



FUNDACIÓN
UNIVERSITARIA

ISBN: 978-958-8943-05-3

ELEMENTOS BÁSICOS DE ESTADÍSTICA DESCRIPTIVA PARA EL ANÁLISIS DE DATOS

Gabriel Jaime Posada Hernández



Fundación Universitaria Luis Amigó

ELEMENTOS BÁSICOS DE ESTADÍSTICA DESCRIPTIVA PARA EL ANÁLISIS DE DATOS

Gabriel Jaime Posada Hernández

2016

519.53 P855

Posada Hernández, Gabriel Jaime

Elementos básicos de estadística descriptiva para el análisis de datos [recurso electrónico] / Gabriel Jaime Posada Hernández. -- Medellín : Funlam, 2016

158 p : il., col.

ISBN: 978-958-8943-05-3

Incluye referencias bibliográficas

ESTADÍSTICA; MUESTREO (ESTADÍSTICA); VARIABLES (ESTADÍSTICA); ANÁLISIS DE DATOS

ELEMENTOS BÁSICOS DE ESTADÍSTICA DESCRIPTIVA PARA EL ANÁLISIS DE DATOS

© Fundación Universitaria Luis Amigó

Transversal 51 A 67 B 90. Medellín, Antioquia, Colombia.

Tel: (574) 448 76 66 (Ext. 9711 Departamento de Fondo Editorial)

www.funlam.edu.co - fondoeditorial@funlam.edu.co

ISBN: 978-958-8943-05-3

FECHA DE EDICIÓN: 16 de junio de 2016

AUTOR: Gabriel Jaime Posada Hernández

CORRECCIÓN DE ESTILO: Rodrigo Gómez Rojas

DIAGRAMACIÓN Y DISEÑO: Diana Ortiz Muñoz

EDICIÓN: Fondo Editorial Luis Amigó

COORDINADORA DEPARTAMENTO EDITORIAL: Carolina Orrego Moscoso

Hecho en Medellín - Colombia / Made in Medellín - Colombia

Financiación Realizada por la Fundación Universitaria Luis Amigó.

El autor es moral y legamente responsable de la información expresada en este libro, así como del respeto a los derechos de autor; por tanto, no compromete en ningún sentido a la Fundación Universitaria Luis Amigó.

Prohibida la reproducción total o parcial, por cualquier medio, sin autorización escrita de la Fundación Universitaria Luis Amigó.

Licencia Atribución - No comercial - Sin Derivar de Creative Commons.



A mi esposa Ruby
y mis hijas Laura y Sara

CONTENIDO

INTRODUCCIÓN.....	IX
1. CONCEPTOS GENERALES SOBRE ESTADÍSTICA	10
1.1 Historia.....	11
1.2 Definición.....	13
1.3 Categorías.....	14
1.4 Unidad de análisis o de investigación.....	14
1.5 Población.....	14
1.6 Muestra.....	15
1.7 Parámetros y estadígrafos.....	15
1.8 Variables.....	16
1.8.1 Variables cualitativas.....	16
1.8.2 Variables cuantitativas.....	16
1.9 Escalas de medición.....	18
1.9.1 Escala nominal.....	18
1.9.2 Escala ordinal.....	18
1.9.3 Escala de intervalo.....	19
1.9.4 Escala de razones.....	19
1.10 Ejercicios de aplicación.....	20

2. INTRODUCCIÓN A LOS TIPOS DE MUESTREO.....	22
2.1 Métodos de muestreo probabilístico.....	23
2.1.1 Muestreo aleatorio simple.....	23
2.1.2 Muestreo aleatorio sistemático.....	24
2.1.3 Muestreo aleatorio estratificado.....	24
2.1.4 Muestreo aleatorio por conglomerados.....	26
2.2 Aspectos importantes del muestreo probabilístico.....	27
2.2.1 Error de cobertura o sesgo en la selección.....	27
2.2.2 Error o sesgo de no respuesta.....	28
2.2.3 Error de muestreo.....	28
2.2.4 Error de medición.....	28
2.3 Métodos de muestreo no probabilístico.....	28
2.3.1 Muestreo por cuotas.....	29
2.3.2 Muestreo intencional u opinático.....	29
2.3.3 Muestreo bola de nieve.....	29
2.4 Ejercicios de aplicación.....	30
3. TABULACIÓN DE DATOS ESTADÍSTICOS.....	33
3.1 Rango o recorrido.....	35
3.2 Número de intervalos o clases.....	36
3.3 Amplitud del intervalo de clase.....	37
3.4 Límites de los intervalos.....	37
3.5 Tabulación.....	38
3.6 Marca de clase.....	39
3.7 Ejercicios de aplicación.....	39

4. DISTRIBUCIÓN DE FRECUENCIAS.....	44
4.1 Frecuencia absoluta.....	45
4.2 Frecuencia relativa.....	45
4.3 Frecuencia absoluta acumulada.....	46
4.4 Frecuencia relativa acumulada.....	47
4.5 Ejercicios de aplicación.....	48
5. GRÁFICAS O DIAGRAMAS.....	52
5.1 Histogramas.....	53
5.2 Polígono de frecuencias.....	54
5.3 Ojivas o polígonos de frecuencias acumuladas.....	56
5.4 Diagrama de barras.....	57
5.5 Diagrama circular.....	59
5.6 Ejercicios de aplicación.....	60
6. TABULACIÓN DE DATOS BINARIOS O CRUZADOS.....	64
6.1 Tabla de contingencia de porcentaje de fila.....	67
6.2 Tabla de contingencia de porcentaje de columna.....	67
6.3 Ejercicios de aplicación.....	68
7. MEDIDAS DE TENDENCIA CENTRAL.....	72
7.1 Media aritmética.....	73
7.2 Mediana.....	76
7.3 Moda.....	80
7.4 Ejercicios de aplicación.....	83

8. MEDIDAS DE POSICIÓN.....	86
8.1 Cuartiles.....	87
8.2 Deciles.....	91
8.3 Percentiles.....	92
8.4 Ejercicios de aplicación.....	94
9. MEDIDAS DE DISPERSIÓN.....	96
9.1 Rango.....	97
9.2 Rango intercuartil.....	98
9.3 Varianza.....	98
9.4 Desviación estándar.....	102
9.5 Coeficiente de variación.....	104
9.6 Ejercicios de aplicación.....	106
10. MEDIDAS DE FORMA.....	108
10.1 Coeficiente de asimetría o sesgo.....	109
10.1.1 Coeficiente de asimetría de Pearson.....	110
10.1.2 Coeficiente de asimetría de Bowley.....	110
10.1.3 Coeficiente de asimetría de Fisher.....	111
10.2. Coeficiente de apuntamiento o curtosis.....	112
10.2.1 Distribución mesocúrtica.....	113
10.2.2 Distribución leptocúrtica.....	113
10.2.3 Distribución platicúrtica.....	113
10.3 Ejercicios de aplicación.....	115

11. ESTADÍSTICA DESCRIPTIVA MEDIANTE MICROSOFT EXCEL.....	116
1.1 Activación de la función “Análisis de datos”	117
11.2 Estadística descriptiva.....	120
11.3 Caracterización e interpretación de resultados.....	122
11.4 Ejercicios de aplicación.....	124
12. ANÁLISIS ESTADÍSTICO BAJO SPSS.....	126
12.1 Ingreso al programa SPSS.....	127
12.2 Preparación de variables.....	128
12.3 Ingreso de datos.....	135
12.4 Resultados.....	137
12.5 Ejercicio de aplicación.....	150
REFERENCIAS.....	153
INFORMACIÓN DEL AUTOR.....	155

INTRODUCCIÓN

La estadística ha cobrado gran importancia en el análisis de los datos, no sólo en los asuntos del Estado (de ahí su nombre), sino también en las facetas del comportamiento humano, expandiendo su aplicación en las diferentes ciencias y disciplinas tales como la administración, economía, comunicación, agricultura, medicina, física, ciencias políticas, psicología, sociología, ingeniería, entre otras.

Por tanto, este texto servirá de apoyo para estudiantes y profesionales de áreas que en su quehacer necesitan de la estadística, puesto que el contenido del libro soporta los conceptos generales de la estadística descriptiva, muestreo, ordenación de datos, representación gráfica, medidas de tendencia central, de posición, de dispersión y de forma. Adicionalmente, en los capítulos 11 y 12 se explican las aplicaciones para el análisis estadístico mediante “análisis de datos” de Microsoft Excel y “estadísticos descriptivos” de SPSS.

El propósito del texto es ofrecer una guía para la comprensión de los elementos básicos que permiten analizar un conjunto de datos. Es por ello que, en cada uno de los capítulos, se desarrollan los temas desde lo conceptual, acompañados de ejemplos prácticos y culminando con ejercicios de aplicación; de esta manera, siendo una herramienta didáctica, el libro genera motivación y posibilita el aprendizaje del tema en aquellas personas que ven en la estadística una herramienta para la toma de decisiones.

1 CONCEPTOS GENERALES SOBRE ESTADÍSTICA

OBJETIVOS:

- » Conocer los orígenes y conceptos fundamentales de la estadística.
- » Diferenciar los tipos de variables y sus escalas de medición.

CONTENIDO:

- 1.1 Historia
- 1.2 Definición
- 1.3 Categorías
- 1.4 Unidad de análisis o de investigación
- 1.5 Población
- 1.6 Muestra
- 1.7 Parámetros y estadígrafos
- 1.8 Variables
- 1.9 Escalas de medición
- 1.10 Ejercicios de aplicación

1. CONCEPTOS GENERALES SOBRE ESTADÍSTICA

1.1 Historia

La estadística puede ser considerada como una de las ciencias más antiguas entre las matemáticas. Según lo relata Bonnet (2003), en el antiguo Egipto, en el año 3050 antes de Cristo, los faraones recopilaban información relacionada con la población y la riqueza que poseía el país, con el objetivo de preparar la construcción de las pirámides. En este mismo territorio, Ramsés II aplicó un censo sobre las tierras para realizar un nuevo reparto; práctica de registro que efectuaron 69 veces, como se narra en la historia, para la estimación de impuestos, derechos de voto, proyección de la guerra, entre otros propósitos. Así mismo, cuarenta siglos atrás, los chinos realizaron censos, mientras que los griegos lo hacían periódicamente con fines tributarios, sociales o militares.

Narra David Ruíz Muñoz (2005) que los romanos, considerados maestros de la organización política, fueron los que mejor emplearon la estadística. Realizaban censos de la población cada cinco años para identificar nacimientos, defunciones y matrimonios, para inventariar el ganado y las riquezas en las tierras adquiridas después de la guerra. Luego de la caída del Imperio Romano, durante los mil años siguientes se hicieron pocos cálculos estadísticos, exceptuando las tierras relacionadas con la Iglesia, adquiridas por Pipino el Breve en el año 758, y por Carlomagno en el 762 después de Cristo.

Por su parte, durante los siglos XV, XVI, y XVII grandes pensadores, entre ellos Leonardo da Vinci, Nicolás Copérnico, Galileo, Neper, William Harvey, Sir Francis Bacon y René Descartes, desarrollaron técnicas que aportaron al método científico, de tal manera que al crearse los Estados Nacionales ya se contaba con un método para el análisis de datos económicos y surgió con mayor fuerza el comercio internacional (Bonnet, 2003).

En el siglo XVII se realizaron aportes más concretos sobre métodos de observación y análisis cuantitativo, lo cual permitió ampliar los conceptos de la inferencia y la teoría estadística. Los estudiosos de esta época se interesaron por la estadística y la demografía como resultado de la incertidumbre en relación a la dinámica de la población; es decir, si aumentaba, disminuía o permanecía constante. A finales del siglo XVII e inicios del siglo XVIII, algunos matemáticos como Bernoulli, Francis Maseres, Lagrange y Laplace introdujeron la teoría de probabilidades, que en principio era usada en los juegos de azar y a finales del siglo XVIII inicia su aplicabilidad en los grandes problemas científicos (Ruíz Muñoz, 2005).

Bonnet (2003), citado por Posada y Buitrago (2006), plantea lo siguiente:

Godofredo Achenwall, profesor de la Universidad de Gotinga, acuñó en 1760 la palabra estadística, que extrajo del término italiano *statista* (estadista). Creía, y con sobrada razón, que los datos de la nueva ciencia serían el aliado más eficaz del gobernante consciente. La raíz remota de la palabra se halla, por otra parte, en el término latino *status*, que significa estado o situación. Esta etimología aumenta el valor intrínseco de la palabra, por cuanto la estadística revela el sentido cuantitativo de las más variadas situaciones (p. 4).

Con la incursión de la estadística en las ciencias sociales, Jacques Quételet interpreta la teoría de la probabilidad y aplica el principio de “promedios y de variabilidad” a los fenómenos sociales. Quételet fue considerado como el primero en llevar los métodos estadísticos a las diferentes ramas de la ciencia. Posteriormente, entre 1800 y 1820 Laplace y Gauss, mediante la teoría de los mínimos cuadrados, lograron un avance conceptual y matemático para la estadística (Bonnet, 2003).

Por su parte, Ruíz (2005), citado por Posada y Buitrago (2006), plantea que:

A finales del siglo XIX, Sir Francis Gaston ideó el método conocido por Correlación, que tenía por objeto medir la influencia relativa de los factores sobre las variables. De aquí partió el desarrollo del coeficiente de correlación creado por Karl Pearson y otros cultivadores de la ciencia biométrica como J. Pease Norton, R. H. Hooker y G. Udny Yule, que efectuaron amplios estudios sobre la medida de las relaciones (p. 10).

1.2 Definición

Para definir la Estadística es preciso mencionar que su objetivo es recopilar información de orden cualitativa o cuantitativa, perteneciente a individuos, grupos, hechos o fenómenos, y deducir a partir del análisis de los datos respuestas a interrogantes o proyecciones futuras.

La estadística, en general, estudia los métodos empleados en la recolección, organización, resumen, análisis e interpretación de datos, con el fin de obtener validez en las conclusiones y tomar decisiones de manera razonable y efectiva (Spiegel, 2013).

En el mundo académico se confunden varios términos relacionados con la estadística; por tal razón, Alzate (2004) aclara que esta palabra tiene tres significados:

La palabra estadística, en primer término, se usa para referirse a la información estadística; también se utiliza para referirse al conjunto de técnicas y métodos que se utilizan para analizar la información estadística; y el término estadístico, en singular y en masculino, se refiere a una medida derivada de una muestra (p. 3).

1.3 Categorías

La estadística persigue una finalidad orientada al análisis de datos de forma secuencial; sin embargo, para efectos de su estudio, se habla de dos tipos: estadística descriptiva y estadística inferencial (Lind, Mason y Marchal, 2000).

La estadística descriptiva está orientada a la presentación de datos mediante tablas y gráficas que permiten resumir o describir el comportamiento de los mismos, sin realizar inferencias sobre ellos debido a que son obtenidos de una parte de la población.

La estadística inferencial, en cambio, se deriva de la muestra, de mediciones u observaciones que se han hecho a una parte representativa de la población, por lo cual busca establecer generalidades para la población a partir de los datos obtenidos. En consecuencia, la estadística inferencial investiga o analiza una población a partir de la muestra seleccionada, utilizando técnicas adecuadas del muestreo.

1.4 Unidad de análisis o de investigación

La unidad de análisis o de investigación es el elemento al cual se investiga. Éste puede ser un objeto, una persona, un grupo de individuos o un acontecimiento. La unidad debe ser coherente con el tipo de investigación que se está realizando y se debe tener en cuenta que las características que posea el elemento no sean ambiguas, sean entendibles y que permitan mediciones y comparaciones (Ruíz Muñoz, 2005).

1.5 Población

La población o universo es un conjunto de elementos a los cuales se le estudian algunas características comunes; por ejemplo, los docentes de una institución educativa, las empresas de un sector productivo, los barrios de una ciudad, los artículos vendidos en un supermercado, las calificaciones de una prueba de aptitud, entre otros.

La población puede ser finita o infinita. Se estima que una población es *finita* cuando el número de los elementos que la integran es conocido por el investigador; tal es el caso de los barrios de una ciudad, los docentes de una universidad, los operarios de una industria, etc., mientras que para la población *infinita* no se conoce el número de elementos, ya sea porque es muy grande o porque se sabe que existe pero no se conoce el tamaño, por ejemplo: los lanzamientos de un dado, el número de veces que una persona puede pasar por un sitio y demás.

1.6 Muestra

La muestra se define como un conjunto de elementos seleccionados adecuadamente, que pertenecen a una población determinada, o sea que es una parte de la población o universo. Al seleccionar una muestra se pretende que el análisis realizado en ella pueda proporcionar conclusiones similares a las que se lograrían si se hubiese estudiado la totalidad de elementos de la población; por tal razón, la muestra debe ser representativa.

En la mayoría de los estudios se procura que el número de elementos de la muestra sea cercano al número de la población para evitar errores generados por el muestreo; sin embargo, con el fin de optimizar recursos de tiempo, dinero, etc., se asumen los errores generados por la diferencia en el número de unidades entre la muestra y la población y se acude al muestreo. En los casos en los cuales el número de elementos de la muestra es igual al de la población, el estudio se denomina *censo*.

1.7 Parámetros y estadígrafos

Los parámetros son medidas cuantitativas que describen una característica de la población, entre ellas están: media aritmética, varianza y coeficiente de variación. En una población se presentan muchas características y, en consecuencia tendrá varios parámetros. Los estadígrafos o estadísticas son medidas cuantitativas que describen una característica de la muestra y se consideran estimadores para la población.

1.8 Variables

Se considera como variable cualquier característica o propiedad general de una población que sea posible medir con distintos valores o describir con diferentes modalidades, por ejemplo: el coeficiente intelectual de los estudiantes de un grupo puede tener diferentes valores, o el estado civil de los empleados de una organización puede estar caracterizado como soltero, casado, separado, entre otros. Así, estas dos características se consideran variables porque, como el término lo indica, varían al medirse o caracterizarse de una unidad de análisis a otra.

En algunos casos, las características de las unidades de análisis pueden ser medidas, mientras que en otros solo es posible describirlas. Para el ejemplo anterior, el coeficiente intelectual es posible medirse en los estudiantes, lo cual obedece a una característica cuantitativa, y el estado civil en los empleados solo se puede describir (no medir), por ser una característica cualitativa. En este sentido, las variables pueden diferenciarse en dos grupos: cualitativas y cuantitativas.

1.8.1 Variables cualitativas

Las variables cualitativas son aquellas que representan atributos de los elementos y no permiten una representación numérica definida. Sin embargo, algunas cualidades pueden ser representadas por códigos numéricos que, en el fondo, generan categorías de orden cualitativo. Entre las variables cualitativas están: el estrato socioeconómico, el estado civil, la profesión, el color de una flor, entre otras.

1.8.2 Variables cuantitativas

Estas variables permiten una escala numérica y las características de los elementos son observados cuantitativamente a través de una medida y una escala definidas. Entre las variables cuantitativas se

encuentran: el salario de los empleados, la talla de una persona, el peso, el número de hijos en una familia, el número de artículos vendidos en un almacén, entre otros.

Las variables cualitativas y cuantitativas se representan con letras mayúsculas del alfabeto (X, Y, Z...) y los atributos de cada variable se simbolizan con letras minúsculas en compañía de subíndices. Por ejemplo, la variable estado civil de los empleados en una empresa puede ser representada por la letra X y sus posibles atributos de “soltero, casado, separado” se representan por, x_1 : soltero, x_2 : casado, x_3 : separado.

Las variables cuantitativas pueden ser clasificadas en dos grupos: continuas y discretas.

Una variable es **cuantitativa continua** si a lo largo de un intervalo puede tomar cualquier valor; es decir, entre uno y otro valor de la variable siempre puede existir otro valor intermedio. Son variables cuantitativas continuas la talla o altura de personas, el peso de objetos, el salario de empleados, el tiempo dedicado a realizar una actividad, la temperatura de un lugar, entre otras.

Tomando específicamente por ejemplo la talla en metros de un grupo de personas, entre 1,60 m y 1,70 m, la variable puede tomar múltiples valores, es decir no se brinca de 1,60 a 1,70 m, sino que es posible que se ubique en 1,61; 1,62; 1,63; 1,68; 1,69 m.

Una variable es **cuantitativa discreta** si solo puede tomar un valor de un conjunto de números; existen separaciones entre dos valores sucesivos que no pueden llenarse con valores intermedios; en este caso la variable toma valores aislados.

Por ejemplo, los empleados de una organización, artículos vendidos en un almacén, instituciones educativas de un sector; en estos casos, solo es posible medir la variable con valores como 15, 16, 17 u otro número entero y no con valores intermedios, tales como 15,7 o 16,8.

1.9 Escalas de medición

La escala de medición es considerada como un sistema que asigna valores numéricos a características susceptibles de medir. Normalmente, las escalas pueden ser de cuatro tipos: nominal, ordinal, de intervalos y de razón.

1.9.1 Escala nominal

La escala nominal se utiliza para representar a las variables cualitativas (también llamadas categóricas) y determina múltiples categorías identificadas por un nombre, que bien pudieron estar fijadas previamente o ser precisadas por el investigador según sus necesidades, manteniendo rigurosidad en su definición y convirtiéndolas en elementos mutuamente excluyentes, pues las categorías son exclusivas y solo existe una para cada elemento de la población (Triola, 2000); algunos ejemplos: color del cabello (negro, rubio, castaño, otro); estado de un artículo (bueno, imperfecto); género de los estudiantes (masculino, femenino).

1.9.2 Escala ordinal

Esta escala se caracteriza por presentar niveles con un rango determinado, lo que facilita la comparación entre ellos y es posible diferenciarlos como “mayor que” o “menor que”. Es importante resaltar que la distancia entre un nivel y otro adyacente no es la misma (Suárez y Tapia, 2012). Ejemplos: estado de salud de una persona: sano, ligeramente afectado, enfermo, muy enfermo. Producción en una empresa: alta, media, baja.

A pesar de que los niveles se pueden representar por un número, éste cumple la función de etiqueta y no es posible usarlo como cantidad numérica en operaciones matemáticas, tal es el caso del estrato socioeconómico, donde los números representan un nivel satisfacción de necesidades y un orden definido en los sectores a los cuales pertenece: estrato $1 < 2 < 3 < 4 < 5 < 6$.

1.9.3 Escala de intervalo

La escala de intervalo presenta mayor información que las escalas nominal y ordinal. Se caracteriza por establecer de forma ordenada los niveles y si la distancia entre uno y otro es la misma, lo cual conlleva a usar una unidad de distancia de referencia. Por esta razón, esta escala permite relacionar intervalos y de esta forma se puede observar, por ejemplo, que la distancia entre 5 y 6 es la misma que existe entre 23 y 24. En esta escala se asigna el punto cero como una medida arbitraria y no implica ausencia de la característica que se está midiendo (Posada y Buitrago, 2008).

Un ejemplo típico para esta escala es la medición de la temperatura; para ello se pueden usar varios sistemas: el Celsius, Kelvin o Fahrenheit. En cualquiera de estos sistemas se observa que la distancia entre un grado y el consecutivo es la misma; además, el cero en cualquiera de ellos no implica ausencia de temperatura.

1.9.4 Escala de razones

La escala de razones es aquella que posee más información en relación a las escalas anteriores. Toma un cero no arbitrario (absoluto) que significa ausencia del atributo o la característica; esto facilita la comparación, tanto en intervalos como en razones, en cualquier sistema de medición que se utilice (Posada y Buitrago, 2008).

Por ejemplo, si un elemento posee una longitud de 4,6 cm tendrá el doble de extensión al compararse con otro elemento que mide 2,3 cm, en cualquier tipo de sistema en que se registre la longitud.

A esta escala pertenecen todas aquellas mediciones que están relacionadas con el tiempo, longitud, superficie (áreas), capacidad (volúmenes), peso, dinero, entre otras.

En términos generales, se denomina para cualquier escala de medición como débil si contiene poca información, razón por la cual restringe la aplicación de los métodos estadísticos. Mientras que las

escalas con mayor información son consideradas como fuertes y es posible analizarlas por medio de métodos específicos diseñados para explicar su comportamiento.

Si se clasifica de una manera más amplia, las variables cualitativas utilizan la escala nominal, mientras que las cuantitativas emplean las escalas de razones o de intervalo. Las variables que usan la escala ordinal se ubican en una transición entre cualitativas y cuantitativas.

1.10 Ejercicios de aplicación

1. Establezca la diferencia entre estadística descriptiva y estadística inferencial.

2. ¿Qué diferencia hay entre estadígrafo y parámetro?

3. ¿Cuándo se justifica realizar un censo en un estudio estadístico?

4. En la tabla 1 se presentan las variables analizadas en un accidente ocurrido a un grupo de empleados en una industria. Clasifique en cada uno de los casos el tipo de variable y la escala de medición.

Tabla 1. Variables para el ejercicio de aplicación 1.10

Variable	Codificación	Tipo de variable	Escala de medición
Edad	En años cumplidos		
Tiempo de antigüedad	En años cumplidos		
Estrato socioeconómico	Bajo, medio, alto		
Estado civil	Soltero, casado, viudo, separado, otro		
Escolaridad	Ninguna, primaria, secundaria, universitario, posgrado, otra		
Tipo de lesión presentada	Múltiples opciones		
Área del cuerpo afectada	Extremidades, pecho, cara, otra		
Grado de la lesión	Leve, moderada, grave		
Requirió atención médica	Sí, no		
Tiempo de incapacidad	En días		
Ha presentado previamente accidentes laborales	Sí, no		
Utiliza implementos de bioseguridad en su trabajo	Sí, no		
Fecha del accidente	Ubicación en el calendario (dd/mm/aaaa)		
En qué jornada ocurrió el accidente	Mañana, tarde, noche		
Temperatura del sitio del accidente	Temperatura en °C		
Fuma	Sí, no		
Número de cigarrillos consumidos diariamente	Nº de cigarrillos		
Consumo de alcohol	Sí, no		
Frecuencia de consumo de alcohol	Diario, semanal, quincenal, mensual		
Personas con quienes consume licor	Amigos, familiares, pareja, compañeros de trabajo, otros		
Cociente intelectual (CI)	Medido en escala de CI		
Capacidad para el estudio	Puntuación en un test		
Barrio de residencia	Múltiples opciones		

Elaboración propia

2 INTRODUCCIÓN A LOS TIPOS DE MUESTREO

OBJETIVOS:

- » Diferenciar los métodos de muestreo.
- » Identificar los posibles errores al aplicar un muestreo probalístico.

CONTENIDO:

- 2.1 Métodos de muestreo probabilístico
- 2.2 Aspectos importantes en el muestreo probabilístico
- 2.3 Métodos de muestreo no probabilístico
- 2.4 Ejercicios de aplicación

2. INTRODUCCIÓN A LOS TIPOS DE MUESTREO

El muestreo es utilizado en la estadística con la finalidad de optimizar recursos (tiempo, mano de obra, materiales e insumos). Para lograrlo, los métodos estadísticos ofrecen dos tipos de muestreo: probabilístico y no probabilístico (Martínez, 2007). En este capítulo se pretende esbozar de forma general, los conceptos relacionados con los métodos de muestreo usados en estadística.

2.1 Métodos de muestreo probabilístico

Los métodos de muestreo probabilístico buscan que todos los elementos que conforman la población tengan igual probabilidad al ser seleccionados en la muestra. Es decir, todas las muestras de un tamaño determinado que se puedan sacar de una población tendrán la misma probabilidad de ser elegidas. Este método de muestreo es el más recomendable, dado que aseguran de alguna manera la representatividad de la muestra que se extrae; se dice “de alguna manera” porque pueden emplearse técnicas para determinar el tamaño representativo de la muestra para poblaciones finitas e infinitas, a partir del nivel de confianza, error de muestreo y tamaño de la población. A continuación se describen los principales métodos de muestreo probabilístico.

2.1.1 Muestreo aleatorio simple

El muestreo aleatorio simple es más utilizado por su agilidad, sin embargo es poco útil cuando se tienen poblaciones muy grandes. Los pasos a seguir en este tipo de muestreo son los siguientes: se asigna un número a cada elemento de la población y, por medio de un mecanismo tal como balotas dentro de un recipiente o el uso de números aleatorios, se elige el número de elementos requeridos para la muestra (Martínez, 2000).

2.1.2 Muestreo aleatorio sistemático

El muestreo aleatorio sistemático consiste en asignar números a la totalidad de elementos de la población. Martínez (2000, p. 2) describe el proceso para aplicar este tipo de muestreo:

Inicialmente se toma un número aleatorio i (elegido al azar) y los elementos que conforman la muestra son aquellos que se ubican en los sitios $i, i+k, i+2k, i+3k, \dots, i+(n-1)k$, o sea que se toman los individuos de k en k , donde k es el resultado de dividir el número de elementos de la población entre el número de elementos de la muestra: $k=N/n$. El número i que se emplea como punto de partida será un número al azar entre 1 y k .

Al emplear el muestreo aleatorio sistemático se debe tener cuidado con las poblaciones en las cuales se dan periodicidades, pues al elegir elementos con periodicidad constante (k), es posible que se obtenga una muestra con diferencia constante entre los datos, mientras que en la población no se presenta este comportamiento.

2.1.3 Muestreo aleatorio estratificado

Este tipo de muestreo busca simplificar los procesos, con el objeto de minimizar el error muestral al determinar el tamaño óptimo de la muestra, básicamente cuando la población es heterogénea y presenta gran variabilidad. El procedimiento consiste en determinar categorías (o estratos) diferentes entre los elementos de la población (Martínez, 2000). Los estratos se definen con respecto a la similitud entre las características de los elementos, por ejemplo: estado civil, género y nivel salarial.

El objeto del muestreo estratificado garantiza que todos los estratos definidos por el investigador estarán representados en la muestra. Cada estrato es independiente y en cada uno de ellos se realiza el muestreo aleatorio simple para seleccionar los elementos que conformarán la muestra.

La distribución de los elementos que serán muestreados en cada estrato se conoce como afijación, y puede ser de diferentes tipos:

Afijación simple: cuando el número de elementos muestreados se reparte en iguales cantidades en todos los estratos.

Afijación proporcional: el número de elementos muestreados en cada estrato es la proporción de su tamaño con respecto al de la población.

Afijación óptima: además de la proporción en cada estrato, se tiene en cuenta la variación (desviación típica). Este tipo de afijación no es muy usada, debido a que en la mayoría de los procesos muestrales no se conoce la desviación.

Posada y Buitrago (2008) plantean el siguiente ejemplo ilustrativo:

Se está interesado en estudiar el grado de aceptación que la implantación de la reforma educativa ha tenido entre los padres de un municipio. A tal efecto se seleccionó una muestra de 600 padres de familia. Se conoce por los datos del Ministerio de Educación que de los 10.000 niños escolarizados en la básica, 7.000 acuden a colegios públicos y 3.000 a colegios privados. Como el interés es que en la muestra estén representados todos los tipos de colegio, se realiza un muestreo estratificado empleando como variable de estratificación el tipo de colegio.

Si se emplea una afijación simple serían 300 niños de cada tipo de centro, pero en este caso parece más razonable utilizar una afijación proporcional pues hay bastante diferencia en el tamaño de los estratos. Por consiguiente, se calcula la proporción para cada uno de los estratos respecto de la población, para poder reflejarlo en la muestra.

Colegios públicos: $7.000/10.000 = 0.70$

Colegios privados: $3.000/10.000 = 0.30$

Para conocer el tamaño de cada estrato en la muestra se multiplica la proporción por el tamaño muestral.

Colegios públicos: $0.70 \times 600 = 420$ padres de familia
Colegios privados: $0.30 \times 600 = 180$ padres de familia
La selección de los 420 y 180 padres de familia se realizará utilizando el muestreo aleatorio simple (p. 36).

2.1.4 Muestreo aleatorio por conglomerados

En este tipo de muestreo, el marco muestral está integrado por varios elementos que forman una unidad, la cual se conoce como conglomerado. Por ejemplo: las unidades de salud, los departamentos de una empresa, los municipios de un departamento. Este tipo de muestreo se realiza mediante la selección de un número de conglomerados de forma aleatoria y posteriormente se analizan todos los elementos que pertenecen a los conglomerados seleccionados (Martínez, 2000).

El muestreo por conglomerados suele confundirse con el muestreo estratificado. La diferencia radica en que el estratificado presenta de manera natural las agrupaciones, mientras que el muestreo por conglomerados ya están definidos los grupos de acuerdo a la finalidad del estudio.

En algunos casos, el proceso de muestreo puede llegar a ser complejo y se acude al *muestreo polietápico*, el cual consiste en realizar etapas sucesivas, utilizando en cada una de ellas alguno de los métodos de muestreo probabilístico ya mencionados. Normalmente, la primera etapa utilizada da lugar a la división de las unidades en una segunda etapa y así sucesivamente, hasta completar las unidades requeridas en la muestra.

A manera de ejemplo se menciona un caso en el cual es posible utilizar el muestreo polietápico: se requiere tomar una muestra de 400 estudiantes de grado once de los colegios públicos ubicados en la zona urbana de la ciudad de Medellín, para indagar sobre los programas e instituciones de educación superior preferidos por ellos. Para seleccionar la muestra se utilizarán cuatro etapas, las cuales se describen a continuación:

Primera etapa: selecciona al azar 8 de las 16 comunas que conforman la zona urbana del municipio de Medellín.

Segunda etapa: tomar aleatoriamente 5 instituciones educativas en cada una de las 8 comunas.

Tercera etapa: en cada institución educativa, se toma aleatoriamente un grupo de grado once.

Cuarta etapa: en cada grupo de grado once seleccionado, se toman al azar 10 estudiantes.

Finalmente, la muestra estará constituida por 400 estudiantes (8x5x1x10), a los cuales les será aplicado el instrumento.

2.2 Aspectos importantes del muestreo probabilístico

Los métodos de muestreo probabilístico, en su mayoría, utilizan como instrumento de recolección de información la encuesta, la cual posibilita errores como los que se mencionan a continuación.

2.2.1 Error de cobertura o sesgo en la selección

El éxito de lograr una selección adecuada de la muestra está determinado por la claridad en los datos de la población que entran a participar en el muestreo. El error de cobertura se presenta cuando se omiten algunos elementos que hacen parte de la población, de forma que no tienen posibilidad de entrar en la muestra, conduciendo la selección de la muestra a un proceso sesgado. Si en la lista de la población no se incluyen todos los elementos, la muestra orientará la estimación a una parte de la población y no a la población real (Martínez, 2007).

2.2.2 Error o sesgo de no respuesta

El error de no respuesta se presenta al no obtener los datos de todos los elementos de la muestra. En este caso debe intentarse varias veces usando diferentes medios (por teléfono o correo electrónico), para la aplicación de la encuesta (Martínez, 2007).

2.2.3 Error de muestreo

El error de muestreo es usual cuando se aplica el instrumento a una muestra y no a la totalidad de la población; es decir, cuando no se realiza un censo. A pesar de que este error no se puede evitar, es posible controlarlo mediante la selección de un diseño de muestreo adecuado. Este tipo de error representa la variación o las diferencias aleatorias entre las posibles muestras de la población (Martínez, 2007).

2.2.4 Error de medición

El error de medición está asociado con la baja precisión de las respuestas obtenidas en la aplicación del instrumento. Básicamente se debe a errores cometidos al formularse las preguntas o al grado de incidencia que pudo tener el entrevistador sobre el entrevistado (Martínez, 2007).

2.3. Métodos de muestreo no probabilístico

Para algunos tipos de estudios, el muestreo probabilístico puede llegar a generar altos costos, y es en estos casos donde se acude a métodos no probabilísticos. Pese a que estos métodos no permiten generalización, porque todos los elementos de la población no poseen la misma probabilidad de ser seleccionados, es posible elegir las unidades muestrales bajo determinados criterios buscando mayor representatividad en la muestra. Los métodos más usados en el muestreo probabilístico son:

2.3.1 Muestreo por cuotas

También se conoce como muestreo “accidental”, por ser muy utilizado en las encuestas de opinión. Tiene como premisa el conocimiento amplio de los estratos o grupos relevantes de la población, es por eso que se asemeja al muestreo estratificado, pero difiere al no tener el carácter aleatorio (Posada y Buitrago, 2008).

Para el muestreo se fijan “cuotas” que corresponden a un número de elementos que cumplen ciertos requerimientos, un ejemplo podría ser: 50 personas entre los 18 y 30 años, estudiantes de pregrado y residentes en una misma ciudad. En este caso, la cuota son los estudiantes de pregrado residentes en una misma ciudad y serán seleccionados los primeros 50 estudiantes que cumplan la condición.

2.3.2 Muestreo intencional u opinático

Este tipo de muestreo se utiliza cuando la población es desconocida y se requiere obtener “representativas”. Para ello, se incluyen en la muestra grupos típicos que permitan suministrar la información requerida, por ejemplo sondeos preelectorales o gustos sobre un evento (Martínez, 2007).

2.3.3 Muestreo bola de nieve

Este tipo de muestreo es utilizado en estudios con poblaciones en las cuales no se conocen los elementos que la integran, pero se sabe que existen. En este proceso se localizan algunos individuos que permitan la vinculación con otros, y así hasta obtener una muestra representa-

tiva, por ejemplo: estudios de graduados de una institución, de personas que han cometido un determinado delito, de miembros de una secta religiosa, entre otros.

2.4 Ejercicios de aplicación

1. Mencione los beneficios del muestreo en los análisis estadísticos.

2. Establezca la diferencia entre el muestreo probabilístico y no probabilístico.

3. Realice un paralelo entre las características principales de los tipos de muestreo probabilístico.

4. Realice un paralelo entre las características principales de los tipos de muestreo no probabilístico.

5. En cada uno de los siguientes enunciados, determine cuál tipo de muestreo es más recomendable aplicar y justifique las razones por las cuales se recomienda.

- a) Se debe seleccionar una muestra de 150 personas de una comunidad de 2.000 habitantes, con el objeto de hacerles una encuesta sobre los servicios públicos de salud que reciben. Los habitantes están distribuidos en 5 barrios con tamaños diferentes.

- b) Se requiere realizar una encuesta sobre las políticas públicas de un municipio, para ello se divide el municipio en 16 comunas, de estas se toman al azar la 4, 6, 7, 11, 13 y 16, y sólo se concentran en estas comunas. Se toma una muestra aleatoria de habitantes de cada una de ellas para encuestarlos.

- c) Un docente se encuentra realizando una investigación en una universidad, para constituir la muestra usa estudiantes voluntarios de la Institución.

- d) Un reportero selecciona tres o cuatro congresistas considerando que éstos reflejan la opinión general de todos los congresistas sobre una ley en particular.

- e) Se presentó una lista de los 1.500 restaurantes de una ciudad, se elige uno al azar y a partir de esta primera selección, y contando de 30 en 30, se ha ido seleccionando una muestra de 40 restaurantes.

3 TABULACIÓN DE DATOS ESTADÍSTICOS

OBJETIVOS:

- » Construir intervalos que permitan agrupar un conjunto de datos numéricos.
- » Tabular la información de un conjunto de datos agrupados en intervalos.

CONTENIDO:

- 3.1 Rango o recorrido
- 3.2 Número de intervalos o clases
- 3.3 Amplitud del intervalo de clase
- 3.4 Límites de los intervalos
- 3.5 Tabulación
- 3.6 Marca de clase o punto medio
- 3.7 Ejercicios de aplicación

3. TABULACIÓN DE DATOS ESTADÍSTICOS

Al aplicar las encuestas o acudir a una base de datos para obtener la información, es posible que a simple vista se dificulte la interpretación por la cantidad de datos que pudieron ser generados en el proceso investigativo. Por esta razón, es conveniente agrupar los datos y construir tablas que faciliten la interpretación.

La tabulación de datos es el proceso mediante el cual se toman los diferentes valores o atributos de la variable y se ubican en una columna, según el criterio de ordenación definido por el investigador, y al frente de cada valor o atributo se coloca la frecuencia.

Cuando se trata de variables cualitativas, para la tabulación de datos se recomienda seguir un patrón con orden cronológico, alfabético o con nivel jerárquico y, posteriormente, se realiza el conteo; por ejemplo, si el interés es conocer el nivel de escolaridad de un grupo de personas, se ordena de la siguiente manera (ver tabla 2):

Tabla 2. Ordenación de datos para la variable cualitativa

Nivel	Nº de personas
Primaria	20
Bachillerato	18
Técnica	22
Tecnológica	15
Profesional	25

Elaboración propia

Para la tabulación de datos que corresponden a la variable cuantitativa discreta se emplean escalas numéricas y se ordena en forma creciente; por ejemplo, si se consulta por el número de hijos de un grupo de personas, estos deben ordenarse de forma creciente (ver tabla 3):

Tabla 3. Ordenación de datos para la variable cuantitativa discreta

N° de hijos	N° de personas
0	17
1	15
2	12
3	4
4	2

Elaboración propia

Cuando los datos son numerosos o corresponden a la variable cuantitativa continua, la ordenación en forma creciente ya no es funcional y, por lo tanto, se deben agrupar los datos en intervalos, mediante el siguiente procedimiento: calcular rango o recorrido, número de intervalos o clases, amplitud del intervalo de clase, límites de los intervalos y tabulación.

3.1 Rango o recorrido

El rango o recorrido (R) se define como la variación numérica de la variable, es el recorrido que toma la variable desde el valor más pequeño hasta el valor más alto. Se calcula mediante la siguiente ecuación:

$$R = [L_s - L_i]$$

Donde, R: rango o recorrido.

L_i : límite inferior (menor valor de la variable).

L_s : límite superior (mayor valor de la variable).

Por ejemplo, en la tabla 4 se presenta el tiempo en minutos requerido por un grupo de personas para realizar una actividad.

Tabla 4. Tiempo requerido en minutos por un grupo de personas para realizar una actividad.

Persona	Tiempo (min)	Persona	Tiempo (min)	Persona	Tiempo (min)	Persona	Tiempo (min)	Persona	Tiempo (min)
1	70	11	47	21	57	31	52	41	51
2	71	12	68	22	55	32	63	42	50
3	62	13	60	23	55	33	65	43	60
4	63	14	54	24	57	34	50	44	56
5	67	15	63	25	59	35	53	45	67
6	65	16	60	26	74	36	59	46	59
7	75	17	69	27	56	37	45	47	68
8	62	18	54	28	59	38	72	48	61
9	65	19	73	29	71	39	64	49	51
10	56	20	55	30	50	40	69	50	64

Elaboración propia

Al observar los datos se identifica que el valor más alto y el más bajo son 75 y 45 minutos, respectivamente; por lo tanto:

$$L_i = 45 \text{ minutos}, L_s = 75 \text{ minutos y } R = [75 - 45] = 30 \text{ minutos.}$$

3.2 Número de intervalos o clases

El número de intervalos (m) se define de forma convencional, procurando mantenerse entre 5 y 20 intervalos o clases. Sin embargo, para evitar caer en los extremos, es posible obtener una aproximación usando la fórmula propuesta por Sturges (Martínez, 2007), la cual es:

$$m = 1 + 3,3 * \log(n)$$

Donde n es el número total de datos. En caso de que se analice la totalidad de elementos de la población, ésta se representa por N .

Para el ejemplo de la tabla 4, el número de intervalos es:

$$m = 1 + 3.3 \times \log(50)$$

$$m = 1 + 3.3 \times 1,6989$$

$$m = 1 + 5,60637$$

$$m = 6,60637$$

En este caso es posible construir 6 o 7 intervalos, pero éste debe ser definido en el paso siguiente, al calcular la amplitud del intervalo.

3.3 Amplitud del intervalo de clase

La amplitud de los intervalos (C) no es necesaria que sea igual para todos; sin embargo, con el objeto de simplificar y volverlo más funcional, es conveniente que todos los intervalos tengan igual amplitud. Para el cálculo de la amplitud del intervalo se toma el cociente entre la amplitud del rango (AR) y el número de intervalos (m) que se considere más adecuado, teniendo en cuenta que este resultado (C) debe ser una cantidad exacta.

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

Siguiendo el ejemplo presentado en la tabla 4, la amplitud del intervalo sería:

Si $m = 6$, la amplitud del intervalo es $C = 30/6$, $C = 5,0$ minutos

Si $m = 7$, la amplitud del intervalo es $C = 30/7$, $C = 4.285714286...$ minutos

En este caso se recomienda usar 6 intervalos ($m = 6$), con una amplitud de 5 minutos ($C = 5$).

Si la *amplitud del intervalo* (C) no es exacta con ninguno de los dos valores arrojados por la fórmula de Sturges, el número de intervalos se puede incrementar hasta hacer la división exacta.

3.4. Límites de los intervalos

Al construir los intervalos, cada uno de ellos está determinado por dos extremos: límite inferior (l_i) y límite superior (l_s). Para el primer intervalo, el límite inferior es igual al límite inferior del rango L_i y el límite superior de este intervalo se conforma sumando la amplitud (C) al límite inferior. El segundo intervalo parte del límite superior del primer intervalo y se le suma la amplitud para obtener el límite superior. Este proceso se repite para el total de intervalos en los cuales se agrupó el conjunto de datos. Los intervalos para el ejemplo de la tabla 4 serían (ver tabla 5):

Tabla 5. Intervalos para el tiempo (en minutos) que requiere un grupo de personas para realizar una actividad

Nº. de Intervalo	Límites de Clase l_i - l_s	Intervalos o Clase l_i - l_s
1	45 - 45 + 5 = 50	[45 - 50]
2	50 - 50 + 5 = 55	(50 - 55]
3	55 - 55 + 5 = 60	(55 - 60]
4	60 - 60 + 5 = 65	(60 - 65]
5	65 - 65 + 5 = 70	(65 - 70]
6	70 - 70 + 5 = 75	(70 - 75]

Elaboración propia

Es importante aclarar que, en el primer intervalo se toma el límite inferior y el superior, es decir cerrado en ambos extremos [45 - 50], mientras que los demás intervalos no toman el extremo inferior pero sí toman el superior, o sea abierto en extremo inferior y cerrado en el extremo superior (50 - 55]. De esta forma se da continuidad a los valores asumidos por la variable.

3.5. Tabulación

Luego de construir los intervalos, se procede al conteo o frecuencia de la información, ubicando cada dato en el intervalo al cual corresponde, como se muestra en la tabla 6.

Tabla 6. Tabulación de los datos del tiempo (en minutos) que requiere un grupo personas para realizar una actividad

Nº. de Intervalo	Intervalo (minutos)	Tabulación	Frecuencia (Nº. de personas)
1	[45 - 50]	II	2
2	(50 - 55]	IIII IIII	9
3	(55 - 60]	IIII IIII II	12
4	(60 - 65]	IIII IIII I	11
5	(65 - 70]	IIII IIII	9
6	(70 - 75]	IIII II	7

Elaboración propia

3.6 Marca de clase

La marca de clase, conocida también como punto medio, es el valor representativo para cada intervalo. Se representa por $\dot{x}_i$ y se calcula promediando el límite inferior l_i y superior l_s en cada intervalo. La fórmula para el cálculo es:

$$\dot{x}_i = \frac{l_s + l_i}{2}$$

Tabla 7. Marca de clase para los intervalos del tiempo (en minutos) requerido por un grupo de personas para realizar una actividad

Nº. de Intervalo	Intervalo (minutos)	Marca de Clase ($\dot{x}_i$)
1	[45 - 50]	47,5
2	(50 - 55]	52,5
3	(55 - 60]	57,5
4	(60 - 65]	62,5
5	(65 - 70]	67,5
6	(70 - 75]	72,5

Elaboración propia

Es importante resaltar que la marca de clase al pasar de un intervalo a otro, aumenta la misma cantidad del intervalo C.

3.7 Ejercicios de aplicación

1. Explique la razón por la cual se recomienda agrupar un conjunto de datos para el análisis estadístico.

2. La siguiente información corresponde a la evaluación del nivel de afectación que generó la construcción de una vía sobre los recursos naturales. Para ello se tomaron como referencia diez puntos de control ubicados cada dos kilómetros. Los resultados fueron:

Alta	Sin afección
Sin afección	Leve
Moderada	Leve
Moderada	Alta
Moderada	Moderada

Realice la tabulación y el conteo o frecuencia para la variable correspondiente.

- En una empresa manufacturera se realizó un control de calidad sobre la producción. Fueron seleccionados diez productos y se les identificó el número de imperfectos que presentaban. Los resultados fueron los siguientes:

1	0	2	3	2
0	2	1	4	1

Realizar la tabulación y el conteo de la variable correspondiente.

- Las siguientes son las calificaciones obtenidas por 100 aspirantes que se presentaron a un concurso sobre oratoria:

38	51	32	65	25	28	34	12	29	43
71	62	50	37	8	24	19	47	81	53
16	62	50	37	4	17	75	94	6	25
55	38	46	16	72	64	61	33	59	21
13	92	37	43	58	52	88	27	74	66
63	28	36	19	56	84	38	6	42	50
94	51	62	3	17	43	47	54	58	26
12	42	34	68	77	45	60	31	72	23
18	22	70	34	5	59	20	68	55	49
33	52	14	40	38	54	50	11	41	76

- a) Realice todos los pasos para la construcción de intervalos:
- Límites inferior y superior
 - Rango
 - Número de intervalos
 - Amplitud del intervalo
 - Marca de clase
- b) Construya la tabulación y el conteo o frecuencias de datos para cada intervalo.

- c) Responda las siguientes preguntas:

- ¿Entre qué puntuaciones está la mayor cantidad de aspirantes?

- ¿Qué porcentaje de aspirantes obtuvo los puntajes más altos?

- ¿Qué porcentaje de aspirantes obtuvo los puntajes más bajos?

- ¿Cuántos aspirantes obtuvieron los puntajes más bajos?

- ¿Cuántos aspirantes obtuvieron los puntajes más altos?

- d) Para pasar a la siguiente ronda se requiere un puntaje superior a 55 puntos. ¿Cuántos aspirantes pasarán a la siguiente ronda?

-
5. En un grupo de deportistas se requiere conocer la talla para seleccionar los integrantes del equipo de baloncesto. La talla (en metros) de los 40 estudiantes que integran el grupo es la siguiente:

1,64	1,60	1,54	1,55
1,73	1,61	1,65	1,73
1,55	1,61	1,62	1,73
1,56	1,65	1,66	1,75
1,57	1,63	1,76	1,67
1,58	1,79	1,70	1,78
1,58	1,63	1,69	1,63
1,52	1,59	1,62	1,71
1,64	1,53	1,60	1,72
1,54	1,60	1,64	1,72

- a) Realice todos los pasos para la construcción de intervalos:
- Límites inferior y superior
 - Rango
 - Número de intervalos
 - Amplitud del intervalo
 - Marca de clase
- b) Construya la tabulación y el conteo o frecuencias de datos para cada intervalo.

c) Responda las siguientes preguntas:

- ¿Entre qué tallas está la mayor cantidad de deportistas?

- ¿Qué porcentaje de deportistas obtuvo las tallas más altas?

- ¿Qué porcentaje de deportistas obtuvo las tallas más bajas?

- ¿Cuántos deportistas obtuvieron las tallas más bajas?

- ¿Cuántos deportistas obtuvieron las tallas más altas?

d) Si para ser integrante del equipo de baloncesto se requiere una talla superior a 1,70 m. ¿Cuántos deportistas integrarán el equipo?

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.

- 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).

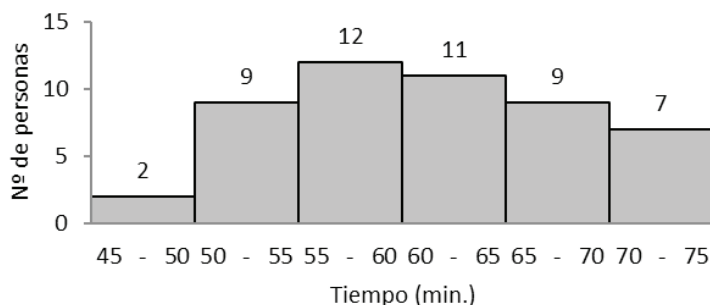
5.1 Histogramas

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).

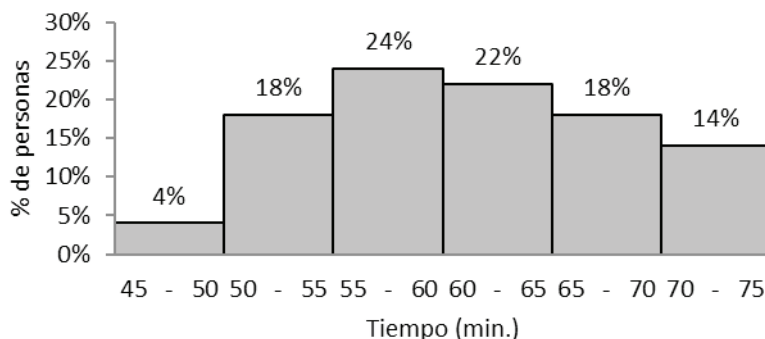
Los histogramas pueden representar a las frecuencias absolutas o relativas, dependiendo de la ubicación que se le dé a éstas sobre el eje vertical del plano cartesiano. De esta manera se obtiene el *histograma de frecuencias absolutas* o el *histograma de frecuencias relativas*, como se muestra en las 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).

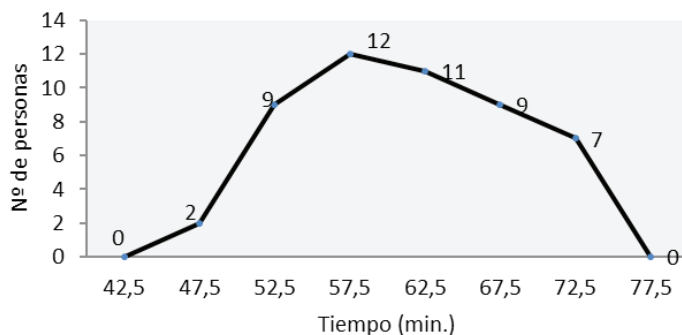
5.2 Polígono de frecuencias

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*

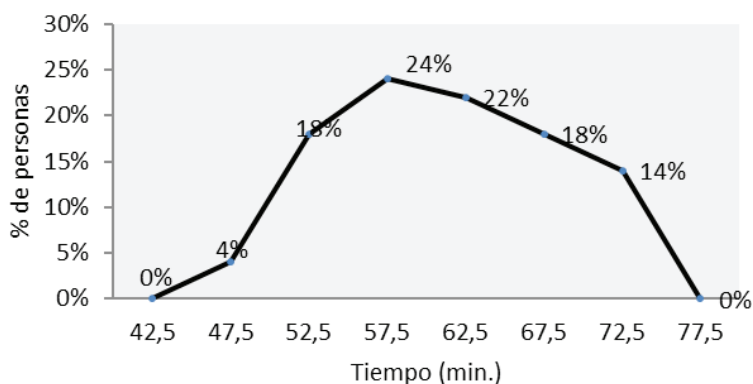
de frecuencias absolutas o relativas, dependiendo de las frecuencias utilizadas. En las gráficas 3 y 4 se ilustran estos tipos de polígonos para el ejemplo que representa a un grupo de personas que realizan una actividad.

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



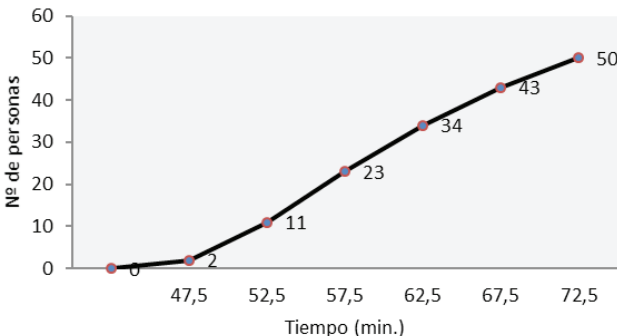
Gráfica 4. Comportamiento del número de personas según el tiempo que tardan en realizar una actividad.
(Elaboración propia).

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

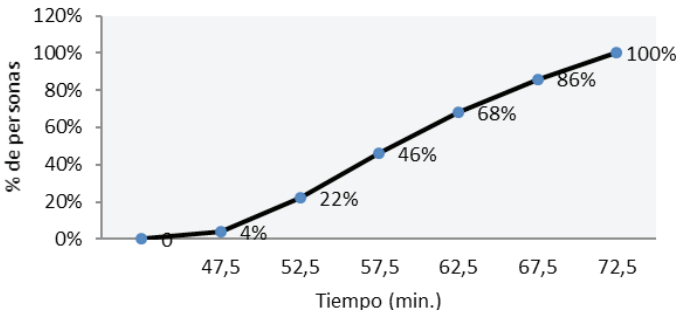
La ojiva representa el comportamiento acumulado de las unidades de investigación en relación a la variable analizada. Al igual que en los polígonos, las ojivas pueden ser construidas con las frecuencias absolutas o relativas. En las gráficas 5 y 6 se ilustran las ojivas para el ejemplo del grupo de personas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

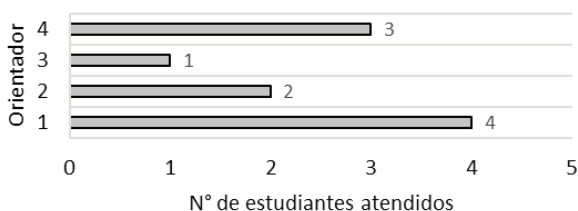
Al analizar la información se observa que los datos corresponden a una variable cuantitativa discreta. Para este tipo de variable, la tabulación se realiza ordenando en forma ascendente los valores que arroja la característica y se procede al cálculo de las frecuencias absolutas y relativas, tal como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

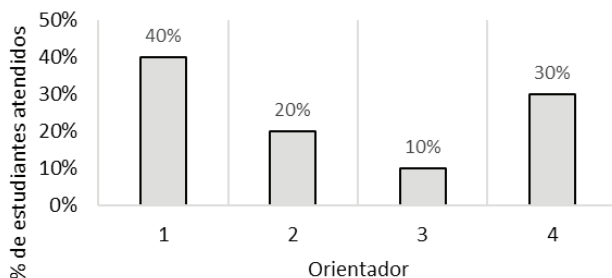


Gráfica 7. Distribución del número de estudiantes atendidos por cada orientador. (Elaboración propia).

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):

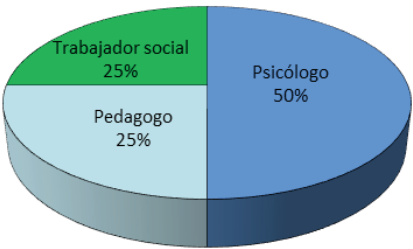
Tabla 18. Distribución de frecuencias para la variable cualitativa

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 $\bar{X}$.

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

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

Donde

$\bar{X}$: 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	$0 \times 1 = 0$
1	2	$1 \times 2 = 2$
2	4	$2 \times 4 = 8$
3	2	$3 \times 2 = 6$
4	1	$4 \times 1 = 4$
$n = \sum n_i = 10$		$\sum x_i * n_i = 20$

Elaboración propia

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

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:

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

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

$$n = \sum n_i = 50 \quad \sum \dot{x}_i * n_i = 3060$$

Elaboración propia

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

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:

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

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 $\frac{N}{2}$
2. Ubicar el intervalo cuya frecuencia absoluta acumulada N_i contiene a $\frac{N}{2}$
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 $\frac{N}{2}$

N : número total de datos de la población

N_{i-1} : frecuencia absoluta acumulada anterior al intervalo que contiene a $\frac{N}{2}$

n_i : frecuencia absoluta del intervalo que contiene a $\frac{N}{2}$

c : amplitud del intervalo que contiene a $\frac{N}{2}$

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 $\bar{X}$ 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:

$$\neq 55$$

$$A_1 12 - 9 = 3$$

$$A_2 12 - 11 = 1$$

$$c = 60 - 55 = 5$$

Por lo tanto, la moda sería:

$$Mol = i + \left(\frac{\Delta_1}{\Delta_{12} + \Delta} \right) * C = 55 + \left(\frac{3}{3+1} \right) * 5 = 55 + \frac{15}{4} = 58,75 \text{ minutos}$$

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,

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

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, $Q_3 = \frac{37+38}{2} = 37,5$ 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):

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

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 contiene 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 $k(n/4)$. ($1 \cdot 35/4 = 8,75$)
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:

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

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

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

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.

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

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

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 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 ($k = 1, 2, 3, 4, 5 \dots 10, 11, 12 \dots 98, 99$) 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:

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

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 intercuartil 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 ($\bar{x}$ 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 $(x_i - \bar{x})$; 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:

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

Donde:

$\bar{x}$: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	Desviación ($x_i - \bar{x}$)	Desviación al cuadrado ($(x_i - \bar{x})^2$)
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

$$\sum (x_i - \bar{x}) = 0 \quad \sum (x_i - \bar{x})^2 = 2,6$$

Elaboración propia

Luego, la varianza será:

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

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:

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

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

Donde:

$\bar{x}$: 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)	$\mathcal{X}$	$\mathcal{X}^2$	n_i	$\mathcal{X}_i^2 * n_i$
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
					$\sum \mathcal{X}_i^2 * n_i = 926.768,75$

Elaboración propia

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

Luego, la varianza será:

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

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:

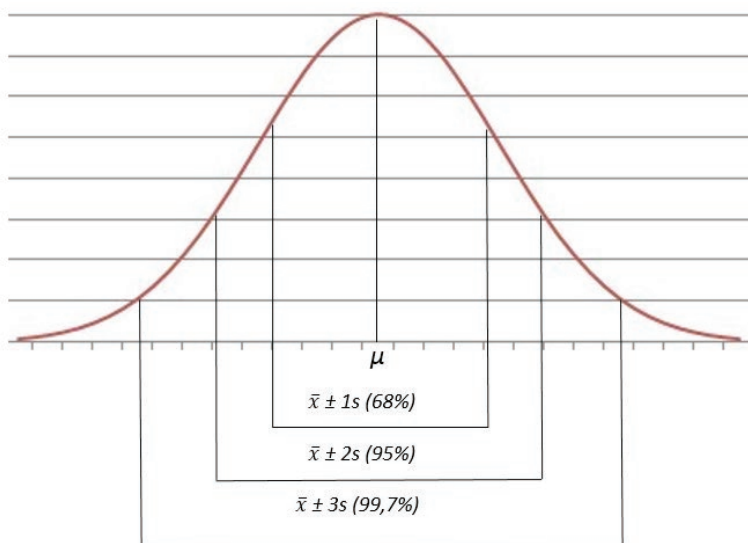
$$s = \sqrt{s^2}$$

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

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:

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

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:

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

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

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

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.

10 MEDIDAS DE FORMA

OBJETIVOS:

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

CONTENIDO:

- 10.1 Coeficiente de asimetría o sesgo
- 10.2 Coeficiente de apuntamiento o curtosis
- 10.3 Ejercicios de aplicación

10. MEDIDAS DE FORMA

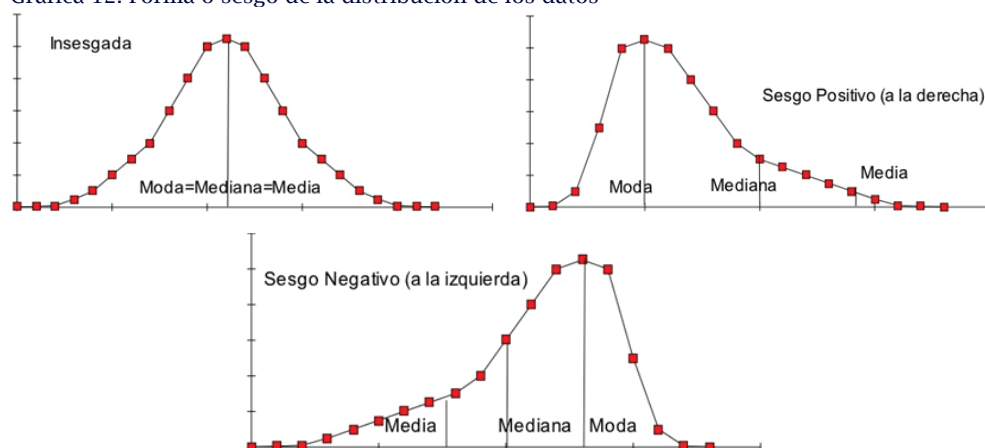
Las medidas de forma son usadas para identificar valores particulares de un conjunto de datos. Para ello toman como referencia la media aritmética y la desviación estándar de la población o la muestra.

10.1 Coeficiente de asimetría o sesgo

Al analizar la distribución de los datos es posible que exista una tendencia de estos hacia uno de los extremos (derecho o izquierdo). Esta tendencia se denomina sesgo y permite mostrar la inclinación de los datos hacia los extremos (Pierdant y Rodríguez, 2006).

Para describir el sesgo o la forma de la distribución de los datos, se comparan la media aritmética, la mediana y la moda. Si estas medidas son exactamente iguales, se considera que la distribución de los datos es insesgada o simétrica (con *sesgo cero*). En otro caso, cuando la media aritmética es superior a la mediana, la distribución de los datos estará *sesgada a la derecha* (o con *sesgo positivo*) y si la media aritmética es inferior a la mediana, los datos tendrán *sesgo a la izquierda* (o con *sesgo negativo*) (Spiegel, 2013), tal como se muestra en la gráfica 12.

Gráfica 12. Forma o sesgo de la distribución de los datos



Elaboración propia

El sesgo mantiene relación directa con la media aritmética, es decir, si la media se afecta por valores extremos, esto se verá reflejado en el sesgo. Si no hay valores extremos (muy pequeños o muy grandes) la distribución se comporta de forma simétrica, en tal forma existe una compensación entre los valores grandes y los pequeños (Martínez, 2000).

La asimetría de un conjunto de datos se puede calcular mediante varios coeficientes, entre ellos están:

10.1.1 Coeficiente de asimetría de Pearson

Relaciona la diferencia entre media aritmética y la moda con la desviación. Pese a que este coeficiente es fácil de calcular, no se utiliza con frecuencia en la práctica, ya que la distribución de los datos debe ser unimodal y moderada o ligeramente asimétrica, condiciones que no se observan de forma directa en la distribución, por lo que resultan muy exigentes (Suárez y Tapia, 2012). El Coeficiente de Pearson varía entre -3 y 3 y la fórmula es:

$$Ap = \frac{\bar{x} - M_o}{s}$$

10.1.2 Coeficiente de asimetría de Bowley

Este coeficiente es el menos usado por sus altas exigencias. Para emplearlo se requiere que tanto al extremo izquierdo como al derecho de la distribución de los datos, se presente un comportamiento similar, de lo contrario no es imposible estimar la asimetría. El cálculo se basa en la posición que presentan los cuartiles y la mediana (Suárez y Tapia, 2012). La Medida de Bowley varía entre -1 y 1 y se calcula de acuerdo con la siguiente expresión:

$$Ab = \frac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1}$$

10.1.3 Coeficiente de asimetría de Fisher

El coeficiente de Fisher está basado en las desviaciones que presentan los datos con respecto a la media. Es el coeficiente más usado para determinar la asimetría de un conjunto de datos, debido a que no es afectado por valores extremos y sólo vincula la media aritmética y la desviación (Suárez y Tapia, 2012). La fórmula para su cálculo es la siguiente:

$$g_1 = \frac{\sum (x_i - \bar{x})^3}{n * s^3} ; \text{ para datos sin agrupar.}$$

$$g_1 = \frac{\sum n_i (x_i - \bar{x})^3}{n * s^3} ; \text{ para datos agrupados en frecuencias.}$$

Es importante resaltar que para determinar el sesgo se sugiere utilizar el coeficiente de asimetría de Fisher, el cual es más confiable para analizar la similitud de la distribución de los datos con la Distribución Normal; además el valor obtenido es muy similar con el estimado por el Excel.

Debe tenerse en cuenta que el análisis del sesgo se realiza a partir del signo que arroja cualquiera de los coeficientes mencionados y, particularmente para el coeficiente de Fisher, mientras más se aleje de cero, mayor es el sesgo de la distribución de los datos, tanto a la derecha como a la izquierda. En síntesis:

Sesgo > 0: sesgo positivo o a la derecha.

Sesgo = 0: simetría en la distribución de los datos.

Sesgo < 0: sesgo negativo o a la izquierda.

Por ejemplo, el sesgo para la puntuación de la evaluación de desempeño de siete empleados del área de mercadeo de una empresa (ver tabla 34), con media aritmética de 3,6 y desviación de 0,66, calculado mediante el coeficiente de asimetría de Fisher, será:

Tabla 36. Sesgo para para la puntuación de la evaluación de desempeño de siete empleados del área de mercadeo de una empresa

Empleado	Calificación (x_i)	Media de la muestra ($\bar{x}$)	Diferencia ($x_i - \bar{x}$)	$(x_i - \bar{x})^3$
1	3,5	3,6	-0,1	-0,001
2	4,5	3,6	0,9	0,729
3	4,2	3,6	0,6	0,216
4	3,0	3,6	-0,6	-0,216
5	2,7	3,6	-0,9	-0,729
6	3,3	3,6	-0,3	-0,027
7	4,0	3,6	0,4	0,064
TOTAL				0,036

Elaboración propia

Aplicando la fórmula del coeficiente de asimetría de Fisher se tiene:

$$g_1 = \frac{\sum (x_i - \bar{x})^3}{N * s^3} = \frac{0,036}{7 * 0,66^3} = \frac{0,036}{2,01} = 0,0179$$

Dado que el valor es positivo, el sesgo es a la derecha o positivo, es decir, la calificación de la mayoría de los empleados tiende a estar por debajo del promedio de 3,6.

10.2 Coeficiente de apuntamiento o curtosis

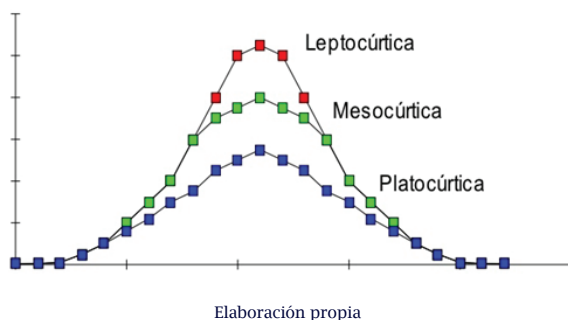
La curtosis es una medida que permite analizar la concentración de los datos alrededor de los valores medios de la muestra (Mendenhall, 1995). Se calcula con el coeficiente de Fisher para la curtosis; la ecuación es la siguiente:

$$g_2 = \frac{\sum (x_i - \bar{X})^4}{n * s^4} - 3 \quad \text{para datos sin agrupar.}$$

$$g_2 = \frac{\sum n_{ii}(x - \bar{X})^4}{n * s^4} - 3 \quad \text{para datos agrupados en frecuencias.}$$

El coeficiente de curtosis diferencia tres clases de distribuciones (Suárez y Tapia, 2012), que se ilustran en la gráfica 13:

Gráfica 13. Tipos de distribución según el coeficiente de curtosis



10.2.1 Distribución mesocúrtica

El grado de concentración alrededor de los valores centrales de la variable es moderado. Este mismo comportamiento se presenta en una distribución simétrica o normal (Alzate, 2004). El coeficiente presenta valor igual a cero.

10.2.2 Distribución leptocúrtica

El grado de concentración alrededor de los valores centrales de la variable es elevado, lo que la hace ver de forma puntiaguda, dado que las frecuencias altas están alrededor de la media (Martínez, 2007). El valor del coeficiente es mayor a cero.

10.2.3 Distribución platocúrtica

El grado de concentración alrededor de los valores centrales de la variable es reducido, mostrándose de forma aplanada dado que las frecuencias bajas están alrededor de la media (Suárez y Tapia, 2012). El valor del coeficiente es menor a cero.

Para el ejemplo sobre la puntuación de la evaluación de desempeño de siete empleados del área de mercadeo de una empresa con media aritmética de 3,6 y desviación estándar de 0,66, la curtosis será (ver tabla 37):

Tabla 37. Curtosis para para la puntuación de la evaluación de desempeño de siete empleadosdel área de mercadeo de una empresa

Empleado	Calificación (x_i)	Media de la muestra ($\bar{x}$)	Diferencia ($x_i - \bar{x}$)	$(x_i - \bar{x})^4$
1	3,5	3,6	-0,1	-0,001
2	4,5	3,6	0,9	0,6561
3	4,2	3,6	0,6	0,1296
4	3,0	3,6	-0,6	0,1296
5	2,7	3,6	-0,9	0,6561
6	3,3	3,6	-0,3	0,0081
7	4,0	3,6	0,4	0,0256
Total				1,6052

Elaboración propia

Aplicando la fórmula del coeficiente de asimetría de Fisher se tiene:

$$g_2 = \frac{\sum (x_i - \bar{X})^4}{N * s^4} - 3 = \frac{1,6052}{7 * 0,66^4} - 3 = \frac{1,6052}{1,3282} - 3 = 1,208 - 3 = -1,792$$

Dado que el coeficiente de Fisher para la curtosis es -1,792, se considera una distribución platicúrtica, lo cual significa que existe reducida concentración de los datos alrededor de los valores centrales de la distribución.

10.3 Ejercicios de aplicación

A partir de los datos presentados en los capítulos 8 y 9, sobre la prueba de comprensión lectora para aspirantes de ingreso a la universidad para los programas de Odontología y Trabajo social, se pide:

1. Calcular el coeficiente de asimetría de Fisher y analizar el sesgo de la distribución de los puntos obtenidos por los estudiantes.
2. Calcular el coeficiente de Fisher para la curtosis y analizar la concentración de los datos alrededor de las medidas centrales.

11 ESTADÍSTICA DESCRIPTIVA MEDIANTE MICROSOFT EXCEL

OBJETIVOS:

- » Conocer la herramienta de Microsoft Excel para el análisis descriptivo de un conjunto de datos.
- » Calcular e interpretar los estadígrafos de un conjunto de datos por medio de Microsoft Excel.

CONTENIDO:

- 11.1 Activación de la función “Análisis de datos”
- 11.2 Estadística descriptiva
- 11.3 Caracterización e interpretación de los resultados
- 11.4 Ejercicios de aplicación

11. ESTADÍSTICA DESCRIPTIVA MEDIANTE MICROSOFT EXCEL

El programa de Microsoft Excel ofrece una herramienta, diferente a la función de fórmulas, que permite realizar todos los cálculos correspondientes a la estadística descriptiva. Por razones desconocidas, esta herramienta se encuentra inactiva en la mayoría de los equipos y es preciso activarla para lograr un mayor beneficio.

11.1 Activación de la función “Análisis de datos”

A continuación se detallan los pasos para activar la función “Análisis de datos”, en la cual se encuentra la herramienta “Estadística descriptiva”.

- » Inicialmente, al ubicar en Excel la pestaña DATOS se puede evidenciar si la función se encuentra activa o inactiva en el equipo. Normalmente no aparece en la parte superior derecha, como se ilustra con un círculo en la gráfica 14.

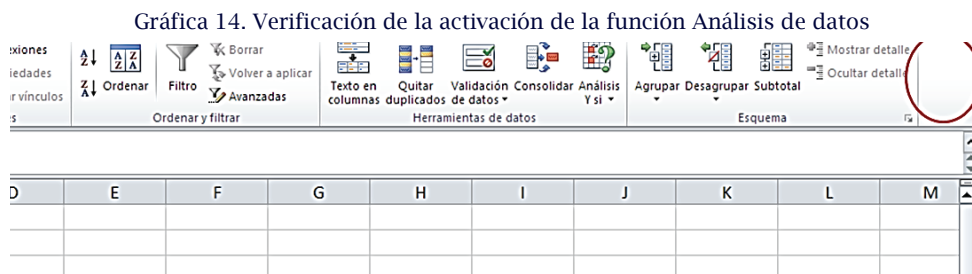


Imagen real del programa Excel

- » En la pestaña superior izquierda, al dar clic en el menú ARCHIVO se ubica el icono de OPCIONES, señalado con un círculo en la gráfica 15.

Gráfica 15. Ubicación del icono “Opciones” para la activación de la función “Análisis de datos”

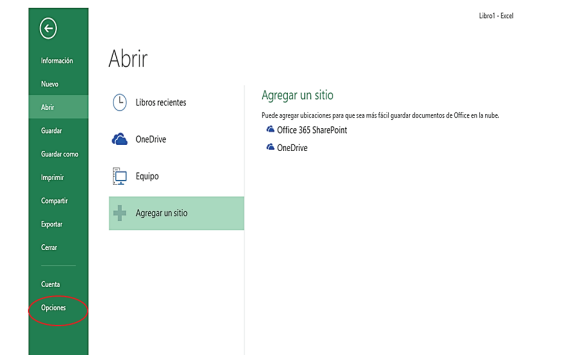


Imagen real del programa Excel

- » Al dar clic en OPCIONES, aparece un cuadro de diálogo “Opciones de Excel”. Se da clic en COMPLEMENTOS (ver gráfica 16).

Gráfica 16. Ubicación de la opción “Complementos” para la activación de la función “Análisis de datos”

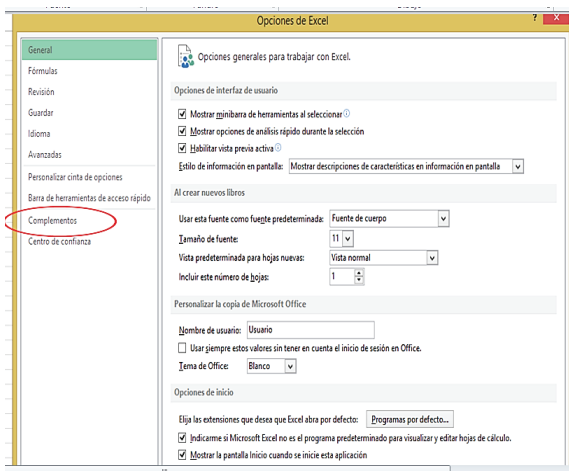


Imagen real del programa Excel

- » El icono de COMPLEMENTOS ofrece un cuadro de diálogo en el cual se debe seleccionar “Complementos de Excel” y dar clic en “ir...” (ver gráfica 17).

Gráfica 17. Direccionamiento de la opción “Complementos de excel”.

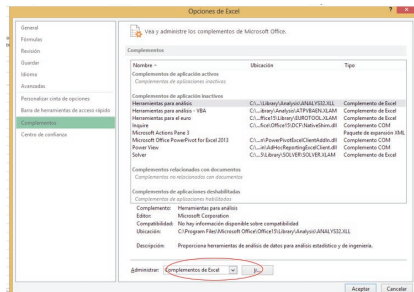


Imagen real del programa Excel

- » Esta opción ofrece un nuevo cuadro de diálogo denominado “Complementos”, en el cual se debe seleccionar “Herramientas para análisis” y luego dar clic en “Aceptar” (ver gráfica 18).

Gráfica 18. Activación de la opción “Herramientas para análisis”.

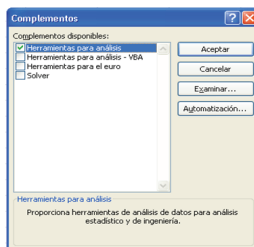


Imagen real del programa Excel

- » Finalmente, aparece en la parte superior derecha de la pestaña DATOS la función activa “Análisis de datos”, como se ilustra en la gráfica 19.

Gráfica 19. Confirmación de la activación de la herramienta “Análisis de datos análisis”

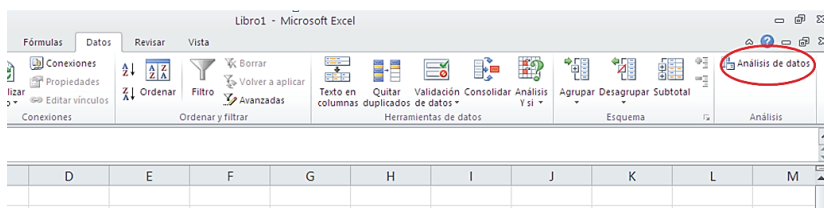


Imagen real del programa Excel

De esta manera queda activa la función “Análisis de datos” en el equipo y podrá ser usada en los análisis estadísticos.

11.2 Estadística descriptiva

Para realizar el análisis estadístico de un conjunto de datos mediante la función “Análisis de datos” se debe generar la base de datos, teniendo en cuenta que éstos deben ser valores numéricos. Posteriormente, se ingresa desde la pestaña DATOS a “Análisis de datos”, que conduce al cuadro de diálogo que contiene varias opciones para analizar un conjunto de datos. Para el análisis estadístico se selecciona la opción “Estadística descriptiva” y “Aceptar” (ver gráfica 20).

Gráfica 20. Ubicación de la opción “Estadística descriptiva”

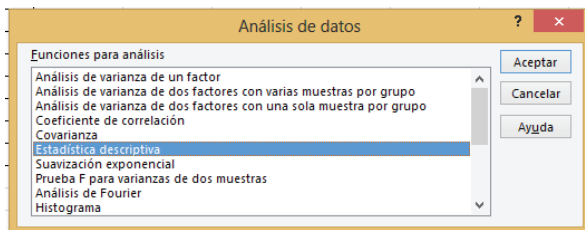


Imagen real del programa Excel

Para ingresar la información al cuadro de diálogo “Estadística descriptiva” debe disponerse de la información estructurada en una base de datos. A continuación se describen las opciones requeridas en el cuadro de diálogo (ver gráfica 21):

- » “Rango de entrada”: debe seleccionarse la columna de datos numéricos con su respectivo título.
- » “Agrupar por”: seleccionar columnas.
- » Seleccionar “Rótulos en la primera fila”.
- » “Opción de salida”: puede ser “Rango de salida” seleccionando una celda para los resultados o “En una hoja nueva” o “En un libro nuevo”.
- » Seleccionar “Resumen de estadísticas”.
- » Dar clic en “Aceptar”.

Gráfica 21. Opciones requeridas en “Estadística descriptiva” para el análisis de datos.

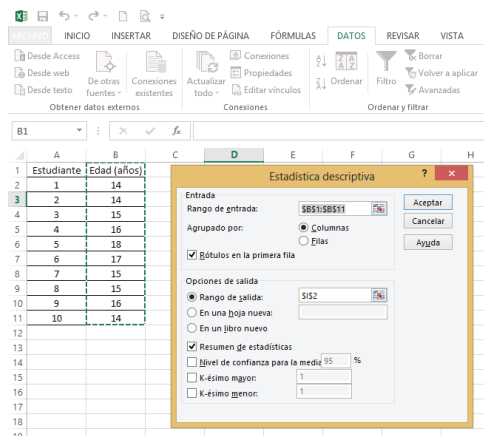


Imagen real del programa Excel

Finalmente aparecen los resultados para el análisis estadístico.

Gráfica 22. Resultados obtenidos en “Estadística descriptiva” para el análisis de datos

ARCHIVO INICIO INSERTAR DISEÑO DE PÁGINA FÓRMULAS DATOS REVISAR VISTA						
Obtener datos externos Conexiones Actualizar todo Conexiones Ordenar y filtrar						
L15						
A	B	C	D	E	F	G
1	Estudiante	Edad (años)				
2	1	14				
3	2	14				
4	3	15				
5	4	16				
6	5	18				
7	6	17				
8	7	15				
9	8	15				
10	9	16				
11	10	14				
12						
13						
14						
15						
16						
17						

Edad (años)		
Media		15,4
Error típico		0,42687495
Mediana		15
Moda		14
Desviación estándar		1,34989712
Varianza de la muestra		1,82222222
Curtosis		-0,12524433
Coefficiente de asimetría		0,77241659
Rango		4
Mínimo		14
Máximo		18
Suma		154
Cuenta		10

Imagen real del programa Excel

11.3 Caracterización e interpretación de resultados

Una vez obtenidos los resultados para el análisis de un conjunto de datos, por medio de la herramienta Estadística Descriptiva, es necesario interpretarlos para comprender el comportamiento de la variable o característica que se está estudiando y facilitar la toma de decisiones. A continuación se describe y se interpreta cada uno de los estadígrafos obtenidos en el ejemplo ilustrado sobre la edad en años de un grupo de estudiantes:

- » **Media:** representa el promedio de los argumentos, o sea que la edad promedio de los estudiantes es de 15,4 años.
- » **Error típico:** es una medida del promedio de las desviaciones de las posibles muestras que se pueden obtener en una población. Mientras más pequeño sea el error típico, más se parecen los datos al valor medio de la población. El error típico para la edad es de 0,42 años, lo cual significa que la diferencia del promedio de la muestra se aleja 0,42 años del promedio de la población.
- » **Mediana:** es el valor central de los datos, de tal manera que la mitad de los datos son superiores a la mediana y la otra mitad son menores a ella. La mediana para el ejemplo significa que el 50% de los estudiantes tienen una edad inferior o igual a 15 años, y el otro 50% superior a 15 años.
- » **Moda:** es el dato que presenta la frecuencia más alta en el colectivo. La edad que más se repite en el grupo de estudiantes es 14 años.
- » **Desviación estándar:** es el valor promedio de dispersión de los datos respecto a la media. La variación de la edad de los estudiantes es de 1,3 años alrededor de la media de 15,4 años. Podría decirse que la mayoría de los estudiantes están entre 14,1 ($15,4 - 1,3$) y 16,7 ($15,4 + 1,3$) años.

- » **Varianza de la muestra:** es el cuadrado de la desviación estándar. Este estadígrafo no se interpreta debido a que representaría las unidades cuadradas de medición de la variable.
- » **Curtosis:** describe la elevación o aplanamiento de los datos. Para el ejemplo, el valor de la curtosis de $-0,12$ indica una distribución platicúrtica, es decir, baja concentración de la edad de los estudiantes alrededor de los valores centrales, con tendencia a ser una distribución relativamente plana.
- » **Coefficiente de asimetría:** caracteriza el grado de asimetría o sesgo de la distribución de los datos con respecto a la media. La asimetría de la edad de los estudiantes es positiva ($0,77$); indica una distribución unilateral con sesgo positivo o a la derecha, es decir, la mayoría de las edades están por debajo de la media de $15,4$ años.
- » **Rango:** muestra el recorrido de la variable, es decir la diferencia entre el máximo valor y el mínimo valor. La edad de los estudiantes presenta un recorrido de 4 años, el cual oscila entre 14 y 18 años.
- » **Mínimo:** es el valor más pequeño del conjunto de datos. La edad más mínima que hay en el grupo de estudiantes es de 14 años.
- » **Máximo:** es el dato de mayor valor que se presenta en el colectivo. La edad máxima que hay en el grupo de estudiantes es de 18 años.
- » **Suma:** es el resultado de la suma de los valores del conjunto de datos. El total de las edades del grupo de estudiantes es de 154 años.
- » **Cuenta:** es el número de datos que se presenta en el colectivo. El número de estudiantes a los que se les calculó la edad fue de 10 .

11.4 Ejercicios de aplicación

El siguiente conjunto de datos corresponde al resultado de la competencia genérica razonamiento cuantitativo de 200 aspirantes a un programa de educación superior.

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

Con las herramientas “Análisis de datos” y “Estadística descriptiva”, realice el análisis estadístico para los resultados de los aspirantes, calculando e interpretando las siguientes medidas:

- a) Media
- b) Error típico
- c) Mediana
- d) Moda
- e) Desviación estándar
- f) Varianza

- g) Curtosis
- h) Coeficiente de asimetría
- i) Rango
- j) Mínimo
- k) Máximo
- l) Suma
- m) Cuenta

12 ANÁLISIS ESTADÍSTICO BAJO SPSS

OBJETIVOS:

- » Conocer algunos componentes del software SPSS para el análisis descriptivo de un conjunto de datos.
- » Calcular e interpretar los estadígrafos de un conjunto de datos por medio de SPSS.

CONTENIDO:

- 12.1 Ingreso al programa SPSS
- 12.2 Preparación de variables
- 12.3 Ingreso de datos
- 12.4 Resultados
- 12.5 Ejercicios de aplicación

12. ANÁLISIS ESTADÍSTICO BAJO SPSS

IBM SPSS es un software estadístico para el análisis de datos, conocido como SPSS. En un principio se diseñó para las ciencias sociales y actualmente se utiliza en las diferentes ciencias (Pérez, 2005).

Para el alcance del texto sólo se abordará la aplicación del SPSS en el procesamiento de los datos, la visualización de los resultados mediante tablas y gráficas, y el análisis descriptivo de los datos. Para ello se parte de conceptos expuestos en unidades anteriores.

12.1 Ingreso al programa SPSS

Al seleccionar SPSS se abrirá una ventana; en la parte inferior aparece la barra de estado y en ella, dos pestañas esenciales para el ingreso de los datos: “Vista de datos” y “Vista de variables”. En la parte superior se encuentra la barra de menú, con las operaciones posibles a realizar (Pérez, 2005), como se indica en la gráfica 22.

Archivo: permite crear, abrir o guardar archivos de SPSS, muestra información de datos, imprime, entre otros.

Edición: deshace, borra, copia, corta, pega, entre otros.

Ver: barra de estado, barras de herramientas, fuente, etc.

Datos: define fechas y efectúa modificaciones en el archivo, selecciona, añade, pondera, ordena, etc.

Transformar: modifica y genera nuevas variables.

Analizar: análisis estadísticos.

Gráficos: edición y creación de gráficas.

Utilidades: cambiar fuentes, información sobre archivos.

Ventana: funciones habituales para controlar las ventanas.

Gráfica 22. Componentes del menú de SPSS

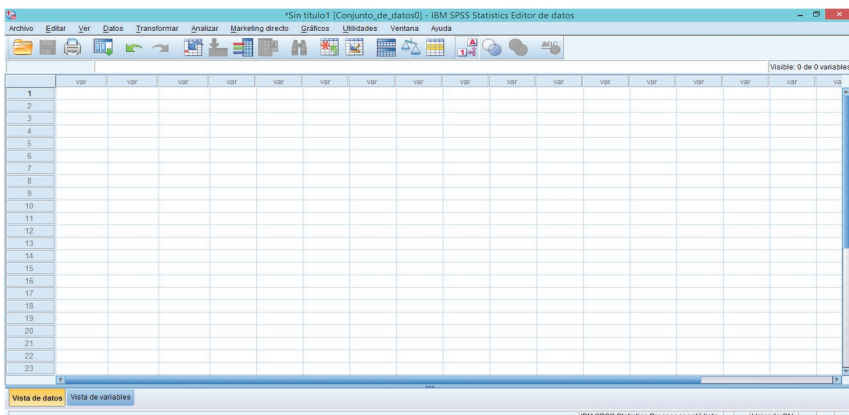


Imagen real del programa SPSS

12.2 Preparación de variables

El primer paso para ingresar los datos en un archivo es definir las variables; esto se puede realizar por cualquiera de los siguientes procedimientos: hacer doble clic sobre el encabezamiento de la variable o seleccionar, en la parte inferior, la pestaña “Vista de variables”, como se muestra en la gráfica 23.

Gráfica 23. Ingreso de datos en SPSS

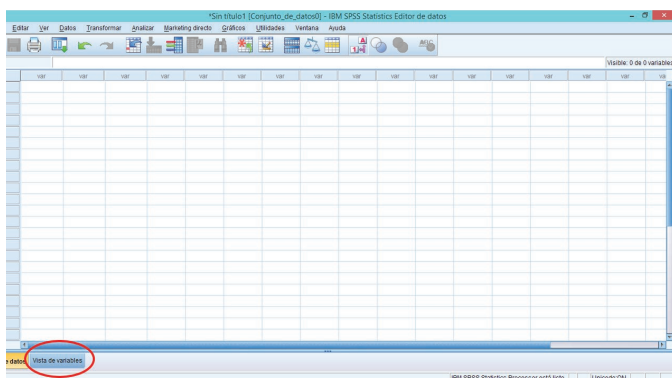


Imagen real del programa SPSS

Posteriormente aparece una ventana, en la cual se deben definir cada una de las variables, basándose en las opciones que se ofrecen en las columnas con los siguientes rótulos: Nombre, Tipo, Anchura, Decimales, Etiqueta, Valores, Perdidos, Columnas, Alineación y Medida y Rol (ver gráfica 24).

Gráfica 24. Definición de variables en SPSS

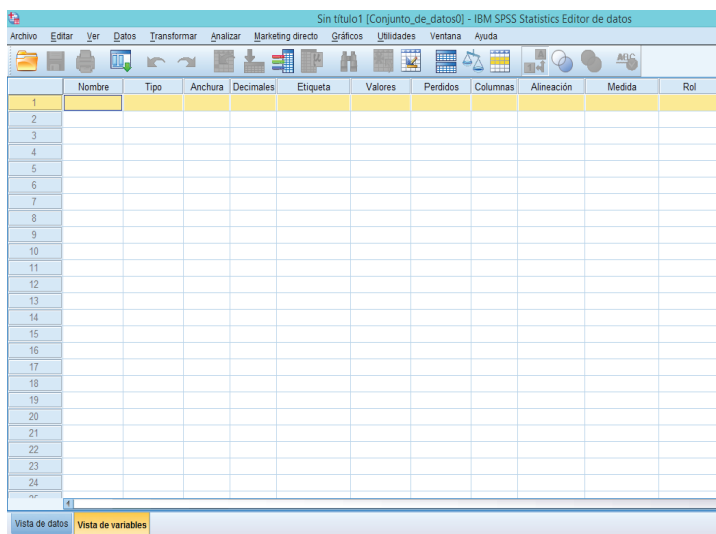


Imagen real del programa SPSS

La definición de cada uno de los rótulos de las variables se describe a continuación (Pérez, 2005):

Nombre: asignar el nombre a cada variable, cumpliendo las siguientes reglas:

- » Nombres con máximo 8 caracteres (siendo siempre el primero una letra).
- » No utilizar símbolos como &, /, \$, etc.
- » No utilizar nunca espacios en blanco.
- » No utilizar expresiones: ALL, AND, BY, EQ, GE, GT, LE, NE, NOT, OR, TO, o WITH.

Tipo: indicar el máximo número de dígitos para la variable y el tipo de la variable con la que corresponde: alfanumérica, fecha, moneda o numérica, indicando en este último caso el número de cifras decimales. Los tipos de variable pueden ser:

- » Numéricas: formato numérico estándar.
- » Coma: comas de separación cada tres posiciones.
- » Punto: un punto para indicar el decimal.
- » Notación científica: uso de la E para exponente. Ejemplo 2,34E2
- » Cadena: variable alfanumérica (de más de 8 caracteres se considera larga).
- » Además están los formatos de fecha, dólar y moneda personalizada.

Si no se asigna el tipo de variable, el sistema las asigna automáticamente, siendo el formato por defecto: numérica, anchura 8 y decimales 2.

Etiqueta: es la denominación de la variable con el nombre que aparecerá en los resultados. Cada variable debe tener entre 30 y 40 caracteres que permita tener más información sobre ésta.

Valores: se trata de asignar etiquetas a los valores de cada variable. No es obligatorio, pero es útil en algunos casos.

Perdidos: permite definir los valores de los datos especificados como “perdidos por el usuario”. Es útil para saber por qué se pierde información. Por ejemplo, se quiere distinguir el dato perdido correspondiente a un entrevistado que se niega a responder. Los valores de datos especificados como perdidos normalmente se catalogan con 9; 99 o 999, aparecen marcados para un tratamiento especial y se excluyen de la mayoría de los cálculos.

Columnas: consiste en especificar la amplitud de la columna.

Alineación: seleccionar la disposición de las entradas de la columna: izquierda, derecha y centrado.

Medida: se puede seleccionar uno de los tres niveles de medida:

- » Escala: los valores de datos son numéricos en una escala de intervalo. Las variables de escala deben ser numéricas.
- » Ordinal: los valores de datos representan categorías con un cierto orden intrínseco (bajo, medio, alto; totalmente de acuerdo, de acuerdo, en desacuerdo). Las variables ordinales pueden ser de cadena o valores numéricos. Notar que para variables de cadena ordinales, se asume que el orden alfabético de los valores de cadena indica el orden correcto de las categorías; en el caso de bajo, medio y alto el orden sería alto, bajo y medio (orden que no es correcto); por lo que es más fiable utilizar códigos numéricos para representar datos ordinales, que usar etiquetas de estos códigos.
- » Nominal: los valores de datos representan categorías sin un cierto orden intrínseco. Las variables nominales pueden ser de cadena o valores numéricos que representan categorías diferentes, por ejemplo 1 = Masculino y 2 = Femenino.

Rol: funciones o roles que se pueden asignar a las variables para el análisis. Estos pueden ser:

- » Entrada: la variable se utiliza como una entrada (por ejemplo: predictor, variable independiente).
- » Objetivo: la variable se utiliza como salida o de destino (por ejemplo: variable dependiente).
- » Ambos: la variable se utiliza como entrada y salida.
- » Ninguna: la variable no tiene una asignación de funciones.
- » Partición: la variable se utiliza para dividir los datos en muestras separadas.
- » Segmentar: las variables con este papel no se utilizan como variables generales del archivo que ha sido fraccionado.

Por defecto, a todas las variables se les asigna el rol de “entrada”.

A continuación se realiza la preparación de variables para el estudio realizado a un grupo de personas de la ciudad de Medellín que se han realizado tatuajes en el cuerpo. En la tabla 38 se presentan las variables analizadas con algunos resultados obtenidos.

Tabla 38. Resultados de las personas de la ciudad de Medellín que se han estampado tatuajes en el cuerpo.

Género	Edad	Labora	Cuántos tatuajes tiene	Parte del cuerpo tatuada	Por medio de qué elige el lugar donde tatuarse	Verifica las normas de sanidad del lugar	Consecuencias en la vida social	Se arrepiente de tatuarse	Costo del tatuaje
Masculino	19	No	2	Espalda	Amigo	No	No	No	\$ 50.000
Femenino	20	Sí	1	Pierna	Amigo	Sí	No	No	\$ 470.000
Masculino	26	Sí	1	Brazo	Amigo	No	No	No	\$ 70.000
Masculino	27	Sí	4	Hombros	Internet	Sí	No	No	\$ 400.000
Masculino	26	Sí	1	Brazo	Familia	No	Sí	Sí	\$ 50.000
Femenino	20	Sí	2	Espalda	Amigo	Sí	No	No	\$ 300.000
Masculino	17	No	1	Pierna	Amigo	Sí	Tal vez	No	\$ 190.000
Masculino	27	Sí	2	Brazo	Amigo	Sí	Sí	No	\$ 210.000

Elaboración propia

Al ubicar el nombre de la variable “Género”, automáticamente se definen los demás rótulos de las columnas, como aparece en la gráfica 25.

Gráfica 25. Ingreso del nombre de la variable en SPSS.

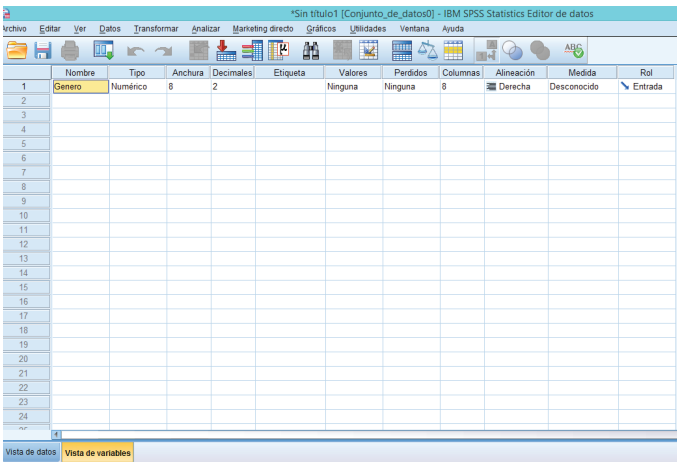


Imagen real del programa SPSS

Posteriormente, se deben ajustar los demás rótulos según las características de la variable. En el caso de la variable Género, al dar clic en la celda “Numérico” correspondiente al tipo de variables, aparece un cuadro de diálogo en el cual se debe seleccionar “Cadena”, debido a que la variable es cualitativa o categórica (ver gráfica 26).

Gráfica 26. Definición del tipo de variable en SPSS.

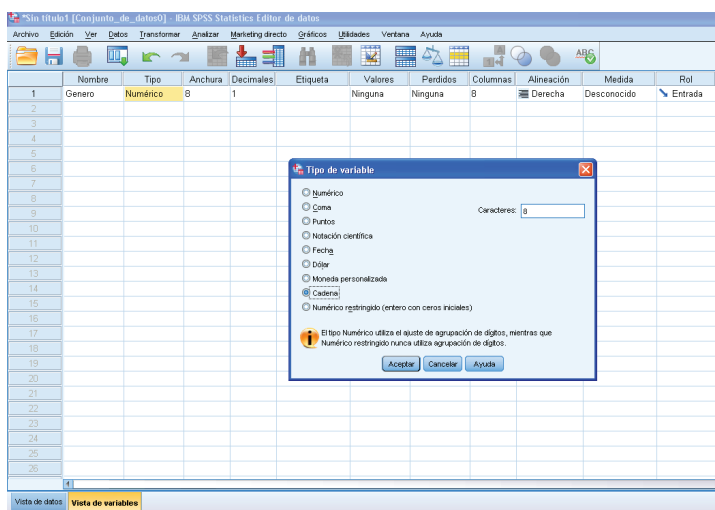


Imagen real del programa SPSS

Al dar clic en “Aceptar”, ubica el número de decimales en “0” y la medida en “Nominal”, por ser variable categórica (ver gráfica 27).

Gráfica 27. Definición de decimales y medida de la variable en SPSS.

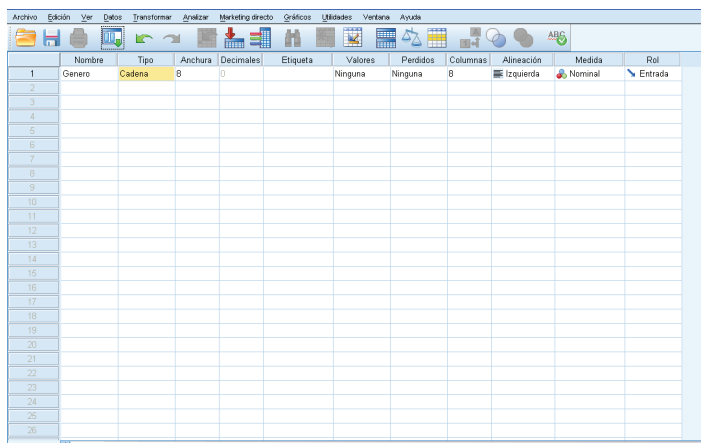


Imagen real del programa SPSS

La “Etiqueta” se define con el nombre de la variable, en este caso se asume la misma: “Género”.

Dando clic en “Ninguna”, que corresponde a “Valores”, se pueden etiquetar las opciones de género “Femenino” y “Masculino” con los valores “1” y “2”, respectivamente, y luego “Aceptar”, como aparece en la gráfica 28.

Gráfica 28. Definición de etiquetas de la variable en SPSS.

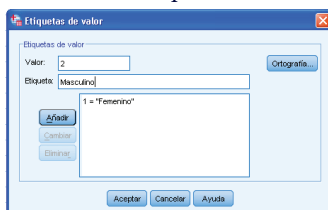


Imagen real del programa SPSS

En “Perdidos” se define el código con el cual se van a identificar los valores que tienen esta característica. En este caso, se asigna “99” (ver gráfica 29).

Gráfica 29. Definición de valores perdidos de la variable en SPSS.

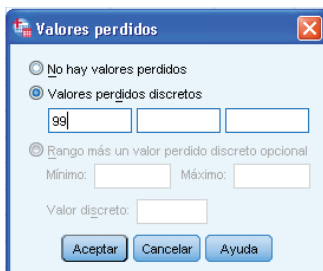


Imagen real del programa SPSS

La “Amplitud” es el ancho de la columna. Por defecto, el sistema asume un ancho de 8. La “Alineación” puede ubicarse a la izquierda, derecha o centrado. Finalmente, el sistema asume el “Rol” como “Entrada”.

Para ingresar una variable cuantitativa, en este caso la “Edad”, los diferentes rótulos deben ser ajustados a las características correspondientes a este tipo de variable. Igualmente, este procedimiento se realiza para cada una de las variables aplicadas en el estudio, tal como se ilustra en la gráfica 30.

Gráfica 30. Ajuste de todas las variables para el ingreso de datos en SPSS.

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	Género	Cadena	8	0	Género	(1, Femeni...	99	9	Centrado	Nominal	Entrada
2	Edad	Númerico	8	0	Edad (Años)	Ninguna	99	8	Centrado	Escala	Entrada
3	Labora	Cadena	8	0	Actividad laboral	(1, Si)...	99	8	Centrado	Nominal	Entrada
4	Tatuajes	Númerico	8	0	Nº de tatuajes ...	Ninguna	99	8	Centrado	Escala	Entrada
5	Cuerpo	Cadena	8	0	Parte del cuerpo...	Ninguna	99	8	Centrado	Nominal	Entrada
6	Elección	Cadena	11	0	Mediación para...	Ninguna	99	8	Centrado	Nominal	Entrada
7	Verificación	Cadena	8	0	Verificación de ...	(1, Si)...	99	8	Centrado	Nominal	Entrada
8	Implicaciones	Cadena	8	0	Ha tenido impli...	(1, Si)...	99	8	Centrado	Nominal	Entrada
9	Arepiente	Cadena	8	0	Se arrepiente d...	(1, Si)...	99	8	Centrado	Nominal	Entrada
10	Costo	Númerico	8	0	Costo del tatuaje	Ninguna	99	8	Centrado	Escala	Entrada
11											
12											
13											
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											

Imagen real del programa SPSS

12.3 Ingreso de datos

Para ingresar los datos al programa, en la pestaña “Vista de datos” ubicada en la parte inferior izquierda de la pantalla, se ingresa la información de cada variable, teniendo en cuenta los valores asignados para las etiquetas. Por ejemplo: “Género” se etiquetó 1: “Femenino” y 2: “Masculino”, y con estos códigos se debe ingresar. Igual sucede con “Labora” 1: “Sí” y 2: “No”, y así con las demás variables que fueron etiquetadas (ver tabla 39).

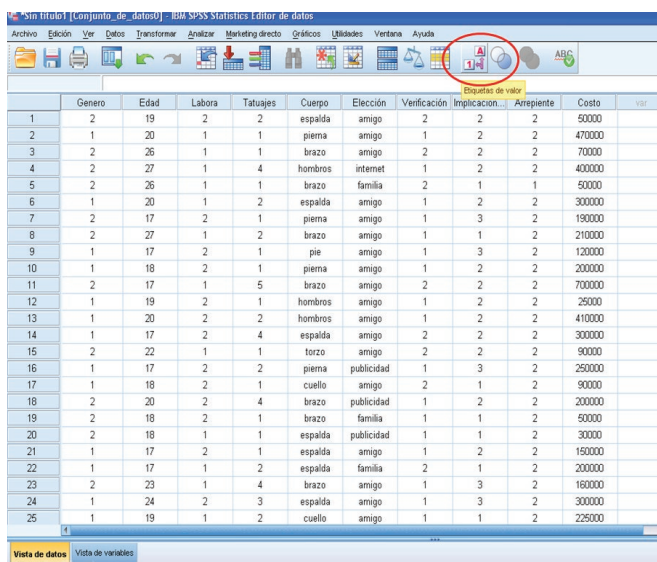
Tabla 39. Etiquetas para variables cualitativas

Género	Edad	Labora	Cuántos tatuajes tiene	Parte del cuerpo tatuada	Por medio de qué elige el lugar donde tatuarse	Verifica las normas de sanidad del lugar	Consecuencias en la vida social	Se arrepiente de tatuarse	Costo del tatuaje
2	19	2	2	Espalda	Amigo	3	2	2	\$ 50.000
1	20	1	1	Pierna	Amigo	1	2	2	\$ 470.000
2	26	1	1	Brazo	Amigo	2	2	2	\$ 70.000
2	27	1	4	Hombros	Internet	1	2	2	\$ 400.000
2	26	1	1	Brazo	Familia	2	1	1	\$ 50.000
1	20	1	2	Espalda	Amigo	1	2	2	\$ 300.000
2	17	2	1	Pierna	Amigo	1	3	2	\$ 190.000
1	27	2	2	Brazo	Amigo	1	1	2	\$ 210.000

Elaboración propia

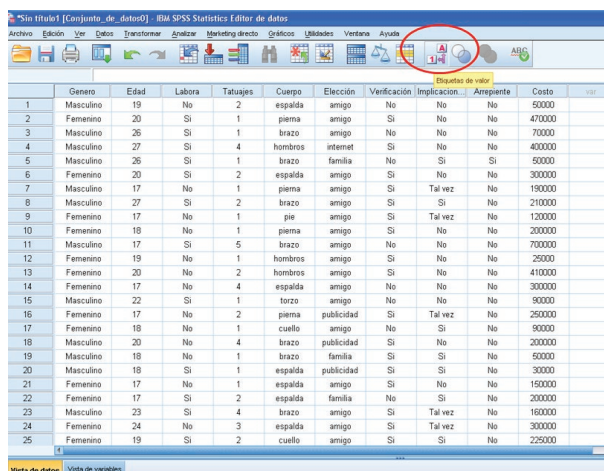
Al ingresar los datos en el sistema, se observa la apariencia de las etiquetas (ver gráfica 31). Sin embargo, al dar clic en “Etiquetas de valor”, se muestran los resultados originales que fueron etiquetados (ver gráfica 32).

Gráfica 31. Visualización de las variables con etiquetas en SPSS



	Genero	Edad	Labora	Tatuajes	Cuerpo	Elección	Verificación	Implicacion	Arrepiente	Costo	var
1	2	19	2	2	espalda	amigo	2	2	2	50000	
2	1	20	1	1	pierna	amigo	1	2	2	470000	
3	2	26	1	1	brazo	amigo	2	2	2	70000	
4	2	27	1	4	hombros	internet	1	2	2	400000	
5	2	26	1	1	brazo	familia	2	1	1	50000	
6	1	20	1	2	espalda	amigo	1	2	2	300000	
7	2	17	2	1	pierna	amigo	1	3	2	190000	
8	2	27	1	2	brazo	amigo	1	1	2	210000	
9	1	17	2	1	pie	amigo	1	3	2	120000	
10	1	18	2	1	pierna	amigo	1	2	2	200000	
11	2	17	1	5	brazo	amigo	2	2	2	700000	
12	1	19	2	1	hombros	amigo	1	2	2	25000	
13	1	20	2	2	hombros	amigo	1	2	2	410000	
14	1	17	2	4	espalda	amigo	2	2	2	300000	
15	2	22	1	1	torzo	amigo	2	2	2	90000	
16	1	17	2	2	pierna	publicidad	1	3	2	250000	
17	1	18	2	1	cuello	amigo	2	1	2	90000	
18	2	20	2	4	brazo	publicidad	1	2	2	200000	
19	2	18	2	1	brazo	familia	1	1	2	50000	
20	2	18	1	1	espalda	publicidad	1	1	2	30000	
21	1	17	2	1	espalda	amigo	1	2	2	150000	
22	1	17	1	2	espalda	familia	2	1	2	200000	
23	2	23	1	4	brazo	amigo	1	3	2	160000	
24	1	24	2	3	espalda	amigo	1	3	2	300000	
25	1	19	1	2	cuello	amigo	1	1	2	225000	

Gráfica 32. Visualización de las variables con valores o atributos en SPSS.



	Genero	Edad	Labora	Tatuajes	Cuerpo	Elección	Verificación	Implicación	Anepeñete	Costo	var
1	Masculino	19	No	2	espalda	amigo	No	No	No	50000	
2	Femenino	20	Si	1	pierna	amigo	Si	No	No	470000	
3	Masculino	26	Si	1	brazo	amigo	No	No	No	70000	
4	Masculino	27	Si	4	hombros	internet	Si	No	No	400000	
5	Masculino	26	Si	1	brazo	familia	No	Si	Si	50000	
6	Femenino	20	Si	2	espalda	amigo	Si	No	No	300000	
7	Masculino	17	No	1	pierna	amigo	Si	Tal vez	No	190000	
8	Masculino	27	Si	2	brazo	amigo	Si	Si	No	210000	
9	Femenino	17	No	1	pie	amigo	Si	Tal vez	No	120000	
10	Femenino	18	No	1	pierna	amigo	Si	No	No	200000	
11	Masculino	17	Si	5	brazo	amigo	No	No	No	700000	
12	Femenino	19	No	1	hombros	amigo	Si	No	No	25000	
13	Femenino	20	No	2	hombros	amigo	Si	No	No	410000	
14	Femenino	17	No	4	espalda	amigo	No	No	No	300000	
15	Masculino	22	Si	1	torzo	amigo	No	No	No	90000	
16	Femenino	17	No	2	pierna	publicidad	Si	Tal vez	No	250000	
17	Femenino	18	No	1	cuello	amigo	No	Si	No	90000	
18	Masculino	20	No	4	brazo	publicidad	Si	No	No	200000	
19	Masculino	18	No	1	brazo	familia	Si	Si	No	50000	
20	Masculino	18	Si	1	espalda	publicidad	Si	Si	No	30000	
21	Femenino	17	No	1	espalda	amigo	Si	No	No	150000	
22	Femenino	17	Si	2	espalda	familia	No	Si	No	200000	
23	Masculino	23	Si	4	brazo	amigo	Si	Tal vez	No	160000	
24	Femenino	24	No	3	espalda	amigo	Si	Tal vez	No	300000	
25	Femenino	19	Si	2	cuello	amigo	Si	Si	No	225000	

12.4 Resultados

SPSS presenta diferentes procedimientos estadísticos para el análisis descriptivo de la información, los cuales dependen del tipo de variable: Categórica o Escala, (Pérez, 2005). Para describir estos procedimientos se utilizará el ejemplo de la preparación de variables e ingreso de datos, sobre el estudio realizado a un grupo de personas de la ciudad de Medellín que se han efectuado tatuajes en el cuerpo.

Los procedimientos estadísticos que ofrece SPSS se encuentran ubicados en el menú “Analizar”. Al hacer clic en él se despliega una lista, en la cual se observan los diferentes procedimientos del programa, como Informes, Estadísticos descriptivos, Tablas, Comparar medias, Modelo lineal general, Modelos mixtos, Correlación, Regresión, entre otros. En la gráfica 33 se observa que cada una de estas opciones presenta una flecha al lado derecho, que conducen a submenús que contienen diferentes opciones, entre ellas Frecuencias, Descriptivos, Explorar, etc.

Gráfica 33. Ubicación del menú “Análisis” para los “Estadísticos descriptivos” en SPSS

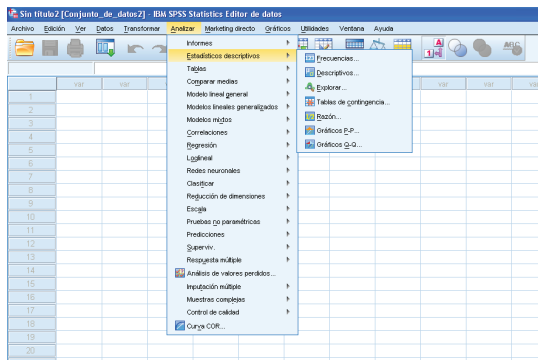


Imagen real del programa SPSS

Para el alcance del contenido desarrollado en el texto, se usará la opción “Frecuencias” del submenú “Estadísticos descriptivos”. Al dar clic en esta opción, se despliega un cuadro de diálogo donde aparecen todas las variables (ver gráfica 34).

Gráfica 34. Visualización de variables para el análisis descriptivo en SPSS

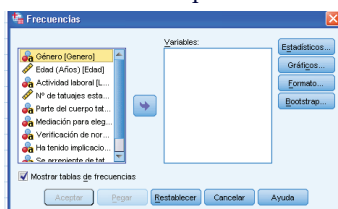


Imagen real del programa SPSS

Las variables deben ser trasladadas al costado derecho, por medio de la flecha ubicada entre los dos espacios; este procedimiento se puede realizar pasando cada una de las variables o seleccionándolas en su totalidad (ver gráfica 35).

Gráfica 35. Selección de las variables con etiquetas en SPSS

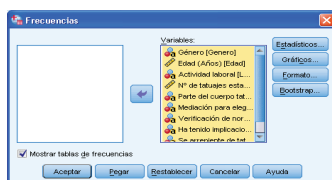


Imagen real del programa SPSS

En la parte superior derecha del cuadro de diálogo aparece la opción “Estadísticos”. Al dar clic en éste, despliega las medidas descriptivas requeridas para el análisis estadístico. Se recomienda seleccionar: cuartiles, media, mediana, moda, desviación típica, varianza, rango, mínimo, máximo, asimetría y curtosis. Posteriormente, se da clic en “Continuar” (ver gráfica 36).

Gráfica 36. Selección de Estadísticos para el análisis de variables en SPSS

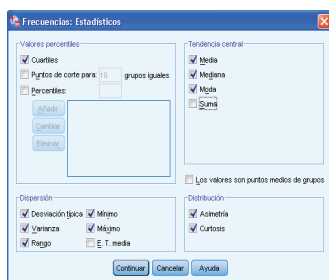


Imagen real del programa SPSS

Al dar clic en “Continuar”, regresa de nuevo al cuadro de diálogo de “Frecuencias”. Allí se selecciona la opción “Gráficos” y el tipo de gráfico deseado; en este caso se tomarán “Gráficos de barras”, y en “Valores del gráfico” se indicará “Porcentaje”, posteriormente “Continuar” (ver gráfica 37).

Gráfica 37. Selección de gráficos para el análisis de variables en SPSS

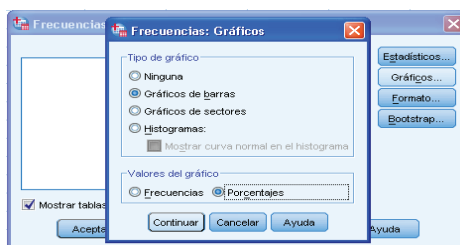


Imagen real del programa SPSS

En la opción “Formato” se debe seleccionar Ordenar por “Valores ascendentes”, y en Variables múltiples “Organizar resultados según variables”. Esto para que en los resultados aparezcan consecutivamente las tablas y las gráficas por cada variable, luego se da clic en “Continuar” (ver gráfica 38).

Gráfica 38. Configuración para la obtención de resultados por variable en SPSS

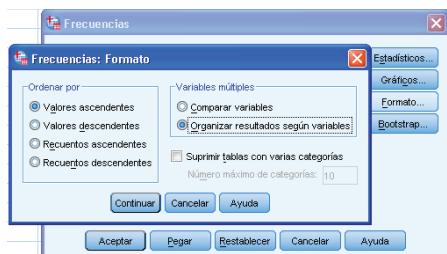


Imagen real del programa SPSS

Se llega de nuevo al cuadro de diálogo de “Frecuencias”, donde se da clic en “Aceptar” para obtener los resultados (ver gráfica 39).

Gráfica 39. Obtención de resultados para las variables en SPSS

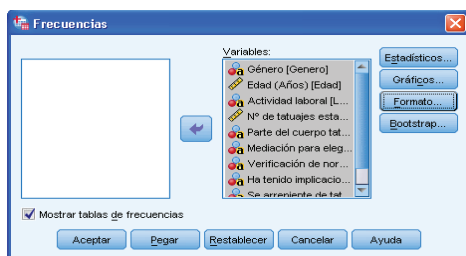


Imagen real del programa SPSS

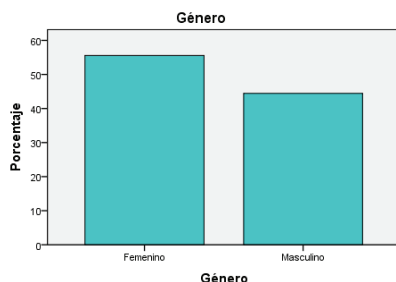
Finalmente, aparecen los resultados solicitados para el respectivo análisis estadístico, los cuales muestran para cada variable el resumen de datos válidos, las frecuencias absolutas, relativas (porcentaje) y relativas acumuladas, con la respectiva gráfica, esta última puede ser de barras, diagrama circular o histograma, según la selección que se realizó en los pasos anteriores. Además, para las variables cuantitativas, arroja las medidas de tendencia central (media, mediana y moda), de dispersión (varianza, desviación, rango), de forma (asimetría y curtosis), entre otros.

A continuación se presentan los resultados obtenidos para el ejemplo del estudio realizado a un grupo de personas de la ciudad de Medellín que se han efectuado tatuajes en el cuerpo.

Género

Estadísticos		
Género		
N	Válidos	81
	Perdidos	0

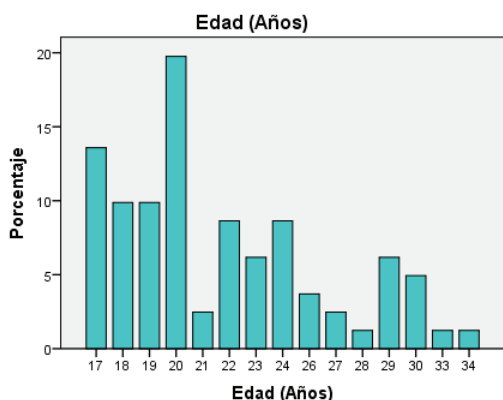
Género					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Femenino	45	55,6	55,6	55,6
	Masculino	36	44,4	44,4	100,0
	Total	81	100,0	100,0	



Edad (Años)

Estadísticos		
Edad (Años)		
N	Válidos	81
	Perdidos	0
Media		21,90
Mediana		20,00
Moda		20
Desv. típ.		4,291
Varianza		18,415
Asimetría		,932
Error típ. de asimetría		,267
Curtosis		,031
Error típ. de curtosis		,529
Rango		17
Mínimo		17
Máximo		34
Percentiles	25	19,00
	50	20,00
	75	24,00

Edad (Años)				
	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	17	11	13,6	13,6
	18	8	9,9	23,5
	19	8	9,9	33,3
	20	16	19,8	53,1
	21	2	2,5	55,6
	22	7	8,6	64,2
	23	5	6,2	70,4
	24	7	8,6	79,0
	26	3	3,7	82,7
	27	2	2,5	85,2
	28	1	1,2	86,4
	29	5	6,2	92,6
	30	4	4,9	97,5
	33	1	1,2	98,8
	34	1	1,2	100,0
Total	81	100,0	100,0	



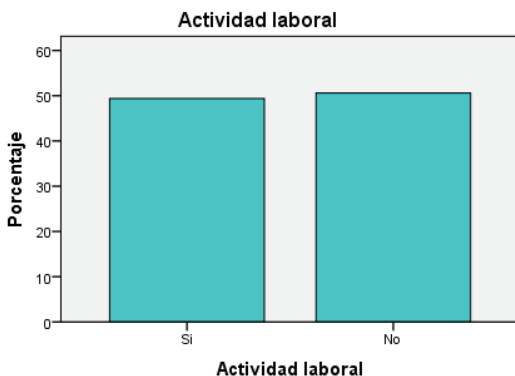
Actividad laboral

Estadísticos

Actividad laboral

N	Válidos	81
	Perdidos	0

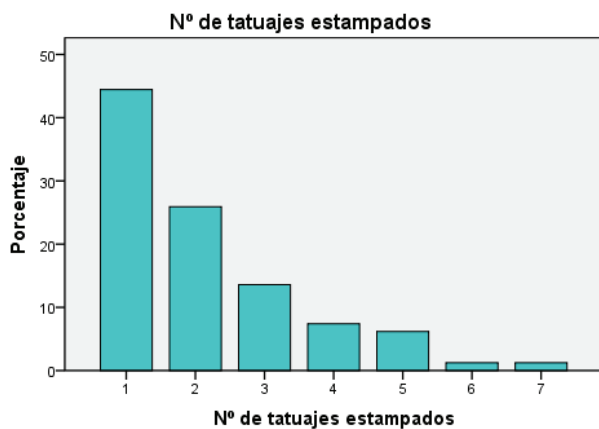
Actividad laboral				
	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Sí	40	49,4	49,4
	No	41	50,6	100,0
	Total	81	100,0	



Nº. de tatuajes estampados

Estadísticos		
Nº. de tatuajes estampados		
N	Válidos	81
	Perdidos	0
Media		2,14
Mediana		2,00
Moda		1
Desv. típ.		1,394
Varianza		1,944
Asimetría		1,340
Error típ. de asimetría		,267
Curtosis		1,369
Error típ. de curtosis		,529
Rango		6
Mínimo		1
Máximo		7
Percentiles	25	1,00
	50	2,00
	75	3,00

Nº de tatuajes estampados					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	1	36	44,4	44,4	44,4
	2	21	25,9	25,9	70,4
	3	11	13,6	13,6	84,0
	4	6	7,4	7,4	91,4
	5	5	6,2	6,2	97,5
	6	1	1,2	1,2	98,8
	7	1	1,2	1,2	100,0
	Total	81	100,0	100,0	



Parte del cuerpo tatuada

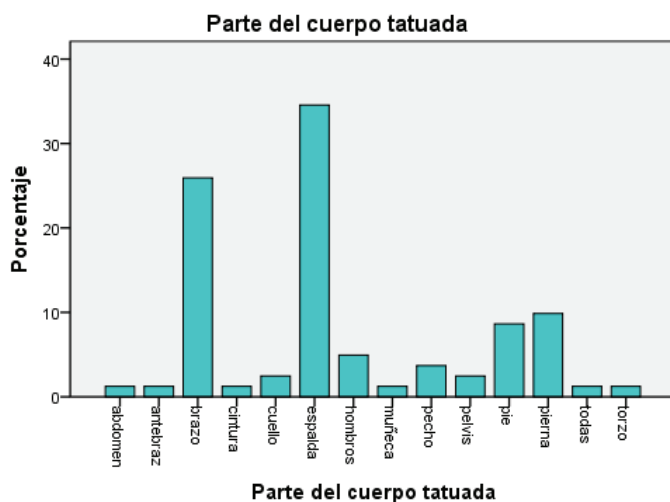
Estadísticos

Parte del cuerpo tatuada

N Válidos 81
 Perdidos 0

Parte del cuerpo tatuada

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Abdomen	1	1,2	1,2	1,2
	Antebrazo	1	1,2	1,2	2,5
	Brazo	21	25,9	25,9	28,4
	Cintura	1	1,2	1,2	29,6
	Cuello	2	2,5	2,5	32,1
	Espalda	28	34,6	34,6	66,7
	Hombros	4	4,9	4,9	71,6
	Muñeca	1	1,2	1,2	72,8
	Pecho	3	3,7	3,7	76,5
	Pelvis	2	2,5	2,5	79,0
	Pie	7	8,6	8,6	87,7
	Pierna	8	9,9	9,9	97,5
	Todas	1	1,2	1,2	98,8
	Torzo	1	1,2	1,2	100,0
	Total	81	100,0	100,0	



Mediación para elegir el sitio

Estadísticos

Mediación para elegir el sitio

N	Válidos	81
	Perdidos	0

Mediación para elegir el sitio

	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Amigo	60	74,1	74,1	74,1
Familia	7	8,6	8,6	82,7
Internet	1	1,2	1,2	84,0
Publicidad	12	14,8	14,8	98,8
Volante	1	1,2	1,2	100,0
Total	81	100,0	100,0	



Verificación de normas de sanidad del sitio

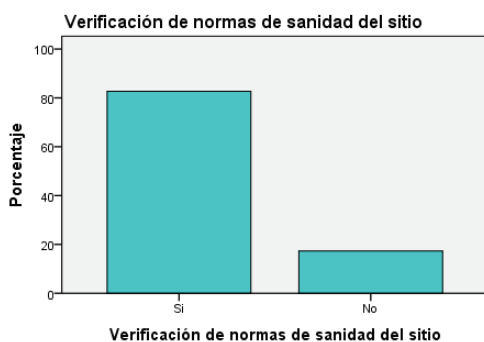
Estadísticos

Verificación de normas
de sanidad del sitio

N	Válidos	81
	Perdidos	0

Verificación de normas de sanidad del sitio

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Sí	67	82,7	82,7	82,7
	No	14	17,3	17,3	100,0
	Total	81	100,0	100,0	



Ha tenido implicaciones en la vida social

Estadísticos

Ha tenido implicaciones
en la vida social

N	Válidos	81
	Perdidos	0

Ha tenido implicaciones en la vida social

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Sí	28	34,6	34,6	34,6
	No	43	53,1	53,1	87,7
	Tal vez	10	12,3	12,3	100,0
	Total	81	100,0	100,0	



Se arrepiente de tatuarse

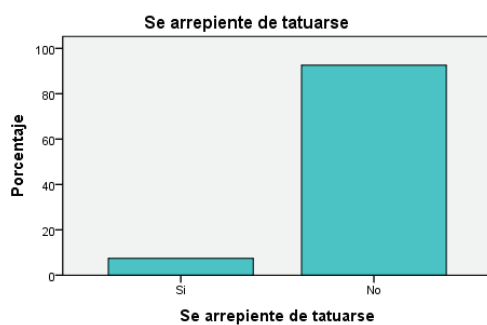
Estadísticos

Se arrepiente de tatuarse

N	Válidos	81
	Perdidos	0

Se arrepiente de tatuarse

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Sí	6	7,4	7,4	7,4
	No	75	92,6	92,6	100,0
	Total	81	100,0	100,0	

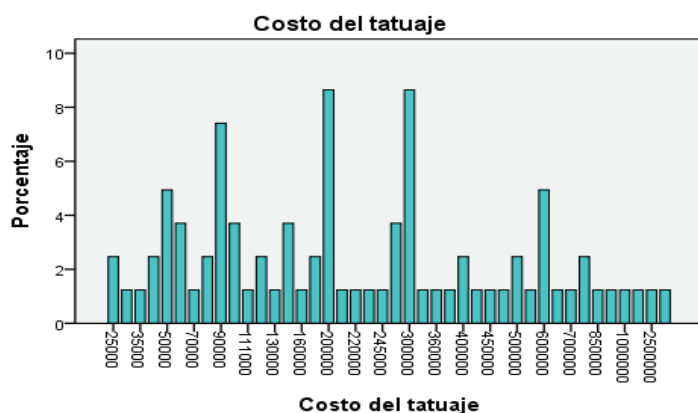


Costo del tatuaje

Estadísticos		
Costo del tatuaje		
N	Válidos	81
	Perdidos	0
Media		365259,26
Mediana		200000,00
Moda		200000 ^a
Desv. típ.		563041,801
Varianza		317016069444,444
Asimetría		4,461
Error típ. de asimetría		0,267
Curtosis		24,117
Error típ. de curtosis		0,529
Rango		3975000
Mínimo		25000
Máximo		4000000
Percentiles	25	90000,00
	50	200000,00
	75	405000,00

a. Existen varias modas. Se mostrará el menor de los valores

Costo del tatuaje				
	Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
25000	2	2,5	2,5	2,5
30000	1	1,2	1,2	3,7
35000	1	1,2	1,2	4,9
40000	2	2,5	2,5	7,4
50000	4	4,9	4,9	12,3
60000	3	3,7	3,7	16,0
70000	1	1,2	1,2	17,3
80000	2	2,5	2,5	19,8
90000	6	7,4	7,4	27,2
100000	3	3,7	3,7	30,9
111000	1	1,2	1,2	32,1
120000	2	2,5	2,5	34,6
130000	1	1,2	1,2	35,8
150000	3	3,7	3,7	39,5
160000	1	1,2	1,2	40,7
190000	2	2,5	2,5	43,2
200000	7	8,6	8,6	51,9
Válidos 210000	1	1,2	1,2	53,1
220000	1	1,2	1,2	54,3
225000	1	1,2	1,2	55,6
245000	1	1,2	1,2	56,8
250000	3	3,7	3,7	60,5
300000	7	8,6	8,6	69,1
320000	1	1,2	1,2	70,4
360000	1	1,2	1,2	71,6
370000	1	1,2	1,2	72,8
400000	2	2,5	2,5	75,3
410000	1	1,2	1,2	76,5
450000	1	1,2	1,2	77,8
470000	1	1,2	1,2	79,0
500000	2	2,5	2,5	81,5
520000	1	1,2	1,2	82,7
600000	4	4,9	4,9	87,7
620000	1	1,2	1,2	88,9
700000	1	1,2	1,2	90,1
800000	2	2,5	2,5	92,6
850000	1	1,2	1,2	93,8
950000	1	1,2	1,2	95,1
1000000	1	1,2	1,2	96,3
2000000	1	1,2	1,2	97,5
2500000	1	1,2	1,2	98,8
4000000	1	1,2	1,2	100,0
Total	81	100,0	100,0	



12.5 Ejercicio de aplicación

La siguiente información corresponde al análisis realizado a estudiantes del grado cuarto de la Escuela Tulio Botero Salazar de la ciudad de Medellín (Posada, 2014). Se analizaron las siguientes variables: género, materia favorita, edad, número de hermanos, peso y estatura. En la tabla 40 se presenta la información consolidada.

Tabla 40. Variables para el ejercicio de aplicación 12.5

Estudiante	Género	Materia favorita	Edad (años)	N° de hermanos	Peso (kg)	Estatura (m)
1	Masculino	Geometría	9	0	30	1,40
2	Masculino	Matemáticas	9	1	33	1,35
3	Femenino	Artística	10	2	37	1,47
4	Masculino	Educación física	9	1	32	1,32
5	Femenino	Educación física	9	2	34	1,32
6	Femenino	Educación física	9	1	34	1,34
7	Masculino	Matemáticas	9	1	30	1,35
8	Femenino	Geometría	9	3	30	1,37
9	Masculino	Educación física	10	1	37	1,37
10	Masculino	Artística	9	1	35	1,37
11	Femenino	Matemáticas	9	3	32	1,35
12	Femenino	Matemáticas	8	1	32	1,37
13	Masculino	Artística	10	1	33	1,30
14	Femenino	Matemáticas	9	2	35	1,38

Continúa en la siguiente página

Continúa en la anterior página

Estudiante	Género	Materia favorita	Edad (años)	N° de hermanos	Peso (kg)	Estatura (m)
15	Masculino	Artística	12	4	37	1,38
16	Masculino	Geometría	11	2	38	1,30
17	Masculino	Artística	9	2	34	1,41
18	Femenino	Matemáticas	9	1	30	1,32
19	Femenino	Educación física	9	4	30	1,40
20	Masculino	Artística	9	4	26	1,30
21	Masculino	Artística	9	1	33	1,34
22	Femenino	Educación física	10	1	30	1,35
23	Masculino	Matemáticas	9	1	25	1,33
24	Masculino	Matemáticas	9	3	35	1,37
25	Femenino	Educación física	9	0	28	1,29
26	Femenino	Educación física	9	1	33	1,31
27	Masculino	Artística	9	1	34	1,33
28	Femenino	Artística	10	1	45	1,50
29	Masculino	Artística	9	2	25	1,25
30	Femenino	Artística	9	2	31	1,30
31	Masculino	Artística	9	1	32	1,30
32	Femenino	Educación física	9	2	36	1,40
33	Masculino	Educación física	9	3	33	1,35
34	Masculino	Artística	9	0	30	1,35
35	Femenino	Educación física	9	3	30	1,31
36	Masculino	Educación física	9	1	27	1,26
37	Femenino	Educación física	9	1	30	1,40
38	Masculino	Educación física	9	1	35	1,38
39	Masculino	Español	9	1	32	1,35
40	Femenino	Español	10	1	30	1,32
41	Femenino	Español	10	0	34	1,36
42	Femenino	Educación física	9	1	31	1,30
43	Masculino	Matemáticas	9	1	35	1,33
44	Femenino	Artística	8	1	30	1,32

Elaboración propia a partir de encuesta aplicada a estudiantes del grado cuarto de la Escuela Tulio Botero Salazar de la ciudad de Medellín

1. Usando SPSS, construir las frecuencias para las variables género, materia favorita, edad, número de hermanos, peso y estatura.
2. Elaborar las gráficas correspondientes para cada una de las variables mencionadas anteriormente.
3. Calcular media aritmética, mediana, moda, varianza, desviación, sesgo, curtosis y cuartiles para cada una de las variables cuantitativas. Interpretar los resultados obtenidos.

REFERENCIAS

- Alzate, J. C. (2004). *Estadística Descriptiva*. Medellín: Iniremington.
- Berenson, M., Levine, D., y Krehbiel, T. (2006). *Estadística para administración*. México: Pearson.
- Bonnet, J. (2003). *Lecciones de estadística*. España: Editorial Club Universitario.
- Departamento Administrativo Nacional de Estadística (DANE, 2008). Estimación interpretación del coeficiente de variación de la encuesta Cencensal. Recuperado de: https://www.dane.gov.co/files/investigaciones/boletines/censo/est_interp_coefvariacion.pdf
- Hernández, S. (2005). *Historia de la estadística*. México: La ciencia y el hombre. Recuperado de: <http://148.226.12.104/bitstream/123456789/5640/2/20052P5.pdf>
- Lind, D., Mason, R. D., y Marchal, W. G. (2000). *Estadística para administración y economía*. España: McGraw-Hill.
- Martínez, C. (2000). *Estadística y muestreo*. Bogotá: Ecoe Ediciones.
- Martínez, C. (2007). *Estadística básica aplicada*. Bogotá: Ecoe Ediciones.
- Mendenhall, W. (1995). *Estadística para administradores*. México: Iberoamerica.
- Pierdant, A., y Rodríguez, J. (2006). *Elementos básicos de estadística para Ciencias Sociales*. México DF: Universidad Autónoma Metropolitana.

- Pérez, C. (2005). *Técnicas estadísticas con SPSS 12. Aplicaciones al análisis de datos*. Madrid: Pearson Educación.
- Posada, G.J. (2014). *Proyecto de aula: aplicación de conceptos estadísticos*. Estudiantes del curso de Estadística, semestre 01 de 2014. Fundación Universitaria Luis Amigó, Medellín.
- Posada, G. J., y Buitrago, M. V. (2006). *Módulo de Estadística*. Colombia: Fundación Universitaria Luis Amigó.
- Spiegel, M. (2013). *Estadística*. España: McGraw-Hill.
- Suárez, M., y Tapia, F. (2012). *Interaprendizaje de estadística básica*. Ecuador: Universidad Técnica del Norte.
- Ruíz Muñoz, D. (2005). *Manual de Estadística*. Universidad Pablo de Olavide. Recuperado de: <http://www.eumed.net/cursecon/libreria/drm/index.htm>
- Triola, M. F. (2000). *Estadística elemental*. México: Pearson Educación.
- Walpole, R. E., y Myers, R. H. (2012). *Probabilidad y estadística*. México: McGraw-Hill.

INFORMACIÓN DEL AUTOR

GABRIEL JAIME POSADA HERNÁNDEZ

Es Ingeniero Forestal, Especialista en Gestión Pública de la Escuela Superior de Administración Pública (ESAP) y Magister en Estudios Urbano-Regionales de la Universidad Nacional de Colombia.

Entre sus publicaciones se encuentran: Contribuciones de la economía solidaria al desarrollo local: el caso del Altiplano norte del departamento de Antioquia, publicado en la Revista Semestre Económico; Línea de investigación en matemáticas aplicadas, 2008; Ecuaciones Lineal y Cuadrática, 2012; Análisis comparativo de la competencia genérica Razonamiento Cuantitativo de la prueba Saber Pro, 2013; Ámbitos socioeconómicos de los municipios colombianos: una mirada desde el ordenamiento territorial, 2014; Estado de la competencia genérica Razonamiento Cuantitativo en los aspirantes a la educación superior, 2014; Relación entre las competencias genéricas Razonamiento Cuantitativo y Lectura Crítica en las pruebas Saber Pro para el departamento de Antioquia, 2015.

Obtuvo Menciones Especiales por producción intelectual de los módulos de Cálculo Diferencial e Integral (Funlam, 2004) y por contribuir al pensamiento en cooperación Francisco Luis Jiménez Arcila (Asociación Antioqueña de Cooperativas y Confederación de Cooperativas de Antioquia, 2005).

Actualmente se desempeña como docente de tiempo completo del Departamento de Ciencias Básicas en la Fundación Universitaria Luis Amigo y es integrante de los grupos de investigación de Economía Solidaria (ECOSOL) y Sistemas de Información y Sociedad del

Conocimiento (SISCO), donde ha participado en diversos proyectos, entre ellos: Descripción y análisis de variables financieras y socioeconómicas del sector de la economía solidaria en Antioquia (2001-2002), Variables financieras y socioeconómicas del sector de la economía solidaria en Antioquia (2002-2003), Aporte de la economía solidaria al desarrollo local del altiplano del norte de Antioquia (2005), Responsabilidad social en cooperativas de reciclaje del oriente cercano del departamento de Antioquia y el Área Metropolitana del Valle de Aburrá (2007), Análisis comparativo de la competencia genérica razonamiento cuantitativo de la prueba Saber Pro y su relación con los cursos ofrecidos por el Departamento de Ciencias Básicas en los programas de pregrado de la Funlam y la UdeM (2012).



FUNDACIÓN
UNIVERSITARIA



Medellín
2016