsiones, al quedar a criterio del investigador la selección de las categorías.

Muestreo bola de nieve

Se localiza a algunos individuos, los cuales conducen a otros, y estos a otros, y así hasta conseguir una muestra suficiente. Este tipo se emplea muy frecuentemente cuando se hacen estudios con poblaciones "marginales", delincuentes, sectas, determinados tipos de enfermos, etc.

Muestreo mixto

Se combinan diversos tipos de muestreo. Por ejemplo: se puede seleccionar las unidades de la muestra en forma aleatoria y después aplicar el muestreo por cuotas.

2. UN CONCEPTO IMPORTANTE: VARIABLE

2.1. Concepto

Una **variable** es una característica que varía de un elemento a otro de la población o de la muestra.

Lo que se estudia en cada individuo o elemento de la muestra son las **variables** (edad, sexo, peso, talla, tensión arterial sistólica, etcétera). Los datos son los valores que toma la variable en cada caso. Se asignan valores a las variables incluidas en el estudio. Se debe además concretar la escala de medida que se aplicará a cada variable.

La naturaleza de las observaciones será de gran importancia a la hora de elegir el método estadístico más apropiado para abordar su análisis.

2.2. Clasificación de las variables según su naturaleza

Según su naturaleza las variables se clasifican en **cualitativas** *y* **cuantitativas**.

Son **variables cualitativas** aquellas que no son susceptibles de medición numérica. Representan cualidades y atributos que se expresan en categorías, por eso, estas variables también se llaman categóricas. Por ejemplo, son variables cualitativas el color de las flores, cuyas categorías pueden ser rojo,

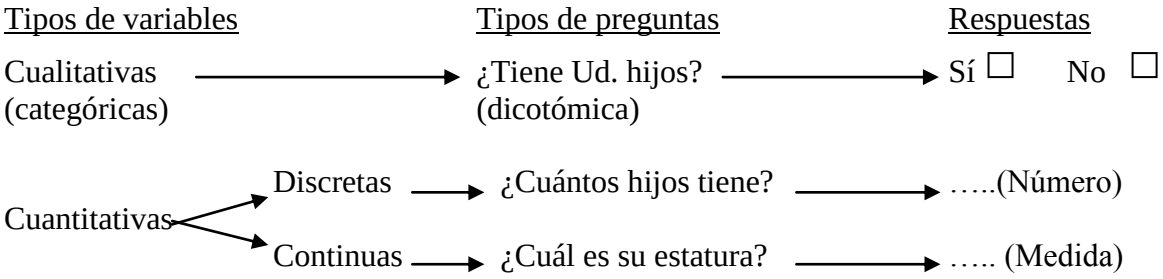
rosado, blanco; el tamaño de las empresas, cuyas categorías pueden ser pequeñas, medianas y grandes; los días de la semana, las estaciones del año, el color del cabello y de los ojos de las personas, etc. En esta clase de variables se encuentran las dicotómicas, que son aquellas variables cualitativas que solo admiten dos categorías, por ejemplo, Sí y No, correcto e incorrecto, frío y calor, femenino y masculino.

Son **variables cuantitativas** aquellas susceptibles de medición numérica. Sus valores provienen de medir o de contar los elementos de la población o de la muestra. Según que se generen contando o midiendo, estas variables se clasifican en ***discretas*** y ***continuas***.

Son **variables cuantitativas discretas** aquellas cuyos valores provienen de contar, por ejemplo, cantidad de hijos por familia, cantidad de alumnos por aula, número de errores de facturación por mes, número de ausentes por día en una empresa. Sus valores asumen números enteros.

Son **variables cuantitativas continuas** las que provienen de efectuar mediciones. Se caracterizan porque entre dos valores cualesquiera de la variable, existen infinitos otros valores. Por ejemplo, la altura y el peso de las personas, los valores monetarios en cualquier tipo de moneda, la edad de las personas, el tiempo de espera para ser atendidos, los precios de los artículos, y tantos otros ejemplos. Sus valores pueden asumir números con cifras decimales.

A modo de resumen se puede presentar lo siguiente:



2.3. Escalas de medición de las variables

Se entenderá por medición al proceso de asignar el valor a una variable de un elemento en observación. Este proceso utiliza diversas escalas: nominal, ordinal, de intervalo y de razón.

Escalas	Ejemplos
Nominal:	Lugar de nacimiento, temperatura (frío, calor)
Ordinal:	Nivel de instrucción (Primario, Medio, Superior, etc.)
De intervalo:	Temperatura (5°, 10°, etc.). El cero es convencional.
De razón:	Cantidad de dinero por persona, N° de hijos por flia. El cero es natural.

En general:

Variables cualitativas → Escalas nominal y ordinal.
Variables cuantitativas → Escalas de intervalo y de razón.

La **escala nominal** se utiliza cuando las categorías de una variable cualitativa no tienen naturalmente un orden establecido. Los siguientes son ejemplos de variables con este tipo de escala: Nacionalidad, Uso de anteojos, Número de camiseta en un equipo de fútbol, Número de Cédula Nacional de Identidad.

La **escala ordinal**, en cambio, es útil cuando las categorías de una variable cualitativa tienen naturalmente un orden o jerarquía preestablecidos, siendo un ejemplo claro las categorías ocupacionales de las personas: jefe, subjefe, empleado, etc.; categorías de los profesores de la Universidad: Titular, Asociado, Adjunto, y de los Auxiliares de docencia, Jefe de Trabajos Prácticos, Ayudante de Primera y Ayudante de segunda.

La escala ordinal, además de las propiedades de la escala nominal, permite establecer un orden entre los elementos medidos. Otros ejemplos de variables con escala ordinal son: Preferencia a productos de consumo, Etapa de desarrollo de un ser vivo, Clasificación de películas por una comisión especializada, Madurez de una fruta al momento de comprarla.

La **escala de intervalo**, además de todas las propiedades de la escala ordinal, hace que tenga sentido calcular diferencias entre las mediciones.

Los siguientes son ejemplos de variables con esta escala: Temperatura de una persona, Ubicación en una carretera respecto de un punto de referencia (Kilómetro 85 Ruta 5), Sobrepeso respecto de un patrón de comparación, Nivel de aceite en el motor de un automóvil medido con una vara graduada.

Finalmente, la **escala de razón** permite, además de lo de las otras escalas, comparar mediciones mediante un cuociente.

Algunos ejemplos de variables con la escala de razón son los siguientes: Altura de personas, Cantidad de litros de agua consumido por las personas en un día, Velocidad de los autos en la ruta, Número de goles marcados por un jugador de básquetbol en los partidos de un año.

Las escalas de intervalo y de razón se diferencian fundamentalmente por dos razones: 1) por la existencia del cero natural, que significa "ausencia de…" (razón), y el cero convencional que no significa ausencia de … (intervalo); 2) porque la escala de razón permite establecer proporciones entre los valores de las variables, mientras que la escala de intervalo no lo admite.

2.4. Simbología de las variables

El símbolo para una variable cualquiera será una letra mayúscula, y los valores individuales que puede asumirse simbolizan con la misma letra, minúscula, con un subíndice.

Por ejemplo:

Variables →	X	Y	Z
Valores individuales →	x_i	y_i	z_i
	x_1	y_1	z_1
	x_2	y_2	z_2
	x_3	y_3	z_3
	…	…	…
	x_n	y_n	z_n

siendo n el tamaño de la muestra.

Ejemplo: X: Cantidad de hijos por familia en una muestra de 12 familias.

$x_1 = 2$ → indica que la familia 1 tiene 2 hijos.

$x_2 = 0$ → indica que la familia 2 no tiene hijos

y así sucesivamente.

3. TABULACIÓN DE DATOS

3.1. Serie simple y distribución de frecuencias

3.1.1. Serie simple: es un conjunto de pocos datos (generalmente n < 30 datos).

¿Cómo es el tratamiento adecuado de estos datos?

Generalmente, la primera forma como deben analizarse o explorarse los datos es mediante un gráfico que permita descubrir un patrón de comportamiento, tendencias, variaciones estacionales o simplemente las variaciones aleatorias. Igualmente, el análisis gráfico permite, mediante una simple ojeada, dar una idea de la información y sus características básicas.

Los métodos gráficos se pueden usar para visualizar la información bruta (sin ningún tipo de organización o análisis previo) o la información ya resumida y/o consolidada. En este sentido adquiere plena validez la frase "Una imagen vale más que mil palabras".

Una forma adecuada de representar y ordenar una serie simple es mediante el **diagrama de tallo y hojas.**

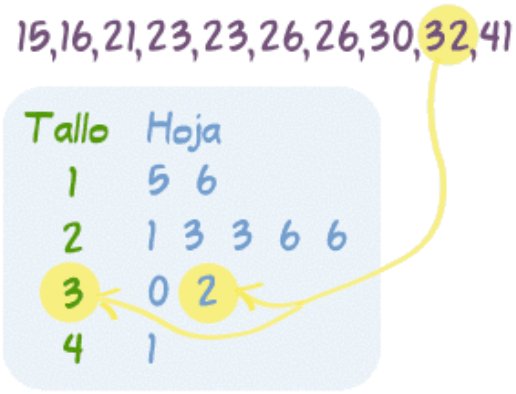
Es un diagrama donde cada valor de datos es dividido en una "hoja" (normalmente el último dígito) y un "tallo" (los otros dígitos). Por ejemplo "32" sería dividido en "3" (tallo) y "2" (hoja). Los valores del "tallo" se escriben hacia abajo y los valores "hoja" van a la derecha (o izquierda) del los valores tallo. El "tallo" es usado para agrupar los puntajes y cada "hoja" indica los puntajes individuales dentro de cada grupo.

Objetivos

- Representación visual de la información
- Descubrir un patrón de comportamiento de los datos, es decir, qué distribución pueden seguir los datos
- Identificar si hay valores extremos o datos anormales en la muestra

Es aplicable para valores formados por al menos dos cifras.

Por ejemplo:



Principio: Cada número se divide en dos partes, una que llamaremos "Tallo" y la otra denominada "ramas u hojas".

Tallo	Formado por uno o más dígitos principales (cifras mas significativas), ubicados a la izquierda del número.
Ramas u hojas	Resto de los números (cifras secundarias) ubicadas a la derecha.

Otro ejemplo. Considere los siguientes números: 65, 57, 79, 69, 53, 63, 71. Los tallos serán las decenas, y las ramas serán las unidades, de la siguiente manera:

Tallo	Ramas
5	73
6	593
7	91

Y con las hojas ordenadas queda:

Tallo	Hojas
5	37
6	359
7	19

Procedimiento:

1. Se define cómo se van a dividir los números en tallos y ramas, es decir, se identifican cuales van a ser los tallos, y cuales va a ser las ramas.
2. En una columna se listan los tallos en orden ascendente.

3. Se recorren los datos y se colocan, en la columna siguiente, las hojas de acuerdo al tallo que tengan.

Observaciones:

- Se recomienda que el número de tallos esté entre 5 y 20.
- A veces, de acuerdo con la información que se tenga, pueden resultar muy pocos tallos, con lo cual las ramas quedan muy concentradas, y realmente no se obtiene mucha información. En estos casos, puede ser conveniente partir los tallos en dos: Un tallo inferior (que tenga, por ejemplo, las ramas menores que 5), y un tallo superior (que tenga las ramas mayores o iguales a cinco). Así, por ejemplo, el tallo 6 puede dividirse en 6I, para los valores entre 60 y 64, y el tallo 6S, para los valores entre 65 y 69.
- Cuando se parten los tallos en dos, todos los tallos deben partirse en dos. Solamente el primero y el último tallo podrían dejarse sin partir, en caso de que en el primer tallo sólo haya información para el tallo superior, y cuando para el último tallo sólo haya información para el tallo inferior.

Otro Ejemplo

Considere la siguiente información sobre duración de baterías de carro, en años. Se pide:

- Construir el diagrama de tallos y hojas usando como tallos la parte entera.
- Construir el diagrama de tallos y hojas partiendo cada tallo en dos.

Duración de baterías (en años)							
2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

Solución

- Usando como tallos la parte entera

Tallos: Dígitos principales (Parte entera).

Ramas: Dígitos secundarios (Parte decimal)

Tallo	**Ramas**	**Frecuencia**
1	9	1
2	2 6 5 9 6	5
3	5 2 7 0 4 1 3 8 1 7 4 6 3 9 1 3 1 7 2 4 8 2 9 0 5	25
4	1 5 6 7 3 4 1 7 2	9
Total		40

- Partiendo cada tallo en dos

En este caso el tallo 1 únicamente tendría la parte superior, y el tallo 4 tendría tanto la parte inferior como la superior

Tallo	Hojas	Frecuencia
1 S	9	1
2 I	2	1
2 S	6 5 9 6	4
3 I	2 0 4 1 3 1 4 3 1 3 1 2 4 2 0	15
3 S	5 7 8 7 6 9 7 8 9 5	10
4 I	1 3 4 1 2	5
4 S	5 6 7 7	4
Total		40

Ordene los datos de las ramas u hojas en los dos diagramas anteriores y analice la diferencia entre los dos diagramas.

3.1.2. **Distribuciones de frecuencia**: es una tabla de resumen en la que los datos se agrupan o arreglan en clases o categorías ordenadas en forma numérica, establecidas de modo conveniente. También se les dice "Datos agrupados".

Datos agrupados sin intervalos: se utiliza cuando la variable, sea discreta o continua, presenta pocos valores diferentes entre sí, repetidos muchas veces cada uno. La tabla se presenta así:

Variable: x_i : x_1 x_2 x_3 ..... x_k
Frecuencia: f_i : f_1 f_2 f_3 ..... f_k siendo $\Sigma_i f_i = n$ (cantidad de datos)
Para $i = 1, 2, 3, ..., k$

Donde f_i se llama frecuencia absoluta e indica la cantidad de veces que se presenta o se repite cada valor de la variable.

La tabla se presenta generalmente en forma vertical.

Ejemplo

X: cantidad de materias aprobadas de los estudiantes que cursaron Estadística en 2009.

Cant. de mat. aprobadas	Frecuencia
0	11
1	18
2	28
3	39
4	33
5	29
6	7
7	8
8	7
9	4
12	2
13	1
14	1
Total	188

Significa que hay 11 estudiantes que no tienen materias aprobadas, 18 estudiantes que tienen una materia aprobada, 29 estudiantes que tienen cinco materias aprobadas, y así sucesivamente.

Datos agrupados en intervalos: se utiliza esta forma de distribución de frecuencias, cuando la variable, sea discreta o continua, presenta muchos valores diferentes entre sí repetidos muchas veces.

El objetivo es distribuir los datos en intervalos de clase, preferiblemente del mismo tamaño, y verificar cuantas observaciones se presentan en cada intervalo (frecuencia absoluta).

El procedimiento para encontrar la distribución de frecuencia es el siguiente:

1. **Encontrar el rango de variación de los datos**. Para ello se requiere calcular los valores mínimo y máximo de la muestra,

   $X_{mín}$ = Mínimo $\{x_i\}$
   $X_{máx}$ = Máximo {xi}
   Rango = R = $x_{máx}$-x_{min}

2. **Definir el número de intervalos de clase (k).** Se recomienda que el número de intervalos de clase esté entre 5 y 15, dependiendo del tamaño de la muestra disponible. Si se usa un número muy bajo, los valores quedan muy concentrados y se pierde mucha precisión, mientras que si se emplea un número muy alto y la muestra es muy pequeña, los datos quedan muy dispersos y realmente no se obtiene mucha información. Como una guía para escoger el número de intervalos puede usarse la fórmula de Sturgess, dada por:

   $$k = 1 + 3.32 \log_{10} n$$

3. **Calcular el tamaño del intervalo de clase o amplitud de clase (a).** Para ello se debe calcular la relación entre el rango de los datos y el número de intervalos. Se tomará como tamaño del intervalo a un valor ligeramente superior a esta relación, es decir,

   $$a > (x_{máx}-x_{mín}) / k$$

4. **Construir los intervalos**. cada intervalo de clase i, está definido mediante un límite inferior (Lim $Inf_i = b_{i-1}$) y por un límite superior (Lim $Sup_i = b_i$). Para el primer intervalo de clase, el límite inferior corresponde al valor más pequeño de la muestra o menor (Lim $Inf_1 \leq b_0 = x_{mín}$), y el límite superior de cada intervalo siempre será igual al límite inferior más el ancho del intervalo de clase (Lim $Sup_i = b_{i-1} + a$).
   Para los demás intervalos diferentes al primero, el límite inferior será igual al limite superior del intervalo inmediatamente anterior (Lim Inf_i = Lim Sup_{i-1}).

De acuerdo con lo anterior se calculan los límites de los intervalos de clase, los cuales estarán dados de la siguiente manera, según se muestra en la tabla

Intervalo	**Límite Inferior b_{i-1}**	**Límite Superior b_i**
1	b_0 = bmin	$b_1 = b_0 + a$
2	$b_1 = b_0 + a$	$b_2 = b_1 + a$
3	$b_2 = b_1 + a$	$b_3 = b_2 + a$
....	....	....
i	$b_{i-1} = b_{i-2} + a$	$b_i = b_{i-1} + a$
....	....	....
k	$b_{k-1} = b_{k-2} + a$	$b_k = b_{k-1} + a$

5. Se toman los valores de la muestra, y se define a qué intervalo corresponde. El intervalo i comprenderá aquellos valores que son mayores o iguales al límite inferior de dicho intervalo (b_{i-1}) y

estrictamente menores que el respectivo límite superior (b_i). Es decir, el valor x quedará en el intervalo i si cumple la siguiente condición.

$$\mathbf{b_{i-1} \leq x < b_i}$$

Es decir, si un valor es igual al límite superior de un intervalo, entonces la observación corresponde al intervalo siguiente.

Para ello se toma cada valor y se compara sucesivamente con el límite superior del primer intervalo, luego con el del segundo, y así sucesivamente hasta que caiga en alguno. Si el valor x queda en el intervalo i, entonces se aumenta en uno la frecuencia del respectivo intervalo.

Ejemplo de aplicación

La inversión real anual de 60 empresas es la siguiente:

10	12	8	40	16	28	10	30	2	8	6	14	16	20	25	36	39	52	30	0
30	4	6	10	18	17	13	17	21	7	6	8	14	7	15	26	14	28	30	26
6	8	39	11	13	15	18	20	30	60	6	12	25	45	26	8	37	12	19	27

Siguiendo los pasos para construir la distribución de frecuencias:

1) Rango: $r = 60 - 0 = 60$ (amplitud total de la serie)
2) Nº de clases: $k = 1 + 3.3 \log 60 = 6.87 \cong 7$
3) Amplitud de clase : $a = 60/7 = 8.57 \cong 9$

Para simplificar la construcción de los intervalos se tomará $a = 10$

4) Formación de los intervalos y 5) registro de datos:

Intervalos	Registros	f_i	x_i
[0 – 10)	///// ///// /////	15	5
[10 – 20)	///// ///// ///// ///// /	21	15
[20 – 30)	///// ///// /	11	25
[30 – 40)	///// ////	9	35
[40 – 50)	//	2	45
[50 – 60)	/	1	55
[60 – 70)	/	1	65
		Total: 60	

← **Marcas de clase**: son los puntos medios de los intervalos. Representan a todos los valores de la variable comprendidos en el intervalo.

$$x_i = (L_i + L_s) / 2$$

3.2. Representación gráfica

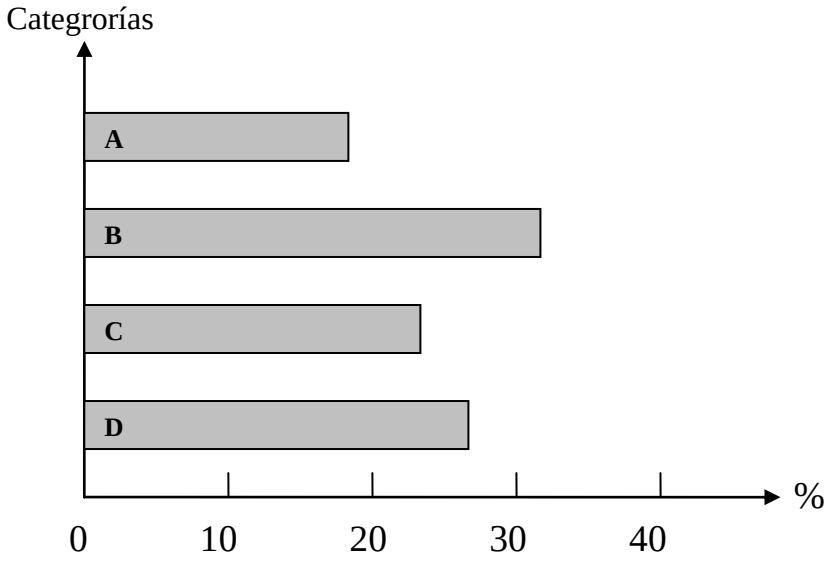
3.2.1. Variable cualitativa o categórica

Existen diversas formas de representar gráficamente una variable cualitativa, pero generalmente se utilizan las barras, y de entre ellas se prefieren las barras horizontales.

Por ejemplo, si se deben representar gráficamente los datos siguientes:

Categorías de la variable:	A	B	C	D
Frecuencias (%):	18	32	23	27

El gráfico adecuado es el de barras horizontales, como se indica a continuación:



Cada barra tiene la longitud del porcentaje que representa.

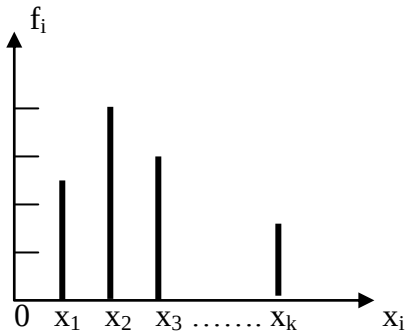
3.2.2. Variable cuantitativa

Serie simple o datos no agrupados: no tiene representación gráfica.

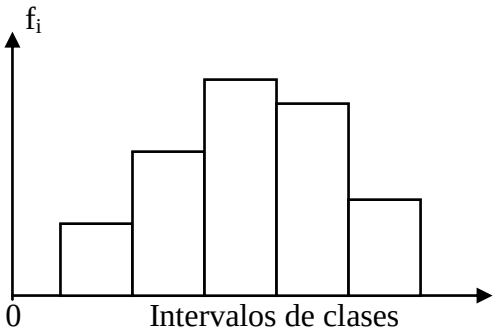
Serie de frecuencias o datos agrupados: en este caso deben distinguirse dos casos diferentes, según que los datos hayan sido agrupados con o sin intervalos.

Datos agrupados sin intervalos → Gráfico de bastones
Datos agrupados en intervalos → Histograma y polígono de frecuencias.

Gráfico de bastones



Histograma y polígono de frecuencias



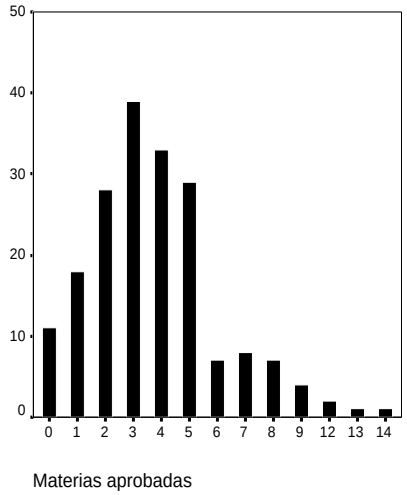
El gráfico de barras adyacentes constituye el **histograma de frecuencias absolutas**, y la línea quebrada que une los puntos medios de los lados superiores de los rectángulos, es el **polígono de frecuencias absolutas.**

En el histograma la frecuencia está representada por el área de los rectángulos, no por la altura de los mismos, por lo tanto, si los intervalos son de amplitud no constante, deberá ajustarse la altura proporcional a las bases distintas de los rectángulos.

En la abscisa se colocan los límites de los intervalos de clase b_0, b_1, b_2,.., b_k , y en la ordenada se dibuja, bien sea la frecuencia absoluta, o la frecuencia relativa. Para cada intervalo se levanta una barra cuya longitud es proporcional a la frecuencia (absoluta, o relativa). La forma que toma el gráfico es la misma, bien sea que se trabaje con frecuencia absoluta o relativa, ya que la diferencia entre las dos es

simplemente un cambio de escala. A veces se dibuja una ordenada izquierda con la frecuencia absoluta, y una ordenada derecha con la frecuencia relativa.

El gráfico de bastones resultante de representar las frecuencias absolutas del ejemplo de la cantidad de materias aprobadas por los estudiantes de Estadística, es el siguiente:



Un ejemplo de histograma y polígono de frecuencias con datos agrupados en intervalos.

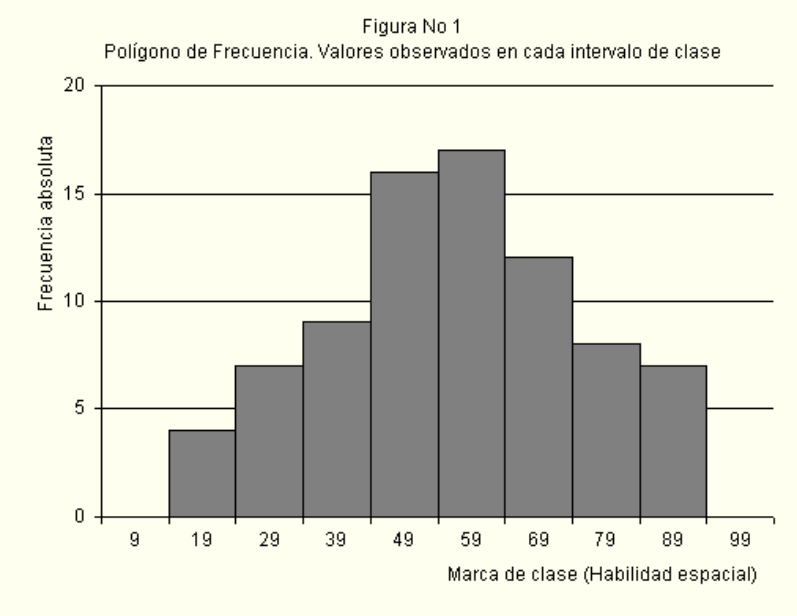
Ejemplo. A un grupo de 80 empleados se les ha aplicado una prueba de habilidad espacial. En una graduación de 0 a 100 han obtenido las puntuaciones dadas en la tabla siguiente. Se pide: Distribuir los datos en intervalos de clase y construir el histograma de frecuencias.

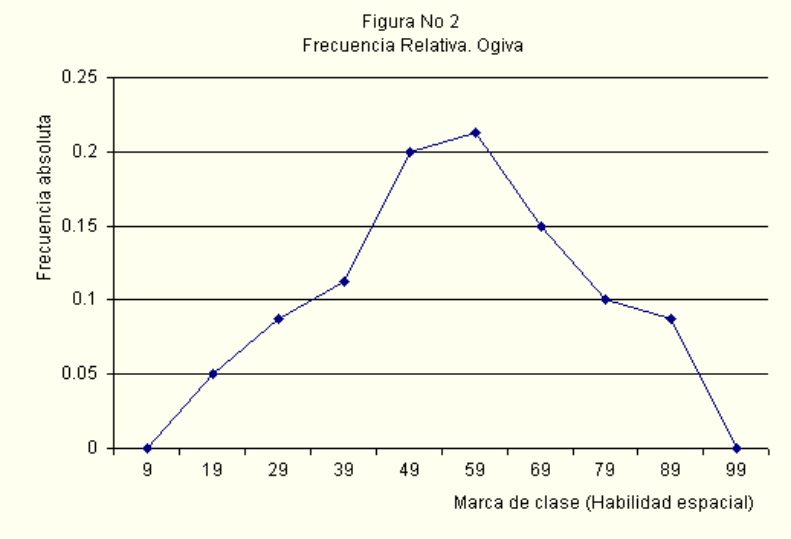
Pruebas de habilidad espacial (puntajes)															
29	78	48	29	30	44	72	73	46	82	84	71	75	84	45	45
47	35	33	54	56	33	62	63	64	36	38	53	54	38	40	57
42	51	52	53	56	57	58	71	76	77	58	60	60	62	65	65
14	16	73	74	45	21	23	66	67	42	43	51	67	70	57	78
55	27	78	48	49	50	51	86	58	59	89	36	37	91	92	93

Conteo de los valores en los intervalos de clase			
Intervalo	**Límite inferior** b_{i-1}	**Límite Superior** b_i	**Conteo**
1	14	24	|||| = 4
2	24	34	||||| || = 7
3	34	44	||||| |||| = 9
4	44	54	||||| ||||| ||||| | = 16
5	54	64	||||| ||||| ||||| || = 17
6	64	74	||||| ||||| || = 12
7	74	84	||||| ||| = 8
8	84	94	||||| || = 7

Intervalo de clase	Límite Inferior b_{i-1} (1)	Límite Superior b_i (2)	Marca de clase MC_i (3)	Frecuencia Absoluta f_i (4)
1	14	24	19	4
2	24	34	29	7
3	34	44	39	9
4	44	54	49	16
5	54	64	59	17
6	64	74	69	12
7	74	84	79	8
8	84	94	89	7
Total				**80**

El histograma de frecuencias absolutas y el polígono de frecuencias correspondiente, se muestran en los gráficos siguientes:





4. OTRAS DISTRIBUCIONES DE FRECUENCIAS

4.1. Distribución de frecuencias relativas

Se simboliza r_i y se obtiene dividiendo la frecuencia absoluta por n.

$r_i = f_i / n$ Así: $r_1 = f_1 / n$; $r_2 = f_2 / n$, etc.

Además, $\sum_i r_i = 1$ (i = 1, 2, 3, …, k)

O bien $\sum_i r_i = 100\%$ si r_i está expresada en porcentaje.

Las frecuencias relativas se utilizan para saber qué proporción o porcentaje de observaciones tiene un determinado valor, o están comprendidas en un intervalo determinado. Su representación gráfica es igual a la de las frecuencias absolutas, sólo cambia la escala del eje de ordenadas, en el cual se representan las frecuencias relativas.

La importancia de la frecuencia relativa radica en que indica la proporción de observaciones referida al total de observaciones realizadas, y esta es una interpretación más completa y más precisa que la de las frecuencias absolutas.

4.2. Distribución de frecuencias acumulativas

Se simbolizan $F_i\downarrow$ o $F_i\uparrow$ según que las frecuencias se acumulen de la forma "Menor que" (Ls) o "Mayor o igual que" (Li), en el caso de que los datos sean agrupados en intervalos, o de la forma "< que" ó " ≥ que" cuando los datos se agruparon sin intervalos.

Cuando los datos han sido agrupados en intervalos de clase, las frecuencias acumuladas se calculan como se indica a continuación:

Menor que	$\mathbf{f_i}$	$\mathbf{F_i}\downarrow$ (la flechita hacia abajo indica el sentido de la acumulación)
Li_1	0	0
Ls_1	f_1	$F_1 = f_1$
Ls_2	f_2	$F_2 = f_1 + f_2 = F_1 + f_2$
Ls_3	f_3	$F_3 = f_1 + f_2 + f_3 = F_2 + f_3$
Ls_4	f_4	$F_4 = f_1 + f_2 + f_3 + f_4 = F_3 + f_4$
. .	.	. . . . .
Ls_k	f_k	$F_k = f_1 + \ldots + f_k = F_{k-1} + f_k$ ⟹ $F_k = n$

La representación gráfica es un diagrama con una línea curva siempre creciente llamado polígono de frecuencias acumuladas u "ojiva". Cuando las frecuencias son acumuladas de la forma "Mayor que" ($F_i\uparrow$) la línea es decreciente. Si se genera un gráfico con ambos tipos de frecuencias acumulativas, el punto de intersección de las ojivas corresponde a la Mediana, una medida de posición. (Ver gráficos de página 11).

Cuando la agrupación de los datos se realiza sin intervalos, entonces:

≤ que	f_i	$F_i\downarrow$
x_1	f_1	$F_1 = f_1$
x_2	f_2	$F_2 = f_1 + f_2 = F_1 + f_2$
x_3	f_3	$F_3 = f_1 + f_2 + f_3 = F_2 + f_3$
x_4	f_4	$F_4 = f_1 + f_2 + f_3 + f_4 = F_3 + f_4$
.	.	. . . . .
x_k	f_k	$F_k = f_1 + \ldots + f_k = F_{k-1} + f_k$ ⟹ $F_k = n$

La representación gráfica es un diagrama escalonado, en este caso el escalón más alto le corresponde a una ordenada igual a n.

$F_i \downarrow$ genera un gráfico escalonado creciente, mientras que $F_i\uparrow$ genera una escalera descendente. El punto de intersección de ambas curvas corresponde a la Mediana, una medida de posisión. (Ver gráficos en página 10).

Las $F_i \downarrow$ se utilizan cuando se desea averiguar cuántas observaciones de la variable son menores o iguales que una de ellas determina, mientras que las $F_i \uparrow$ son más apropiadas cuando se necesita saber qué cantidad de observaciones de la variable son mayores o iguales que alguna de ellas.

≥ que	$\mathbf{f_i}$	$\mathbf{F_i \uparrow}$ (la flechita hacia abajo indica el sentido de la acumulación)
Li_1	f_1	$F_1 = n$
Li_2	f_2	$F_2 = F_1 - f_1$
Li_3	f_3	$F_3 = F_2 - f_2$
Li_4	f_4	$F_4 = F_3 - f_3$
Li_5	f_5	$F_5 = F_4 - f_4$
. .	.	. . . . .
Li_k	f_k	$F_k = F_{k-1} - f_{k-1} = f_k$

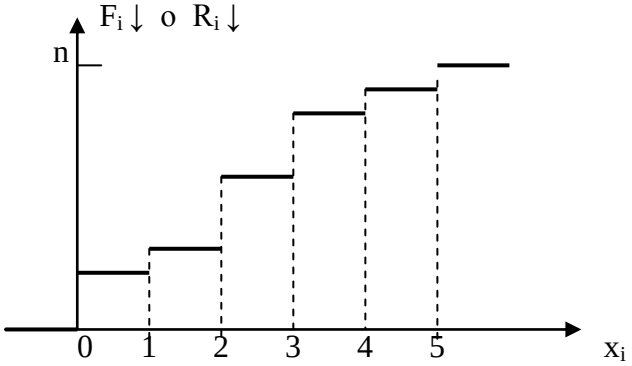
4.3. Distribución de frecuencias relativas acumuladas

Las frecuencias relativas acumuladas se obtienen acumulando las frecuencias relativas, o bien relativizando las frecuencias acumuladas.

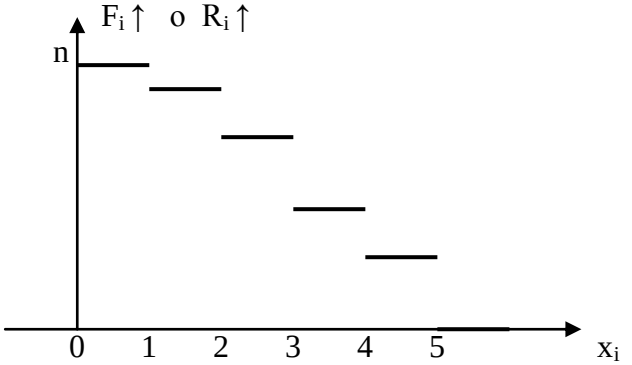
Se simbolizan R_i , con la flechita indicando el sentido de la acumulación.

4.4. Representación gráfica

Datos agrupados sin intervalos: ***Gráfico escalonado***

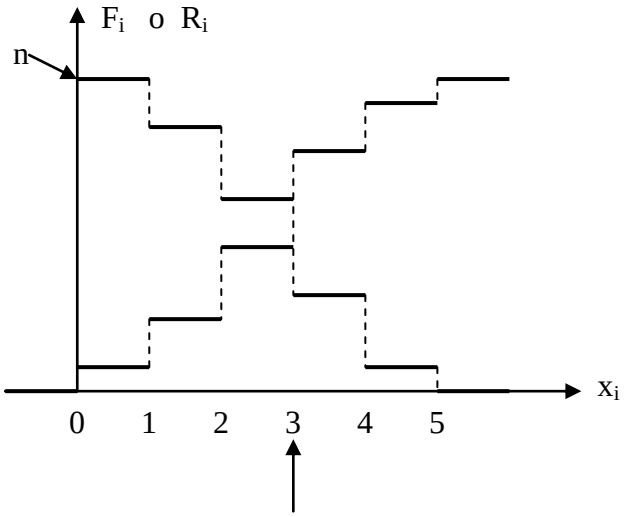


Frecuencias absolutas o relativas acumuladas de la forma “menor que”.



Frecuencias absolutas o relativas de la forma “mayor que”.

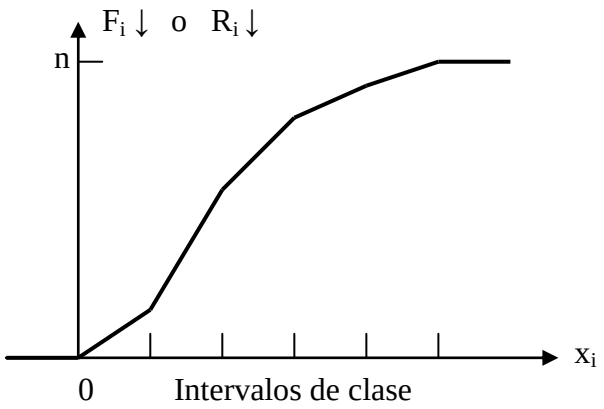
Combinando ambas representaciones en un solo gráfico, se obtiene:



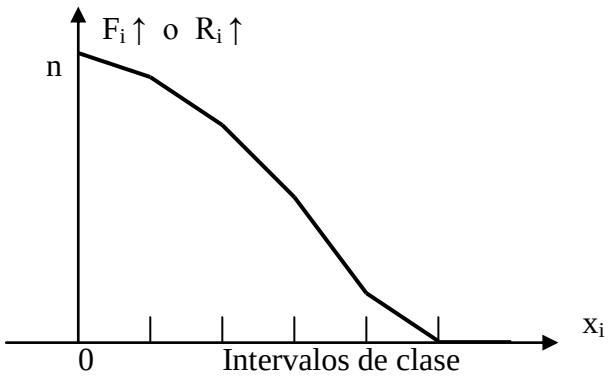
Mediana = 3 (valor de X donde se interceptan ambos gráficos)

Variable continua o discreta agrupada con intervalos de clase

Cuando la variable está agrupada en intervalos de clase, la representación gráfica se llama **polígono de frecuencias acumulativas** u ***"ojiva"***, y toma las formas siguientes:

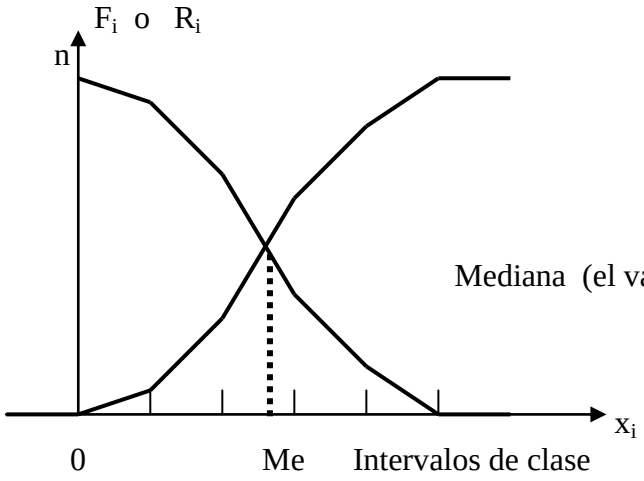


Frecuencias absolutas o relativas acumuladas de la forma "menor que".



Frecuencias absolutas o relativas acumuladas de la forma "mayor que".

Combinando ambas representaciones en un solo gráfico, se obtiene:



Mediana (el valor de X de intersección de ambas curvas)

.

Considerando nuevamente el ejemplo de los puntajes en habilidad espacial de los 80 empleados, la distribución de frecuencias absolutas, relativas y acumulativas, es la siguiente:

Intervalo de clase	**Límite Inferior b_{i-1} (1)**	**Límite Superior b_i (2)**	**Marca de clase MC_i (3)**	**Frecuencias Absolutas f_i (4)**	**Frecuencias Relativas r_i (5)**	**Frecuencias Acumuladas $F_i\downarrow$ (6)**
1	14	24	19	4	0,05	4
2	24	34	29	7	0,0875	11
3	34	44	39	9	0,1125	20
4	44	54	49	16	0,2	36
5	54	64	59	17	0,2125	53
6	64	74	69	12	0,15	65
7	74	84	79	8	0,1	73
8	84	94	89	7	0,0875	80
Total				**80**	**1**	

Interpretación:

Por ejemplo, $r_6 = 0,15$ o bien 15%, indica que la proporción de empleados que obtuvieron un puntaje comprendido entre 64 y 74 puntos es 0,15, o también que el 15% de los empleados obtuvieron puntajes comprendidos entre 64 y 74 puntos.

Y $F_6 = 65$ indica que 65 empleados tienen menos de 74 puntos en la prueba de habilidad espacial.

Si los datos están agrupados en una tabla de frecuencias sin intervalos, como en el ejemplo de la cantidad de materias aprobadas por los estudiantes de Estadistica, la tabla de frecuencias (obtenida utilizando el software SPSS) tiene el aspecto siguiente:

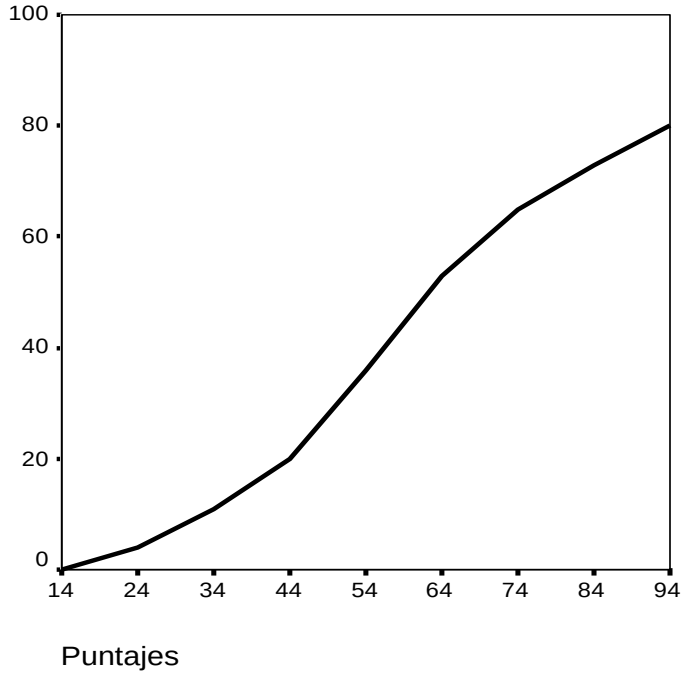
Materias aprobadas

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	11	5,8	5,9	5,9
	1	18	9,5	9,6	15,4
	2	28	14,8	14,9	30,3
	3	39	20,6	20,7	51,1
	4	33	17,5	17,6	68,6
	5	29	15,3	15,4	84,0
	6	7	3,7	3,7	87,8
	7	8	4,2	4,3	92,0
	8	7	3,7	3,7	95,7
	9	4	2,1	2,1	97,9
	12	2	1,1	1,1	98,9
	13	1	,5	,5	99,5
	14	1	,5	,5	100,0
	Total	188	99,5	100,0	
Perdidos	Sistema	1	,5		
Total		189	100,0		

Las frecuencias relativas y acumulativas están expresadas en porcentaje. Por ejemplo, r_i = 20,6 indica que el 20,6% de los estudiantes tiene 3 materias aprobadas. Si el porcentaje se calcula sobre el total de casos válidos, resulta que 20,7% es el porcentaje de estudiantes que tiene 3 materias aprobadas.

Y F_i = 92,0% significa que el 92% de los estudiantes tiene 7 o menos materias aprobadas.

La representación gráfica de las frecuencias acumuladas (ojiva) para el ejemplo de los puntajes de los empleados, es la siguiente:



Para el ejemplo de la cantidad de materias aprobadas, correspondería representar las frecuencias acumuladas mediante los gráficos escalonados.

5. MEDIDAS DESCRIPTIVAS

Para completar la descripción de los datos recopilados se determinan diferentes medidas que caracterizan al conjunto de observaciones desde distintos aspectos. Estas medidas pueden ser: de posición o tendencia central, de dispersión o variabilidad, de asimetría y de kurtosis o agudeza.

Medidas descriptivas

Medidas descriptivas	Nombre de la medida
De posición o tendencia central	Media aritmética
	Moda o modo
	Mediana y cuantiles
	Media geométrica
	Media armónica
De dispersión o variabilidad	Rango o recorrido
	Desviación semiintercuartilar
	Desviación media y desviación mediana
	Varianza o variancia y desviación estándar
	Coeficiente de variación
De asimetría	Coeficiente de asimetría
De kurtosis o agudeza	Coeficiente de kurtosis

Interpretación

Medidas de tendencia central: indican los valores centrales de la variable hacia los cuales tienden a agruparse las observaciones. Comúnmente se los llama ***promedios***.

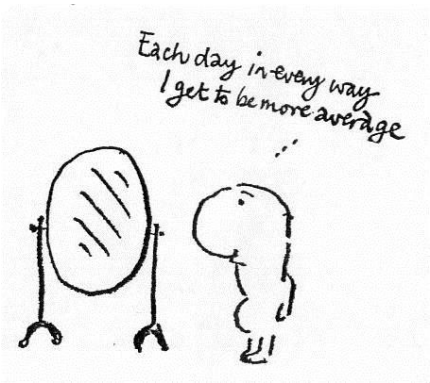
Medidas de dispersión: miden la cantidad de variación, desperdigamiento o diseminación de los datos alrededor de los valores centrales.

Medidas de asimetría: determinan si la distribución de los valores de la variable es simétrica con respecto a los valores centrales, o si existe un sesgamiento hacia la derecha o hacia la izquierda.

Medidas de kurtosis: miden el grado de apuntamiento o agudeza de la distribución de los valores de la variable.

5.1. Medidas de tendencia central

Al describir grupos de observaciones, con frecuencia se desea describir el grupo con un solo número. Para tal fin, desde luego, no se usará el valor más elevado ni el valor más pequeño como único representante, ya que solo representan los extremos más bien que valores típicos. Entonces sería más adecuado buscar un valor central. Las medidas que describen un valor típico en un grupo de observaciones suelen llamarse medidas de tendencia central. Es importante tener en cuenta que estas medidas se aplican a grupos más bien que a individuos. Un promedio es una característica de grupo, no individual.



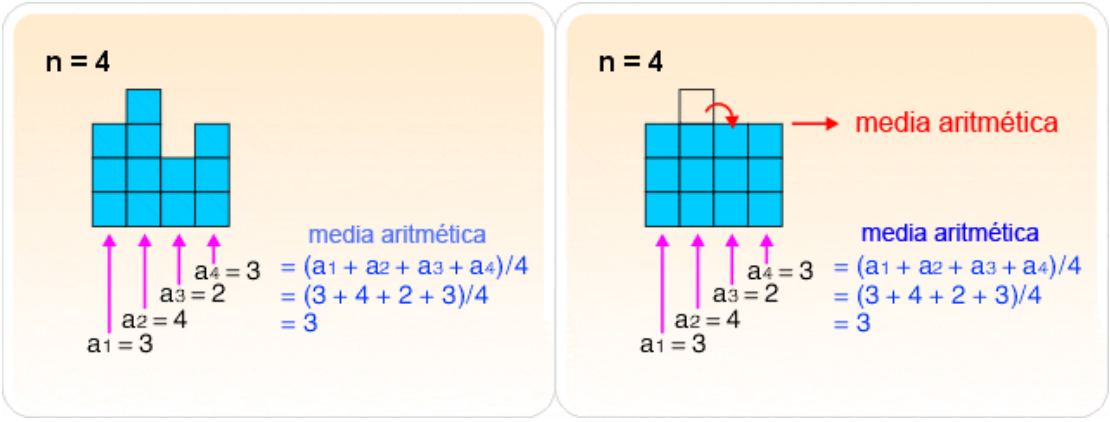
5.1.1. La media aritmética

La medida de tendencia central más obvia que se puede elegir, es el valor obtenido sumando las observaciones y dividiendo esta suma por el número de observaciones que hay en el grupo. La media resume en un valor las características de una variable teniendo en cuenta a todos los casos. Solamente puede utilizarse con variables cuantitativas. Es el promedio más conocido y de mayor uso.

Dada una **serie simple** de observaciones de la variable X: $x_1 \quad x_2 \quad x_3 \ldots\ldots \quad x_n$, la media aritmética es:

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

La media aritmética de un conjunto de n valores es el resultado de la suma de todos ellos dividido entre n. Actúa como *punto de equilibrio*, de modo que las observaciones que son mayores que la media equilibran a las que son menores.



La fórmula para la media aritmética de una serie simple es la siguiente: $\bar{X} = (\sum_{i}^{n} x_i) / n$

Ejemplo de aplicación

De serie simple o datos no agrupados: la inversión real (en miles de dólares) anual de un grupo de 24 pequeñas empresas fue: 12 8 40 6 8 10 30 2 8 6 14 16 20 25 28 30 26 30 26 30 4 6 10.

$\bar{X}$ = (10 + 12 + 8 + . . . + 10) / 24 = 405 / 24 = 16,875 miles de dólares

La inversión real promedio es de 16.875 dólares.

La media de datos agrupados o de una serie de frecuencias

Si los datos están agrupados en una tabla de frecuencias, por ejemplo:

x_i : x_1 x_2 x_3 ….. x_k

f_i : f_1 f_2 f_3 ….. f_k la media aritmética es:

$$\overline{X} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \ldots\ldots + x_k f_k}{f_1 + f_2 + f_3 + \ldots\ldots + f_k}$$

$$\overline{X} = \frac{\sum_i^k x_i f_i}{\sum_i^k f_i}$$

Ejemplo de media aritmética con datos agrupados

En una prueba de aptitud realizada a un grupo de 42 personas se han obtenido las puntuaciones que muestra la tabla siguiente. Calcular la puntuación media.

Intervalos	x_i	f_i	$x_i.f_i$
[10, 20)	15	1	15
[20, 30)	25	8	200
[30,40)	35	10	350
[40, 50)	45	9	405
[50, 60	55	8	440
[60,70)	65	4	260
[70, 80)	75	2	150
		42	1.820

$$\overline{X} = 1820/42 = 43{,}33$$

Si los datos están agrupados en una tabla de frecuencias sin intervalos, los valores x_i son directamente los que asume la variable, los que en el ejemplo anterior se obtuvieron calculando las marcas de clase.

Propiedades de la media aritmética

1. Puede ser calculada en distribuciones con escala relativa e intervalar.

2. Todos los valores son incluidos en el cómputo de la media.

3. Una serie de datos solo tiene una media.

4. Es una medida muy útil para comparar dos o más poblaciones.

5. Es la única medida de tendencia central donde la suma de las desviaciones de cada valor respecto a la media es igual a cero. Por lo tanto podemos considerar a la media como el punto de balance de una serie de datos.

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0$$

Demostración: Basta desarrollar la sumatoria para obtener

$$\sum_{i=1}^{n}(x_i - \overline{x}) = (x_1 - \overline{x}) + \ldots + (x_n - \overline{x}) = (x_1 + \ldots + x_n) - n\,\overline{x} = n\,\overline{x} - n\,\overline{x} = 0$$

Este resultado nos indica que el error cometido al aproximar un valor cualquiera de la variable, por ejemplo x_1, mediante el valor central $\overline{x}$, es compensado por los demás errores:

La suma de las desviaciones de los números 8, 3, 5, 12, 10, de su media aritmética 7,6 es igual a cero.

(8 – 7,6) + (3 – 7,6) + (5 – 7,6) + (12 – 7,6) + (10 – 7,6) = 0,4 – 4,6 – 2,6 + 4,4 + 2,4 = 0

Otro ejemplo con datos agrupados

Obtener las desviaciones con respecto a la media en la siguiente distribución y comprobar que su suma es cero.

l_{i-1} - l_i	n_i
0 - 10	1
10 - 20	2
20 - 30	4
30 - 40	3

Solución:

l_{i-1} - l_i	n_i	x_i	$x_i\,n_i$	$x_i - \overline{x}$	$(x_i - \overline{x})n_i$
0 - 10	1	5	5	-19	-19
10 - 20	2	15	30	-9	-18
20 - 30	4	25	100	+1	+4
30 - 40	3	35	105	+11	+33
	n=10		$\sum x_i n_i = 240$		$\sum = 0$

La media aritmética es:

$$\overline{x} = \frac{1}{n}\sum x_i n_i = \frac{240}{10} = 24$$

Como se puede comprobar sumando los elementos de la última columna,

$$\sum (x_i - \overline{x}) \cdot n_i = 0$$

6. La suma de los cuadrados de las desviaciones de los valores de la variable con respecto a la media aritmética, es un mínimo. Esto significa que si se calcula esa suma tomando otro valor cualquiera distinto de la media aritmética, el resultado siempre será mayor que cuando se toman las desviaciones con respecto a la media.

$$\Sigma(X_i - \overline{X})^2 \ Mínimo$$

$$\sum_{i=1}^{n}(x_i-\overline{x})^2 < \sum_{i=1}^{n}(x_i-k)^2 \qquad \text{si } k \neq \overline{x}$$

Demostración:

Sea $k \neq \overline{x}$. Se verá que el error cuadrático cometido por *k* es mayor que el de $\overline{x}$.

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-k)^2 &= \sum_{i=1}^{n}[x_i-(k-\overline{x}+\overline{x})]^2 \qquad \text{(sumando y restando } \overline{x}\text{)} \\
&= \sum_{i=1}^{n}[(x_i-\overline{x})-(k-\overline{x})]^2 \qquad \text{(y usando el binomio de Newton...)} \\
&= \sum_{i=1}^{n}(x_i-\overline{x})^2 - 2(\overline{x}-k)\underbrace{\sum_{i=1}^{n}(x_i-\overline{x})}_{0} + \underbrace{\sum_{i=1}^{n}(k-\overline{x})^2}_{n(k-\overline{x})^2>0} \\
&> \sum_{i=1}^{n}(x_i-\overline{x})^2
\end{aligned}$$

7. Si a todos los valores de la variable se les suma una constante, la media aritmética queda aumentada en dicho número.

Demostración:

Sea la variable Y = *a* + X, siendo *a* una constante (positiva o negativa).

$\overline{y} = (\sum_{i}^{n} a + x_i) / n = \{an + (\sum_{i}^{n} x_i)\} / n = a + \overline{x}$

8. Si todos los valores de la variable se multiplican por una constante, la media aritmética queda multiplicada por dicho número.

Demostración:

Sea la variable Y = *a*X, siendo *a* una constante (puede ser *a* o 1/*a*).

$\overline{y} = (\sum_{i}^{n} ax_i) / \text{n} = a\,(\sum_{i}^{n} x_i) / \text{n} = a\,\overline{x}$

9. Propiedad de linealidad de la media (resultante de las dos propiedades anteriores)

$$Y = a + b\,X \Longrightarrow \overline{y} = a + b\,\overline{x}$$

10. Dados *r* grupos con $n_1, n_2, ..., n_r$ observaciones y siendo $\overline{x}_1$, $\overline{x}_2$, ..., $\overline{x}_r$ las respectivas medias de cada uno de ellos. Entonces la media de las $n = n_1 + \cdots + n_r$ observaciones es

$$\overline{x} = \frac{n_1\,\overline{x}_1 + \ldots + n_r\,\overline{x}_r}{n_1 + \ldots + n_r}$$

Demostración

Llamando x_{ij} a la *j*-ésima observación del grupo *i*; Entonces es

$$\left.\begin{array}{lcccc} 1^{er}\text{ grupo} & \longrightarrow & x_{11} & \ldots & x_{1n_1} \\ 2^{o}\text{ grupo} & \longrightarrow & x_{21} & \ldots & x_{2n_2} \\ & \ldots & & & \\ r^{esimo}\text{ grupo} & \longrightarrow & x_{r1} & \ldots & x_{rn_r} \end{array}\right\} \Longrightarrow \begin{array}{l} \overline{x}_1 = \left(\sum_{j=1}^{n_1} x_{ij}\right)/n_1 \\ \overline{x}_2 = \left(\sum_{j=1}^{n_2} x_{ij}\right)/n_2 \\ \ldots \\ \overline{x}_r = \left(\sum_{j=1}^{n_r} x_{ij}\right)/n_r \end{array}$$

Así, agrupando convenientemente las observaciones se llega a que

$$\begin{aligned} \overline{x} &= \frac{(x_{11}+\ldots+x_{1n_1})+(x_{22}+\ldots+x_{2n_2})+\ldots+(x_{r1}+\ldots+x_{rn_r})}{n_1+n_2+\ldots+n_r} \\ &= \frac{n_1\overline{x}_1+\ldots+n_r\overline{x}_r}{n} \end{aligned}$$

Observaciones sobre la media aritmética

1. La media se puede hallar solo para variables cuantitativas.

2. La media es independiente de las amplitudes de los intervalos.

3. La media es muy sensible a las observaciones extremas. Si se cuenta con los siguientes valores de la variable peso:

65kg 69kg 65kg 72kg 66kg 75kg 70kg 110kg

La media es igual a 74kg, que es una medida de tendencia central poco representativa de la distribución.

4. La media no se puede calcular si hay un intervalo abierto (con amplitud indeterminada).

	x_i	f_i
[60, 63)	61.5	5
[63, 66)	64.5	18
[66, 69)	67.5	42
[69, 72)	70.5	27
[72, ∞)		8
		100

En este caso no es posible hallar la media porque no se puede calcular la marca de clase del último intervalo.

Ventajas de la media aritmética

- Es la medida de tendencia central más usada.
- El promedio es estable en el muestreo.
- Es sensible a cualquier cambio en los datos (puede ser usado como un detector de variaciones en los datos).
- Se emplea a menudo en cálculos estadísticos posteriores.
- Presenta rigor matemático.
- En la gráfica de frecuencia representa el centro de gravedad.

Desventajas

- Es sensible a los valores extremos. Si alguno de los valores es extremadamente grande o extremadamente pequeño, la media no es el promedio apropiado para representar la serie de datos.
- No es recomendable emplearla en distribuciones muy asimétricas.
- Si se emplean variables discretas o cuasi-cualitativas, la media aritmética puede no pertenecer al conjunto de valores de la variable.

La media aritmética ponderada

Se denomina **media (aritmética) ponderada** de un conjunto de números al resultado de multiplicar cada uno de los números por un valor particular para cada uno de ellos, llamado su *peso*, obteniendo a continuación la suma de estos productos, y dividiendo el resultado de esta suma de productos entre la suma de los pesos + la masa según la característica de cada número inicial. Este "peso" depende de la importancia de cada uno de los valores. O dicho de otro modo es un promedio en el que cada valor de observación se pondera con algún índice de su importancia.

Para una serie de datos $X = \{ x_1, x_2, ..., x_n\}$

a la que corresponden los pesos $W = \{ w_1, w_2, ..., w_n\}$

la media ponderada se calcula como:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

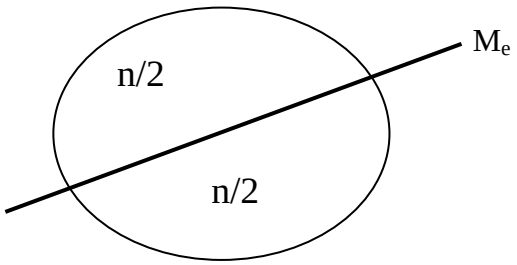
O bien:

$$\bar{x} = \frac{x_1 w_1 + x_2 w_2 + x_3 w_3 + ... + x_n w_n}{w_1 + w_2 + w_3 + ... + w_n}$$

Un ejemplo es la obtención de la **media ponderada** de las notas de una oposición en la que se asigna distinta importancia (*peso*) a cada una de las pruebas de que consta el examen.

5.1.2. La mediana

La mediana de un conjunto finito de valores es aquel valor que divide al conjunto en dos partes iguales, de forma que el número de valores mayor o igual a la mediana es igual al número de valores menores o igual a estos. Su aplicación se ve limitada ya que solo considera el orden jerárquico de los datos y no alguna propiedad propia de los datos, como en el caso de la media.



La notación mas usual que se utiliza para representar a la mediana es $\tilde{\bar{x}}$, Md , M_e ó M_{ed} .

La mediana para datos no agrupados

Lo primero que se requiere es ordenar los datos en forma ascendente o descendente, cualquiera de los dos criterios conduce al mismo resultado.

Sean ordenados lo datos en orden ascendente $x_1, x_2, x_3, \ldots, x_n$

Si el número de valores es impar, la mediana es el valor medio, el cual corresponde al dato $x_{\frac{n}{2}}$.

Ejemplo:

Dados los siguientes datos: 1, 2, 3, 4, 0, 1, 4, 3, 1, 1, 1, 1, 2, 1, 3, para la obtención de la mediana se deberán de ordenar. Tomemos el criterio de orden ascendente con lo que, tendremos:

0, 1, 1, 1, 1, 1, 1, 1, 2, 2, 3, 3, 3 4, 4,

por otro lado, el número de datos es igual a 15 datos, siendo el número de datos impar se elige el dato que se encuentra a la mitad, una vez ordenados los datos, en este caso es **Me = 1**.

Cuando el número de valores en el conjunto es par, no existe un solo valor medio, si no que existe dos valores medios, en tal caso, la mediana es el promedio de los valores, es decir, la mediana es numéricamente igual a

$$Md = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

Si el ejemplo anterior tuviera una observación más, es decir, n = 16, los datos serían:

0, 1, 1, 1, 1, 1, 1, 1, 2, 2, 3, 3, 3 4, 4, 5

Entonces la mediana es: $M_{ed} = (1+2)/2 = 1,5$

La mediana para datos agrupados

Datos agrupados sin intervalos

En este caso la mediana es el valor de la variable al cual le corresponde la frecuencia acumulada, de la forma "menor que", inmediatamente superior a la mitad de las observaciones (n/2).

En el ejemplo de la cantidad de materias aprobadas por alumno en la Cátedra Estadística, cuya distribución de frecuencias se muestra en el cuadro siguiente:

Materias aprobadas

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	11	5,8	5,9	5,9
	1	18	9,5	9,6	15,4
	2	28	14,8	14,9	30,3
	3	39	20,6	20,7	51,1
	4	33	17,5	17,6	68,6
	5	29	15,3	15,4	84,0
	6	7	3,7	3,7	87,8
	7	8	4,2	4,3	92,0
	8	7	3,7	3,7	95,7
	9	4	2,1	2,1	97,9
	12	2	1,1	1,1	98,9
	13	1	,5	,5	99,5
	14	1	,5	,5	100,0
	Total	188	99,5	100,0	
Perdidos	Sistema	1	,5		
Total		189	100,0		

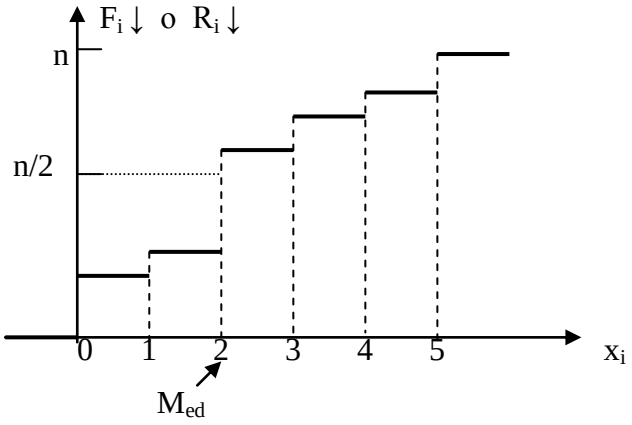
La última columna indica las frecuencias acumuladas porcentuales, por lo tanto n/2 = 50%. La frecuencia acumulada inmediatamente superior a 50% es 51,1%, por lo tanto,

M_{ed} = 3 materias aprobadas

Este resultado se interpreta diciendo que "la mitad de los estudiantes que cursaban Estadística en 2009 tenían 3 o menos materias aprobadas y la otra mitad tenía 3 o más materias aprobadas".

Cálculo gráfico

En el gráfico escalonado de frecuencias absolutas o relativas acumuladas de la forma "menor que":



Se traza una línea paralela al eje de abscisas hasta cortar el gráfico escalonado, por esa intersección se baja una línea perpendicular al mismo eje, y allí se encuentra la mediana.

Datos agrupados en intervalos

La extensión para el cálculo de la mediana en el caso de datos agrupados en intervalos se realiza a continuación:

$$Md = Li + \frac{\frac{n}{2} - f_{acum(i-1)}}{f_{mediana}} A$$

Donde:

M_{ed} = Mediana.

Li = Limite inferior del intervalo donde se encuentra la mediana, la forma de calcularlo es a través de encontrar la posición n/2. En ocasiones en el intervalo donde se encuentra la mediana se conoce como intervalo mediano.

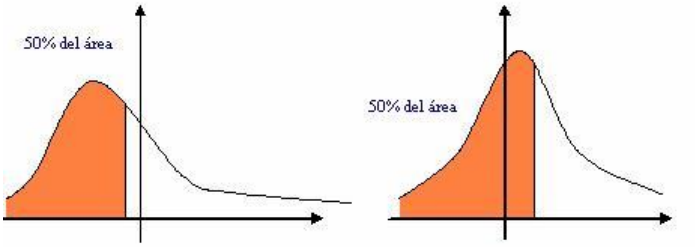
n = Número de observaciones o frecuencia total.

$f_{acum(i-1)}$ = frecuencia acumulada hasta el intervalo anterior al intervalo mediano.

$f_{mediana}$ = Frecuencia del intervalo mediano.

A = Amplitud del intervalo en el que se encuentra la mediana.

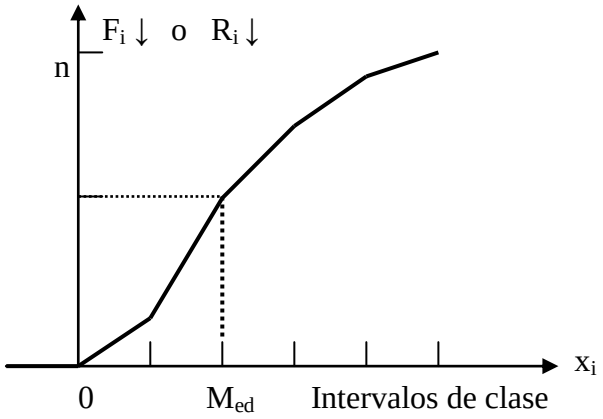
Geométricamente la mediana se encuentra en el valor X que divide al histograma en dos partes de áreas iguales.



Descripción geométrica para dos distribuciones. Se observa que la mediana divide la curva suave que suple al histogramas

Cálculo gráfico

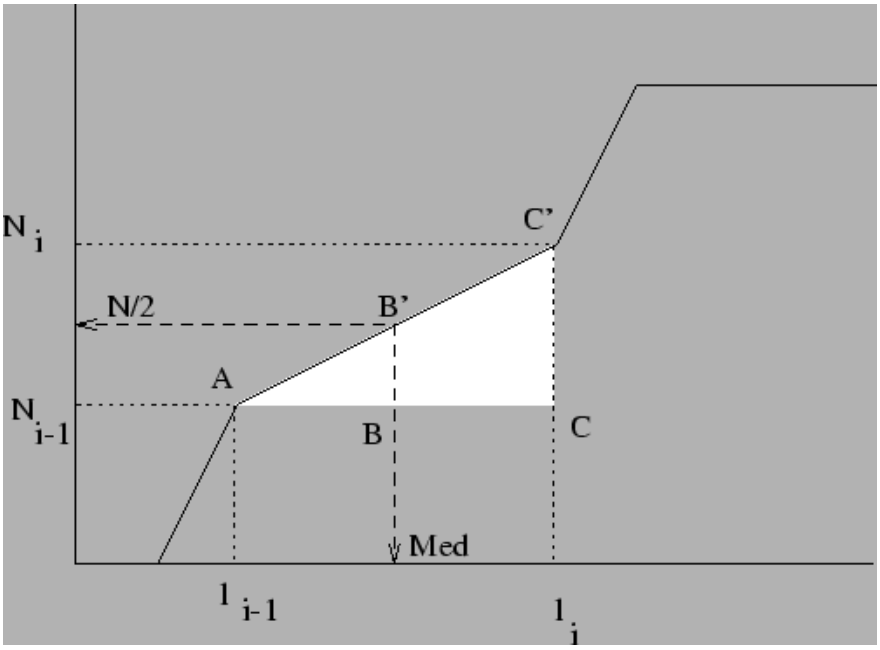
En el gráfico escalonado de frecuencias absolutas o relativas acumuladas de la forma "menor que":



Se traza una línea paralela al eje de abscisas hasta cortar el polígono de frecuencias acumuladas, por esa intersección se baja una línea perpendicular al mismo eje, y allí se encuentra la mediana.

Cómo obtener la fórmula de la mediana con datos agrupados en intervalos

En un gráfico de frecuencias acumuladas de datos agrupados en intervalos,



Sea $(l_{i-1},l_i]$ el intervalo donde hemos encontrado que por debajo están el 50% de las observaciones. Entonces se obtiene la mediana a partir de las frecuencias absolutas acumuladas, mediante interpolación lineal (teorema de Thales) como sigue:

$$\frac{CC'}{AC} = \frac{BB'}{AB} \implies \frac{n_i}{a_i} = \frac{\frac{n}{2} - N_{i-1}}{M_{ed} - l_{i-1}}$$

$$\implies \boxed{M_{ed} = l_{i-1} + \frac{\frac{n}{2} - N_{i-1}}{n_i} \cdot a_i}$$

Ejemplo:

La tabla siguiente muestra la edad de las personas que recibieron atenciones médicas brindadas por el hospital,

Tabla de frecuencias de edades reportadas por la clínica

Clases (Datos en años)	Punto medio de cada clase x_i	Frecuencias de cada clase f_i	Frecuencias acumulada $f_{acumulada}$
$10 \le x < 20$	15	8	8
$20 \le x < 30$	25	20	28
$30 \le x < 40$	35	14	42
$40 \le x < 50$	45	8	50
$50 \le x < 60$	55	2	52
$60 \le x < 70$	65	2	54
$70 \le x < 80$	75	1	55
		55 enfermos atendidos	

Se determina n/2, como n = 55 entonces $n/2=27.5$

El intervalo mediano o la clase donde se encuentra la mediana es la segunda clase, porque le corresponde la frecuencia acumulada inmediatamente superior a la mitad de los datos.

$Li = 20;$

$facum(i-1) = 8$

$fmediana = 20$

$A = 10$

sustituyendo en la ecuación se obtiene

$$Md = Li + \frac{\frac{n}{2} - facum(i-1)}{fmediana} A = 20 + \frac{\frac{55}{2} - 8}{20} 10 = 29.75$$

por lo que se puede concluir que el 50% de las personas atendidas en un fin de semana por el hospital tienen una edad inferior o igual a los 29,75 años, y el otro 50% tiene una edad igual o superior a los 29,75 años.

Propiedades de la mediana

1.- Es única y simple.

2.- Los valores extremos no tienen efectos importantes sobre la mediana, lo que si ocurre con la media. Como medida descriptiva, tiene la ventaja de no estar afectada por las observaciones extremas, ya que no depende de los valores que toma la variable, sino del orden de las mismas. Por ello es adecuado su uso en distribuciones asimétricas.

$$X \rightsquigarrow 2,5,7,9,12 \Longrightarrow \overline{x} = 7, \qquad M_{ed} = 7$$

Si se cambia la última observación por otra anormalmente grande, esto no afecta a la mediana, pero si a la media:

$$X \rightsquigarrow 2,5,7,9,125 \Longrightarrow \overline{x} = 29,6; \qquad M_{ed} = 7$$

En este caso la media no es un posible valor de la variable, y se ha visto muy afectada por la observación extrema. Este no ha sido el caso para la mediana.

3.- Es de cálculo rápido y de interpretación sencilla.

4.- Si una población está formada por 2 subpoblaciones de medianas M_{ed1} y M_{ed2}, sólo se puede afirmar que la mediana, M_{ed}, de la población está comprendida entre M_{ed1} y M_{ed2}

$$M_{ed1} \leq M_{ed} \leq M_{ed2}$$

5.- Puede ser calculada aunque el intervalo inferior o el superior no tenga límites.

6.- La suma de las diferencias de los valores absolutos de *n* puntuaciones respecto a su mediana es menor o igual que cualquier otro valor.

$$\sum_{i=1}^{n} | x_i - M_{ed} |$$

Esta expresión es un mínimo.

7.- El mayor defecto de la mediana es que tiene unas propiedades matemáticas complicadas, lo que hace que sea muy difícil de utilizar en *inferencia estadística*.

Otro ejemplo

Obtener la media aritmética y la mediana en la distribución siguiente. Determinar gráficamente cuál de los dos promedios es más significativo.

l_{i-1} - l_i	n_i
0 - 10	60
10 - 20	80
20 - 30	30
30 - 100	20
100 - 500	10

Solución:

l_{i-1} - l_i	n_i	a_i	x_i	$x_i\, n_i$	N_i	n_i'
0 - 10	60	10	5	300	60	60
10 - 20	80	10	15	1.200	140	80
20 - 30	30	10	25	750	170	30
30 - 100	20	70	65	1.300	190	2,9
100 - 500	10	400	300	3.000	200	0,25
	n=200			$\sum x_i n_i = 6.550$		

La media aritmética es:

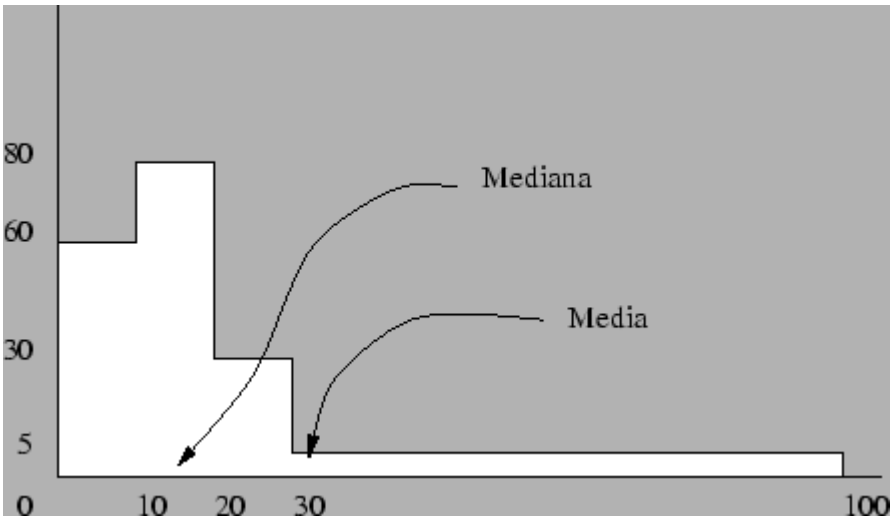
$$\overline{x} = \frac{1}{n}\sum x_i = \frac{6.550}{200} = 32,75$$

La primera frecuencia absoluta acumulada que supera el valor $n/2$=100 es N_i=140. Por ello el intervalo mediano es [10;20). Así:

$$M_{ed} = l_{i-1} + \frac{n/2 - N_{i-1}}{n_i} \cdot a_i = 10 + \frac{100 - 60}{80} \times 10 = 15$$

Para ver la representatividad de ambos promedios, se realiza el histograma de los datos, y se observa que dada la forma de la distribución, la mediana es más representativa que la media.

Para esta distribución de frecuencias es más representativo usar como estadístico de tendencia central la mediana que la media.



5.1.3. La moda o modo (M_o)

Es el valor más frecuente.

Su cálculo es el más simple de los tres correspondientes a estadísticos de centralidad pero la moda es el estadístico de mayor varianza.

La moda puede no existir y cuando existe no es necesariamente única. No tiene sentido en muestras pequeñas en las que la aparición de coincidencias en los valores es con gran frecuencia más producto del azar que de otra cosa.

La media es el estadístico de centralidad más usado cuando uno espera que la población tenga una distribución más o menos simétrica, sin estar clasificada en grupos claramente diferenciados.

En el caso de distribuciones muy asimétricas, con una cola muy larga, la mediana es, normalmente, el valor de elección dado que la media suele estar desplazada respecto al núcleo principal de observaciones de la variable. En estos casos, la mediana es el valor que mejor expresa el punto donde se acumulan mayoritariamente las observaciones de la variable.

En el caso de poblaciones o muestras subdivididas en grupos claramente definidos la media y la mediana carecen, normalmente, de sentido y los valores que más claramente reflejan el comportamiento de las observaciones de la variable son las modas.

La moda de una serie simple (o datos no agrupados)

Dados los siguientes datos: 1, 2, 3, 4, 0, 1, 4, 3, 1, 1, 1, 1, 2, 1, 3, para la obtención de la moda se debe detectar cual es el valor que se repite mayor cantidad de veces. En este caso es:

$$M_o = 1$$

La moda de una serie de frecuencias (o datos agrupados)

Para datos agrupados sin intervalos

En este caso la Moda es el valor de la variable que tiene mayor frecuencia absoluta o relativa. En el ejemplo del número de materias aprobadas por estudiante,

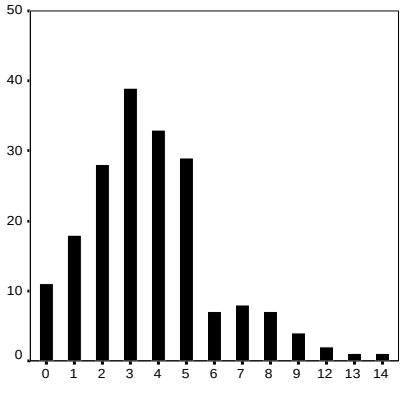
Materias aprobadas

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	0	11	5,8	5,9	5,9
	1	18	9,5	9,6	15,4
	2	28	14,8	14,9	30,3
	3	39	20,6	20,7	51,1
	4	33	17,5	17,6	68,6
	5	29	15,3	15,4	84,0
	6	7	3,7	3,7	87,8
	7	8	4,2	4,3	92,0
	8	7	3,7	3,7	95,7
	9	4	2,1	2,1	97,9
	12	2	1,1	1,1	98,9
	13	1	,5	,5	99,5
	14	1	,5	,5	100,0
	Total	188	99,5	100,0	
Perdidos	Sistema	1	,5		
Total		189	100,0		

La moda es 3, porque es el valor de la variable que tiene la mayor frecuencia absoluta y/o relativa.

Se interpreta diciendo que hay mayor cantidad de estudiantes que tienen 3 materias aprobadas.

Gráficamente, se detecta la moda porque es el valor de la variable al cual, en el gráfico de bastones, le corresponde el bastón más alto.



Para datos agrupados con intervalos

En este caso habrá un intervalo al cual le corresponde la máxima frecuencia absoluta y/o relativa, el **intervalo modal**. En ese intervalo se aplica la fórmula de interpolación para calcular el valor modal.

$$Moda = \left(\frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \right) a_i + l_{i-1}$$

Donde, f_i es la frecuencia absoluta del intervalo modal; f_{i-1} es la frecuencia absoluta del intervalo premodal; f_{i+1} es la frecuencia absoluta del intervalo posmodal; a_i es la amplitud del intervalo modal y l_i es el límite inferior del intervalo modal.

En el ejemplo de las edades de los pacientes atendidos en la clínica durante un fin de semana,

Tabla de frecuencias de edades reportadas por la clínica

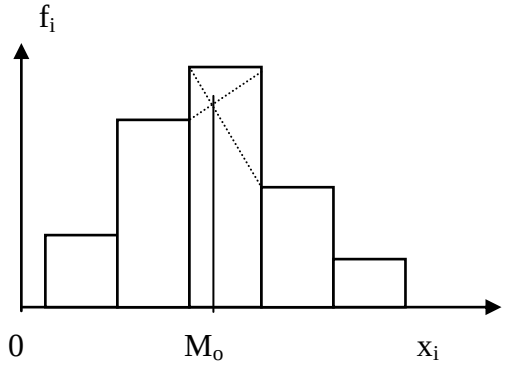
Clases (Datos en años)	Punto medio de cada clase x_i	Frecuencias de cada clase f_i	Frecuencias acumulada $f_{acumulada}$
$10 \leq x < 20$	15	8	8
$20 \leq x < 30$	25	20	28
$30 \leq x < 40$	35	14	42
$40 \leq x < 50$	45	8	50
$50 \leq x < 60$	55	2	52
$60 \leq x < 70$	65	2	54
$70 \leq x < 80$	75	1	55
		55 enfermos atendidos	

La mayor frecuencia absoluta es 20, por lo tanto, el intervalo modal es $20 \leq x < 30$, entonces, aplicando la fórmula en ese intervalo, se obtiene la Mo.

$M_o = 20 + \{(20-8)/ [(20-8) + (20-14)]\}10 = 26{,}67 \approx 27$ años

Significa que, entre los pacientes atendidos, hay mayor cantidad que tiene 27 años.

Gráficamente, la moda se calcula en el histograma de frecuencias absolutas o relativas, como se indica en el gráfico siguiente:



Se hablará de una distribución bimodal de los datos, cuando se encuentren dos modas, es decir, dos datos que tengan la misma frecuencia absoluta máxima. Una distribución trimodal de los datos es en la que se encuentran tres modas. Si todas las variables tienen la misma frecuencia es que no hay moda.

Otras medidas de posición

Cuartiles

La mediana, como se vió, separa en dos mitades el conjunto ordenado de observaciones. Se puede aún dividir cada mitad en dos de tal manera que resulten cuatro partes iguales. Cada una de esas divisiones se conoce como Cuartil y se simboliza mediante la letra *Q* agregando un subíndice según a cual de los cuatro cuartiles se estemos refiera. Se llama primer cuartil Q_1 a la mediana de la mitad que contiene los datos más pequeños. Este cuartil, corresponde al menor valor que supera – o que deja por debajo de él - a la cuarta parte de los datos. Se llama tercer cuartil Q_3 a la mediana de la mitad formada por las observaciones más grandes. El tercer cuartil es el menor valor que supera – o que deja por debajo de él - a las tres cuartas partes de las observaciones. Con esta terminología, la mediana es el segundo cuartil Q_2 y el cuarto cuartil Q_4 coincide con el valor que toma el último dato, luego de ordenados.

Cuartiles para datos sin agrupar

Tal como se concluye de lo anterior, el cálculo será idéntico al de la mediana para el segundo cuartil. El primer cuartil será

$$Q_1 = x_{\left(\frac{N+1}{4}\right)}$$ en caso de que N sea impar y

$$Q_1 = \frac{x_{\left(\frac{N}{4}\right)} + x_{\left(\frac{N}{4}+1\right)}}{2}$$ en caso de que N sea par

Y el tercer cuartil será

$$Q_3 = x_{\left(\frac{N+1}{4}3\right)}$$ en caso de que N sea impar y

$$Q_3 = \frac{x_{\left(\frac{N}{4}3\right)} + x_{\left(\frac{N}{4}3\right)+1}}{2}$$ en caso de que N sea par

Cuartiles para datos agrupados

Sin duda el cálculo para el cuartil dos es idéntico al de la mediana.

Solo quedan por ver los otros dos cuartiles, que serán análogos a los cálculos de la mediana, pero con las salvedades correspondientes

$$Q_1 = \left(\frac{0{,}25 - F_{i-1}}{f_i}\right)a_i + l_{i-1} \quad Q_3 = \left(\frac{0{,}75 - F_{i-1}}{f_i}\right)a_i + l_{i-1}$$

Quintiles

Los quintiles son valores que resultan de dividir la población (el N de las observaciones) en cinco partes iguales (20% en c/u)

Cálculo para datos sin agrupar

El **quintil$_g$** se obtiene identificando el valor que para la variable en cuestión tiene el individuo que ocupa la posición que corresponde al (g.20) % de la población.

Cálculo para datos agrupados a partir de la frecuencia absoluta

$$\text{quintil}_g = \left(\frac{g.\frac{N}{5} - N_{i-1}}{n_i} \right) a_i + l_{i-1}$$

Deciles

Los deciles son valores que resultan de dividir la población (el N de las observaciones) en diez partes iguales (10% en c/u)

Cálculo para datos sin agrupar

El **decil$_h$** se obtiene identificando el valor que para la variable en cuestión tiene el individuo que ocupa la posición que corresponde al (h.10) % de la población.

Cálculo para datos agrupados a partir de la frecuencia absoluta

$$decil_h = \left(\frac{h\frac{N}{10} - N_{i-1}}{n_i} \right) a_i + l_{i-1}$$

Percentiles

Los percentiles son valores que resultan de dividir la población (el N de las observaciones) en cien partes iguales (1% en cada una).

Cálculo para datos sin agrupar

El **percentil$_j$** se obtiene identificando el valor que para la variable en cuestión tiene el individuo que ocupa la posición j%.

Cálculo para datos agrupados a partir de la frecuencia absoluta

$$percentil_j = \left(\frac{j.\frac{N}{100} - N_{i-1}}{n_i} \right) a_i + l_{i-1}$$

5.2. Medidas de dispersión o variabilidad

Las **medidas de dispersión**, también llamadas medidas de variabilidad, muestran la variabilidad de una distribución, indicando por medio de un número, si las diferentes puntuaciones de una variable están muy alejadas de la media. Cuanto mayor sea ese valor, mayor será la variabilidad, cuanto menor sea, más homogénea será. Así se sabe si todos los casos son parecidos o varían mucho entre ellos.

5.2.1. El **rango** o recorrido estadístico es la diferencia entre el valor mínimo y el valor máximo en un grupo de números aleatorios. Se le suele simbolizar con **R**.

Requisitos del rango

- Se ordenan los números según su tamaño.
- Se resta el valor mínimo del valor máximo.

Ejemplo: Para una muestra (8,7,6,9,4,5), el dato menor es 4 y el dato mayor es 9. Sus valores se encuentran en un rango de:

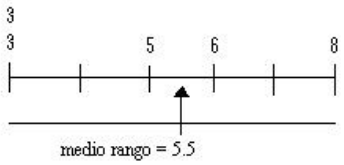
Rango = (9-4) =5

El **medio rango** de un conjunto de valores numéricos es la media del menor y mayor valor, o la mitad del camino entre el dato de menor valor y el dato de mayor valor. En consecuencia el **medio rango** es:

$$medioRango = \frac{(Min + Max)}{2}$$

Ejemplo: Para una muestra de valores (3, 3, 5, 6, 8), el dato de menor valor Min= **3** y el dato de mayor valor Max= **8**. El **medio rango** resolviendolo mediante la correspondiente fórmula sería:

$$medioRango = \frac{(3 + 8)}{2} = 5.5$$



El **rango intercuartílico,** RI es, sencillamente, la diferencia entre el tercer y el primer cuartil, es decir

$$RI = Q_3 - Q_1$$

Esto dice en cuántas unidades de los valores que toma la variable se concentra el cincuenta por ciento central de los casos. Mide la variabilidad de la mitad central de los datos.

Para calcular la variabilidad que una distribución tiene respecto de su media, se calcula la media de las desviaciones de las puntuaciones respecto a la media aritmética. Pero la suma de las desviaciones es siempre cero, así que se adoptan dos clases de estrategias para salvar este problema. Una es tomando las desviaciones en valor absoluto (Desviación media) y otra es tomando las desviaciones al cuadrado (Varianza).

5.2.2. Varianza y desviación estándar

La **varianza** (también denominada **variancia**, aunque esta denominación es menos utilizada) es una medida estadística que mide la dispersión de los valores respecto a un valor central (media), es decir, la media de las diferencias cuadráticas de las puntuaciones respecto a su media aritmética. Suele ser representada con la letra griega **σ** o una **V** en mayúscula.

$$S_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

La expresión de la varianza muestral, en su fórmula de trabajo, es la siguiente:

$$S_X^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1} = \frac{\sum X_i^2 - n\overline{X}^2}{n-1}$$

Y la expresión de la varianza poblacional, es:

$$\sigma^2 = \frac{\sum_{i=1}^{N}\left(X_i - \mu\right)^2}{N} = \frac{\sum X_i^2 - N\mu^2}{N}$$

Propiedades

- La varianza es siempre positiva o 0: $S_X^2 \geq 0$

Cuando todos los datos de la distribución son iguales, la varianza y la desviación típica son iguales a 0.

- Para su cálculo se utilizan todos los datos de la distribución; por tanto, cualquier cambio de valor será detectado.
- Son índices que describen la variabilidad o dispersión y por tanto cuando los datos están muy alejados de la media, el numerador de sus fórmulas será grande y la varianza y la desviación típica lo serán.
- Al aumentar el tamaño de la muestra, disminuye la varianza y la desviación típica. Para reducir a la mitad la desviación típica, la muestra se tiene que multiplicar por 4.
- Si a los datos de la distribución les sumamos una cantidad constante la varianza no se modifica.

$Y_i = X_i + k$

$$S_Y^2 = \frac{\sum(Y_i - \bar{Y})^2}{n} = \frac{\sum[(X_i + k) - (\bar{X} + k)]^2}{n} = \frac{\sum(X_i + k - \bar{X} - k)^2}{n} = \frac{\sum(X_i - \bar{X})^2}{n} = S_X^2$$

- Si a los datos de la distribución les multiplicamos una constante, la varianza queda multiplicada por el cuadrado de esa constante.

$$Y_i = X_i \cdot k$$

$$S_Y^2 = \frac{\sum(Y_i - \bar{Y})^2}{n} = \frac{\sum(X_i \cdot k - \bar{X} \cdot k)^2}{n} = \frac{\sum[k \cdot (X_i - \bar{X})]^2}{n} = \frac{\sum[k^2 \cdot (X_i - \bar{X})^2]}{n} = k^2 \cdot \frac{\sum(X_i - \bar{X})^2}{n} = k^2 \cdot S_X^2$$

- Propiedad distributiva: $V(X \pm Y) = V(X) + V(Y)$

Esta varianza muestral se obtiene como la suma de las diferencias de cuadrados y por tanto tiene como unidades de medida el cuadrado de las unidades de medida en que se mide la variable estudiada.

Como ejemplo, se consideran 10 personas de edades 21 años, 32, 15, 59, 60, 61, 64, 60, 71, y 80. La media de edad de estos sujetos será de:

$$\bar{X} = \frac{21 + 32 + 15 + 59 + 60 + 61 + 64 + 60 + 71 + 80}{10} = 52.3\ años$$

la varianza sería:

$$S^2 = \frac{(15 - 52.3)^2 + (21 - 53{,}2)^2 \ldots (80 - 53{,}2)^2}{10} = 427{,}61$$

La varianza a veces no se interpreta claramente, ya que se mide en unidades cuadráticas. Para evitar ese problema se define otra medida de dispersión, que es la **desviación típica**, o **desviación estándar**, que se halla como la raíz cuadrada positiva de la varianza. La desviación típica informa sobre la dispersión de los datos respecto al valor de la media; cuanto mayor sea su valor, más dispersos estarán los datos. Esta medida viene representada en la mayoría de los casos por **S**, dado que es su inicial de su nominación en inglés.

Desviación típica muestral

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

En el ejemplo anterior es: $S = \sqrt{427,61} = 20,68$ años

Se interpreta diciendo que "la dispersión de los datos mayores que la media por encima de la media, y de los valores menores que la media por debajo de la media, es de 20,68 años.

Desviación típica poblacional

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{N}}$$

Cuando los datos están agrupados, sea con o sin intervalos, cada desviación al cuadrado deberá multiplicarse por la correspondiente frecuencia absoluta antes de realizar la suma.

La desviación estándar es una medida del grado de dispersión de los datos del valor promedio. Una desviación estándar grande indica que los puntos están lejos de la media, y una desviación pequeña indica que los datos están agrupados cerca a la media.

Por ejemplo, las tres muestras (0, 0, 14, 14), (0, 6, 8, 14) y (6, 6, 8, 8) cada una tiene una media de 7. Sus desviaciones estándar son 7, 4 y 1, respectivamente. La tercera muestra tiene una desviación mucho menor que las otras dos porque sus valores están más cerca de 7.

La desviación estándar puede ser interpretada como una medida de incertidumbre. La desviación estándar de un grupo repetido de medidas nos da la precisión de éstas. Cuando se va a determinar si un grupo de medidas está de acuerdo con el modelo teórico, la desviación estándar de esas medidas es de vital importancia: si la media de las medidas está demasiado alejada de la predicción (con la distancia medida en desviaciones estándar), entonces se considera que las medidas contradicen la teoría. Esto es coherente, ya que las mediciones caen fuera del rango de valores en el cual sería razonable esperar que ocurrieran si el modelo teórico fuera correcto. La desviación estándar muestra la agrupación de los datos alrededor de un valor central (la media o promedio).

5.2.3. Coeficiente de variación

Otra medida que se suele utilizar es el **coeficiente de variación** (CV). Es una medida de dispersión relativa de los datos y se calcula dividiendo la desviación típica muestral por la media y multiplicando el cociente por 100. Su utilidad estriba en que permite comparar la dispersión o variabilidad de dos o más grupos. Así, por ejemplo, si tenemos el peso de 5 pacientes (70, 60, 56, 83 y 79 Kg) cuya media es de 69,6 kg. y su desviación típica (s) = 10,44 y la TAS de los mismos (150, 170, 135, 180 y 195 mmHg) cuya media es de 166 mmHg y su desviación típica de 21,3. La pregunta sería: ¿qué distribución es más dispersa, el peso o la tensión arterial? Si comparamos las desviaciones típicas observamos que la desviación típica de la tensión arterial es mucho mayor; sin embargo, no podemos comparar dos variables que tienen escalas de medidas diferentes, por lo que calculamos los coeficientes de variación:

$$CV = \frac{S}{|\bar{x}|}$$

CV de la variable peso = $\frac{10,44}{69,6} = 15\%$

CV de la variable TAS = $\frac{21,30}{166} = 12,8\%$

A la vista de los resultados, se observa que la variable peso tiene mayor dispersión.

5.2.4. Desviación media y desviación mediana

La **desviación media** (DM) es la media aritmética de las desviaciones absolutas de los valores de la variable con respecto a la media.

Para serie simple la fórmula es: $DM = [\sum_{i=1}^{n} |x_i - \bar{x}|] / n$

Para serie de frecuencias la fórmula es: $DM = [\sum_{i=1}^{k} |x_i - \bar{x}| f_i] / n$ siendo $n = \sum_{i=1}^{k} f_i$

La interpretación es, por ejemplo, si la variable peso de los estudiantes presenta una DM = 12 kg, significa que, en promedio, el peso se desvía del peso promedio en 12 kg.

La **desviación mediana** (DM_e) es la media aritmética de las desviaciones absolutas de los valores de la variable con respecto a la mediana.

Para serie simple la fórmula es: $DM = [\sum_{i=1}^{n} |x_i - M_e|] / n$

Para serie de frecuencias la fórmula es: $DM = [\sum_{i=1}^{k} |x_i - M_e| f_i] / n$ siendo $n = \sum_{i=1}^{k} f_i$

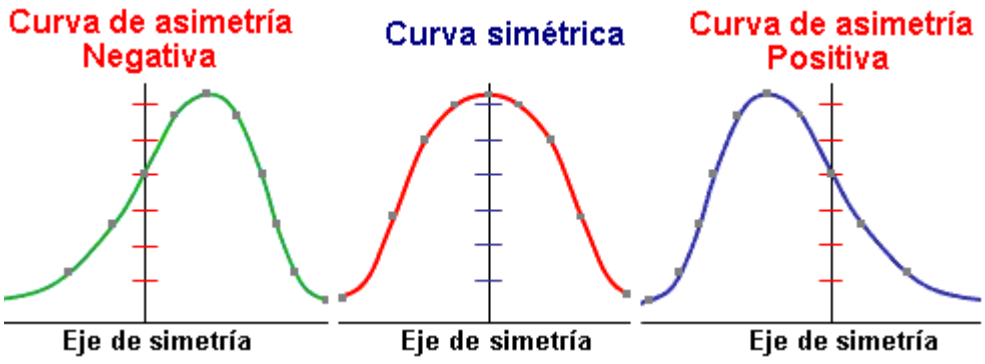
La interpretación es, por ejemplo, si la variable peso de los estudiantes presenta una DM_e = 10,5kg, significa que, en promedio, el peso se desvía del peso mediano en 10,5 kg.

5.3. Medidas de distribución: Asimetría y Kurtosis

Las medidas de distribución permiten identificar la forma en que se separan o aglomeran los valores de acuerdo a su representación gráfica. Estas medidas describen la manera como los datos tienden a reunirse de acuerdo con la frecuencia con que se hallen dentro de la información. Su utilidad radica en la posibilidad de identificar las características de la distribución sin necesidad de generar el gráfico. Sus principales medidas son la *Asimetría* y la *Curtosis*.

5.3.1. Asimetría

Esta medida permite identificar si los datos se distribuyen de forma uniforme alrededor del punto central (Media aritmética). La asimetría presenta tres estados diferentes, cada uno de los cuales define de forma concisa como están distribuidos los datos respecto al eje de asimetría. Se dice que la ***asimetría es positiva*** cuando la mayoría de los datos se encuentran por encima del valor de la media aritmética, la curva es ***Simétrica*** cuando se distribuyen aproximadamente la misma cantidad de valores en ambos lados de la media y se conoce como ***asimetría negativa*** cuando la mayor cantidad de datos se aglomeran en los valores menores que la media.



El ***Coeficiente de asimetría***, se representa mediante la ecuación matemática,

$A_s = (\bar{x} - M_o)/ S$ cuyo campo de variación es: $-1 \leq A_s \leq 1$

- (A_s = 0): Se acepta que la distribución es Simétrica, es decir, existe aproximadamente la misma cantidad de valores a los dos lados de la media. Este valor es difícil de conseguir por lo que se tiende a tomar los valores que son cercanos ya sean positivos o negativos (± 0.5).
- (A_s > 0): La curva es asimétrica positiva por lo que los valores se tienden a reunir más en la parte izquierda que en la derecha de la media.
- (A_s < 0): La curva es asimétrica negativa por lo que los valores se tienden a reunir más en la parte derecha de la media.

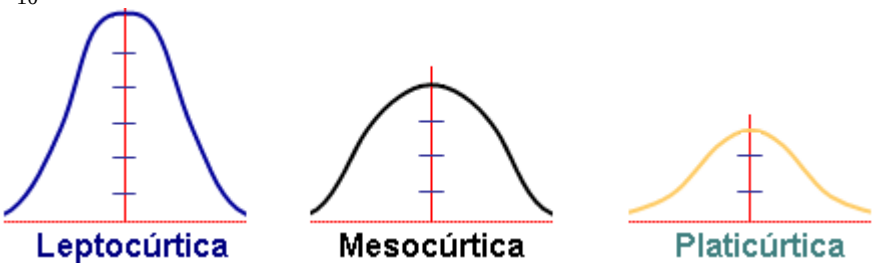
Desde luego entre mayor sea el número (Positivo o Negativo), mayor será la distancia que separa la aglomeración de los valores con respecto a la media.

5.3.2. Curtosis

Esta medida determina el grado de concentración que presentan los valores en la región central de la distribución. Por medio del ***Coeficiente de Curtosis***, se puede identificar si existe una gran concentración de valores (***Leptocúrtica***), una concentración normal (***Mesocúrtica***) ó una baja concentración (***Platicúrtica***).

Para calcular el coeficiente de Curtosis (K) se utiliza la ecuación:

$$K = \frac{\frac{1}{2}(Q_3 - Q_1)}{P_{90} - P_{10}}$$ su campo de variación es $0 \leq K \leq 0{,}5$



- **(K** $\Rightarrow$ 0) la distribución es *Platicúrtica*
- (K $\Rightarrow$ 0,5) la distribución es *Leptocúrtica*
- (K $\Rightarrow$ 0,25) la distribución es *Mesocúrtica*

5.4. Media geométrica

La media geométrica (MG), de un conjunto de *n* números positivos se define como la raíz enésima del producto de los *n* números. Por tanto, la fórmula para la media geométrica es dada por

$$MG = \sqrt[n]{(X_1)(X_2)\cdots(X_n)} \qquad \bar{x} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

Existen dos usos principales de la media geométrica:

1. Para promediar porcentajes, índices y cifras relativas y

2. Para determinar el incremento porcentual promedio en ventas, producción u otras actividades o series económicas de un periodo a otro.

Ejemplo

Supóngase que las utilidades obtenidas por una compañía constructora en cuatro proyectos fueron de 3, 2, 4 y 6%, respectivamente. ¿Cuál es la media geométrica de las ganancias?

En este ejemplo y así la media geométrica es determinada por

$$MG = \sqrt[4]{(3)(2)(4)(6)}$$
$$= 3.464101615$$

y así la media geométrica de las utilidades es el 3.46%.

La media aritmética de los valores anteriores es 3.75%. Aunque el valor 6% no es muy grande, hace que la media aritmética se incline hacia valores elevados. La media geométrica no se ve tan afectada por valores extremos.

Propiedades

El logaritmo de la media geométrica es igual a la media aritmética de los logaritmos de los valores de la variable.

Ventajas:

- considera todos los valores de la distribución y
- es menos sensible que la media aritmética a los valores extremos.

Desventajas:

- es de significado estadístico menos intuitivo que la media aritmética,
- su cálculo es más difícil y
- en ocasiones no queda determinada; por ejemplo, si un valor $x_i=0$ entonces la media geométrica se anula.

Solo es relevante la media geométrica si todos los números son positivos. Como hemos visto, si uno de ellos es 0, entonces el resultado es 0. Si hubiera un número negativo (o una cantidad impar de ellos) entonces la media geométrica sería o bien negativa, o bien inexistente en los números reales.

En muchas ocasiones se utiliza su trasformación en el manejo estadístico de variables con distribución no normal.

La media geométrica es relevante cuando varias cantidades son multiplicadas para producir un total.

Media geométrica ponderada

Al igual que en una media aritmética pueden introducirse pesos como valores multiplicativos para cada uno de los valores con el fin de ponderar o hacer pesar más en el resultado final ciertos valores, en la media geométrica pueden introducirse pesos como exponentes:

$$\bar{x} = \left(\prod_{i=1}^{n} {x_i}^{\alpha_i}\right)^{\frac{1}{\sum_i \alpha_i}} = \left({x_1}^{\alpha_1} {x_2}^{\alpha_2} \dots {x_n}^{\alpha_n}\right)^{\frac{1}{\alpha_1+\dots+\alpha_n}}$$

Donde las α_i son los «pesos».

5.4. Media armónica

La **media armónica**, simbolizada **H**, de una cantidad finita de números es igual al recíproco, o inverso, de la media aritmética de los recíprocos de dichos valores

Así, dados los números $a_1, a_2, \dots, a_n$, la media armónica será igual a:

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{a_i}} = \frac{n}{\left(\frac{1}{a_1} + \cdots + \frac{1}{a_n}\right)}$$

La media armónica resulta poco influida por la existencia de determinados valores mucho más grandes que el conjunto de los otros, siendo en cambio sensible a valores mucho más pequeños que el conjunto.

La media armónica no está definida en el caso de la existencia en el conjunto de valores nulos.

Propiedades

1. La inversa de la media armónica es la media aritmética de los inversos de los valores de la variable.
2. Siempre se puede pasar de una media armónica a una media aritmética transformando adecuadamente los datos.

Ventajas

- Considera todos los valores de la distribución y
- en ciertos casos, es más representativa que la media aritmética.

Desventajas

- La influencia de los valores pequeños y
- El hecho que no se puede determinar en las distribuciones con algunos valores iguales a cero; por eso no es aconsejable su empleo en distribuciones donde existan valores muy pequeños.

Se utiliza para promediar velocidades, tiempos, rendimientos, en general promedios por unidad.

Media Armónica ponderada

Ejemplo: calcular la media armónica de la siguiente distribución:

$\mathbf{x_i}$	$\mathbf{n_i}$
100	10
120	5
125	4
140	3

Para poder hallarla, es necesario que calculemos el inverso de x y el inverso de la frecuencia por lo que ampliaremos la tabla con 2 columnas adicionales:

$\mathbf{x_i}$	$\mathbf{n_i}$	$\mathbf{1/x_i}$	$\mathbf{n_i/x_i}$	$\mathbf{x_i n_i}$
100	10	1/100	0.1	1000
120	5	1/120	0.042	600
125	4	1/125	0.032	500
140	3	1/140	0.021	420
	N= 22		0.195	2520

$$H = \frac{n}{\sum \frac{n_i}{x_i}} = \frac{22}{0{,}195} = 112{,}82 \qquad \overline{X} = \frac{\sum x_i n_i}{n} = \frac{2520}{22} = 114{,}545$$

Entre la media aritmética, la media geométrica y la media armónica se presenta la siguiente relación:

$$\mathbf{H < G <} \ \overline{x}$$

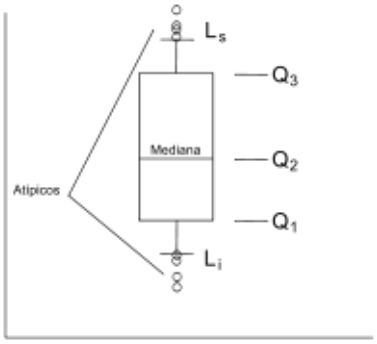
6. UN GRÁFICO MUY DESCRIPTIVO

Diagramas de caja o boxplots

Un **diagrama de caja** es un gráfico, basado en cuartiles, mediante el cual se visualiza un conjunto de datos. Es un gráfico que suministra información sobre la mediana, el cuartil Q_1 y Q_3, sobre la existencia de valores atípicos y la simetría de la distribución.

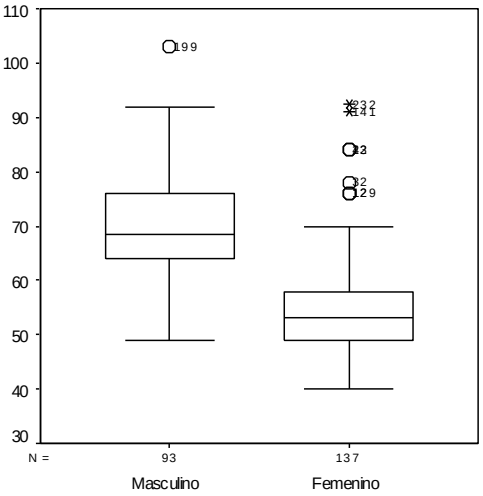
Este diagrama se usa cuando se necesita la mayor información acerca de la distribución de los datos, la ventaja que posee con respecto a los demás diagramas es que este gráfico posee características como centro y dispersión de los datos, y la principal desventaja que posee es que no presenta ninguna información acerca de las frecuencias que presentan los datos.

x_i



Nombre de la Variable

Ejemplo: de una base de datos de 230 estudiantes de Estadística, se representó gráficamente el peso de los estudiantes, diferenciados por sexo, obteniendo el siguiente boxplot:



Sexo de los estudiantes

Interpretación

Por la ubicación de las cajas en el diagrama se deduce que el peso de los varones es bastante mayor que el peso de las mujeres. El grupo está constituido por 93 varones y 137 mujeres. Las medianas ascienden aproximadamente a 69kg en los varones y a 53kg en las mujeres. Para los varones Q_1 es 64kg y Q_3 es 76kg, mientras que las mujeres presentan Q_1 igual a 49kg y Q_3 de 57kg, aproximadamente. Además, el peso de los varones registra mayor dispersión que el de las mujeres (porque la caja es más alta). La distribución del peso de las mujeres es casi simétrica, mientras que la del peso de los varones tiene asimetría positiva (mayor concentración en los menores valores de la variable). También puede verse que el peso de las mujeres tiene mayor kurtosis. Existe mayor cantidad de valores atípicos en los pesos de las mujeres que en los pesos de los varones.

Como puede apreciarse por los comentarios anteriores, este diagrama brinda información sobre las medidas de posición, de dispersión, de asimetría y kurtosis. También sobre diferentes categorías de alguna variable cualitativa (como el sexo de los estudiantes), sobre la cantidad de individuos en cada grupo, y sobre los valores atípicos.

En síntesis, el boxplot proporciona una visión general de la distribución de la variable en estudio.

Como dibujarlo

- Ordenar los datos y obtener el valor mínimo, el máximo, y los cuartiles Q_1, Q_2 y Q_3.
- Dibujar un rectángulo con Q_1 y Q_3 como extremos e indicar la posición de la mediana (Q_2) mediante una línea.
- Calcular los límites superior e inferior, Li y Ls, que identifiquen a los valores atípicos.

  $Li = Q_1 - 1{,}5(Q_3 - Q_1)$ y $Ls = Q_3 + 1{,}5(Q_3 - Q_1)$
- Considerar como atípicos los puntos localizados fuera del intervalo (Li, Ls).
- Dibujar las líneas que van desde cada extremo del rectángulo central hasta el valor más alejado no atípico.
- Marcar como atípicos todos los datos que están fuera del intervalo (Li, Ls).

Referencias

Pita Fernández S, Pértega Díaz, S. (2001). *Estadística descriptiva de los datos*. Unidad de Epidemiología Clínica y Bioestadística. Complexo Hospitalario Juan Canalejo. A Coruña (España).

Universidad de Antioquia. *Estadística Descriptiva.* Estadística Matemática I. Facultad de Ingeniería.

http://ftp.medprev.uma.es/libro/node15.htm

http://dieumsnh.qfb.umich.mx/estadistica/mediana.htm

http://www.bioestadistica.uma.es/libro/node16.htm

http://www.ucm.es/info/genetica/Estadistica/estadistica_basica%202.htm

http://www.spssfree.com/spss/analisis3.html