Análisis exploratorio de datos

Una introducción a la estadística descriptiva y probabilidad

Leandro González Támara

EDITORIAL UTADEO

Análisis exploratorio de datos

Una introducción a la estadística descriptiva y probabilidad

UTADEO
UNIVERSIDAD DE BOGOTÁ JORGE TADEO LOZANO
FACULTAD DE CIENCIAS NATURALES E INGENIERÍA
DEPARTAMENTO DE CIENCIAS BÁSICAS Y MODELADO

González Támara, Leandro
Análisis exploratorio de datos. Una introducción a la estadística descriptiva y probabilidad / Leandro González Támara. – Bogotá : Universidad de Bogotá Jorge Tadeo Lozano. Facultad de Ciencias Naturales e Ingeniería. Departamento de Ciencias Básicas y Modelado, 2017.
342 p. : il., grfcas. ; 22 cm.

ISBN 978-958-725-226-2

1. ESTADÍSTICA 2. PROBABILIDADES. I. tit.

CDD519.5

Análisis exploratorio de datos. Una introducción a la estadística descriptiva y probabilidad
ISBN impreso: 978-958-725-226-2
ISBN digital: 978-958-725-227-9

Rectora: Cecilia María Vélez White
Vicerrectora Académica: Margarita María Peña Borrero
Vicerrectora Administrativa: Nohemy Arias Otero
Decano de la Facultad de Ciencias Naturales e Ingeniería:
Isaac Dyner Rezonzew
Director Departamento de Ciencias Básicas y Modelado: Favio Cala Vitery

Editorial Utadeo
Jefe de Publicaciones: Daniel Mauricio Blanco Betancourt
Coordinación gráfica y diseño: Luis Carlos Celis Calderón
Coordinación editorial: Mary Lidia Molina Bernal
Coordinación revistas científicas: Juan Carlos García Sáenz
Distribución y ventas: Sandra Guzmán
Asistente administrativa: Blanca Esperanza Torres

Edición:
Diseño de carátula y pauta gráfica: Juanita Giraldo
Adecuación pauta gráfica: Luis Carlos Celis Calderón
Corrección de estilo: Juan David Ardila Suárez
Coordinación editorial: Mary Lidia Molina Bernal
Diagramación: Francisco Jiménez
Impresión: Disonex S.A.

Análisis exploratorio de datos

Una introducción a la estadística descriptiva y probabilidad

Leandro González Támara

Contenido

Presentación 11

1 Estadística Descriptiva **15**

¿Qué es la estadística? 15
Población, muestras y variables 17
Gráficos para resumir datos cualitativos 20
Diagrama de barras y gráfico de sectores 20
Rascacielos 23
Diagrama de barras y mosaicos 24
Gráficos para resumir datos cuantitativos 41
Diagrama de tallo y hojas 41
Diagrama de dispersión y gráfico de puntos 43
Diagrama de líneas 44
Histogramas 46
Pirámides de población 52

2 Estadísticos de centro y variabilidad **65**

Estadísticos de centro 65
La media o promedio $\bar{x}$ 65
Estadísticas de orden 68

La mediana *m* 70
La moda 72
Estadísticos de variabilidad 91
El rango 92
La varianza 92
La desviación estándar muestral 93
El coeficiente de variación CV 97
Conclusiones acerca de la distribución de una variable 113
La desigualdad de Chevyshev 113
Regla para las distribuciones acampanadas o regla empírica 116
Otras medidas de localización 124
Percentiles muestrales 124
Cuartiles 127
Gráfico de caja 129

3 Correlación y regresión 143

Regresión lineal simple y correlación 144
La covarianza 148
Coeficiente de correlación 149
El modelo de la regresión lineal simple 152
Modelos alternativos a la regresión lineal simple 157

4 Introducción a la probabilidad 179

Experimentos aleatorios 179
Espacio muestral 182
Diagrama de árbol 184

Técnicas de conteo 188
Regla de la multiplicación 188
Permutaciones 192
Combinaciones 196
Conceptos básicos de probabilidad 202
Probabilidad 205
Cálculo de probabilidades 216
Eventos mutuamente excluyentes 220
Complemento de un evento 221
Probabilidad condicional y teorema de Bayes 227
Independencia 229
Teorema de Bayes 232
Variables aleatorias 242
Variables aleatorias discretas 243
Variables aleatorias continuas 251

5 Distribuciones discretas de probabilidad. Algunos casos prácticos 263

Distribución de probabilidad binomial 264
Situación inicial: aprobar un examen sin estudiar 264
Características de una distribución de probabilidad binomial 266
Distribución de probabilidad de Poisson 275
Distribución de probabilidad hipergeométrica 281

6 Distribución de probabilidad normal 289

Características de la distribución de probabilidad normal 289
Distribución de probabilidad normal estándar *Z*. 293

Esquema para un proyecto de estadística descriptiva **303**

Pruebas de conocimientos **306**

Modelo Identificación, Argumentación, Formulación, Resolución e Interpretación para el análisis e interpretación de un problema de distribuciones de probabilidad **315**

Referencias **321**

Anexo 1. Código de Buenas Prácticas de las Estadísticas Europeas" (Eurostat, 2011) **323**

Principio 1. Independencia profesional 324
Indicadores 324
Principio 2. Mandato de recogida de datos 324
Indicadores 325
Principio 3. Adecuación de los recursos 325
Indicadores 325
Principio 4. Compromiso de calidad 325
Indicadores 326
Principio 5. Confidencialidad estadística 326
Indicadores 326
Principio 6. Imparcialidad y objetividad 327
Indicadores 327
Principio 7. Metodología sólida 327
Indicadores 327

Principio 8. Procedimientos estadísticos adecuados 328
Indicadores 328
Principio 9. Una carga para los encuestados que no sea excesiva 329
Indicadores 329
Principio 10. Relación costo-eficacia 329
Indicadores 329
Principio 11. Pertinencia 330
Indicadores 330
Principio 12. Precisión y fiabilidad 330
Indicadores 330
Principio 13. Oportunidad y puntualidad 331
Indicadores 331
Principio 14. Coherencia y comparabilidad 331
Indicadores 331
Principio 15. Accesibilidad y claridad 332
Indicadores 332

Anexo 2. *Función de distribución binomial* **333**

Anexo 3. *Función de distribución de Poisson* **337**

Anexo 4. *Función de distribución normal estándar* **340**

Presentación

El libro de *Análisis exploratorio de datos* es la evolución del libro de *Estadística descriptiva y probabilidad* (González, 2013), que ha sido la referencia principal de un curso dirigido a estudiantes de ciencias económico-administrativas durante cuatro años en la Universidad de Bogotá Jorge Tadeo Lozano. Esta propuesta pone mayor acento en la exploración de datos y su interpretación. En lugar de hacer un compendio exhaustivo de la estadística descriptiva y la probabilidad, en el libro se han seleccionado temas generales de estadística descriptiva y probabilidad que les permitan a los futuros profesionales producir y comprender un reporte estadístico por medio de la exploración de datos y su interpretación.

Tal como su predecesor, este es el libro guía de un curso universitario introductorio sobre estadística descriptiva y probabilidad. Está escrito con un lenguaje sencillo y claro para favorecer el encuentro inicial con los conceptos básicos y también propone ideas de tipo didáctico a diferentes tipos de profesionales que se dedican a la enseñanza de esta disciplina. Es un libro diseñado para ser seguido en una clase presencial que motive la discusión de diferentes temáticas; puede ser considerado como texto de transición porque permite acercar a los estudiantes a los conceptos por medio de aplicaciones, mucho antes de concentrarse en aspectos teóricos. Contiene más de 100 ejemplos solucionados en detalle para guiar a los estudiantes en su aprendizaje y casi 300 ejercicios prácticos para motivar la discusión de variadas temáticas de estadística y probabilidad.

El libro contiene seis capítulos. El primero, sobre estadística descriptiva, introduce el lenguaje de la estadística y muestra diferentes formas gráficas para resumir datos. El segundo capítulo gira en torno a los estadísticos de centro y dispersión con énfasis en su interpretación. El tercer capítulo hace una introducción al concepto de *regresión* y muestra al estudiante cómo puede usar esta técnica. El capítulo cuatro aborda el concepto de *probabilidad* como fundamento para el estudio de las variables aleatorias. Las distribuciones de probabilidad binomial, Poisson e hipergeométrica se tratan en el capítulo cinco. Por último, en el capítulo seis, se estudia la distribución normal de probabilidad.

Cada capítulo consta de un ejemplo o situación inicial que motiva el estudio de algunos de los contenidos. En lugar de un desarrollo teórico detallado de los temas, se presentan explicaciones breves para que los detalles sean tema de discusión en las clases. Hay más de un centenar de ejemplos con explicaciones minuciosas y estrategias para resolver problemas de análisis de datos y de probabilidad, los cuales en su

mayoría abordan situaciones de contextos significativos para los estudiantes con el propósito de despertar su interés. Estos ejemplos también ayudan a quienes de manera individual se acercan al análisis de datos, pues sirven de modelo para tratar situaciones en las cuales la exploración y la interpretación de datos son necesarias. Cada concepto tratado se cierra con una sección de ejercicios para afianzar habilidades y para generar discusión en las clases. Los ejercicios desarrollan en los estudiantes diversas habilidades de análisis estadístico y también pueden ser utilizados por los profesores de un primer curso de estadística y probabilidad para evaluar el aprendizaje. Se espera que los estudiantes puedan realizar la mayoría de los cálculos aritméticos y gráficas con un programa de computador o una aplicación estadística para teléfono celular, para que tanto profesores como estudiantes superen con facilidad los procesos de cálculo y puedan enfatizar en los significados y las interpretaciones.

Al final del libro se incluye una sección con seis pruebas de conocimiento que son útiles para evaluar lo aprendido.

Los datos presentados en el libro son en su mayoría reales y fueron obtenidos en páginas web de entidades reconocidas. Entre ellas se anotan:

- Index Mundi. Disponible en http://www.indexmundi.com
- Instituto Nacional de Estadísticas (España). Disponible en www.ine.es
- Departamento Nacional de Estadísticas (Colombia). Disponible en www.dane.gov.co
- Oficina del Censo de Estados Unidos. Disponible en http://www.census.gov/
- Agencias de Estadísticas Internacionales. Disponible en http://www.census.gov/aboutus/stat_int.html
- Estadísticas internacionales del U. S. Census Bureau. Disponible en http://www.census.gov/population/international/data/idb/country.php
- Bases de datos internacionales de U. S. Census Bureau. Disponible en http://www.census.gov/population/international/data/idb/informationGateway.php
- Banco de la República (Colombia). Disponible en http://www.banrep.gov.co/
- Organización Mundial de la Salud. Disponible en http://www.who.int/research/en/
- Estadística de búsquedas de Google. Disponible en http://www.google.com/insights/search/?hl=es#
- Tendencias de Google. Disponible en http://www.google.es/trends

Varios ejercicios hacen referencia a bases de datos que se pueden consultar en la plataforma virtual Avata de la Universidad de Bogotá Jorge Tadeo Lozano. Su tamaño obliga a un procesamiento con un programa de computador.

Se advierte que se sigue la convención de escribir la coma como separador decimal y el punto como separador de miles, utilizada en Centroamérica, Suramérica y Europa. No obstante, el lector encontrará algunas gráficas en las cuales esto no ocurre porque el *software* con el que fueron generadas proviene de un país donde el punto es el separador decimal.

1
Estadística Descriptiva

¿Qué es la estadística?

No hace falta ser un experto para estar en contacto con la estadística. Con frecuencia, los medios de comunicación presentan información de naturaleza estadística, como encuestas de intención de voto y de opinión; información sobre el *rating*; evolución de las tasas de cambio de diferentes divisas y otros indicadores económicos; datos del clima, etc. Los resultados que muestran los medios solo son una parte del trabajo estadístico y detrás de ellos existe una amplia disciplina, un forma de razonar y un método de generación de conocimiento. Los problemas que resuelve la estadística surgen de preguntas en diversas áreas que se fundamentan en el conocimiento inductivo. En un ámbito más pragmático, la estadística tiene que ver con la forma más conveniente de obtener, resumir y analizar información. Algunos autores la definen como el arte de aprender de los datos. En la tabla 1.1 se muestran algunas preguntas, de diversas áreas, en las cuales la estadística resulta muy útil.

Tabla 1.1. Algunas preguntas en las cuales resulta útil la estadística

Economía	¿Cuál es el ingreso promedio de un profesional egresado en Cartagena?
	¿Cuántas personas viajan en avión en Colombia?
	¿Cómo han variado los precios de los alimentos el último mes?
	¿Cuántas personas desempleadas hay en el país?
	¿Cuál será la tasa de cambio del dólar el próximo mes?
Publicidad	¿Cuántas personas ven un determinado programa de televisión?
	¿Los adolescentes en Nariño leen prensa escrita?
	¿Qué percepción tienen los consumidores de cierta marca?
Mercadeo	¿Cuántas tiendas hay en Medellín?
	¿A qué precio se debe vender un producto?
	¿Cuánto vende una marca de bebida gaseosa en Ibagué?
Política	¿Cuál es la favorabilidad de un gobernante?
	¿Cuántas personas pueden participar en las siguientes elecciones?
Sector público	¿Cuántos niños menores de cinco años hay en actualidad en Boyacá?
	¿Cuántos ancianos hay en Santa Marta?
	¿Cuánto le cuesta al Estado una elección popular?
	¿Qué características tienen los evasores del impuesto predial de Bogotá?

Fuente: elaboración propia.

La estadística se ha convertido en el lenguaje aceptado por la comunidad científica para la puesta a prueba, validación o rechazo de hipótesis de investigación; una descripción estadística es una forma de comunicación con un lenguaje especial. Los términos estadísticos tienen significados precisos y se asimilan paso a paso. La estadística se puede clasificar por su intención como descriptiva o inferencial. La primera tiene que ver con la mención de los hechos observados o la descripción de características de un conjunto de datos. La segunda se refiere a la obtención de propiedades generales basada en una parte de un conjunto de datos. En este libro se estudian algunas técnicas para hacer descripciones estadísticas y se realiza una introducción a la probabilidad, que es la base de la inferencia estadística.

Para desarrollar las ideas anteriores, considere la siguiente situación: imagine que se desea estudiar cuáles son los hábitos y los gastos de los estudiantes de una universidad en telefonía celular. En algún momento posterior esto requerirá realizar un sondeo dentro de la población de estudiantes. La tarea inicial se debe concentrar en determinar un objetivo y definir claramente el concepto central del estudio. En este caso, definir con exactitud qué se entiende por hábitos y gastos. Las características del problema particular determinarán si basta con una descripción de la información recolectada o si con esta se pueden realizar conclusiones más generales o atribuir causas a los comportamientos observados. Este proceso está más allá del alcance de este libro, pero los lectores interesados en el tema pueden remitirse a Mendenhall (2013) y Lohr (2010) para un estudio detallado.

Población, muestra y variables

En estadística, se llama *población* a una colección total de elementos que tienen algunas características comunes y sobre la que se desea obtener alguna información o realizar algún análisis. El tamaño de la población es el número de elementos de esta colección y generalmente se denota con la letra *N*. Con frecuencia el tamaño de la población es demasiado grande como para intentar examinar a todos sus elementos, por esta razón es conveniente seleccionar solo una parte de los elementos de la población, es decir, una *muestra*, que refleje el comportamiento general de la población con respecto a lo que se desea estudiar. A las medidas que se refieren a los elementos de toda la población se les llama *parámetros*, mientras que a las que son obtenidas de los elementos de la muestra se les llama *estadísticos*.

Sobre los elementos de la población o de la muestra se miden características que recogen información del concepto en estudio. Como estas características son diferentes en cada elemento se denominan *variables*. Por ejemplo, si se está interesado en averiguar sobre la actividad laboral en un grupo de individuos en Bucaramanga, se pueden medir variables como nivel de estudio o salario en los individuos empleados, o edad y tiempo de inactividad entre los desempleados.

Aquella variable que mida un atributo o característica que se pueda clasificar en categorías se le denomina *cualitativa*. Si examina una característica que se pueda medir o contar, se le llama *cuantitativa*.

Existen diversas clasificaciones de las variables según su naturaleza y escala. Una de estas es la división de las variables cualitativas en nominales y ordinales, y las cuantitativas, en discretas y continuas:

Variable cualitativa ordinal: identifica características que se pueden clasificar en grupos que admiten orden. Ejemplos de esto son el resultado de una entrevista (rechazado, normal y sobresaliente) o el premio en una prueba olímpica (oro, plata, bronce).

Variable cualitativa nominal: identifica características que se pueden clasificar en grupos que difieren o son iguales en una propiedad, por ejemplo, el municipio de nacimiento o el color de los ojos. Observe que en las variables cualitativas nominales las características no admiten un orden natural.

Variable cuantitativa discreta: es aquella que toma un número finito o contable[1]* de valores. El número de respuestas correctas en un examen de cinco preguntas {0,1,2,3,4,5} o el número de hijos de una persona {0,1,2,⋯} son ejemplos de estas variables.

Variable cuantitativa continua: es aquella que puede tomar infinitos valores en cualquier intervalo razonable de la variable. La estatura de un hombre adulto es una variable aleatoria continua. Note que existe una infinidad de posibilidades para la estatura en un intervalo arbitrario, por ejemplo, entre los 160 y 161 centímetros. Aunque las personas suelen responder con un número entero cuando se les pregunta por su estatura (en centímetros), esta variable se puede medir con la precisión deseada utilizando el instrumento adecuado.

Tabla 1.2. Tipos de variables

		Los valores que toma la variable:
Cualitativa	**Ordinal**	Se pueden clasificar en grupos ordenados
	Nominal	Se pueden clasificar en grupos pero no hay un orden entre ellos
Cuantitativa	**Discreta**	Son finitos o contables
	Continua	Son infinitos en cualquier intervalo razonable

Fuente: elaboración propia.

Adicionalmente, algunos autores consideran una clasificación adicional de variables según la escala de medición. Una variable cuantitativa en cuya elección del punto correspondiente al cero no es arbitraria y significa la ausencia de la característica que se mide se dice que está medida en *escala de razón*, por ejemplo, el peso o la estatura. Si el cero de la escala es arbitrario, se dice que es de *escala de intervalo***.** Son ejemplo de esto la temperatura en grados centígrados o el resultado de conocimientos en una prueba de matemáticas.

1 En términos matemáticos, un conjunto se dice contable si es posible construir una función biyectiva entre dicho conjunto y los números naturales.

Ejemplo 1.1

Enunciado:

¿Cree que el color rosado es femenino? Existe una percepción generalizada acerca de que efectivamente lo es. Es posible que en el pasado esto no fuera así o que ciertas culturas actuales no lo vean como un color femenino. Para estudiar esta idea se llevó a cabo un experimento con 65 integrantes de la comunidad embera, cuarenta hombres y veinticinco mujeres. A cada uno se le pidió señalar su predilección entre tres prendas idénticas salvo por el color (azul, verde o rosado). Se obtuvieron los siguientes resultados:

Tabla 1.3. Respuestas del ejemplo 1.1

	Color escogido		
	Azul	**Verde**	**Rosado**
Hombres	22	12	6
Mujeres	4	9	12

Fuente: elaboración propia.

En general, ¿las mujeres embera se sienten más atraídas por el color rosado que los hombres?

Solución:

Observe que la pregunta hace referencia a la población embera, cuyo tamaño puede llegar a los 70.000 individuos. En el estudio se consideró una muestra de 65 elementos, es decir, 65 miembros de esta comunidad. La pregunta hace referencia a la población, es decir, se intenta realizar una inferencia estadística con base en la muestra.

No es posible responder a la pregunta a partir de esta información. Se desconoce la razón por la cual se seleccionó una muestra de 65 individuos o si este número es suficiente para tener una visión de las preferencias de toda la comunidad. No se menciona si la muestra se eligió al azar o si, por el contrario, se autoelegió (los que decidieron participar por voluntad propia). Tampoco se revela la manera como se permitió a los individuos la elección del color, si fue de manera individual o grupal. Este aspecto es importante porque si se hiciera de manera grupal, las últimas elecciones se pueden ver influenciadas por las primeras. Todo esto sugiere que para realizar una inferencia estadística de esta naturaleza se requiere un diseño o plan mucho más cuidadoso. Por lo pronto solo es posible hacer descripciones estadísticas como por ejemplo:

La mayoría de los hombres embera de la muestra escogieron el color azul (55%), mientras que la mayoría de las mujeres optaron por el color rosado (48%). Tan solo el 15% de los hombres de la muestra fueron atraídos por el color rosado, porcentaje similar al de las mujeres que escogieron el color azul, 16%.■

Gráficos para resumir datos cualitativos

Diagrama de barras y gráfico de sectores

En esta parte del libro se ilustra la utilidad de varias gráficas estadísticas a la hora de extraer información de los datos. Se mencionan los gráficos de barras, de sectores y de puntos. Se acompañan los ejemplos con el análisis de la situación considerada. Como regla general, cuando se presenta algún tipo de información estadística, las gráficas deben ir acompañadas de un comentario o explicación.

Ejemplo 1.2

Enunciado:

Construya una gráfica para representar la distribución del sexo en un salón con 10 hombres y 6 mujeres.

Solución:

Primero se resume la información en una tabla de frecuencias:

Tabla 1.4. Tabla de frecuencias del ejemplo 1.2

Sexo	Frecuencia absoluta	Frecuencia relativa
Hombres	10	$\frac{10}{16} \times 100 = 62{,}5\%$
Mujeres	6	$\frac{6}{16} \times 100 = 37{,}5\%$

Fuente: elaboración propia.

Cuando se desea comparar algún valor en diferentes categorías se utiliza un *diagrama de barras* (figura 1.1) y, cuando lo importante es resaltar la contribución de cada categoría, se utiliza un gráfico de sectores (figura

1.2). Las dos gráficas son igualmente informativas, pero con los énfasis mencionados. Observe que las barras aparecen separadas en el diagrama porque no existen ni orden ni distancia numérica entre los sexos.■

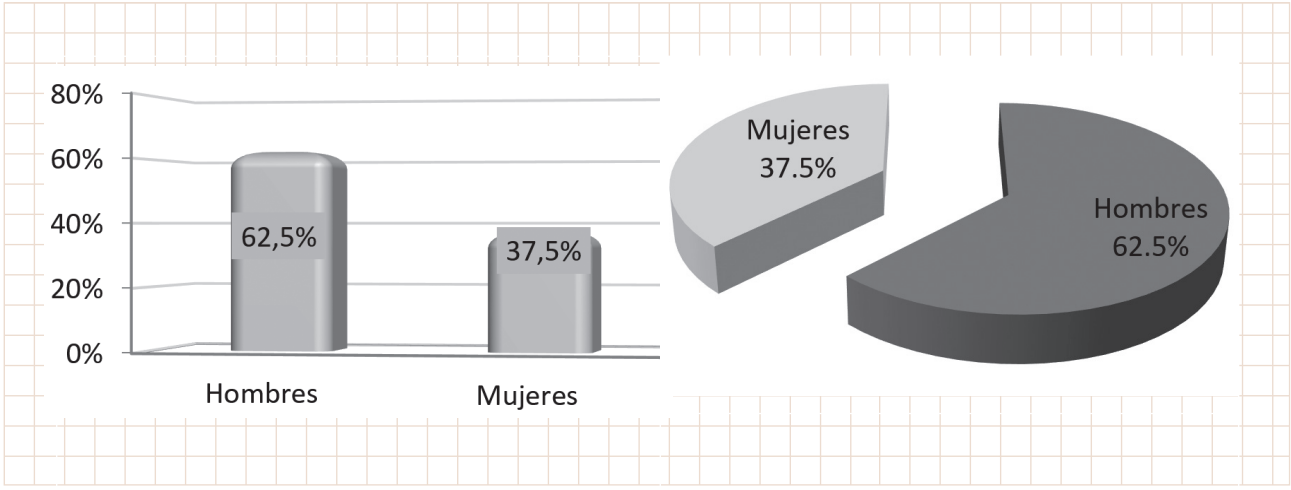


Figura 1.1. Diagrama de Barras. Distribución del sexo. Fuente: elaboración propia.

Figura 1.2. Gráfico de sectores. Distribución del sexo. Fuente: elaboración propia.

Ejemplo 1.3

Enunciado:

En cierto año, el número global de quejas a tres operadores de telefonía móvil fueron los siguientes:

Tabla 1.5. Número de quejas a operadores móviles

Empresa	Quejas
Claro	15.008
Movistar	5.825
Tigo	1.997
Usuarios: Claro (28.495.932), Movistar (9.537.552),Tigo (4.429.875)	

Fuente: elaboración propia.

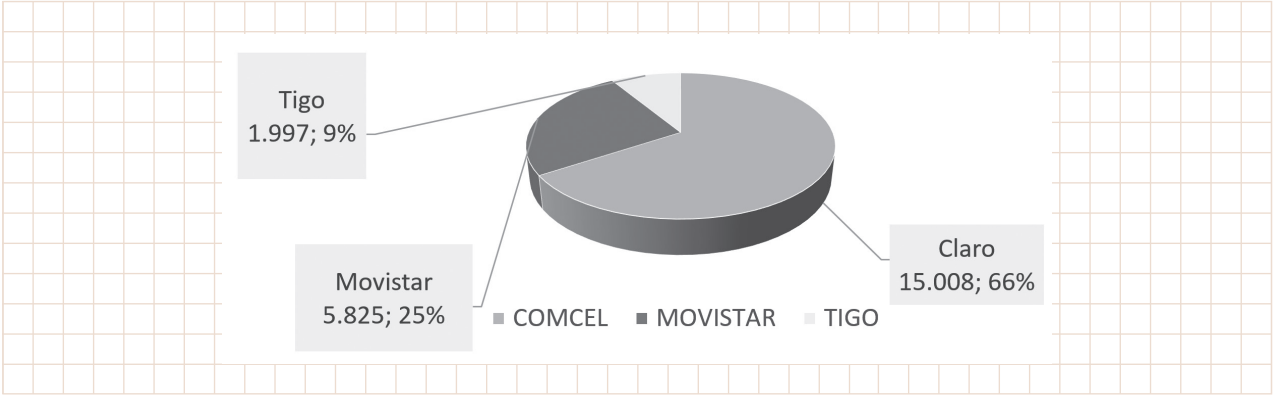


Figura 1.3. Quejas a operadores de telefonía. Fuente: elaboración propia.

Analice la gráfica y la situación de quejas de estos tres operadores.

Solución:

El gráfico de sectores mostrado no representa la realidad porque fue construido con el número global de quejas. Es de esperar que las empresas con más usuarios tengan más quejas, y esto no implica que tengan un peor servicio.

Si se calculan las quejas por cada 100.000 usuarios, tal como lo muestra el diagrama de barras, se puede observar que la empresa de telefonía con más quejas es Movistar (61 por cada 100.000 usuarios), le sigue Claro con 53, y, por último, está Tigo con 45.■

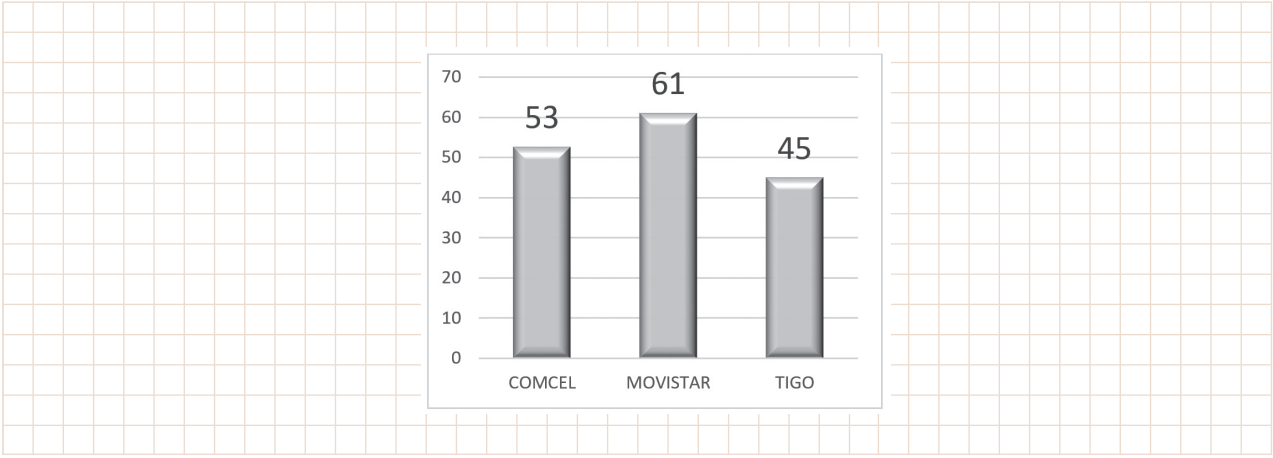


Figura 1.4. Diagrama de barras. Quejas a operadores de telefonía. Fuente: elaboración propia.

Rascacielos

Un rascacielos es un diagrama que permite ver la distribución conjunta de dos variables de tipo cualitativo.

Ejemplo 1.4

Enunciado:

Construya una gráfica para representar las variables *sexo* y *condición de fumador*, consignadas en la tabla 1.6.

Tabla 1.6. Frecuencias de sexo y condición de fumador

	Hombres	Mujeres	Total
Fuma	6	5	11
No fuma	4	1	5
Total	10	6	16

Fuente: elaboración propia.

Solución:

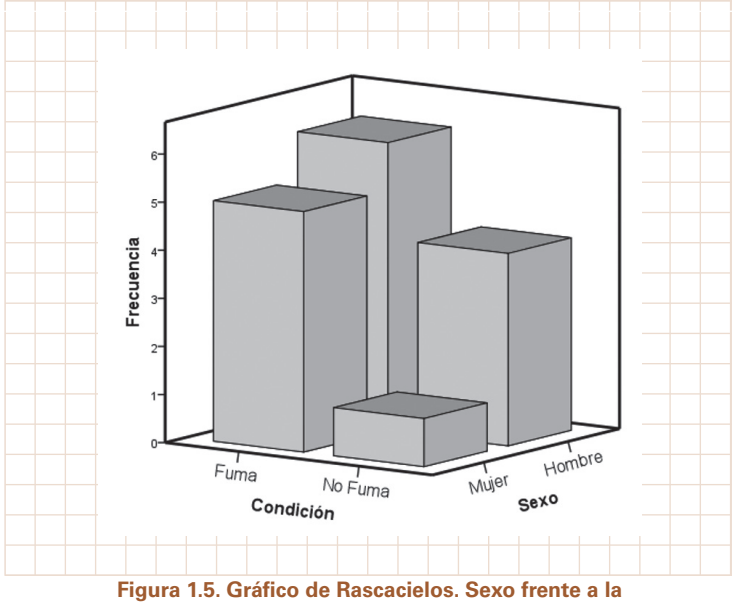


Figura 1.5. Gráfico de Rascacielos. Sexo frente a la condición de fumador. Fuente: elaboración propia.

Un gráfico de *rascacielos* como el de la figura 1.5 es adecuado para estos datos. Las alturas de los prismas rectangulares corresponden a las frecuencias de cada uno de los cuatro grupos formados por las dos variables:

6 hombres fumadores.
4 hombres no fumadores.
5 mujeres fumadoras.
1 mujer no fumadora.■

Diagramas de barras y mosaicos

Otra gráfica para el mismo conjunto de datos es la de la figura 1.6. A esta se le llama diagrama de *barras bivariado*, o de *columnas apiladas*, y representa los mismos cuatro grupos de la figura 1.5.

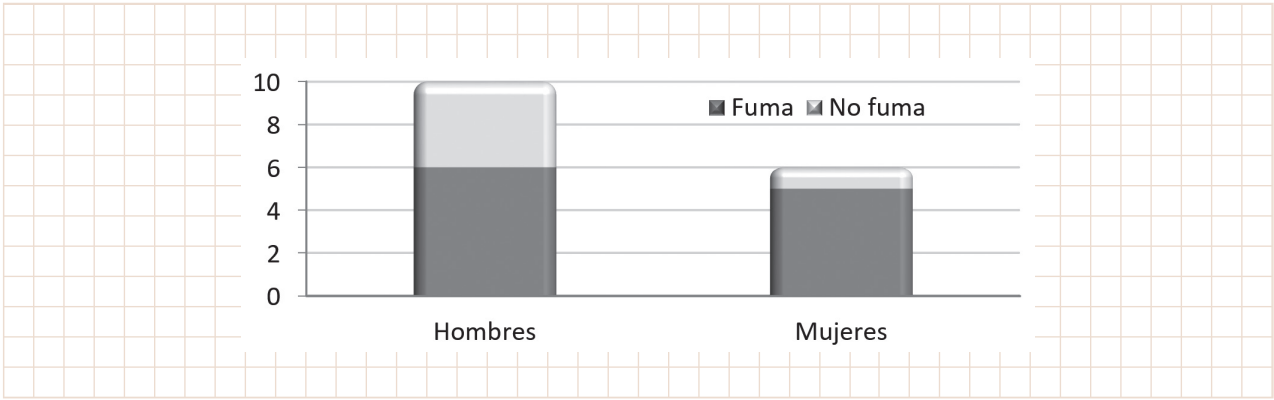


Figura 1.6. Gráfico de barras bivariado. Sexo frente a condición de fumador. Fuente: elaboración propia.

En los dos últimos gráficos sobresale más el grupo de hombres fumadores. Esto puede sugerir, erróneamente, que hay mayor número de fumadores entre los hombres. En cifras absolutas esto es cierto, pero en términos relativos existe un mayor porcentaje de fumadores en las mujeres. Para evitar malentendidos con las cifras absolutas se construye un gráfico de barras de frecuencias relativas, o *mosaico*, que representa cada sexo en columnas de la misma altura. Para ello se parte de la tabla 1.6. Por su parte, en la tabla 1.7 se muestran las frecuencias relativas a la tabla, la fila (*condición de fumador*) y la columna (*sexo*).

Tabla 1.7. Frecuencias relativas de sexo y condición de fumador

	Hombres	Mujeres
Fuma	6	5
	37,5%	31,3%
	54,5%	45,5%
	60,0%	83,3%
No fuma	4	1
	25,0%	6,3%
	80,0%	20,0%
	40,0%	16,7%

Celda
Frecuencia absoluta
Porcentaje de la tabla
Porcentaje de la fila
Porcentaje de la columna

Fuente: elaboración propia.

A manera de ejemplo se muestran los significados de los valores en la última celda (tabla 1.8.):

Tabla 1.8. Ejemplo de interpretación de la tabla 1.7

1	Número de mujeres no fumadoras
6,3%	Porcentaje de mujeres no fumadoras
20,0%	Porcentaje de mujeres dentro de los no fumadores
16,7%	Porcentaje de no fumadoras dentro de las mujeres

Fuente: elaboración propia.

Con base en la información de las celdas correspondientes a los porcentajes columna de la tabla 1.6. se construye la figura 1.7:

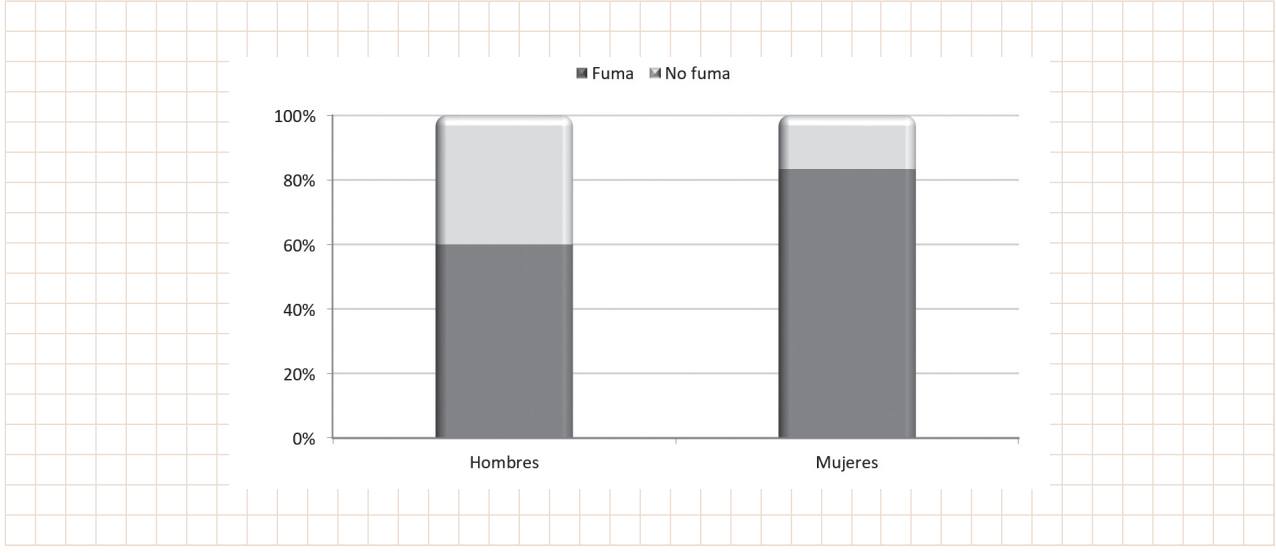


Figura 1.7 Mosaico. Sexo frente a condición de fumador. Fuente: elaboración propia.

En este último gráfico se observa que existe mayor porcentaje de fumadores en las mujeres que en los hombres en la muestra de 16 personas de la tabla 1.6.

Ejemplo 1.5

Enunciado:

A un grupo de participantes de un congreso sobre minorías indígenas se le realizó una encuesta. Una de las preguntas era si las tareas del hogar son responsabilidad exclusiva del hombre, de la mujer o son una responsabilidad compartida. El siguiente gráfico (figura 1.8) resume los resultados e incluye el país de origen de los entrevistados.

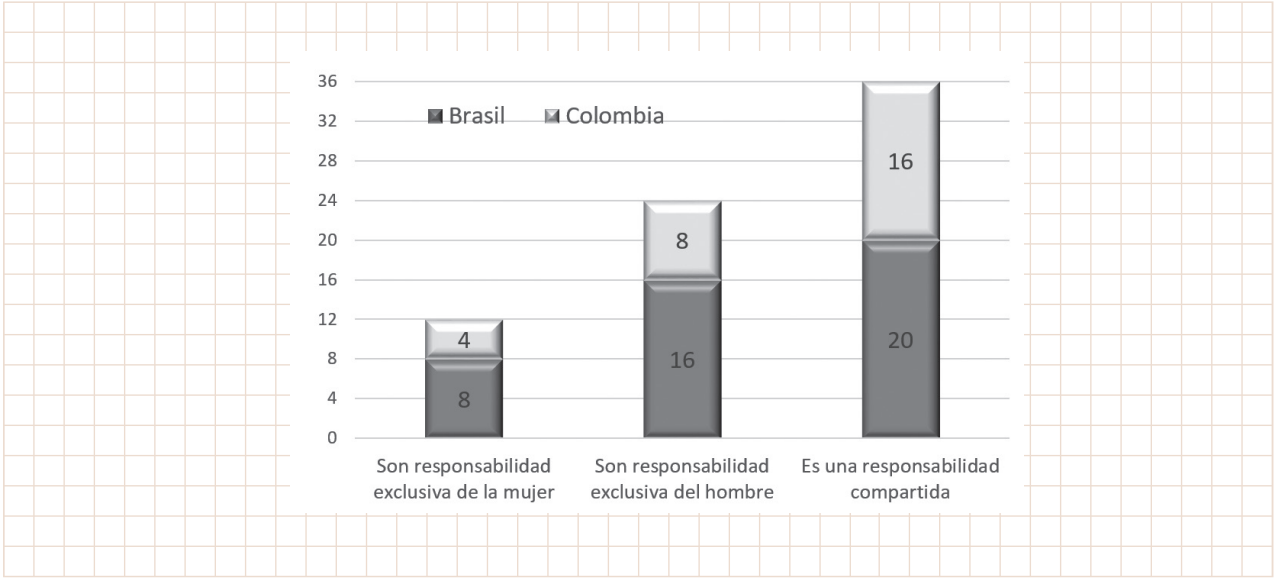


Figura 1.8. Responsabilidad de las tareas de hogar. Fuente: elaboración propia.

a. ¿Qué porcentaje de entrevistados opina que es una responsabilidad compartida?
b. De los que opinan que es responsabilidad exclusiva de la mujer, ¿qué porcentaje son colombianos?
c. De los brasileños, ¿qué porcentaje cree que es una responsabilidad exclusiva del hombre?
d. ¿Quiénes tienen una posición menos sexista?

Solución:

a. El 50% de los entrevistados opina que es una responsabilidad compartida.
b. De los que opinan que es una responsabilidad exclusiva de la mujer, el 33,33% son colombianos
c. De los brasileños, el 36,36% cree que es una responsabilidad exclusiva del hombre.
d. La posición sexista de los colombianos es de 42,86%, mientras que la de los brasileños es de 54,55%. Por lo tanto, en esta muestra, los colombianos tienen una posición menos sexista.■

Ejemplo 1.6

Enunciado:

La mortalidad perinatal (MP) está constituida por todas las muertes ocurridas entre la semana 18 de gestación y el séptimo día de nacimiento. El desarrollo social, acceso, cobertura y calidad en los servicios de salud en una población puede ser evaluado por medio de este tipo de información. En la tabla 1.9 se presentan los casos de MP en cada uno de los departamentos del país durante el 2009 según el Ministerio de Salud (2017)

Construya una gráfica con las estadísticas de MP en Colombia contenidas en la tabla 1.9.

Solución:

No es posible realizar una comparación de MP entre los departamentos colombianos solo con el número de casos registrados durante el 2009. Esta cifra absoluta no es comparable ya que resulta lógico que se presenten más casos en aquellos departamentos en donde hay mayor población, y esto puede suceder de manera independiente de las condiciones de los servicios de salud. Para realizar una comparación objetiva, la estadística MP se debe acompañar del tamaño de la población en cada departamento. De esta manera, se calculan los casos de MP por cada cien mil habitantes, cifra que resulta comparable y da un indicio de lo que ocurre en cada departamento. Como en otras cifras de salud pública, no se muestra las estadísticas por cada cien habitantes debido a la baja incidencia del fenómeno. Esto se debe a que, tratándose de casos, sea deseable expresar los resultados en valores enteros. Se agregó a la tabla la población total y el cálculo del número de casos de MP por cada 100.000 habitantes.

Tabla 1.9. Mortalidad perinatal en Colombia (Ministerio de Salud, 2017)

Depto.	Antioquia	Arauca	Atlántico	Bogotá	Bolívar	Boyacá	Caldas	Caquetá	Casanare	Cauca
Casos de MP	1.059	49	231	1.353	292	283	173	41	73	199
Población total	5.988.458	244.507	2.284.840	7.259.597	1.958.224	1.265.517	976.438	442.033	319.502	1.308.090
MP x 100.000	18	20	10	19	15	22	18	9	23	15

Depto.	Cesar	Chocó	Córdoba	C/marca	Huila	Guajira	Magdalena	Meta	Nariño	Santander N
Casos de MP	255	109	372	146	328	112	277	219	298	189
Población total	953.827	471.601	1.558.793	2.437.151	1.068.820	791.027	1.190.585	853.115	1.619.464	1.286.728
MP x 100.000	27	23	24	6	31	14	23	26	18	15

Depto.	Putumayo	Quindío	Risaralda	San Andrés	Santander	Sucre	Tolima	Valle	Amazonia
Casos de MP	69	101	139	22	382	183	168	490	61
Población total	322.681	546.566	919.653	72.735	2.000.045	802.733	1.383.323	4.337.909	313.796
MP x 100.000	21	18	15	30	19	23	12	11	19

Fuente: elaboración propia.

De esta manera, se representa en la figura 1.9 cada uno de los departamentos junto con el número de casos de MP por cada 100.000 habitantes. Para que la lectura de la gráfica sea más sencilla se ordenaron los departamentos según la mortalidad perinatal. Note que los dos departamentos con mayor MP son Huila y San Andrés, mientras que los dos con menor MP son Caquetá y Cundinamarca. Bogotá es una región en la que se presenta una mortalidad perinatal intermedia.■

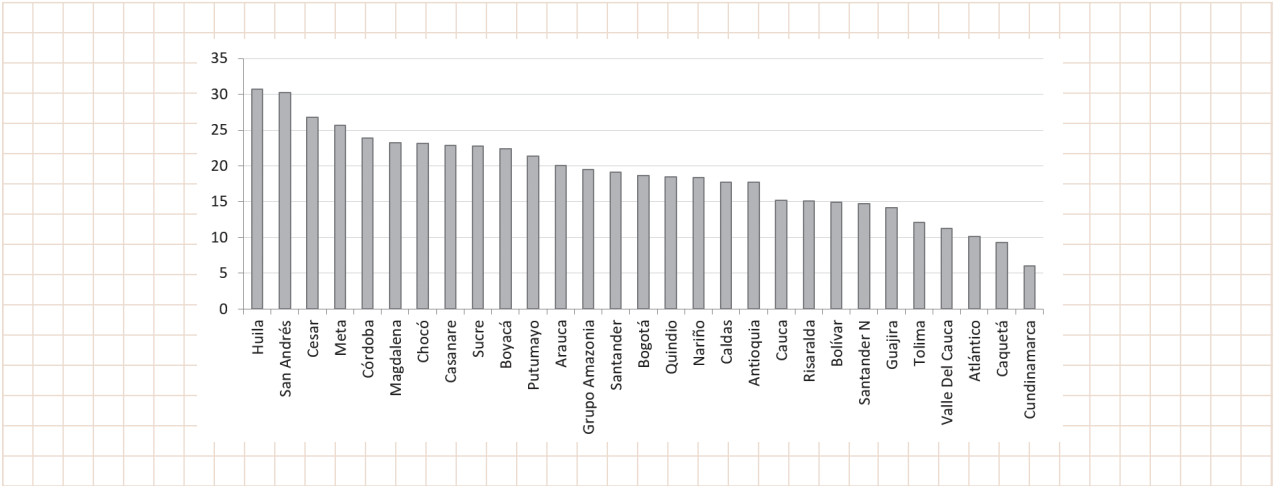


Figura 1.9. Mortalidad perinatal en Colombia durante el 2009. Casos por cada 100.000 habitantes. Fuente: elaboración propia.

Ejercicio 1.1 Descriptiva 1

En cuanto sea posible, utilice una hoja de cálculo o un programa de computador especializados para construir las gráficos de los siguientes ejercicios.

1. Plantee una pregunta de investigación en la que resulte útil un estudio estadístico.

2. Discuta el proceso que se debería adelantar para estudiar cuáles son los hábitos y gastos de los estudiantes de la universidad en telefonía celular.

3. Suponga que desea averiguar cuáles son las empresas en el país con mejores salarios. Discuta cuál es la población y cómo se debería seleccionar una muestra.

4. Determine el elemento sobre el que se deben medir las siguientes variables:
   a. Candidato por el que se votará en una elección.
   b. Números de goles de un jugador.
   c. Calificación de un estudiante en un parcial.
   d. Resultado de una prueba de embarazo.
   e. Precipitación (la cantidad de lluvia se mide en mm y equivale al espesor de la lámina de agua que se formaría por su causa sobre una superficie de $1m^2$ plana e impermeable).

5. Diga si las siguientes variables son cuantitativas o cualitativas:
   a. Saldo en las cuentas de ahorro de un banco.
   b. Goles marcados en los partidos de fútbol de un campeonato.
   c. Estado en el que inician sesión los usuarios de Messenger.
   d. Nacimientos mensuales en el Hospital San Ignacio.
   e. Tiempo de duración de una llamada.
   f. Porcentaje de estudiantes fumadores en las universidades de Bogotá.

6. Explique la diferencia entre variables cuantitativas y cualitativas. Dé un ejemplo de cada una.

7. Diferencie las variables cualitativas nominales de las ordinales. Proporcione un ejemplo de cada una.

8. Diga si las siguientes variables son discretas o continuas:
   a. Número de dulces en un paquete.
   b. Saldo en una cuenta corriente.
   c. Calificación de un estudiante en el primer corte del semestre.
   d. Número de titulares de una tarjeta de crédito que se encuentran en mora.

9. Explique la diferencia entre variables discretas y continuas. Dé un ejemplo de cada una.

10. Los siguientes datos corresponden a titulares de créditos de un banco:

Datos de los titulares de crédito de un banco del ejercicio

Identificación	Nombre	Edad	Profesión	Ingresos	Vivienda	Vehículo
802342...	María C...	23	Abogado	6.540.000	Sí	CRE543
791234...	Juan An..	54	Traductor	5.000.400	Sí	SAP232
102345...	Ricardo...	35	Ingeniero	4.850.000	No	-
323212...	Mario...	29	-	2.350.000	No	HDR342

Fuente: elaboración propia.

   a. ¿Cuáles son los elementos de la población?
   b. ¿Cuáles son las variables?
   c. ¿Qué tipo de variable es cada una?

11. Una muestra de amas de casa del sur de la ciudad escogió el supermercado de su preferencia, con los siguientes resultados:

Supermercado de preferencia de amas de casa del ejercicio 1.1.11

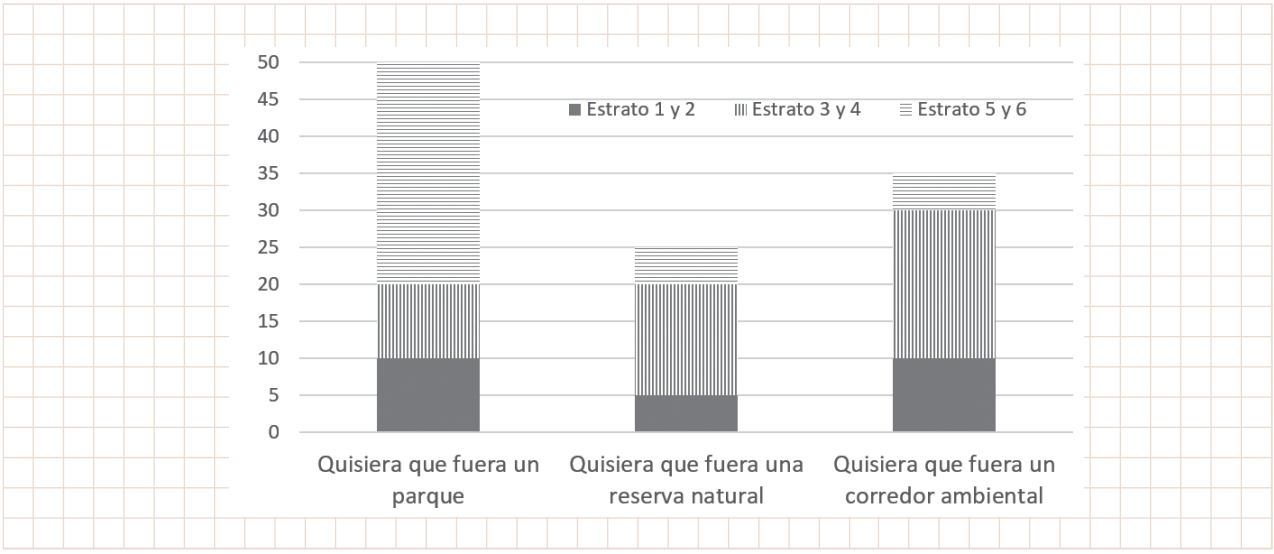
Supermercado	Amas de casa
Tiendas D1	741
Justo y Bueno	382
Ara	100
Jumbo	143

Supermercado	Amas de casa
Éxito	254
Carulla	84
Olímpica	179
Colsubsidio	340

Fuente: elaboración propia.

a. ¿Cuáles son los elementos de la población?
b. ¿Cuál es la variable que se mide?; ¿es cualitativa o cuantitativa?
c. Construya un gráfico de barras para describir los datos.
d. Construya un gráfico de sectores para describir los datos.
e. ¿Cuál es la proporción de amas de casa que tienen preferencia por el Éxito?
f. ¿Cuál es la proporción de amas de casa que no prefieren Olímpica o Colsubsidio?

12. A un grupo de habitantes de Bogotá se le realizó una encuesta. Una de las preguntas consistía en elegir cómo desearían ver en el futuro los alrededores del río Fucha. El siguiente gráfico resume los resultados e incluye el estrato socioeconómico de los entrevistados.



Futuro del río Fucha. Fuente: elaboración propia.

a. ¿Qué porcentaje de entrevistados quisiera ver los alrededores del río Fucha como una reserva natural?
b. De quienes quisieran que fuera un corredor ambiental, ¿qué porcentaje son de estrato 3 y 4?
c. De los habitantes de estrato 1 y 2, ¿qué porcentaje quisieran que fuera un corredor ambiental?
d. ¿Cuáles estratos se inclinan más por ver los alrededores de río Fucha como un parque?

13. En un salón de clases los estudiantes registraron el lugar de fabricación de sus teléfonos celulares, con los siguientes resultados:

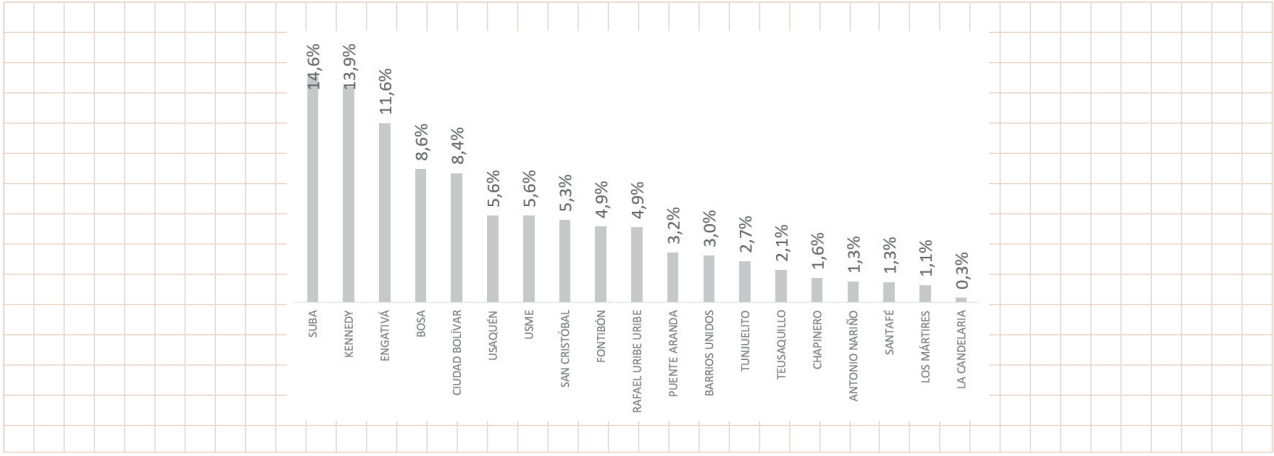
Lugar de fabricación de los teléfonos celulares del ejercicio 1.1.13

SHANGHAI	EE. UU.	SHANGHAI	MÉXICO
SHANGHAI	SHANGHAI	HONG KONG	TAIWAN
SHANGHAI	MEXICO	TAIWAN	TAIWAN
TAIWAN	SHANGHAI	TAIWAN	SHANGHAI
HONG KONG	EE. UU.	HONG KONG	EE. UU.
SHANGHAI	SHANGHAI	TAIWAN	TAIWAN
SHANGHAI	SHANGHAI	TAIWAN	TAIWAN
HONG KONG	SHANGHAI	MÉXICO	TAIWAN

Fuente: elaboración propia.

a. ¿Cuáles son los elementos de la población?
b. ¿Cuál es la variable que se mide?
c. Construya un gráfico para describir los datos.
d. ¿Qué proporción de teléfonos fueron construidos fuera de China?
e. ¿Qué proporción de teléfonos fueron construidos en América?

14. La figura muestra uno de los resultados publicados en Barragán y González (2017) sobre la proporción de la población estudiantil de 5 años en Bogotá, con base en la encuesta multipropósito del 2014.



Proporción de población estudiantil por localidad. Fuente: Barragán y González (2017).

a. ¿Existe un orden natural entre las localidades de Bogotá?
b. ¿Qué nombre recibe el anterior gráfico?
c. ¿Cuántas variables están representadas en el gráfico?
d. Averigüe el número de habitantes por localidad en Bogotá y determine si hay algún tipo de relación entre la proporción de la población estudiantil y el número de habitantes por localidad.

15. Lo que un grupo de mujeres colombianas respondió a la pregunta "¿qué condición prefiere en un hombre?" de la encuesta "Lo que las mujeres quieren" ("Cien mujeres dicen", 2017) se resume en la siguiente tabla:

100 colombianas dicen: ¿Qué condición prefiere en un hombre?

Que sea sociable	22%
Que sea talentoso	60%
Que tenga dinero	11%
Que tenga pinta	5%

Fuente: adaptación de "Cien mujeres dicen" (2007).

a. ¿Están todas las condiciones enumeradas en la tabla de resultados? Al agregar otra condición, ¿cambiarían los porcentajes de las demás?
b. Construya un gráfico de barras para describir los datos.
c. Si usted le realiza esta misma pregunta a una nueva mujer y le da las mismas opciones, ¿cuál cree que sea la respuesta más probable?

16. Imagine que se desea averiguar cuál es el lugar que eligieron los estudiantes de una universidad en sus últimas vacaciones. Redacte la pregunta que debería incluirse en una encuesta dirigida a estudiantes de diferentes facultades, jornadas y semestres de esa universidad de tal manera que los encuestados puedan elegir una y solo una de las opciones incluidas en ella. ¿Qué características debe tener la pregunta? Pida a su profesor un ejemplo de categorías de respuestas exhaustivas y excluyentes.

17. En estudios de opinión, es frecuente la utilización de la escala Likert para determinar la orientación de las respuestas de un grupo de personas con respecto a un tema. Por ejemplo, en un estudio acerca de la posición de estudiantes de Administración y Relaciones Internacionales frente a la posibilidad de incluir en su carrera la certificación en un tercer idioma, se eligió la siguiente escala de respuesta:

Opciones de respuesta del estudio de opinión del ejercicio 1.1.17

1	Bastante en desacuerdo
2	Algo en desacuerdo
3	Neutral
4	Algo de acuerdo
5	Bastante de acuerdo
6	No aplica

Fuente: elaboración propia.

Un resumen de las respuestas dadas por una muestra de estudiantes se presenta en la siguiente tabla:

Resumen de las respuesta de los estudiantes del ejercicio 1.1.17

	Administración	Relaciones	Total
Bastante en desacuerdo	93	9	102
	19,6%	1,9%	21,5%
	91,2%	8,8%	
	31,5%	5,0%	
Algo en desacuerdo	136	16	152
	28,6%	3,4%	32,0%
	89,5%	10,5%	
	46,1%	8,9%	
Neutral	30	26	56
	6,3%	5,5%	11,8%
	53,6%	46,4%	
	10,2%	14,4%	

Frecuencia
% tabla
% fila
% columna

	Administración	Relaciones	Total
Algo de acuerdo	22	57	79
	4,6%	12,0%	16,6%
	27,8%	72,2%	
	7,5%	31,7%	
Bastante de acuerdo	8	67	75
	1,7%	14,1%	15,8%
	10,7%	89,3%	
	2,7%	37,2%	
No aplica	6	5	11
	1,3%	1,1%	2,3%
	54,5%	45,5%	
	2,0%	2,8%	
Total	295	180	475
	62,1%	37,9%	100,0%

Fuente: elaboración propia.

a. ¿Cuántos estudiantes de Administración tuvieron una posición neutral?
b. ¿Qué porcentaje de respuestas corresponde a estudiantes de Relaciones?
c. ¿Qué porcentaje de respuestas corresponde a estudiantes de Relaciones y que están algo de acuerdo?
d. Entre los estudiantes de Administración, ¿cuál es el porcentaje de los que estuvieron bastante en desacuerdo?
e. Entre quienes estuvieron bastante de acuerdo, ¿cuál es el porcentaje de estudiantes de Relaciones?
f. Si se elige un estudiante al azar, ¿cuál es su opinión más probable?
g. Si se escoge un estudiante aleatoriamente y se sabe que estudia Relaciones, ¿cuál es su opinión más probable?
h. Desde un punto de vista intuitivo, ¿parece probable que los estudiantes de Administración y de Relaciones difieran sustancialmente en su opinión frente al tema consultado?

18. La siguiente tabla muestra la distribución de la edad de las personas de 5 años o más que se encontraban estudiando en el momento de aplicación de la encuesta multipropósito del 2014 en Bogotá.

Distribución de edades del ejercicio 1.1.18

¿En qué nivel está matriculado?

Edad	Preescolar	Primaria	Secudaria	Técnico	Tecnológico	Universitario	Especialización	Maestría	Doctorado
5-9	160.711	437.379	1.400	0	0	0	0	0	0
10-14	601	208.834	359.300	0	0	0	0	0	0
15-19	205	2.257	326.550	31.911	17.105	105.480	129	0	0
20-24	0	0	10.305	24.476	22.025	204.649	3.498	2.569	34
25-29	37	246	5.090	15.689	11.357	73.251	11.607	6.136	381
30-34	13	148	2.370	11.370	6.517	33.106	8.922	6.786	1.369
35-39	0	721	1.371	4.717	3.775	14.241	6.605	4.144	860
40-44	0	232	2.471	2.846	2.332	4.684	3.684	4.055	1.638
45-49	0	91	1.635	2.112	794	3.711	2.724	2.105	1.015
50-54	0	334	1.115	784	172	3.453	985	1.011	359
55-59	13	54	269	772	377	443	459	317	413
60-64	0	195	9	377	251	344	297	87	0
65-69	0	227	223	319	0	0	265	138	0
70-74	0	64	0	0	0	0	51	0	0
75-79	7	0	298	0	0	0	0	0	0
Total	**161.586**	**650.779**	**712.406**	**95.372**	**64.705**	**443.362**	**39.226**	**27.348**	**6.070**

Fuente: elaboración propia.

De este grupo de población (bogotanos de 5 años o más que se encuentran estudiando):

a. ¿Qué porcentaje está matriculado en la Universidad y está entre los 40 y 44 años?
b. De los matriculados en secundaría, ¿cuál es la proporción predominante en edad?
c. ¿Cuál es porcentaje de adultos mayores (65 años o más) que se encuentra matriculado en la secundaría?
d. ¿Qué porcentaje está matriculado en preescolar y está entre los 5 y 9 años?
e. De los matriculados en maestrías, ¿cuál es la proporción predominante en edad?
f. ¿Cuál es porcentaje de jóvenes de 20 a 24 años que se encuentra matriculado en la Universidad?

19. En un banco, a final de mes, se registró el sexo y el estado de cumplimiento de un grupo de usuarios de tarjetas de crédito.

Datos de los usuarios de tarjetas de crédito del ejercicio 1.1.19

Sexo	Estado	Sexo	Estado	Sexo	Estado
F	Incumplido	F	Cumplido	F	Incumplido
F	Cumplido	F	Incumplido	M	Cumplido
F	Incumplido	M	Cumplido	F	Incumplido
M	Incumplido	F	Incumplido	F	Cumplido
F	Cumplido	F	Incumplido	F	Incumplido
F	Cumplido	M	Cumplido	M	Cumplido
F	Incumplido	F	Incumplido	F	Incumplido
M	Cumplido	M	Cumplido	F	Cumplido

Fuente: elaboración propia.

a. Construya una tabla del *sexo* frente al *estado*.
b. Comente cualquier diferencia aparente que observe.

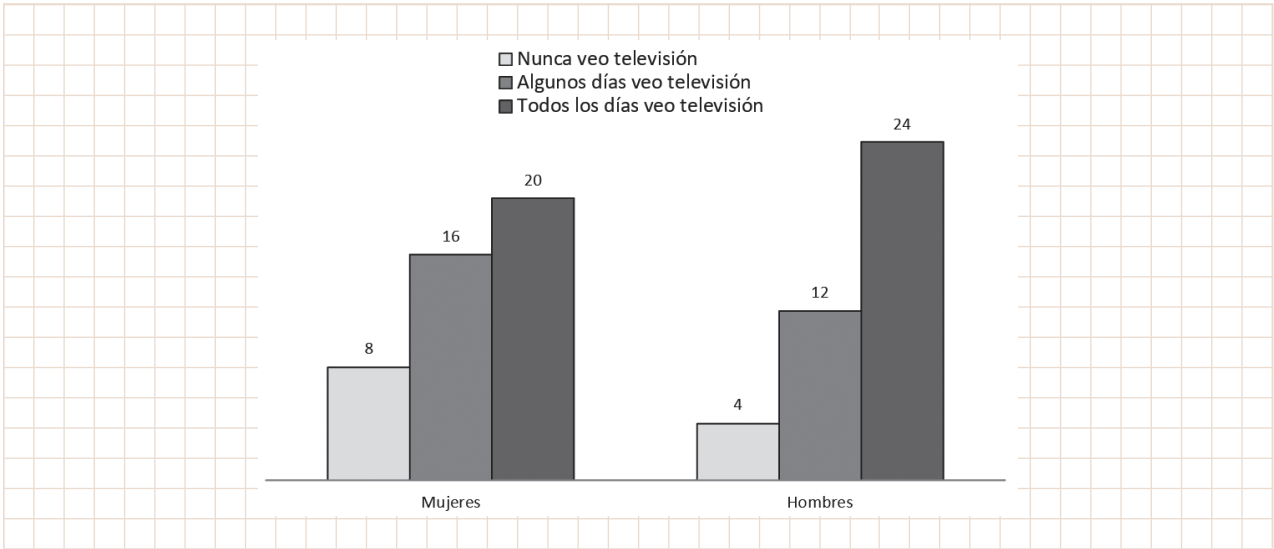
20. La paradoja de Simpson, nombrada en honor del estadístico que la planteó (Malinas y Bigelow, 2017), es un ejemplo de cómo la estadística pueden mostrar resultados contradictorios: se realiza un ensayo clínico para determinar la eficacia de dos drogas cuyos resultados se muestran en la tabla 1.18. ¿Cuál es la paradoja?; ¿cuál de las dos drogas es mejor?

La paradoja de Simpson

	Mujeres		Hombres	
	Droga 1	Droga 2	Droga 1	Droga 2
Surtió efecto	200	10	19	1.000
No surtió efecto	1.800	190	1	1.000

Fuente: adaptación de Malinas y Bigelow (2007).

21. En el 2002 se adelantó un estudio para examinar, entre otros aspectos, la relación del sexo con el gusto por la televisión en los estudiantes de colegios públicos en La Mesa. El gráfico muestra los resultados de las respuestas de una muestra de estudiantes acerca del sexo y si ven televisión.



¿Qué tanto ve televisión? Fuente: elaboración propia.

a. ¿Cuál es el número de estudiantes en la muestra?
b. ¿Qué porcentaje de estudiantes nunca ve televisión?
c. ¿Qué porcentaje de mujeres nunca ve televisión?
d. De quienes todos los días ven televisión, ¿qué porcentaje son hombres?

22. Los diez términos más buscados en Google en el 2017 en Colombia son:

Los términos más buscados en Google durante el 2017 en Colombia

1	Facebook	100
2	Youtube	65
3	Colombia	65
4	Traducción	40
5	Bogotá	35
6	Video	30
7	Imagen	30
8	Juego	25
9	Gmail	25
10	Idioma español	3

Fuente: Google Trends (2017).

Estos términos pueden corresponder a un grupo menor de categorías. Por ejemplo, Gmail puede pertenecer a la categoría de correo electrónico.

a. Construya un grupo reducido de categorías que reúna en grupos los términos de la tabla.
b. Explique por qué es de utilidad una agrupación de este tipo.
c. Dé un ejemplo en el que sea útil crear categorías para resumir información.

23. Los siguientes datos corresponden a las edades y el sexo de 20 asistentes a un curso de inglés avanzado:

Edad y sexo de una muestra de los asistentes a un curso avanzado de inglés

(M: Masculino, F: Femenino)									
15	15	18	25	22	42	40	32	24	28
F	M	M	F	M	F	M	F	M	F
18	18	28	20	16	26	45	30	25	20
M	F	F	M	M	M	F	M	F	F

Fuente: elaboración propia.

a. Construya un diagrama de barras para el sexo.
b. Suponga que se desea clasificar a cada persona como "joven" o "mayor", y se determina que una persona es "joven" si tiene menos de 21 años, y "mayor", si tiene 21 años o más. Construya una tabla que resuma esta clasificación junto con la variable *sexo*.

24. Con la información contenida en Avata "Resultados del Mundial", ¿qué gráfica podría construir con el país y el número de partidos jugados?

25. De acuerdo con las cifras entregadas por la Secretaría de Movilidad de Bogotá, las infracciones o comparendos más comunes cometidos por conductores y motociclistas en la capital de enero a marzo del 2011 son:

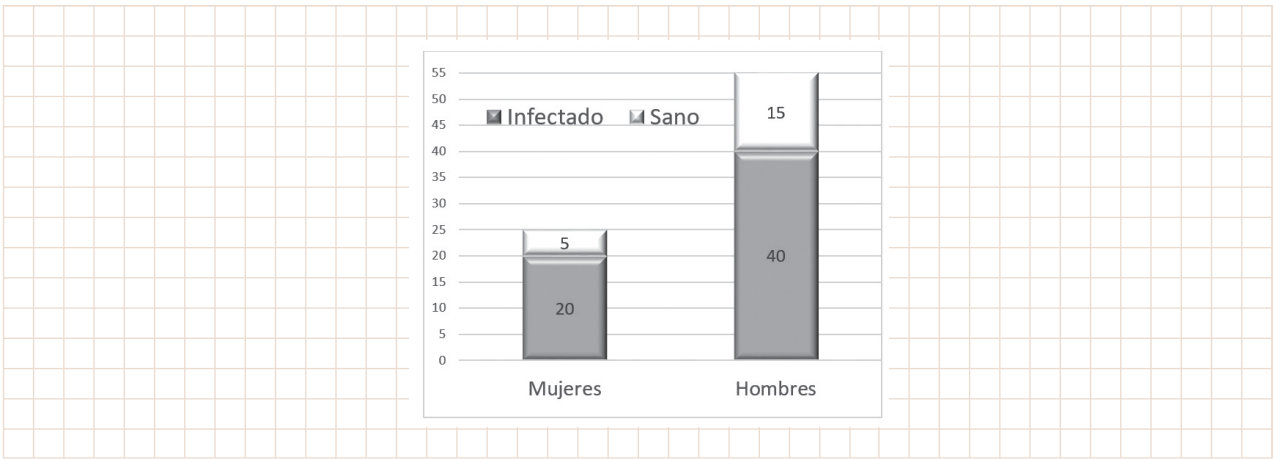
Comparendos más comunes cometidos por conductores y motociclistas en Bogotá de enero a marzo del 2011

Infracción	Comparendos
Incumplimiento de normas para motocicletas.	13.361
Estacionar un vehículo en sitios prohibidos.	9.998
Usar el celular al momento de conducir.	5.829
No realizar la revisión técnico-mecánica o de emisión de gases.	5.535
Conducir un vehículo sin portar la licencia de conducción.	4.682
No utilizar el cinturón de seguridad.	3.593
Transitar por sitios restringidos o en horas prohibidas.	3.342
Bloquear una calzada o intersección con un vehículo.	2.470
Conducir un vehículo de transporte público de pasajeros incumpliendo las normas.	1.360
Dejar o recoger pasajeros en sitios prohibidos.	180

Fuente: Secretaría Distrital de Movilidad (2011).

a. ¿Cuáles son los elementos de la población?
b. ¿Estos datos corresponden a una población o a una muestra?
c. ¿Cuál es la variable que se mide?; ¿es cualitativa o cuantitativa?
d. Construya un gráfico adecuado para describir los datos.

26. Se examinó a un grupo de pasajeros provenientes de la Costa en un hospital de Medellín para determinar si estaba infectado con el virus del chikunguña. En el siguiente gráfico se muestra la relación que existe entre la presencia del virus y el sexo de estos pasajeros.



Virus del chikunguña con relación al sexo. Fuente: elaboración propia.

a. ¿Qué porcentaje de mujeres tenía el virus?
b. Del grupo de pasajeros, ¿cuál es el porcentaje de hombres infectados?
c. De los infectados, ¿qué porcentaje son mujeres?
d. De los hombres, ¿qué porcentaje están infectados?
e. En esta muestra, ¿el virus infectó más a hombres o más a mujeres?

Gráficos para resumir datos cuantitativos

Diagrama de tallo y hojas

Una técnica útil para la observación inicial de los datos y la comprensión de lo que representan y cómo se distribuyen es un *diagrama de tallo y hojas*. En este diagrama los datos se reproducen fielmente y consta de varias columnas, la primera, o tallo, representa en orden la mayor porción de los datos y las siguientes, separadas por una línea vertical, representan la menor porción de cada dato en particular. Para ver cómo se construye y cuál es su utilidad, se incluye el siguiente ejemplo.

Ejemplo 1.7

Enunciado:

Construya un diagrama de tallo y hojas con los datos de la tabla 1.10 del reporte de la Organización Mundial de la Salud, la cual presenta la fertilidad de una muestra de 109 países. La columna *fer*. (fertilidad) es el número promedio de hijos que una cohorte hipotética de mujeres tendría al final de su periodo reproductivo.

Tabla 1.10. Fertilidad mundial en el 2010. Hijos promedio por mujer

Ubicación	Fer.	Ubicación	Fer.	Ubicación	Fer.	Ubicación	Fer.	Ubicación	Fer.
Bosnia	1,2	Russia	1,4	Korea N	1,9	S. Africa	2,5	Samoa	4,0
Korea S	1,2	Spain	1,4	Lebanon	1,9	Venezuela	2,5	Guatemala	4,1
Andorra	1,3	Barbados	1,5	Norway	1,9	Ecuador	2,6	Iraq	4,1
Germany	1,3	Cuba	1,5	Sweden	1,9	Panama	2,6	Sudan	4,2
Japan	1,3	Switzerland	1,5	Costa Rica	2,0	Peru	2,6	Congo	4,4
Malta	1,3	Canada	1,6	Ireland	2,0	Dominican	2,7	Cameroon	4,6
Poland	1,3	Serbia	1,6	Maldives	2,0	India	2,7	Kenya	4,9
Romania	1,3	Trinidad	1,6	Mongolia	2,0	Nicaragua	2,7	Senegal	5,0
Singapore	1,3	Armenia	1,7	N. Zealand	2,0	Israel	2,8	Liberia	5,1

Ubicación	Fer.
Slovakia	1,3
Ukraine	1,3
Austria	1,4
Bulgaria	1,4
Croatia	1,4
Czech	1,4
Greece	1,4
Hungary	1,4
Italy	1,4
Latvia	1,4
Lithuania	1,4
Macedonia	1,4
Portugal	1,4

Ubicación	Fer.
Estonia	1,7
Netherlands	1,7
Australia	1,8
Belgium	1,8
China	1,8
Denmark	1,8
Finland	1,8
Iran	1,8
Thailand	1,8
UK	1,8
Brazil	1,9
Chile	1,9
France	1,9

Ubicación	Fer.
Dominica	2,1
Turkey	2,1
Uruguay	2,1
USA	2,1
Argentina	2,2
Indonesia	2,2
Kuwait	2,2
Mexico	2,2
El Salvador	2,3
Sri Lanka	2,3
Colombia	2,4
Jamaica	2,4
Qatar	2,4

Ubicación	Fer.
Cambodia	2,9
Egypt	2,9
Nepal	2,9
Paraguay	3,1
Philippines	3,1
S. Arabia	3,1
Honduras	3,3
Syrian	3,3
Namibia	3,4
Zimbabwe	3,4
Bolivia	3,5
Haiti	3,5
Pakistan	4,0

Ubicación	Fer.
S. Leone	5,2
Yemen	5,2
Ethiopia	5,3
Nigeria	5,3
Angola	5,8
Zambia	5,8
Congo	6,0
Chad	6,2
Uganda	6,3
Somalia	6,4
Afghanistan	6,6
Niger	7,1

Fuente: elaboración propia.

Solución:

El tallo del diagrama (Figura 1.10) está constituido por los números acompañados de las comas decimales. Esta es la parte común de los números de la correspondiente fila. Por ejemplo, los 24 números en la primera fila van desde 1,2 hasta 1,4, y los 18 de la tercera fila, del 2,0 al 2,4. En la primera columna está el número de datos hasta la correspondiente fila contando de izquierda a derecha o de derecha izquierda salvo en la fila que contiene al dato en la posición central. En la siguiente sección se discutirá en más detalle este último hecho.

24	1,	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4
48	1,	5	5	5	6	6	6	7	7	7	8	8	8	8	8	8	8	8	9	9	9	9	9	9	9
(18)	2,	0	0	0	0	0	1	1	1	1	2	2	2	2	3	3	4	4	4						
43	2,	5	5	6	6	6	7	7	7	8	9	9	9												
31	3,	1	1	1	3	3	4	4																	
24	3,	5	5																						
22	4,	0	0	1	1	2	4																		
16	4,	6	9																						
14	5,	0	1	2	2	3	3																		
8	5,	8	8																						
6	6,	0	2	3	4																				
2	6,	6																							
1	7,	1																							

Figura 1.10. Diagrama de tallo y hojas del ejemplo 1.7. Fuente: elaboración propia.

Se puede ver que existe gran concentración de países donde el número promedio de hijos por mujer es menor a 3,5 (primeras cinco filas). Este diagrama de tallo y hojas es útil para determinar la posición exacta de cada una de las observaciones; así, el máximo de las tasas de fertilidad es 7,1, y el mínimo es 1,2, que se repite dos veces.■

Diagrama de dispersión y gráfico de puntos

Si no resulta sencillo agrupar los datos en categorías o intervalos, resultan muy útiles los dos siguientes tipos de gráficos: *diagrama de dispersión*, para una variable, y *gráfico de puntos*.

Ejemplo 1.8

Enunciado:

Elabore un diagrama de dispersión de los datos de fertilidad.

Solución:

Para construir el diagrama, primero se ubica una escala en el eje *X*, que corresponde al rango de los datos de fertilidad. Cada uno de los puntos de la figura 1.11 representa un país, y su ubicación refleja el número de hijos promedio por mujer. La altura de los puntos se decide al azar para tener una mejor visualización de los datos. Se puede ver una gran concentración de países donde las mujeres tienen en promedio entre 1,2 y 3,4 hijos. También se destaca Niger con un número promedio de hijos por mujer muy alto.■

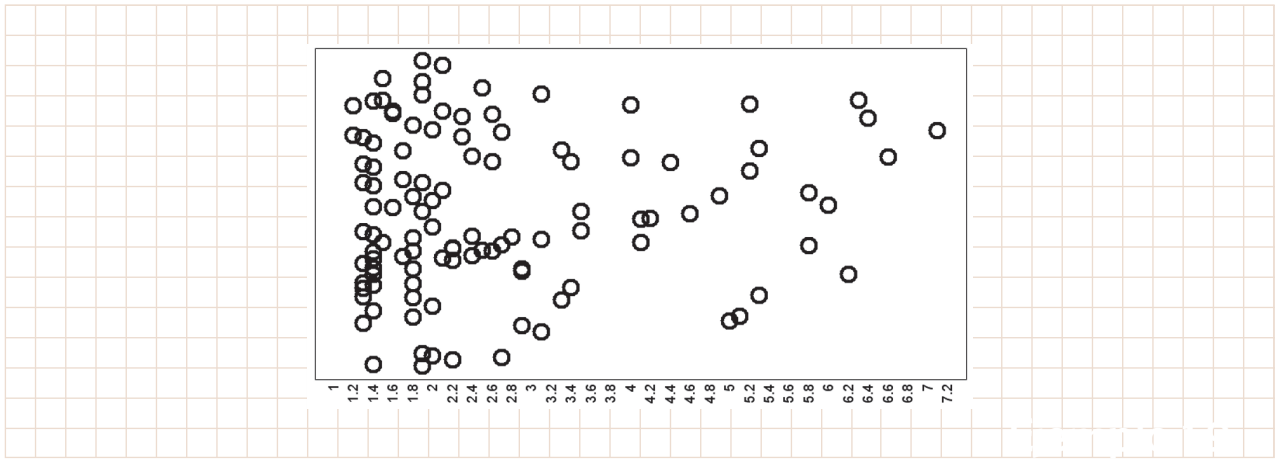


Figura 1.11. Diagrama de dispersión de la fertilidad mundial. Fuente: elaboración propia.

Ejemplo 1.9

Enunciado:

Construya una gráfica de puntos con los datos de fertilidad del ejemplo 1.7.

Solución:

En la figura 1.12 cada dato se representa con un punto ubicado sobre el eje *X*. Valores coincidentes se ubican uno sobre otro. En este caso, por comodidad, la primera pila de puntos corresponde a valores 1,2 o 1,3, la segunda a valores entre 1,4 y 1,5, y así sucesivamente.

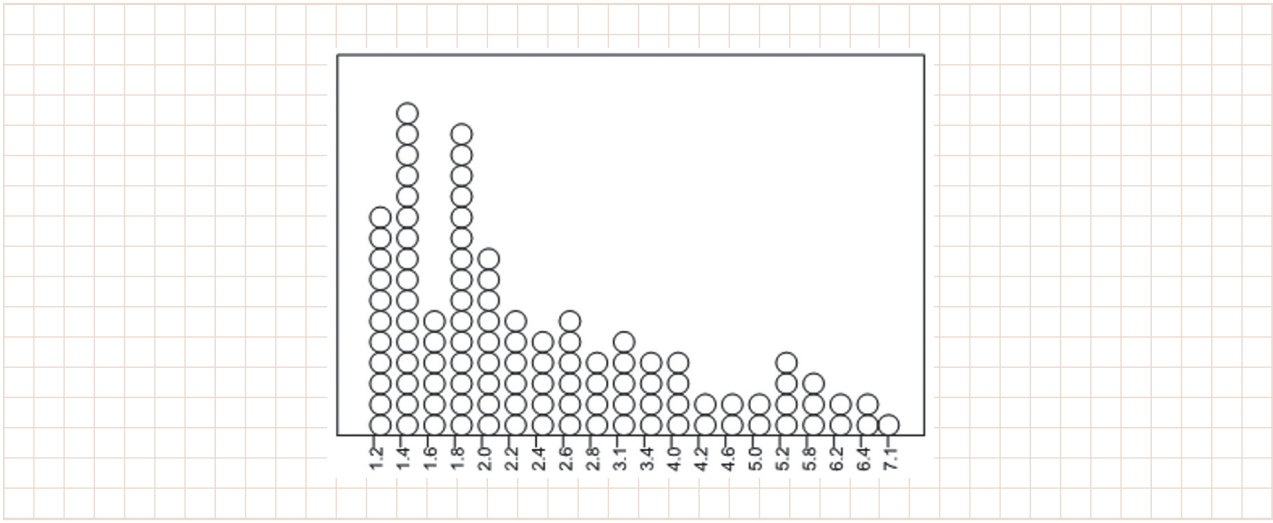


Figura 1.12. Gráfica de puntos de la fertilidad mundial. Fuente: elaboración propia.

Se vuelve a notar la concentración de países con un número promedio de hijos por mujer en los valores más bajos. Pero se resalta que hay un número considerable de ellos en donde las mujeres tienen en promedio entre 1,4 y 1,5 hijos.■

Diagrama de líneas

Un diagrama de líneas es especialmente util para mostrar la evolución de una variable en términos de otras, en especial, del tiempo.

Ejemplo 1.10

Enunciado:

En la tabla 1.11 se muestra un resumen de la base de datos de cédulas de la Registraduría Nacional del Estado Civil recopilados por Meisel y Vega (2004).

Tabla 1.11. Estatura en Colombia 1910 a 1984

Año de nacimiento	Estatura promedio hombres	Estatura promedio mujeres	Número de hombres	Número de mujeres	Total
1910-1914	163,48	150,78	1.751	2.197	3.948
1915-1919	163,61	151,49	4.582	4.993	9.575
1920-1924	164,16	152,38	9.086	9.779	18.865
1925-1929	164,7	153,06	15.659	16.299	31.958
1930-1934	165,17	153,48	22.219	24.619	46.838
1935-1939	165,76	154,21	34.637	33.820	68.457
1940-1944	166,26	154,69	40.186	45.623	85.809
1945-1949	167,1	155,59	53.164	60.723	113.887
1950-1954	167,84	156,4	73.835	80.863	154.698
1955-1959	168,07	156,81	101.613	111.278	212.891
1960-1964	168,47	157,17	142.054	157.950	300.004
1965-1969	169	157,34	177.464	221.024	398.488
1970-1974	168,91	157,21	298.908	426.660	725.568
1975-1979	169,66	157,81	1.639.499	1.646.878	3.286.377
1980-1984	170,64	158,65	1.528.875	1.468.110	2.996.985
Total	**169,664**	**157,778**	**4.143.532**	**4.310.816**	**8.454.348**

Fuente: Meisel y Vega (2004).

Construya y analice una gráfica para resumir la información de las estaturas de los hombres.

Solución:

En este ejemplo resulta importante resaltar la tendencia de la estatura promedio en el tiempo. En general, cuando este es el objetivo, se elige un *gráfico de líneas*, como el que se muestra enseguida (figura 1.13).

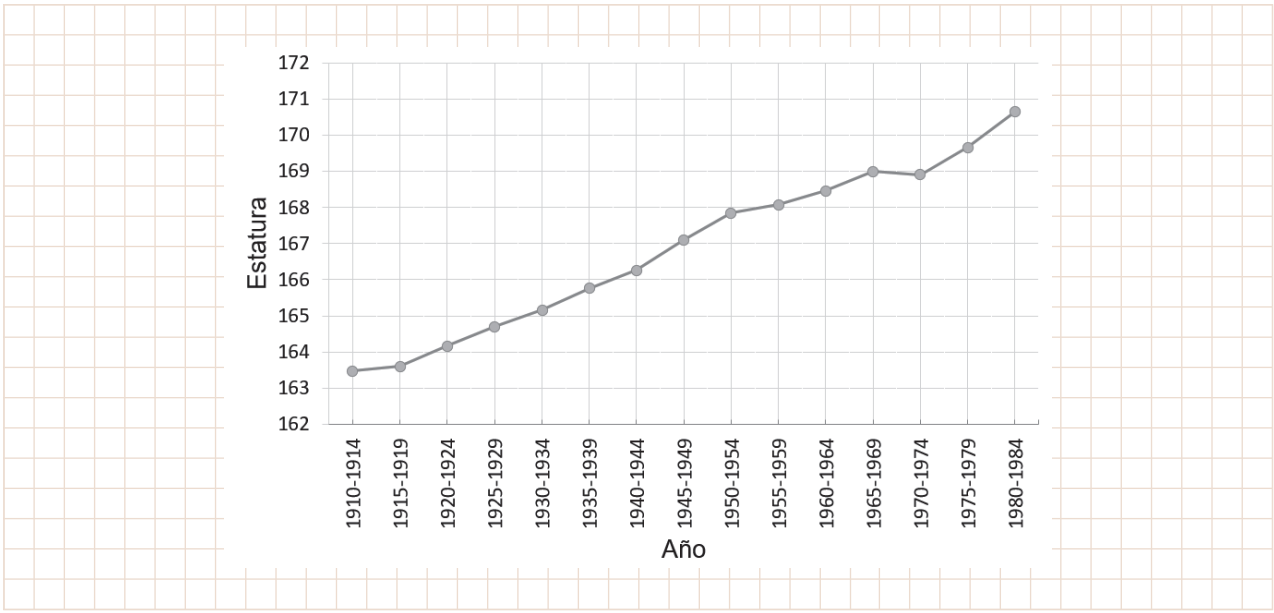


Figura 1.13. Evolución de la estatura promedio de los hombres en Colombia. Fuente: Meisel y Vega (2004).

En la figura 1.13 se puede ver que existe una tendencia de aumento de la estatura a lo largo del tiempo. Algunos autores utilizan la variable *estatura adulta* como medida indirecta de la calidad de vida, porque esta refleja la situación nutricional durante los años en los que la persona está en crecimiento. Si esto es así, es posible concluir que se evidencia un aumento en la calidad de vida de este grupo de hombres en el periodo considerado.■

Histogramas

Cuando una variable es cuantitativa de tipo continuo, la representación gráfica del conteo de frecuencias resulta inadecuada, debido a la gran cantidad o infinidad de valores que puede tomar la variable. En este caso es útil dividir en intervalos los valores de la variable, *intervalos de clase*, para luego graficar las frecuencias observadas en cada uno de ellos. A la anterior representación se le denomina *histograma*. Es aconsejable que el número de clases no sea muy pequeño, a fin de no perder información, y que no sea demasiado grande, para poder mostrar un patrón distinguible en la distribución de los datos. El número adecuado de clases para la construcción de un histograma es una elección subjetiva y resulta conveniente probar con diferentes números

de clases para determinar cuál de los histogramas descubre algún patrón en los datos. No obstante, una propuesta para especificar el número de clases, sugerida por Herbert Sturges, es incluida en los siguientes pasos para la construcción de un histograma.

Histograma: representación gráfica de una distribución de frecuencias por medio de rectángulos, cuyas anchuras representan intervalos de la clasificación y cuyas alturas representan las correspondientes frecuencias.

La forma como se reparten los porcentajes a través del histograma se suele llamar informalmente la **distribución de la variable o distribución de frecuencias**.

Pasos para la construcción de un histograma:

1. Determine el número de clases de la distribución de frecuencias. La sugerencia de Sturges es:

Tabla 1.12. Número de clases en un histograma

Número de datos	Número de clases
De 1 a 15	Muy pocos datos
De 16 a 31	5
De 32 a 63	6
De 64 a 127	7
⋮	⋮
n	$1 + 3{,}322 log_{10} n$

Fuente: elaboración propia.

2. Calcule el rango de los datos, que corresponde a la diferencia entre el dato mayor y el menor:

$$Rango = Máximo - Mínimo$$

3. Calcule la amplitud de las clases y redondear por exceso a la misma unidad de los datos:

$$Amplitud = \frac{Rango}{Número\ de\ clases}$$

La razón de redondear por exceso es que con esto se garantiza cubrir por completo el rango de los datos. Si la amplitud llega a tener la misma precisión de los datos, esta se debe aumentar en una unidad completa.

4. Determine el límite inferior de la primera clase. Para garantizar que ningún dato de los observados caiga en un límite de clase, estos contendrán un decimal más que ellos. Es así como la primera clase comenzará media unidad por debajo del valor mínimo.

Tabla 1.13. Unidades y medias unidades según la precisión de los datos

Precisión de los datos	Unidad	Unidad
Enteros	1	0,5
Décimas	0,1	0,05
Centésimas	0,01	0,005
Milésimas	0,001	0,0005
Diezmilésimas	0,0001	0,00005
⋮	⋮	⋮

Fuente: elaboración propia.

5. Determine los límites de cada uno de los intervalos de clases y realice el conteo de frecuencias de los datos en cada una de estas. Al promedio de los límites de cada clase se la denomina *punto medio de la clase* o *marca de clase*.

Para ilustrar estos pasos se incluye el siguiente ejemplo:

Este libro se terminó de editar en
la Editorial UTADEO en el mes de
diciembre de 2017

El libro *Análisis exploratorio de datos* es una introducción a la estadística descriptiva y probabilidad, con aplicaciones en áreas económico-administrativas. Su énfasis está puesto en la aplicación de conceptos estadísticos en contextos significativos de aprendizaje. El libro contiene más de 100 ejemplos resueltos y 300 ejercicios prácticos con información estadística aplicada. La lectura de este material revela paulatinamente una metodología que aborda conceptos estadísticos desde contextos cercanos a los estudiantes, ya que hace referencia a su realidad, su país o su profesión. El libro se enriquece con información proveniente de reconocidos organismos como la Organización Mundial de la Salud, el Departamento Administrativo Nacional de Estadística, la Oficina del Censo de Estados Unidos, entre otros, y promueve dos tipos de habilidades, la argumentación y la interpretación. Está dividido en seis capítulos: el primero, sobre estadística descriptiva, introduce el lenguaje de la estadística y muestra diferentes formas gráficas de resumir datos. El segundo capítulo explica las estadísticas de centro y dispersión, con énfasis en su interpretación. El tercer capítulo es una introducción conceptual a la regresión y la correlación. El capítulo cuatro aborda el concepto de *probabilidad* como fundamento para el estudio de las variables aleatorias y se cierra con dos ejemplos de distribuciones discretas y continuas: la distribución geométrica y la exponencial. Las distribuciones de probabilidad binomial, de Poisson e hipergeométrica se tratan en el capítulo cinco, y, por último, en el capítulo seis se estudia la distribución de probabilidad normal.

9 789587 252262