LUIS AMIGÓ
FUNDACIÓN UNIVERSITARIA

ISBN: 978-958-8943-05-3

ELEMENTOS BÁSICOS DE ESTADÍSTICA DESCRIPTIVA
PARA EL ANÁLISIS DE DATOS

Gabriel Jaime Posada Hernández

FONDO
Editorial
LUIS AMIGÓ

4 DISTRIBUCIÓN DE FRECUENCIAS

Objetivos:

- Elaborar la tabla de frecuencias para un conjunto de datos.
- Interpretar la distribución de frecuencias para datos agrupados y sin agrupar.

Contenido:

4.1 Frecuencia absoluta
4.2 Frecuencia relativa
4.3 Frecuencia absoluta acumulada
4.4 Frecuencia relativa acumulada
4.5 Ejercicios de aplicación

4. DISTRIBUCIÓN DE FRECUENCIAS

4.1 Frecuencia absoluta

Se denomina frecuencia absoluta (n_i) a la cantidad de veces que se presenta el valor x_i de la variable X en la muestra o la población. Las frecuencias absolutas para el grupo de personas que realizan una actividad se ilustran en la tabla 8.

Tabla 8. Frecuencias absolutas para el ejemplo del tiempo (en minutos) requerido por un grupo de personas para realizar una actividad

Nº. de Intervalo	Intervalo (minutos)	Frecuencia absoluta (n_i)
1	[45 - 50]	2
2	(50 - 55]	9
3	(55 - 60]	12
4	(60 - 65]	11
5	(65 - 70]	9
6	(70 - 75]	7
Total		50

Elaboración propia

Se debe tener en cuenta que el total de datos debe corresponder a la suma de las frecuencias absolutas.

4.2 Frecuencia relativa

La frecuencia relativa (h_i) se define como el porcentaje de frecuencia absoluta en relación al total de datos de la muestra (n). Se obtiene con el cociente entre la frecuencia absoluta y el total de datos, usando la siguiente ecuación:

$$h_i = \frac{n_i}{n} * 100$$

Siendo n el total de datos.

El cálculo de frecuencias relativas estimadas para el grupo de personas que realizan una actividad se presenta en la tabla 9.

Tabla 9. Frecuencias relativas para el ejemplo del tiempo (en minutos) requerido por un grupo de personas para realizar una actividad

Nº. de Intervalo	Intervalo (minutos)	Frecuencia relativa (h_i)
1	[45 - 50]	(2/50)*100 = 4%
2	(50 - 55]	(9/50)*100 = 18%
3	(55 - 60]	(12/50)*100 = 24%
4	(60 - 65]	(11/50)*100 = 22%
5	(65 - 70]	(9/50)*100 = 18%
6	(70 - 75]	(7/50)*100 = 14%
Total		100%

Elaboración propia

El total de frecuencias relativas debe ser igual al 100%, o aproximadamente igual a este valor cuando se presentan decimales y se redondean los valores del porcentaje.

4.3 Frecuencia absoluta acumulada

La frecuencia absoluta acumulada (N_i) para un valor x_i de una variable X es la adición de las frecuencias absolutas n_i hasta alcanzar la totalidad de los datos. Se calcula mediante la siguiente ecuación:

$$N_i = \sum_{k=1}^{i} n_k$$

Tabla 10. Frecuencias absolutas acumuladas para el ejemplo del tiempo (en minutos) requerido por un grupo de personas para realizar una actividad

Nº. de intervalo	Intervalo (minutos)	Frecuencia absoluta (n_i)	Frecuencia absoluta acumulada (N_i)
1	[45 - 50]	2	2
2	(50 - 55]	9	2 + 9 = 11
3	(55 - 60]	12	2 + 9 + 12 = 23
4	(60 - 65]	11	2 + 9 + 12 + 11 = 34
5	(65 - 70]	9	2 + 9 + 12 + 11 + 9 = 43
6	(70 - 75]	7	2 + 9 + 12 + 11 + 9 + 7 = 50

Elaboración propia

4.4 Frecuencia relativa acumulada

La frecuencia relativa acumulada H_i para un valor x_i de una variable X es la adición de las frecuencias relativas h_i hasta alcanzar la totalidad de los datos. Se calcula mediante la siguiente ecuación:

$$H_i = \sum_{k=1}^{i} h_k$$

Tabla 11. Frecuencias relativas acumuladas para el ejemplo del tiempo (en minutos) requerido por un grupo de personas para realizar una actividad

Nº. de Intervalo	Intervalo (minutos)	Frecuencia relativa (h_i)	Frecuencia absoluta acumulada (H_i)
1	[45 - 50]	4%	4%
2	(50 - 55]	18%	4% + 18% = 22%
3	(55 - 60]	24%	4% + 18% + 24% = 46%
4	(60 - 65]	22%	4% + 18% + 24% + 22% = 68%
5	(65 - 70]	18%	4% + 18% + 24% + 22% + 18% = 86%
6	(70 - 75]	14%	4% + 18% + 24% + 22% + 18% + 14% = 100%

Elaboración propia

Al construir las frecuencias acumuladas, se debe considerar que éstas no aplican a la variable cualitativa.

En la tabla 12 se muestra el consolidado de frecuencias correspondientes al ejemplo del tiempo que requiere un grupo de personas en realizar una actividad.

Tabla 12. Intervalos y frecuencias para el ejemplo del tiempo (en minutos) requerido por un grupo de personas para realizar una actividad

Nº. de Intervalo	Intervalo (Tiempo en minutos)	n_i	h_i	N_i	H_i	$\dot{x}_i$
1	[45 - 50]	2	4%	2	4%	47,5
2	(50 - 55]	9	18%	11	22%	52,5
3	(55 - 60]	12	24%	23	46%	57,5
4	(60 - 65]	11	22%	34	68%	62,5
5	(65 - 70]	9	18%	43	86%	67,5
6	(70 - 75]	7	14%	50	100%	72,5

Elaboración propia

A partir de la tabla anterior, pueden realizarse los siguientes análisis:

» Las frecuencias absolutas y relativas se interpretan a partir del intervalo.

Ejemplo: 2 personas se demoran entre 45 y 50 minutos en realizar la actividad o el 4% de las personas se demoran entre 45 y 50 minutos en realizar la actividad; 9 personas se demoran entre 50 y 55 minutos en realizar la actividad o el 18% de las personas se demoran entre 50 y 55 minutos en realizar la actividad, así sucesivamente.

» Las frecuencias acumuladas (absolutas y relativas) se interpretan utilizando la expresiones *"máximo"*, *"hasta"* o *"a lo sumo"*.

Por ejemplo: 2 personas se demoran máximo 50 minutos en realizar la actividad o 4% de las personas se demoran hasta 50 minutos en realizar la actividad; 11 personas se demoran a lo sumo 55 minutos en realizar la actividad o 22% de las personas se demoran máximo 55 minutos en realizar la actividad, así sucesivamente.

Es importante tener en cuenta que la marca de clase, por ser representante del intervalo, sólo será usada para el cálculo de algunos parámetros como la media aritmética y la desviación estándar, los cuales se explicarán en próximos capítulos.

4.5 Ejercicios de aplicación

1. En un estudio realizado sobre la aplicación de tatuajes en el cuerpo, a estudiantes universitarios se les consultaron aspectos como los que se ilustran en la tabla 13.

Tabla 13. Variables para el ejercicio de aplicación 4.5

Género	Edad (años)	Actividad laboral	N°. De tatuajes	Parte del cuerpo tatuada
Masculino	19	No	2	Espalda
Femenino	20	Sí	1	Pierna
Masculino	26	Sí	1	Brazo
Masculino	27	Sí	4	Hombros
Masculino	26	Sí	1	Brazo
Femenino	20	Sí	2	Espalda
Masculino	17	No	1	Pierna
Masculino	27	Sí	2	Brazo
Femenino	17	No	1	Pie
Femenino	18	No	1	Pierna
Masculino	17	Sí	5	Brazo
Femenino	19	No	1	Hombros
Femenino	20	No	2	Hombros
Femenino	17	No	4	Espalda
Masculino	22	Sí	1	Espalda
Femenino	17	No	2	Pierna
Femenino	18	No	1	Cuello
Masculino	20	No	4	Brazo
Masculino	18	No	1	Brazo
Masculino	24	No	1	Pierna

Elaboración propia

Clasifique cada una de las variables (género, edad, actividad laboral, N°. de tatuajes, parte del cuerpo tatuada) según su tipo (cualitativas, cuantitativas discretas o continuas), y para cada una de ellas realice la tabulación y la distribución de frecuencias.

2. Analice cada una de las variables tabuladas y responda las siguientes preguntas:

a) ¿Qué género presenta mayor aplicación de tatuajes en el cuerpo?

____

b) ¿Qué porcentaje de hombres prefieren tatuarse?

____

c) ¿Cuántas mujeres se tatúan?

____

d) ¿Qué porcentaje de personas más jóvenes se tatúan?

____

e) ¿Cuántas personas tatuadas presentan la edad más alta?

____

f) ¿De las personas tatuadas, cuántas tienen la edad máxima de 22 años?

____

g) ¿Cuántas personas tatuadas tiene una edad mínima de 20 años?

____

h) ¿De las personas tatuadas, qué porcentaje labora?

____

i) ¿Cuántas personas presentan el mayor número de tatuajes?

____

j) ¿Qué porcentaje de personas poseen de dos a cuatro tatuajes?

____

k) ¿Cuál es la parte del cuerpo más preferida para tatuarse?

____

l) ¿Qué porcentaje de personas se tatúan el cuello?

____

5 GRÁFICAS O DIAGRAMAS

Objetivos:

» Diferenciar las formas de representación gráfica para un conjunto de datos.

» Construir gráficas e interpretarlas según el tipo de variable o característica.

Contenido:

5.1 Histogramas
5.2 Polígonos de frecuencia
5.3 Ojivas o polígonos de frecuencias acumuladas
5.4 Diagramas de barras
5.5 Diagrama circular
5.6 Ejercicios de aplicación

5. GRÁFICAS O DIAGRAMAS

La organización de los datos obtenidos en una investigación mediante tablas de frecuencias no es suficiente para analizar el comportamiento de la variable. En la mayoría de los casos, las tablas ofrecen varias opciones de ser abordadas, es decir, posibilitan distintas entradas por medio de filas o columnas. Para una comprensión más efectiva del comportamiento de la variable, se hace útil el empleo de gráficas, dado que éstas permiten describir rápidamente las características del grupo.

Para representar el comportamiento de una variable se pueden usar varios tipos de gráficas, entre ellas están los histogramas, polígonos, ojivas, diagramas de barras y circulares (Berenson, Levine y Krehbiel, 2006).

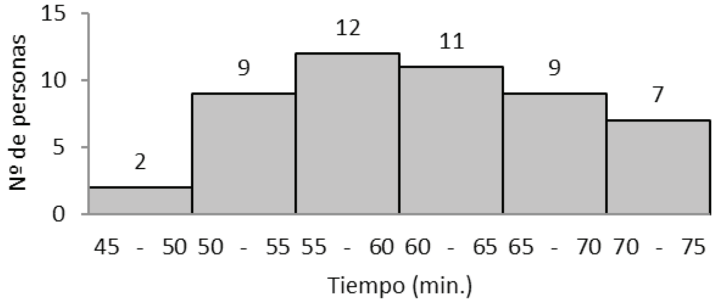
5.1 Histogramas

El histograma de frecuencias está conformado por un conjunto de rectángulos distribuidos en un plano cartesiano. Los histogramas representan variables cuantitativas continuas.

En el plano cartesiano, sobre el eje horizontal se distribuyen los intervalos de la variable y sobre el eje vertical se ubican las frecuencias. La base de los rectángulos está determinada por la amplitud del intervalo, y la altura de cada rectángulo corresponde a la frecuencia que presenta cada intervalo (Martínez, 2007).

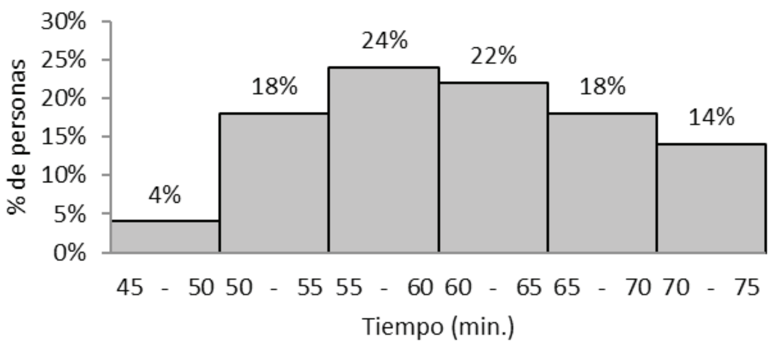
Los histogramas pueden representar a las frecuencias absolutas o relativas, dependiendo de la ubicación que se le dé a éstas sobre el eje vertical del plano cartesiano. De esta manera se obtiene el *histograma de frecuencias absolutas o* el *histograma de frecuencias relativas*, como se muestra en las gráficas 1 y 2 para los datos de la tabla 4, que representan el tiempo que tarda un grupo de personas en realizar una actividad.

Gráfica 1. Histograma de frecuencias absolutas para el ejemplo del tiempo requerido por un grupo de personas para realizar una actividad



Gráfica 1. Distribución del número de personas según el tiempo que tardan en realizar una actividad. (Elaboración propia).

Gráfica 2. Histograma de frecuencias relativas para el ejemplo del tiempo requerido por un grupo de personas para realizar una actividad.



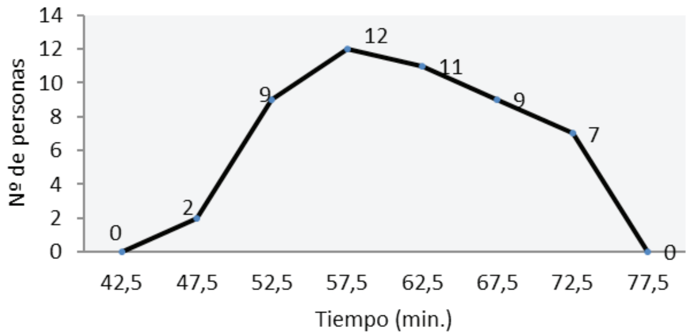
Gráfica 2. Distribución del porcentaje de personas según el tiempo que tardan en realizar una actividad. (Elaboración propia).

5.2 Polígono de frecuencias

El polígono de frecuencias es un conjunto de líneas sobre un plano cartesiano que representan el comportamiento de la característica en la población. Al igual que el histograma, el polígono se aplica a la variable cuantitativa continua.

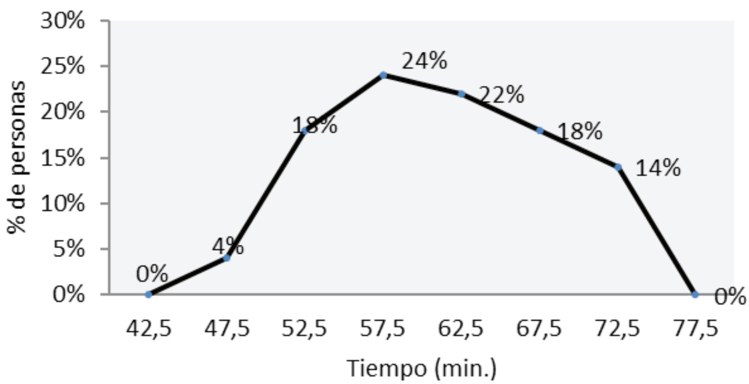
Se construye de forma similar al histograma. En el eje horizontal se ubican las marcas de clase y en el eje vertical las frecuencias absolutas o relativas. De esta forma es posible construir el *polígono* *de frecuencias absolutas* o *relativas*, dependiendo de las frecuencias utilizadas. En las gráficas 3 y 4 se ilustran estos tipos de polígonos para el ejemplo que representa a un grupo de personas que realizan una actividad.

Gráfica 3. Polígono de frecuencias absolutas para el ejemplo del tiempo requerido por un grupo de personas para realizar una actividad



Gráfica 3. Comportamiento del número de personas según el tiempo que tardan en realizar una actividad. (Elaboración propia).

Gráfica 4. Polígono de frecuencias relativas para el ejemplo del tiempo requerido por un grupo de personas para realizar una actividad



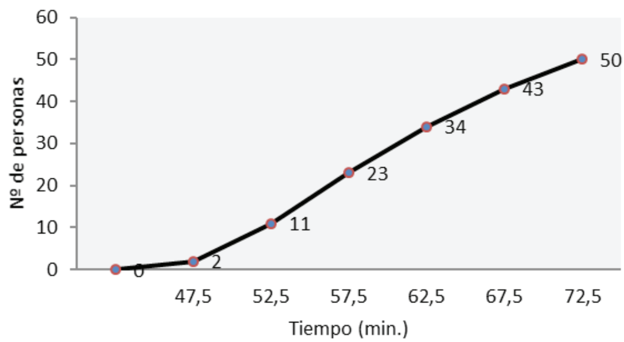
Gráfica 4. Comportamiento del número de personas según el tiempo que tardan en realizar una actividad. (Elaboración propia).

Para cerrar el polígono, se debe ampliar la marca de clase en los extremos manteniendo la misma amplitud y ubicando la frecuencia con el valor de cero.

5.3 Ojivas o polígonos de frecuencias acumuladas

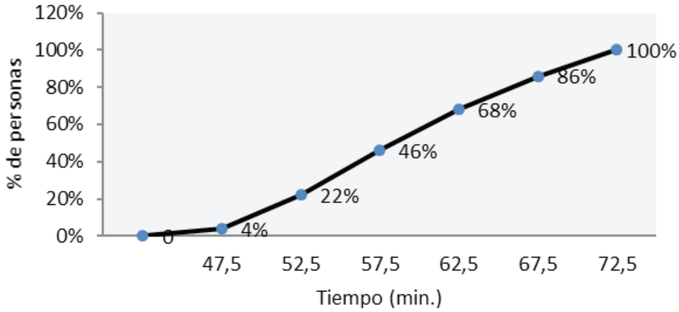
La ojiva representa el comportamiento acumulado de las unidades de investigación en relación a la variable analizada. Al igual que en los polígonos, las ojivas pueden ser construidas con las frecuencias absolutas o relativas. En las gráficas 5 y 6 se ilustran las ojivas para el ejemplo del grupo de personas que realizan una actividad.

Gráfica 5. Polígono de frecuencias absolutas acumuladas para el ejemplo del tiempo requerido por un grupo de personas para realizar una actividad



Gráfica 5. Comportamiento del número de personas según el tiempo que tardan en realizar una actividad. (Elaboración propia).

Gráfica 6. Polígono de frecuencias relativas acumuladas para el ejemplo del tiempo requerido por un grupo de personas para realizar una actividad



Gráfica 6. Comportamiento del número de personas según el tiempo que tardan en realizar una actividad. (Elaboración propia).

5.4 Diagrama de barras

El diagrama de barras es de las gráficas más utilizadas en los diferentes tipos de informes debido a que dan a conocer de forma fácil y sencilla las características de un grupo de elementos de una muestra o una población, especialmente cuando están asociadas a variables cualitativas o cuantitativas discretas.

El diagrama de barras consiste en líneas gruesas que constituyen rectángulos de anchura variable que representan los valores que toma la variable, y de longitud definida por las frecuencias absolutas o relativas. Las barras se construyen de forma horizontal o vertical y cada una puede ser representada con frecuencias absolutas o relativas.

Ejemplo: una institución educativa generó una campaña sobre orientación vocacional para los estudiantes del grado 11; para ello dispuso de un grupo de profesionales para que, de forma preliminar, orientaran a los alumnos sobre la elección de la carrera a seguir. Los datos siguientes representan el número de estudiantes que cada orientador atiende en una hora.

Tabla 15. Número de estudiantes que cada orientador atiende en una hora

Orientador	1	2	3	4
N°. de estudiantes	4	2	1	3

Elaboración propia

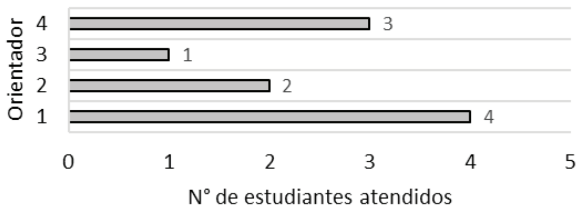
Al analizar la información se observa que los datos corresponden a una variable cuantitativa discreta. Para este tipo de variable, la tabulación se realiza ordenando en forma ascendente los valores que arroja la característica y se procede al cálculo de las frecuencias absolutas y relativas, tal como se muestra en la tabla 16.

Tabla 16. Distribución de frecuencias para la variable cuantitativa discreta

Orientador	n_i	h_i	N_i	H_i
1	4	40%	4	40%
2	2	20%	6	60%
3	1	10%	7	70%
4	3	30%	10	100%

Elaboración propia

Gráfica 7. Diagrama de barras horizontal para los estudiantes atendidos por los orientadores

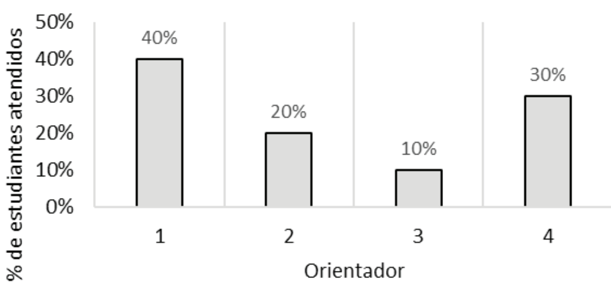


Gráfica 7. Distribución del número de estudiantes atendidos por cada orientador. (Elaboración propia).

En la gráfica anterior se puede observar que el orientador 1 es el que más estudiantes atiende en una hora, mientras que el orientador 3 solamente atiende a un estudiante en una hora.

Si al diagrama de barras horizontal se invierten los ejes (ver Gráfica 7), se obtiene el diagrama vertical (ver Gráfica 8). En cualquiera de los dos diagramas se pueden representar las frecuencias absolutas o relativas; la elección queda a discreción de quien este elaborando la gráfica.

Gráfica 8. Diagrama de barras vertical para los estudiantes atendidos por los orientadores



Gráfica 8. Distribución del número de estudiantes atendidos por cada orientador. (Elaboración propia).

5.5 Diagrama circular

El diagrama circular se fundamenta en la distribución de un círculo por fracciones que representan de forma proporcional los porcentajes de la característica objeto de análisis. Se usa para representar variables de tipo cualitativas o cuantitativas discretas. Si el número de categorías sobre las cuales se está realizando la distribución de los porcentajes es superior a 4, no se recomienda la construcción del diagrama circular y, en su defecto, se debe usar el diagrama de barras.

El diagrama circular se construye tomando los 360° de la circunferencia y se divide conforme a las frecuencias relativas de la característica. Retomando el ejemplo de la tabla 16 sobre el grupo de profesionales que asisten a los estudiantes del grado 11 en orientación vocacional, se consultó la formación de cada uno de ellos y se obtuvieron los siguientes datos (ver tabla 17):

Tabla 17. Formación de los orientadores vocacionales

Orientador	1	2	3	4
Formación	Psicólogo	Pedagogo	Trabajador Social	Psicólogo

Elaboración propia

La información anterior corresponde a una variable cualitativa. Al tabular los datos y calcular las frecuencias absolutas y relativas se obtienen los siguientes resultados (ver tabla 18):

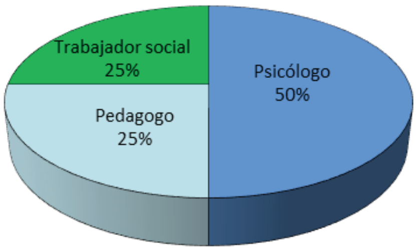
Tabla 18. Distribución de frecuencias para la variable cualitativa

Formación	n_i	h_i
Psicólogo	2	50%
Pedagogo	1	25%
Trabajador Social	1	25%

Elaboración propia

Obsérvese que para la variable cualitativa no aplican las frecuencias acumuladas, debido a que no es posible agrupar cualidades de forma simultánea una unidad de análisis o de investigación.

Gráfica 9. Diagrama circular para la formación de los orientadores



Gráfica 9. Distribución porcentual de la formación del orientador. (Elaboración propia).

5.6 Ejercicios de aplicación

1. En un estudio realizado en la Comuna Centro-Oriental de la ciudad de Medellín se indagó a jóvenes sobre los métodos anticonceptivos. Los resultados fueron los siguientes (ver tabla 19):

Tabla 19. Datos para el ejercicio de aplicación 5.6

Edad (años)	Género	Estrato socioec.	Nivel de estudio	Tipo de anticonceptivo	Tiempo de uso (meses)
23	Femenino	2	Universitario	Inyección	36
19	Femenino	1	Bachillerato	Ritmo	12
21	Femenino	2	Bachillerato	Píldora	4
19	Femenino	3	Bachillerato	Dispositivo	2
23	Femenino	1	Bachillerato	Inyección	3
20	Femenino	2	Bachillerato	Píldora	72
20	Masculino	3	Bachillerato	Preservativo	72
23	Femenino	3	Universitario	Píldora	60
20	Masculino	1	Bachillerato	Preservativo	48
19	Masculino	1	Universitario	Preservativo	24

Continúa en la siguiente página

Edad (años)	Género	Estrato socioec.	Nivel de estudio	Tipo de anticonceptivo	Tiempo de uso (meses)
15	Femenino	2	Bachillerato	Ritmo	6
18	Femenino	2	Universitario	Inyección	36
21	Masculino	2	Bachillerato	Píldora	8
18	Femenino	2	Bachillerato	Píldora	12
21	Femenino	1	Universitario	Dispositivo	36
23	Femenino	3	Universitario	Inyección	24
23	Femenino	3	Universitario	Inyección	12
20	Femenino	3	Universitario	Inyección	10
23	Femenino	3	Universitario	Inyección	60
23	Femenino	3	Universitario	Inyección	12
23	Masculino	2	Bachillerato	Preservativo	48
20	Femenino	3	Universitario	Píldora	48
20	Femenino	3	Universitario	Dispositivo	48
19	Masculino	2	Universitario	Preservativo	12
22	Femenino	3	Bachillerato	Píldora	24
20	Masculino	2	Universitario	Preservativo	8
18	Masculino	3	Bachillerato	Preservativo	12
19	Femenino	3	Universitario	Inyección	3
17	Femenino	2	Universitario	Dispositivo	24
19	Femenino	3	Universitario	Píldora	12
18	Masculino	2	Universitario	Preservativo	72
22	Femenino	2	Universitario	Inyección	12
17	Femenino	2	Bachillerato	Píldora	12
22	Femenino	2	Bachillerato	Inyección	48
20	Femenino	3	Universitario	Píldora	12
15	Femenino	2	Bachillerato	Ritmo	4
14	Femenino	2	Bachillerato	Píldora	2
18	Femenino	2	Bachillerato	Píldora	12

Continúa en la siguiente página

Edad (años)	Género	Estrato socioec.	Nivel de estudio	Tipo de anticonceptivo	Tiempo de uso (meses)
21	Masculino	2	Universitario	Preservativo	36
18	Femenino	2	Bachillerato	Dispositivo	5
21	Femenino	2	Universitario	Píldora	48
23	Femenino	3	Bachillerato	Dispositivo	36
21	Femenino	3	Bachillerato	Dispositivo	24
16	Masculino	3	Bachillerato	Preservativo	12
22	Femenino	3	Universitario	Píldora	36
23	Femenino	2	Bachillerato	Dispositivo	24
20	Masculino	3	Universitario	Preservativo	36
19	Femenino	1	Bachillerato	Inyección	12
21	Femenino	3	Bachillerato	Inyección	24
23	Masculino	2	Bachillerato	Preservativo	48
18	Masculino	2	Bachillerato	Preservativo	12
19	Masculino	3	Bachillerato	Preservativo	24
20	Femenino	3	Universitario	Inyección	24
22	Femenino	2	Universitario	Píldora	12
20	Femenino	1	Bachillerato	Inyección	12
18	Masculino	2	Bachillerato	Preservativo	24
23	Femenino	3	Universitario	Inyección	12
20	Masculino	3	Universitario	Preservativo	36
21	Femenino	3	Universitario	Inyección	24
20	Masculino	3	Universitario	Preservativo	24

Elaboración propia

Para las variables reportadas en el estudio, realice las gráficas que correspondan a cada una de ellas y analice su comportamiento, teniendo en cuenta la clasificación de la tabla 20:

Tabla 20. Clasificación de variables para el ejercicio de aplicación 5.6

Variable	Tipo de gráfica que aplica
Edad	Histograma Polígono Ojiva
Género	Diagrama de barras Diagrama circular
Estrato socioeconómico	Diagrama de barras Diagrama circular
Nivel de estudio	Diagrama de barras Diagrama circular
Tipo de anticonceptivo	Diagrama de barras Diagrama circular
Tiempo de uso	Histograma Polígono Ojiva

Elaboración propia

Es preciso recordar que para realizar las gráficas, previamente se debe hacer la tabulación y la distribución de frecuencias para cada variable.

6 TABULACIÓN DE DATOS BINARIOS O CRUZADOS

Objetivos:

» Relacionar dos variables por medio de la tabulación.

» Construir e interpretar tablas cruzadas a partir de la fila o de la columna.

Contenido:

6.1 Tabla de contingencia de porcentaje de fila
6.2 Tabla de contingencia de porcentaje de columna
6.3 Ejercicios de aplicación

6. TABULACIÓN DE DATOS BINARIOS O CRUZADOS

En la mayoría de los estudios estadísticos se emplea el análisis unidimensional para interpretar su comportamiento de forma aislada o individualmente. Sin embargo, los vínculos que tienen las diferentes personas, objetos o fenómenos, facultan el establecimiento de relaciones entre las características o variables que ellas presentan. Estas relaciones permiten analizar simultáneamente el comportamiento de dos variables, ya sean cualitativas o cuantitativas, usando para ello la tabulación cruzada o tablas de contingencia.

En el siguiente ejemplo se demuestra el procedimiento para la elaboración de una tabla de contingencia:

Un informe sobre instituciones de educación superior de una región muestra, entre sus resultados, la calidad académica de la institución y la antigüedad en años de funcionamiento. La calidad académica corresponde a una variable cualitativa, calificada como excelente, muy buena y buena. La antigüedad, como variable cuantitativa continua, oscila entre 10 y 49 años. La muestra fue de 300 instituciones. En la tabla 21 sólo se muestran los datos para las primeras 10 instituciones.

Tabla 21. Calidad académica y antigüedad de las instituciones de educación superior.

Institución	Calidad académica	Antigüedad en años
1	Buena	18
2	Muy buena	22
3	Buena	28
4	Excelente	38
5	Muy buena	33
6	Buena	28
7	Muy buena	19

Elaboración propia

El formato general para la tabla de contingencia o tabulación cruzada se describe en la tabla 22, con la síntesis de los datos para las instituciones de educación superior. En los costados izquierdo y superior se ubican los encabezados de las variables y en las demás posiciones, el número de instituciones que presentan simultáneamente la calidad y antigüedad correspondientes. De esta manera, en cada intervalo de antigüedad se cuenta el número de instituciones que presentan evaluación en cada categoría (buena, muy buena y excelente).

En la tabla 22 se observa que la mayor cantidad de instituciones educativas (64) tienen calidad "muy buena", y que la antigüedad está entre 20 y 29 años. Sólo hay dos instituciones con calidad "excelente" y antigüedad entre 10 y 19 años. De forma análoga se interpretan las demás frecuencias. Como se observa en la tabla 22, los totales de los costados derecho e inferior indican la distribución de frecuencias de la calidad académica y de la antigüedad de la institución, respectivamente. Es posible observar que hay 84 instituciones con buena calidad académica, 150 muy buena y 66 excelente. De forma similar, se puede observar en la margen inferior la distribución de frecuencias de la antigüedad: 78 instituciones tienen entre 10 y 19 años, 118 entre 20 y 29, 76 entre 30 y 39 y 28 entre 40 y 49.

Tabla 22. Tabulación cruzada o tabla de contingencia de la calidad académica y la antigüedad de las instituciones de educación superior

Calidad académica	Antigüedad (años)				Total
	10 - 19	20 - 29	30 - 39	40 - 49	
Buena	42	40	2	0	84
Muy buena	34	64	46	6	150
Excelente	2	14	28	22	66
Total	78	118	76	28	300

Elaboración propia

A partir de los resultados de la tabla 22, la mayor antigüedad parece estar asociada con una mayor calidad académica de la institución, y la antigüedad más baja con una menor calidad académica.

6.1 Tabla de contingencia de porcentaje de fila

Al convertir las frecuencias de la tabla 22 a porcentajes de fila o de columna, es posible tener un panorama más amplio de la relación existente entre las variables. Para obtener los porcentajes de fila, se divide cada frecuencia entre su respectivo total de la fila. Por ejemplo, el porcentaje de instituciones con buena calidad y antigüedad entre 10 y 19 años (50%), se obtiene dividiendo 42 entre 84 (ver tabla 23).

Tabla 23. Tabulación cruzada o tabla de contingencia de porcentaje de fila para la calidad académica y la antigüedad de las instituciones de educación superior.

Calidad académica	Antigüedad (años)				Total
	10 - 19	20 - 29	30 - 39	40 - 49	
Buena	50,0%	47,6%	2,4%	0,0%	100%
Muy buena	22,7%	42,7%	30,6%	4,0%	100%
Excelente	3,0%	21,2%	42,4%	33,4%	100%

Elaboración propia

6.2. Tabla de contingencia de porcentaje de columna

Los porcentajes de columna se obtienen de forma similar, es decir, dividiendo cada frecuencia de la columna entre el total de cada una de ellas. Por ejemplo, el porcentaje de instituciones con antigüedad entre 10 y 19 años con buena calidad académica (53,8%), se consigue dividiendo 42 entre 78; estos resultados se ilustran en la tabla 24.

Tabla 24. Tabulación cruzada o tabla de contingencia de porcentaje de columna para la calidad académica y la antigüedad de las instituciones de educación superior.

Calidad académica	Antigüedad (años)			
	10 - 19	20 - 29	30 - 39	40 - 49
Buena	53,8%	33,9%	2,6%	0,0%
Muy buena	43,5%	54,2%	60,6%	21,4%
Excelente	2,7%	11,9%	36,8%	78,6%
Total	100%	100%	100%	100%

Elaboración propia

De las tablas anteriores se deduce que en la categoría de buena calidad académica, el 50% de las instituciones tiene antigüedad entre 10 y 19 años, el 47,6% entre 20 y 29 años, el 2,4% entre 30 y 39 años y ninguna entre 40 y 49 años (tabla 24). Mientras que en las instituciones menos antiguas (ubicadas entre 10 y 19 años), el 53,8% tuvieron buena calidad académica, 43,5% como muy buena y 2,7% como excelente (ver tabla 24).

6.3 Ejercicios de aplicación

En un ejercicio de indagación sobre la preferencia de los estudiantes de Psicología de la Fundación Universitaria Luis Amigó de Medellín por las corrientes psicológicas y el rendimiento académico en los cursos relacionados con éstas, se obtuvo la siguiente información:

Tabla 25. Datos para la tabulación cruzada del ejercicio de aplicación 6.3.

N°. Estudiante	Nivel	Corriente preferida	Nota promedio cursos humanistas	Nota promedio ursos dinámicos	Nota promedio cursos cognitivos
1	8	Dinámica	4,2	4,6	4,5
2	5	Humanista	4,0	3,9	3,6
3	9	Humanista	4,3	3,8	4,0
4	6	Cognitiva	3,5	3,6	3,9
5	6	Dinámica	3,7	4,2	3,8
6	7	Cognitiva	3,9	3,8	4,3
7	8	Cognitiva	4,2	4,0	4,2
8	8	Cognitiva	4,4	3,9	4,4
9	8	Humanista	4,3	3,5	3,2
10	6	Dinámica	3,8	3,6	3,5
11	9	Humanista	3,8	3,2	3,6
12	7	Cognitiva	3,2	3,8	4,0
13	8	Dinámica	3,6	3,9	3,5
14	8	Dinámica	3,4	4,2	3,2

Continúa en la siguiente página

N°. Estudiante	Nivel	Corriente preferida	Nota promedio cursos humanistas	Nota promedio ursos dinámicos	Nota promedio cursos cognitivos
15	8	Cognitiva	3,5	3,4	3,8
16	5	Humanista	3,5	3,4	3,6
17	8	Cognitiva	3,1	3,4	3,8
18	5	Cognitiva	3,6	3,4	4,0
19	8	Humanista	3,9	3,6	4,0
20	7	Dinámica	3,6	4,2	3,4
21	9	Cognitiva	3,5	3,5	3,8
22	6	Dinámica	3,7	4,2	3,6
23	8	Cognitiva	3,6	3,5	3,9
24	5	Humanista	4,4	3,5	3,6
25	8	Humanista	4,0	3,2	3,4
26	6	Dinámica	3,4	3,8	3,2
27	9	Cognitiva	3,0	3,5	3,9
28	7	Cognitiva	3,5	3,4	3,8
29	5	Dinámica	3,1	3,9	3,6
30	6	Cognitiva	3,8	3,2	3,6
31	8	Cognitiva	4,0	3,6	3,9
32	7	Dinámica	3,4	3,6	4,2
33	9	Dinámica	3,6	3,4	4,0
34	6	Cognitiva	3,6	3,9	4,2
35	5	Cognitiva	3,9	3,5	4,2
36	5	Humanista	4,0	3,6	3,8
37	6	Humanista	4,5	3,2	3,5
38	7	Dinámica	3,6	3,9	3,6
39	8	Cognitiva	3,8	3,4	3,9
40	7	Cognitiva	3,2	3,5	4,2

Elaboración propia

Realice las tablas de contingencia de porcentaje de fila y de columna relacionando las siguientes variables e interprete los resultados:

1. Nivel y corriente preferida.

2. Corriente preferida y nota promedio de cursos humanistas.

3. Corriente preferida y nota promedio de cursos dinámicos.

4. Corriente preferida y nota promedio de cursos cognitivos.

5. Responda las siguientes preguntas:

   - ¿Qué porcentaje de estudiantes del nivel 9 prefieren la corriente cognitiva?

   ____

- ¿Qué porcentaje de estudiantes del nivel 9 prefieren la corriente dinámica?

___

- ¿Qué porcentaje de estudiantes del nivel 9 prefieren la corriente humanística?

___

- ¿Cuántos estudiantes presentan la nota promedio más alta en la corriente humanista?

___

- ¿Cuántos estudiantes presentan la nota promedio más baja en la corriente cognitiva?

___

- ¿Cuántos estudiantes presentan la nota promedio más alta en la corriente dinámica?

___

7 MEDIDAS DE TENDENCIA CENTRAL

Objetivos:

- Conocer las medidas de tendencia central para un conjunto de datos.
- Calcular e interpretar las medidas de tendencia central para datos agrupados y sin agrupar.

Contenido:

7.1 Media aritmética
7.2 Mediana
7.3 Moda
7.4 Ejercicios de aplicación

7. MEDIDAS DE TENDENCIA CENTRAL

En los estudios estadísticos es importante el análisis de la información que corresponde a variables cualitativas y cuantitativas, a partir de la tabulación y la representación de los datos por medio de gráficas. Además de esto, es necesario analizar los datos por medio de cálculos matemáticos que resuman el comportamiento de las características del objeto de estudio.

En la mayoría de los casos, el conjunto de datos obtenidos, ya sea de una muestra o de una población, tienden a reunirse alrededor de un valor central. De esta manera, es posible obtener un valor típico o representativo de todo el conjunto de datos, el cual se denomina medida de tendencia central (Walpole y Myers, 2012). Las medidas de tendencia central más representativas son: media aritmética, mediana y moda.

7.1 Media aritmética

La media aritmética es la medida de tendencia central más utilizada y la de mayor representatividad en los análisis estadísticos. Representa el promedio del conjunto de datos de la muestra. Su cálculo se realiza con la suma de todos los valores de los datos, dividida entre el número de datos que componen la muestra. Si la variable de estudio está representada por X, la media aritmética se representa por $\overline{X}$.

Cuando los datos son pocos y no se han agrupado en clases o intervalos, la media aritmética sería:

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

Donde

$\overline{X}$: media aritmética de la muestra

n : total de datos de la muestra
x_i : dato de la variable
$\sum_{i=1}^{n} x_i$: suma de todos los valores de la muestra

Por ejemplo, sea X el tiempo que tarda en horas un grupo de 4 estudiantes en realizar una actividad, cuyos valores son: 2, 4, 3 y 5.

La media aritmética es $\overline{X} = \frac{2+4+3+5}{4} = \frac{14}{4} = 3,5$ horas.

En este caso, el tiempo promedio que tardó el grupo de estudiantes en realizar la actividad fue 3,5 horas.

Cuando se agrupan los datos en una tabla de frecuencias, sin construir intervalos, se calcula la media aritmética mediante la siguiente formula:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i * n_i}{n}$$

Donde n_i es frecuencia absoluta para cada valor de la variable.

Por ejemplo, sea X el número de hijos de los empleados de una organización, los cuales se representan en la tabla 26.

Tabla 26. Número de hijos de los empleados

Número de hijos x_i	Frecuencia n_i	$x_i * n_i$
0	1	0 x 1 = 0
1	2	1 x 2 = 2
2	4	2 x 4 = 8
3	2	3 x 2 = 6
4	1	4 x 1 = 4
$n = \sum n_i = 10$	$\sum x_i * n_i = 20$	

Elaboración propia

$$\overline{X} = \frac{\sum_{i=1}^{m} x_i * n_i}{n} = \frac{20}{10} = 2 \textit{ hijos.}$$

Lo que significa que el promedio de hijos para el grupo de empleados es 2.

Si el conjunto de datos se han agrupado en intervalos, el cálculo de la media aritmética se realiza mediante la siguiente fórmula:

$$\overline{X} = \frac{\sum_{i=1}^{m} \dot{x}_i * n_i}{n}$$

Donde $\dot{x}_i$ es la marca de clase de cada intervalo.

Retomando la información de la tabla 12, donde se analiza el tiempo que tarda un grupo de personas en realizar una actividad, la media aritmética será:

Tabla 27. Media aritmética para el ejemplo del tiempo (en minutos) que tarda un grupo de personas en realizar una actividad

Nº. de Intervalo	Minutos	$\dot{x}$	n_i	$\dot{x}_i * n_i$
1	[45 - 50]	47,5	2	95
2	(50 - 55]	52,5	9	472,5
3	(55 - 60]	57,5	12	690
4	(60 - 65]	62,5	11	687,5
5	(65 - 70]	67,5	9	607,5
6	(70 - 75]	72,5	7	507,5
			$n = \sum n_i = 50$	$\sum \dot{x}_i * n_i = 3060$

Elaboración propia

$$\overline{X} = \frac{\sum_{i=1}^{N} \dot{x}_i * n_i}{n} = \frac{3060}{50} = 61{,}2 \text{ minutos.}$$

Lo que significa que el tiempo promedio que tarda el grupo de personas en realizar la actividad es 61,2 minutos.

7.2 Mediana

La mediana en un conjunto de datos es el valor que ocupa el lugar central, de tal forma que aquel valor deja el 50% de las observaciones por debajo de él y el otro 50% por encima de él. Para la ubicación de la posición de la mediana se deben ordenar los datos de forma ascendente. La mediana es representada por Me.

Si el conjunto de datos no se han agrupado, la posición i de la mediana se ubica según los siguientes criterios:

Cuando el total de datos (n) es impar, la posición de la mediana estará determinada por la fórmula:

$$i = X_{\frac{n+1}{2}}$$

Mientras que si el total de datos (n) es par, la posición de la mediana estaría determinada por:

$$i = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}$$

Por ejemplo, sea X el número de errores por página cometidos por un grupo de digitadores, los cuales se presentan en la tabla 28.

Tabla 28. Número de errores por página cometidos por un grupo de digitadores.

Digitador	A	B	C	D	E
Nº de errores	3	6	4	5	8

Elaboración propia

Inicialmente se deben ordenar los datos en forma ascendente, esto es: 3, 4, 5, 6, 8. Por tratarse de una muestra, se asume Me como un estimador de la mediana para la población. Esto es, el total de datos es n = 5 y la posición para el estimador será:

$$i = X_{\frac{n+1}{2}} = X_{\frac{5+1}{2}} = X_3$$

El valor que corresponde a la posición i_3 en el conjunto de datos ordenados es 5.

Nº. de errores	3	4	5	6	8
Posición i	i_1	i_2	i_3	i_4	i_5

Elaboración propia

En consecuencia, el estimador para la mediana Me es 5. Es decir, el 50% de los digitadores cometen 5 errores o menos por página, y el otro 50% cometen 5 errores o más por página.

Suponiendo que en el ejemplo anterior se toma un nuevo grupo de digitadores y se obtienen los siguientes resultados ordenados de forma ascendente: 5, 5, 7, 9, 11, 13, 13, 15.

En este caso, el total de datos es n = 8. Al calcular la posición para la mediana se tendrá:

$$i = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} = \frac{X_{\frac{8}{2}} + X_{\frac{8}{2}+1}}{2} = \frac{X_4 + X_5}{2}$$

Los valores que corresponden a las posiciones i_4 y i_5 en el conjunto de datos, luego de ser ordenados, son 9 y 11, respectivamente.

Nº de errores	5	5	7	9	11	13	13	15
Posición i	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8

Elaboración propia

En consecuencia, el valor de la mediana será:

$$Me = \frac{9+11}{2} = \frac{20}{2} = 10$$ errores.

Es decir, el 50% de los digitadores cometen menos de 10 errores por página, y el otro 50% cometen 10 o más errores por página.

En caso de que el conjunto de datos esté agrupado en intervalos, el cálculo de la mediana se realiza mediante el siguiente procedimiento:

1. Hallar $N/2$
2. Ubicar el intervalo cuya frecuencia absoluta acumulada N_i contiene a $N/2$
3. Calcular la mediana mediante la fórmula

$$Me = l_i + \left(\frac{\frac{N}{2} - N_{i-1}}{n_i} \right) * c$$

Donde:

l_i : límite inferior del intervalo que contiene a $N/2$

N : número total de datos de la población

N_{i-1} : frecuencia absoluta acumulada anterior al intervalo que con tiene a $N/2$

n_i : frecuencia absoluta del intervalo que contiene a $N/2$

c : amplitud del intervalo que contiene a $N/2$

Es importante aclarar que cuando el conjunto de datos corresponden a una muestra, la fórmula para la mediana se asume como un estimador y, en consecuencia, el total de datos se representa por n.

Para el ejemplo mencionado anteriormente en la tabla 12, sobre el tiempo que tarda un grupo de personas en realizar una actividad, se tiene la información de la tabla 29.

Tabla 29. Cálculo de la mediana para el ejemplo del tiempo (minutos) requerido por un grupo de personas para realizar una actividad

Nº. de Intervalo	Minutos	n_i	f_i	N_i	F_i	$\dot{x}$
1	[45 - 50]	2	4%	2	4%	47,5
2	(50 - 55]	9	18%	11	22%	52,5
3	(55 - 60]	12	24%	23	46%	57,5
4	(60 - 65]	11	22%	34	68%	62,5
5	(65 - 70]	9	18%	43	86%	67,5
6	(70 - 75]	7	14%	50	100%	72,5

Elaboración propia

Para calcular el estimador de la mediana, se utilizan los pasos descritos en el enunciado anterior, esto es:

1. El total de personas que realizaron la actividad es 50, por lo tanto, $n/2 = \frac{50}{2} = 25$ personas.
2. Al analizar la frecuencia absoluta acumulada, se encuentra que 25 se ubica en el 4° intervalo (no es posible ubicar el valor de 25 en el tercer intervalo, debido a que solo acumula 23 personas).

3. Los datos para el cálculo de la mediana serán:

$$l_i = 60$$
$$n/2 = 25$$
$$N_{i-1} = 23$$
$$n_i = 11$$
$$c = 65 - 60 = 5$$

Luego,

$$Me = l_i + \left(\frac{\frac{n}{2} - N_{i-1}}{n_i} \right) * c = 60 + \left(\frac{25-23}{11} \right) * 5 = 60 + \left(\frac{2}{11} \right) * 5 = 60 + 0.9 = 60,9 \text{ minutos}$$

Significa que el 50% de las personas realizaron la actividad en 60,9 minutos o menos y el otro 50% tardaron más de 60,9 minutos.

En los análisis estadísticos, la medida de tendencia central más representativa es la media aritmética. Sin embargo, en aquellos casos en los cuales se presentan valores extremos, es preferible usar la mediana en vez de la media, debido a que ésta no es afectada por valores extremos y por lo tanto, no es tan sensible como la media aritmética.

Por ejemplo: sea X la edad (en años) de un grupo de personas pertenecientes a un club de actividades lúdicas, estas son: 17, 16, 17, 18, 17, 16, 17, 18, 35.

Al calcular la media aritmética $\overline{X}$ se tendría un promedio de 19 años y la mediana Me de 17 años. Sin embargo, al analizar el comportamiento de las edades de los deportistas, se observa que estas tienden a agruparse más alrededor de 17 que a 19 años. Además, la media aritmética se afecta directamente con la presencia del valor extremo de 35 años, mientras que la mediana se mantiene en su valor, independiente de los valores extremos que se presenten en el conjunto de datos. En estos casos, es decir cuando se presentan valores extremos que afectan visiblemente el promedio en el conjunto de datos, se prefiere como medida de tendencia central a la mediana y no a la media aritmética.

7.3 Moda

En la vida cotidiana se escucha la expresión “está de moda” cuando algo se observa o se presenta repetidamente. En estadística, el concepto de la moda no se aleja de esta apreciación y, efectivamente, se denomina moda de un conjunto de datos al valor que más se presenta, es decir, el atributo o el valor de mayor frecuencia. La moda se representa por Mo y puede ser aplicada a las variables cualitativas y cuantitativas discretas o continuas.

Para obtener la moda de un conjunto de datos que están sin agrupar, se construyen las frecuencias y se ubica el valor o la característica que corresponde a la frecuencia mayor. Por ejemplo, los siguientes son los resultados obtenidos al indagar a varias personas por el color de preferencia: blanco, azul, rosado, azul, negro, azul, morado, azul, negro y blanco; al construir las frecuencias en la preferencia de las personas, se tiene (ver tabla 30):

Tabla 30. Moda para la variable cualitativa

Color	n_i
Blanco	2
Azul	4
Rosado	1
Negro	2
Morado	1

Elaboración propia

En la tabla 30 se observa que el color con mayor frecuencia es el azul, por tanto, la moda en el color de preferencia de las personas es el azul.

En el ejemplo anterior se presenta una sola moda, razón por la cual se denomina a este conjunto de datos como una distribución unimodal. Cuando existen varias modas, se denomina distribución multimodal y en caso de no existir moda, se denomina amodal. Veamos, el número de cursos matriculados por varios estudiantes en un semestre son: 6, 7, 6, 6, 7, 8, 9, 7; en este caso existen dos modas: 6 y 7, dado que cada uno de estos cursos presentan igual número de frecuencia y, además es la máxima; por tal razón, se denomina una distribución multimodal (particularmente bimodal).

Cuando los datos han sido agrupados en clases o intervalos, la moda se calcula utilizando la ponderación en el intervalo, con el siguiente procedimiento (Posada y Buitrago, 2008):

1. Ubicar el intervalo (o los intervalos) con mayor frecuencia absoluta n_i

2. Calcular la moda (o las modas) con la fórmula:

$$Mo = l_i + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) * C$$

Donde:

l_i : límite inferior del intervalo con mayor frecuencia absoluta

Δ_1: diferencia entre la mayor frecuencia absoluta y la anterior

Δ_2: diferencia entre la mayor frecuencia absoluta y la siguiente

c : amplitud del intervalo con mayor frecuencia absoluta

Al retomar el ejemplo mencionado anteriormente en la tabla 12, sobre el tiempo que tarda un grupo de personas en realizar una actividad, se toma la siguiente información para el cálculo de la moda (ver tabla 31):

Tabla 31. Moda para el ejemplo del tiempo (en minutos) requerido por un grupo de personas para realizar una actividad

Nº. de Intervalo	Minutos	n_i	f_i	N_i	F_i	$\dot{x}$
1	[45 - 50]	2	4%	2	4%	47,5
2	(50 - 55]	9	18%	11	22%	52,5
3	(55 - 60]	12	24%	23	46%	57,5
4	(60 - 65]	11	22%	34	68%	62,5
5	(65 - 70]	9	18%	43	86%	67,5
6	(70 - 75]	7	14%	50	100%	72,5

Elaboración propia

1. El intervalo de mayor frecuencia absoluta es el 3.

2. Los valores para el cálculo de la moda son:

$l_i = 55$

$\Delta_1 = 12 - 9 = 3$

$\Delta_2 = 12 - 11 = 1$

c = 60 – 55 = 5

Por lo tanto, la moda sería:

$$Mo = l_i + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) * C = 55 + \left(\frac{3}{3+1}\right) * 5 = 55 + \frac{15}{4} = 58,75 \text{ minutos}$$

Es decir, el tiempo que la mayoría de personas invierten para realizar la actividad es 58,75 minutos.

7.4. Ejercicios de aplicación

Se aplicó una encuesta a estudiantes de grado once de un colegio femenino ubicado en la ciudad de Medellín (Posada, 2014). El tema central fue la educación sexual, con preguntas orientadas a la formación, conocimiento y prácticas sobre la sexualidad. Algunas de las preguntas fueron extraídas para realización de este ejercicio y aparecen en la tabla 32:

Tabla 32. Datos para el ejercicio de aplicación 7.4.

N° de Estudiante	Tiene herman@s	Estrato socioec.	Edad	N° de métodos anticoncep. que conoce	Protección en la primera relación	Importancia de relaciones sexuales en el noviazgo	Influencia del alcohol en las relaciones	N° de relaciones que ha tenido	Relaciones sexuales con alguien diferente a su novio
1	Sí	3	15	5	No	Sí	No	7	Sí
2	Sí	3	17	1	Sí	Sí	No	1	No
3	Sí	3	16	5	No responde	Sí	No	0	No
4	No	3	15	6	Sí	Sí	Sí	3	No
5	No	3	15	3	No responde	No	Sí	0	No
6	No	4	15	8	Sí	Sí	Sí	2	No
7	Sí	3	16	7	Sí	Sí	No	8	Sí
8	No	4	17	9	Sí	Sí	Sí	9	Sí
9	Sí	3	18	8	No	Sí	Sí	3	No
10	Sí	3	16	10	Sí	Sí	Sí	8	No
11	Sí	3	14	11	Sí	Sí	Sí	7	No
12	Sí	2	16	4	Sí	Sí	Sí	7	Sí
13	Sí	3	18	5	Sí	Sí	No	2	No
14	Sí	3	15	6	No responde	Sí	Sí	0	No
15	Sí	3	16	8	Sí	Si	Sí	3	No
16	No	3	17	2	Sí	Si	No	2	No
17	Sí	3	17	9	Sí	No	Sí	3	No
18	Sí	4	17	3	Sí	Sí	No	3	No
19	No	3	14	7	Sí	Sí	Sí	3	No
20	Sí	2	15	9	Sí	No	Sí	2	No
21	Sí	3	16	8	Sí	Sí	Sí	7	Sí
22	Sí	3	18	6	No	Sí	Sí	1	No
23	No	3	19	7	Sí	Sí	Sí	8	No
24	No	3	17	8	No responde	No	Sí	0	No
25	No	2	14	2	No responde	Sí	Sí	0	No
26	Sí	3	15	3	No responde	Sí	Sí	0	No
27	No	3	15	4	No responde	Sí	Sí	0	No
28	No	3	16	3	No responde	No	Sí	0	No
29	Sí	3	19	9	No responde	Sí	Sí	0	No
30	Sí	2	16	7	No responde	No	Sí	0	No

Basada en la encuesta a estudiantes de grado once de un colegio femenino ubicado en la ciudad de Medellín, semestre 01 de 2014.

1. Calcule e interprete la media aritmética, mediana y moda para las variables "N°. de métodos anticonceptivos que conoce y N°. de relaciones que ha tenido".

2. Agrupe en intervalos la variable "edad" y calcule la media aritmética, mediana y moda. Interprete los resultados.

3. Calcule la moda para las variables cualitativas e interprete los resultados.

8 MEDIDAS DE POSICIÓN

Objetivos:

- Conocer las medidas de posición que se aplican a un conjunto de datos.
- Calcular e interpretar las medidas de posición para los datos agrupados y sin agrupar.

Contenido:

8.1 Cuartiles
8.2 Deciles
8.3 Percentiles
8.4 Ejercicios de aplicación

8. MEDIDAS DE POSICIÓN

Las medidas de posición, también llamadas cuantiles, son aquellas que permiten calcular valores en la distribución de los datos y que la dividen en partes iguales, de tal forma que los intervalos generados por los cuantiles contienen el mismo número de datos. Los cuantiles más usados son los cuartiles, deciles y percentiles. Cuando se tienen datos agrupados en intervalos, estas medidas se consideran en cierta forma como una extensión de la mediana.

8.1 Cuartiles

Los cuartiles (Q_k) son valores que fraccionan la distribución de los datos en cuatro partes iguales (Ruíz Muñoz, 2005). Existen tres cuartiles y cada una de las partes representa un 25% de los datos.

El primer cuartil Q_1 deja por debajo el 25% de la distribución de los datos o el 75% por encima de él. El segundo cuartil (Q_2) acumula el 50% de los datos por debajo y el otro 50% por encima de él (por tal razón es igual a la mediana); y el tercer cuartil (Q_3) deja por debajo el 75% de los datos y por encima el 25% (Ruíz Muñoz, 2005).

El cálculo de los cuartiles se realiza mediante el siguiente procedimiento:

1. Ordenar los datos de forma ascendente.
2. Calcular la posición i con la ecuación: $i = \left(\frac{k}{4}\right) n$. Donde *K* es el número del cuartil (*k* = 1, 2, 3) y *n* el número total de datos.
3. Si i no es un número entero, se debe redondear al entero siguiente y el valor que ocupa esta posición será el cuartil requerido. Si i es un número entero, el cuartil es el promedio de los valores e . i $i+1$.

Ejemplo: se le consultó a un grupo de siete estudiantes el número de horas semanal que dedican para el repaso de los temas vistos en clase, obteniendo los siguientes resultados: 3, 5, 2, 7, 6, 4, 9 horas. Para el cálculo de los cuartiles, se empleará el procedimiento descrito anteriormente.

1. Ordenar los datos en forma ascendente: 2, 3, 4, 5, 6, 7, 9.
2. Para el cuartil Q_1 la posición i sería: $i = \left(\frac{1}{4}\right)7 = 1{,}75$
3. Dado que i no es un entero, se redondea al entero siguiente, es decir a 2. En este caso, el cuartil Q_1 corresponde al valor ubicado en la posición 2, el cual es 3 horas. Su interpretación significa que el 25% de los estudiantes dedican máximo 3 horas semanales para el repaso a los temas vistos en clase.

De forma similar, para el cuartil Q_2 la posición i sería: $i = \left(\frac{2}{4}\right)7 = 3{,}5$

Como i no es un entero, se redondea al entero siguiente, es decir a 4. Por tanto, el cuartil Q_2 será el valor correspondiente a la posición 4, el cual es 5 horas. Esto es, el 50% de los estudiantes dedican máximo 5 horas semanales para el repaso a los temas vistos en clase. Nótese que este valor corresponde a la mediana.
En este caso, para el cuartil Q_3 la posición i sería:

$$i = \left(\frac{3}{4}\right)7 = 5{,}25$$

Al redondearla quedaría en 6, y el valor del cuartil Q_3 es 7 horas. Indica que el 75% de los estudiantes dedican máximo 7 horas semanales para el repaso a los temas vistos en clase.

Ejemplo: la talla de los neonatos prematuros nacidos en los partos durante una noche en un hospital fueron: 40, 37, 29, 31, 32, 38, 38, 38 cm.; para el cálculo de los cuartiles se empleará el procedimiento del ejemplo anterior, teniendo en cuenta el resultado obtenido al calcular la posición i.

Primer paso: ordenar los datos en forma ascendente: 29, 31, 32, 37, 38, 38, 38, 40.

Segundo paso: para el cuartil Q_1 la posición i sería: $i = \left(\frac{1}{4}\right)8 = 2$

Tercer paso: dado que i es un entero, el cuartil Q_1 corresponde al promedio entre los valores ubicados en las posiciones 2 y 3. Esto es,

$Q_1 = \frac{31+32}{2} = 31{,}5$ cm. Su interpretación significa que el 25% de los neonatos prematuros presentaron una talla máxima de 31,5 cm.

El cuartil Q_2 tendría la posición $i = \left(\frac{2}{4}\right)8 = 4$ y sería el promedio entre los valores ubicados en las posiciones 4 y 5. Esto es,

$Q_2 = \frac{37+38}{2} = 37{,}5$ cm.

Su interpretación significa que el 50% de los neonatos prematuros presentaron una talla máxima de 37,5 cm, igual a la mediana.

Para el Q_3, la posición será: $i = \left(\frac{3}{4}\right)8 = 6$ y sería el promedio entre los valores ubicados en las posiciones 6 y 7. Esto es, $Q_2 = \frac{37+38}{2} = 37,$ cm. Su interpretación significa que el 75% de los neonatos prematuros presentaron una talla máxima de 38 cm.

Si los datos se han agrupado en clases o intervalos, los cuartiles se calculan mediante la siguiente ecuación (Posada y Buitrago, 2008):

$$Q_k = l_i + \left[\frac{k(\frac{n}{4}) - N_{i-1}}{n_i}\right] * C$$

Donde:

k : número del cuartil, k= 1, 2, 3.

n : número total de datos.

l_i : límite inferior del intervalo que contiene a k(n/4).

N_{i-1} : frecuencia absoluta acumulada anterior al intervalo que con tiene a k(n /4).

n_i : frecuencia absoluta del intervalo que contiene a k(n /4).

C : amplitud del intervalo.

Ejemplo: en la tabla 33 se presentan los datos ordenados de la estatura, en centímetros, de un grupo de mujeres que asisten al gimnasio.

Tabla 33. Cuartiles para la estatura en centímetros de un grupo de mujeres que asisten al gimnasio

Nº. de Intervalo	Intervalo (Estatura en cm)	n_i	f_i	N_i	F_i	$\dot{x}$
1	[150 – 155]	1	3%	1	3%	152.5
2	(155 – 160]	11	31%	12	34%	157.5
3	(160 – 165]	13	37%	25	71%	162.5
4	(165 – 170]	6	17%	31	89%	167.5
5	(170 – 175]	4	11%	35	100%	172.5

Elaboración propia

El cuartil uno se calcula mediante el siguiente procedimiento:

1. Se halla k(n /4). (1*35/4 = 8,75)
2. Se ubica el intervalo que contiene a k(n /4) en la frecuencia absoluta acumulada N_i. (El segundo intervalo contiene a 8,75 en la frecuencia absoluta acumulada).
3. El primer cuartil se obtiene mediante la fórmula:

$$Q_1 = l_i + \left[\frac{1(\frac{n}{4}) - N_{i-1}}{n_i} \right] * C$$

Nota: la descripción de los componentes de la fórmula es la misma que se realizó en la mediana.

$$Q_1 = 155 + \left[\frac{1(\frac{35}{4}) - 1}{11} \right] * 5 = 159,4 \text{ centímetros}$$

Se estima que el 25% de las mujeres que asisten al gimnasio presentan una estatura máxima de 159,4 cm.

De forma similar se obtienen los cuartiles dos y tres.

$$Q_2 = l_i + \left[\frac{2(\frac{n}{4}) - N_{i-1}}{n_i} \right] * C \qquad Q_2 = 160 + \left[\frac{2(\frac{35}{4}) - 12}{13} \right] *5 = 162,1 \text{ cm}$$

$$Q_3 = l_i + \left[\frac{3(\frac{n}{4}) - N_{i-1}}{n_i} \right] * C \qquad Q_3 = 165 + \left[\frac{3(\frac{35}{4}) - 25}{6} \right] *5 = 166 \text{ cm}$$

El 50% de las mujeres presentan una estatura máxima de 162,1 cm (cuartil dos) y el 75% tienen una estatura máxima de 166 cm (cuartil tres).

8.2 Deciles

Los deciles (D_k) son valores que fraccionan la distribución de los datos en diez partes iguales (Ruíz Muñoz, 2005). En la distribución se presentan nueve deciles: el D_1 acumula el 10% del conjunto de datos, el

D_2 deja el 20%, y así sucesivamente hasta el D_9, que acumula el 90% de los datos. Para el cálculo de los deciles se usa un procedimiento similar al de los cuartiles:

1. Ordenar los datos de forma ascendente.

2. Calcular la posición i con la ecuación: $i = \left(\frac{k}{10}\right)n$. Donde K es el número del decil (k = 1, 2, 3, 4, 5, 6, 7, 8, 9) y n el número total de datos.

3. Si la posición i no es un número entero, se debe redondear al entero siguiente y el valor que ocupa esta posición será el cuartil requerido. Si la posición i es un número entero, el decil es el promedio de los valores i e $i+1$.

Para datos agrupados en intervalos: $D_k = l_i + \left[\frac{k(\frac{n}{10}) - N_{i-1}}{n_i}\right] * C$

El cálculo de los deciles uno y nueve para el ejemplo de la estatura de las mujeres, presentado en la tabla 33, se detalla a continuación:

$$D_1 = l_i + \left[\frac{1(\frac{n}{10}) - N_{i-1}}{n_i}\right] * C = 155 + \left[\frac{3,5-1}{11}\right] * 5 = 155 + 1,1 = 156,1 \text{ centímetros}$$

$$D_9 = l_i + \left[\frac{9(\frac{n}{10}) - N_{i-1}}{n_i}\right] * C = 170 + \left[\frac{31,5-31}{4}\right] * 5 = 170 + 0,6 = 170,6 \text{ centímetros}$$

8.3 Percentiles

Los percentiles (P_k) son valores que fraccionan la distribución de los datos en cien partes iguales (Ruíz Muñoz, 2005). En la distribución se presentan 99 percentiles: el primer percentil P_1 acumula el 1% del conjunto de datos, el percentil P_2 deja el 2%, y de forma similar los demás percentiles hasta llegar al percentil P_{99} que acumula el 99% de los datos. Para el cálculo de los percentiles se usa un procedimiento similar al empleado para los cuartiles y deciles:

1. Ordenar los datos de forma ascendente.

2. Calcular la posición i con la ecuación: $i = \left(\frac{k}{100}\right)n$. Donde K es el número del percentil (k = 1, 2, 3, 4, 5… 10, 11, 12… 98, 99) y n el número total de datos.

3. Si i no es un número entero, se debe redondear al entero siguiente y el valor que ocupa esta posición será el cuartil requerido. Si i es un número entero, el percentil es el promedio de los valores i e $i+1$.

Para datos agrupados en intervalos:

$$P_K = l_i + \left[\frac{K(\frac{n}{100}) - N_{i-1}}{n_i} \right] * C$$

Al analizar los cuartiles, deciles y percentiles se pueden deducir las siguientes relaciones:

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

$Q_1 = P_{25}$

$Q_3 = P_{75}$

$D_1 = P_{10}$

$D_2 = P_{20}$

$D_3 = P_{30}$

$D_4 = P_{40}$

$D_6 = P_{60}$

$D_7 = P_{70}$

$D_8 = P_{80}$

$D_9 = P_{90}$

8.4 Ejercicios de aplicación

En un examen de ingreso a la universidad se aplicó una prueba sobre comprensión lectora. Los resultados obtenidos por los estudiantes que aspiraban a los programas de Odontología y Trabajo social, fueron los siguientes:

Aspirantes a Odontología

3,0	3,7	2,5	2,2	3,5	3,0
2,7	3,8	3,8	2,2	2,5	3,8
3,3	2,8	2,7	3,5	2,2	2,7
2,2	2,3	2,3	2,3	2,0	3,8
2,7	2,2	3,2	2,5	3,7	2,3
3,5	2,2	2,0	2,5	3,0	2,8
3,8	2,3	3,0	2,0	2,3	3,7
2,7	2,2	2,3	2,7	2,7	2,2
3,3	2,0	3,0	2,2	2,3	2,3
3,5	2,8	3,0	3,0	2,2	4,0

Aspirantes a Trabajo Social

3,8	3,5	2,7	2,2	3,0	2,0
3,5	3,7	2,0	3,8	2,0	2,5
2,7	2,0	2,7	3,8	2,3	3,2
3,7	3,3	3,3	2,2	2,3	2,5
2,0	2,7	3,0	3,8	3,3	3,8
2,7	2,3	2,2	3,5	3,5	2,5
2,5	2,7	3,2	3,3	2,3	2,2

2,3	3,0	3,5	3,5	2,2	2,7
2,2	2,7	2,5	3,8	2,7	2,2
2,3	2,5	2,7	2,7	2,2	3,8

1. Calcular cuartiles y deciles para cada uno de los programas.

2. Interpretar los resultados en cada uno de los programas.

3. Comparar los cuartiles entre los programas.

9 MEDIDAS DE DISPERSIÓN

Objetivos:

- » Conocer las medidas de dispersión para un conjunto de datos.
- » Calcular las medidas de dispersión y su interpretación para datos agrupados y sin agrupar.

Contenido:

9.1 Rango
9.2 Rango intercuartil
9.3 Varianza
9.4 Desviación estándar
9.5 Coeficiente de variación
9.6 Ejercicios de aplicación

9. MEDIDAS DE DISPERSIÓN

Además de las medidas de tendencia central que posibilitan la representación del conjunto de datos por medio de un valor, es necesario conocer la variabilidad o la dispersión que los datos pueden tener en relación a una medida central.

En los análisis estadísticos, las medidas de dispersión que presentan más representatividad son: rango, rango intercuartil, varianza, desviación estándar y coeficiente de variación (Posada y Buitrago, 2008).

9.1 Rango

El rango es considerado como la medida de dispersión más simple para el análisis de los datos. No ofrece mucha información sobre la variabilidad de los datos por estar basada sólo en los valores extremos, razón por la cual debe ser usada como complemento de otras medidas de dispersión. Para el cálculo del rango se utiliza la siguiente ecuación:

$$\text{Rango} = \text{valor máximo} - \text{valor mínimo}$$

Por ejemplo, para los datos de la tabla 33, sobre la talla (en centímetros) de un grupo de mujeres que asisten al gimnasio, el rango sería:

$$\text{Rango} = 175 - 150 = 25 \text{ cm.}$$

Al interpretar el rango se deben relacionar los valores mínimo y máximo; es decir, resaltar las cantidades entre las cuales se encuentra el rango. Para el ejemplo mencionado anteriormente, la variación de la talla de las mujeres que asisten al gimnasio es de 25 cm, la cual oscila entre 150 y 175 cm. Si no se hace claridad que el rango está entre los valores 150 y 175 cm, puede generar confusión debido a que pueden existir muchos valores extremos con rango de 25 cm.

9.2 Rango intercuartil

El rango intercuartil (RIC) se denomina de esta manera porque es una medida de dispersión que evita que los valores extremos influyan en el conjunto de datos. Se calcula mediante la diferencia entre el cuartil tres (Q_3) y el cuartil uno (Q_1). Es decir, el rango intercuartil corresponde al rango del 50% ubicado en el centro de los datos. El RIC se calcula por medio de la siguiente ecuación (Triola, 2000):

$$\text{Rango intercuartil (RIC)} = Q_3 - Q_1$$

Para la talla (en cm) de un grupo de mujeres que asisten al gimnasio, presentada en la tabla 33, los cuartiles son Q_1 = 159,4 cm y Q_3 = 166 cm. Así, el rango intecuartil es:

$$\text{RIC} = 166 - 159{,}4 = 6{,}6 \text{ cm.}$$

En este caso, el intervalo entre 159,4 y 166 cm se denomina *mitad central*, es decir el 50% que contiene la información central; y 6,6 cm representa la dispersión media o rango intercuartil de la talla del grupo de mujeres que asisten al gimnasio.

9.3 Varianza

La varianza es una medida de dispersión basada en la diferencia de cada dato con la media aritmética. Posada y Buitrago (2008) plantean que "la diferencia entre cada x_i y el promedio ($\overline{x}$ para una muestra y μ para una población) se llama *desviación respecto al promedio*. Para una muestra, la desviación respecto a la media se expresa como $(x_i - \overline{x})$; para una población es $(x_i - \mu)$" (p. 86). Al sumar el total de las desviaciones respecto al promedio, éste tiende a cero por la compensación de las desviaciones positivas (cuando los datos están por encima del promedio), con las desviaciones negativas (cuando los datos están por debajo del promedio). De esta manera, no es posible obtener efectivamente la desviación de los datos respecto del promedio, por lo cual se hace necesario elevar cada desviación al cuadrado, garantizando así que todas las desviaciones obtenidas presenten cantidades positivas; el resultado entonces quedará en unidades cuadradas.

Cuando se tiene la totalidad de los datos de la población, el promedio de las desviaciones elevadas al cuadrado se denomina *varianza poblacional* y se representa con la letra del alfabeto griego *sigma* (σ^2). Para una población con total de datos N y promedio μ, el parámetro para la varianza se calcula mediante la siguiente ecuación:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

La *varianza de la muestra* (s^2) tiene como objetivo convertirse en un estimador de la variación para la población; por tal razón, se define como la suma de las desviaciones elevadas al cuadrado, distribuidas entre el tamaño de la muestra, menos uno. El estimador para la varianza muestral se calcula mediante la siguiente ecuación:

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$$

Donde:

$\overline{x}$:media aritmética de la muestra

n :total de los datos de la muestra

x_i :cada dato u observación de la variable X

En relación al denominador (n–1), Posada y Buitrago (2008, p. 94) manifiestan que:

> Si el denominador fuera n en lugar de (n–1), se obtendría el promedio de los cuadrados de las diferencias con respecto a la media. Sin embargo, se utiliza (n–1) debido a ciertas propiedades matemáticas deseadas que tiene el estadístico s^2, las cuales lo hacen apropiado para hacer inferencias estadísticas. Al aumentar el tamaño de la muestra, la diferencia entre n y (n–1) disminuye cada vez más.

Al calcular la varianza, los datos se elevan al cuadrado, por tanto, las unidades con las cuales se midieron también se elevan al cuadrado, imposibilitando la interpretación. En consecuencia, en la mayoría de los análisis estadísticos se emplea la varianza como una medida que permite comparar la dispersión entre dos o más variables, identificando la de mayor varianza como aquella que posee mayor dispersión o variabilidad. La importancia de la varianza está en que es una medida transitoria para el cálculo de la desviación típica o estándar de un conjunto de datos.

Por ejemplo, en la tabla 34 se presenta la puntuación de la evaluación de desempeño de siete empleados del área de mercadeo de una empresa. La puntuación es valorada en la escala de 1 a 5. Se requiere conocer la varianza de la calificación de los empleados.

Tabla 34. Varianza para la evaluación de desempeño de siete empleados del área de mercadeo de una empresa

Empleado	Calificación (x_i)	Media de la muestra $(\overline{x})$	Desviación $(x_i - \overline{x})$	Desviación al cuadrado $(x_i - \overline{x})^2$
1	3,5	3,6	-0,1	0,01
2	4,5	3,6	0,9	0,81
3	4,2	3,6	0,6	0,36
4	3,0	3,6	-0,6	0,36
5	2,7	3,6	-0,9	0,81
6	3,3	3,6	-0,3	0,09
7	4,0	3,6	0,4	0,16
			$\sum(x_i - \overline{x}) = 0$	$\sum(x_i - \overline{x})^2 = 2,6$

Elaboración propia

Luego, la varianza será:

$$s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} = \frac{2,6}{6} = 0,43$$

Nótese que si se interpreta la varianza se estaría diciendo que la variación en la calificación de desempeño de los empleados es de 0,43 puntos cuadrados, lo cual no es lógico. En este sentido, cobra importancia la varianza como medida de transición para la desviación típica o estándar.

Si los datos se agruparon en frecuencias o en intervalos, la varianza puede ser calculada mediante las siguientes formulas:

$$\sigma^2 = \frac{\sum x_i^2 * n_i}{N} - \mu^2$$ Como parámetro para la población.

$$s^2 = \frac{\sum x_i^2 * n_i}{n-1} - \overline{x}^2$$ Como estimador para la muestra.

Donde:

$\overline{x}$: media aritmética

n : total de datos de la muestra

N : total de datos de la población

x_i : cada dato de la variable o marca de clase si es intervalo

n_i : frecuencia absoluta

Para los datos del ejemplo de la estatura en centímetros de un grupo de mujeres que asisten al gimnasio, presentados en la tabla 33, la varianza sería (ver tabla 35):

Tabla 35. Varianza para la estatura en centímetros de un grupo de mujeres que asisten al gimnasio.

Nº. de Intervalo	Intervalo (Estatura en cm)	x	x^2	n_i	$x_i^2 * n_i$
1	[150 – 155]	152,5	23.256,25	1	23.256,25
2	(155 – 160]	157,5	24.806,25	11	272.868,75
3	(160 – 165]	162,5	26.406,25	13	343.281,25
4	(165 – 170]	167,5	28.056,25	6	168.337,50
5	(170 – 175]	172,5	29.756,25	4	119.025,00
					$\sum x_i^2 * n_i = 926.768,75$

Elaboración propia

Al calcular el promedio de la estatura para las 34 mujeres se obtiene: $\bar{x} = 162,6$ cm.

Luego, la varianza será:

$$s^2 = \frac{\sum x_i^2 * n_i}{n-1} - \bar{x}^2 = \frac{926.768,75}{35-1} - (162,6)^2 = 27.257,9 - 26.438,8 = 819,1$$

9.4 Desviación estándar

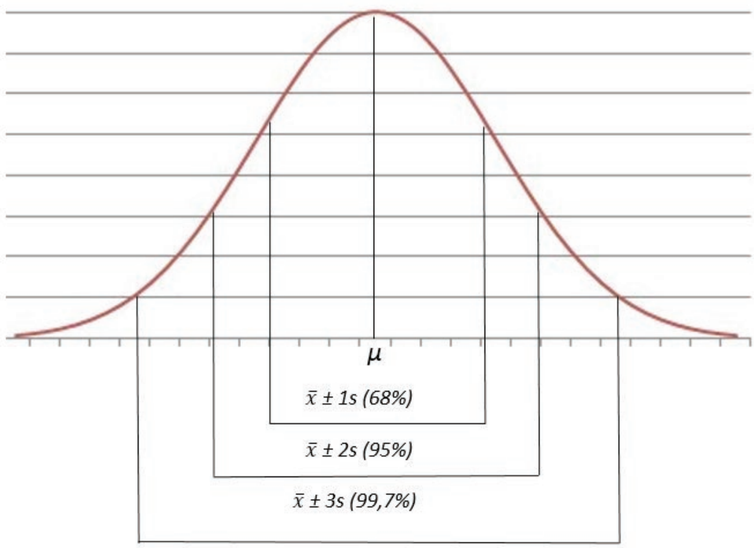
La desviación estándar es considerada la medida de dispersión con mayor representatividad para un conjunto de datos. Matemáticamente se calcula como la raíz cuadrada positiva de la varianza, y se denota por (*s*) cuando se estima para la muestra y por (σ) si se calcula para la población:

$$s = \sqrt{s^2}$$

$$\sigma = \sqrt{\sigma^2}$$

La desviación estándar indica la distribución de los datos alrededor de la media aritmética o promedio. Cuando la distribución de los datos se aproxima a una forma de campana o es simétrica, como se ilustra en la gráfica 11, la desviación estándar puede interpretarse mediante la regla empírica, esta es: el 68% de los datos se agrupan alrededor de la media, entre el intervalo $(\bar{x}-1s)$ y $(\bar{x}+1s)$, el 95% entre $(\bar{x}-2s)$ y $(\bar{x}+2s)$, el 99,7% entre $(\bar{x}-3s)$ y $(\bar{x}+3s)$ (Triola, 2000). Para los análisis estadísticos sólo se analiza la dispersión de los datos a partir de una variación de la desviación alrededor de la media aritmética, es decir, el intervalo que cubre aproximadamente el 68% de los datos, teniendo en cuenta que la distribución de éstos debe ser simétrica.

Gráfica 11. Variación de la desviación estándar alrededor de la media aritmética.



Elaboración propia.

Retomando la información de la tabla 35 sobre la estatura en centímetros de un grupo de mujeres que asisten al gimnasio, la desviación estándar sería:

$$s=\sqrt{s^2}=\sqrt{819,1}=28,6 \text{ cm.}$$

Al interpretar la desviación estándar, significa que la estatura varía 28,6 cm alrededor de la media (162,6 cm). Por la regla empírica, podría decirse que el 68% de las estaturas está dentro de una desviación estándar de la media, se estima que el 95% de las estaturas estará entre $(\bar{x}\pm 2s)$ y el 99,7% estará entre $(\bar{x}\pm 3s)$.

Es importante resaltar que las medidas del rango, rango intercuartil, varianza y desviación estándar nunca asumen valores negativos. La relación de estas medidas con la dispersión es directa, es decir, si los valores de las medidas son altos, la dispersión también será alta y viceversa.

9.5 Coeficiente de variación

El coeficiente de variación (*CV*) es una medida que relaciona la desviación estándar con la media aritmética para determinar qué tan homogénea o dispersa es la información. Expresa el porcentaje que representa la desviación con relación a la media aritmética y se calcula por medio de la siguiente ecuación:

$$CV = \frac{S}{\overline{\overline{X}}} * 100$$

Cuando se tiene una muestra, el coeficiente de variación puede ser utilizado para calificar estadísticamente la calidad de las estimaciones. Para ello se consideran los siguientes criterios (Departamento Administrativo Nacional de Estadística, DANE, 2008, p. 5):

CV menor o igual al 7%, las estimaciones se consideran precisas.
CV entre el 8% y el 14%, las estimaciones tienen precisión aceptable.
CV entre el 15% y 20%, la precisión es regular.
CV mayor del 20% indica que la estimación es poco precisa.

Para el ejemplo de la tabla 35 sobre la talla de un grupo de mujeres que asisten al gimnasio, la media aritmética fue 162,6 cm y la desviación estándar 28,6 cm. Al calcular el coeficiente de variación, se obtiene:

$$CV = \frac{S}{\overline{\overline{X}}} * 100 = \frac{28,6}{162,6} * 100 = 17,6\%$$

Al interpretar los datos, es posible establecer que la desviación representa el 17,6% de la media. En términos del ejercicio, podría interpretarse que los datos varían 17,6% alrededor de la media, lo cual intuye que la precisión de estimación de los parámetros para esta población es regular.

El coeficiente de variación, por ser una medida de dispersión relativa, se utiliza para comparar la variabilidad de distintas muestras o poblaciones, aunque tengan unidades de medida diferentes (Triola, 2000). En el siguiente ejemplo se muestra esta situación:

Una persona desea realizar una inversión en un negocio que tenga buena rentabilidad, para ello se le presentan dos proyectos con posibilidades diferentes. El primer proyecto ha presentado utilidades promedio en el último año de $150 millones y desviación de $50 millones. En el mismo año, el promedio de utilidades para el segundo proyecto fueron de $120 millones con una desviación estándar de $12 millones. ¿Cuál proyecto presenta más estabilidad para generar confianza al inversionista?

Al analizar la desviación estándar, el primer proyecto es más variable que el segundo proyecto. Sin embargo, como el promedio de las utilidades de los proyectos es diferente, se recomienda considerar la variación de la utilidad con respecto al promedio, para observar la estabilidad de ambos proyectos.

Los coeficientes de variación para los proyectos serían:

Primer proyecto: $$CV_1 = \frac{S}{\overline{\overline{X}}} * 100 = \frac{\$50}{\$150} * 100 = 33,3\%$$

Segundo proyecto: $$CV_2 = \frac{S}{\overline{\overline{X}}} * 100 = \frac{\$12}{\$120} * 100 = 10\%$$

En consecuencia, en relación con la media, la utilidad del primer proyecto es más variable que la del segundo. Por tanto, a pesar de presentar el segundo proyecto menor utilidad promedio, es más estable que el primero, lo cual puede generar mayor confianza para el inversionista.

9.6 Ejercicios de aplicación

Retomando el ejercicio de aplicación propuesto en el capítulo 8 sobre la prueba de comprensión lectora para aspirantes de ingreso a la universidad para los programas de Odontología y Trabajo social, se tiene:

Aspirantes a Odontología

3,0	3,7	2,5	2,2	3,5	3,0
2,7	3,8	3,8	2,2	2,5	3,8
3,3	2,8	2,7	3,5	2,2	2,7
2,2	2,3	2,3	2,3	2,0	3,8
2,7	2,2	3,2	2,5	3,7	2,3
3,5	2,2	2,0	2,5	3,0	2,8
3,8	2,3	3,0	2,0	2,3	3,7
2,7	2,2	2,3	2,7	2,7	2,2
3,3	2,0	3,0	2,2	2,3	2,3
3,5	2,8	3,0	3,0	2,2	4,0

Aspirantes a Trabajo Social

3,8	3,5	2,7	2,2	3,0	2,0
3,5	3,7	2,0	3,8	2,0	2,5
2,7	2,0	2,7	3,8	2,3	3,2
3,7	3,3	3,3	2,2	2,3	2,5
2,0	2,7	3,0	3,8	3,3	3,8
2,7	2,3	2,2	3,5	3,5	2,5
2,5	2,7	3,2	3,3	2,3	2,2
2,3	3,0	3,5	3,5	2,2	2,7
2,2	2,7	2,5	3,8	2,7	2,2
2,3	2,5	2,7	2,7	2,2	3,8

1. Analizar la dispersión para cada uno de los programas mediante el cálculo de las siguientes medidas: rango, rango intercuartil, varianza y desviación estándar. Interpretar los resultados.

2. Calcular el coeficiente de variación para cada uno de los programas y compararlos, estableciendo los niveles de dispersión en cada uno de ellos.