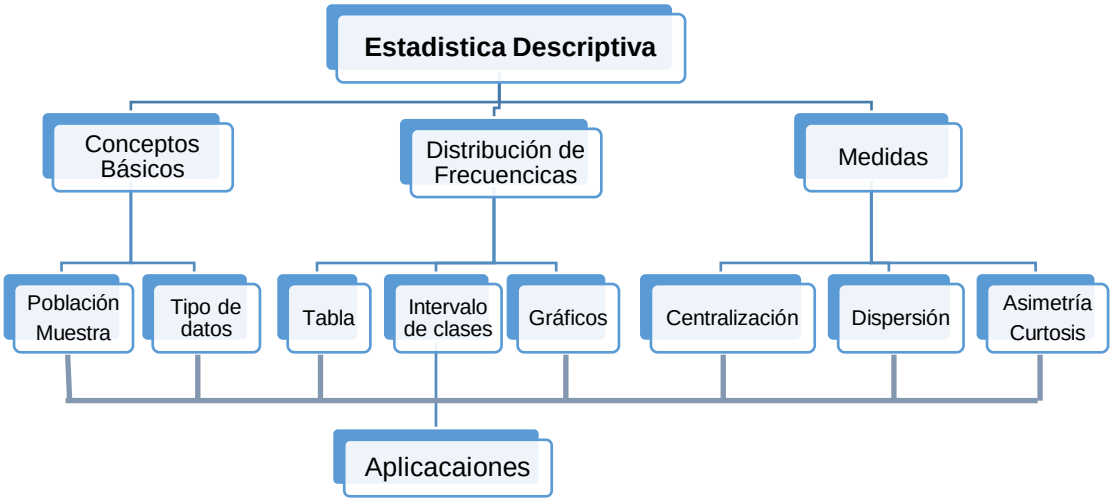
Estadística Descriptiva


Estadistica Descriptiva
Conceptos Básicos
Distribución de Frecuencicas
Medidas
Población Muestra
Tipo de datos
Tabla
Intervalo de clases
Gráficos
Centralización
Dispersión
Asimetría Curtosis
Aplicacaiones

OBJETIVOS

- ✓ Comprender y dominar los conceptos básicos de estadística descriptiva.
- ✓ Adquirir aptitudes para aplicar eficazmente conceptos y procedimientos estadísticos en el planteamiento y la resolución de problemas propios de la informática.
- ✓ Conocer y utilizar software específico para analizar, modelar, manipular y diseñar soluciones para una gran cantidad de datos.
- ✓ Conocer las técnicas descriptivas de clasificación y obtención de información a través de parámetros característicos de la muestra o población analizada.
- ✓ Conocer y utilizar adecuadamente el lenguaje estadístico.
- ✓ Asumir la necesidad y utilidad de la Estadística como herramienta en su ejercicio profesional.
- ✓ Sintetizar y describir una gran cantidad de datos seleccionando los estadísticos adecuados al tipo de variables y analizar las relaciones existentes entre ellas.

Estadística Descriptiva

1.- Introducción

La Estadística es la ciencia de los datos, ya que se encarga de recogerlos, organizarlos e interpretarlos. La vida diaria está compuesta por datos estadísticos: encuestas electorales, economía, deportes, datos meteorológicos, calidad de los productos, audiencias de televisión, etc. Por lo tanto se hace necesario una formación básica en Estadística para evaluar toda esta información.

La utilidad de la Estadística va mucho más allá de estos ejemplos, ya que es fundamental para muchas ramas de la ciencia, desde las Ciencias Sociales hasta la Ingeniería. Pero sobre todo es una herramienta de trabajo profesional.

En un principio la Estadística se ocupaba sobre todo de la descripción de los datos, fundamentalmente sociológicos, demográficos y económicos (censos de población, producciones agrícolas, riquezas, etc.), principalmente por razones fiscales. En el siglo XVII el cálculo de probabilidades se consolida como disciplina independiente, aplicándose sobre todo a los juegos de azar. Luego en el siglo XVIII el uso de la Estadística se extiende a problemas físicos (principalmente de Astronomía) y actuariales (seguros marítimos). Posteriormente se hace imprescindible en la investigación científica y es ésta la que la hace avanzar. Finalmente, en el siglo XIX, nace la Estadística como ciencia.

Actualmente la Estadística está tan difundida y sus méritos tan aceptados que prácticamente no existe actividad que no la utilice de una u otra manera, a tal punto que cualquier investigación que genere datos y no la utilice en la forma adecuada para su análisis, corre el riesgo que sus conclusiones no sean consideradas científicamente válidas.

2.- Conceptos Básicos

El tratamiento estadístico tiene básicamente dos fases: la organización y análisis inicial de los datos recogidos y la extracción de conclusiones válidas y toma de decisiones razonables a partir de ellos.

Los objetivos de la Estadística Descriptiva son los que se abordan en la primera de estas fases. Es decir, su misión es ordenar, describir y sintetizar la información recogida. En este proceso será necesario establecer medidas cuantitativas que reduzcan a un número manejable de parámetros del conjunto (en general grande) de datos obtenidos.

La elaboración de gráficas (visualización de los datos en diagramas) también forma parte de la Estadística Descriptiva dado que proporciona una manera visual directa de organizar la información.

La finalidad de la Estadística Descriptiva no es, entonces, extraer conclusiones generales sobre el fenómeno que ha producido los datos bajo estudio, sino solamente su descripción (de ahí el nombre).

2.1.- Población y Muestra

Al realizar un estudio es necesario tener bien en claro la diferencia entre población y muestra.

Definición:

Se llama población al conjunto completo de elementos que se estudia y que tienen una característica en común, que es el objeto de estudio.

Esta definición incluye, por ejemplo, a todos los sucesos en que podría concretarse un fenómeno o experimento cualesquiera. Una población puede ser finita o infinita.

Los habitantes de un país, los planetas del Sistema Solar, las estrellas en la Vía Láctea, son elementos de una población finita. Sin embargo, el número de posibles medidas que se puedan hacer de la velocidad de la luz, o de tiradas de un dado, forman poblaciones infinitas.

Aunque la población sea finita, el número de elementos que posee puede ser elevado, entonces es necesario trabajar con solo una parte de dicha población.

Definición:

A un subconjunto de elementos de la población se le conoce como muestra.

Si se quiere estudiar las propiedades de las estrellas en una Galaxia, no se tiene la oportunidad de observarlas todas; se debe estudiar esas características sobre una muestra representativa.

Nótese que elegir de forma representativa los elementos de una muestra es algo muy importante. De hecho existe un grave problema, conocido como efecto de selección, que puede condicionar el resultado de un estudio si no se realiza una selección correcta de los elementos que forman parte de una muestra.

El número de elementos de la muestra recibe el nombre de: tamaño de la muestra. Es fácil deducir que para que los resultados del estudio estadístico sean fiables es necesario que la muestra tenga un tamaño mínimo.

El caso particular de una muestra que incluye a todos los elementos de la población es conocido como censo.

2.2.- Característica de los conjuntos de datos

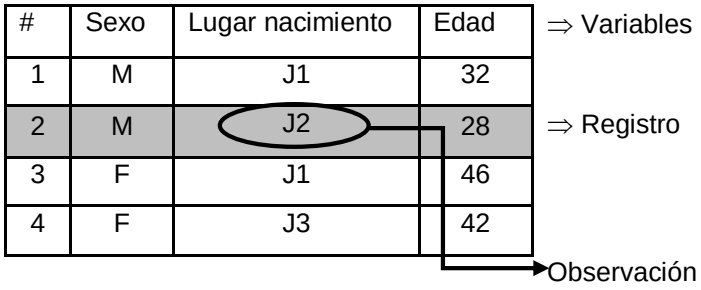
El objeto a ser medido pueden ser caracteres de tipos muy diferentes, de allí que se denomine: *Unidad de análisis o de observación,* al objeto bajo estudio. Puede ser una persona, una familia, un país, una región, una institución, en general cualquier objeto.

Variable, a cualquier característica de la unidad de observación que interese registrar y que al momento de ser registrada puede ser transformada en un número.

Valor de una variable, *Observación o Medición*, al número que describe a la característica de interés en una unidad de observación particular.

Registro, al conjunto de mediciones realizadas sobre una unidad de observación.

Por ejemplo, considérese que se desea registrar el sexo, lugar de nacimiento y edad de los habitantes de una región del norte del país.



Sexo, lugar nacimiento, edad, son variables que describen a una persona, pero el sexo de esa persona, su lugar de nacimiento y su edad son los valores que estas variables toman para esa persona.

Es importante, al comenzar a manejar un conjunto de datos, identificar cuántas variables se han registrado y cómo fueron registradas esas variables, esto permitirá definir la estrategia de análisis.

En el ejemplo anterior algunas de las variables son números mientras que otras son letras que indican categorías. A continuación se presenta una clasificación de los distintos tipos de datos que se puede encontrar. Debe observarse que distintos autores usan distintos criterios para clasificar datos por lo que se presentará un criterio que resulta útil desde el punto de vista de seleccionar el método de análisis estadístico más apropiado para los mismos.

2.3.- TIPO DE DATOS

Datos categóricos o cualitativos

Las variables categóricas o cualitativas resultan de registrar la presencia de un atributo. Las categorías de este tipo de variables deben ser mutuamente excluyentes y exhaustivas, es decir que, cada unidad de observación debe ser clasificada sin ambigüedad en una y sólo una de las categorías posibles y que existe una categoría para clasificar a todo individuo.

Es importante contemplar todas las posibilidades cuando se construyen variables categóricas, incluyendo una categoría tal como No sabe / No contesta, o No registrado u Otras, que asegura que todos los individuos observados serán clasificados con el criterio que define la variable.

Los datos categóricos pueden ser

a) Dicotómicos, la unidad de observación puede ser asignada a solo una de dos categorías. En general, se trata de la presencia ó ausencia de un atributo. La ventaja de este tipo de datos es la de poder asignar código 0 a la ausencia y 1 a la presencia.

Ejemplos

Varón / Mujer Embarazada / No embarazada

b) Más de dos categorías.

En este tipo de datos, si no existe un orden obvio entre las categorías, se denominan *nominales*. Por ejemplo: país de origen, estado civil. De existir un orden natural entre las categorías se denominan *ordinales*. Un ejemplo clásico es cuando se debe manifestar el acuerdo o no, respecto de una cuestión: Totalmente en desacuerdo, En desacuerdo, Indiferente, De acuerdo y Totalmente de acuerdo.

Datos numéricos

Una variable es numérica cuando el resultado de la observación o medición es un número. Estas variables pueden ser

a) Discretas, cuando solo pueden tomar una cantidad (finita o infinita) numerable de valores, es decir pueden tomar un cierto conjunto de valores posibles. En general, aparecen por conteo. Por ejemplo el número de electrones de un átomo, cantidad de personas que viven en un departamento.

b) Continuas, generalmente son el resultado de una medición que se expresa en unidades. Las mediciones pueden tomar teóricamente un conjunto infinito de valores posibles dentro de un rango. En la práctica los valores posibles de esta variable están limitados por la precisión del método de medición o por el modo de registro.

Por ejemplo la velocidad o altura de un móvil, peso de una persona.

La distinción entre datos discretos y continuos es la diferencia básica que existe entre contar y medir.

Considérese por ejemplo, la variable edad. Edad es continua, pero si se la registra en años resulta ser discreta. En estudios con adultos, en que la edad va de 20 a 70 años, por ejemplo, no hay problemas en tratarla como continua, ya que el número de valores posibles es muy grande. Pero en el caso de niños en edad preescolar, si la edad se registra en años debe tratarse como discreta, en tanto que si se la registra en meses puede tratarse como continua.

Por otra parte, las variables numéricas se pueden clasificar en unidimensionales, cuando solo se mide un carácter o dato de los elementos de la muestra, o bidimensionales, tridimensionales, y en general n–dimensionales, cuando se estudian simultáneamente varios caracteres de cada elemento.

Por ejemplo, la temperatura o la presión atmosférica (por separado), son variables unidimensionales. La temperatura y la presión atmosférica (estudiadas conjuntamente), o la longitud y el peso de una barra conductora, son ejemplos de variables bidimensionales. La velocidad, carga eléctrica y masa de un ión es tridimensional.

3.- DISTRIBUCIONES DE FRECUENCIAS

El primer paso para el estudio estadístico de una muestra es su ordenación y presentación en una tabla de frecuencias.

3.1.- TABLA DE FRECUENCIA DE UNA VARIABLE DISCRETA

Supóngase que se tiene una muestra de tamaño N, donde la variable estadística x toma los valores distintos $x_1,\ x_2, \ldots\ x_k$. En primer lugar hay que ordenar los diferentes valores que toma la variable estadística en orden (normalmente creciente). La diferencia entre el valor mayor y menor que toma la variable se conoce como *recorrido o rango*.

En el caso de variables discretas, generalmente, un mismo valor de la variable aparecerá repetido más de una vez (es decir $k < N$). De forma que el siguiente paso es la construcción de una tabla en la que se indiquen los valores posibles de la variable y su frecuencia de aparición. Esta es la tabla de frecuencias de una variable discreta:

Valores de la variable estadística x_i	Frecuencias absolutas n_i	Frecuencias relativas f_i	Frecuencias absolutas acumuladas N_i	Frecuencias relativas acumuladas F_i
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
. . .	. . .	. . .	. . .	. . .

x_k	n_k	f_k	N_k	F_k

Si la variable es categórica, es posible también hacer una tabla de frecuencia. En este caso en la primera columna se deben escribir las distintas cualidades o atributos que puede tomar la variable.

En las siguientes columnas se escriben para cada valor de la variable:

Frecuencia absoluta n_i, definida como el número de veces que aparece repetido el valor en cuestión de la variable estadística; en el conjunto de las observaciones realizadas.

Si N es el tamaño de la muestra, las frecuencias absolutas cumplen las siguientes propiedades:

$$0 \leq n_i \leq N \quad y \quad \sum_{i=1}^{k} n_i = N.$$

Frecuencia relativa f_i, definida como el cociente entre la frecuencia absoluta y el tamaño de la muestra. Es decir que $f_i = \frac{n_i}{N}$

Las frecuencias relativas cumplen con las siguientes propiedades:

$$0 \leq f_i \leq 1 \quad y \quad \sum_{i=1}^{k} f_i = 1 \Leftrightarrow \sum_{i=1}^{k} \frac{n_i}{N} = 1 \Leftrightarrow \frac{\sum_{i=1}^{k} n_i}{N} = 1$$

Estas frecuencias también pueden expresarse en tantos por ciento del tamaño de la muestra, para lo cual simplemente debe multiplicarse por 100.

Por ejemplo, si el valor x_i de la variable x tiene por frecuencia relativa $f_i = 0{,}2$; significa que el valor x_i se repite en el 20% de la muestra.

Frecuencia absoluta acumulada N_i, definida como la suma de las frecuencias absolutas de los valores inferiores o iguales a x_i. Es decir que $N_i = \sum_{j=1}^{i} n_j$

Esta frecuencia se puede definir en forma recursiva: $N_i = N_{i-1} + n_i \ \ con \ \ N_1 = n_1$. Además el valor de la frecuencia acumulada del último valor será igual al tamaño de la muestra, vale decir que: $N_k = N$.

Frecuencia relativa acumulada F_i, definida como la suma de las frecuencias relativas de los valores inferiores o iguales a x_i. Es decir que $F_i = \sum_{j=1}^{i} f_j$.

Otra forma de definir esta frecuencia es mediante el cociente entre la frecuencia absoluta acumulada y el tamaño de la muestra.

$$F_i = \frac{N_i}{N} \Leftrightarrow F_i = \frac{\sum_{j=1}^{i} n_j}{N} \Leftrightarrow \sum_{j=1}^{i} \frac{n_j}{N} \Leftrightarrow \sum_{j=1}^{i} f_j$$

El valor de la frecuencia relativa acumulada del último valor será 1, o sea, $F_k = 1$. Esta frecuencia se puede expresar como un porcentaje y su significado será el tanto por ciento de medidas con valores por debajo o igual que x_i .

Ejemplo 1

Se registró el número de hijos de una muestra de 20 familias:

2 1 1 3 1 2 5 1 2 3

4 2 3 2 1 4 2 3 2 1.

Elaborar la tabla de frecuencias.

Variable: número de hijos que tiene la familia (x_i)

Tamaño de la muestra: 20 (N)

Número de valores posibles que puede asumir x_i: 5 (k).

Recorrido: 5 – 1 = 4

x_i	n_i	f_i $n_i/20$	N_i $\sum_{j=1}^{i} n_j$	F_i $\sum_{j=1}^{i} f_j$
1	6	0,30	6	0,30
2	7	0,35	13	0,65
3	4	0,20	17	0,85
4	2	0,10	19	0,95
5	1	0,05	20	1,00

$n_2 = 7$ significa que 7 familias tienen 2 hijos, es decir que el 35% $(f_i = 0{,}35)$ de las 20 familias tienen 2 hijos.

$N_3 = 17$ significa que 17 familias tienen 3 o menos hijos, es decir que el 85% $(F_i = 0{,}85)$ de las 20 familias tienen 3 o menos hijos.

Ejemplo 2

En la siguiente tabla se registró el género de 70 libros nuevos que ingresaron a una biblioteca.

x_i	n_i	f_i $n_i/70$	N_i $\sum_{j=1}^{i} n_j$	F_i $\sum_{j=1}^{i} f_j$
Narrativa	12	0,17	12	0,17
Biografía	5	0,07	17	0,24
Poesía	20	0,29	37	0,53
Cuento	23	0,33	60	0,86
Teatro	10	0,14	70	1,00

Variable: género del libro (x_i)

Tamaño de la muestra: 70 (N)

Número de categorías posibles que puede asumir x_i: 5 (k).

$n_2 = 5$ significa que 5 libros son del género biográfico, es decir que el 7% $(f_i = 0{,}07)$ de las 70 libros son del género biografía.

$N_4 = 60$ significa que 60 libros son ó bien del género narrativo ó del biográfico ó del poético ó de cuento, es decir que estas cuatro categorías conforman el 86% $(F_i = 0{,}86)$ de los libros.

3.2.- AGRUPAMIENTO EN INTERVALOS DE CLASES

Cuando el número de valores distintos que toma la variable estadística es demasiado grande o la variable es continua, no resulta útil elaborar una tabla de frecuencias como la vista anteriormente. En estos casos se puede realizar un *agrupamiento de los datos en intervalos* y se hace un recuento del número de observaciones que caen dentro de cada uno de ellos.

Estos intervalos se denominan *intervalos de clase* y al valor de la variable en el centro de cada intervalo se denomina *marca de clase*. De esta forma se sustituye cada medida por la marca de clase del intervalo a que corresponda. La diferencia entre el extremo superior e inferior de cada intervalo se denomina *amplitud del intervalo*. Normalmente se trabaja con intervalos de amplitud constante.

La tabla de frecuencias resultante es similar a la vista anteriormente. En el caso de una distribución en *k* intervalos ésta sería:

Intervalos de clase $a_i - a_{i+1}$	Marca de clase c_i	Frecuencias absolutas n_i	Frecuencias relativas f_i	Frecuencias absolutas acumuladas N_i	Frecuencias relativas acumuladas F_i
$a_1 - a_2$	c_1	n_1	f_1	N_1	F_1
$a_2 - a_3$	c_2	n_2	f_2	N_2	F_2
. . .	. . .	. . .	. . .	. . .	. . .
$a_k - a_{k+1}$	c_k	n_k	f_k	N_k	F_k

La ventaja de realizar el agrupamiento en intervalos de clase es la simplificación del trabajo pero, esto tiene por contrapartida la perdida de información ya que no se tiene en cuenta cómo se distribuyen los datos dentro de cada intervalo.

Para que dicha pérdida sea mínima es necesario elegir con cuidado los intervalos. Aunque no existen reglas estrictas para la elección de estos, los pasos a seguir son los siguientes:

a) Determinar el recorrido o rango de los datos. Vale decir, calcular la diferencia entre el mayor y el menor de los valores que toma la variable.

b) Decidir el número de intervalos de clase (k) en que se van a agrupar los datos. Por lo general $5 \leq k \leq 20$ dependiendo del caso que se estudia, k será más grande cuanto más datos posea la muestra. Una regla que se suele seguir es elegir k como el entero más próximo a $\sqrt{N}$, recordando que N es el tamaño de la muestra.

c) Determinar la amplitud (constante) de cada intervalo, dividiendo el recorrido o rango de los datos entre el número de intervalos (k). No es necesario que esta amplitud sea exactamente el resultado de esa división, sino que normalmente se puede redondear hacia un número ligeramente mayor.

d) Determinar los extremos de los intervalos de clase. Evidentemente el extremo superior de cada intervalo ha de coincidir con el extremo inferior del siguiente. Es importante que ninguna observación coincida con alguno de los extremos, para evitar así una ambigüedad en la clasificación de este dato. Una forma de conseguir esto es asignar a los extremos de los intervalos una cifra decimal más que las medidas de la muestra. Por ejemplo, si la variable estadística toma valores enteros: 10, 11, 12, etc., los extremos de los intervalos podrían ser: 9.5 – 11.5, 11.5 – 13.5, etc.

e) Calcular las marcas de clase de cada intervalo como el valor medio entre los límites inferior y superior de cada intervalo de clase. Aquí se debe intentar que las marcas de clase coincidan con medidas de la muestra, disminuyéndose así la pérdida de información debida al agrupamiento.
Una vez determinados los intervalos se debe hacer un recuento cuidadoso del número de observaciones que caen dentro de cada intervalo, para construir así la tabla de frecuencias.

Ejemplo
Se registró el peso de 80 alumnos de un curso perteneciente a un colegio del nivel medio de la localidad de San Salvador de Jujuy. Elaborar una tabla de frecuencias con datos agrupados en intervalos de clases.
60; 66; 77; 70; 66; 68; 57; 70; 66; 52; 75; 65; 69; 71; 58; 66; 67; 74; 61; 63; 69; 80; 59; 66; 70; 67; 78; 75; 64; 71; 81; 62; 64; 69; 68; 72; 82; 56; 65; 74; 67; 54; 65; 65; 69; 61; 67; 73; 57; 62; 67; 68; 63; 67; 71; 68; 76; 61; 62; 63; 76; 61; 67; 67; 64; 72; 64; 73; 79; 58; 67; 71; 68; 59; 69; 70; 66; 62; 63; 66.

a) Recorrido: $82 - 52 = 30$

b) $k = \sqrt{80} \cong 8{,}94 \implies k = 9$ Como se redondea por exceso, la amplitud del intervalo multiplicada por el número de intervalos será mayor que el recorrido y no se tendrá problemas en los extremos.

c) Amplitud del intervalo: $30/9 \cong 3,\hat{3} \cong 3{,}4$

d) Extremos de los intervalos. Para evitar coincidencias se toma un decimal más. El primer extremo se toma algo menor que el valor mínimo, pero calculándolo de forma que el último extremo sea algo mayor que el valor máximo.

Si se toma $a_1 = 51{,}5$ se verifica que es menor que 52 (valor mínimo) y el último extremo será $51{,}5 + 9 * 3{,}4 = 82{,}1$ que resulta ser mayor que el valor máximo, 82.

$a_i - a_{i+1}$	c_i	n_i	$f_i = n_i/N$	N_i	F_i
$51{,}5 - 54{,}9$	$53{,}2$	2	$0{,}025$	2	$0{,}025$
$54{,}9 - 58{,}3$	$56{,}6$	5	$0{,}0625$	7	$0{,}0875$
$58{,}3 - 61{,}7$	60	7	$0{,}0875$	14	$0{,}175$
$61{,}7 - 65{,}1$	$63{,}4$	16	$0{,}2$	30	$0{,}375$
$65{,}1 - 68{,}5$	$66{,}8$	21	$0{,}2625$	51	$0,6375$
$68{,}5 - 71{,}9$	$70{,}2$	13	$0{,}1625$	64	$0,8$
$71{,}9 - 75{,}3$	$73{,}6$	8	$0{,}1$	72	$0,9$
$75{,}3 - 78{,}7$	77	4	$0{,}05$	76	$0,95$
$78{,}7 - 82{,}1$	$80{,}4$	4	$0{,}05$	80	1

3.3.- Representaciones Gráficas

Luego de haber construido la tabla de frecuencias de una muestra, es conveniente la representación gráfica de la distribución de los datos. Esto permite una visualización rápida de la información recogida.

Dependiendo del tipo de datos y de cómo estén organizados, se pueden utilizar distintos tipos de representaciones gráficas.

a) Si se trata de una variable discreta sin agrupar, se usa principalmente el diagrama de barras. En este diagrama se representan sobre el eje de las abscisas los distintos valores de la variable y sobre cada uno de ellos se levanta una barra de longitud igual a la frecuencia correspondiente. Se pueden representarse tanto las frecuencias absolutas n_i como las relativas f_i. En la práctica se puede graduar simultáneamente el eje de las ordenadas tanto para frecuencias absolutas como para las relativas, estas últimas en tantos por ciento.

El diagrama anterior puede completarse con el polígono de frecuencias. Éste se obtiene uniendo con rectas los puntos medios de los extremos superiores de las barras del diagrama de barras. De la misma forma, pueden representarse frecuencias absolutas, relativas, o ambas a la vez.
En el ejemplo 1 del apartado "Tabla de frecuencia de una variable discreta" se elaboró la siguiente tabla, correspondiente al número de hijos de una muestra de 20 familias.

x_i	n_i	f_i	N_i	F_i
1	6	0,30	6	0,30
2	7	0,35	13	0,65
3	4	0,20	17	0,85
4	2	0,10	19	0,95
5	1	0,05	20	1,00

El diagrama de barras y polígono de frecuencias correspondientes, es el siguiente.

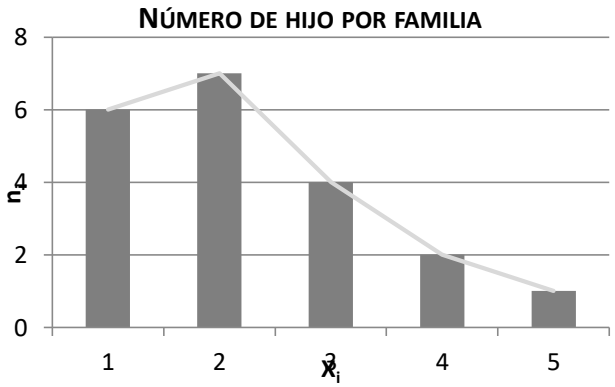


Grafico 1. Diagrama de barra y polígono de frecuencia. Ejemplo 1 del apartado "Tabla de frecuencia de una variable discreta".

Para representar las frecuencias (absolutas ó relativas) acumuladas se usa el diagrama de frecuencias acumuladas. Este gráfico, en forma de escalera se construye representando en el eje de las abscisas los distintos valores de la variable y levantando sobre cada x_i una perpendicular cuya longitud será la frecuencia acumulada $(N_i \text{ ó } F_i)$ de ese valor. Los puntos se unen con tramos horizontales y verticales. Evidentemente la escalera resultante ha de ser siempre ascendente.

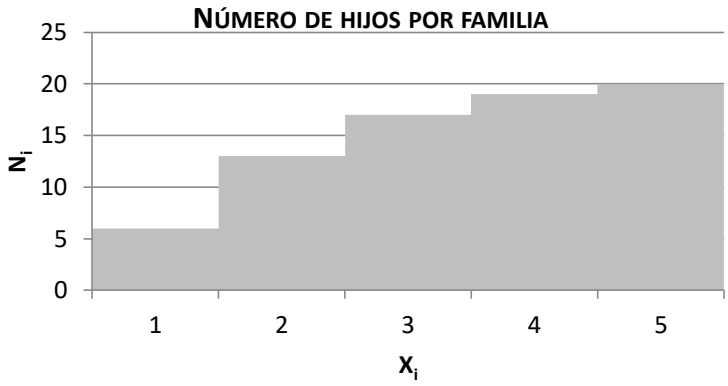


Grafico 2. Diagrama de frecuencias acumuladas (N_i). Ejemplo 1 del apartado "Tabla de frecuencia de una variable discreta".

b) Si se trata de datos agrupados la representación gráfica más utilizada es el histograma de frecuencias absolutas o relativas.

Un histograma es un conjunto de rectángulos adyacentes, cada uno de los cuales representa un intervalo de clase. La base de cada rectángulo es proporcional a la amplitud del intervalo. Por lo tanto el centro de la base de cada rectángulo corresponde a la marca de clase del intervalo que representa. La altura se suele determinar para que el área de cada rectángulo sea igual a la frecuencia de la marca de clase correspondiente.

En consecuencia, la altura de cada rectángulo se puede calcular como el cociente entre la frecuencia (absoluta o relativa) y la amplitud del intervalo. En el caso de que la amplitud de los intervalos sea constante, la representación es equivalente a usar como altura la frecuencia de cada marca de clase, siendo este método más sencillo para dibujar rápidamente un histograma.

Al igual que en las variables no agrupadas, otro tipo de representación es el polígono de frecuencias. Este se obtiene uniendo con líneas rectas los puntos medios de cada segmento superior de los rectángulos en el histograma.

En el ejemplo del apartado "Agrupamiento en intervalos de clases" se elaboró la siguiente tabla, correspondiente al peso de 80 alumnos de un curso perteneciente a un colegio del nivel medio de la localidad de San Salvador de Jujuy.

$a_i - a_{i+1}$	c_i	n_i	$f_i = n_i/N$	N_i	F_i
$51{,}5 - 54{,}9$	53,2	2	0,025	2	0,025
$54{,}9 - 58{,}3$	56,6	5	0,0625	7	0,0875

$58{,}3 - 61{,}7$	60	7	0,0875	14	0,175
$61{,}7 - 65{,}1$	63,4	16	0,2	30	0,375
$65{,}1 - 68{,}5$	66,8	21	0,2625	51	$0,6375$
$68{,}5 - 71{,}9$	70,2	13	0,1625	64	$0,8$
$71{,}9 - 75{,}3$	73,6	8	0,1	72	$0,9$
$75{,}3 - 78{,}7$	77	4	0,05	76	$0,95$
$78{,}7 - 82{,}1$	80,4	4	0,05	80	1

El histograma y polígono de frecuencias correspondientes, es el siguiente.

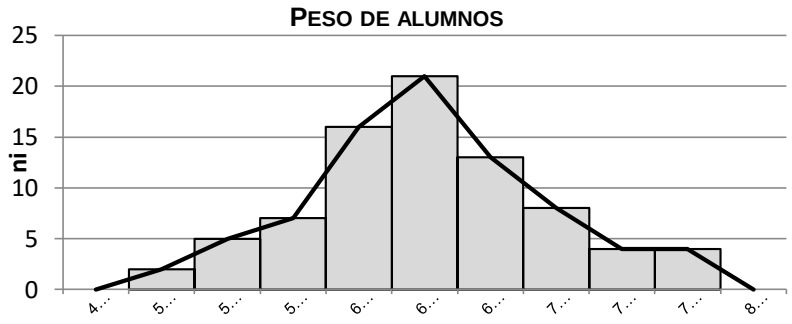


Grafico 3. Histograma y polígono de frecuencia. Ejemplo del apartado "Agrupamiento en intervalos de clases".

El polígono de frecuencias acumuladas sirve para representar las frecuencias acumuladas de datos agrupados por intervalos. En el eje de las abscisas se representan los diferentes intervalos de clase. Sobre el extremo superior de cada intervalo se levanta una línea vertical de altura igual a la frecuencia (absoluta o relativa) acumulada de ese intervalo. A continuación se unen por segmentos rectos los extremos de las líneas anteriores. El polígono parte de una altura cero para el extremo inferior del primer intervalo. Evidentemente, la altura que se alcanza al final del polígono es N, para frecuencias absolutas, o 1, para frecuencias relativas.

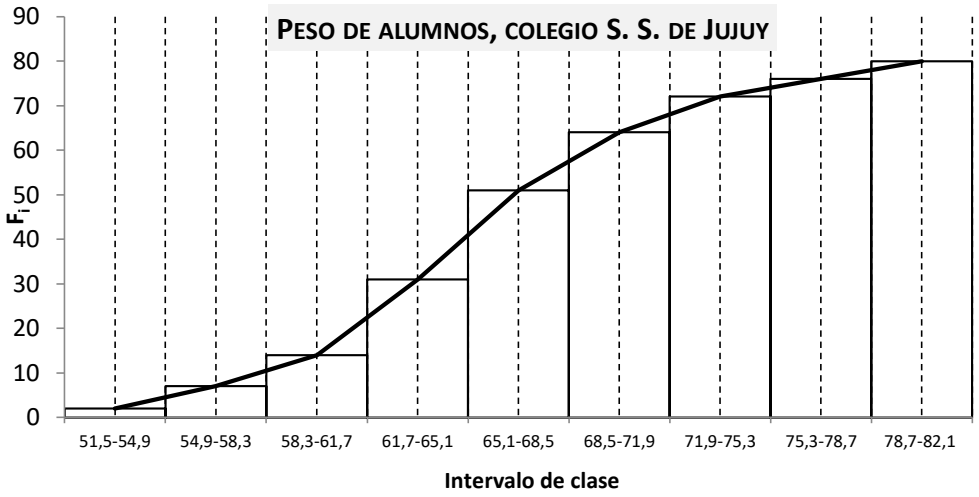


Grafico 4. Polígono de frecuencias acumuladas (N_i). Ejemplo del apartado "Agrupamiento en intervalos de clases".

c) Existe una gran variedad de representaciones gráficas para variables cualitativas, pero son dos las más usadas. El diagrama de rectángulos, que es similar al diagrama de barras y el histograma para las variables cuantitativas y consiste en representar en el eje de abscisas los diferentes caracteres cualitativos y levantar sobre cada uno de ellos un rectángulo (de forma no solapada) cuya altura sea la frecuencia (absoluta o relativa) de dicho carácter.

El otro diagrama muy usado es el diagrama de sectores, también llamado diagrama de torta. En él se representa el valor de cada carácter cualitativo como un sector de un círculo completo, siendo el área de cada sector (ó lo que es lo mismo, el arco subtendido) proporcional a la frecuencia del carácter en cuestión. Es habitual escribir dentro, o a un lado, de cada sector la frecuencia correspondiente. Este tipo de diagrama proporciona una idea visual muy clara de cuáles son los caracteres que más se repiten.

En el ejemplo 2 del apartado "Tabla de frecuencia de una variable discreta" se elaboró la tabla, correspondiente al género de 70 libros nuevos que ingresaron a una biblioteca.

El diagrama de rectángulos y diagrama de sectores correspondientes, son los siguientes.

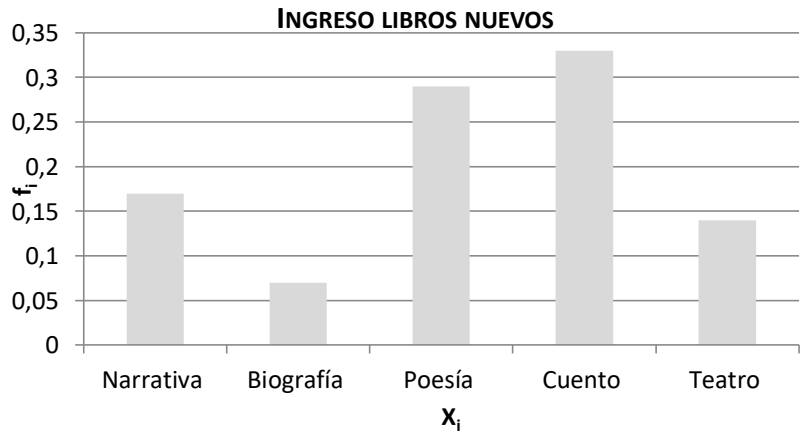


Grafico 5. Diagrama de rectángulos. Ejemplo 2 del apartado "Tabla de frecuencia de una variable discreta".

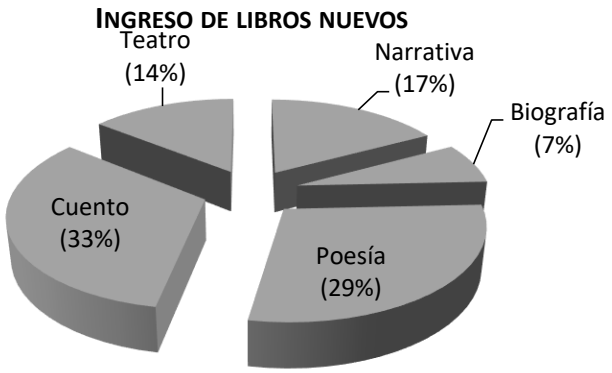


Grafico 6. Diagrama de sectores. Ejemplo 2 del apartado "Tabla de frecuencia de una variable discreta".

4.- MEDIDAS CARACTERÍSTICAS DE UNA DISTRIBUCIÓN

Después de haber construido tablas de frecuencias y haber realizado alguna representación gráfica, el siguiente paso para llevar a cabo un estudio de los datos recogidos; es el cálculo de diferentes medidas características de la distribución.

Se pueden calcular diversas medidas que son capaces de resumir toda la información recogida en un pequeño número de valores.

Resumir un conjunto de datos significa pasar de una visión detallada a una generalización simple e informativa tratando de preservar las características esenciales. Este proceso permite simplificar la comprensión y la comunicación de los datos.

Estas medidas resumen; van a permitir comparar distintas muestras y dar una idea rápida de cómo se distribuyen los datos. Es evidente que todas estas medidas solo pueden definirse para variables cuantitativas.

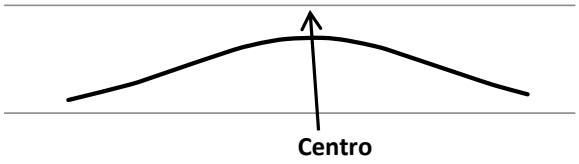
4.1.- MEDIDAS DE CENTRALIZACIÓN

Entre las medidas características de una distribución se destacan las llamadas medidas de centralización, que indican el valor promedio de los datos, o en torno a qué valor se distribuyen estos. Es decir que estas medidas describen un valor alrededor del cual se encuentran las observaciones.

Por lo tanto, una medida de centralización es un valor que pretende indicar dónde se encuentra el centro de la distribución de un conjunto de datos. Pero, ¿cómo identificar el centro de una distribución?

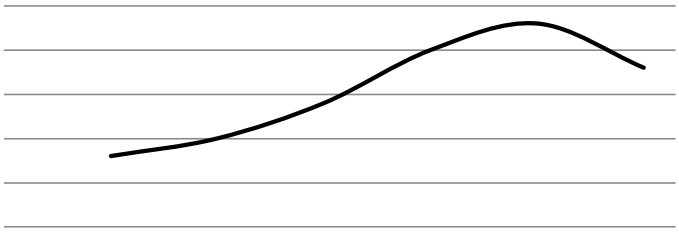
El centro es fácil de identificar si la distribución es simétrica.

DISTRIBUCIÓN SIMÉTRICA



Pero si la distribución es asimétrica, resulta difícil identificar el centro.

DISTRIBUCIÓN ASIMÉTRICA



Por esta razón, no existe una única medida de centralización para resumir una distribución. Si la distribución es simétrica diferentes medidas conducirán a resultados similares. Si la distribución es claramente asimétrica diferentes propuestas apuntarán a distintos concepto de "centro" y por lo tanto los valores serán diferentes.

Para salvar este inconveniente es necesario analizar las distintas medidas calculadas y ver cuál de ellas es la que mejor se adapta a la distribución de datos, que se analiza.

a) Media aritmética.

Supóngase que se tiene una muestra de tamaño N, donde la variable estadística x toma los valores $x_1,\ x_2, \ldots\ ; x_n$. Se define la *media aritmética* $\overline{x}$, o simplemente media, de la muestra como $\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$

Vale decir que la media es básicamente un promedio y se calcula sumando los distintos valores de la variable x y dividiendo por la cantidad de datos. En el caso de que los diferentes valores de la variable aparezcan repetidos, tomando los valores $x_1,\ x_2, \ldots\ ; x_k$ con frecuencias absolutas $n_1,\ n_2, \ldots\ n_k$, la media se determina como $\overline{x} = \frac{\sum_{i=1}^{k} x_i . n_i}{N}$, pudiéndose expresar también en función de las frecuencias relativas como: $\overline{x} = \sum_{i=1}^{k} x_i . f_i$

Continuando con el ejemplo 1 del apartado "Tabla de frecuencia de una variable discreta" calcúlese la media aritmética de los datos.

x_i	n_i	f_i	$x_i \,.\, n_i$	$x_i \,.\, f_i$
1	6	0,30	6	0,30

2	7	0,35	14	0,70
3	4	0,20	12	0,60
4	2	0,10	8	0,40
5	1	0,05	5	0,25
Total	20	1,00	45	2,25

$\overline{x} = \frac{\sum_{i=1}^{k} x_i . n_i}{N} = \frac{45}{20} = 2{,}25$ ó bien $\overline{x} = \sum_{i=1}^{k} x_i . f_i = 2{,}25.$

$\overline{x} = 2{,}25$ significa que en promedio las familias que intervinieron en la muestra tienen dos hijos.

En el caso de tener una muestra agrupada en k intervalos de clase la media se puede calcular, a partir de las marcas de clase c_i y el número n_i de datos en cada intervalo, utilizando la expresión: $\overline{x} = \frac{\sum_{i=1}^{k} c_i . n_i}{N}$. Nótese que la expresión anterior es solamente aproximada. Es más exacto para el cálculo de la media, no realizar el agrupamiento en intervalos y usar alguna de las expresiones anteriores.

Calcúlese la media aritmética del ejemplo del apartado "Agrupamiento en intervalos de clases".

c_i	n_i	$c_i . n_i$
53,2	2	106,4
56,6	5	283
60	7	420
63,4	16	1014,4
66,8	21	1402,8
70,2	13	912,6
73,6	8	588,8
77	4	308
80,4	4	321,6
Total	80	5357,6

$\overline{x} = \frac{\sum_{i=1}^{k} c_i . n_i}{N} = \frac{5357{,}6}{80} = 66{,}97$

Nótese la diferencia si se emplea la expresión:

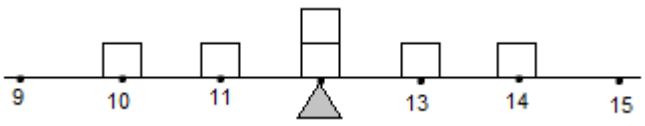
$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{5361}{80} = 67{,}0125$

Esto significa que en promedio el peso de los alumnos pertenecientes a un curso de un colegio del nivel medio de la localidad de San Salvador de Jujuy es de 67 kg.

Una propiedad importante de la media aritmética es que la suma de las desviaciones (ó distancias) de un conjunto de datos respecto a su media es cero. Es decir, la media equilibra las desviaciones positivas y negativas respecto a su valor. Esto se expresa como: $\sum_{i=1}^{N}(x-\overline{x})=0$
Por ejemplo, sean los datos: 10, 11, 12, 12, 13 y 14 cuya media es $\overline{x}=12$. En la siguiente tabla comprobamos, para este ejemplo, la propiedad enunciada.

x_i	$x_i-\overline{x}$
10	-2
11	-1
12	0
12	0
13	1
14	2
Total	0

Por lo tanto, una segunda propiedad de la media aritmética es que representa una especie de centro de gravedad, o centro geométrico, del conjunto de datos.
Se puede imaginar a los datos como un sistema físico en el que cada uno tiene una "masa" unitaria. Si se ubican los datos sobre una barra horizontal en la posición correspondiente a su valor; la media representa la posición en que se deberá ubicar el punto de apoyo para que el sistema esté en equilibrio.



Una tercera propiedad de la media como medida de tendencia central es que es poco "*robusta*", es decir depende mucho de valores atípicos de los datos. Si por ejemplo, en una muestra se introduce un nuevo dato con un valor mucho mayor que el resto, la media aumenta apreciablemente. Continuando con el ejemplo anterior 10, 11, 12, 12, 13 y 14 y $\overline{x}=12$, si ahora se tiene 10, 11, 12, 12, 13, 14 y 68 la media es $\overline{x}=20$. La media aritmética es por tanto muy dependiente de observaciones extremas.

Existen otras definiciones de media que pueden tener su utilidad en algunos casos. La primera de éstas es la *media geométrica* x_G. En el caso de una muestra con valores diferentes de la variable se define como la raíz enésima (N es el tamaño de la muestra) del producto de los valores de la variable

$\overline{x_G} = \sqrt[N]{x_1 \, . \, x_2 \, . \, . \, . \, x_N}$.

La *media armónica* $\overline{x_A}$ se define como la inversa de la media aritmética de las inversas de los valores de la variable. Es decir, para variables no agrupadas y agrupadas respectivamente, sería:

$\overline{x_A} = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$; $\overline{x_A} = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$

Una tercera definición, que tiene su utilidad con frecuencia en la aplicación a fenómenos físicos, corresponde a la *media cuadrática* $\overline{x_Q}$. Que se define como la raíz cuadrada de la media aritmética de los cuadrados de los valores.

$\overline{x_Q} = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}}$; $\overline{x_Q} = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 . n_i}{N}}$

En general ninguna de estas medias es muy robusta, aunque esto depende de cómo se distribuyan las variables. Por ejemplo, la media armónica es muy poco sensible a valores muy altos de la variable, mientras que a la media cuadrática apenas le afectan los valores muy bajos de la variable.

b) Mediana

Una medida de tendencia central importante es la mediana M_e; que se define como una medida central tal que, con los datos ordenados de menor a mayor, la mitad de estos son inferiores a su valor y la otra mitad tienen valores superiores. Es decir, la mediana divide en dos partes iguales la distribución de frecuencias ó, gráficamente, divide el histograma en dos partes de áreas iguales. Se distinguirá distintos casos para el cálculo de la mediana.

Supóngase, en primer lugar, que los diferentes valores de la variable no aparecen, en general, repetidos.

En este caso y suponiendo que se tienen los datos ordenados, la mediana será el valor central, si el tamaño de la muestra N, es impar ó la media aritmética de los dos valores centrales, si N es par.

Por ejemplo, si $x = 1, 4, 6, 7, 9$ entonces $M_e = 6$; por otro lado, si $x = 1, 4, 6, 7$ la mediana es $M_e = \frac{4+6}{2} = 5$.

En segundo lugar, supóngase que se tiene una variable discreta con valores repetidos sobre la cual se ha elaborado una tabla de frecuencias; se calcula en primer lugar el número de observaciones, N, dividido entre 2.

Se pueden distinguir aquí dos subcasos. El primero de ellos es cuando el valor $N/2$ coincide con la frecuencia absoluta acumulada N_j de un valor x_j de la variable ó, lo que es lo mismo, cuando la frecuencia relativa acumulada $F_j = 0{,}5$. En este caso la mediana se ha de situar entre este valor de la variable y el siguiente ya que de esta forma dividirá la distribución de frecuencias en dos partes. Es decir, se calcula como la media aritmética de dicho valor de la variable y su superior $M_e = \frac{x_j + x_{j+1}}{2}$

Se modificará levemente el ejemplo 1 del apartado "Tabla de frecuencia de una variable discreta" para calcular la mediana, acorde a lo enunciado precedentemente.

x_i	N_i
1	6
2	10
3	15
4	17
5	20

Se tiene 1 1 1 1 1 1 2 2 2 2 3 3 3 3 3 4 4 5 5 5 como $\frac{N}{2} = 10 = N_2$ entonces la mediana se calculará como: $M_e = \frac{x_2 + x_{2+1}}{2} = \frac{2+3}{2} = 2{,}5$

Si el valor $N/2$ no coincidiese con ningún valor de la columna de frecuencias acumuladas (segundo subcaso) la mediana sería el primer valor de x_j con frecuencia absoluta acumulada N_j mayor que $N/2$, ya que el valor central de la distribución correspondería a una de las medidas englobadas en ese x_j.

Continuando con el ejemplo 1 del apartado "Tabla de frecuencia de una variable discreta" calcúlese la mediana.

x_i	N_i
1	6
2	13
3	17
4	19
5	20

Se tiene 1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 4 4 5 como $\frac{N}{2} = 10$ entonces la mediana será el primer valor de x_i con frecuencia absoluta acumulada $N_i > 10$, es decir: $M_e = x_2 = 2.$

En tercer lugar, supóngase que se tiene una muestra de una variable continua cuyos valores están agrupados en intervalos de clase. En este caso pueden ocurrir dos situaciones. En primer lugar, si $N/2$ coincide con la frecuencia absoluta acumulada N_j de un intervalo $(a_j,\ a_{j+1})$ (con marca de clase c_j), la mediana será sencillamente el extremo superior a_{j+1} de ese intervalo. En el caso general de que ninguna frecuencia absoluta acumulada coincida con $N/2$ será necesario interpolar en el polígono de frecuencias acumuladas.

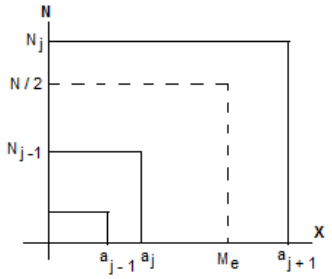


Grafico 7: Cálculo de la mediana.

Supóngase que el valor $N/2$ se encuentra entre las frecuencias N_{j-1} y N_j correspondientes a los intervalos $(a_{j-1},\ a_j)\ y\ (a_j,\ a_{j+1})$ respectivamente, la mediana se situará en algún lugar del intervalo superior $(a_j,\ a_{j+1})$. Según se muestra en el gráfico. Para calcular el valor exacto se debe interpolar[1]:

$\frac{a_{j+1}-a_j}{N_j-N_{j-1}} = \frac{M_e-a_j}{\frac{N}{2}-N_{j-1}}$ de donde se despeja M_e,

entonces

$M_e = \frac{a_{j+1}-a_j}{N_j-N_{j-1}} \cdot \left(\frac{N}{2} - N_{j-1}\right) + a_j$reordenando $M_e = a_j + \frac{\frac{N}{2}-N_{j-1}}{N_j-N_{j-1}}\left(a_{j+1}-a_j\right)$ y finalmente

$M_e = a_j + \frac{\frac{N}{2}-N_{j-1}}{n_j}\left(a_{j+1}-a_j\right)$

Calcúlese la mediana para el ejemplo del apartado “Agrupamiento en intervalos de clases”.

$a_i - a_{i+1}$	n_i	N_i
$51{,}5 - 54{,}9$	2	2
$54{,}9 - 58{,}3$	5	7

[1] En el análisis numérico, se denomina interpolación a la obtención de nuevos puntos partiendo del conocimiento de un conjunto discreto de puntos.

$58{,}3-61{,}7$	7	14
$61{,}7-65{,}1$	16	30
$65{,}1-68{,}5$	21	51
$68{,}5-71{,}9$	13	64
$71{,}9-75{,}3$	8	72
$75{,}3-78{,}7$	4	76
$78{,}7-82{,}1$	4	80

$\frac{N}{2}=40\neq N_i$

$(N_4=30)<\left(\frac{N}{2}=40\right)<(N_5=51)$ Por lo tanto la mediana se situará en el intervalo $65{,}1-68{,}5$ es decir que $65{,}1<M_e<68{,}5$.

$$M_e=a_j+\frac{\frac{N}{2}-N_{j-1}}{n_j}\left(a_{j+1}-a_j\right)=a_5+\frac{40-N_4}{n_5}(a_6-a_5)=65{,}1+\frac{40-30}{21}(68{,}5-65{,}1)$$

$$M_e=65{,}1+0{,}4762\ .3{,}4=66{,}72$$

Compárese este valor con el de la media aritmética $\overline{x}\cong 67$

Una de las propiedades de la mediana es que puede ser utilizada no sólo para datos numéricos sino además para datos ordinales, ya que para calcularla sólo es necesario establecer un orden en los datos.

Una segunda propiedad es que la mediana es insensible a la distancia de las observaciones al centro, ya que solamente depende del orden de los datos. Esta característica si bien la hace robusta, es una desventaja de la mediana.

Por ejemplo, todos los siguientes conjuntos de datos siguientes tienen mediana 12.

i) 10 11 12 13 14 ii) 10 11 12 13 100

iii) 0 11 12 12 12 iv) 10 11 12 100 100

Si se comparan las dos medidas de tendencia central estudiadas, la mediana tiene propiedades muy distintas respecto de la media aritmética, presentando sus ventajas e inconvenientes respecto de esta.

En primer lugar si la distribución de los datos es aproximadamente simétrica la media y la mediana serán aproximadamente iguales.

Pero si la distribución de los datos es asimétrica, la media y la mediana diferirán según el siguiente patrón:

Asimetría derecha (cola larga hacia la derecha), entonces $\overline{x}>M_e$

Asimetría izquierda (cola larga hacia la izquierda), entonces $\overline{x}<M_e$

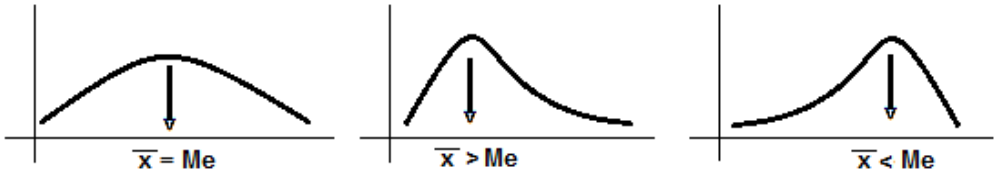
Por ejemplo, si la variable x toma los valores:

i) 12, 13, 14, 15, 16, entonces $\overline{x} = M_e = 14$

ii) 12, 13, 14, 15, 20, entonces $\overline{x} \cong 15 > M_e = 14$

iii) 2, 13, 14, 15, 16, entonces $\overline{x} = 12 < M_e = 14$

Gráficamente:



Por otro lado, la mayor ventaja de la media es que utiliza toda la información de la distribución de frecuencias (todos los valores particulares de la variable), en cambio la mediana solo utiliza el orden en que se distribuyen los valores de la variable. Podría pues considerarse, desde este punto de vista, que la media aritmética es una medida más fiable del valor central de los datos. Sin embargo recuérdese que la media es muy poco robusta, en el sentido de que es muy sensible a valores extremos de la variable y, en consecuencia, a posibles errores en las medidas.

La mediana, es una medida robusta, ya que no es afectada por valores que se desvíen mucho ó que sean atípicos.

Por ejemplo, supóngase que la variable x toma los valores: 2, 4, 5, 7 y 8, la media aritmética y la mediana serían en este caso muy parecidas: $\overline{x} = 5{,}2 \ \ y \ \ M_e = 5$. Pero sustitúyase el último valor 8 por 30, la nueva media se ve muy afectada $\overline{x} = 9{,}6$ no siendo en absoluto una medida de la tendencia central, mientras que el valor de la mediana no cambia. Pudiese ocurrir también el caso inverso.

Por ejemplo para el caso de las longitudes (en cm) de barras de hierro, inicialmente idénticas calentadas a temperaturas desconocidas en distintos recipientes: 1,80; 1,82; 1,85; 1,90 y 2,00, cuya media y mediana son $\overline{x} = 1{,}874 \ \ y \ \ M_e = 1{,}85$ respectivamente. Si la temperatura de uno de esos recipientes varía y la longitud mayor aumenta de 2,00 a 2,20 cm, la mediana no varía, pero la media ahora es $\overline{x} = 1{,}914$.

En general, lo mejor es considerar media aritmética y mediana como medidas complementarias. Es más, la comparación de sus valores puede suministrar información muy útil sobre la distribución de los datos.

c) Moda

Se define moda M_0 de una muestra como aquel valor de la variable que tiene una frecuencia máxima, es decir que la moda es el valor que más se repite. Hay que indicar que puede suceder que la moda no sea única, o sea que aparezcan varios máximos en la distribución de frecuencias, en ese caso se dice que la distribución es bimodal, trimodal, etc. Evidentemente, en el caso de una variable discreta que no tome valores repetidos, la moda no tiene sentido. Cuando sí existen valores repetidos su cálculo es directo ya que puede leerse directamente de la tabla de distribución de frecuencias.

Continuando con el ejemplo 1 del apartado "Tabla de frecuencia de una variable discreta", calcúlese la moda.

x_i	n_i	f_i	N_i	F_i
1	6	0,30	6	0,30
2	7	0,35	13	0,65
3	4	0,20	17	0,85
4	2	0,10	19	0,95
5	1	0,05	20	1,00

El valor que más se repite es 2 hijos, que ocurre en siete familias de la muestra $(n_i = 7)$. Por lo tanto la moda es $M_0 = 2$ y en este ejemplo coincide con la mediana.

En el caso de variables continuas agrupadas en intervalos de clase, existirá un intervalo en el que la frecuencia sea máxima, llamado intervalo modal. Es posible asociar la moda a un valor determinado de la variable dentro del intervalo modal. Para esto, supóngase que sea $(a_j;\ a_{j+1})$ el intervalo modal cuya frecuencia máxima es n_j. Si $n_{j-1}\ y\ n_{j+1}$ son las frecuencias de los intervalos anterior y posterior al modal, se define $\delta_1 = n_j - n_{j-1}\ y\ \delta_2 = n_j - n_{j+1}$ como se muestra en el siguiente gráfico.

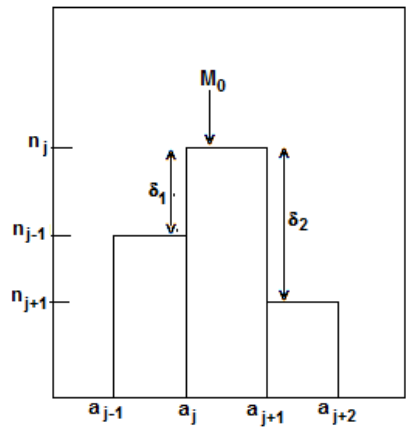


Grafico 8: Cálculo de la moda.

Puede demostrarse que el valor exacto de la moda es $M_0 = a_j + \frac{\delta_1}{\delta_1+\delta_2}(a_{j+1} - a_j)$. Es decir que la moda estará más próxima a a_j cuanto menor sea la diferencia de frecuencias con el intervalo anterior y al revés. Si, por ejemplo $n_{j-1} = n_j$ $(\delta_1 = 0)$ la moda será efectivamente a_j. Por el contrario si ocurre que $n_{j+1} = n_j$ $(\delta_2 = 0)$ la moda será a_{j+1} estando situada entre dos intervalos.

Calcúlese la moda del ejemplo del apartado "Agrupamiento en intervalos de clases".

$a_i - a_{i+1}$	c_i	n_i
$51{,}5 - 54{,}9$	$53{,}2$	2
$54{,}9 - 58{,}3$	$56{,}6$	5
$58{,}3 - 61{,}7$	60	7
$61{,}7 - 65{,}1$	$63{,}4$	16
$65{,}1 - 68{,}5$	$66{,}8$	21
$68{,}5 - 71{,}9$	$70{,}2$	13
$71{,}9 - 75{,}3$	$73{,}6$	8
$75{,}3 - 78{,}7$	77	4
$78{,}7 - 82{,}1$	$80{,}4$	4

El intervalo modal $(a_j;\ a_{j+1}) = (65{,}1 - 68{,}5)$

$j = 5,\ n_{j-1} = 16,\ n_j = 21\ \ y\ \ n_{j+1} = 13$

$\delta_1 = n_j - n_{j-1} = 21 - 16 = 5\ \ y\ \ \delta_2 = n_j - n_{j+1} = 21 - 13 = 8$

$M_0 = a_j + \frac{\delta_1}{\delta_1+\delta_2}(a_{j+1} - a_j) = 65{,}1 + \frac{5}{5+8}(68{,}5 - 65{,}1)$

$M_0 = 66{,}40$

Si la distribución de datos que se analiza fuese perfectamente simétrica, las tres medidas de tendencia central, media aritmética, mediana y moda coincidirían en el mismo valor. Sin embargo, cuando la distribución es claramente asimétrica, en general, la posición relativa entre las tres medidas suele ser la siguiente: la mediana se sitúa entre la moda y la media $M_0 < M_e < \overline{x}$.

Compruébese lo enunciado anteriormente para el ejemplo del apartado "Agrupamiento en intervalos de clases", para el cual se obtuvieron los siguientes valores $M_0(66{,}40) < M_e(66{,}72) < \overline{x}(66{,}97)$.

d) Cuartiles, deciles y percentiles

El concepto de mediana puede ser generalizado. Se vió que esta, es el valor de la variable que divide a la muestra, ordenada, en dos partes iguales.

Se puede definir de manera similar los *cuartiles* como aquellos tres valores que dividen a la muestra en cuatro partes iguales. De esta manera el primer cuartil $Q_{1/4}$ será la medida tal que el 25% de los datos sean inferiores a su valor y el 75% de los mismos sean superiores. El segundo cuartil $Q_{1/2}$ coincide con la mediana, mientras que el tercer cuartil $Q_{3/4}$ determinará el valor tal que las tres cuartas partes de las observaciones sean inferiores a él y una cuarta parte sea superior. La forma de calcular los cuartiles es igual a la ya vista para la mediana pero sustituyendo $\frac{N}{2}$ por $\frac{N}{4}\ y\ \frac{3N}{4}$ para $Q_{1/4}\ y\ Q_{3/4}$ respectivamente.

Continuando con el ejemplo 1 del apartado "Tabla de frecuencia de una variable discreta", calcúlese los cuartiles.

x_i	N_i
1	6
2	13
3	17
4	19
5	20

$\frac{N}{4} = \frac{20}{4} = 5 \implies Q_{1/4} = 1$ (1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 4 4 5)

$\frac{N}{2} = \frac{20}{2} = 10 \implies Q_{1/2} = M_e = 2$ (1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 4 4 5)

$\frac{3\,N}{4} = \frac{60}{4} = 15 \implies Q_{3/4} = 3$ (1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 4 4 5)

En el caso de las medidas agrupadas en intervalos de clase se trabaja de la misma manera que para determinar la mediana.

Calcúlese los cuartiles para el ejemplo del apartado "Agrupamiento en intervalos de clases".

$a_i - a_{i+1}$	n_i	N_i
$51{,}5 - 54{,}9$	2	2
$54{,}9 - 58{,}3$	5	7
$58{,}3 - 61{,}7$	7	14

$61{,}7-65{,}1$	16	30
$65{,}1-68{,}5$	21	51
$68{,}5-71{,}9$	13	64
$71{,}9-75{,}3$	8	72
$75{,}3-78{,}7$	4	76
$78{,}7-82{,}1$	4	80

$\frac{N}{4}=20<30$ por lo tanto $Q_{1/4}$ se sitúa en el intervalo $61{,}7-65{,}1$

$\frac{3\,N}{4}=60<64$ por lo tanto $Q_{3/4}$ se sitúa en el intervalo $68{,}5-71{,}9$

$$Q_{1/4}=a_j+\frac{\frac{N}{4}-N_{j-1}}{n_j}\left(a_{j+1}-a_j\right)=61{,}7+\frac{20-14}{16}(65{,}1-61{,}7)=62{,}975$$

$$Q_{3/4}=a_j+\frac{\frac{3\,N}{4}-N_{j-1}}{n_j}\left(a_{j+1}-a_j\right)=68{,}5+\frac{60-51}{13}(71{,}9-68{,}5)=70{,}854$$

De forma similar se puede definir los *deciles* como aquellos valores de la variable que dividen la muestra, ordenada, en diez partes iguales. Estos valores, denotados por $D_k,\ con\ k=1,\ 2,\ \ldots\ 9$, tienen un valor tal que el decil k–esimo deja por debajo de él al $10k$ por ciento de los datos de la muestra. De la misma manera se definen los *percentiles*, también llamados *centiles*, como aquellos valores de la variable denotados por $P_k,\ con\ k=1,\ 2,\ldots\ ,99$ que dividen a la muestra en cien partes iguales. Esto equivale a decir que el percentil P_k deja por debajo de él al k por ciento de la muestra ordenada.

La forma de calcular deciles y percentiles es igual a la de la mediana y los cuartiles, sustituyendo $\frac{N}{2}$ por la fracción del número total de datos correspondiente. Evidentemente algunos valores de cuartiles, deciles y centiles coinciden, cumpliéndose por ejemplo: $P_{50}=D_5=Q_{1/2}=M_e$.

4.2.- MEDIDAS DE DISPERSIÓN

Las medidas de tendencia central reducen la información recogida de la muestra a un solo valor; dando una idea de dónde se encuentra el centro de la distribución. Sin embargo, dicho valor central o medio, será más o menos representativo de los valores de la muestra dependiendo de la dispersión que las medidas individuales tengan respecto a dicho centro. Es decir que las medidas de tendencia central no indican cuán disperso es el conjunto de datos. Por ejemplo considérese los siguientes conjuntos de datos:

Muestra A: 55 55 55 55 55 55 55

Muestra B: 47 51 53 55 57 59 63

Muestra C: 39 47 53 55 57 63 71

En los tres casos $\overline{x}=M_e=55$, pero, como es evidente, las muestras difieren notablemente.

Para analizar la representatividad de las medidas de centralización se definen las llamadas medidas de dispersión. Estas indican la variabilidad de los datos en torno a su valor promedio, es decir si se encuentran muy o poco esparcidos en torno a su centro.
Se pueden definir diversas medidas de desviación o dispersión, siendo éstas fundamentales para la descripción estadística de la muestra.

a) Rango o recorrido

Una evaluación rápida de la dispersión de los datos se puede realizar calculando el *rango ó recorrido* ó diferencia entre el valor máximo y mínimo que toma la variable estadística.
El rango de n observaciones $x_1,\ x_2, \ldots\ ,\ x_n$ es la diferencia entre el valor máximo y mínimo que toma la variable. $R = max(x_i) - min(x_i)$
Calcúlese el rango para cada una de las muestras dadas en el ejemplo anterior.

Una de las propiedades del rango es de ser una medida extremadamente sensible a la presencia de datos atípicos, de existir estos datos, estarán en los extremos que son los datos que se usan para calcular el rango.
Una segunda característica es la de ignorar la mayoria de los datos puesto que solo usa solo dos observaciones: la mayor y la menor.
Con el fin de eliminar la excesiva influencia de los valores extremos en el recorrido, se puede definir el recorrido intercuartílico como la diferencia entre el tercer y primer cuartil: $R_I = Q_{3/4} - Q_{1/4}$
Está claro que este recorrido brinda, entonces, el rango que ocupan el 50% central de los datos.
En ocasiones se puede tambien, utilizar el recorrido semiintercuartílico, o mitad del recorrido intercuartílico: $R_I = \frac{Q_{3/4} - Q_{1/4}}{2}$

b) Desviación media

Otra manera de estimar cuan dispersos estan los valores de la muestra, es comparar cada uno de ellos con el valor de una medida de centralización. Una de las medidas de dispersión más usada es la *desviación media*, también llamada con más precisión *desviación media respecto a la media aritmética*. Ésta medida se define como la media aritmética de las diferencias absolutas entre los valores de la variable y la media aritmética de la muestra. Suponiendo que en una muestra de tamaño N los k distintos valores x_i de la variable tengan frecuencias absolutas n_i, la expresión de la desviación media será: $D_{\overline{x}} = \frac{\sum_{i=1}^{k} | x_i - \overline{x} | n_i}{N}$
Si la variable no toma valores repetidos ni esta agrupada en intervalos, la expresión anterior se reduce a: $D_{\overline{x}} = \frac{\sum_{i=1}^{k} | x_i - \overline{x} |}{N}$

Destáquese la importancia de tomar el valor absouto de las desviaciones, ya que si no se hiciese así, unas desviaciones se anularían con otras alcanzando finalmente la desviación media un valor 0, debido a la propiedad de la media aritmética ya vista.

Se puede definir una desviación media en términos de desviaciones absolutas en torno a una medida de centralización diferente de la media aritmética. Cuando se utiliza la mediana, se obtiene la llamada desviación media respecto a la mediana, definida como: $D_{M_e} = \frac{\sum_{i=1}^{k} | x_i - M_e | n_i}{N}$

Continuando con el ejemplo 1 del apartado "Tabla de frecuencia de una variable discreta" calcúlese la desviación media de los datos, cuya media aritmética es $\overline{x} = 2{,}25$

x_i	n_i	$\lvert x_i - \overline{x} \rvert n_i$
1	6	7,5
2	7	1,75
3	4	3
4	2	3,5
5	1	2,75
Total	20	18,5

$$D_{\overline{x}} = \frac{\sum_{i=1}^{k} | x_i - \overline{x} | n_i}{N} = \frac{18{,}5}{20} = 0{,}925$$

Continuando con el ejemplo del apartado "Agrupamiento en intervalos de clases" calcúlese la desviación media de los datos, cuya media aritmética es $\overline{x} = 67{,}0125$. Como los datos estan agrupados en intervalos, x_i representa la marca de clase del i-ésimo intervalo.

c_i	n_i	$\lvert x_i - \overline{x} \rvert n_i$
53,2	2	27,625
56,6	5	52,0625
60	7	49,0875
63,4	16	57,8
66,8	21	4,4625
70,2	13	41,4375
73,6	8	52,7
77	4	39,95
80,4	4	53,55

Total	80	378,675

$$D_{\overline{x}} = \frac{\sum_{i=1}^{k} | x_i - \overline{x} | n_i}{N} = \frac{378{,}675}{80} = 4{,}7334$$

El valor absoluto de la desviación, respecto a la media, de un dato en particular indica lo lejos que este dato está del valor de la media. Un valor igual a cero indica que el dato coincide con la media, mientras que un valor elevado con respecto a las demás desviaciones informa de que el dato está alejado de los demás.

Ahora bien, como la desviación media es la media aritmética de los valores absolutos de las desviaciones respecto a la media, esta medida informa de lo muy dispersados – ó no – que están los datos. Una desviación media elevada implica mucha variabilidad en los datos, mientras que una desviación media igual a cero implica que todos los valores son iguales y por lo tanto coinciden con la media.

c) Varianza y desviación estándar

Sin lugar a dudas la medida más usada para estimar la dispersión de los datos es la desviación estándar. Ésta es especialmente aconsejable cuando se usa la media aritmética como medida de tendencia central. Al igual que la desviación media, está basada en un valor promedio de las desviaciones respecto a la media.

En este caso, en vez de tomar valores absolutos de las desviaciones, para evitar así que se compensen desviaciones positivas y negativas, se usan los cuadrados de las desviaciones. Esto hace además que los datos con desviaciones grandes influyan mucho en el resultado final.

Por lo tanto se define, en primer lugar, *la varianza* de una muestra con datos repetidos de la siguiente manera: $s^2 = \frac{\sum_{i=1}^{k} (x_i - \overline{x})^2 n_i}{N-1}$

Evidentemente la varianza no tiene las mismas unidades que los datos de la muestra. Para conseguir las mismas unidades se define la *desviación estándar* (algunas veces llamada desviación típica) como la raíz cuadrada de la varianza, o sea que: $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \overline{x})^2 n_i}{N-1}}$

Si los datos no se repiten, estas definiciones se simplifican a:

$$s^2 = \frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N-1} \quad y \quad s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N-1}}$$

Se puede definir varianza y desviación estandar utilizando N en vez de $N-1$ en el denominador, representando entonces la varianza una verdadera media aritmética del cuadrado de las desviaciones. Está claro que ambas definiciones llevan a valores muy parecidos cuando N es grande. El motivo de haber optado por la definición con $N-1$ es que ésta da una mejor estimación de la dispersión de los datos.

Téngase en cuenta que como la suma de las desviaciones $x_i - \overline{x}$ es siempre 0, la desviación del último dato puede calcularse una vez que se conozcan las $N-1$ anteriores. Esto quiere decir que sólo se tienen $N-1$ desviaciones independientes y se promedia entonces dividiendo por $N-1$, ya que no tiene mucho sentido promediar N números no independientes. Notesé además, cuando solo se tiene un dato $(N = 1)$, en el caso de la definición con N en el denominador se obtendría una varianza 0, que no tiene mucho sentido, mientras que en la definición con $N-1$ la varianza estaría indeterminada.

Ejemplos

Continuando con el ejemplo 1 del apartado "Tabla de frecuencia de una variable discreta" calcúlese la varianza y la desviación estándar de los datos, cuya media aritmética es $\overline{x} = 2{,}25$

x_i	n_i	$(x_i - \overline{x})^2 n_i$
1	6	9,375
2	7	0,4375
3	4	2,25
4	2	6,125
5	1	7,5625
Total	20	25,75

$s^2 = \frac{\sum_{i=1}^{k}(x_i-\overline{x})^2 n_i}{N-1} = \frac{25{,}75}{19} = 1{,}355$ por lo tanto $s = \sqrt{1{,}355} = 1{,}16$

Continuando con el ejemplo del apartado "Agrupamiento en intervalos de clases" calcúlese la varianza y desviación estándar de los datos, cuya media aritmética es $\overline{x} = 67{,}0125$. Como los datos estan agrupados en intervalos, x_i representa la marca de clase del i-ésimo intervalo.

c_i	n_i	$(x_i - \overline{x})^2 n_i$
53,2	2	381,57
56,6	5	542,10
60	7	344,23
63,4	16	208,80
66,8	21	0,9483

70,2	13	132,08
73,6	8	347,16
77	4	399,00
80,4	4	716,90
Total	80	3073,51

$s^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i}{N-1} = \frac{3073,51}{79} = 38,90$ por lo tanto $s = \sqrt{38,90} = 6,24$

La desviación estándar s es útil para comparar la variabilidad de dos conjuntos de datos en los que la variable a sido medida en las mismas unidades. Por ejemplo si en una muestra $s = 2,3$ y en otra $s = 8,4$ se puede asegurar que los datos de la segunda muestra están más dispersos que los de la primera. Pero ¿cómo se interpreta el valor $s = 2,3$?

La desviación estándar da la idea de la distancia promedio de los datos a la media (estrictamente hablando no es el promedio). Pero la interpretación de s requiere algún conocimiento de la distribución de los datos. Es por ello que se puede dar la siguiente regla empírica.

Si el histograma de los datos es aproximadamente simétrico y acampanado entonces:

Aproximadamente el 68% de las observaciones caen en el intervalo $\overline{x} - s \quad y \quad \overline{x} + s$

Aproximadamente el 95% de las observaciones caen en el intervalo $\overline{x} - 2s \; y \; \overline{x} + 2s$

Prácticamente todas las observaciones caen en el intervalo $\overline{x} - 3s \quad y \quad \overline{x} + 3s$

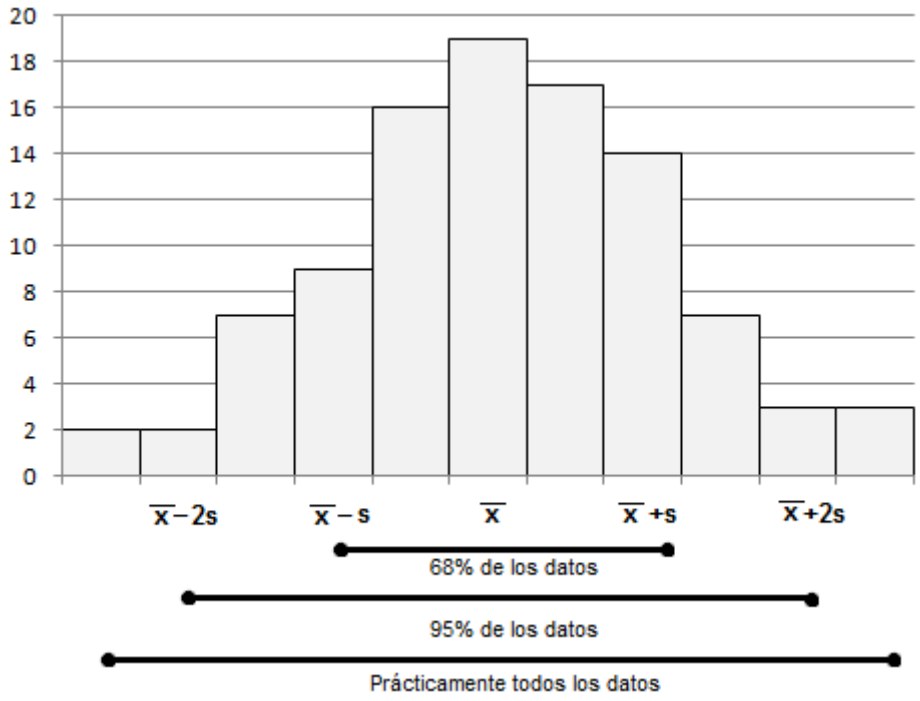


Grafico 9: Regla práctica para la interpretación de la desviación estándar.

Esta regla es válida para distribuciones no necesariamente acampanadas, pero puede ser errónea cuando se aplica a distribuciones fuertemente asimétricas.

Nótese que la desviación estándar no es una medida robusta de la dispersión. El hecho de que se calcule evaluando los cuadrados de las desviaciones hace que sea muy sensible a observaciones extremas, bastante más que la desviación media (dado que aparece un cuadrado). O sea que, la desviación estándar no es una buena medida de dispersión cuando existe algún dato muy alejado de la media. El rango intercuartílico daría, en ese caso, una idea más aproximada de cuál es la dispersión de los datos. El que la desviación estándar sea la medida de dispersión más común se debe a su íntima conexión con la distribución normal.
Por último la desviación estándar valdrá cero solamente cuando todos los datos son iguales, de otro modo es positiva.

d) Coeficientes de variación

Las medidas de dispersión vistas presentan un inconveniente ya que vienen expresadas en las unidades en que se ha medido la variable. Es decir, son medidas absolutas y con el único dato de su valor no es posible decir si se tiene una dispersión importante o no. Para solucionar esto, se definen unas medidas de dispersión relativas, independiente de las unidades usadas. Estas dispersiones relativas van a permitir además comparar la dispersión entre diferentes muestras (con unidades diferentes). Entre estas medidas hay que destacar el *coeficiente de variación de Pearson,* definido como el cociente entre la desviación estándar y la media aritmética: $CV = \frac{s}{|\overline{x}|}$

Obviamente este coeficiente no se puede calcular cuando $\overline{x} = 0$. Normalmente CV se expresa en porcentaje, multiplicando su valor por 100. Evidentemente, cuanto mayor sea CV, mayor dispersión tendrán los datos.

Continuando con el ejemplo 1 del apartado "Tabla de frecuencia de una variable discreta" calcúlese el coeficiente de variación de los datos, cuya media aritmética es $\overline{x} = 2{,}25$ y desviación estándar $s = 1{,}16$

$$CV = \frac{s}{|\overline{x}|} = \frac{1{,}16}{2{,}25} = 0{,}515 \cong 52\%$$

Para el ejemplo del apartado "Agrupamiento en intervalos de clases" calcúlese el coeficiente de variación de los datos, cuya media aritmética es $\overline{x} = 67{,}0125$ y desviación estándar $s = 6{,}24$

$$CV = \frac{s}{|\overline{x}|} = \frac{6{,}24}{67{,}0125} = 0{,}093 \cong 9{,}3\%$$

También se pueden definir otras medidas de dispersión relativas, como el *coeficiente de variación media*. Éste es similar al coeficiente de variación de Pearson, pero empleando una desviación media en vez de la desviación estándar. Por lo tanto se tienen dos coeficientes de variación media

dependiendo de que se calcule la desviación media respecto a la media aritmética ó respecto a la mediana:

$$CVM_{\overline{x}} = \frac{D_{\overline{x}}}{|\overline{x}|} \quad y \quad CVM_{Me} = \frac{D_{Me}}{|M_e|}$$

4.3.- ASIMETRÍA Y CURTOSIS

La descripción estadística de una muestra de datos incluye además de las medidas de tendencia central y de dispersión, el grado de simetría de los datos respecto a su medida central y la concentración de estos alrededor de dicho valor. De esta forma se dará una descripción completa de la muestra.

a) Coeficientes de asimetría

Se dice que una distribución de medidas es simétrica cuando, valores de la variable equidistantes del valor central, tienen la misma frecuencia. Es decir, en este caso se tendría simetría en el histograma (ó en el diagrama de barras) alrededor de una recta vertical trazada por el punto central. En el caso de una distribución perfectamente simétrica los valores de la media aritmética, mediana y moda coinciden $(\overline{x} = M_e = M_o)$. Esto se muestra en el siguiente gráfico.

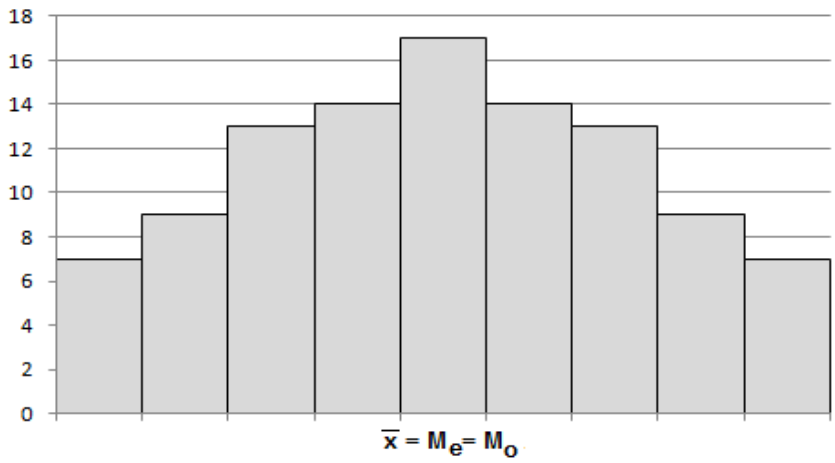


Grafico 10: Distribución simétrica.

En el caso de no tener simetría, se tiene entonces, asimetría a la derecha (ó positiva) ó a la izquierda (ó negativa) dependiendo de que el histograma muestre una cola de medidas hacia valores altos o bajos de la variable, respectivamente. También se puede decir que la distribución está sesgada a la derecha (sesgo positivo) ó a la izquierda (sesgo negativo). En el caso de una distribución asimétrica, la media, mediana y moda no coinciden, siendo $\overline{x} \geq M_e \geq M_o$ para una asimetría positiva y $\overline{x} \leq M_e \leq M_o$ para una asimetría negativa. Como se muestra en el siguiente gráfico.

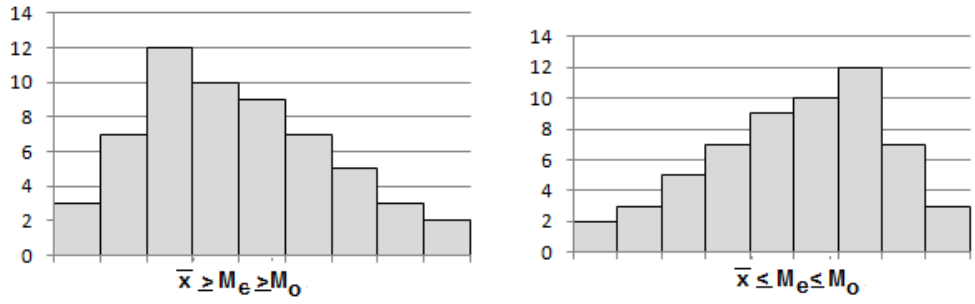


Grafico 11: Distribución asimétrica a la derecha y a la izquierda.

Con el fin de cuantificar el grado de asimetría de una distribución se pueden definir los coeficientes de asimetría. Aunque no son los únicos, existen dos coeficientes principales:

Coeficiente de asimetría de Fisher, que define como: $A_F = \frac{\frac{\sum_{i=1}^{k}(x_i-\overline{x})^3 n_i}{N}}{s^3}$

En el caso de una distribución simétrica, las desviaciones respecto a la media aritmética se anularán y el coeficiente de asimetría será nulo. En caso contrario, A_F tendrá valores positivos para una asimetría positiva (a la derecha) y negativos cuando la asimetría sea en el otro sentido. Nótese que la división por el cubo de la desviación estándar se hace para que el coeficiente sea adimensional y, por lo tanto, comparable entre diferentes muestras.

Coeficiente de asimetría de Pearson. Este coeficiente también adimensional se define como: $A_P = \frac{\overline{x}-M_O}{s}$

Su interpretación es similar a la del coeficiente de Fisher, siendo nulo para una distribución simétrica y tanto más positivo, ó negativo, cuando más sesgada esté la distribución hacia la derecha, ó hacia la izquierda.

Continuando con el ejemplo 1 del apartado "Tabla de frecuencia de una variable discreta" calcúlese el coeficiente de asimetría de los datos, cuya media aritmética es $\overline{x} = 2{,}25$, desviación estándar $s = 1{,}16$ y moda es $M_O = 2$

x_i	n_i	$(x_i - \overline{x})^3 n_i$
1	6	-11, 7188
2	7	-0,1094
3	4	1,6875
4	2	10, 7188
5	1	20,7969
Total	20	21,375

$$A_F = \frac{\frac{\sum_{i=1}^{k}(x_i-\overline{x})^3 n_i}{N}}{s^3} = \frac{\frac{21{,}375}{20}}{1{,}5609} = 0{,}685 \qquad A_P = \frac{\overline{x}-M_O}{s} = \frac{2{,}25-2}{1{,}16} = 0{,}215$$

b) Coeficientes de curtosis

Otra característica importante de la forma en la que se distribuyen los datos de la muestra, además de la simetría, es cómo se agrupan en torno al valor central. Los datos se pueden distribuir de forma que se tenga un gran apuntamiento ó pico, alrededor del valor central, en cuyo caso se dice que la distribución es *leptocúrtica*[2], ó en el extremo contrario, la distribución puede ser muy aplanada, lo que se caracteriza diciendo que es *platicúrtica*[3]. En el caso intermedio, se dice que la distribución es *mesocúrtica*[4] y el agrupamiento corresponderá al de una distribución llamada normal, o en forma de campana de Gauss.

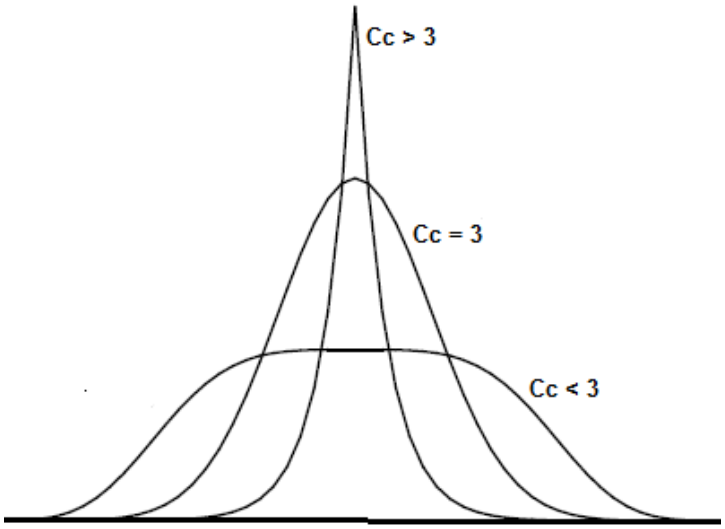


Grafico 12: Distribuciones con diferentes grados de apuntamiento.

Esta característica del agrupamiento de los datos se denomina curtosis y para cuantificarla se define el *coeficiente de curtosis* C_c de la siguiente manera:

$$C_c = \frac{\frac{\sum_{i=1}^{k}(x_i-\overline{x})^4 n_i}{N}}{s^4}$$

Este coeficiente sin dimensión alcanza valores mayores cuanto más puntiaguda es la distribución, teniendo un valor de 3 para la distribución mesocúrtica (o normal), mayor que 3 para la leptocúrtica y menor que 3 para la platicúrtica.

[2] El prefijo griego “lepto” significa delgado, fino. Curtosis o apuntamiento. Leptocurtica, un apuntamiento alargado.

[3] Plati, prefijo procedente del griego “platys” que significa ancho.

[4] Meso, prefijo procedente del griego “mésos” que signfica medio.