UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

ESTADÍSTICA DESCRIPTIVA

Jorge Federico Paniagua Ballinas
Alejandra Vargas Espinoza de los Monteros

Colaboración:
Dra. Isabel Patricia Aguilar Juárez

DIVISIÓN DE CIENCIAS BÁSICAS

PANIAGUA BALLINAS, Jorge Federico y
VARGAS ESPINOZA de los MONTEROS, Alejandra
Estadística descriptiva
Universidad Nacional Autónoma de México
Facultad de Ingeniería, 2024, 172 págs.

ESTADÍSTICA DESCRIPTIVA

Primera edición electrónica de un ejemplar (12 MB) en formato PDF
Publicado en línea: en octubre de 2024

FACULTAD DE INGENIERÍA
http://www.ingenieria.unam.mx/

Hecho en México.

UNIDAD DE APOYO EDITORIAL
Cuidado de la edición: Elvia Angélica Torres Rojas
Diseño editorial : Luis Enrique Vite Rangel

Prólogo

La aparición de la Matemática en la historia del ser humano está asociada al concepto de número, una circunstancia que se dio en las primeras comunidades primitivas. Es posible que, desde sus inicios, el ser humano haya tenido noción de la dimensión o tamaño de las cosas y de la cantidad de ellas; ideas que se fueron desarrollando con el transcurso del tiempo.

Más adelante, los grupos de personas y los conjuntos de cosas fueron cada vez más grandes. Así, con la formación de sociedades más complejas y la aparición del comercio, el concepto de número se fue mejorando para contar y operar múltiples cantidades, como mercancías, habitantes de una población, miembros del ejército y muchas otras más. Esta necesidad de contar impulsó la creación de asignar nombres a los símbolos numéricos y a las operaciones con ellos. Los vestigios más antiguos encontrados hasta ahora que dan testimonio de ello son la tablilla de barro de Plimpton 322 (1800 a.C.) y el papiro de Moscú (1850 a.C.).

Los estudiosos de la historia de la Matemática coinciden en que esta ciencia básica surgió como consecuencia de necesidades humanas que debían irse resolviendo, tales como medir predios o áreas de tierra, hacer cálculos de la compra y venta de mercancías, y tener herramientas conceptuales para construir aparatos y obtener conocimientos astronómicos, producir armamento bélico, entre otras.

La Matemática, en su desarrollo tan extenso e importante para los diferentes fines de la sociedad, ha tenido que subdividirse en varias ramas, como la Aritmética, el Álgebra, el Cálculo Diferencial e Integral, el Álgebra Lineal, las Ecuaciones Diferenciales, el Cálculo Vectorial, la Topología, la Probabilidad y, desde luego, la Estadística.

Es común considerar que la Estadística se inicia con el sustento de los principios de la teoría de la Probabilidad, merced al intercambio de información entre Pierre Fermat y Blaise Pascal a mediados del siglo XVII, o con los conteos que se hacían en los juegos de azar. Sin embargo, se han encontrado juegos de azar con más de 40,000 años de antigüedad. Se tienen registros de que los sumerios, que inventaron la escritura cuneiforme, ya utilizaban huesos en forma semejante a la de un dado para los juegos de azar. La Estadística es una ciencia aplicada que toma su nombre de “Ciencia del Estado”, pues eran los gobernantes quienes empezaron a llevar el registro de sus cosechas, población, recaudación fiscal, mili-

cia, etc., y más adelante, las metodologías estadísticas se ampliaron a conjuntos de datos de cualquier otro campo o actividad del conocimiento humano, como el comportamiento del clima, el nivel de producción de toneladas de acero, o la paridad de una moneda con respecto a otra, solo por mencionar unos ejemplos entre millones que pueden existir. Actualmente, el desarrollo de la estadística tiene su principal aplicación en el análisis e interpretación de datos, así como en la conclusión de parámetros o distribuciones poblacionales a partir de la información contenida en una muestra.

La Probabilidad y la Estadística cobraron importancia en el mundo científico cuando Albert Einstein, en 1905, presentó su famoso artículo sobre el movimiento browniano, denominado así porque se le atribuye al biólogo Robert Brown (1827). En ese artículo, Einstein explica de manera específica que Brown había observado en el microscopio cómo el polen era movido por moléculas individuales de agua, lo que constituyó para Einstein un testimonio convincente de que existen átomos y moléculas. El movimiento browniano es resultado de procesos estocásticos o aleatorios simples, y está relacionado con procesos estocásticos más simples como el conocido límite aleatorio planteado en el Teorema de Donsker, apoyado en la universalidad de la distribución normal.

La ciencia estadística, desde mediados del siglo XX, ha sido requerida para conocer las necesidades de alimentos, el control de inventarios de todo tipo, desde los bélicos hasta los civiles. Asimismo, coadyuva en el desarrollo industrial, tecnológico y de servicios (médicos, educativos, ambientales, deportivos, electorales, etc.). En el siglo XXI, continúa su desarrollo, ahora orientado a la inteligencia artificial y el análisis de grandes datos, que empezamos a observar a partir de la enorme cantidad de información que se tiene de cualquier ámbito.

La ingeniería de hoy nos maravilla con las comunicaciones de facto instantáneas. Podemos entrar en contacto con nuestro centro de trabajo, familiares y amigos desde dos puntos muy lejanos del planeta, e incluso fuera de él, en una fracción de segundo. La automatización de los procesos de fabricación de automóviles, aviones, trenes y barcos rebasa la ficción de hace unos pocos años. La mínima intervención humana, la toma de decisiones y los estudios de mercado de esos productos no serían posibles sin la recopilación y el análisis de información que se genera en los procesos, en los mercados y otras variables clave de las industrias. Estos grandes avances son posibles, entre otros factores, gracias a la enorme recopilación de datos y el análisis estadístico de la información existente, así como a su retroalimentación y evaluación continuas.

En nuestro país, contamos desde 1983 con un órgano autónomo, el Instituto Nacional de Estadística y Geografía (INEGI), que se encarga oficialmente de normar y coordinar el Sistema Nacional de Información, Estadística y de Geografía a diferentes niveles, así como de realizar cada diez años el censo nacional de los datos más relevantes para conocer la situación poblacional, económica y las necesidades sociales de la sociedad mexicana. Esto es de suma utilidad para la formulación de los programas de desarrollo económico y social en los tres niveles de gobierno del país, una base para la toma de decisiones de los inversionistas nacionales y extranjeros, y para la investigación científica que se lleva a cabo en universidades e institutos principalmente en nuestro territorio.

La asignatura de Probabilidad y Estadística se imparte formalmente en la Facultad de Ingeniería desde hace más de cincuenta años en todas las carreras de Ingeniería y sirve como antecedente de materias importantes de nivel profesional, dentro de los planes de estudio de las 15 carreras que se ofrecen actualmente en la Facultad de Ingeniería de la UNAM. Recientemente, los contenidos de esa asignatura se han agrupado en dos asignaturas, una denominada Probabilidad y la otra Estadística.

El documento "Estadística Descriptiva" es el resultado de varios años de trabajo e investigación disciplinar de dos profesoras y un profesor de la planta académica de la División de Ciencias Básicas de la Facultad de Ingeniería de la UNAM. Seguramente será de gran apoyo para los estudiantes de ingeniería que por primera vez llevan en su plan de estudios la asignatura de Estadística.

El libro ofrece una opción didáctica y conceptual para entender los principios teóricos y las metodologías de la denominada Estadística Descriptiva. Es un texto que acerca al estudiantado a conocer los sucesos de carácter aleatorio, la utilización de técnicas y recursos tecnológicos para la generación de bancos de información ordenados que le serán de utilidad en posteriores asignaturas del ámbito de la ingeniería, de sus respectivos planes de estudio profesional y de posgrado, como lo son la inferencia estadística, la estadística aplicada, la hidrología, el análisis de señales, la inteligencia artificial, la planeación, la teoría económica, entre otras.

M.I. Ángel Leonardo Bañuelos Saucedo
2024

Contenido

1 Conceptos básicos de la estadística

En los medios de comunicación radio, televisión, prensa y otros, leemos o escuchamos diariamente noticias como éstas:

- El candidato A aventaja al candidato B por diez puntos de diferencia.
- Los pumas llevan tres triunfos en el inicio de la temporada.
- El promedio de años para terminar un plan de estudios de una licenciatura en Ingeniería es de siete años.
- El número de contingencias ambientales fase 2 en la CDMX en los últimos 5 años.

Son expresiones de la investigación de hechos o fenómenos físicos o humanos que contienen información extraída de la realidad, que pueden pasar desapercibidas por algunos o causar preocupación para otros. Como sea, esa información es producto de un trabajo de índole estadístico.

1.1 ¿QUÉ ES LA ESTADÍSTICA?

> **Estadística.** Es la disciplina de las matemáticas que tiene por objeto recolectar, organizar, analizar e interpretar información obtenida de determinados fenómenos de interés para el investigador o una organización. Su finalidad es describir comportamientos de un conjunto de datos obtenido acerca de un objeto de estudio y hacer inferencias de un conjunto mayor.

La estadística tiene tres grandes *funciones*:

- Describir los resultados de una investigación científica.
- Apoyar la toma de decisiones basadas en los resultados de la investigación.
- Estimar cantidades desconocidas y plantear proyecciones de cierta información.

1.2 CONCEPTOS DE POBLACIÓN Y MUESTRA

La estadística se nutre de datos obtenidos de la observación o de pruebas aleatorias previamente determinadas.

> **Población.** Conjunto de todos los resultados posibles de un fenómeno aleatorio, observables por medio de la experimentación. Corresponde al objeto de estudio.

> **Muestra.** Es el conjunto de datos observados de la población.

Estos datos pueden obtenerse de toda la población o de una parte de ella, según se requieran. De aquí surgen los siguientes conceptos necesarios para comprender de manera ordenada los dimensionamientos grupales de los datos:

Figura 1. **Población y muestra**

- **Parámetro.** Es una cantidad numérica que caracteriza a la población.
- **Estadística.** Es una característica numérica de la muestra.

Ejemplo 1.1. Un ejemplo de un parámetro es el promedio del número de habitantes por vivienda en nuestro país, en el año 2010. El Anuario Estadístico de la República Mexicana elaborado por el INEGI indica que es de 3.9 personas/vivienda. Por su parte, una empresa de publicidad aplicó una encuesta a 5 viviendas y obtuvo un promedio de 4.2 personas/vivienda, esto constituye una estadística.

Ejemplo 1.2. La Ciudad de México (CDMX) es una de las urbes con mayor número de museos en el mundo, con gran diversidad de todo tipo (historia, artes, tecnología, economía, astronomía, deportes, etc.) a la altura de ciudades como Londres, París, Madrid. Según datos del Sistema de Información Cultural (SIC) de la Secretaría de Cultura del gobierno mexicano (2023) existen en total 185 museos, este número es un ejemplo de un parámetro. Se define como el objeto de estudio: todos los museos de la Ciudad de México. Ahora bien, dentro de la población de la CDMX, en algunas de sus 16 demarcaciones, como la Alcaldía Cuauhtémoc se han contabilizado 94 museos, en la Alcaldía Miguel Hidalgo se tienen 23 museos y en la Alcaldía Xochimilco solo están ubicados en su territo-

rio 3 museos; estos tres últimos números de las mencionadas alcaldías, por separado, se catalogan como estadísticas, en el contexto mencionado de la CDMX.

1.3 ESTADÍSTICA DESCRIPTIVA E INFERENCIAL

¿A qué edad se casan?

¿Cuál es su escolaridad?

¿En qué trabajan?

Figura 2. Estadística descriptiva e inferencial

Para cumplir sus objetivos, la estadística se divide en dos campos de estudio: la descriptiva y la inferencial.

> **Estadística descriptiva.** Conjunto de técnicas encaminadas a organizar y presentar de manera clara para su análisis, la información contenida en una muestra.

Los métodos de la estadística descriptiva nos ayudan a expresar como se presenta la realidad que nos rodea.

Las siguientes son situaciones que utilizan la estadística descriptiva:

- Un jugador de baloncesto quiere conocer su promedio de anotaciones en los últimos 10 juegos.
- Un político desea saber el porcentaje de votantes en los comicios para elegir gobernador de una región, en cada una de las últimas tres elecciones recientes.
- Un profesor de ingeniería expresa en una tabla las evaluaciones parciales, entrega de tareas y el resultado final de calificaciones de sus alumnos de la asignatura Probabilidad y Estadística, que comprende la totalidad del semestre. Con base en esta información, hace un estudio de su porcentaje de aprobación y promedio de calificaciones.
- El secretario de Agricultura requiere saber el crecimiento promedio de la producción de trigo en México, durante los últimos 5 años.

> **Estadística inferencial.** Área de la estadística que tiene por objeto la obtención de conclusiones acerca de las características de una población, a partir de la información contenida en una muestra, midiendo, desde el punto de vista probabilista, la validez de dichas conclusiones.

Las decisiones e inferencias se basan en información limitada o incompleta; los métodos de la estadística inferencial y el conocimiento obtenido nos permiten usar información disponible para entender y tratar con las incertidumbres de la realidad en un contexto aleatorio y que cambia con el tiempo.

Los siguientes ejemplos son una ampliación de ejemplos mencionados anteriormente y requieren de un tratamiento por los métodos que provee la estadística inferencial:

- El jugador de baloncesto quiere estimar la oportunidad que tiene de ganar el campeonato de anotaciones en el actual torneo, con base en su promedio actual y el promedio de sus futuros contrincantes.
- Con base en una encuesta de opinión, al político le gustaría estimar la oportunidad que tiene de ganar las próximas elecciones para gobernador de la región.
- En la facultad de Ingeniería, con base en el resultado de aprobación en la asignatura Álgebra, se deberá prever el número necesario de grupos para la asignatura consecuente Álgebra Lineal para el próximo semestre.
- El secretario de Agricultura quiere estimar lo más aproximado posible la producción de trigo en toneladas para los próximos diez años, con base en las tendencias que muestra la producción obtenida en los cinco últimos años.

No es posible realizar un análisis estadístico si no se cuenta con información que, dentro del proceso de análisis, se organizará de manera que permita presentarse de forma clara y facilite la generalización de lo observado a la población.

A la información utilizada en estadística se le denomina ***datos***. Para su adecuada utilización, los datos deben de organizarse y mostrarse apropiadamente. El tipo de datos indicará el método a seguir para el análisis correspondiente.

> **Dato** es una unidad de información obtenida del experimento bajo estudio. A pesar de que en ocasiones el término "dato" se usa como sinónimo de "muestra", consideraremos que una muestra es un conjunto de datos.

1.4 CLASIFICACIÓN DE DATOS

Una forma de clasificación de los datos es por el tipo de escala de medición.

> **Datos cuantitativos** son aquellas observaciones, necesariamente numéricas, que provienen de una escala de medición intervalar o de razón, es decir, representan valores que son comparables por su magnitud

Ejemplos de este tipo de datos pueden ser la edad en años, precio en pesos, estatura en centímetros, etc.

Ejemplo 1.3. La siguiente tabla indica información sobre el tiempo en minutos propuesto para diferentes tareas, se escribe el porcentaje que cubren determinados usuarios de un programa de cómputo al emplearlo para realizar completamente diversas tareas:

Tabla 1. **Tabla comparativa de tiempo de tareas**

Porcentaje de la tarea completada por cada usuario en diferentes tareas						
	TAREA 1 (1 min) 60 s	TAREA 2 (3 min) 180 s	TAREA 3 (1 min) 60 s	TAREA 4 (1:30 min) 90 s	TAREA 5 (1:30 min) 90 s	TOTAL
Usuario 1	100	25	100	100	0	325
Usuario 2	100	10	0	0	25	135
Usuario 3	100	50	0	0	30	180
Usuario 4	100	100	0	100	0	300
Usuario 5	100	100	0	0	15	215
Usuario 6	100	100	0	0	0	200
Usuario 7	0	0	100	0	100	200
TOTAL	600	385	200	200	170	1555

> **Datos cualitativos** representan atributos que pueden clasificarse bajo un criterio o cualidad. Son datos que provienen de escalas de medición nominal u ordinal.

Ejemplo 1.4. La siguiente tabla proporciona información cualitativa sobre caracteres utilizados para el análisis en el campo de la biología.

Tabla 2. Caracteres cualitativos usados en el análisis de similitud de las cuatro especies de Neobuxbaumia (cactus mexicano)

	Carácter	0	1	2
1	Hábito	columnar	ramificador	
2	Tonalidad del color verde del tallo	limón	oscuro	grisáceo
3	Sección transversal de costilla joven	triangular aguda	triangular redonda	
4	Depresión interareolar	ausente	presente	
5	Tricomas persistentes en la aréola	ausente	presente	
6	Abundancia de tricomas	escasa	regular	
7	Arreglo de las espinas radiales	5 - 7	3 - 9	< 9
8	Color de espinas radiales	blanco-crema	amarillo	obscuro
9	Forma de la espina central	acicular	tubulada	
10	Disposición de la espina central	recta	curva	
11	Consistencia de las espinas	flexible	rigida	
12	Ubicación de las flores en el tallo	longitudinal	apical	
13	Color de los tépalo	blanco-verdoso	verde rojizo	blanco
14	Nectarios extra florales	ausente	presente	
15	Escamas papiráceas de la flor	ausente	presente	
16	Color del fruto	verde claro	verde rojizo	obscuro

Ejemplo 1.5. El manejo de los trámites escolares de los estudiantes en la facultad de ingeniería contiene variables cualitativas en la información para cada alumno.

Tabla 3. Variables cualitativas para cada alumno

Características	Valoración cualitativa
Número de cuenta	0398654
Género	F
Escolaridad	LICENCIATURA
Carrera	Industrial
Trabaja	No

2 Organización de datos

2.1 DISTRIBUCIÓN DE FRECUENCIAS DE DATOS NO AGRUPADOS

La ***organización de la información*** consiste en acomodar un conjunto de datos, en forma conveniente y útil, para apreciar las características esenciales de los mismos.

Los datos que no están organizados **en clases o categorías** se les denomina **datos no agrupados**.

Cuando se manejan los datos no agrupados, se conserva el valor de cada uno de los datos. Según se necesite, se pueden ordenar de manera diferente:

- Del menor al mayor
- Del mayor al menor

> **Frecuencia.** Es el número de veces que aparece una medida o un valor dentro del conjunto de datos. Se utiliza el símbolo f_i para denotar la frecuencia del i-ésimo valor.

Ejemplo 2.1. El número de tiros a gol fallados por los jugadores de los Pumas en los últimos 8 partidos fueron 8, 10, 12, 8, 9, 9, 7, 9. Una manera de ordenar los datos se muestra en la tabla.

X: Número de tiros a gol fallados

Tabla 4. Número de tiros a gol fallados por los Pumas

X	Frecuencia f_i	Lectura:
7	1	Frecuencia de 7 es 1
8	2	Frecuencia de 8 es 2
9	3	Frecuencia de 9 es 3
10	1	Frecuencia de 10 es 1
12	1	Frecuencia de 12 es 1
Suma	**8**	

Distribución de frecuencias para datos no agrupados

Nótese que la suma de frecuencias es igual al total de juegos en los que han participado los Pumas.

2.2 DISTRIBUCIÓN DE FRECUENCIAS DE DATOS AGRUPADOS

2.2.1 CONCEPTOS BÁSICOS PARA DATOS AGRUPADOS

Nos referimos a datos agrupados cuando la información no se presenta dato a dato, sino organizada en clases o categorías. Por ejemplo, la información solicitada de los ingresos de las familias puede agruparse en rangos como menos de un salario mínimo, entre uno y tres salarios mínimos o más de tres salarios mínimos.

Las **tablas de distribución de frecuencias** empíricas se usan para resumir grandes cantidades de datos que contienen relativamente pocas repeticiones. Probablemente una de las razones para hacer este tipo de resúmenes es que la experiencia ha mostrado que para identificar los patrones o características de un conjunto grande de datos es **conveniente** agrupar las observaciones en **intervalos de clase**.

> **Tabla de distribución de frecuencias.** Agrupación de los datos en clases o categorías, dependiendo de su valor.

Existen diversos métodos para construir tablas de distribución de frecuencias, que difieren entre sí solamente en pequeños detalles. Aunque algunos toman más consideraciones que otros, los resultados que producen son semejantes y la interpretación de la información también lo es.

Estamos seguros de que, quien es capaz de comprender la información contenida en una tabla de distribución de frecuencias, independientemente de la metodología utilizada para construirla, es capaz de comprender otra tabla construida con alguna metodología distinta. En este trabajo se muestran dos métodos, seleccionados por su amplia aceptación entre diversos autores.

El primer método es una tabla clásica de distribución de frecuencias como la que se muestra en la tabla 5, que consta de seis columnas (límites de clase, marcas de clase, frecuencia, frecuencia relativa, frecuencia acumulada y frecuencia acumulada relativa).

Tabla 5. Tabla clásica de distribución de frecuencias. La clasificación se hace solamente usando límites de clase

Límites de clase		Marca de clase	Frecuencia	Frecuencia relativa	Frecuencia acumulada	Frecuencia acumulada relativa
Inferior	Superior					
13,050	84,960	49,005	89	0.4684	89	0.4684
84,960	156,870	120,915	38	0.2000	127	0.6684
156,870	228,780	192,825	14	0.0737	141	0.7421
228,780	300,690	264,735	11	0.0579	152	0.8000
300,690	372,600	336,645	1	0.0053	153	0.8053
372,600	444,510	408,555	33	0.1737	186	0.9789
444,510	516,420	480,465	3	0.0158	189	0.9947
516,420	588,330	552,375	1	0.0053	190	1.0000
			190			

La primera columna de dicha tabla corresponde a los ***límites de clase***, los cuales establecen el criterio que se utilizará para clasificar a los datos. Los intervalos de clase, en esta construcción particular, son intervalos cerrados por la izquierda y abiertos por la derecha. Esto significa que un dato en la muestra que sea mayor o igual al límite inferior de la clase y menor que el límite superior, pertenecerá a la clase en cuestión. Es importante observar que, de acuerdo con este método, los intervalos de clase definidos por los límites de clase son intervalos unidos, pero no traslapados, esto es, el límite superior de una clase es al mismo tiempo, el límite inferior de la clase siguiente.

> **Límite de clase.** Son intervalos cerrados por la izquierda y abiertos por la derecha que contienen los valores menor y mayor que, de encontrarse como datos en la muestra pertenecen a la clase en cuestión.

> **Fronteras de clase.** Son valores teóricos que permiten eliminar los espacios que hay entre los límites. Se calculan promediando el valor del límite inferior de la clase y el valor del límite superior de la clase siguiente.

De acuerdo con el segundo método, una tabla de distribución de frecuencias consta de siete columnas, como se muestra en la tabla 6.

De acuerdo con el segundo método, en dicha tabla aparecen dos criterios de clasificación, los límites de clase y las fronteras de clase, sin embargo, ambos criterios producen una clasificación idéntica. En esta construcción, los intervalos definidos por los límites de clase son cerrados, es decir, cualquier dato mayor o igual al límite inferior de la clase y menor o igual al límite superior de la misma, se contabilizará en la clase en cuestión.

Como se puede ver en esta construcción, las clases definidas por los límites de clase están separadas y la distancia entre ellas es una unidad de precisión de la medida en los datos; por su parte, las fronteras o límites reales de clase, que se muestran en la segunda columna, tienen la función de unir los intervalos eliminando así los huecos existentes entre categorías consecutivas.

Tabla 6. Tabla clásica de distribución de frecuencias. Considera límites y fronteras de clase

Límites de clase		Fronteras de clase		Marca de clase	Frecuencia	Frecuencia relativa	Frecuencia acumulada	Frecuencia acumulada relativa
Inferior	Superior	Inferior	Superior					
13,050.5	84,959.4	13,050.45	84,959.45	49,004.95	89	0.4684	89	0.4684
84,960.5	156,869.4	84,959.45	156,868.45	120,913.95	38	0.2000	127	0.6684
156,870.5	228,779.4	156,868.45	228,777.45	192,822.95	14	0.0737	141	0.7421
228,780.5	300,689.4	228,777.45	300,686.45	264,731.95	11	0.0579	152	0.8000
300,690.5	372,599.4	300,686.45	372,595.45	336,640.95	1	0.0053	153	0.8053
372,600.5	444,509.4	372,595.45	444,504.45	408,549.95	33	0.1737	186	0.9789
444,510.5	516,419.4	444,504.45	516,413.45	480,458.95	3	0.0158	189	0.9947
516,420.5	588,329.4	516,413.45	588,322.45	552,367.95	1	0.0053	190	1.0000
					190			

En ambos métodos de construcción, la columna que sigue a aquella(s) de la clasificación corresponde a las ***marcas de clase***, que es un valor representativo de los datos contenidos en ella, de tal manera que se utilizará como dato observado, ya que una vez que se tienen los datos agrupados en una tabla de distribución de frecuencias, se pierden los datos originales.

> **Marca de clase.** La marca de la clase se calcula como el punto medio de la clase en cuestión, es decir, la suma de los límites de la clase, dividida por dos. Se acostumbra a denotar a la marca de la clase i como X_i.

A continuación, en la misma tabla, se presenta la columna de "frecuencias" o "frecuencias absolutas". La información contenida en esta columna corresponde al número de datos en la muestra que pertenecen a la clase i. Los valores asentados en esta columna se obtienen mediante un procedimiento de conteo.

> **Frecuencia de clase.** La frecuencia de la clase i, o frecuencia absoluta de la clase i, es el número de datos en la muestra que pertenece a la clase i, es decir, cuyo valor es mayor o igual al límite inferior y menor que el superior de la clase si la construcción incluye solamente límites de clase.
>
> Si la construcción considera fronteras de clase, la frecuencia de la clase i es el número de datos en la muestra que pertenecen a la clase i, es decir, cuyo valor es mayor o igual al límite inferior y menor o igual que el superior de la misma clase.
>
> Se denota a la frecuencia de la clase i como f_i.

Las siguientes tres columnas en la tabla 5, solamente muestran la información que ya se ha plasmado en ella, pero en términos distintos. Estas tres columnas corresponden a la frecuencia relativa, la frecuencia acumulada y la frecuencia acumulada relativa.

> **Frecuencia relativa.** La frecuencia de la clase es la proporción de los datos en la muestra que se encuentran en el intervalo de clase i. La forma de denotar a la frecuencia relativa es f_i^*.

Esta frecuencia relativa se calcula como el cociente entre la frecuencia de la clase, entre el número de datos en la muestra.

$$f_i^* = \frac{f_i}{\sum_{i=1}^{c} fi} = \frac{f_i}{n}$$

En donde f_i es la frecuencia de la clase i, n es el número de datos en la muestra y c es el número de intervalos de clase que constan en la tabla de distribución de frecuencias.

Frecuencia acumulada. La frecuencia acumulada de la clase k es el número de datos en la muestra que son menores o iguales a la frontera superior de la misma clase, o menores que el límite superior de la clase k, en los casos en los que la construcción no considera fronteras.

Esto es:

$$F_k = f_1 + f_2 + f_3 + \ldots + f_k = \sum_{i=1}^{k} f_i$$

Frecuencia acumulada relativa. Indica la proporción de los datos en la muestra, que son menores que el límite superior de la clase . Esta frecuencia acumulada relativa se calcula como el cociente entre la frecuencia acumulada de la clase, entre el número de datos en la muestra.

$$F_k^* = \frac{F_k}{n}$$

O también

$$F_k^* = f_1^* + f_2^* + f_3^* + \ldots + f_k^* = \sum_{i=1}^{k} f_i^*$$

Número de intervalos. Se denota por c al número de intervalos en la tabla de distribución de frecuencias. Una forma empírica de determinar un valor conveniente de c se calcula como la raíz cuadrada del número de datos en la muestra, redondeada al entero más cercano, si la muestra es menor a 400 elementos. Sin embargo, si el número de

datos es muy grande, este criterio puede proponer la construcción de un número también grande de intervalos, perdiéndose así el objetivo de la agrupación En estos casos, se puede utilizar la Regla de Sturges.

Sturges (1926) establece como un número adecuado de clases el que resulta de la operación:

$$c = 1 + 3.3 \log_{10} (n)$$

redondeada al entero más cercano, en donde n es el número de datos en la muestra.

Dado que el objetivo de la agrupación es hacer un resumen de la información, muchos autores consideran conveniente la utilización de un número de intervalos no menor a 5 para que no se pierdan las características esenciales de la distribución y no mayor a 20 para no perder el sentido práctico de la agrupación (Spiegel, 1974).

Tabla 7. Sturges para determinar el número de intervalos en la tabla de frecuencias

Número de datos (n)	Número recomendado de intervalos de clase (c)		Número de datos (n)	Número recomendado de intervalos de clase (c)		Número de datos (n)	Número recomendado de intervalos de clase (c)	
	$\sqrt{n}$	$1+3.3 \log_{10} n$		$\sqrt{n}$	$1+3.3 \log_{10} n$		$\sqrt{n}$	$1+3.3 \log_{10} n$
15	4	5	250	16	9	40,000	200	16
20	4	5	300	17	9	50,000	224	17
25	5	6	400	20	10	60,000	245	17
30	5	6	500	22	10	70,000	265	17
35	6	6	600	24	10	80,000	283	17
40	6	6	700	26	10	90,000	300	17
45	7	6	800	28	11	100,000	316	18
50	7	7	900	30	11	200,000	447	19
55	7	7	1,000	32	11	6,000,000	2,449	24
56	7	7	1,100	33	11	7,000,000	2,646	24
57	8	7	1,200	35	11	8,000,000	2,828	24
60	8	7	1,500	39	12	9,000,000	3,000	24
65	8	7	2,000	45	12	300,000	548	19
70	8	7	2,500	50	12	400,000	632	20
75	9	7	3,000	55	13	500,000	707	20
80	9	7	4,000	63	13	1,000,000	1,000	21
85	9	7	5,000	71	13	2,000,000	1,414	22
90	9	7	6,000	77	14	3,000,000	1,732	23
95	10	8	7,000	84	14	4,000,000	2,000	23
100	10	8	8,000	89	14	5,000,000	2,236	23
125	11	8	9,000	95	14	10,000,000	3,162	24
150	12	8	10,000	100	14	50,000,000	7,071	27
175	13	8	20,000	141	15	100,000,000	10,000	28
200	14	9	30,000	173	16	200,000,000	14,142	29

Rango. Diferencia entre el dato mayor y el dato menor en la muestra.

Tamaño del intervalo o amplitud del intervalo. Diferencia entre límites inferiores consecutivos o límites superiores consecutivos. Este tamaño se calcula mediante la expresión:

$$w = \frac{rango}{número\ de\ intervalos}$$

Unidad de precisión de la medida de clase UPMC. Es la unidad (metros, kilómetros, pesos, barriles de petróleo, etc.) en la que están expresados los datos, es un ajuste en el valor del extremo superior del intervalo, equivale a agregar decimales a ese extremo.

$$\text{lím}_{\text{inf}} = \text{lím}_{\text{sup}} + UPMC$$

Fronteras o límites reales es un ajuste en el valor en los extremos del intervalo, se calcula mediante

$$Front_{\text{inf}(i)} = \text{lím}_{\text{inf}} - \frac{1}{2}(UPMC) \quad \text{y} \quad Front_{\text{sup}(i)} = \text{lím}_{\text{sup}} + \frac{1}{2}(UPMC)$$

que son valores teóricos muy útiles para los fines de análisis de la continuidad de los datos y su representación gráfica de las distribuciones de frecuencias.

Ejemplo 2.2. En la tabla 8 se muestran 48 calificaciones de estudiantes de la asignatura Probabilidad, se desea construir una tabla de distribución de frecuencias que represente la información de las calificaciones:

Tabla 8. Promedios de calificaciones de 48 alumnos

Calificaciones					
8.3	9.0	8.2	8.4	8.8	7.8
7.9	8.0	8.3	8.2	9.0	8.1
8.6	8.2	8.6	9.5	9.5	7.7
9.1	7.7	8.3	8.9	8.5	8.3
9.1	7.6	8.3	9.9	8.2	8.5
9.1	7.6	8.3	8.2	8.8	8.5
7.4	8.1	9.2	9.3	8.3	8.3
8.4	7.4	8.5	7.7	8.9	9.6

Solución:

a) Se obtienen los valores mínimo = 7.4 y máximo = 9.9
b) Para construir la tabla de distribución de frecuencias, inicialmente se establece el **número de intervalos C.**
c) El número de intervalos recomendado, con el criterio de **Sturges** es

$$c = 1 + 3.3 \log_{10}(48) = 6.585 \cong 7$$

Si se utiliza el criterio de la raíz cuadrada para comparar con el criterio de Sturges, en el ejemplo con 48 elementos, se tiene: $C = \sqrt{48} = 6.9 \cong 7$**. En la tabla se indicarán 7 intervalos.**

Una vez que se tiene el número de intervalos en los que se repartirán los datos, se debe calcular el **tamaño del intervalo** o **amplitud del intervalo** w. Este tamaño se calcula mediante:

$$w = \frac{rango}{número\ de\ intervalos}$$

$$rango = 9.9 - 7.4 = 2.5$$

Tomando el dato de la tabla de Sturges para el número de intervalos, la amplitud de clase es:

$$w = \frac{2.5}{7} = 0.357 \cong \mathbf{0.4}$$

Se redondea a la UPMC mayor más próxima.

Con esta información se procede a construir la tabla de frecuencias. Se inicia con el valor del límite inferior de la primera clase que puede coincidir con el valor más pequeño de los datos. Sin embargo, para evitar inducir un sesgo en el análisis, se recomienda recorrer ese valor a uno un poco más pequeño que facilite los cálculos y garantice una frecuencia diferente de cero para esa clase. El valor del límite inferior de la siguiente clase que corresponderá es 7.7, ya que se suman 0.4 unidades al primer límite inferior de la clase anterior, que en este caso consideraremos 7.3.

Tabla 9. Frecuencias con límites

	Límites	
Intervalo	Inferior	Superior
1	7.3	7.6
2	7.7	8.0
3	8.1	8.4
4	8.5	8.8
5	8.9	9.2
6	9.3	9.6
7	9.7	10.0

En la tabla 9 puede observarse que la diferencia entre dos límites inferiores contiguos es la amplitud de clase ($w = 0.4$), y lo mismo sucede en los límites superiores contiguos, el intervalo superior de la clase 7, o de la última clase, deberá incluir el valor máximo.

En el caso del ejemplo 2.1, es una décima de punto de calificación (en una escala del 0 a 10), esto es UPMC=0.10. Lo anterior explica que, una vez definido el primer intervalo de clase ($i = 1$), el límite inferior del siguiente intervalo de clase ($i = 2$) se puede definir aumentando una UPMC al límite superior clase ($i = 1$):

$$\text{lím}_{\text{inf}(2)} = \text{lím}_{\text{sup}\,(1)} + UPMC = 7.6 + 0.10 = 7.7$$

Donde

$\text{lím}_{\text{inf}(i)}$: límite inferior de la clase i
$\text{lím}_{\text{sup}(i)}$: límite superior de la clase i

De manera general, también puede obtenerse la $UPMC = \text{lím}_{\text{inf}(i)} - \text{lím}_{\text{sup}(i-1)}$. Solo para ilustrar lo mencionado considérese el intervalo de clase $i = 4$, entonces:

$$UPMC = \text{lím}_{\text{inf}(4)} - \text{lím}_{\text{sup}\,(3)} = 8.5 - 8.4 = 0.10$$

lo cual es correcto.

Se procede entonces a calcular las **marcas de clase y las fronteras**, para el primer intervalo.

Las fronteras para el intervalo de clase ($i = 1$) se obtienen mediante:

$$front_{\text{inf}} = 7.3 - \frac{1}{2}(0.10) = 7.25$$

$$front_{\text{sup}} = 7.6 + \frac{1}{2}(0.10) = 7.65$$

La marca de clase del primer intervalo será entonces $x_i = \frac{7.3+7.6}{2} = 7.45$; y así sucesivamente se procederá para todas las fronteras y marcas de clase del resto de los intervalos. La tabla quedará:

Tabla 10. Frecuencias con límites, fronteras y marcas de clase

	Límite		Fronteras		Marca de clase
intervalo	inferior	superior	inferior	superior	X_i
1	7.3	7.6	7.25	7.65	7.45
2	7.7	8.0	7.65	8.05	7.85
3	8.1	8.4	8.05	8.45	8.25
4	8.5	8.8	8.45	8.85	8.65
5	8.9	9.2	8.85	9.25	9.05
6	9.3	9.6	9.25	9.65	9.45
7	9.7	10.0	9.65	10.05	9.85

El siguiente punto en la tabla debe ser la frecuencia de datos encontrados en cada intervalo, al contar el número de datos que se encuentra dentro del primer intervalo de 7.3 a 7.6, se observa que son 4, en el siguiente intervalo de 7.8 a 8.1 se tienen 5. En la tabla 10 se muestran las marcas correspondientes a todas las clases.

Tabla 11. **Frecuencias con límites, fronteras, marcas de clase y frecuencias absolutas**

	Límite		Fronteras		Marca de clase	Frecuencia
intervalo	inferior	superior	inferior	superior	x_i	f_i
1	7.3	7.6	7.25	7.65	7.45	4
2	7.7	8.0	7.65	8.05	7.85	6
3	8.1	8.4	8.05	8.45	8.25	17
4	8.5	8.8	8.45	8.85	8.65	8
5	8.9	9.2	8.85	9.25	9.05	8
6	9.3	9.6	9.25	9.65	9.45	4
7	9.7	10.0	9.65	10.05	9.85	1
Suma						**48**

La suma de todas las frecuencias es igual al número de datos proporcionados en la muestra n, en este caso deberá ser 48.

Una vez que se tiene la tabla de frecuencias es posible hacer la construcción de distintos gráficos para representar la información. Se necesitan solamente las marcas de clases o las fronteras de clase en el eje de las abscisas, y las frecuencias para el eje de las ordenadas.

Para ilustrar la parte operativa de estos conceptos fundamentales en la estadística, se retoma el ejemplo 2.2.

Tabla 12. Frecuencias con límites, fronteras, marcas de clase, frecuencias absolutas, frecuencias relativas, frecuencias acumuladas y frecuencias relativas acumuladas

	Límites		Fronteras		Marca de clase	Frecuencia	Frec. Rel.	Frec. acumulada	Frec. Acum. Rel.
intervalo	inferior	superior	inferior	superior	x_i	f_i	f_i^*	F_i	F_i^*
1	7.3	7.6	7.25	7.65	7.45	4	0.0388	4	0.0833
2	7.7	8.0	7.65	8.05	7.85	6	0.1250	10	0.2083
3	8.1	8.4	8.05	8.45	8.25	17	0.3542	27	0.5625
4	8.5	8.8	8.45	8.85	8.65	8	0.1667	35	0.7292
5	8.9	9.2	8.85	9.25	9.05	8	0.1667	43	0.8958
6	9.3	9.6	9.25	9.65	9.45	4	0.0833	47	0.9792
7	9.7	10.0	9.65	10.05	9.85	1	0.0208	48	1.000
Suma						48	1		

Donde c es el número de intervalos de clase en la distribución de frecuencias. Así, si se requiere conocer el número de datos acumulados hasta el término del tercer intervalo de clase, esto es $k = 3$

$$F_3 = f_1 + f_2 + f_3 = 4 + 6 + 17 = 27$$

Lo anterior indica que en la distribución se tienen 27 datos iguales o menores a 8.45.

También en la columna de F_i se aprecia que la frecuencia acumulada en los seis intervalos de clase es de 48, que es el número total de datos n.

La suma de todas las frecuencias relativas siempre debe ser uno. Se denota por f_i^* y se determina de la siguiente manera:

$$f_i^* = \frac{f_i}{\sum_{i=1}^{c} f_i} = \frac{f_i}{n}$$

En el caso del ejemplo, se observa la frecuencia relativa en la sexta columna de la tabla 12. Asimismo, para calcular el porcentaje de datos que hay en cada intervalo de clase, se multiplica a la frecuencia relativa correspondiente por 100:

$$\text{Porcentaje} = (f_i^*)\,(100)\%$$

Si se observa la tabla de frecuencias del ejemplo, en la clase $i = 2$ se ha acumulado el 20.83 % de los datos y, en la última clase $i = 6$, se ha acumulado el 100 % de los datos. Generalizando, el cálculo del acumulado de datos de manera porcentual hasta una clase k predeterminada, se tiene:

$$F_k^* = f_1^* + f_2^* + f_3^* + \ldots + f_k^* = \sum_{i=1}^{k} f_i^*$$

A manera de resumen, para construir la tabla de frecuencias se deben tomar en cuenta los siguientes pasos:

1. Localizar el valor mínimo y el valor máximo.
2. Calcular el rango.
3. Determinar el número de intervalos.
4. Determinar la amplitud del intervalo.
5. Identificar y contabilizar el número de datos que corresponden a cada intervalo de clase.
6. La tabla de frecuencias debe iniciar con el valor mínimo de los datos o de acuerdo con las fronteras o límites reales.
7. En caso de considerar la construcción con fronteras, agregar en la tabla una columna que las muestre.

2.2.2 REPRESENTACIÓN GRÁFICA DE DATOS

Una forma de representar la información de manera que se pueda observar su comportamiento es a través de gráficas.

Los tipos de gráficos más comunes son:

- Histograma
- Polígono de frecuencias
- Ojiva, ojiva porcentual
- Grafica de sectores o sectores circulares
- Diagrama de tallo y hojas

El histograma es una gráfica de barras verticales en la que la base de cada barra está centrada en la marca de clase correspondiente, dicha base tiene como ancho la longitud de clase. La altura de la barra es la frecuencia de la misma clase.

Para ilustrar lo anterior, se muestra el histograma de frecuencia relativa del ejemplo 2.2 (figura 3):

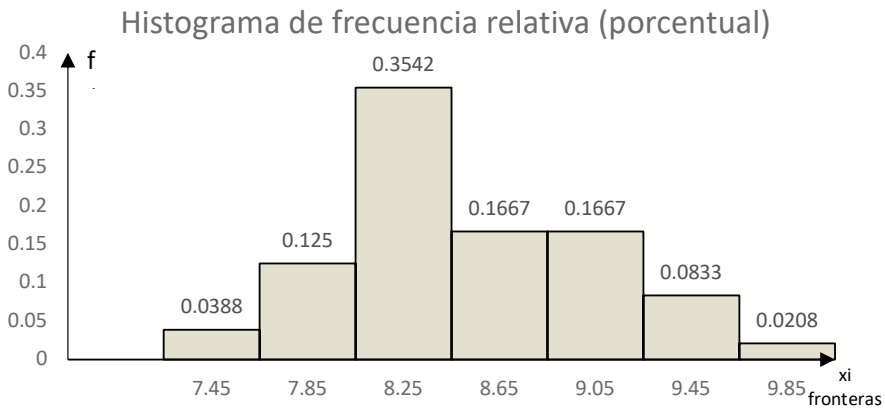


Figura 3. Histograma de fronteras y frecuencias relativas

El **polígono de frecuencias relativas** muestra la forma en que el porcentaje de los datos se va distribuyendo en cada uno de los intervalos de clase.

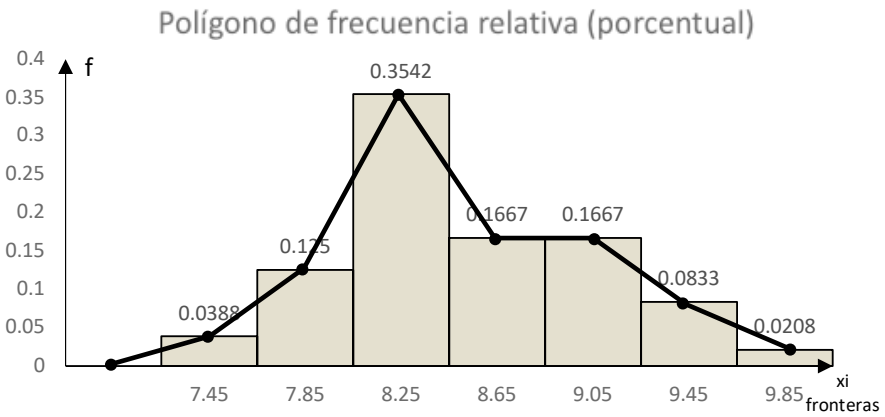


Figura 4. Polígono de fronteras y frecuencias relativas

La **ojiva de frecuencia relativa acumulada** muestra la forma en que el porcentaje de los datos, en cada uno de los intervalos de clase, se va acumulando de manera ascendente.

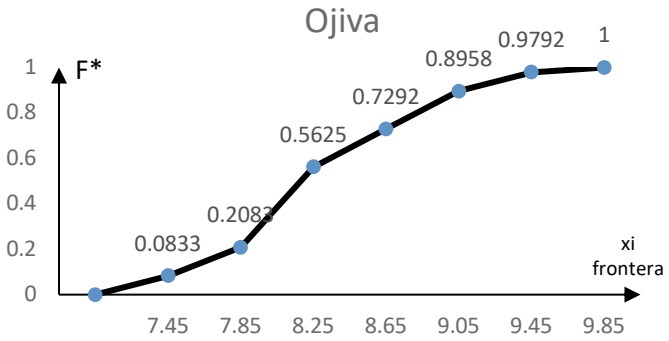


Figura 5. Ojiva de fronteras y frecuencia relativa acumulada

La **gráfica de sectores circulares o gráfica de pastel** presenta de otra forma la frecuencia relativa en porciones equivalentes a dicha frecuencia.

Las gráficas circulares o de pastel son muy utilizadas en el campo de la estadística aplicada, con el objeto de expresar de manera ilustrativa, para lectores o analistas, de diferentes niveles de estudio, la forma en que se distribuye porcentualmente la información del asunto de interés; como puede ser en el campo científico básico, en las ciencias sociales, en el ámbito deportivo o en áreas donde el análisis de la información existente es de importancia.

Tabla 13. Partición y valores para el sector circular

Sector i	Frecuencia relativa f_i^*	Ángulos (grados) (f^*) (360)
1	0.1333	97.988
2	0.2667	96.012
3	0.4000	144.000
4	0.1333	47.988
5	0.0667	24.012
Suma		360

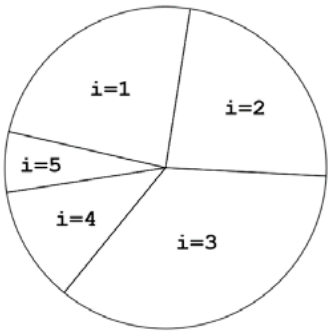


Figura 6. Gráfica de sector circular

> **Diagrama de tallo y hojas.** En términos generales, el diagrama de tallo y hojas se forma de una serie de renglones (filas horizontales) constituidas por números. Al primer número de cada renglón se le conoce como tallo, y al resto de los números de éste, se les llama hojas del diagrama.

En la observación de la naturaleza, se puede constatar que la mayoría de los árboles tienen un tronco, un número mayor de ramas, y aún más grande el número de hojas; posiblemente en esta apreciación se inspiró John Tubey (1977) para ofrecer a la Ciencia Estadística una de las más originales representaciones gráficas de datos que la denominó diagrama de tallo y hojas.

Para construir un diagrama de tallo y hojas no se determinan reglas rígidas; sino más bien, un conjunto de criterios que hay que tomar en cuenta para que el diagrama final permita apreciar mejor la esencia y características de la información en estudio. Así, algunos autores consideran los siguientes:

- El primer dígito o los dos primeros dígitos de los datos considerados, se selecciona o seleccionan siempre como tallos, estos desde luego se deben de alinear en una primera columna que se conoce como la columna de los tallos; seguida de una línea vertical para separarla del resto de la información: hojas.
- Para ilustrar los pasos que permiten elaborar un diagrama de este tipo, se explicará cada uno de ellos con el siguiente ejemplo:

Ejemplo 2.3. La tabla muestra la información obtenida por la Secretaría de Salud en 20 hospitales de la zona sur de una gran capital de un cierto país:

Tabla 14. Información de hospitales de la Secretaría de Salud

Hospital	Número de camas disponibles	Número de pacientes internados	Cajones de estacionamiento	Número de médicos
1	52	48	45	4
2	86	68	55	4
3	77	69	60	5
4	66	45	35	4
5	117	92	60	8
6	91	76	56	7
7	99	91	40	9
8	108	102	30	9
9	119	83	68	8
10	83	83	45	7
11	65	61	65	6
12	94	72	35	7
13	98	54	24	8
14	104	93	54	9
15	98	78	63	6
16	82	78	36	7
17	80	79	75	4
18	107	86	56	9
19	92	92	68	7
20	74	71	50	6

De la anterior información, la Secretaría decide analizar el número de lugares disponibles en los hospitales de esta zona para ver la cobertura que pueden absorber de pacientes que necesitan ser internados y para planear, en caso de que la demanda sea mayor, nuevos hospitales para este tipo de requerimientos. Para lograr una mejor apreciación de la información existente, los analistas deciden presentarla por medio de un diagrama de tallo y hojas.

Paso uno: se determinan los valores de los tallos de los datos de la segunda columna, los cuales están formados por dos y tres cifras, lo que indica que se tienen decenas y centenas y eso dificulta seleccionar los tallos. Lo que procede en estos casos es identificar el menor y el mayor valor de ellos; así, en este caso, el menor es 52 y el mayor es 119 de camas disponibles. Entonces, para el fin de homogeneizar las unidades de los tallos, en este caso, conviene formarlos con las cifras de las decenas y centenas; esto es, el valor del menor se puede expresar como 052 y la columna de tallos queda:

Tabla 15. Diagrama de tallo para los dígitos más significativos para las camas disponibles

Tallo
05
06
07
08
09
10
11

Paso dos: para cada dato se separan tallo y hojas, y se colocan las hojas a la derecha de una línea vertical separadora; de esta manera, para el hospital número 2, le corresponderá el tallo 08 y la hoja 6.

Tabla 16. Diagrama de tallo y hojas para los dígitos menos significativos en camas disponibles

Tallo	Hojas
05	
06	
07	
08	6
09	
10	
11	

De manera similar para el mismo tallo 08, la siguiente hoja es la del hospital número 10, que es el dígito 3. Con ello el diagrama queda:

Tabla 17. Diagrama de tallo y hojas para los dígitos menos significativos en el número de camas disponibles

Tallo	Hojas
05	
06	
07	
08	6 3
09	
10	
11	

Paso tres: se reproduce esta acción de manera ordenada para todos y cada uno de los datos de la columna de la tabla; se registran cada uno de los números hojas (unidades) que le siguen a cada uno de los tallos (decenas y centenas):

Tabla 18. Diagrama de tallo y hojas para el registro de valores menos significativos en las camas disponibles

Tallo	Hojas
05	2
06	6 5
07	7 4
08	6 3 2 0
09	1 9 4 8 8 2
10	8 4 7
11	7 9

Paso cuatro: ahora deben en forma ascendente, ordenarse los valores de las hojas:

Tabla 19. Diagrama de tallo y hojas con los valores ordenados, vista horizontal

Tallo	Hojas
05	2
06	5 6
07	4 7
08	0 2 3 6
09	1 2 4 8 8 9
10	4 7 8
11	7 9

Este arreglo diagramático, ya nos muestra, que en la zona sur de esta gran capital la disponibilidad de lugares para internar pacientes está mayormente entre 80 y 100. Ahora, si se gira el diagrama de tallos y hojas conservando los tallos (en posición horizontal) y las hojas (en posición vertical), con los valores de ambos de manera ascendente en su nueva posición se tendrá:

Tabla 20. Diagrama de tallo y hojas con los valores ordenados, vista vertical

				9		
				8		
			6	8		
			3	4	8	
	6	7	2	2	7	9
2	5	4	0	1	4	7
05	06	07	08	09	10	11

Como puede apreciarse, esta disposición de datos es prácticamente la misma que tiene el histograma de la distribución de frecuencias correspondiente a estos datos (se muestra abajo); con la ventaja del diagrama de árbol que, en cada caso, muestra el valor original de los datos.

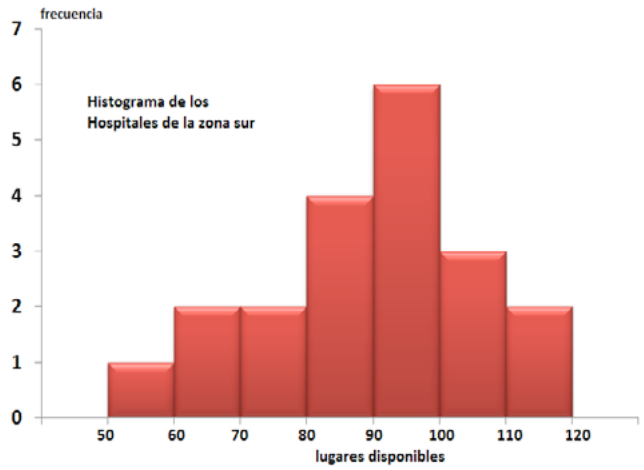


Figura 7. Histograma para el número de camas disponibles

2.3 EJERCICIOS RESUELTOS No. 1

1. Considérese la siguiente distribución de frecuencias:

Tabla 21. Distribución de frecuencias

Clase i	Intervalo de clase		Frecuencia
	inferior	superior	f_i
1	1	5	1
2	6	10	3
3	11	15	4
4	16	20	5
5	21	25	2
	Suma		15

a) Obtener fronteras de clase y amplitud
b) Marcas de clase
c) Histograma y polígono de frecuencia
d) Histograma de frecuencias relativas
e) Ojiva de frecuencia acumulada
f) Ojiva de frecuencia relativa acumulada

Solución:

a) $UPMC = \text{lím } inf_{(i)} - \text{lím } sup_{(i-1)}$

si $i = 3$

$UPMC = \text{lím } inf_3 - \text{lím } sup_2 = 11-10 = 1.0$

$w = \text{lím } inf_3 - \text{lím } inf_2 = 11-6 = 5$

Tabla 22. Distribución de frecuencia relativa

Clase i	Intervalo de clase		Fronteras		Frecuencia
	inferior	superior	inferior	superior	f_i
1	1	5	0.5	5.5	1
2	6	10	5.5	10.5	3
3	11	15	10.5	15.5	4
4	16	20	15.5	20.5	5
5	21	25	20.5	25.5	2
	Suma				15

$$Front\,inf_1 = \text{lím } inf_1 - \frac{1}{2}(UPMC) = 1 - \frac{1}{2}(1) = 1 - 0.5 = 0.5$$

$$Front\,sup_1 = \text{lím } sup_1 + \frac{1}{2}(UPMC) = 5 + \frac{1}{2}(1) = 5 + 0.5 = 5.5$$

Estos dos últimos valores corresponden a la frontera inferior y frontera superior clase 1, en la tabla 22 se indican los valores de las fronteras de las otras 5 clases restantes.

b) Marcas de clase

$$x_i = \frac{\text{lím } inf_i + \text{lím } sup_i}{2} \qquad x_i = \frac{front\ inf_i + front\ sup_i}{2}$$

Para la clase 1:

$$x_1 = \frac{1+5}{2} = 3$$

Para la clase 5:

$$x_5 = \frac{20.5 + 25.5}{2} = 23$$

c) Histograma y polígonos de frecuencias:

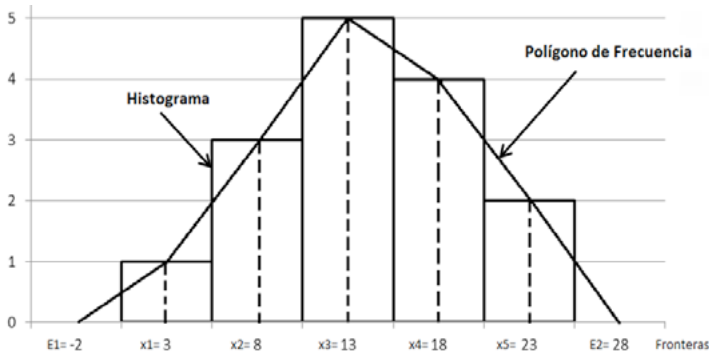


Figura 8. **Gráfica de histograma con polígono de frecuencias absolutas**

Valores extremos E_1 y E_2 en el polígono de frecuencias absolutas (figura 8)

Con las marcas de clase:

$$E_1 = x_1 - w = 3 - 5 = -2$$

$$E_2 = x_5 + w = 23 + 5 = 28$$

Considerando las fronteras:

$$E_1 = Front\ inf_1 - \frac{1}{2}w = 0.5 + 2.5 = -2$$

$$E_2 = Front\ sup_5 + \frac{1}{2} = 25.5 + 2.5 = 28$$

d) Histograma de frecuencias relativas:

En la tabla 23, se calcula la frecuencia relativa para cada clase:

$$f^* = \frac{f_i}{\sum_{i=1}^{c} f_i} = \frac{f_i}{n}$$

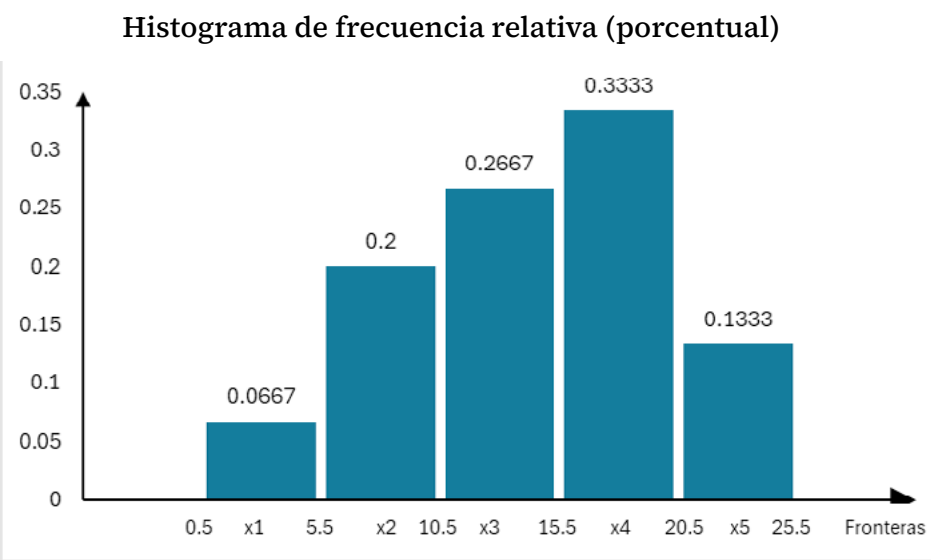


Figura 9. Histograma de fronteras y frecuencias relativas

Tabla 23. Distribución de frecuencias con intervalos de clase, fronteras, marcas de clase y frecuencias absolutas

Clase i	Límites		Fronteras		Marca de clase	Frecuencia	Frecuencia relativa	Frecuencia relativa acumulada
	inferior	superior	inferior	superior	x_i	f_i	f_i^*	Fi
1	1	5	0.5	5.5	3	1	0.0667	0.0667
2	6	10	5.5	10.5	8	3	0.2000	0.2667
3	11	15	10.5	15.5	13	4	0.2667	0.5334
4	16	20	15.5	20.5	18	5	0.3333	0.8667
5	21	25	20.5	25.5	23	2	0.1333	1.0000
	Suma					15	1.0	

Ojivas:

e) Ojiva de frecuencias acumuladas

En la frontera inferior de la clase uno no hay acumulación y en la frontera superior clase cinco se han acumulado todos los datos (15).

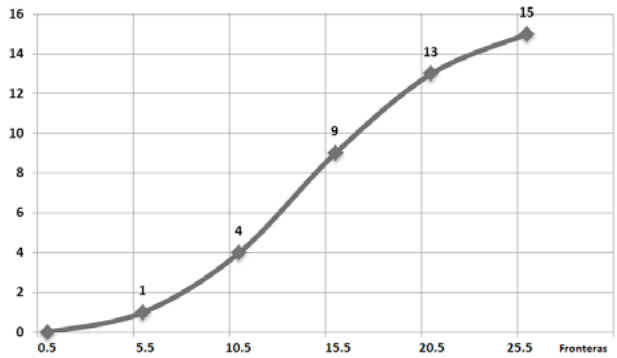


Figura 10. Ojiva de fronteras y frecuencias acumuladas

f) Ojiva de frecuencias relativas acumuladas

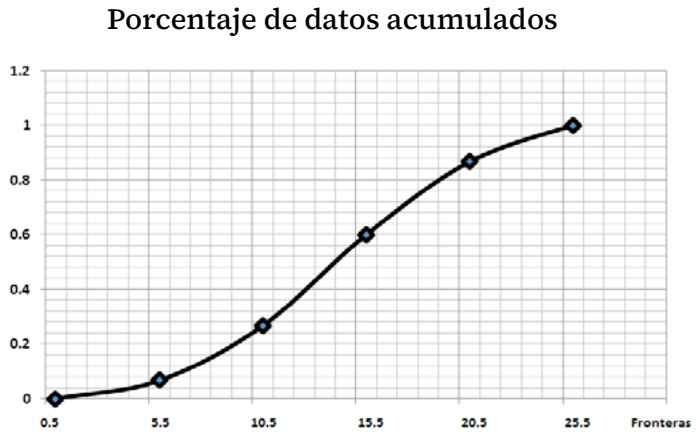


Figura 11. Ojiva de fronteras y frecuencias relativas acumuladas

2. En una consulta, a los principales funcionarios públicos y representantes por elección popular del país, sobre las posibilidades de construir una nueva Capital Federal, fuera del Valle de México, en algún lugar del territorio Nacional, se obtuvo los siguientes resultados:

Tabla 24. Respuestas para la construcción de una nueva capital federal

Funcionario o representante	Lo considera viable (*a*)	Lo considera no viable (*b*)	Indeciso (*c*)	Total
1. Secretario de Estado	15	3	2	20
2. Gobernador	20	6	1	27
3. Asesor Presidencial	6	8	4	18
4. Senador	35	35	18	88
5. Diputado	80	45	48	173
TOTAL	156	97	73	326

Completar la tabla 24 y responder las preguntas siguientes:

a) ¿Qué porcentaje de estos personajes están a favor de una nueva capital federal?
b) ¿Qué porcentaje de los senadores están a favor?
c) ¿Qué porcentaje del total de estos personajes corresponde a los gobernadores?
d) ¿Qué porcentaje de estos personajes no están a favor de una nueva capital federal?

Solución:

a) La siguiente tabla muestra los valores de las frecuencias relativas a favor de la nueva capital federal.

x_i	Frecuencia f_i	Frecuencia relativa f_i	Porcentaje
a: A favor	156	0.4785	47.85 %
b: No a favor	97	0.2975	29.39 %
c: Indeciso	73	0.2239	22.39 %
Sumas	326	1.0	100 %

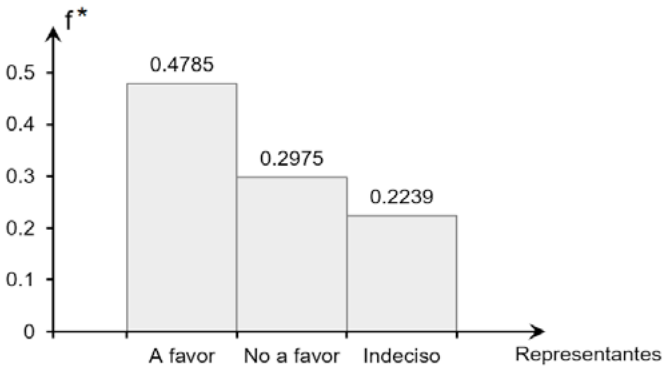


Figura 12. Histograma de las respuestas para la construcción de una nueva capital federal

b) ¿Qué porcentaje de los senadores están a favor?

x_i	Frecuencia f_i	Frec. relat. f_i^*	Porcentaje
a: A favor	35	0.3977	39.77 %
b: No a favor	35	0.3977	39.77 %
c: Indeciso	18	0.2045	20.45 %
Sumas	88	1.0	100 %

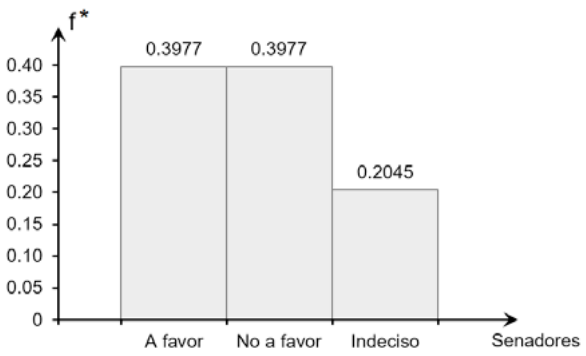


Figura 13. Histograma de senadores a favor de una nueva capital federal

El porcentaje de senadores a favor es del 39.77 %.

c) ¿Qué porcentaje del total de estos personajes corresponde a los gobernadores?

x_i	Frecuencia f_i	Frec. relat. f_i^*	Porcentaje
1	20	0.0613	6.13 %
2	27	0.0828	8.28 %
3	18	0.0552	5.52 %
4	88	0.2699	26.99 %
5	173	0.5307	53.07 %
Sumas	326	1.0	100 %

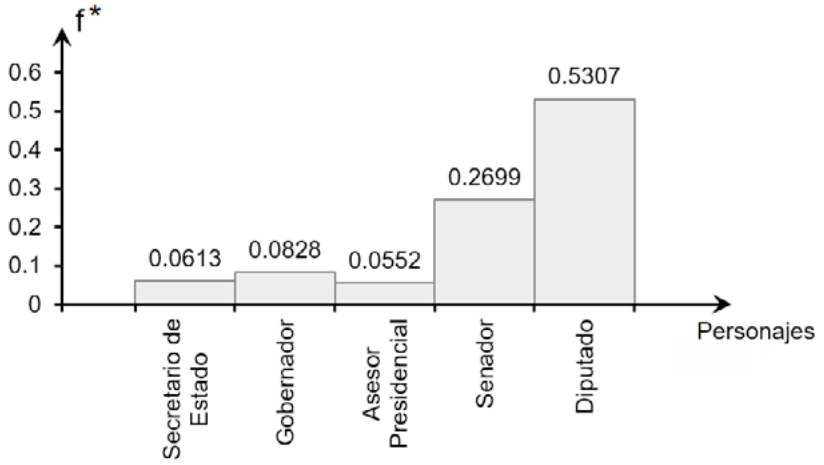


Figura 14. **Histograma de respuestas relativas a favor de una nueva capital federal**

El porcentaje de gobernadores es del 8.28 %.

d) ¿Qué porcentaje de los representantes considerarán no viable dicha construcción de una nueva capital federal?

Tabla 25. **Frecuencia para la no viabilidad de la construcción de una nueva capital federal**

x_i	Frecuencia f_i	Frec. relat. f_i^*	Porcentaje
a: A favor	156	0.4785	47.85 %
b: No a favor	97	0.2975	29.75 %
c: Indeciso	73	0.2239	22.9 %
Sumas	326	1.0	100 %

De la tabla 25 se observa que el porcentaje de representantes que consideran no viable la construcción de una nueva capital federal es del 29.75 %.

2.4 EJERCICIOS PROPUESTOS No. 1

1. Los datos que aparecen a continuación corresponden a los precios en el mercado de valores de las acciones de BBVA en los últimos 30 días. Completar la tabla asociada con dichos datos.

Tabla 26. **Precios de las acciones de BBVA**

Intervalo	Marcas de clase	Frecuencias	Frecuencias relativas	Frecuencias acumuladas	Frecuencias acumuladas relativas
1			0.033		
2	8.79				
3			0.167	8	
4					0.500
5	9.18		0.200		
6		8			
7					

2. Con la finalidad de optimizar el proceso de distribución y entrega a las refinerías en cierta región, Petróleos Mexicanos obtuvo información acerca de la

producción bruta m^3/día en 36 pozos seleccionados al azar dentro de la misma región. Los datos obtenidos fueron los siguientes (m^3/día):

21	75	42	36	12	132	9	39	30	12	60	12	39	9	33	21	12	66
27	39	0	6	63	33	57	6	105	15	105	24	21	12	72	12	15	72

Obtener la tabla de frecuencias, frecuencias relativas, frecuencias acumuladas y frecuencias relativas acumuladas para los datos anteriores. Hacerlo con seis intervalos de clase, con un ancho de clase de 23 unidades e iniciando en -0.5 como frontera inferior. Dibujar el histograma y la ojiva de frecuencias.

3. El gasto en pesos por consumo de energía eléctrica de cada uno de los 50 departamentos de un edificio en la colonia Narvarte se lista a continuación:

49.16	58.19	60.32	51.15	55.14	63.61	62.3	48.09	58.7	71.28
60.08	58.1	68.81	65.07	57.14	47.16	58.63	62.38	47.73	57.16
59.46	52.49	74.13	57.14	46.37	43.43	54.16	55.01	67.08	59.17
70.89	66.94	46.39	46.02	50.82	58.76	58.14	52.41	47.75	59.62
57.1	56.17	50.25	69.48	58.32	51.42	55.45	60.37	64	48.1

a) Obtenga una tabla de distribución de frecuencia relativa.
b) Elaborar un polígono de frecuencias.
c) ¿Qué porcentaje de departamentos pagan más de 55 pesos por consumo de energía?
d) Elaborar la ojiva de frecuencias relativas acumuladas.
e) ¿Qué porcentaje de datos hay menor que 47.16?

3 Análisis de datos univariados

Como se dijo antes, el objetivo de la Estadística Descriptiva es "explicar el comportamiento de los datos en la muestra", y es importante hacerlo de la manera más detallada posible ya que, siendo la única información disponible de la población, lo único que se podría inferir a ésta será lo que es posible decir acerca de la muestra.

Ciertamente, a través de las ordenaciones o agrupaciones de los datos y, especialmente mediante las formas gráficas, se puede hacer cierta descripción del comportamiento de la muestra; sin embargo, las percepciones obtenidas suelen ser inciertas, ya que dependen tanto de las escalas utilizadas en las gráficas, como de la habilidad que tenga el analista para observar y aproximar valores y tendencias a partir de una gráfica. Es necesario entonces, contar con formas descriptivas que, por estar asociadas solamente a los valores observados en la muestra, permiten hacer una descripción más precisa y certera. Estas herramientas o formas descriptivas reciben el nombre de ***medidas descriptivas de la muestra o estadísticas.***

Las medidas numéricas de una muestra se clasifican en tres tipos:

1) Medidas de tendencia central o medidas de posición
2) Medidas de dispersión o medidas de variabilidad de la muestra
3) Medidas de forma

Los datos contenidos en una muestra se pueden analizar ya sea sin agrupar o agrupándolos en tablas de distribución de frecuencias. Por esta razón, las medidas numéricas de la muestra se pueden calcular ya sea utilizando todos los datos en la muestra sin considerar agrupación alguna, o bien, a partir de la agrupación de ellos en una tabla.

En las secciones que siguen se definen y detallan las características de las medidas descriptivas más ampliamente utilizadas en la práctica y se explica la forma de interpretar cada una de ellas.

3.1 DATOS NO AGRUPADOS

Se entiende por una muestra con datos no agrupados aquella que basa su análisis en absolutamente todos los datos observados.

3.1.1 MEDIDAS DE TENDENCIA CENTRAL O DE POSICIÓN

Las medidas de tendencia central o posición son medidas características de los datos que siempre toman su valor dentro del rango de las observaciones y nunca fuera de él, de ahí su nombre de "medidas de tendencia central". Algunas de estas medidas se ubicarán ya sea en el punto medio del rango de la muestra o cercano a él, sin embargo, de la mayoría de dichas medidas solo se puede decir que se encontrarán dentro del rango de los datos, pero su posición posiblemente sea lejana al punto medio.

Las principales medidas de tendencia central son las siguientes:

a) Media aritmética
b) Media geométrica
c) Media armónica
d) Mediana
e) Moda

Media aritmética: Es el promedio de los datos contenidos en la muestra. Se define a la media aritmética como la suma de todos los datos en la muestra dividida entre el número de datos. Esto es,

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

en donde
x_i : i-ésimo dato en la muestra
n : número de datos en la muestra
Es importante resaltar que el cálculo de esta medida solo es aplicable para el análisis de datos cuantitativos.

Ejemplo 3.1. En un estudio realizado por la Procuraduría Federal del Consumidor (Profeco) en la Ciudad de México en 2003, en relación con el precio de un perfil tiroideo que incluyera la prueba de anticuerpos antitiroideos, en seis distintos laboratorios clínicos ubicados en la colonia Roma, se observaron los siguientes datos en pesos: 1810, 1670, 1300, 1100, 1405 y 750. Determine el precio promedio de dicho estudio.

Solución:

$$\begin{aligned}\bar{x} &= \frac{1}{6}\sum_{i=1}^{6} x_i \\ &= \frac{1}{6}\left(1810 + 1670 + 1300 + 1100 + 1405 + 750\right) \\ &= \frac{1}{6}\left(8035\right) \\ \bar{x} &= 1339.17\end{aligned}$$

Media geométrica: Dada una muestra con n elementos distintos, se llama media geométrica de esa muestra a la raíz n-ésima del producto de sus elementos, es decir

$$G = \bar{x}_g = \sqrt[n]{x_1\, x_2 \ldots x_n}$$

En general, cuando se repiten los datos, la fórmula de la media geométrica queda:

$$G = \bar{x}_g = \sqrt[n]{x_1^{n_1}\, x_2^{n_2} \ldots x_k^{n_k}}$$

en donde $n_1 + n_2 + \ldots + n_k = n$

Es importante hacer notar que la media geométrica tiene las siguientes restricciones:

1) No es útil si algún valor es nulo.
2) No debe calcularse si hay valores negativos.

Ejemplo 3.2. La media geométrica del siguiente conjunto de datos: 3, 7, 11 y 18 es:

$$G = \bar{x}_g = \sqrt[4]{(3)(7)(11)(18)} = \sqrt[4]{4158} = 8.0301$$

Ejemplo 3.3. La media geométrica del siguiente conjunto de datos: 3, 3, 3, 3, 7, 7, 11, 11, 11, 18, 18, 18, 18 y 18 es:

$$G = \bar{x}_g = \sqrt[14]{\left(3^4\right)\left(7^2\right)\left(11^3\right)\left(18^5\right)} = \sqrt[14]{\left(9.9820945\right)\left(10^{12}\right)} = 8.4823$$

Otra medida de tendencia central que es útil para el análisis de ciertos tipos de datos tales como aquellos que provienen de mediciones que cambian en función del tiempo, por ejemplo, en la productividad de un proceso o en datos hidrológicos es la media armónica.

Media armónica: La media armónica de un conjunto de valores x_1, x_2, x_3, ... , x_n correspondiente a una población o muestra, se determina como el recíproco de la media aritmética de los recíprocos de los valores, es decir:

$$H = \bar{x}_H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\ldots+\frac{1}{x_n}\right)}$$

Dada la definición de esta medida, su valor es siempre menor que el de la media aritmética.

Es importante resaltar que el cálculo de esta medida solo es aplicable para el tratamiento de datos cuantitativos.

Ejemplo 3.4. Determine la media armónica de la serie de números: 2,3,4,5,9.

Solución:

$$\bar{x}_H = \frac{1}{\frac{1}{5}\left(\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{9}\right)} = 3.5857$$

Según el tipo de datos que se analice será más apropiado utilizar la media aritmética, la media geométrica o la media armónica. Como se mencionó antes, algunos autores refieren que la media armónica suele usarse en datos cuya variabilidad depende del tiempo, en tanto que se recomienda el uso de la media geométrica cuando los datos corresponden a índices, porcentajes o proporciones. En estos casos, se considera que estas medidas son mejores alternativas que la media aritmética para analizar el conjunto de datos.

Mediana. La mediana de una muestra es el valor para el cual el 50% de los datos en ésta son menores o iguales a dicho valor. Se denota a la mediana de una muestra como $\tilde{x}$

Para calcular la mediana con datos no agrupados, es necesario ordenar los datos en forma ascendente. Por tanto, una forma metódica de hacer el cálculo debe realizar los pasos siguientes:

1. Ordenar los datos en forma ascendente.
2. Observar el tamaño de la muestra (n).
   2.1 Si n es par, la mediana será el promedio de los dos datos centrales en la ordenación.

$$x = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right)$$

   2.2 Si n es impar, la mediana será el dato central en la ordenación.

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

La mediana puede ser un dato observado si el número de datos en la muestra es impar, pero si dicho número de datos es par, la mediana será un valor teórico, es decir, no coincidirá con dato alguno.

Una característica de la mediana es que no está influida por los valores extremos ya que solamente considera uno o dos valores centrales, como muestra la metodología de cálculo presentada antes.

Ejemplo 3.5. Se observa cada minuto, durante un periodo de once, la entrada del número de clientes que acuden a una tienda que vende refacciones para automóvil. La tabla muestra los resultados:

Minuto	1	2	3	4	5	6	7	8	9	10	11
Número de cliente	25	34	16	9	13	27	32	18	15	21	20

Determinar la mediana del número de clientes por minuto que asisten a la tienda.

Solución:

1) Se ordenan los datos de manera ascendente: 9, 13, 15, 16, 18, 20, 21, 25, 27, 32, 34.

2) Dado que el número de observaciones es impar, entonces la mediana es el dato en la posición $\frac{11+1}{2} = 6$, por lo tanto, $\tilde{x} = 20$

Ejemplo 3.6. Un maratonista tiene los siguientes registros de tiempos (minutos), en las últimas seis competencias en las que ha participado: 132, 141, 382, 145, 148, 139. Calcular la media aritmética y la mediana de los tiempos registrados en las seis competencias.

Solución:

La media aritmética es la siguiente:

$$\bar{x} = \frac{132 + 141 + 382 + 145 + 148 + 139}{6} = 181.16$$

Para la mediana se ordenan los datos de manera ascendente:

132, 139. 141, 145, 148, 382

El número de datos es par

Por lo tanto, la mediana: $\tilde{x} = \frac{141+145}{2} = 143$

Conclusión: Puede observarse que el tiempo de la tercera competencia (382), está fuera de lo normal (es atípico), lo cual eleva el valor promedio o media de los tiempos empleados para recorrer los más de 42 kilómetros de un maratón. Por lo tanto, la mediana podría reflejar mejor el rendimiento del atleta, porque no está influida por los valores extremos.

> **Moda:** es el dato que aparece con mayor frecuencia en la muestra. Se denota a la moda como x_{moda}, x_{mo} o x_0 .

Ejemplo 3.7. Considere el siguiente conjunto de números: 1, 1, 1, 3, 4, 3, 4, 2, 3, 5, 9, 3. Determine el valor de la moda de esta muestra.

Solución:

Ordenando los datos en forma creciente se tiene 1, 1, 1, 2, 3, 3, 3, 3, 4, 4, 5, 9

La frecuencia mayor es 4 y el dato que aparece con mayor frecuencia es el 3. Por lo tanto, la moda es $x_{moda} = 3$.

Ejemplo 3.8. Determine la moda del conjunto de datos: 1, 1, 1, 2, 3, 2, 3, 2, 5, 8, 13, 13.

Solución:

Ordenando los datos en forma creciente se tiene 1, 1, 1, 2, 2, 2, 3, 3, 5, 8, 13,13 .

Claramente, los datos 1 y 2 aparecen el mismo número de veces, y son los que se presentan con mayor frecuencia. Por tal razón, se concluye que la muestra tiene dos modas:

$$x_{mo_1} = 1 \quad \text{y} \quad x_{mo_2} = 2$$

Por tal razón se dice que la distribución del conjunto de datos es bimodal.

Ejemplo 3.9. En la Ciudad de México el 12 % de los automóviles son de marcas europeas, el 52 % de marcas japonesas, el 33 % de marcas norteamericanas y el 3 % restante de otras marcas. Determinar la moda.

Solución:

Dado que el porcentaje mayor es 52 % y corresponde a marcas japonesas, por lo tanto, los automóviles de moda en la Ciudad de México son de marcas japonesas.

A pesar de no ser consideradas como medidas de tendencia central, existen otras medidas de las cuales la mediana es caso particular, que también se ubican dentro del rango de la muestra y que permiten hacer una mejor descripción de ella. Estas medidas son los percentiles o fractiles.

> **Percentiles o fractiles.** El percentil $(1-\alpha) \times 100$ % $(0 \leq \alpha \leq 1)$ de una muestra es el valor para el cual el $(1-\alpha) \times 100$ % de los datos en ésta son menores o iguales a dicho valor. Se denota al percentil $(1-\alpha) \times 100$ % de una muestra como $x_{(1-\alpha)}$. También se conoce a este percentil como fractil de $(1-\alpha)$. Claramente cuando $\alpha = 0.5$, probabilidad del 50 %, el valor del percentil es justamente la mediana de la muestra.

Para su cálculo, se pueden tomar en cuenta los datos agrupados en una tabla de distribución de frecuencias o se puede trabajar con datos no agrupados.

Para datos no agrupados una forma metódica para calcular el percentil $(1-\alpha) \times 100\ \%$ es seguir los pasos que se indican a continuación:

1. Ordenar los datos en forma creciente.
2. Si el tamaño de la muestra es n, determinar la posición en que se encuentra el percentil, mediante la operación $(1-\alpha)\ (n+1)$.
   2.1 Si $(1-\alpha)\ (n+1)$ es entero, $x_{(1-\alpha)}$ será el dato que se encuentre en el lugar $(1-\alpha)\ (n+1)$ dentro de la ordenación.
   2.2 Si $(1-\alpha)\ (n+1)$ es fraccionario, se identifican los datos adyacentes inferior (a_1), aquel que se encuentra en la posición $(1-\alpha)\ (n+1)$ y adyacente superior (a_s) el dato que esté en el lugar $[(1-\alpha)\ (n+1)]+1$ dentro de la ordenación. En donde, $[(1-\alpha)\ (n+1)]$ es la parte entera de $(1-\alpha)\ (n+1)$.

3. Determinar el valor del percentil $x_{(1-\alpha)}$ como

$$x_{(1-\alpha)} = a_1 + \textit{fracción de la posición}\ (a_s - a_1)$$

En donde *fracción de la posición* $= (1-\alpha)\ (n+1) - [(1-\alpha)\ (n+1)]$

De acuerdo con la forma de cálculo presentada, es claro que el percentil $x_{(1-\alpha)}$ puede ser un dato observado o no, dependiendo del percentil deseado y del tamaño de la muestra.

Existen algunos casos de percentiles o fractiles que son más ampliamente usados en la práctica. Estos son lo cuartiles primero, segundo y tercero ($x_{0.25} = q_1$, $x_{0.50} = q_2$ y $x_{0.75} = q_3$, respectivamente), que dividen a la muestra en cuartas partes, o bien los deciles primero o inferior, segundo, tercero, etc. ($x_{0.10}$, $x_{0.20}$, $x_{0.30}$, ... ,$x_{0.90}$) los cuales dividen a la muestra en décimas partes.

3.1.2 MEDIDAS DE DISPERSIÓN O VARIABILIDAD

Las medidas de tendencia central son muy importantes, pero no suficientes cuando se trata de describir el comportamiento de una muestra, ya que se pue-

den encontrar muestras que tengan las mismas medidas de tendencia central, pero la distancia entre los valores en la muestra no coincida, o bien que dichos valores se alejen más o menos de la media común. Por tanto, es importante la observación de las diferencias de los valores a la media $(x_i - \bar{x})$ con el fin de lograr obtener una mejor descripción de los datos en la muestra. A la diferencia $x_i - \bar{x}$ le denomina ***desviación***.

Ejemplo 3.10. Un supervisor en una fábrica de automóviles, después de revisar cinco de ellos, encontró el siguiente número de defectos por cada uno:

Tabla 27. Desviación de defectos en autos

Defectos x_i	Desviación $x_i - \bar{x}$
1	1 - 5 = - 4
4	4 - 5 = - 1
6	6 – 5 = 1
6	6 – 5 = 1
8	8 – 5 = 3
$\sum_i (x_i - \bar{x})$	0

$$\bar{x} = \frac{1+4+6+6+8}{5} = \frac{25}{5} = 5$$

Se puede observar:

1. Que la diferencia $(x_i - \bar{x})$ de las medidas mayores a la media tiene signo positivo
2. Que la diferencia $(x_i - \bar{x})$ de las medidas menores a la media tiene signo negativo
3. La suma de las desviaciones es cero, es decir, $\sum_i (x_i - \bar{x}) = 0$

Ejemplo 3.11. Sean dos conjuntos de datos:

						Medias
A:	8	9	10	11	12	$\bar{x}_A = 10$
B:	1	3	9	11	26	$\bar{x}_B = 10$

Los datos de los conjuntos A y B tienen la **misma media** pero, como se puede observar en la figura 15, muy **diferentes desviaciones**.

No obstante, en ambos casos la suma de las desviaciones es cero:

Para A: $(8-10)+(9-10)+(10-10)+(11-10)+(12-10) = \mathbf{-1} + (-1) + 0 + 1 + 1 = 0$

Para B: $(1-10)+(3-10)+(9-10)+(11-10)+(26-10) = \mathbf{-9} + (-7) + (-1) + 1 + 16 + = 0$

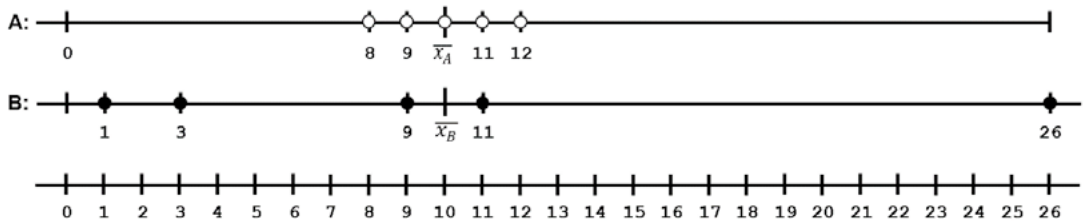


Figura 15. Diagrama de desviaciones de los conjuntos de datos *A* y *B*

En realidad, en todos los casos, $\sum_{i=1}^{n} \left(x_i - \bar{x} \right) = 0$, independientemente de cuáles sean los datos. Por tal razón, esta suma en ningún caso es útil como una medida de dispersión, y como se puede observar en los ejemplos anteriores, la cancelación se debe a la existencia de diferencias positivas unas y negativas otras que, en todos los casos, produce una compensación, hasta llevar la suma a cero. Por tal razón, varias de las medidas de dispersión de uso más generalizado en estadística, se basan en el uso de funciones que resuelvan el problema que ocasiona la diferencia de signos y, por tanto, la compensación.

Las medidas de dispersión más usadas en ingeniería son:

a) Rango
b) Rango intercuartil
c) Desviación media
d) Desviación mediana
e) Varianza
f) Desviación estándar
g) Coeficiente de variación

Rango: Como se definió en el inciso 2.2.1, el rango de la muestra es una medida de dispersión:

$$R = \text{Dato mayor} - \text{dato menor}$$

El rango es una primera medida de aproximación de la variabilidad, pero se ve afectado por los valores extremos.

Ejemplo 3.12. Las edades en años de un grupo familiar son: 30, 2, 1, 7, 32 y 10 años. El rango del grupo familiar es: $R = 32 - 1 = 31$ años

En el ejemplo anterior si agregamos a una persona (el abuelo) de 85 años, el rango se modifica drásticamente: $R = 85 - 1 = 84$ años.

Rango intercuartil o recorrido intercuartil. Esta medida es indiferente a los valores extremos. Se define como la diferencia entre el cuartil 3 y el cuartil 1:

$$RIq = q_3 - q_1$$

q_1 : aquel valor para el cual se tienen acumulados el 25% de los datos en él o debajo de él.
q_1: Es aquel valor que en él o debajo de él se tiene el 75% de los datos.

Ejemplo 3.13. Considérese la siguiente serie de datos: 14, 5, 6, 3, 1, 9, 15, 9, 11. Determinar los cuartiles: $x_{0.25}$, $x_{0.50}$, $x_{0.75}$ y el recorrido intercuartil.

Solución:

Se ordenan los datos en forma ascendente: 1, 3, 5, 6, 9, 9, 11, 14, 15

Primer cuartil: Se ubica la posición: $= 0.25\ (n+1) = 0.25\ (9+1) = 2.5$

Lo que indica que la posición de q_1, está entre los datos $a_I = 3$ (adyacente inferior) y $a_s = 5$ (adyacente superior). Por lo tanto:

$$q_1 = a_I + \textit{fracción de la posición}\ (a_s - a_I)$$
$$q_1 = 3 + 0.50\ (5 - 3) = 4$$

Segundo cuartil: Se ubica la posición: = 0.50 (n + 1) = 0.50 (9 + 1) = 5.00

Lo que indica que la posición de q_2, está entre los datos a_I=9 (adyacente inferior) y a_s = 9 (adyacente superior). Por lo tanto:

$$q_2 = a_I + \textit{fracción de la posición}\ (a_s - a_I)$$
$$q_2 = 9 + 0\ (\ 9 - 9\ ) = 9$$

Nótese que el segundo cuartil coincide con el valor de la mediana de los datos.
Tercer cuartil: Se ubica la posición = 0.75 (n + 1) = 0.75 (9 + 1) = 7.5

Lo que indica que la posición de q_3 está entre los datos a_1=11 (adyacente inferior) y a_s = 14 (adyacente superior). Por lo tanto:

$$q_3 = a_1 + \textit{fracción de la posición}\ (a_s - a_1)$$

$$q_3 = 11 + 0.50\ (14 - 11) = 11 + 2.25 = 12.5$$

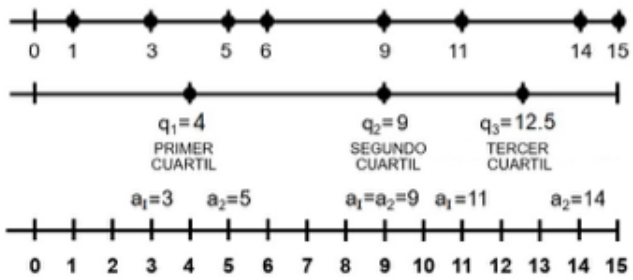


Figura 16. Diagrama de cuartiles

El recorrido intercuartil es: $RIq = q_3 - q_1 = 12.5 - 4 = 8.5$

Ejemplo 3.14. Considérese la siguiente serie de datos: 24, 38, 19, 35, 56, 35, 32, 40, 35, 33, 44, 35. Determinar los percentiles: $x_{0.53}$, $x_{0.82}$

Solución:

Se ordenan los datos en forma ascendente: 19, 24, 32, 33, 35, 35, 35, 35, 38, 40, 44, 56

Percentil 53: Se ubica la posición: = 0.53 (n + 1) = 0.53 (12 + 1) = 6.89

Lo que indica que la posición de $x_{0.53}$ está entre los datos $a_I = 35$ (adyacente inferior) y $a_s = 35$ (adyacente superior). Por lo tanto:

$$x_{0.53} = a_I + \textit{fracción de la posición}\ (a_s - a_I)$$

$$x_{0.53} = 35 + 0.89\ (35 - 35) = 35$$

Percentil 82: Se ubica la posición: $= 0.82(n+1) = 0.82\ (12 + 1) = 10.66$
Lo que indica que la posición de $x_{0.82}$ está entre los datos $a_I = 40$ (adyacente inferior) y $a_s = 44$ (adyacente superior). Por lo tanto:

$$x_{0.82} = a_I + \textit{fracción de la posición}\ (a_s - a_I)$$

$$x_{0.82} = 40 + 0.66\ (44 - 40) = 42.64$$

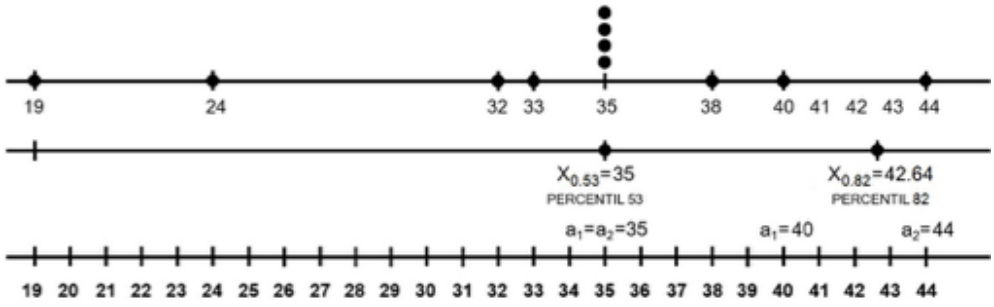


Figura 17. Diagrama de percentiles

Desviación media: La desviación media de una muestra, denotada como ***d.m.*** , es el promedio de las desviaciones absolutas de los datos a la media. Es decir,

$$d.m. = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

Desviación mediana: La desviación mediana de una muestra, denotada como ***d.md.***, es el promedio de las desviaciones absolutas de los datos a la mediana. Es decir,

$$d.md. = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \tilde{x}\right|$$

La suma de los cuadrados de las desviaciones de los datos a la media, ***Sum of the Square*** (SS por sus siglas en inglés), para una muestra constituida por ***n*** datos, corresponde a la suma que se muestra a continuación.

$$SS = \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$$

Ejemplo 3.15. Encontrar la *SS* de los resultados de puntaje obtenido de la muestra de cinco alumnos: 35, 42, 65, 70, 73 en un examen de Cálculo III.

Solución:

$$\bar{x} = \frac{1}{5}\left(35 + 42 + 65 + 70 + 73 \right) = 57$$

$$SS = \sum_{i=1}^{5} = \left(x_i - \bar{x} \right)^2 = (35-57)^2 + (42-57)^2 + (65-57)^2 + (70-57)^2 + (73-57)^2$$

$$= 484 + 225 + 64 + 169 + 256 = 1198$$

A partir del concepto de la suma de cuadrados de las diferencias de la media, se obtiene la varianza que es una medida muy importante de dispersión.

Varianza o variancia: La varianza de una muestra se puede definir como:

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

$n =$ número de datos de la muestra

Es de llamar la atención que un número grande de autores y profesores deciden utilizar la expresión de $S^2_{x(n-1)}$ como definición de la varianza de una muestra, en lugar de $S^2_{x(n)}$, sin mediar para ello una explicación.

En aras de la comprensión de la situación, y con fines explicativos, recordaremos que la varianza es una medida de la dispersión de los datos a la media. Claramente, esta distancia difiere de un dato a otro, por lo que un valor representativo es el promedio de los cuadrados[1] de las dispersiones.

Desde este punto de vista, una definición adecuada de la varianza de una muestra sería la que se explicó primero, aquella que corresponde a la suma de los cuadrados de las diferencias dividida entre ***n***, el número de datos. Si el objetivo es solamente describir el comportamiento de los datos en la muestra, esta es la definición correcta de la varianza. Sin embargo, generalmente, el objetivo de describir el comportamiento de la muestra es utilizar esta información para aproximar los valores de los parámetros de la población.

En este caso, en la teoría de la Inferencia Estadística, se demuestra que los valores de $S^2_{x(n-1)}$ son más cercanos al valor de la varianza de la población que aquellos obtenidos mediante el uso de la expresión $S^2_{x(n)}$. Por lo que se prefiere usar la primera expresión como medida de la dispersión de los datos de la muestra y aproximación a la varianza poblacional.

Independientemente de cuál de las dos expresiones se utilice para determinar el valor de la varianza de la muestra, el resultado tendrá unidades cuadráticas, lo cual dificulta su interpretación. Es necesario entonces, contar con una medida que refleje la variabilidad de los datos y, al mismo tiempo, sea fácil de interpretar una medida que tenga las mismas unidades que los datos en la muestra. Para ello, contamos ya con la desviación media, sin embargo, dado que involucra a la función valor absoluto, puede resultar un poco complicada para los usuarios.

Por lo anterior, se considera como una opción más sencilla, a una medida definida como la raíz cuadrada de la varianza. Esta medida se llama ***desviación estándar.***

[1] Recordar que previamente se explicó que en este trabajo se utilizan los cuadrados de las dispersiones, ya que por haber dispersiones positivas y negativas en una misma muestra, la suma de las dispersiones lineales es cero y, por tanto, la media también lo es, razón por la cual no permite medir la dispersión de los datos.

Desviación estándar: Se define como la raíz cuadrada positiva de la varianza, se denota por:

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

La desviación estándar tiene las mismas unidades que los datos y la expresión que se use para determinarla, responde a las circunstancias antes descritas.

Ejemplo 3.16. Según los datos tomados de la página del Instituto Nacional de Estadística y Geografía, INEGI, en los meses de mayo y junio de 2014 se han registrado 13 temblores en la ciudad de México, las intensidades medidas en grados Richter fueron:

6.4	6.9	5.0	4.1	4.0	4.0	4.0	4.1	6.2	4.4	4.0	4.0	4.5

Determinar la desviación estándar de las intensidades de los temblores mencionados.

Solución:

Para determinar la desviación estándar de la información se utiliza la expresión

$$S_{x(n\text{-}1)} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}$$

Se requiere calcular la media de los datos

$$\bar{x} = \frac{6.4 + 6.9 + 5.0 + 4.1 + 4.0 + 4.0 + 4.0 + 4.1 + 6.2 + 4.4 + 4.0 + 4.0 + 4.5}{13} = 4.738$$

$$SS = \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 = \left(6.4 - 4.7 \right)^2 + \left(6.9 - 4.7 \right)^2 + \left(5.0 - 4.7 \right)^2 + \left(4.1 - 4.7 \right)^2 + \left(4.0 - 4.7 \right)^2 +$$

$$+ \left(4.0 - 4.7 \right)^2 + \left(4.0 - 4.7 \right)^2 + \left(4.1 - 4.7 \right)^2 + \left(6.2 - 4.7 \right)^2 + \left(4.4 - 4.7 \right)^2 +$$

$$+ \left(4.0 - 4.7 \right)^2 + \left(4.0 - 4.7 \right)^2 + \left(4.5 - 4.7 \right)^2 = 13.4$$

$$S_{x(n-1)} = \sqrt{\frac{13.4}{12}} = \sqrt{1.11} = 1.055$$

Coeficiente de variación: Es una medida de dispersión de una muestra que se define como el cociente de la desviación estándar con respecto a la media de la muestra. Y se denota como *C.V.* es decir

$$C.V. = \frac{S_x}{\bar{x}}$$

El *C.V.* es una medida adimensional y se puede interpretar como la proporción de la media que representa la desviación estándar.

Valores cercanos a cero indican baja dispersión independientemente de lo que se esté midiendo.

Ejemplo 3.17. Frecuentemente se requiere la estimación de crecientes en ríos, producto de las lluvias en una zona donde posiblemente se construirán presas, derivadoras u otras obras hidráulicas para generar energía eléctrica, regar tierras o controlar avenidas; generalmente se recurre a las estaciones pluviométricas existentes, dentro del área de estudio, para obtener las características de las tormentas de la zona y luego transformarlas en tormentas de diseño que establecen los gastos máximos mediante una relación lluvia-escurrimiento. La información que se presenta en las dos primeras columnas del cuadro corresponde a la que está disponible sobre precipitación máxima diaria anual (mm) en el sistema ERIC II (IMTA, 2000) de 15 estaciones pluviométricas del altiplano Potosino del estado de San Luis Potosi, dentro de la Región Hidrológica No. 37 (El Salado), con más de 30 años de datos, la cual fue publicada en el artículo: "Contraste de la distribución TERC en el Altiplano Potosino", Vol XII, Núm. 2, abril-junio de 2011, en la revista "Ingeniería" de la Facultad de Ingeniería, UNAM.

A partir de la información presentada, determinar la media, la desviación estándar y el coeficiente de variación de la precipitación máxima diaria (mm) en la Región Hidrológica No. 37, del Altiplano Potosino.

Tabla 28. Observaciones de precipitación pluvial

Estación pluviométrica	Precipitación máxima observada (mm)	Estación pluviométrica	Precipitación máxima observada (mm)
1	315.7	9	66.5
2	145.0	10	100.0
3	86.5	11	200.6
4	140.1	12	120.0
5	81.5	13	77.0
6	122.0	14	200.0
7	84.0	15	101.0
8	97.5		

Solución:

$$\bar{x} = \frac{315.7+145+86.5+140.1+81.5+122+84+97.5+66.5+100+200.6+120+77+200+101}{15}$$

$$\bar{x} = \frac{1937.4}{15} = 129.16 \text{ mm}$$

$$SS = \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 = \left(315.7 - 129.16 \right)^2 + \left(145 - 129.16 \right)^2 + \left(86.6 - 129.16 \right)^2 +$$

$$+ \left(140.1 - 129.16 \right)^2 + \left(81.5 - 129.16 \right)^2 + \left(122 - 129.16 \right)^2 + \left(84 - 129.16 \right)^2 +$$

$$+ \left(97.5 - 129.16 \right)^2 + \left(66.5 - 129.16 \right)^2 + \left(100 - 129.16 \right)^2 + \left(200.6 - 129.16 \right)^2 +$$

$$+ \left(120 - 129.16 \right)^2 + \left(77 - 129.16 \right)^2 + \left(200 - 129.16 \right)^2 + \left(101 - 129.16 \right)^2$$

$$SS = \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 = 60848.276$$

$$S^2_{x(n)} = 4056.551733$$

$$S^2_{x(n-1)} = 4346.305429$$

$$S_{x(n)} = 63.691065$$

$$S^2_{x(n-1)} = 65.926515$$

$$c.v. = \frac{s}{\bar{x}} = \frac{65.926515}{129.16} = 0.5104$$

Se puede observar que la dispersión en la precipitación máxima observada es de aproximadamente 51 % de la media, es decir, no es una variabilidad pequeña, sino que las mediciones pueden variar de manera significativa. Probablemente, esta variabilidad en las mediciones esté asociada a la diferencia en las zonas en las que están ubicadas las estaciones de medición, que en algunos casos es bosque y en otra, zona árida.

Tal vez se podría concluir que, aunque la media es una medida obtenida de la precipitación máxima observada, en caso de utilizarla como parámetro de diseño o con el fin de tomar decisiones, es necesario ser cuidadoso y considerar que los valores esperados para dicha medida de precipitación estarán en un rango de más o menos 65.93 mm alrededor de la media.

3.1.3 MOMENTOS

En estadística, los momentos son herramientas que permiten describir el comportamiento de una muestra, ya que algunos momentos corresponden a medidas de tendencia central y otros, a medidas de dispersión. Dado que son medidas que la Estadística toma de la Mecánica, la forma de cálculo de los momentos es la misma que la que determina esta última, solamente que, en el ámbito de dicha asignatura, estas medidas siempre tienen una interpretación física, en tanto que, en estadística, algunos momentos particulares tienen una interpretación y otros, la mayoría no; aunque solo se utilizan como herramienta para la determinación de otros parámetros.

Es posible definir los momentos con respecto a cualquier punto “a”, pero en estadística los más usuales son con respecto al origen y con respecto a la media.

Momentos con respecto al origen: Se define el *r*-ésimo momento de la muestra con respecto al origen como:

$$m'_r = \frac{\sum_{i=1}^{n} x_i^r}{n}$$

Momentos con respecto a la media: El *r*-ésimo momento o momento de orden r, de la muestra con respecto a la media se define como:

$$m_r = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^r}{n}$$

Ejemplo 3.18. Sea una muestra cuyos valores son 5,7,12 y 16. Determinar el tercer momento con respecto a la media, si se sabe que el primer momento con respecto al origen es: $m_1 = 10$

Solución:

$$m_1 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^1}{n} = \frac{(5-10)+(7-10)+(12-10)+(16-10)}{4} = 0$$

$$m_2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2}{n} = \frac{(5-10)^2+(7-10)^2+(12-10)^2+(16-10)^2}{4} = 18.5$$

$$m_3 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^3}{n} = \frac{(5-10)^3+(7-10)^3+(12-10)^3+(16-10)^3}{4} = 4$$

3.1.4 MEDIDAS DE FORMA[2]

Coeficiente de sesgo de Fisher: El coeficiente de sesgo o coeficiente de asimetría es un parámetro de forma de la muestra que permite identificar comportamientos simétricos o asimétricos de los datos. Se denota al coeficiente de sesgo de una muestra por a_3 y se define como:

$$a_3 = \frac{m_3}{s_x^3}$$

Donde m_3 es el tercer momento respecto a la media.
De esta forma, si:
$a_3 < 0$ sesgo negativo
$a_3 > 0$ sesgo positivo
$a_3 = 0$ simétrica

Coeficiente de curtosis: es un parámetro de forma de la muestra que permite medir el grado de apuntalamiento o pronunciamiento de la distribución de los datos. Se denota al coeficiente de sesgo de una muestra por a_4 y se define como:

$$a_4 = \frac{m_4}{s_x^4}$$

Donde m_4 es el cuarto momento respecto a la media.
De esta forma, si:

$a_4 < 0$ se dice que la distribución es platicúrtica o achatada
$a_4 > 0$ se habla de que la distribución es leptocúrtica, es decir, puntiaguda
$a_4 = 0$ se concluye que la distribución de la muestra es mesocúrtica, como una meseta.

[2] El 3 sale de la comparación de la curva patrón que es la normal, ya que en ella el coeficiente de sesgo es cero y curtosis es 3.

Ejemplo 3.19. La densidad de población de algunas entidades seleccionadas al azar reportada por el INEGI en el año 2010 se muestra en la siguiente tabla:

Tabla 29. Densidad de población en México en 2010

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
Aguascalientes	Colima	Jalisco	Guanajuato	Hidalgo	México	Morelos	Puebla	Querétaro	Tlaxcala
211	116	94	179	128	679	364	168	156	293

Describir el comportamiento de la muestra.

Solución:

El cálculo de la media:

$$\bar{x} = \frac{211+116+94+179+128+679+364+168+156+293}{10} = \frac{2388}{10}$$

$\bar{x} = 238.8$ este valor corresponde al primer momento respecto al origen.

La varianza de una muestra se obtiene mediante $s^2_{x(n-1)} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$

$$\sum_{i=1}^{10}\left(x_i - \bar{x}\right)^2 = \left(211-238.8\right)^2 + \left(116-238.8\right)^2 + \left(94-238.8\right)^2 + \left(179-238.8\right)^2 + \left(128-238.8\right)^2$$
$$+ \left(679-238.8\right)^2 + \left(364-238.8\right)^2 + \left(168-238.8\right)^2 + \left(156-238.8\right)^2 + \left(293-238.8\right)^2$$

$$\sum_{i=1}^{10}\left(x_i - \bar{x}\right)^2 = 276929.6$$

$$s^2_{x(n)} = \frac{276929}{10} = 27692.9 \quad ; \quad s_{x(n)} = 166.41$$

$$s^2_{x(n-1)} = \frac{SS}{n-1} = \frac{276929}{10-1} = 30769.88 \; ; \quad s_{x(n-1)} = 175.41$$

Este valor no necesariamente es el mismo que el obtenido mediante momentos debido a que a través del método de los momentos, la varianza calculada es la SS dividida entre n.

$$s^2_{x(n)} = m'_2 - \left(m'_1\right)^2 = \frac{1}{n}\sum_{i=1}^{10}\left(x_i\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{10} x_i\right)^2$$

Para obtener los parámetros de forma mediante momentos, se recurre a la expresión:

$$m_3 = \frac{1}{10}\sum_{i=1}^{10}\left(x_i - \bar{x}\right)^3 =$$

$$= \frac{1}{10}\left[\left(211-238.8\right)^2 + \left(116-238.8\right)^2 + \left(94-238.8\right)^2 + \left(179-238.8\right)^2 + \left(128-238.8\right)^2 + \left(679-238.8\right)^2 + \left(364-238.8\right)^2 + \left(168-238.8\right)^2 + \left(156-238.8\right)^2 + \left(293-238.8\right)^2\right]$$

$$m_3 = 8001597.4$$

Cálculo del sesgo:

$$a_3 = \frac{m_3}{s^3_{x\,(n)}} = \frac{8001597.4}{\left(175.41\right)^3} = 1.48$$

$$\sum_{i=1}^{10}\left(x_i - \bar{x}\right)^4 = \left(211-238.8\right)^4 + \left(116-238.8\right)^4 + \left(94-238.8\right)^4 + \left(179-238.8\right)^4 + \left(128-238.8\right)^4 + \left(679-238.8\right)^4 + \left(364-238.8\right)^4 + \left(168-238.8\right)^4 + \left(156-238.8\right)^4 + \left(293-238.8\right)^4$$

$$m_4 = \frac{1}{10}\sum_{i=1}^{10}\left(x_i - \bar{x}\right)^4 = \frac{1}{10}\left(38706738960\right) = 3870673896$$

Cálculo de la curtosis:

$$a_4 = \frac{m_4}{s^4_x} = \frac{3870673896}{\left(175.41\right)^4} = 4.09$$

Los resultados mostrados se pueden interpretar de la siguiente manera:

Se puede observar que, en promedio, hay 238.8 habitantes por km2, es decir, no parecen ser zonas muy pobladas aunque la densidad de cada una, en promedio, podría variar entre $\bar{x} - s_x = 238.8 - 166.41$ y $\bar{x} + s_x = 238.8 + 166.41$; esto es, una alta proporción de las ciudades consideradas tienen una densidad de población entre 72.39 y 405.21 habitantes/km^2.

Estos valores muestran que hay una fuerte dispersión entre los datos. En relación con la forma de la distribución, el coeficiente de sesgo de 1.48 muestra que se tiene un sesgo positivo, es decir, la curva se extiende más a la derecha de la media y son más frecuentes las densidades pequeñas que las grandes, y el coeficiente de curtosis de 4.09 > 3 muestra que la curva es pronunciada hacia arriba, lo cual se expresaría diciendo que la distribución es leptocúrtica.

Cuando los conjuntos de datos obtenidos en un experimento tienen una distribución de frecuencias simétrica, coinciden la media, la mediana y la moda.

Ejemplo 3.20. Considérese que la mayoría de los fabricantes de llantas, para automóviles que utilizan rin 14, especifica que la presión óptima para éstas, en la parte delantera, es de 30 psi (pounds-force per square inch). Un equipo de investigación del Departamento de Tránsito de cierta ciudad, a fin de hacer una evaluación de la responsabilidad de los usuarios de los vehículos ante esta norma de eficiencia de gasto de gasolina y estabilidad estructural del automóvil mide, en una determinada estación de monitoreo, la presión de una llanta delantera para una muestra aleatoria de 8 vehículos con el mencionado tipo de rin. Los datos obtenidos de este ejercicio fueron los siguientes: 29.1, 28.7, 24.3, 33.1, 29.3. 17.1, 27.6, 31.2.

Describir el comportamiento de la muestra:

Determinar las siguientes medidas características que describe el comportamiento de la muestra:

a) Media
b) Desviación estándar
c) Desviación media
d) Desviación mediana

Solución:

Tabla 30. **Frecuencias para la eficiencia de gasto de gasolina y estabilidad estructural del auto**

	Datos					
	x_i	$x_i - \bar{x}$	$x_i - \tilde{x}$	$\left(x_i - \bar{x}\right)^2$	$\lvert x_i - \bar{x} \rvert$	$\lvert x_i - \tilde{x} \rvert$
	29.1	1.55	0.4	2.4025	1.55	0.4
	28.7	1.15	0	1.3225	1.15	0
	24.3	-3.25	-4.4	10.5625	3.25	4.4
	33.1	5.55	4.4	30.825	5.55	4.4
	29.3	1.75	0.6	3.0625	1.75	0.6
	17.1	−10.45	-5.8	109.2025	10.45	11.6
	27.6	0.05	-1.1	0.0025	0.05	1.1
	31.2	3.65	2.5	13.3225	3.65	2.5
Sumas	220.4	2.84217E-14	-2.1	170.68	27.40	25

a) Media:

$$\bar{x} = \frac{\sum_{i=1}^{n}\left(x_i\right)}{n} = \frac{220.4}{8} = 27.55$$

b) Desviación estándar:

$$s_{x(n)} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}} = \sqrt{\frac{170.68}{8}} = 4.8169$$

$$s_{x(n-1)} = \sqrt{\frac{170.68}{7}} = 4.9379$$

c) Desviación media:

$$d.m. = \frac{\sum_{i=1}^{n} \left| x_i - \bar{x} \right|}{n} = \frac{27.4}{8} = 3.425$$

d) Desviación mediana:

$$d.md = \frac{\sum_{i=1}^{n} \left| x_i - \tilde{x} \right|}{n} = \frac{25}{8} = 3.125$$

3.1.5 MEDIDAS DESCRIPTIVAS CON DATOS QUE SE REPITEN

Si en la muestra de tamaño n se tienen datos repetidos: x_1, x_2, x_3,..., x_n, donde x_1 se repite f_1 veces (frecuencia f_1), x_2 se repite f_2 veces, x_3 se repite f_3 veces,..., x_k se repite f_k veces; resumiendo, las medidas que se pueden obtener con la información:

Medida	Expresión
Media	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Varianza	$S^2_{x(n-1)} = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$
Desviación media	$d.m. = f_i \sum_{i=1}^{n} \left| x_i - \bar{x} \right|$
Desviación mediana	$d.md. = f_i \sum_{i=1}^{n} \left| x_i - \tilde{x} \right|$
Sesgo	$a_3 = \frac{m_3}{s_x^3}$
Curtosis	$a_4 = \frac{m_4}{S_x^4}$

son los momentos principales descritos al principio de este apartado, bajo la condición de datos repetidos, están dados por:

El *r-ésimo* momento con respecto al origen:

$$m'_r = \frac{\sum_{i=1}^{k} f_i \, x_i^r}{n}$$

El *r-ésimo* momento con respecto a la media:

$$m_r = \frac{\sum_{i=1}^{k} f_i \left(x_i - \bar{x} \right)^r}{n}$$

Las expresiones matemáticas anteriores pueden utilizarse para datos no agrupados considerando a x_i con su valor original y en datos agrupados como el valor de las marcas de clase correspondiente a cada clase.

Ejemplo 3.21. Sea una muestra de 12 valores: 1, 1, 3, 3, 3, 5, 5, 5, 5, 7, 7, 9. Determinar:

a) Rango
b) Desviación estándar
c) Recorrido intercuartil

Solución:

Frecuencias, marcas de clase y desviación al cuadrado

x_i	f_i	$x_i - \bar{x}$	$\left(x_i - \bar{x}\right)^2$	$f_i \left(x_i - \bar{x}\right)^2$
1	2	-3.5	12.25	24.5
3	3	-1.5	2.25	6.75
5	4	0.5	0.25	1.00
7	2	2.5	6.25	12.50
9	1	4.5	20.25	20.25
Sumas	12		41.25	65.00

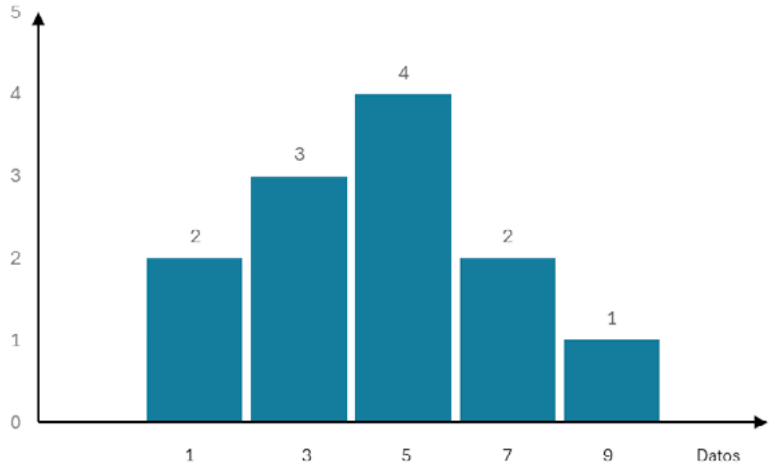
a) Rango = 9 − 1 = 8

b) Desviación estándar

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i\, x_i}{n} = \frac{(2)(1)+(3)(3)+(5)(4)+(7)(2)+(9)(1)}{12} = 4.5$$

$$s_{x(n)} = \sqrt{\frac{65}{11}} = 2.327$$

$$s_{x(n-1)} = \sqrt{\frac{\sum_{i=0}^{n} f_i \left(x_i - \bar{x}\right)^2}{n-1}} = \sqrt{\frac{41.25}{11}} = 3.75$$



c) Recorrido intercuartil

$RIq = q_3 - q_1$

$q_3 = a_1 +$ *fracción de la posición* $(a_s - a_i)$

El tercer cuartil se ubica en la posición 0.75 $(n+1)$ = 0.75 (13) = 9.75 lo que indica que la posición de q_3 está entre los datos $a_1 = 5$ (adyacente inferior) y $a_5 = 7$ (adyacente superior), por lo tanto, $q_3 = 5 + 0.75\ (7 - 5) = 6.5$

El primer cuartil se ubica en la posición 0.25 $(n+1)$ = 0.25 (13) = 3.25, lo que indica que la posición de q_1 está entre los datos $a_1 = 3$ (adyacente inferior) y $a_3 = 3$ (adyacente superior), por lo tanto, $q_1 = 3 + 0.25\ (3 - 3) = 3$

$$RIq = 6.5 - 3 = 3.50$$

Ejemplo 3.22. De la siguiente muestra de datos: 5, 7, 12, 16, determinar el primero, el segundo y el tercer momento con respecto al origen:

Solución:

$$m_1' = \frac{\sum_{i=1}^{n} \left(x_i^1\right)}{n} = \frac{5+7+12+16}{4} = 10.0$$

$$m_2' = \frac{\sum_{i=1}^{n} \left(x_i^2\right)}{n} = \frac{5^2+7^2+12^2+16^2}{4} = 118.5$$

$$m_3' = \frac{\sum_{i=1}^{n} \left(x_i^3\right)}{n} = \frac{5^3+7^3+12^3+16^3}{4} = 1573$$

Ejemplo 3.23. En la tabla se muestra una distribución de frecuencias correspondientes a las edades de una muestra de 17 personas del poblado San Pedro Mártir en la Ciudad de México, que se presentó el primer día de campaña al hospital regional de la zona para recibir una determinada vacuna que promueve el Sector Salud del Gobierno Federal:

i	Intervalo de clase	Frecuencia (f_i)	Marca de clase x_i
1	1 – 8	1	4.5
2	9 – 16	4	12.5
3	17 – 24	7	20.5
4	25 – 32	3	28.5
5	33 – 40	2	36.5
	Sumas	17	102.5

a) Obtener los tres primeros momentos con respecto al origen de la distribución de edades.
b) Determinar el segundo y cuarto momento con respecto a la media de la distribución de las edades

Solución:

i	Intervalo de clase	Frecuencia (f_i)	Marca de clase x_i	$f_i\ x_i$	$f_i\ x_i^2$	$f_i\ x_i^3$	$f_i(x_i-\bar{x})^2$	$f_i(x_i-\bar{x})^4$
1	1 - 8	1	4.5	4.5	20.25	91.125	271.2609	73582.475
2	9 - 16	4	12.5	50.0	625.00	7812.500	286.9636	20587.026
3	17 - 24	7	20.5	143.5	2941.75	60305.875	1.5463	0.341
4	25 - 32	3	28.5	85.5	2436.75	69447.375	170.1027	9644.976
5	33 - 40	2	36.5	73.0	2664.50	97254.250	482.3618	116336.453
	sumas:	17	102.5	356.5	8688.25	234911.125	1212.2353	220151.272

a.1) El primer momento con respecto al origen:

$$\bar{x} = m_1' = \frac{\sum_{i=1}^{c} f_i\left(x_i\right)}{n} = \frac{356.5}{17} = 20.97$$

donde c es el número de intervalos de clase.

a.2) El segundo momento con respecto al origen:

$$m_2' = \frac{\sum_{i=1}^{c} f_i\left(x_i^2\right)}{n} = \frac{8688.25}{17} = 511.07$$

a.3) El tercer momento con respecto al origen:

$$m_3' = \frac{\sum_{i=1}^{c} f_i\left(x_i^3\right)}{n} = \frac{234911.125}{17} = 13818.30$$

b.1) El segundo momento con respecto a la media (varianza de x).

$$m_2 = \frac{\sum_{i=1}^{c} f_i \left(x_i - \bar{x} \right)^2}{n} = \frac{1212.2353}{17} = 71.307$$

b.2) El cuarto momento con respecto a la media

$$m_4 = \frac{\sum_{i=1}^{c} f_i \left(x_i - \bar{x} \right)^4}{n} = \frac{220151.272}{17} = 12950.074$$

3.2 DATOS AGRUPADOS

Cuando se tiene una gran cantidad de datos y se ha decidido agruparlos en una tabla de frecuencias, es posible calcular a partir de ella las mismas medidas que se obtuvieron para los datos no agrupados.

3.2.1 MEDIDAS DE TENDENCIA CENTRAL PARA DATOS AGRUPADOS

En algunas ocasiones, la única información disponible es una tabla de distribución de frecuencias. En estos casos, solo es posible obtener valores aproximados para la media, la mediana y la moda.

Media. Es la medida de tendencia central que se espera obtener en un experimento estadístico, también se conoce como promedio, valor esperado o esperanza matemática.

Se determina con base en las marcas de clase y la ponderación (frecuencias) de cada una de ellas por cada intervalo de clase:

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

Donde:

x_i: Marca de clase del intervalo i
f_i: Frecuencia de clase i
n: Número total de datos en la muestra
c: Número de clases

Media geométrica: Dada una muestra con n elementos distintos, su media geométrica cuando los datos están agrupados se calcula mediante

$$\log G = \frac{\sum_{i=1}^{c} f_i \log (x_i)}{\sum_{i=1}^{c} f_i}$$

Donde G es la media geométrica (Murray Pi, 1974), al despejar a G queda:

$$G = \log^{-1} \frac{\sum_{i=1}^{c} f_i \log (x_i)}{\sum_{i=1}^{c} f_i}$$

Donde:

x_i: Marca de clase del intervalo i
c : Número de intervalos
f_i: Frecuencia de la clase i

Es importante hacer notar que la media geométrica tiene las siguientes restricciones:

1) No es útil si algún valor es nulo.
2) No es posible su cálculo cuando hay un número par de datos y el radicando es negativo.

Media armónica. El concepto de media armónica es otro de los tipos de media que puede encontrarse en estadística. La media armónica tiene como particularidad que parte del principio de reciprocidad, el cual tiene que ver con la inversa de la inversión multiplicativa $H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$.

Esto significa que, si tenemos un valor cualquiera, podemos reconocer su inverso multiplicativo como $\frac{1}{x}$. A través del principio de reciprocidad podemos definir también a x como el inverso de dicho inverso multiplicativo, es decir $x = \frac{1}{\frac{1}{x}}$

Moda. En el caso de datos agrupados, la moda se puede definir como la marca de la clase con mayor frecuencia. A la moda calculada de esta manera también se le conoce como ***moda cruda***.

Es importante notar que en una misma tabla de distribución de frecuencias puede haber una o más clases modales. En el caso en que todas las clases tengan la misma frecuencia, se dirá que todas las clases son modales o bien que no existe clase modal, esto de la misma manera que se concluyó para el caso de los datos no agrupados.

Cuando en un conjunto de datos agrupados hay dos o más marcas de clase con la misma frecuencia, se dice que es multimodal.

Clase modal. Se conoce como ***clase modal*** en una tabla de distribución de frecuencias, al intervalo de clase que tiene la frecuencia de clase más alta.

Forma de cálculo de la moda, por interpolación

Paso 1
Identifique la clase modal, aquel intervalo con la mayor frecuencia de clase f_i

Paso 2
Determine la frontera inferior $front\ inf_{(moda)}$ y la frontera superior $front\ sup_{(moda)}$ de la clase modal.

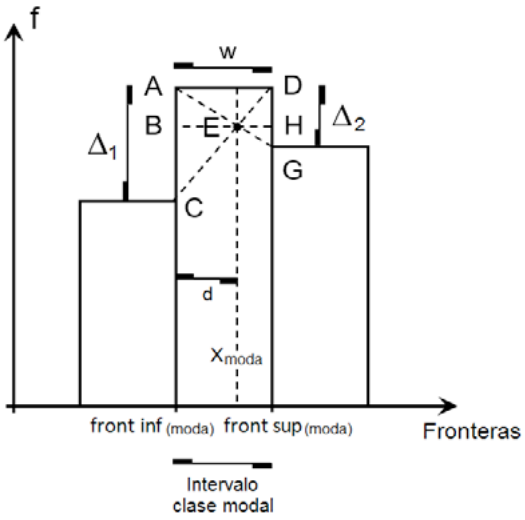


Figura 18. Cálculo de moda

Paso 3

Como se muestra en la figura 18, identifique los siguientes puntos en el histograma de frecuencias.

$$A = \left(front\ inf_{(moda)},\ f_k \right)$$
$$C = \left(front\ inf_{(moda)},\ f_{k-1} \right)$$
$$D = \left(front\ sup_{(moda)},\ f_k \right)$$
$$G = \left(front\ sup_{(moda)},\ f_{k+1} \right)$$

en donde,

f_k : Frecuencia de la clase modal
f_{k-1}: Frecuencia de la clase anterior a la clase modal
f_{k+1}: Frecuencia de la clase siguiente a la clase modal

Paso 4

También en la figura 18, trace el segmento de recta que une los puntos A y G, así como el segmento de recta que une los puntos C y D. Llame E al punto de intersección de los segmentos $\overline{AC}$ y $\overline{CD}$. Las coordenadas del punto E serán E=(x_{mo}, f_k), en donde f_k es la frecuencia de la moda.

Paso 5
Identifique los puntos

$$B = \left(front\ inf_{(moda)},\ f_k \right)$$
$$H = \left(front\ sup_{(moda)},\ f_k \right)$$

en la figura 18.

Paso 6
Mediante el procedimiento que se muestra a continuación, realice la interpolación para determinar la magnitud $d = |\overline{BE}|$ que se muestra en la figura 18.

Observe que los triángulos Δ_1 ACE y Δ_2 DGE son triángulos semejantes. Por lo tanto, se cumple la ecuación

$$\frac{\Delta_1}{|BE|} = \frac{\Delta_2}{|EH|} \qquad ...(1)$$

en donde

y
$$\Delta_1 = |\overline{AB}| + |\overline{BC}|$$
$$\Delta_2 = |\overline{DH}| + |\overline{HG}|$$

Note que

$$w = |\overline{BE}| + |\overline{EH}|$$

es decir,

$$w = d + |\overline{EH}| \qquad ...(2)$$

De las ecuaciones (1) y (2) se tiene

$$\frac{\Delta_1}{\Delta_2} = \frac{|\overline{BE}|}{|\overline{EH}|} = \frac{d}{|\overline{EH}|} \qquad ...(3)$$

Pero como se puede observar en la figura 18

$$\left|\overline{EH}\right| = w - d$$

$$d = \frac{\Delta_1 \left|\overline{\mathrm{EH}}\right|}{\Delta_2}$$

Sustituyendo en (3):

$$d = \frac{\Delta_1 \left(w - d \right)}{\Delta_2}$$

Desarrollando

$$d\Delta_2 = \Delta_1 \left(w - d \right)$$

$$d\Delta_2 = \Delta_1 w - d\Delta_1$$

$$d\Delta_2 + d\Delta_1 = \Delta_1 w$$

$$d\left(\Delta_1 + \Delta_2 \right) = \Delta_1 w$$

Despejando a d, se tiene

$$d = w\left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right)$$

Finalmente, el valor de la moda, obtenido por interpolación es

$$x_{mo} = front\ inf_{(moda)} + d$$

Mediana: Es aquel valor que divide en dos partes iguales la distribución de frecuencia. Se obtiene mediante una interpolación en la ojiva.

Cálculo de la mediana para datos agrupados en una tabla de distribución de frecuencias:

1) Identificar la clase en la que se alcanza el 50 % de los datos. Esta clase recibe el nombre de clase mediana.

2) Trazar la gráfica de la porción de la ojiva correspondiente a la clase mediana.

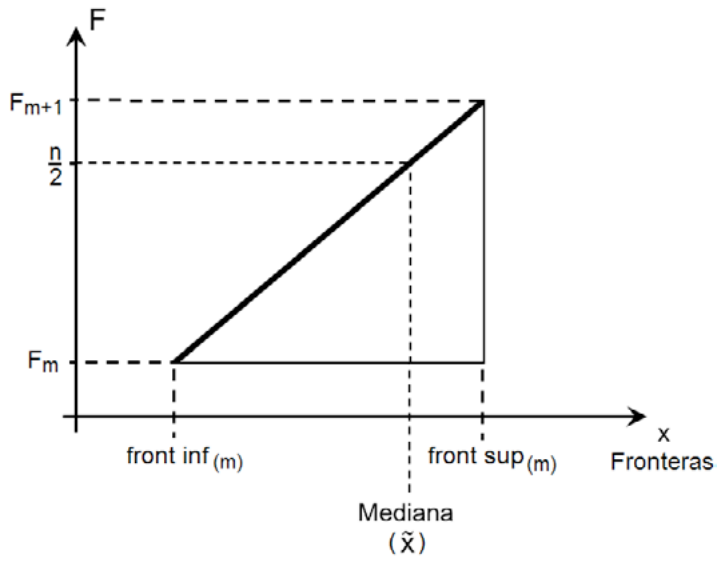


Figura 19. Identificación de la posición de la mediana

Interpolando en la ojiva:

$$\frac{\tilde{x} - front\ inf_{(m)}}{front\ inf_{(m)}} = \frac{\frac{n}{2} - F_m}{F_{m+1} - F_m}$$

$$\tilde{x} = \frac{\left(\frac{n}{2} - F_m\right)\left(front\ sup_{(m)} - front\ inf_{(m)}\right)}{F_{m+1} - F_m} + front\ inf_{(m)}$$

en donde:

$front\ inf_{(m)}$: Frontera inferior de la clase mediana
$front\ sup_{(m)}$: Frontera superior de la clase mediana

F_m : Frecuencia acumulada hasta $front\ inf_{(m)}$
F_{m+1} : Frecuencia acumulada hasta $front\ sup_{(m)}$ de la clase anterior a la mediana
n : Tamaño de la muestra
m : Clase mediana

Simplificando la expresión anterior, se tiene:

$$x = front\ inf_{(m)} + w\left(\frac{\frac{n}{2} - F_{m-1}}{f_m}\right)$$

Donde:

$front\ inf_{(m)}$ = Frontera inferior del intervalo de clase donde está la mediana
F_{m-1}: Frecuencia acumulada de la clase mediana hasta frontera superior de la clase anterior a la mediana
f_m: Frecuencia del intervalo de clase donde está la mediana

Ejemplo 3.24. Sea la siguiente tabla de distribución de frecuencia sobre cuotas anuales (dólares) que cobran 40 compañías de un seguro de 25000 dólares para personas de 25 a 35 años, en el mes de enero de 2019.

a) Elaborar el histograma de frecuencia relativa
b) Obtener la media
c) Obtener la mediana
d) Obtener la moda

Tabla 31. Frecuencias para cuotas anuales de un seguro de 25 000 dólares

i	Límites de clase Cuota anual (dólares)	Fronteras (dólares)	Marca de clase x_i	f_i	F_i	f_i^*	F_i^*
1	82 – 86	81.5 – 86.5	84	3	3	0.075	0.075
2	87 – 91	86.5 – 91.5	89	7	10	0.175	0.250
3	92 – 96	91.5 – 96.5	94	8	18	0.200	0.450
4	97 – 101	96.5 – 101.5	99	8	26	0.200	0.650
5	102 – 106	101.5 – 106.5	104	7	33	0.175	0.825
6	107 – 111	106.5 – 111.5	109	7	40	0.175	1.0
Sumas				40			

Solución:

a) Histograma de frecuencia relativa:

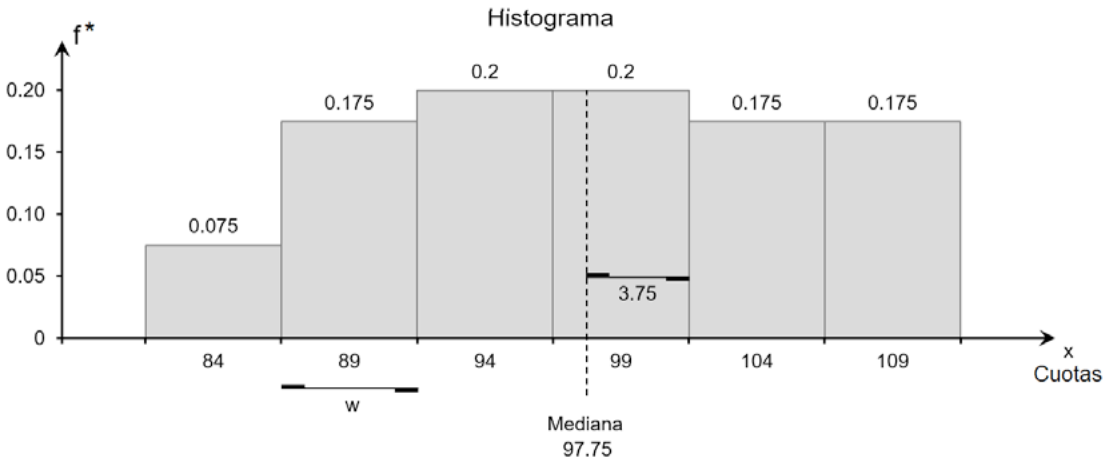


Figura 20. Histograma de cuotas y frecuencias relativas

b) Media:

$$\bar{x} = \frac{\sum_{i=1}^{c} f_i \, x_i}{n} = \frac{3910}{40} = 97.75 \text{ dólares}$$

c) Mediana:

$$\tilde{x} = front\ inf_{(m)} + w \left(\frac{\frac{n}{2} - F_{m-1}}{f_m} \right)$$

En la tabla 31 se observa que la mediana está comprendida entre las fronteras de la clase 4 debido a que $F_3^* = 0.45 < 0.5$ y $F_4^* = 0.65 > 0.5$

$$f_4 = 8$$

$$front\ inf_{(m)} = 96.5$$

$$w = 5$$

$$\tilde{x} = 96.5 + 5 \left(\frac{\frac{40}{2} - 18}{8} \right) = 96.5 + 5 \left(\frac{2}{8} \right) = 97.75 \text{ dólares}$$

Comprobación: Como se observa en el histograma de frecuencia relativa existe un equilibrio de las áreas que integran el histograma a partir del valor de la mediana; el valor del área a la izquierda de la mediana es igual al valor del área a la derecha de ésta.

$$A_1 + A_2 + A_3 + A_4 = A_5 + A_6 + A_7$$

$$A_1 + A_2 + A_3 + A_4 = 0.075\,(5) + 0.175\,(5) + (0.2)\,(5) + (0.200)\,(1.25) = 2.5$$

$$A_5 + A_6 + A_7 = 3.75\,(0.200) + (5)\,(0.175) + (5)\,(0.175) = 2.5$$

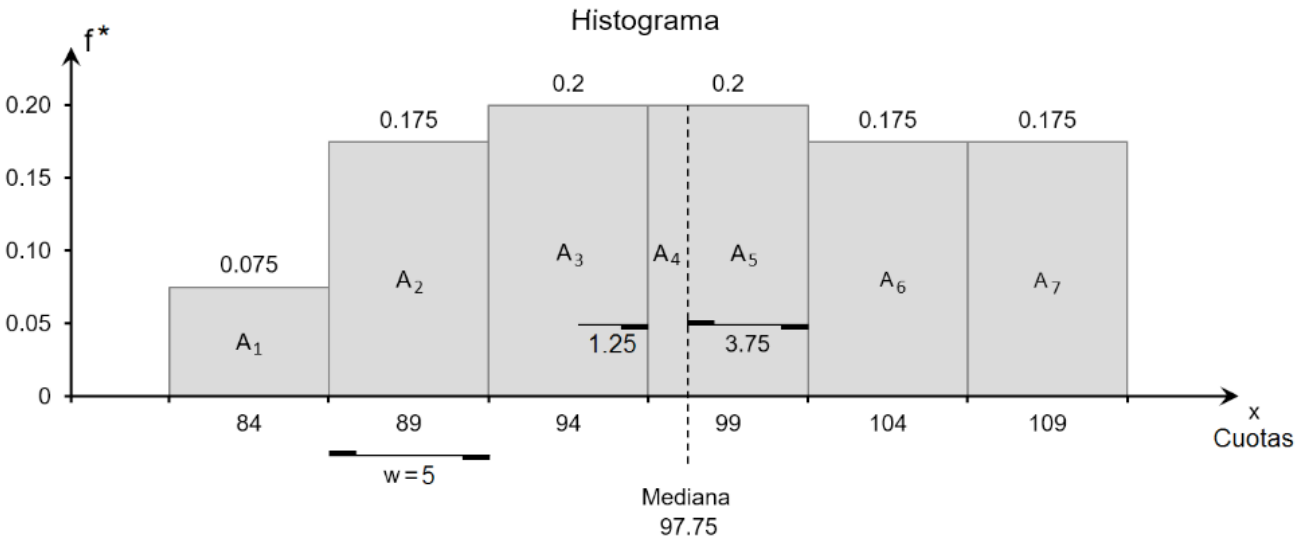


Figura 21. Histograma que muestra la mediana

d) Moda. En este caso, se tienen dos valores de cuotas que tienen mayor frecuencia: 94 y 99, además de que se tienen los valores contiguos Δ_1 y Δ_2, iguales, se considera x_{moda} a ambos valores de las marcas de clase 3 y 4. Si se toma el criterio de la moda cruda, se observa que hay dos modas a ambos valores de las marcas de clase

$$x_{m_1} = 94 \text{ (marca de la clase 3)}$$

$$x_{m_2} = 99 \text{ (marca de la clase 4)}$$

Ejemplo 3.25. Según información reportada por el INEGI sobre características de los hogares que tienen a una mujer como cabeza de familia, sin incluir la Ciudad de México, se tiene por entidad:

Tabla 32. Hogares con mujeres cabeza de familia en México. Fuente INEGI

Entidad federativa	Hogares con jefatura femenina	Entidad federativa	Hogares con jefatura femenina
01 Aguascalientes	87,578	18 Nayarit	94,011
02 Baja California	301,576	19 Nuevo León	329,031
03 Baja California Sur	58,071	20 Oaxaca	307,919
04 Campeche	71,485	21 Puebla	447,681
05 Coahuila de Zaragoza	194,562	22 Querétaro	151,732
06 Colima	60,061	23 Quintana Roo	119,482
07 Chiapas	300,561	24 San Luis Potosí	291,193
08 Chihuahua	309,570	25 Sinaloa	254,560
09 Ciudad de México	929,120	26 Sonora	258,562
10 Durango	130,995	27 Tabasco	196,896
11 Guanajuato	403,769	28 Tamaulipas	279,700
12 Guerrero	293,086	29 Tlaxcala	82,600
13 Hidalgo	218,866	30 Veracruz de Ignacio de la Llave	692,882
14 Jalisco	579,707	31 Yucatán	161,507
16 Michoacán de Ocampo	332,433	32 Zacatecas	100,031
17 Morelos	168,716	15 CDMX	1,158,268

Tabla 33. Frecuencias para hogares con mujeres cabeza de familia incluyendo la CDMX

Intervalo	Frontera inferior	Frontera superior	x_i	f_i	F_i	f_i^*	F_i^*
1	58070.5	241437.5	149754	15	15	0.47	0.47
2	241437.5	424804.5	333121	12	27	0.38	0.84
3	424804.5	608171.5	516488	2	29	0.06	0.91
4	608171.5	791538.5	699855	1	30	0.03	0.94
5	791538.5	974905.5	883222	1	31	0.03	0.97
6	974905.5	1158272.5	1066589	1	32	0.03	1.00

Solución:

Para determinar la mediana, se identifica en la ojiva la posición de ésta. Lo cual se puede corroborar en la última columna de la tabla 33, en la que se observa que es en la clase 2.

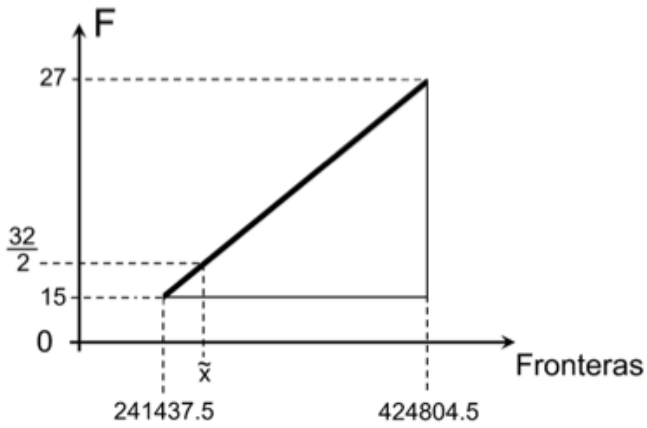


Figura 22. Gráfica del segmento de ojiva para el cálculo de la mediana

Por semejanza de triángulos:

$$\frac{424804.5 - 241437.5}{27 - 15} : \frac{\tilde{x} - 241437.5}{\frac{32}{2} - 15}$$

$$\left(\frac{32}{2}-15\right)(424804.5-241437.5)=(x-241437.5)(27-15)$$

$$(\tilde{x}-241437.5)(12)=(1)(183367)$$

$$(\tilde{x}-241437.5)=\left(\frac{1}{12}\right)(183367)$$

$$\tilde{x}=241437.5+15280.6$$

$$\tilde{x}=256718.1$$

Otra forma:

Con la expresión matemática de la página 85

$$\tilde{x}=front\ inf_{(m)}+w\left(\frac{\frac{n}{2}-F_{m-1}}{f_m}\right)$$

$$\tilde{x}=241437.5+183367\left(\frac{\frac{32}{2}-15}{12}\right)=256718.1$$

La tabla correspondiente en donde no se contempla a la Ciudad de México:

MIN=58071 MAX=929120 RANGO=871049 C=6 w=145175

Intervalo	Frontera inferior	Frontera superior	x_i	f_i	F_i	f_i*	F_i*
1	58070.5	203245.5	130658	14	14	0.45	0.45
2	203245.5	348420.5	275833	12	26	0.39	0.84
3	348420.5	493595.5	421008	2	28	0.06	0.90
4	493595.5	638770.5	566183	1	29	0.03	0.94
5	638770.5	783945.5	711358	1	30	0.03	0.97
6	783945.5	929120.5	856533	1	31	0.03	1

Por semejanza de triángulos:

$$\frac{348420.5 - 203245.5}{26 - 14} : \frac{\tilde{x} - 203245.5}{\frac{31}{2} - 14}$$

$$\left(\frac{31}{2} - 14\right)(348420.5 - 203245.5) = (\tilde{x} - 203245.5)(26 - 14)$$

$$(\tilde{x} - 203245.5)(12) = (0.5)(145175)$$

$$(\tilde{x} - 203245.5) = \left(\frac{1}{12}\right)(72587.5)$$

$$\tilde{x} = 203245.5 + 6048.96$$

$$\tilde{x} = 209294.45$$

Al comparar las medias de las muestras con la CDMX y sin la CDMX, resulta:

Media $\bar{x}$		Mediana $\tilde{x}$	
CON CDMX	SIN CDMX	CON CDMX	SIN CDMX
307449.62	355653.67	256718.1	209294.45

Se puede observar que la diferencia entre los valores de la mediana es pequeña en comparación con los de la media.

Se propone al alumno calcular la media y la mediana para la Ciudad de México con el dato de 1 158 268 hogares.

Comparando los resultados obtenidos, ¿Por qué considera que la diferencia de valores en las medianas es pequeña y en las medias es grande?

3.2.2 MEDIDAS DE DISPERSIÓN DE DATOS AGRUPADOS

Como los datos agrupados se encuentran en intervalos de clase desplegados en una tabla de distribución de frecuencias, para encontrar la suma de cuadrados: $SS = (x_i - \bar{x})^2$ que tienen medidas repetidas, determinamos primero la frecuencia de cada medida y utilizamos la marca de clase x_i correspondiente a cada intervalo de clase.

VARIANZA Y DESVIACIÓN ESTANDAR DE DATOS AGRUPADOS

Varianza o variancia: La varianza de una muestra se define como:

$$s^2_{x\,(n-1)} = \frac{SS}{n-1} = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 f_i}{n-1}$$

en donde n es el número de datos de la muestra.

También se puede definir a la varianza de la población como

$$s^2_{x\,(n)} = \frac{SS}{n} = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 f_i}{n}$$

n = número de datos de la población

En la teoría de la ***inferencia estadística*** se demuestra que los valores de $s^2_{x\,(n-1)}$ son más cercanos al valor de la varianza de la población que aquellos obtenidos mediante el uso de la expresión $s^2_{x(n)}$. Por lo que se prefiere usar la primera expresión como medida de la dispersión de los datos de la muestra y aproximación a la varianza poblacional.

Desviación estándar: Se define como la raíz cuadrada positiva de la varianza, se denota por:

$$s_{x(n-1)} = \sqrt{\frac{SS}{n}} = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 f_i}{n-1}}$$ para la muestra

$$s_{x(n)} = \sqrt{\frac{SS}{n}} = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 f_i}{n}}$$ para la población

La desviación estándar se expresa en las mismas unidades que los datos y la expresión que se use para determinarla, responde a las circunstancias antes descritas.

Ejemplo 3.26. Los datos de la tabla representan las frecuencias de unidades vendidas por día de un determinado producto de una compañía de electrodomésticos en un periodo de 40 días. Obtener:

a) La media y la mediana de las unidades vendidas.
b) La varianza y la desviación estándar aproximada del conjunto de datos.
c) Recorrido intercuartil.

Tabla 34. Frecuencias de unidades vendidas por día

Clase i	Fronteras	Marca de clase x_i	Frecuencia de clase f_i	F_i
1	81.5 – 86.5	84	3	3
2	86.5 – 91.5	89	7	10
3	91.5 – 96.5	94	8	18
4	96.5 – 101.5	99	8	26
5	101.5 – 106.5	104	7	33
6	106.5 – 111.5	109	7	40
	SUMAS		40	

Solución:

Tabla 35. Distribución de frecuencias para el número de unidades vendidas del producto

Clase	Fronteras	Marca de clase x_i	Frec. de clase f_i	F_i	$f_i\,(x_i)$	$x_i-\overline{x}$	$(x_i-\overline{x})^2$	$f_i\,(x_i-\overline{x})^2$	f_i^*	F_i^*
1	81.5 – 86.5	84	3	3	252	-13.5	189.0625	567.1875	0.075	0.075
2	86.5 – 91.5	89	7	10	623	-8.75	76.5625	535.9375	0.175	0.250
3	91.5 – 96.5	94	8	18	752	-3.75	14.0625	112.5	0.2	0.45
4	96.5 – 101.5	99	8	26	792	1.25	1.5625	12.5	0.2	0.65
5	101.5 – 106.5	104	7	33	728	6.25	39.0625	273.4375	0.175	0.825
6	106.5 – 111.5	109	7	40	763	11.25	126.5625	885.9375	0.175	1
	SUMAS		40		3910	0		2387.5		

a.1) Media

$$\overline{x} = \frac{\sum_{i=1}^{c} f_i\, x_i}{\sum_{i=1}^{c} f_i} = \frac{3910}{40} = 97.75$$

a.2) Mediana

$i = 4$ es la clase donde está la mediana, se visualiza en la columna F_i^* de la tabla 35.

$$\tilde{x} = front\ inf_{(m)} + w\left(\frac{\frac{n}{2} - F_{m-1}}{f_m}\right) = 96.5 + 5\left(\frac{2}{8}\right) = 97.75$$

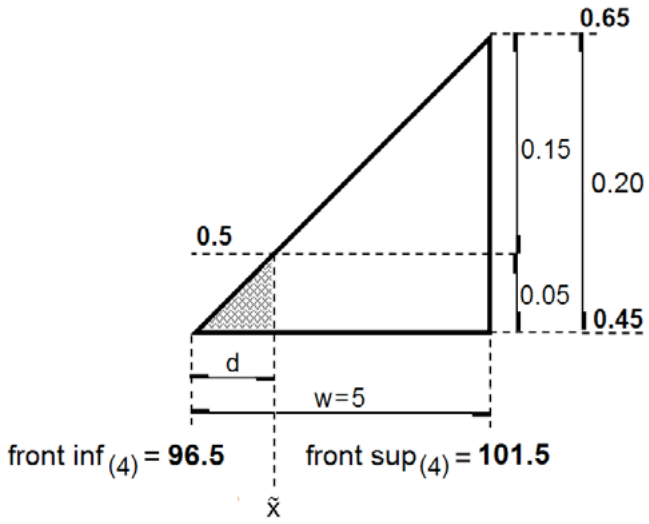


Figura 23. Semejanza de triángulos para el cálculo de la mediana

b) Para calcular la varianza

$$S^2_{x\,(n-1)} = \frac{\sum_{i=1}^{c} f_i\left(x_i - \bar{x}\right)^2}{n-1} = \frac{2387.5}{39} = 61.2179 \left(\frac{\text{unidades}}{\text{día}}\right)^2$$

c) Para la desviación estándar

$$S_{x\,(n-1)} = 7.8241 \ \frac{\text{unidades}}{\text{día}}$$

d) Recorrido intercuartil. Para determinar el valor del cuartil 3(x_{q3}) se realiza una interpolación lineal entre las fronteras 101.5 y 106.5 donde se ubica éste, como se puede apreciar en la ojiva de frecuencia relativa acumulada que se muestra en la gráfica siguiente (figura 23). Asimismo, el triángulo que se forma con las mencionadas fronteras y los niveles de acumulación porcentual de los datos, en cada uno de ellos, se reproduce de mayor tamaño a fin de contar con las diversas relaciones en los puntos A, B, y C entre las fronteras y los niveles de acumulación (figuras 24 y 25) y así, por triángulos semejantes, puede determinarse el valor de "d" que es la distancia que existe entre la frontera 106.5 y x_{q3} que corresponde al 0.75 de acumulación de datos del tercer cuartil. De manera semejante, se obtiene el

primer cuartil (x_{q1}) que debe coincidir con el 0.25 de datos acumulados. En la gráfica de la ojiva mencionada, se ubica el recorrido intercuartil (RIC). Los acumulados de datos 0.75 y 0.25 se pueden ubicar en el eje de las ordenadas, siguiendo líneas de referencias desde estos puntos hasta la ojiva y luego mediante líneas de referencia verticales, se puede observar la posición de los dos cuartiles y del recorrido intercuartil (*RIq*).

$$\text{Recorrido intercuartil} = RI_q = x_{q3} - x_{q1}$$

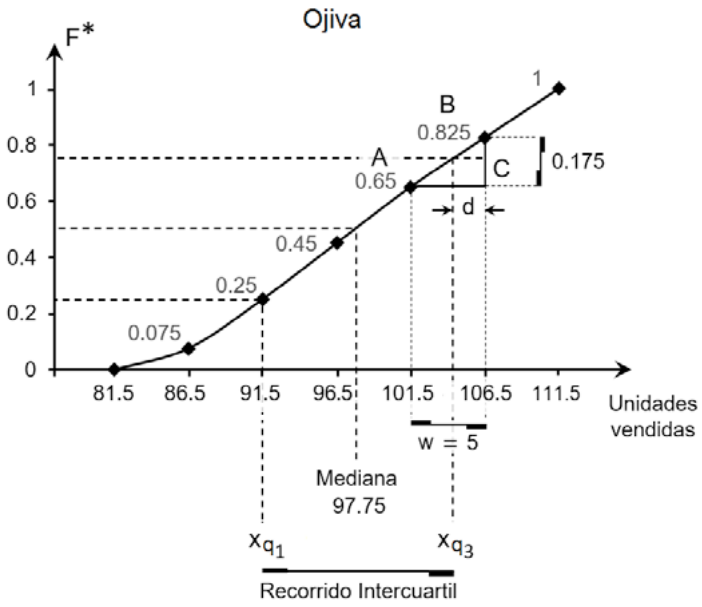


Figura 24. Recorrido intercuartil en la gráfica de distribución relativa acumulada

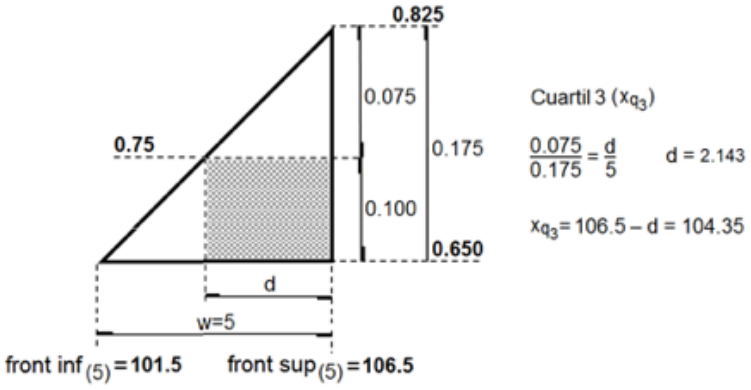


Figura 25. Semejanza de triángulos para el cálculo del tercer cuartil

De manera similar:

$$x_{q1} = 91.5$$

$$RIq = x_{q3} - x_{q1} = 104.35 - 91.5 = 12.85 \text{ unidades vendidas}$$

Percentiles o fractiles. El percentil $(1-\alpha) \times 100\ \%$ $(0 \leq \alpha \leq 1)$ de una muestra es el valor para el cual el $(1-\alpha) \times 100\ \%$ de los datos en ésta son menores o iguales a dicho valor. Se denota al percentil $(1-\alpha) \times 100\ \%$ de una muestra como $x_{(1-\alpha)}$. También se conoce a este percentil como fractil $(1-\alpha)$. Claramente, cuando $\alpha = 0.5$, el valor del percentil es justamente la mediana de la muestra.

Una forma metódica para calcular el percentil $(1-\alpha) \times 100\ \%$ es seguir los pasos que se indican a continuación:

1. Ordenar los datos en forma creciente.

2. Si el tamaño de la muestra es n, determinar la posición en que se encuentra el percentil, mediante la operación $(1-\alpha)\ (n+1)$.
   2.1 Si $(1-\alpha)\ (n+1)$ es entero, $x_{(1-\alpha)}$ será el dato que se encuentre en el lugar $(1-\alpha)\ (n+1)$ dentro de la ordenación.
   2.2 Si $(1-\alpha)\ (n+1)$ es fraccionario, se identifican los datos, adyacente inferior (a_I), aquel que se encuentra en la posición $[(1-\alpha)\ (n+1)]$, y adyacente superior (a_s), el dato que está en el lugar $[(1-\alpha)(n+1)]+1$ dentro de la ordenación. En donde, $[(1-\alpha)\ (n+1)]$ es la parte entera de $(1-\alpha)\ (n+1)$.

3. Determinar el valor del percentil como

$$x_{(1-\alpha)} = a_I + \textit{fracción de la posición}\ (a_s - a_I)$$

En donde *fracción de la posición* $= (1-\alpha)\ (n+1) - [(1-\alpha)\ (n+1)]$

De acuerdo con la forma de cálculo presentada, es claro que el percentil $x_{(1-\alpha)}$ puede ser un dato observado o no, dependiendo del percentil deseado y del tamaño de la muestra.

> Existen algunos casos de percentiles o fractiles que son más ampliamente usados en la práctica. Estos son lo cuartiles primero, segundo y tercero ($x_{0.25} = q_1$, $x_{0.50} = q_2$ y $x_{0.30} = q_3$), respectivamente, que dividen a la muestra en cuartas partes, o bien los deciles primero o inferior, segundo, tercero, etc. ($x_{0.10}$, $x_{0.20}$, $x_{0.30}$,..., $x_{0.90}$) los cuales dividen a la muestra en décimas partes.

3.3 MOMENTOS Y MEDIDAS DE FORMA

Los momentos en estadística son una herramienta para analizar más ampliamente la tendencia, la dispersión y la forma de la distribución de una muestra de datos. Sin embargo, es la misma herramienta definida en la Mecánica para investigar la tendencia al giro y el centroide, por lo tanto, la definición de los momentos de una muestra coincide con aquella de los momentos de masa, usados en Mecánica, en donde el momento de orden r con respecto a un punto "***a***" se determina como

$$\frac{\sum_{i=1}^{n} \left(x_i - a \right)^r m_i}{\sum_{i=1}^{n} m_i}$$

en donde x_i es la posición sobre el eje real, en el que se ubica la masa m_i.

Es claro en la expresión anterior, que el momento de masa de orden r es el promedio de las diferencias de los puntos x_i al punto ***a,*** elevadas a la r – ésima potencia, ponderadas por la masa existente en cada punto.

Si en esta expresión, se sustituyen las masas por las frecuencias de clase y los puntos x_i por las marcas de clase en una tabla de distribución de frecuencias, es claro que se pueden definir los momentos de una muestra con respecto a cualquier punto y de cualquier orden. Sin embargo, en estadística, los momentos que resultan de gran utilidad son aquellos con respecto al origen y los que se calculan con respecto a la media que, cabe decir, corresponden a los momentos con respecto al centroide.

Momento respecto al origen: Consideremos una muestra conformada por n datos, organizados en una tabla de distribución de frecuencias que consta de c intervalos de clase. Sean x_i y f_i, respectivamente, la marca y la frecuencia de la clase i ($i = 1, 2,...,c$). Se define el r – ésimo momento o momento de orden r, con respecto al origen como

$$m'_r = \frac{\sum_{i=1}^{c} x_i^r \, f_i}{n}$$

O, de forma equivalente

$$m'_r = \sum_{i=1}^{c} x_i^r \, f_i^*$$

Claramente, la media $\bar{x}$ de la muestra es el primer momento con respecto al origen. Es decir,

$$\bar{x} = m'_r = \frac{\sum_{i=1}^{c} x_i^r \, f_i}{n}$$

Otros momentos importantes son aquellos en los que el punto de referencia es la media $\bar{x}$, ya que algunos de ellos permiten medir la dispersión de los datos en la muestra y otros ayudan a describir la forma que tiene la distribución de la muestra, ya sea su forma sesgada o simétrica, o el grado de apuntamiento (curtosis) de la distribución.

Momento respecto a la media: Consideremos una muestra conformada por n datos, organizados en una tabla de distribución de frecuencias que consta de c intervalos de clase. Sean x_i y f_i, respectivamente, la marca y la frecuencia de la clase $i = (i = 1, 2, ..., c)$. Se define el r – ésimo momento o momento de orden r, con respecto a la media como

$$m_r = \frac{\sum_{i=1}^{c} \left(x_i - \bar{x}\right)^r \, f_i}{n}$$

O, de forma equivalente

$$m_r = \sum_{i=1}^{c} \left(x_i - \bar{x}\right)^r f_i^*$$

3.4 MEDIDAS DE FORMA

Las medidas de forma permiten describir el comportamiento de la distribución de datos de la muestra. En concreto, podemos estudiar las siguientes características de la curva en cuanto al sesgo y grado de apuntamiento de la gráfica:

> **Coeficiente de asimetría o sesgo:** Este coeficiente mide si la curva tiene una forma simétrica, es decir, si respecto al centro de ésta (centro de simetría) los segmentos de curva que quedan a la derecha e izquierda son similares. Se denota con a_3

Para medir el nivel de asimetría, se utiliza el llamado **coeficiente de asimetría de Fisher o sesgo,** que se define cuando trazamos una vertical paralela al eje de las ordenadas y sobre el valor de la media de una variable en el diagrama de barras o histograma, según sea discreta o continua, ésta se transforma en eje de simetría, si decimos que la distribución es simétrica entonces existe el mismo número de valores a la izquierda y a la derecha de la media. En caso contrario, dicha distribución será asimétrica, como se aprecia en la figura 26.

Se calcula mediante

$$a_3 = \frac{\frac{1}{n}\sum_{i=1}^{c}\left(x_i - \bar{x}\right)^3 f_i}{S_x^3}$$

Los resultados pueden ser los siguientes:

a_3	Interpretación
$a_3 = 0$	Se considera una distribución simétrica; existe la misma concentración de valores a la derecha y a la izquierda de la media, véase la figura 26a.
$a_3 > 0$	Distribución asimétrica positiva; existe mayor concentración de valores a la izquierda de la media que a su derecha, esto es, los valores de los datos se extienden más hacia la derecha, véase la figura 26b.
$a_3 < 0$	Distribución asimétrica negativa; existe mayor concentración de valores a la derecha de la media que a su izquierda, esto es, los valores de los datos se extienden más hacia la izquierda, véase la figura 26c.

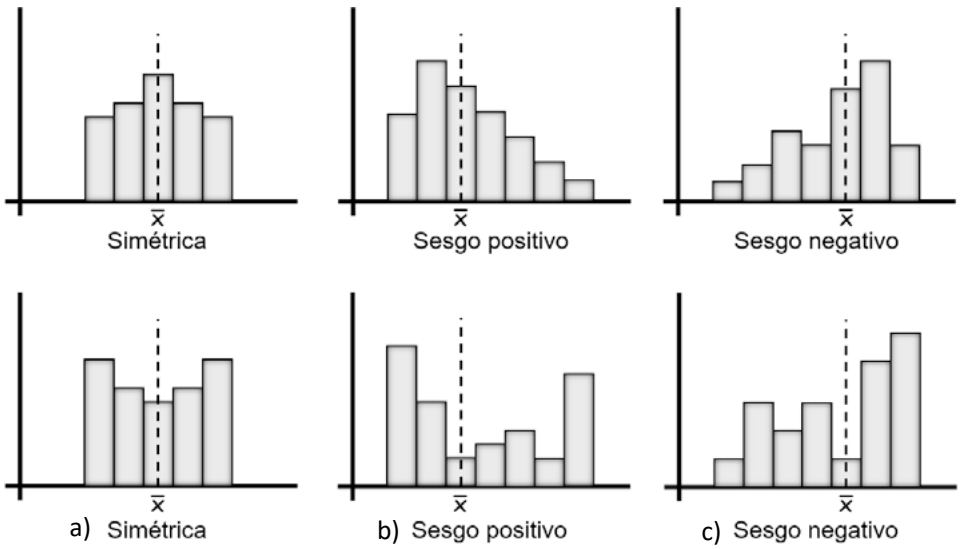


Figura 26. Gráficas de sesgo o simetría

Curtosis

El Coeficiente de Curtosis Como se definió en el caso de datos no agrupados, el coeficiente de curtosis es un parámetro de forma de la muestra que permite medir el grado de pronunciamiento de la punta más alta del polígono de frecuencias asociado a la distribución de los datos. Se denota al coeficiente de curtosis de una muestra por a_4 y se define como:

$$a_4 = \frac{m_4}{s^4} = \frac{\frac{1}{n}\sum_{i=1}^{c}\left(x_i - \bar{x}\right)^4 f_i}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{c}\left(x_i - \bar{x}\right)^2 f_i}\right)^4}$$

Donde m_4 es el cuarto momento respecto a la media.
De esta forma, si:

$a_4 < 3$	Se dice que la distribución es platicúrtica o achatada, véase la figura 27.
$a_4 > 3$	Se habla de que la distribución es leptocúrtica, es decir, puntiaguda, véase la figura 27.
$a_4 = 3$	Se concluye que la distribución de la muestra es mesocúrtica, véase la figura 27.

Se definen 3 tipos de distribuciones, según su grado de curtosis:

Distribución mesocúrtica: presenta un grado de concentración medio alrededor de los valores centrales de la variable (el mismo que presenta una distribución normal).

Distribución leptocúrtica: presenta un elevado grado de concentración alrededor de los valores centrales de la variable.

Distribución platicúrtica: presenta un reducido grado de concentración alrededor de los valores centrales de la variable.

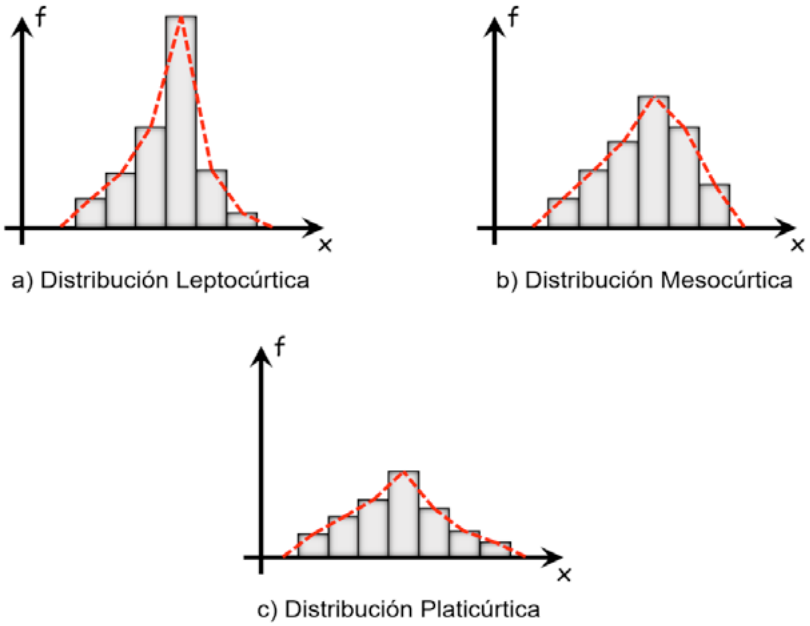


Figura 27. Gráficas de curtosis

Ejemplo 3.27. Sea la distribución de frecuencia de la estatura de los estudiantes de cierta universidad, en pulgadas:

Tabla 36. Frecuencia para la estatura de estudiantes

Alturas	Frecuencia (f_i)	Marcas de clase Alturas (x_i)
60 – 62	5	61
63 – 65	18	64
66 – 68	42	67
69 – 71	27	70
72 - 74	8	73
Sumas	100	

Obtener:

a) Media
b) Varianza
c) Coeficiente de sesgo. Proporcionar una interpretación
d) Curtosis. Proporcionar una interpretación

Solución:

Alturas i	Frecuencia (f_i)	Marcas de clase (x_i)	$(x_i-\bar{x})$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})^2 f_i$	$(x_i-\bar{x})^3 f_i$	$(x_i-\bar{x})^4 f_i$
60 – 62	5	61	-6.45	41.6035	208.0175	-1341.6806	8653.8400
63 – 65	18	64	-3.45	11.9025	214.2450	-739.1453	2550.0511
66 – 68	42	67	-0.45	0.2025	8.5050	-3.8273	1.7223
69 – 71	27	70	2.55	6.5025	175.5675	447.6971	1141.6277
72 - 74	8	73	5.55	30.8025	246.4200	1367.6310	7590.3521
Sumas	100				852.7500	-269.3249	19937.5929

a) Media:

$$m_1' = \bar{x} = \frac{\sum_{i=1}^{c} x_i f_i}{\sum_{i=1}^{c} f_i} = \frac{5(61)+(64)+42(67)+27(70)+8(73)}{100} = \frac{6745}{100} = 67.45$$

b) Varianza:

$$m_2 = S^2_{(n)} = \frac{\sum_{i=1}^{c} f_i \left(x_i - \bar{x} \right)^2}{n} = \frac{852.75}{100} = 8.5275$$

Otra forma de calcular la varianza es utilizando momentos con respecto al origen:

$$\text{Varianza} = m_2 = m'_2 - \left(m'_1 \right)^2$$

$$m'_2 = \frac{\sum_{i=1}^{c} f_i \, x_i^2}{n} = \frac{5\,(61)^2 + 18\,(64)^2 + 42\,(67)^2 + 27\,(70)^2 + 8\,(73)^2}{100} = 4558.03$$

$$S^2_n = 4558.03 - \left(67.45 \right)^2 = 8.5275$$

c) Coeficiente de sesgo:

El tercer momento con respecto a la media o tercer momento central es:

$$m_3 = \frac{\sum_{i=1}^{c} f_i \left(x_i - \bar{x} \right)^3}{n} = \frac{-269.3249}{100} = -26932$$

Por lo tanto, el coeficiente de sesgo es:

$$a_3 = \frac{m_3}{\left(m_2 \right)^{\frac{3}{2}}} = \frac{-2.6932}{\left(8.5275 \right)^{\frac{3}{2}}} = -01081 < 0$$

la gráfica es casi simétrica y tiene un pequeño sesgo a la izquierda.

d) Curtosis

El cuarto momento con respecto a la media es:

$$m_4 = \frac{\sum_{i=1}^{c} f_i \left(x_i - \bar{x} \right)^4}{n} = \frac{19937.5929}{100} = 199.3759$$

La curtosis resulta:

$$a_4 = \frac{m_4}{(m_2)^2} = \frac{199.3759}{(8.5275)^2} = 2.74175 < 3$$

la gráfica es casi mesocúrtica y ligeramente achatada.

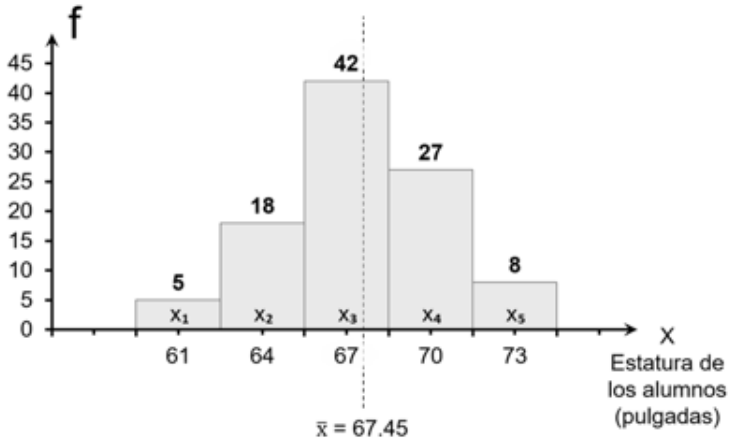


Figura 28. Histograma de distribución de estaturas de los estudiantes. Puede observarse un ligero sesgo a la izquierda y una curtosis cercana a una curva mesocúrtica

Ejemplo 3.28. Los datos de la tabla representan las frecuencias de unidades vendidas por día de un determinado producto de una compañía de electrodomésticos, en un periodo de 40 días. Obtener:

a) La media y la desviación estándar de las unidades vendidas del producto
b) Graficar la ojiva de frecuencia relativa acumulada
c) El recorrido interdecil
d) El recorrido intercuartil
e) La mediana

Tabla 37. Frecuencias de unidades vendidas por día

Clase i	Número de unidades vendidas	Frecuencia (días) f_i
1	82 – 86	3
2	87 – 91	7
3	92 – 96	8
4	97 – 101	8
5	102 – 106	7
6	107 – 111	7
sumas		40

Solución:

Tabla 38. Tabla completa de frecuencias de unidades vendidas por día

Clase	Número de unidades vendidas		Frecuencia de clase (días) f_i	Marcas de clase	Frecuencia relativa f_i^*	Frecuencia Acumulada F_i	Frecuencia Acumulada Relativa F^*	$x_i\ f_i$	$(x_i - \bar{x})$	$(xi - \bar{x})^2\ fi$
1	82	86	3	84	0.075	3	0.075	252	-13.75	567.19
2	87	91	7	89	0.175	10	0.25	623	-8.75	535.94
3	92	96	8	94	0.2	18	0.45	752	-3.75	112.50
4	97	101	8	99	0.2	26	0.65	792	1.25	12.50
5	102	106	7	104	0.175	33	0.825	728	6.25	273.44
6	107	111	7	109	0.175	40	1	763	11.25	885.94
Sumas			40					3910		2387.50

$$w = 5 \qquad \sum_{i=1}^{c} f_i = 40 \qquad UPMC = 1.0$$

a) Media:

$$\bar{x} = \frac{\sum_{i=1}^{c} x_i\ f_i}{\sum_{i=1}^{c} f_i} = \frac{3910}{40} = 97.75 \text{ dólares}$$

b) Varianza:

$$S^2_{x(n)} = \frac{\sum_{i=1}^{c} f_i \left(x_i - \bar{x}\right)^2}{n} = \frac{2387.50}{40} = 59.6875$$

$$S^2_{x(n-1)} = \frac{\sum_{i=1}^{c} f_i \left(x_i - \bar{x}\right)^2}{n-1} = \frac{2387.50}{39} = 61.2179$$

Desviación estándar:

$$S_{x(n)} = 7.7257$$

$$S_{x(n-1)} = 7.8241$$

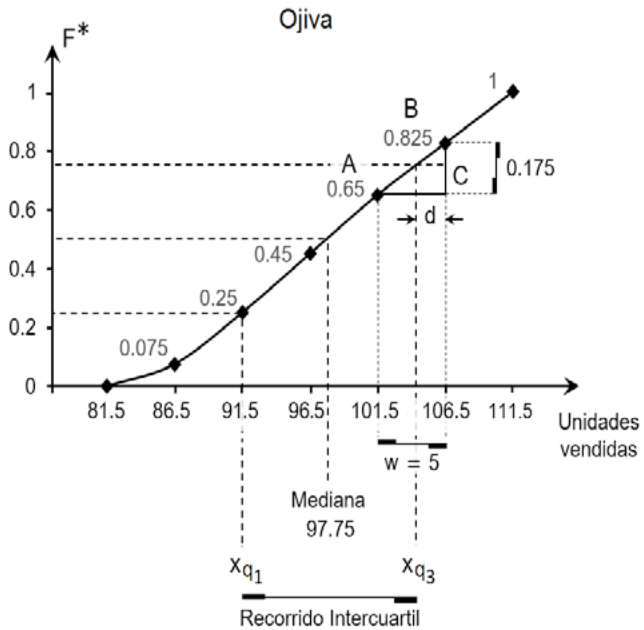


Figura 29. Gráfica de la ojiva

c) Recorrido interdecil

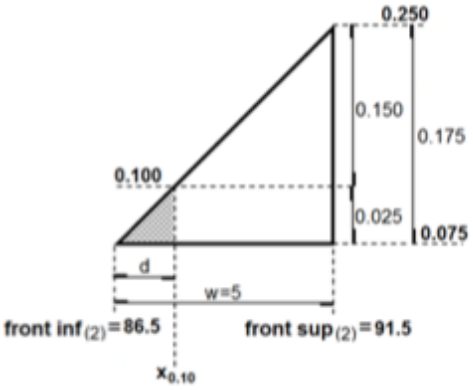


Figura 30. Esquema para el primer decil

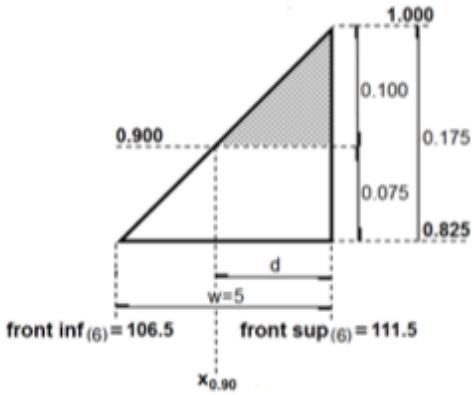


Figura 31. Esquema para el noveno decil

Decil 1: Con base en la tabla 38, se determina el primer decil en la figura 30, y en la figura 31, el noveno decil.

$$\frac{0.175}{0.025} = \frac{5.0}{d} \qquad d = 0.7142$$

$$x_{0.10} = front\ inf_{(2)} + d$$
$$x_{0.10} = 86.5 + d = 87.2142$$

Decil 9: Se encuentra en la clase 6, de acuerdo con tabla 38.

$$\frac{0.175}{0.100} = \frac{5.0}{d} \quad ; \quad d = \frac{5.0\,(0.100)}{0.175} = 2.8571$$

$$x_{0.90} = 111.5 - 2.8571 = 108.6429$$

$$RId = x_{0.90} - x_{0.10} = 108.6428 - 87.2142 = 21.4286$$

d) Recorrido intercuartil (*RIq*):

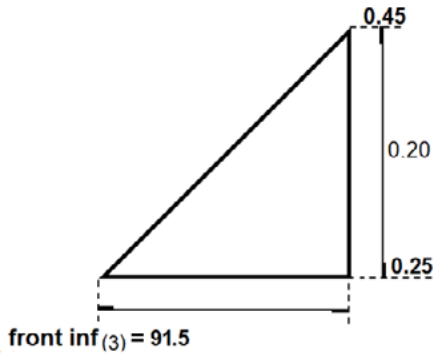


Figura 32. Esquema para el primer cuartil

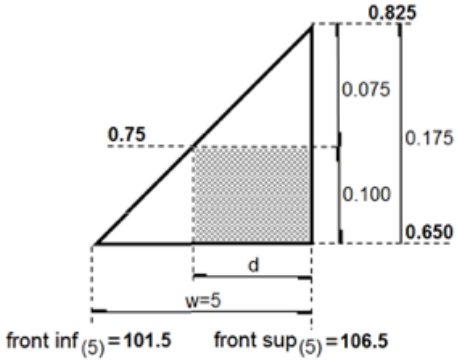


Figura 33. Esquema para el tercer cuartil

Cuartil 1: Con base en la tabla 38, se muestra el primer cuartil en la figura 32 y en la figura 33, el tercer cuartil.

$$x_{0.25} = 91.5$$

Cuartil 3: Revisando la tabla 38, está ubicado en el intervalo de clase 5 de la figura 33

$$\frac{0.175}{0.100} = \frac{5.0}{d} \quad ; \quad d = \frac{5.0\,(0.100)}{0.175} = 2.8571$$

$$x_{0.75} = 101.5 + 2.8571 = 104.357$$

$$RIq = x_{0.75} - x_{0.25} = 104.357 - 91.5 = 12.857$$

e) Mediana: La clase mediana (m) se determina en la clase 4, con base en la tabla 38.

$$\tilde{x} = front\ inf_{(m)} = w\left(\frac{k}{f_m}\right) = 96.5 + 5.0\left(\frac{k}{8}\right)$$

$$k = \frac{n}{2} - F_{(m-1)} = \frac{40}{2} - 18 = 2.0$$

$$\tilde{x} = 96.5 + 5.0\left(\frac{2}{8}\right) = 97.75$$

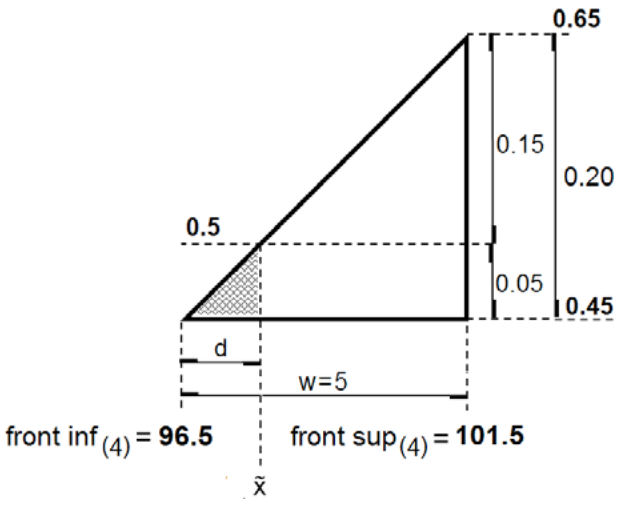


Figura 34. Esquema para el cálculo de la mediana

Otro método, interpolación:

$$\frac{0.200}{0.050} = \frac{5}{d} \qquad d = \frac{5\,(0.050)}{0.200} = 1.25 \qquad \rightarrow \qquad x = 96.5 + 1.25 = 97.75$$

Como puede observarse en este caso, coinciden la media y la mediana.

3.5 EJERCICIOS RESUELTOS No. 2

1. Las medidas siguientes representan los días que tarda el correo para llegar y entregar los envíos de una determinada zona del norte del país, enviado desde la Ciudad de México. Los últimos diez envíos tardaron: 2, 2, 2, 3, 3, 4, 4, 5, 5 y 10 días.

Determine:

a) La media, la media geométrica y la mediana
b) La suma de los cuadrados de las desviaciones
c) La desviación estándar

Solución:

Con el objetivo de simplificar los cálculos, es posible organizar los datos de la muestra en una tabla como la que se muestra a continuación

Dato (x_i)	Frecuencia (f_i)
2	3
3	2
4	2
5	2
10	1

De donde:

a) La media

$$\tilde{x} = \frac{\sum_{i=1}^{n} f_i \, x_i}{n}$$

$$\bar{x} = \frac{2(3) + 3(2) + 4(2) + 5(2) + 10(1)}{10} = 4$$

La media geométrica para datos agrupados se calcula mediante:

$$\log G = \frac{\sum_{i=1}^{n} f_i \log\left(x_i\right)}{\sum_{i=1}^{n} f_i}$$

utilizando la tabla, se obtiene la expresión:

$$\log G = \frac{(3 \log(2))(2 \log(3))(2 \log(4))(2 \log(5))(1 \log(10))}{10} = 0.5459$$

Al calcular el antilogaritmo considerando la base 10

$$G = \textit{anti} \log (0.5459) = 3.515$$

En Excel se puede calcular el antilogaritmo mediante la expresión

$$= \text{POTENCIA}(10, \textit{valor})$$

b) Para la suma de los cuadrados de las desviaciones, en este caso se está en presencia de una distribución de frecuencias de datos no agrupados x = número de días que tarda el correo en llegar a la Zona Norte desde la Ciudad de México.

Tabla 39. Frecuencias con datos repetidos no agrupados

x	$x-\bar{x}$	$(x-\bar{x})^2$
2	-2	4
2	-2	4
2	-2	4
3	-1	1
3	-1	1
4	0	0
4	0	0
5	1	1
5	1	1
10	6	36
$\sum(x-\bar{x})=0$		

Tabla 40. Tabla de frecuencias para datos repetidos no agrupados

x_i	f_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$f_i(x_i-\bar{x})^2$	$f_i(x_i-\bar{x})$
2	3	-2	4	12	-6
3	2	-1	1	2	-2
4	2	0	0	0	0
5	2	1	1	2	2
10	1	6	36	36	6
Sumas			42	52	0

Por lo tanto, la suma de los cuadrados de las desviaciones se tiene en la quinta columna:

$$SS = \sum_{i=1}^{c} f_i \left(x_i - \bar{x}\right)^2 = 3\left(2-4\right)^2 + 2\left(3-4\right)^2 + 2\left(4-4\right)^2 + 2\left(5-4\right)^2 + 1\left(10-4\right)^2$$

$$= 12 + 2 + 0 + 2 + 36$$

$$= 52$$

c) En datos agrupados, la varianza de una muestra puede calcularse de manera aproximada mediante la expresión:

$$s^2_{x\,(n-1)} = \frac{SS}{n-1} = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 f_i}{n-1}$$

Donde n: es el número de datos de la muestra.

x_i: valor real del dato
f_i: frecuencia, número de veces que se repite el valor del dato

También se puede definir a la varianza de la población como:

$$s^2_{x(n)} = \frac{SS}{n} = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 f_i}{n}$$

Por lo tanto, la desviación estándar es:

$$s_{x\,(n-1)} = \sqrt{\frac{52}{9}} = 2.4037$$

$$s_{x\,(n)} = \sqrt{\frac{52}{10}} = 2.2804$$

2. En el grupo de Probabilidad y Estadística hay 60 alumnos y de acuerdo con sus edades presentan la siguiente distribución de frecuencias:

Tabla 41. Frecuencias para alumnos de Probabilidad y Estadística

Intervalo de clase	Fronteras	Frecuencia absoluta	Frecuencia absoluta acumulada	Frecuencia relativa	Frecuencia relativa acumulada	
$(li - ls)$	$(Fi - Fs)$	(f)	F_i	(f_i^*)	(F_i^*)	
17-19	16.5 – 19.5	12	12	0.20	0.20	
20-22	19.5 – 22.5	25	37	0.4166	0.6166	$\leftarrow X_{P50}$
23-25	22.5 – 25.5	15	52	0.25	0.8666	$\leftarrow X_{d8}$
26-28	25.5 – 28.5	6	58	0.10	0.9666	
29-31	28.5 – 31.5	2	60	0.0333	1.0	
Sumas		60		1.0		

a) Elaborar el histograma
b) Trazar la ojiva
c) El diagrama de frecuencias relativas
d) Trazar la ojiva porcentual
e) Determinar el 8º decil
f) Calcular el cincuentavo percentil (2º cuartil, 5º décil o mediana).

Solución:

UPMC = 1.0 año, $w = 3$

a)

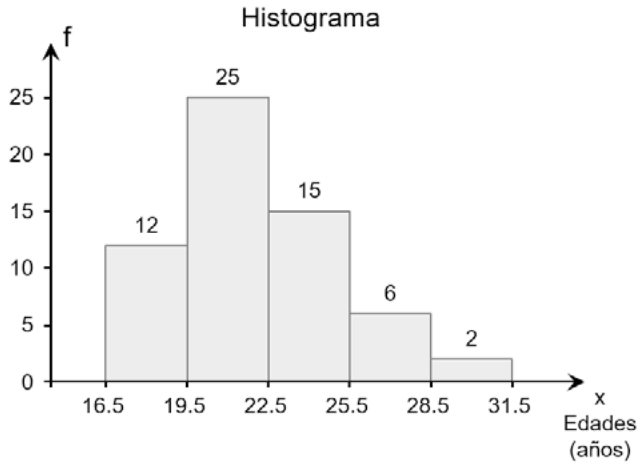


Figura 35. Histograma para la distribución de frecuencias de alumnos

b) La ojiva de frecuencias acumuladas se obtiene aplicando la siguiente expresión (los valores en cada intervalo de clase se muestran en la cuarta columna de la tabla):

$$F_k = \sum_{i=1}^{k} f_1 + f_2 + f_3 + \ldots + f_k$$

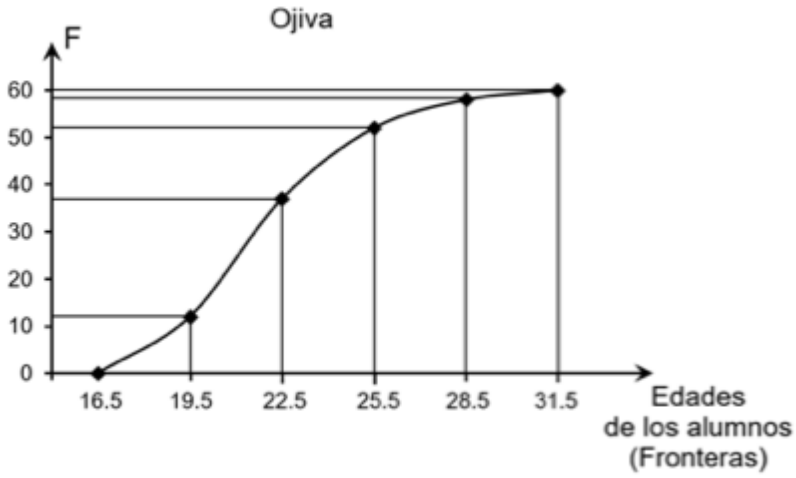


Figura 36. Ojiva para las edades de alumnos con frecuencias absolutas

c) El diagrama se forma con los valores de la 2ª columna y los de la 4ª columna de la tabla.

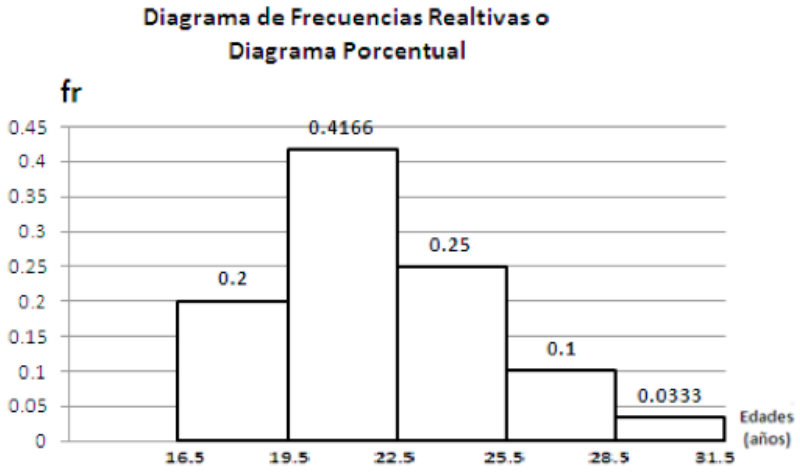


Figura 37. Histograma para las edades de alumnos

d) La ojiva se forma con los valores de la 2ª columna y los de la 6 ª columna de la tabla.

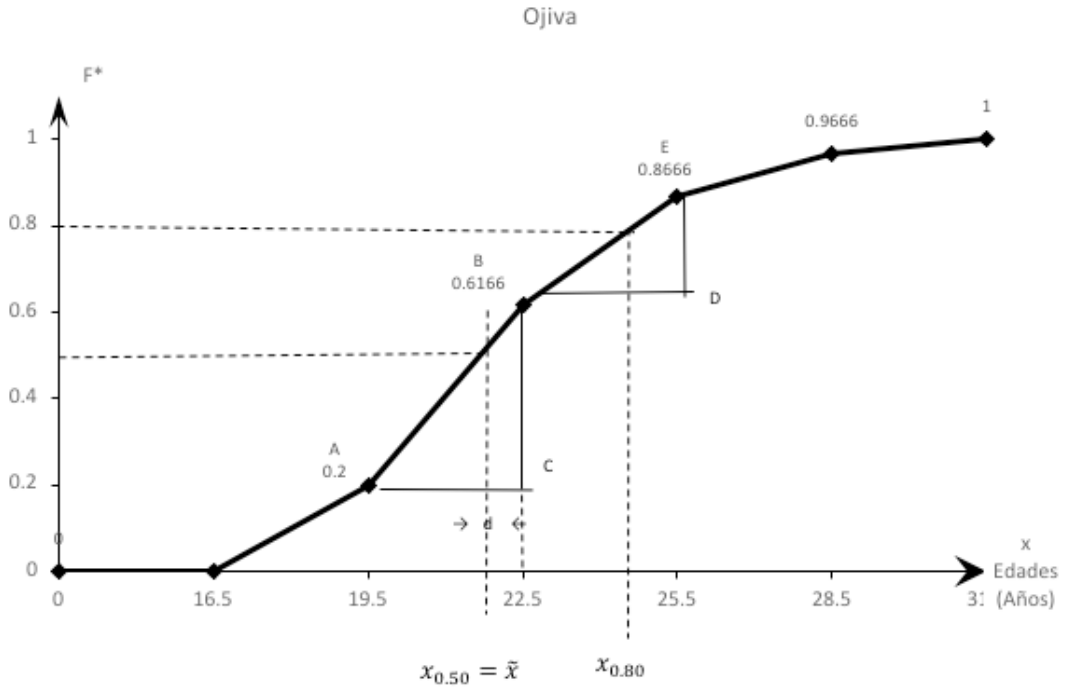


Figura 38. Ojiva para las edades de alumnos con frecuencia relativa acumulada

e) Determinar el 8° decil que se ubica en el intervalo de clase 3 y puede observarse ampliando el triángulo BDE que se forma con la ojiva porcentual.

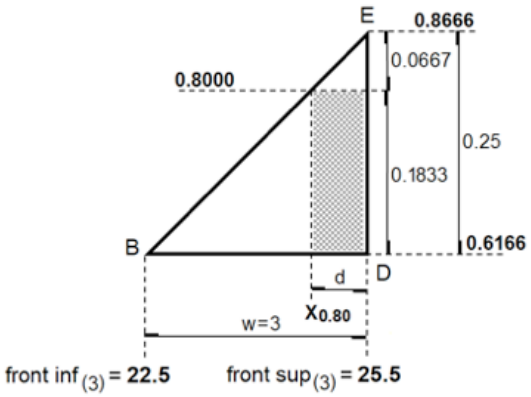


Figura 39. Diagrama para calcular el 8° decil

$$\frac{d}{3.0} = \frac{0.0667}{0.25} \qquad d = 0.8004$$

$$X_{0.80} = 25.5 - 0.8004 = 24.6996$$

f) Mediana. De manera similar, la mediana se localiza en el intervalo de clase 2.

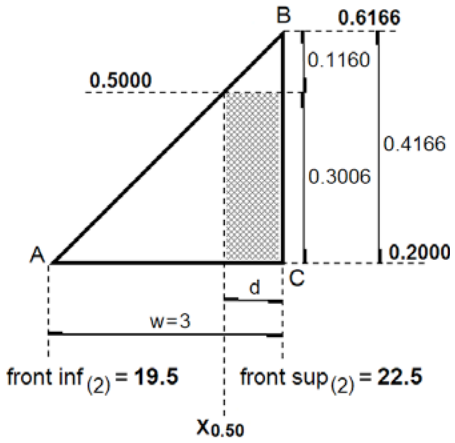


Figura 40. Diagrama para calcular la mediana, 2° cuartil o 50° percentil

$$\frac{d}{w} = \frac{0.1166}{0.4166} \qquad d = \frac{w\,(0.116)}{0.4166}$$

$$d = 0.8396$$

$$X_{0.50} = \tilde{x} = 22.5 - 0.8396 = 21.66$$

3. En una investigación por parte de la Secretaría del Medio Ambiente se enfocó el análisis de las partículas PM2.5 en la Zona Metropolitana del Valle de México por el daño que causan a la población, sin embargo, este contaminante no tiene norma de protección a la salud, por lo que futuros estudios serán de gran importancia para su regulación.

   Los datos fueron obtenidos con equipo TEOM en las estaciones de Azcapotzalco y Santa Úrsula de la RAMA, del 7 de diciembre de 1998 al 20 de mayo de 1999 y del 2 de febrero al 20 de mayo de 1999, respectivamente.

   Al comparar los resultados de ambas estaciones, se observó que Santa Úrsula registró los niveles máximos de partículas PM2.5 en el periodo de análisis, obtenidos en mayo; sin embargo, en Azcapotzalco se presentaron las concentraciones más altas en diciembre y decrecieron a lo largo del periodo. Los coeficientes de correlación más altos fueron con NOx (óxido de nitrógeno) y el NO_2 (dióxido de nitrógeno).

Las PM10 se definen como partículas con diámetro de masa aerodinámica media menor a 10 micrómetros (μm). Para poder imaginar el tamaño de una partícula PM10, basta con señalar que el diámetro promedio de un cabello es de 50μm. Las PM10 se subdividen en dos tipos, fino y grueso; el tipo fino define partículas con diámetro menor a 2.5μm (PM2.5), el tipo grueso se refiere a las partículas con diámetro entre 2.5 y 10μm1.

Las partículas gruesas tienen una composición de origen terrestre, mientras que la composición química de las partículas finas muestra una abundancia mayor de derivados de azufre y nitrógeno, así como un 20 % de material orgánico. Dichas partículas se forman a partir de la interacción química o física entre los contaminantes presentes en el aire.

Entre más pequeño sea el diámetro de la partícula más profundo puede penetrar en el pulmón y mayor será el impacto sobre la salud. En efecto, desde el punto de vista de las funciones pulmonares, las partículas con diámetro mayor a 10 μm son menos peligrosas, debido a su peso caen al suelo rápidamente, si llegan a ser inhaladas se detienen en la nariz y garganta siendo fácilmente eliminadas a través de la tos, el estornudo o bien del sistema digestivo.

Para normar la concentración permisible de este contaminante en México, la Secretaría de Salud hizo un estudio preliminar; se seleccionaron aleatoriamente 36 días de un periodo de 4 meses, para los cuales se encontraron las siguientes concentraciones diarias (en $\mu g/m^3$, promedio de 24 horas).

Tabla 42. Lectura de concentración de contaminantes

14	15	11	10	7	5
12	15	17	18	20	18
16	13	11	9	15	8
10	12	15	16	17	17
13	9	6	20	18	3
14	13	10	15	14	13

a) Obtener la tabla de distribución de frecuencias, dividiendo los datos de la mejor manera. Calcular la media y la desviación estándar para los datos agrupados.

b) Dibujar el histograma y polígono de frecuencias.

Solución:

a) Para construir la tabla de distribución de frecuencias es necesario determinar el número de intervalos en los que se basará la construcción. Si se cuenta con 36 datos, entonces el número de intervalos será $c = \sqrt{36} = 6$

Se necesita saber qué tamaño tendrán los intervalos, para esto se necesita conocer el rango que es la diferencia entre el valor máximo y el valor mínimo dados en la tabla:

$$\text{rango} = 20 - 3 = 17$$

El tamaño del intervalo es: $w = \frac{rango}{c} = \frac{17}{6} \approx 3$

La tabla de frecuencias es:

Tabla 43. Frecuencias para la lectura de concentración de contaminantes

	Intervalo de clase							
Clase	lím inf	lím sup	front int	front sup	Frecuencia f_i	Marca de clase x_i	$x_i \quad f_i$	$f_i(x_i - \overline{x})^2$
1	3	5	2.5	5.5	2	4	8	165.0139
2	6	8	5.5	8.5	3	7	21	111.0208
3	9	11	8.5	11.5	7	10	70	66.5486
4	12	14	11.5	14.5	9	13	117	0.625
5	15	17	14.5	17.5	10	16	160	85.0694
6	18	20	17.5	20.5	5	19	95	175.0347

Para calcular la media, la varianza y la desviación estándar a partir de la tabla de frecuencias con los datos agrupados, se utilizan las expresiones:

$$\bar{x} = \frac{\sum_{i=1}^{c} x_i f_i}{n} = 13.08$$

$$S^2_{(n)} = \frac{\sum_{i=1}^{c} \left(x_i - \bar{x} \right)^2 f_i}{n} = 16.7431$$

$$S_{(n)} = \sqrt{\frac{\sum_{i=1}^{c} f_i \left(x_i - \bar{x} \right)^2}{n}} = \sqrt{16.7431} = 4.0918$$

b)

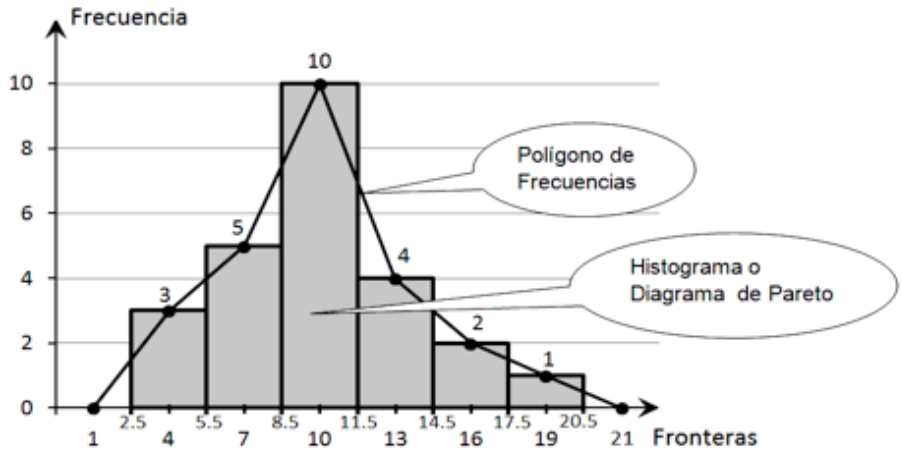


Figura 41. Histograma para la distribución de contaminantes

4. Una compañía de entregas foráneas determinó la cantidad de kilómetros por litro asociadas a 25 viajes realizados por sus camiones. En la tabla que se muestra a continuación, se registraron los siguientes datos:

Tabla 44. Cantidad de kilómetros recorridos y la cantidad de entregas

km por litro	5 - 7	8 – 10	11 – 13	14 – 16	17 – 19	20 – 22
No. de entregas	3	5	10	4	2	1

a) Elaborar una tabla de frecuencia relativa acumulada y trazar su gráfica.
b) Determinar la probabilidad de que en un viaje se tenga un kilometraje por litro menor a 13.
c) Calcular el promedio de kilómetros por litro.

Solución:

a) Se tienen 6 intervalos de valores en la tabla, los valores de los kilómetros por litro corresponden a los valores de los intervalos inferior y superior, el número de entregas corresponde a la frecuencia, cuya suma dará el total de entregas.

Para realizar la gráfica es necesario conocer el valor de la marca de clase que se obtienen mediante:

$$x_i = \frac{\text{front inf}_{(i)} + \text{front sup}_{(i)}}{2}$$

Con $i = 1$ $\quad x_i = \frac{5+7}{2} = 6$

La frecuencia relativa se obtiene con $f_i^* = \frac{f_i}{n}$

Tabla 45. Frecuencias relativas para la cantidad de kilómetros recorridos y la cantidad de entregas

	km por litro		Número de entregas f_i	Marca de clase x_i	Frecuencia relativa f_i^*	Frecuencia acumulada	Frecuencia relativa acumulada F_i^*
Intervalo	inf	sup					
1	5	7	3	6	0.12	3	0.12
2	8	10	5	9	0.2	8	0.32
3	11	13	10	12	0.4	18	0.72
4	14	16	4	15	0.16	22	0.88
5	17	19	2	18	0.08	24	0.96
6	20	22	1	21	0.04	25	1
	TOTAL		25				

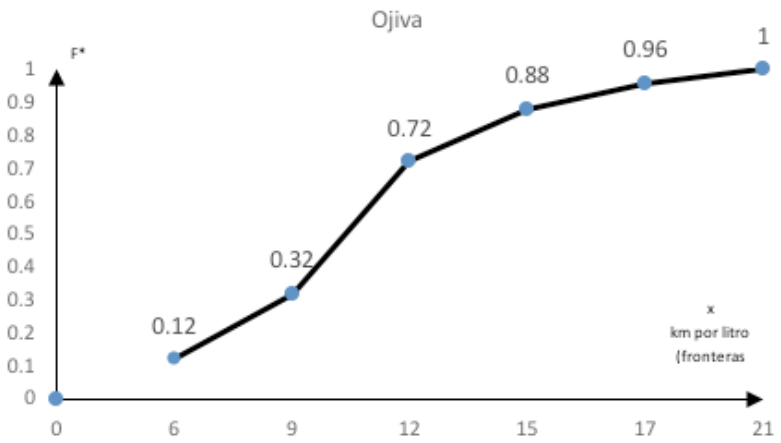


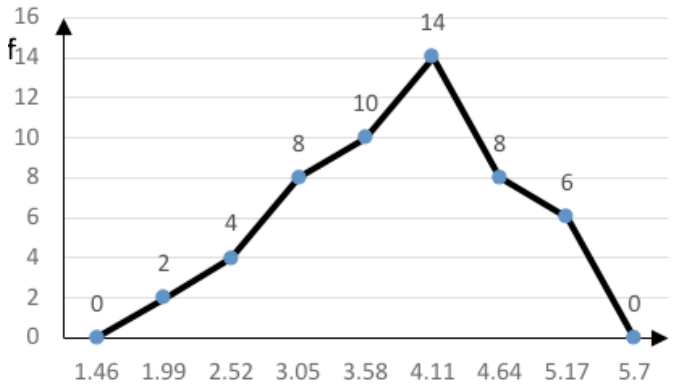
Figura 42. Ojiva para los kilómetros recorridos

b) Para la probabilidad de que en un viaje se tenga un kilometraje por litro menor a 13, se observa en la tabla que el valor de 13 km por litro cae en el tercer intervalo con una probabilidad acumulada de 0.72, y que es la probabilidad buscada: $p = 0.75$.
c) El promedio de kilómetros por litro se obtiene con:

$$\bar{x} = \frac{\sum_{i=1}^{c} x_i f_i}{n} = \frac{\sum_{i=1}^{6} x_i f_i}{25} = 12$$

5. En la figura 43 se muestra el polígono de frecuencias de las resistencias obtenidas en una muestra de suelo.

a) Formar la tabla de frecuencia correspondiente y encontrar el valor de la media, la desviación estándar y el coeficiente de variación.
b) Dibujar el polígono de frecuencia relativa acumulada (ojiva porcentual).
c) Si el valor de diseño se define como el percentil veinte del conjunto de datos, encontrar ese valor de diseño.



Marcas extremas del polígono de frecuencias

$x_1 = 1.46$
$x_9 = 5.7$

Figura 43. Polígono de frecuencias para la muestra de suelo

Solución:

a) Se observa en la gráfica que los datos fluctúan en un intervalo de valores entre poco más de 1 y un poco menos de 6. Asimismo, se señala que hay 9 marcas de clase y que la primera es $x_1 = 1.46$.

También se observa que la frecuencia mínima es cero en los extremos y la máxima es 14 en el intervalo de clase 6, y que el tamaño de la muestra acumulada de las frecuencias es 52.

La amplitud del intervalo se obtiene restando los valores de las marcas de clase consecutivos $w = 5.75 - 5.18 = 0.53$, siendo todos los intervalos con la misma amplitud.

El valor de la primera frontera inferior se obtiene dividiendo la amplitud del intervalo en dos partes iguales y restándola a la marca de clase

$$f_1 = 1.46 - \left(\frac{0.53}{2}\right) = 1.20$$

Tabla 46. Frecuencias para la muestra de suelo

Intervalo	Fronteras		Marca de clase x_i	Frecuencia f_i	Frecuencia relativa f_i^*	Frecuencia acum F_i^*	Frecuencia relativa acum. F_i^*
	Inf	sup					
1	1.20	1.73	1.46	0	0	0	0
2	1.73	2.26	1.99	2	0.038	2	0.038
3	2.26	2.79	2.52	4	0.077	6	0.115
4	2.79	3.32	3.05	8	0.154	14	0.269
5	3.32	3.85	3.58	10	0.192	24	0.462
6	3.85	4.38	4.11	14	0.269	38	0.731
7	4.38	4.91	4.64	8	0.154	46	0.885
8	4.91	5.44	5.17	6	0.115	52	1
9	5.44	5.97	5.7	0	0	52	1
TOTAL				52			

b) Dibujar el polígono de frecuencia relativa acumulada (ojiva porcentual).

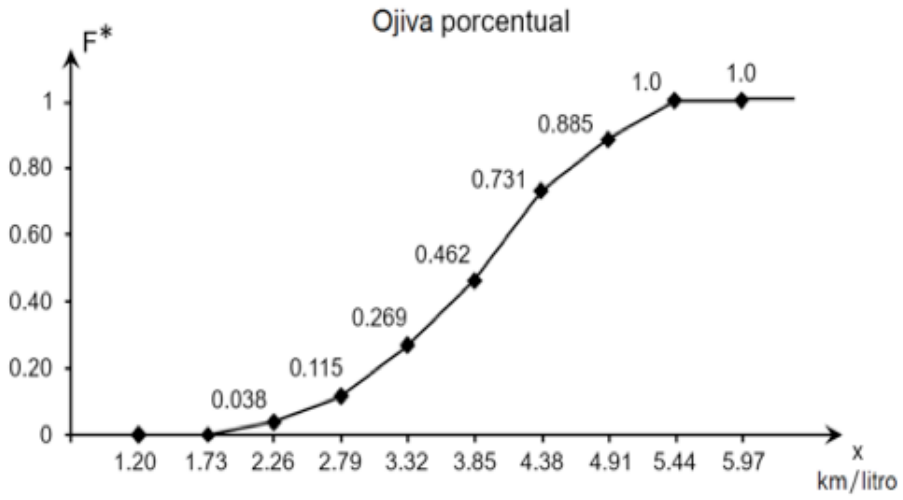


Figura 44. Ojiva para la muestra de suelo

c) Cálculo del percentil veinte del conjunto de datos.

La posición para el percentil 20 utiliza la expresión $\frac{(20)(52)}{100} = 10.4$

Se encuentra en el intervalo de clase 4, cuyo segmento de la tabla es

Tabla 47. Segmento de la tabla de frecuencias para localizar 20° percentil

Intervalo	Fronteras		Marca de clase x_i	Frecuencia f_i	Frecuencia relativa f_i^*	Frecuencia acumulada F_i^*	Frecuencia relativa acumulada F_i^*
	Inferior	Superior					
3	2.27	2.81	2.54	4	0.077	6	0.115
4	2.81	3.35	3.08	8	0.154	14	0.269

$$\frac{8}{0.54} = \frac{4.4}{x_{0.20} - 2.81} \quad ; \quad x_{0.20} = 2.81 + \frac{(4.4)(0.54)}{8} = 3.107$$

6. En la tabla se muestra la distribución de frecuencias de las medidas de resistencia a la fractura en ***MPa*** (megapascal), para barras de cerámica quemadas en un horno.

Tabla 48. Frecuencias para la resistencia a la fractura en megapascales

MPa	[81,83)	[83,85)	[85,87)	[87,89)	[89,91)	[91,93)	[93,95)	[95,97)	[97,99]
Frecuencia	6	7	17	30	43	28	22	13	3

a) Trazar un histograma y describir el comportamiento de los datos.
b) ¿Qué proporción de las observaciones son cuando menos 85 MPa?
c) ¿Qué proporción de las observaciones son menores que 95 MPa?

Solución:

a) Para trazar el histograma es necesario conocer las marcas de clase por lo que con los datos de las clases en la tabla se tiene:

Tabla 49. Frecuencias para la resistencia a la fractura

Int clase	Fronteras		Marcas de clase	Frecuencia de clase	Frecuencia relativa	Frecuencia acumulada	Frecuencia acumulada relativa
	Inferior	Superior					
1	81	83	82	6	0.036	6	0.036
2	83	85	84	7	0.041	13	0.077
3	85	87	86	17	0.101	30	0.178
4	87	89	88	30	0.178	60	0.355
5	89	91	90	43	0.254	103	0.609
6	91	93	92	28	0.166	131	0.775
7	93	95	94	22	0.130	153	0.905
8	95	97	96	13	0.077	166	0.982
9	97	99	98	3	0.018	169	1.000
			Total	169			

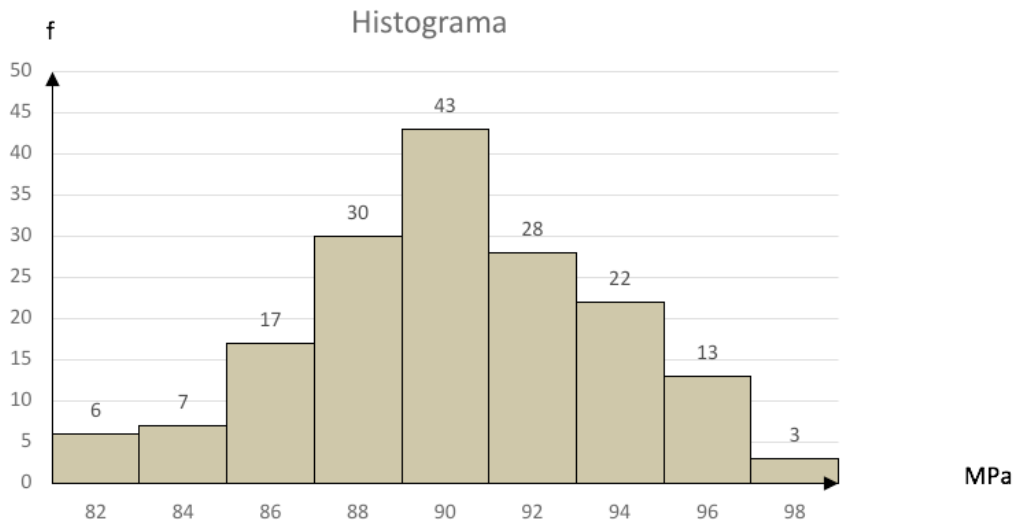


Figura 45. Histograma de resistencia a la fractura

Media:

$$\bar{x} = \frac{\sum_{i=1}^{c} x_i (f_i)}{n} = 90.17$$

La mediana se encuentra en $\frac{n}{2} = \frac{169}{2} = 84.5$ que se localiza en el intervalo de clase 5

$$x = front\ inf_{(m)} + w\left(\frac{\frac{n}{2} - F_{m-1}}{f_m}\right)$$

$$x = 89 + 2\left(\frac{\frac{169}{2} - 60}{30}\right) = 90.063$$

Moda:

$$x_{mo} = front\, inf_{(moda)} + w\left[\frac{f_k - f_{k-1}}{(f_k - f_{k-1}) + (f_k - f_{k+1})}\right] = 89 + 2\left[\frac{43 - 30}{(43 - 30) + (43 - 28)}\right] = 90.73$$

Varianza:

$$S^2_{(n)} = \frac{\sum_{i=1}^{c} (x_i - \bar{x})^2 f_i}{n} = 12.7063$$

Desviación estándar:

$$S_{(n)} = \sqrt{\frac{\sum_{i=1}^{c} (x_i - \bar{x})^2 f_i}{n}} = \sqrt{12.7063} = 3.565$$

Sesgo:

$$a_3 = \frac{\frac{1}{n}\sum_{i=1}^{c} (x_i - \bar{x})^3 f_i}{S_x^3} \quad , \quad a_3 = \frac{\frac{1}{169}(-1004.0956)}{(3.565)^3} = -0.1311$$

Curtosis:

$$a_4 = \frac{m_4}{s^4} = \frac{\frac{1}{n}\sum_{i=1}^{c}\left(x_i - \bar{x}\right)^4 f_i}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{c}\left(x_i - \bar{x}\right)^2 f_i}\right)^4}, \qquad a_4 = \frac{m_4}{s^4} = \frac{\frac{1}{169}\left(74007.639\right)}{\left(3.565\right)^4} = 2.7111$$

De acuerdo con los cálculos realizados, la resistencia promedio de las barras de cerámica quemadas en el horno es de 90.17 MPa, en tanto que la mitad de las barras tiene una resistencia mayor o igual a 90.063 MPa, valor un poco menor al promedio y el valor medido más frecuentemente es 90.73.

Con base en los valores calculados de las medidas de tendencia central, todos ellos parecidos entre sí, se puede concluir que la distribución de los datos tiene una pequeña asimetría, lo cual se puede constatar a través del coeficiente de sesgo (-0.13) que, como se puede observar es cercano a cero.

Al ordenar la media, la mediana y la moda de acuerdo con su valor, la moda es la más pequeña y la media la menor de las tres medidas. Con base en esta distribución de dichos parámetros, se podría intuir que la distribución tendría sesgo positivo.

Finalmente, el coeficiente de curtosis tiene un valor menor a 3, pero muy cercano a ese número, por lo que es posible pensar que la distribución es un poco achatada, si se le compara con la forma de la gráfica de la distribución normal.

b) Proporción de las observaciones que son cuando menos 85 MPa.

En el intervalo de clase 6 de la tabla, se localiza la frecuencia acumulada relativa del 77.5 %, para calcular cuando menos el 85 % se interpreta como el 85 % o más y se requiere del complemento

$$P_{0.85}, = 1 - 0.775 = 0.225$$

es decir, el 22.5 % de las observaciones son cuando menos 85 MPa.

c) Proporción de las observaciones que son menores que 95 MPa.

En el intervalo de clase 7 de la tabla, se localiza la frecuencia acumulada relativa del 90.5 %; como se solicita el 90 %, el 0.905 abarca ese valor.

7. Se escogió una muestra de 705 conductores de camiones de carga pesada y se registró en una tabla el número de accidentes de tránsito que tuvieron durante 4 años.

Tabla 50. Número de accidentes de tránsito

Número de accidente	0	1	2	3	4	5	6	7	8	9	10	11
Frecuencia	114	157	158	115	78	44	21	7	6	1	3	1

a) ¿Cuál es la moda?
b) Determine la media.
c) Obtenga la mediana.
d) Determine el sesgo.

Solución:

a) La mayor frecuencia de accidentes por conductor de carga pesada, en 4 años, es de 2.

$$x_{\text{moda}} = 2$$

b) Media

$$\bar{x} = \frac{\sum_{i=1}^{c} x_i f_i}{\sum_{i=1}^{c} f_i} = \frac{1623}{705} = 2.3$$

Tabla 51. Frecuencias para el número de accidentes de tránsito

x_i	f_i	F_i	$x_i^* f_i$	$(x_i - \bar{x})$	$f_i^*(x_i - \bar{x})^2$	$f_i^*(x_i - \bar{x})^3$
0	114	114	0	-2.3	603.06	-1387.04
1	157	271	157	-1.3	265.33	-344.929
2	158	429	316	-0.3	14.22	-4.266
3	115	544	345	0.7	56.35	39.445
4	78	622	312	1.7	225.42	383.214
5	44	666	220	2.7	320.76	866.052
6	21	687	126	3.7	287.49	1063.713
7	7	694	49	4.7	154.63	726.761
8	6	700	48	5.7	194.94	1111.158
9	1	701	9	6.7	44.89	300.763
10	3	704	30	7.7	177.87	1369.599
11	1	705	11	8.7	75.69	658.503
Sumas	705		1623		2420.65	4782.973

Como puede observarse los datos están muy concentrados en los cuatro primeros valores.

c) Se sugiere al estudiante calcular las columnas de frecuencia relativa (f^*) y frecuencia relativa acumulada (F^*). A partir de ahí, podrá determinar que la mediana ($\tilde{x}$) se encuentra entre los valores 1 y 2 accidentes. Por interpolación el valor teórico de la mediana es $\tilde{x} = 1.5162$ (resultado que el estudiante deberá comprobar).

d) Sesgo: El tercer momento central es:

$$m_3 = \frac{\sum_{i=1}^{c} f_i \left(x_i - \bar{x} \right)^3}{n} = \frac{4782.583}{705} = 6.7842$$

$$a_3 = \frac{m_3}{\left(m_2 \right)^{\frac{3}{2}}} = \frac{6.7842}{\left(3.434 \right)^{\frac{3}{2}}} = 1.066 > 0$$ La distribución tiene sesgo a la derecha

3.6 ACTIVIDADES DE AUTOEVALUACIÓN

1. Crucigrama

VERTICALES

1. Es una representación gráfica de una tabla de frecuencias para datos, donde uno de los ejes representa frecuencias y el otro representa marcas de clase.

5. Número que representa la mitad en un conjunto ordenado de mediciones.

6. Número que representa la mitad en un conjunto ordenado de mediciones.

HORIZONTALES

2. Primera palabra: Es una forma geométrica obtenida de segmentos de recta que unen los puntos medios de los intervalos de clase adyacentes en un histograma.

3. Tipo de medida con el que se conoce a los valores de posición: media, mediana y moda.

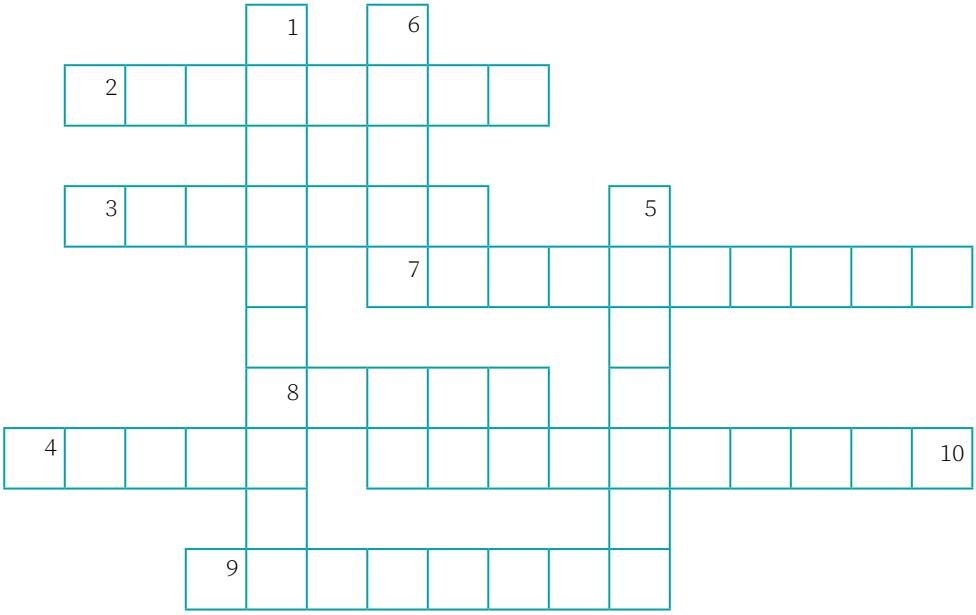
4. Valor que corresponde al punto de equilibrio de los datos y es una medida de tendencia central.

7. Sinónimo de medida de variabilidad. Nombre del conjunto de medidas que indican qué tan alejados o cercanos se encuentran los datos con respecto a la media o entre sí.

8. Mide la amplitud de todos los datos.

9. Es un valor que da una idea de la variabilidad de los datos, pero no se puede representar gráficamente y corresponde al cuadrado de la desviación estándar.

10. Primera palabra: Es la medida que describe la extensión o dispersión en un conjunto de datos, alrededor de la media. Escribirla al revés.



			H		M									
P	o	l	í	g	o	n	o							
			s		d									
C	e	n	t	r	a	l			M					
			o		d	i	s	p	e	r	s	i	o	n
			g						d					
			r	a	n	g	o		i					
M	e	d	i	a		n	o	i	c	a	i	v	s	e	d
			m						n					
		v	a	r	i	a	n	z	a					

2. Escribe **V** si la oración es verdadera o **F** si es falsa:

1. La mediana es un valor en el rango de la muestra que distribuye de manera equilibrada los datos alrededor de él. ( )

2. La frecuencia relativa asociada a una marca de clase muestra el número de elementos acumulados o datos contados de manera acumulada, entre el total de datos. ( )

3. La estadística descriptiva utiliza solamente valores enteros. ( )

4. La muestra son todos los valores que se desea estudiar en la estadística descriptiva. ( )

5. Para realizar la gráfica de la ojiva, se necesitan los valores de las fronteras y de la frecuencia acumulada. ( )

6. Los valores de la media, la mediana y la moda siempre deben coincidir. ( )

7. La mediana se calcula mediante el promedio de todos valores de los datos. ( )

8. El segundo cuartil corresponde al cincuentavo percentil. ( )

9. Los deciles se calculan dividiendo el intervalo en cien partes iguales. ( )

10. El coeficiente de variación tiene las mismas unidades que los datos trabajados en el experimento. ( )

11. Cuando la medida del sesgo es positiva, se sabe que los valores se encuentran concentrados a la derecha en la gráfica. ( )

12. Si la curtosis resulta un valor menor de 3, la gráfica de la distribución es achatada. ( )

3.7 APLICACIONES CON SOTFWARE ESPECIALIZADO

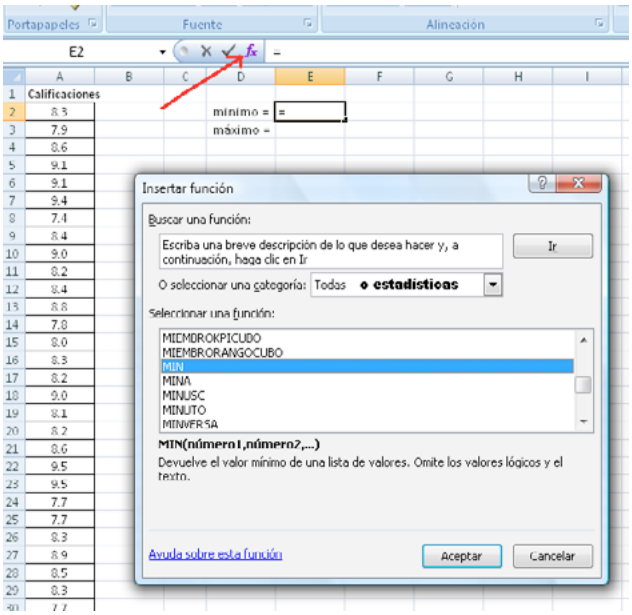
Ejemplo realizado con Excel

Elaborar la tabla de distribución de frecuencias para la siguiente tabla de valores que representa el promedio de calificaciones de 48 alumnos:

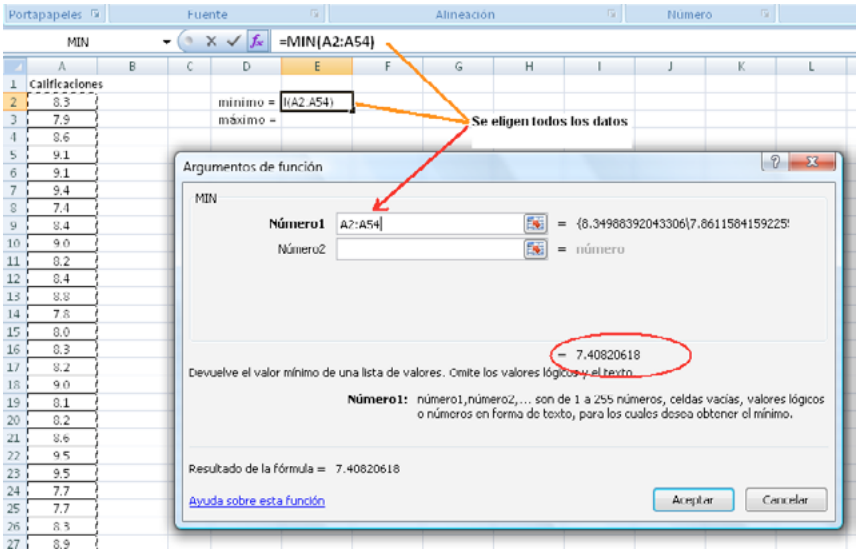
Tabla 52. Datos de estaturas en metros

Calificaciones					
8.3	9.0	8.2	8.4	8.8	7.8
7.9	8.0	8.3	8.2	9.0	8.1
8.6	8.2	8.6	9.5	9.5	7.7
9.1	7.7	8.3	8.9	8.5	8.3
9.1	7.6	8.3	9.7	8.2	8.5
9.4	8.0	8.3	8.2	8.8	8.5
7.4	8.1	9.2	9.3	8.3	8.3
8.4	7.4	8.5	7.7	8.9	9.6

Para formar la tabla de frecuencias con Excel, primero se deberán buscar el número mayor, el número menor y calcular con ellos el rango. Para ello se selecciona

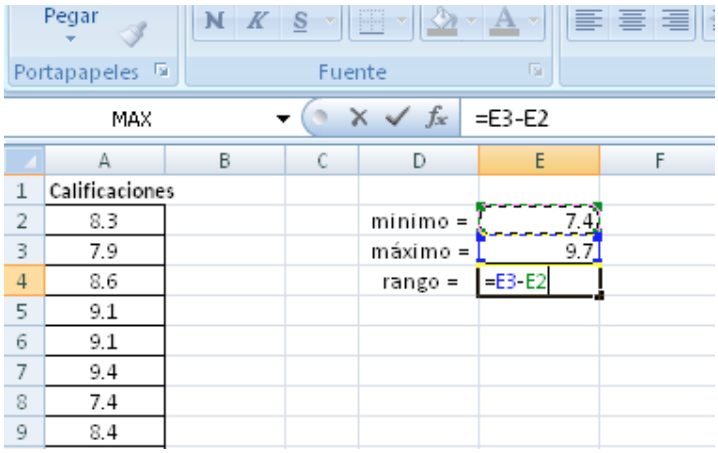


Al abrirse la ventana de selección de datos, se elegirán todos los datos

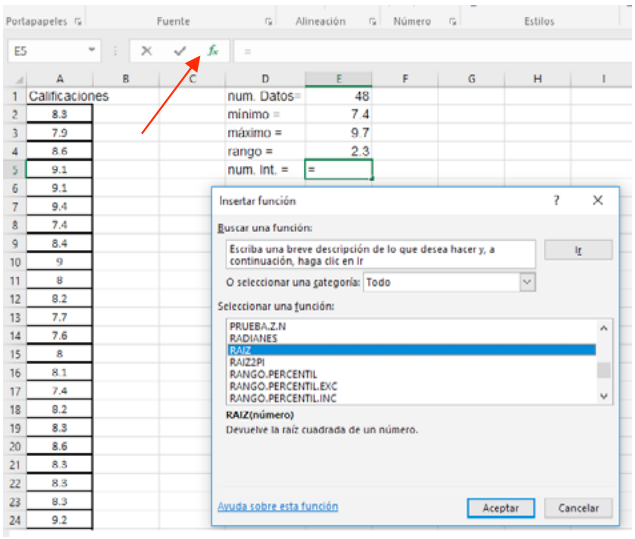


Muestra el valor mínimo que será colocado en la celda.

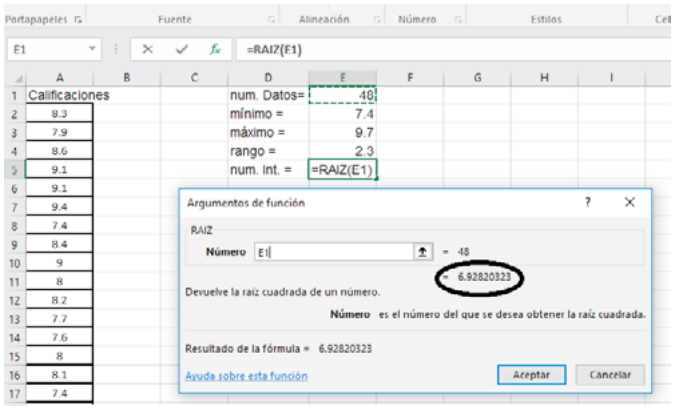
Se procederá de la misma manera para el valor máximo eligiendo en la función *fx* MAX.



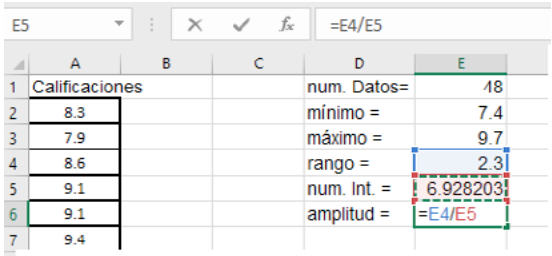
Para formar la tabla de frecuencias es necesario calcular la raíz del total de datos.



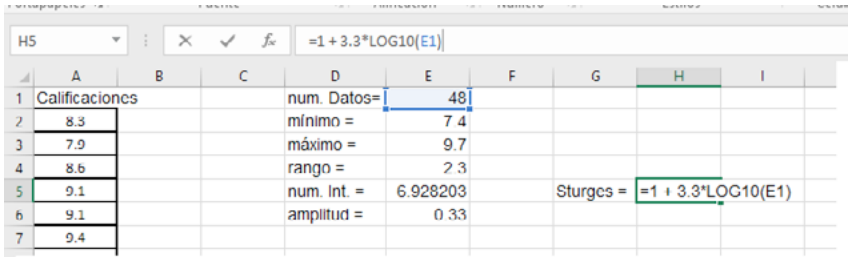
Mostrando el resultado



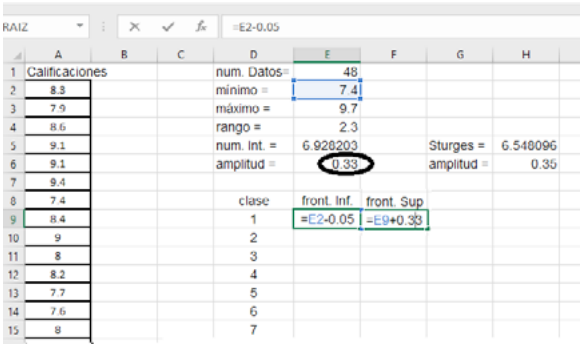
Se calcula también el tamaño del intervalo, que se redondeará para que sea un número entero.



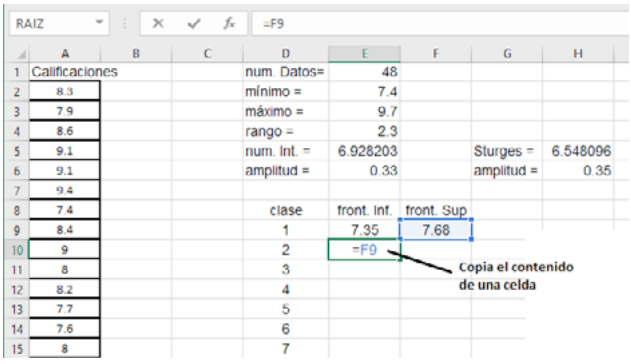
Con esta información se construirá la tabla de frecuencias, se comienza por numerar las clases y se coloca el valor de la frontera inferior comenzando con el valor mínimo menos 0.05 y el superior se obtendrá sumando al inmediato inferior el tamaño del intervalo obtenido.



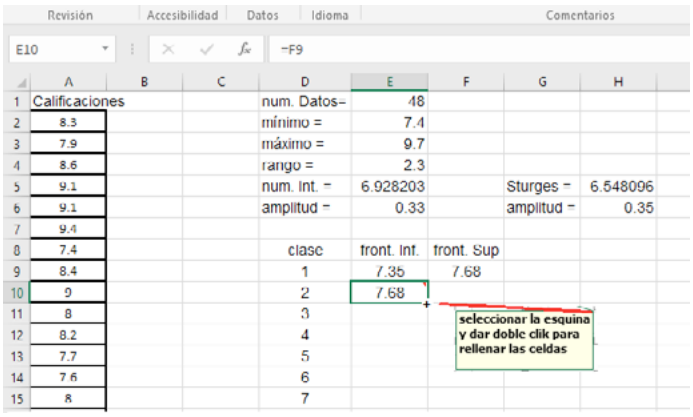
	A	B	C	D	E	F	G	H
1	Calificaciones			num. Datos=	48			
2	8.3			mínimo =	7.4			
3	7.9			máximo =	9.7			
4	8.6			rango =	2.3			
5	9.1			num. Int. =	6.928203		Sturges =	6.548096
6	9.1			amplitud =	0.33		amplitud =	0.35



Rellenar las siguientes celdas hacia abajo.

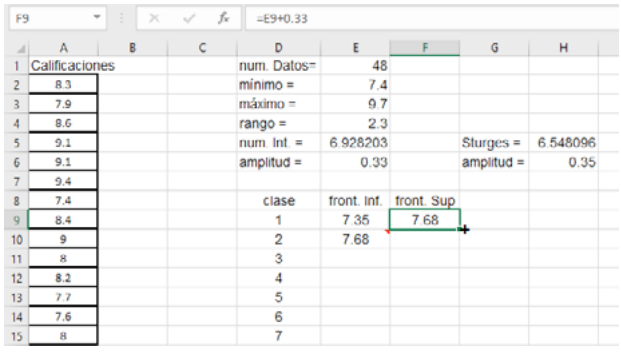


Antes de repetir el proceso para hacer lo mismo en el lado del intervalo superior, deberá fijarse la amplitud del intervalo.



	A	D	E	F	G	H
1	Calificaciones	num. Datos=	48			
2	8.3	mínimo =	7.4			
3	7.9	máximo =	9.7			
4	8.6	rango =	2.3			
5	9.1	num. Int. =	6.928203		Sturges =	6.548096
6	9.1	amplitud =	0.33		amplitud =	0.35
7	9.4					
8	7.4	clase	front. Inf.	front. Sup		
9	8.4	1	7.35	7.68		
10	9	2	7.68			
11	8	3				
12	8.2	4				
13	7.7	5				
14	7.6	6				
15	8	7				

Repetir el proceso en la siguiente celda a la derecha.



	A	D	E	F	G	H
1	Calificaciones	num. Datos=	48			
2	8.3	mínimo =	7.4			
3	7.9	máximo =	9.7			
4	8.6	rango =	2.3			
5	9.1	num. Int. =	6.928203		Sturges =	6.548096
6	9.1	amplitud =	0.33		amplitud =	0.35
7	9.4					
8	7.4	clase	front. Inf.	front. Sup		
9	8.4	1	7.35	7.68		
10	9	2	7.68			
11	8	3				
12	8.2	4				
13	7.7	5				
14	7.6	6				
15	8	7				

También la tabla se llenará al sostener el botón izquierdo y deslizar hacia abajo manteniendo la cruz marcada de la celda elegida. Se puede observar que el último intervalo superior es menor al valor máximo de la información, por lo que se debe incrementar la amplitud del intervalo en una unidad decimal más.

	C	D	E
	num. Datos=	48	
	mínimo =	7.4	
	máximo =	9.7	
	rango =	2.3	
	num. Int. =	6.9	
	amplitud =	0.33	
	clase	front. Inf.	front. Sup
	1	7.35	7.68
	2	7.68	8.01
	3	8.01	8.34
	4	8.34	8.67
	5	8.67	9
	6	9	9.33
	7	9.33	9.66

=(D4/D5)+0.1

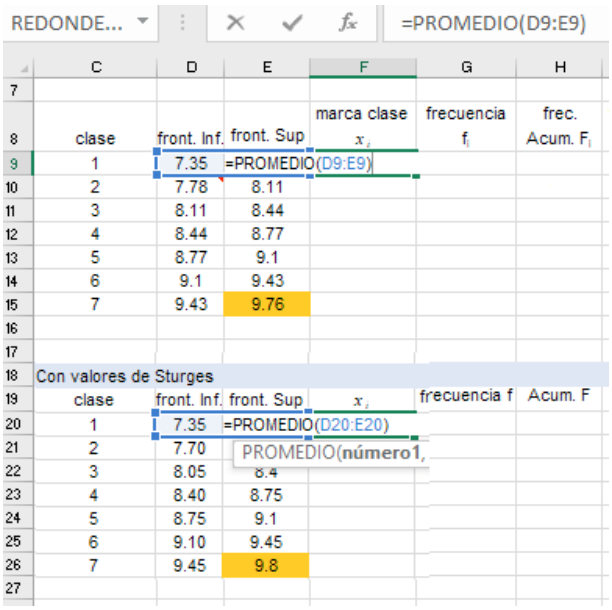
	C	D	E
	num. Datos=	48	
	mínimo =	7.4	
	máximo =	9.7	
	rango =	2.3	
	num. Int. =		
	amplitud =	=(D4/D5)+0.1	
	clase	front. Inf.	front. Sup
	1	7.35	7.78
	2	7.78	8.11
	3	8.11	8.44
	4	8.44	8.77
	5	8.77	9.1
	6	9.1	9.43
	7	9.43	9.76

Para comparar, se crea la tabla utilizando los valores obtenidos con Sturges

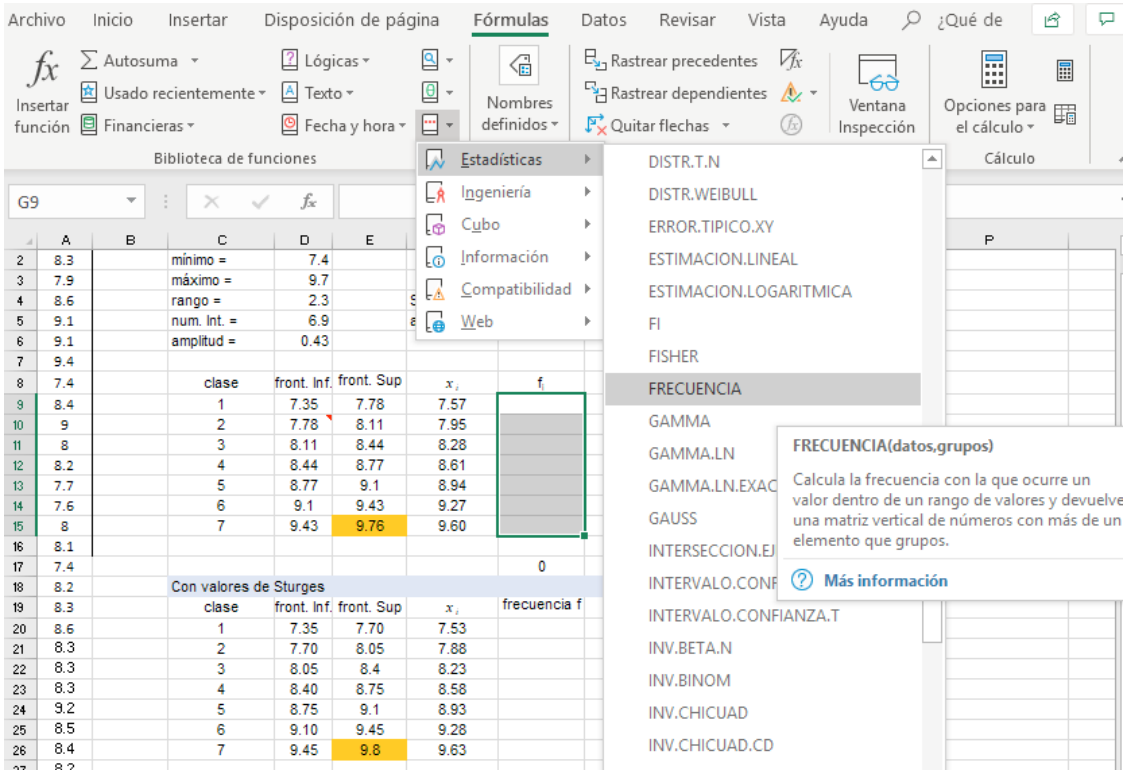
A	B	C	D	E	F	G
Calif.		num. Datos=	48			
8.3		mínimo =	7.4			
7.9		máximo =	9.7			
8.6		rango =	2.3		Sturges =	6.5
9.1		num. Int. =	6.9		amplitud =	0.35
9.1		amplitud =	0.43			
9.4						
7.4		clase	front. Inf.	front. Sup		
8.4		1	7.35	7.78		
9		2	7.78	8.11		
8		3	8.11	8.44		
8.2		4	8.44	8.77		
7.7		5	8.77	9.1		
7.6		6	9.1	9.43		
8		7	9.43	9.76		
8.1						
7.4						
8.2		Con valores de Sturges				
8.3		clase	front. Inf.	front. Sup		
8.6		1	7.35	7.70		
8.3		2	7.70	8.05		
8.3		3	8.05	8.4		
8.3		4	8.40	8.75		
9.2		5	8.75	9.1		
8.5		6	9.10	9.45		
8.4		7	9.45	9.8		

El alumno puede darse cuenta de que es posible aumentar la amplitud del intervalo en el caso de requerirlo debido a que el último valor en la tabla no abarca al valor máximo obtenido de los datos. En este caso, con la amplitud del intervalo obtenida con los valores de Sturges sí se abarcan todos los datos originales.

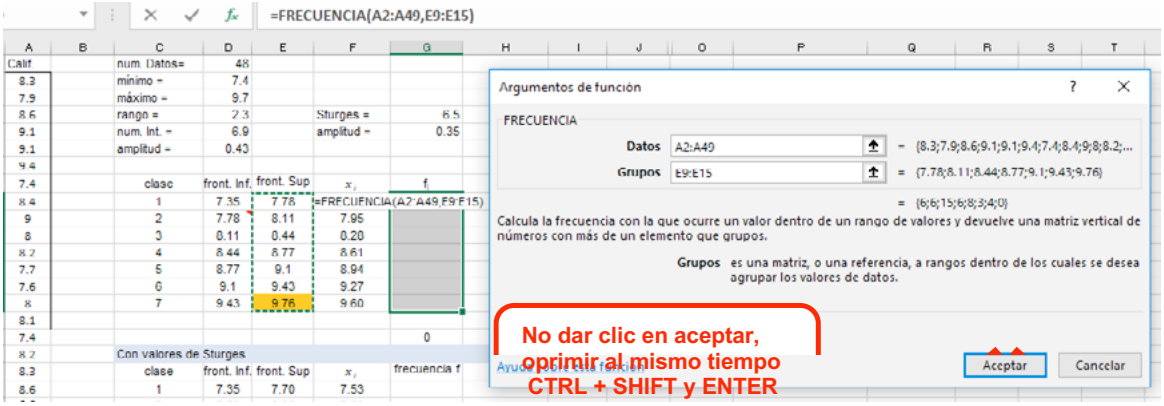
Una vez que se tienen las fronteras se procede a obtener la marca de clase x_i, que servirán para representar lo que en probabilidad se vio como función de distribución de probabilidad, donde los valores de la variable aleatoria corresponden a la marca de clase, estos valores sirven para representar gráficamente la tabla, junto con las frecuencias.



Columnas de frecuencia, frecuencia acumulada y sus valores relativos. Una vez agregada la columna y con la misma seleccionada, se proceden a contar las frecuencias mediante la función ***frecuencia***.



En *datos*, se debe elegir el conjunto de datos no agrupados, en *grupos* se deben elegir los datos de la frontera superior previamente obtenida. Al final, se deben presionar al mismo tiempo las teclas ***Ctrl+Shift+Enter***. En automático se hará el conteo de la frecuencia dentro de las fronteras, y se comprobará que la suma de las frecuencias sea igual al número de datos.



Los valores de las frecuencias acumuladas, de las relativas y relativas acumuladas se obtienen haciendo las operaciones correspondientes a las que se estudiaron en la teoría, con las frecuencias y el número de datos proporcionados.

clase	front. Inf.	front. Sup	marca clase x_i	frecuencia f_i	frec. Acum. F_i	frec. Rel. f_i*	frec. Acum. F_i*
1	7.35	7.78	7.57	6	=G9		
2	7.78	8.11	7.95	6			
3	8.11	8.44	8.28	15			
4	8.44	8.77	8.61	6			
5	8.77	9.1	8.94	8			
6	9.1	9.43	9.27	3			
7	9.43	9.76	9.60	4			

num. Datos=	48						
mínimo =	7.4						
máximo =	9.7						
rango =	2.3		Sturges =	6.5			
num. Int. =	6.9		amplitud =	0.35			
amplitud =	0.43						
clase	front. Inf.	front. Sup	marca clase x_i	frecuencia f_i	frec. Acum. F_i	frec. Rel. f_i*	frec. Acum. F_i*
1	7.35	7.78	7.57	6	6	G9 D1	=H9/D1
2	7.78	8.11	7.95	6	=H9+G10		
3	8.11	8.44	8.28	15			
4	8.44	8.77	8.61	6			
5	8.77	9.1	8.94	8			
6	9.1	9.43	9.27	3			
7	9.43	9.76	9.60	4			

Las tablas de frecuencias quedan como se muestra en la siguiente tabla, considerando el número de intervalos obtenido a partir de la raíz cuadrada del tamaño de la muestra y a partir de la regla de Sturges:

O32

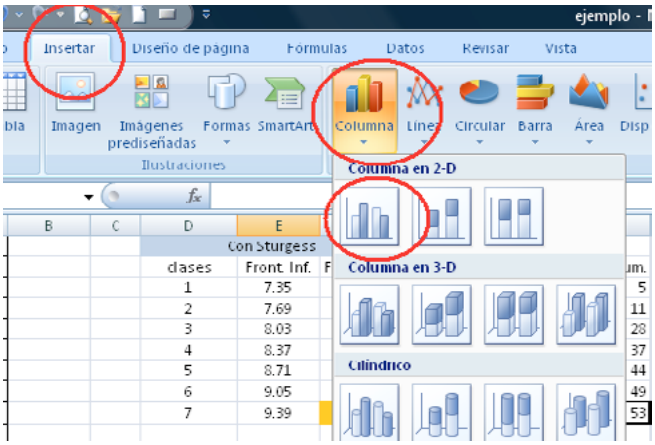
	C	D	E	F	G	H	I	J
7								
8	clase	front. Inf.	front. Sup	marca clase x_i	frecuencia f_i	frec. Acum. F_i	frec. Rel. f_i*	frec. Acum. F_i*
9	1	7.35	7.78	7.57	6	6	0.125	0.125
10	2	7.78	8.11	7.95	6	12	0.125	0.250
11	3	8.11	8.44	8.28	15	27	0.313	0.563
12	4	8.44	8.77	8.61	6	33	0.125	0.688
13	5	8.77	9.1	8.94	8	41	0.167	0.854
14	6	9.1	9.43	9.27	3	44	0.063	0.917
15	7	9.43	9.76	9.60	4	48	0.083	1.000
16								
17								
18	Con valores de Sturges							
19	clase	front. Inf.	front. Sup	x_i	frecuencia f	Acum. F	f_i*	Acum. F_i*
20	1	7.35	7.70	7.53	6	6	0.13	0.13
21	2	7.70	8.05	7.88	4	10	0.08	0.21
22	3	8.05	8.4	8.23	17	27	0.35	0.56
23	4	8.40	8.75	8.58	6	33	0.13	0.69
24	5	8.75	9.1	8.93	8	41	0.17	0.85
25	6	9.10	9.45	9.28	3	44	0.06	0.92
26	7	9.45	9.8	9.63	4	48	0.08	1.00

A partir de la tabla, se podrán obtener las gráficas que se deseen.

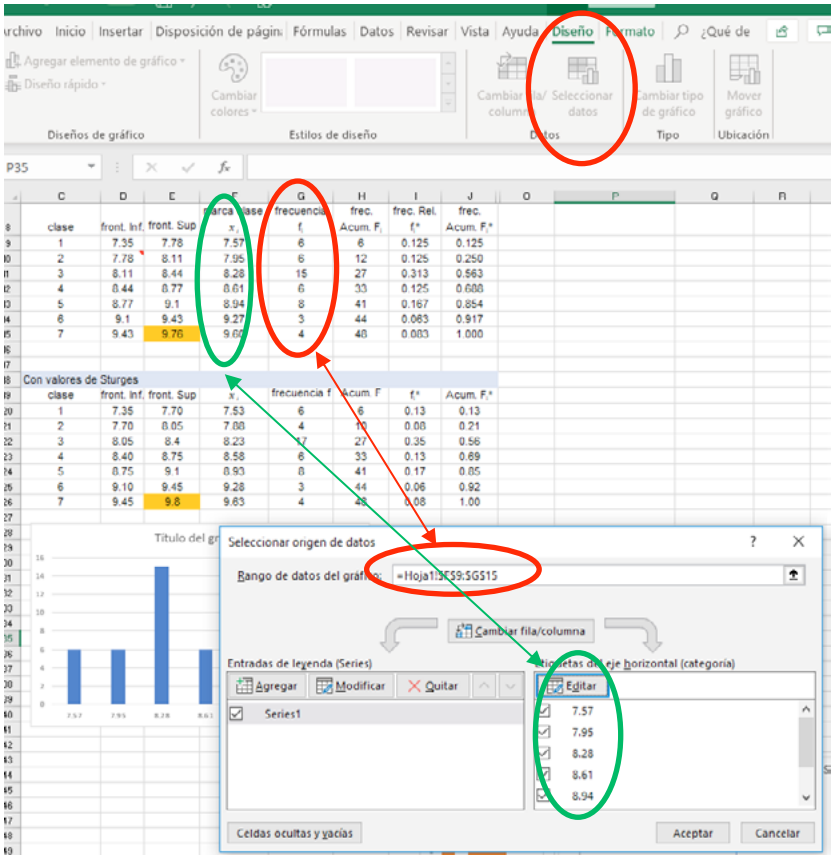
R36

	C	D	E	F	G	H	I	J
7								
8	clase	front. Inf.	front. Sup	marca clase x_i	frecuencia f_i	frec. Acum. F_i	frec. Rel. f_i*	frec. Acum. F_i*
9	1	7.35	7.78	7.57	6	6	0.125	0.125
10	2	7.78	8.11	7.95	6	12	0.125	0.250
11	3	8.11	8.44	8.28	15	27	0.313	0.563
12	4	8.44	8.77	8.61	6	33	0.125	0.688
13	5	8.77	9.1	8.94	8	41	0.167	0.854
14	6	9.1	9.43	9.27	3	44	0.063	0.917
15	7	9.43	9.76	9.60	4	48	0.083	1.000
16								
17								
18	Con valores de Sturges							
19	clase	front. Inf.	front. Sup	x_i	frecuencia f	Acum. F	f_i*	Acum. F_i*
20	1	7.35	7.70	7.53	6	6	0.13	0.13
21	2	7.70	8.05	7.88	4	10	0.08	0.21
22	3	8.05	8.4	8.23	17	27	0.35	0.56
23	4	8.40	8.75	8.58	6	33	0.13	0.69
24	5	8.75	9.1	8.93	8	41	0.17	0.85
25	6	9.10	9.45	9.28	3	44	0.06	0.92
26	7	9.45	9.8	9.63	4	48	0.08	1.00

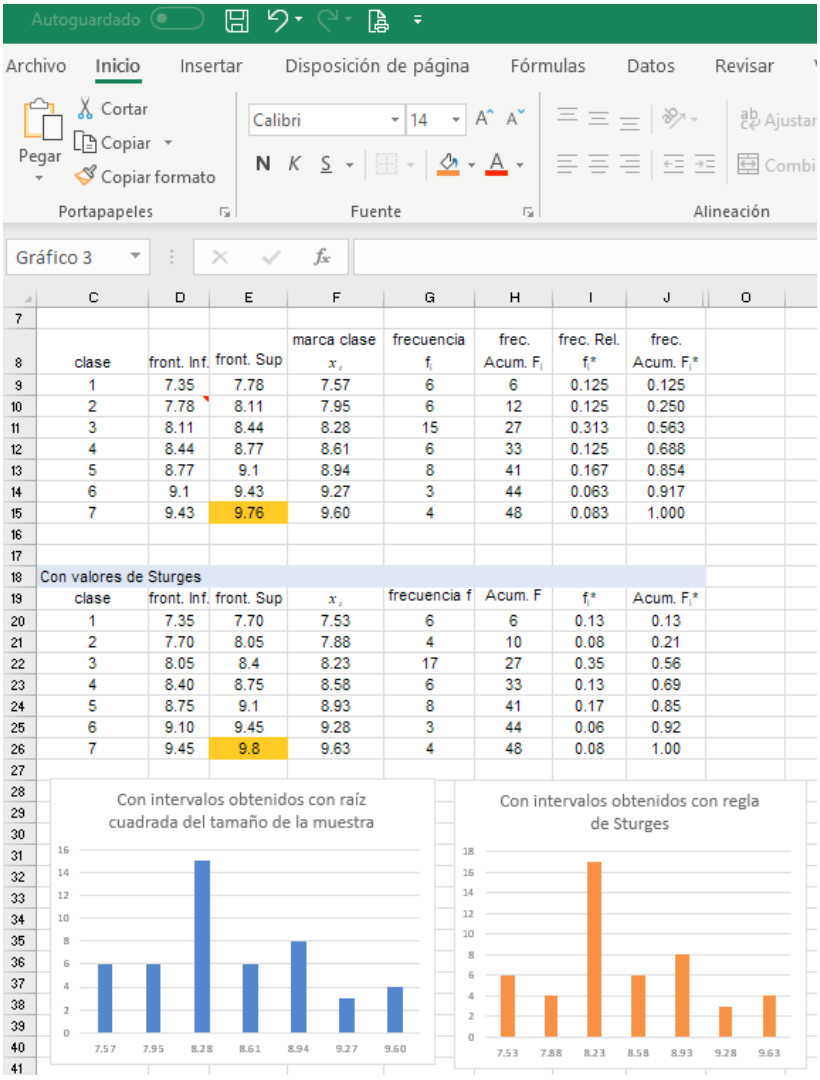
Para trazar el histograma, se elegirán las columnas: x_i (marcas de clase) y frecuencia. Seleccionar *Insertar* del menú de insertar, elegir *Columna*, y aparecerá una ventana con distintas opciones; elegir la de barras sencillas.



Al dar *clic*, aparecerá una ventana en blanco, se deberá elegir la información que se va a trazar; en *rango de datos* se seleccionan todos los datos de la frecuencia (los valores aparecerán en el eje vertical), después se elige *editar* en el lado derecho para después seleccionar todos los datos de la marca de clase (los valores aparecerán en el eje horizontal).

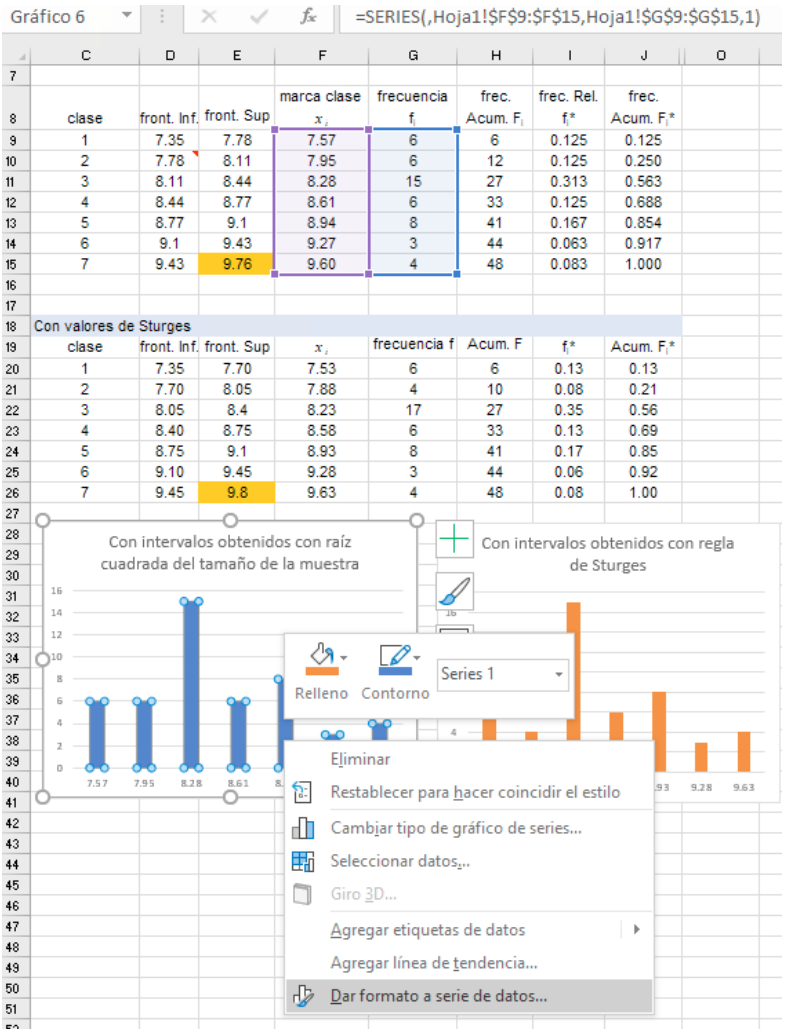


Los histogramas para cada tabla de frecuencias quedan:



	C	D	E	F	G	H	I	J	O
8	clase	front. Inf.	front. Sup	marca clase x_i	frecuencia f_i	frec. Acum. F_i	frec. Rel. f_i*	frec. Acum. F_i*	
9	1	7.35	7.78	7.57	6	6	0.125	0.125	
10	2	7.78	8.11	7.95	6	12	0.125	0.250	
11	3	8.11	8.44	8.28	15	27	0.313	0.563	
12	4	8.44	8.77	8.61	6	33	0.125	0.688	
13	5	8.77	9.1	8.94	8	41	0.167	0.854	
14	6	9.1	9.43	9.27	3	44	0.063	0.917	
15	7	9.43	9.76	9.60	4	48	0.083	1.000	
18	Con valores de Sturges								
19	clase	front. Inf.	front. Sup	x_i	frecuencia f	Acum. F	f_i*	Acum. F_i*	
20	1	7.35	7.70	7.53	6	6	0.13	0.13	
21	2	7.70	8.05	7.88	4	10	0.08	0.21	
22	3	8.05	8.4	8.23	17	27	0.35	0.56	
23	4	8.40	8.75	8.58	6	33	0.13	0.69	
24	5	8.75	9.1	8.93	8	41	0.17	0.85	
25	6	9.10	9.45	9.28	3	44	0.06	0.92	
26	7	9.45	9.8	9.63	4	48	0.08	1.00	

Si se desea cerrar los espacios entre las barras se deberá situar el ratón dentro de una barra y hacer clic con el botón derecho. Se abrirá una ventana para elegir *dar formato a la serie de datos*, y se hace clic con el botón izquierdo para hacer el cambio en el eje.

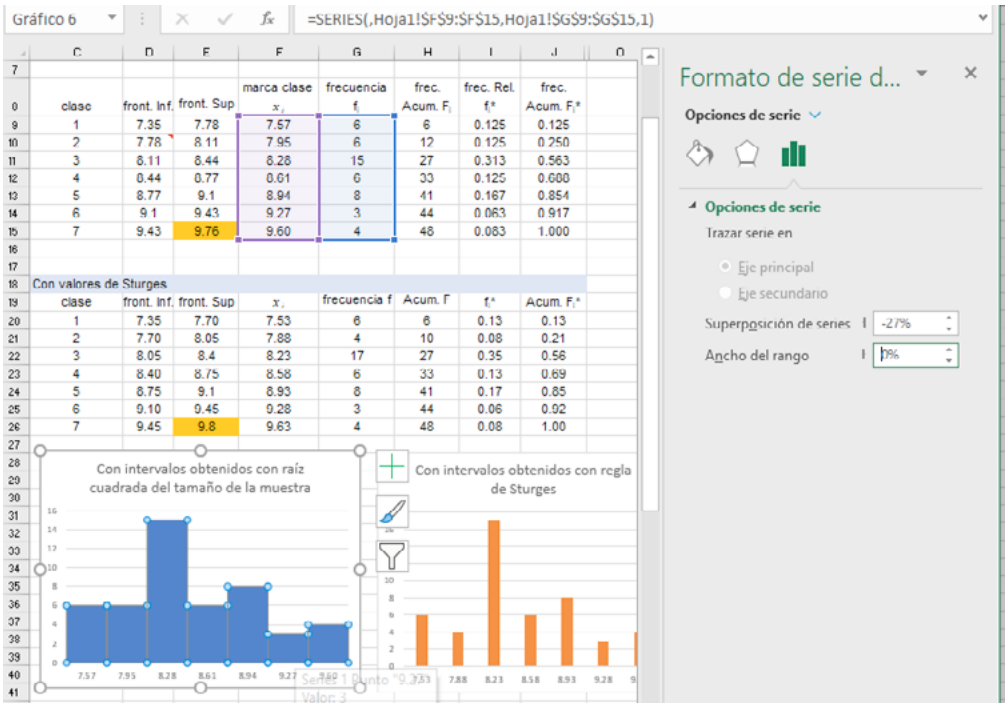


clase	front. Inf.	front. Sup	marca clase x_i	frecuencia f_i	frec. Acum. F_i	frec. Rel. f_i*	frec. Acum. F_i*
1	7.35	7.78	7.57	6	6	0.125	0.125
2	7.78	8.11	7.95	6	12	0.125	0.250
3	8.11	8.44	8.28	15	27	0.313	0.563
4	8.44	8.77	8.61	6	33	0.125	0.688
5	8.77	9.1	8.94	8	41	0.167	0.854
6	9.1	9.43	9.27	3	44	0.063	0.917
7	9.43	9.76	9.60	4	48	0.083	1.000

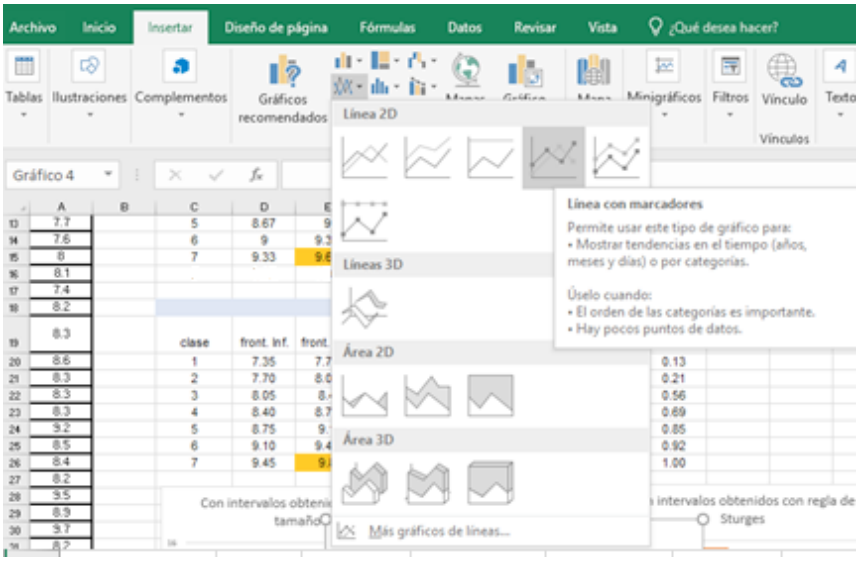
Con valores de Sturges

clase	front. Inf.	front. Sup	x_i	frecuencia f	Acum. F	f_i*	Acum. F_i*
1	7.35	7.70	7.53	6	6	0.13	0.13
2	7.70	8.05	7.88	4	10	0.08	0.21
3	8.05	8.4	8.23	17	27	0.35	0.56
4	8.40	8.75	8.58	6	33	0.13	0.69
5	8.75	9.1	8.93	8	41	0.17	0.85
6	9.10	9.45	9.28	3	44	0.06	0.92
7	9.45	9.8	9.63	4	48	0.08	1.00

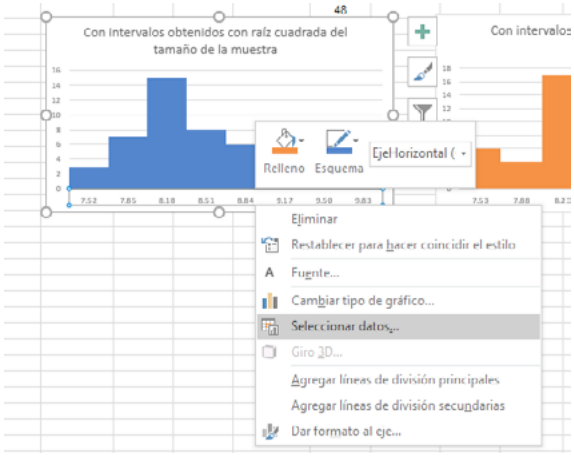
En la ventana, manteniendo oprimido el botón izquierdo, recorrer la barra hasta el cero, y se verá el cambio en la gráfica, finalmente cerrar la ventana.



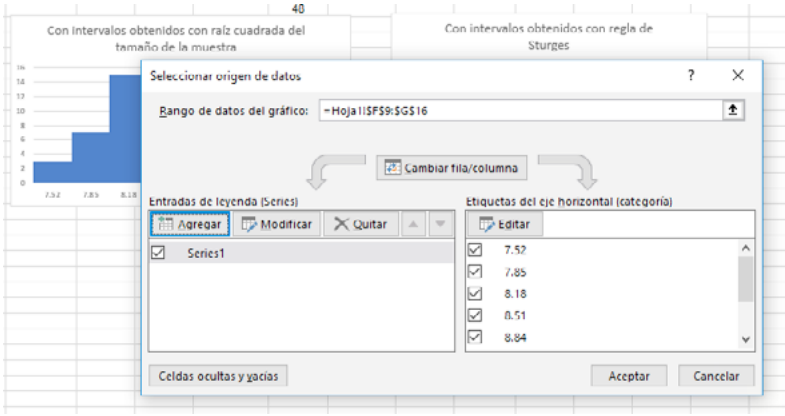
Para trazar el polígono de frecuencias, se puede repetir el proceso descrito en el histograma, y se elige como tipo de gráfico el de línea con marcadores.



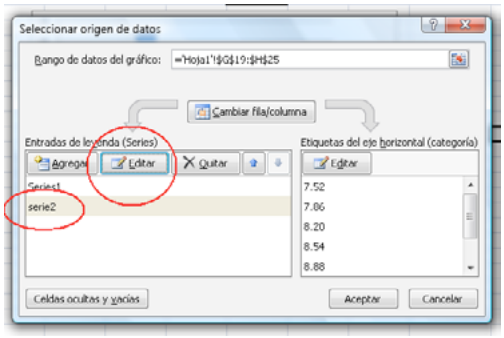
O también, se puede colocar el cursor dentro del histograma que se acaba de elaborar y dar clic con botón derecho, con lo que se mostrará una ventana en la que se deberá elegir *Seleccionar datos*.



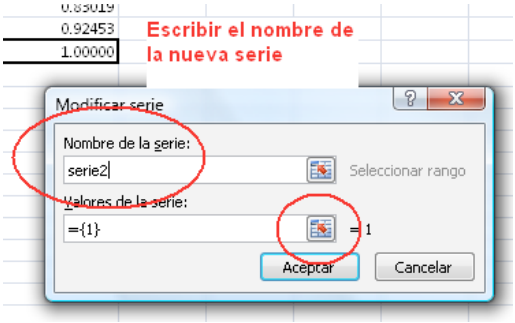
Aparecerá una ventana donde se elegirá *Agregar* para insertar una nueva serie de datos.



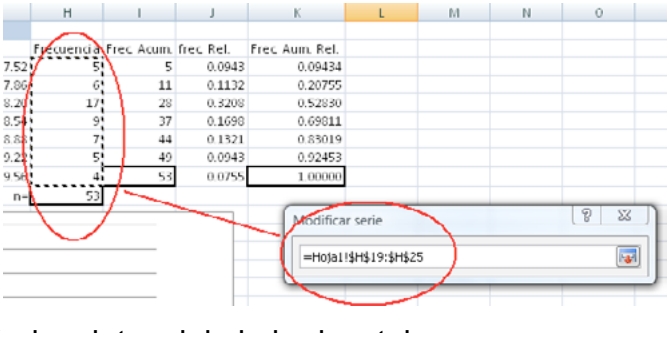
Elegir los datos que estarán en el eje vertical en la nueva serie a través de *Editar*.



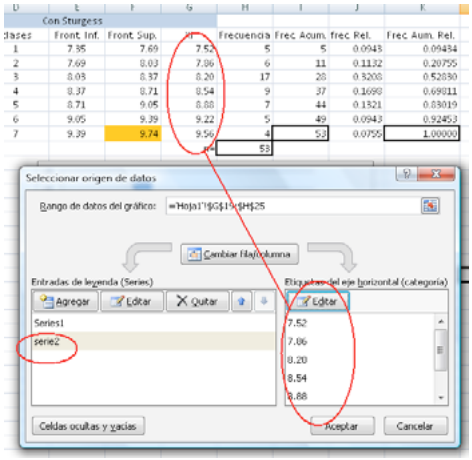
En el cuadro escribir el nombre de la nueva serie y elegir los datos para el eje vertical.



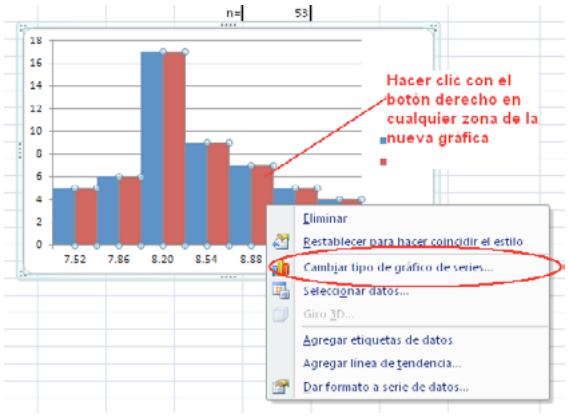
Elegir la información correspondiente al eje de las ordenadas.



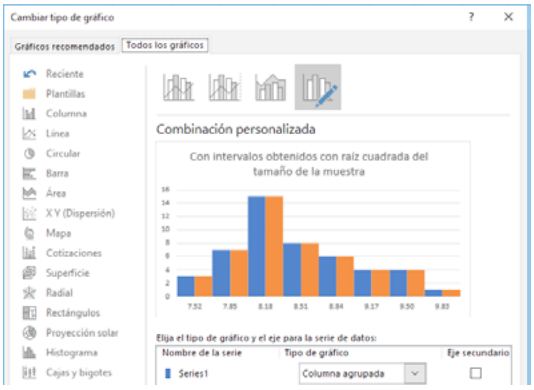
Después de ajustarán los datos del eje horizontal.



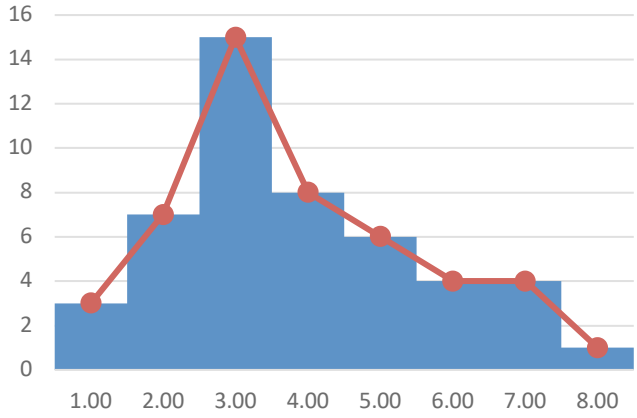
Al dar Aceptar, la gráfica correspondiente estará de otro color, pero también en forma de barras, para cambiarla se ubicará el cursor en cualquier zona de la nueva gráfica y con clic derecho se abrirá una ventana para hacer el ajuste.



Cambiar el tipo de gráfico.

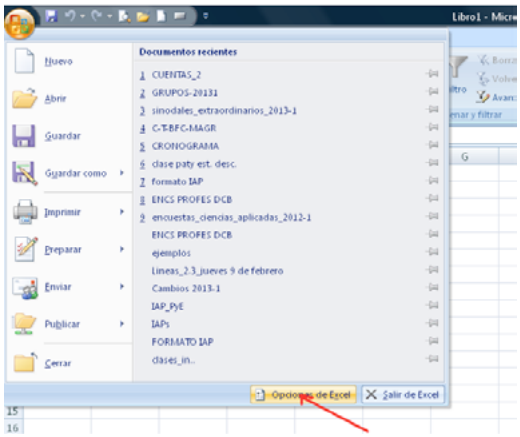


Al aceptar quedará la gráfica del polígono de frecuencias junto con la de barras.

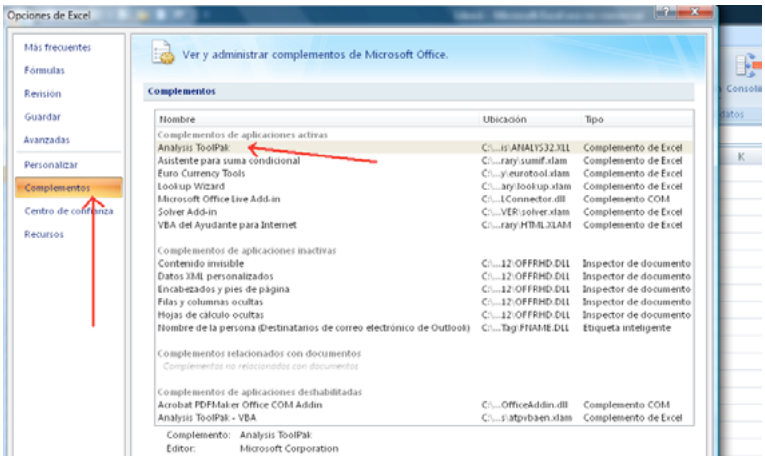


Otra forma para realizar el análisis de la información es con *Análisis de datos* incluido en Excel.

Para realizar este análisis de la información en Excel se debe tener habilitada la opción de *Análisis de datos*, dentro de las opciones en la ventana de datos. Si esta opción no se encuentra será necesario habilitarla desde opciones de Excel.

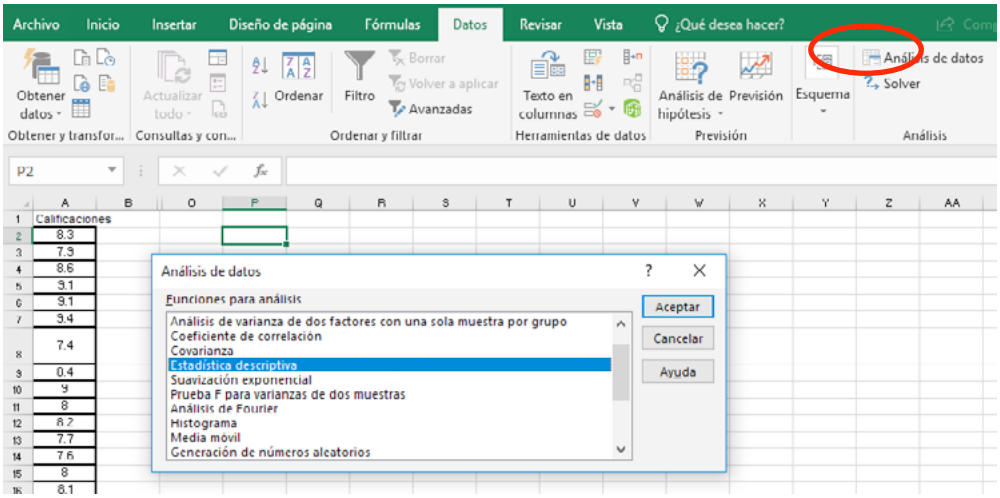


Seleccionar *Complementos* y después *Análisis*

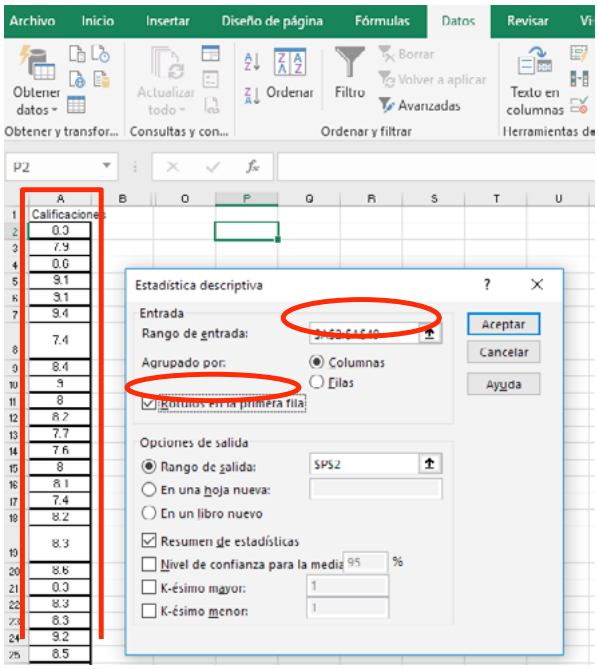


Dar clic izquierdo en Aceptar y quedará instalado, en algunos casos hay que reiniciar el equipo para que quede habilitada la opción.

Una vez instalado en la ventana de datos, se visualizará el ícono de *Análisis*, al elegirlo, se abrirá una ventana que muestra un menú de opciones para analizar la información. Seleccionar *Estadística descriptiva*.



Se abrirá otra ventana en la que se solicita la información que se analizará, primero el rango de datos, seleccionar toda la tabla de datos con el ratón, incluyendo el título, mismo que se indicará en *rótulos en la primera fila*. Posteriormente, seleccionar dónde se desea colocar la salida de información, se puede elegir *output range* e ingresar la celda donde se desee que se comience a colocar la salida de la información. Elegir también el *resumen estadístico* y finalmente Aceptar. Los datos deben estar o en un renglón o en una columna para facilitar su análisis, el resultado del análisis es:



Calificaciones	
Media	8.470833333
Error típico	0.083949407
Mediana	8.3
Moda	8.3
Desviación estándar	0.581618554
Varianza de la muestra	0.338280142
Curtosis	-0.464050565
Coeficiente de asimetría	0.320061456
Rango	2.3
Mínimo	7.4
Máximo	9.7
Suma	406.6
Cuenta	48

3.8 EJERCICIOS PROPUESTOS No. 2

1. En la tabla se muestran los datos de las estaturas en metros de 40 alumnos de un grupo de Estadística del presente semestre.

Tabla 53. Datos de estaturas en metros

1.54	1.69	1.68	1.69	1.55	1.59	1.68	1.56	1.59	1.68
1.65	1.56	1.72	1.54	1.67	1.64	1.63	1.70	1.71	1.65
1.71	1.72	1.74	1.49	1.70	1.72	1.79	1.70	1.74	1.76
1.71	1.75	1.73	1.81	1.77	1.73	1.80	1.67	1.68	1.82

a) Elaborar un histograma.
b) ¿Hasta qué valor se tiene el 25 % de los más bajos de estatura?
c) ¿A partir de qué valor se tiene la cuarta parte de los más altos?
d) ¿A partir de qué estatura se tiene el 40 % de los más altos?
e) Elaborar un histograma en donde se muestre lo anterior.
f) ¿Cuál es el valor de la mediana?

2. Considere la siguiente distribución de frecuencias con datos agrupados:

Tabla 54. Distribución de frecuencias

i	Lim. Inf	Lim.sup	f_i	F_i	Front. inf	Front. sup	Marca x_i	f_i*	F_i*
1	1	10	2	2	0.5	10.5	5.5	0.1333	0.13
2	11	20	4	6	10.5	20.5	15.5	0.2667	0.40
3	21	30	6	12	20.5	30.5	25.5	0.4000	0.80
4	31	40	2	14	30.5	40.5	35.5	0.1333	0.93
5	41	50	1	15	40.5	50.5	45.5	0.0667	1.0
	Suma		15					1.0	

a) Calcular la media y la moda.
b) Determine la mediana. Localícela en la ojiva porcentual.
Resp.: a) $X_{Media} = 22.83$, $X_{Moda} = 25.5$; b) $X_{mediana} = 23.0$, demostración $A_{izq} = A_{der} = 5.0$

3. Investigación documental: Realice la siguiente investigación documental, con datos de los últimos 20 presidentes de los Estados Unidos Mexicanos.

a) Cuadro de ordenación con sus nombres, edades en la que iniciaron su mandato y periodos en años, que ocuparon el cargo de presidente de México,
b) Edad promedio en la que iniciaron el mandato.
c) ¿A qué personaje de los enlistados corresponde la mediana de la edad mencionada (inicio de mandato)?
d) Calcule la desviación estándar de las edades.
e) ¿A quiénes (personajes) corresponde el cuartil más bajo y a cuál el más alto de la edad del inicio de mandato?

3.9 MEDICIONES Y REPRESENTACIONES DE POSICIÓN RELATIVA

3.9.1 Aplicación empírica del teorema de Tchebycheff

TEOREMA DE TCHEBYCHEFF

En algunos casos es conveniente tener un método que conduzca a identificar, de manera rápida, los valores de datos comunes o frecuentes y aquellos poco frecuentes o atípicos en las distribuciones. La conclusión del Teorema de Tchebycheff [3] se puede tomar como orientación para describir el comportamiento de un conjunto de datos. Se expresa de la siguiente forma:

> **La proporción de cualquier distribución que esté a menos de k desviaciones estándar de la media es por lo menos $1 - \frac{1}{k^2}$ donde k es cualquier número positivo mayor que 1. Este teorema es válido para todas las distribuciones de datos.**

Esto es, el valor de k puede ser cualquier número real igual o mayor a uno. El planteamiento de Tchebycheff se ilustra en la gráfica que se muestra a continuación:

[3] Desarrollado en 1867 por P. L. Tchebysheff (1821–1894) matemático de origen ruso.

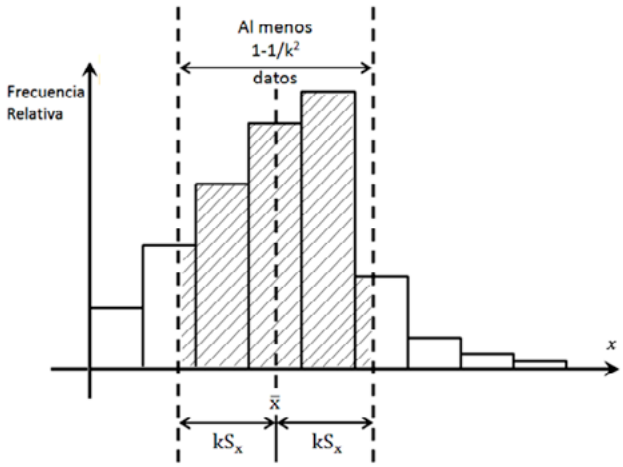


Figura 46. Gráfica para el Teorema de Tchevycheff

Con base en lo anterior, si se considera a manera de ejemplo los valores de $k = 1$, 1.5, 2, 2.5 y 3, se tendrá, de acuerdo con el mencionado teorema:

Tabla 55. Teorema de Tchebycheff

k	$1-\frac{1}{k^2}$	Intervalo $\bar{x}-ks_x<x<\bar{x}+ks_x$	Mínimo porcentaje de datos contenidos en el intervalo
1.0	0.00	$\bar{x}-s_x<x<\bar{x}+s_x$	0
1.5	0.55	$\bar{x}-1.5s_x<x<\bar{x}+1.5s_x$	55
2.0	0.75	$\bar{x}-2s_x<x<\bar{x}+2s_x$	75
2.5	0.81	$\bar{x}-2.5s_x<x<\bar{x}+2.5s_x$	81
3.0	0.88	$\bar{x}-3s_x<x<\bar{x}+3s_x$	88

Como puede observarse, el resultado para el primer intervalo no aporta información, sin embargo, para los otros cuatro valores de k se muestran los porcentajes mínimos de datos en el intervalo correspondiente. Así, a una y media desviación estándar, a la izquierda y a la derecha, a partir de la media se tendrá al menos un 55 % del total de los n datos; y en un intervalo de dos desviaciones estándar alrededor de la media como mínimo, se localizará el 75 % del total de los n datos.

Ejemplo 3.29. Los siguientes datos pertenecen a 20 sondeos en una zona agrícola en el centro del país, que registran la profundidad del nivel freático:

Tabla 56. Sondeo zona agrícola

38.1	62.4	66.7	36.3	48.0
49.3	41.3	37.4	27.1	39.4
51.7	54.1	39.2	50.6	47.2
48.9	46.8	50.3	37.4	42.2

A partir de esta información, determinar el mínimo de datos que se pueden tener de acuerdo con la tabla 51 el Teorema de Tchebycheff en los intervalos que se forman para los siguientes valores de k= 1, 1.5. 2, 2.5 y 3.

Solución:

Tabla 57. Zona agrícola con Tchebycheff

k	Intervalo $\bar{x} - k s_x < x < \bar{x} + k s_x$	Frecuencia del intervalo	Frecuencia relativa	Porcentaje mínimo de datos en el intervalo, según el Teorema de Tchebycheff
1.0	36.50 - 54.89	16	0.80	0 %, al menos cero datos
1.5	31.96 - 59.47	17	0.85	55 %, al menos 11.1 datos
2.0	27.38 - 64.06	18	0.90	75 %, al menos 15 datos
2.5	27.79 - 68.64	19	0.95	84 %, al menos 16.8 datos
3.0	18.21 - 59.47	20	1.00	88 %, al menos 17.7 datos

De los resultados que se muestran en la tabla 53, en la tercera columna se tiene el número de datos reales (frecuencia) y en la cuarta columna la proporción que representan en toda la muestra (frecuencia relativa). Por lo que se observa que, la aplicación empírica es útil para conocer en una primera aproximación la acumulación de los datos en intervalos predeterminados.

El Teorema de Tchebycheff y la Regla Empírica (o de la probabilidad normal) pueden apoyar para entender los principios básicos de los métodos de estimación, en temas posteriores de la estadística inferencial.

3.9.2 DIAGRAMA DE CAJA

Para que un experimento arroje información valiosa, debe siempre obtener observaciones (datos) en igualdad de condiciones en cada prueba, no obstante, los resultados suelen ser diferentes; aunque se pueden repetir. Además de esto, se pueden tener dos categorías diferentes de resultados, los que están dentro de ciertos límites determinados alrededor de la mediana que se denominan ***datos típicos*** y algunos extremos demasiado alejados de la media, muy pequeños o grandes que se les conoce como datos atípicos.

Un diagrama de caja puede permitir visualizar las características típicas y atípicas de un conjunto de datos, producto de las observaciones resultantes de un determinado experimento; pero además nos proporcionará elementos para conocer aproximadamente los valores de tendencia central, dispersión y sesgo.

La construcción de un diagrama de caja se logra con cinco valores: el primer cuartil , el segundo cuartil o mediana , el tercer cuartil , y los datos extremos, el más pequeño front. inf y el más grande front. sup del conjunto de las observaciones resultantes.

El trazo de un diagrama de caja se realiza considerando lo siguiente:

- Identificar en el conjunto los datos extremos front. Inf y front. Sup (menor y mayor respectivamente); así como los primeros tres cuartiles del conjunto de datos en estudio.
- Trazar una recta horizontal, a una escala conveniente, que abarque mínimamente el rango de los datos.
- Por arriba de la recta horizontal a escala, dibujar una caja delimitada por los cuartiles primero y tercero; esto muestra el recorrido intercuartil $\boldsymbol{RI}_q$. Dentro de la caja trazar también una recta vertical que pase por el segundo cuartil (mediana x_{Q2} o $\tilde{x}$).
- Para detectar resultados atípicos de las pruebas realizadas en el experimento, deben establecerse los límites internos inferior y superior. La mayor parte de los autores coinciden en que deben fijarse a una distancia 1.5 unidades del recorrido intercuartil,
- Límite interno inferior: $x_{Q1} - 1.5\ (RIq)$
- Límite interno superior: $x_{Q3} + 1.5\ (RIq)$
- Aquellos datos menores al límite interno inferior o mayores al límite interno superior, se consideran medidas atípicas.

Los datos atípicos se marcan con un asterisco en el contexto del diagrama de caja, como puede verse a continuación:

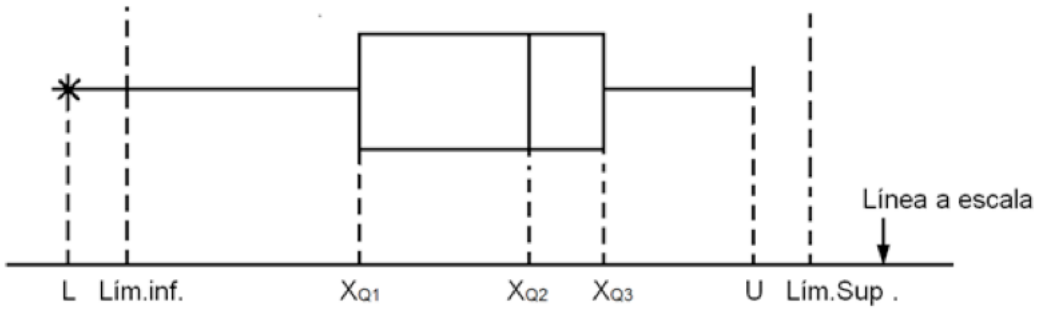


FIGURA 47. Diagrama de caja y bigotes

Los llamados bigotes son las líneas que van desde la caja, se extienden hasta los valores máximo y mínimo de la serie o hasta 1,5 veces el recorrido intercuartil.

Finalmente, en un diagrama de caja se pueden establecer los **límites extremos** de la siguiente forma:

Los límites extremos son los valores más pequeños y los valores más grandes de los datos.

$$\textit{límite extremo inferior} = x_{Q1} - 2(1.5)(RIq)$$
$$\textit{límite extremo superior} = x_{Q3} + 2(1.5)(RIq)$$

Ejemplo 3.30. Un productor que cultiva miles de árboles de navidad toma una muestra aleatoria de diez de ellos, listos para comercializarse por la temporada, con la finalidad de obtener conclusiones sobre el comportamiento de la altura de este producto natural. Esta variable influye en relación directa al precio de venta, a mayor altura del árbol mayor es su precio. Elaborar un diagrama de caja, y plantear tres conclusiones relevantes que se evidencien en la mencionada muestra, que consta de los siguientes datos 226, 231, 136, 198, 202, 214, 228, 199, 221, 216 [cm].

Solución:

Primero se ordenan los datos de manera ascendente:

Tabla 58. Datos ordenados

Altura del árbol
136
198
199
202
214
216
221
226
228
231

Cálculo de los cuartiles:

La ubicación de x_{Q1}, x_{Q2} y x_{Q3}, es:

0.25 (n+1) = 0.25 (10+1) = 2.75 valor correspondiente de la tabla 198
0.50 (n+1) = 0.50 (10+1) = 5.5 valor correspondiente de la tabla 214
0.75 (n+1) = 0.75 (10+1) = 8.25 valor correspondiente de la tabla 226

Por lo tanto:

Q_1, = a_I + fracción de la posición ($a_S - a_I$)
$x_{Q1} = 198 + 0.75\ (1.0) = 198.75$
$x_{Q2} = 214 + 0.50\ (2.0) = 215$
$x_{Q3} = 226 + 0.25\ (2.0) = 226.5$

Por lo tanto: Riq = 226.5 − 198.75 = 27.75

Límite interno inferior = $x_{Q1} - 1.5(RIq) = 198.25 - 1.5(27.75) = 156.325$
Límite interno superior = $x_{Q3} + 1.5(RIq) = 226.5 + 1.5(27.75) = 268.125$

3.10 EJERCICIOS RESUELTOS

La edad de treinta profesores de una cierta universidad, que fueron mejor evaluados por sus alumnos, se presenta en la siguiente tabla:

Tabla 59. Edad de los profesores

63	67	56	33	51	53
59	48	58	56	35	39
48	74	49	55	72	56
54	58	60	59	44	52
61	57	52	21	36	53

a) A partir de una distribución de frecuencias con datos agrupados, determinar los tres cuartiles y el recorrido intercuartil.
b) Calcular la media, la mediana, la moda, la desviación estándar, desviación media y desviación mediana.
c) Construir un diagrama de caja.

Solución:

a) Rango = dato mayor – dato menor = 74 – 21 = 53

$$c = 3.3 * \log_{10}(30) + 1 = 5.87 \approx 6.0$$

$$w = \frac{R}{c} = \frac{53}{60} = 8.83 \approx 9.0$$

TABLA 60. Frecuencias para edad de los profesores

i	Intervalos de clase	Fronteras	Marca (x_i)	Frecuencia (f_i)	Frecuencia acumulada (F_i)	Frecuencia relativa (f_i *)	Frecuencia relativa acumulada (F_i *)
1	21 – 29	20.5 – 29.5	25	1	1	0.0333	0.0333
2	30 – 38	29.5 – 38.5	34	3	4	0.1000	0.1333
3	39 – 47	38.5 – 47.5	43	2	6	0.0667	0.2000
4	48 – 56	47.5 – 56.5	52	13	19	0.4333	0.6333
5	57 – 65	56.5 – 65.5	61	8	27	0.2667	0.9000
6	66 – 74	65.5 – 74.5	70	3	30	0.1000	1.0000
Suma				30			

$$x_{Q1} = Front\ inf_{(Q1)} + w\left(\frac{k}{frec_{Q1}}\right) = 20.5 + 9\left(\frac{0.9}{1}\right) = 28.6$$

$$x_{Q2} = Front\ inf_{(Q2)} + w\left(\frac{k}{frec_{Q2}}\right) = 47.5 + 9\left(\frac{1.5}{13}\right) = 48.53$$

$$x_{Q3} = Front\ inf_{(Q3)} + w\left(\frac{k}{frec_{Q3}}\right) = 56.5 + 9\left(\frac{3.5}{8}\right) = 60.44$$

$$Riq = x_{Q3} - x_{Q1} = 11.90$$

Entre los llamados ***límites internos,*** tanto inferior: $LI_{(i)}$, como superior $LI_{(s)}$, se ubican los ***datos típicos*** o más predecibles. Fuera de estos límites se consideran los denominados datos atípicos considerados aquellos que son mayores que Q_3 en, por lo menos, 1.5 veces el rango intercuartil y menores que Q_1 en, al menos, 1.5 veces el rango intercuartil, los cuales se indican en el diagrama con un círculo vacío. Los límites internos se calculan de la siguiente manera:

Limite interno inferior: 48.53 – 1.5(11.90)=38.68

Limite interno superior: 60.43 + 1.5 (11.90)=78.28

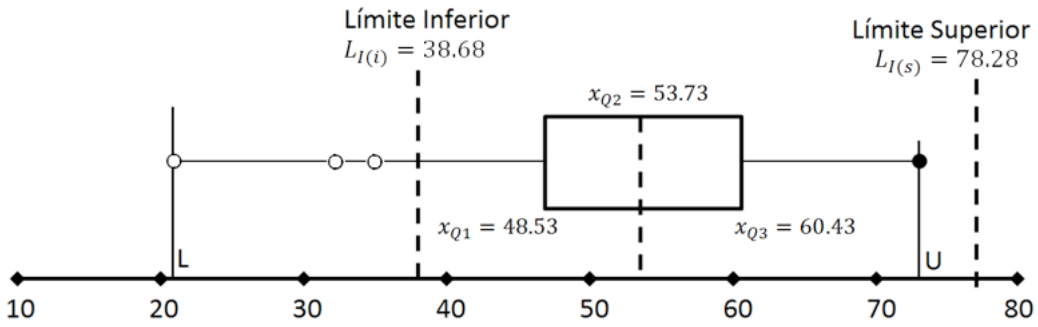


Figura 48. Diagrama de caja y bigotes para las edades de profesores

Finalmente, en un diagrama de caja se pueden establecer los ***límites extremos*** de la siguiente forma:

$$límite\ extremo\ inferior = 48.53 - 2(1.5)(11.90) = 19.97$$
$$límite\ extremo\ superior = 60.43 + 2(1.5)(11.90) = 96.13$$

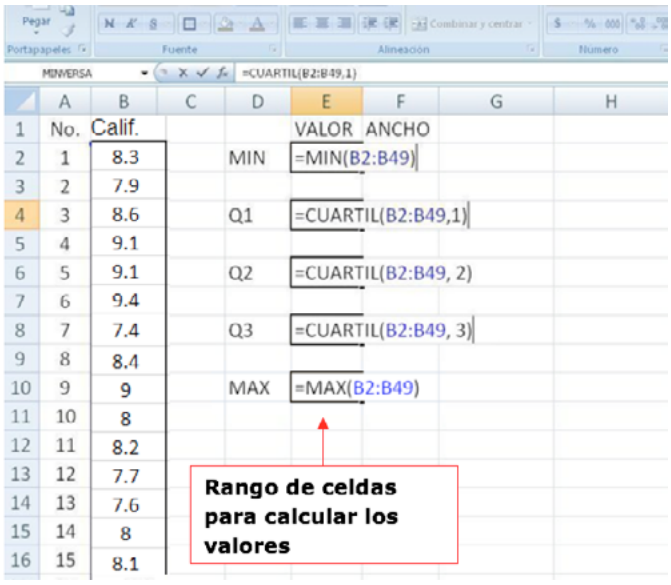
Como puede apreciarse, ningún valor es más pequeño que el límite extremo inferior, ni más grande que el límite extremo superior. Si en algún caso se llegara a tener un valor con esas características correspondería a los llamados ***valores atípicos extremos*** y se deben indicar con un asterisco en el diagrama de caja.

Es importante señalar que los ***valores atípicos*** provienen de dos causas principales. La primera corresponde a un resultado inusual, pero legítimo, esto es, la medición es correcta y debe tomarse en cuenta como un dato más en el cálculo de medidas características como la media o la varianza. La segunda causa es cuando se han cometido errores al realizar las mediciones y suelen llamarse datos aberrantes, por lo tanto, el resultado sale de los rangos límites internos y externos; en consecuencia, estos valores producto de error deben desecharse para el cálculo de medidas características, y así evitar distorsiones e imprecisiones en el análisis general de la información proporcionada por los datos.

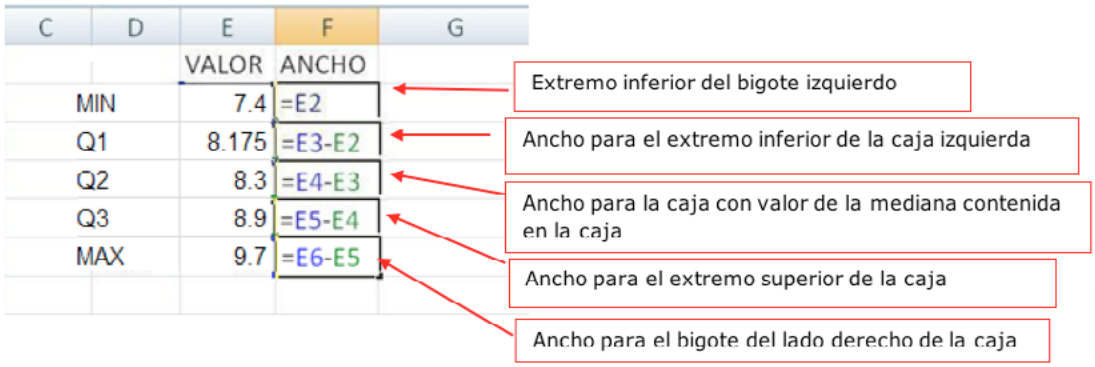
3.11 EJERCICIOS DE APLICACIÓN CON SOFTWARE ESPECIALIZADO

Utilizando Excel se puede construir un diagrama de caja para una muestra de 80 datos con las calificaciones de alumnos del tercer semestre de la carrera de Ingeniería. Para ello, se requieren el valor mínimo, el valor máximo y los valores de los tres primeros cuartiles.

Primero se obtiene la tabla con los datos:



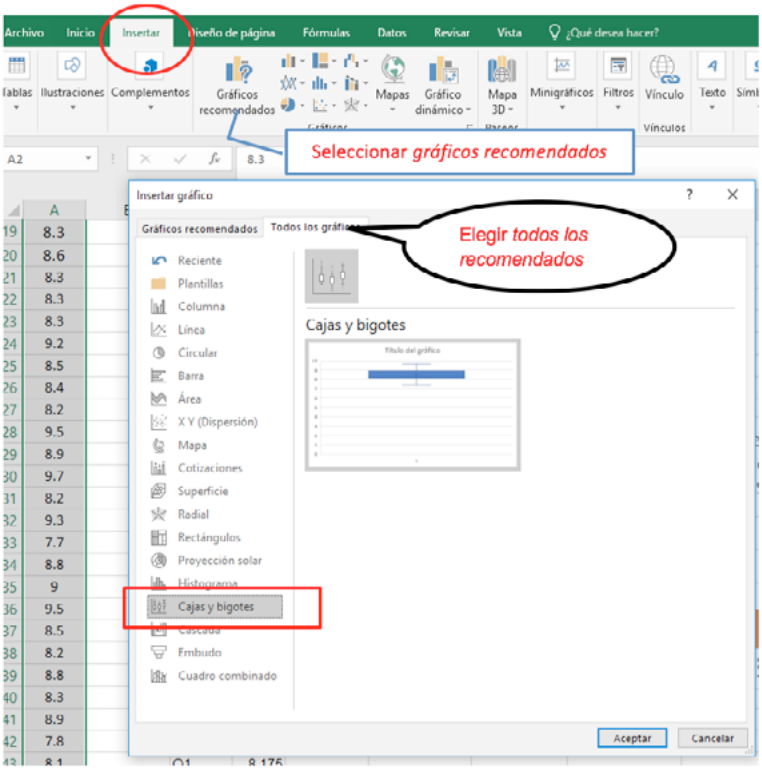
La tabla con los valores de los cuartiles queda como se muestra enseguida:

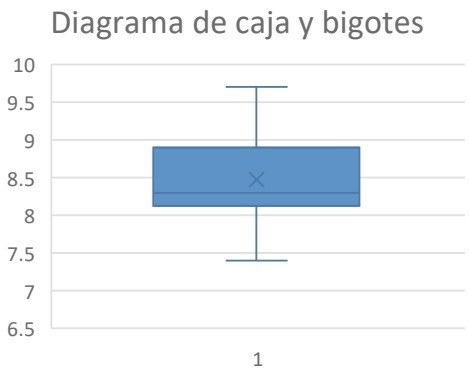


Para elaborar el diagrama de caja y bigotes, se recurre a las funciones gráficas ya programadas en Excel,

Se seleccionan primero todos los datos no agrupados.

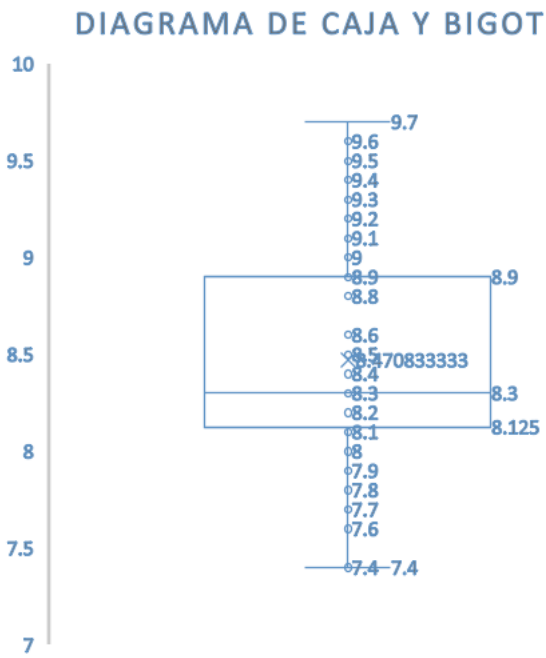
Se recurre al menú ***Insertar*** y allí se selecciona ***Gráficos recomendados,*** en ese submenú se elige ***Todos los recomendados*** y, finalmente, se selecciona el diagrama de ***Cajas y bigotes***, el gráfico aparecerá de inmediato.





En la gráfica se pueden agregar los valores de la serie mediante:

1. Dar doble clic en la figura, con botón derecho se mostrará una ventana de *formato de serie de datos*.
2. Aparecerá a la derecha una ventana con opciones, hacer clic en *mostrar puntos interiores* y *marcadores de la media*.
3. Dar clic en cualquier punto de los mostrados con botón derecho, aparecerá una ventana con un menú para colocar las *etiquetas de datos*.
4. Se pueden observar los valores mínimo, máximo, Q1, Q2, Q3 entre otros.



3.12 EJERCICIOS PROPUESTOS No. 3

1. Un grupo de 100 alumnos es seleccionado para jugar en el equipo de básquetbol de la universidad. Por cuestiones de carácter estratégico deportivo, el entrenador pide a su auxiliar que mida las estaturas de los seleccionados, con precisión de décimas de centímetro y con esa información:

   a) Establecer una tabla de distribución de frecuencias
      Con base en la tabla de distribución de frecuencias obtener:
   b) La mediana ($\tilde{x}$) o percentil 50
   c) El sesentavo percentil
   d) El treintavo percentil
   e) ¿Qué porcentaje mide entre 159.5 y 189.5 cm de estatura?
   f) ¿Qué porcentaje de seleccionados mide menos de 169.5 y más de 189.5 cm?
   g) ¿Qué valor corresponde al percentil 89?

150.8	175.1	194.8	190.6	155.2	162.8	190.9	176.8	152	159.8
171.9	153	192	188.7	186.1	192.1	180.7	194.4	172.6	160.8
171.2	168.9	169.7	189.8	152.6	184.1	156.6	172.2	175.3	194.3
161.4	187.6	195	165	176.2	176.9	171	192.6	173.3	160.1
174.3	160.7	170.5	183.9	161.3	191.4	193.2	158.6	188.4	160
166.8	192.7	183.2	179.3	187.6	194.7	180	188.8	158.3	155.5
175.1	151.8	171.2	156.6	189.2	193.6	186	160.1	199.2	197.3
179.3	157.3	174.4	150.6	191.6	198.2	163.5	194.5	159.4	170.3
195.2	173.6	152	167.1	193.6	170.9	186.5	169.2	151.2	180
189.5	156.8	165	167.4	191.9	198.2	151.6	168.8	171.7	168.9

2. Por las obras de un tramo del Sistema de Transporte Metropolitano (Metro) de la Ciudad de México, se midieron las velocidades (km/h) de los automóviles que circularon a las 19:00 h, por dos cruceros A y B contiguos a este lugar. La siguiente tabla proporciona la información de algunos percentiles de los dos cruceros.

Tabla 61. Percentiles

Percentil	Crucero	
	A	B
Dieciocho ($x_{0.18}$)	25.9	23.7
Cincuenta ($x_{0.50}$)	36.5	28.4
Ochenta ($x_{0.80}$)	42.0	39.5

a) ¿Qué signo debe tener el coeficiente de sesgo (asimetría), en cada crucero?
b) El recorrido intercuartil es mayor, igual o menor a 16.1.

3. Se observa el flujo vehicular por hora de ocho casetas de cobro de la entrada de una autopista en el estado de Querétaro, entre las 7:30 a 8:30 hrs de un día determinado, en un periodo vacacional.

Tabla 62. Flujo vehicular

Número de caseta	1	2	3	4	5	6	7	8
Flujo Vehicular por hora	63	87	116	128	135	108	96	81

a) Elabore un diagrama de caja.
b) Si el orden de las casetas en la tabla es el mismo que éstas tienen físicamente en la entrada de la autopista, plantea dos conclusiones sobre el funcionamiento del conjunto de casetas de cobro.

4. La Secretaría de Comercio revisó 25 gasolineras de la Ciudad de México, previamente seleccionadas, para conocer la cantidad de mililitros que despacha una de las bombas de cada uno de esos expendios de combustible para automóviles; los resultados se muestran en la tabla:

Tabla 63. Datos de cantidad de mililitros despachados

845	773	705	873	888
909	870	914	916	923
924	897	927	935	936
938	901	951	952	954
953	908	1010	982	954

a) Elaborar un diagrama de caja.
b) Plantear al menos dos conclusiones que en este estudio pueden ser relevantes.

SOLUCIÓN A LOS EJERCICIOS PROPUESTOS

EJERCICIOS PROPUESTOS No. 1

1.

Intervalo	Marcas de clase	Frecuencias	Frecuencias relativas	Frecuencias acumuladas	Frecuencias acumuladas relativas
1	8.66	1	0.033	1	0.033
2	8.79	2	0.067	3	0.1
3	8.92	5	0.167	8	0.267
4	9.05	7	0.233	15	0.5
5	9.18	6	0.2	21	0.7
6	9.31	8	0.267	29	0.967
7	9.44	1	0.033	30	1

2.

Intervalo	Marcas de clase	Frecuencias	Frecuencias relativas	Frecuencias acumuladas	Frecuencias acumuladas relativas
1	8.66	1	0.033	1	0.033
2	8.79	2	0.067	3	0.1
3	8.92	5	0.167	8	0.267
4	9.05	7	0.233	15	0.5
5	9.18	6	0.2	21	0.7
6	9.31	8	0.267	29	0.967
7	9.44	1	0.033	30	1

3.

Intervalo	Frontera		Marcas de clase	Frecuencias	Frecuencias relativas	Frecuencias acumuladas	Frecuencias acumuladas relativas
	inferior	superior					
1	43.425	47.825	45.625	7	0.19	7	0.14
2	47.825	52.225	50.025	7	0.19	14	0.28
3	52.225	56.625	54.425	7	0.19	21	0.42
4	56.625	61.025	58.825	17	0.47	38	0.76
5	61.025	65.425	63.225	5	0.14	43	0.86
6	65.425	69.825	67.625	4	0.11	47	0.94
7	69.825	74.2	72.025	3	0.08	50	1
				50			

EJERCICIOS PROPUESTOS No. 2

1.

b) 1.6475
c) 1.73
d) 1.68
f) 1.695

2.

a) X_{Media}= 22.83; X_{Moda}= 25.5
b) $X_{mediana}$= 23.0

EJERCICIOS PROPUESTOS No. 3

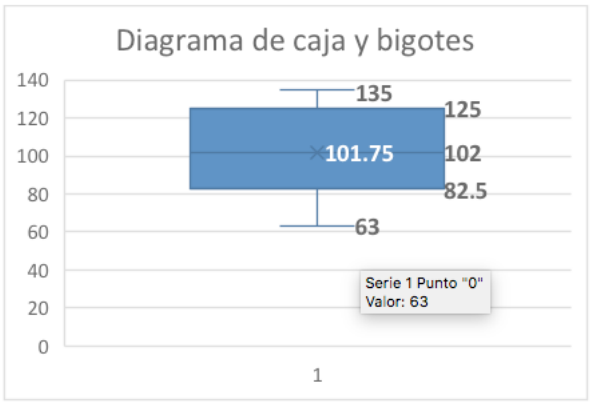
1.

	Clase	front_inf	front_sup	*Xi*	Frecuencia	*fi**	*Fi*	*Fi**
	1	150.55	155.45	153	10	0.1	10	0.1
	2	155.45	160.35	157.9	12	0.12	22	0.22
30° porciento	3	160.35	165.25	162.8	8	0.08	30	0.3
mediana	4	165.25	170.15	167.7	8	0.08	38	0.38
	5	170.15	175.05	172.6	14	0.14	52	0.52
60° porciento	6	175.05	179.95	177.5	8	0.08	60	0.6
	7	179.95	184.85	182.4	6	0.06	66	0.66
	8	184.85	189.75	187.3	10	0.1	76	0.76
	9	189.75	194.65	192.2	16	0.16	92	0.92
	10	194.65	199.55	197.1	8	0.08	100	1

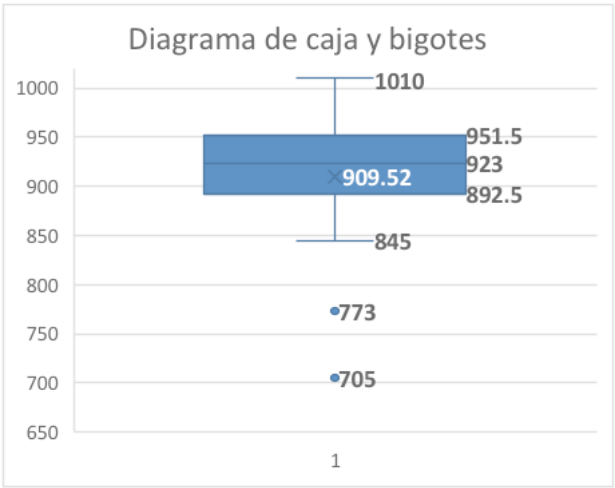
2.

a) Negativo

3.



4.



BIBLIOGRAFÍA

AGUILAR M., A. (2010). *Introducción a la inferencia estadística,* México: Pearson, 210 pp.

BENNET, Jeffrey O. (2011). R*azonamiento estadístico,* México: Pearson Educación, 494 pp.

DEVORE, J. (2015). *Probabilidad y la estadística para ingeniería y ciencias*, 8ª ed., México: Cengage Learning, 776 pp.

HINES, W., Montgomery, D., Goldsman, D., Borror, C. (2009). *Probabilidad y estadística para ingeniería,* 3ª ed., México: Grupo Editorial Patria.

JOHNSON, Richard A. (2011). *Probabilidad y estadística para ingenieros*, 8ª ed., México: Pearson Educación, 552 pp.

MENDENHALL, W. (2010). *Introducción a la probabilidad y la estadística,* México: Cengage Learning, 776 pp.

_____, Wackerly, D., Scheaffer, R. (1990). *Estadística matemática con aplicaciones*, 2ª ed., México: Grupo Editorial Patria.

MILTON, Susan. (2003). *Probabilidad y estadística aplicada a la ingeniería y ciencias de la computación*, 4ª ed., México: Mc Graw Hill, 804 pp.

MONTGOMERY, Douglas C. (2008). *Probabilidad y estadística aplicada a la ingeniería*, 4ª ed., México: Mc Graw Hill, 816 pp.

_____. (2008). *Diseño y análisis de experimentos,* México: Limusa, 589 pp.

SPIEGEL, M. (1970). *Estadística,* México: McGraw Hill.

WACKERLY, Dennis D. (2010). *Estadística matemática con aplicaciones,* 7ª ed., México: Cengage Learning, 991 pp.

WALPOLE, Ronald. (2012). *Probabilidad y estadística para ingeniería y ciencias*, 9ª ed., México: Pearson Educación, 816 pp.

WEIMER, Richard C. (1996). *Estadística*. México: CECSA, 838 pp.

ÍNDICE ANALÍTICO

udae
PUBLICACIONES FI

ESTADÍSTICA DESCRIPTIVA
se publicó digitalmente en el repositorio de
la Facultad de Ingeniería
en octubre de 2024. Primera edición
electrónica de un ejemplar
(12 MB) en formato PDF.

El cuidado de la edición y diseño
estuvieron a cargo de la Unidad de Apoyo
Editorial de la Facultad de Ingeniería.
La familia tipográfica utilizada fue
Source Serif 4 para títulos y textos con sus
respectivas variantes.