

PROBABILIDADES Y ESTADÍSTICAS



CONCEPTOS BÁSICOS

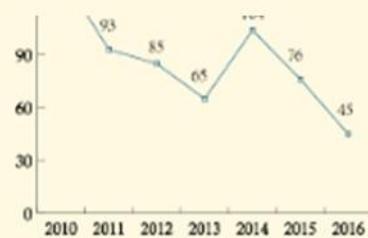


**26% MÁS INGRESANTES
 EN 2016 QUE EN 2015.**

Esa es la mayor cantidad en los últimos 18 años.

**CASI UN QUINTO DE LOS
 INGRESANTES ES MUJER.**

Sin embargo hasta ahora el 14% de los ingresantes 2017 son mujeres.



**LA MITAD DE LOS
 PRIMERIZOS REGRESA.**

El 47% de los ingresantes 2015 volvieron este año.

AÑO 2015

Prof. Cdra. Gladys M. Rouadi



La Estadística constituye una disciplina científica que trata de la selección, análisis y uso de datos con el fin de resolver problemas. A toda persona, tanto en su ejercicio profesional como en su actividad diaria en contacto con diferentes medios, se le ofrece información en forma de datos. Consecuentemente, algunos conocimientos de Estadística le serán de utilidad a la población en general, pero en particular en conocimiento estadístico será de vital importancia para ingenieros de todas las especialidades, científicos y administradores, debido a que manejan y analizan datos cotidianamente. En consecuencia las herramientas básicas de la Estadística les resultan de gran importancia a la hora del ejercicio profesional.

Las aplicaciones de la Probabilidad y la Estadística son numerosas en todos los casos de la ciencia aplicada en donde existan variaciones y donde las conclusiones acerca de un sistema están basadas en datos observados.

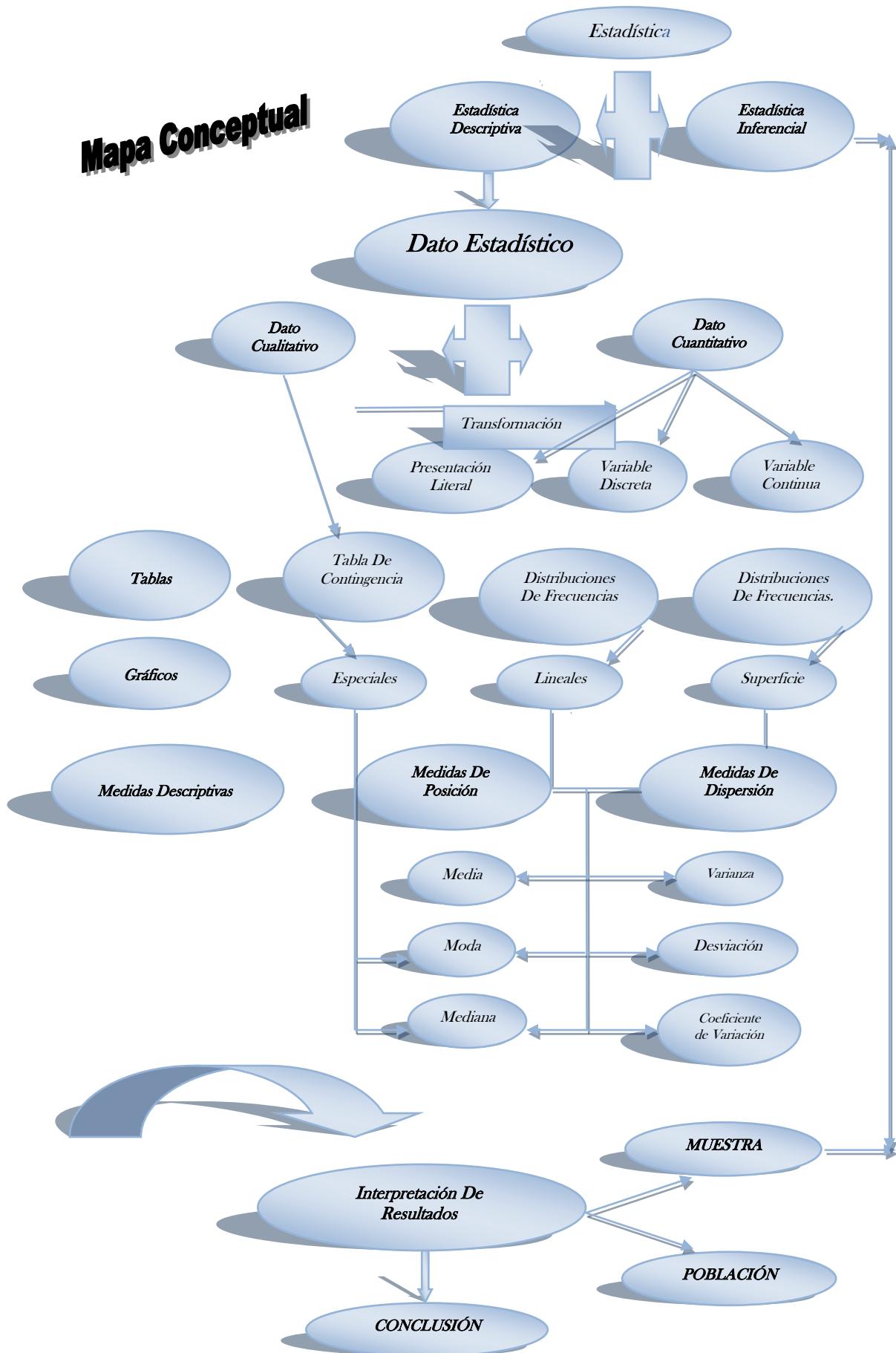
Por Estadística y Probabilidad se entiende los métodos para describir y modelar la variabilidad, además de permitir la toma de decisiones cuando la variabilidad está presente.

Del diseño a la producción, los procesos tienen que ser permanentemente mejorados. Con sus conocimientos técnicos y dotados de habilidades estadísticas básicas para la recolección y representación gráfica de datos, ingenieros y científicos podrán desenvolverse eficientemente. Agradezco a los integrantes de la Cátedra y a los alumnos que colaboraron en la detección de errores que permiten año tras año mejorar al presente material.

Probabilidades y Estadística: conceptos básicos. 1º ed.-Córdoba. ROUADI, Gladys Margarita. Eudecor. 2013. ISBN 978-987-1536-38-2. Fecha de catalogación: 12/04/2013



Mapa Conceptual





CAPITULO N° 1: Datos Estadísticos y Etapas para su Análisis

Objetivos Específicos

Que el estudiante:

Comprenda el concepto y la utilidad de la estadística

Distinga la diferencia entre censo y muestreo

Comprenda los diferentes tipos de datos o variables y sus distintas unidades de medida

Distinga entre variable, unidad de análisis y unidad de relevamiento

Comprenda el procedimiento sistemático que le permita llegar a conclusiones confiables de manera ordenada

Conozca las distintas formas de captación de datos

Contenidos

1. Significado de estadística

1.1. Estadística descriptiva.

1.2. Estadística inferencial.

1.3. Población y muestra.

2. Datos estadísticos

2.1. Variables cuantitativas.

2.1.1. Discretas.

2.1.2. Continuas.

2.2. Variables cualitativas.

2.3. Unidad estadística o unidad de análisis.

2.4. Unidad de relevamiento.

2.5. Escalas de medida.

2.5.1. Escala nominal.

2.5.2. Escala ordinal.

2.5.3. Escala de intervalos.

2.5.4. Escala de razón.

2.6. Las variables según el tipo de unidad de referencia.

2.6.1. Individuales

2.6.2. Agregadas.

2.6.3. Mixtas.

2.7. Las variables según el papel que cumplen en la investigación.

2.7.1. Independientes.

2.7.2. Dependientes.

2.7.3. De Control.

3. Etapas del método científico en el análisis de datos

3.1. Formulación o definición del problema.

3.2. Diseño del experimento.

3.3. Recopilación de datos estadísticos o relevamiento estadístico.

3.3.1. Tipos de fuentes de datos.

3.3.2. Datos secundarios y datos primarios.

3.3.3. Técnicas de recogida de datos primarios: grupos de interés, teléfono, cuestionarios por correo, de puerta a puerta, abordaje en centros comerciales, registros, observaciones, entrevistas, experimentos. Ventajas y desventajas de las técnicas de recogida de datos.

3.3.4. Relevamiento estático y dinámico.

3.4. Clasificación, tabulación y descripción de los resultados.

3.5. Generalización o inferencia final.



1. SIGNIFICADO DE ESTADÍSTICA

La noción de Estadística se derivó originalmente del vocablo *estado*, porque ha sido función tradicional de los gobiernos centrales llevar registros de población, nacimientos, defunciones, etc. Contar y medir estos hechos genera muchas clases de datos.

Diariamente se nos “bombardea” con datos. Las “estadísticas” se nutren de los números generados por los espacios informativos, la publicidad, los sondeos de opinión y los debates públicos.

La persona común concibe la Estadística como columnas de cifras o gráficos. Este concepto se asemeja a la definición tradicional de Estadística: “La compilación, organización, resumen, presentación y análisis de datos numéricos”.

Las organizaciones modernas tienen miles de conceptos de datos en sus archivos de documentos y en las computadoras. Cientos o miles de valores de datos se agregan a ese total todos los días. Algunos de los datos nuevos se generan normalmente durante el registro de las actividades; otros son el resultado de estudios en investigaciones especiales.

Sin los procedimientos estadísticos, ninguna organización podría entender la gigantesca cantidad de datos generados por su actividad.

Es importante la recopilación y el estudio de datos; por eso los conocimientos de Estadística son valiosos para una gran variedad de actividades.

Las oficinas de Estadística del gobierno publican información numérica sobre la inflación y el desempleo.

Quienes se dedican a realizar previsiones, los economistas, los asesores financieros y los que determinan la política de una empresa y del gobierno estudian estos datos para tomar decisiones basadas en la información obtenida.

Las organizaciones, cualquiera sea su naturaleza, generan una gran cantidad de datos que les permite definir indicadores para medir su gestión.

Con el fin de ofrecer un tratamiento adecuado a sus pacientes, los profesionales de la salud deben entender las estadísticas de las investigaciones que se publican en revistas médicas. En política, los funcionarios que ocupan cargos directivos se basan en las estadísticas de la opinión pública para medir su imagen y definir las exigencias de sus votantes.

Las empresas fundan sus decisiones en estudios de mercado sobre los patrones de compra de los consumidores.

Los granjeros y campesinos registran datos para estudiar nuevas composiciones de alimento o nuevas variedades de siembra.

La Estadística se estructuró, como una disciplina científica, en el siglo pasado, pero ya se conocía y se aplicaba en forma rudimentaria desde la antigüedad.



La configuración actual de la Estadística significa la culminación de un proceso en el que pueden distinguirse antecedentes que se desarrollaron en forma independiente y que luego confluyeron, mediante la obra de Laplace y sus continuadores, hacia un solo cuerpo de doctrina y una metodología.

Los estudios de probabilidad de mediados del siglo XVIII, motivados en gran medida por los juegos de azar, dieron lugar al tratamiento matemático de los errores de medición y a la teoría en la que hoy se sustenta la Estadística.

Iniciados por la escuela de estadísticos ingleses y luego continuados en otros países, se desarrollaron entre los últimos años del siglo pasado y lo que va del presente, los sectores modernos de la Estadística dando como resultado que esta disciplina llegara a constituirse en uno de los métodos más potentes de investigación tanto en las Ciencias Sociales como en las Físico-Naturales.

Se llega así al estado actual. En todas las naciones científicamente desarrolladas, se trabaja intensamente en investigaciones teóricas y de aplicación.

Todos los capítulos de la Estadística se renuevan y amplían diariamente, al mismo tiempo que se perfeccionan sus procedimientos de aplicación en diversos campos del conocimiento.

La Estadística trata de la selección, análisis y uso de datos con el fin de resolver problemas. A toda persona, tanto en su ejercicio profesional como en su actividad diaria en contacto con diferentes medios, se le ofrece información en forma de datos.

Consecuentemente, algunos conocimientos de Estadística serán de utilidad a la población en general, pero en particular, el conocimiento estadístico será de vital importancia para Ingenieros, en todas las especialidades, Científicos y Administradores, debido a que de manera rutinaria manejan y analizan datos. Es por ello la necesidad de aportar las herramientas básicas de la Estadística para ejercer sus profesiones.

Además, la Probabilidad, que estudia las variaciones al azar en diversos sistemas, se presentará necesariamente para el estudio de la Estadística Inferencial y para dar sustento a otras aplicaciones de la probabilidad y la Estadística en Ingeniería y Ciencias.

La Estadística moderna ofrece una gran variedad de herramientas analíticas en la toma de decisiones bajo la incertidumbre, es decir cuando no es factible medir con exactitud, relacionándose de esta forma con la Teoría de Probabilidades.

La Estadística propicia un criterio para lograr mejoras, debido a que sus técnicas se pueden usar para describir y comprender la variabilidad, y ésta existe en todo tipo de procesos.

Las aplicaciones de la Probabilidad y la Estadística son numerosas en todos los casos de la ciencia aplicada en donde existan variaciones y donde las conclusiones acerca de un sistema están basadas en datos observados.

En realidad, todo el trabajo experimental tiene esta naturaleza y la variabilidad es el común denominador de estos problemas.



Por Estadística y Probabilidad se entienden los métodos para describir y modelar la variabilidad además de permitir la toma de decisiones cuando la variabilidad está presente.

La Estadística Inferencial, más reciente que la Descriptiva, aporta a la Ingeniería moderna la mayoría de sus aplicaciones, incluyendo la Inferencia y la Toma de Decisiones.

Pocas áreas han experimentado tan poderosamente el impacto del desarrollo reciente de la Estadística como la Ingeniería y Administración Industrial, a través de sus contribuciones a los problemas de la producción, al uso eficaz de los materiales y fuerza de trabajo, a la investigación básica y al desarrollo de nuevos productos, convirtiéndose en una herramienta vital para los ingenieros, ya que les permite comprender fenómenos sujetos a variación y predecirlos o controlarlos eficazmente.

Otro aspecto que no podemos dejar de considerar es el referido al mejoramiento de la calidad. El mundo se encuentra en la actualidad en medio de una revolución internacional en cuanto a mejora de la calidad.

Las enseñanzas e ideas de W. Edwards Deming fueron de gran utilidad para el rejuvenecimiento de Japón. El propio Deming ha señalado (en relación a la industria Estadounidense) que, si se desea sobrevivir, se debe asumir un incesante compromiso con el mejoramiento de la calidad. Del diseño a la producción, los procesos tienen que ser permanentemente mejorados. Con sus conocimientos técnicos y dotados de habilidades estadísticas básicas para la recolección y representación gráfica de datos, ingenieros y científicos pueden participar decisivamente en el cumplimiento de esta meta.

Entonces, la función principal de la Estadística es elaborar principios y métodos que nos ayuden a tomar decisiones frente a la incertidumbre, por lo cual decimos que es el **MÉTODO PARA LA TOMA DE DECISIONES FRENTE A LA INCERTIDUMBRE**.

Se emplea hoy en toda clase de estudios científicos, en toda situación en la cual deba sacarse una conclusión, tomarse una decisión o realizar una predicción, basada en datos empíricos.

Diremos entonces que *Estadística* es un método que, a través de la recolección en masa y el agrupamiento racional de los hechos, permite reseñar y observar los fenómenos colectivos, obtener relaciones numéricas sensiblemente independientes de las anomalías del azar y poner de manifiesto la regularidad de las variaciones.

La *Estadística* puede dividirse en dos ramas:

1.1. *Estadística Descriptiva*

Es la que se va a encargar de la recopilación, tratamiento, presentación y análisis de los datos, con el objeto de resumirlos y describirlos para su interpretación.

Utiliza tablas y gráficos para sintetizarlos y parámetros (si se trabaja con la población) o estadísticos (si se trabaja con una muestra) para describirlos.



1.2. Estadística Inferencial

Es la técnica que permite sacar conclusiones u obtener generalizaciones acerca de un parámetro de población a partir del correspondiente estadígrafo, calculado con la información provista por una muestra de esa población.

Es la que se va a encargar del proceso de utilización de los datos muestrales para la generalización sobre el, o los parámetros de la población de la cual forman parte los datos analizados.

1.3. Población Y Muestra

A fin de entender cómo se pueden aplicar los métodos estadísticos, se debe distinguir entre *Población* y *Muestra*.

Una *Población, Colectivo ó Universo*, es la totalidad de individuos u objetos de interés, acerca de los cuales se desea información, según el objetivo del estudio.

Una Población Es El Conjunto Completo De Individuos O Elementos De Interés

Una Censo Es La Medición De Todos Los Elementos De Una Población De Interés

Los elementos de la población en sentido estadístico pueden:

1. Ser de existencia Real: automóvil, persona, casa
2. Ser de existencia Abstracta: Temperatura, voto
3. Coincidir con unidades naturales, como obreros, turistas, neumáticos.
4. Ser creados artificialmente con el propósito de la investigación. Así, cuando se analiza un campo sembrado de trigo, es común dividir el campo en cuadrados o rectángulos; en este caso los elementos de la población están dados por los cuadrados o rectángulos y no por cada planta de trigo.
5. Ser una entidad simple. Un hombre, una cosa
6. Ser una entidad compleja. Una familia, una escuela

Con respecto al tamaño de las poblaciones, diremos que pueden ser:

<i>Infinitas</i>	<i>Comprende Un Número Infinitamente Grande De Elementos (Unidades De Análisis, Unidades Elementales).</i>
<i>Finitas</i>	<i>Sólo Contiene Un Número Finito De Elementos (Unidades De Análisis, Unidades Elementales).</i>

Una *Muestra* es la parte de la población que se ha seleccionado para el análisis.



En la práctica es a menudo costoso, largo y en algunos casos físicamente imposible realizar un censo, en cuyo caso se recurre al muestreo como medio más práctico para realizar el estudio. Se está tratando de tomar una decisión acerca de la población, en base a los datos provenientes de la muestra, razón por la cual es necesario contar con elementos representativos del total que sólo se obtendrán si la muestra en sí, es representativa.

*Una Muestra Es Un Subconjunto
Seleccionado De La Población*

*Un Muestreo Es La Medición De
Todos Los Elementos De Una*

Elegir una muestra representativa es un problema importante en las investigaciones estadísticas. A menos que sea sencillo y rentable medir cada elemento de la población a través de un censo, el investigador se encuentra con el problema de cómo seleccionar una muestra representativa entre toda la población objeto de análisis. Una muestra representativa puede proporcionar una visión útil de la naturaleza de la población que se estudia, mientras que una muestra no representativa puede sugerir conclusiones totalmente incorrectas sobre esa población. La Estadística ofrece los llamados Procedimientos de Muestreo que nos indican las formas para seleccionar muestras.

Además de las observaciones anteriores, hay otros conceptos que debemos considerar para analizar información: *Dato Estadístico, Unidad de Análisis, Unidad de Relevamiento, Escalas De Medición, y Clasificación De Las Variables Según El Tipo De Unidad De Referencia Y Según El Papel Que Cumplen En La Investigación.*

2- DATOS ESTADÍSTICOS

No toda información es considerada dato estadístico. Para cumplir con este requisito debe tratarse de un conjunto o conjuntos de valores factibles de ser comparados, analizados e interpretados.

Así, el peso de una sola persona, no permite comparación, en cambio, el peso de 1.000 personas sí lo permite.

Antes de poder procesar un conjunto de datos para la toma de decisiones, se deben encontrar los datos apropiados, que por lo general se obtienen contando, midiendo o clasificando individuos u objetos.

Estas medidas se llaman variables porque pueden tomar muchos valores diferentes. En contraposición, una constante tiene un valor fijo.

Entonces, una variable es toda característica o dimensión de un individuo u objeto susceptible de adoptar distintos valores o nombres (categorías).



Por ejemplo, una variable como el peso, toma valores (65 kgs., 70 kgs., etc.), mientras que otra como el sexo toma nombres, categorías (Masculino, Femenino).

En base a ello, las variables pueden clasificarse en:

2.1. Variables Cuantitativas

Son aquellas que arrojan valores numéricos, es decir, surgen de un proceso de conteo ó medición.

Pueden ser, a su vez:

2.1.1. Discretas

Son respuestas numéricas que surgen de un proceso de conteo, es decir, la unidad no es divisible, sólo puede ser definida en términos enteros o ciertos valores fraccionarios especificados.

El número de alumnos en un aula es un ejemplo, ya que la respuesta toma uno de un número finito de valores que se pueden contar. El aula tiene 1, 2, 3,..., 40, etc. alumnos.

2.1.2. Continuas

Son respuestas numéricas que surgen de un proceso de medición. Pueden asumir cualquier valor numérico (cualquier número real) dentro de una amplitud específica. En tal serie, valores sucesivos pueden diferir en cantidades infinitesimales.

Una serie continua es aquella en la que las unidades pueden dividirse en fracciones de cualquier tamaño, por pequeñas que sean, de modo que haya un flujo continuo de valores con graduaciones infinitamente pequeñas.

La estatura de una persona es un ejemplo de variable continua, ya que la respuesta puede tomar cualquier valor dentro de un intervalo, según sea la precisión del instrumento de medición.

El tiempo que utiliza un alumno para concluir sus estudios, es también una variable continua.

Es interesante observar que mientras el peso, la longitud, la altura, el tiempo y la temperatura son variables continuas, sus mediciones son discretas, porque el instrumento de medición tiene algún límite de precisión.

No obstante, para cálculos y análisis estadísticos las consideramos por su naturaleza, como continuas.

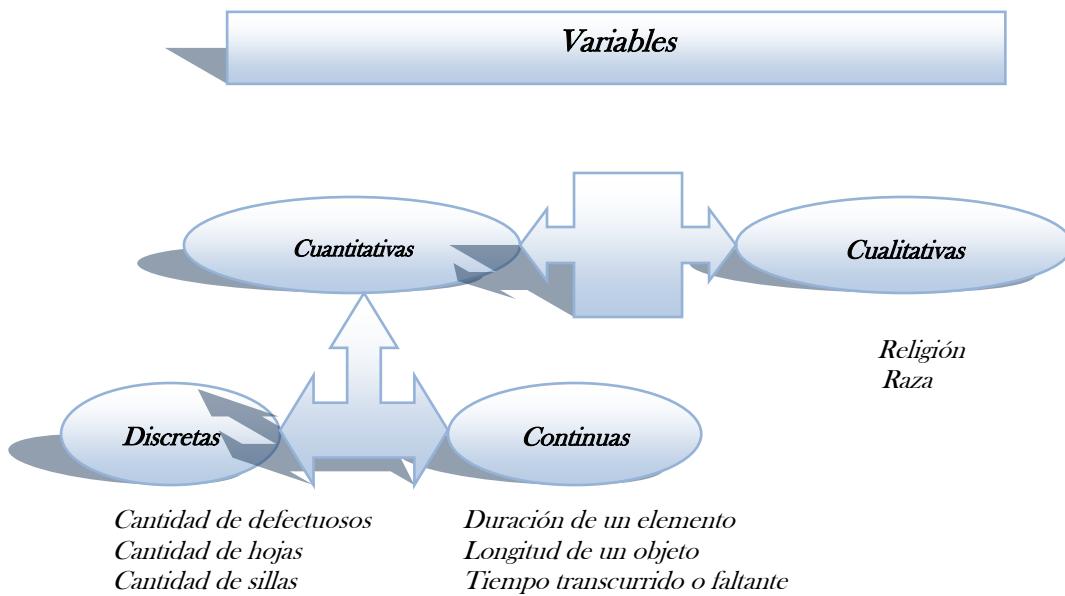
2.2. Variables Cualitativas

También llamadas variables categóricas, son aquellas que arrojan respuestas que se describen por palabras. Los individuos u objetos son poseedores o no poseedores de cierta cualidad o propiedad. Sólo pueden clasificarse, no medirse. A menudo pueden ser



expresadas numéricamente fijando valores que representen a cada una de las categorías consideradas. Así, si por ejemplo trabajamos con el sexo, asignamos el valor 1 al elemento que posee la característica de nuestro interés y 0 al que no la posee. En otros casos podremos establecer otra escala de medición según sea conveniente y representativa para nuestro objetivo.

Resumen



2.3. Unidad Estadística o Unidad De Análisis

Constituida por cada uno de los elementos individuales, que en forma conjunta constituyen la población sobre la que se quiere realizar un relevamiento. En otras palabras, dentro de un relevamiento estadístico es el elemento del conjunto o universo poseedor de la característica o variable, sobre la cual queremos hacer el análisis.

Si queremos estudiar cuál es la práctica religiosa de un grupo de personas, estamos haciendo referencia a una variable: la práctica religiosa, y a unas unidades de análisis, es decir, cada una de las personas.

Cuando por ejemplo, queremos estudiar cómo cambia la práctica religiosa con el paso del tiempo, tendremos que medir a los mismos (o diferentes) individuos en distintos momentos del tiempo.

Si estudiamos la inflación en cada una de las provincias de nuestro país, la inflación será la variable y las provincias la unidad estadística o de análisis.

En los ejemplos dados, las dos variables y las dos unidades, difieren entre sí, no sólo por su contenido sino también por sus propiedades técnicas, lo que implicará la aplicación de técnicas diferentes.



Así, si 100 individuos son católicos y otros 100 no lo son, podremos decir que hay un 50% en cada categoría, pero no podremos calcular una religión promedio.

En cambio, si una provincia tiene una tasa de inflación de 10 puntos y otra la tiene de 15 puntos, podremos decir que una provincia tiene más inflación que otra, con una diferencia de 5 puntos y que la inflación promedio es de $(10 + 15) / 2$.

De esto se desprende, que las variables tienen un nivel de medida, que hace que en algunos casos puedan calcularse promedios y en otros no.

El número de características que puede tener un individuo u objeto es infinito, por lo tanto, el número de variables es infinito, y los puntos de vista desde los que pueden ser observados son también infinitos.

2.4. Unidad De Relevamiento

Están formadas por unidades estadísticas. Es el lugar donde se encuentran las unidades estadísticas o de análisis. Así, en un Censo Ganadero, la unidad de relevamiento la constituye la estancia o cabaña y en un Censo de Población, la familia.

2.5. Escalas De Medida

Para preparar datos para el análisis se debe estar familiarizado con cuatro escalas de medida: *Nominal*, *Ordinal*, *De Intervalos* y *De Razón*. Los datos medidos en una escala de razón contienen más información que los datos medidos en una escala de intervalos, que, a su vez, contienen más información que los medidos en una escala ordinal. Los que menos información contienen son los datos medidos en una escala nominal.

2.5.1. Escala Nominal

Los datos medidos en una escala nominal representan el nivel más bajo de la jerarquía y consisten en categorías en las que se registra el número de observaciones.

Estas categorías no tienen un orden lógico ni una relación específica. Se dice que las categorías son *Mutuamente Excluyentes* puesto que un individuo, objeto o medida queda incluida sólo en una de ellas. El resultado que se obtiene incluye datos cualitativos casi siempre medidos por recuento.

Ejemplo: Número de empresas según tipo de actividad

Tipo De Actividad	Número De Empresas
Servicios	250
Construcción	120
Financiamiento, Seguros Y Bienes Raíces	90
Venta Minorista	85
Venta Mayorista	65
No Clasificados	50
Producción	45
Transporte Y Servicios Públicos	40
Agricultura Y Minería	35
Total	780



Los colores son otro ejemplo de datos nominales, donde las categorías podrían ordenarse según orden de preferencia. Además, si un color es rojo, no puede ser amarillo o verde. Al no importar el orden en que se presenten las categorías y ser mutuamente excluyentes, los datos se clasifican como nominales.

Otros ejemplos son: Sexo (Hombre-Mujer); Categorización por Provincias (Córdoba, Buenos Aires, Santa Fe, Mendoza); Categorización de medios de transporte (colectivo, tren, avión); Categorización de Carreras Universitarias (Medicina, Abogacía, Administración, Psicología)

Una Escala Nominal Consiste En Categorías Mutuamente Excluyentes Que No Implican Ningún Orden Lógico.

2.5.2. Escala Ordinal

Los datos medidos en una escala *Ordinal* consisten en categorías cualitativas en las que hay una progresión en orden.

Ejemplo: Calificaciones expuestas por concepto a un grupo de alumnos.

Calificación	Número De Alumnos
Excelente	4
Muy Bueno	15
Bueno	25
Suficiente	10
Insuficiente	2
Total	56

La calificación de “Excelente” es superior a la de “Muy Bueno” y así sucesivamente, es decir, existe un orden en las categorías.

Los datos medidos en una escala ordinal contienen más información que los medidos en una escala nominal, debido a que las categorías están ordenadas: los valores en una categoría son mayores o menores que los valores en otras categorías, es decir, que las categorías pueden ser ascendentes o descendentes.

Otro ejemplo que mencionaremos es el de los niveles de escolaridad, pues cada categoría implica un nivel educativo más alto que el anterior.

Entonces, podemos categorizar, de la siguiente manera:

Sin Estudios	Primario	Secundario	Terciario	Universitario			
				Grado	Posgrado		
					Especialidad	Maestría	Doctorado
	Incompleto	Incompleto	Incompleto	Incompleto	Incompleta	Incompleta	Incompleto
	Completo	Completo	Completo	Completo	Completa	Completa	Completo

Registrando para un grupo sometido al estudio, el número de individuos, en cada categoría, se obtendrán datos ordinales.



Una Escala Ordinal Se Compone De Distintas Categorías En Las Que Hay Implícito Un Orden.

Tanto las variables nominales como ordinales tienen categorías, que han de ser:

- a) *Colectivamente Exhaustivas*, es decir, cuando permiten clasificar a todas las unidades que estamos investigando. Para ello, se debe incluir una categoría adicional “otros”, en la que se clasifican todos los individuos que no caben en una categoría específica.
 - b) *Mutuamente Excluyentes*, es decir, cuando están definidas sin ambigüedad, pudiendo sólo ser incluidas en una sola categoría.
 - c) Basadas en un *Único Principio Clasificatorio*. Si consideramos “católicos”, “protestantes”..., “asiste regularmente a misa”, las dos primeras son variable religión, mientras que la última se relaciona con la práctica religiosa.

2.5.3. Escala De Intervalos

Los siguientes dos tipos de esquemas de clasificación manejan datos cuantitativos. La escala de intervalos, se produce cuando se toman medidas numéricas sobre algunos elementos y se pueden determinar con exactitud los intervalos entre esas medidas.

Los datos que damos seguidamente, están medidos en una escala de intervalos: la distancia entre cualesquiera dos unidades de temperatura es de tamaño constante y conocida.

Ejemplo: Temperatura que desean los empleados de una fábrica

A este conjunto de datos, lo resumiremos en la siguiente tabla, donde aparecen el número de ocurrencias en cada categoría.

<i>Temperatura</i>	<i>Número De Entrevistados</i>
65-66	6
67-68	14
69-70	27
71-72	22
73-74	6
Total	75

Para este resumen, los datos en la escala de intervalos se convirtieron a una escala ordinal (las categorías están en orden ascendente).

Esta escala es una forma de medida más completa que las anteriores, ya que permite discernir no sólo qué valor observado es el más grande, sino también por cuanto.



Esto se debe a que se mide el ancho del intervalo entre dos valores, en lugar de limitarse a jerarquizarlos. Por ejemplo, el intervalo frío/calor constituye una escala ordinal, mientras que el intervalo 65°F/70°F está basado en una escala de intervalos.

Esta es la diferencia más importante entre los datos de intervalo y los ordinales; con los datos ordinales no se puede medir las distancias entre las categorías, mientras que con los datos de intervalo sí se puede.

En consecuencia, los datos medidos en una escala de intervalos tiene un punto cero arbitrario; es decir, la persona que diseña la escala de manera arbitraria decide dónde colocar el punto cero. Para calificar como una escala de intervalos sólo tiene que definirse la distancia entre los valores numéricos.

Por ejemplo, en el índice de precios al consumidor, si el año base es 1982, el nivel de precios durante 1982 estará en 100. Aunque ésta sea una escala de medición de intervalos iguales, el punto 0 es arbitrario.

Un ejemplo clásico de datos de intervalo es la medida en grados Fahrenheit (dada en el ejemplo anterior). Se puede calcular la cantidad de calor necesario para elevar la temperatura de un cuarto de 40°F a 60°F. Los valores de los datos 40 y 60 no son etiquetas arbitrarias para las categorías. Se trata de valores numéricos definidos con precisión. Además, hay un punto cero arbitrario para la escala de Fahrenheit; la escala de temperatura de grados CELSIUS usa un punto cero diferente. Cada una de estas escalas constituye una escala de intervalos, ya que se puede especificar con precisión la distancia entre cualesquiera dos valores numéricos, y cada escala tiene un punto arbitrario que se define como cero.

La Escala De Intervalos Es Un Conjunto De Valores Numéricos Para Los Que La Distancia Entre Números Sucesivos Es De Tamaño Constante Y Medible Y Cada Escala Tiene Un Punto Arbitrario Que Se Define Como Cero.

2.5.4. Escala De Razón

Por el contrario, los datos medidos en una escala de razón tienen un punto cero fijo o no arbitrario.

Ejemplo: Encuesta sobre la opinión de los votantes, según edad de los entrevistados.

18-19-19-19-20-20-21-21-22-25-25-28-28-29-29-30-32-32
33-33-35-37-37-39-41-41-42-45-45-48-48-49-49-49-50
55-55-57-60-62-65-71-72

Los valores dados se resumen con su media de 38,7.

La mayor parte de las medidas numéricas en las situaciones prácticas dan como resultado datos medidos en una escala de razón. Como ejemplo cabe mencionar la vida útil del cinescopio de un televisor. La diferencia entre 500 y 250 días es una diferencia



medible. Además, puede decirse que un cinescopio de 500 días duró el doble que uno de 250 días.

Por comparación, no diríamos que un día de temperatura de 60° es el doble de calor que uno de 30° . Para los datos de intervalos, la razón de dos números no es apropiada.

Por otro lado, existe un punto cero para la escala de razón: una vida útil cero significa que el cinescopio nunca trabajó, todos entienden con claridad este valor cero.

Otros ejemplos de datos con escala de razón incluyen el peso de los automóviles, los salarios anuales, los períodos que transcurren, las distancias de embarque y las tasas de interés.

La Escala De Razón Consiste En Medidas Numéricas Para Las Cuales Las Distancias Entre Números Tienen Tamaño Constante Y Conocido, Y Donde La Razón Entre Los Números Tiene Algun Significado; Además Existe Un Punto Cero Fijo, No

Entonces, antes de analizar los datos, es importante determinar primero si se recogieron datos cuantitativos o cualitativos. En otras palabras, es vital reconocer de qué tipo de datos se dispone: a) datos de escala nominal u ordinal o b) datos de escala de intervalos o de razón. Se usan técnicas estadísticas diferentes para los tipos básicos de escalas, por lo que se pueden esperar resultados erróneos si se aplica una técnica inapropiada.

La mayor parte de las técnicas estadísticas que presentaremos se usan con datos cuantitativos. Algunas técnicas están diseñadas para datos nominales u ordinales.

Por último haremos una importante observación: Es posible, y en ocasiones recomendable, trabajar en los niveles inferiores de la escala jerárquica. Imaginemos, por ejemplo, que se han recogido datos con una escala de razón referente a la edad de un grupo de individuos. Se han medido las edades numéricas y, obviamente, existe un punto cero fijo: se trata, por lo tanto, de una escala de razón. Puede ser útil convertirla en categorías, por ejemplo:

*Menos de 20
De 20 a menos de 40
De 40 a menos de 60
De 60 a menos de 80
De 80 a menos de 100*

Se obtienen de esta forma datos medidos en una escala ordinal.

Sin embargo, suponga que la recogida de datos original se hizo dentro de estas categorías, entonces, no se pueden convertir a las edades reales, ya que la edad real de una persona que está en la categoría “de 20 a menos de 40” no se conoce.

En otras palabras, es posible descender en la jerarquía de datos pero no ascender.



Una excepción a esta regla unidireccional se emplea con frecuencia en situaciones prácticas.

Los datos para medir el rendimiento de los alumnos, utilizan la siguiente escala:

1. Excelente
2. Muy Bueno
3. Bueno
4. Suficiente
5. Insuficiente

Estos datos se midieron con una escala ordinal, ya que se registraron las frecuencias en cada una de las categorías y se ordenaron.

Así, los datos se analizarán usando técnicas diseñadas para datos cualitativos, por lo cual no se podrá promediar. Promediar es una técnica para datos cuantitativos (sean de intervalos o de razón).

No obstante, puede argumentarse que los intervalos entre las categorías son todos iguales, es decir existe la misma diferencia en el rendimiento entre excelente y muy bueno, que entre muy bueno y bueno. Si este argumento es correcto, se cumple la definición de datos con escala de intervalos: hay distancias medibles e iguales entre valores sucesivos de datos. Por lo tanto, podrían aplicarse técnicas numéricas como la de promediar. Estos métodos sólo pueden usarse después de asignar los números (1, 2, 3,...) a las categorías.

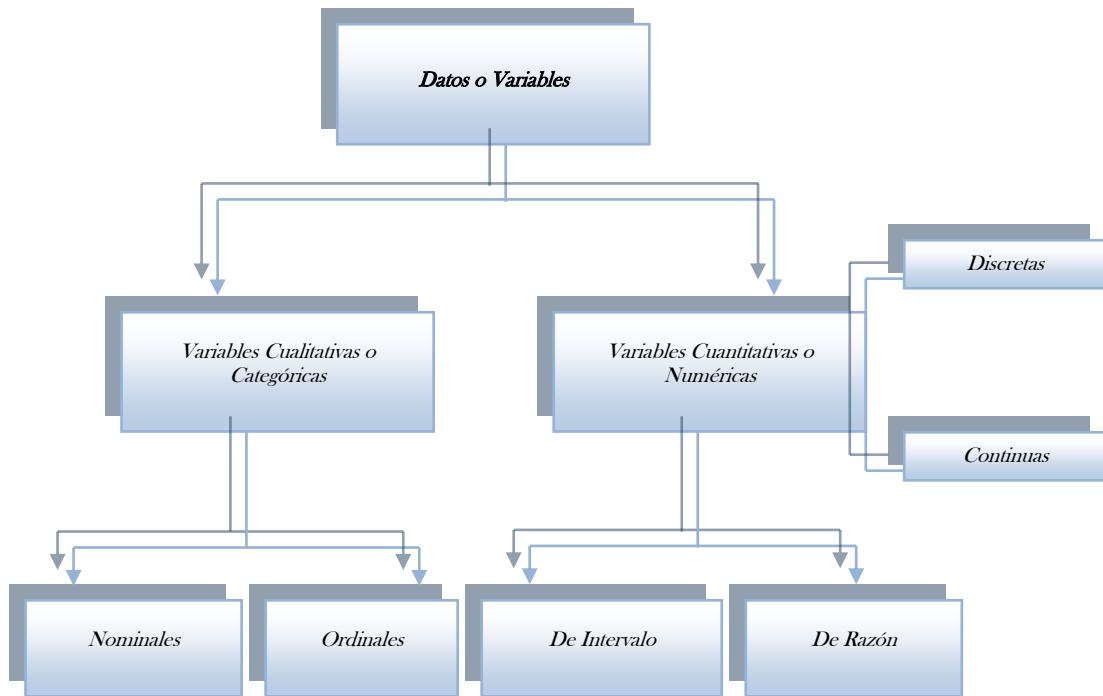
Con frecuencia, los analistas utilizan técnicas numéricas con datos ordinales sin tener en cuenta el hecho de que existan intervalos iguales entre las categorías. Esto puede originar errores graves si en realidad existen intervalos desiguales.

Resumen

<i>Escala de Medida</i>	<i>Características</i>
<i>Nominal</i>	<i>Clasificación única</i>
<i>Ordinal</i>	<i>Jerarquización o calificación</i>
<i>De Intervalos</i>	<i>Diferencia conocida entre dos puntos cualesquiera. Cero arbitrario</i>
<i>De Razón</i>	<i>Diferencia conocida entre dos puntos cualesquiera Cero único o verdadero</i>



Tipos De Datos y Escalas De Medida





2.6. Las Variables Según El Tipo De Unidad De Referencia

También las diferencias se plantean en la unidades, no es lo mismo trabajar con unidades referidas a individuos (por ejemplo, cada integrante de la familia), que con unidades referidas a agregados de individuos (por ejemplo, la familia), o con una mezcla de ambas (cada integrante y la familia). A partir de las diferencias en las unidades de referencia (unidad de análisis o unidad de relevamiento), podemos realizar la siguiente clasificación para las variables:

2.6.1. Individuales

Hacen referencia a características de un individuo o unidad individual de análisis.

Ejemplo:

En un estudio sobre fracaso escolar, el número de horas dedicadas al estudio o las notas obtenidas son variables de tipo individual.

2.6.2. Agregadas

Hacen referencia a características del agregado en el que actúan los individuos.

Ejemplo:

Siguiendo con el fracaso escolar, la titularidad de la institución educativa (pública o privada) serviría como ejemplo de variable agregada. Su valor viene referido a unidades de nivel superior (Institución Educativa), y es independiente de las características de las unidades de nivel inferior (individuos): la institución es pública o privada (variable agregada) por constitución, independientemente de que los alumnos estudien mucho o poco, o saquen buenas o malas notas (variables individuales).

2.6.3. Mixtas

Hacen referencia a unidades de nivel superior (agregadas), pero formadas a partir de características de las unidades de nivel inferior (individuos).

Ejemplo:

Si los alumnos de una institución son de clase media (variable individual), podemos decir que la institución es de clase media (variable agregada-individual). Si cambian los alumnos que van a la institución, también cambia el valor de la variable a nivel de institución, cosa que no ocurría con las variables agregadas.



2.7. Las Variables Según El Papel Que Cumplen En La Investigación

Según el papel, posición o función que cumplen las variables en la investigación, se habla de variables independientes, dependientes y de control. Mientras que el nivel de medida de una variable es algo intrínseco a ella misma, se refiere a la naturaleza de la variable, la distinción entre variables independientes, dependientes o de control se basa en la relación que establecen entre sí: si una variable es independiente, sólo por relación a otra(s) que es (son) dependiente (s), y viceversa, es decir que como mínimo se necesitan dos variables; mientras que una variable es de control por relación a otras dos que actúan como independiente y dependiente, es decir que como mínimo se necesitan tres variables. Ninguna variable en sí misma es independiente, dependiente o de control. La relación con otras variables es lo que determina su calificación.

2.7.1. Variables Independientes (Predictoras - Explicativas)

Se llaman variables independientes a aquellas que toman ó tienen valores ó categorías que influyen en otras variables. En la investigación experimental es el investigador el que da valores (categorías) a una variable para ver cómo influyen sobre otras variables. Por ejemplo, el investigador somete a un conjunto de personas a mirar durante un cierto tiempo una publicidad (variable independiente: tiempo que se ve la publicidad) para observar sus reacciones.

En la investigación no experimental los valores (categorías) no son controlados (asignados) por el investigador, sino que los tienen los individuos (unidades, objetos de análisis). Los estudios que tiene una persona (variable independiente) seguro que influyen en sus actitudes, opiniones, etc., sin que el entrevistador se los haya dado. Las variables independientes también reciben el nombre de predictoras, puesto que a partir de su conocimiento vamos a tratar de predecir los valores de otras variables, y explicativas, puesto que van a ser utilizadas para explicar otras variables.

2.7.2. Variables Dependientes (Criterio - Explicadas)

El valor (categoría) de las variables dependientes depende del valor que hayan tomado (investigación experimental) o tengan (investigación no experimental) las variables independientes. Por ejemplo, según sea el nivel de estudios de las personas (variable independiente), así serán, o puede que sean, sus ingresos (variable dependiente).

También reciben el nombre de explicadas, por ser las variables que hay que explicar en la investigación. Se trata de las variables que dan origen a la investigación.

2.7.3. Variables De Control

Las variables de control sirven para comprender mejor la relación entre una variable independiente y otra dependiente. Hay tres situaciones típicas donde las variables de control son absolutamente necesarias:

- Cuando una técnica estadística muestra que dos variables están relacionadas, y dudamos si entre ambas existe una relación, no sólo estadística, sino causal o de dependencia.
- No existe duda sobre la relación de dependencia entre dos variables, pero nos preguntamos porque.



- En muchas ocasiones se nos presenta la duda de saber si la relación que se establece entre dos variables funciona en todas las circunstancias o, por el contrario, tan sólo se manifiesta en determinadas ocasiones.

Resumen





3- ETAPAS DEL MÉTODO CIENTÍFICO EN EL ANÁLISIS DE DATOS

En Estadística, como método para analizar la información, es necesario cumplimentar con una serie de etapas a los fines de lograr resultados congruentes, homogéneos y fáciles de interpretar.

Para ello, consideraremos cinco etapas:

Formulación o Definición del Problema

Diseño del Experimento

Recopilación de Datos Estadísticos

Clasificación, Tabulación y Descripción de los resultados

Generalización o Inferencia Final

Desarrollaremos a continuación cada una de ellas.

3.1. Formulación o Definición Del Problema

En Estadística, la primera tarea es conocer exactamente que ha de ser investigado, es decir formular el problema o pregunta lo más clara y precisamente posible. Sólo entonces, puede decidir el investigador cuáles datos son relevantes al problema. Si se fracasa en esta formulación los datos compilados pueden ser irrelevantes o inadecuados.

La calidad de las conclusiones estadísticas depende de la corrección y precisión de los datos, que a su vez dependen de la exactitud en la formulación del problema. Las técnicas estadísticas, por muy refinadas y precisas que sean, no pueden ayudar a alcanzar decisiones si son aplicadas a datos inapropiados.

3.2. Diseño Del Experimento

Una vez formulado con precisión el problema que requiere análisis estadístico, el investigador debe decidir si estudiar a todos los individuos u objetos involucrados, o sólo a una parte de ellos, extraída del total. En la práctica es a menudo costoso, lento o aún físicamente imposible efectuar un censo, debiendo por ello recurrir al muestreo. A partir de una muestra, inferimos sobre la población de la cual ha sido extraída, es, por este motivo, que la muestra debe ser representativa de la población, es decir, debe representar adecuadamente a la población. Supone preguntas como éstas: ¿Qué tipos de datos deben recogerse?, ¿Cómo deben ser compilados?, ¿De qué tamaño debe ser la muestra? Estas preguntas corresponden al Diseño Experimental o Diseño de Muestras.

3.3. Recopilación de Datos Estadísticos o Relevamiento Estadístico

Se refiere a los métodos usados para obtener información pertinente de los individuos u objetos involucrados en el estudio, sea éste un censo ó un muestreo.

En otras palabras, es aquella etapa que tiene por objeto la extracción y recolección de datos a partir de las fuentes que lo suministran.



3.3.1. Tipos de Fuentes de Datos

3.3.2. Datos Secundarios y Datos Primarios

Los datos necesarios para elaborar un análisis estadístico o bien se encuentran disponibles o deben recogerse. Los datos que se encuentran disponibles se denominan **datos secundarios** y los datos que deben recogerse se llaman **datos primarios**.

Los Datos Primarios Se Recogen Específicamente Para El Análisis Deseado. Los Datos Secundarios Ya Se Han Compilado Y Están Disponibles Para El Análisis Estadístico.

Existen muchas fuentes de datos secundarios. Las bibliotecas son, quizás, el ejemplo más obvio. Algunas fuentes de datos secundarios incluyen redes de computadoras. Los suscriptores a estos servicios pueden tener acceso a grandes cantidades de datos secundarios por medio del teléfono o su propia computadora. Los gobiernos generan cantidades ingentes de datos cada año. También pueden encontrarse en publicaciones nacionales o internacionales.

La ventaja al usar datos secundarios para una investigación estadística es que ya se dispone de ellos y no es necesario recogerlos para un proyecto específico. Incluso la compra de ellos es, por lo general, menos costosa que recoger datos primarios. La desventaja radica en que no siempre cubren las necesidades específicas del análisis.

3.3.3. Técnicas De Recogida De Datos Primarios

Las técnicas que se presentan seguidamente son las más usadas en la práctica para reunir la información necesaria para el análisis y la toma de decisiones. Se requiere experiencia y habilidad para determinar qué técnica o combinación de técnicas se adecuan mejor al estudio específico. La clave para realizar una buena investigación reside, en gran medida, en la pericia a la hora de elegir la técnica adecuada.

Grupos de interés

Se usan como un método preliminar de recogida de datos. Están integrados por un reducido número de personas que se reúnen para debatir qué datos son importantes en la investigación. Un moderador conduce las sesiones y encauza el debate hacia el área de interés. Las sesiones duran entre una y dos horas y por lo general participan en ellas entre 8 a 12 personas. Como el grupo de interés es pequeño, los resultados se utilizan como guía para la investigación más profunda.

Es importante elegir participantes que reúnan las mismas características que el grupo mayor sometido a estudio.

Por ejemplo, si se desea realizar una encuesta sobre la calidad de las prestaciones ofrecidas por una Obra Social para jubilados, sería apropiado incluir en el grupo de interés sólo participantes de 65 ó más años, así como elegir igual número de mujeres que de hombres, considerando diferentes áreas de la región que se quiere estudiar. Se trata



entonces de obtener un perfil más o menos representativo en el grupo de interés para que sus comentarios puedan tomarse como indicadores fiables de las opiniones del conjunto de la población estudiada.

Por lo general, es aconsejable tener al menos tres grupos de interés para un tema específico. Cada grupo de interés tiene su propio carácter dependiendo de sus miembros, y es después de dos o tres sesiones cuando surgen temas comunes. Un solo grupo de interés puede ir por un camino equivocado, en especial si algunas personas con opiniones muy firmes dominan la sesión.

Teléfono

La entrevista por teléfono es otra técnica habitual para recoger opiniones de un grupo.

Las ventajas de este método son: rapidez, bajo costo, es relativamente sencillo y proporciona una tasa de respuesta bastante alta.

Las desventajas son: que sólo pueden hacerse preguntas sencillas, la entrevista debe ser breve y algunas personas consideran que esas llamadas son una invasión a su vida privada y no quieren responder.

Cuestionarios por Correo

Se usan con frecuencia para recopilar datos cuando se cuenta con una lista o cuando los entrevistados se encuentran dispersos en un área muy grande. En ellos, se pueden incluir preguntas detalladas, ya que los encuestados dispondrán de tiempo para releerlas y pensar la respuesta. Por otro lado, si el cuestionario es demasiado largo, no se molestarán en contestarlo. El mayor problema que plantea es hacer que los destinatarios lo contesten y vuelvan a remitirlo. La tasa de respuesta es casi siempre muy baja y esto puede producir resultados erróneos.

De Puerta a Puerta

Se utilizan debido a que es relativamente fácil llevarlas a cabo y generan altas tasas de respuesta. Debe tenerse cuidado al seleccionar y adiestrar a los encuestadores para garantizar la comprensión de las preguntas. No debe manejarse un cuestionario muy largo. Con este método es posible cubrir un área geográfica grande y es fácil obtener una buena distribución de los ingresos percibidos por los entrevistados, ya que los ingresos de las personas normalmente se reflejan en el tipo de casas que habitan.

Abordaje en Centros Comerciales

Se usa con frecuencia cuando los investigadores de mercado están interesados en obtener opiniones de compradores. Los entrevistadores se instalan en áreas de mucho movimiento e invitan a las personas elegidas a contestar algunas preguntas. Se pueden usar ayudas visuales o de otro tipo, como pedir a los entrevistados que prueben un nuevo producto y comenten sobre él, o mostrarles un nuevo envase y pedir su opinión.



Registros

En ocasiones, los datos se recogen mediante el registro de nuevos productos.

Se pide a los consumidores que llenen un cuestionario cuando “registran” un producto que acaban de comprar al sellar la garantía. No todos devuelven el cuestionario, pudiendo introducirse un sesgo que puede orientar mal a la empresa que reúne los datos.

Observaciones

Los esfuerzos en este sentido incluyen el diseño de un experimento en el que se controlan las condiciones con precisión, de manera que se puedan observar y analizar los efectos al introducir cambios. Los experimentos son comunes en el campo de la ciencia, pero suelen utilizarse en algunas ocasiones en los negocios.

Entrevistas

La entrevista personal se usa cuando el entrevistador necesita determinar en forma profunda las opiniones y actitudes. Aunque este enfoque proporciona datos de calidad, el costo y el tiempo necesarios para programar y hacer las entrevistas limita su utilidad.

Todos los métodos anteriores de recogida de datos extraen opiniones o actitudes de las personas. Sin embargo, muchos problemas se refieren a mediciones de objetos, como cinescopios, maderos o ensambles soldados. La medición de objetos evita la interacción y comunicación con la gente, pero debe resolver aspectos como el tamaño de la muestra y otras consideraciones.

El método de reunir datos de poblaciones no humanas a menudo es sencillo y directo. Así, si deseamos estudiar si la longitud de determinados tornillos está dentro de ciertos límites, sólo necesitamos medirlos. En otros problemas, los procedimientos para observar unidades elementales de poblaciones no humanas son muy complicados y técnicos. Así, para descubrir si existen fallas en funciones de aluminio, se requieren rayos X, en este caso la elección del método está más allá de la capacidad del estadístico, por lo tanto lo hace el especialista.

Los datos sobre poblaciones humanas, según analizamos, pueden ser compilados haciendo observaciones directas ó formulando preguntas (entrevistas personales, cuestionarios enviados por correo, llamadas telefónicas). Luego se hacen registros apropiados de los resultados, que constituyen los datos originales de los que el estadístico prepara sus cuadros y gráficos para el análisis posterior.

Las ventajas y desventajas de los diferentes métodos varían según las circunstancias.

En general, un método es considerado el mejor, en un estudio determinado, si obtiene la información necesaria con la máxima precisión, con los costos más bajos y en el menor tiempo.

En cualquiera de los métodos el primer paso es planear el cuestionario. La mejor forma de asegurar que las preguntas sean formuladas apropiadamente es someterlas a



prueba previamente con una muestra pequeña (encuesta piloto) y hacer las mejoras necesarias basándose en las respuesta antes de ser usadas en gran escala.

Deben tenerse presente algunos principios generales de buena técnica de interrogación:

- 1- Las preguntas deben formularse sencilla y claramente.
- 2- pueden formularse preguntas que el interrogado pueda responder correctamente.

A menudo se obtienen resultados erróneos e información incorrecta cuando se formulan preguntas sobre prestigio individual (ingresos, posición social), cuestiones emocionales o inclinaciones o prejuicios personales. Cuando se necesita información sobre estos temas, las preguntas deben ser formuladas para reducir la respuesta emocional y la turbación. También es de utilidad, asegurar al interrogado que la información permanecerá completamente confidencial.

- 3- Las preguntas deben ser formuladas lógicamente y orientadas de modo que faciliten la tabulación de las respuestas.

Experimento

Con frecuencia, la investigación experimental genera conjuntos de datos de interés para la comunidad empresarial. Los experimentos difieren de otras técnicas de recogida de datos en términos del grado de control sobre la situación que se investiga. En un experimento se manipula una variable y se mide su efecto sobre otra, mientras que se controlan todas las demás variables.



Resumen

Ventajas Y Desventajas De Las Técnicas De Recogida De Datos

Técnica de Recogida de Datos	Ventajas	Desventajas
Grupo de Interés	<ul style="list-style-type: none">Buena técnica preliminar	<ul style="list-style-type: none">Muestra pequeñaNo se pueden proyectar resultados
Entrevistas por Teléfono	<ul style="list-style-type: none">Rápida.Poco costosa.Fácil de llevar a cabo.Alta tasa de respuesta.Flexibilidad para el entrevistador.	<ul style="list-style-type: none">Deben hacerse preguntas sencillas.La entrevista debe ser breve.
Cuestionarios por Correo	<ul style="list-style-type: none">Puede cubrir un área geográfica grandePoco costosaPreguntas estandarizadas	<ul style="list-style-type: none">Tasas bajas de respuestaSe emplea mucho tiempo
De puerta a puerta	<ul style="list-style-type: none">Puede cubrir un área geográfica grandeFácil de llevar a caboAlta tasa de respuesta	<ul style="list-style-type: none">Se emplea mucho tiempoCostosa
Abordaje en un centro comercial	<ul style="list-style-type: none">RápidaPoco costosaFácil de llevar a caboPueden usarse ayudas visualesFlexibilidad para el entrevistador	<ul style="list-style-type: none">No se pueden proyectar los resultadosLa entrevista debe ser breve
Entrevistas personales	<ul style="list-style-type: none">Pueden usarse ayudas visualesFlexibilidad para el entrevistadorLas preguntas se pueden analizar en profundidad	<ul style="list-style-type: none">CostosaSe emplea mucho tiempoSe obtienen muestras pequeñas



3.3.4. Relevamiento Estático y Dinámico

En un relevamiento estático, los datos son obtenidos a una fecha determinada, en un día fijo, obteniéndose ese día todos los datos necesarios para la investigación. Ejemplo: Censo poblacional.

En un relevamiento dinámico, los datos corresponden a aquellas operaciones que se realizan en forma continuada a través del tiempo y de forma sistemática, no en un solo día, sino en forma sucesiva. Ejemplo: Datos sobre nacimientos, defunciones, casamientos, etc.

Considerando los anteriores conceptos, podemos dividir en dos a los métodos de captación (relevamiento, recogida, levantamiento) de la información:

Métodos de Captación Completo o Exhaustivo

Censo o Relevamiento Estático

Consiste en estudiar a la Población a través de todos sus elementos en un momento determinado. Ejemplo: Censo Ganadero. Censo Poblacional.

Registro Exhaustivo o Relevamiento Dinámico

Consiste en estudiar a la Población a través de todos sus elementos en forma continua. Ejemplo: Registro Civil. Registro Prendario. Registro del Automotor.

Métodos de Captación Parcial

Consiste en tomar información sobre una parte representativa de la Población que constituye la muestra.

3.4. Clasificación, Tabulación Y Descripción De Los Resultados

Se refiere a la organización, presentación y descripción de los datos recopilados a los fines de facilitar la interpretación y el análisis de los mismos.

Cuando los datos con los que se cuentan son pocos, es posible presentar la información en forma de explicación escrita, llamada **presentación literal**.

Pero, cuando se trabaja con un gran número de datos, se hace necesaria la presentación en forma ordenada, lográndose tal objetivo al clasificarlos de manera sistemática y presentarlos en cuadros ó tablas, que llamaremos **Distribuciones de Frecuencias**, ó bien, construyendo gráficos y diagramas.

Es además, posible y fundamental, el cálculo de medidas descriptivas, tales como proporciones, promedios o desviaciones, que permiten a través de unos pocos valores, caracterizar o describir el comportamiento de grandes series de datos.



Las formas de organizar, presentar y describir la información, dependerá, según veremos, no sólo del tipo de dato con que estemos trabajando, sino del volumen de la misma.

Además, los datos pueden ser recolectados a través de un Censo, es decir, la medición fue realizada para la totalidad de elementos bajo estudio, o a través de un Muestreo, es decir, la medición fue realizada para una parte de la Población. En base a ello, diremos que toda medida basada en datos de muestra se llama *Estadística o Estadígrafo*, mientras que toda medida basada en datos de población, se llama *Parámetro*.

3.5. Generalización o Inferencia Final

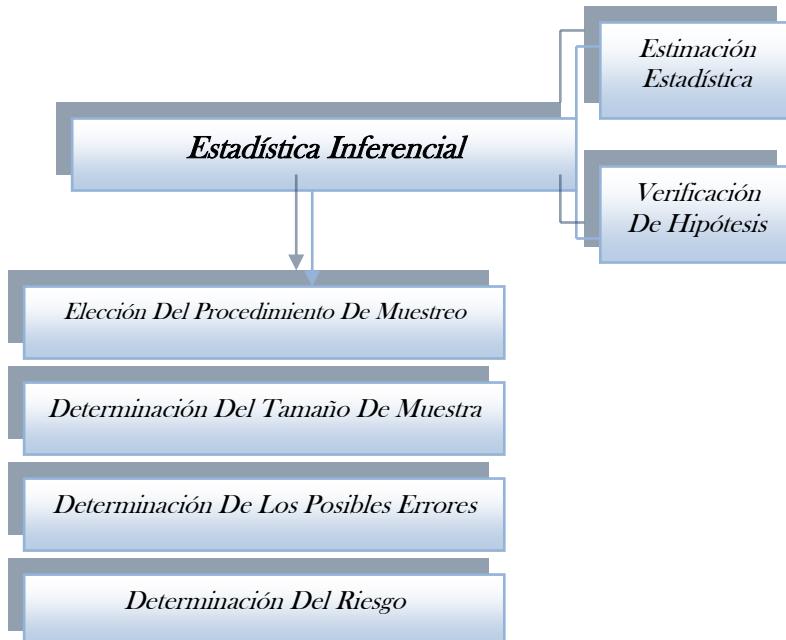
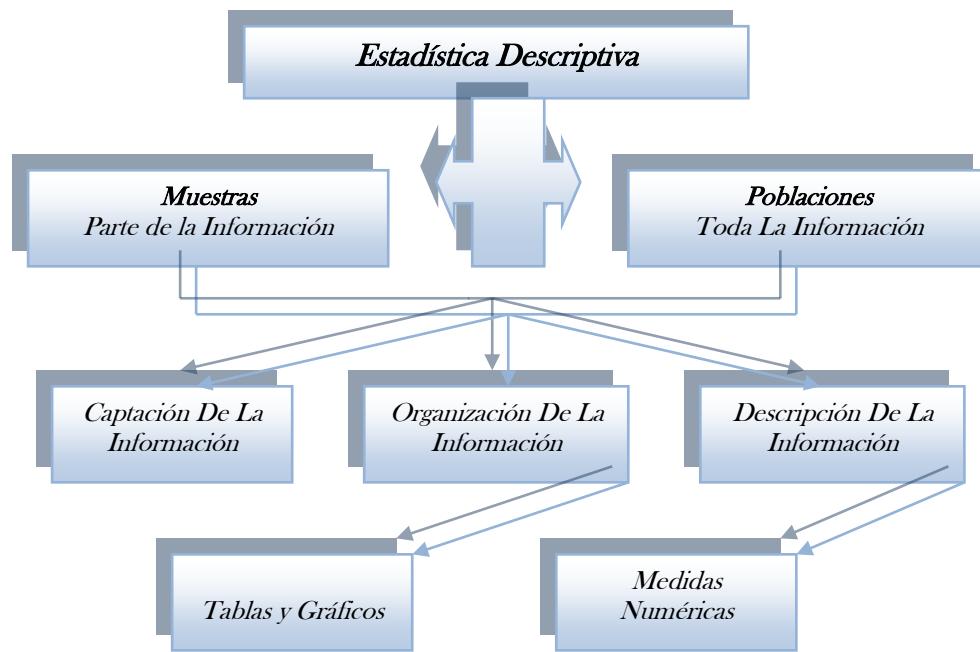
Si se ha trabajado con la totalidad de los elementos, censo ó enumeración completa, un estudio finaliza con el cálculo de medidas descriptivas. Entonces, puede describirse y revelarse las características de toda la población, pudiendo fácilmente arribar a conclusiones o tomar una decisión sobre el problema.

Si por el contrario, el estudio se realizó en base a una muestra, es necesaria una etapa adicional, consistente en tratar de responder, basándose en estadísticas de muestra, al problema o pregunta original formulada, que siempre se refiere a la población, en particular a su distribución y sus parámetros.

Así, el proceso de utilización de datos muestrales para inferir o generalizar sobre la población a la que pertenecen y de la cual fueron extraídos, requiere de los conocimientos y herramientas brindados por la Inferencia Estadística, a través de dos técnicas: la *Estimación Estadística* y la *Docimasia de Hipótesis*.



Resumen





CAPITULO N° 2: Organización y Presentación de Datos Estadísticos

Objetivos Específicos

Que el estudiante:

Reconozca la importancia de los procedimientos de resumen y presentación de datos

Conozca las formas de organizar y presentar datos estadísticos

Identifique y construya tablas estadísticas, según el tipo de dato

Identifique y construya gráficos, según el tipo de dato

Comprendra que las tablas y gráficos construidos aportan un resumen del comportamiento de la variable bajo estudio



Contenidos

1. Tablas estadísticas

- 1.1. Tipos de tablas estadísticas.
- 1.2. Partes principales de una tabla estadística.
- 1.3. Construcción de tablas estadísticas.

2. Formas de agrupar variables cuantitativas

- 2.1. Series simples o datos no agrupados.
- 2.2. Datos agrupados o distribuciones de frecuencias.
 - 2.2.1. Distribuciones de frecuencias en lista.
 - 2.2.1.1. Frecuencias absolutas.
 - 2.2.1.2. Frecuencias relativas.
 - 2.2.1.3. Frecuencias acumuladas.
- 2.2.2. Distribuciones de frecuencias en intervalos.
 - 2.2.2.1. Frecuencias absolutas.
 - 2.2.2.2. Frecuencias relativas.
 - 2.2.2.3. Frecuencias acumuladas.

3. Formas de agrupar variables cualitativas

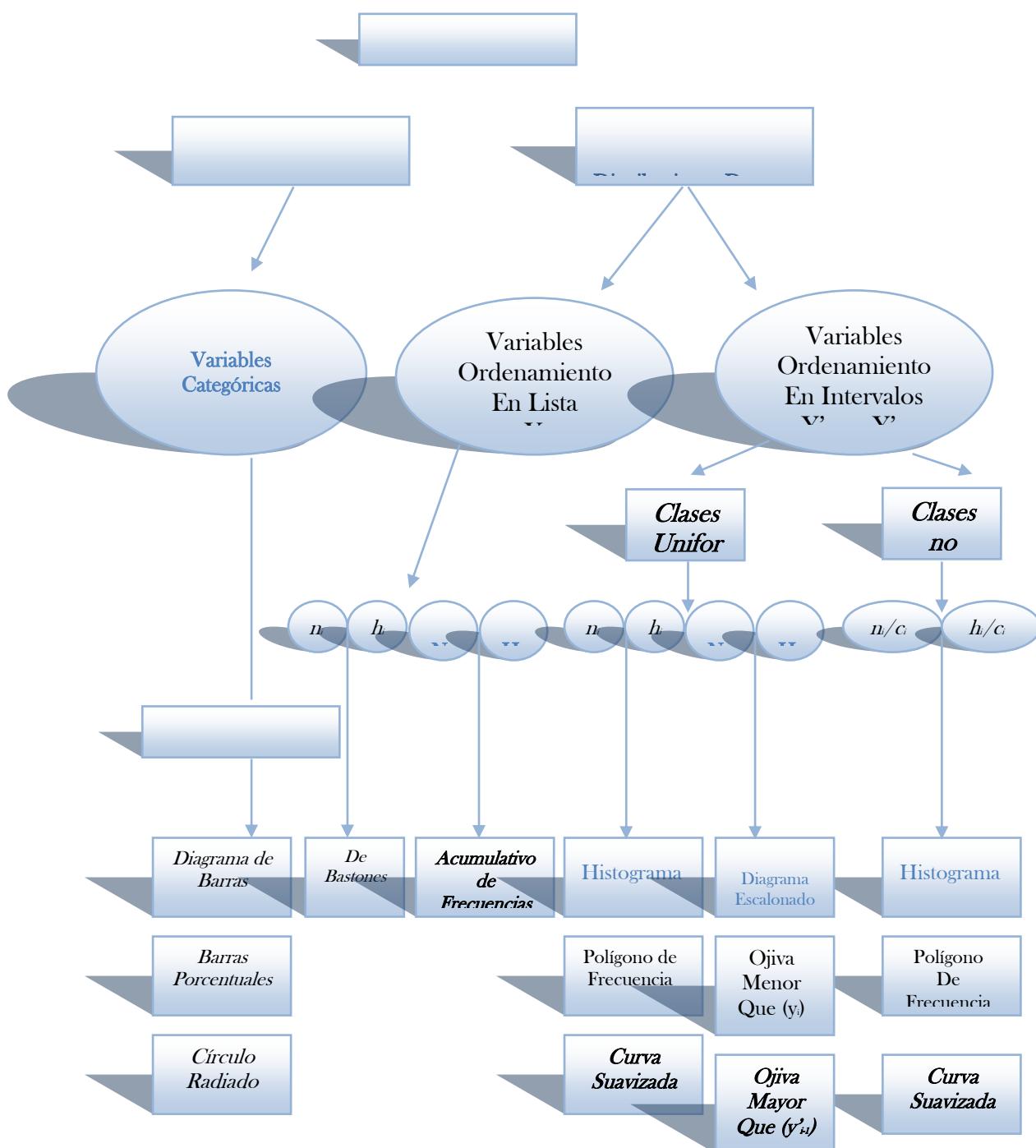
- 3.1. Distribuciones categóricas o tablas de contingencia.

4. Representaciones gráficas

- 4.1 Gráficos lineales.
 - 4.1.1. Gráfico de bastones.
 - 4.1.2. Gráfico acumulativo de frecuencias.
- 4.2. Gráficos de superficie.
 - 4.2.1. Histograma.
 - 4.2.2. Polígono de frecuencias.
 - 4.2.3. Curva suave.
 - 4.2.4. Diagrama escalonado.
 - 4.2.5. Ojivas.
 - 4.2.6. Curva acumulativa.
- 4.3. Gráficos especiales
 - 4.3.1 Variables categóricas.
 - 4.3.1.1. Diagrama de barras horizontales.
 - 4.3.1.2. Gráfica de barras de componentes de porcentajes.
 - 4.3.1.3. Diagrama de pastel o círculo radiado.
 - 4.3.2 Otros fenómenos.
 - 4.3.2.1. Gráfico de zonas.
 - 4.3.2.2. Diagrama de Pareto.
 - 4.3.2.3. Diagrama de tallos y hojas.



Resumen De Tablas Y Gráficos





1. TABLAS ESTADÍSTICAS

Como ya mencionamos, es necesaria la organización de los datos captados, sean éstos poblacionales o muestrales, a los fines de poder interpretarlos para sacar conclusiones o tomar de decisiones.

1.1. Tipos de Tablas Estadísticas

De acuerdo al propósito para las cuales fueran creadas, podemos clasificarlas en:

1- *Tablas para propósitos generales o de referencia:* no se construyen con un fin específico, sino que proporcionan información que puede ser usada como referencia o uso general.

2- *Tablas para propósitos especiales o de resumen, analíticas o de texto:* se confeccionan con el fin de proporcionar información para una exposición particular.

1.2. Partes principales de una Tabla Estadística

a) *Título:* es una descripción del contenido de la tabla. Debe ser preciso y completo y en general contar con los siguientes elementos:

- A qué corresponden los datos incluidos en la tabla
- Dónde está ubicada el área a la que corresponden los datos
- Cómo están clasificados los datos
- Cuando ocurrieron los hechos que dieron origen a la tabla

Cuando se presenta más de una tabla en el análisis, es necesaria su enumeración para que sea más fácil hacer referencia a la misma.

b) *Encabezamiento:* es la parte superior de las columnas, donde se colocan los títulos que indican los conceptos y los datos presentados en la tabla. En algunos casos existen encabezamientos y subencabezamientos.

c) *Conceptos o Columna Matriz:* son las descripciones que aparecen en las fileras de las tablas, en el extremo izquierdo de las mismas. En general, representan las clasificaciones de las cifras incluidas en el cuerpo de la tabla y la naturaleza de estas clasificaciones está incluida en el encabezamiento de esta columna. Cada concepto puede ser dividido en subconceptos si es necesario.

d) *Cuerpo:* está formado por el contenido de los datos estadísticos, los que están agrupados de acuerdo con las descripciones o clasificaciones de los encabezamientos y conceptos.

e) *Notas del encabezamiento:* se ubican sobre los encabezamientos y debajo del título y aclaran ciertos aspectos relacionados con la tabla que no han sido incluidos en el título, ni en los encabezamientos, ni en los conceptos. Por ejemplo, la unidad de medida en la que están expresados los datos, se acostumbra indicarla como una nota del encabezamiento.



f) *Notas al pie:* debajo del cuerpo de la tabla se ubican las notas al pie, que sirven para clarificar algunos aspectos incluidos y que no son explicados en otras partes de la tabla.

g) *Fuente:* debajo de las notas al pie, se establece la fuente que sirvió para obtener la información proporcionada por la tabla.

Ejemplo:

Identificación de los conceptos desarrollados:

Título	I. EVOLUCIÓN DEL COMERCIO EXTERIOR-EXPORTACIONES				
Notas del Encabezamiento	En millones de dólares				
	CONCEPTO	2000	2001	2010
	Productos Primarios	1436	1618	3677

Cuerpo	Sin clasificar	0	5	15
	Total de Exportaciones	2942	3916	8396
Notas al Pie	Nota: Las cifras pueden no coincidir con el total por redondeo				
Fuente	Fuente: Organismo gubernamental encargado de producir esta información				

1.3. Construcción de Tablas Estadísticas

Una vez que los datos recopilados han sido ordenados convenientemente, es necesario antes de la construcción efectiva de la tabla, y con el fin de que la misma cumpla cabalmente con su cometido, considerar una serie de elementos, entre ellos:

- Simplicidad en la presentación de la tabla
- Tratamiento de un tema por vez en cada tabla
- Ordenamiento adecuado de las clasificaciones. Pueden ser ordenados de acuerdo a cuatro bases:

1- *Cronológica:* Los datos clasificados por intervalos de tiempo se ordenan generalmente tomando en consideración el orden cronológico, ya sea iniciando la tabla con el período más antiguo ó el más reciente; no obstante, se acostumbra comenzar con el período más antiguo.

2- *Geográfica:* se acostumbra ordenarlas por orden alfabético o de acuerdo a la importancia de las áreas consideradas.

3- *Cualitativas:* son ordenadas teniendo en cuenta la importancia de las clases individuales, no obstante lo cual, si no interesa enfatizar tal circunstancia, se prefiere el orden alfabético.



4- *Cuantitativas:* Se ordenan, ya sea de menor a mayor o de mayor a menor, teniendo en cuenta el valor relativo de los datos.

- Facilitar al máximo las comparaciones
- Destacar cifras importantes
- Redondear cifras cuando no se requiere mayor detalle

En muchos casos, las tablas se diseñan sin necesidad de establecer cifras exactas, sobre todo si la información básica cuenta con muchos dígitos en sus datos. En tal caso, se utilizan cifras aproximadas y se procede a redondear las cantidades exactas de forma que los dígitos posteriores o las comas o puntos sean aproximaciones a unidades. Existen distintas reglas que se utilizan en el redondeo de cifras, pero la más sencilla y práctica es la siguiente:

- Si la porción de dígitos a ser eliminada comienza con 4 ó menos, se mantiene el dígito precedente a los que se eliminan, sin cambio. Ejemplo: 7.386.432.....7.386 millones.
- Si la porción de dígitos a eliminar comienza con 5 ó más, se agrega una unidad al dígito precedente. Ejemplo: 7.386.532.....7.387 millones.

Lo importante a tener en cuenta, es que las cifras sufran la menor variación posible a consecuencia del redondeo.

Frecuentemente, el total de cifras redondeadas en una tabla no concuerda exactamente con el total redondeado de las cifras originales. En este caso, se puede adoptar uno de los siguientes criterios:

- Agregar una aclaración en la Nota al Pie, en la que se especifique la razón por lo que no concuerdan, como por ejemplo: "La suma de las cifras detalladas no concuerda con los totales indicados debido a redondeo".
- Ajustar uno de los sumandos, de tal manera que el total de los números redondeados concuerde con el total redondeado, teniendo especial cuidado en que el número ó los números a ser ajustados provoquen el menor cambio.

2. FORMAS DE AGRUPAR VARIABLES CUANTITATIVAS

Según ya analizamos, la variable es una entidad que varía y de la que hay que distinguir entre *Valores posibles* y *valores realmente observados*:

Los Valores Posibles, Son Todos Aquellos Que La Variable Puede Asumir



Los Valores Realmente Observados, Son Todos Aquellos, Que De Entre Los Posibles, Han Sido Efectivamente Obtenidos En La Recopilación Efectuada.

Para exemplificar, utilizaremos las notas de un examen. Ellas pueden variar, en números enteros, del 0 al 10, entonces los valores posibles serán: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. Si se han tomado cinco exámenes y las notas fueron 5, 3, 8, 6, 9; éstos serán los valores realmente observados.

Las formas para agrupar datos que desarrollaremos seguidamente, son aplicables para cuando contamos con los valores observados de la variable, sea ésta Discreta o Continua.

La cantidad de valores observados, así como el comportamiento de los mismos, nos indicarán el tipo de agrupamiento que debemos realizar para resumir la información obtenida.

2.1. Series Simples o Datos No Agrupados

Se utilizan cuando la serie de datos con que contamos es pequeña, y en general, todos los valores son distintos, es decir, que cada uno de ellos se presenta una sola vez, con independencia del tipo de variable. En estos casos no se requiere agrupamiento de los datos analizados, es por ello que hablamos también de datos *No Agrupados*.

Cada valor de la serie representa una observación, la que designaremos, para el caso de una muestra, por x_1, x_2, \dots, x_n conforme al orden en que se presenten.

En general $x_i, i = 1, 2, \dots, n$, representa la i -ésima observación.

Entonces:

x_1	Primera medición de la Variable
x_2	Segunda medición de la Variable
x_3	Tercera medición de la Variable
.....
x_n	n -ésima medición de la variable

Es decir, que a través de x , representamos la característica bajo estudio o variable, y con el subíndice, el orden en que esa característica o valor ha sido observado.

Este conjunto de observaciones constituye una muestra de tamaño n procedente de la población que contiene el total de elementos bajo estudio.

Cabe aclarar que si realizáramos un censo las notaciones serían las mismas, solo que usaríamos letras mayúsculas, es decir X para la variable y N para el tamaño de la población.



Ejemplo:

Variable: Número de Hijos por familia

Muestra: $n = 10$ familias

Valores Observados:

$$x_1 = 2; x_2 = 1; x_3 = 3; x_4 = 1; x_5 = 2; x_6 = 1; x_7 = 3; x_8 = 0; x_9 = 2; x_{10} = 1$$

Podemos observar a simple vista, que la mayoría de las familias tienen 1 hijo (4), luego 2 hijos (3), luego 3 hijos (2) y ningún hijo (1).

2.2. Datos Agrupados o Distribuciones de Frecuencias

Cuando el número de datos es muy grande, el simple análisis de ellos no nos permite obtener ninguna conclusión de interés. Es por ello, que prácticamente todo el análisis se practica a través de tablas que se conocen como *Distribuciones de Frecuencias*, cuyo objetivo es condensar y simplificar la información sin perder muchos detalles. Es decir, se hace necesario realizar un agrupamiento de los datos analizados, razón por la cual hablamos de *Datos Agrupados*.

Las Distribuciones de Frecuencias, pueden ser:

- *Unidimensionales*: analizan una variable por vez. Por ejemplo, peso o altura o edad.
- *Bidimensionales*: analizan dos variables por vez, por ejemplo, relación entre peso y altura, peso y edad, altura y edad.
- *Multidimensionales*: analizan más de dos variables por vez, como por ejemplo, relación entre peso, altura y edad.

En este curso trabajaremos solamente, con *Distribuciones Unidimensionales*.

Este agrupamiento puede realizarse de dos maneras:

En Forma De Lista
En Forma De Intervalos

La elección entre ellas, como ya dijimos, dependerá fundamentalmente de la cantidad de valores observados y de su comportamiento.

2.2.1. Distribuciones De Frecuencias En Lista

Consiste en construir una tabla de dos columnas. En la primera de ellas, se listan los diferentes valores que asumió la variable, en las observaciones realizadas, ordenados por su magnitud, en orden descendente o ascendente, con independencia de la naturaleza de la variable. En la segunda columna se indica el número de veces que cada valor distinto se ha repetido, o bien la proporción que cada cantidad de valores distintos representa en el total de las observaciones. Estos números constituyen desde dos ópticas, una en valores absolutos, y la otra en valores relativos, la frecuencia de presentación, según veremos seguidamente.



Simbolizaremos con y_i $i = 0, 1, 2, \dots, m$, a los distintos valores que se hayan presentado en las observaciones realizadas, donde m representa la cantidad de estos valores distintos. Nótese que m será siempre menor que n , número total de observaciones.

2.2.1.1. Frecuencias Absolutas

Cuando la segunda columna de la tabla construida, muestre el número de veces que se han repetido los distintos valores de la variable, la llamaremos Frecuencia Absoluta y la simbolizaremos por n_i $i = 0, 1, 2, \dots, m$.

Necesariamente n_i , será menor ó a lo sumo igual que n , ($n_i \leq n$), y mayor ó igual a 0 ($n_i \geq 0$). Además, la suma de todas las frecuencias absolutas será igual al tamaño de la muestra (o población), es decir al total de las observaciones. Por lo tanto para el caso de una muestra: $\sum_{i=1}^m n_i = n$.

Simbólicamente, la tabla queda construida de la siguiente forma:

y_i	n_i
y_1	n_1
y_2	n_2
...	...
y_m	n_m
	$\sum_{i=1}^m n_i = n$

Entonces, debe verificarse:

$$\begin{array}{l} n_i \leq n \\ n_i \geq 0 \\ \sum_{i=1}^m n_i = n \end{array}$$

2.2.1.2. Frecuencias Relativas

Cuando la segunda columna de la tabla construida, muestre la proporción de los distintos valores de la variable, la llamaremos Frecuencia relativa y la simbolizaremos por h_i $i = 0, 1, 2, \dots, m$.

Su cálculo se reduce a obtener el cociente entre el valor de cada frecuencia absoluta (n_i) por el total de observaciones (n).

Entonces

$$h_i = \frac{n_i}{n}$$

El resultado es una fracción que no puede ser mayor que 1 ($h_i \leq 1$), ni menor que 0 ($h_i \geq 0$), y la suma de todas las frecuencias relativas será siempre igual a 1.

Simbólicamente, la tabla queda construida de la siguiente forma:



y_i	h_i
y_1	h_1
y_2	h_2
...	...
y_m	h_m
	$\sum_{i=1}^m h_i = 1$

Entonces, debe verificarse:

$$h_i \leq 1$$

$$h_i \geq 0$$

$$\sum_{i=1}^m h_i = 1$$

2.2.1.3. Frecuencias Acumuladas

Tanto las frecuencias absolutas como las relativas que hemos definido, pueden acumularse, es decir, que se pueden ir sumando los distintos valores hasta el correspondiente subíndice de la variable que nos interese. De esta manera podemos distinguir entre:

Frecuencias Absolutas Acumuladas

La simbolizaremos por N_i , $i = 0, 1, 2, \dots, m$, valor que corresponde a y_i , siendo por definición:

$$N_i = n_1 + n_2 + \dots + n_i$$

Entonces:

La primera frecuencia acumulada será

$$N_1 = n_1$$

La segunda

$$N_2 = n_1 + n_2$$

La tercera

$$N_3 = n_1 + n_2 + n_3$$

...

...

Y así sucesivamente hasta la última

$$N_m = n_1 + n_2 + n_3 + \dots + n_m = \sum_{i=1}^m n_i = n$$

Frecuencias Relativas Acumuladas

La simbolizaremos por H_i para $i = 0, 1, 2, \dots, m$ valor que corresponde a y_i , siendo por definición:

$$H_i = h_1 + h_2 + \dots + h_i$$

Entonces:

La primera frecuencia acumulada será

$$H_1 = h_1$$

La segunda

$$H_2 = h_1 + h_2$$

La tercera

$$H_3 = h_1 + h_2 + h_3$$

...

...

Y así sucesivamente hasta la última

$$H_m = h_1 + h_2 + h_3 + \dots + h_m = \sum_{i=1}^m h_i = 1$$



Entonces:

y_i	N_i	H_i
y_1	$N_1 = n_1$	$H_1 = h_1$
y_2	$N_2 = n_1 + n_2$	$H_2 = h_1 + h_2$
...
y_m	$N_m = n_1 + n_2 + n_3 + \dots + n_m = \sum_{i=1}^m n_i = n$	$H_m = h_1 + h_2 + h_3 + \dots + h_m = \sum_{i=1}^m h_i = 1$

Ejemplo: Sea la siguiente distribución del número de hijos en $n = 10$ familias observadas, donde y_i representa el número de hijos y n_i la cantidad de familias.

y_i	n_i	h_i	$h_i \times 100$	N_i	H_i
0	1	1/10=0,10	10%	1	0,10
1	4	4/10=0,40	40%	5	0,50
2	3	3/10=0,30	30%	8	0,80
3	2	2/10=0,20	20%	10	1
	10	1	100		

Para ciertos fines, es conveniente disponer de los datos acumulados. Puede que deseemos contestar: ¿Cuántas personas ganan \$ 1.000 o menos al mes? ¿Cuántos vendedores venden una cantidad dada ó más por semana?

Pueden formarse frecuencias acumuladas sobre la base de “menos que” o “más que” y pueden ser absolutas o relativas.

Se ejemplificarán estas acotaciones cuando se desarrolle gráficos estadísticos.

2.2.2. Distribuciones De Frecuencias En Intervalos

Si el número de observaciones es grande y el número de valores distintos que asumió la variable es casi tan grande como el total de observaciones, el recorrido de las variables ya no puede ser descripto por unas pocas categorías distintas de medidas, como lo hicimos en el caso anterior, sea la variable de naturaleza discreta o continua.

Entonces como en el caso anterior, construiremos una tabla de dos columnas, pero en la primera de ellas, los valores de la variable se presentan a través de intervalos o clases. En la segunda columna se indica la cantidad de valores contenidos por intervalos, o bien la proporción que esa cantidad de valores representa en el total de las observaciones. Estos números, al igual que antes, constituyen desde dos ópticas, una en valores absolutos, y otra en valores relativos, las frecuencias de presentación.

Para la construcción de estas tablas, definiremos los siguientes conceptos:

Intervalos De Clases o Clases: Los valores de la variable aparecen separados en intervalos de clases, es decir que para condensar y simplificar los datos, se utilizan las tablas de distribuciones de frecuencias, en las cuales se agrupan convenientemente los valores de la variable, en intervalos.



Intervalos de clases, que simbolizaremos por $y_{i-1}^{\cdot} - y_i^{\cdot}$, donde y_{i-1}^{\cdot} es el extremo inferior o límite inferior del intervalo, es decir los números que aparecen a la izquierda y y_i^{\cdot} , es el extremo superior o límite superior, es decir los números que aparecen a la derecha.

Amplitud Del Intervalo: Está dada por la diferencia entre los valores extremos del intervalo.

Lo simbolizaremos y calcularemos por: $c_i = y_{i-1}^{\cdot} - y_i^{\cdot}$.

Su tamaño depende del problema analizado y también del interés del investigador en condensar más o menos esos datos.

Cuanto menor sea la amplitud del intervalo, mayor será la cantidad de clases que tendremos, y viceversa.

A medida que los datos se agrupen más, se va perdiendo no sólo información, sino precisión en los resultados.

Marcas De Clase: Son los puntos medios del intervalo.

Las simbolizaremos y calcularemos por: $y_i^{\cdot} = \frac{y_{i-1}^{\cdot} + y_i^{\cdot}}{2}$

Es decir, es la media aritmética de los valores extremos del intervalo.

Obsérvese que los distintos subíndices de las marcas de clases, coincidirán con los del extremo superior del intervalo de clase respectivo. Por trabajar con y_i^{\cdot} , se dice que perdemos precisión, pues este valor, representa a los datos originales y los valores de frecuencia absoluta no corresponden exactamente a y_i^{\cdot} , sino a los valores originales comprendidos entre los extremos en los cuales se encuentra y_i^{\cdot} .

Recorrido De La Serie: Es la diferencia entre el valor superior ($y_{máx}$) y el valor inferior ($y_{mín}$) de la serie. Es decir:

$$R = y_{máx} - y_{mín}$$

Al construir los intervalos de clase se amplía el recorrido para evitar que los límites inferior y superior coincidan con los valores extremos que adopta la variable y por convención se establece: "Si un valor de la variable cae en el extremo superior del intervalo, corresponde al intervalo siguiente". Generalmente los intervalos tienen igual amplitud y su número oscila entre 5 y 12.

El número de intervalos se fija arbitrariamente y la amplitud de cada uno de ellos (c), será igual a:

$$c_i = \frac{R}{N^{\circ} \text{ de Intervalos}} = \frac{y_{máx} - y_{mín}}{m}$$

Si el resultado es un número decimal se amplía el intervalo hasta lograr que su amplitud sea una cifra entera y cómoda de utilizar.

La cantidad que se agrega a R para llegar a R' (recorrido ampliado), se divide en partes iguales al comienzo y al final de la distribución.



Presentaremos a continuación las correspondientes frecuencias, mencionando las interpretaciones que le daremos para este tipo de presentación:

2.2.2.1. Frecuencias Absolutas

Indican ahora, la cantidad de valores comprendidos en cada intervalo. Se simbolizan y responden a iguales condiciones que las vistas en Frecuencias Absolutas para Distribuciones de Frecuencias en Lista.

2.2.2.2. Frecuencias Relativas

Indican ahora, la proporción de valores comprendidos en cada intervalo. Se simbolizan y responden a iguales condiciones que las vistas Frecuencias Relativas para Distribuciones de Frecuencias en Lista.

2.2.2.3. Frecuencias Acumuladas

Frecuencias Absolutas Acumuladas

Indican ahora, la cantidad de valores menores desde el límite superior de cada intervalo.

Frecuencias Relativas Acumuladas

Indican ahora, la proporción de valores menores desde el límite superior de cada intervalo.

Se simbolizan y responden a iguales condiciones que las vistas en Frecuencias tanto absolutas como relativas acumuladas vistas en la distribución de Frecuencias en Lista.

Pueden también acumularse considerando al límite inferior de cada clase indicando en este caso, la cantidad o proporción, según corresponda de valores mayores o iguales a partir del límite considerado.

Presentaremos una tabla conteniendo los conceptos analizados

Intervalos De Clase	Marcas De Clase	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Absoluta Acumulada	Frecuencia Relativa Acumulada
$y_{i-1}^+ - y_i^+$	y_i	n_i	h_i	N_i	H_i
$y_0^+ - y_1^+$	y_1	n_1	h_1	N_1	H_1
$y_1^+ - y_2^+$	y_2	n_2	h_2	N_2	H_2
$y_2^+ - y_3^+$	y_3	n_3	h_3	N_3	H_3
...
$y_{m-1}^+ - y_m^+$	y_m	n_m	h_m	$N_m = N$	$H_m = 1$
		$\sum_{i=1}^m n_i = n$	$\sum_{i=1}^m h_i = 1$		

Ejemplo:

Supongamos que tenemos 20 datos para analizar, correspondientes a las edades de un grupo de personas:

$$y_1 = 74; y_2 = 47; y_3 = 67; y_4 = 82; y_5 = 67; y_6 = 65; y_7 = 70; y_8 = 55; y_9 = 57; y_{10} = 85; y_{11} = 69; \\ y_{12} = 58; y_{13} = 71; y_{14} = 61; y_{15} = 52; y_{16} = 76; y_{17} = 79; y_{18} = 77; y_{19} = 88; y_{20} = 94.$$

Contamos con 20 observaciones para la variable, entonces $n = 20$

Convengamos en construir una tabla con 5 intervalos, entonces $m = 5$

Procedimiento:

1- Ordenamos los datos de la serie en forma ascendente o descendente, a los fines de detectar el mínimo (y_{\min}) y el máximo (y_{\max}) valor: 47, 52, 55, 57, 58, 61, 65, 67, 67, 69, 70, 71, 74, 76, 77, 79, 82, 85, 88, 94.

2- Obtenemos el Recorrido de la serie: $R = y_{\max} - y_{\min} = 94 - 47 = 47$

3- Determinamos la amplitud que deberá tener cada intervalo en función del Recorrido (R) y del número de intervalos definidos (m):

4-

$$c_i = \frac{R}{N^{\circ} \text{ de Intervalos}} = \frac{y_{\max} - y_{\min}}{m} = \frac{47}{5} = 9,4$$

Hemos encontrado una amplitud de intervalo que resulta inadecuada. Por lo general, se trata de que sean números enteros y sobre todo divisibles por 2, para que al dividir la amplitud del intervalo por 2, nos dé una marca de clase o punto medio no fraccionario.

Entonces, tomaremos como amplitud a 10, lo cual modifica el recorrido, pasando a ser un recorrido ampliado igual a: $5 \times 10 = 50$ ($m \times c_i = R'$). Para llegar a este recorrido de 50, debemos tomar en la serie, en lugar de 47, 45 (47-2), como punto mínimo y en lugar de 94, 95 (94+1), como punto máximo. Es decir, que hemos ampliado la serie en 3, 1 por arriba y 2 por abajo, que corresponde al incremento del nuevo recorrido (95-45).

Tenemos ya determinados los elementos necesarios para construir la tabla, que en resumen son: $y_{\min} = 45$; $y_{\max} = 95$; $m = 5$ y $c_i = 10$.

(1) $y_{i-1} - y_i$	(2) y_i	n	h	N	H
45-55	50	2	$2/20=0,10$	2	0,10
55-65	60	4	$4/20=0,20$	6	0,30
65-75	70	7	$7/20=0,35$	13	0,65
75-85	80	4	$4/20=0,20$	17	0,85
85-95	90	3	$3/20=0,15$	20	1
		20	1		



Los intervalos se construyen de la siguiente forma: El límite superior del primer intervalo coincide con el valor inferior de la serie ($y_{mín}$), que en este caso es de 45. A este valor se le suma la amplitud del intervalo y se obtiene el límite superior del primer intervalo ($45+10=55$).

Procedemos de igual forma para el resto de los intervalos, hasta llegar al último, en donde el límite superior debe coincidir con el valor máximo de la serie (95).

Es decir, a cada límite inferior le sumaremos la amplitud para obtener el límite superior. (Columna 1).

Una vez determinados los intervalos de clase, obtenemos para cada uno de ellos la marca de clase, Columna (2), haciendo:

$$\text{Primer Intervalo} \quad y_1 = \frac{45 + 55}{2} = 50$$

$$\text{Segundo Intervalo} \quad y_2 = \frac{55 + 65}{2} = 60$$

Y así sucesivamente.

Para distribuir una masa de datos brutos entre las clases que ya han sido establecidas, a los fines de conocer la frecuencia absoluta, podemos usar:

Hoja de Cuenta

Supone establecer clases y representar cada unidad que corresponde a cada clase por una raya diagonal, entonces contamos el número de unidades de cada clase:

<i>Clase</i>		
45-55	//	2
55-65	////	4
65-75	-//////	7
75-85	////	4
85-95	///	3
		20

Es decir, el primer dato es 47, el cual está incluido en el primer intervalo, el segundo dato es 52, también incluido en el primer intervalo, lo que hemos dejado representado por ambas barras, y así sucesivamente se van ubicando el resto de las observaciones. El total debe coincidir con el número de datos.



Forma de Asiento

Las clases se disponen horizontalmente en la parte superior por orden ascendente de izquierda a derecha. Las unidades reales son anotadas en las clases correspondientes. Las respectivas unidades son anotadas en la parte inferior de dicha clase. Las cifras totales de las clases constituyen las frecuencias de clases.

45-55	55-65	65-75	75-85	85-95
47	55	65	76	85
52	57	67	77	88
	58	67	79	94
	61	69	82	
		70		
		71		
		74		
2	4	7	4	3

Es más laboriosa que la anterior, pero ofrece ciertas ventajas:

- Pueden hallarse registros en las columnas inapropiadas con sólo examinarlas.
- Pueden hacerse nuevas clasificaciones si las clases originales son insatisfactorias.
- Puede saberse cuán estrechamente concuerda el valor medio con el promedio de las unidades de dicha clase.

Vamos a considerar ahora, nuevos ejemplos, mostrados en las siguientes tablas, a partir de los cuales realizaremos algunas consideraciones, introduciendo otros conceptos importantes:

$y_{i-1} - y_i$	y_i	n	h	N	H
0 - 4	2	34	0,829	34	0,829
4 - 8	6	4	0,098	38	0,927
8 - 12	10	2	0,049	40	0,976
12 - 16	14	0	0	40	0,976
16 - 20	18	0	0	40	0,976
20 - 24	22	0	0	40	0,976
24 - 28	26	1	0,024	41	1
28 - 32	30	0	0	41	1
		41	1		

En este caso, vemos a simple vista que esta distribución realizada no es la más conveniente, pues la mayoría de los valores están concentrados en el primer intervalo (0-4), representados por $y_i = 2$.

Ejemplo:

Suponemos la distribución de los Salarios para 2.720 trabajadores.



Clase $y_{i-1} - y_i$	Amplitud c	n	Densidad de Frecuencia (Por cada mil) n/c
Menos de 5000	5000	100	100/5,0=20
5000-6000	1000	150	150/1,0=150
6000-7000	1000	200	200/1,0=200
7000-8000	1000	250	250/1,0=250
8000-9500	1500	200	200/1,5=133,33
9500-11000	1500	250	250/1,5=166,67
11000-12500	1500	500	500/1,5=333,33
12500-14500	2000	350	350/2=175
14500-16500	2000	400	400/2=200
16500-19500	3000	200	200/3=66,67
19500-22500	3000	100	100/3=33,33
22500 o más		20	0
		2720	

En situaciones donde hay unos pocos valores extraordinariamente pequeños o extraordinariamente grandes; o todos juntos, en los que los datos poseen grandes vacíos, y cuando el número de observaciones llega a ser de millares y aún de millones, con ventajas pueden utilizarse clases de extremo abierto e intervalos no uniformes.

En estos casos deben considerarse algunos otros conceptos:

- A veces puede formarse una distribución sin el límite inferior para la primera clase o sin el límite superior para la última clase, o sin ambos límites. Se dice entonces, que son *Clases De Extremo Abierto*. La amplitud del intervalo para una clase de extremo abierto es el infinito y su punto medio es $\pm \infty$.
- No es necesario que los intervalos tengan la misma amplitud, es decir sean *Uniformes*.

Pueden usarse intervalos con distinta amplitud, llamados *No Uniformes*.

Cuando se usan intervalos no uniformes, debemos calcular lo que se conoce como *Densidad De Frecuencia*, estimando cuáles serían las frecuencias de clase si se usaran intervalos de clase uniformes.

Las densidades de frecuencia se han calculado en el supuesto de un intervalo de clase uniforme de \$1000. La densidad de frecuencia para la primera clase es 20, porque no es realmente de extremo abierto, ya que se trata de salarios, y su límite inferior es 0.

Entonces, como la amplitud es de 5000, o sea 5 veces 1000, la densidad es $100/5=20$.

La densidad de frecuencia para la última clase es 0, porque hemos supuesto que el límite superior es ∞ .

Cuando se emplean clases de extremo abierto, es recomendable dar los valores mínimo o máximo, o ambos, y el valor o los valores en tal clase, en una nota al pie.

Otra solución es asegurarse que todas las clases que tienen por lo menos una observación, sean cerradas, es decir, asegurarse de que todas las clases de extremo abierto estén vacías de observaciones, entonces, el supuesto del punto medio puede aplicarse a



todos los datos, y los problemas causados por las clases de extremo abierto podrán evitarse.

Podemos utilizar la densidad de frecuencia, expresada por unidad de intervalo.

En ese caso la definimos como el número de casos por unidad de tamaño de clase. Es un promedio en cada clase de cuántas unidades hay por unidad de ancho del intervalo. Se obtiene como la frecuencia de clase dividido por el ancho del verdadero intervalo.

Si la frecuencia considerada es absoluta (n), entonces n/c define la densidad de frecuencia absoluta. Si en cambio, la frecuencia considerada es relativa (h), entonces h/c define la densidad de frecuencia relativa.

Ejemplo:

$y_{i-1} - y_i$	n	c	n/c
50-60	8	10	$8/10=0,8$
60-70	10	10	$10/10=1$
70-90	8	20	$8/20=0,4$
90-150	3	60	$3/60=0,05$

En resumen

<i>Densidad de Frecuencia</i>	<i>Simbología y Cálculo</i>	<i>Interpretación</i>
Absoluta	$dfa = n/c$	Indica el promedio en cada clase de cuántas unidades hay por unidad de ancho del intervalo.
Relativa	$dfr = h/c$	Indica el promedio en cada clase de la proporción de unidades por unidad de ancho del intervalo.

En ambos casos, se ha considerado cada unidad de ancho del intervalo, pudiéndose calcular en el supuesto de intervalos de clase uniformes de 10, 100, 1000, etc., según convenga.

3. FORMAS DE AGRUPAR VARIABLES CUALITATIVAS

3.1. Distribuciones Categóricas o Tablas de Contingencia

Hasta acá hemos trabajado con variables cuantitativas. Veremos ahora, el caso de variables cualitativas o categóricas, donde hablaremos de Distribuciones Categóricas o Tablas de Contingencia.

Una Distribución de Frecuencias Categóricas muestra el número, o la proporción de observaciones que corresponde a cada una de las clases cualitativas, mutuamente excluyentes, que hayamos determinado.



El observador puede limitarse a anotar la presencia o ausencia de un cierto atributo en una serie de objetos o individuos y contar el número de los que lo poseen y el de los que no lo poseen.

Ejemplo:

Argentinos - No Argentinos

Ciegos - Videntes

Estatura, considerando altos a los que exceden cierta estatura y bajos a todos los demás (Transformación de series cuantitativas y series cualitativas).

Es decir, se forman dos clases distintas, siendo éste el caso más simple, pero si se consideran varios atributos, el proceso de clasificación puede extenderse indefinidamente.

La clasificación en la que cada clase se subdivide en dos subclases, se denomina dicotómica, caso contrario, se habla de clasificación múltiple.

El hecho de la clasificación, no implica necesariamente, la existencia de una línea divisoria natural o claramente definida entre ambas clases. Puede ser totalmente arbitraria. La clasificación puede ser vaga o imprecisa. Ejemplo: Demencia y Cordura.

Puede haber discrepancias sobre la clase en que determinado individuo debe incluirse. Por ejemplo, al estudiar las distintas clases de alimentos vendidos por un comercio, podríamos adoptar clases tales como: "carnes y productos cárnicos", "alimentos congelados", etc.; tendríamos dificultad entonces para decidir a qué clase corresponde un producto tal como "carne congelada".

Así, debemos tener cuidado al diseñar un conjunto de clases que deben ser *mutuamente excluyentes*, es decir, no pueden presentarse juntas, pues no tienen elementos en común, y *colectivamente exhaustivas*, pues consideran todo el espectro de posibilidades.

Ejemplo:

Un periódico muy conocido efectuó una encuesta telefónica para estudiar las opiniones de los ciudadanos de Córdoba, en relación a varias problemáticas. Se seleccionó un total de 419 personas, a través de una muestra aleatoria simple. Los siguientes datos reflejan las respuestas a una de las preguntas realizadas sobre lo adecuado de la protección brindada por la policía y los bomberos: "¿Es adecuada la protección brindada por la policía y los bomberos?

Respuestas	n
Sí	80
No	293
No sabe / No contesta	46
Totales	419

A menudo, es necesario examinar variables en forma simultánea



Por ejemplo, en un mazo de cartas francesas, con un total de 52 naipes, elaboramos para las variables “color” y “as”, la siguiente tabla:

	<i>Rojo</i>	<i>Negro</i>	<i>Totales</i>
<i>As</i>	2	2	4
<i>No As</i>	24	24	48
<i>Totales</i>	26	26	52

Estas tablas, de dos sentidos, de clasificación cruzada, se conocen como *Tablas de Contingencia*.

4. REPRESENTACIONES GRÁFICAS

Gráficos Lineales, para agrupamientos en lista

- Gráfico de Bastones (Frecuencias absolutas o relativas)
- Gráfico Acumulativo de Frecuencias (Frecuencias acumuladas, sean absolutas o relativas)

Gráficos de Superficie, para agrupamiento en intervalos

- | | | |
|---|---|---|
| <ul style="list-style-type: none">- Histograma- Polígono de Frecuencias- Curva Suave | } | <p><i>Frecuencias absolutas</i>
<i>Frecuencias relativas</i>
<i>Densidades de Frecuencias</i></p> |
| <ul style="list-style-type: none">- Diagrama Escalonado- Ojiva menor que y Ojiva mayor que- Curva Acumulativa Suavizada | } | <p><i>Frecuencias Acumuladas</i>
<i>Absolutas</i>
<i>Relativas</i></p> |

Gráficos Especiales

Para variables categóricas

- Diagrama de Barras Horizontales
- Gráfica de Barras de Componentes de Porcentajes o Barra Porcentual
- Diagrama de Pastel o Círculo Radiado

Para otros fenómenos

- Gráfico de Zonas
- Diagrama de Pareto
- Diagrama de Tallo y Hoja



4.1. Gráficos Lineales

4.1.1. Gráfico de Bastones

Utilizamos esta gráfica cuando tenemos unos pocos valores distintos correspondientes a un agrupamiento en lista.

Consisten en líneas o segmentos de líneas. Para su construcción trabajaremos con el cuadrante positivo de un Sistema de Coordenadas Cartesianas Ortogonales, cuyos ejes “ x ” e “ y ” deberán ser rotulados claramente. Convenimos que el eje “ x ” representará los valores de la variable y el eje “ y ” los valores correspondientes a las frecuencias, ya sean absolutas o relativas, según se elija una u otra.

Una vez determinadas las escalas de ambas coordenadas, se marcan los puntos en el plano, correspondientes a cada uno de los pares de valores (variable y su respectiva frecuencia) que aparecen en la tabla. Luego, se eleva una perpendicular al eje de las abscisas hasta intersectar el punto marcado en el plano y que indica la observación. La altura de estas ordenadas o bastones, nos indican las frecuencias absolutas o relativas, ya que es indistinto el valor que tomemos en la ordenada, si las escalas son proporcionales a los valores que representan.

Ejemplo:

Trabajaremos con el ejemplo dado en Distribuciones de Frecuencias en lista, o sea el Número de Hijos en 10 familias observadas.

y_i	n_i	h_i
0	1	$1/10=0,10$
1	4	$4/10=0,40$
2	3	$3/10=0,30$
3	2	$2/10=0,20$
	10	1

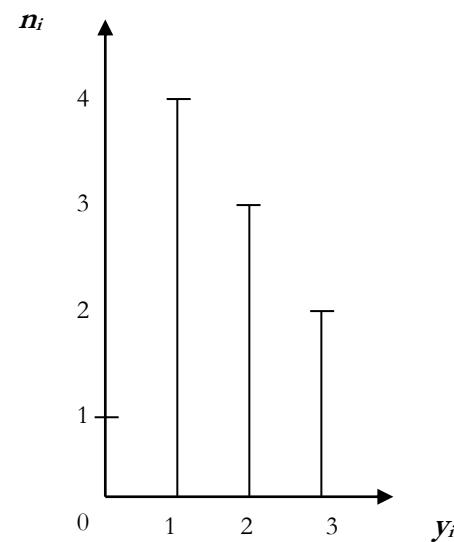


Gráfico de Bastones

Podríamos haber utilizado h_i , rotulando convenientemente.



4.1.2. Gráfico Acumulativo De Frecuencias

Si se consideran las frecuencias absolutas o relativas acumuladas, la representación gráfica, se efectúa mediante el Gráfico Acumulativo de Frecuencias.

Al igual que para el caso anterior, utilizaremos el cuadrante positivo de un Sistema de Coordenadas, cuyos ejes deberán ser rotulados claramente.

También convendremos que en el eje “ x ” representará los valores de la variable y el eje “ y ” mostrará ahora, la frecuencia acumulada absoluta o relativa. Luego de marcar sobre el eje horizontal los valores de la variable, se levanta en cada uno de estos puntos una vertical de longitud igual a la frecuencia acumulada respectiva, completando con tramos horizontales la parte de los intervalos en que no se presentan observaciones, quedando indicada una línea poligonal escalera.

Ejemplo:

Continuamos con el ejercicio planteado en el caso anterior.

y_i	N_i	H_i
0	1	0,10
1	5	0,50
2	8	0,80
3	10	1

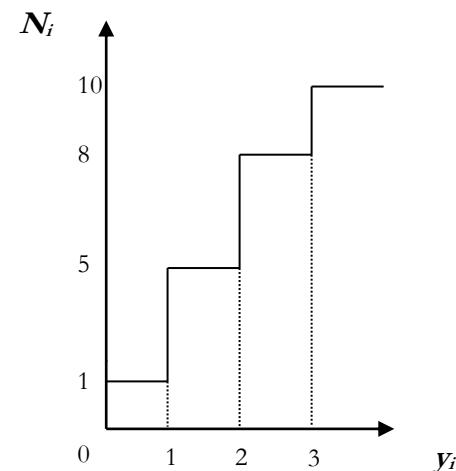


Gráfico Acumulativo de Frecuencias

Podríamos haber utilizado H_i , rotulando convenientemente.



A partir de esta información y recurriendo a la Planilla Excel construiremos el gráfico de bastones y el gráfico acumulativo de frecuencias.

Gráfico de Bastones

En primer lugar utilizaremos la frecuencia absoluta, que indica en este caso el número de familias.

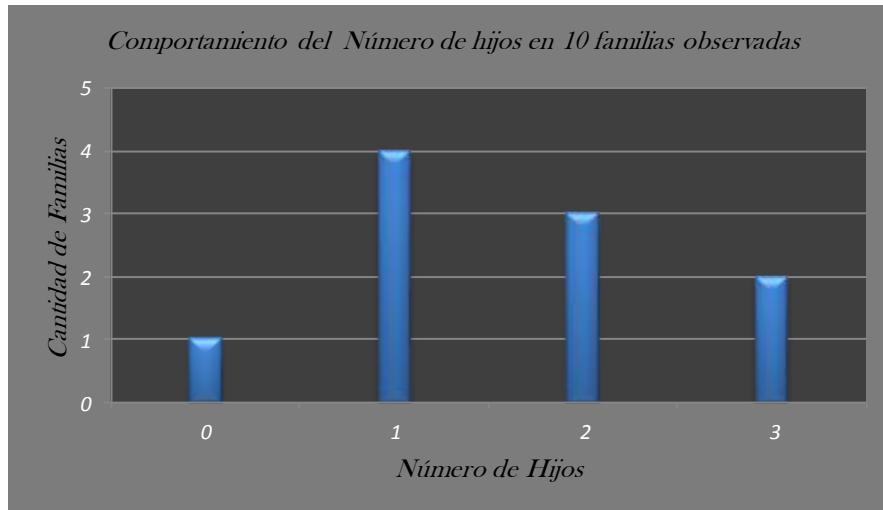


Gráfico de Bastones

Podríamos haber utilizado h_i , rotulando convenientemente, lo que indicará la proporción de familias.

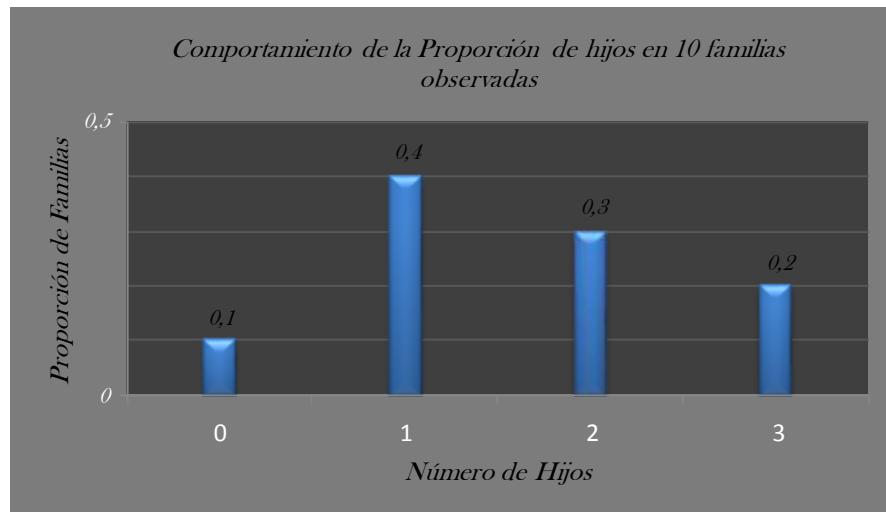


Gráfico de Bastones



Gráfico Acumulativo de Frecuencias

En primer lugar utilizaremos la frecuencia absoluta, que indica en este caso el número de familias:

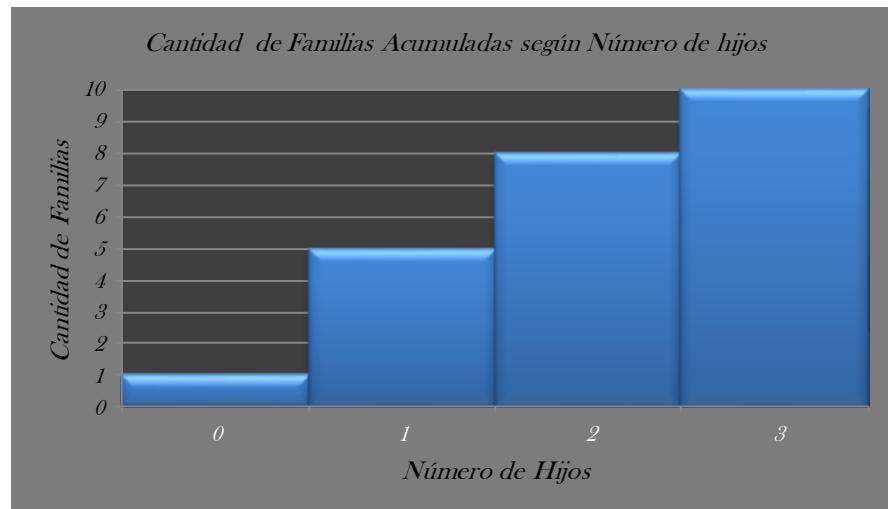


Gráfico Acumulativo de Frecuencias

Podríamos haber utilizado h_i , rotulando convenientemente, lo que indicará la proporción de familias.

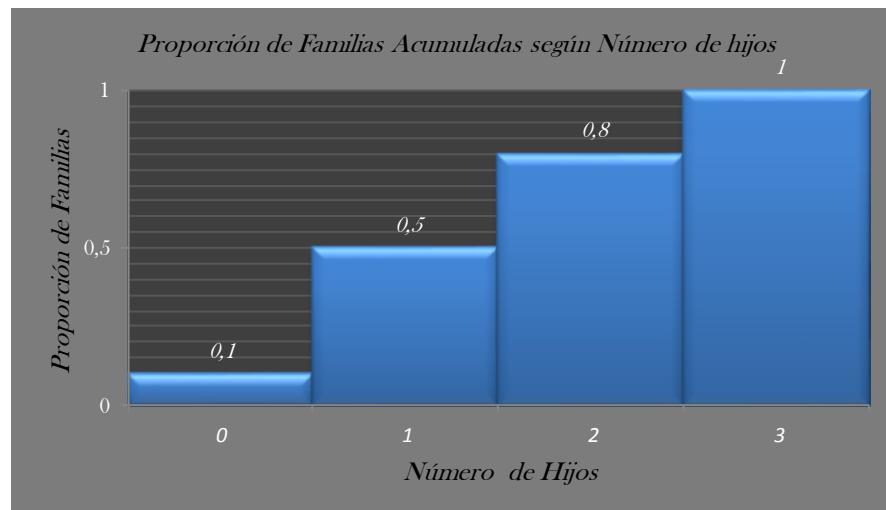


Gráfico Acumulativo de Frecuencias



4.2. Gráficos de Superficie

4.2.1. Histograma

Características especiales para su construcción

1- Las frecuencias de clase suelen representarse en el eje “y”, y la escala de los intervalos en el eje “x”. Los ejes “x” e “y” deben comenzar en 0, con interrupciones de escala si son necesarias. Ambos ejes, deben ser rotulados, clara y completamente.

2- Un espacio de la mitad al tamaño completo del intervalo de clase, se deja en cada extremo del eje “x”.

3- Las designaciones de escala “x” suelen ser colocadas como los verdaderos límites de clase. Las barras deben tocarse unas con otras, sin brechas, excepto para clases vacías. A veces, se rotula la escala “x” colocando el valor medio de cada clase en el centro de la base de la barra.

La escala “x”, es igualmente espaciada cuando los intervalos de clase son uniformes. En una distribución donde los intervalos son variables, la escala “x” debe ser ajustada apropiadamente. Por ejemplo, si se usan dos intervalos, de amplitud 100 y 500, el espacio para clases de 500 debe ser 5 veces más ancho que los espacios con intervalos de amplitud 100.

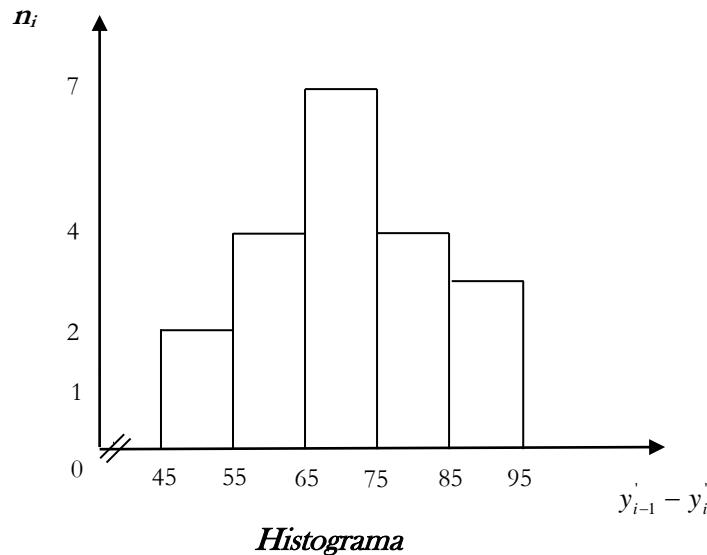
4- El eje “y” debe ser completamente rotulado para mostrar si representa frecuencia o densidad de frecuencia. Si los intervalos son **uniformes**, el patrón visual del gráfico será igual, tanto si se representan frecuencias como densidades de frecuencias. Sin embargo es conveniente rotular el eje “y” de modo que el lector conozca cuál está siendo representado. Si los intervalos son **no uniformes**, el patrón visual del gráfico diferirá según si se representan frecuencias o densidades de frecuencias. Es conveniente utilizar densidades, de manera de homogeneizar las unidades, puesto que en este caso se habla de individuos o proporción de individuos, por unidad de intervalo, o intervalo estándar, mientras que con las frecuencias, se tiene cantidad o proporción por cada intervalo que tiene distinta amplitud.

5- Un histograma se representa siempre como compuesto de barras, tanto si se muestran explícitamente como si no se muestran. Entonces, se construye utilizando un segmento de la escala horizontal para representar cada intervalo de clase, erigiendo un rectángulo igual en altura a la frecuencia absoluta o relativa o densidad correspondiente a esa clase.

Ejemplo:

Utilizaremos el ejercicio desarrollado en Distribuciones de Frecuencias en Intervalos

$y_{i-1} - y_i$	y_i	n_i	n/c	h	h/c
45-55	50	2	2/10	0,10	0,10/10
55-65	60	4	4/10	0,20	0,20/10
65-75	70	7	7/10	0,35	0,35/10
75-85	80	4	4/10	0,20	0,20/10
85-95	90	3	3/10	0,15	0,15/10
		20		1	



Hemos utilizado en este caso, frecuencias absolutas, pero podría haberse construido con frecuencias relativas, o densidades, siguiendo igual procedimiento.

Esta alternativa es válida siempre que los intervalos sean uniformes, pues si la amplitud de clase es distinta deberá construirse con densidad de frecuencia, sea absoluta o relativa.

Analizaremos seguidamente cuál es la interpretación de las partes del Histograma:

Concepto	Interpretación
Base o Ancho de cada Barra	Mide la amplitud del intervalo de clase, o sea c , cualquiera sea la rotulación del eje "y"
Altura de cada Barra	Su significado varía, según como se rotula el eje "y"
n	Cantidad de unidades que hay en dicha clase.
h	Proporción de unidades que hay en dicha clase.
n/c	Promedio en cada clase de cuántas unidades hay por unidad de ancho de intervalo
h/c	Promedio en cada clase de la proporción de unidades hay por unidad de ancho de intervalo

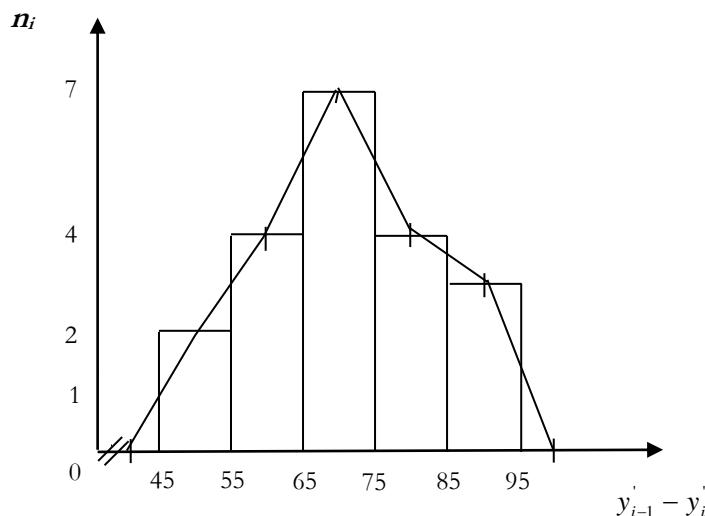


<u>Área de cada Barra</u>	<i>Su significado varía, según como se rotula el eje "y"</i>
n_i	<i>Carece de significado</i>
h_i	<i>Carece de significado</i>
n_i/c_i	<i>La frecuencia absoluta para dicha clase, pues $n_i/c_i \times c_i = n_i$</i>
h_i/c_i	<i>La frecuencia relativa para dicha clase, pues $h_i/c_i \times c_i = h_i$</i>
<u>Área Total del Histograma</u>	<i>Su significado varía, según como se rotula el eje "y"</i>
n	<i>Carece de significado</i>
h	<i>Carece de significado</i>
n_i/c	<i>n, puesto que si el área de cada barra es n_i, la sumatoria de los n_i ($i=1,2,\dots,m$) es igual al tamaño de la muestra.</i>
h_i/c	<i>1, puesto que si el área de cada barra es h_i, la sumatoria de los h_i ($i=1,2,\dots,m$) es igual a 1</i>

4.2.2. Polígono de Frecuencias

Si disponemos de una Histograma, construimos el Polígono uniendo con líneas rectas los puntos medios de cada barra del histograma.

Con mucha frecuencia, se construyen sin trazar los rectángulos. Sin el Histograma obtenemos el Polígono localizando las coordenadas: las ordenadas que son las frecuencias de clases y las abscisas, que son los puntos medios. Estos puntos son unidos después por líneas rectas.



No representa muy bien los datos básicos. La diferencia más notable es que las áreas situadas debajo de él, generalmente no son proporcionales a las frecuencias. Un remedio es cerrar el Polígono en la base prolongando ambos extremos de la curva hasta los puntos medios de dos clases hipotéticas situadas en los extremos de la distribución que tienen frecuencia 0.



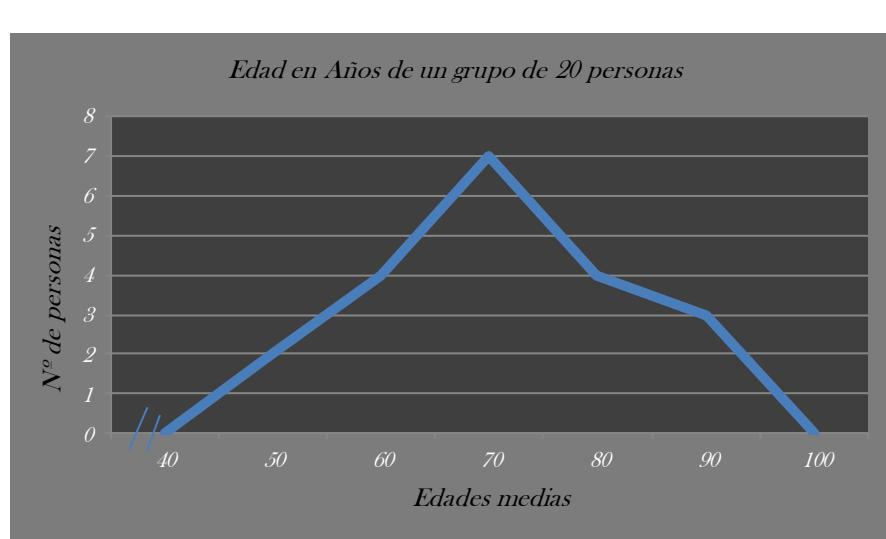
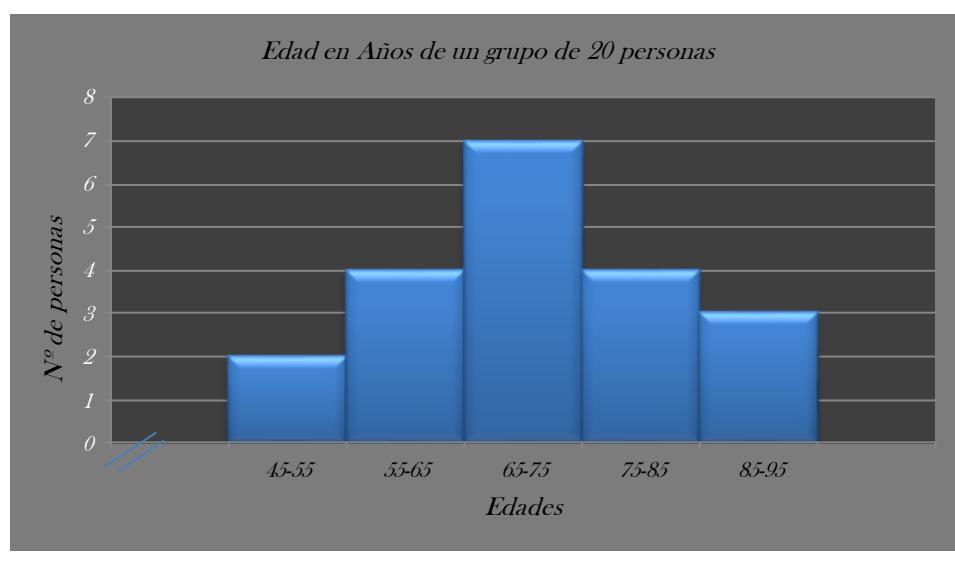
Razones para usar el Polígono:

- 1- Cuando han de compararse varias distribuciones sobre el mismo gráfico, es mucho más claro superponer polígonos que Histogramas, especialmente cuando todas las distribuciones tienen los mismos límites de clase.
- 2- El Polígono sugiere el empleo de una curva suave como una representación idealizada de la distribución de la población.

Una muestra consta de sólo un número limitado de unidades, porque su distribución se caracteriza por irregularidades.

Sin embargo, si las unidades de la muestra se incrementan y la amplitud de los intervalos se disminuye continuamente, podemos esperar que la distribución sea cada vez más suave y cada vez más regular, porque las irregularidades que afectan a una pequeña cantidad de datos, serían eliminadas gradualmente.

Recurriendo Excel, se obtienen los siguientes gráficos:





4.2.3. Curva Suave

Cuando la muestra es muy grande, los intervalos de clase son muy estrechos (c muy pequeño), pero cada uno contendrá un número sustancial de unidades. Al mismo tiempo, si la escala vertical que mide la frecuencia, es reducida de modo que el área del histograma para esta muestra extraordinariamente grande sea igual al área de la pequeña muestra original, el Histograma de la muestra grande formará prácticamente una curva suave.

La Curva Suave adquiere importancia porque se considera que representa la verdadera distribución de la población de la que se extrae la muestra. Pero la derivación de una curva suave ampliando la muestra es generalmente una imposibilidad práctica. Lo que solemos hacer, es aproximar la distribución de la población sobre la base de los datos de la muestra, suavizando las puntas del Polígono de Frecuencia, dibujando a mano o introduciendo una curva suave a los datos de la muestra con alguna fórmula matemática.

Debido a la sorprendente libertad para ajustar la curva, la Curva Suave debe presentarse siempre con el Histograma.

Son llamadas alternativamente Modelos de Población, porque describen las características importantes de las distribuciones de población.

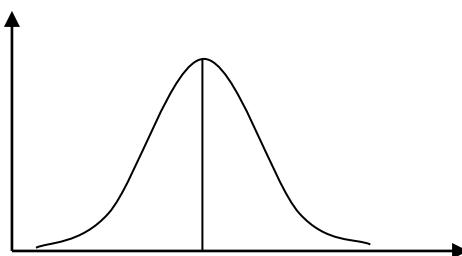
Estas generalizaciones son de gran utilidad en el análisis estadístico porque proporcionan métodos simplificados de describir las características básicas de las poblaciones.

Otras razones de interés por los Modelos de Población:

- 1- Son necesarios para la toma de decisiones.
- 2- Las inferencias estadísticas a menudo requieren que conozcamos Modelos de Población.
- 3- Un Modelo de Población, estando representado por una curva suave, a veces se presta más fácilmente a un tratamiento matemático.

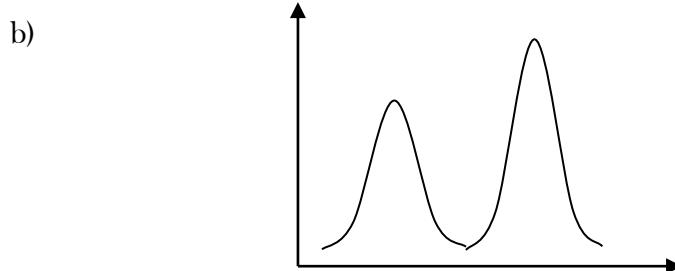
Algunos Modelos de Población son:

a)



Tiene forma de campana, las densidades de frecuencia más grandes están en el centro. Hay densidades muy pequeñas en ambos extremos. Es una Distribución Simétrica, llamada Curva de Distribución Normal o Curva Normal.

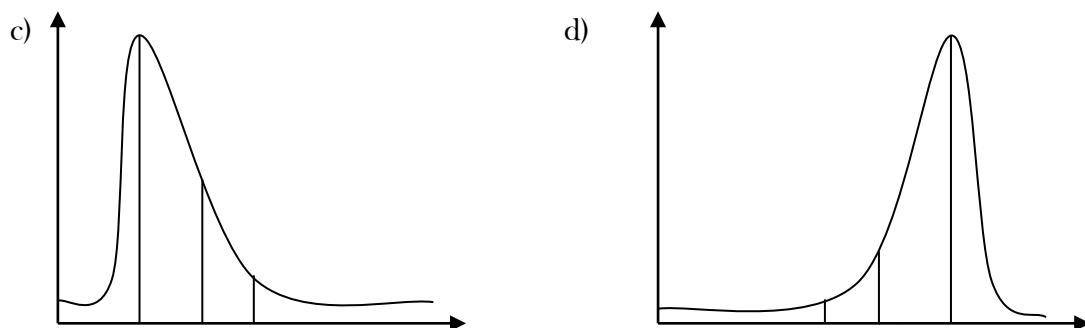
Ejemplo: Estatura, Inteligencia.



Distribución Bimodal. Significa que tiene dos picos. Este caso de distribución aparece cuando una población contiene ciertos elementos que pueden ser divididos en dos clases que difieren entre sí en las características que se miden. La población no es homogénea.

Ejemplo:

Saldos de depósitos a la vista. Un pico destacado se encuentra en un valor relativamente bajo para los saldos mantenidos por unidades de consumo y otro pico distinto, para un valor relativamente alto para saldos mantenidos por empresas comerciales.



Modelos de Distribuciones Asimétricas. Generalmente tienen un solo pico situado en la parte inferior o superior de la curva.

Cuando la cola más larga está a la derecha (c), la distribución es asimétrica derecha ó positivamente asimétrica.

Ejemplo:

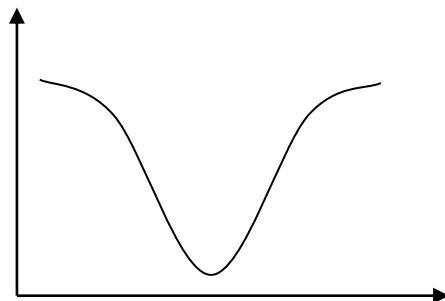
Distribución de sueldos, indicando que algunos empleados (relativamente pocos en comparación con el grupo general) reciben sueldos más altos que los recibidos por la mayoría de los empleados.

Cuando la cola más larga está a la izquierda (d), la distribución es asimétrica izquierda o negativamente asimétrica.

Ejemplo:

Describe bien una población cuyas variables tienen un límite superior. El límite superior de la razón de costo de ventas (Costo de Ventas/Ventas), sería la unidad, o 100x100.

e)

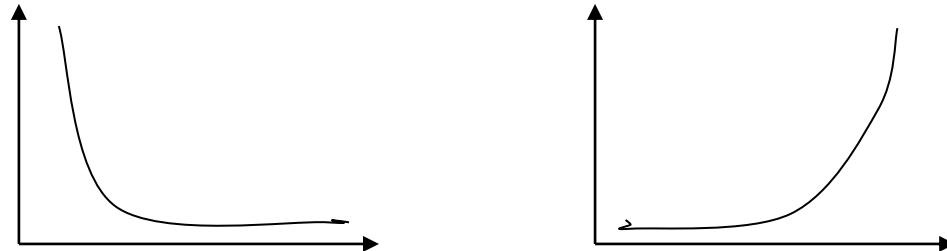


Describe una distribución que contienen predominantemente valores bajos y altos, siendo relativamente escasos los valores intermedios.

Ejemplo:

La distribución de las naciones del mundo según sus etapas de desarrollo económico revelaría abultamientos en dos extremos, con solo unos pocos en las etapas intermedias.

f)



Curva en forma de J ó J invertida, en que las frecuencias de ocurrencia aumentan o disminuyen continuamente a lo largo de la escala horizontal.

Ejemplo:

Distribución de corporaciones clasificadas según el tamaño del activo.

Distribución de quiebras comerciales con el eje x como tiempo de operación.

Si bien la mayor parte de las veces no conocemos las verdaderas distribuciones de población, podemos aproximar sus modelos trazando una curva suave a los datos de la muestra o por puro razonamiento deductivo.



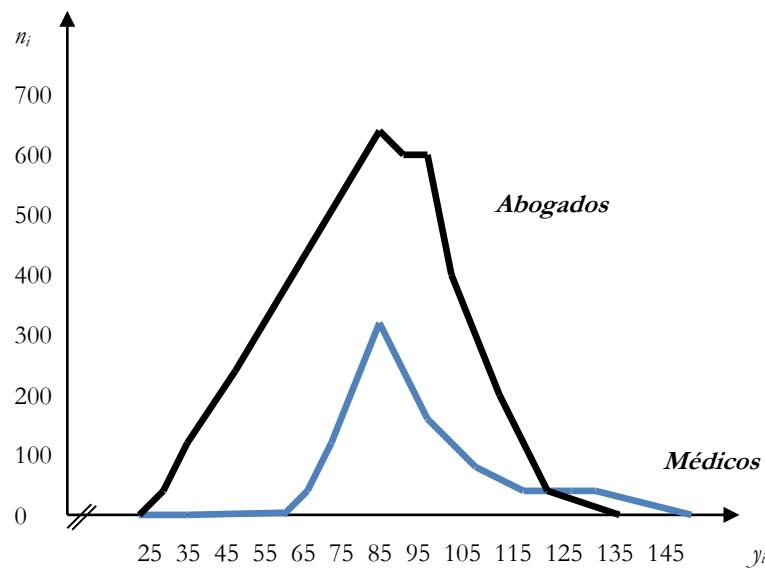
Distribuciones Relativas De Frecuencias

El gráfico de una distribución relativa (Histograma, Polígono, Curva Suave), se presta más fácilmente a la comparación de diferentes distribuciones, especialmente si difieren mucho en el número total de observaciones.

Ejemplo:

Distribución de Médicos y Abogados según ingresos diarios medios.

Ingresos en U\$S	Médicos		Abogados	
	n	n/n	n	n/n
20-30	0	0,000	11	0,003
30-40	0	0,000	135	0,042
40-50	1	0,001	247	0,077
50-60	24	0,024	466	0,145
60-70	150	0,150	658	0,205
70-80	322	0,322	596	0,185
80-90	185	0,185	579	0,180
90-100	120	0,120	379	0,118
100-110	78	0,078	115	0,036
110-120	66	0,066	25	0,008
120-130	22	0,022	3	0,001
130-140	17	0,017	0	0,000
140-150	15	0,015	0	0,000
TOTAL	1000	1	3214	1



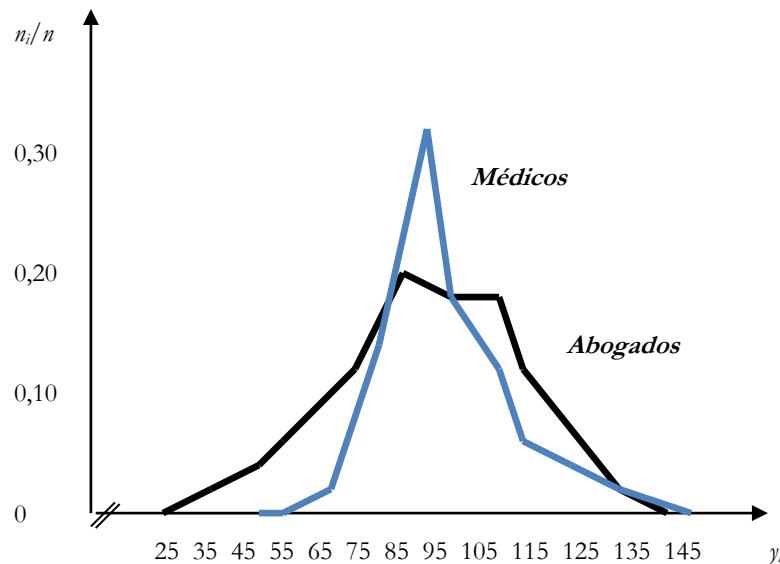


GRÁFICO II

Es muy difícil descubrir en el Gráfico I, las similitudes o diferencias entre las dos distribuciones. Más grave es el hecho de que pueden obtenerse algunas impresiones incorrectas de ellas.

Pero cuando se consultan los polígonos de las dos distribuciones relativas pueden extraerse varias conclusiones fácil y precisamente:

- 1- La distribución para los abogados tiende a localizarse en valores más bajos que para los médicos.
- 2- Ambas distribuciones son positivamente asimétricas, pero el grado de asimetría para los médicos es mayor.
- 3- Aunque la distribución de los médicos está más compactamente distribuida alrededor del pico, la de los abogados semeja una cima plana.
- 4- Una mayor proporción de médicos que de abogados tienen ingresos medios entre U\$S 60 y U\$S 100. Una mayor proporción de médicos tienen ingresos medios diarios inferiores a U\$S 60, mientras que una mayor proporción de médicos que de abogados tienen ingresos medios diarios superiores a U\$S 100.

La comparación de frecuencias relativas resulta aún más reveladora cuando tratamos distribuciones cuyos valores mínimo y máximo, o ambos, difieren mucho entre sí.

4.2.4. Diagrama escalonado

Representa a una distribución de frecuencias acumuladas mediante una serie de líneas horizontales trazadas en las correspondientes clases a la altura de las respectivas frecuencias, con las coordenadas que presentan las escalas de la misma forma que para los Histogramas. Los puntos finales de las líneas horizontales pueden ser unidos o no por líneas verticales.

**Ejemplo:**

Utilizaremos los datos dados para la construcción del Histograma

$y_{i-1} - y_i$	y_i	N_i	H
45-55	50	2	0,10
55-65	60	6	0,30
65-75	70	13	0,65
75-85	80	17	0,85
85-95	90	20	1

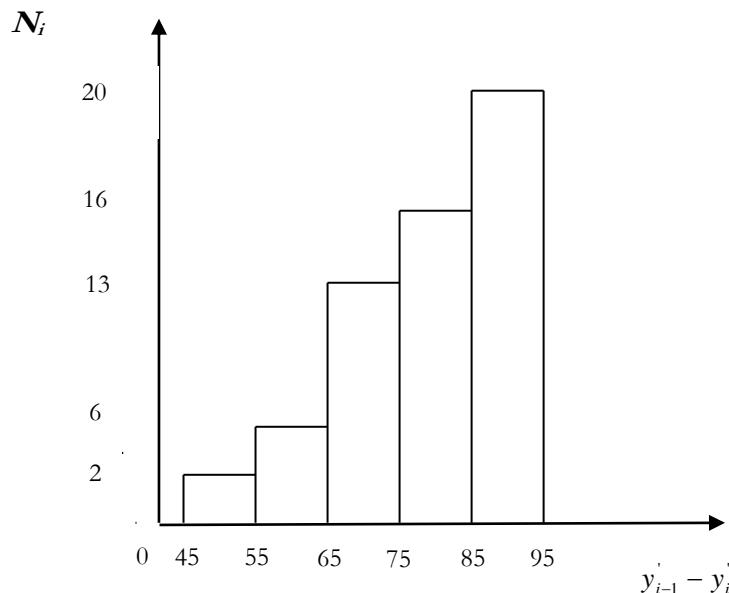


Diagrama Escalonado

Utilizando Excel, la gráfica es la siguiente:

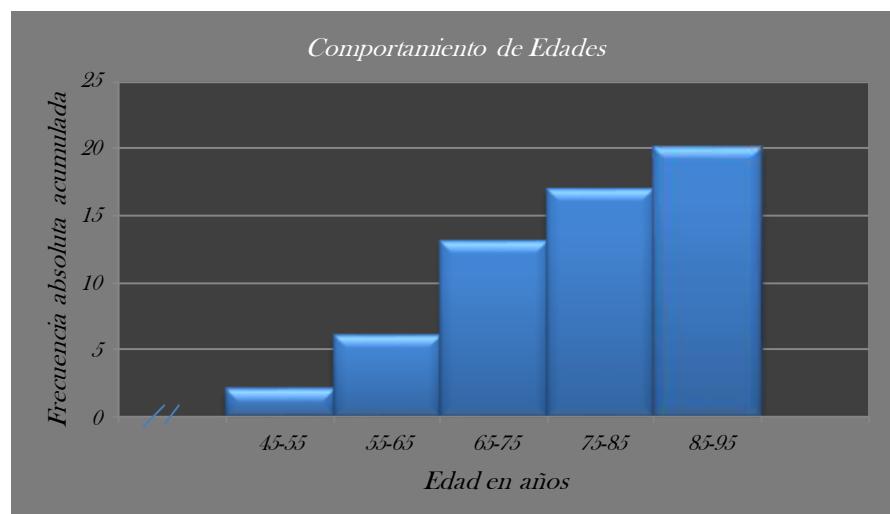


Diagrama Escalonado



4.2.5. Ojiva

Es un polígono que representa una distribución acumulada en forma de un diagrama de líneas.

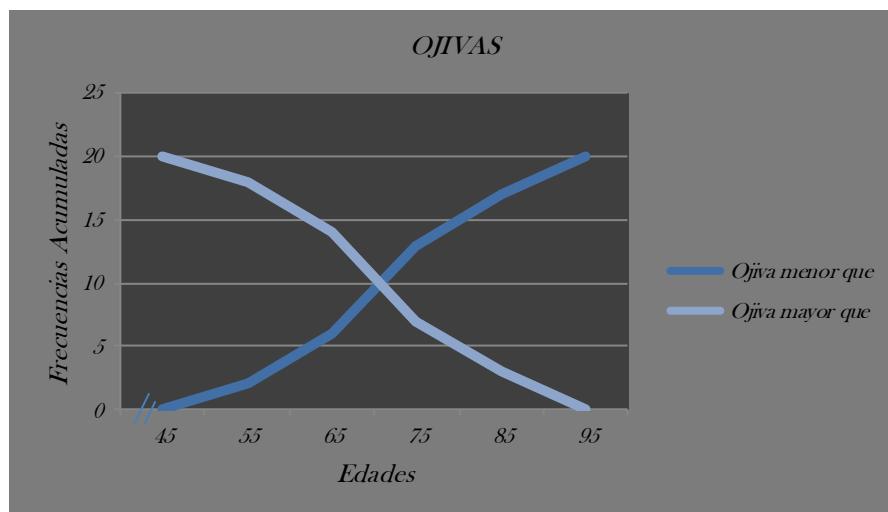
Puede ser:

- Ojiva Mayor que
- Ojiva Menor que

En el caso de “Menor que”, las frecuencias son acumuladas usando límites de clase superior, y_i^+ .

En el caso de “Mayor que”, las frecuencias son acumuladas usando los límites de clase inferior, y_{i-1}^+ .

$y_{i-1}^+ - y_i^+$	y_i^+	N		H	
		<	>	<	>
45-55	50	2	20	0,10	20/20
55-65	60	6	18	0,30	18/20
65-75	70	13	14	0,65	14/20
75-85	80	17	7	0,85	7/20
85-95	90	20	3	1	3/20



Una ojiva se usa principalmente para interpolaciones, que pueden hacerse de dos modos:

1- Menor que: Si se escoge un punto de la escala horizontal, el número o proporción correspondiente de observaciones en la distribución cuyos valores son iguales



o menores que el valor indicado por el punto escogido pueden encontrarse en la escala vertical.

Por ejemplo, si escogemos el punto 60 de la escala horizontal, trazamos una línea vertical hasta cortar la ojiva y de esta intersección trazamos una línea horizontal a la escala vertical de la izquierda, obtenemos el valor 4. Esto significa que aproximadamente 4 observaciones de la muestra tienen valores iguales o menores de 60.

2- Mayor que: Esta vez nos desplazamos del eje vertical al eje horizontal para hallar el valor debajo del cual encontraremos un número, o proporción dado de observaciones. Supongamos que trazamos una línea horizontal desde 10 para que corte la ojiva y, luego, bajamos una perpendicular de la intersección a la escala horizontal (70).

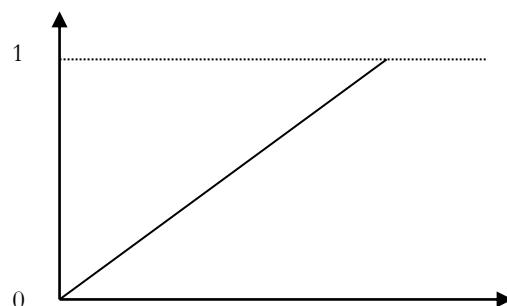
Esto significa que 10 observaciones de la muestra tienen valores iguales o menores de 70 (o también 10 tienen valores iguales o mayores de 70).

4.2.6. Curva Acumulativa Suavizada

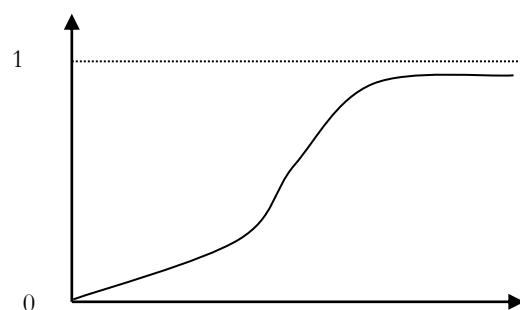
También el Diagrama Escalonado o una Ojiva pueden ser suavizados para representar una Distribución de Población.

Formas:

a- Distribución Uniforme

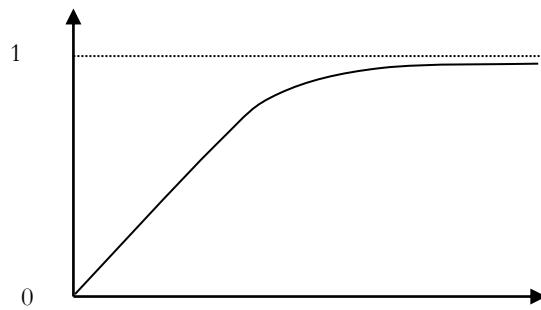


b- Distribución en forma de campana

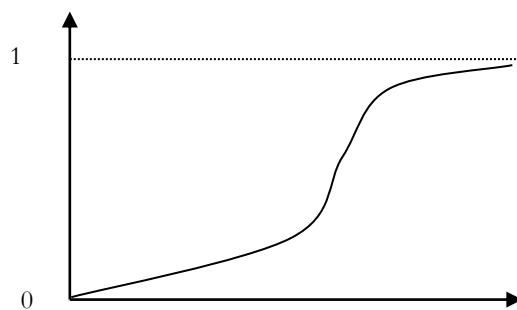




c- Distribuciones Positivamente Asimétricas



d- Distribuciones Negativamente Asimétricas



4.3. Gráficos Especiales

4.3.1. Distribuciones Categóricas

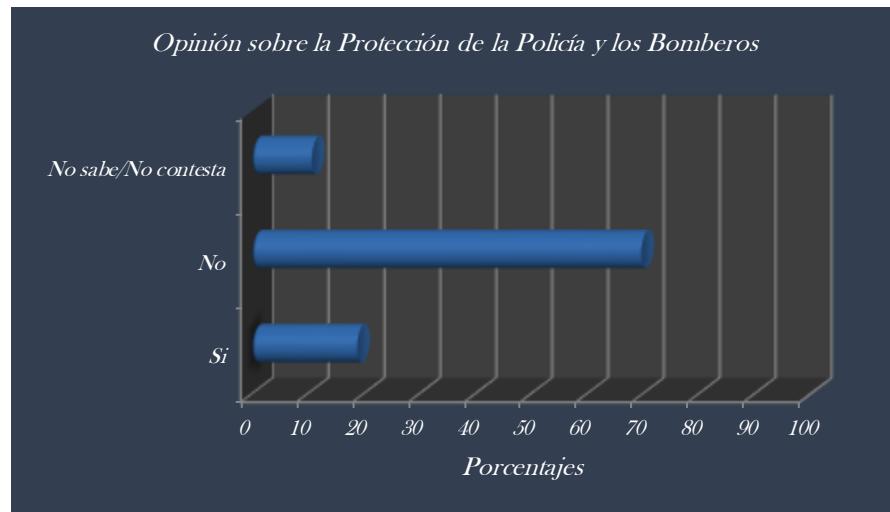
4.3.1.1. *Diagrama de Barras Horizontales*

La escala horizontal mide la frecuencia, absoluta o relativa, de cada categoría. La longitud de cada barra es la frecuencia de dicha categoría. Los anchos de las barras son enteramente arbitrarios y no tienen significado práctico. Las barras pueden ser verticales, en vez de horizontales, con tal que, por supuesto, los ejes sean apropiadamente definidos.

Consideraremos el ejemplo sobre la “*Opinión sobre la Protección de la Policía y los Bomberos*”:

Respuestas	n	%
Sí	80	19,09
No	293	69,93
No sabe/No contesta	46	10,98
	419	100,00

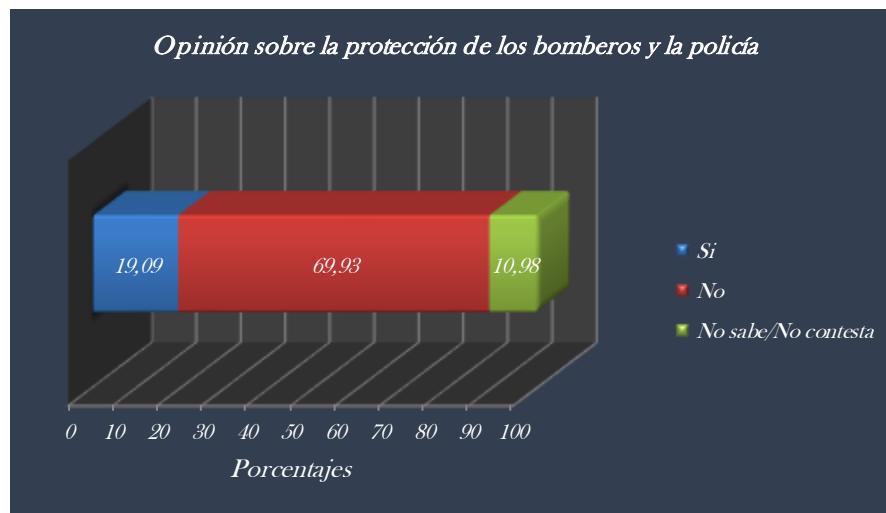
Haciendo uso de la planilla de cálculo Excel, obtenemos la siguiente gráfica:



4.3.1.2. Gráfica de Barras de Componentes de Porcentajes

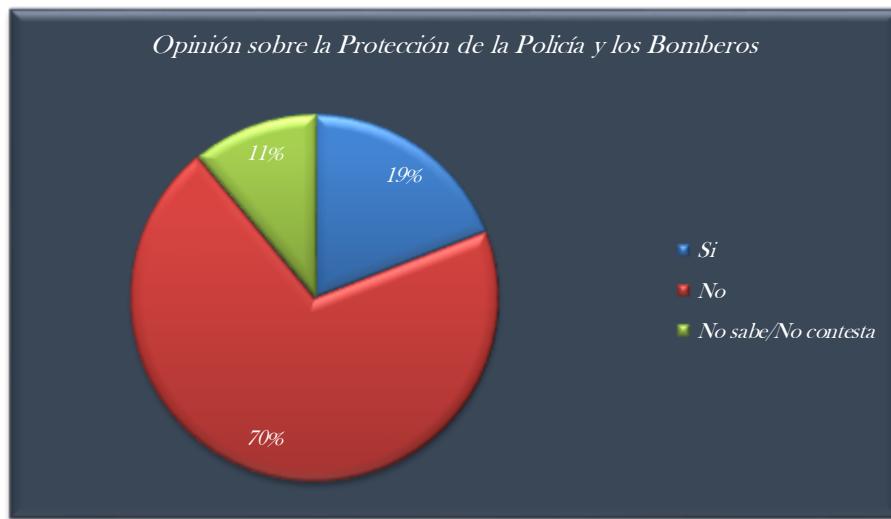
También llamado Barra Porcentual, se representa mediante un rectángulo horizontal, donde la superficie del rectángulo representa el 100% del fenómeno analizado.

Lo graficaremos a través de una planilla de cálculo, encontrando:



4.3.1.3. Diagrama de Pastel o Círculo Radiado

También indica porcentajes del fenómeno estudiado y para su construcción utilizaremos los datos dados en el ejemplo anterior. Utilizando una planilla de cálculo obtenemos:



Ejemplo:

Se está analizando la aplicación de los Ingresos de una Universidad Estatal y de una Universidad Privada.

Supongamos que los Ingresos de la Universidad Estatal son de \$10.000, distribuyéndose de la siguiente manera: Sueldos de Docentes: \$3000; Sueldos de No Docentes: \$2000; y Otros Gastos: \$5000, y que los Ingresos de la Universidad Privada son de \$15000, distribuyéndose de la siguiente manera: Sueldos de Docentes: \$3500; Sueldos de No Docentes: \$4000; y Otros Gastos: \$5500.

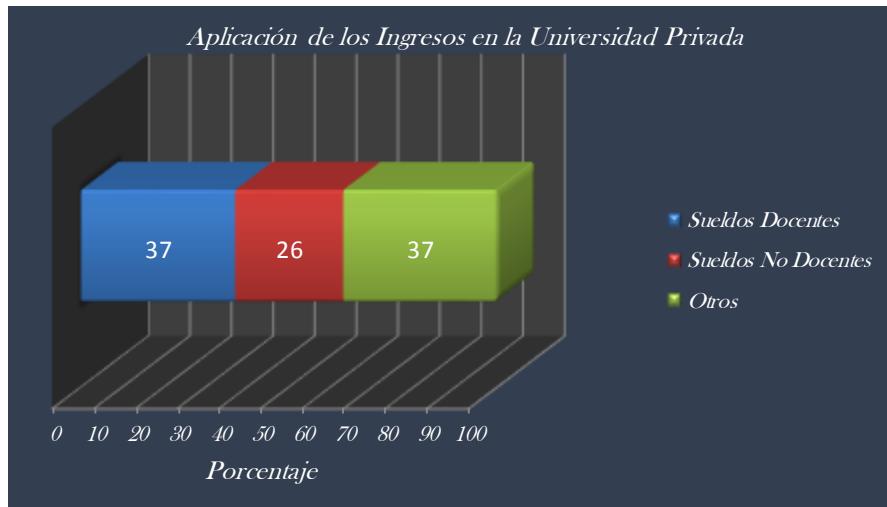
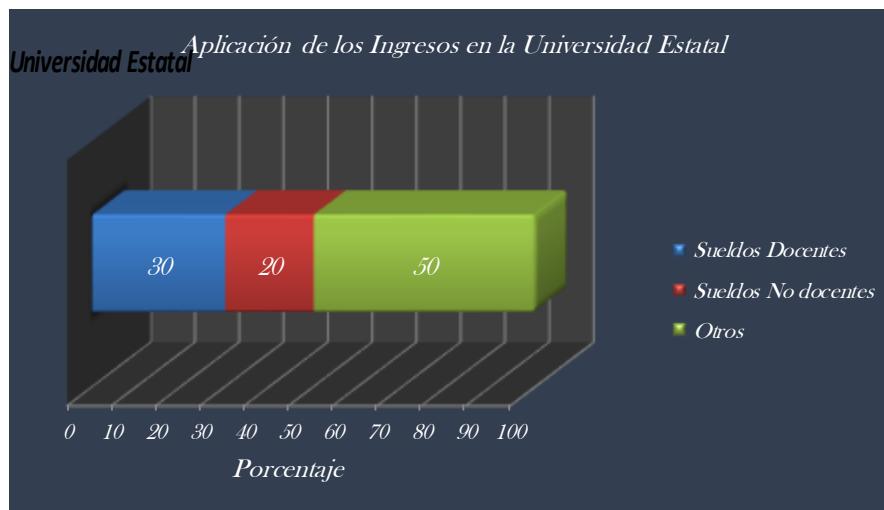
Para graficar la anterior información, construiremos dos Barras, una para la Universidad Estatal y la otra para la Universidad Privada.

Universidad Estatal

<i>Aplicación de Ingresos (1)</i>	<i>Porcentaje (2)</i>
Sueldos Docentes	30
Sueldos No Docentes	20
Otros	50
Total	100

Universidad Privada

<i>Aplicación de Ingresos (1)</i>	<i>Porcentaje (2)</i>
Sueldos Docentes	37
Sueldos No Docentes	26
Otros	37
Total	100



Luego, podemos comparar dentro de cada Barra y entre Barras.

Así, observamos que en la Universidad Privada es mayor el porcentaje de los Ingresos destinados a Sueldos Docentes y No Docentes y menor el destinado a otros gastos, en comparación con la

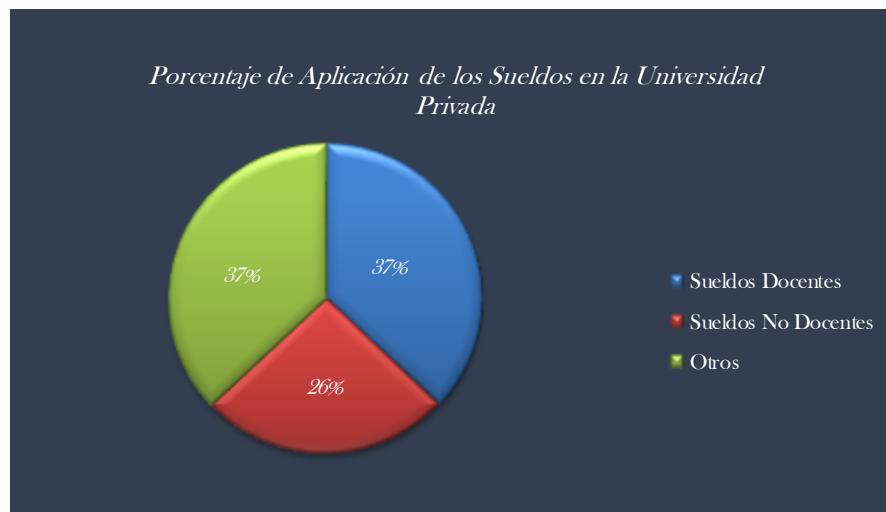
Universidad Estatal. Por otro lado la Universidad Estatal aplica iguales porcentajes a sueldos que a gastos.



Ahora, construiremos el círculo radiado

<i>Aplicación de Ingresos (1)</i>		<i>Porcentaje (2)</i>
<i>Sueldos Docentes</i>	3000	30
<i>Sueldos No Docentes</i>	2000	20
<i>Otros</i>	5000	50
Total	10000	100

Haciendo uso de la Planilla Excel:





4.3.2. Para Otros Fenómenos

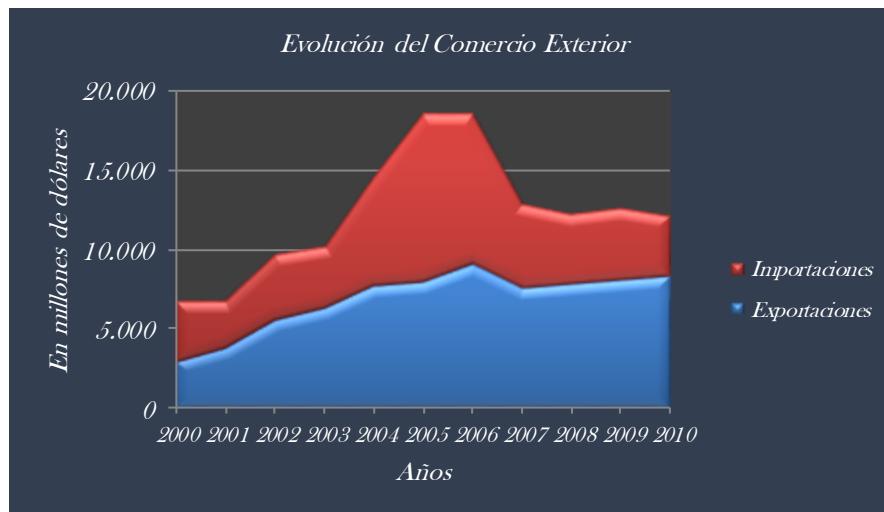
4.3.2.1. Gráfico de Zonas

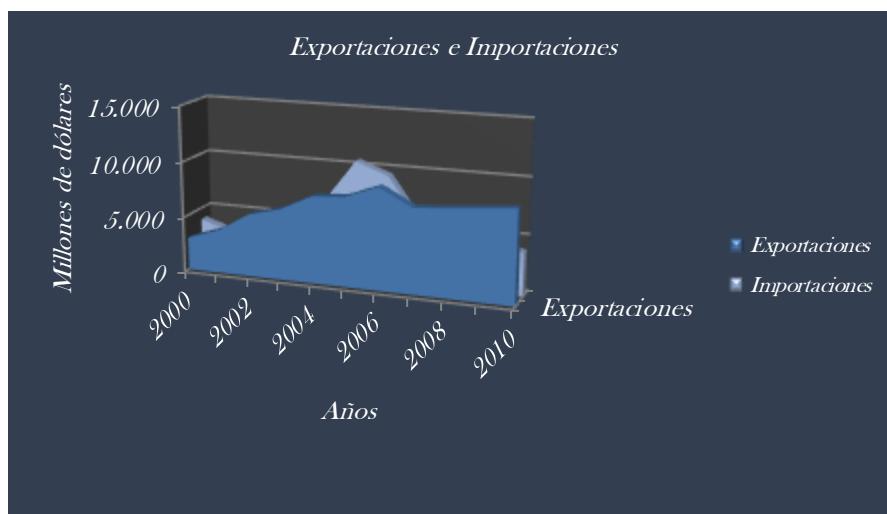
Sirve para representar en un mismo gráfico varios fenómenos con el fin de compararlos entre sí. Se utiliza muy frecuentemente en series cronológicas que analizan aspectos que ocurren a través del tiempo. Se pueden sumar dos o más fenómenos como en el siguiente ejemplo, en el que se adicionan exportaciones e importaciones, mostrando en la ordenada, el comportamiento de cada una, así como su total.

Ejemplo:

Si contamos con la información sobre la “Evolución del Comercio Exterior-Exportaciones”, en millones de dólares y la “Evolución del Comercio Exterior-Importaciones”, en millones de dólares para los años 2000 a 2010, por totales de un determinado país, según se expone en la siguiente tabla:

Años	Exportaciones	Importaciones	Totales
1975	2.942	3.946	6.888
1976	3.916	3.033	6.949
1977	5.650	4.162	9.812
1978	6.396	3.834	10.230
1979	7.810	6.700	14.510
1980	8.021	10.541	18.562
1981	9.143	9.430	18.573
1982	7.625	5.337	12.962
1983	7.836	4.501	13.173
1984	8.107	4.583	12.690
1985	8.396	3.814	12.210





4.3.2.2. Diagrama de Pareto

Reiteramos lo dicho anteriormente, en relación a que los datos deben recolectarse de tal forma que ofrezcan la información vital necesaria para resolver problemas. Es por ello, que deben describirse y analizarse para producir información resumida. Las representaciones gráficas son útiles para el cumplimiento de este objetivo.

Veremos ahora una representación gráfica, en la que cada tipo de falla o defecto se organiza de acuerdo con su frecuencia, ayudando a los ingenieros a identificar importantes problemas y sus causas, dado que es una herramienta que se utiliza para priorizar los problemas o las causas que los generan.

Entonces, el *Diagrama De Pareto* (conocido también como diagrama ABC, 80-20,70-30), es una gráfica para organizar datos de forma que estos queden en un orden descendente, de izquierda a derecha y separados por barras, permitiendo de esta manera, asignar un orden de prioridades.

El diagrama se basa en el *Principio De Pareto* (pocos vitales, muchos triviales) es decir, que hay muchos problemas sin importancia frente a unos pocos graves. Mediante la gráfica colocamos los "pocos vitales a la izquierda" y los "muchos triviales" a la derecha.

El nombre de Pareto fue dado por el Dr. Juran en honor del economista italiano VILFREDO PARETO (1848-1923) quien realizó un estudio sobre la distribución de la riqueza, en el cual descubrió que la minoría de la población poseía la mayor parte de la riqueza y la mayoría de la población poseía la menor parte de la riqueza. El Dr. Juran aplicó este concepto a la calidad, obteniéndose lo que hoy se conoce como la regla 80/20.

Según este concepto, si se tiene un problema con muchas causas, podemos decir que el 20% de las causas resuelven el 80 % del problema y el 80 % de las causas solo resuelven el 20 % del problema.

El diagrama es una buena herramienta de trabajo que facilita el estudio comparativo de los numerosos procesos llevados a cabo en industrias, así como fenómenos naturales que requieran de este tipo de análisis. Hay que tener en cuenta que



tanto la distribución de los defectos como sus posibles causas no presentan un comportamiento lineal, sino que el 20% de las causas totales originan el 80% de los defectos o problemas.

Se recomienda su uso:

- Para identificar oportunidades para mejorar
- Para identificar un *Producto O Servicio* para el análisis de mejora de la calidad.
- Cuando existe la necesidad de llamar la atención a los problemas o causas de una forma sistemática.
- Para analizar las diferentes agrupaciones de datos.
- Al buscar las causas principales de los problemas y establecer la prioridad de las soluciones.
- Para evaluar los resultados de los cambios efectuados a un proceso comparando sucesivos diagramas obtenidos en momentos diferentes, (antes y después).
- Cuando los datos puedan clasificarse en categorías.
- Cuando el rango de cada categoría es importante.
- Para comunicar fácilmente a otros miembros de la organización las conclusiones sobre causas, efectos y costes de los errores.

Los propósitos generales del diagrama de Pareto:

- Analizar las causas
- Estudiar los resultados
- Planear una mejora continua

La Gráfica de Pareto es una herramienta sencilla pero poderosa al permitir identificar visualmente en una sola revisión las minorías de características vitales a las que es importante prestar atención y de esta manera utilizar todos los recursos necesarios para llevar a cabo una acción de mejora sin malgastar esfuerzos ya que con el análisis descartamos las mayorías triviales.

Algunos ejemplos de tales minorías vitales serían:

- La minoría de clientes que representen la mayoría de las ventas.
- La minoría de productos, procesos, o características de la calidad causantes del grueso de desperdicio o de los costos de re trabajos.
- La minoría de rechazos que representa la mayoría de quejas de los clientes.
- La minoría de vendedores que está vinculada a la mayoría de partes rechazadas.
- La minoría de problemas causantes del grueso del retraso de un proceso.
- La minoría de productos que representan la mayoría de las ganancias obtenidas.
- La minoría de elementos que representan la mayor parte del costo de un inventario etc.

Ejemplo:

Un fabricante de accesorios plásticos desea analizar cuáles son los defectos más frecuentes que aparecen en las unidades al salir de la línea de producción. Para esto, empezó por clasificar todos los defectos posibles en sus diversos tipos:



Tipo de Defecto	Detalle del Problema
Mal color	El color no se ajusta a lo requerido por el cliente
Fuera de medida	Ovalización mayor a la admitida
Mal terminación	Aparición de rebabas
Rotura	El accesorio se quiebra durante la instalación
Desbalance	El accesorio requiere contrapesos adicionales
Aplastamiento	El accesorio se aplasta durante la instalación
Incompleto	Falta alguno de los insertos metálico
Mal alabeo	Nivel de alabeo no aceptables
Otros	Otros defectos

Posteriormente, un inspector revisa cada accesorio a medida que sale de producción registrando sus defectos de acuerdo con dichos tipos. Al finalizar la jornada, se obtuvo una tabla como esta:

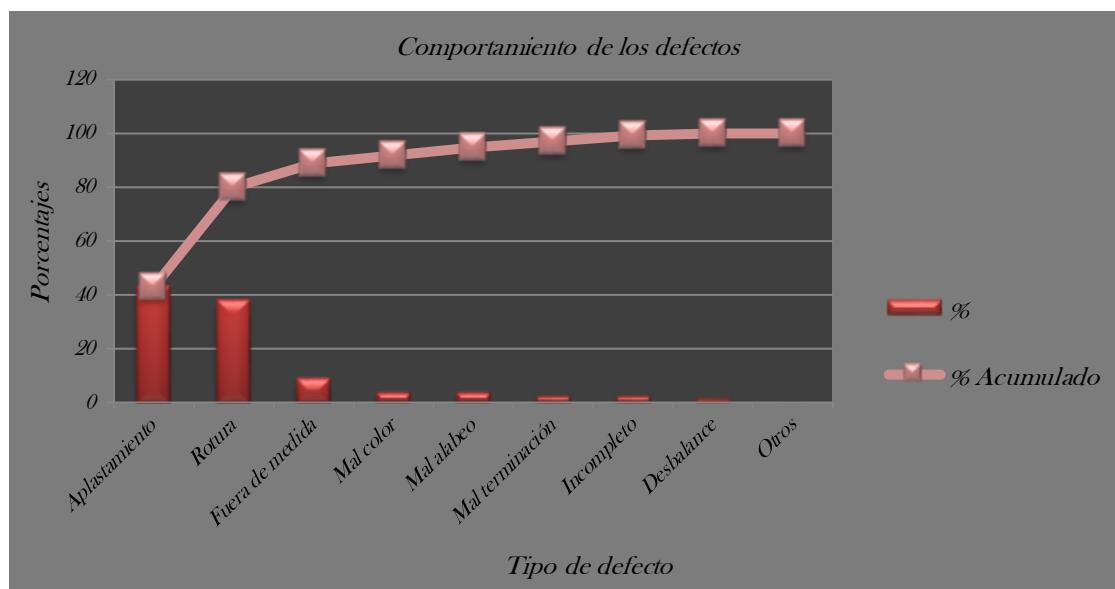
Tipo de defecto	Detalle del problema	n	%	% Acumulado
Aplastamiento	El accesorio se aplasta durante la instalación	40	42.6	42.6
Rotura	El accesorio se quiebra durante la instalación	35	37.2	79.8
Fuera de medida	Ovalización mayor a la admitida	8	8.5	88.3
Mal color	El color no se ajusta a lo requerido por el cliente	3	3.2	91.5
Mal alabeo	Nivel de alabeo no aceptable	3	3.2	94.7
Mal terminación	Aparición de rebabas	2	2.1	96.8
Incompleto	Falta alguno de los insertos metálicos	2	2.1	98.9
Desbalance	El accesorio requiere contrapesos adicionales	1	1.1	100
Otros	Otros defectos	0	0	100
TOTAL		94	100	

La tercera columna muestra el número de accesorios que presentaban cada tipo de defecto, es decir, la frecuencia con que se presenta cada defecto (n_i). En lugar de la frecuencia numérica (n_i) podemos utilizar la frecuencia porcentual, es decir, el porcentaje de accesorios en cada tipo de defecto, lo cual se indica en la cuarta columna. En la última columna vamos acumulando los porcentajes.

Para hacer más evidente los defectos que aparecen con mayor frecuencia hemos ordenado los datos de la tabla en orden decreciente de frecuencia.

Vemos que la categoría "otros" siempre debe ir al final, sin importar su valor. De esta manera, si hubiese tenido un valor más alto, igual debería haberse ubicado en la última fila.

Podemos ahora representar los datos en un histograma como el siguiente:



Ahora resulta evidente cuáles son los tipos de defectos más frecuentes. Podemos observar que los 2 primeros tipos de defectos se presentan en el 79,8 % de los accesorios con fallas. Por el Principio de Pareto, concluimos que: La mayor parte de los defectos encontrados en el lote pertenece sólo a 2 tipos de defectos (los "pocos vitales"), de manera que si se eliminan las causas que los provocan desaparecería la mayor parte de los defectos.

Otro análisis complementario y sumamente útil e interesante, es calcular los costos de cada problema, con lo cual podríamos construir un diagrama similar a partir de ordenar las causas por sus costos.

Este análisis combinado de causas y costos permite obtener la mayor efectividad en la solución de problemas, aplicando recursos en aquellos temas que son relevantes y alcanzando una mejora significativa.

Recuerde Que No Es Sólo Una Herramienta Para Manufactura. Los Problemas Se Presentan Por Igual En Servicios.

4.3.3.2. Diagrama De Tallo Y Hoja

Es una técnica estadística para representar un conjunto de datos. Cada valor numérico u observación, se divide en dos partes, una que se llamará el tallo de la observación y la otra, la hoja. Los tallos están colocados a lo largo del eje vertical, y las hojas de cada observación a lo largo del eje horizontal.

Ejemplo:

La siguiente distribución de frecuencia muestra el número de anuncios comerciales pagados por 45 empresas durante el año anterior. Observemos que 7 de las 45 empresas pagaron entre 90 y 99 anuncios (pero menos de 100). Sin embargo, ¿El número de comerciantes pagados en esta clase se agrupan en alrededor de 90, están dispersos a lo largo de toda clase, o se acumulan alrededor de 99? No podemos saberlo.



Nº de Anuncios Comprados	Frecuencia
80 - 90	2
90 - 100	7
100 - 110	6
110 - 120	9
120 - 130	8
130 - 140	7
140 - 150	3
150 - 160	3
Total	45

En el ejemplo anterior no podemos identificar los valores de la clase de 90 a 100, sólo conocemos que hay 2. Para ilustrar la construcción de un diagrama de tallo y hojas usando el número de comerciales comprados, supongamos que las 7 observaciones en la clase de 90 a 100 sean 96, 94, 93, 94, 95, 96, 97. El valor de tallo es el dígito o dígitos principales, en este caso el 9. Las hojas son los dígitos secundarios. El tallo se coloca a la izquierda de una línea vertical y los valores de las hojas a la derecha.

Los valores de las clases de 90 a 100, aparecerían como sigue:

9 | 6 4 3 4 5 6 7

Por último, ordenamos los valores dentro de cada tallo de menor a mayor. El segundo renglón del diagrama de tallo y hojas aparecería como sigue:

9 | 3 4 4 5 6 6 7

Con el diagrama de tallo y hojas podemos observar rápidamente que hubo 2 empresas que compraron 94 comerciales y que el número de anuncios comprados fue desde 93 hasta 97. Un diagrama de tallo y hojas es semejante a una distribución de frecuencia, pero con más información, esto es, valores de datos en lugar de marcas.

Veamos otro ejemplo:

En la siguiente tabla se observan las evaluaciones matemáticas para 20 Universidades durante una fecha determinada:

Universidad	Calificación	Universidad	Calificación
1	82,45	11	89,26
2	72,81	12	65,70
3	84,84	13	70,62
4	75,19	14	81,02
5	75,10	15	78,08
6	70,03	16	82,12
7	81,09	17	73,64
8	76,98	18	75,61
9	83,74	19	73,41
10	68,90	20	90,38

Así, el tallo y la hoja de la calificación 82,45 son:



Tallo	Hoja
8	2,45

De manera similar, el tallo y hoja de la calificación 72,81 son:

Tallo	Hoja
7	2,81

La representación de tallo y hoja para las 20 calificaciones sería la siguiente:

6	5,70; 8,90
7	0,03; 0,62; 2,81; 3,41; 3,64; 5,10; 5,19; 5,61; 6,98; 8,08
8	1,02; 1,09; 2,12; 2,45; 3,74; 4,84; 9,26
9	0,38

La columna a la izquierda, contiene los tallos, puestos del menor (al principio de la columna) hasta el mayor. Las hojas de cada tallo aparecen después en el renglón correspondiente en orden ascendente de magnitud, de izquierda a derecha. Así el tallo 6, contiene 2 hojas (5,70 y 8,90). De manera semejante el renglón del tallo 7 contiene 10 hojas, desde 0,03 (la menor) hasta 8,08 (la mayor).

Podemos observar ahora, que si la representación se coloca con el tallo en la base, los listados conforman un histograma de frecuencias. En particular, este histograma contiene cuatro clases, de 60 a 70, de 70 a 80, de 80 a 90 y de 90 a 100, con 2 calificaciones en 60; 10 en 70, 7 en 80, y 1 en 90. Luego, podemos decir que la representación de tallo y hoja muestra cómo se distribuyen los datos en clases y presenta una distribución que es muy similar a un histograma de frecuencias, con la ventaja de que permite reconstruir y conservar el conjunto original de los datos, y ubicarlos en orden de magnitud. Es decir, que no solo podemos visualizar que en la primera clase hay dos observaciones, sino también los valores que asumieron: 65,70 y 68,90.

Podemos cambiar el ancho de clase, redefiniendo de la siguiente manera: Para 82,45 el tallo y la hoja son Tallo: 82 Hoja: 0,45. Esta definición para el tallo y la hoja contendrá demasiados tallos (clases), considerando que la menor calificación es 65,70 y la mayor 90,38, e implicará números de tallo de 65 a 90, para un conjunto de datos tan pequeño.

Resumiendo, una representación de tallo y hoja se construye fácilmente y produce el mismo tipo de figuras que una distribución de frecuencias relativas. Además, permite al observador reconstruir y conservar el conjunto de datos e identificar también observaciones ordenadas (como la quinta, de mayor a menor).

Hay tres desventajas:

- 1- Es apropiada para describir un conjunto pequeño de datos. (Si el número de observaciones es grande, el conjunto de hojas en un renglón de tallo puede salirse del papel).
- 2- Existe poca flexibilidad en la elección de número de tallos, o clases, en la representación.
- 3- No permite una lectura rápida de la frecuencia relativa de clase.



CAPITULO N° 3: Descripción de Datos Estadísticos

Objetivos Específicos

Que el Estudiante:

Conozca las herramientas de descripción de datos en términos de posición

Identifique las medidas de posición más frecuentes como síntesis de la información, analizando su adecuación y sentido según los tipos de variables y la forma de distribución de frecuencia

Utilice las medidas de posición para realizar un análisis de la asimetría de la distribución

Conozca las herramientas de descripción de datos en términos de dispersión

Identifique el sentido de las medidas de dispersión y la diferencia entre ellas, a los fines de seleccionar la más apropiada para describir la variación de los datos y obtener conclusiones adecuadas

Utilice las medidas de dispersión para realizar un análisis de la puntiagudez de la curva

Valore la utilidad de estas herramientas para caracterizar un conjunto de datos



Contenidos Parámetros y Estadísticos

1. Medidas de posición

1.1. **Media aritmética:**

Simbología. Definición. Formas de cálculo para series simples y distribuciones de frecuencias en lista y en intervalos. Interpretación. Propiedades.

1.2. **Mediana:**

Simbología. Definición. Formas de cálculo para series simples y distribuciones de frecuencias en lista y en intervalos. Interpretación. Propiedad.

1.3. **Moda, modo o valor modal:**

Simbología. Definición. Formas de cálculo para series simples y distribuciones de frecuencias en lista y en intervalos. Interpretación.

1.4. **Fractiles:**

Cuartiles, deciles y percentiles simbología. Definición. Formas de cálculo para series simples y distribuciones de frecuencias en lista y en intervalos. Interpretaciones.

2. Medidas de asimetría

Simetría. Asimetría positiva o derecha. Asimetría negativa o izquierda.

2. Medidas de dispersión o de concentración

3.1. **Recorrido:**

Simbología. Definición. Formas de cálculo para series simples y distribuciones de frecuencias en lista y en intervalos.

3.2. **Desviación media:**

Simbología. Definición. Formas de cálculo para series simples y distribuciones de frecuencias en lista y en intervalos.

3.3. **Varianza:**

Simbología. Definición. Formas de cálculo para series simples y distribuciones de frecuencias en lista y en intervalos. Propiedades.

3.4. **Desviación estándar:**

Simbología. Definición. Formas de cálculo para series simples y distribuciones de frecuencias en lista y en intervalos. Propiedades.

3.5. **Coeficiente de variación:**

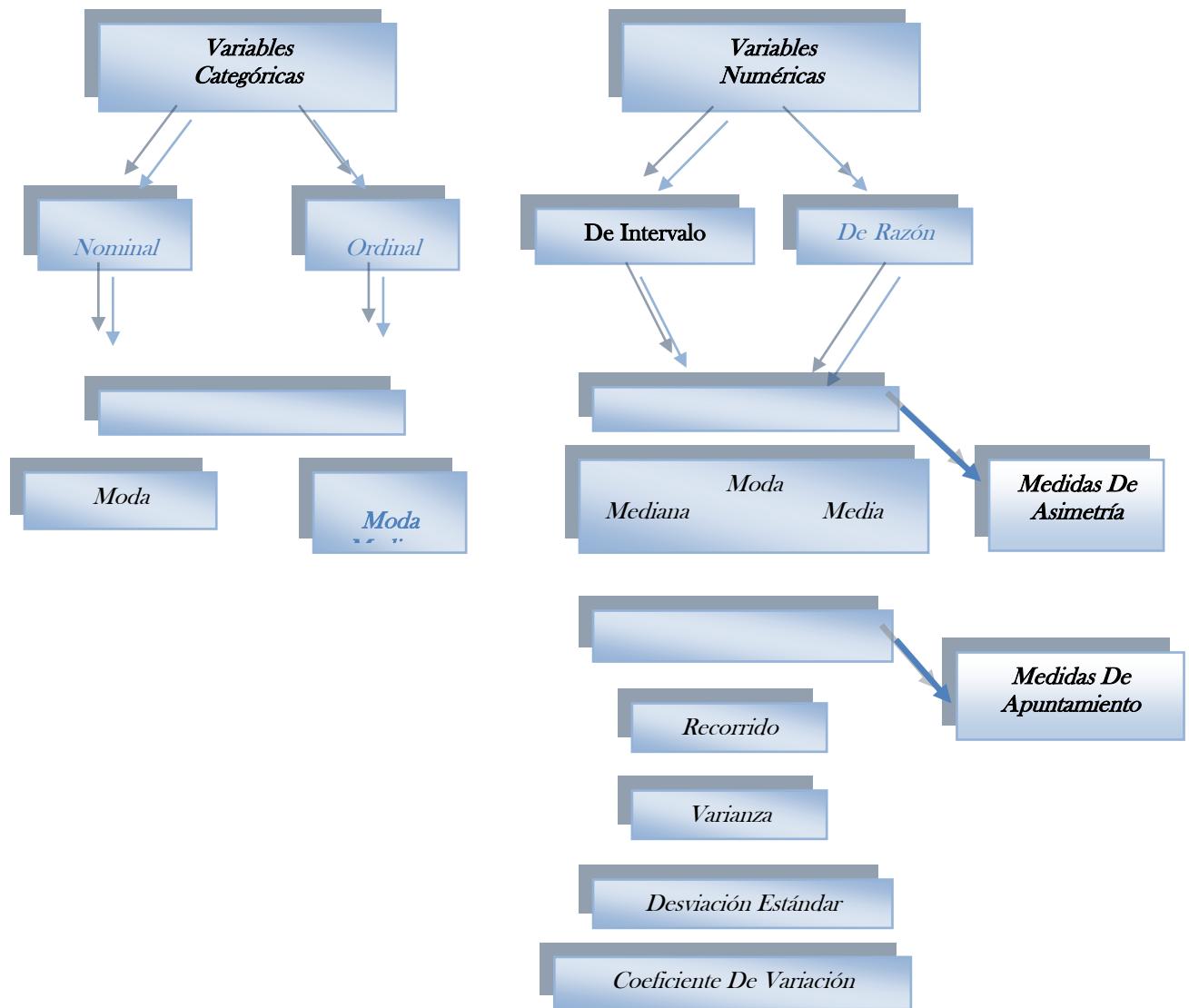
Simbología. Definición. Formas de cálculo para series simples y distribuciones de frecuencias en lista y en intervalos.

4. Medidas de apuntamiento o puntiagudez

Platicúrtica. Mesocúrtica. Leptocúrtica.



DESCRIPCIÓN DE DATOS ESTADÍSTICOS





PARÁMETROS Y ESTADÍSTICOS

Se expuso cómo podían ser reducidos los datos muestrales o poblacionales a forma compacta, comprensible y comunicable a través de las distribuciones de frecuencias. Esto no es sólo un método para organizar datos, sino también una medida descriptiva del modelo de distribución de una variable. Realmente, puede ser considerada como un conjunto de medidas descriptivas, porque cada número que muestra la frecuencia (o densidad) de observaciones de una clase, es una estadística. Pero a menudo se necesitan medidas descriptivas en forma de números que pueden concentrar mejor la atención en varias propiedades de un conjunto de datos que se investiga.

Si estas medidas de resumen descriptivas se calculan con una *Muestra* de datos se llaman *Estadísticos, Estadísticas o Estadígrafos*. Si estas medidas descriptivas se calculan a partir de toda una *Población* de datos, se llaman *Parámetros*.

Estadísticas y parámetros son calculados con las mismas ecuaciones, por lo tanto convendremos:

- 1- Las variables que corresponden a una *Población* se representarán por mayúsculas, tales como X, Y, Z; y las que corresponden a una muestra por letras minúsculas, tales como x, y, z.
- 2- Los tamaños de *Población* Y *Muestra* se representarán por N y n, respectivamente.
- 3- Los parámetros se representarán con letras mayúsculas, tales como \bar{X} o bien M(X), para la media de población; mientras que las estadísticas por letras minúsculas, tales como \bar{x} o bien M(x), para la media de la muestra.

En Resumen

Concepto	Población	Muestra
<i>Variables</i>	X, Y, Z	x, y, z
<i>Cantidad de Observaciones</i>	N	n
<i>Media</i>	$\bar{X} = M(X)$	$\bar{x} = M(x)$

Como por lo general, trabajaremos con muestras, la preferencia será por los estadísticos.

En términos de análisis estadístico y de aplicación, nos interesan cuatro propiedades básicas, que a menudo son suficientes para caracterizar las distribuciones de frecuencias unidimensionales.



Medidas De Posición

Se Refieren Al Valor De La Variable Que Representa Al Conjunto De Datos.

Población: Parámetros De Posición

Medidas De Dispersion Ó De Concentración

Se Refieren Al Grado De Variación De Los Valores Individuales Alrededor Del Centro De La Distribución, o A La Tendencia De Los Valores Individuales A Desviarse De La Medida De Tendencia Central.

Población: Parámetros De Dispersion

Medidas De Asimetría

Se Refieren A La Falta De Asimetría De Ambos Lados Del Pico, Es Decir, Del Punto Con Más Alta Densidad De Frecuencia, De Una Distribución.

Población: Parámetros De Asimetría

Medidas De Puntiagudez

Se Refieren Al Grado De Variación, o La Velocidad Con Que Sube y Baja La



Desarrollaremos seguidamente, cada una de estas medidas, para el caso de trabajar con muestras, recordando que para la población tendrán el mismo tratamiento, sólo que las simbolizaremos con letras mayúsculas.

1. Medidas de Posición

Son valores típicos, en el sentido que se emplean para representar a todos los valores individuales de una serie o de una variable.

No pasan de ser un valor más de la variable, por lo tanto, tendrán las mismas dimensiones que ella.

Deberán tenerse en cuenta los siguientes aspectos:

- a) Deben definirse rigurosamente y no ser susceptibles de diferentes interpretaciones.
- b) Debe basarse en todas las observaciones, de lo contrario no sería una característica de toda la distribución.
- c) Que sea claro y sencillo en su estructura.
- d) Que pueda calcularse con facilidad y rapidez.
- e) Que esté influenciado lo menos posible por “fluctuaciones muestrales”.
- f) Se preste fácilmente al cálculo algebraico.

1.1. Media Aritmética

Indica la localización del centro de la distribución, o la medida de tendencia central.

Definición

Se define y calcula dividiendo la suma de los valores de la variable por el número de observaciones.

Simbología

Población	$M(X) = \bar{X}$
Muestra	$M(x) = \bar{x}$

Formas de Cálculo

a) Series Simples o Datos No Agrupados

$$M(x) = \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

Ejemplo:

Utilizaremos el mismo ejemplo dado para series simples, es decir que consideraremos el número de hijos en 10 familias observadas.

$$n = 10$$

x_i = número de hijos por familia

$$x_1 = 2; x_2 = 1; x_3 = 3; x_4 = 1; x_5 = 2; x_6 = 1; x_7 = 3; x_8 = 0; x_9 = 2; x_{10} = 1$$

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{x_1 + x_2 + \dots + x_{10}}{10} = \frac{2+1+3+1+2+1+3+0+2+1}{10} = \frac{16}{10} = 1,6$$

$$\bar{x} = 1,6 \text{ hijos}$$

*b) Datos Agrupados**b.1) Distribuciones de Frecuencias en Lista*

$$M(y) = \bar{y} = \frac{y_1 n_1 + y_2 n_2 + \dots + y_m n_m}{n} = \frac{\sum_{i=1}^m y_i n_i}{n}$$

O bien, teniendo en cuenta que $\frac{n_i}{n} = h_i$, o sea considerando la frecuencia relativa:

$$M(y) = \bar{y} = y_1 h_1 + y_2 h_2 + \dots + y_m h_m = \sum_{i=1}^m y_i h_i$$

Ejemplo:

Considerando el ejemplo anterior, podemos construir la siguiente tabla:

y	n	h	y n	y h
0	1	0,10	0x1= 0	0x0,10= 0
1	4	0,40	1x4= 4	1x0,40= 0,40
2	3	0,30	2x3= 6	2x0,30=0,60
3	2	0,20	3x2= 6	3x0,20=0,60
	10	1	16	1,60

Utilizando cualquiera de las fórmulas dadas:

$$M(y) = \bar{y} = \frac{y_1 n_1 + y_2 n_2 + y_3 n_3 + y_4 n_4}{n} = \frac{\sum_{i=1}^4 y_i n_i}{n} = \frac{16}{10} = 1,6 \text{ hijos}$$

$$M(y) = \bar{y} = y_1 h_1 + y_2 h_2 + y_3 h_3 + y_4 h_4 = \sum_{i=1}^4 y_i h_i = 1,6 \text{ hijos}$$



Los resultados que se obtienen con a) y b.1) son idénticos, lo cual es lógico ya que en el caso de la tabla de distribución de frecuencias, sólo hemos agrupado a aquellos valores que se repiten. Luego, a) es un caso particular, donde n_i es igual a 1.

b.2) Distribuciones de Frecuencias de Variables Continuas

$$M(y) = \bar{y} = \frac{y_1 n_1 + y_2 n_2 + \dots + y_m n_m}{n} = \frac{\sum_{i=1}^m y_i n_i}{n}$$

O bien, teniendo en cuenta que $\frac{n_i}{n} = h_i$, o sea considerando la frecuencia relativa:

$$M(y) = \bar{y} = y_1 h_1 + y_2 h_2 + \dots + y_m h_m = \sum_{i=1}^m y_i h_i$$

Donde y_i es ahora, la marca de clase o punto medio.

Ejemplo:

Trabajaremos con el caso planteado en distribuciones de frecuencias en intervalos, es decir la edad de un grupo de 20 personas.

$y_{i-1} - y_i$	y_i	n_i	h_i	$y_i n_i$	$y_i h_i$
45-55	50	2	0,10	100	5
55-65	60	4	0,20	240	12
65-75	70	7	0,35	490	24,5
75-85	80	4	0,20	320	16
85-95	90	3	0,15	270	13,5
		20	1,00	1420	71

Entonces:

$$M(y) = \bar{y} = \frac{y_1 n_1 + y_2 n_2 + y_3 n_3 + y_4 n_4 + y_5 n_5}{n} = \frac{\sum_{i=1}^5 y_i n_i}{n} = \frac{1420}{20} = 71 \text{ años}$$

$$M(y) = \bar{y} = y_1 h_1 + y_2 h_2 + y_3 h_3 + y_4 h_4 + y_5 h_5 = \sum_{i=1}^5 y_i h_i = 71 \text{ años}$$

$$\bar{y} = 71 \text{ años}$$

Interpretación

Es un promedio de los valores observados

Así, en el ejemplo del número de hijos en 10 familias observadas, concluimos que hay en promedio 1,6 hijos por familia (entre 1 y 2 hijos por familia).

Luego, en el ejemplo de las edades en grupo de 20 personas, la edad promedio es de 71 años.



Propiedades

1- La suma de las desviaciones con respecto a la media aritmética, es igual a 0, cualquiera sea la distribución. Cuando decimos sumas, es necesario considerar las desviaciones tantas veces como se presenten.

$$\sum_{i=1}^m (y_i - \bar{y}) n_i = 0$$

Ejemplo

y_i	n_i	$y_i - \bar{y}$	$(y_i - \bar{y}) n_i$
0	1	-1,60	-1,60
1	4	-0,60	-2,40
2	3	0,40	1,20
3	2	1,40	2,80
	10		$\sum_{i=1}^m (y_i - \bar{y}) n_i = 0$

2- La suma de los cuadrados de las desviaciones es mínima cuando dichas desviaciones son obtenidas con respecto a la media aritmética.

$$\sum_{i=1}^m (y_i - \bar{y})^2 n_i < \sum_{i=1}^m (y_i - a)^2 n_i$$

Donde a es cualquier valor distinto de \bar{y} .

Ejemplo:

Consideraremos $a = 1$ y $\bar{y} = 1,6$

y_i	n_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2 n_i$	$y_i - a$	$(y_i - a)^2 n_i$
0	1	-1,60	2,56	-1	1
1	4	-0,60	1,44	0	0
2	3	0,40	0,48	1	3
3	2	1,40	3,92	2	8
	10		$\sum_{i=1}^m (y_i - \bar{y})^2 n_i = 8,40$		$\sum_{i=1}^m (y_i - a)^2 n_i = 12$

3- La media de una muestra es igual a la media ponderada de las sub muestras, siendo las ponderaciones los tamaños de dichas sub muestras.

$$\bar{y} = \frac{\bar{y}_{(1)} n_{(1)} + \bar{y}_{(2)} n_{(2)} + \bar{y}_{(3)} n_{(3)} + \dots + \bar{y}_{(k)} n_{(k)}}{n}$$

Ejemplo:



y_i	n_i
0	1
1	4
2	3
3	2
	10

Donde $\bar{y} = 1,6$

Supongamos, además:

$$n_{(1)} = 4$$

$$y_1 = 0; y_2 = 1; y_3 = 1; y_4 = 2$$

$$n_{(2)} = 6$$

$$y_1 = 1; y_2 = 1; y_3 = 2; y_4 = 2; y_{(5)} = 3; y_{(6)} = 3$$

En base a ello, calculamos:

$$\bar{y}_{(1)} = \frac{0+1+1+2}{4} = 1$$

$$\bar{y}_{(2)} = \frac{1+1+2+3+3}{6} = \frac{12}{6} = 2$$

Entonces:

$$\bar{y} = \frac{\bar{y}_{(1)} n_{(1)} + \bar{y}_{(2)} n_{(2)}}{n} = \frac{1(4) + 2(6)}{10} = \frac{4+12}{10} = \frac{16}{10} = 1,6$$

Otra aplicación de esta propiedad podemos encontrarla, considerando una Carrera de 5 años de duración, en donde contamos con los Promedios de notas por año, de materias aprobadas. ¿Cuál es el promedio de la Carrera?

<i>Años de una carrera</i>	<i>Promedios anuales de notas de materias aprobadas</i> \bar{y}_i	<i>Cantidad de materias aprobadas</i> n_i	$\bar{y}_i n_i$
1º	7,54	6	45,24
2º	7,98	7	55,86
3º	8,45	8	67,6
4º	9,09	10	90,9
5º	9,45	12	113,4
		43	373

Luego: $\bar{y} = \frac{373}{43} = 8,67$

4- La media aritmética de una constante es igual a dicha constante.

$$M(k) = k$$

Si $y_i = k \rightarrow \bar{y} = k$
 $i=1, 2, \dots, n$

Ejemplo:



$$y_i = 6,6,6,6,6$$

y_i	n_i	$y_i n_i$
6	5	30
		30

$$\bar{y} = \frac{\sum_{i=1}^5 y_i n_i}{n} = \frac{30}{5} = 6$$

5- La media aritmética del producto de una constante por una variable es igual al producto de la constante por la media aritmética de la variable.

$$M(yk) = k \bar{y}$$

Ejemplo:

y_i <i>Salarios</i>	n_i	$y_i n_i$
200	4	800
300	4	1200
400	2	800
	10	2800

Calculamos: $\bar{y} = \frac{\sum_{i=1}^3 y_i n_i}{n} = \frac{2800}{10} = 280$

Luego, se incrementan al doble todos los sueldos, entonces definimos $\bar{y}' = 2\bar{y}_i$

$\bar{y}' = 2\bar{y}_i$	n_i	$\bar{y}' n_i$
400	4	1600
600	4	2400
800	2	1600
	10	5600

Y calculamos: $\bar{y}' = \frac{\sum_{i=1}^3 \bar{y}' n_i}{n} = \frac{5600}{10} = 560$

Entonces la media de esta nueva variable es igual a 560, resultado al que arribamos por aplicación de la propiedad:

$$M(ky) = M(2y) = 2M(Y) = 2 \times 280 = 560$$

6- La media aritmética de la suma algebraica de dos o más variables es igual a la suma algebraica de las medias aritméticas de dichas variables:

$$M \left[\sum_{j=1}^k x_{(j)i} \right] = \sum_{j=1}^k M(x_j)_i$$

Nota: no es extensible al multiplicatorio.

Ejemplo:

$x_{(1)i}$	1 2 3 2	$M[x_{(1)i}] = \frac{\sum_{i=1}^4 x_{(1)i}}{4} = \frac{1+2+3+2}{4} = 2$
$x_{(2)i}$	1 2 2 3	$M[x_{(2)i}] = \frac{\sum_{i=1}^4 x_{(2)i}}{4} = \frac{1+2+2+3}{4} = 2$
$x_{(3)i}$	2 1 3 2	$M[x_{(3)i}] = \frac{\sum_{i=1}^4 x_{(3)i}}{4} = \frac{2+1+3+2}{4} = 2$
$\sum_{j=1}^3 x_{(j)i}$	4 5 8 7	$\sum_{j=1}^3 x_{(j)i} = 6$

$$M\left[\sum_{j=1}^3 x_{(j)i}\right] = \frac{\sum_{i=1}^4 \left[\sum_{j=1}^3 x_{(j)i} \right]}{4} = \frac{4+5+8+7}{4} = \frac{24}{4} = 6$$

7- Se puede utilizar para estimar una cantidad total en una población. Por ejemplo, la tarifa media de salario por hora para una muestra de $n=6$ secretarias ejecutivas es de \$8,00. Si en esta compañía, hay 200 secretarias, el costo total por hora de la mano de obra secretarial se calcula como:

$$\sum x = N \bar{x} \quad \text{Donde } N: \text{Tamaño de la Población y } \bar{x} : \text{media aritmética de la muestra}$$

$$\sum x = 200 \times 8 = 1.600$$

Llamamos "Total" a la expresión $\sum x$

8- Su cálculo se base en cada observación, por lo tanto la media se ve afectada por cualquier valor o valores extremos.

Ejemplo:

$$y_1 = 2; y_2 = 7; y_3 = 8; y_4 = 7; y_{(5)} = 9$$

$$\bar{y} = \frac{2+7+8+7+9}{5} = 6,6$$

Se observa, que un solo valor extremadamente pequeño da por resultado una gran reducción en \bar{y} . Por supuesto, ocurriría el caso inverso si el valor extremo fuera mucho mayor que las otras observaciones.



1.2. Mediana

Definición

Es el valor central de la variable cuando los datos están ordenados por su magnitud, siendo indistinto que el ordenamiento sea ascendente o descendente.

Dicho de otra manera, es el valor de la variable que supera a no más de la mitad de las observaciones y es superado por no más de la mitad de las observaciones.

Simbología

Población	M_c
Muestra	me

Presenta los siguientes inconvenientes:

- Es necesario ordenar previamente los elementos por su magnitud, tarea que suele insumir mucho tiempo.
- Está influenciada solamente por el número de elementos, no por el valor de los mismos.
- Se usa generalmente cuando se trata de atributos.

Formas de Cálculo

a) Series Simples o Datos No Agrupados

1. Ordenar los datos por su magnitud de menor a mayor o viceversa.
2. Encontrar el orden de la mediana, o sea su posición en la serie, a través de:

$$\left(\frac{n+1}{2}\right)^o$$

Se pueden presentar dos casos, en función de si la cantidad de datos es impar o par, entonces:

a.1) Número de Datos Impar

Ejemplo:

$$y_i = 1, 4, 10, 8, 10$$

1. Ordenamos por su magnitud, en este caso de menor a mayor: 1, 4, 8, 10, 10
2. Establecemos la posición que en la serie tendrá el valor de la mediana:

$$\left(\frac{n+1}{2}\right)^o = \left(\frac{5+1}{2}\right)^o = (3)^o$$

Entonces, la mediana es el valor de la variable que ocupa el tercer lugar, luego:

$$me = 8$$



Vemos que hay dos elementos (1,4) antes de la mediana y dos elementos (10,10) después de la mediana.

Ejemplo:

$$y_i = 1, 2, 3, 4, 5, 5, 5, 6, 7, 8, 9$$

$$\text{Entonces: } \left(\frac{n+1}{2}\right)^{\circ} = \left(\frac{11+1}{2}\right)^{\circ} = (6)^{\circ} \Rightarrow me = 5$$

Puede que la mediana tenga valores idénticos al suyo, a ambos lados. Con estas características se puede definir a la mediana como “Aquel valor de la variable que divide a la serie de tal forma que por lo menos el 50% de los valores sean menores o iguales a él, y por lo menos el 50% de los valores sean mayores o iguales a él”.

a.2) Número de Datos Par

No existe una mediana verdadera, ya que existen dos valores centrales de la variable que cumplen con la condición de la definición.

Por ello, se conviene que la mediana es igual al promedio de los dos valores centrales de la variable, luego de ordenarlos por su magnitud.

Ejemplo

$$y_i = 9, 6, 2, 5, 18, 12$$

- 1- Ordenamos: 2, 5, 6, 9, 12, 18
- 2- Determinamos el orden de la mediana

$$\left(\frac{n+1}{2}\right)^{\circ} = \left(\frac{6+1}{2}\right)^{\circ} = (3,5)^{\circ} \Rightarrow me = \frac{6+9}{2} = \frac{15}{2} = 7,5$$

b) Datos Agrupados

b.1) Distribución de Frecuencias En Lista

Vamos a llamar N_j a la primer frecuencia absoluta acumulada que supere a $\frac{n}{2}$.

De tal manera que $N_{j-1} \leq \frac{n}{2} \leq N_j$.

Entonces, pueden presentarse dos casos:



$$\text{b.1.1)} \quad N_{j-1} < \frac{n}{2} \Rightarrow me = y_j$$

Donde y_j es el valor de y_i , que corresponde a la primer frecuencia absoluta acumulada

que supera a $\frac{n}{2}$ (N_j)

Ejemplo

	y_i	n_i	N_i	
	0	2	2	
y_{j-1}	1	4	6	N_{j-1}
y_j	2	6	12	N_j
	3	3	15	
	4	5	20	
		20		

$$\frac{n}{2} = \frac{20}{2} = 10 \Rightarrow N_{j-1} < \frac{n}{2} \Rightarrow me = y_j = 2$$

Si transformamos los datos en una serie simple, y obtenemos el orden de la mediana, según vimos, $\left(\frac{n+1}{2}\right)^0 = \left(\frac{21}{2}\right)^0 = (10,5)^0$

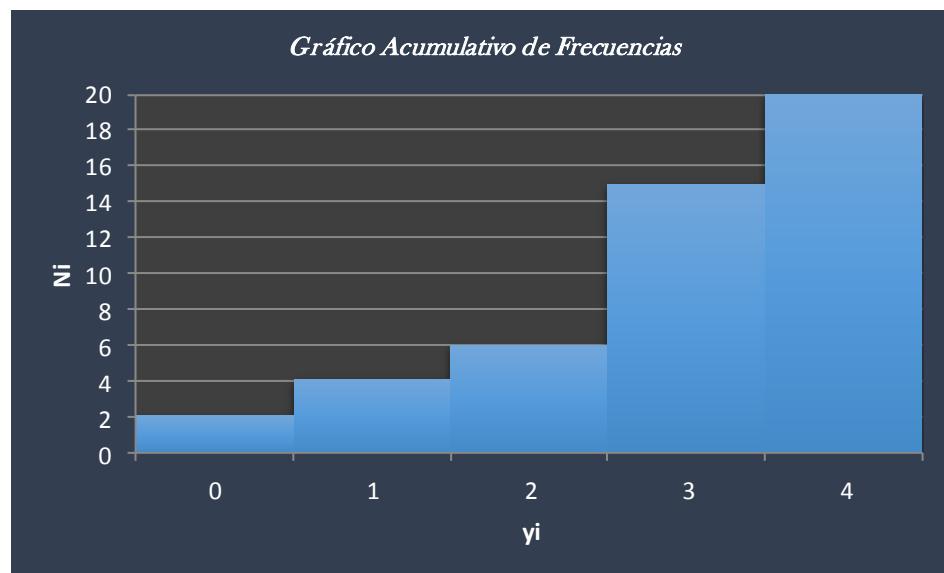
La mediana será el valor de la variable que ocupa el lugar 10,5, por lo tanto resulta del promedio de los elementos centrales de la serie.

00 1111 2222↓ 22 333 44444

$$me = \frac{2+2}{2} = 2$$

Como puede observarse es idéntico al encontrado a través de la Distribución de Frecuencias.

Este valor, puede también determinarse gráficamente, para lo cual debe utilizarse el Gráfico Acumulativo de Frecuencias, ya explicado. Una vez construido, se identifican los valores de N_j , N_{j-1} , y $\frac{n}{2}$, en la ordenada, y se traza una horizontal a la altura de $\frac{n}{2}$, hasta intersectar el polígono y bajando la vertical, se obtiene el valor de la mediana, que en nuestro caso es 2.



$$\text{b.1.2)} \quad N_{j-1} = \frac{n}{2} \quad \Rightarrow \quad me = \frac{y_{j-1} + y_j}{2}$$

Donde todos los valores del intervalo cerrado $y_{j-1} - y_j$ son valores que corresponden a la definición de mediana.

Ejemplo

y _i	n _i	N _i		
0	2	2		
1	4	6		
y _{j-1}	2	4	10	N _{j-1} = $\frac{n}{2}$
y _j	3	7	17	N _j
	4	3	20	
		20		

$$\frac{n}{2} = \frac{20}{2} = 10 \Rightarrow N_{j-1} = \frac{n}{2} \Rightarrow me = \frac{y_{j-1} + y_j}{2} = \frac{2+3}{2} = 2,5$$

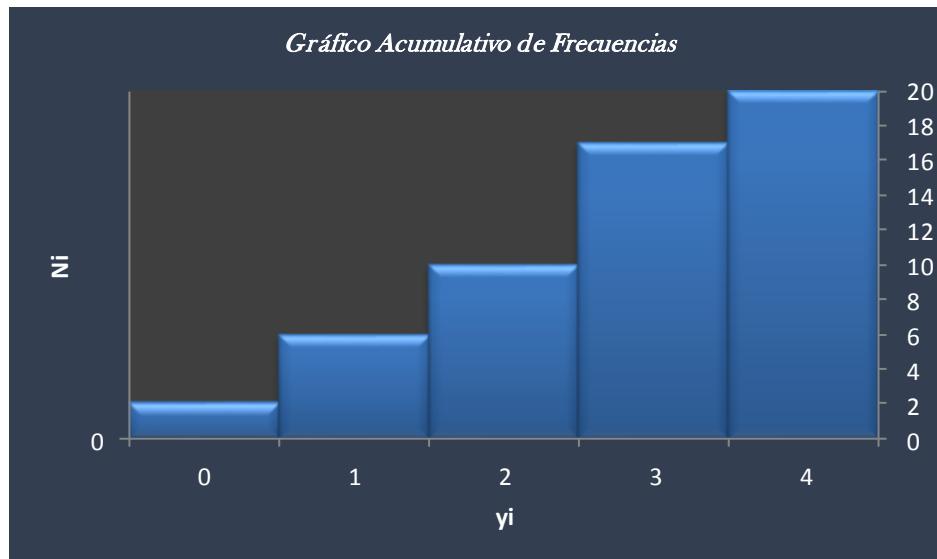
Si trabajamos con serie simple, el orden de la misma es $\left(\frac{n+1}{2}\right)^0 = \left(\frac{21}{2}\right)^0 = (10,5)^0$

00 1111 2222 ↓ 3333333 444

Entonces $me = \frac{2+3}{2} = 2,5$, igual que antes.



Gráficamente, también se utiliza el Gráfico Acumulativo de Frecuencias, con igual procedimiento que el aplicado en el caso anterior. De esta manera la $me = 2,5$.



b.2) Distribución De Frecuencias En Intervalos

Pasos a seguir:

- 1- Establecer la columna en la tabla donde se registran las frecuencias absolutas acumuladas.
- 2- Obtener la clase mediana.
- 3- Interpolar la mediana a partir de los valores incluidos en la clase mediana.

Pueden también presentarse dos casos:

$$\text{b.2.1)} \quad N_{j-1} < \frac{n}{2} \quad me = y_{(i-1)_j} + c_j \frac{\frac{n}{2} - N_{j-1}}{n_j}$$

Donde $y_{(i-1)_j}$ es el valor que corresponde al límite inferior del intervalo donde está ubicada la mediana, es decir el intervalo j-ésimo.

Ejemplo:

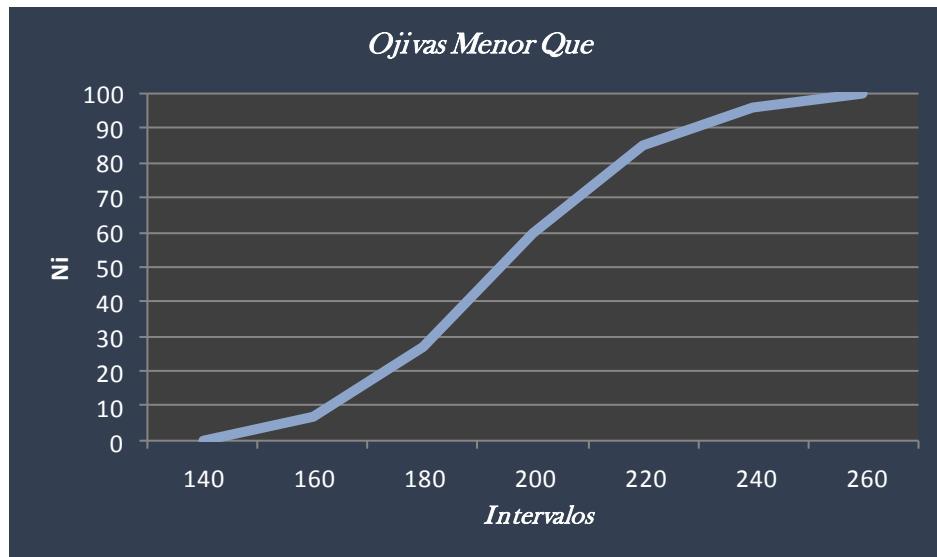
$y_{i-1} - y_i$	n_i	N_i	
140-160	7	7	
160-180	20	27	N_{j-1}
180-200	$33 = n_J$	60	N_J
200-220	25	85	
220-240	11	96	
240-260	4	100	
	100		



Entonces:

$$N_{j-1} \left\langle \frac{n}{2} \right\rangle \Rightarrow me = y_{(i-1)_j} + c_j \frac{\frac{n}{2} - N_{j-1}}{n_j} = 180 + 20 \frac{50 - 27}{33} = 193,94$$

Gráficamente, corresponde el uso de la Ojiva. Se identifica en la ordenada el valor correspondiente a $n/2$, trazando a esta altura una horizontal hasta intersectar la curva, punto donde bajamos una vertical al eje de las x , determinando el valor de la mediana.



b.2.2) $N_{j-1} = \frac{n}{2} \Rightarrow me = y_{(i-1)_j}$, puesto que el segundo término se anula.

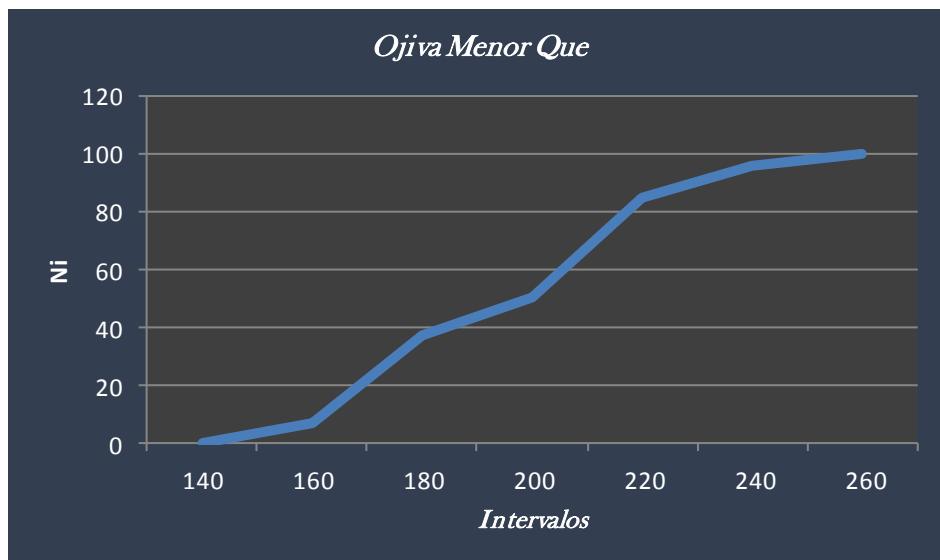
Ejemplo:

$y_{i-1} - y_i$	n_i	N_i	
140 - 160	7	7	
160 - 180	30	37	
180 - 200	13	50	$N_{j-1} = \frac{n}{2}$
200 - 220	35	85	N_j
220 - 240	11	96	
240 - 260	4	100	
	100		

$$me = y_{(i-1)_j} = 200$$



Gráficamente, también se utiliza la Ojiva, con igual procedimiento que el visto anteriormente.



Una ventaja definitiva del uso de la mediana en relación con la media, como medida representativa con datos agrupados, es que se puede tener aproximación a la mediana con la fórmula anterior, sin importar que la distribución de frecuencia tenga o no anchos iguales de intervalos de clase o clases con un extremo abierto.

Interpretación

El 50% de los valores son menores a la mediana y el otro 50% son mayores a la mediana.

Propiedades

La suma de los valores absolutos de las desviaciones con respecto a la mediana es mínima.

$$\left| \sum_{i=1}^m y_i - me \right| n_i < \sum_{i=1}^m |y_i - a| n_i$$

Donde a es un valor cualquiera.

Aplicación: Decisión sobre la localización óptima de una fábrica.



Ejemplo:

Supongamos que una organización en cadena de restaurantes tiene siete de ellos en cierta carretera, como se indica en el cuadro:

(1)	(2)	De A	De B	De C	De D	De E	De F	De G
A	12	0	28	64	84	112	120	136
B	40	28	0	36	56	84	92	108
C	76	64	36	0	20	38	56	72
D	96	84	56	20	0	28	36	52
E	124	112	84	48	28	0	8	24
F	132	120	92	56	36	8	0	16
G	148	136	108	72	52	24	16	0
Totales		544	404	296	276	294	328	408

(1) Restaurante

(2) Distancia En Kilómetros De La Entrada Al Restaurante.

Diariamente, cada restaurante debe buscar alimentos frescos de un almacén central.

Además, supongamos que el servicio requiera dos camiones cada día, entonces para minimizar la distancia total de los siete restaurantes al almacén central, ¿dónde debe estar ubicado éste?

Recordando que las desviaciones absolutas con relación a la mediana es un mínimo, podemos decir fácilmente que el almacén central debe estar situado en el Restaurante “D” o en su vecindad. La distancia mediana del restaurante a la entrada de la carretera es de 96 kms., y del restaurante “D” al recorrido total para reabastecer a cada restaurante una vez, se minimizaría en 276 kms., en cada sentido. Si el almacén central está ubicado en cualquier otro punto, la suma de la distancia será mayor.

La mediana de la serie es “D” = 96 kms., pues se trata de una serie impar, por lo que el valor de variable que ocupa el 4º lugar, corresponde a la mediana.

1.3. Moda, Modo o Valor Modal

Definición

Es el valor más frecuente, más comúnmente observado en un conjunto de datos.

Simbología

Población	Md
Muestra	md

Presenta los siguientes inconvenientes:

- Puede que dos o más valores se repitan un número iguales de veces. En tal situación, no hay forma lógica de determinar qué valor debe ser escogido como la moda.



- Puede que no hallemos ningún valor que aparezca más de una vez.
- Es un valor muy inestable, puede cambiar con el método de redondeo de los datos.

Peso Real	Redondeo hasta el Kg. más próximo	Truncado
3,000	3	3
3,500	4	3
3,571	4	3
3,784	4	3
4,500	5	4
4,831	5	4
6,115	6	6
6,115	6	6
$md = 6,115$	$md = 4$	$md = 3$

Formas de Cálculo

a) Series Simples o Datos No Agrupados

Ejemplos:

$x_i = 1, 4, 10, 8, 10 \quad md = 10$, pues ocurre más veces

$x_i = 1, 3, 3, 7, 7, 8 \quad md = 3 \quad md = 7$, puesto que cada uno ocurre dos veces

$x_i = 1, 2, 4, 9 \quad$ No hay moda, puesto que ninguno de los valores ocurre más de un vez

b) Datos Agrupados

b.1) Distribuciones de Frecuencias en Lista

$$md = y_j \quad Si \ se \ cumple \ que \quad n_{j-1} \langle n_j \rangle n_{j+1}$$

Con lo cual son valores de variable que presentan picos en sus frecuencias.

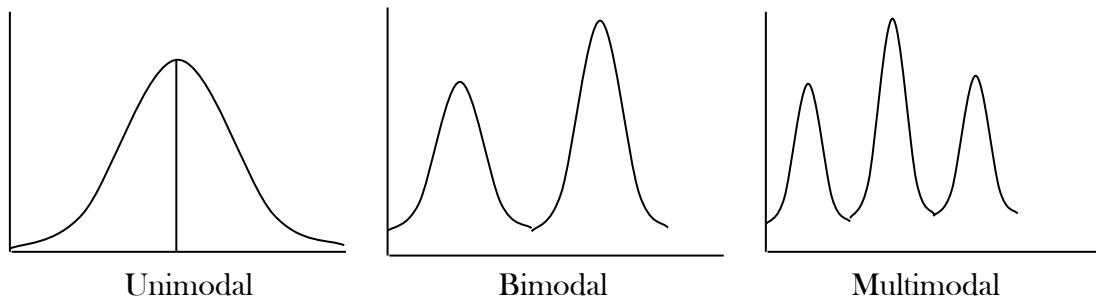
Hablando en sentido riguroso, cualquier valor se llama moda, si aparece más a menudo que cualquiera de los valores adyacentes. Sin embargo, mientras la frecuencia de los valores modales no sea igual, podríamos decidir escoger el valor con la frecuencia más alta como la moda para la serie.

Ejemplo:

y_i	n_i	
0	2	
1	3	
2	6	$n_{j-1} \langle n_j \rangle n_{j+1} \Rightarrow md = 2$
3	5	
4	4	



En una distribución de frecuencias de gran recorrido, pueden presentarse una ó más modas, que corresponderán a un ó más máximos que se presenten, es decir, que pueden existir distribuciones:



Otros ejemplos:

1-

y_i	n_i	
0	2	
1	3	$n_{j-1} < n_j > n_{j+1} \Rightarrow md = 1$
2	1	
3	5	$n_{j-1} < n_j > n_{j+1} \Rightarrow md = 3$
4	2	
5	1	

2-

y_i	n_i	
0	2	
1	2	
2	2	No hay moda

3-

y_i	n_i	
0	2	
1	5	$n_{j-1} < n_j > n_{j+1} \Rightarrow md = 1$
2	2	
3	5	$n_{j-1} < n_j > n_{j+1} \Rightarrow md = 3$
4	3	
5	1	



4-

y_i	n_i	
0	16	$n_{j-1} \langle n_j \rangle n_{j+1} \Rightarrow md = 0$, ya que la frecuencia hacia arriba es 0.
1	5	
2	4	
3	3	

5-

y_i	n_i	
0	4	
1	4	
2	3	
3	5	$n_{j-1} \langle n_j \rangle n_{j+1} \Rightarrow md = 3$
4	1	
5	9	$n_{j-1} \langle n_j \rangle n_{j+1} \Rightarrow md = 5$, ya que la frecuencia hacia abajo es 0.

En los dos últimos casos, la moda puede corresponder a valores extremos, y entonces difícilmente podría ser considerada como medida de tendencia central.

b.2) Distribuciones de Frecuencias en Intervalos

Puede ser obtenida por varios métodos, cada uno de los cuales puede dar un valor diferente de moda:

- b.2.1) Método crudo o de la marca de clase
- b.2.2) Interpolación por método gráfico
- b.2.3) Interpolación mediante fórmula
- b.2.4) Método empírico

Para los datos agrupados, con intervalos iguales, la moda se encuentra en el intervalo j -ésimo de la variable, o sea, $(y_{i-1}^* - y_i^*)_j$, si se cumple que: $n_{j-1} \langle n_j \rangle n_{j+1}$.

Dado que no se define exactamente el valor modal, sino que se determina el intervalo en el cual está ubicado, o sea la clase modal, es necesario plantear algún método de cálculo que nos permita obtener el modo.

b.2.1) Método crudo o de la marca de clase

Utiliza como moda al valor correspondiente al punto medio del intervalo modal. Si $(y_{i-1}^* - y_i^*)_j$ es el intervalo modal, entonces:

$$md = y_j = \frac{(y_{i-1}^*)_j + (y_i^*)_j}{2}$$



En este caso, se considera que la frecuencia de clase se concentra alrededor de la marca de clase.

b.2.2) Interpolación por método gráfico

Se utilizan:

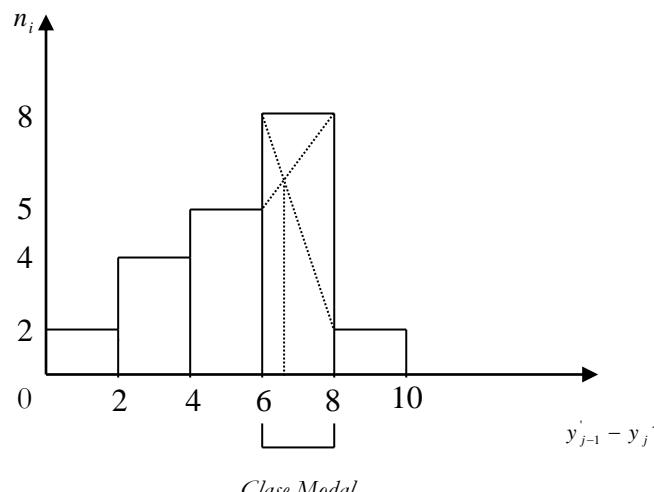
1. Los valores del intervalo modal
2. La frecuencia del intervalo modal
3. Las frecuencias de las clases inmediatas anterior y posterior a la clase modal.

Con estos datos se opera de la siguiente forma:

1. Se construye el Histograma correspondiente a los datos dados.
2. Se trazan dos líneas diagonales que cortan el rectángulo de la clase modal, partiendo de las esquinas superiores de dicho rectángulo a las esquinas superiores de los rectángulos adyacentes.
3. Se traza una línea perpendicular desde la intersección de las diagonales hasta la abscisa y en el punto donde la corta estará ubicada la moda.

Ejemplo:

$y_{i-1} - y_i$	n_i	
Kms. recorridos	Nº de estudiantes	
0-2	2	
2-4	4	
4-6	5	n_{j-1}
6-8	8	n_j
8-10	1	n_{j+1}
	20	



b.2.3) Interpolación mediante fórmula



Se usan también la frecuencia de clase modal y las frecuencias en las clases inmediata anterior y posterior a dicha clase. Los resultados son coincidentes cuando se utilizan los métodos b.2.2) y b.2.3)

Llamemos

$$d_1 = n_j - n_{j-1}$$
$$d_2 = n_j - n_{j+1}$$

Dónde:

- n_j : Frecuencia absoluta del intervalo modal
 n_{j-1} : Frecuencia absoluta anterior al intervalo modal
 n_{j+1} : Frecuencia absoluta posterior al intervalo modal

La fórmula de interpolación, es entonces:

$$md = (y_{i-1})_j + c \frac{d_1}{d_1 + d_2}$$

En el ejemplo dado anteriormente:

$$md = (y_{i-1})_j + c \frac{d_1}{d_1 + d_2} = 6 + 2 \frac{(8-5)}{(8-5)+(8-1)} = 6 + 2 \frac{3}{3+7} = 6,6$$

b.2.4) Método empírico

$$md = \bar{y} - 3(\bar{y} - me)$$

1.4. Fractiles

En una distribución de frecuencias, cierta cantidad de los datos cae en un fractil o por debajo de éste. Hemos definido anteriormente a la Mediana como el valor central de la variable una vez que los datos han sido ordenados por su magnitud, dividiendo de esta manera a la serie en dos partes iguales. Es decir, que la mediana, es el fractil que queda en medio, puesto que la mitad de los datos está por debajo de este valor y la otra mitad está por arriba de él.

Otros fractiles, como los Cuartiles, Deciles o Percentiles, se parecen mucho a la Mediana porque también subdividen una distribución de mediciones de acuerdo con la proporción de frecuencias observadas. Mientras que la Mediana divide a la distribución en dos mitades, los Cuartiles la dividen en cuatro cuartos, los Deciles la dividen en diez décimos y los Percentiles la dividen en cien partes.

Se emplean con más frecuencia en variables continuas, siempre que el número de intervalos sea grande y se desee obtener un promedio que corresponda a una determinada parte de la distribución.



Cuartiles

Los cuartiles de una distribución, son los valores de la variable que dividen a los datos (ordenados de menor a mayor) en cuatro partes iguales.

Para su cálculo procederemos de manera idéntica que para el cálculo de la Mediana, sólo que ahora dividiremos a la serie en cuatro partes iguales, de tal modo que cada parte contenga el mismo número de observaciones, es decir un 25%.

Los simbolizaremos por Q_k para $k = 1, 2, 3$, que indica el número de Cuartil

Símbolo	Nombre	Definición
Q_1	Primer Cuartil	Valor de la variable que supera al 25% de las observaciones y es superado por el 75% restante.
Q_2	Segundo Cuartil Mediana	Valor de la variable que supera al 50% de las observaciones y es superado por el 50% restante.
Q_3	Tercer Cuartil	Valor de la variable que supera al 75% de las observaciones y es superado por el 25% restante.

Formas de cálculo

a) Datos no agrupados

Cuartiles	Orden de cada cuartil	
	Nº de datos Par	Nº de datos Impar
<i>General</i>		
Q_1	$\left[k \left(\frac{n}{4} \right) \right]^\circ$	$\left[k \left(\frac{n+1}{4} \right) \right]^\circ$
<i>K: Número de cuartil: 1, 2, 3</i>		
Q_1	$\left[1 \left(\frac{n}{4} \right) \right]^\circ$	$\left[1 \left(\frac{n+1}{4} \right) \right]^\circ$
Q_2	$\left[2 \left(\frac{n}{4} \right) \right]^\circ = \left(\frac{n}{2} \right)^\circ$	$\left[2 \left(\frac{n+1}{4} \right) \right]^\circ = \left(\frac{n+1}{2} \right)^\circ$
Q_3	$\left[3 \left(\frac{n}{4} \right) \right]^\circ$	$\left[3 \left(\frac{n+1}{4} \right) \right]^\circ$

**Ejemplo 1:**

Supongamos que la altura de 8 árboles, ordenadas de menor a mayor son las siguientes:

$$150 - 151 - 152 - 154 - 155 - 156 - 157 - 159 \text{ cms.}$$

1- Determinación de Q_1

Determinamos el Orden: $\left[1\left(\frac{8}{4} \right) \right]^\circ = 2^\circ$

El primer cuartil, Q_1 , es el valor de la variable que divide a la primera mitad de la serie en dos partes iguales. Luego:

$$Q_1 = \frac{151+152}{2} = 151,5$$

2- Determinación de Q_2

Determinamos el Orden: $\left[2\left(\frac{8}{4} \right) \right]^\circ = 4^\circ$

El segundo cuartil, Q_2 es el valor de la variable que divide a la serie en dos partes iguales. Luego:

$$Q_2 = \frac{154+155}{2} = 154,5$$

Obsérvese que la mediana de este conjunto de datos, es el valor de la variable que ocupa el siguiente el orden:

$$\left(\frac{n+1}{2} \right)^\circ = \left(\frac{8+1}{2} \right)^\circ = 4,5^\circ$$

Por lo tanto es el valor de la variable ubicado entre el 4º y 5º lugar, es decir $Me = 154,5$ cms. Entonces Q_2 es la Mediana.

3- Determinación de Q_3

Determinamos el Orden: $\left[3\left(\frac{8}{4} \right) \right]^\circ = 6^\circ$

El tercer cuartil, Q_3 es el valor de la variable que divide a la segunda mitad en dos partes iguales. Luego:

$$Q_3 = \frac{156+157}{2} = 156,5$$



Resumiendo:

1º Cuarto	2º Cuarto	3º Cuarto	4º Cuarto
150 - 151	152 - 154	155 - 156	157 - 159
$Q_1 = 151,5$	$Q_2 = 154,5$	$Q_3 = 156,5$	Me

Ejemplo 2:

Si ahora consideramos la altura de 9 árboles, ordenados de menor a mayor:

150 - 151 - 152 - 153 - 154 - 155 - 156 - 157 - 159 cms.

1- Determinación de Q_1

Determinamos el Orden: $\left[1 \left(\frac{9+1}{4} \right) \right]^\circ = 2,5^\circ$

El primer cuartil, Q_1 , es el valor de la variable que divide a la primera mitad de la serie en dos partes iguales. Luego:

$$Q_1 = \frac{151+152}{2} = 151,5$$

2- Determinación de Q_2

Determinamos el Orden: $\left[2 \left(\frac{9+1}{4} \right) \right]^\circ = \left(\frac{9+1}{2} \right)^\circ = 5^\circ$

El segundo cuartil, Q_2 es el valor de la variable que divide a la serie en dos partes iguales. Luego:

$$Q_2 = 154$$

Obsérvese que la mediana de este conjunto de datos, es el valor de la variable que ocupa el siguiente el orden:

$$\left(\frac{n+1}{2} \right)^\circ = \left(\frac{9+1}{2} \right)^\circ = 5^\circ$$

Por lo tanto es el valor de la variable ubicado en el 5º lugar, es decir Me = 154 cms. Entonces Q_2 es la Mediana.

3- Determinación de Q_3

Determinamos el Orden: $\left[3 \left(\frac{9+1}{4} \right) \right]^\circ = 7,5^\circ$

El tercer cuartil, Q_3 es el valor de la variable que divide a la segunda mitad en dos



partes iguales. Luego:

$$Q_1 = \frac{156+157}{2} = 156,5$$

Resumiendo:

1º Cuarto	2º Cuarto	3º Cuarto	4º Cuarto
150 - 151	1525 - 153	154	155 - 156
$Q_1 = 151,5$	$Q_2 = 154$	Me	$Q_3 = 156,5$

b) Datos Agrupados

Seguiremos igual procedimiento que para el cálculo de la mediana. Considerando ahora que para ubicar el orden de Q_1 se debe considerar $\frac{n}{4}$ y para ubicar el orden de Q_3 se debe considerar $3\frac{n}{4}$.

b.1) Distribución de Frecuencias en Lista

Cuartiles	Formas de Cálculo	
General	$N_{j-1} < k \frac{n}{4}$	$N_{j-1} = k \frac{n}{4}$
	$Q_k = y_j$	$Q_k = \frac{y_{j-1} + y_j}{2}$
	K: Número de cuartil: 1, 2, 3	
1º Cuartil Q_1	$N_{j-1} < 1 \frac{n}{4}$ $Q_1 = y_j$	$N_{j-1} = 1 \frac{n}{4}$ $Q_1 = \frac{y_{j-1} + y_j}{2}$
2º Cuartil Q_2	$N_{j-1} < 2 \frac{n}{4}$ $Q_2 = y_j$	$N_{j-1} = 2 \frac{n}{4}$ $Q_2 = \frac{y_{j-1} + y_j}{2}$
3º Cuartil Q_3	$N_{j-1} < 3 \frac{n}{4}$ $Q_3 = y_j$	$N_{j-1} = 3 \frac{n}{4}$ $Q_3 = \frac{y_{j-1} + y_j}{2}$

Ejemplo:

Tabla de frecuencias para el Número de hijos de 56 familias observadas

	Número de Hijos	n _i	N _i	Q _i	Q _s
	0	5	5	N _{j-1}	
Q _i	1	11	16	N _j	N _{j-1}
Q _s	2	35	51		N _j
	3	2	53		
	4	2	55		
	5	1	56		
	Total	56			

Cálculo del primer cuartil Q_i

$$\text{Orden: } \frac{n}{4} = \frac{56}{4} = 14$$

$$N_{j-1} = 5; \text{ por lo cual } N_{j-1} < \frac{n}{4} = 14$$

$$Q_i = y_j = 1 \text{ Hijo}$$

Cálculo del tercer cuartil Q_s

$$\text{Orden: } 3\frac{n}{4} = 3\frac{(56)}{4} = 42$$

$$N_{j-1} = 16$$

$$Q_s = y_j = 2 \text{ Hijos}$$



b.2) Distribución de Frecuencias en Intervalos

También son de aplicación las fórmulas de interpolación:

Cuartiles	Formas de Cálculo	
General	$N_{j-1} < k \frac{n}{4}$	$N_{j-1} = k \frac{n}{4}$
	$Q_k = (y'_{i-1})_j + c \frac{k(\frac{n}{4}) - N_{j-1}}{n_j}$	$Q_k = (y'_{i-1})_j$
	$K: \text{Número de cuartil: } 1, 2, 3$	
1°Cuartil Q_1	$N_{j-1} < 1 \frac{n}{4}$	$N_{j-1} = 1 \frac{n}{4}$
	$Q_1 = (y'_{i-1})_j + c \frac{1(\frac{n}{4}) - N_{j-1}}{n_j}$	$Q_1 = (y'_{i-1})_j$
2°Cuartil Q_2	$N_{j-1} < 2 \frac{n}{4}$	$N_{j-1} = 2 \frac{n}{4}$
	$Q_2 = (y'_{i-1})_j + c \frac{2(\frac{n}{4}) - N_{j-1}}{n_j} = (y'_{i-1})_j + c \frac{\frac{n}{2} - N_{j-1}}{n_j}$	$Q_2 = (y'_{i-1})_j$
3°Cuartil Q_3	$N_{j-1} < 3 \frac{n}{4}$	$N_{j-1} = 3 \frac{n}{4}$
	$Q_3 = (y'_{i-1})_j + c \frac{3(\frac{n}{4}) - N_{j-1}}{n_j}$	$Q_3 = (y'_{i-1})_j$

$$\text{El Cuartil } 2^{\circ} \quad Q_2 = Me$$

**Ejemplo:**

Tabla de frecuencias para la longitud de los tornillos fabricados por una máquina.

	Intervalos de Clase	n _i	N _i	
	6-7	11	11	
	7-8	9	20	N_{j-1}
y_{j-1}^+	8 9	14	34	N_j
	9-10	11	45	
	10-11	22	67	N_{j-1}
y_{j-1}^+	11 -12	14	81	N_j
	12-13	7	88	
	13-14	5	93	
	14-15	4	97	
	15-16	3	100	
	Total	100		

Cálculo del 1º cuartil

$$Q_1 = 8 + 1 \frac{\left(\frac{100}{4}\right) - 20}{14} = 8,36 \text{ mm}$$

Concluimos entonces, en que el 25% de los tornillos producidos mide menos de 8,36 mm y el 75% mide más de 8,36 mm.

Cálculo del 3º cuartil

$$Q_3 = 11 + 1 \frac{\left[\frac{3(100)}{4}\right] - 67}{14} = 11,57 \text{ mm}$$

Este resultado indica que el 75% de los tornillos mide menos de 11,57 mm., y el 25% restante mide más de 11,57 mm.

Podemos utilizar la forma gráfica a través de un polígono de frecuencias relativas a los efectos de visualizar la información aportada por los cuartiles.



Deciles

En relación a los Deciles de una distribución, diremos que, como su nombre lo indica, son valores de la variable, que dividen al conjunto de observaciones (ordenadas de menor a mayor) en diez partes iguales. Tienen el mismo significado y la misma forma de cálculo que los cuartiles.

Los simbolizaremos por D_k para $k = 1, 2, \dots, 9$ que indica el número de Decil

Simbología	Definición
D_1	Valor de la variable que agrupa el 10% de los datos. (10% de los datos a la izquierda y 90% de los datos a la derecha)
D_2	Valor de la variable que agrupa el 20% de los datos.
D_3	Valor de la variable que agrupa el 30% de los datos.
D_4	Valor de la variable que agrupa el 40% de los datos.
D_5 <i>Mediana</i>	Valor de la variable que agrupa el 50% de los datos.
D_6	Valor de la variable que agrupa el 60% de los datos.
D_7	Valor de la variable que agrupa el 70% de los datos.
D_8	Valor de la variable que agrupa el 80% de los datos.
D_9	Valor de la variable que agrupa el 90% de los datos.
D_{10}	Valor de la variable que agrupa el 100% de los datos.

Formas de Cálculo

a) *Datos No Agrupados*



Deciles	Orden de cada Decil	
	Número de datos par	Número de datos impar
<i>General</i>		
D_1	$\left[k \left(\frac{n}{10} \right) \right]^\circ$	$\left[k \left(\frac{n+1}{10} \right) \right]^\circ$
<i>K: Número de Decil: 1, 2, ..., 9</i>		
D_1	$\left[1 \left(\frac{n}{10} \right) \right]^\circ$	$\left[1 \left(\frac{n+1}{10} \right) \right]^\circ$
D_5	$\left[5 \left(\frac{n}{10} \right) \right]^\circ = \left(\frac{n}{2} \right)^\circ$	$\left[5 \left(\frac{n+1}{10} \right) \right]^\circ = \left(\frac{n+1}{2} \right)^\circ$
b) D_7	$\left[7 \left(\frac{n}{10} \right) \right]^\circ$	$\left[7 \left(\frac{n+1}{10} \right) \right]^\circ$



b) Datos Agrupados

Seguiremos igual procedimiento que para el cálculo de la mediana. Considerando ahora que para ubicar el orden de D_i se debe considerar $\frac{n}{10}$ y para ubicar el orden de D_5 se debe considerar $5\frac{n}{10}$.

b.1) Distribuciones de Frecuencias en Lista

Deciles	Formas de Cálculo	
General	$N_{j-1} < k \frac{n}{10}$	$N_{j-1} = k \frac{n}{10}$
	$D_k = y_j$	$D_k = \frac{y_{j-1} + y_j}{2}$
	$K: \text{Número de Decil: } 1, 2, \dots, 9$	
1° Decil D_1	$N_{j-1} < 1 \frac{n}{10}$ $D_1 = y_j$	$N_{j-1} = 1 \frac{n}{10}$ $D_1 = \frac{y_{j-1} + y_j}{2}$
5° Decil D_5	$N_{j-1} < 5 \frac{n}{10}$ $D_5 = y_j$	$N_{j-1} = 5 \frac{n}{10}$ $D_5 = \frac{y_{j-1} + y_j}{2}$
7° Decil D_7	$N_{j-1} < 7 \frac{n}{10}$ $D_7 = y_j$	$N_{j-1} = 7 \frac{n}{10}$ $D_7 = \frac{y_{j-1} + y_j}{2}$



b.2) Distribuciones de Frecuencias en Intervalos

Deciles	Formas de Cálculo	
General	$N_{j-1} < k \frac{n}{10}$	$N_{j-1} = k \frac{n}{10}$
	$D_k = (y'_{i-1})_j + c \frac{k(\frac{n}{10}) - N_{j-1}}{n_j}$	$D_k = (y'_{i-1})_j$
	$K:$ Número de Decil: 1, 2, ..., 9	
1º Decil D_1	$N_{j-1} < 1 \frac{n}{10}$	$N_{j-1} = 1 \frac{n}{10}$
	$D_1 = (y'_{i-1})_j + c \frac{1(\frac{n}{10}) - N_{j-1}}{n_j}$	$D_1 = (y'_{i-1})_j$
5º Decil D_5	$N_{j-1} < 5 \frac{n}{10}$	$N_{j-1} = 5 \frac{n}{10}$
	$D_5 = (y'_{i-1})_j + c \frac{5(\frac{n}{10}) - N_{j-1}}{n_j} = (y'_{i-1})_j + c \frac{\frac{n}{2} - N_{j-1}}{n_j}$	$D_5 = (y'_{i-1})_j$
7º Decil D_7	$N_{j-1} < 7 \frac{n}{10}$	$N_{j-1} = 7 \frac{n}{10}$
	$D_7 = (y'_{i-1})_j + c \frac{7(\frac{n}{10}) - N_{j-1}}{n_j}$	$D_7 = (y'_{i-1})_j$

El Decil 5º $D_5 = Q_2 = Me$



Percentiles

Por último, y en relación a los Percentiles de una distribución, también, como su nombre lo indica, son valores de la variable, que dividen al conjunto de observaciones (ordenadas de menor a mayor) en cien partes iguales. Tienen el mismo significado y la misma forma de cálculo que los cuartiles.

Así, cuando se habla del percentil 15, se quiere expresar que es el valor de la variable que deja el 15% de los datos a su izquierda y el 85% de los mismos a su derecha.

Los simbolizaremos por P_k para $k = 1, 2, \dots, 99$ que indica el número de Percentil

Formas de Cálculo

a) Datos No Agrupados



Percentiles	Orden de cada Percentil	
	Número de datos par	Número de datos impar
General	$\left[k \left(\frac{n}{100} \right) \right]^\circ$	$\left[k \left(\frac{n+1}{100} \right) \right]^\circ$
<i>K: Número de Percentil: 1, 2, ..., 99</i>		
1° Percentil P_1	$\left[1 \left(\frac{n}{100} \right) \right]^\circ$	$\left[1 \left(\frac{n+1}{100} \right) \right]^\circ$
50° Percentil P_{50}	$\left[50 \left(\frac{n}{100} \right) \right]^\circ = \left(\frac{n}{2} \right)^\circ$	$\left[50 \left(\frac{n+1}{100} \right) \right]^\circ = \left(\frac{n+1}{2} \right)^\circ$
70° Percentil P_{70}	$\left[70 \left(\frac{n}{100} \right) \right]^\circ$	$\left[70 \left(\frac{n+1}{100} \right) \right]^\circ$



b) Datos Agrupados

Seguiremos igual procedimiento que para el cálculo de la mediana. Considerando ahora que para ubicar el orden de P_i se debe considerar $\frac{n}{100}$ y para ubicar el orden de D_{50} se debe considerar $50 \frac{n}{100}$.

b.1) Distribución de Frecuencias en Lista

Percentiles	Formas de Cálculo	
General	$N_{j-1} < k \frac{n}{100}$	$N_{j-1} = k \frac{n}{100}$
	$P_k = y_j$	$P_k = \frac{y_{j-1} + y_j}{2}$
	$K:$ Número de Percentil: 1, 2,..., 99	
1° Percentil P_1	$N_{j-1} < 1 \frac{n}{100}$ $P_1 = y_j$	$N_{j-1} = 1 \frac{n}{100}$ $P_1 = \frac{y_{j-1} + y_j}{2}$
50° Percentil P_{50}	$N_{j-1} < 50 \frac{n}{100}$ $P_5 = y_j$	$N_{j-1} = 50 \frac{n}{100}$ $P_5 = \frac{y_{j-1} + y_j}{2}$
70° Percentil P_{70}	$N_{j-1} < 70 \frac{n}{100}$ $P_7 = y_j$	$N_{j-1} = 70 \frac{n}{100}$ $P_7 = \frac{y_{j-1} + y_j}{2}$



b.2) Distribución de Frecuencias en Intervalos

Percentiles	Formas de Cálculo	
General	$N_{j-1} < k \frac{n}{100}$	$N_{j-1} = k \frac{n}{100}$
	$P_k = (y'_{i-1})_j + c \frac{k(\frac{n}{100}) - N_{j-1}}{n_j}$	$P_k = (y'_{i-1})_j$
	K: Número de Percentil: 1, 2, ..., 99	
1° Percentil P_1	$N_{j-1} < 1 \frac{n}{100}$	$N_{j-1} = 1 \frac{n}{100}$
	$P_1 = (y'_{i-1})_j + c \frac{1(\frac{n}{100}) - N_{j-1}}{n_j}$	$P_1 = (y'_{i-1})_j$
50° Percentil P_{50}	$N_{j-1} < 50 \frac{n}{100}$	$N_{j-1} = 50 \frac{n}{100}$
	$P_{50} = (y'_{i-1})_j + c \frac{50(\frac{n}{100}) - N_{j-1}}{n_j} = (y'_{i-1})_j + c \frac{(\frac{1}{2}) - N_{j-1}}{n_j}$	$P_{50} = (y'_{i-1})_j$
70° Percentil P_{70}	$N_{j-1} < 70 \frac{n}{100}$	$N_{j-1} = 70 \frac{n}{100}$
	$P_{70} = (y'_{i-1})_j + c \frac{70(\frac{n}{100}) - N_{j-1}}{n_j}$	$P_{70} = (y'_{i-1})_j$

$$\text{El Percentil } 50^{\circ}: P_{50} = Q_2 = D_5 = Me.$$

La mediana es el 2° Cuartil, 5° Decil y 50° Percentil, así:

$$Me = Q_2 = D_5 = P_{50}$$



2. Medidas de Asimetría

Cuando dos distribuciones coinciden en sus medidas de posición y dispersión, no tenemos datos analíticos para ver si son distintas. Una manera de compararlas es mediante su forma.

Entonces, las medidas de asimetría, son medidas de forma de una distribución.

Permiten identificar y describir la manera en que los datos tienden a reunirse, en función de su frecuencia, dentro de la distribución. Presenta las siguientes formas, en comparación con la distribución normal estandarizada, con media 0 y desviación estándar 1:

Simetría

Se da cuando en una distribución se reúne aproximadamente la misma cantidad de los datos a ambos lados de la media aritmética. No tiene alargamiento o sesgo. Se representa por una curva normal en forma de campana llamada campana de Gauss (matemático Alemán 1777-1855) o también conocida como de Laplace (1749-1827).

Igualmente se dice que una distribución es simétrica cuando su media aritmética, su mediana y su moda son iguales, en símbolos

$$\bar{x} = m_e = m_d$$

Una distribución simétrica significa que la clase central tiene la mayor frecuencia y que las frecuencias, arriba de la clase central de la distribución, corresponden exactamente a las frecuencias debajo de la clase central.

Si la distribución de frecuencias se muestra mediante una gráfica de barras (Histograma), o un gráfico de líneas (Polígono), una recta vertical a través del punto medio de la clase central, puede dividir la gráfica en dos mitades iguales. La intersección de la recta vertical y el eje de las x , da los valores de la Media, Mediana y Moda.

Ejemplo

Para la siguiente distribución, correspondiente a las calificaciones de 28 alumnos:

$y_{i-1} - y_i$	y_i	n_i	$y_i \cdot n_i$	N_i	N_{st}	N
30 - 40	35	2	70	2		
40 - 50	45	3	135	5		
50 - 60	55	5	275	10		
60 - 70	65	8 = n	520	18		
70 - 80	75	5	275	23		
80 - 90	85	3	255	26		
90 - 100	95	2	190	28		
		28	1.820			



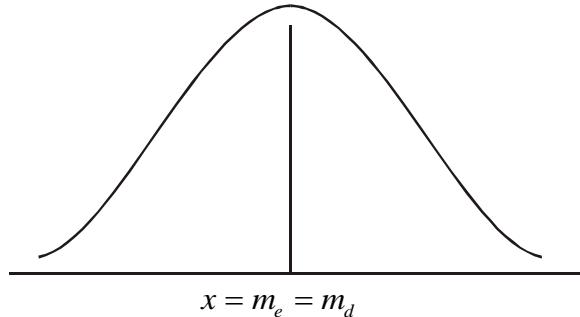
Calculamos:

$$\bar{y} = \frac{1.820}{28} = 65$$

$$m_e = \frac{n}{2} = \frac{28}{2} = 14 \Rightarrow N_{j-1} < \frac{n}{2} \Rightarrow M_e = 60 + 10 \cdot \frac{14 - 10}{8} = \underline{\underline{65}}$$

$$m_d = 60 + 10 \cdot \frac{3}{6} = \underline{\underline{65}}$$

Gráficamente



Si las frecuencias por arriba de la clase modal, no son las mismas que las frecuencias por debajo de dicha clase, la distribución no es simétrica y los valores de las medidas vistas no son iguales.

Cuando la distribución no es simétrica, el polígono de frecuencias se tornará asimétrico, ya sea hacia el lado derecho sobre la escala de las x (positivamente asimétrico), o hacia el lado izquierdo del eje de las x (negativamente asimétrico).

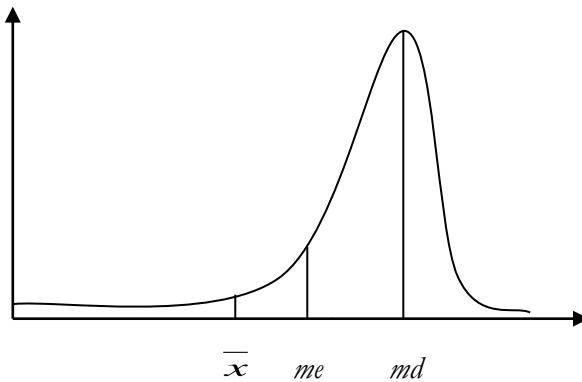
Asimetría Negativa o a la Izquierda

Se da cuando en una distribución la minoría de los datos está en la parte izquierda de la media. Este tipo de distribución presenta un alargamiento o sesgo hacia la izquierda, es decir, la distribución de los datos tiene a la izquierda una cola más larga que a la derecha.

También se dice que una distribución es simétrica a la izquierda o tiene sesgo negativo cuando el valor de la media aritmética es menor que la mediana y ésta es menor que la moda, en símbolos

$$\bar{x} \langle m_e \langle m_d$$

Gráficamente:



La moda es el valor que más frecuentemente se presenta, por lo tanto está siempre bajo el punto más alto de la curva.

La media aritmética está afectada por los valores bajos de la distribución, y su posición sobre la escala de las x es estirada hacia la izquierda o hacia valores más bajos de la escala.

La mediana no es afectada tan grandemente por los valores bajos como lo es la media, y por definición se encuentra al medio de la distribución. Si bien tiene un valor menor a la moda, se ubicar entre ella y la media.

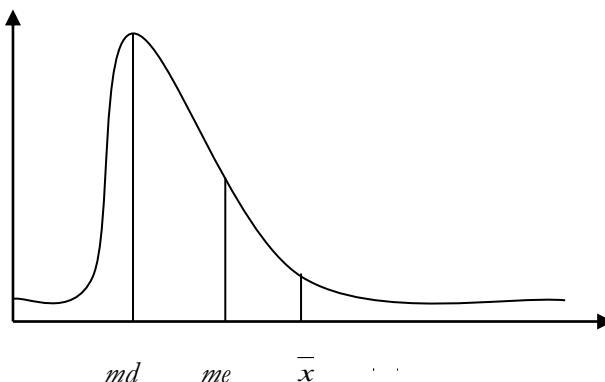
Asimetría Positiva o a la Derecha

Se da cuando en una distribución la minoría de los datos está en la parte derecha de la media aritmética. Este tipo de distribución presenta un alargamiento o sesgo hacia la derecha, es decir, la distribución de los datos tiene a la derecha una cola más larga que a la izquierda.

También se dice que una distribución es simétrica a la derecha o tiene sesgo positivo cuando el valor de la media aritmética es mayor que la mediana y ésta es mayor que la moda, en símbolos

$$\boxed{\bar{x} > m_e > m_d.}$$

Gráficamente:





La moda es el valor que más frecuentemente se presenta, y está siempre bajo el punto más alto de la curva.

La media, en este caso está afectada por los valores altos de la distribución, y su posición sobre la escala de las x es estirada hacia la derecha o hacia valores más altos de la escala.

La mediana, como ya dijimos, no es afectada tan grandemente por los valores altos como lo es la media, puesto que la mediana es el elemento central y divide a la curva en dos áreas iguales. Ahora es un valor más alto que la moda, pero siempre se ubica entre la moda y la media.

Nota: Sesgo es el grado de asimetría de una distribución, es decir, cuánto se aparta de la simetría.

Coefficiente de asimetría de Karl Pearson

Esta medida, la medida pearsoniana de asimetría, se basa en las relaciones entre la media, la mediana y la moda. Según lo visto anteriormente, para una distribución unimodal simétrica, estas medidas son iguales, pero en las distribuciones asimétricas la media se aleja de la moda según la asimetría, manteniéndose la mediana entre ambas.

Entonces, una medida de la asimetría sería la diferencia entre la media y la moda. Mientras mayor sea esa distancia, positiva o negativa, más asimétrica será la distribución.

Pero esta medida tiene dos defectos en su aplicación:

- 1- Al ser una medida dada en valor absoluto, mantiene sus unidades de medida.
- 2- La misma cantidad absoluta de asimetría tiene un diferente significado para distintas series con distintos grados de variabilidad.

Para eliminar ambos defectos se trabaja con una medida relativa de asimetría, llamado coeficiente de Pearson, que simbolizamos y calculamos de la siguiente manera:

$$k_p = \frac{3(\bar{x} - m_d)}{\hat{s}}$$

Dónde:

\bar{x} : media aritmética

m_d : moda

\hat{s} : desviación estándar muestral corregida

Como el valor modal de muchas distribuciones solo es una aproximación, en su lugar se utiliza la mediana



$$k_p = \frac{3(\bar{x} - m_d)}{\hat{s}}$$

Como vimos anteriormente la moda por el método empírico se determina:

$$m_d = \bar{x} - 3(\bar{x} - m_e)$$

Reemplazando

$$\bar{x} - m_d = \bar{x} - [\bar{x} - 3(\bar{x} - m_e)] = 3(\bar{x} - m_e)$$

Entonces, podemos escribir:

$$k_p = \frac{3(\bar{x} - m_e)}{\hat{s}}$$

Por esta medida: $3(\bar{x} - m_e)$, el coeficiente de Pearson puede asumir los siguientes resultados:

Si $k_p < 0$ la distribución será asimétrica negativa.

Si $k_p = 0$ la distribución será simétrica.

Si $k_p > 0$ la distribución será asimétrica positiva.

El Coeficiente de Pearson varía entre -3 y 3. Sin embargo, en raras ocasiones supera los límites de ± 1 .

Coeficiente de Asimetría de Fisher

$$k_3 = \frac{\mu_3}{\hat{s}^3}, \quad \text{donde} \quad \mu_3 = \frac{\sum_{i=1}^m (y_i - \bar{y})^3 n_i}{n} = \sum_{i=1}^m (y_i - \bar{y})^3 h_i$$

Pudiendo k arrojar los siguientes resultados:

Cuando $\bar{x} = m_d = m_e$, $\mu_3 = 0$ y k_3 podrá asumir los siguientes resultados:

$k < 0$; indica asimetría negativa

$k = 0$; indica simetría

$k > 0$; indica asimetría positiva



3. Medidas de Dispersion o De Concentración

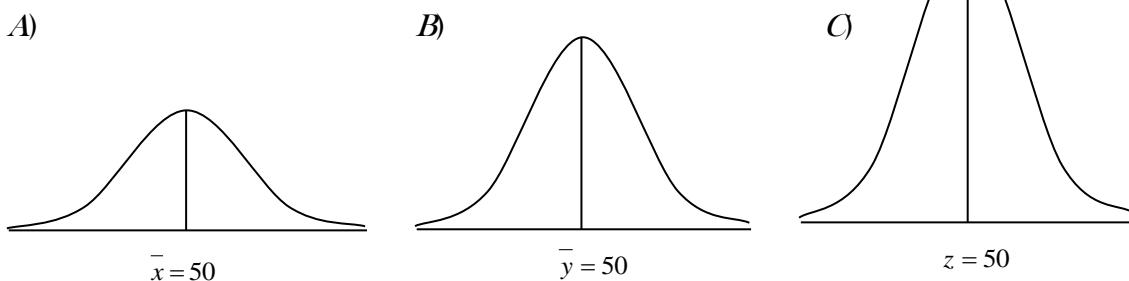
Llamadas también medidas de variabilidad

Son útiles porque:

1. Permiten juzgar la confiabilidad de la medida de tendencia central.
2. Los datos demasiados dispersos tienen un comportamiento especial.
3. Es posible comparar la dispersión de diversas distribuciones.

Los parámetros o estadígrafos de posición no son suficientes para caracterizar totalmente una cierta distribución de frecuencias, ya que los valores incluidos en un grupo de datos en general varían significativamente en magnitud; es decir, algunos de ellos son pequeños y otros grandes, estableciéndose lo que se llama dispersión o concentración de los valores. Dos ó más conjuntos de datos pueden diferir tanto en tendencia central como en dispersión o pueden tener las mismas medidas de tendencia central, pero pueden tener grandes diferencias de dispersión.

Ejemplo:



Las figuras *A*, *B* y *C*, si bien son iguales en cuanto a su promedio, los datos en *C* están más concentrados alrededor del promedio que la figura *B* y los de ésta más que los de la figura *A*), o dicho de otro modo, los datos de la figura *A*) están más dispersos que los de la figura *B*) y éstos más que los de la *C*.

Una medida de dispersión, es importante por dos razones:

- 1- Indica el grado de variación entre los valores de la serie de datos recopilados, en relación a alguna medida de posición. Utilizaremos generalmente a la media aritmética.
- 2- Es utilizada como complemento de la medida de posición para caracterizar un conjunto de datos o para compararlos con respecto a otro conjunto cualquiera. Cuando la dispersión es grande, el promedio tiene muy poca significación, en cambio si la dispersión es baja, es decir los datos están muy concentrados alrededor el promedio, éste se vuelve altamente significativo, ya que en este caso la medida de posición corresponde a un valor muy representativo del conjunto.

Toda medida de dispersión es un número real y nunca puede ser menor a 0. Si es 0 todos los datos son idénticos. Va en aumento según los datos se hacen más diversos.



Las Medidas De Tendencia Central Tienen Como Objetivo El Sintetizar Los Datos En Un Valor Representativo, Las Medidas De Dispersión Nos Dicen Hasta Qué Punto Estas Medidas De Tendencia Central Son Representativas Como Síntesis De La Información. Las Medidas De Dispersion Cuantifican La Separación, La Dispersion, La Variabilidad De Los Valores De La Distribución Respecto Al Valor Central.

3.1. Recorrido

También llamado Amplitud o Campo de Variación, es la medida más simple y más bruta de dispersión.

Se define como la diferencia entre el valor máximo y el valor mínimo de los valores observados. Mide la distancia entre un punto y otro. Es una medida posicional de dispersión ya que está basada en las posiciones de ciertos elementos.

Simbología

R

Formas De Cálculo

a) Series Simples o Datos No Agrupados

$$R = x_{\max} - x_{\min} \quad R = x_n - x_1$$

$$x_i \quad (i = 1, 2, \dots, n); \quad x_1, x_2, \dots, x_n$$

x_{\max} : valor máximo observado
 x_{\min} : valor mínimo observado

b) Datos Agrupados

Distribuciones De Frecuencias En Lista o En Intervalos

$$R = y_{\max} - y_{\min} \quad R = y_m - y_1$$

$$y_i \quad (i = 1, 2, \dots, m); \quad y_1, y_2, \dots, y_m$$

y_{\max} : valor máximo observado
 y_{\min} : valor mínimo observado



Presenta los siguientes inconvenientes:

- Se ve muy influenciado por valores no usuales de los datos. Si aparece un valor fuera de lo común ya sea muy grande o muy pequeño, no es posible utilizarlo como medida de dispersión, porque su resultado está deformado por estos valores fuera de serie.
- No es una medida de dispersión de los datos intermedios con relación al valor típico.
- Es sensible al tamaño de muestra. Tiende a cambiar aunque no proporcionalmente, en la misma dirección en que varía el tamaño de la muestra. Cuando aumenta el número, es posible que algún dato pueda tener mayor valor que el máximo y algún otro dato un valor que el mínimo de la muestra anterior.

A pesar de sus numerosas deficiencias, puede ser usado muy provechosamente como medida de dispersión para muchos fines:

- En situaciones en las que se desea conocer sólo la extensión de la dispersión extrema en condiciones ordinarias. Si el dato máximo o mínimo no es usual, la amplitud no revela nada acerca de la distribución ordinaria de los datos.

Ejemplo:

Los informes de mercado de acciones se expresan frecuentemente en términos de su amplitud, cotizando precios altos y bajos de acciones durante un período de tiempo.

Cuando no se producen movimientos excepcionales de los precios de las acciones, la amplitud cotizada puede medir la variación ordinaria. Pero, cuando ocurren movimientos excepcionales, la amplitud revela los efectos de condiciones trastornadoras temporales en el mercado.

- Para medir la dispersión de una serie simétrica y casi continua. En tal caso puede ser aproximada la medida, tomando el promedio de los dos valores extremos, llamado *Amplitud Media*.

Ejemplo:

Temperatura media diaria. Se promedia la máxima y la mínima, en vez de usar las veinticuatro lecturas horarias del día.

- Cuando la muestra es pequeña, especialmente cuando la misma operación de muestreo se repite a menudo y es utilizado un promedio de cada resultado sucesivo.

Ejemplo:

Control Estadístico de la Calidad

En general, se desea una medida de variabilidad que dependa de todas las observaciones y no de unas pocas. Por ello es razonable medir la variación en términos de las desviaciones con respecto a alguna medida de localización. En base a ello, podemos definir:



3.2. Desviación Media

Definición

Es la media aritmética de los valores absolutos de las desviaciones entre los valores de la variable y una medida de tendencia central. Las desviaciones deben considerarse tantas veces como se presentan. Entonces, podemos identificar tres desviaciones medias:

1- Desviación Media Respecto a La Media:

$$DM_{\bar{X}} = \frac{\sum_{i=1}^m |y_i - \bar{x}| n_i}{n}$$

2- Desviación Media Respecto a La Mediana:

$$DM_{me} = \frac{\sum_{i=1}^m |y_i - me| n_i}{n}$$

3- Desviación Media Respecto a La Moda:

$$DM_{md} = \frac{\sum_{i=1}^m |y_i - md| n_i}{n}$$

La desviación media da cuenta de la distancia promedio que existe entre los valores de la variable y la medida de tendencia central.

Generalmente para determinar las desviaciones se toma como medida de localización a la media aritmética.

Entonces la Desviación Media con respecto a la media aritmética se define como la media aritmética de los valores absolutos de las desviaciones entre los valores de la variable con respecto a la media aritmética.

De esta manera esta medida se basa en todos los elementos y permite medir la dispersión alrededor de un promedio.

Simbología

Población	$DM(X)$
Muestra	$DM(x)$



Formas De Cálculo

a) Series Simples o Datos No Agrupados

$$DM(x) = M \left[|x_i - \bar{x}| \right] = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

Se ignoran los signos positivos o negativos de las desviaciones ya que como se vio al estudiar las propiedades de la media:

$$\sum_{i=1}^n (x_i - \bar{x}) = 0 \quad \text{y} \quad \sum_{i=1}^m (y_i - \bar{y}) n_i = 0$$

a) Datos Agrupados

Distribuciones De Frecuencias En Lista o En Intervalos

$$DM(y) = M \left[|y_i - \bar{y}| \right] = \frac{\sum_{i=1}^m |y_i - \bar{y}| n_i}{n}$$

Donde y_i será:

- Los distintos valores que asumió la variable si se trata de una Distribución de Frecuencias en Lista
- Las marcas de clase si se trata de una Distribución de Frecuencias en Intervalos

Procedimiento:

- 1- Obtener la media aritmética
- 2- Obtener las desviaciones de cada valor de la variable o cada marca de clase, según corresponda, con respecto a la media aritmética
- 3- Obtener la suma de los valores absolutos de las desviaciones (multiplicada por n para el caso de datos agrupados).
- 4- Dividir la suma por el número de datos.



3.3. Varianza

Definición

Se define como la Media Aritmética de los cuadrados de las desviaciones con respecto a la media aritmética.

Simbología

Población	$V(X) = \sigma_x^2$
Muestra	$V(x) = s_x^2$

Como ya vimos $\frac{\sum_{i=1}^n (x_i - \bar{x})}{n} = 0$, en series simples, y $\frac{\sum_{i=1}^m (y_i - \bar{y}) n_i}{n} = 0$, en datos agrupados, cualquiera sea la distribución. El que esta media valga 0, se debe a las compensaciones de las desviaciones positivas con las negativas. Esta compensación, ya fue evitada considerando el valor absoluto de las desviaciones.

La otra forma de evitarlas es tomar una potencia par para dichas desviaciones. Si tomamos la potencia 2, se transforma en la varianza, haciendo uso de esta manera de la propiedad de los mínimos cuadrados de la media:

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \text{mínimo} \quad \text{y} \quad \sum_{i=1}^m (y_i - \bar{y})^2 n_i = \text{mínimo}$$

Por lo cual, como medidas de las diferencias promedio al cuadrado en torno a la media, la varianza y la desviación estándar, deben ser menores que cualquier otra medida de diferencias promedio al cuadrado en torno a cualquier otro indicador de tendencia central.

Dado que en el cálculo precedente se elevó al cuadrado, ni la varianza, ni la desviación estándar, podrán ser nunca negativas y solo alcanzarán el valor 0 si cada observación realizada es exactamente igual.

Formas De Cálculo

a) Series Simples o Datos No Agrupados

$$V(x) = M \left[(x_i - \bar{x})^2 \right] = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$



b) Datos Agrupados

Distribuciones De Frecuencias En Lista y En Intervalos

$$V(y) = M[(y_i - \bar{y})^2] = \frac{\sum_{i=1}^n (y_i - \bar{y})^2 n_i}{n}$$

Donde y_i , como antes, será:

- Los distintos valores que asumió la variable si se trata de una Distribución de Frecuencias en Lista
- Las marcas de clase si se trata de una Distribución de Frecuencias en Intervalos

Procedimiento:

1. Obtener la media aritmética
2. Obtener las desviaciones de cada valor de la variable con respecto a la media aritmética
3. Elevar al cuadrado cada desviación obtenida
4. Obtener la suma del cuadrado de las desviaciones (multiplicada por n para el caso de datos agrupados).
5. Dividir la suma por el número de datos.

Como puede observarse el cálculo de la varianza a través de esta fórmula definicional es tedioso, además de insumir tiempo, sobre todo, si trabajamos con una gran cantidad de datos y el cálculo es manual.

Para obviar estas dificultades, puede utilizarse la llamada fórmula de cálculo rápido o método abreviado.

A tal fórmula arribamos a partir de:

a) Series Simples o Datos No Agrupados

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



Resolviendo el cuadrado del numerador:

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

Introduciendo el sumatorio:

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

Recordamos que:

El sumatorio de una constante por una variable es la constante por el sumatorio de la variable. Así, en el segundo término del numerador $2\bar{x}$ son constantes. La media de una muestra, es, para la muestra, una constante.

El sumatorio de una constante es n veces la constante.

Luego:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \Rightarrow \sum_{i=1}^n x_i = \bar{x}n \quad , \text{ reemplazando:}$$

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

Entonces:

$$V(x) = \frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2 = M(x^2) - [M(x)]^2$$

Ejemplo:

Sean:

$n = 10$

x = número de hijos por familia = 2, 1, 3, 1, 2, 1, 3, 0, 2, 1

Aplicamos la fórmula de cálculo

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

Obtenemos



$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{16}{10} = 1,6$$

Elevamos al cuadrado cada x_i y sumamos

$$\bar{x}^2 = 4, 1, 9, 1, 4, 1, 9, 0, 4, 1 \quad y \quad \sum_{i=1}^{10} x_i^2 = 34$$

Luego reemplazamos:

$$V(x) = \frac{34}{10} - (1,6)^2 = 0,84$$

b) Datos Agrupados

$$V(y) = M[(y_i - \bar{y})^2] = \frac{\sum_{i=1}^m (y_i - \bar{y})^2 n_i}{n}$$

Que corresponde a la fórmula definicional, a partir de la cual, y realizando igual procedimiento que para series simples, arribamos a:

$$V(y) = \frac{\sum_{i=1}^m y_i^2 n_i}{n} - \bar{y}^2 = M(y^2) - [M(y)]^2$$

O bien, considerando las frecuencias relativas:

$$V(y) = \sum_{i=1}^m (y_i - \bar{y}) h_i = \sum_{i=1}^m y_i^2 h_i - \bar{y}^2$$

Nota: En la Inferencia Estadística para el cálculo de la varianza muestral y en consecuencia, de la desviación estándar muestral, se utiliza como denominador a $n-1$ en vez de n . Esto se debe a que con una muestra tomada de una población grande se pretende descubrir cuánto varían los datos alrededor de la media poblacional. Sin embargo cuando no se conoce la media de la población se estima a través de la media aritmética de la muestra y esto hace que parezca menos variable de lo que es en realidad. Al dividir por $n-1$ se está compensando por la variabilidad más pequeña que se observa en la muestra. Por este motivo la varianza muestral calculada con $n-1$ es considerada un estimador más eficiente para la varianza poblacional.



Ejemplo:

y_i	n_i	$y_i n_i$	$y_i(y_i n_i) = y_i^2 n_i$	h_i	$y_i^2 h_i$
0	1	0	0	0,10	0
1	4	4	4	0,40	0,40
2	3	6	12	0,30	1,20
3	2	6	18	0,20	1,80
	10	16	34	1	3,40

b.1) Distribuciones De Frecuencias En Lista

Con los datos dados en Serie Simple, construimos la siguiente Tabla:

$$V(y) = \frac{\sum_{i=1}^m y_i^2 n_i}{n} - \bar{y}^2 = \frac{34}{10} - (1,6)^2 = \underline{0,84}$$

O bien

$$V(y) = \sum_{i=1}^m y_i^2 h_i - \bar{y}^2 = 3,40 - (1,6)^2 = \underline{0,84}$$

b.2) Distribuciones De Frecuencias En Intervalos

$y_{i-1} - y_i$	y_i	n_i	$y_i n_i$	$y_i^2 n_i$	h_i	$y_i^2 h_i$
45-55	50	2	100	5.000	0,10	250
55-65	60	4	240	14.400	0,20	720
65-75	70	7	490	34.300	0,35	1.715
75-85	80	4	320	25.600	0,20	1.280
85-95	90	3	270	24.300	0,15	1.215
		20	1.420	103.600	1	5.180

$$\bar{y} = \frac{\sum_{i=1}^m y_i n_i}{n} = \frac{1.420}{20} = 71$$

$$V(y) = \frac{\sum_{i=1}^m y_i^2 n_i}{n} - \bar{y}^2 = \frac{103.600}{20} - (71)^2 = \underline{139}$$

$$V(y) = \sum_{i=1}^m y_i^2 h_i - \bar{y}^2 = 5.180 - (71)^2 = \underline{139}$$



Propiedades

1- La Varianza es siempre una cantidad no negativa.

$$V(y) \geq 0$$

2- La Varianza de una constante es 0, cualquiera sea la constante.

$$V(k) = 0$$

Si todas las observaciones son iguales (k), es decir iguales, la media coincide con el valor común y las desviaciones son todas nulas.

3- La Varianza del producto de una constante por una variable, es igual al producto del cuadrado de la constante por la varianza de la variable.

$$V(ky) = k^2 V(y)$$

Ejemplo:

y_i	n_i
0	1
1	4
2	3
3	2
	10

$$V(y) = 0,84$$

Supongamos que multiplicamos a la variable por 2. Simbolizaremos por y' a la nueva variable y obtendremos la varianza.

$y_i \times 2 = y'_i$	n_i	$y'_i \cdot n_i$	$y'_i (y'_i \cdot n_i)$
0	1	0	0
2	4	8	16
4	3	12	48
6	2	12	72
	10	32	136

$$\bar{y}' = \frac{\sum_{i=1}^m y'_i \cdot n_i}{n} = \frac{32}{10} = 3,20$$



$$V(y') = \frac{\sum_{i=1}^m y_i'^2 n_i}{n} - \bar{y}'^2 = \frac{136}{10} - (3,20)^2 = 3,36$$

Luego: $V(2y) = 2 V(Y) = 4 \times 0,84 = 3,36$

- 4- La Varianza de la suma de una variable y una constante es igual a la varianza de la variable. Si a todos los valores de la variable se le suma un número igual, la varianza no se altera.

$$V(k + y) = V(y)$$

Ejemplo:

Supongamos que en el caso dado anteriormente, sumamos 2 a los valores originales

$y_i + 2 = y'_i$	n_i	$y'_i n_i$	$y'_i (y'_i n_i)$
2	1	2	4
3	4	12	36
4	3	12	48
5	2	10	50
	10	36	138

$$\bar{y}' = \frac{\sum_{i=1}^m y'_i n_i}{n} = \frac{36}{10} = 3,60$$

$$V(y') = \frac{\sum_{i=1}^m y_i'^2 n_i}{n} - \bar{y}'^2 = \frac{138}{10} - (3,60)^2 = 0,84$$

Esta propiedad se hace extensiva a la diferencia:

$$V(y - k) = V(y)$$

- 5- Si “x” e “y” son dos variables independientes, entonces la varianza de la suma o diferencia, es igual a la suma de las varianzas.

$$V(x + y) = V(x) + V(y)$$

y

$$V(x - y) = V(x) + V(y)$$



- 6- Si “ x ” e “ y ” son dos variables dependientes, entonces la varianza de la suma es igual a la suma de las varianzas más dos veces la covarianza de x, y .

$$V(x + y) = V(x) + V(y) + 2 \operatorname{Cov}(x, y)$$

y

$$V(x - y) = V(x) + V(y) - 2 \operatorname{Cov}(x, y)$$

Observaciones

- 1) La varianza, al igual que la media, es un índice muy sensible a las puntuaciones extremas.
- 2) En los casos en que no se pueda hallar la media, tampoco será posible hallar la varianza.
- 3) La varianza no viene expresada en las mismas unidades que los datos, ya que las desviaciones están elevadas al cuadrado.

3.4. Desviación Estándar

También llamada Desviación Típica ó Desviación Cuadrática Media.

Definición

Se define como el valor positivo de la raíz cuadrada de la Varianza.

Simbología

Población	$DS(X) = \sigma_x$
Muestra	$DS(x) = s_x$

Esta medida ha sido definida con el fin de hacer comparable su resultado o valor con los valores de la variable, dado que en el caso de la Varianza, por ser su valor de un orden superior a los valores de la variable, debido a que es un promedio de cuadrados, no es comparable. A los fines de reducir dicho valor a la misma dimensión que los valores de la variable, se le extrae la raíz cuadrada y se define la Desviación Estándar.

Mide, al igual que la Varianza, la dispersión de los valores de la variable respecto a su media aritmética. Es simplemente el promedio o variación esperada con respecto a la media aritmética.

Ninguna, Varianza y Desviación Estándar, tiene una interpretación intuitivamente obvia.

Cuando comparamos dos o más conjuntos de datos cuyas unidades de medición son idénticas, podemos decir que una muestra tiene un menor grado de dispersión que



otra, si la primera tiene una menor Varianza, ó Desviación Estándar. Sin embargo, dudaríamos de hacer una declaración precisa acerca de un conjunto de datos cuando se da un valor específico de una u otra medida.

Por ejemplo, si hallamos la desviación estándar de n observaciones de gastos de consumo familiar de alimentos, por día, de \$ 10, no podemos decir si este valor implica un grado de variabilidad alto, moderado o bajo de gastos de consumo de alimentos. Para vencer esta dificultad es necesario utilizar la regla empírica para la interpretación de la Desviación Estándar.

Los problemas experimentales van a ser estudiados mediante las distribuciones de probabilidades, como modelos teóricos explicativos de los hechos reales. Así en el tema Distribución Normal que veremos más adelante introduciremos la regla referida para interpretar la Desviación Estándar

Formas de Cálculo

a) Series Simples o Datos No Agrupados

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

b) Datos Agrupados

Distribuciones De Frecuencias En Lista y En Intervalos

$$DS(y) = s_y = \sqrt{V(y)} = \sqrt{\frac{\sum_{i=1}^m y_i^2 n_i}{n} - \bar{y}^2}$$

Propiedades

1- La desviación típica será siempre un valor positivo.

$$D(y) \geq 0$$

2- La desviación típica será siempre cero, en el caso de que las puntuaciones sean iguales.

$$D(k) = 0$$

3- Si a todos los valores de la variable se les suma un mismo número la desviación típica no varía

$$D(k+y) = D(y)$$



Esta propiedad se hace extensiva a la diferencia

$$D(y - k) = D(y)$$

Si todos los valores de la variable se multiplican por un mismo número la desviación típica queda multiplicada por dicho número.

$$D(k y) = k D(y)$$

- 4- Si “ x ” e “ y ” son dos variables independientes, entonces la desviación típica de la suma o diferencia, es igual a la suma de las desviaciones.

$$D(x + y) = \sqrt{V(x) + V(y)}$$

$$D(x - y) = \sqrt{V(x) + V(y)}$$

Observaciones

- 1- La desviación típica, al igual que la media, es un índice muy sensible a las puntuaciones extremas.
- 3- En los casos en que no se pueda hallar la media, tampoco será posible hallar la desviación típica.
- 4- Cuanto menor sea la desviación típica mayor será la concentración de los datos alrededor de la media.

3.5. Coeficiente de Variación

Definición

El Coeficiente de Variación es en realidad una Medida de Dispersion o de variabilidad relativa, pero de gran importancia, y de gran versatilidad, ya que su interpretación está basada en porcentajes, y nos da la relación existente entre la medida de posición y su precisión. Se suele expresar en "tanto" por ciento.

El coeficiente de variación es, al igual que la desviación estándar, una medida de dispersión o variabilidad y análogamente indica que tan dispersos o no, se encuentran los datos con respecto al promedio.

Es más precisa que la misma desviación estándar ya que está en proporción a la media, es decir señala qué tan grande es la magnitud de la desviación estándar respecto al promedio del conjunto de datos que se examina.

Por otro lado, las medidas de dispersión expresadas en valores absolutos son útiles para describir la dispersión de un solo conjunto de valores, mientras que el



coeficiente de variación, nos permite comparar la variabilidad de dos o más conjuntos de datos.

Si se comparan dos conjuntos de cifras, los valores absolutos son convenientes sólo si los promedios de los dos conjuntos son aproximadamente del mismo tamaño y las unidades de medida son iguales.

Por ejemplo, si nos piden comparar la dispersión de los pesos de las poblaciones de elefantes de dos circos diferentes, la desviación estándar nos dará información útil.

¿Pero, qué ocurre si lo que comparamos es la altura de unos elefantes con respecto a su peso? Tanto la media como la desviación típica, se expresan en las mismas unidades que la variable. Por ejemplo, en la variable altura podemos usar como unidad de longitud el metro y en la variable peso, el kilogramo. Comparar una desviación (con respecto a la media) medida en metros con otra en kilogramos no tiene ningún sentido.

El problema no deriva sólo de que una de las medidas sea de longitud y la otra sea de masa. La misma dificultad se plantea si medimos cierta cantidad, por ejemplo la masa, de dos poblaciones, pero con distintas unidades. Este es el caso en que comparamos el peso en *toneladas* de una población de 100 elefantes con el correspondiente en *miligramos* de una población de 50 hormigas.

Este inconveniente no se resuelve tomando las mismas escalas para ambas poblaciones.

Por ejemplo, se nos puede ocurrir medir a las hormigas con las mismas unidades que los elefantes (toneladas). Lo lógico es que la dispersión de la variable *peso de las hormigas* sea prácticamente nula (¡Aunque haya algunas que sean 1.000 veces mayores que otras!)

En los dos primeros casos mencionados anteriormente, el problema viene de la *dimensionalidad* de las variables, y en el tercero de la diferencia enorme entre las medias de ambas poblaciones. El *coeficiente de variación* es lo que nos permite evitar estos problemas, pues elimina la dimensionalidad de las variables y tiene en cuenta la proporción existente entre medias y desviación típica.

En síntesis, si los promedios son diferentes, aunque las unidades sean las mismas, o bien, si los promedios son similares, pero las unidades de medida son distintas, la comparación de grados de dispersión basados en valores absolutos no se puede realizar.

En estos casos, es necesario un valor relativo, a fin de obtener una cifra independiente de las unidades empleadas, es decir, un número abstracto, que puede incluso establecerse en forma de porcentaje.

Puesto que tanto la desviación estándar como la media se miden en las unidades originales, el *CV* es una medida independiente de las unidades de medición.

Debido a la propiedad anterior el *CV* es la cantidad más adecuada para comparar la variabilidad de dos conjuntos de datos.



Ejemplo:

Si tengo que el peso promedio de los elefantes es de 7500 kg con una desviación estándar de 500kg y también tengo que el peso promedio de los ratones es de 30g con una desviación estándar de 5 g, podré suponer (erróneamente) que los elefantes tienen mayor desviación que los ratones, sin embargo, si calculamos el coeficiente de variación de cada muestra se tendrá que el de los elefantes es de 6,7% y el de los ratones de 16,67%, es decir, es mayor la variación entre los pesos de los ratones.

De esta manera, se define la dispersión relativa ó coeficiente de variación.

Simbología

Población	$CV(X)$
Muestra	$CV(x)$

Formas de Cálculo

a) Series Simples o Datos No Agrupados

$$CV(x) = \frac{DS(x)}{M(x)}$$

b) Datos Agrupados

$$CV(y) = \frac{DS(y)}{M(y)}$$

Basta dar una rápida mirada a la definición del coeficiente de variación, para ver que las siguientes consideraciones deben ser tenidas en cuenta:

- Sólo se debe calcular para variables con todos los valores positivos. Todo índice de variabilidad es esencialmente no negativo. Las observaciones pueden ser positivas o nulas, pero su variabilidad debe ser siempre positiva. De ahí que sólo debemos trabajar con variables positivas, para la que tenemos con seguridad que la media debe ser mayor que 0.
- No es invariante ante cambios de origen. Es decir, si a los resultados de una medida le sumamos una cantidad positiva, $b > 0$, para tener $Y = X + b$, entonces $CV(y) < CV(x)$, ya que la desviación típica no es sensible ante cambios de origen, pero si la media. Lo contrario ocurre si restamos ($b < 0$).

$$CV(y) = \frac{DS(y)}{M(y)} = \frac{DS(x)}{M(x)+b} \left(\frac{DS(x)}{M(x)} \right)$$



- Es invariante a cambios de escala. Si multiplicamos x por una constante a , para obtener $y = ax$, entonces

$$CV(y) = \frac{DS(y)}{M(y)} = \frac{DS(ax)}{aM(x)} \left(\frac{a DS(x)}{a M(x)} \right) = CV(x)$$

Interpretación del Coeficiente de Variación

El Coeficiente de Variación, mide la variabilidad relativa a la Media. Expresa la proporción de variabilidad de una característica por cada unidad de la Media.

Mencionaremos algunos criterios encontrados, para interpretar los valores que alcanza este coeficiente:

- Si $CV \leq 20\%$ se dice que el promedio es representativo, o que los datos son homogéneos

Si el CV es mayor al 20%, el promedio no es representativo de los datos, o los mismos no son homogéneos.

2)

Coeficiente de Variación	Apreciación
26% o más	Muy heterogéneo
Del 16% a menos del 26%	Heterogéneo
Del 11% a menos del 16%	Homogéneo
0% a menos del 11%	Muy Homogéneo

3)

Valor del Coeficiente de Variación (%)	Interpretación del coeficiente	
	Variabilidad	Estabilidad
Igual a 0	Nula	Muy alta
Mayor de 0 hasta 20	Baja	Alta
Mayor de 20 hasta 60	Moderada	Moderada
Mayor de 60 hasta 90	Alta	Baja
Mayor de 90	Muy alta	Nula

4)

Coeficiente de Variación	Apreciación
$\leq 10\%$	Poca variabilidad
$>10 \text{ y } \leq 33\%$	Variabilidad Aceptable
$>33 \text{ y } \leq 50\%$	Mucha variabilidad pero tolerable
$>50\%$	Variabilidad excesiva y pérdida de su naturaleza

Es importante destacar lo mencionado anteriormente, en relación a que el *coeficiente de variación* sirve para comparar las variabilidades de dos conjuntos de valores (muestras o poblaciones). Si deseamos comparar a dos *individuos* de cada uno de esos conjuntos, deberá emplearse el *Puntaje típico o Puntaje estándar*, es decir *valores tipificados*. Por ello, es necesario introducir una nueva variable llamada Variable Desvió



Estandarizada o Tipificada que se define como el cociente de un valor aislado de la variable (y_i) menos su media (\bar{y}), dividido por su desviación estándar ($DS(y)$).

Simbólicamente:

$$z_i = \frac{y_i - \bar{y}}{DS(y)}$$

Entonces tendremos las siguientes variables:

		Media
Variable Natural	y_i	$M(y) = \bar{y}$
Variable Desvío	$y_i - \bar{y}$	$M(y_i - \bar{y}) = 0$
Variable Tipificada	$z_i = \frac{y_i - \bar{y}}{DS(y)}$	0

Nos interesa la posibilidad que existe de llevar todas las distribuciones de distintas variables a una única escala de valores relativos, que por ser tales pierden ya sus respectivas unidades de medida y pasan a ser valores abstractos, y como consecuencia, podrán ser analizadas con una única distribución de probabilidad, como modelo teórico explicativo, que es la conocida Distribución Normal Estandarizada, que veremos más adelante.

4. Medidas De Apuntamiento o Puntiagudez

Según ya mencionáramos, estas medidas miden el grado de variación, o la velocidad con que sube o baja la curva de izquierda a derecha.

El concepto de curtosis o apuntamiento de una distribución surge al comparar la forma de dicha distribución con la forma de la distribución normal. De esta forma, clasificaremos las distribuciones según sean más o menos apuntadas que la distribución normal.

Dicha variación, es medida a través del coeficiente de curtosis de Fisher, que analiza el grado de concentración que presentan los valores alrededor de la zona central de la distribución. Basándose en el dato de que en una distribución normal se verifica que:

$$\frac{\mu_4}{s^4} = 3$$

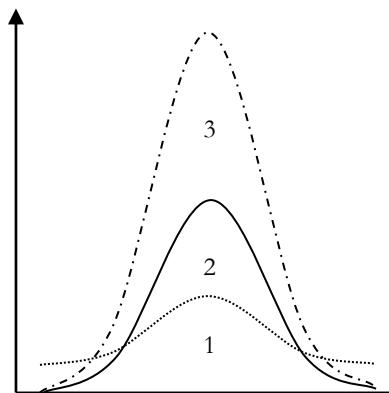
$$k_4 = \frac{\mu_4}{s^4} - 3, \text{ donde } \mu_4 = \frac{\sum_{i=1}^m (y_i - \bar{y})^4 n_i}{n} = \sum_{i=1}^m (y_i - \bar{y})^4 h_i$$



A partir de los resultados que pueden obtenerse, se definen tres tipos de distribuciones según su grado de curtosis:

- a. **Mesocúrtica:** Si el resultado del coeficiente de curtosis es igual a 0, lo que indica que presenta un grado de concentración medio alrededor de los valores centrales de la variable;
- b. **Leptocúrtica:** Si el resultado del coeficiente de curtosis es mayor que 0, lo que indica que presenta un elevado grado de concentración alrededor de los valores centrales de la variable;
- c. **Platicúrtica:** Si el resultado del coeficiente de curtosis es menor que 0, lo que indica que presenta un reducido grado de concentración alrededor de los valores centrales de la variable.

Gráficamente:



1-Platicúrtica

2- Mesocúrtica

3- Leptocúrtica