Introducción a la Estadística Descriptiva

Javier Alberto Mesa Guerrero
Segundo Javier Caicedo Zambrano

Editorial
Universidad de Nariño

Editorial
Universidad de **Nariño**

Introducción a la Estadística Descriptiva

Introducción a la Estadística Descriptiva

Javier Alberto Mesa Guerrero

Segundo Javier Caicedo Zambrano

Editorial
Universidad de Nariño

Mesa Guerrero, Javier Alberto

Introducción a la estadística descriptiva / Javier Alberto Mesa Guerrero, Segundo Javier Caicedo Zambrano. San Juan de Pasto : Editorial Universidad de Nariño, 2020

123 p.

Incluye bibliografía

ISBN: 978-958-5123-11-3 digital

1. Estadística descriptiva 2. Variables (estadística) 3. Estadística – problemas, ejercicios, etc. 4. Estadística - enseñanza I. Caicedo Zambrano, Segundo Javier

519.53 M578 – SCDD-Ed. 22 Biblioteca Alberto Quijano Guerrero

Introducción a la Estadística Descriptiva

Autores: Javier Alberto Mesa G.

Segundo Javier Caicedo Z.

ISBN: 978-958-5123-11-3

Diagramación: Segundo Javier Caicedo Zambrano

Diana Sofía Salas Chalapud

Universitaria de la Universidad de Nariño.

San Juan de Pasto – Nariño - Colombia

índice de Contenido

CAPÍTULO 2.
MEDIDAS ESTADÍSTICAS..... 30

INTRODUCCIÓN

Esta obra surge por el interés de los autores de publicar un libro de texto a nivel introductorio sobre fundamentos de estadística descriptiva, que se constituya en fuente de consulta y de nivelación sobre conceptos básicos de estadística. Por el enfoque, está orientado a estudiantes de los primeros semestres de programas universitarios relacionados con las Ciencias Básicas, Técnicas e Ingenierías, aunque también lo pueden utilizar estudiantes de otras áreas del conocimiento, incluso de programas de educación no formal.

Los temas que se abordan en esta obra, han sido seleccionados con base en la experiencia de los autores, quienes reconocen que se constituye en soporte importante para estudiantes que inician el estudio de la estadística. Para el efecto, se presenta, en forma resumida, la conceptualización, ejemplos y se proponen ejercicios para reforzar lo aprendido. Si bien existen muchos programas para el cálculo de estadísticas, los autores consideran que es importante que los estudiantes realicen los cálculos paso a paso, tal como se ilustra en los ejemplos, porque ayuda para la comprensión e interpretación de los resultados.

El libro está organizado en cuatro (4) capítulos. El primero, “Conceptos Básicos”, presenta generalidades de estadística y organización de datos en tablas de frecuencia. En el segundo capítulo, “Medidas estadísticas”, se trabaja las medidas de tendencia central, medidas de posición, medidas de variabilidad, momentos, relación y correlación simple, y análisis de series cronológicas. En el tercer capítulo, “Regresión y correlación”, se incluye coeficiente de correlación, regresión lineal, regresión no lineal: función potencial, función exponencial y función cuadrática. En el cuarto capítulo, “Series cronológicas”, se aborda el análisis de estas series.

Se sugiere que el estudio del libro se realice en forma secuencial y se desarrollen todos los talleres que se proponen, con lo cual, el estudiante avanzará seguro y tendrá la posibilidad de finalizar su estudio con una sólida comprensión de los conceptos básicos de estadística descriptiva.

Los autores, Marzo 2020

CAPÍTULO 1.
CONCEPTOS BÁSICOS

CAPÍTULO 1.
CONCEPTOS BÁSICOS

1.1 GENERALIDADES

En términos generales, se considera que la finalidad de la Estadística es suministrar información acerca de un determinado hecho o fenómeno; su utilidad depende, en gran parte, de los fines que se propone y de la forma cómo se obtienen los datos.

Por medio de la Estadística se puede lograr los siguientes propósitos:

Conocer la realidad de una observación. Si mediante la investigación se logra CUANTIFICAR un fenómeno, se conoce la situación real o una estimación del mismo.

Determinar lo típico o normal de las observaciones. Cuando se cuantifican las características de un fenómeno, se está determinando el comportamiento general del grupo, el grado de uniformidad o variabilidad, la asimetría y el tipo de distribución de la variable.

Estimar y proyectar el comportamiento futuro de un hecho observado.

Determinar las causas que originan un fenómeno.

Comprobar hipótesis planteadas en una investigación.

Cruzar dos o más variables con el fin de conocer el grado de relación existente entre ellas y determinar la ecuación que las relaciona.

Hacer inferencias basándose en resultados muestrales y en las leyes del Cálculo de Probabilidades. Teniendo en cuenta que la Estadística estudia los fenómenos colectivos, es necesario conocer ciertas técnicas que permitan agrupar la información con el fin de facilitar el procesamiento, presentación, análisis y publicación de resultados.

Dependiendo de si el estudio se realiza con base en una muestra o en una población, la Estadística se clasifique en Descriptiva e Inferencial, respectivamente.

1.1.1 Estadística Descriptiva

La **Estadística Descriptiva** comprende la recolección, organización, presentación, análisis y publicación de los resultados observados. Su finalidad es describir las características principales de una muestra, lo cual se puede realizar mediante cuadros, gráficos o índices.

1.1.2 Estadística Inferencial

La **Estadística Inferencial** se apoya en el Cálculo de Probabilidades y usa los resultados de la estadística descriptiva con el fin de generalizar y aplicar los conceptos a la población.

1.1.3 Variables

Las **variables** son las características de la muestra o población que se está estudiando y los datos son los valores de las variables; corresponden a los resultados de la medición.

Ejemplo:

La variable salarios de los trabajadores de una empresa, se mide en pesos.

1.1.4 Escalas de medición

De acuerdo a la clasificación de Steven, las variables pueden ser:

A) Nominales.

B) Ordinales.

C) De Intervalo.

D) De Razón.

Escala nominal: *u*na variable está medida en escala nominal cuando asigna nombres a la característica.

Ejemplo:

Género, estado civil, profesión.

Escala ordinal: se asigna nombres a la característica y además estos representan un orden de menor a mayor o viceversa.

Ejemplo:

La opinión de los estudiantes acerca de la administración de la universidad, puede ser: excelente, buena, regular, mala; el estrato social puede ser bajo, medio, alto.

Escala de intervalo: se presenta cuando se toman medidas numéricas sobre algunos elementos y se puede determinar con exactitud los intervalos entre esas medidas. La distancia entre números sucesivos es de tamaño constante y medible. Los datos medidos en escala de intervalos tienen un punto cero arbitrario. Es decir, la persona que diseña la escala de manera arbitraria decide donde poner el punto cero.

Ejemplo:

Frio/Caliente es una escala ordinal. Pero si medimos la temperatura en grados 65°F/70°F es una escala de intervalo. Por ejemplo, un día con 60 grados de temperatura no es el doble de caliente que otro día con 30 grados. Para los datos de intervalos la razón de números no es apropiada. El valor cero no indica ausencia de temperatura. Los índices de precios al consumidor pueden tomar como año base (0) de acuerdo al criterio del investigador.

La escala de razón: consiste en medidas numéricas, para las cuales, las distancias entre los números tienen tamaño constante y conocido, y donde la razón entre los números tiene significado; además, existe un `punto cero (0) fijo y no arbitrario.

Ejemplo:

La vida útil de un artículo en días. La diferencia entre 500 y 250 días indica que la duración del primero es el doble del segundo,

1.1.5 Población

El término población se refiere al conjunto de todas las observaciones posibles en una situación dada. Cuando el conjunto es muy numeroso o infinito, es imposible incluir todos los elementos en el estudio, siendo necesario tomar una muestra representativa para la investigación. Una muestra es representativa si cumple las siguientes condiciones:

- Está formada por la diversidad de elementos existentes en la población.
- El tamaño es adecuado según la variable en estudio.

Cuando se analiza las variables en una población se obtiene ciertas características llamadas *parámetros* o *valores verdaderos*. Los parámetros son constantes y para representarlos, generalmente, se utiliza letras del alfabeto griego; en cambio, esas características en una muestra son variables y se denominan *estadígrafos o estadísticas* y se representan con letras del alfabeto latino.

Se define la estadística como "La tecnología del método científico"; por lo cual, constituye una herramienta muy importante en la investigación.

1.1.6 Etapas de la investigación estadística

Una **investigación estadística** es un proceso dinámico mediante el cual se observa un fenómeno, se diseña un experimento, se recolecta datos para probar una hipótesis y se analizan y obtienen conclusiones útiles para la población objeto de estudio.

La investigación estadística se desarrolla en tres etapas: Planeación, Ejecución y Evaluación.

1.1.6.1 Primera etapa: planeación

Como producto de esta etapa, se debe elaborar un documento que contenga los siguientes aspectos:

- Formulación del problema.
- Fijación de objetivos, marco teórico y justificación.
- Planteamiento de hipótesis.
- Revisión bibliográfica.
- Metodología de la investigación (definiciones de población, muestra, unidad, diseño muestral, diseño de instrumentos, planes de recolección, recursos, etc.).

- Cronograma de actividades.
- Presupuesto

1.1.6.2 Segunda etapa: ejecución

Se deben desarrollar las siguientes actividades:

- Recolección de información (Trabajo de campo).
- Revisión o crítica.
- Codificación.
- Procesamiento de la información (Elaborar una base de datos, tablas de frecuencia y gráficos).
- Análisis, conclusiones y recomendaciones.
- Presentación y publicación.

1.1.6.3 Tercera etapa: evaluación

En esta etapa se deben desarrollar las siguientes actividades:

- Verificar el cumplimiento de los objetivos.
- Determinar la utilidad de los resultados en la sociedad.

1.2 ORGANIZACIÓN DE DATOS

Debido a que los datos en forma individual no tienen utilidad práctica, es necesario organizarlos de manera sistemática para facilitar su interpretación y análisis. La manera más sencilla de organizar la información es en una base de datos, la cual posibilita realizar clasificaciones simples o cruzadas, según las variables o series.

1.2.1 Serie estadística

Es una colección de datos estadísticos, clasificados y ordenados por un determinado criterio. Las series se clasifican en: atemporales y temporales.

1.2.1.1 Atemporales

Los datos se clasifican de acuerdo a una cualidad o magnitud, en un espacio determinado y en tiempo constante; pueden ser cualitativas o cuantitativas. Las series cualitativas por sí solas no tiene mérito. Para realizar un análisis descriptivo profundo se puede hacer una representación tabular, gráfica y cruces de variables; para analizar la relación, se puede utilizar diagramas de barras o circulares. Las series atemporales tienen aplicación en la estadística inferencial; por ejemplo, para el análisis de encuestas de opinión, encuestas electorales, entre otros.

En las series cuantitativas los datos se clasifican de acuerdo a una magnitud o medición; según la naturaleza de la variable pueden ser discretas o continuas. Es en este tipo de series es donde se puede aplicar la mayoría de técnicas estadísticas, haciendo un análisis individual de cada una de las variables o cruzando dos o más de ellas, con el fin de realizar un estudio conjunto de las mismas.

1.2.1.2 Temporales o cronológicas

Las series cronológicas se refieren a datos de un mismo tema, en un espacio determinado y en épocas diferentes; por ejemplo, el número de alumnos egresados de la Universidad de Nariño en los últimos 5 años, organizados por programas.

Ejemplo:

Suponga que después de recolectar la información para un estudio acerca del rendimiento académico y las carreras preferidas por 30 estudiantes de los colegios de la ciudad de Pasto, se obtuvo la siguiente información:

Tabla 1. Ejemplo de Rendimiento académico y selección de carrera de un grupo de estudiantes

No	Género	Edad	Faltas	Nota	ICFES	Carrera	Jornada	Modalidad
01	M	17	0	7.8	320	Ing. Industrial	D	PR
02	M	18	1	8.0	300	C. Pública	D	PR
03	M	18	2	8.4	268	Derecho	D	PR
04	M	19	3	9.0	289	Derecho	N	SE
05	F	20	3	9.2	245	Agronomía	D	PR
06	F	18	3	4.6	250	Ing. Sistemas	N	SE
07	M	18	3	3.5	290	Comunicaciones	D	SE
08	M	17	3	6.8	300	Psicología	N	PR
09	F	23	5	7.0	305	Zootecnia	D	PR
10	F	22	4	7.0	304	Veterinaria	N	PR
11	F	18	4	7.0	300	Derecho	N	SE
12	F	19	3	6.8	298	Matemáticas	D	PR
13	M	18	0	7.8	310	Física	D	PR
14	M	19	0	8.1	290	Derecho	N	PR
15	M	18	2	6.7	340	Música	N	PR
16	F	17	2	7.8	320	Arquitectura	D	SE
17	F	16	3	8.4	315	Arquitectura	D	PR
18	F	17	3	7.6	287	Bacteriología	D	PR
19	F	17	3	8.6	300	Medicina	D	PR
20	M	17	3	7.3	289	Medicina	D	SE
21	M	17	4	4.5	290	Medicina	D	PR
22	F	18	5	5.0	310	Odontología	D	PR

No	Género	Edad	Faltas	Nota	ICFES	Carrera	Jornada	Modalidad
23	F	17	6	6.7	320	Derecho	D	SE
24	F	19	2	9.0	298	Música	D	PR
25	F	18	1	9.5	345	Ing. Sistemas	N	SE
26	M	17	1	8.9	325	Ing. Industrial	N	PR
27	M	17	1	8.9	324	Zootecnia	D	PR
28	F	19	0	7.8	314	Derecho	N	SE
29	F	18	0	8.2	325	Matemáticas	D	SE
30	M	17	2	9.0	316	Agronomía	D	PR

Fuente: elaboración propia con datos hipotéticos de 30 estudiantes que aspiran ingresar a la Universidad de Nariño

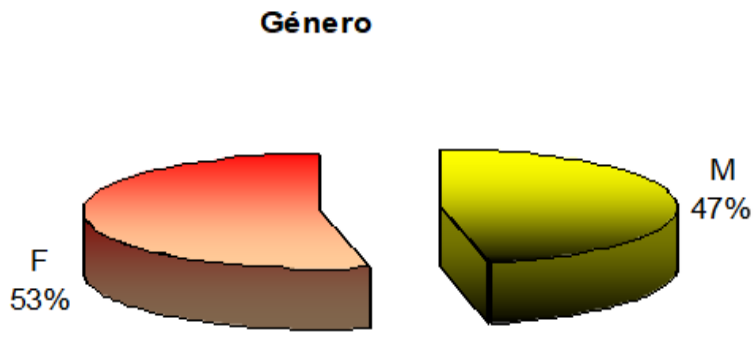
La columna Género, corresponde a una variable cualitativa; genera la tabla 2:

Tabla 2. Resumen de datos de la variable género de la tabla 1

Género	Frecuencia	Porcentaje
M	14	46.67%
F	16	53.33%
Total	30	100%

Fuente: elaboración propia con los datos obtenidos de la tabla 1

Gráfico 1. Representación gráfica de la variable género



Fuente: elaboración propia con los datos de la tabla 2

Como se puede observar en el gráfico 1, los porcentajes de las categorías de la variable son aproximadamente iguales; por lo cual, no hay diferencia significativa entre los porcentajes de la variable género.

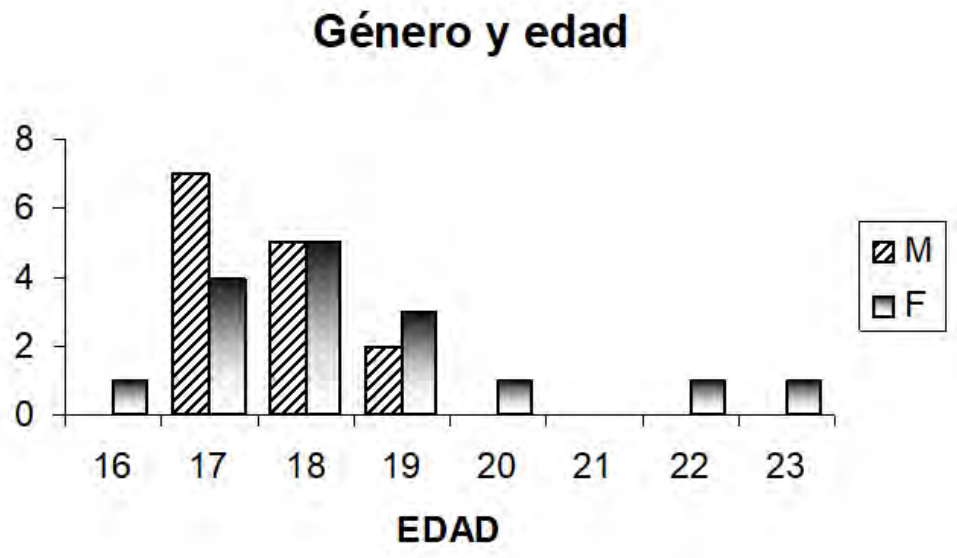
A continuación, se presenta una tabla resumen y una gráfica con los datos de edad y género de los estudiantes.

Tabla 3. Edad y género con los datos de la tabla 1

Edad	Género		Total
	M	F	
16	0	1	1
17	7	4	11
18	5	5	10
19	2	3	5
20	0	1	1
21	0	0	0
22	0	1	1
23	0	1	1
Total	14	16	30

Fuente: elaboración propia con los datos obtenidos de la tabla 1

Gráfico 2. Representación gráfica de género y edad



Fuente: elaboración propia con los con los datos de la tabla 3

Se observa en el gráfico, que la edad de la mayoría de las personas encuestadas está entre 17 y 18 años; además, que no hay hombres de 16 años ni con edades mayores de 20 años.

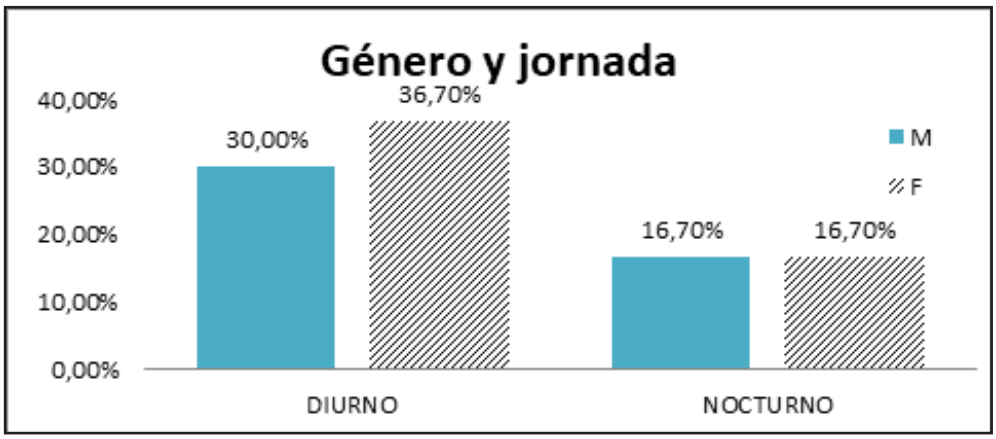
La tabla que sigue, resume los datos de género y jornada de los estudiantes.

Tabla 4. Género y jornada con datos de la tabla 1

Género	Jornada			%	Total
	Diurno	%	Nocturno		
M	9	30,00%	5	16,70%	14
F	11	36,70%	5	16,70%	16
Total	20	66,70%	10	33,30%	30

Fuente: elaboración propia con los datos obtenidos de la tabla 1

Gráfico 3. Género y jornada



Fuente: elaboración propia con los datos obtenidos de la tabla 4

1.2.2 Distribución de frecuencias

Una vez producidos o recolectados los datos de un estudio se organizan en tablas, denominadas distribuciones de frecuencia.

1.2.2.1 Distribución de frecuencias para variable discreta

Si los resultados de las observaciones toman valores enteros únicamente, se trata de variables discretas. Los resultados según el orden en que se presenten se pueden simbolizar por ; es probable que muchos datos se repitan. Para ordenar los datos en una distribución de frecuencias, se designa por al mínimo de los ; al siguiente, y así sucesivamente, hasta llegar al máximo, haciendo corresponder cada dato con el número de veces que aparece (), como se muestra en la tabla 5.

Tabla 5. Modelo de tabla de frecuencias absolutas

Valores de X_i	Frecuencia
X_1	f_1
X_2	f_2
...	...
X_m	f_m

Fuente: elaboración propia

En esta tabla, se cumple la relación que sigue, donde corresponde al tamaño de la muestra.

$$\sum f_i = n$$

Por comodidad, cuando en la sumatoria se incluyan todos los términos, se suprimirán los subíndices.

Una vez estén agrupados los datos en una distribución de frecuencias, es mucho más fácil realizar cálculos que permitirán hacer análisis porcentual y descriptivo de la variable.

Se denota por F_i la i-ésima frecuencia acumulada (ascendente) la cual se obtiene sumando las frecuencias observadas desde hasta .

$$F_i = f_1 + f_2 + \cdots + f_i$$

$$F_{i+1} = F_i + f_{i+1}$$

h_i se llama frecuencia relativa o porcentual y se denomina frecuencia relativa acumulada.

La frecuencia se obtiene dividiendo cada frecuencia observada por el tamaño de la muestra, así:

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

La frecuencia corresponde a la i-ésima frecuencia relativa acumulada (ascendente), y se obtiene así:

Tabla 6. Modelo de distribución de frecuencias absolutas y relativas

X_i	f_i	F_i	h_i	H_i
X_1	f_1	$F_1 = f_1$	$h_1 = \frac{f_1}{n}$	$H_1 = \frac{F_1}{n}$
X_2	f_2	$F_2 = F_1 + f_2$	$h_2 = \frac{f_2}{n}$	$H_2 = \frac{F_2}{n}$
⋮	⋮	⋮	⋮	⋮
X_m	f_m	$F_m = n$	$h_m = \frac{f_m}{n}$	$H_m = 1$
$\sum f_i$	n		1	

Fuente: elaboración propia

Ejemplo:

Construir una distribución de frecuencias con los datos hipotéticos de inasistencia a clases de un grupo escolar, relacionados en el cuadro 1:

Cuadro 1. Datos hipotéticos de inasistencia a clases de un grupo de estudiantes

2	3	4	0	3	3	3	1	2	3
2	1	2	3	2	3	3	5	5	4
2	4	1	6	1	6	4	0	5	3
3	3	1	1	6	0	3	3	4	4
3	2	3	2	4	4	2	5	2	1

Fuente: elaboración propia

Cada uno de los datos, Y_i , de la tabla anterior, son los resultados del control de asistencia de un grupo de estudiantes durante un año escolar. Con los datos suministrados se ha construido la distribución de frecuencias de la tabla 7.

Tabla 7. Distribución de frecuencias con los datos de inasistencia a clases de un grupo escolar representados en el cuadro 1

Faltas (X_i)	f_i	F_i	h_i	H_i
0	3	3	6%	6%
1	7	10	14%	20%
2	10	20	20%	40%
3	**15**	**35**	**30%**	**70%**
4	8	43	16%	86%
5	4	47	8%	94%
6	3	50	6%	100%
Total	50		100%	

Fuente: elaboración propia con los datos del cuadro 1

La interpretación de los datos de la cuarta fila, es como sigue:

A) 15 estudiantes faltaron 3 veces a clase.

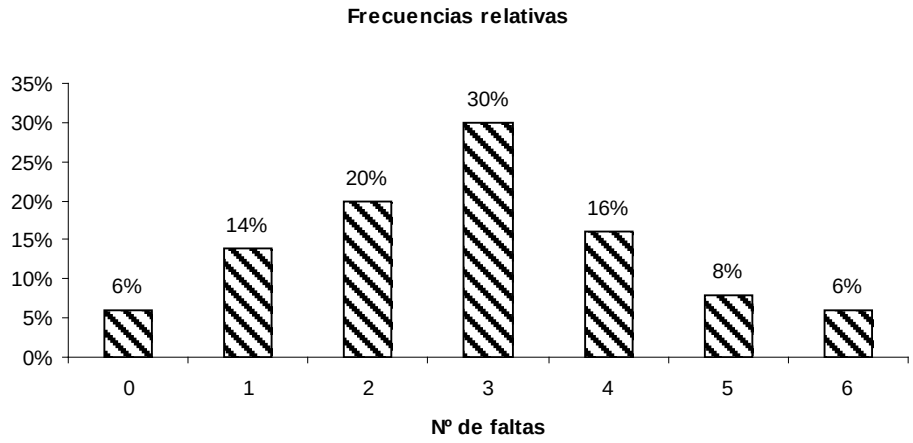
B) El 30% de los estudiantes tienen 3 faltas.

C) 35 estudiantes tienen menos de 4 faltas.

D) EL 70% de los estudiantes tienen menos de 4 faltas

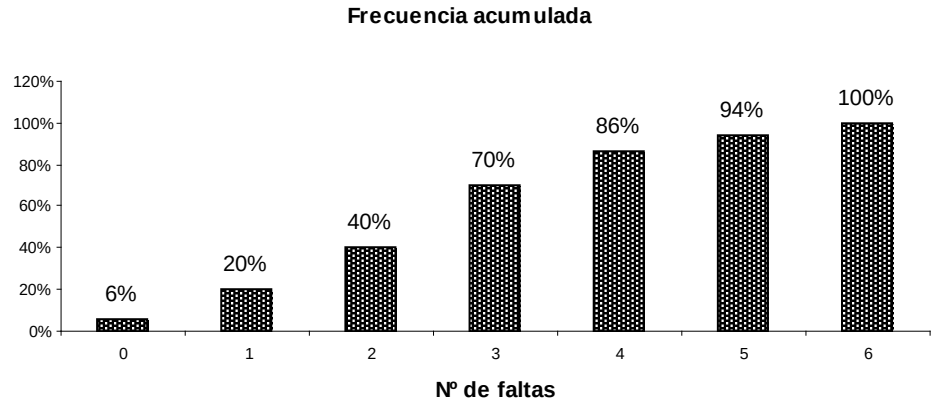
Esta información se representa en los gráficos de barras 4 y 5, que siguen.

Gráfico 4. Frecuencias relativas con los datos de la tabla 7



Fuente: elaboración propia con los datos de la tabla 7

Gráfico 5. Frecuencias acumuladas con los datos de la tabla 7



Fuente: elaboración propia con los datos de la tabla 7

1.2.2.2 Distribución de frecuencias para variable continua

Cuando se trata de una variable continua, es necesario agrupar los datos en intervalos, llamados categorías o clases, los cuales pueden tener igual o distinta amplitud; inclusive, pueden existir intervalos semiabiertos. Por ejemplo:

A) Si se trata de medir rendimiento académico, independientemente del número de estudiantes, se acostumbra a utilizar 5 grupos de igual amplitud que representan rendimiento: excelente, bueno, regular, malo y pésimo.

B) En estudios demográficos, en donde se considere la variable edad, los valores se pueden organizar en intervalos de diferente amplitud, así: menores de 15 años, de 15 a 20, de 20 a 45 y de 45 años o más.

Es recomendable utilizar intervalos de igual amplitud con el fin de facilitar los cálculos estadísticos.

Cuando una variable discreta toma valores muy dispersos, se sugiere organizar los datos en intervalos, mediante una distribución de frecuencias, procediendo de la siguiente manera:

1. Aproximar y ordenar los datos de la variable X.

2. Determinar el recorrido o rango de la variable: $R = X_{max} - X_{min}$.

3. Determinar el número de clases; número que depende de la cantidad de datos y del interés del investigador de agruparlos en más o menos clases; sin embargo, se debe evitar que ocurra lo siguiente:

- Clases con frecuencia nula.
- Número reducido de intervalos, porque se pierde información original.
- Número excesivo de clases, puesto que se pierde el sentido de la agrupación.

En términos generales, es ideal tomar entre 5 y 15 intervalos, para lo cual, se puede utilizar la fórmula: , donde $m = 1 + 3.3 * log(n)$, corresponde a la cantidad de datos y al número de intervalos.

A manera de ejemplo, se presentan los siguientes casos:

- Para menos de 100 datos, se puede escoger 5 o 6 intervalos.
- Para 100 datos: $m = 1 + 3{,}3 * \log(100) = 1 + 3{,}3 * 2 = 7{,}6$
- Para muestras entre 100 y 1000 datos, pueden ser necesarios de 7 a 9 grupos.
- Para 1000 datos: $m = 1 + 3{,}3 * \log(1000) = 1 + 3{,}3 * 3 = 10{,}9$
- Para muestras entre 1000 y 10000 datos se puede utilizar entre 10 y 13 grupos.
- Para 10000 datos: $m = 1 + 3.3 * \log(10000) = 1 + 3.3 * 4 = 14.2$
- Para 10000 o más datos, 14 o 15 grupos.

4. ¿Determinar la amplitud del intervalo mediante la siguiente fórmula:

$$C = \frac{R}{m}$$

5. Formar las clases, empezando con el dato mínimo y aumentar cada vez la amplitud del intervalo.

Ejemplo:

Suponga que los puntajes que se presentan en el cuadro 2 corresponden a los resultados obtenidos en una prueba de conocimientos, calificados en una escala de 100 a 400 puntos.

Cuadro 2. Datos hipotéticos de una prueba de conocimientos de 50 estudiantes calificados en una escala de 100 a 400 puntos

190	*190*	*192*	*198*	*200*	*200*	*203*	*205*	*208*	*208*
209	*214*	*219*	*220*	*225*	*225*	*226*	*227*	*227*	*227*
230	*230*	*230*	*230*	*235*	*240*	*240*	*245*	*246*	*247*
250	*260*	*267*	*268*	*275*	*278*	*280*	*280*	*295*	*296*
297	*300*	*310*	*330*	*331*	*332*	*333*	*335*	*338*	*338*

Fuente: elaboración propia.

Aplicando el procedimiento indicado, se obtiene:

Tamaño de muestra: $n = 50$

Máximo puntaje: 338

Mínimo puntaje: 190

Rango: $R = X_{max} - X_{min} = 338 - 190 = 148$

Número de grupos:

Amplitud: $C = \frac{R}{m} = \frac{148}{7} = 21{,}14 \sim 22$

Se observa que el valor 21,14 se aproxima al entero siguiente, porque si se redondea al entero menor, es posible que en el último intervalo no queden incluidos todos datos, requiriendo adicionar otro intervalo. Los grupos también se pueden escribir como intervalos de la forma [, así: [190, 212).

Tabla 8. Distribución de frecuencias con los datos hipotéticos de una prueba de conocimientos de 50 estudiantes, calificada en una escala de 100 a 400, representados en el cuadro 2.

L_i	L_s	f	h	F	H	X
190	211	11	22%	11	22%	200,5
212	**233**	**13**	**26%**	**24**	**48%**	**222,5**
234	255	7	14%	31	62%	244,5
256	277	4	8%	35	70%	266,5
278	299	6	12%	41	82%	288,5
300	321	2	4%	43	86%	310,5
322	343	7	14%	50	100%	332,5
Total		50	100%			

Fuente: elaboración propia con los datos del cuadro 2

La columna contiene las marcas de clase de cada intervalo y corresponde al valor que representa los datos contenidos en cada intervalo, es decir, en cada grupo.

Para la aplicación de las fórmulas estadísticas en datos agrupados se utilizan las marcas de clase, con el riesgo de que se pierda precisión en el resultado final, puesto que no se trabaja con los datos originales. Si para la distribución de frecuencias se toma un número adecuado de intervalos, las diferencias pueden no ser significativas.

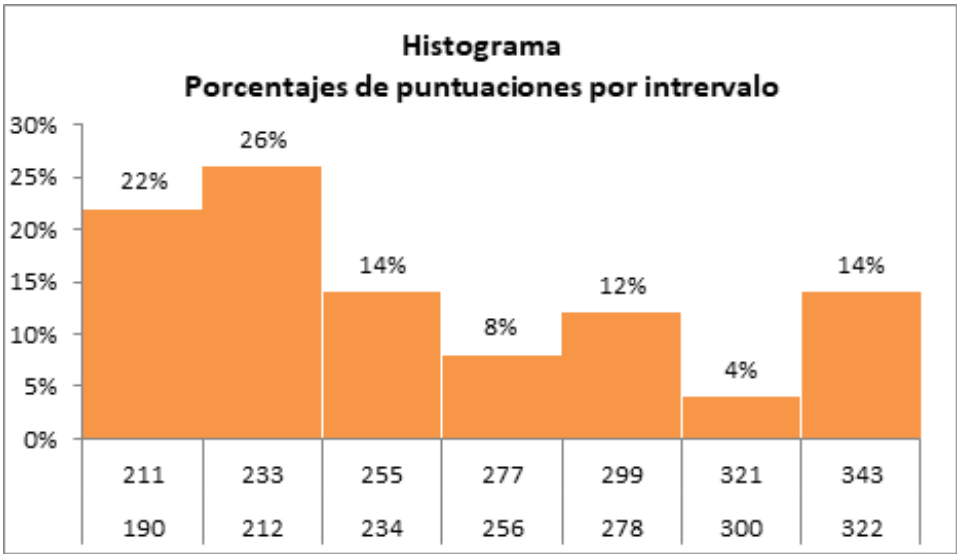
La interpretación de los datos de la segunda fila, es la siguiente:

- 13 estudiantes obtuvieron un puntaje comprendido entre 212 y 233 puntos.
- El 26% de los estudiantes obtuvo un puntaje comprendido entre 212 y 233 puntos.
- 24 estudiantes tienen un puntaje inferior a 234 puntos.
- El 48% de los estudiantes obtuvo un puntaje inferior a 234 puntos.

1.2.2.3 Representación gráfica de una distribución de frecuencias

Una forma de representar gráficamente una distribución de frecuencias es por medio de un gráfico de barras llamado *histograma*, el cual, consiste en una serie de rectángulos cuya base es igual a la amplitud de los intervalos, y la altura es proporcional a la frecuencia respectiva. También se pueden representar por medio de un gráfico de líneas llamado *polígono de frecuencias*, que se construye con las marcas de clase y la frecuencia respectiva.

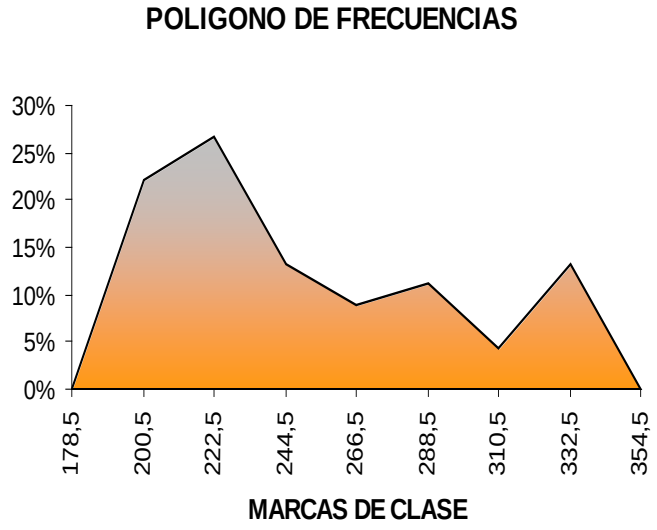
Gráfico 6. Histograma de frecuencias relativas con los datos de la tabla 8



Fuente: elaboración propia con los datos de la tabla 8

La segunda barra se interpreta de la siguiente manera: el 26% de los estudiantes que presentaron la prueba de conocimientos, obtuvieron puntuaciones entre 212 y 233 puntos.

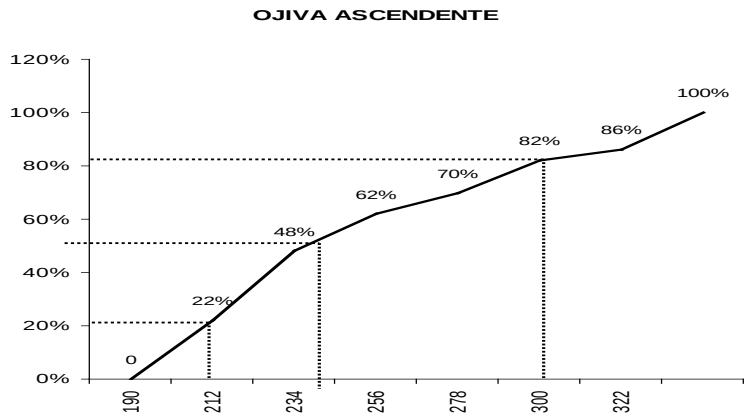
Gráfico 7. Polígono de frecuencias relativas con los datos de la tabla 8



Fuente: elaboración propia con los datos de la tabla 8

Un gráfico que muestra las frecuencias acumuladas (menor que), se denomina *polígono de frecuencias acumuladas u Ojiva.*

Gráfico 8. Frecuencia acumulada ascendente con los datos de la tabla 8



Fuente: elaboración propia con los datos de la tabla 8

Se puede observar que el 22% de los estudiantes tiene un puntaje inferior a 212 puntos, lo cual coincide con los datos de la tabla 9; de igual forma, el 82% tiene puntuaciones menores que 300 puntos.

Del gráfico de una Ojiva ascendente, se puede determinar, de manera aproximada, cualquier porcentaje de alumnos que se ubique por debajo de un valor indicado o encontrar el valor (percentil) por debajo del cual queda un determinado porcentaje (rango percentil). Por ejemplo, el 50% de los estudiantes tienen puntajes inferiores a 243 puntos.

Ejemplo:

Las notas correspondientes al primer parcial de 44 estudiantes de Estadística Descriptiva en el período A-2019, fueron las siguientes:

Cuadro 3. Notas hipotéticas del primer parcial de Estadística Descriptiva en el período A-2019 de la Universidad de Nariño

1,8	*2,5*	*5,0*	*3,0*	*5,0*	*3,0*	*3,8*	*2,8*	*0,8*
2,8	*3,0*	*1,0*	*5,0*	*4,0*	*3,0*	*3,5*	*4,5*	*2,5*
4,0	*1,0*	*2,0*	*4,5*	*2,5*	*2,0*	*2,5*	*1,5*	*0,9*
4,5	*4,0*	*4,0*	*0,5*	*3,2*	*4,0*	*5,0*	*4,5*	*2,5*
1,8	*3,0*	*2,5*	*3,8*	*3,2*	*4,6*	*4,6*	*1,0*	

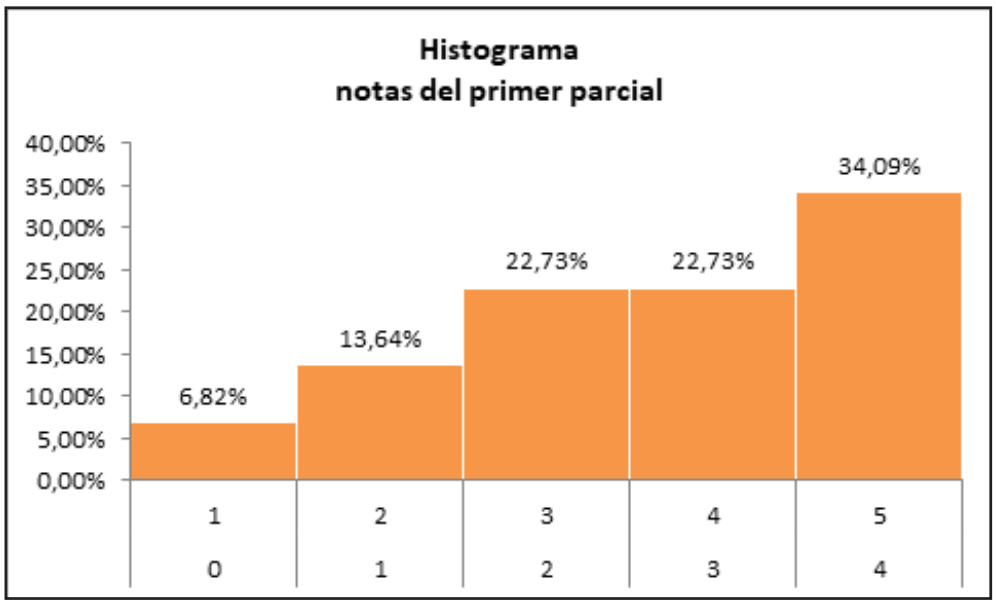
Los intervalos utilizados son intervalos semiabiertos, de la siguiente forma: [. Por ejemplo, en el intervalo se incluye las calificaciones mayores o iguales a cero y menores que uno.

Tabla 9. Notas parciales hipotéticas de Estadística Descriptiva del primer parcial de Estadística Descriptiva en el período A-2019 de la Universidad de Nariño, representados en el cuadro 3

L_i	L_s	f	h	F	H	X
0	1	3	6,82%	3	6,82%	0,495
1	2	6	13,64%	9	20,45%	1,495
2	3	10	22,73%	19	43,18%	2,495
3	4	10	22,73%	29	65,91%	3,495
4	5	15	34,09%	44	100,00%	4,5
		44	100%			

Fuente: elaboración propia con datos del cuadro 3

Gráfico 9. Histograma de frecuencias



Fuente: elaboración propia con los datos de la tabla 9

1.3 TALLER

a) Elaborar un proyecto de investigación con el fin de analizar el desempeño de los estudiantes en Estadística Descriptiva de la Universidad de Nariño, durante el período Académico A-2018. Para el efecto, diseñar un formulario adecuado mediante el cual recolectar la información necesaria.

b) Con la información recolectada en el literal anterior, construir una tabla de modo que en las columnas se ubiquen las preguntas (variables), y en las filas los registros o respuestas a cada una de las variables.

c) Realizar una representación tabular y gráfica de la información recolectada e interpretar los resultados.

CAPÍTULO 2.
MEDIDAS ESTADÍSTICAS

CAPÍTULO 2.
MEDIDAS ESTADÍSTICAS

Son valores que permiten conocer las características de una variable, también se conocen con el nombre de estadígrafos; se destacan las medidas de tendencia central o promedios, de posición, de variabilidad, de asimetría y curtosis.

Todas las medidas estadísticas aquí estudiadas, se pueden calcular de dos maneras:

a) Utilizando funciones estadísticas de Microsoft Excel: se trabaja directamente en la base de datos, de lo cual se obtiene los resultados únicamente.

b) Aplicando fórmulas en las distribuciones de frecuencias: en este caso, se obtiene los resultados de manera didáctica.

2.1 MEDIDAS DE TENDENCIA CENTRAL

Las distribuciones de frecuencia permiten resumir los datos y, a la vez, apreciar las variaciones que experimenta una variable; sin embargo, no son suficientes para un análisis completo; en consecuencia, es necesario conocer los procedimientos estadísticos que permiten obtener información amplia y suficiente para la toma de decisiones.

Las medidas de tendencia central -MTC- o promedios, son valores típicos y representativos de un conjunto de datos; los más importantes son: promedio aritmético, promedio ponderado, mediana, moda, media geométrica y media armónica.

El uso de una u otra medida depende de las características de los datos, los cuales pueden ser: temporales, atemporales, de proporcionalidad; pero también depende de la conveniencia y de la experiencia del investigador.

2.1.1 Promedio aritmético

La media aritmética de una variable X se representa por $\overline{X}$ y se define de la siguiente manera:

a) Para datos no agrupados $X_1, X_2, \ldots, X_n$ se suma todos los datos y se divide por el tamaño de la muestra:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

b) Para datos agrupados en una distribución de frecuencias, se calcula así:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i f_i}{n}$$

Donde X_i representa las de clase y f_i las frecuencias observadas.

En los siguientes casos **NO** es recomendable utilizar el promedio aritmético:

a) Los datos son muy heterogéneos.

b) Hay presencia de valores extremos muy altos o muy bajos.

c) Se desea conocer el promedio de una variable a través del tiempo; por ejemplo, costo de vida, crecimiento de población, operaciones financieras.

Ejemplo:

Calcular el promedio de los puntajes de una prueba de conocimientos aplicada a 50 estudiantes, cuyos datos se presentan en el cuadro 4.

Cuadro 4. Puntajes hipotéticos de una prueba de conocimientos aplicada a 50 estudiantes en una escala de 1 a 400

190	190	192	198	200	200	203	205	208	208
209	214	219	220	225	225	226	227	227	227
230	230	230	230	235	240	240	245	246	247
250	260	267	268	275	278	280	280	295	296
297	300	310	330	331	332	333	335	338	338

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = 251{,}58$$

En seguida se calcula el promedio aritmético aplicando la fórmula para datos agrupados; para lo cual se organizan en una distribución de frecuencias con siete (7) intervalos (clases), tal como se indica en la tabla 10.

Tabla 10. Puntajes hipotéticos de una prueba de conocimientos aplicada a 50 estudiantes en una escala de 1 a 400, representados en el cuadro 4.

Grupos	L_i	L_s	f	X	Xf
1	190	211	11	200,5	2205,5
2	212	233	13	222,5	2892,5
3	234	255	7	244,5	1711,5
4	256	277	4	266,5	1066
5	278	299	6	288,5	1731
6	300	321	2	310,5	621
7	322	343	7	332,5	2327,5
	Total		50		12555

Fuente: elaboración propia con los datos del cuadro 4

Se puede observar que los promedios calculados con los datos **NO** agrupados y con los mismos datos agrupados en clases, son aproximadamente iguales; pues, si bien, en la agrupación se pueden producir diferencias, en general estas no son importantes.

2.1.1.1 Propiedades de la media aritmética

Se denominan *desviaciones* a las diferencias de una variable con respecto a un valor particular de la misma, el cual se toma como referencia. En lo que sigue, se consideran las desviaciones respecto al promedio aritmético.

1) La suma algebraica de todas las desviaciones (diferencias) respecto al promedio aritmético es igual a cero.

a) Para **datos no agrupados:**

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0$$

Ejemplo:

Suponga que los siguientes datos corresponden a las calificaciones de cinco estudiantes, evaluados en una escala de 1 a 10: 2, 4, 6, 8, 10.

El promedio de las calificaciones es:

$$\overline{X} = \frac{2+4+6+8+10}{5} = 6$$

Tabla 11. Ejemplo de desviación de datos no agrupados respecto a la media

X	$X - \overline{X}$
2	*-4*
4	*-2*
6	*0*
8	*2*
10	*4*
Suma	**0**

Fuente: elaboración propia con los datos del ejemplo anterior

b) Para **datos agrupados**

$$\sum_{i=1}^{n}(X_i - \overline{X})f_i = 0$$

Tabla 12. Ejemplo de cálculo de desviación en datos agrupados respecto a la media con los datos de la tabla 10

Grupos	L_i	L_s	f	X	Xf	$\left(X_i - \overline{\overline{X}}\right) f_i$
1	190	211	11	200,5	2205,5	-556,6
2	212	233	13	222,5	2892,5	-371,8
3	234	255	7	244,5	1711,5	-46,2
4	256	277	4	266,5	1066	61,6
5	278	299	6	288,5	1731	224,4
6	300	321	2	310,5	621	118,8
7	322	343	7	332,5	2327,5	569,8
	Total		50		12555	**0**

Fuente: elaboración propia con los datos del cuadro 4

2) El promedio aritmético de una constante es igual a la misma constante:

$$\overline{K} = K$$

3) El promedio del producto de una constante por una variable, es igual al producto de la constante por el promedio de la variable:

$$\overline{KX} = K\overline{X}$$

4) El promedio de la suma de una variable y una constante, es igual a la media de la variable más la constante:

$$\overline{X + K} = \overline{X} + K$$

Si una muestra de observaciones con promedio $\overline{X}$, se divide en dos o más submuestras de datos con promedio $\overline{X_1}$ n_2, datos con promedio $\overline{X_2}, \ldots, n_k$ datos con promedio $\overline{X_k}$, entonces la media de todos los datos es:

$$\overline{X} = \frac{n_1\overline{X}_1 + n_2\overline{X}_2 + \cdots + n_k\overline{X}_k}{n_1 + n_2 + \ldots + n_k}$$

Ejemplo:

En un curso de Estadística hay 60 estudiantes, de los cuales 20 son mujeres; la nota promedio de los hombres es 4,0 y de las mujeres 3,5. Determinar la nota nota promedio de todo el grupo.

El promedio del grupo se obtiene aplicando la fórmula para el cálculo del promedio ponderado, así:

$$\overline{X} = \frac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1 + n_2}$$

$$\overline{X} = \frac{20 * 3{,}5 + 40 * 4}{20 + 40} = 3{,}83$$

Se observa que el promedio ponderado es 3,83 puntos, un poco menos que el promedio de los hombres, y a la vez, un poco más que el promedio de las mujeres.

2.1.1.2 Método abreviado para calcular el promedio aritmético

Este método se aplica en distribuciones de frecuencias con intervalos de igual amplitud. Sea la siguiente transformación lineal:

$$u = \frac{X - A}{c}$$

$\boldsymbol{X}$ representa las marcas de clase, es cualquiera de las marcas de clase y, corresponde a la amplitud del intervalo. Dado que las diferencias $X - A$ son múltiplos de C, la variable U toma valores enteros consecutivos. Despejando $\boldsymbol{X}$ de la ecuación anterior, se obtiene la siguiente expresión:

$$X = A + cu$$

Aplicando las propiedades 3 y 4 del promedio aritmético, se obtiene la siguiente ecuación:

$$\overline{X} = A + c\overline{u}$$

Ejemplo:

Utilizando los datos de la tabla 10, sea ; determinar los valores de *u*:

Tabla 13. Distribución de frecuencias para cálculo de promedio por método abreviado con los datos de la tabla 10

Grupos	L_i	L_s	f	x	u	uf
1	190	211	11	200,5	-3	-33
2	212	233	13	222,5	-2	-26
3	234	255	7	244,5	-1	-7
4	256	277	4	266,5	0	0
5	278	299	6	288,5	1	6
6	300	321	2	310,5	2	4
7	322	343	7	332,5	3	21
	Total		50		0	-35

Fuente: elaboración propia con los datos de la tabla 10

$$A = 266{,}5;\ c = 22$$

$$u = \frac{X - A}{c}$$

$$u = \frac{200{,}5 - 266{,}5}{22} = -3$$

$$\overline{u} = \frac{\sum u_i f_i}{\sum f_i} = -\frac{35}{50} = 0{,}7$$

$$\overline{X} = A + c\overline{u} = 266{,}5 + 22 * (-07) = 251{,}1$$

$$\overline{X} = 251{,}1$$

2.1.1.3 Promedio aritmético ponderado

Hay casos en que los valores de una variable no tienen la misma importancia o peso, por lo cual, es necesario asignarles un determinado factor o ponderación, que se simboliza

por $\overline{X}_p$

Si X_1 tiene un peso W_1, X_2 tiene un peso W_2, etc., entonces, la media aritmética ponderada se define por:

$$\overline{X_p} = \frac{\sum X_i W_i}{\sum W_i}$$

Los pesos W_i pueden tomar cualquier valor numérico.

Ejemplo:

Suponga que en un curso de administración se efectúan cuatro evaluaciones parciales, asignando los siguientes pesos: 1 para la primera evaluación, 2 para la segunda, 5 para la tercera y 8 para la última evaluación. Calcular la nota promedio.

Tabla 14. Cálculo del promedio con pesos ponderados

No. de Evaluación	Notas (X)	Peso (W)	XW
1	3,5	1,0	3,5
2	4,0	2,0	8,0
3	4,5	5,0	22,5
4	2,0	8,0	16,0
	Total	**16**	**50**

Fuente: elaboración propia

Aplicando ponderaciones a las evaluaciones, se obtiene lo siguiente:

$$\overline{X_p} = \frac{\sum X_i W_i}{\sum W_i} = \frac{50}{16} = 3{,}125$$

Por su parte, el promedio aritmético sin aplicar ponderaciones es:

$$\overline{X} = \frac{3{,}5 + 4 + 4{,}5 + 2}{4} = 3{,}5$$

Se observa que el promedio ponderado difiere del promedio aritmético no ponderado. La aplicación de una u otra fórmula depende del enunciado del problema y de la naturaleza del mismo.

2.1.2 Mediana.

Es el valor que divide en dos partes iguales una serie ordenada de datos; es decir, es el valor central de la serie. Se puede utilizar la mediana como un valor representativo de un conjunto de datos cuando no sea recomendable utilizar el promedio aritmético. Se denota por M_e

Ejemplos:

a) Determinar la mediana del siguiente conjunto de datos: $2, 4,6,8,10$

Ordenando los datos, se tiene: $2, 4,6,8,10$.

valor que ocupa la tercera posición en una lista ordenada de 5 datos.

b) Determinar la mediana de los siguientes datos: $2, 5, 6, 7, 8, 9$..

Cuando el número de datos es par, la mediana es el promedio de los valores centrales.

La lista tiene 6 datos y los dos valores centrales son 6 y 7, por lo cual, la mediana se calcula así:

$$M_e = \frac{6+7}{2} = 6{,}5$$

2.1.2.1 Cálculo de la mediana en datos agrupados para una variable discreta

Tabla 15. Modelo de tabla para cálculo de mediana en datos agrupados

X_i	f_i	F_i	h_i	H_i
X_1	f_1	F_1	h_1	H_1
X_2	f_2	F_2	h_2	H_2
X_m	f_m	$F_m = n$	h_m	$H_m = 1$
$\sum f_i$	n		1	

Fuente: elaboración propia

Pasos:

1) Determinar el lugar central aplicando la siguiente fórmula:

$$LMe = \frac{n+1}{2}$$

2) Determinar el valor el cual corresponde a la primera frecuencia relativa acumulada, que contiene el 50% de los datos.

Ejemplo:

Determinar la mediana de los datos de la tabla que sigue, la cual contiene datos hipotéticos de puntuaciones en una prueba de conocimientos:

X_i	f_i
2	5
4	3
5	8
6	**10**
7	12
8	13
Total	51

Tabla 16. Ejemplo de cálculo de mediana en datos agrupados sin intervalos

X_i	f_i	F_i	h_i	H_i
2	5	5	9,80%	9,80%
4	3	8	5,90%	15,70%
5	8	16	15,70%	31,40%
6	**10**	**26**	**19,60%**	**51,00%**
7	12	38	23,50%	74,50%
8	13	51	25,50%	100,00%
Total	51		1	

Fuente: elaboración propia.

$$LM_e = \frac{n+1}{2} = \frac{51+1}{2} = 26$$

Por lo tanto, el lugar de la mediana es ; en consecuencia, observando la columna de los datos se tiene que la mediana es .

2.1.2.2 Cálculo de la mediana para datos agrupados en intervalos

Se aplica una fórmula de interpolación, ubicando previamente la clase median, la cual corresponde al lugar de la mediana LM_e.

$$M_e = L_{ir} + \frac{\left(\frac{n}{2} - F_a\right) * c}{f_0}$$

Donde:

L_{ir}: límite inferior real de la clase mediana.

n: tamaño de muestra.

F_a : frecuencia acumulada anterior a la clase mediana.

f_0 : frecuencia observada de la clase.

c : amplitud del intervalo.

Ejemplo:

Calcular la mediana con los datos de la tabla 17, que contiene los puntajes obtenidos por 50 estudiantes en una prueba de conocimientos.

Tabla 17. Ejemplo de cálculo de mediana en datos agrupados en intervalos

L_i	L_s	f	h	F	H
190	211	11	22%	11	22%
212	233	13	26%	24	48%
234	255	7	14%	31	62%
256	277	4	8%	35	70%
278	299	6	12%	41	82%
300	321	2	4%	43	86%
322	343	7	14%	50	100%
Total		50	100%		

Fuente: elaboración propia.

$$n = 50$$

$$LM_e = \frac{n+1}{2} = 25{,}5$$

$$c = 22$$

$$L_{ir} = 233{,}5$$

$$F_a = 24$$

$$f_0 = 7$$

$$M_e = L_{ir} + \frac{\left(\frac{n}{2} - F_a\right) * c}{f_0} = 233{,}5 + \frac{(\frac{50}{2} - 24) * 22}{7} = 236{,}6$$

Se obtiene que $M_e = 236{,}6$; en consecuencia, el 50% de los estudiantes obtuvo un puntaje inferior a puntos.

2.1.3 Media geométrica

La media geométrica, M_g de una serie $X_1, X_2, \ldots, X_n$ se define como la raíz nésima del producto de los datos; es decir:

$$M_g = \sqrt[n]{\prod_{i=1}^{n} X_i}$$

Cuando los datos están agrupados en una distribución de frecuencias, la media geométrica se obtiene con la siguiente fórmula:

$$M_g = \sqrt[n]{\prod_{i=1}^{n} X_i f_i}$$

Si la muestra es grande, la fórmula anterior presenta algunos inconvenientes de cálculo, generados por el producto de potencias; lo cual se puede evitar aplicando logaritmos, así:

$$Log(M_g) = \frac{\sum f_i log(X_i)}{n}$$

Se puede observar que el logaritmo de la media geométrica es igual a la media aritmética de los logaritmos de los datos.

La media geométrica se utiliza para medir los cambios de una variable a través del tiempo; por ejemplo, para promediar tasas de cambio, analizar el incremento de una población o para la obtención de índices, siempre y cuando las variaciones sean directamente proporcionales.

Ejemplo:

Suponga que la demanda de cupos escolares entre los años 1999 y 2004, corresponde a los datos que muestra el cuadro que sigue. Calcular la tasa media de crecimiento porcentual, por año.

Año	Demanda
1999	5.000
2000	6.000
2001	9.000
2002	15.000
2003	30.000
2004	50.000

Tabla 18. Ejemplo de cálculo de media geométrica

Año	Demanda	Factor de crecimiento (X_i)	$Log(X_i)$
1999	5.000	-	-
2000	6.000	1,2	0,07918
2001	9.000	1,5	0,17609
2002	15.000	1,6666667	0,22272
2003	30.000	2	0,30103
2004	50.000	1,6666667	0,22272
		$\prod X_i = 10$	$\sum Log(X_i) = 1{,}00174$

Fuente: elaboración propia

Los factores de crecimiento (índices) se obtienen dividiendo cada dato por el anterior; de este modo, el factor de crecimiento para el año 2000 es:

$$\frac{6000}{5000} = 1{,}20$$

El año 1999 no tiene factor de crecimiento por ser el primero, el cual se toma como referencia.

El factor de crecimiento del año 2000 es 1,20 valor que indica un crecimiento del 20% con respecto al año 1999.

La tasa promedio de crecimiento de la demanda, se obtiene aplicando la fórmula de la media geométrica a los factores de crecimiento, así:

$$M_g = \sqrt[5]{1{,}20x1{,}50x1{,}67x2{,}0x1{,}67} = \sqrt[5]{10.0400} = 1{,}585$$

Por lo tanto, la tasa promedio de crecimiento de la demanda de estudiantes entre 1999 y 2004 es del 58,5% anual.

Determinemos ahora el valor de M_g aplicando logaritmos:

$$log(M_g) = \frac{\sum log(X_i)}{n} = \frac{1{,}00174}{5} = 0{,}20034$$

$$M_g = anti\ log(0{,}20034) = 1{,}585$$

Este resultado coincide con el calculado anteriormente.

Cuando los datos están agrupados, la fórmula para el cálculo del promedio geométrico es la siguiente:

$$M_g = \sqrt[n]{\prod x_i^{f_i}}$$

Se puede observar que cada factor representa números muy grandes y el producto de todos los factores es muy dispendioso, por lo cual, es conveniente utilizar logaritmos de los datos.

Ejemplo:

Calcular la media geométrica con los datos de la siguiente tabla, que contiene los puntajes obtenidos por 50 estudiantes en una prueba de conocimientos.

Grupos	L_i	L_s	f	X_i	$Log(X_i)$	$f * Log(X_i)$
1	190	211	11	200,5	2,302	25,322
2	212	233	13	222,5	2,347	30,511
3	234	255	7	244,5	2,388	16,716
4	256	277	4	266,5	2,426	9,704
5	278	299	6	288,5	2,46	14,76
6	300	321	2	310,5	2,492	4,984
7	322	343	7	332,5	2,522	17,654
	Total		50			119,651

Fuente: elaboración propia

$$Log(M_g) = \frac{\sum f_i\ log\ X_i}{n} = \frac{119{,}651}{50} = 2{,}39302$$

$$Mg = anti\ log(2{,}39302) = 247{,}2$$

Se indicó que la media geométrica se utiliza para determinar el promedio de crecimiento de una variable a través del tiempo. En este sentido, la siguiente expresión determina el crecimiento de una cantidad después de períodos de tiempo, con una tasa constante de crecimiento :

$$F = A(1 + r)^n$$

F : cantidad final

A : cantidad inicial

$1 + r$: factor de crecimiento

n : unidades de tiempo.

De la fórmula anterior, se obtiene que:

$$1 + r = \sqrt[n]{\frac{F}{A}}$$

El cociente $\frac{F}{A}$ se denomina *índice acumulado* y es igual al producto de los factores de crecimiento.

Ejemplo:

Si en un país, la población universitaria en el año 2000 era de 4.000 estudiantes y en el 2019 es de 13.000. ¿Cuál es el factor y la tasa de crecimiento anual?

$$1 + r = \sqrt[n]{\frac{F}{A}}$$

$$1 + r = \sqrt[19]{\frac{13.000}{4000}} = 1{,}064$$

Este valor es el factor de crecimiento anual; significa que en cada año la población universitaria se va multiplicando por 1,064 y que la rasa de crecimiento es del 6,4% anual.

2.1.4 Media armónica

La media armónica de una serie de observaciones $X_1, X_2, \ldots, X_n$ se define como el recíproco de la media aritmética de los recíprocos. Se simboliza por M_h y se utiliza para promediar cantidades inversamente proporcionales.

Por ejemplo, se usa para promediar velocidades en tiempos diferentes, así:

$$S = VT$$

Al trabajar con dos espacios S_1 y S_2, se tiene lo siguiente:

$$S_1 = V_1T_1 \text{ y } S_2 = V_2T_2$$

$$S = S_1 + S_2 = V_1T_1 + V_2T_2$$

Dividiendo esta expresión por el tiempo total, se obtiene:

$$\frac{S}{T} = \frac{S_1 + S_2}{T} = \frac{V_1T_1 + V_2T_2}{T_1 + T_2} = V_m$$

Entonces:

$$V_m = \frac{S}{T} = \frac{S}{T_1 + T_2} = \frac{S}{\sum \frac{S_i}{V_i}}$$

2.1.4.1 Cálculo en datos no agrupados

$$M_h = \frac{1}{\sum\left(\frac{\frac{1}{X_i}}{n}\right)}$$

2.1.4.2 Cálculo en datos agrupados

$$M_h = \frac{n}{\sum\left(\frac{f_i}{X_i}\right)}$$

Ejemplo:

Suponga que la distancia entre dos ciudades A y B es de 80 kilómetros y que entre B y C es de 120m km. Si un automóvil gasta una hora en cada uno de los recorridos, determinar la velocidad promedio del recorrido.

En este caso, la solución se obtiene aplicando la media aritmética, porque el tiempo permanece constante.

$$V_1 = 80\ km/h$$

$$V_2 = 120\ km/h$$

$$V_m = \frac{80 + 120}{2} = 100\ km/h$$

El resultado es similar si se utiliza la media armónica.

$$Vm = \frac{80 + 120}{\frac{80}{80} + \frac{120}{120}} = 100km/h$$

Pero si el automóvil recorre los 80 km. a una velocidad de $100\ kph$. y los $120\ km$ a una velocidad de $80\ kph$, sería incorrecto utilizar el promedio aritmético para calcular la velocidad promedio; en este caso se tendría que aplicar la media armónica.

Tabla 19. Datos para ejemplo de cálculo de media armónica

X (Vel)	F (Km)
100	80
80	120
Total	200

Fuente: elaboración propia

$$M_h = \frac{n}{\sum\left(\frac{f}{n}\right)} = \frac{200}{\frac{80}{100} + \frac{120}{80}} = 86{,}956\ kph$$

Por tanto, el vehículo recorrió los dos tramos con una velocidad promedio de 86,956 kph.

2.1.5 Moda

Se define como el valor de mayor frecuencia en un conjunto de datos; es decir, corresponde al valor que más se repite; por ejemplo, la nota predominante en un examen.

2.1.5.1 Cálculo en datos agrupados

$$M_o = L_{ir} + \frac{\Delta_1 * c}{\Delta_1 + \Delta_2}$$

L_i : límite inferior real de la clase de mayor frecuencia (clase modal)

Δ_1 : frecuencia de la clase modal, menos la frecuencia anterior

Δ_2 : frecuencia de la clase modal, menos la frecuencia posterior

c : amplitud del intervalo

Ejemplo:

Calcular la moda de la siguiente distribución de datos:

Tabla 20. Datos para ejemplo de cálculo de moda en datos agrupados

Puntaje	f_i
190-220	13
220-250	17
250-280	6
280-310	6
310-340	8
Total	50

Fuente: elaboración propia

$$\Delta_1 = 17 - 13 = 4$$

$$\Delta_2 = 17 - 6 = 11$$

$$M_o = 220 + \frac{4 * 30}{4 + 11} = 228$$

2.2 MEDIDAS DE POSICIÓN

La mediana divide una serie de datos ordenados en dos partes iguales, dejando un 50% de información por debajo de este valor y un 50% por encima. Pero la serie puede dividirse en cuatro, diez o cien partes iguales, dando lugar a cuartiles, deciles y percentiles.

2.2.1 Cuartiles

El cuartil uno Q_1 o cuartil inferior, es el valor que supera el 25% de la información y a la vez es superado por el 75% restante. El cuartil dos Q_2 es la misma mediana. El cuartil tres Q_3 o cuartil superior, es aquel valor que supera el 75% de la información y es superado por el 25%. Para su cálculo, se procede de igual forma que en la mediana; primero se lo ubica en H (frecuencia relativa acumulada) y luego se aplica la siguiente fórmula:

$$Q_k = L_{ir} + \frac{\left(\frac{kn}{4} - F_a\right) * c}{f_0}$$

$$k = 1, 2, 3$$

2.2.2 Deciles

De igual forma que para el caso de los cuartiles, se puede dividir una serie ordenada de datos en 10 pares iguales, denominadas deciles. El primer decil D_1 deja por debajo el 10% de información. El D_2 el 20% y así sucesivamente. Para su cálculo se utiliza el mismo procedimiento que en los cuartiles, solo que en lugar de dividir entre 4, se divide entre 10.

$$D_k = L_{ir} + \frac{\left(\frac{kn}{10} - F_a\right) * c}{f_0}$$

$$k = 1, 2, 3, \ldots, 9$$

2.2.3 Percentiles

Si se quiere dividir la distribución en 100 partes iguales, se procede de la misma forma que para los deciles. El primer percentil P_1 deja por debajo el 1% de información y es superado por el 99%. El percentil dos P_2 supera al 2% y es superado por el 98% de los datos, y así sucesivamente.

Entre cuartiles, deciles y percentiles, se cumplen las siguientes relaciones:

$$D_1 = P_{10};\ D_2 = P_{20};\ Q_1 = P_{25};\ Q_2 = D_{25} = P_{50};\ Q_3 = P_{75}$$

Para determinar los percentiles, primero se encuentra su lugar en H, luego se aplica la siguiente fórmula de interpolación:

$$P_k = L_{ir} + \frac{\left(\frac{kn}{100} - F_a\right) * c}{f_0}$$

$$k = 1, 2, 3, \ldots, 99$$

Ejemplo:

La tabla 21 presenta la distribución por edades de 200 estudiantes, con base en la cual se pide calcular cuartiles, deciles y percentiles.

Tabla 21. Ejemplo de cálculo de medidas de posición

Edad	f	F	H
10-12	10	10	5%
12-14	35	45	22,50%
14-16	95	140	70%
16-18	35	175	87,50%
18-20	15	190	95%
20-22	10	200	100%
Total	200		-

Fuente: elaboración propia

Calculemos el cuartil Q_1 que es equivalente al percentil P_{25}.

$$Q_1 = P_{25} = L_{ir} + \frac{\left(\frac{25n}{100} - F_a\right) * c}{f_0}$$

$$P_{25} = L_{ir} + \frac{\left(\frac{25n}{100} - F_a\right) * c}{f_0} = 14 + \frac{\left(\frac{25 * 200}{100} - 45\right) * 2}{95} = 14{,}1$$

$$P_{25} = 14{,}1 \text{ años.}$$

Este valor indica que el 25% de los estudiantes tiene menos de 14,1 años de edad.

$$Q_3 = P_{75} = L_{ir} + \frac{\left(\frac{75n}{100} - F_a\right) * c}{f_0} = 16 + \frac{\left(\frac{75 * 200}{100} - 140\right) * 2}{35} = 16{,}6$$

$$P_{75} = 16{,}6 \text{ años.}$$

Este valor indica que el 75% de los estudiantes tiene menos de 16,6 años de edad.

Cuando se realiza análisis de frecuencias, surgen interrogantes que se pueden resolver en forma inmediata a partir de los datos de la distribución de frecuencias. Por ejemplo, el porcentaje de estudiantes con menos de 16 años de edad, según lo muestra la tabla, es 70%.

Tabla 22. Distribución de frecuencias relativas de los datos de la tabla 21

Edad	f_i	F_i	h_i (%)	H_i (%)
10 - 12	10	10	5	5
12 - 14	35	45	17,5	22,5
14 - 16	95	140	47,5	70,0
16 - 18	35	175	17,5	87,5
18 -20	15	190	7,5	95,0
20 -22	10	200	5,0	100,0
Total	200		100	

Fuente: elaboración propia

Para determinar el porcentaje de estudiantes con menos de 15 años, dado que *no aparece* directamente en la tabla, se tiene que calcular.

El porcentaje de datos que está por debajo de un determinado valor, se conoce como *rango percentil* y se obtiene despejando , de la fórmula del percentil.

2.2.4 Rango percentil

$$k = \left[\frac{(p_k - L_{ir})f_0}{c} + F_a\right] * \frac{100}{n}$$

Aplicando esta fórmula, se obtiene que, el porcentaje de estudiantes con menos de 15 años de edad es:

$$k = \left[\frac{(15 - 14)95}{2} + 45\right] * \frac{100}{200} = 46{,}25\%$$

En este ejemplo, se tiene que el porcentaje de estudiantes con menos de 15 años de edad es $k = 46{,}25\%$.

2.3 MEDIDAS DE VARIABILIDAD

Estas medidas se utilizan para conocer el grado de dispersión o variación que presentan los datos al rededor del promedio. Cuando comparamos dos (2) o más distribuciones de frecuencia, las medidas de tendencia central no son suficientes para hacer las comparaciones, ya que dos distribuciones pueden tener la misma media y ser diferentes. Para lograr un mejor conocimiento acerca del comportamiento de los datos, se utiliza las medidas de variabilidad, entre las cuales se encuentran el recorrido o rango, rango intercuartílico, desviación media, varianza, desviación típica o estándar, y coeficiente de variabilidad.

2.3.1 Recorrido o rango

Aunque el recorrido es una medida muy sencilla de la dispersión de los datos, se puede utilizar cuando se desea obtener rápidamente el grado de variabilidad; por ejemplo, cuando se quiere comparar las notas de un examen en dos grupos diferentes.

$$\text{Recorrido } R = X_{max} - X_{min}$$

Ejemplo:

Suponga quela tabla 23 contiene las notas máximas y mínimas de dos cursos de Estadística Descriptiva y se desea comparar el rendimiento de los grupos.

Tabla 23. Ejemplo de cálculo del rango en distribuciones de frecuencia

Grupos	Nota máxima	Nota mínima	Rango
A	*9,8*	*2,0*	*7,8*
B	*9,0*	*6,0*	*3,0*

Fuente: elaboración propia

De la tabla se puede decir que el rendimiento del Grupo B, fue más homogéneo, puesto que el rango del Grupo B es menor que el rango de los datos del Grupo A.

Aunque el recorrido es una medida significativa, tiene el inconveniente de estar afectado solo por los valores extremos. Con el fin de superar este problema se utiliza el Rango Intercuartílico, que es la diferencia entre el cuartil Q_3 y el cuartil Q_1 u otra medida de variabilidad.

$$\text{Rango intercuartílico} = P_{75} - P_{25}$$

$$P_{75} = 16{,}6 \text{ años}$$

$$P_{25} = 14{,}1 \text{ años}$$

Entonces $Rango\ Intercuartílico = 2{,}5\ años$

2.3.2 Desviación media

Se entiende por variación la diferencia de un dato con respecto a un punto de referencia, que bien puede ser el promedio. Para disponer de una medida de variabilidad que tenga en cuenta todas las diferencias o desviaciones, lo más común es calcular el promedio aritmético de todas las desviaciones; sin embargo, como se demostró anteriormente, la suma de las desviaciones respecto a la media es cero, ya que el promedio aritmético es el punto de equilibrio de la distribución, por lo cual, las desviaciones a la izquierda de la media se compensan con las desviaciones a la derecha. Una manera de evitar el cero es trabajar con los valores absolutos de las desviaciones respecto a la media.

2.3.2.1 Cálculo de la desviación media en datos NO agrupados

$$D_m = \frac{\sum |X - \overline{X}|}{n}$$

2.3.2.2 Cálculo de la desviación media en datos agrupados

$$D_m = \frac{\sum |X - \overline{X}| * f}{n}$$

Ejemplo:

Calcular la desviación media en la tabla 24, donde $\overline{X} = 252{,}4$.

Tabla 24. Ejemplo de cálculo de la desviación media en una distribución de frecuencias

Puntajes	f_i	X_i	$X_i - \overline{X}$	$\lvert X_i - \overline{X} \rvert$	$\lvert X_i - \overline{X} \rvert * f_i$
190-220	13	205	-47,4	47,4	616,2
220-250	17	235	-17,4	17,4	295,8
250-280	6	265	12,6	12,6	75,6
280-310	6	295	42,6	42,6	255,6
310-340	8	325	72,6	72,6	580,8
Total	50				1824

Fuente: elaboración propia

$$\overline{X} = 252{,}4$$

$$D_m = \frac{\sum |X - \bar{X}| * f}{n} = 36{,}48$$

Otra forma de evitar el cero de las desviaciones respecto a la media, es tomar los cuadrados de tales desviaciones; en este caso, el promedio aritmético de estos nuevos valores se denomina varianza.

2.3.3 Varianza

Por definición, la varianza es el promedio aritmético de los cuadrados de las desviaciones respecto a la media de los datos.

2.3.3.1 Cálculo de varianza en datos no agrupados

$$S^2 = \frac{\sum (X - \bar{X})^2}{n}$$

Esta fórmula es equivalente a:

$$S^2 = \frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2$$

2.3.3.1 Cálculo de varianza en datos agrupados

$$S^2 = \frac{\sum (X - \overline{X})^2 f}{n}$$

Esta fórmula es equivalente a:

$$S^2 = \frac{\sum X^2 f}{n} - \left(\frac{\sum X f}{n}\right)^2$$

2.3.4 Desviación típica o desviación estándar

Es la raíz cuadrada no negativa de la varianza; es decir:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2 * f}{n}}$$

Ejemplo:

Tabla 25. Ejemplo de cálculo de varianza en datos agrupados

Puntajes	f_i	X_i	$X_i * f$	$X_i^2 * f$
190-220	13	205	2665	546325
220-250	17	235	3995	938825
250-280	6	265	1590	421350
280-310	6	295	1770	522150
310-340	8	325	2600	845000
Total	50		12620	3273650

Fuente: elaboración propia

$$S^2 = \frac{\sum X^2 f}{n} - \left(\frac{\sum X f}{n}\right)^2$$

$$S^2 = \frac{\sum X^2 f}{n} - \left(\frac{\sum X f}{n}\right)^2 = \frac{3.273.650}{50} - \left(\frac{12.620}{50}\right)^2 = 1.767{,}24$$

2.3.5 Coeficiente de variación

Las medidas de variabilidad expresadas en valores absolutos, son bastante útiles para describir la dispersión de un conjunto de datos. Sin embargo, cuando se trata de comparar dos conjuntos de valores, estas medidas son convenientes sólo cuando las unidades de medida son las mismas y los conjuntos tienen promedios aproximadamente iguales, en caso contrario, no son comparables. Cuando no son compa-

rables, se utiliza una medida de variación relativa denominada coeficiente de variación y se define como el cociente resultante entre la desviación estándar y el promedio aritmético:

$$C_v = \frac{S}{\overline{X}}$$

Ejemplo:

Calcular la varianza, desviación estándar y el coeficiente de variación de la distribución de la siguiente tabla, cuya media es: .

Tabla 26. Ejemplo de cálculo desviación estándar y coeficiente de variación en datos agrupados

Puntajes	f_i	X_i	$X_i - \overline{X}$	$(X_i - \overline{X})$	$(X_i - \overline{X})^2 * f_i$
190-220	13	205	-47.4	2.246,76	29.207,88
220-250	17	235	-17.4	302,76	5.146,92
250-280	6	265	12.6	158,76	952,56
280-310	6	295	42.6	1.814,76	10.888,56
310-340	8	325	72.6	5.270,76	42.166,08
Total	50				88.361,94

Fuente: elaboración propia

$$S^2 = \frac{\sum(X - \overline{X})^2 f}{n} = \frac{88.361,94}{50} = 1.767,24$$

La varianza se interpreta como el promedio de los cuadrados de las desviaciones; por su parte, la desviación estándar es la raíz cuadrada no negativa de la varianza.

$$S = \sqrt{1.767,24} = 42.$$

La desviación estándar se interpreta como la diferencia promedio que presentan los datos con respecto al promedio aritmético; esto es, la variable puntajes puede variar hacia arriba o hacia abajo del promedio en esa cantidad, hasta 3,5 veces. En decir $3,5 * S$.

En una distribución normal, se presenta lo siguiente:

- El 68,26% de la información se encuentra entre el promedio o una desviación estándar.
- El 95,44% entre el promedio $+ o -$ dos (2) desviaciones estándar.
- El 99,74% entre el promedio $+ o -$ tres (3) desviaciones estándar.
- El 99,96% entre el promedio $+ o -$ tres punto cinco (3,5) desviaciones estándar.

$$C_v = \frac{S}{\bar{X}} = 0{,}166 = 16{,}6\% \sim 17\%$$

El coeficiente de variabilidad indica el porcentaje que la desviación estándar representa del promedio; en este caso, la desviación estándar es el 17% del promedio y se utiliza para comparar dos o más grupos, siendo más homogéneo el grupo de menor coeficiente de variabilidad.

2.4 MOMENTOS (MEDIDAS DE FORMA)

Sean $X_1, X_2, \ldots, X_n$ los valores de una variable X y sea $d = X - A$, las desviaciones de la variable X respecto a la constante A.

El momento de orden r respecto al punto A se denota por m_r y se define como sigue:

$$m_r = \frac{\sum (X - A)^r f}{n}$$

Si $A = 0$ se obtiene los momentos con respecto al origen.

En este caso m_r es igual al promedio de X; y se determina como sigue:

$$m_1 = \frac{\sum (X - 0)^1 f}{n} = \frac{\sum Xf}{n}$$

$$m_2 = \frac{\sum (X - 0)^2 f}{n} = \frac{\sum X^2 f}{n}$$

Sí en la fórmula de los momentos se reemplaza $A = \overline{X}$, los momentos se denominan centrales; en este caso, m_r queda determinado así:

$$M_r = \frac{\sum (X - \overline{X})^r f}{n}$$

Se puede probar que:

$$M_1 = 0$$

$$M_2 = m_2 - (m_1)^2$$

$$M_3 = m_3 - 3m_1m_2 + 2(m_1)^3$$

Para evitar unidades particulares, se define los momentos adimensionales a_r de la siguiente manera:

$$a_r = \frac{M_r}{\sqrt{(M_2)^r}}$$

Para el cálculo de los momentos en datos **NO** agrupados, se reemplaza f por en la fórmula de datos agrupados, quedando así:

$$M_r = \frac{\sum(X - \overline{X})^r f}{n} = \frac{\sum(X - \overline{X})^r * 1}{n} = \frac{\sum(X - \overline{X})^r}{n}$$

2.4.1 Coeficiente de asimetría

Con el cálculo de los momentos, además de conocer la media, la varianza y la desviación estándar, también se puede medir el grado de sesgo o de asimetría de la distribución. Se define así:

$$a_3 = \frac{M_3}{\sqrt{(M_2)^3}}$$

Si $a_3 = 0$, la distribución es simétrica; si $a_3 > 0$, la distribución tiene sesgo positivo, es decir la gráfica de la distribución tiene una cola a la derecha; y si $a_3 < 0$, la asimetría es negativa, o sea, la gráfica de la distribución tiene una cola a la izquierda.

2.4.2 Coeficiente de curtosis

El coeficiente de curtosis mide el grado de apuntamiento de la distribución y se mide utilizando el momento M_4, de la siguiente manera:

$$a_4 = \frac{M_4}{(M_2)^2} - 3$$

Si $a_4 = 0$, la gráfica de la distribución es mesocúrtica, o sea con elevación normal; si $a_4 > 0$, la distribución es leptocúrtica, es decir, puntiaguda; y si $a_4 < 0$, la distribución es platicúrtica, o sea, la gráfica de la distribución se presenta achatada.

Ejemplo:

La tabla que sigue presenta el valor de matrícula en miles de pesos pagada por 50 estudiantes de la Universidad de Nariño, en un período académico.

Cuadro 5. Datos hipotéticos de pago de matrícula de 50 estudiantes de la Universidad de Nariño

100	130	140	165	166	180	180	230	268	270
278	280	290	300	320	323	325	328	340	350
360	361	362	362	365	368	369	369	370	370
380	385	386	392	395	400	410	415	416	417
420	450	458	459	500	520	521	524	574	594

Fuente: elaboración propia

Tabla 27. Ejemplo de cálculo de momentos con datos de pago de matrícula de 50 estudiantes de la Universidad de NariñoWW

L_i	L_s	f	x	xf	$(x-\overline{x})$	$(x-\overline{x})f$	$(x-\overline{x})^2 f$	$(x-\overline{x})^3 f$	$(x-\overline{x})^4 f$
100	170	5	135	675	-215,84	-1079,2	232935	-50276589	10851698867
171	241	3	206	618	-144,84	-434,52	62936	-9115632	1320308196
242	312	6	277	1662	-73,84	-443,04	32714	-2415607	178368435,3
313	383	17	348	5916	-2,84	-48,28	137	-389	1105,916357
384	454	11	419	4609	68,16	749,76	51104	3483224	237416562,3
455	525	6	490	2940	139,16	834,96	116193	16169423	2250136843
526	596	2	561	1122	210,16	420,32	88334	18564368	3901487634
						0,00	584352,7	-23591202	18739417644
						0,00	11687,1	-471824,0	374788352,9
						M_1	M_2	M_3	M_4

Fuente: elaboración propia

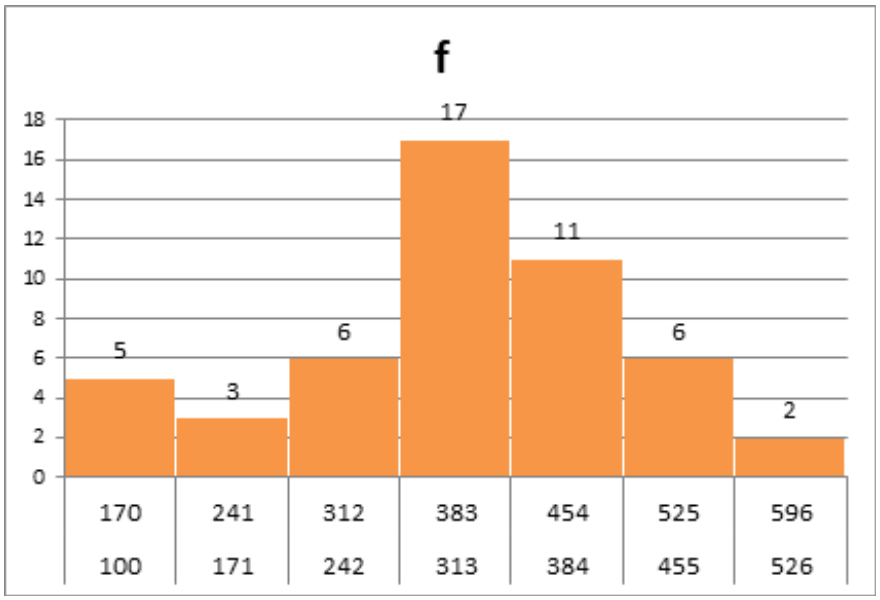
$$\overline{x} = 350{,}84 \qquad C_v = 31\%$$

$$S^2 = 11687{,}1 \qquad a_3 = -0{,}37$$

$$S = 108{,}11 \qquad a_4 = -0{,}26$$

Interpretación: dado que el coeficiente de asimetría está contenido en el intervalo abierto $(-0{,}5, 0{,}5)$ se puede concluir que los datos correspondientes a los valores de matrícula tienen una distribución simétrica. De igual manera, el coeficiente de curtosis cumple esta condición, por lo tanto, la distribución es mesocúrtica.

Gráfico 10. Distribución de con datos de la tabla 27



Fuente: elaboración propia

2.5 ESTANDARIZACIÓN DE UNA VARIABLE

Si una distribución es simétrica, es decir, aproximadamente normal, se espera que el promedio aritmético coincida con la mediana y la moda. Cualquier valor de X se puede transformar en una nueva variable estandarizada Z, utilizando la siguiente expresión:

$$Z = \frac{X - \overline{X}}{S}$$

Mediante la distribución normal, se puede calcular cuartiles, deciles percentiles, rangos o cualquier otro cálculo porcentual. A la variable se la denomina **variable tipificada o variable estandarizada.**

El proceso de determinación de la variable Z correspondiente a una variable aleatoria normal X, se denomina estandarización de la variable.

Ejemplo:

Suponga que 3.000 estudiantes de bachillerato presentaron las Pruebas de Estado Saber 11, en esta la ciudad de Pasto y que la calificación promedio fue de 280 puntos con una desviación estándar de 20 puntos. Suponga que un alumno obtuvo 300 puntos. Se pide lo siguiente:

A) Calcular el porcentaje de estudiantes que están por debajo de 300 puntos.

B) Determinar la nota mínima del 10% superior.

Solución:

Primero se trasforma el valor de en ; luego, en la tabla de la distribución normal, se busca el área correspondiente.

$$Z = \frac{X - \overline{X}}{S} = \frac{300 - 280}{20} = 1$$

- En la tabla de la distribución normal se encuentra que el área hasta es , lo cual indica que el de estudiantes obtuvo menos de puntos.
- Si en la cola derecha de la gráfica se ubica el 10%, entonces, a la izquierda queda el 90%. Se busca este valor en la tabla de la distribución normal, encontrándose que $Z = 1{,}28$..

De la fórmula de Z, se obtiene X:

$$Z = \frac{X - \overline{X}}{S} \rightarrow X = ZS + \overline{X}$$

$$X = ZS + \overline{X} = 1{,}28(20) + 280 = 305{,}6 = 306$$

Interpretación: el 10% de los estudiantes obtuvo un puntaje de 306 puntos o más.

Ejercicio:

Considerando los datos de la tabla 19, determinar los valores de las medidas de tendencia central, variabilidad, posición y forma, aplicando los conceptos de momentos de primero, segundo, tercero y cuarto orden.

CAPÍTULO 3.
REGRESIÓN Y CORRELACIÓN

CAPÍTULO 3.
REGRESIÓN Y CORRELACIÓN

Todas las medidas estadísticas descritas anteriormente permiten hacer un análisis de una variable, pero hay casos en los cuales se necesita analizar valores apareados, correspondientes a dos variables, con el fin de determinar si existe o no relación entre ellas y, en caso afirmativo, determinar el tipo de relación.

Si por cada medida de la variable X, existe un valor correspondiente en la variable Y, entonces, el conjunto resultante de parejas de valores se denomina distribución bivariable.

Ejemplos:

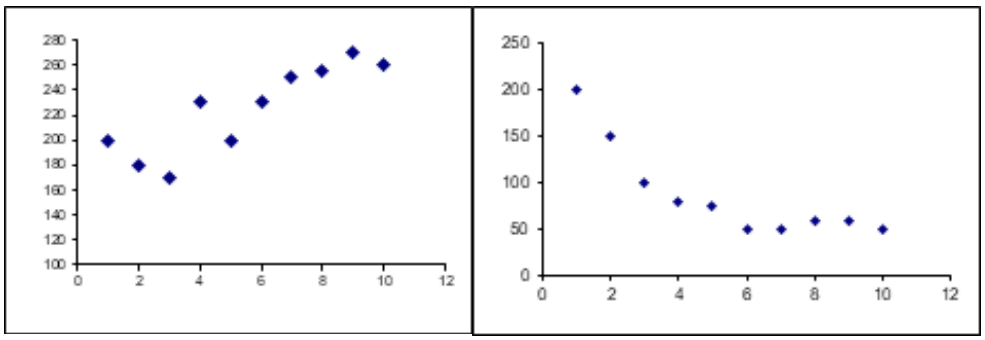
- La edad de los estudiantes y su rendimiento académico: X edad, Y rendimiento.
- Las notas de primer y segundo parcial: X notas primer parcial, Y notas segundo parcial.
- Estatura y peso de un grupo de personas: X estatura, Y peso.

Cuando se trata únicamente de dos variables, se aplica la regresión y correlación simple; y para más de dos variables se utiliza regresión y correlación múltiple.

Una manera sencilla para determinar algún tipo de correlación, es representar las parejas en un plano cartesiano y observar el diagrama de puntos; si estos tienden a ubicarse alrededor de una recta, la correlación es lineal; además, si Y tiende a aumentar cuando X aumenta, la correlación es positiva o directa; en cambio, si la variable Y disminuye cuando aumenta X, la relación es negativa o inversa. Por su parte, si los puntos están cerca de una curva, la relación es no lineal; y si no se observa trayectoria alguna en los puntos, no hay relación entre las variables.

En el Gráfico 11, se puede observar que la primera gráfica indica una trayectoria lineal y positiva, y la segunda muestra una trayectoria no lineal y negativa.

Gráfico 11. Análisis de correlación



Fuente: elaboración propia

3.1 COEFICIENTE DE CORRELACIÓN

En forma cuantitativa, se puede medir el grado de relación entre dos variables X, Y a través del coeficiente de correlación definido por Pearson de la siguiente manera:

$$r(x,y) = \frac{cov(x,y)}{S_x S_y}$$

Donde, el numerador $Cov(x,y)$ indica la covarianza entre las variables, la cual se determina con la siguiente fórmula:

$$Cov(x,y) = M_{xy} - M_x M_y$$

Donde:

$M_{xy} = media\ del\ procucto\ XY$

$M_x = media\ de\ X$

$M_y = media\ de\ Y$

En términos de sumatorias, el coeficiente de correlación se obtiene de la siguiente manera:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Ejemplo:

Sean X y Y las notas obtenidas por seis (6) estudiantes en los cursos de matemáticas y física, respectivamente.

Estudiantes	x	y	xy	x^2	y^2
A	3	5	15	9	25
B	4	4	16	16	16
C	5	5	25	25	25
D	6	5	30	36	25
E	7	6	42	49	36
F	8	5	40	64	25
Total	**33**	**30**	**168**	**199**	**152**
Media	**5,5**	**5**	**28**	**33,17**	**25,33**

$$M_x = 5{,}5$$

$$M_y = 5{,}0$$

$$M_{xy} = 28$$

$$Cov(x,y) = M_{xy} - M_x M_y = 28 - 5{,}5 * 5 = 0{,}5$$

$$Var(x) = 33{,}17 - (5{,}5)^2 = 2{,}92$$

$$S_x = 1{,}70$$

$$Var(y) = 25{,}33 - 5^2 = 0{,}33$$

$$S_x = 0{,}57$$

$$r = \frac{0{,}5}{1{,}.70 - 0{,}.57} = 0.515$$

El coeficiente de correlación es un número comprendido entre -1 y 1. Se interpreta de la siguiente manera:

Tabla 28.Interpretación de coeficientes de correlación

Coeficiente de correlación	Interpretación
0,0 - 0,2	*Insignificante*
0,2 - 0,4	*Baja*
0,4 - 0,6	*Moderada*
0,6 - 0,8	*Alta*
0,8 - 1,0	*Muy alta*

Cuando el coeficiente de correlación es -1 o 1, la correlación entre las variables es perfecta; gráficamente se puede observar que todos los puntos se ubican en una misma línea. En la medida en que r se acerca a cero por la derecha o por la izquierda, los puntos se alejan de la línea.

Las fórmulas vistas anteriormente son válidas únicamente para la correlación lineal, pero si se trata de correlación no lineal es necesario otra fórmula que incluya la ecuación de la curva a la cual se desea ajustar los datos. El estudio de las ecuaciones que se ajustan a datos observados se llama **Regresión.**

3.2 REGRESIÓN LINEAL

Cuando se trata de ajustar los datos de las variables x, y a una recta de regresión, se procede así:

Sea $y = a + bx$ la ecuación de la recta buscada.

Se plantea las ecuaciones de la recta:

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

Resolviendo el sistema de ecuaciones planteado, se obtiene los valores de los coeficientes a y b, donde:

$$a = \frac{\sum y - b\sum x}{n} \rightarrow a = \overline{y} - b\overline{x}$$

$$b = \frac{cov(x,y)}{var(x)}$$

El coeficiente b se denomina coeficiente de regresión lineal; indica los cambios de la variable y por cada cambio unitario de la variable x.

Ejemplo:

La tabla 29 contiene la demanda de cupos escolares entre 1982 y 1988 en una institución de educación media.

Tabla 29.Demanda hipotética de cupos escolares 1982-1988

AÑO	DEMANDA DE CUPOS
1.982	200
1.983	180
1.984	230
1.985	230
1.986	255
1.987	270
1.988	260

Una vez conocidos los valores estimados de la variable y, se puede medir la dispersión alrededor de la recta de regresión utilizando un método similar al de método de la desviación típica o estándar, sólo que, en este caso, la recta de regresión hace las veces de promedio del fenómeno observado.

La desviación típica de la estimación y sobre x, se obtiene así:

$$S_{y.x} = \sqrt{\frac{\Sigma(y - y_e)^2}{n - 2}}$$

De esta manera, se construye intervalos de $68\%, 95\%$ o 99% para la estimación, sumando y restando una, dos o tres desviaciones estándar a los valores estimados, respectivamente.

Hasta el momento se ha utilizado solamente la regresión lineal, es decir, se ha ajustado los valores a una recta. Si el coeficiente de correlación lineal es próximo a cero, entonces, casi no hay correlación lineal entre las variables; sin embargo, puede existir una alta correlación NO lineal de los datos; por lo tanto, es necesario otra fórmula para medir la correlación en términos de la ecuación de regresión que se utilice. Tal expresión está dada por:

$$r = \sqrt{\frac{\Sigma(y_e - \bar{y})^2}{\Sigma(y - \bar{y})^2}}$$

Los valores estimados se pueden calcular mediante diversos tipos de ecuaciones; las más comunes son regresión lineal, regresión parabólica, regresión exponencial y relación geométrica (potencial), cuyas ecuaciones son las siguientes:

a) Regresión lineal: $Y = a + bx$

b) Regresión parabólica: $Y = a + bx + cx^2$

c) Regresión exponencial: $Y = a * b^x$

Regresión geométrica o potencial: $Y = a * x^b$

El método de ajuste de cualquiera de estas ecuaciones es el mismo de la recta; con la excepción que las dos últimas se deben expresar en términos de logaritmos.

Ejemplo:

Determinar la ecuación de la recta para los datos de la tabla 30.

Tabla 30. Ejemplo para cálculo de regresión

	x (años)	y (demanda)	xy	x^2
	1	*200*	*200*	*1*
	2	*180*	*360*	*4*
	3	*230*	*690*	*9*
	4	*230*	*920*	*16*
	5	*255*	*1.275*	*25*
	6	*270*	*1.620*	*36*
	7	*260*	*1.820*	*49*
Total	**28**	**1.625**	**6.885**	**140**
Promedio	**4**	**232,14**	**983,57**	**20**

Fuente: elaboración propia

$$b = \frac{cov(x,y)}{var(x)}$$

$$Cov(x,y) = 983{,}57 - 4 * 232{,}14 = 55{,}01$$

$$Var(x) = 20 - 16 = 4$$

$$b = \frac{55{,}01}{4} = 13{,}75$$

$$y = a + bx$$

$$a = 232{,}14 - 13{,}75 * 4 = 177{,}14$$

Por lo tanto, la ecuación de la recta es: $y = 177{,}14 + 13{,}75x$

Esta ecuación permite estimar valores de la variable y en términos de x. Por ejemplo, la demanda estimada de cupos para 1.989 y 1.990 es:

$$y_{e\ para\ 1989} = 177{,}14 + 13{,}75(8) = 287{,}14 \sim 287$$

$$y_{e\ para\ 1990} = 177{,}14 + 13{,}75(9) = 301$$

De igual manera, se puede calcular los valores estimados para cada uno de los años anteriores.

x	y	y_e	$(y-y_e)^2$	$(y-M_y)^2$	$(y_e-M_y)^2$
1	200	190,89	82,9921	1.032,98	1.701,56
2	180	204,64	607,1296	2.718,58	756,25
3	230	218,39	134,7921	4,5796	189,0625
4	230	232,14	4,5796	4,5796	0
5	255	245,89	82,9921	522,5796	189,0625
6	270	259,64	107,3296	1.433,38	756,25
7	260	273,39	179,2921	776,1796	1.701,56
Total	**1625**	**1625**	**1199,11**	**6492,86**	**5293,7**

$$M_y = 232{,}14$$

$$r = \sqrt{\frac{\Sigma(y_e - M_y)^2}{\Sigma(y - M_y)^2}}$$

$$r = \sqrt{\frac{5.293{,}7}{6.492{,}86}} = 0{,}90$$

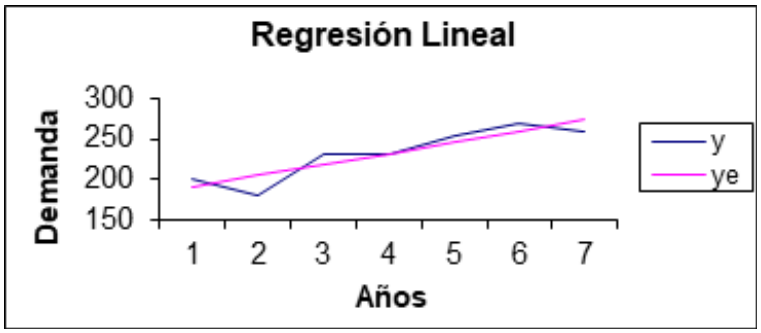
$$S_{y.x} = \sqrt{\frac{\Sigma(y - y_e)^2}{n-2}} = \sqrt{\frac{1199{,}1}{5}} = 15$$

Si al valor estimado para 1.989 se le suma y se le resta una desviación estándar, queda determinado un intervalo del 68% de confianza para la estimación.

$$y_e + S_{yx} = 287 \pm 15$$

Este resultado indica que la demanda de cupos para el año 1.989 oscila entre 272 y 302.

Gráfico 12. Ejemplo de regresión lineal



Fuente: elaboración propia

3.3 REGRESIÓN NO LINEAL

A continuación se indica el procedimiento para determinar la ecuación de regresión no lineal.

3.3.1 Función potencial

$$y = a(x^b)$$

Aplicando logaritmos, se obtiene:

$$log(y) = log(a) + b\big(log(x)\big)$$

Para la denotación de los logaritmos se suele utilizar *mayúsculas*, pero aquí se usará minúsculas

$$Y = A + bX$$

Tabla 31. Producción de café en Kg. Cafecol 1997

Mes	x	y	$log(x)$	$log(y)$	$(log(x)) * (log(y))$	$(log(x))^2$	$y_e = ax^b$
			X	Y	XY	X^2	
Enero	*1*	*25*	*0*				*17,4*
Febrero	*2*	*15*	*0,301*	*1,18*	*0,354*	*0,091*	*24,7*
Marzo	*3*	*28*	*0,477*	*1,45*	*0,69*	*0,228*	*30,3*
Abril	*4*	*35*	*0,602*	*1,54*	*0,93*	*0,362*	*35,1*
Mayo	*5*	*30*	*0,699*	*1,48*	*1,032*	*0,489*	*39,3*
Junio	*6*	*70*	*0,778*	*1,85*	*1,436*	*0,606*	*43,1*
Total	**21**	**203**	**2,857**	**8,887**	**4,442**	**1,775**	

Fuente: Cafecol 1997

$$b = \frac{n\sum log\, x\, log\, y - \sum log\, x \sum log\, y}{n\sum(log\, x)^2 - (\sum log\, x)^2}$$

$$b = \frac{6(4{,}442) - 2{,}857(8{,}887)}{6(1{,}775) - (2{,}857)^2}$$

$$b = 0{,}507$$

$$A = \frac{\sum log\, y - b\sum log\, x}{n}$$

$$A = \frac{8.887 - 0.507 * 2.857}{6}$$

$$A = 1{,}24$$

$$a = 10^{1{,}24} = 17{,}38$$

$$y = ax^b$$

$$y_{e1} = 17{,}38(1)^{0{,}507} = 17{,}4$$

$$y_{e2} = 17{,}38(2)^{0{,}507} = 24{,}7$$

$$y_{e3} = 17{,}38(3)^{0{,}507} = 30{,}3$$

$$y_{e4} = 17{,}38(4)^{0{,}507} = 35{,}1$$

$$y_{e5} = 17{,}38(5)^{0{,}507} = 39{,}3$$

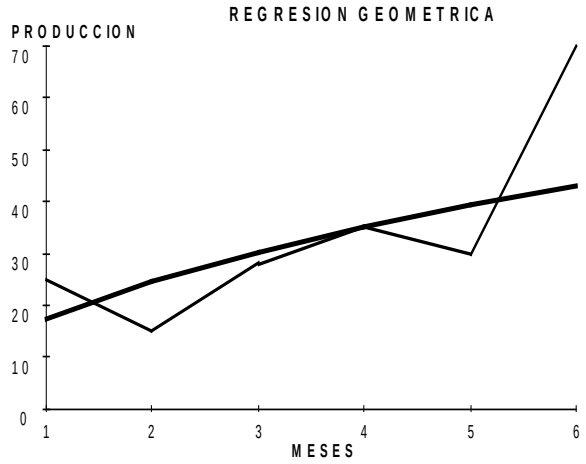
$$y_{e6} = 17{,}38(6)^{0{,}507} = 43{,}1$$

La producción de café en el mes de marzo, fue de 28.000 kilos; se esperaba una producción de 30.300 kilos; por lo tanto, la producción real no fue como se proyectaba.

En septiembre, la producción esperada e kilos, será:

$$y = 17{,}38(9^{0{,}507}) = 52{,}948$$

Gráfico 13. Ejemplo de Regresión Geométrica



Fuente: elaboración propia

3.3.2. Función exponencial

$$y = a(b^x)$$

Aplicando logaritmos, se obtiene:

$$log(y) = log(a) + x\big(log(b)\big) \quad (*)$$

Observe que para la denotación de logaritmo se utiliza minúsculas.

Sean las siguientes relaciones: $Y = log(y);\ A = log(a);\ B = log(b)$

Reemplazando en la ecuación (*), se obtiene:

$$Y = A + Bx$$

Tabla 32. Producción de café en Kg. Cafecol 1997

Mes	x	y	$log(y)$	$x\big(log(y)\big)$	x^2	$y_e = a(b)^x$
			Y	xY		
Enero	1	25	1,4	1,4	1	17,84
Febrero	2	15	1,18	2,35	4	22,14
Marzo	3	28	1,45	4,34	9	27,48
Abril	4	35	1,54	6,18	16	34,1
Mayo	5	30	1,48	7,39	25	42,32
Junio	6	70	1,85	11,07	36	52,52
Total	21	203	8,887	32,72	91	

Fuente: Cafecol 1997

$$B = \frac{n\sum x \log y - \sum x \sum \log y}{n\sum x^2 - (\sum x)^2}$$

$$B = \frac{6(32{,}72) - 21(8{,}887)}{6(91) - (21^2)} = 0{,}0924$$

$$B = log(b) = 0{,}0924$$

$$b = antilog(B) = antilog(0{,}924)$$

$$b = 10^{0{,}0924} = 1{,}24$$

$$A = \frac{\sum log\, y - B\sum x}{n}$$

$$A = \frac{8{,}887 - 0{,}0924(21)}{6} = 1{,}158$$

$$a = antilog(A)$$

$$a = 10^{1{,}58} = 14{,}377$$

$$y = a(b^x)$$

$$y_{e1} = 14{,}377(1{,}24)^1 = 17{,}84$$

$$y_{e2} = 14{,}377(1{,}24)^2 = 22{,}14$$

$$y_{e3} = 14{,}377(1{,}24)^3 = 27{,}48$$

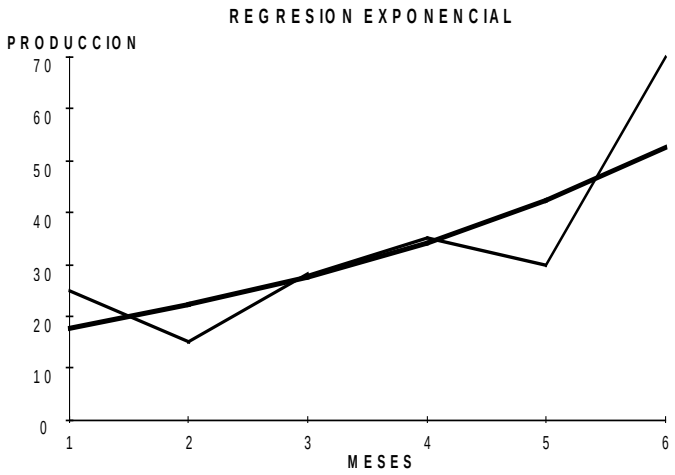
$$y_{e4} = 14{,}377(1{,}24)^4 = 34{,}10$$

$$y_{e5} = 14{,}377(1{,}24)^5 = 42{,}32$$

$$y_{e6} = 14{,}377(1{,}24)^6 = 52{,}52$$

Se puede observar que en el mes de abril la producción de café en Cafecol fue de 35.000 kilos y se esperaba una producción de 34.100 kilos.

Gráfico 14. Ejemplo de Regresión Exponencial



Fuente: elaboración propia

Para calcular la producción esperada en el mes de diciembre, se procede así:

$$y = 14{,}377(1{,}241^{12}) = 191{,}836$$

Entonces, para el mes de diciembre se espera que la producción de café en Cafecol sea de 191.835 kilos.

3.3.3 Función cuadrática

$$Y = a + bX + cX^2$$

Para aplicar el sistema de ecuaciones simplificadas, es necesario construir una nueva variable x la cual se obtiene restando el promedio aritmético a la variable original; es decir:

$$x = X - \overline{X}$$

Tabla 33. Ejemplo de Regresión Cuadrática

Mes	X	x	y	xy	x^2	x^2y	x^4	$y = a + bx + cx^2$
Enero	1	-2.5	25	-62.5	6.25	156.25	39.06	24.62
Febrero	2	-1.5	15	-22.5	2.25	33.75	5.06	19.85
Marzo	3	-0.5	28	-14	0.25	7	0.06	21.42
Abril	4	0.5	35	17.5	0.25	8.75	0.06	29.32
Mayo	5	1.5	30	45	2.25	67.5	5.06	43.57
Junio	6	2.5	70	175	6.25	437.5	39.06	64.16
Total	21	0	203	138.5	17.5	710.75	88.36	202.94

Fuente: Cafecol 1997

$$y = a + bx + cx^2$$

Se plantean el siguiente sistema de ecuaciones, aplicando sumatoria, multiplicando por x y por x^2, de la siguiente manera:

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

Sea $\sum x = 0$,entonces $\sum x^3 = 0$, con lo cual, el sistema de ecuaciones queda así:

$$\sum y = na + \sum x^2$$

$$\sum xy = b\sum x^2$$

$$\sum x^2y = a\sum x^2 + c\sum x^4$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{138,5}{17,5} = 7,91$$

$$c = \frac{n\sum x^2y - \sum x^2 \sum y}{n\sum x^4 - (\sum x^2)^2}$$

$$c = \frac{6(710,75) - 17,5(203)}{6(88,36) - (17,5)^2} = 3,17$$

$$a = \frac{\sum y - c\sum x^2}{n}$$

$$a = \frac{203 - 3.17(17,5)}{6} = 24,58$$

Dado que:

$$Y = a + bX + cX^2$$

Entonces, se tiene la siguiente ecuación:

$$y_e = 24{,}58 + 7{,}91(x) + 3{,}17(x^2)$$

$$y_{e1} = 24{,}58 + 7{,}91(-2{,}5) + 3{,}17(-2{,}5)^2 = 24{,}62$$

$$y_{e2} = 24{,}58 + 7{,}91(-1{,}5) + 3{,}17(-1{,}5)^2 = 19{,}85$$

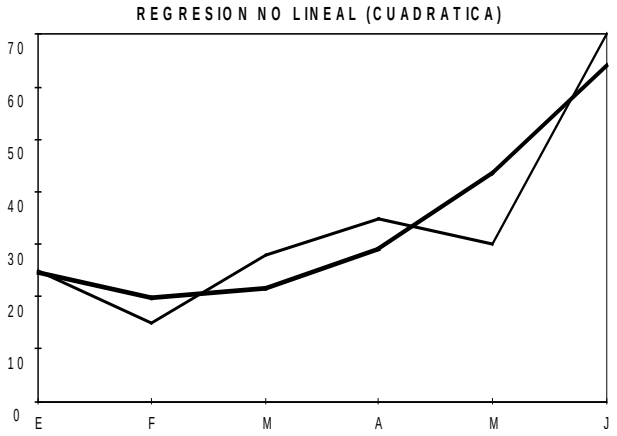
$$y_{e3} = 24{,}58 + 7{,}91(-0{,}5) + 3{,}17(-0{,}5)^2 = 21{,}42$$

$$y_{e4} = 24{,}58 + 7{,}91(0{,}5) + 3{,}17(0{,}5)^2 = 29{,}32$$

$$y_{e5} = 24{,}58 + 7{,}91(1{,}5) + 3{,}17(1{,}5)^2 = 43{,}53$$

$$y_{e6} = 24{,}58 + 7{,}91(2{,}5) + 3{,}17(2{,}5)^2 = 64{,}16$$

Gráfico 15. Ejemplo de Regresión Cuadrática



Fuente: elaboración propia

CAPÍTULO 4.
SERIES CRONOLÓGICAS

CAPÍTULO 4.
SERIES CRONOLÓGICAS

Son herramientas que se utilizan para el análisis de datos de variables que cambian a través del tiempo.

4.1 ANÁLISIS DE SERIES CRONOLÓGICAS

Cuando se observa una variable a través del tiempo, se obtiene una serie cronológica.

Para realizar el análisis, la serie se descompone en cuatro factores: Tendencia (T), Variaciones Estacionales (S), Variaciones Cíclicas (C) y Variaciones Irregulares (I); de tal manera que la variable dependiente se puede expresar como el producto de los cuatro factores anteriores, así:

$$Y = T * S * C * I$$

$T * C$ = promedio móvil centrado

$S * I$ = porcentaje del promedio móvil

$T * S$ = tendencia estacional

$C * I$ = porcentajes irregulares y cíclicos.

Generalmente, la serie cronológica muestra los datos por trimestres, en diferentes años.

Utilizando una hoja electrónica, los datos se pueden organizar como se indica en el ejemplo que sigue.

Columna A: Años

Columna B: Trimestres

Columna C: Enumeración correlativa de trimestres (X)

Columna D: Variable a estudiar (Y)

Una vez introducida la información, se procede a descomponer la serie, de la siguiente manera:

1) En la *Columna* E, se determina la tendencia (T), aplicando una ecuación de regresión.

2) *Columna* F, se determina el promedio móvil de Y en bloques de cuatro trimestres.

3) *Columna* G: se centra el promedio móvil $(T * C)$.

4) *Columna* H: se calcula $S * I$ dividiendo Y entre $T * C$ (Columna G entre Columna H).

5) Se construye una nueva tabla cruzando años y trimestres, y se pone los valores de $S * I$ en el sitio correspondiente. Al promedio trimestral de S*I se lo denota por S, puesto que al promediar se ha eliminado la variación irregular.

6) *Columna* I: se transcribe los valores de S, en el trimestre correspondiente de cada año.

7) *Columna* J: se multiplican los valores $T * S$.

8) *Columna* K: se determina $C * I$ dividiendo Y entre $T * S$.

9) *Columna* L: se elimina lo irregular, calculando el promedio centrado de $C * I$ y se lo denota por C.

10) *Columna* M: se divide $C * I$ entre C y se determina las variaciones irregulares.

La tabla que sigue muestra los encabezados que se pueden utilizar en una hoja de cálculo.

Tabla 34. Modelo de tabla para el análisis de series cronológicas

A	B	C	D	E	F	G	H	I	J	K	L	M
Año	Tri	X	Y	T	Móvil	T*C	S*I	S	T*S	C*I	C	I

Fuente: elaboración propia.

Ejemplo

La tabla 35, presenta el número de viviendas construidas en Pasto de 2010 a 2014.

Tabla 35.Número de viviendas construidas en Pasto del año 2010 al 2014

Trimestre	2010	2011	2012	2013	2014
I	102	110	111	115	122
II	120	126	128	135	144
III	90	95	97	103	110
IV	78	83	86	91	98

Fuente: elaboración propia

			No Vi-viendas	Tendencia	Promedio Móvil	Promedio centrado	% del pro.M	I. Est.	Proyecc
AÑOS	TRIM	X	Y	T	Vrs. de Y	T*C	S*I	S	T*S
2010	I	1	102	99,7				1,06	106,0
	II	2	120	100,5				1,23	123,4
	III	3	90	101,3	100,9	101,3	0,89	0,91	91,9
	IV	4	78	102,0	101,7	102,0	0,76	0,79	80,7
2011	I	5	110	102,8	102,4	102,8	1,07	1,06	109,4
	II	6	126	103,6	103,2	103,6	1,22	1,23	127,3
	III	7	95	104,4	104,0	104,4	0,91	0,91	94,8
	IV	8	83	105,2	104,8	105,2	0,79	0,79	83,2
2012	I	9	111	106,0	105,6	106,0	1,05	1,06	112,8
	II	10	128	106,8	106,4	106,8	1,20	1,23	131,2
	III	11	97	107,6	107,2	107,6	0,90	0,91	97,6
	IV	12	86	108,4	108,0	108,4	0,79	0,79	85,7
2013	I	13	115	109,2	108,8	109,2	1,05	1,06	116,2
	II	14	135	110,0	109,6	110,0	1,23	1,23	135,1
	III	15	103	110,8	110,4	110,8	0,93	0,91	100,5
	IV	16	91	111,6	111,2	111,6	0,82	0,79	88,2
2014	I	17	122	112,4	112,0	112,4	1,09	1,06	119,5
	II	18	144	113,1	112,7	113,1	1,27	1,23	139,0
	III	19	110	113,9	113,5			0,91	103,4
	IV	20	98	114,7				0,79	90,7
2015	I	21		115,5				1,06	122,9
	II	22		116,3				1,23	142,9
	III	23		117,1				0,91	106,3
	IV	24		117,9				0,79	93,2

Fuente: elaboración propia.

4.2 ECUACIÓN DE TENDENCIA

Pendiente $b = 0{,}79$

Intersección con el eje Y, $a = 98{,}9$

La ecuación de la recta para estimar el número de viviendas construidas en cada trimestre es $y = 98{,}9 + 0{,}79X$

Una vez encontrada la tendencia (T) y los Índices Estacionales (S), el producto T*S da una proyección muy cercana a la realidad. Este método permite proyectar, con mucha exactitud, el número de viviendas a construirse en el año siguiente.

Gráfico 16. Ejemplo de viviendas construida. Construcción propia.

Fuente: elaboración propia.

4.3 TALLER

1. Señale la respuesta correcta.

El valor de cada observación se tiene en cuenta para calcular el aritmético	**V**	**F**
Una muestra representativa es cualquier subconjunto de la población	**V**	**F**
Los valores extremos de un conjunto de datos tienen un fuerte efecto sobre la mediana	**V**	**F**
La suma de las frecuencias relativas es igual a n	**V**	**F**
Podemos calcular el promedio aritmético en datos desordenados	**V**	**F**
Un parámetro es una característica de la muestra	**V**	**F**
La media geométrica se utiliza para medir los cambios de una variable a través del tiempo	**V**	**F**
Las medidas de tendencia central se utilizan para medir la variabilidad de los datos.	**V**	**F**
La media aritmética es el punto de equilibrio de una serie de datos	**V**	**F**
La media geométrica es una medida de asimetría.	**V**	**F**

2. Los puntajes en una prueba de aptitud matemática de 50 estudiantes de la universidad de Nariño, son:

100	130	140	165	166	180	180	230	268	270
278	280	290	300	320	323	325	328	340	350
360	361	362	362	365	368	369	369	370	370
380	385	386	392	395	400	410	415	416	417
420	450	458	459	500	520	521	524	574	594

Forme una distribución de frecuencias con intervalos de igual amplitud; determine e interprete: el promedio aritmético, la mediana, los cuarteles, el percentil 80, la varianza, la desviación estándar y el coeficiente de variabilidad. Determine el porcentaje de estudiantes que tienen puntajes en el intervalo:

$$(\overline{x} - s, \overline{x} + s)$$

3. El conteo bacterial de cierto cultivo pasó de 1000 a 4000 en tres días. Determinar lo siguiente:

a) El incremento promedio porcentual por día.

b) El factor de crecimiento diario.

c) El número de días en el cual se duplicaría el valor inicial

4.-Al señor Pérez, empleado de la compaña Ecopetrol, le consignaron su cesantía en Horizonte, por valor de $18.000.000,oo. Si el rendimiento es del 3% mensual, calcular:

a) El valor de su cesantía después de un año.

b) E factor de crecimiento.

c) El tiempo en el cual se duplicaría el valor inicial de su cesantía.

5. Las exportaciones de Café, en millones de dólares en los últimos años fueron las siguientes:

Año	Miles de dólares
1998	3.000
1999	4.000
2000	3.800
2001	5.000
2002	6.500
2003	8.000
2004	7.200
2005	8.800
2006	9300

Determinar y representar gráficamente lo siguiente:

a) Factores de crecimiento, índice acumulado y la tasa promedio de crecimiento anual.

b) La ecuación de la recta $Y = a + bX$ y estime el valor de las exportaciones para los años 2007 y 2008.

c) La ecuación geométrica o potencial: $Y = a * X^b$ y los valores estimados para los años 2007 y 2008.

d) La ecuación exponencial $Y = a * b^x$ y los valores estimados para los años 2007 y 2008.

e) La ecuación de regresión cuadrática $Y = a + bX + cX^2$

6. Defina los conceptos de Estadística, población, muestra, datos, variable, estadígrafo y parámetro. ¿Cuántos tipos de variable se manejan en Estadística?. Indique un ejemplo de cada uno. Elabore una lista de las medidas estadísticas: de tendencia central, posición, variabilidad, asimetría y apuntamiento.

7. La siguiente tabla muestra la distribución salarial de 65 empleados

Salario (miles de pesos)	No. de empleados
500 – 600	*20*
600 – 700	*18*
700 – 800	*14*
800 – 900	*8*
900 – 1000	*5*
Total	**65**

Encontrar: todas las medidas estadísticas.

8. El fondo de empleados de una empresa desea realizar un estudio estadístico de las cuotas mensuales pagadas por 50 deudores. Los datos en miles de pesos son los siguientes:

150	*150*	*180*	*165*	*166*	*167*	*255*	*267*	*268*	*270*
278	*280*	*290*	*300*	*320*	*323*	*325*	*328*	*340*	*350*
360	*361*	*362*	*362*	*365*	*368*	*369*	*369*	*370*	*370*
380	*385*	*386*	*392*	*395*	*400*	*410*	*415*	*416*	*417*
420	*450*	*458*	*459*	*500*	*520*	*521*	*524*	*574*	*597*

Crear una distribución de frecuencias absolutas, relativas y acumuladas.

a) Encontrar e interpretar el promedio aritmético, la varianza y la desviación estándar, los cuartiles, el percentil 40 y percentil 80, coeficiente de asimetría y coeficiente de curtosis.

b) Representar gráficamente la distribución de frecuencias correspondiente.

Para calcular la varianza utilice cada uno de los siguientes métodos:

a) $S^2 = \frac{\sum(x-\bar{x})^2 f}{n}$

b) $S^2 = \sum \frac{x^2 f}{n} - \left(\frac{\sum xf}{n}\right)^2$

c) $S^2 = c^2 \left[\sum \frac{u^2 f}{n} - \left(\frac{\sum uf}{n}\right)^2\right]$

9. Si el precio de un artículo se triplica en un periodo de 4 meses, ¿cuál es el incremento promedio mensual?

10. Las calificaciones de un estudiante en la asignatura de Estadística son las siguientes: 7,1;7,8 y 5,9. Si los pesos asignados a estas notas, son respectivamente:2; 4 y 5, determinar lo siguiente:

a) El promedio adecuado.

b) El promedio si todos los pesos fueran iguales.

11. Tres profesores de Economía registraron calificaciones medias de sus exámenes, así: 79; 82 y 84. Sus clases estaban conformadas por 32, 25 y 17 estudiantes. ¿Cuál es la calificación media del grupo?

12. Los siguientes datos corresponden a los porcentajes autorizados por el gobierno nacional para el incremento de salarios a trabajadores oficiales en el presente año. Determinar el promedio ponderado.

Porcentaje autorizado, x	Salario
20%	*$ 200.000*
24%	*$ 150.000*
28%	*$ 100.000*
30%	*$ 80.000*
Suma	**$530.000**

13. Sea la siguiente fórmula:

$$S^2 = \frac{\sum(x-\bar{x})^2 f}{n}$$

Demostrar que la varianza también se puede obtener de la siguiente manera:

$$S^2 = \sum \frac{x^2 f}{n} - \left(\frac{\sum xf}{n}\right)^2$$

14. La tabla que sigue presenta el volumen de ventas diarias en un supermercado de la ciudad, durante los últimos días.

Ventas (millones de pesos	Días
2 - 4	5
4 - 6	10
6 - 8	21
10 - 12	4
Total	**40**

Se pide lo siguiente:

a) Determinar los cuatro primeros momentos con respecto al origen y los cuatro momentos centrales.

b) Expresar los momentos centrales en función de los momentos con respecto al origen.

15. La siguiente tabla contiene las utilidades mensuales en miles de pesos, de 200 socios de una empresa comunitaria de la ciudad de Pasto en el año 2007. Determinar los siguiente:

a) La varianza, aplicando tres métodos diferentes.

b) La desviación estándar y el coeficiente de variabilidad.

c) El percentil 75.

d) El porcentaje de socios que tienen utilidades inferiores a $ 35.000.

Miles de pesos	F
20 - 24	12
24 – 28	36
28 - 32	84
32 - 36	52
36 - 40	16
Total	**200**

16. Los precios de un artículo durante los primeros meses del año, fueron los siguientes:

Meses	Precio	Índice	Variación
Enero	1500		
Febrero	1700		
Marzo	1750		
Abril	1650		
Mayo	1800		
Junio	2100		

Determinar:

a) Índices de crecimiento, índice acumulado, índice promedio mensual, el aumento promedio mensual.

b) Las cuatro ecuaciones de regresión mencionadas en el punto 4.

17. La siguiente tabla muestra la altura del padre (X) y del hijo (Y), en pulgadas.

X	65	63	67	64	62	70	72	68	66	69	71
Y	68	66	68	65	66	68	65	71	67	68	70

Determinar:

a) La ecuación de la recta de regresión $Y = a + bX$ y los valores estimados Y_e.

b) El error típico de estimación S_{yx}.

c) Sume y reste el error típico de estimación a todos los valores estimados y construya e interprete un gráfico de líneas.

18. La tabla siguiente muestra la distribución de frecuencias de la vida media de 400 válvulas de radio probadas en la empresa L & M; determinar:

a) La frecuencia relativa de la sexta clase.

b) El porcentaje de válvulas cuya vida media, no pasa de 600 horas.

c) El porcentaje de válvulas cuya vida media es mayor o igual a 900

d) El porcentaje de válvulas cuya vida media es superior al Promedio Aritmético.

Vida media	**Número**
300- 400	14
400-500	46
500-600	58
600-700	98
700-800	74
800-900	62
900-100	48

19. La tabla que sigue contiene la distancia de frenado D en metros y la velocidad V en kilómetros por hora, de un automóvil.

Velocidad	Distancia
20	4
30	7
40	9
50	12
60	14
70	16
80	20

Determinar la ecuación de la parábola $D = a + bV + cV^2$.

20. Utilizando la fórmula abreviada de la variable $u = \frac{x-A}{c}$ calcular el coeficiente de variabilidad para los datos de la tabla 2.

21. La siguiente tabla muestra el índice de desempleo en Colombia, durante la última década.

Años	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
Desempleo	12.7	11.0	10.0	8.4	12.5	13.8	15.7	18.0	19.8	23.4

Se pide lo siguiente:

a) Determinar la ecuación de la recta $Y = a + bX$ y los valores estimados, el error de estimación y los límites del 68% de confianza. Realizar un gráfico de líneas.

b) Determinar las ecuaciones las siguientes ecuaciones:

$$Y = a * b^X \quad Y = a * X^b \quad Y = a + bX + cX^2$$

c) Calcular el coeficiente de correlación en cada uno de los casos, utilizando la siguiente fórmula:

$$r = \sqrt{\frac{\Sigma(Ye-\bar{Y})^2}{\Sigma(Y-\bar{Y})^2}}$$

d) Comparar los resultados del literal c) y decidir cuál es la mejor ecuación de estimación.

22. El salario anual pagado a los empleados de una compañía es de $ 4.500.000.

Los salarios medios pagados a los hombres y a las mujeres fueron de $ 4.950.000 y $ 4.450.000 respectivamente. Determinar el porcentaje de hombres y mujeres empleados en la compañía.

23. La siguiente tabla presenta las utilidades mensuales en miles de pesos en 1998, de 200 socios de una empresa comunitaria de la ciudad de Pasto.

Miles de $	**f**								
20 - 24	*12*								
24 - 28	*36*								
28 - 32	*84*								
32 - 36	*52*								
36 - 40	*16*								
Total	**200**								

Calcular e interpretar lo siguiente:

a) El promedio aritmético y la mediana.

a) La varianza y la desviación estándar.

b) El coeficiente de asimetría.

Acerca de los Autores

Alberto Javier Mesa Guerrero.

Docente adscrito al Departamento de Matemáticas y Estadística, Facultad de Ciencias Exactas y Naturales, Universidad de Nariño. Licenciado en Matemáticas y Física, Universidad de Nariño. Especialista en Computación para la Docencia, Universidad Mariana. Estadístico en Salud, Universidad de Antioquia. Profesor en categoría Asociado de la Universidad de Nariño.

Correo electrónico: soundmesa@yahoo.com.

Segundo Javier Caicedo-Zambrano.

Docente adscrito al Departamento de Matemáticas y Estadística, Facultad de Ciencias Exactas y Naturales, Universidad de Nariño. Doctor en Ciencias de la Educación, Universidad del Tolima. Licenciado en Matemáticas y Física, Universidad de Nariño. Ingeniero de Sistemas, Universidad Antonio Nariño. Especialista en Computación para la Docencia, Universidad Mariana. Especialista en Multimedia Educativa, Universidad Antonio Nariño. Magister en Software Libre, Universidad Autónoma de Bucaramanga. Asesor de Desarrollo Académico, Universidad de Nariño. Profesor en categoría Asociado de la Universidad de Nariño.

Correo electrónico: jacaza l@gmail.com; jacaza l@udenar.edu.co.

índice de Tablas

índice de Cuadros

índice de Gráficos

REFERENCIAS BIBLIOGRÁFICAS

BENJAMIN, J.R. Probabilidad y Estadística en Ingeniería Civil. Ed. McGraw-Hill. Bogotá. 1981

CHAO, Lincoln, Estadística aplicada a las ciencias administrativas. Ed. Mc Graw Hill. México 1.985

CHOU, Ya Lun. Análisis Estadístico. Ed. Interamericana. México. 1977

CANAVOS, George C. Probabilidad y Estadística Ma. Graw Hill-México 1984

HANKE John E./ Reitsh. Estadística para negocios. Mc Graw Hill. México 1994.

HERNÁNDEZ, Roberto y otros. Metodología de la Investigación. Ed. McGraw-Hill. México. 1999.

KAZMIER , Leonard. Estadística para la Administración y la Economía. Serie Schaum. Mc Graw Hill. México 1998.

LEVIN Ricahrd & RUBIN David. Estadística para administradores. Prentice Hall. México 1994.

MARTOS, José. Statgraphics Conceptos y Aplicaciones. Ed. Paraninfo. Madrid. 2001.

MARTINEZ BENCARDINO, Ciro. Estadística. Ed. ECOE. Bogotá 1978.

MENDENHALL, William. Introducción a la probabilidad yu Estadística. Wastworth Internacional.E.E.UU. 1.979

MILLER Y FREUND. JOHNSON RICHARD Probabilidad y Estadística para Ingenieros. 5a. Ed. Prentice May Hispanoamericana S.A. México. 1964.

SPIEGEL, Murray. Estadística. Teoría y 875 problemas resueltos. Serie Schaum. Ed. McGraw-Hill. México. 1970.

Editorial
Universidad de **Nariño**

Introducción a la Estadística Descriptiva

Esta obra surge por el interés de los autores de publicar un libro de texto a nivel introductorio sobre fundamentos de estadística descriptiva, que se constituya en fuente de consulta y de nivelación sobre conceptos básicos de estadística. Por el enfoque, está orientado a estudiantes de los primeros semestres de programas universitarios relacionados con las Ciencias Básicas, Técnicas e Ingenierías, aunque también lo pueden utilizar estudiantes de otras áreas del conocimiento, incluso de programas de educación no formal.

Los temas que se abordan en esta obra, han sido seleccionados con base en la experiencia de los autores, quienes reconocen que se constituye en soporte importante para estudiantes que inician el estudio de la estadística. Para el efecto, se presenta, en forma resumida, la conceptualización, ejemplos y se proponen ejercicios para reforzar lo aprendido. Si bien existen muchos programas para el cálculo de estadísticas, los autores consideran que es importante que los estudiantes realicen los cálculos paso a paso, tal como se ilustra en los ejemplos, porque ayuda para la comprensión e interpretación de los resultados.

El libro está organizado en cuatro (4) capítulos. El primero, "Conceptos Básicos", presenta generalidades de estadística y organización de datos en tablas de frecuencia. En el segundo capítulo, "Medidas estadísticas", se trabaja las medidas de tendencia central, medidas de posición, medidas de variabilidad, momentos, relación y correlación simple, y análisis de series cronológicas. En el tercer capítulo, "Regresión y correlación", se incluye coeficiente de correlación, regresión lineal, regresión no lineal: función potencial, función exponencial y función cuadrática. En el cuarto capítulo, "Series cronológicas", se aborda el análisis de estas series.

ISBN: 978-958-5123-11-3 digital

9 789585 123113

Editorial
Universidad de Nariño